

Lecture Notes on Statistical Modelling

Qingyuan Zhao

December 4, 2024

Website for this course: <http://www.statslab.cam.ac.uk/~qz280/teaching/modelling-2021/>.

Copyright ©2024 Dr Qingyuan Zhao (qyzhao@statslab.cam.ac.uk)

This document should be used for educational purposes only. Please contact me if you find any mistakes or have any comments.

Contents

1	Scope and approach	1
2	Linear models	4
2.1	The normal linear model	4
2.2	Ordinary least squares and its geometry	6
2.2.1	Derivation of ordinary least squares	6
2.2.2	Orthogonal projections	7
2.2.3	Projection onto nested models	7
2.3	Exact inference for the normal linear model	9
2.3.1	Multivariate normal and related distributions	9
2.3.2	Distribution of $\hat{\beta}$ and $\hat{\sigma}^2$	11
2.3.3	Confidence sets	11
2.3.4	Hypothesis tests and analysis of variance	12
2.4	Linear conditional expectation model	13
2.4.1	Generalized least squares	13
2.4.2	Heteroscedasticity	14
2.4.3	Misspecified linear models	15
2.4.4	Omitted-variables bias and Simpson's paradox	16
2.5	Model diagnostics and model selection	17
2.5.1	Linear model diagnostics	18
2.5.2	The bias-variance decomposition	21
2.5.3	Quantitative criteria for model selection	22
2.5.4	Algorithms for model selection	25
2.5.5	*Regularization	26
2.5.6	*Inference after model-selection (not covered this year)	27
3	Exponential families	29
3.1	Definition and examples	29
3.1.1	Exponential tilting	29
3.1.2	Examples	30
3.2	Properties of exponential families	32

3.2.1	Cumulants	32
3.2.2	Mean value parametrization	33
3.2.3	IID sampling	34
3.2.4	Bayesian posterior distribution	35
3.2.5	*Empirical Bayes	35
3.3	Likelihood inference	36
3.3.1	Maximum likelihood estimator	36
3.3.2	Asymptotic inference	37
3.3.3	Hypothesis testing	38
3.3.4	Deviance	39
3.3.5	Deviance residual	40
3.4	Exponential dispersion families	41
3.4.1	Motivation and definition	41
3.4.2	Examples	42
4	Generalized linear models	44
4.1	From linear models to generalized linear models	44
4.1.1	Non-normal noise and the Box-Cox transformation	44
4.1.2	Three components of a generalized linear model	44
4.2	The canonical form	45
4.3	Linkage and over-dispersion	47
4.3.1	Estimation	47
4.3.2	Asymptotic normality and confidence intervals	48
4.3.3	Overdispersion due to clustering	49
4.4	Analysis of deviance	49
4.4.1	Nested models	50
4.4.2	The deviance additivity theorem	51
4.4.3	Analysis of deviance	51
4.5	Numerical computation	52
4.5.1	Newton-Raphson	52
4.5.2	Fisher scoring	53
4.5.3	Iteratively reweighted least squares	53
4.6	Model diagnostics and model selection	54
4.6.1	The general idea	55
4.6.2	Redefining residuals	55
4.6.3	Model selection	56
4.7	Binomial regression	56
4.7.1	Common link functions	57
4.7.2	Latent variable interpretation	57
4.7.3	Logistic regression and odds ratio	58
4.8	Poisson regression	58
4.8.1	Models for count data	58
4.8.2	*Variance stabilizing transform	59
4.8.3	Poisson regression	59
4.8.4	Multinomial models and the Poisson trick	61

4.9	Contingency tables	62
4.9.1	Two-way contingency tables	62
4.9.2	Three-way contingency tables	64
4.9.3	*Graphical models (not covered this year)	65
5	Review and look forward	67
5.1	Review	67
5.2	Look forward	70

Chapter 1

Scope and approach

This course requires a good understanding of the Part IB course *Statistics*. This course complements the Part II courses *Principles of Statistics* and *Mathematics of Machine Learning* by providing a more applied and computational perspective.

This year we will take a slightly different approach to statistical modelling. On the course website you will find the lecture notes from 2019, which take a more classical approach. Additionally, you might find the following books useful:

- A. Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley 2015. (Especially Chapters 2, 3, 4, 7.)
- P. McCullagh, J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989. (A classic but lacks details and examples.)
- G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning (with Applications in R)*. Springer 2013. (Provides perspectives from machine learning.)
- D. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009. (Provides perspectives from causal inference and scientific applications.)

For mathematics students, it might not be obvious that *statistics is not a branch of mathematics*.¹ There is no consensus on the definition of statistics (especially with the rise of machine learning and data science), but the following definition in Wikipedia cannot be too wrong:

- *Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.*

Compare this with the following definition of mathematical statistics:

- *Mathematical statistics is the study of statistics from a mathematical standpoint, using probability theory as well as other branches of mathematics such as linear algebra and analysis.*

Another way to think about the difference is mathematics is mostly about deductive reasoning from a set of axioms and assumptions, while statistics is mostly concerned with inductive reasoning from empirical data.² Through exploring different statistical models and learning R, a great programming language for statistical computing, you will be exposed to both the mathematical and non-mathematical elements of statistics.

To understand how statistics are used in practice, the following quote by G. Box³ may be illuminating:

Scientific research is usually an iterative process. The cycle: conjecture–design–experiment–analysis leads to a new cycle of conjecture–design–experiment–analysis and so on.... The experimental environment ... and techniques appropriate for design and analysis tend to change as the investigation proceeds.

At one point, the dominant view was that statistical modelling is a critical step of “analysis” and the model is built after data are collected. However, modern statisticians (and in fact, many pioneers like Box and Fisher) view statistical model as an essential component of the scientific process that guides all steps of the cycle and is being continuously updated.

Another important realization in modern statistics is that statistical models may come at different levels:

- (i) Models for conditional moments. For example, a *linear model for conditional expectation* assumes $\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] = \mathbf{x}^T \boldsymbol{\beta}$.
- (ii) Models for joint or conditional distributions. For example, the *classical normal linear model* assumes $Y = \mathbf{X}^T \boldsymbol{\beta} + \epsilon$ where the noise variable $\epsilon \perp \mathbf{X}$ and $\epsilon \sim N(0, \sigma^2)$.
- (iii) Structural or causal models that not only describe (associational) relationship for the data at hand but also (causal) relationship under counterfactual interventions. For example, the *linear structural equation model* assumes $Y^{(\mathbf{x})} = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$, where $Y^{(\mathbf{x})}$ is the counterfactual value of Y under the intervention that sets \mathbf{X} to \mathbf{x} and ϵ is an independent noise variable.

This course will not consider the third type of statistical model; see Freedman’s book for some good introduction to it.

This course will discuss the first two types of statistical models, which are often called *regression models*.⁴ In particular, our focus will be on a class of models called *generalized linear models* (GLM), which extends the classical linear model by using a beautiful theory for exponential family distributions. In essence, a GLM assumes that the conditional distribution of Y given \mathbf{X} is (almost) determined by the conditional expectation $\mathbb{E}[Y \mid \mathbf{X}]$.

Why do we care about regression problems? One obvious reason is their nearly ubiquitous presence in applications. Another reason is divide-and-conquer: the joint distribution of some random variables can always be factorized as a product of conditional distributions.

Why do we still care about (generalized) linear models, given the rise of machine learning algorithms that almost always have better prediction accuracy? Because GLMs

are simple, elegant, and interpretable. Moreover, more complex models are often constituted by GLMs. For example, a neural network is essentially the composition of numerous GLMs (with the distributional assumptions stripped away).

Notation

Upper-case letters indicate matrices or random variables. Lower-case letters indicate fixed quantities. We use \mathbf{I}_p to denote the $p \times p$ identity matrix, $\mathbf{1}_p$ to denote the p -vector of ones, and $\mathbf{0}_p$ the p -vector of zeros. Bolded symbols are vectors or matrices. Independent random variables (or vectors) X and Y are denoted as $X \perp\!\!\!\perp Y$. As a convention, we usually use subscript $i \in \{1, \dots, n\}$ to index observations and $j \in \{1, \dots, p\}$ to index variables. “Independent and identically distributed” is abbreviated as “i.i.d.”. The Euclidean norm of a vector \mathbf{Y} is denoted as $\|\mathbf{Y}\|$. Convergence in distribution (weak convergence) is denoted as \xrightarrow{d} .

Notes

¹Perhaps this is true for any non-statistician. When I told my neighbours that I am a statistician, most of their first reaction is that I do mathematics.

²Mathematics also involves induction, see G. Pólya’s book *Mathematics and Plausible Reasoning*, but mathematical induction is a deductive method. Statistics also involves deductive reasoning (which is basically mathematical statistics).

³Abstracts. (1957). *Biometrics*, 13(2), 238–246.

⁴The terminology “regression” was derived from a statistical phenomenon called “regression toward the mean” discovered by F. Galton. The original meaning of regression is no longer relevant today, but the terminology was kept for historical reasons.

Chapter 2

Linear models

Suppose our data are comprised of observations

$$(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n).$$

For any observation $i \in \{1, \dots, n\}$, $\mathbf{X}_i \in \mathbb{R}^p$ is a p -dimensional random vector whose entries are often called the regressors, covariates, predictors, explanatory variables, or independent variables. The random variable $Y_i \in \mathbb{R}$ is often called the response, outcome, target, or dependent variable. To emphasize that we are considering regression problems (as opposed to causal problems or purely predictive tasks), we will refer to \mathbf{X}_i as the regressor and Y_i as the response.¹

Unless otherwise stated, we assume $n > p$ throughout the course.

2.1 The normal linear model

The classical normal linear model assumes that $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)$ are independent and the conditional distribution of Y_i given \mathbf{X}_i satisfies²

$$Y_i = \sum_{j=1}^p X_{ij}\beta_j + \epsilon_i, \quad \epsilon_i \perp \mathbf{X}_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n. \quad (2.1)$$

Equation (2.1) is rather cumbersome and can be simplified by introducing the following vector/matrix notation:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_n^T \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Now we can rewrite (2.1) as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \mid \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p). \quad (2.2)$$

The matrix \mathbf{X} is known as the *design matrix* or *model matrix*. The former terminology was derived from the classical setting in experimental design in which \mathbf{X} is chosen by the

experimenter. This is rarely the case in modern applications. For this reason, we will refer to \mathbf{X} as the model matrix in this course.

Example 2.1 (Normal measurements). This model assumes $Y_i \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$, $i = 1, \dots, n$. The model matrix $\mathbf{X} = \mathbf{1}_n$ is a matrix with just one column and the regression coefficient $\beta = \mu$ is one-dimensional.

Example 2.2 (ANOVA). Let $F_i \in \{1, \dots, l\}$ be a *categorical variable* with l levels (also called a *factor*). The classical ANalysis Of VAriance (ANOVA) assumes $Y_i = \beta_{F_i} + \epsilon_i$, $\epsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ where β is a l -dimensional parameter vector. The i th row \mathbf{X}_i of the corresponding model matrix \mathbf{X} is an indicator vector whose F_i th entry is 1 and all other entries are 0.

To distinguish (2.1) and (2.2) with models for the conditional expectation, let $\mu_i = \mathbb{E}[Y_i | \mathbf{X}_i]$. Then (2.2) contains three different types of assumptions:

(i) The conditional expectation satisfies

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \mathbf{X}\boldsymbol{\beta};$$

(ii) The noise $\boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}$ satisfies $\boldsymbol{\epsilon} \perp \mathbf{X}$;

(iii) The noise $\boldsymbol{\epsilon}$ is distributed as $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

You may have noticed that (2.1) does not make any assumptions on the distribution of the regressors \mathbf{X} . This is intentional, because we are only concerned with the conditional distribution of Y given \mathbf{X} . In most applications, the distribution of \mathbf{X} is unknown. However, this does not matter in the classical linear model because the assumption $\boldsymbol{\epsilon} \perp \mathbf{X}$ allows us to factorize the likelihood function as

$$L(\boldsymbol{\beta}) = f(x_1, \dots, x_n, y_1, \dots, y_n; \boldsymbol{\beta}) = f(x_1, \dots, x_n) \cdot \prod_{i=1}^n f(y_i | \mathbf{x}_i; \boldsymbol{\beta}), \quad (2.3)$$

where f is a generic symbol for density functions and $f(y_i | \mathbf{x}_i; \boldsymbol{\beta})$ is the density function of a normal random variable:

$$f(y_i | \mathbf{x}_i; \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 / (2\sigma^2)}.$$

Because $f(\mathbf{x}_1, \dots, \mathbf{x}_n)$ does not depend on $\boldsymbol{\beta}$, whether the distribution of \mathbf{X} is known or not does not affect the inference for $\boldsymbol{\beta}$.³

2.2 Ordinary least squares and its geometry

2.2.1 Derivation of ordinary least squares

Following (2.3), the log-likelihood function is given by

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma^2) &= \log \prod_{i=1}^n f(Y_i | \mathbf{X}_i; \boldsymbol{\beta}) + \text{constant} \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \text{constant} \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \text{constant} \end{aligned}$$

Therefore, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is given by the solution of the *ordinary least squares* (OLS) problem

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2. \quad (2.4)$$

Notice that this holds regardless of whether σ^2 is known or not.

We may obtain a closed-form solution to (2.4) by using the following identities for matrix calculus:

$$\frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{a}^T \boldsymbol{\beta}) = \mathbf{a}, \text{ and } \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta}) = (\mathbf{A} + \mathbf{A}^T) \boldsymbol{\beta}.$$

Therefore the OLS estimator satisfies

$$\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}. \quad (2.5)$$

Equation (2.5) is called the *normal equations* because it requires the vector of *residuals* $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ to be orthogonal to \mathbf{X} .

The linear equations (2.5) have a unique solution if $\mathbf{X}^T \mathbf{X}$ is invertible (or equivalently, because $n > p$, \mathbf{X} has full rank). In this case, we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

Unless otherwise stated, we assume \mathbf{X} has rank p throughout this course.

The maximum likelihood estimator of σ^2 can be obtained by differentiating $l(\boldsymbol{\beta}, \sigma^2)$ with respect to σ^2 :

$$\frac{\partial}{\partial \sigma^2} l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

By solving $l(\hat{\boldsymbol{\beta}}, \sigma^2)$, we obtain

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n} \|\mathbf{R}\|^2.$$

The quantity $\|\mathbf{R}\|^2$ is often referred to as the *residual sum of squares* (RSS). Because $\hat{\sigma}_{\text{MLE}}^2$ is biased (see Section 2.3), it is more common to use the following unbiased estimator of σ^2 :

$$\hat{\sigma}^2 = \frac{n}{n-p} \hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n-p} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n-p} \|\mathbf{R}\|^2.$$

2.2.2 Orthogonal projections

Before discussing the statistical properties of the OLS estimator, it is useful to get a geometric understanding of what it does. By definition, the *fitted values* in the linear model are given by

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

which is a linear transformation of the original response vector \mathbf{Y} . Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, which is often called the *hat matrix* in statistics literature for the obvious reason. Geometrically, the least squares problem (2.4) implies that the vector of fitted values $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{Y}$ is the projection of the response vector \mathbf{Y} onto the column space of \mathbf{X} .

We briefly review some basic results about orthogonal projections. Two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are *orthogonal* if $\mathbf{u}^T\mathbf{v} = 0$. For a linear subspace \mathcal{W} of \mathbb{R}^n , its *orthogonal complement* is defined as $\mathcal{W}^\perp = \{\mathbf{v} \mid \mathbf{v}^T\mathbf{u} = 0 \text{ for all } \mathbf{u} \in \mathcal{W}\}$. Any vector $\mathbf{y} \in \mathbb{R}^n$ admits a unique decomposition $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$ where $\mathbf{y}_1 \in \mathcal{W}$ and $\mathbf{y}_2 \in \mathcal{W}^\perp$, and the Pythagoras theorem says $\|\mathbf{y}\|^2 = \|\mathbf{y}_1\|^2 + \|\mathbf{y}_2\|^2$. Moreover, $\dim(\mathcal{W}) + \dim(\mathcal{W}^\perp) = n$.

Let $\mathcal{C}(\mathbf{X}) = \{\mathbf{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}$ denote the column space of \mathbf{X} . Consider the decomposition

$$\mathbf{Y} = \underbrace{\mathbf{X}\hat{\boldsymbol{\beta}}}_{\hat{\boldsymbol{\mu}} \text{ (fitted values)}} + \underbrace{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}_{\mathbf{R} \text{ (residuals)}}$$

Note that $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} \in \mathcal{C}(\mathbf{X})$. Furthermore, the normal equations (2.5) can be written as $\mathbf{X}^T\mathbf{R} = 0$, so $\mathbf{R} \in \mathcal{C}(\mathbf{X})^\perp$. Again we see that $\hat{\boldsymbol{\mu}}$ is the projection of \mathbf{Y} onto $\mathcal{C}(\mathbf{X})$ and we have $\|\mathbf{Y}\|^2 = \|\hat{\boldsymbol{\mu}}\|^2 + \|\mathbf{R}\|^2$.

The hat matrix \mathbf{H} is a *projection matrix* onto $\mathcal{C}(\mathbf{X})$ that satisfies the following properties:

- (i) $\mathbf{H}\mathbf{u} = \mathbf{u}$ if $\mathbf{u} \in \mathcal{C}(\mathbf{X})$; $\mathbf{H}\mathbf{u} = 0$ if $\mathbf{u} \in \mathcal{C}(\mathbf{X})^\perp$.
- (ii) $\mathbf{I}_n - \mathbf{H}$ is the projection matrix onto $\mathcal{C}(\mathbf{X})^\perp$.
- (iii) \mathbf{H} is symmetric (i.e. $\mathbf{H}^T = \mathbf{H}$) and idempotent (i.e. $\mathbf{H}^2 = \mathbf{H}$).
- (iv) Orthonormal bases of $\mathcal{C}(\mathbf{X})$ and $\mathcal{C}(\mathbf{X})^\perp$ are eigenvectors of \mathbf{H} with eigenvalues 1 and 0, respectively.
- (v) $\text{tr}(\mathbf{H}) = \text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = p$.

We can define the projection matrix \mathbf{P} for an arbitrary subspace \mathcal{W} of \mathbb{R}^n by replacing $\mathcal{C}(\mathbf{X})$ with \mathcal{W} in property (i). Moreover, \mathbf{P} is a projection matrix for some subspace of \mathbb{R}^n if and only if property (iii) is satisfied. An immediate consequence of property (i) is that $\mathbf{H}\mathbf{X} = \mathbf{X}$.

2.2.3 Projection onto nested models

Consider a partition of the regressors:

$$\mathbf{X} = (\mathbf{X}_0 \ \mathbf{X}_1), \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},$$

where $\mathbf{X}_0 \in \mathbb{R}^{n \times p_0}$, $\mathbf{X}_1 \in \mathbb{R}^{n \times (p-p_0)}$, $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0 \times 1}$, and $\boldsymbol{\beta}_1 \in \mathbb{R}^{(p-p_0) \times 1}$. We are often interested in comparing the *full model* $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ with the *submodel* $\boldsymbol{\mu} = \mathbf{X}_0\boldsymbol{\beta}_0$ (possibly under additional independence and distributional assumptions, see Section 2.1).

Let \mathbf{P} denote the projection matrix onto $\mathcal{C}(\mathbf{X})$ (so $\mathbf{P} = \mathbf{H}$) and \mathbf{P}_0 denote the projection matrix onto $\mathcal{C}(\mathbf{X}_0)$. They satisfy two important properties:

- (i) $\mathbf{P}\mathbf{P}_0 = \mathbf{P}_0\mathbf{P} = \mathbf{P}_0$; see Figure 2.1.
- (ii) $\mathbf{P} - \mathbf{P}_0$ is also a projection matrix.

Exercise 2.3. Prove the second property. Which subspace does $\mathbf{P} - \mathbf{P}_0$ project onto?

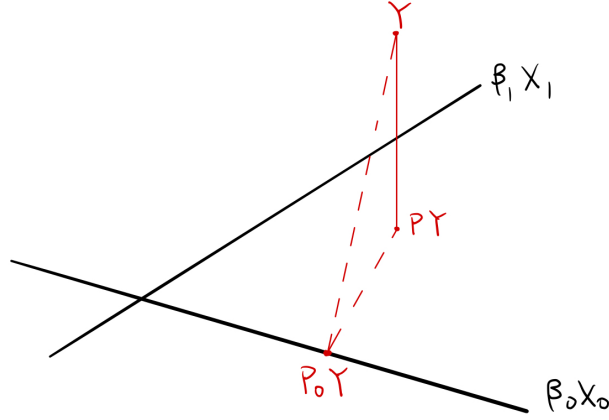


Figure 2.1 Nested model projections.

The first property implies the following identity. Let

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix}$$

be the partition of the OLS estimator $\hat{\boldsymbol{\beta}}$. Let $\tilde{\mathbf{X}}_1 = (\mathbf{I} - \mathbf{P}_0)\mathbf{X}_1$ and $\tilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{P}_0)\mathbf{Y}$ be the residuals of \mathbf{X}_1 and \mathbf{Y} after projecting onto $\mathcal{C}(\mathbf{X}_0)$. Then $\hat{\boldsymbol{\beta}}_1$ is equal to the OLS estimator for a linear regression of $\tilde{\mathbf{Y}}$ on $\tilde{\mathbf{X}}_1$:

$$\hat{\boldsymbol{\beta}}_1 = (\tilde{\mathbf{X}}_1^T \tilde{\mathbf{X}}_1)^{-1} \tilde{\mathbf{X}}_1^T \tilde{\mathbf{Y}}. \quad (2.6)$$

This is a generalization of the Gram-Schmidt process in linear algebra.⁴ To prove this, consider any $\mathbf{X}_2 \in \mathbb{R}^{n \times (n-p)}$ such that $(\mathbf{X}_0 \ \mathbf{X}_1 \ \mathbf{X}_2)$ is a full-rank $n \times n$ matrix. By applying Gram-Schmidt, we obtain matrices \mathbf{X}_0 , $\tilde{\mathbf{X}}_1 = (\mathbf{P} - \mathbf{P}_0)\mathbf{X}_1$, and $\tilde{\mathbf{X}}_2 = (\mathbf{I} - \mathbf{P})\mathbf{X}_2$ that are orthogonal to each other. In consequence, $\mathbf{P} - \mathbf{P}_0 = \mathbf{P}_{\tilde{\mathbf{X}}_1} = \tilde{\mathbf{X}}_1(\tilde{\mathbf{X}}_1^T \tilde{\mathbf{X}}_1)^{-1} \tilde{\mathbf{X}}_1^T$

is the projection matrix onto the column space of $\tilde{\mathbf{X}}_1$. Correspondingly, \mathbf{Y} can be decomposed as

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}_0\hat{\beta}_0 + \mathbf{X}_1\hat{\beta}_1 + \mathbf{R} \\ &= \underbrace{(\mathbf{X}_0\hat{\beta}_0 + \mathbf{P}_0\mathbf{X}_1\hat{\beta}_1)}_{\mathbf{P}_0\mathbf{Y}} + \underbrace{(\mathbf{I} - \mathbf{P}_0)\mathbf{X}_1\hat{\beta}_1}_{(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}} + \underbrace{\mathbf{R}}_{(\mathbf{I} - \mathbf{P})\mathbf{Y}}.\end{aligned}$$

Therefore,

$$\tilde{\mathbf{X}}_1\hat{\beta}_1 = (\mathbf{P} - \mathbf{P}_0)\mathbf{Y} = \mathbf{P}_{\tilde{\mathbf{X}}_1}\mathbf{Y}.$$

Because $\tilde{\mathbf{X}}_1$ has full rank, this shows (2.6).

An important special case is $p_0 = p - 1$, where \mathbf{X}_1 is a single regressor. In this case, some refer to $\hat{\beta}_1$ as the *partial regression coefficient* to distinguish from the *marginal regression coefficient* in a regression of \mathbf{Y} on just \mathbf{X}_1 .

Example 2.4 (Simple linear regression). When $p = 1$, the OLS estimator is given by the simple formula:

$$\hat{\beta} = \frac{\mathbf{X}^T\mathbf{Y}}{\mathbf{X}^T\mathbf{X}}.$$

When $p = 2$ and the model matrix is

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix},$$

the coefficient β_1 is called the *intercept* and β_2 is called the *slope*. By treating the first column of \mathbf{X} as \mathbf{X}_0 in the above partition, we obtain

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

where $\bar{X} = \sum_{i=1}^n X_i/n$ and $\bar{Y} = \sum_{i=1}^n Y_i/n$.

2.3 Exact inference for the normal linear model

Besides motivating the OLS problem (2.4) as finding the MLE of β in the normal linear model, the rest of (2.2) was entirely algebraic. In this section, we discuss statistical properties of the OLS estimator and how to use it to make exact inference under the normal linear model (2.1).

2.3.1 Multivariate normal and related distributions

We first review the properties of some common probability distributions.

A d -dimensional random vector \mathbf{Z} is said to follow the *multivariate normal* distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, written as $\mathbf{Z} \sim N_d(\mu, \Sigma)$, if its probability density function is given by

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-(\mathbf{z}-\mu)^T \Sigma^{-1}(\mathbf{z}-\mu)/2}.$$

The multivariate normal distribution has two important properties:

- (i) If $\mathbf{Z} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for any fixed matrix $\mathbf{A} \in \mathbb{R}^{k \times d}$ and vector $\mathbf{b} \in \mathbb{R}^k$, $\mathbf{AZ} + \mathbf{b} \sim N_k(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$.
- (ii) If \mathbf{Z}_1 and \mathbf{Z}_2 are two random vectors and $\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix}$ follows a multivariate normal distribution, then $\mathbf{Z}_1 \perp \mathbf{Z}_2$ if and only if $\text{Cov}(\mathbf{Z}_1, \mathbf{Z}_2) = \mathbf{0}$.

We often omit the subscript d in $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if the dimension is clear from the context.

Let $\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I})$. Then we say

$$\|\mathbf{Z}\|^2 = \sum_{i=1}^d Z_i^2 \sim \chi_d^2$$

follows the *chi-square distribution* with d degrees of freedom. The following result will be useful for us. Suppose $\mathbf{P} \in \mathbb{R}^{d \times d}$ be a projection matrix and $\text{rank}(\mathbf{P}) = r$, then $\|\mathbf{PZ}\|^2 \sim \chi_r^2$.

Exercise 2.5. Prove the last result.

Suppose $Z \sim N(0, 1)$, $S \sim \chi_d^2$, and $Z \perp S$. Then we say

$$\frac{Z}{\sqrt{S/d}} \sim t_d$$

follows the (*Student's*) *t-distribution*⁵ with d degrees of freedom.

Suppose $S_1 \sim \chi_{d_1}^2$, $S_2 \sim \chi_{d_2}^2$, and $S_1 \perp S_2$. Then we say

$$\frac{S_1/d_1}{S_2/d_2} \sim F_{d_1, d_2}$$

follows the *F-distribution* with degrees of freedom d_1 and d_2 .

Informally, the above definitions can be summarized as follows:

$$\begin{aligned} \chi_d^2 &= \underbrace{N(0, 1)^2 + \cdots + N(0, 1)^2}_{d \text{ times}}; \\ t_d &= \frac{N(0, 1)}{\sqrt{\chi_d^2/d}}; \\ F_{d_1, d_2} &= \frac{\chi_{d_1}^2/d_1}{\chi_{d_2}^2/d_2}. \end{aligned}$$

In this informal notation, two random variables (as indicated by their distributions) are independent whenever they appear in the same expression. It is obvious that $t_d^2 = F_{1, d}$. Moreover, $\mathbb{E}[\chi_d^2] = d$.

2.3.2 Distribution of $\hat{\beta}$ and $\hat{\sigma}^2$

Since $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ is a linear transformation of \mathbf{Y} , it has a multivariate normal distribution conditional on \mathbf{X} :

$$\begin{aligned}\hat{\beta} &\sim \mathcal{N}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\mu}, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I}_n \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}) \\ &= \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}).\end{aligned}$$

The estimator $\hat{\sigma}^2$ of the noise variance σ^2 can be written as

$$\hat{\sigma}^2 = \frac{1}{n-p} \|\mathbf{Y} - \mathbf{X} \hat{\beta}\|^2 = \frac{1}{n-p} \|(\mathbf{I}_n - \mathbf{H}) \mathbf{Y}\|^2.$$

Because $\mathbf{I}_n - \mathbf{H}$ is also a projection matrix, this implies that

$$\hat{\sigma}^2 \mid \mathbf{X} \sim \sigma^2 \chi_{n-p}^2 / (n-p).$$

This shows that $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$ and thus $\hat{\sigma}^2$ is unbiased.⁶

Exercise 2.6. Show $\hat{\beta}$ and $\hat{\sigma}^2$ are still unbiased without the normality assumption, that is, by only assuming ϵ given \mathbf{X} has mean $\mathbf{0}$ and covariance matrix $\sigma^2 \mathbf{I}_n$.

Finally, $\hat{\beta}$ and $\hat{\sigma}^2$ are independent under the normal linear model, because, given \mathbf{X} , $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ and $(\mathbf{I}_n - \mathbf{H}) \mathbf{Y}$ are jointly normal and

$$\text{Cov}((\mathbf{I}_n - \mathbf{H}) \mathbf{Y}, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}) = (\mathbf{I}_n - \mathbf{H}) \sigma^2 \mathbf{I}_n \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{0}.$$

2.3.3 Confidence sets

For the rest of this section, we view \mathbf{X} as fixed (in other words, the inference is conditional on \mathbf{X}). The key to exact inference is to find *pivotal quantities* whose distribution does not depend on unknown parameters. For example,

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2 \quad (2.7)$$

is pivotal, but

$$\hat{\beta} - \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}) \quad (2.8)$$

is not pivotal because the distribution depends on σ^2 . Instead, we can use the following pivotal quantity

$$\frac{\hat{\beta} - \boldsymbol{\beta}}{\hat{\sigma}} \sim \frac{\mathcal{N}(\mathbf{0}, (\mathbf{X}^T \mathbf{X})^{-1})}{\sqrt{\chi_{n-p}^2 / (n-p)}}.$$

Element-wise, we have

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}} \sim \frac{\mathcal{N}(0, (\mathbf{X}^T \mathbf{X})_{jj}^{-1})}{\sqrt{\chi_{n-p}^2 / (n-p)}} = \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}} \cdot t_{n-p}, \quad j = 1, \dots, p. \quad (2.9)$$

By using (2.9), we can immediately construct a $(1 - \alpha)$ -confidence interval for β_j :

$$\mathcal{CI}_j(\alpha) = \left[\hat{\beta}_j - \hat{\sigma} \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}} t_{n-p}(\alpha/2), \hat{\beta}_j + \hat{\sigma} \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}} t_{n-p}(\alpha/2) \right],$$

where $t_{n-p}(\alpha/2)$ is the upper $(\alpha/2)$ -quantile of t_{n-p} . By $(1 - \alpha)$ -confidence interval, we mean the following probabilistic statement is true:

$$\mathbb{P}(\beta_j \in \mathcal{CI}_j(\alpha)) = 1 - \alpha.$$

To construct a confidence region for the p -dimensional vector $\boldsymbol{\beta}$, a simple approach is to take the product of univariate confidence intervals $\prod_{j=1}^p \mathcal{CI}_j(\alpha/p)$. (*Exercise:* Show that this set covers $\boldsymbol{\beta}$ with probability at least $1 - \alpha$.)

However, this product set is usually quite conservative because it does not take into account the dependence between the entries of $\hat{\boldsymbol{\beta}}$. A better solution is to use the following pivotal quantity

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\mathbf{X}^T \mathbf{X}) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p \hat{\sigma}^2} \sim F_{p, n-p}. \quad (2.10)$$

So the following ellipsoid is a $(1 - \alpha)$ -confidence region of $\boldsymbol{\beta}$:

$$\mathcal{CI}(\alpha) = \left\{ \boldsymbol{\beta} \in \mathbb{R}^p \mid \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\mathbf{X}^T \mathbf{X}) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p \hat{\sigma}^2} \leq F_{p, n-p}(\alpha) \right\},$$

where $F_{p, n-p}(\alpha)$ is the upper α -quantile of $F_{p, n-p}$.

Exercise 2.7. Use (2.7) to construct a $(1 - \alpha)$ -confidence interval for σ^2 .

Exercise 2.8. Let $(\mathbf{X}^*, Y^*) \in \mathbb{R}^p \times \mathbb{R}$ be a new observation of the normal linear model. That is, suppose $Y^* = (\mathbf{X}^*)^T \boldsymbol{\beta} + \epsilon^*$ where $\epsilon^* \perp (\mathbf{X}, \boldsymbol{\epsilon}, \mathbf{X}^*)$ and $\epsilon^* \sim N(0, \sigma^2)$. Construct a $(1 - \alpha)$ -confidence interval for $(\mathbf{X}^*)^T \boldsymbol{\beta}$ and Y^* . The latter is called a $(1 - \alpha)$ -*prediction interval*.

2.3.4 Hypothesis tests and analysis of variance

By using the duality between hypothesis testing and confidence interval, we can easily construct level- α tests for

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0 \quad \text{and} \quad H_0 : \boldsymbol{\beta} = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\beta} \neq \mathbf{0}.$$

That is, we reject $\beta_j = 0$ if $0 \notin \mathcal{CI}_j(\alpha)$ and reject $\boldsymbol{\beta} = \mathbf{0}$ if $\mathbf{0} \notin \mathcal{CI}(\alpha)$.

More generally, we may be interested in comparing nested linear models. As before, consider the following partition

$$\mathbf{X} = (\mathbf{X}_0 \ \mathbf{X}_1) \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix},$$

where $\mathbf{X}_0 \in \mathbb{R}^{n \times p_0}$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$. We are interested in comparing the full model $\boldsymbol{\mu} = \mathbf{X} \boldsymbol{\beta}$ with the submodel $\boldsymbol{\mu} = \mathbf{X}_0 \boldsymbol{\beta}_0$, which amounts to testing $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ vs. $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$. The (*generalized*) *likelihood ratio statistic* is given by

$$\frac{\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta}, \sigma^2)}{\sup_{\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}, \boldsymbol{\beta}_1 = \mathbf{0}} L(\boldsymbol{\beta}, \sigma^2)} = \left\{ 1 + \frac{\|(\mathbf{P} - \mathbf{P}_0) \mathbf{Y}\|^2}{\|(\mathbf{I} - \mathbf{P}) \mathbf{Y}\|^2} \right\}^{n/2}$$

Exercise 2.9. Prove the above equality.

Thus, the likelihood ratio test rejects $H_0 : \beta_1 = \mathbf{0}$ if $\|(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}\|^2 / \|(\mathbf{I} - \mathbf{P})\mathbf{Y}\|^2$ is large. Note that $\|(\mathbf{I} - \mathbf{P})\mathbf{Y}\|^2$ is the residual sum of squares (RSS) of the full model, while $\|(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}\|^2$ is the reduction of RSS when we enlarge the submodel to the full model. This ratio has obvious geometric interpretations; see Figure 2.1.

To determine the critical value, we need to derive the distribution of the test statistic under $H_0 : \beta_1 = \mathbf{0}$. Under this null hypothesis, we have $\mathbf{Y} = \mathbf{X}\beta + \epsilon = \mathbf{X}_0\beta_0 + \epsilon$. Therefore,

$$\frac{\|(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}\|^2}{\|(\mathbf{I} - \mathbf{P})\mathbf{Y}\|^2} = \frac{\|(\mathbf{P} - \mathbf{P}_0)\epsilon\|^2}{\|(\mathbf{I} - \mathbf{P})\epsilon\|^2}$$

Because ϵ follows a multivariate normal distribution and

$$\text{Cov}((\mathbf{P} - \mathbf{P}_0)\epsilon, (\mathbf{I} - \mathbf{P})\epsilon) = (\mathbf{P} - \mathbf{P}_0)\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{P}) = \mathbf{0},$$

we have $(\mathbf{P} - \mathbf{P}_0)\epsilon \perp (\mathbf{I} - \mathbf{P})\epsilon$. Because $\mathbf{P} - \mathbf{P}_0$ and $\mathbf{I} - \mathbf{P}$ are projection matrices, $\|(\mathbf{P} - \mathbf{P}_0)\epsilon\|^2 \sim \chi_{p-p_0}^2$ and $\|(\mathbf{I} - \mathbf{P})\epsilon\|^2 \sim \chi_{n-p}^2$. Therefore,

$$F = \frac{\|(\mathbf{P} - \mathbf{P}_0)\mathbf{Y}\|^2 / (p - p_0)}{\|(\mathbf{I} - \mathbf{P})\mathbf{Y}\|^2 / (n - p)} \sim F_{p-p_0, n-p} \text{ under } H_0.$$

The level- α likelihood ratio test rejects $H_0 : \beta_1 = \mathbf{0}$ when $F > F_{p-p_0, n-p}(\alpha)$.

Exercise 2.10. Show that the t -test and F -test for $H_0 : \beta_j = 0$ vs. $H_0 : \beta_j \neq 0$ are equivalent.

2.4 Linear conditional expectation model

As discussed in Section 2.1, the normal linear model (2.2) contains three assumptions: the conditional expectation follows a linear model $\mu = \mathbf{X}\beta$, the noise ϵ is independent of \mathbf{X} , and the noise $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. The last two distributional assumptions play an essential role in the exact statistical inference discussed in Section 2.3 but are often too restrictive in applications. Next, we briefly discuss relaxations of these assumptions.

2.4.1 Generalized least squares

One possible relaxation is to assume ϵ follows a non-isotropic normal distribution:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon | \mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{\Sigma}),$$

where $\sigma^2 \in \mathbb{R}$ is unknown and $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ is a known positive-definite matrix that may depend on \mathbf{X} .

Theoretical results in Sections 2.2 and 2.3 can be easily extended to this model by using the transformation $\mathbf{X} \rightarrow \mathbf{\Sigma}^{-1/2}\mathbf{X}$ and $\mathbf{Y} \rightarrow \mathbf{\Sigma}^{-1/2}\mathbf{Y}$. The maximum likelihood estimator of β in this model is given by the *generalized least squares* (GLS) estimator:

$$\hat{\beta}_{\text{GLS}} = (\mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Sigma}^{-1} \mathbf{Y}.$$

Exercise 2.11. Derive the above formula for $\hat{\beta}_{\text{GLS}}$ from the definitions. Then derive it again using the formula for $\hat{\beta}_{\text{OLS}}$.

An important special case of GLS is the *weighted least squares* (WLS). Given a vector of weights $\mathbf{w} = (w_1, \dots, w_n)$, the WLS estimator is given by

$$\hat{\beta}_{\text{WLS}} = \arg \min_{\beta} \sum_{i=1}^n w_i (Y_i - \mathbf{X}_i^T \beta)^2.$$

This is equivalent to choosing $\Sigma = \text{diag}(w_1^{-1}, \dots, w_n^{-1})$ in GLS.

2.4.2 Heteroscedasticity

Consider the following less restrictive linear model:

$$Y_i = \mathbf{X}_i^T \beta + \epsilon_i, \quad i = 1, \dots, n,$$

where

- $(\epsilon_i, \mathbf{X}_i)$, $i = 1, \dots, n$, are *independent and identically distributed* (IID);
- $\mathbb{E}(\epsilon_i \mid \mathbf{X}_i) = 0$;
- $\text{Var}(\epsilon_i \mid \mathbf{X}_i) = \sigma^2(\mathbf{X}_i)$.

Compared to the classical normal linear model (2.1), it no longer assumes $\epsilon_i \perp \mathbf{X}_i$, the distribution of ϵ_i is normal, or the variance of ϵ_i is a constant.⁷ When $\sigma^2(\mathbf{X}_i) = \sigma^2$ is a constant, we say the noise is *homoscedastic*; otherwise, we say the noise is *heteroscedastic*.

Due to the lack of distributional assumptions, exact statistical inference is no longer possible. However, we can rely on asymptotic arguments:

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &= \sqrt{n} \{ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} - \beta \} \\ &= \sqrt{n} \{ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \beta + \epsilon) - \beta \} \\ &= \sqrt{n} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \\ &= \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{X}_i \epsilon_i \right). \end{aligned}$$

Under suitable regularity conditions, the first term converges in probability to $\Sigma_X = \mathbb{E}[\mathbf{X}_i \mathbf{X}_i^T]$ by the weak law of large numbers, and the second term converges in distribution to $N(\mathbf{0}, \Omega)$, where $\Omega = \mathbb{E}(\mathbf{X}_i \mathbf{X}_i^T \epsilon_i^2) = \mathbb{E}\{\sigma^2(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^T\}$. Therefore, by Slutsky's lemma,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, \Sigma_X^{-1} \Omega \Sigma_X^{-1}), \quad \text{as } n \rightarrow \infty. \quad (2.11)$$

The form of matrix $\Sigma_X^{-1} \Omega \Sigma_X^{-1}$ is common in misspecified maximum likelihood and is often called the *sandwich variance* (for obvious reasons) or the inverse *Godambe information*. When the noise is homoscedastic, i.e. $\sigma^2(\mathbf{X}_i) = \sigma^2$, this reduces to $\sqrt{n}(\hat{\beta} -$

$\beta) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \Sigma_X^{-1})$, which is consistent with the exact distribution (2.8) obtained under normality.

Equation (2.11) is not an (asymptotic) pivotal quantity yet because the distribution depends on Σ and Ω . These unknown quantities can be estimated by

$$\hat{\Sigma}_X = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \text{ and } \hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T e_i^2.$$

Under suitable regularity conditions, they converge to Σ_X and Ω in probability. By Slutsky's lemma,

$$\sqrt{n} \hat{\Sigma}_X \hat{\Omega}^{-1/2} (\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p), \text{ as } n \rightarrow \infty.$$

It is then straightforward to construct confidence intervals or hypothesis tests for β .

2.4.3 Misspecified linear models

One may further question the validity of the linear model $\mu = \mathbf{X}\beta$ itself. To emphasize that the linear model could be misspecified, we sometimes call $\mu = \mathbf{X}\beta$ a linear *working model*. Consider the following setting

$$Y_i = g(\mathbf{X}_i) + \epsilon_i, \text{ where } (\mathbf{X}_i, \epsilon_i) \text{ are IID, } \epsilon_i \perp \mathbf{X}_i, \mathbb{E}(\epsilon_i) = 0, i = 1, \dots, n. \quad (2.12)$$

In *nonparametric regression*, the goal is to estimate the regression function $g(\cdot)$. A parametric model such as the linear model assumes $g(\cdot)$ belongs to a class of function $\{g(\cdot; \beta) \mid \beta \in \mathbb{R}^p\}$ that is indexed by some finite-dimensional parameter β . Here, our interest is to understand how the linear working model behaves when the truth is (2.12).⁸

Recall that the OLS estimator $\hat{\beta}$ minimizes $\sum_{i=1}^n (Y_i - \mathbf{X}_i^T \beta)^2$. Therefore, it is expected that, as $n \rightarrow \infty$, $\hat{\beta}$ will converge to

$$\begin{aligned} \beta_{\text{OLS}} &= \arg \min_{\beta} \mathbb{E} \{ (Y_i - \mathbf{X}_i^T \beta)^2 \} \\ &= \arg \min_{\beta} \mathbb{E} \{ (g(\mathbf{X}_i) - \mathbf{X}_i^T \beta + \epsilon_i)^2 \} \\ &= \arg \min_{\beta} \mathbb{E} \{ (g(\mathbf{X}_i) - \mathbf{X}_i^T \beta)^2 \} + \underbrace{\mathbb{E} \{ (g(\mathbf{X}_i) - \mathbf{X}_i^T \beta) \epsilon_i \}}_{=0} + \underbrace{\mathbb{E}(\epsilon_i^2)}_{=\text{constant}} \\ &= \arg \min_{\beta} \mathbb{E} \{ (g(\mathbf{X}_i) - \mathbf{X}_i^T \beta)^2 \}. \end{aligned}$$

Therefore, $\mathbf{X}_i^T \beta_{\text{OLS}}$ may be viewed as the projection of $g(\mathbf{X}_i)$ onto the space of linear functions of \mathbf{X}_i .

We make two remarks on misspecified linear models. First, the “true” value of the parameter β_{OLS} depends on the distribution of \mathbf{X}_i ; see Figure 2.2 for an illustration. Second, the definition of the population regression coefficient β also generally depends on the estimator we use. For example, the *least absolute deviation* (LAD) estimator⁹

$$\hat{\beta}_{\text{LAD}} = \arg \min_{\beta} \sum_{i=1}^n |Y_i - \mathbf{X}_i^T \beta|$$

does not generally converge to β_{OLS} , unless the linear model is correctly specified (i.e. $g(\mathbf{x})$ is indeed linear in \mathbf{x}).

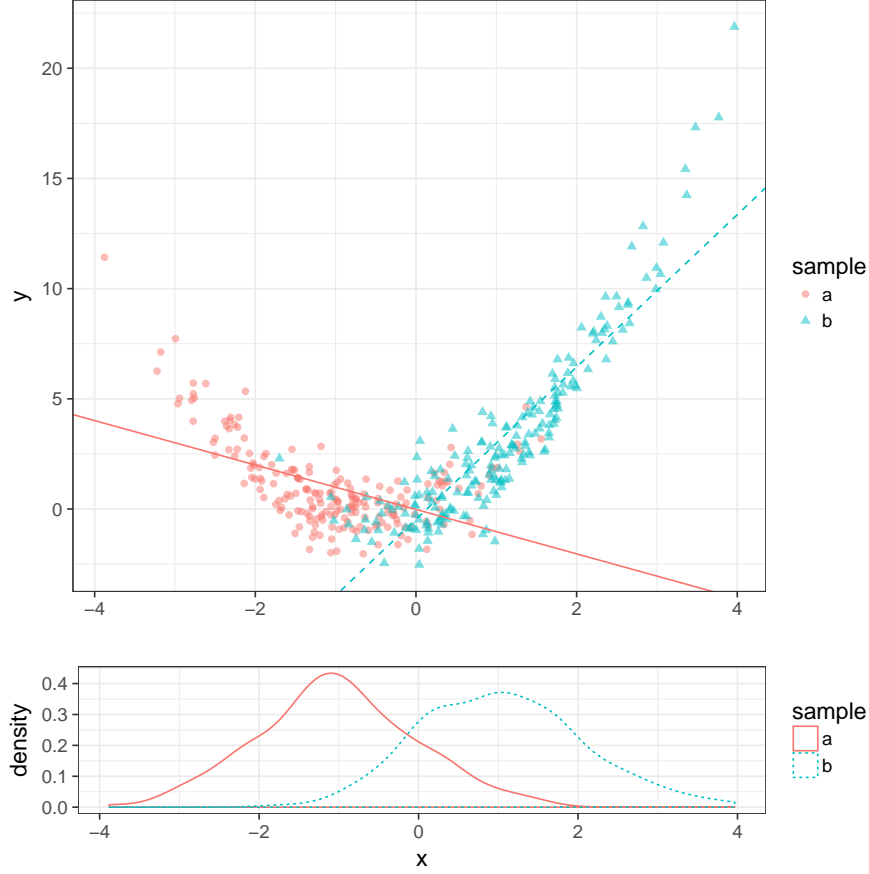


Figure 2.2 The “true” value of β_{OLS} depends on the distribution of the regressors. This figure shows two samples generated from the same conditional distribution of Y_i given X_i but different marginal distributions of X_i . In both samples, $Y_i = X_i^2 + X_i + \epsilon_i$ where $\epsilon_i \sim N(0, 1)$. In sample a , $X_i \sim N(-1, 1)$; in sample b , $X_i \sim N(1, 1)$. The value of β_{OLS} is negative in sample a but positive in sample b .

2.4.4 Omitted-variables bias and Simpson’s paradox

Misspecified models may also arise if some covariates are omitted in the regression. Consider two linear models:

$$\text{Model 1: } Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n;$$

$$\text{Model 1: } Y_i = \mathbf{X}_i^T \boldsymbol{\beta}^* + \mathbf{Z}_i^T \boldsymbol{\gamma}^* + \epsilon_i^*, \quad i = 1, \dots, n.$$

In general, $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$, a phenomenon often referred to as the *omitted-variable bias*.

Exercise 2.12. Show that $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ if $\mathbf{X}_i \perp \mathbf{Z}_i$.

In its extreme form, omitted-variable bias is known as Simpson’s paradox, which was initially discovered by K. Pearson and U. Yule. One of the best-known examples is the 1973 Berkeley admission data; see Table 2.1. Overall, men appear to be more likely to be admitted than women. However, if we look at the department-level statistics, in most cases women have a higher admission rate. This apparent paradox can be explained by the observation that there appear to be more men applications to departments with a higher admission rate. Whether this is also a kind of “gender bias” is another matter of debate.

Fundamentally, the reason behind Simpson’s paradox is that a regression coefficient only measures (conditional) association and does not necessarily indicate causation. A rigorous discourse on causation is beyond the scope of this course, but you might find the cartoon in Figure 2.3 illuminating (or not).

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
⋮	⋮	⋮	⋮	⋮
Total	8442	44%	4321	35%

Table 2.1 Berkeley admission data.¹⁰

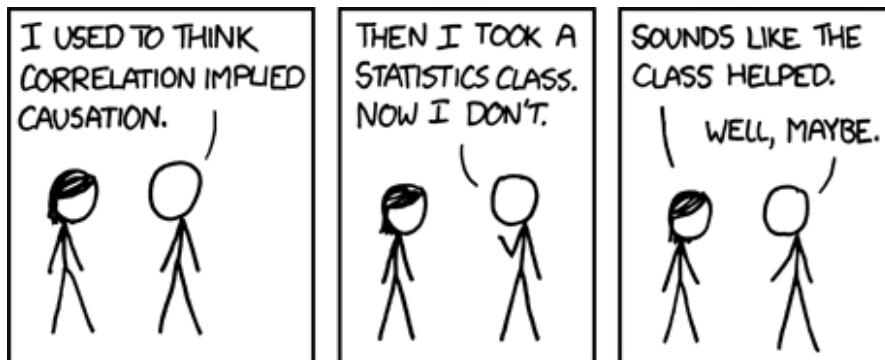


Figure 2.3 Correlation does not imply causation.

2.5 Model diagnostics and model selection

Although the normal linear model makes several restrictive assumptions, it remains the default choice for many applications used due to its simplicity. In practice, a common

task is to select a linear working model according to one or some of the following criteria:

- (i) Does the model appears to provide a good fit to the observed data?
- (ii) How large is the model's prediction error?
- (iii) How likely is the true model covered, assuming the data are indeed generated from it?
- (iv) How interpretable is the model?

This section will provide some theoretical insights for the first three considerations.

2.5.1 Linear model diagnostics

One nice thing about making restrictive assumptions is that we can often check them empirically. Here we provide some useful diagnostic quantities and plots for the normal linear model.

To measure how well the linear model fits the observed data, a widely used value is the *coefficient of determination*, defined as

$$R^2 = \frac{\|\hat{\boldsymbol{\mu}} - \bar{Y}\mathbf{1}\|^2}{\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2} = 1 - \frac{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2}{\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2},$$

where $\bar{Y} = \sum_{i=1}^n Y_i/n$. In words, R^2 is a measure of the proportion of variance of Y_i that can be explained by the linear model. A common mistake in practice is to interpret the absolute value of R^2 out of context, which depends crucially on the level of noise in the observations. So a linear model with $R^2 = 1\%$ is not necessarily a poor model.

The *leverage* of the i th observation is defined as H_{ii} , the i th diagonal element of the hat matrix. Recall that the fitted value for Y_i is

$$\hat{\mu}_i = (\mathbf{H}\mathbf{Y})_i = H_{ii}Y_i + \sum_{k \neq i} H_{ik}Y_k.$$

So the leverage H_{ii} is how much the observed value Y_i determines the fitted value $\hat{\mu}_i$. Another motivation for leverage is the following result (recall $\mathbf{R} = \mathbf{Y} - \hat{\boldsymbol{\mu}}$ is the vector of residuals)

$$\text{Var}(R_i \mid \mathbf{X}) = \sigma^2(1 - H_{ii}). \quad (2.13)$$

So the residual R_i is close to 0 if the leverage H_{ii} is close to 1.

Exercise 2.13. Prove (2.13).

Next we describe the diagnostic plots produced by the R function `plot.lm` by default.

The first is the *residual vs. fitted* plot, which plots the studentized residual \tilde{R}_i against the predicted value $\hat{\mu}_i$. We can visually assess the assumption $\mathbb{E}(\epsilon_i \mid \mathbf{X}_i) = 0$, by checking if there is any obvious trend (e.g. a quadratic trend) in the plot.

The second is the *quantile-quantile (Q-Q) plot*, which is used to visually check normality of the noise ϵ_i . Motivated by (2.13), the *studentized* or *standardized residual* of the i th observation is defined as

$$\tilde{R}_i = \frac{R_i}{\hat{\sigma}\sqrt{1-H_{ii}}}.$$

If the normal linear model is correct, \tilde{R}_i should be close to ϵ_i/σ , which follows a standard normal distribution. We may check this assumption by plotting the sample quantiles of $(\tilde{R}_1, \dots, \tilde{R}_n)$ against the theoretical quantiles of $N(0, 1)$; see Figure 2.4 for an illustration.

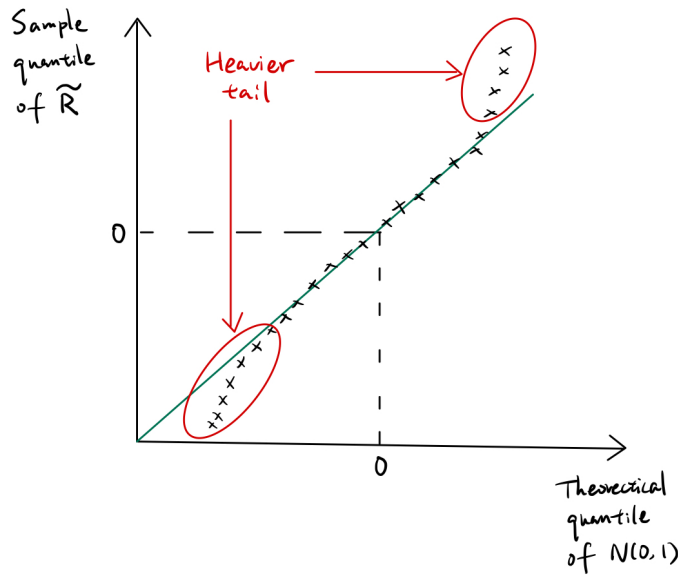


Figure 2.4 Quantile-quantile (Q-Q) plot.

Exercise 2.14. In the normal linear model, show that $\tilde{R}_i \sim t_{n-p-1}$ if we replace $\hat{\sigma}$ in the definition of \tilde{R}_i by $\hat{\sigma}_{(-i)}$, which is the estimator of σ using all observations besides (X_i, Y_i) .

The third diagnostic plot is the *scale-location plot*, which shows the square root of the absolute value of the standardized residual $\sqrt{|\tilde{R}_i|}$ against the fitted value $\hat{\mu}_i$. This plot is used to check the homoscedasticity assumption $\text{Var}(\epsilon_i | \mathbf{X}_i) = \sigma^2$, under which $\sqrt{|\tilde{R}_i|}$ should have an average value around 1.

The fourth and final one is a plot of *residuals vs. leverage*. More precisely, this plot shows \tilde{R}_i against H_{ii} and is used to identify outliers with a large leverage. We say an observation (\mathbf{X}_i, Y_i) is an *outlier* if $|R_i|$ is much larger than what is expected if $\epsilon_i \sim N(0, \sigma^2)$. In other words, these observations differ substantially from model-predicted values. Especially of concern are outliers with a high leverage, because just one or a few of them can severely bias a regression model. Note that the definition of “outlier” depends on the model. It is not rare to have one observation that is not an outlier originally

become an outlier when some other apparently outlying observations are removed. See Figure 2.5 for an illustration.

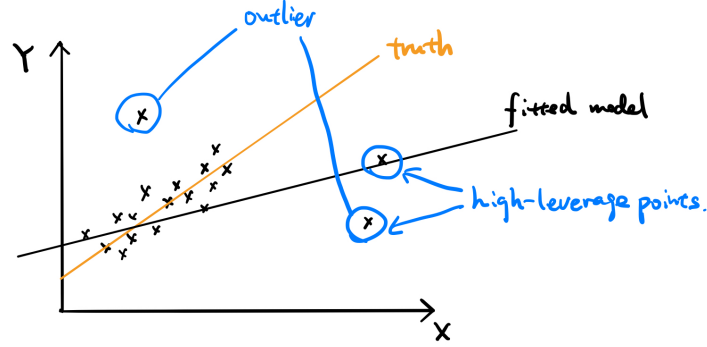


Figure 2.5 Outliers with a high leverage can severely bias a regression model.

A useful quantity for outlier detection is Cook’s distance:

$$D_i = \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})\|^2}{p\hat{\sigma}^2} = \frac{1}{p} \frac{H_{ii}}{1 - H_{ii}} \tilde{R}_i^2, \quad (2.14)$$

where $\hat{\boldsymbol{\beta}}_{(-i)}$ is the “leave-one-out” OLS estimator of $\boldsymbol{\beta}$ when (\mathbf{X}_i, Y_i) is removed from the dataset. By definition, D_i is a standardized change of the fitted values when the i th observation is removed. Therefore, a large value of D_i indicates that the i th observation have a large influence on the fitted values. Some clever algebra produces the formula in (2.14), so in order to compute Cook’s distance, it is unnecessary to repeatedly solve least squares problems.

Recall that in the normal linear model, a $(1 - \alpha)$ -confidence ellipsoid for $\boldsymbol{\beta}$ is given by

$$\mathcal{CI}(\alpha) = \left\{ \boldsymbol{\beta} \in \mathbb{R}^p \mid \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2}{p\hat{\sigma}^2} \leq F_{p, n-p}(\alpha) \right\},$$

Motivated by this, a rule of thumb in practice is that a Cook’s distance $D_i > F_{p, n-p}(0.5)$ indicates an outlier of concern.

Exercise 2.15. Prove (2.14) using the Sherman-Morrison formula

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}$$

that holds for any non-singular $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ such that $\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u} \neq -1$.

As a final remark on model diagnostics, the above quantities and plots should be regarded as visual “falsification tests” of the various assumptions made by the normal

linear model. This means that even when all the diagnostic plots look exactly like what are expected, we cannot conclude that the linear model must be correct. These tools depart from rigorous theorems in mathematical statistics but are immensely useful in practice. They provide empirical evidence to improve a statistical model and fit in nicely with Box's cycle of scientific research discussed in the very beginning of the course.

2.5.2 The bias-variance decomposition

Next, we consider the prediction error of a linear working model when it is possibly misspecified. Consider the nonparametric regression model (2.12) which is repeated below:

$$Y_i = g(\mathbf{X}_i) + \epsilon_i, \text{ where } (\mathbf{X}_i, \epsilon_i) \text{ are IID, } \epsilon_i \perp \mathbf{X}_i, \mathbb{E}(\epsilon_i) = 0, i = 1, \dots, n.$$

We further assume $\text{Var}(\epsilon_i) = \sigma^2$ exists. In Section 2.4.3, we saw that the OLS estimator $\hat{\beta}$ estimates β_{OLS} , the projection of $g(\mathbf{X}_i)$ onto the space of linear functions of \mathbf{X}_i in the population.

Let $\beta_n = \mathbb{E}(\hat{\beta} \mid \mathbf{X}_1, \dots, \mathbf{X}_n)$ be the expected value of $\hat{\beta}$. Notice that β_n depends on the model matrix \mathbf{X} , so β_n is a random quantity. This is why a subscript n is included. Let $(\mathbf{X}_{n+1}, \mathbf{Y}_{n+1})$ be a new independent observation from the same distribution.

The *mean squared prediction error* (MSPE) at a fixed value $\mathbf{x} \in \mathbb{R}^n$ of the regressors is defined as¹¹

$$\begin{aligned} \text{MSPE}(\mathbf{x}) &= \mathbb{E} \left[\{Y_{n+1} - \mathbf{X}_{n+1}^T \hat{\beta}\}^2 \mid \mathbf{X}_{n+1} = \mathbf{x} \right] \\ &= \mathbb{E} \left[\{g(\mathbf{x}) - \mathbf{x}^T \hat{\beta} + \epsilon_{n+1}\}^2 \right] \\ &= \mathbb{E} \left[\{g(\mathbf{x}) - \mathbf{x}^T \hat{\beta}\}^2 \right] + \underbrace{\mathbb{E} \left[\{g(\mathbf{x}) - \mathbf{x}^T \hat{\beta}\} \epsilon_{n+1} \right]}_{=0 \text{ because } \epsilon_{n+1} \perp \hat{\beta} \text{ and } \mathbb{E}[\epsilon_{n+1}] = 0} + \mathbb{E}(\epsilon_{n+1}^2) \\ &= \mathbb{E} \left[\{g(\mathbf{x}) - \mathbf{x}^T \beta_n + \mathbf{x}^T \beta_n - \mathbf{x}^T \hat{\beta}\}^2 \right] + \mathbb{E}(\epsilon_{n+1}^2) \\ &= \mathbb{E} \left[\{g(\mathbf{x}) - \mathbf{x}^T \beta_n\}^2 \right] + \underbrace{\mathbb{E} \left[\{g(\mathbf{x}) - \mathbf{x}^T \beta_n\} \{\mathbf{x}^T \beta_n - \mathbf{x}^T \hat{\beta}\} \right]}_{=0 \text{ because } \mathbb{E}[\hat{\beta} - \beta_n] = 0} \\ &\quad + \mathbb{E} \left[\{\mathbf{x}^T \beta_n - \mathbf{x}^T \hat{\beta}\}^2 \right] + \mathbb{E}(\epsilon_{n+1}^2) \\ &= \mathbb{E} \left[\{g(\mathbf{x}) - \mathbf{x}^T \beta_n\}^2 \right] + \mathbb{E} \left[\{\mathbf{x}^T \beta_n - \mathbf{x}^T \hat{\beta}\}^2 \right] + \mathbb{E}(\epsilon_{n+1}^2). \end{aligned}$$

To summarize, we have obtained the following *bias-variance* decomposition of MSPE:

$$\text{MSPE}(\mathbf{x}) = \underbrace{\mathbb{E} \left[\{g(\mathbf{x}) - \mathbf{x}^T \beta_n\}^2 \right]}_{\text{bias}^2} + \underbrace{\text{Var} \left(\mathbf{x}^T \hat{\beta} \right)}_{\text{variance}} + \underbrace{\sigma^2}_{\text{irreducible}}. \quad (2.15)$$

Equation (2.15) plays a central role in understanding the predictive behaviour of regression models, as its derivation does not rely on how $\hat{\beta}$ is obtained. For the OLS estimator $\hat{\beta}$, it can be shown that

$$\sum_{i=1}^n \text{Var}(\mathbf{X}_i^T \hat{\beta} \mid \mathbf{X}) = p\sigma^2. \quad (2.16)$$

Therefore, the average MSPE over the observed regressors is given by

$$\frac{1}{n} \sum_{i=1}^n \text{MSPE}(\mathbf{X}_i) = \frac{1}{n} \sum_{i=1}^n \{g(\mathbf{X}_i) - \mathbf{X}_i^T \boldsymbol{\beta}_n\}^2 + \frac{p\sigma^2}{n} + \sigma^2. \quad (2.17)$$

Exercise 2.16. Prove (2.16).

Equation (2.17) illustrates a fundamental phenomenon called the *bias-variance trade-off*. In order to make the bias term $\sum_{i=1}^n \{g(\mathbf{X}_i) - \mathbf{X}_i^T \boldsymbol{\beta}_n\}^2$ smaller, we can increase model complexity and include more regressors in the linear model. However, this comes at a price: the variance term $p\sigma^2/n$ will become larger. This trade-off of bias and variance applies to not only the least squares estimator but also many other statistical tasks;¹² see Figure 2.6 for a nice schematic illustration.

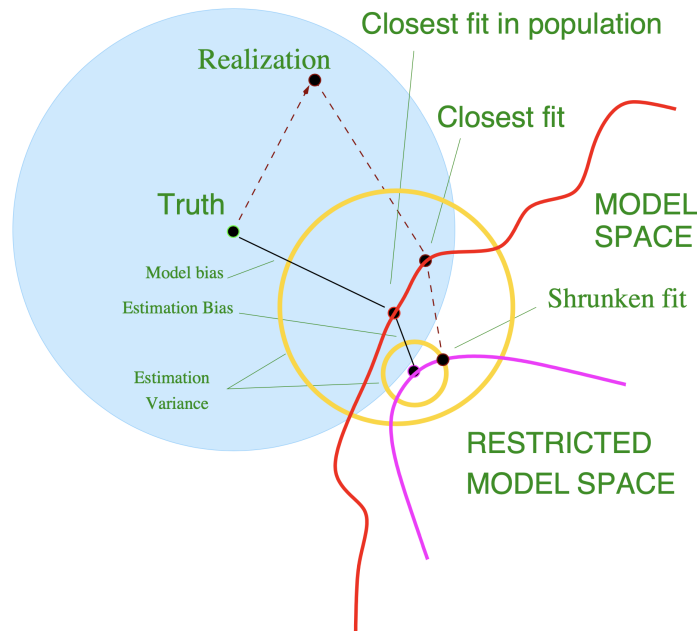


Figure 2.6 Schematic of the behavior of bias and variance.¹³.

In linear models, the complexity of the least squares estimator is measured by p , which coincides with the degrees of freedom. In more complex models, it is not always straightforward to come up with a good measure of model complexity.

2.5.3 Quantitative criteria for model selection

Next, we review some commonly used criteria for model selection. A better idea is to estimate the prediction error of the working model.

The first criterion is *Mallows'* C_p , which is an unbiased estimator of the average MSPE in (2.17) (up to a constant scaling). To derive C_p , we first compute the expected value of

the RSS under the nonparametric regression model (2.12):

$$\begin{aligned}
\mathbb{E}(\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2 \mid \mathbf{X}) &= \mathbb{E}\{\|(\mathbf{I} - \mathbf{H})\mathbf{Y}\|^2 \mid \mathbf{X}\} \\
&= \mathbb{E}\{\|(\mathbf{I} - \mathbf{H})(\boldsymbol{\mu} + \boldsymbol{\epsilon})\|^2 \mid \mathbf{X}\} \\
&= \|(\mathbf{I} - \mathbf{H})\boldsymbol{\mu}\|^2 + \mathbb{E}\{\|(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}\|^2 \mid \mathbf{X}\} \\
&= \|(\mathbf{I} - \mathbf{H})\boldsymbol{\mu}\|^2 + (n - p)\sigma^2.
\end{aligned}$$

Notice that for the OLS estimator $\hat{\boldsymbol{\beta}}$,

$$\mathbf{X}\boldsymbol{\beta}_n = \mathbf{X}\mathbb{E}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbb{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{H}\boldsymbol{\mu}.$$

Therefore, by comparing with (2.17), we see that

$$C_p = \|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2 + 2p\sigma^2 \quad (2.18)$$

is an unbiased estimator of $\sum_{i=1}^n \text{MSPE}(\mathbf{X}_i)$.

In practice, in order to use Mallows' C_p the noise variance σ^2 needs to be estimated. One common choice is to use the $\hat{\sigma}^2$ obtained from the full working model that uses all the regressors.

Heuristically, because the *training error rate* $\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2/n$ evaluates the predictive performance of the model using the same data that the model is fitted to, it underestimates the *prediction error rate* $\sum_{i=1}^n \text{MSPE}(\mathbf{X}_i)/n$. The gap between the training error and prediction error is sometimes referred to as the *optimism* of the training error rate. In the case of OLS, the amount of optimism is $2p\sigma^2/n$, which is proportional to the degrees of freedom p , a measure of model complexity. In general, the optimism tends to become larger when the working model becomes more complex.

Exercise 2.17. Consider any linear estimator $\hat{\boldsymbol{\mu}} = \mathbf{M}\mathbf{Y}$ of $\boldsymbol{\mu}$ where $\mathbf{M} \in \mathbb{R}^{n \times n}$ only depends on the data through \mathbf{X} . Show that $\text{tr}(\mathbf{M})$ is a “generalized degrees of freedom” in the sense that

$$C_M = \|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2 + 2\sigma^2\text{tr}(\mathbf{M})$$

is an unbiased estimator of $\sum_{i=1}^n \text{MSPE}(\mathbf{X}_i)$ for $\hat{\boldsymbol{\mu}}$.

Our second criteria is *leave-one-out cross-validation (LOO-CV)*, which is defined as

$$\text{LOO-CV} = \sum_{i=1}^n (Y_i - \hat{\mu}_{(-i)})^2, \quad \hat{\mu}_{(-i)} = \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_{(-i)},$$

where $\hat{\boldsymbol{\beta}}_{(-i)}$ is the leave-one-out OLS estimator that is computed using all observations besides (\mathbf{X}_i, Y_i) . The idea of leave-one-out was used in the definition of Cook's distance (2.14) previously. Likewise, it is not necessary to actually compute the LOO OLS estimators repeatedly. Indeed, it can be shown that

$$\hat{\mu}_i = H_{ii}Y_i + (1 - H_{ii})\hat{\mu}_{(-i)}.$$

Therefore, we have the following simple formula

$$\text{LOO-CV} = \sum_{i=1}^n \left(Y_i - \frac{\hat{\mu}_i - H_{ii}Y_i}{1 - H_{ii}} \right)^2 = \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{(1 - H_{ii})^2}.$$

For linear models, Mallows' C_p and LOO-CV often lead to very similar estimates of the prediction error. The advantage of cross validation is that it is often easier to use in more complex problems, although it may be no longer possible to obtain a closed-form formula.

Another two commonly used criteria for model selection are *Akaike's information criterion* (AIC) and the *Bayesian information criterion* (BIC). Because these information criteria are based on the likelihood function, they can be applied to a wide range of statistical problems. To illustrate this flexibility, we describe these criteria in more general setups.

Suppose $Y_i \stackrel{\text{IID}}{\sim} f(y), i = 1, \dots, n$, but a parametric model $Y_i \stackrel{\text{IID}}{\sim} f(y; \theta)$ is fitted instead over an Euclidean model space Θ . Under suitable regularity conditions, the MLE is expected to converge to

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log f(Y_i; \theta) \\ &= \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log \frac{f(Y_i; \theta)}{f(Y_i)} \\ &\xrightarrow{p} \arg \max_{\theta \in \Theta} \mathbb{E}_f \left\{ \log \frac{f(Y_i; \theta)}{f(Y_i)} \right\} \\ &= \arg \min_{\theta \in \Theta} \mathbb{E}_f \{ -\log f(Y_i; \theta) \}, \end{aligned}$$

where the subscript f under \mathbb{E} means that the expectation is computed over the density $f(\cdot)$. To select an appropriate model space Θ , AIC attempts to estimate $\mathbb{E}_f \left\{ -2 \log f(Y_{n+1}; \hat{\theta}) \mid \hat{\theta} \right\}$ where the expectation is taken over a new observation Y_{n+1} . It does so by making a correction to the log-likelihood of the observed data at the MLE $\hat{\theta}$:

$$\text{AIC} = -2 \sum_{i=1}^n \log f(Y_i; \hat{\theta}) + 2 \dim(\Theta). \quad (2.19)$$

The correction term $2 \dim(\Theta)$ penalizes the log-likelihood function evaluated at the same data used to fit the model. This closely resembles the idea of estimating the optimism of the training error in Mallows' C_p . It can be shown that AIC (divided by n) is a consistent estimator of its target as $n \rightarrow \infty$, but that proof is beyond the scope of this course.

To simplify the presentation, we assumed above that the data are IID. The same idea can be easily extended to regression problems by replacing $f(Y_i; \hat{\theta})$ with the conditional likelihood given X_i .

Exercise 2.18. Show that for the normal linear model with known α^2 , AIC coincides with Mallows' C_p .

Let $\{\Theta_1, \dots, \Theta_m\}$ be a collection of Euclidean model spaces. Then BIC is defined as

$$\text{BIC}(\Theta_k) = -2 \sum_{i=1}^n \log f(Y_i; \hat{\theta}_k) + \dim(\Theta_k) \log n,$$

where $\hat{\theta}_k$ is the MLE over Θ_k . BIC, as its name indicates is motivated by the Bayesian perspective on model selection. If we assign a uniform prior on the model spaces,

$$\mathbb{P}(\Theta_k) = \frac{1}{m}, \quad k = 1, \dots, m,$$

Then it can be shown that, as $n \rightarrow \infty$, the posterior probability for a model is approximately given by

$$\mathbb{P}(\Theta_k \mid \text{Data}) \propto e^{-\text{BIC}(\Theta_k)/2}.$$

This provides a much more principled way to select “true” regressors than naive selection rules based on statistical significance of regression coefficients.

Compared with AIC, BIC puts a larger penalty on model complexity and thus selects a smaller model. In practice, a rule of thumb is that AIC is more suitable for predictions and BIC is more suitable for selecting the “correct” model.¹⁴

2.5.4 Algorithms for model selection

Besides the statistical considerations discussed above, there are also computational challenges in model selection, as the number of submodels grows exponentially as the number of regressors increases. This section describes some algorithms that explore a large number of models more efficiently.

Our discussion thus far provides two useful insights for model selection. First, the RSS $\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\mu}}\|^2$ coincides with the (negative) log-likelihood function in the classical normal linear model and is indicative of predictive performance. Thus, it is reasonable to compare different models (especially those with the same complexity) by their RSS. Second, a good measure of model complexity is the degrees of freedom, i.e. the number of parameters in the model that are allowed to vary freely. These insights motivate the *best subset* algorithm, which selects the submodel with the smallest RSS for every degrees of freedom k .

However, the best subset algorithm is still a computationally intensive algorithm that requires us to compute the RSS for all 2^p submodel. Two greedy algorithms are commonly used to reduce the number of search paths. The first is the *forward stepwise* algorithm, which starts from the null model and greedily adds one unselected regressor at a time that reduce RSS the most. The second is the *backward stepwise* algorithm, which starts from the full model and greedily removes one unselected regressor at a time that increases the RSS the least. There is of course no guarantee that these greedy algorithms will select the absolute best submodel for each degrees of freedom k . But they often select a reasonably good submodel by examining only $O(p^2)$ submodels.

Example 2.19. Consider Figure 2.7, which shows the RSS for every submodel represented by a set of indices of regressors for $p = 3$. For $k = 0, 1, 2$, and 3,

- The best subset algorithm selects \emptyset , $\{3\}$, $\{1, 2\}$, and $\{1, 2, 3\}$;
- The forward stepwise algorithm selects \emptyset , $\{3\}$, $\{2, 3\}$, and $\{1, 2, 3\}$;
- The backward stepwise algorithm selects \emptyset , $\{2\}$, $\{1, 2\}$, and $\{1, 2, 3\}$.

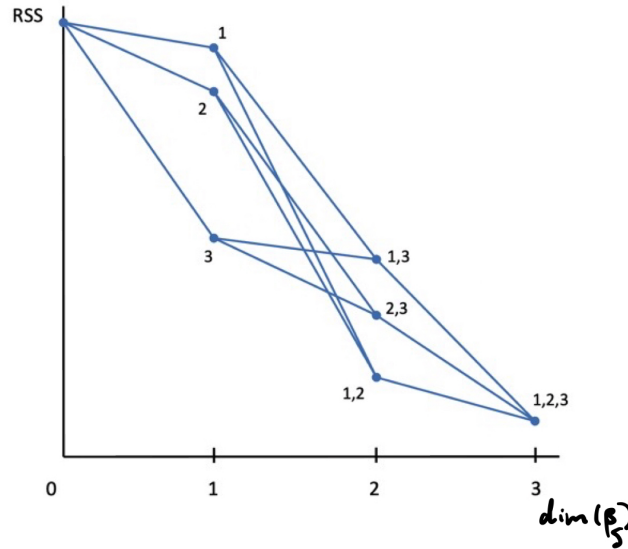


Figure 2.7 An illustration of model selection.

A common feature of these algorithms is that they produce a path of algorithm that is indexed by model complexity. Once such a path is obtained, we can then select a single model by using one of the quantitative criteria introduced above. We can also resort to the model diagnostics and select a model that passes the visual checks or add new regressors to the search space (e.g. add a quadratic term when the residual vs. fitted plot shows a quadratic trend). There is no need to feel too uncomfortable about the ad hoc nature of model selection. As G. Box summarized in a famous aphorism, “All models are wrong, but some are useful.”

2.5.5 *Regularization

Thus far, our discussion on model selection has been fairly “discrete”. A single subset of regressors is selected, and model complexity is measured by an integer (the number of selected regressors). It is possible and in fact often desirable to “smoothen” this process via an important idea called *regularization*. Briefly speaking, regularization tries to stabilize the fitted model (or in statistical terms, reduce the variance of the estimator) by penalizing model complexity.¹⁵

Our first example is the best subset algorithm, which can be rewritten as the solution to the following optimization problem

$$\begin{aligned} & \text{minimize} && \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ & \text{subject to} && \|\boldsymbol{\beta}\|_0 \leq k, \end{aligned}$$

where $\|\boldsymbol{\beta}\|_0 = |\{j \mid \beta_j \neq 0\}|$ is the number of non-zero entries in $\boldsymbol{\beta}$ (the ℓ_0 -“norm”). Because the minimal value of this problem is decreasing in k , the solution path for $k = 0, \dots, p$ can be reconstructed by the solution to the following unconstrained optimization problem

$$\text{minimize} \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_0,$$

where $\lambda\|\boldsymbol{\beta}\|_0$ is the regularizing penalty to the least squares objective $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$ and $\lambda \geq 0$ is a tuning parameter that controls the amount of regularization.

However, the ℓ_0 -“norm” $\|\boldsymbol{\beta}\|_0$ is a difficult penalty to work with computationally. The most widely used alternatives are the *ridge regression* (ℓ_2 -norm penalty) that solves

$$\text{minimize} \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{2.20}$$

and the *lasso* (ℓ_1 -norm penalty) that solves

$$\text{minimize} \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Exercise 2.20. Show that the ridge regression estimator is given by

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

using the following two methods:

- (i) Matrix calculus (see Section 2.2.1); and
- (ii) Transforming (2.20) to an ordinary least squares problem of the form in (2.4).

2.5.6 *Inference after model-selection (not covered this year)

After a model is selected (e.g. by any of the algorithms described in Section 2.5.4), a common pitfall is to pretend that the selected regressors are determined a priori and apply the standard inference procedures (e.g. those described in Section 2.3). This is problematic because the selected subset of regressors is not fixed; in fact, it depends on the realized \mathbf{Y} and incurs selection bias.

There are two common solutions to inference after model-selection:

- (i) One can split the sample and use some observations for model selection and the others for statistical inference;
- (ii) One can try to account for model selection, by excluding the information used by model selection from statistical inference.

The second solution is indeed an active research area in statistics.

Notes

¹The regressor \mathbf{X}_i may be a subset or a transformation of the covariates that are actually observed. For example, suppose Z_i is an observed covariate. Then we may let $\mathbf{X}_i = (1, Z_i, Z_i^2)$ in a linear model to capture trends that are quadratic in Z_i . See also footnote 8.

²Equation (2.1) does not necessarily describe the causal relationship between \mathbf{X}_i and Y_i . That is, if an external force sets \mathbf{X}_i to \mathbf{x}_i (instead of its “natural” value), (2.1) does not make any assumptions on what the resulting Y_i would become. In contrast, a linear structural equation model assumes that the counterfactual value of Y_i , often denoted as $Y_i^{(\mathbf{x}_i)}$, is $\mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$. What has confused generations of statisticians and scientists is that many people also use (2.1) to indicate a linear structural equation model. For more discussion on the distinction between regression and causation, see Section 6.4 of Freedman’s book.

³This is why some texts assume \mathbf{X} is “fixed”. A better way to think about this is that the inference for the normal linear model (2.1) is conditional on the model matrix \mathbf{X} . This is an instance of the *conditionality principle*, which says that the unconditional distribution and the conditional distribution given an ancillary statistic carry the same information for statistical inference.

⁴In econometrics, this result is known as the Frisch–Waugh–Lovell theorem.

⁵Named after the statistician W. S. Gosset who used the pseudonym “Student” to publish his method.

⁶If we are more rigorous, we should write the estimator of σ^2 as $\widehat{\sigma^2}$ instead of $\hat{\sigma}^2$. But it is widely understood that we estimate σ by first estimating σ^2 and $\hat{\sigma}^2$ does not mean “ $\hat{\sigma}$ square”. Notice that unbiasedness of $\hat{\sigma}^2$ does not translate to unbiasedness of $\sqrt{\hat{\sigma}^2}$ as an estimator of σ .

⁷This model makes an extra assumption on the distribution of \mathbf{X} , but this is mostly needed to simplify the presentation.

⁸The word “linear” in “linear model” refers to the modelling assumption that $g(\mathbf{x}; \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$ (instead of \mathbf{x}). So the regression model $Y_i = \beta_1 + \beta_2 Z_i + \beta_3 Z_i^2 + \epsilon_i$, although quadratic in the covariate Z_i , is still a linear model.

⁹The LAD estimator is a *robust regression* method that tries to limit the influence of outliers.

¹⁰Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. W W Norton, p.18.

¹¹Some other texts define MSPE as $\mathbb{E}[\{g(\mathbf{x}) - \mathbf{x}^T \hat{\boldsymbol{\beta}}\}^2]$, which we shall refer to as the *mean squared error*.

¹²One simple instance is *Stein’s paradox*, which is discussed in the *Principles of Statistics* course in detail.

¹³Taken from Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer, Figure 7.2.

¹⁴In a Bayesian setup, it is not necessary to select a single model. An alternative and perhaps more desirable approach is called *Bayesian model averaging*.

¹⁵The same idea is frequently used to solve ill-posed inverse problems in applied mathematics and engineering, often under the name *Tikhonov regularization*.

Chapter 3

Exponential families

This Chapter provides an introduction to the theory of exponential families, which greatly expand the classical statistical theory based upon normality. Exponential families are basic building blocks of the generalized linear models discussed in the next Chapters and more complex statistical models.

3.1 Definition and examples

3.1.1 Exponential tilting

Exponential families are obtained by exponentially “tilting” any density function. Suppose $f_0(\mathbf{y})$, $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^d$ is a density function with respect to a dominating measure $m(d\mathbf{y})$. By *exponential tilting*, we mean a collection of density functions given by

$$f(\mathbf{y}; \boldsymbol{\theta}) \propto e^{\boldsymbol{\theta}^T \mathbf{T}(\mathbf{y})} f_0(\mathbf{y}).$$

By normalizing the density functions, we obtain

$$f(\mathbf{y}; \boldsymbol{\theta}) = e^{\boldsymbol{\theta}^T \mathbf{T}(\mathbf{y}) - K(\boldsymbol{\theta})} f_0(\mathbf{y}), \quad (3.1)$$

where

$$K(\boldsymbol{\theta}) = \log \int_{\mathcal{Y}} e^{\boldsymbol{\theta}^T \mathbf{T}(\mathbf{y})} f_0(\mathbf{y}) m(d\mathbf{y}).$$

Some terminologies for the terms in (3.1):

- $\boldsymbol{\theta} \in \mathbb{R}^p$ is called the *natural parameter* or *canonical parameter*.
- $\mathbf{T}(\mathbf{y}) \in \mathbb{R}^p$ is called the *sufficient statistic*.
- $f_0(\mathbf{y})$ is called the *carrying density*.
- $K(\boldsymbol{\theta})$ is called the *cumulant function*.
- $\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^p \mid \int_{\mathcal{Y}} e^{\boldsymbol{\theta}^T \mathbf{T}(\mathbf{y})} f_0(\mathbf{y}) m(d\mathbf{y}) < \infty \right\}$ is called the *natural parameter space*.

- The exponential family $\{f(\mathbf{y}; \boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}$ is called *minimal* if $\mathbf{T}(\mathbf{y})$ is linearly independent.

Obviously, $f(\mathbf{y}; \mathbf{0}) = f_0(\mathbf{y})$, so $\mathbf{0} \in \Theta$. Furthermore, for any $\boldsymbol{\theta}_0 \in \Theta$, we may rewrite the density function as

$$f(\mathbf{y}; \boldsymbol{\theta}) = e^{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{T}(\mathbf{y}) - \{K(\boldsymbol{\theta}) - K(\boldsymbol{\theta}_0)\}} f(\mathbf{y}; \boldsymbol{\theta}_0),$$

So comparing to (3.1), we see that the same exponential family can be obtained by exponentially tilting any density function $f(\mathbf{y}; \boldsymbol{\theta}_0)$ within it.

Exercise 3.1. Show that Θ is a convex set and $K(\boldsymbol{\theta})$ is a convex function on Θ . [Hint: Use Hölder's inequality.]

3.1.2 Examples

The first motivation to study exponential families is that they contain many important probability distributions. Next we go through some examples.

Example 3.2 (Normal distribution). The density function of $N(\mu, 1)$ is given by

$$\begin{aligned} f(y; \mu) &= \frac{1}{\sqrt{2\pi}} e^{-(y-\mu)^2/2} \\ &= \exp \left(\underbrace{\mu}_{\boldsymbol{\theta}} \underbrace{y}_{\mathbf{T}(y)} - \underbrace{\mu^2/2}_{K(\boldsymbol{\theta})} \right) \underbrace{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}_{f(y;0)}. \end{aligned}$$

Example 3.3 (Poisson distribution). The Poisson distribution with rate λ can be obtained by exponentially tilting the probability mass function of $\text{Poisson}(1)$:

$$f_0(y) = e^{-1} \frac{1}{y!}, \quad y = 0, 1, \dots$$

We can first compute the cumulant function

$$K(\theta) = \log \sum_{y=0}^{\infty} e^{\theta y} e^{-1} \frac{1}{y!} = -1 + \log \sum_{y=0}^{\infty} \left(e^{\theta}\right)^y \frac{1}{y!} = e^{\theta} - 1$$

The exponentially tilted density is then given by

$$\begin{aligned} f(y; \theta) &= e^{\theta y - K(\theta)} f_0(y) \\ &= e^{\theta y - e^{\theta}} \frac{1}{y!} \\ &= \left(e^{\theta}\right)^y e^{-e^{\theta}} \frac{1}{y!} \\ &= \lambda^y e^{-\lambda} \frac{1}{y!}, \end{aligned}$$

where $\lambda = e^{\theta}$. Thus, the natural parameter θ is related to the mean parameter λ via $\theta = \log \lambda$ in the Poisson exponential family.

Example 3.4 (Binomial distribution). The probability mass function of a Binomial(n, π) with fixed n is given by

$$f(y; \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} = e^{y \log \frac{\pi}{1-\pi} + n \log(1-\pi)} \binom{n}{y}, \quad y = 0, 1, \dots, n.$$

So the natural parameter is the so-called *logit function* or *log odds*

$$\theta(\pi) = \log \frac{\pi}{1 - \pi}$$

that maps $(0, 1)$ to \mathbb{R} . By inverting this, we can obtain the usual parameter for binomial by the *expit function*

$$\pi(\theta) = \frac{e^\theta}{1 + e^\theta}.$$

The cumulant function is given by

$$K(\theta) = -n \log(1 - \pi) = n \log(1 + e^\theta).$$

More rigorously, we should further normalize the $\binom{n}{y}$ term as it does not add up to 1. But for all practical purposes, it is enough to obtain a cumulant function $K(\theta)$ up to a constant difference.

Example 3.5 (Multinomial distribution). The probability mass function of a Multinomial($n, \boldsymbol{\pi}$) for $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)$ with fixed n is given by

$$f(\mathbf{y}; \boldsymbol{\pi}) = \frac{n!}{y_1! \dots y_L!} \pi_1^{y_1} \dots \pi_L^{y_L} = \frac{n!}{y_1! \dots y_L!} e^{(\log \boldsymbol{\pi})^T \mathbf{y}}.$$

So it is tempting to treat $\log \boldsymbol{\pi} = (\log \pi_1, \dots, \log \pi_L)$ as the natural parameter. However, this is not minimal because of the constraints $\pi_1 + \dots + \pi_L = 1$ and $y_1 + \dots + y_L = n$.

To overcome this, one possibility is to set the last level as the baseline by rewriting the density as

$$\begin{aligned} f(y; \boldsymbol{\pi}) &= \frac{n!}{y_1! \dots y_L!} \exp \left\{ \sum_{l=1}^{L-1} y_l \log \pi_l + \left(n - \sum_{l=1}^{L-1} y_l \right) \log \left(1 - \sum_{l=1}^{L-1} \pi_l \right) \right\} \\ &= \frac{n!}{y_1! \dots y_L!} \exp \left\{ \sum_{l=1}^{L-1} y_l \log \left(\frac{\pi_l}{1 - \sum_{l=1}^{L-1} \pi_l} \right) + n \log \left(1 - \sum_{l=1}^{L-1} \pi_l \right) \right\}. \end{aligned}$$

So the $(L - 1)$ -dimensional minimal natural parameter is given by

$$\theta_l = \log \frac{\pi_l}{\pi_L} = \log \frac{\pi_l}{1 - \sum_{l=1}^{L-1} \pi_l}, \quad l = 1, \dots, L - 1.$$

Inverting this, we obtain the so-called *multinomial logit* or *softmax function*

$$\pi_l(\boldsymbol{\theta}) = \frac{e^{\theta_l}}{\sum_{l=1}^L e^{\theta_l}}$$

if we define $\theta_L = 0$. The cumulant function is given by

$$K(\boldsymbol{\theta}) = -\log \left(1 - \sum_{l=1}^{L-1} \pi_l \right) = \log \left(\sum_{l=1}^L e^{\theta_l} \right).$$

Exercise 3.6. Show the following distributions are exponential families and find their natural parameter, sufficient statistic, and cumulant function:

(i) The normal distribution

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, y \in \mathbb{R}.$$

(ii) The Gamma distribution

$$f(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, y > 0.$$

(iii) The negative binomial distribution with fixed k

$$f(y; \pi) = \binom{y+k-1}{y} \pi^k (1-\pi)^y, y = 0, 1, 2, \dots$$

3.2 Properties of exponential families

Next, we introduce some important properties about univariate exponential families. Many of the results below can be extended to multivariate exponential families, but such extension is beyond the scope of this course.

3.2.1 Cumulants

The moments of an exponential family distribution (3.1) can be easily computed by its cumulant function. Recall that for a random variable Y , its *moment generating function* is given by

$$M(t) = \mathbb{E}(e^{tY}),$$

and its *cumulant generating function* is defined as

$$K(t) = \log M(t).$$

Suppose $M(t)$ is infinitely differentiable at 0. Then we have the following Maclaurin expansions

$$\begin{aligned} M(t) &= \sum_{r=0}^{\infty} \mathbb{E}(Y^r) \frac{t^r}{r!}, \\ K(t) &= \sum_{r=0}^{\infty} \kappa_r \frac{t^r}{r!}, \end{aligned}$$

where $\mathbb{E}(Y^r) = M^{(r)}(0)$ and $\kappa_r = K^{(r)}(0)$. The values $\kappa_1, \kappa_2, \dots$ are called *cumulants* of the probability distribution and are closely related to the moments. In particular, the first two cumulants are the mean and variance.

Exercise 3.7. Verify that $\kappa_1 = \mathbb{E}(Y)$, $\kappa_2 = \text{Var}(Y)$, $\kappa_3 = \mathbb{E}(Y - \kappa_1)^3$, and $\kappa_4 = \mathbb{E}(Y - \kappa_1)^4 - 3\kappa_2^2$.

The exponential family $\{f(y; \theta) \mid \theta \in \Theta\}$ is called *regular* if Θ is an open set. Nearly all exponential families and certainly all the exponential families we will consider are regular. For a regular exponential family with a one-parameter natural parameter θ , the moment generating function is given by

$$\begin{aligned} M_\theta(t) &= \mathbb{E}_\theta(e^{tY}) \\ &= \int e^{ty} e^{\theta y - K(\theta)} f_0(y) m(dy) \\ &= e^{K(\theta+t) - K(\theta)} \int e^{(t+\theta)y - K(\theta+t)} f_0(y) m(dy) \\ &= e^{K(\theta+t) - K(\theta)}. \end{aligned}$$

So the cumulant generating function is given by

$$K_\theta(t) = \log M_\theta(t) = K(\theta + t) - K(\theta).$$

Therefore, the mean and variance are given by the first two derivatives of the cumulant function $K(\cdot)$ at θ :

$$\mu(\theta) = \mathbb{E}_\theta(Y) = \left. \frac{d}{dt} K_\theta(t) \right|_{t=0} = K'(\theta), \quad (3.2)$$

$$V(\theta) = \text{Var}_\theta(Y) = \left. \frac{d^2}{dt^2} K_\theta(t) \right|_{t=0} = K''(\theta). \quad (3.3)$$

This is why we often only need to determine $K(\theta)$ up to an additive constant.

We refer to $\mu(\theta)$ as the *mean function* and $V(\theta)$ the *variance function*. The above derivation shows that they are related through the following key identity

$$\mu'(\theta) = K''(\theta) = V(\theta) \geq 0. \quad (3.4)$$

This shows that, apart from pathological cases with zero variance, the mean function $\mu(\theta)$ is strictly increasing and the cumulant function $K(\theta)$ is strictly convex.

3.2.2 Mean value parametrization

Because $\mu(\theta)$ is strictly increasing in θ , we can also parameterize a univariate exponential family by its mean value. Suppose the inverse function of $\mu(\theta)$ is $\theta(\mu)$. By the inverse function theorem,

$$\theta'(\mu) = \frac{1}{V(\theta)}.$$

The exponential family can be alternatively written as

$$f(y; \mu) = e^{\theta(\mu)y - K(\theta(\mu))} f_0(y)$$

for $\mu \in \mathcal{M} = \{\mu(\theta) \mid \theta \in \Theta\}$. The set \mathcal{M} is usually referred to as the *mean space*. When using the mean-value parameterization, we often write the variance function as $V(\mu)$.

Example 3.8. Continuing from Examples 3.2 to 3.4, the natural parameter of $N(\mu, 1)$ is $\theta(\mu) = \mu$ and the cumulant function is $K(\theta) = \theta^2/2$. Therefore, the mean and variance functions are

$$\mu(\theta) = \theta, \quad V(\theta) = 1.$$

For $\text{Poisson}(\lambda)$, the natural parameter is $\theta = \log \lambda$ and $K(\theta) = e^\theta - 1$. Therefore, its mean and variance functions are given by

$$\mu(\theta) = V(\theta) = e^\theta = \lambda.$$

For $\text{Bernoulli}(\pi) = \text{Binomial}(1, \pi)$, the natural parameter is $\theta(\mu) = \log\{\pi/(1 - \pi)\}$ and the cumulant function is $K(\theta) = \log(1 + e^\theta)$. Therefore, its mean and variance functions are given by

$$\begin{aligned} \mu(\theta) &= \frac{e^\theta}{1 + e^\theta} = \frac{1}{1 + e^{-\theta}} = \pi, \\ V(\theta) &= \frac{e^\theta}{(1 + e^\theta)^2} = \pi(1 - \pi). \end{aligned}$$

Exercise 3.9. Derive the mean and variance of the negative binomial distribution.

3.2.3 IID sampling

Suppose $Y_1, \dots, Y_n \stackrel{\text{IID}}{\sim} f(y; \theta)$ where $f(y; \theta)$ is a one-parameter exponential family with sufficient statistic Y . Then their joint density is given by

$$\begin{aligned} f(y_1, \dots, y_n; \theta) &= \prod_{i=1}^n f(y_i; \theta) \\ &= \prod_{i=1}^n e^{\theta y_i - K(\theta)} f_0(y_i) \\ &= e^{n\{\theta \bar{y} - K(\theta)\}} \prod_{i=1}^n f_0(y_i). \end{aligned}$$

This is a new exponential family with

- Natural parameter $\theta^{(n)} = n\theta$;
- Sufficient statistic $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$;
- Cumulant function $K^{(n)}(\theta^{(n)}) = nK(\theta) = nK(\theta^{(n)}/n)$;

- Carrying density $\prod_{i=1}^n f_0(y_i)$.

This property allows us to easily extend results for a single random variable from exponential families to IID sampling.

Exercise 3.10. Use the cumulant function above to show that $\mu^{(n)} = \mu$ and $V^{(n)} = V/n$.

3.2.4 Bayesian posterior distribution

Because the exponential family density function (3.1) is symmetric in the natural parameter and sufficient statistic, the Bayesian posterior distribution of the natural parameter is also an exponential family. To see, suppose we observe Y from an univariate exponential family

$$f(y; \theta) = e^{\theta y - K(\theta)} f_0(y),$$

and $\theta \in \Theta \subseteq \mathbb{R}$ itself has a prior density $\theta \sim \pi(\theta)$. Let $f(y)$ be the marginal density

$$f(y) = \int_{\Theta} \pi(\theta) f(y; \theta) d\theta.$$

By using the Bayes formula, the posterior distribution of θ is given by

$$\begin{aligned} \pi(\theta | Y = y) &= \frac{\pi(\theta) f(y; \theta)}{f(y)} \\ &= \frac{\pi(\theta) e^{\theta y - K(\theta)} f_0(y)}{f(y)} \\ &= e^{y\theta - \log\{f(y)/f_0(y)\}} \pi(\theta) e^{-K(\theta)}. \end{aligned}$$

This is an exponential family with natural parameter y , sufficient statistic θ , and cumulant function $\log\{f(y)/f_0(y)\}$ (up to a constant).

As an application of this, suppose $Y \sim N(\mu, \sigma^2)$, where σ^2 is known and μ has a prior density $\pi(\mu)$. Then we have the following *Tweedie's formula*

$$\mathbb{E}(\mu | Y) = Y + \sigma^2 \frac{f'(Y)}{f(Y)}. \quad (3.5)$$

Exercise 3.11. Prove Tweedie's formula.

3.2.5 *Empirical Bayes

A consequence of Tweedie's formula is that the posterior mean of μ only depends on the prior distribution $\pi(\mu)$ through the marginal density $f(y)$. This gives rise to the empirical Bayes methods.

Suppose we observe independent variables $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, n$, where the mean parameters are generated by $\mu_i \stackrel{\text{IID}}{\sim} \pi(\mu)$, $i = 1, \dots, n$ but the density $\pi(\mu)$ is unknown. Suppose σ^2 is known and let $\mathbf{Y} = (Y_1, \dots, Y_n)$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$.

In this model, the MLE of μ is given by $\hat{\boldsymbol{\mu}} = \mathbf{Y}$, with risk

$$\mathbb{E}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2) = \mathbb{E}(\|\mathbf{Y} - \boldsymbol{\mu}\|^2) = n\sigma^2.$$

But we cannot apply the usual asymptotic efficiency theory for MLE here because the dimension of the parameter is not fixed. In fact, the James-Stein estimator

$$\hat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{Y}\|^2}\right) \mathbf{Y}.$$

has a strictly smaller mean squared error than the MLE $\hat{\boldsymbol{\mu}}$ for all values of $\boldsymbol{\mu}$, a result that should be proved in *Principles of Statistics*.

This phenomenon can be best understood in the empirical Bayes framework. If $\pi(\boldsymbol{\mu})$ is known, the optimal estimator of $\boldsymbol{\mu}$ under the mean squared error is given by the posterior mean (the “Bayes estimator”):

$$\hat{\mu}_{\text{Bayes},i} = \mathbb{E}(\mu_i \mid Y_i) = Y_i + \sigma^2 \frac{f'(Y_i)}{f(Y_i)}.$$

The problem is the density $\pi(\boldsymbol{\mu})$ is unknown.

Empirical Bayes departs from the usual Bayesian analysis in the last paragraph by estimating the prior $\pi(\boldsymbol{\mu})$ empirically. This implies that empirical Bayes is indeed a frequentist method. In the normal means problem, empirical Bayes is made simple by Tweedie’s formula because we can estimate the marginal density $f(y)$ directly. In more complicated problems, we may need to solve a deconvolution problem to estimate the prior directly.

Exercise 3.12. Derive the James-Stein estimator by assuming $\mu_i \stackrel{\text{iid}}{\sim} N(0, \tau^2)$ and using the fact that $\mathbb{E}\{(p-2)/\chi_p^2\} = 1$.

What is remarkable about the James-Stein estimator is that it dominates the MLE even though the normal prior on $\boldsymbol{\mu}$ may be wrong.

Tweedie’s formula (3.5) demonstrates a statistical concept called *shrinkage*, which is also closely to regularization. The posterior mean $E(\mu \mid Y)$ is given by the MLE Y (the optimal unbiased estimator) plus a correction term $\sigma^2 f'(Y)/f(Y)$ which increases bias but decreases variance. When $f(\cdot)$ is unimodal, this correction term can be seen as a kind of “regression toward the mean” or a correction to “winner’s curse”; see Figure 3.1 for an illustration.

3.3 Likelihood inference

3.3.1 Maximum likelihood estimator

Consider the setting of IID sampling in Section 3.2.3. That is, suppose $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} f(y; \theta)$ where $f(y; \theta)$ is a one-parameter exponential family with sufficient statistic Y , so the joint density function is given by

$$f(y_1, \dots, y_n; \theta) = e^{n\{\theta\bar{y} - K(\theta)\}} \prod_{i=1}^n f_0(y_i).$$

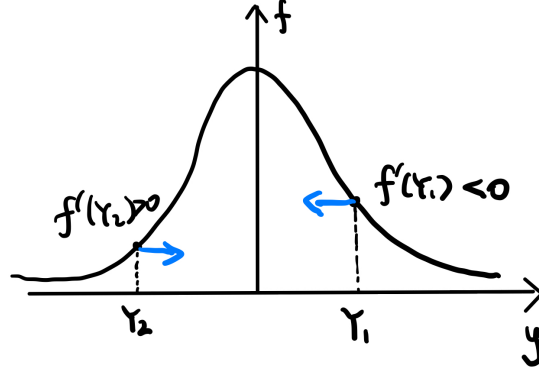


Figure 3.1 Tweedie's formula and shrinkage.

Thus, the log-likelihood function is given by

$$l(\theta) = n \{ \theta \bar{Y} - K(\theta) \} + \text{constant}. \quad (3.6)$$

The *score* is defined as the gradient of the log-likelihood, which in this case is given by

$$U(\theta) = l'(\theta) = n \{ \bar{Y} - K'(\theta) \} = n \{ \bar{Y} - \mu(\theta) \}. \quad (3.7)$$

The last equality uses (3.2), i.e. the first cumulant of a distribution is its mean.

The MLE $\hat{\theta} = \arg \max_{\theta} l(\theta)$ should satisfy the first-order condition $U(\hat{\theta}) = 0$, which means that

$$\hat{\mu} = \mu(\hat{\theta}) = \bar{Y}, \text{ or equivalent } \hat{\theta} = \theta(\bar{Y}).$$

In other words, the MLE simply matches the theoretical mean $\mu(\theta)$ with the observed mean \bar{Y} .

Example 3.13. For $\text{Poisson}(\mu)$, $\hat{\theta} = \log(\hat{\mu}) = \log(\bar{Y})$. For $\text{Binomial}(n, \pi)$ with fixed n , $\hat{\theta} = \log\{\hat{\pi}/(1 - \hat{\pi})\}$ where $\hat{\pi} = \hat{\mu}/n = \bar{Y}/n$.

3.3.2 Asymptotic inference

The large-sample distribution of $\hat{\theta}$ can be obtained by the standard asymptotic theory for MLE. Next we outline some main results in this theory; you are referred to *Principles of Statistics* for rigorous proofs.

By differentiating the identity $\int f(y; \theta) dy = 1$ with respect to θ , under suitable regularity conditions we obtain the *second Bartlett identity*,

$$i^{(n)}(\theta) = \text{Var}\{l'(\theta)\} = \mathbb{E}\{-l''(\theta)\}, \quad (3.8)$$

where we use $i^{(n)}(\theta)$ to denote the *Fisher information* of θ in an IID sample of size n . By using (3.7), we obtain the following formula for exponential families

$$i^{(n)}(\theta) = nK''(\theta) = nV(\theta).$$

Thus in exponential families, the Fisher information $i^{(n)}(\theta)$ is simply the variance function $V(\theta)$ times n .

Exercise 3.14. Prove (3.8).

Consistency of the MLE $\hat{\theta}$ follows from standard arguments and the proof is omitted. To obtain the asymptotic distribution, the general approach is to take the first-order Taylor expansion of the score equation at $\hat{\theta} = \theta$:

$$0 = U(\hat{\theta}) \approx U(\theta) + U'(\theta)(\hat{\theta} - \theta).$$

By using (3.7) and the central limit theorem, we have

$$\frac{U(\theta)}{\sqrt{n}} = \sqrt{n}\{\bar{Y} - \mu(\theta)\} \xrightarrow{d} N(0, V(\theta)).$$

Moreover, (3.7) implies that $U'(\theta) = nK''(\theta) = nV(\theta)$. Thus,

$$\sqrt{n}(\hat{\theta} - \theta) \approx -\frac{U(\theta)/\sqrt{n}}{U'(\theta)/n} \xrightarrow{d} -\frac{N(0, V(\theta))}{V(\theta)} = N\left(0, \frac{1}{V(\theta)}\right). \quad (3.9)$$

Informally, we sometimes write this as

$$\hat{\theta} \dot{\sim} N\left(\theta, \frac{1}{i^{(n)}(\theta)}\right).$$

The above calculations are simplified by the fact that $U'(\theta)$ is a constant for exponential families. In the more general case, one can invoke the law of large numbers and Slutsky's lemma to derive the same asymptotic distribution. That is, (3.9) is generally true for finite-dimensional parametric models (not just one-dimensional exponential families) with the limit covariance matrix given by the inverse of the Fisher information matrix

Because the MLE of θ is given by $\hat{\theta} = \theta(\bar{Y})$, one can also use the delta method to obtain its asymptotic distribution. Roughly speaking, the delta method says that $g(\hat{\eta})$ is asymptotically normal if $\hat{\eta}$ is asymptotically normal and $g(\eta)$ is a smooth function. More specifically, suppose that

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} N(0, \tau^2),$$

where τ^2 may depend on η and $g(\eta)$ is continuously differentiable at η . Then

$$\sqrt{n}\{g(\hat{\eta}) - g(\eta)\} \xrightarrow{d} N(0, \tau^2 g'(\eta)^2).$$

This result can be shown by considering the Taylor expansion of $g(\hat{\eta})$ at η and can be easily extended to multivariate settings.

Exercise 3.15. Prove (3.9) by applying the delta method to $\hat{\theta} = \theta(\bar{Y})$.

3.3.3 Hypothesis testing

Consider testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$ for some $\theta_1 > \theta_0$. By (3.6), the likelihood-ratio statistic is given by

$$l(\theta_1) - l(\theta_0) = n\{(\theta_1 - \theta_0)\bar{Y} - K(\theta_1) + K(\theta_0)\},$$

which is increasing in \bar{Y} . Thus, by the Neyman-Pearson Lemma, the most powerful level- α test rejects H_0 if $\bar{Y} > C_{1-\alpha}$, where $C_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of \bar{Y} under $\theta = \theta_0$. Because this test is independent of θ_1 and controls the type I error for any null parameter value smaller than θ_0 , it is indeed the uniformly most powerful test for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

To test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, one can also resort to asymptotic arguments. The likelihood-ratio statistic is given by $l(\hat{\theta}) - l(\theta_0)$, which converges in distribution to $\chi_1^2/2$ as $n \rightarrow \infty$ by Wilks' theorem.

In general, Wilks' theorem says that, as $n \rightarrow \infty$ and under the null hypothesis, the log-likelihood ratio times two should converge to a χ^2 distribution with a degrees of freedom being the difference in the dimensions of the full parameter space for $H_0 \cup H_1$ and the dimension of the null parameter space for H_0 . A rigorous proof of Wilks' theorem should be covered in *Principles of Statistics*.

3.3.4 Deviance

Deviance is a measure of how one distribution in an exponential family differs from another:

$$\begin{aligned} D(\theta_1, \theta_2) &= 2 \mathbb{E}_{\theta_1} \left\{ \log \frac{f(Y; \theta_1)}{f(Y; \theta_2)} \right\} \\ &= 2 \mathbb{E}_{\theta_1} \{ (\theta_1 - \theta_2)Y - K(\theta_1) + K(\theta_2) \} \\ &= 2 \{ (\theta_1 - \theta_2)\mu_1 - K(\theta_1) + K(\theta_2) \}. \end{aligned} \tag{3.10}$$

If you are familiar with information theory, deviance is simply twice the Kullback-Leibler divergence.

Example 3.16 (Continuing Example 3.2). For the family of normal distributions $N(\mu, 1)$, the natural parameter is $\theta = \mu$ and the cumulant function is $K(\theta) = \theta^2/2$. Therefore,

$$D(\mu_1, \mu_2) = 2 \left\{ (\mu_1 - \mu_2)\mu_1 - \frac{\mu_1^2}{2} + \frac{\mu_2^2}{2} \right\} = (\mu_1 - \mu_2)^2$$

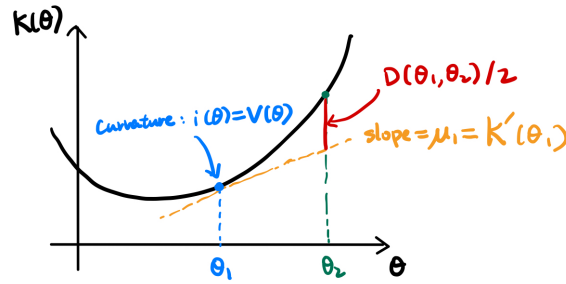
coincides with squared Euclidean distance.

Heuristically, deviance can be thought of as an extension of the Euclidean geometry to exponential families, although generally it is not a distance metric (it is not symmetric and does not obey the triangle inequality). By rewriting μ_1 as $K'(\theta_1)$, we have

$$\frac{D(\theta_1, \theta_2)}{2} = K(\theta_2) - K(\theta_1) - (\theta_2 - \theta_1)K'(\theta_1).$$

Recall that the cumulant function $K(\theta)$ is convex. Thus, the last identity can be informatively represented by the picture in Figure 3.2, which is closely related to duality theory in convex analysis. In particular, this picture shows that the deviance can be locally approximated by the squared Euclidean distance times the Fisher information $i(\theta_1) = i^{(1)}(\theta_1) = V(\theta_1)$, which is curvature of the cumulant function $K(\theta)$ at θ_1 :

$$D(\theta_1, \theta_2) \approx i(\theta_1)(\theta_2 - \theta_1)^2 \text{ for } \theta_2 \approx \theta_1. \tag{3.11}$$



$$\frac{D(\theta_1, \theta_2)}{2} = K(\theta_2) - K(\theta_1) - (\theta_2 - \theta_1)\mu_1$$

Figure 3.2 An informative picture about the deviance in exponential families.

Exercise 3.17. Verify the following formulae:

- For Poisson(λ), the deviance is given by

$$D(\lambda_1, \lambda_2) = 2 \left\{ \lambda_1 \log \frac{\lambda_1}{\lambda_2} - \lambda_1 + \lambda_2 \right\}.$$

- For Binomial(n, π) with fixed n , the deviance is given by

$$D(\pi_1, \pi_2) = 2n \left\{ \pi_1 \log \frac{\pi_1}{\pi_2} + (1 - \pi_1) \log \frac{1 - \pi_1}{1 - \pi_2} \right\}.$$

Deviance also behaves nicely under IID sampling:

$$\begin{aligned} D^{(n)}(\theta_1, \theta_2) &= 2 \mathbb{E}_{\theta_1} \left\{ \log \prod_{i=1}^n \frac{f(Y_i; \theta_1)}{f(Y_i; \theta_2)} \right\} \\ &= \sum_{i=1}^n 2 \mathbb{E}_{\theta_1} \left\{ \log \frac{f(Y_i; \theta_1)}{f(Y_i; \theta_2)} \right\} \\ &= nD(\theta_1, \theta_2). \end{aligned}$$

Exercise 3.18. Show that, for one-parameter exponential family, the likelihood-ratio statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ is given by $D^{(n)}(\hat{\theta}, \theta_0)$. Show that this statistic has a χ_1^2 asymptotic distribution under the null by using (3.11).

3.3.5 Deviance residual

Because deviance can be viewed as an extension to Euclidean distance, it allows us to extend the definition of residuals to exponential families. With an abuse of notation, we use $D(\mu_1, \mu_2)$ to denote the deviance between two distributions in the exponential family with mean μ_1 and μ_2 .

In the normal linear model, $D(y, \mu) = (y - \mu)^2$ is the squared residual. On the other hand, the exponential family analogue of $y - \mu$ is given by

$$\text{sign}(y - \mu)\sqrt{D(y, \mu)}.$$

With IID sampling, the total deviance is given by $D^{(n)}(\hat{\mu}, \mu) = nD(\bar{Y}, \mu)$. This motivates us to define the *deviance residual* by

$$R = \text{sign}(\bar{Y} - \mu)\sqrt{D^{(n)}(\bar{Y}, \mu)}.$$

Exercise 3.19. Use Wilks' theorem to show that $R^2 \xrightarrow{d} \chi_1^2$ as $n \rightarrow \infty$.

In practice, deviance residual R is generally preferred over the Pearson residual

$$R_P = \frac{\bar{Y} - \mu}{\sqrt{V(\mu)/n}},$$

because the distribution of R is much less skewed and closer to the standard normal distribution.

*Bartlett correction

Moreover, it is possible to give a better approximation to the distribution of R via the *Bartlett correction*. Recall that the *skewness* of the distribution of a random variable Y is defined as

$$\frac{\mathbb{E}\{Y - \mathbb{E}(Y)\}^3}{\{\text{Var}(Y)\}^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}},$$

and the *kurtosis* is defined as

$$\frac{\mathbb{E}\{Y - \mathbb{E}(Y)\}^4}{\{\text{Var}(Y)\}^2} - 3 = \frac{\kappa_4}{\kappa_2^2}.$$

By using the Taylor expansion, it can be shown that

$$R = N(-a_n, (1 + b_n)^2) + O_p(n^{-3/2}),$$

where $a_n = O(1/\sqrt{n})$ depends on the skewness of Y and $b_n = O(1/n)$ depends on the skewness and kurtosis. The $O_p(n^{-3/2})$ error term means that

$$\mathbb{P}\left(\frac{R + a_n}{1 + b_n} > z_\alpha\right) = \alpha + O(n^{-3/2}),$$

where z_α is the upper- α quantile of $N(0, 1)$. Therefore, it is possible to obtain inexact but very accurate inference for small n by staying within the exponential family framework.

3.4 Exponential dispersion families

3.4.1 Motivation and definition

The one-parameter exponential family can be restrictive in modelling the distribution of random variables because the distribution is uniquely determined by the mean parameter

μ . In many applications, it is useful to introduce an additional parameter to model how the variance deviates from its theoretical value $V(\theta) = \mu'(\theta)$.

Example 3.20. The normal distribution $N(\mu, \sigma^2)$ is an exponential family only if the variance parameter σ^2 is fixed. As another example, it is undesirable to use the Poisson distribution to model a count variable Y that satisfies $\text{Var}(Y) > \mathbb{E}(Y)$. This is called *overdispersion* and is common in count data.

Rather than considering the full generalization to multi-parameter exponential families, we will make a compromise. An *exponential dispersion family* is a collection of density functions of the form

$$f(y; \theta, \sigma^2) = e^{\{\theta y - K(\theta)\}/\sigma^2} f_0(y; \sigma^2), \quad (3.12)$$

where θ is called the natural parameter, $\sigma^2 > 0$ is called a *dispersion parameter*, and $f_0(y; \sigma^2)$ is a carrying density function. It is straightforward to show that the cumulant generating function for $f(y; \theta, \sigma^2)$ is given by

$$K(t; \theta, \sigma^2) = \mathbb{E}_{\theta, \sigma^2}(e^{tY}) = \frac{1}{\sigma^2} \{K(\sigma^2 t + \theta) - K(\theta)\}. \quad (3.13)$$

Exercise 3.21. Prove (3.13).

Therefore, the mean function would match that of the corresponding exponential family:

$$\mu(\theta, \sigma^2) = \mathbb{E}_{\theta, \sigma^2}(Y) = \left. \frac{\partial}{\partial t} K(t; \theta, \sigma^2) \right|_{t=0} = K'(\theta),$$

and the variance function would have an extra factor of σ^2 :

$$V(\theta, \sigma^2) \text{Var}_{\theta, \sigma^2}(Y) = \left. \frac{\partial^2}{\partial t^2} K(t; \theta, \sigma^2) \right|_{t=0} = \sigma^2 K''(\theta).$$

Compared to the standard exponential family, $\mu(\theta, \sigma^2) = \mu(\theta)$ and $V(\theta, \sigma^2) = \sigma^2 V(\theta)$.

3.4.2 Examples

Example 3.22 (Normal distribution). The density function of $N(\mu, \sigma^2)$ is given by

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = e^{\frac{1}{\sigma^2} \{\mu y - \mu^2/2\}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}.$$

So as expected, $\theta = \mu$ and σ^2 is the dispersion parameter.

Example 3.23 (Poisson distribution). The density function of $\text{Poisson}(\lambda)$ is given by

$$f(y; \lambda) = e^{-\lambda} \frac{\lambda^y}{y!} = e^{y \log \lambda - \lambda} \frac{1}{y!}, \quad y = 0, 1, \dots$$

So $\theta = \log(\lambda)$ and $\mu = \lambda$ as before, and the dispersion parameter $\sigma^2 = 1$ is fixed.

Example 3.24 (Binomial distribution). The density function for $Y \sim \text{Binomial}(n, \pi)/n$ is given by

$$\begin{aligned} f(y; n, \pi) &= \binom{n}{ny} \pi^{ny} (1 - \pi)^{n(1-y)} \\ &= e^{n\{y \log \frac{\pi}{1-\pi} + \log(1-\pi)\}} \binom{n}{ny}, \quad y = 0, 1/n, \dots, 1. \end{aligned}$$

So $\theta = \log \frac{\pi}{1-\pi}$ as before and the dispersion parameter is given by $\sigma^2 = 1/n$.

Example 3.25 (Gamma distribution). The density function for $\Gamma(\alpha, \beta)$ is given by

$$f(y; \alpha, \beta) = \frac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)}, \quad y > 0.$$

Is this an exponential dispersion family? Suppose it is, then the last term in the numerator suggests that $-\beta = \theta/\sigma^2$. Moreover, the mean and variance of the Gamma distribution is given by

$$\mu(\theta) = \alpha/\beta, \quad V(\theta, \sigma^2) = \sigma^2 V(\theta) = \alpha/\beta^2.$$

By using the mean-variance relationship (3.4),

$$\mu'(\theta) = -\frac{1}{\sigma^2} \frac{\partial}{\partial \theta} \frac{\alpha(\theta, \sigma^2)}{\theta} = V(\theta) = \frac{1}{\sigma^2} \frac{\alpha(\theta, \sigma^2)}{\theta^2}.$$

This implies that

$$\frac{\partial}{\partial \theta} \alpha(\theta, \sigma^2) = 0,$$

so $\alpha(\theta, \sigma^2) = \alpha(\sigma^2)$ does not depend on θ . Because the mean

$$\mu(\theta) = \frac{\alpha}{\beta} = \frac{\alpha(\sigma^2)}{-\theta/\sigma^2}$$

only depends on θ , we may take $\alpha = 1/\sigma^2$ and $\beta = -\theta/\sigma^2$. Conversely, $\theta = -\beta/\alpha$ and $\sigma^2 = 1/\alpha$.

Exercise 3.26. Verify that the Gamma density can indeed be written in the form in (3.12) and the cumulant function is given by $K(\theta) = -\log(-\theta)$.

Chapter 4

Generalized linear models

4.1 From linear models to generalized linear models

We are now ready to introduce generalized linear models (GLMs) that expand the classical normal linear models.

4.1.1 Non-normal noise and the Box-Cox transformation

One of the main motivations for considering GLMs is to relax the assumption that the noise variable is normally distributed. This is certainly not true if the response Y is categorical. In many other applications, the distribution of Y may have a heavy tail. For example, how long we spend on social media roughly follows a log-normal distribution (i.e. $\log Y$ is normally distributed).¹ In other cases such as the distribution of wealth, the tail may exhibit a power law.² We can use the central limit theorem to deal with model misspecification (Section 2.4.2). However, with heavy-tailed distributions the central limit theorem may not hold or may only provide a poor approximation for a moderate sample size.

One solution to non-normal noise is to model a transformation of the response. The *Box-Cox transformation* applies the following function to a positive response Y :

$$Y \mapsto Y^{(\lambda)} = \begin{cases} (Y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log Y, & \text{if } \lambda = 0. \end{cases}$$

One can then fit the usual normal linear model. The tuning parameter λ that defines the transformation can be selected using the maximum likelihood estimator (jointly over β and λ) or some visualization tool such as the Q-Q plot.

The Box-Cox transformation attempts to “kill three birds with one stone” in the sense that it uses a transformation indexed by a single parameter λ to achieve normality, linearity in the regressors, and variance stability. In practice, these goals are often difficult to achieve together.

4.1.2 Three components of a generalized linear model

The key idea of GLMs is to use the theory for exponential families, which offer more flexibility than the Box-Cox transformation. GLMs consist of three components:

- (i) Modelling the distribution of Y_i given \mathbf{X}_i using an exponential (dispersion) family;
- (ii) A linear *predictor* $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$;
- (iii) A strictly increasing and smooth *link function* $g(\cdot)$ that relates the predictor η_i with the conditional expectation $\mu_i = \mathbb{E}(Y_i \mid \mathbf{X}_i)$: $\eta_i = g(\mu_i)$ and $\mu_i = g^{-1}(\eta_i)$.

As an example, the familiar normal linear model corresponds to assuming

- (i) $Y_i \mid \mathbf{X}_i \sim N(\mu_i, \sigma^2)$;
- (ii) $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$;
- (iii) $\mu_i = \eta_i$, so $g(\cdot)$ is the identity function.

To simplify the exposition, we will adopt the same vector/matrix notation as in linear models:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_n^T \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \text{ and } \boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}.$$

Unless noted otherwise, the regressors \mathbf{X} will be treated as fixed. In other words, the inference for GLMs will be conditioned on \mathbf{X} . See Section 2.1 for discussion on this in the normal linear model. **In this Chapter, we will always assume the generalized linear model is correct when considering statistical inference (e.g. finding MLE of $\boldsymbol{\beta}$).**

4.2 The canonical form

The canonical form of a generalized linear model corresponds to setting the natural parameter $\theta = \eta$ and identity dispersion $\sigma^2 = 1$. It provides most of the insights into the general theory for GLMs without getting into too much technical details.

More concretely, let $\{f(y; \theta) \mid \theta \in \Theta\}$ be a one-parameter exponential family as defined in the previous Chapter. A *canonical form GLM* assumes that the responses Y_1, \dots, Y_n are independent and

$$Y_i \mid \mathbf{X}_i \sim f(y; \theta_i), \quad i = 1, \dots, n,$$

where the natural parameter is given by

$$\theta_i = \eta_i = \mathbf{X}_i^T \boldsymbol{\beta}.$$

An immediate consequence is that the mean parameter is given by

$$\mu_i = \mathbb{E}(Y_i \mid \mathbf{X}_i) = \mu(\theta_i) = \mu(\eta_i),$$

so the *canonical link function* is given by $g(\mu) = \theta(\mu)$.

The joint density of \mathbf{Y} is given by

$$\begin{aligned} f(\mathbf{y}; \beta) &= \prod_{i=1}^n f(y_i; \theta_i) \\ &= e^{\sum_{i=1}^n \theta_i y_i - K(\theta_i)} \prod_{i=1}^n f_0(y_i) \\ &= e^{\beta^T \mathbf{X}^T \mathbf{y} - \sum_{i=1}^n K(\mathbf{X}_i^T \beta)} \prod_{i=1}^n f_0(y_i). \end{aligned}$$

This is a p -parameter exponential family, with

- Natural parameter β ;
- Sufficient statistic $\mathbf{Z} = \mathbf{X}^T \mathbf{Y}$; and
- Cumulant function $\phi(\beta) = \sum_{i=1}^n K(\mathbf{X}_i^T \beta)$.

Therefore, the canonical form GLMs can be studied using the theory for multi-parameter exponential families and have many nice properties that generalize the theory in Section 3.2. For example, it can be shown that the expectation and covariance matrix of \mathbf{Z} are given by the gradient and Hessian matrix of the cumulant function:

$$\begin{aligned} \mathbb{E}_{\beta}(\mathbf{Z}) &= \nabla \phi(\beta) = \sum_{i=1}^n K'(\mathbf{X}_i^T \beta) \mathbf{X}_i = \mathbf{X}^T \boldsymbol{\mu}(\beta), \\ \text{Cov}_{\beta}(\mathbf{Z}) &= \nabla^2 \phi(\beta) = \sum_{i=1}^n K''(\mathbf{X}_i^T \beta) \mathbf{X}_i \mathbf{X}_i^T = \mathbf{X}^T \mathbf{V}(\beta) \mathbf{X}, \end{aligned}$$

where $\boldsymbol{\mu}(\beta) = (\mu_1(\beta), \dots, \mu_n(\beta))^T$ and $\mathbf{V}(\beta) = \text{diag}(K''(\mathbf{X}_1^T \beta), \dots, K''(\mathbf{X}_n^T \beta)) = \text{diag}(\text{Var}(Y_1), \dots, \text{Var}(Y_n))$.

The log-likelihood function of β is given by

$$l(\beta) = \beta^T \mathbf{X}^T \mathbf{Y} - \sum_{i=1}^n K(\mathbf{X}_i^T \beta) + \text{constant}.$$

The score function is given by

$$\mathbf{U}(\beta) = \nabla l(\beta) = \mathbf{X}^T \mathbf{Y} - \sum_{i=1}^n K'(\mathbf{X}_i^T \beta) \mathbf{X}_i = \mathbf{X}^T \{\mathbf{Y} - \boldsymbol{\mu}(\beta)\}.$$

Thus, the MLE $\hat{\beta}$ satisfies the normal equations

$$\mathbf{X}^T \{\mathbf{Y} - \boldsymbol{\mu}(\hat{\beta})\} = \mathbf{0}. \quad (4.1)$$

Geometrically, the MLE is obtained by projecting \mathbf{Y} onto $\{\boldsymbol{\mu}(\beta) \mid \beta \in \mathbf{R}^p\}$, a p -dimensional manifold in \mathbb{R}^n . Of course, the GLM normal equations (4.1) reduce to the normal equations (2.5) for the normal location family.

For asymptotic inference of GLMs, the one-dimensional theory in Section 3.3.2 can be extended in a straightforward manner. The *Fisher information matrix* for $\boldsymbol{\beta}$ is defined as

$$\mathbf{I}^{(n)}(\boldsymbol{\beta}) = \text{Cov}\{\mathbf{U}(\boldsymbol{\beta})\} = \mathbb{E}\{-\nabla^2 l(\boldsymbol{\beta})\} = \sum_{i=1}^n K''(\mathbf{X}_i^T \boldsymbol{\beta}) \mathbf{X}_i \mathbf{X}_i^T = \mathbf{X}^T \mathbf{V}(\boldsymbol{\beta}) \mathbf{X}.$$

The asymptotic theory for the MLE suggests that, under suitable regularity conditions,

$$\hat{\boldsymbol{\beta}} \sim \text{N}(\boldsymbol{\beta}, \mathbf{I}^{(n)}(\boldsymbol{\beta})^{-1}).$$

This is an informal way of writing the convergence in distribution

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \text{N}(\mathbf{0}, \mathbf{I}(\boldsymbol{\beta})^{-1}), \text{ as } n \rightarrow \infty,$$

where $\mathbf{I}(\boldsymbol{\beta}) = \lim_{n \rightarrow \infty} \mathbf{I}^{(n)}(\boldsymbol{\beta})/n$ is assumed to exist.

4.3 Linkage and over-dispersion

Next, we introduce linkage and dispersion to GLMs. In this more general setup, it is assumed that Y_1, \dots, Y_n are independent and $Y_i \mid \mathbf{X}_i \sim f(y; \theta_i, \sigma_i^2)$ follows a distribution from an exponential dispersion family

$$f(y; \theta, \sigma^2) = e^{\{\theta y - K(\theta)\}/\sigma^2} f_0(y; \sigma^2),$$

where the natural and dispersion parameters are modelled by

$$\theta_i = \theta(\mu_i) = \theta(g^{-1}(\eta_i)) = \theta(g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta})),$$

where $g(\cdot)$ is strictly monotone (usually increasing) and twice differentiable, and

$$\sigma_i^2 = \sigma^2 / w_i,$$

where σ^2 is possibly unknown and w_i is some known weight.

4.3.1 Estimation

The log-likelihood function of this model is given by

$$l(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^n \frac{1}{\sigma_i^2} \{\theta_i Y_i - K(\theta_i)\} + \log f_0(Y_i; \sigma_i^2). \quad (4.2)$$

This function depends on the coefficients $\boldsymbol{\beta}$ through $\theta_1, \dots, \theta_n$. By differentiating $l(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$, the $\boldsymbol{\beta}$ -score is given by

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) = \nabla_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n w_i \frac{(Y_i - \mu_i) \mathbf{X}_i}{V(\mu_i) g'(\mu_i)} = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{W}(\mathbf{Y} - \boldsymbol{\mu}),$$

where

$$\mathbf{W} = \mathbf{W}(\boldsymbol{\beta}) = \text{diag} \left(\frac{w_i}{V(\mu_i)\{g'(\mu_i)\}^2} \right).$$

Thus, the MLE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ should solve

$$\sum_{i=1}^n w_i \frac{(Y_i - \hat{\mu}_i) \mathbf{X}_i}{V(\hat{\mu}_i)g'(\hat{\mu}_i)} = \mathbf{0}, \quad (4.3)$$

where $\hat{\mu}_i = g^{-1}(\mathbf{X}_i^T \hat{\boldsymbol{\beta}})$. Equation (4.3) does not have a clear geometric interpretation like (4.1), but, crucially, it does not depend on σ^2 . This means that including the dispersion parameter σ^2 in the model does not change the MLE $\hat{\boldsymbol{\beta}}$.

Exercise 4.1. Show (4.3) reduces to (4.1) when $g(\cdot)$ is the canonical link function and $w_1 = \dots = w_n = 1$.

To estimate σ^2 , in the normal linear model we use

$$\hat{\sigma}^2 = \frac{1}{n-p} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2.$$

In GLMs, recall that $\mathbb{E}\{(Y_i - \mu_i)^2\} = \text{Var}(Y_i) = (\sigma^2/w_i)V(\mu_i)$. This motivates us to estimate σ^2 (if it is unknown) by

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n w_i \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

To use a consistent notation, we set $\hat{\sigma}^2 = \sigma^2$ when σ^2 is known (e.g. $\sigma^2 = 1$ in the canonical form GLM).

4.3.2 Asymptotic normality and confidence intervals

By computing the Hessian of $l(\boldsymbol{\beta}, \sigma^2)$, it can be shown that the Fisher information matrix for $(\boldsymbol{\beta}, \sigma^2)$ is block-diagonal

$$\mathbf{I}^{(n)}(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{(n)}(\boldsymbol{\beta}, \sigma^2) & 0 \\ 0 & I_{\sigma^2\sigma^2}^{(n)}(\boldsymbol{\beta}, \sigma^2) \end{pmatrix}. \quad (4.4)$$

Therefore, including the dispersion parameter σ^2 does not change the information for $\boldsymbol{\beta}$, which is given by

$$\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{(n)}(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{W} \mathbf{X}. \quad (4.5)$$

Exercise 4.2. Prove (4.4) and (4.5).

The standard asymptotic theory shows that, under suitable regularity conditions, $\hat{\beta}$ has an asymptotic normal distribution

$$\hat{\beta} \sim N\left(\beta, I_{\beta\beta}^{(n)}(\beta, \sigma^2)^{-1}\right) = N\left(\beta, \sigma^2(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}\right).$$

The asymptotic variance on the right hand side depends on the unknown parameters β and σ^2 . To construct a confidence interval/region, they can be replaced by their estimators $\hat{\beta}$ and $\hat{\sigma}^2$. For example, an equi-tailed asymptotic $(1 - \alpha)$ confidence interval for $\beta_j, j \in \{1, \dots, p\}$ is given by

$$\left[\hat{\beta}_j - z_{\alpha/2} \sqrt{I_{\beta\beta}(\hat{\beta}, \hat{\sigma}^2)^{-1}_{jj}}, \hat{\beta}_j + z_{\alpha/2} \sqrt{I_{\beta\beta}(\hat{\beta}, \hat{\sigma}^2)^{-1}_{jj}} \right].$$

With a moderate sample size n , a slightly more accurate inference may be obtained by replacing the normal upper-quantile $z_{\alpha/2}$ with $t_{n-p}(\alpha/2)$, the upper-quantile of the t distribution, which is motivated by the exact theory for the normal linear model.³

4.3.3 Overdispersion due to clustering

At this point, you might wonder why we sometimes want to model the dispersion in GLMs. Overdispersion, the phenomenon that the variance of Y is larger than what is expected from a theoretical model, often occurs in practice. A common mechanism for overdispersion and underdispersion is unaccounted structure in the sample. This is illustrated by the next sample.

Example 4.3. Suppose a sample of size n has n/k clusters, each of size k . The observations are distributed as $Z_{ij} \sim \text{Bernoulli}(\pi_i), i = 1, \dots, n/k, j = 1, \dots, k$. The response Y is the total $Y = \sum_{i=1}^{n/k} \sum_{j=1}^k Z_{ij}$, which is often modelled by a binomial distribution. This is reasonable when $\pi_i = \pi$ for all i , as Y then follows a $\text{Binomial}(n, \pi)$ distribution with

$$\mathbb{E}(Y) = n\pi, \text{ Var}(Y) = n\pi(1 - \pi).$$

However, if the probabilities π_i themselves are IID and

$$\mathbb{E}(\pi_i) = \pi, \text{ Var}(\pi_i) = \tau^2\pi(1 - \pi),$$

it is straightforward to show that

$$\mathbb{E}(Y) = n\pi, \text{ Var}(Y) = \sigma^2 n\pi(1 - \pi), \text{ where } \sigma^2 = 1 + \tau^2(k - 1).$$

That is, the mean of Y is unchanged but the variance is increased by a factor of σ^2 .

4.4 Analysis of deviance

The deviance is a key concept in exponential families and GLMs. It extends the RSS/variance in normal linear models as a way to measure the goodness-of-fit of a GLM. Analysis of deviance is an extension of the ANOVA in the normal linear model

(Section 2.3.4). Recall that in a one-parameter exponential family $\{f(y; \theta) \mid \theta \in \Theta\}$, the deviance between $f(y; \theta_1)$ and $f(y; \theta_2)$ is defined as

$$\begin{aligned} D(\theta_1, \theta_2) &= 2 \mathbb{E}_{\theta_1} \{\log f(Y; \theta_1) - \log f(Y; \theta_2)\} \\ &= 2 \{(\theta_1 - \theta_2)\mu_1 - K(\theta_1) + K(\theta_2)\}. \end{aligned}$$

As discussed in Section 3.3.4, deviance extends the Euclidean geometry to exponential families. With an abuse of notation, it is often convenient to parameterize an exponential family distribution by its mean and write the deviance as $D(\mu_1, \mu_2)$. For two n -vectors of mean-value parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, their *total deviance* is defined as

$$D^{(n)}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \sum_{i=1}^n D(\mu_{1i}, \mu_{2i}).$$

4.4.1 Nested models

The *full model* is given by

$$\boldsymbol{\theta} = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. As in Section 2.2.3, suppose the design matrix \mathbf{X} and coefficient vector $\boldsymbol{\beta}$ are partitioned as

$$\mathbf{X} = (\mathbf{X}_0 \ \mathbf{X}_1), \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix},$$

where $\mathbf{X}_0 \in \mathbb{R}^{n \times p_0}$, $\mathbf{X}_1 \in \mathbb{R}^{n \times (p-p_0)}$, $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0 \times 1}$, and $\boldsymbol{\beta}_1 \in \mathbb{R}^{(p-p_0) \times 1}$. The *submodel* or *null model* we consider is

$$\boldsymbol{\theta} = \boldsymbol{\eta} = \mathbf{X}_0\boldsymbol{\beta}_0.$$

In other words, we are interested in testing the hypothesis $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ against $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$.

According to the GLM normal equations (4.1), the full model MLE $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ and submodel MLE $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p_0}$ satisfy

$$\begin{aligned} \mathbf{X}^T \{\mathbf{Y} - \hat{\boldsymbol{\mu}}\} &= \mathbf{0}, \text{ where } \hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \mu(\mathbf{X}_1^T \hat{\boldsymbol{\beta}}) \\ \vdots \\ \mu(\mathbf{X}_n^T \hat{\boldsymbol{\beta}}) \end{pmatrix}; \text{ and} \\ \mathbf{X}_0^T \{\mathbf{Y} - \hat{\boldsymbol{\mu}}_0\} &= \mathbf{0}, \text{ where } \hat{\boldsymbol{\mu}}_0 = \boldsymbol{\mu}(\mathbf{X}_0\hat{\boldsymbol{\beta}}_0) = \begin{pmatrix} \mu(\mathbf{X}_1^T \hat{\boldsymbol{\beta}}_0) \\ \vdots \\ \mu(\mathbf{X}_n^T \hat{\boldsymbol{\beta}}_0) \end{pmatrix}. \end{aligned}$$

See Figure 4.1 for an geometric illustration of nested GLM fits.

A GLM is called *saturated* if \mathbf{X} has rank n (which implies $p \geq n$). In this case, $\hat{\boldsymbol{\mu}} = \mathbf{Y}$. Assuming the intercept is also included in the model, the smallest GLM is given by $p = 1$ and $\mathbf{X} = \mathbf{1}_n$. In this case, $\hat{\boldsymbol{\mu}} = \bar{Y}\mathbf{1}_n$.

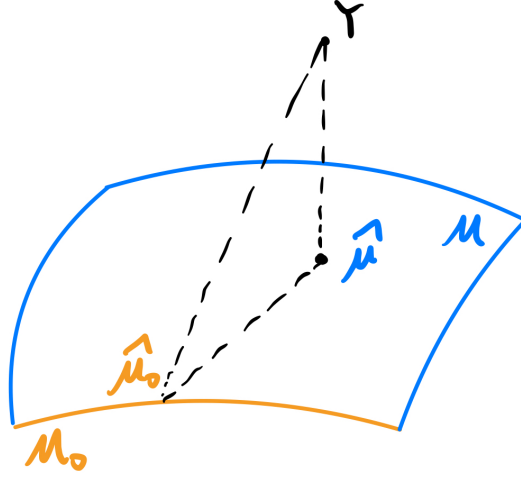


Figure 4.1 Illustration of nested GLMs. The full model space is given by $\mathcal{M} = \{\mu(X\beta) \mid \beta \in \mathbb{R}^p\}$ and the submodel space is given by $\mathcal{M}_0 = \{\mu(X_0\beta_0) \mid \beta_0 \in \mathbb{R}^{p_0}\}$.

4.4.2 The deviance additivity theorem

The *deviance additivity theorem* applies to the setup of nested canonical form GLMs.⁴ It says that the deviance between the observations (or equivalently the saturated model) and the submodel can be decomposed as

$$D^{(n)}(Y, \hat{\mu}_0) = D^{(n)}(Y, \hat{\mu}) + D^{(n)}(\hat{\mu}, \hat{\mu}_0). \quad (4.6)$$

This equation follows immediately from the following relation between the deviance and log-likelihood:

$$D^{(n)}(\hat{\mu}, \hat{\mu}_0) = 2 \{l(\hat{\mu}) - l(\hat{\mu}_0)\}, \quad (4.7)$$

where $l(\mu) = \sum_{i=1}^n \log f(Y_i; \theta(\mu_i))$ is the log-likelihood function. In Example 3.16, we saw that in normal linear models with $\sigma^2 = 1$, we have $D^{(n)}(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|^2$. So in this case the deviance additivity theorem reduces to Pythagoras' theorem.

Exercise 4.4. Show (4.7), then use it to prove (4.6).

The key identity (4.7) can be extended to GLMs with a canonical link function but a undetermined dispersion σ^2 . In this case, it can be shown that⁵

$$D^{(n)}(\hat{\mu}, \hat{\mu}_0) = 2\sigma^2 \{l(\hat{\mu}) - l(\hat{\mu}_0)\}. \quad (4.8)$$

Exercise 4.5. Prove (4.8).

4.4.3 Analysis of deviance

By Wilks' theorem and (4.7), we have $D^{(n)}(\hat{\mu}, \hat{\mu}_0) \xrightarrow{d} \chi_{p-p_0}^2$ as $n \rightarrow \infty$ under the null $H_0 : \beta_1 = 0$. So we reject H_0 if

$$D^{(n)}(\hat{\mu}, \hat{\mu}_0) > \chi_{p-p_0}^2(\alpha).$$

In GLMs with the canonical link function and a dispersion parameter σ^2 , the deviance should be divided by an estimator of σ^2 following (4.8). Motivated by the exact F -test for normal linear models, a slightly more accurate test in small samples rejects H_0 if

$$\frac{D^{(n)}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_0)}{\hat{\sigma}^2} > (p - p_0)F_{p-p_0, n-p}(\alpha).$$

With a sequence of nested GLMs, one can further perform a chain of analyses of deviance. **The interpretation is most straightforward when the link function is canonical so deviance is additive.**

4.5 Numerical computation

Up till now, we have not said anything about how the MLE $\hat{\boldsymbol{\beta}}$ can be computed. Unlike in the normal linear model where $\hat{\boldsymbol{\beta}}$ can be found by solving some linear equations, the score equations (4.1) or (4.3) for GLMs are not linear in $\boldsymbol{\beta}$. Thus, some iterative algorithms are needed.

4.5.1 Newton-Raphson

The Newton-Raphson algorithm is a general algorithm for optimization or root finding problems. We illustrate this with a classical problem in statistics—finding the MLE. Consider the optimization problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{maximize}} \quad l(\boldsymbol{\beta}),$$

where $l(\boldsymbol{\beta})$ is the log-likelihood function for some statistics problem. Let $\mathbf{U}(\boldsymbol{\beta})$ and $\mathbf{H}(\boldsymbol{\beta})$ be the gradient and Hessian matrix of $l(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$. That is,

$$U_k(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_k} l(\boldsymbol{\beta}), \quad k = 1, \dots, p,$$

$$H_{jk}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial \beta_j \partial \beta_k} l(\boldsymbol{\beta}), \quad j, k = 1, \dots, p.$$

The key idea of the Newton-Raphson algorithm is that the objective function $l(\boldsymbol{\beta})$ can be locally approximate near $\boldsymbol{\beta}^* \in \mathbb{R}^p$ by its second-order Taylor expansion (assuming the function is sufficiently smooth):

$$l(\boldsymbol{\beta}) \approx l(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \mathbf{U}(\boldsymbol{\beta}^*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \mathbf{H}(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*).$$

Because the local approximation is a quadratic function of $\boldsymbol{\beta}$, we can easily find its maximizer. By differentiating with respect to $\boldsymbol{\beta}$, the maximizer should satisfy

$$\mathbf{U}(\boldsymbol{\beta}^*) + \mathbf{H}(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{0}.$$

This motivates the following iterative algorithm (see Figure 4.2):

- (i) Start at an initial parameter value $\beta^{(0)}$.
- (ii) For $t = 1, 2, \dots$, update the parameter by

$$\beta^{(t)} = \beta^{(t-1)} - \left\{ \mathbf{H}(\beta^{(t-1)}) \right\}^{-1} \mathbf{U}(\beta^{(t-1)}).$$

- (iii) Stop the algorithm until the sequence $\beta^{(t)}$ converges in a numerical sense (e.g. if $l(\beta^{(t)}) - l(\beta^{(t-1)}) < \tau$ where τ is some tolerance level).

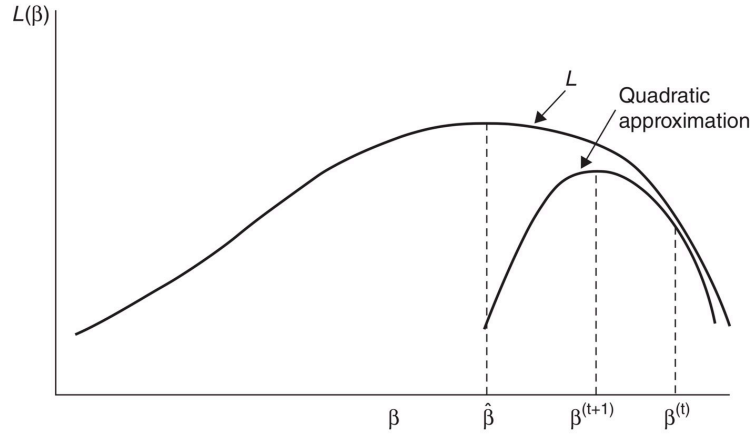


Figure 4.2 An illustration of the Newton-Raphson algorithm.⁶

4.5.2 Fisher scoring

A drawback of the Newton-Raphson algorithm is that the Hessian matrix $\mathbf{H}(\beta^{(t-1)})$ is sometimes close to singularity, making its inverse numerically unstable.

When the objective function $l(\beta)$ is the log-likelihood of some IID data, the Fisher information matrix is the expectation of the negative Hessian matrix (which is sometimes called the observed information):

$$\mathbf{I}(\beta) = \mathbb{E}_{\beta} \{ -\mathbf{H}(\beta) \}.$$

The Fisher information matrix is guaranteed to be positive definite. *Fisher scoring* refers to the modification of the Newton-Raphson algorithm where $-\mathbf{H}(\beta^{(t-1)})$ is replaced by $\mathbf{I}(\beta^{(t-1)})$.⁷

4.5.3 Iteratively reweighted least squares

Let us now apply the general algorithms above to GLMs. For the most general form of GLM described in Section 4.3, the log-likelihood function is given by (4.2) and is repeated below:

$$l(\beta, \sigma^2) = \sum_{i=1}^n \frac{1}{\sigma_i^2} \{ \theta_i Y_i - K(\theta_i) \} + \log f_0(Y_i; \sigma_i^2),$$

where $\theta_i = \theta(g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta}))$ and $\sigma_i^2 = \sigma^2 / w_i$. The $\boldsymbol{\beta}$ -score is given by

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) = \nabla_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n w_i \frac{(Y_i - \mu_i) \mathbf{X}_i}{V(\mu_i) g'(\mu_i)},$$

and the Hessian matrix can be obtained by further differentiating $\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$. In general, this is a complicated matrix, but the calculations greatly simplify if $g(\mu)$ is the canonical link, i.e. $g(\mu) = \theta(\mu)$. In this case, the negative Hessian matrix (a.k.a. the observed information matrix) is indeed equal to the Fisher information matrix:

$$-\mathbf{H}_{\boldsymbol{\beta}, \boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) = \mathbf{I}_{\boldsymbol{\beta}, \boldsymbol{\beta}}^{(n)}(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{W} \mathbf{X}, \quad (4.9)$$

where \mathbf{W} is defined in (4.5). So for GLMs using a canonical link function, the Newton-Raphson algorithm coincides with the Fisher scoring algorithm.

Exercise 4.6. Prove (4.9) when $g(\mu)$ is the canonical link function.

The Fisher scoring algorithm admits a nice representation by defining a “residual” in the predictor

$$R_i = (Y_i - \mu_i) g'(\mu_i), \quad \mathbf{R} = \begin{pmatrix} R_1 \\ \vdots \\ R_n \end{pmatrix}.$$

With this new definition, we can express the $\boldsymbol{\beta}$ -score as

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{W} \mathbf{R}.$$

Let $\eta_i^{(t)} = \mathbf{X}_i^T \boldsymbol{\beta}^{(t)}$, $\mu_i^{(t)} = g^{-1}(\eta_i^{(t)})$, and similarly define $\mathbf{W}^{(t)}$ and $\mathbf{R}^{(t)}$. The Fisher scoring update is then given by

$$\begin{aligned} \boldsymbol{\beta}^{(t)} &= \boldsymbol{\beta}^{(t-1)} + \{\mathbf{I}_{\boldsymbol{\beta}, \boldsymbol{\beta}}^{(n)}(\boldsymbol{\beta}^{(t-1)}, \sigma^2)\}^{-1} \mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(t-1)}, \sigma^2) \\ &= \boldsymbol{\beta}^{(t-1)} + \left(\mathbf{X}^T \mathbf{W}^{(t-1)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(t-1)} \mathbf{R}^{(t-1)} \\ &= \left(\mathbf{X}^T \mathbf{W}^{(t-1)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(t-1)} \left(\boldsymbol{\eta}^{(t-1)} + \mathbf{R}^{(t-1)} \right). \end{aligned}$$

The last expression is the solution to a weighted least squares problem (Section 2.4.1). Therefore, the Fisher scoring algorithm for GLMs is also known as the *iteratively reweighted least squares* that updates the model parameters as follows

$$\boldsymbol{\beta}^{(0)} \rightarrow \boldsymbol{\eta}^{(0)}, \boldsymbol{\mu}^{(0)} \rightarrow \mathbf{W}^{(0)}, \mathbf{R}^{(0)} \xrightarrow{\text{WLS}} \boldsymbol{\beta}^{(1)} \rightarrow \boldsymbol{\eta}^{(1)}, \boldsymbol{\mu}^{(1)} \rightarrow \dots$$

To initiate the algorithm, it is common to choose $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ or $\boldsymbol{\mu}^{(0)} = \mathbf{Y}$.

4.6 Model diagnostics and model selection

The diagnosis of GLMs is very similar to that of linear models, thanks to the iteratively reweighted least squares formulation of the MLE.

4.6.1 The general idea

The key idea is to define the pseudo-response:

$$\mathbf{Z}^{(t)} = \boldsymbol{\eta}^{(t)} + \mathbf{R}^{(t)}$$

So the Fisher scoring update can be written as

$$\boldsymbol{\beta}^{(t)} = \left(\mathbf{X}^T \mathbf{W}^{(t-1)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(t-1)} \mathbf{Z}^{(t-1)}.$$

Let $\hat{\mathbf{Z}}$ be the limit of $\mathbf{Z}^{(t)}$ as $t \rightarrow \infty$. That is,

$$\hat{Z}_i = \lim_{t \rightarrow \infty} \mathbf{X}_i^T \boldsymbol{\beta}^{(t)} + (Y_i - \mu_i^{(t)})g'(\mu_i^{(t)}) = \mathbf{X}_i^T \hat{\boldsymbol{\beta}} + (Y_i - \hat{\mu}_i)g'(\hat{\mu}_i), \quad i = 1, \dots, n.$$

Further, let

$$\hat{\mathbf{W}} = \mathbf{W}(\hat{\boldsymbol{\beta}}) = \text{diag} \left(\frac{\textcolor{red}{w}_i}{V(\hat{\mu}_i)\{g'(\hat{\mu}_i)\}^2} \right) \text{ and } \hat{\mathbf{V}} = \text{diag}(V(\hat{\mu}_i)/\textcolor{red}{w}_i).$$

So we have the following convenient matrix representation

$$\hat{\mathbf{Z}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{R}}, \text{ where } \hat{\mathbf{R}} = \hat{\mathbf{W}}^{-1/2} \hat{\mathbf{V}}^{-1/2} (\mathbf{Y} - \hat{\boldsymbol{\mu}}).$$

The GLM diagnosis proceeds by treating $\hat{\mathbf{W}}^{1/2} \hat{\mathbf{Z}}$, $\hat{\mathbf{W}}^{1/2} \hat{\boldsymbol{\eta}}$, and $\hat{\mathbf{W}}^{1/2} \hat{\mathbf{R}}$ as the “adjusted” responses, fitted values, and residuals (see Section 2.4.1). They often behave like their counterparts in the linear model. For example, the adjusted fitted values are given by

$$\hat{\mathbf{W}}^{1/2} \hat{\boldsymbol{\eta}} = \underbrace{\hat{\mathbf{W}}^{1/2} \mathbf{X} (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}}^{1/2}}_{\mathbf{H}} \hat{\mathbf{W}}^{1/2} \hat{\mathbf{Z}},$$

which motivates the definition of the *adjusted hat matrix* \mathbf{H} (not to be confused with the Hessian matrix in the previous section).

4.6.2 Redefining residuals

It can be shown that

$$\begin{aligned} \text{Var}(\mathbf{Y} - \hat{\boldsymbol{\mu}}) &\approx \sigma^2 \hat{\mathbf{V}}^{1/2} (\mathbf{I} - \mathbf{H}) \hat{\mathbf{V}}^{1/2}, \\ \mathbf{H} \hat{\mathbf{V}}^{-1/2} (\mathbf{Y} - \boldsymbol{\mu}) &\approx \hat{\mathbf{V}}^{-1/2} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}). \end{aligned} \tag{4.10}$$

This motivates us to maintain the definition of the *leverage* of observation i as H_{ii} .

There are several versions of residuals one can use for GLMs. The most common ones are the *Pearson residual*

$$R_{P,i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}^2 V(\hat{\mu}_i)/\textcolor{red}{w}_i}},$$

and the *deviance residual* (see Section 3.3.5)

$$R_{D,i} = \text{sign}(Y_i - \hat{\mu}_i) \sqrt{D(Y_i, \hat{\mu}_i)} / (\sqrt{\hat{\sigma}^2 / \textcolor{red}{w}_i}).$$

The Pearson residual is closely related to the adjusted residual. In fact, it is straightforward to verify that $\mathbf{R}_P = \mathbf{W}^{1/2} \hat{\mathbf{R}} / \hat{\sigma}$. However, the variance of $R_{P,i}$ is usually smaller than 1 upon considering (4.10). So just like in the diagnosis of linear models, the following *standardized Pearson residual* is often used:

$$\tilde{R}_{P,i} = \frac{R_{P,i}}{\sqrt{1 - H_{ii}}}.$$

Similarly, it is common to use the *standardized deviance residual*

$$\tilde{R}_{D,i} = \frac{R_{D,i}}{\sqrt{1 - H_{ii}}}.$$

Cook's distance can be similarly extended to GLMs:

$$D_i = \frac{1}{p} \frac{H_{ii}}{1 - H_{ii}} \tilde{R}_{P,i}^2.$$

4.6.3 Model selection

To select a GLM, we cannot apply Mallows' C_p criterion because it relies on mean squared error. However, we can still use cross-validation by replacing the squared error with the deviance. In other words, we seek a model that minimizes

$$\text{CV}(\text{model}) = \sum_{i=1}^n D(Y_i, \hat{\mu}_{-i}),$$

where $\hat{\mu}_{-i}$ is the leave-one-out fitted value for the i th observation. AIC and BIC can be applied in the same way to GLMs by using the corresponding log-likelihood function.

Regarding algorithms for model selection, the stepwise methods and the best subset method can be applied in the same way as before. Regularization can be achieved by adding the same penalty on certain complexity measure of $\boldsymbol{\beta}$ as before.

4.7 Binomial regression

In the rest of this Chapter, we discuss two of the most widely used families of GLMs: binomial regression and Poisson regression.

In a binomial regression, it is assumed that the responses Y_1, \dots, Y_n are independent and

$$Y_i \sim \frac{1}{n_i} \text{Binomial}(n_i, \mu_i), \quad i = 1, \dots, n,$$

where n_i is known but μ_i is unknown. As we have seen in Example 3.24, $\text{Binomial}(n, \mu)$ is an exponential dispersion family:

$$\begin{aligned} f(y; n, \mu) &= \binom{n}{y} \mu^y (1 - \mu)^{n-y} \\ &= \exp \left\{ \frac{1}{n-1} \left(y \log \frac{\mu}{1 - \mu} + \log(1 - \mu) \right) \right\} \binom{n}{y}. \end{aligned}$$

The natural parameter is $\theta = \log\{\mu/(1 - \mu)\}$, and the cumulant function is given by $K(\theta) = \log(1 + e^\theta)$. In the GLM context, we can fix dispersion parameter at $\sigma^2 = 1$ and use the dispersion weight $w = n$.

4.7.1 Common link functions

Recall that the link function relates the linear predictor with the mean value. Specifically, $g(\mu_i) = \eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$. The canonical link makes η_i equal to the natural parameter θ_i , so for binomial regression the canonical link is given by the logit function

$$g(\mu) = \theta(\mu) = \log \frac{\mu}{1 - \mu}.$$

More generally, we can choose $g(\mu)$ to be any strictly increasing function from $(0, 1)$ to \mathbb{R} .⁸ In other words, we can let g to be the quantile function (inverse of the CDF) of any continuous random variable ϵ . The logit link corresponds to the *logistic distribution*, whose distribution function is simply the expit function:

$$F(\eta) = \frac{e^\eta}{1 + e^\eta}.$$

Another commonly used link is the *probit link* $g(\mu) = \Phi^{-1}(\mu)$, which corresponds to letting $\epsilon \sim N(0, 1)$. Some less common link functions include the *identity link* $g(\mu) = \mu$ and the *complementary log-log (cloglog) link* $g(\mu) = \log\{-\log(1 - \mu)\}$.

4.7.2 Latent variable interpretation

The above quantile function viewpoint provides an interesting interpretation of the link functions for the binomial regression. To illustrate this, suppose $n_i = 1$, $i = 1, \dots, n$. Let

$$Y^* = \eta + \epsilon, \quad \epsilon \sim F(\cdot), \quad Y = 1_{\{Y^* > 0\}},$$

where $F(\cdot)$ is the CDF of some continuous probability distribution. Then the mean value of Y can be given by

$$\mu = \mathbb{E}(Y) = \mathbb{P}(Y^* > 0) = \mathbb{P}(\epsilon > -\eta) = 1 - F(-\eta).$$

Thus, if the distribution is symmetric about 0,

$$\eta = -F^{-1}(1 - \mu) = F^{-1}(\mu).$$

This formulation is quite useful because it allows us to fit a linear model to the latent variable Y^* using just the sign of Y^* , as long as the noise distribution is known.

The cloglog link arises from a similar model in which the latent variable is distributed as

$$Y^* \sim \text{Poisson}(e^\eta).$$

Note that $\mu = e^\eta$ is in fact the canonical link for Poisson regression (see Section 4.8). Suppose the observation is still given by $Y = 1_{\{Y^* > 0\}}$, then

$$1 - \mu = \mathbb{P}(Y = 0) = \mathbb{P}(Y^* = 0) = e^{-e^\eta},$$

which results in the cloglog link $\eta = \log(-\log(1 - \mu))$.

4.7.3 Logistic regression and odds ratio

The logit link (aka the *logistic regression*) is by far the most popular for binomial regression. Beyond the fact that it enjoys some nice properties being the canonical link, it has some other advantages. First, in logistic regression the odds of an observation is given by

$$\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = 0)} = \frac{\mu_i}{1 - \mu_i} = e^{\eta_i} = e^{\mathbf{X}_i^T \boldsymbol{\beta}} = \prod_{j=1}^p \left(e^{\beta_j} \right)^{X_{ij}}.$$

Therefore, e^{β_j} represents a multiplicative change to the odds per unit change of the j th regressor.⁹

Moreover, when we just have a single binary regressor, consider the saturated model

$$\log \frac{\mu}{1 - \mu} = \eta = \beta_0 + \beta_1 X,$$

where $\mu = \mathbb{E}(Y | X) = \mathbb{P}(Y = 1 | X)$. Then the difference in odds ratio for different levels of X is given by

$$\log \frac{\mathbb{P}(Y = 1 | X = 1)}{\mathbb{P}(Y = 0 | X = 1)} - \log \frac{\mathbb{P}(Y = 1 | X = 0)}{\mathbb{P}(Y = 0 | X = 0)} = (\beta_0 + \beta_1) - \beta_0 = \beta_1.$$

Therefore, the *odds ratio* is given by

$$\frac{\mathbb{P}(Y = 1 | X = 1) / \mathbb{P}(Y = 0 | X = 1)}{\mathbb{P}(Y = 1 | X = 0) / \mathbb{P}(Y = 0 | X = 0)} = e^{\beta_1}.$$

The odds ratio is a useful quantity because it enjoys a symmetry:

$$\frac{\mathbb{P}(Y = 1 | X = 1) / \mathbb{P}(Y = 0 | X = 1)}{\mathbb{P}(Y = 1 | X = 0) / \mathbb{P}(Y = 0 | X = 0)} = \frac{\mathbb{P}(X = 1 | Y = 1) / \mathbb{P}(X = 0 | Y = 1)}{\mathbb{P}(X = 1 | Y = 0) / \mathbb{P}(X = 0 | Y = 0)}. \quad (4.11)$$

This neat property implies that we can sample from a population according to Y (suppose $Y = 1$ means a case), and it does not bias the odds ratio. For example, in case-control studies for rare diseases, we can pair each case (e.g. a patient suffering from the disease) with a control (e.g. a healthy individual). This is much more efficient than a random sample from the population, which may contain very few cases. For rare diseases, the odds ratio offers a good approximation to the more interpretable *risk ratio*, defined as $\mathbb{P}(Y = 1 | X = 1) / \mathbb{P}(Y = 1 | X = 0)$, because $\mathbb{P}(Y = 0)$ is very close to 1.

4.8 Poisson regression

4.8.1 Models for count data

Poisson regression is used to model count data: $Y_i \in \{0, 1, 2, \dots\}$, $i = 1, \dots, n$. It is common to model counts by a Poisson distribution, $Y_i \sim \text{Poisson}(\mu_i)$. One rationale for this is the following *law of small numbers*. Consider a triangular array $\{\mu_{n,j} > 0 \mid 1 \leq j \leq$

$n\}$ such that $\sum_{j=1}^n \mu_{n,j} = \mu$. Then under the assumption that $\max_j \mu_{n,j} \rightarrow 0$ as $n \rightarrow \infty$, we have

$$\sum_{j=1}^n \text{Bernoulli}(\mu_{n,j}) \rightarrow \text{Poisson}(\mu) \text{ as } n \rightarrow \infty.$$

In words, if Y is the total count of many small probability events, then Y approximately follows a Poisson distribution.

The Poisson distribution $Y \sim \text{Poisson}(\mu)$ has the mean-variance relation (Example 3.8)

$$\text{Var}(Y) = \mathbb{E}(Y) = \mu.$$

In practice, the data are sometimes overdispersed compared to the theoretical relationship above due to clustering or other reasons (Section 4.3.3).

4.8.2 *Variance stabilizing transform

To deal with overdispersion, one can use the *variance stabilizing transform* that maps Y to $g(Y)$. By the delta method, $\text{Var}(g(Y)) \approx \{g'(\mu)\}^2 \text{Var}(Y)$. Thus, if we take

$$g'(\mu) = \frac{1}{\sqrt{\text{Var}(Y)}},$$

then $\text{Var}(g(Y)) \approx 1$. For Poisson, this is $g(Y) = 2\sqrt{Y}$. We can then fit a linear model for $\mathbb{E}(2\sqrt{Y} \mid X)$ and use the linear model noise variance to probe overdispersion. The drawback of this approach is that \sqrt{Y} might not be the scale we would like to investigate. We can also use a dispersion parameter σ^2 in the (quasi-)Poisson GLM to model overdispersion, which will be discussed in more detail next.¹⁰

4.8.3 Poisson regression

Recall that the probability mass function of $\text{Poisson}(\mu_i)$ is given by

$$f(y_i; \mu_i) = e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!} = e^{y_i \log \mu_i - \mu_i} \frac{1}{y_i!}, \quad y_i = 0, 1, \dots$$

With the dispersion parameter σ^2 included, the probability mass function becomes

$$f(y_i; \mu_i, \sigma^2) = e^{\frac{1}{\sigma^2} \{y_i \log \mu_i - \mu_i\}} f_0(y_i; \sigma^2).$$

So the natural parameter is $\theta_i = \log(\mu_i)$ and the cumulant function is $K(\theta) = e^\theta$.

In Poisson regression, the most common choice of the link function is the canonical log link $g(\mu) = \theta(\mu) = \log(\mu)$, so the model is

$$\log \mu_i = \mathbf{X}_i^T \boldsymbol{\beta}.$$

This is often referred to as the *log-linear model*. This model is straightforward to interpret, as

$$\mu_i = e^{\mathbf{X}_i^T \boldsymbol{\beta}} = \prod_{j=1}^p (e^{\beta_j})^{X_{ij}},$$

So e^{β_j} represents a multiplicative change to the predicted mean value per unit change of the j th regressor.

Other common link functions for Poisson regression including the identity link and the square root link.¹¹ Notice that the square root link assumes that $\sqrt{\mathbb{E}(Y_i | \mathbf{X}_i)} = \mathbf{X}_i^T \boldsymbol{\beta}$, which is different from fitting a linear model after the square root variance stabilizing transform which assumes that $\mathbb{E}(\sqrt{Y_i} | \mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\beta}$.

The deviance in Poisson regression ($\sigma^2 = 1$) is given by (Exercise 3.17)

$$D(Y_i, \hat{\mu}_i) = 2 \left\{ Y_i \log \frac{Y_i}{\hat{\mu}_i} - Y_i + \hat{\mu}_i \right\}.$$

If \mathbf{X} includes intercept (a column $\mathbf{1}$) and the canonical log link is used, the score equation $\mathbf{X}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}$ implies that

$$\sum_{i=1}^n \hat{\mu}_i = \sum_{i=1}^n Y_i.$$

Therefore, by letting $\delta_i = Y_i - \hat{\mu}_i$ and assuming $|\delta_i| \ll \hat{\mu}_i$, the total deviance for the Poisson regression is approximately given by

$$\begin{aligned} D^{(n)}(\mathbf{Y}, \hat{\boldsymbol{\mu}}) &= 2 \sum_{i=1}^n Y_i \log \frac{Y_i}{\hat{\mu}_i} \\ &= 2 \sum_{i=1}^n (\hat{\mu}_i + \delta_i) \log \left(1 + \frac{\delta_i}{\hat{\mu}_i} \right) \\ &\approx 2 \sum_{i=1}^n (\hat{\mu}_i + \delta_i) \left(\frac{\delta_i}{\hat{\mu}_i} - \frac{1}{2} \frac{\delta_i^2}{\hat{\mu}_i^2} \right) \\ &\approx 2 \sum_{i=1}^n \delta_i + \frac{1}{2} \frac{\delta_i^2}{\hat{\mu}_i} \\ &= \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \end{aligned}$$

The last expression is precisely the Pearson χ^2 -statistic from *IB Statistics*:

$$\chi^2 = \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}}.$$

For Poisson regression, Pearson's residual is given by

$$R_{P,i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} = \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i},$$

so Pearson's χ^2 -statistic is given by $\chi^2 = \sum_{i=1}^n R_{P,i}^2$ and converges to χ_{n-p}^2 if the Poisson regression is correctly specified. Note that this convergence does not require n to converge to infinity; in fact, convergence to χ_{n-p}^2 would be ill-defined if n increases and p is fixed. The crucial assumption is that $\min_i \mu_i \rightarrow \infty$ (which can be seen from the assumption that $\delta_i \ll \hat{\mu}_i$). This is the so-called *small dispersion asymptotics*.

4.8.4 Multinomial models and the Poisson trick

Poisson regression can also be used to analyze multinomial data. Suppose $(Y_1, \dots, Y_L) \sim \text{Multinomial}(n, \boldsymbol{\pi})$, where n is known but $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)$ is unknown. The probability mass function is given by

$$f(\mathbf{y}; \boldsymbol{\pi}) = \frac{n!}{y_1! \dots y_L!} \pi_1^{y_1} \dots \pi_L^{y_L}, \text{ for } \sum_{i=1}^L Y_i = n.$$

This is not a minimal exponential family because of the constraint on \mathbf{Y} , which also implies that Y_1, \dots, Y_L are not independent. A potential solution is to set one level as the reference and obtain a $(L-1)$ -parameter exponential family (Example 3.5). However, the symmetry in the parameters is broken.

A more elegant solution is the *Poisson trick*, which refers to the following result in probability. Suppose $Y_i \sim \text{Poisson}(\mu_i)$, $i = 1, \dots, L$ independently. Let $Y_+ = \sum_{i=1}^L Y_i$, then

$$Y_+ \sim \text{Poisson}(\mu_+) \quad \text{and} \quad Y_1, \dots, Y_L \mid Y_+ \sim \text{Multinomial}(Y_+; \boldsymbol{\pi}),$$

where $\pi_i = \mu_i / \mu_+$, $i = 1, \dots, L$ and $\mu_+ = \sum_{i=1}^L \mu_i$.¹²

Exercise 4.7. Verify the Poisson trick.

Consider the log-linear Poisson model

$$Y_i \sim \text{Poisson}(\mu_i) \text{ independently, and } \log \mu_i = \alpha + \mathbf{X}_i^T \boldsymbol{\beta},$$

where the intercept α is distinguished from the rest of the coefficients. Then by the Poisson trick,

$$Y_+ = \sum_{i=1}^L Y_i \sim \text{Poisson}(\mu_+) \quad \text{and} \quad \mathbf{Y} \mid Y_+ \sim \text{Multinomial}(Y_+, \boldsymbol{\pi}),$$

where

$$\begin{aligned} \mu_+ &= \sum_{i=1}^n \mu_i = e^\alpha \sum_{i=1}^n e^{\mathbf{X}_i^T \boldsymbol{\beta}}, \text{ and} \\ \pi_i &= \frac{\mu_i}{\mu_+} = \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}}}{\sum_{i=1}^n e^{\mathbf{X}_i^T \boldsymbol{\beta}}}, \quad i = 1, \dots, n. \end{aligned} \tag{4.12}$$

Importantly, $\boldsymbol{\pi}$ does not depend on the intercept α in the Poisson model. In consequence, the likelihood function for the Poisson model factorizes as

$$\begin{aligned} L_P(\alpha, \boldsymbol{\beta}) &= \prod_{i=1}^n f(Y_i; \mu_i) \\ &= f(Y_1, \dots, Y_L \mid Y_+; \boldsymbol{\beta}) f(Y_+; \alpha, \boldsymbol{\beta}) \\ &= L_M(\boldsymbol{\beta}) f(Y_+; \alpha, \boldsymbol{\beta}), \end{aligned}$$

where $L_M(\boldsymbol{\beta})$ denotes the likelihood for the multinomial model (4.12). Alternatively, because given $\boldsymbol{\beta}$, μ_+ is uniquely determined by α and vice versa, we can write this more concisely as

$$L_P(\mu_+, \boldsymbol{\beta}) = L_M(\boldsymbol{\beta})f(Y_+; \mu_+).$$

The above likelihood factorization implies that we can fit the multinomial model (4.12) using the Poisson log-linear model with an intercept, and the likelihood inference for $\boldsymbol{\beta}$ in the two models will be equivalent. To see this, the MLE $\hat{\boldsymbol{\beta}}$ for the Poisson likelihood $L_P(\mu_+, \boldsymbol{\beta})$ will also maximize the multinomial likelihood $L_M(\boldsymbol{\beta})$. The Fisher information matrix for the Poisson model is block-diagonal:

$$\mathbf{I}^{(n)}(\mu_+, \boldsymbol{\beta}) = \begin{pmatrix} I_{\mu_+ \mu_+}^{(n)}(\mu_+) & \mathbf{0}^T \\ \mathbf{0} & \mathbf{I}_{\boldsymbol{\beta} \boldsymbol{\beta}}^{(n)}(\boldsymbol{\beta}) \end{pmatrix}.$$

Deviance in the Poisson model is the same as deviance in the multinomial model, because $\hat{\mu}_+ = Y_+$.

4.9 Contingency tables

Next we apply the Poisson and multinomial models to analyze contingency tables that display empirical frequencies of random variables.

4.9.1 Two-way contingency tables

Example 4.8. The following contingency table was constructed from a interim release of a Phase-III trial for the Moderna COVID-19 vaccine in November, 2020.¹³ The *

	Not a case	Non-severe case	Severe case
Placebo	*	79	11
Vaccine	*	5	0

cells were not reported, but they are presumably very large because the total number of participants was about 30,000. The press release claims that the vaccine efficacy is about $1 - (5 + 0)/(79 + 11) = 94.5\%$ and the p -value (for no efficacy) is less than 0.0001.

There are two ways to think about the data in contingency tables:

- We observe counts Y_{jk} , $j = 1, \dots, J$, $k = 1, \dots, K$. In the previous example, $J = 2$ and $K = 3$.
- The table is an aggregation of individual observations (A_i, B_i) , $i = 1, \dots, n$. In the previous example, A_i is the treatment received (placebo or vaccine), B_i is the outcome (not a case, non-severe case, or severe case), and $n \approx 30,000$. The observed counts are given by

$$Y_{jk} = \sum_{i=1}^n 1_{\{A_i=j, B_i=k\}}, \quad j = 1, \dots, J, \quad k = 1, \dots, K.$$

A common question in two-way contingency tables is testing the null hypothesis that the rows and columns are independent, $H_0 : A_i \perp B_i$. In the vaccine trial example, this amounts to testing the hypothesis that the vaccine has no effect at all.

Suppose (A_i, B_i) , $i = 1, \dots, n$ are IID. Let $\pi_{jk} = \mathbb{P}(A_i = j, B_i = k)$, $j = 1, \dots, J, k = 1, \dots, K$, so the counts follow a multinomial distribution $\mathbf{Y} \sim \text{Multinomial}(n, \boldsymbol{\pi})$. The null hypothesis can be expressed in terms of $\boldsymbol{\pi}$ as

$$H_0 : \pi_{jk} = \pi_j^A \pi_k^B, \text{ for all } j, k,$$

where $\pi_j^A = \sum_{k=1}^K \pi_{jk}$ and $\pi_k^B = \sum_{j=1}^J \pi_{jk}$ are the marginal distributions of A and B . In the surrogate Poisson model, this can be expressed as

$$H_0 : \mu_{jk} = \mu_+ \pi_j^A \pi_k^B \text{ for all } j, k,$$

which is equivalent to the log-linear model

$$H_0 : \log \mu_{jk} = \alpha + \beta_j^A + \beta_k^B, \text{ for all } j, k, \quad (4.13)$$

This is a submodel of the saturated model that places no restrictions on μ_{jk} :

$$H_1 : \log \mu_{jk} = \alpha + \beta_j^A + \beta_k^B + \beta_{jk}^{AB}, \text{ for all } j, k. \quad (4.14)$$

Therefore, testing independence in contingency tables is equivalent to testing nested models in Poisson regression.

Notice that (4.13) and (4.14) are overparametrized. For identifiability, it is necessary to set some levels as the reference. For example, we may set $\beta_1^A = \beta_1^B = \beta_{1k}^{AB} = \beta_{j1}^{AB} = 0$ for all j, k .

Example 4.9. For a 2×2 table ($J = K = 2$), the null/independence log-linear model assumes

$$\log \boldsymbol{\mu} = \underbrace{\begin{pmatrix} \log \mu_{11} \\ \log \mu_{12} \\ \log \mu_{21} \\ \log \mu_{22} \end{pmatrix}}_{\mathbf{X}_0} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}}_{\mathbf{X}_0} \begin{pmatrix} \alpha \\ \beta_2^A \\ \beta_2^B \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha + \beta_2^B \\ \alpha + \beta_2^A \\ \alpha + \beta_2^A + \beta_2^B \end{pmatrix},$$

and the saturated log-linear model assumes

$$\log \boldsymbol{\mu} = \underbrace{\begin{pmatrix} \log \mu_{11} \\ \log \mu_{12} \\ \log \mu_{21} \\ \log \mu_{22} \end{pmatrix}}_{\mathbf{X}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}}_{\mathbf{X}} \begin{pmatrix} \alpha \\ \beta_2^A \\ \beta_2^B \\ \beta_{22}^{AB} \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha + \beta_2^B \\ \alpha + \beta_2^A \\ \alpha + \beta_2^A + \beta_2^B + \beta_{22}^{AB} \end{pmatrix}.$$

The degrees of freedom of the sub/independence model is $1 + (J-1) + (K-1) = J + K - 1$ and the degrees of freedom of the saturated model is JK . By Wilks' theorem, under H_0 and as $n \rightarrow \infty$, the deviance between the two models or equivalently Pearson's χ^2 -statistic converges in distribution to $\chi_{JK-(J+K-1)}^2 = \chi_{(J-1)(K-1)}^2$. This provides an asymptotic test for the independence hypothesis.

4.9.2 Three-way contingency tables

The discussion above can be extended to three-way contingency tables, although there are more independence and conditional independence hypotheses that can be tested. Individually, the observations come as IID triplets $(A_i, B_i, C_i), i = 1, \dots, n$, which can be aggregated by a three-way table:

$$Y_{jkl} = \sum_{i=1}^n 1_{\{A_i=j, B_i=k, C_i=l\}}, \quad j = 1, \dots, J, \quad k = 1, \dots, K, \quad l = 1, \dots, L.$$

There are several possible models for the joint probability mass $\pi_{jkl} = \mathbb{P}(A_i = j, B_i = k, C_i = l)$. The first model assumes

$$H_1 : \pi_{jkl} = \pi_j^A \pi_k^B \pi_l^C \text{ for all } j, k, l,$$

where $\pi_j^A, \pi_k^B, \pi_l^C$ are the corresponding marginal probabilities (similar conventions are used below). This is equivalent to assuming

$$H_1 : A_i \perp\!\!\!\perp B_i \perp\!\!\!\perp C_i.$$

The second model assumes

$$H_2 : \pi_{jkl} = \pi_j^A \pi_{kl}^{BC} \text{ for all } j, k, l,$$

which amounts to the independence

$$H_2 : A_i \perp\!\!\!\perp (B_i, C_i).$$

The third model assumes

$$H_3 : \pi_{jkl} = \pi_{jk}^{AB} \pi_{kl}^{BC} \text{ for all } j, k, l.$$

It can be shown that this implies

$$\mathbb{P}(A_i = j, C_i = l \mid B_i = k) = \mathbb{P}(A_i = j \mid B_i = k) \mathbb{P}(C_i = l \mid B_i = k). \quad (4.15)$$

So this model amounts to the conditional independence

$$H_3 : A_i \perp\!\!\!\perp C_i \mid B_i.$$

Exercise 4.10. Verify (4.15).

The fourth model assumes

$$H_4 : \pi_{jkl} = \pi_{jk}^{AB} \pi_{kl}^{BC} \pi_{jl}^{AC}.$$

This model does not imply any independence or conditional independence, but it assumes that there is no three-way interaction in the joint distribution.

Finally, the fifth and saturated model assumes

$$H_5 : \pi_{jkl} = \pi_{jkl}^{ABC},$$

where π_{jkl}^{ABC} is completely unrestricted besides the constraints that the marginals need to sum up to 1. Of course, this model also makes no independence or conditional independence assumptions.

The five models above for the three-way contingency table are nested and can be tested using the deviance or Pearson's χ^2 of the corresponding surrogate Poisson models. These Poisson log-linear models differ in whether certain two-way and three-way interaction terms are included.

4.9.3 *Graphical models (not covered this year)

With more variables, it is more convenient to represent independence and conditional independence relationship using a graph.

Consider an undirected graph $(\mathbf{V}, \mathcal{E})$ where $\mathbf{V} = (V_1, \dots, V_p)$ is a discrete random vector and $E \subseteq \{V_1, \dots, V_p\}^2$ is the edge set. We say the distribution of \mathbf{V} *factorizes according to this graph* if the probability mass function of \mathbf{V} can be written as

$$\mathbb{P}(\mathbf{V} = \mathbf{v}) = \prod_{\substack{\mathcal{C} \subseteq \{V_1, \dots, V_p\} \\ (A_1, A_2) \in \mathcal{E} \text{ for all } A_1, A_2 \in \mathcal{C}}} \pi^{\mathcal{C}}(\mathbf{v}_{\mathcal{C}}).$$

Such subset of vertices \mathcal{C} is called a *complete subgraph* or *clique*. Thus, graphical factorization means that the distribution can be decomposed according to the cliques in the graph.

See Figure 4.3 for the graphical models corresponding to the five models for the three-way contingency table. There is a deep connection between graph theory and conditional independence: in an undirected graphical model, if the probability distribution factorizes according to the graph and a subset of variables \mathbf{B} “blocks” all paths between two other non-overlapping subsets \mathbf{A} and \mathbf{C} , then $\mathbf{A} \perp\!\!\!\perp \mathbf{C} \mid \mathbf{B}$.¹⁴

Notes

¹See https://en.wikipedia.org/wiki/Log-normal_distribution#Occurrence_and_applications for more examples.

²See https://en.wikipedia.org/wiki/Pareto_distribution#Occurrence_and_applications for more examples.

³This is indeed what `summary.glm` in R does.

⁴With non-canonical link functions, one can still carry out an analysis of deviance by Wilks' theorem, but the additive relationship (4.6) no longer holds.

⁵The GLM literature is not always consistent with the terminology and notation related to the deviance. In (4.8), the deviance $D^{(n)}(\cdot, \cdot)$ is computed in the original exponential family without overdispersion. Some texts skip the geometric intuition presented here and directly defines the total deviance as the difference in the log-likelihood.

⁶Taken from Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons, Figure 4.2.

⁷In machine learning, this technique is known as the *natural gradient* method.

⁸We may need $g(\mu)$ to be sufficiently smooth (e.g. twice differentiable) for the asymptotic theory to go through.

⁹This is not necessarily a causal effect. See Section 2.4.4.

¹⁰One can also use GLMs with other discrete distributions such as the negative binomial. However, this is beyond the scope of this course.

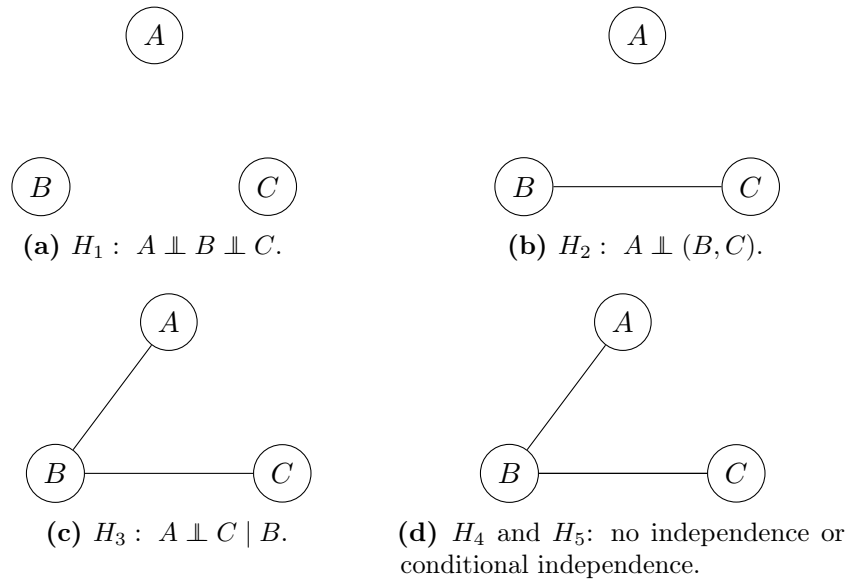


Figure 4.3 Graphical models for three-way contingency tables.

¹¹See `?family` in R.

¹²This is in fact a special instance of a more general result for exponential families. See Brown, L. D. (1986). *Fundamentals of statistical exponential families: With applications in statistical decision theory*. Institute of Mathematical Statistics, Theorem 1.15.

¹³<https://investors.modernatx.com/news-releases/news-release-details/modernas-covid-19-vaccine-candidate-meets-its-primary-efficacy>.

¹⁴This is one direction of the famous Hammersley-Clifford theorem.

Chapter 5

Review and look forward

5.1 Review

Table 5.1 provides a concise summary of the main definitions and results in this course. Some other topics we covered are reviewed below.

Model diagnostics

For linear models:

- Leverage of i th observation: H_{ii} , where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is the hat matrix.
- Coefficient of determination/Variance explained: R^2 .
- Check normality: Q-Q plot using standardized residuals

$$\frac{Y_i - \hat{\mu}_i}{\hat{\sigma} \sqrt{1 - H_{ii}}}.$$

- Check nonlinearity: residual vs. fitted plot.
- An observation with a large residual is called an outlier, which is particularly concerning if the leverage is also large. Check by the residual vs. leverage plot. Cook's distance is another useful diagnostics:

$$D_i = \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})\|^2}{p \hat{\sigma}^2} = \frac{1}{p} \frac{H_{ii}}{1 - H_{ii}} \tilde{R}_i^2.$$

- Check heteroskedasticity: residual scale vs. fitted plot.

For generalized linear models, diagnostics are same as above with some minor modifications:

- Replace the hat matrix by $\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2}$ obtained from iteratively reweighted least squares.

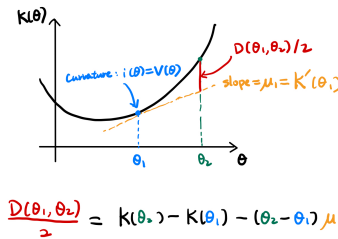
	Distribution of Y	MLE & Geometry	Statistical inference
Chapter 2 Linear model	1. Normal LM: $\mathbf{Y} \mid \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$; 2. Nonparametric relaxation: $Y_i = g(\mathbf{X}_i) + \epsilon_i, \epsilon_i \perp \mathbf{X}_i$.	1. Euclidean: $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \ \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\ ^2$; 2. Nested projections: $\mathbf{X} = (\mathbf{X}_0 \ \mathbf{X}_1) \Rightarrow \mathbf{P}_{\mathbf{X}} \mathbf{P}_{\mathbf{X}_0} = \mathbf{P}_{\mathbf{X}_0}$; 3. Partial regression: $\hat{\beta}_j = \text{lm}(\text{resid}(Y \sim \mathbf{X}_{-j}) \sim \text{resid}(X_j \sim \mathbf{X}_{-j}))$.	1. $\hat{\sigma}^2 = \ \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\ ^2 / (n - p)$; 2. With normality, use pivot $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / \hat{\sigma}$; 3. Without normality, establish $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \text{Normal}$; 4. Nested models: Use ANOVA/LRT.
Chapter 3 Exponential family	1. $Y \sim f(y; \theta) = e^{y\theta - K(\theta)} f_0(y)$; 2. Mean-value parametrization: $\mu = \mu(\theta) = K'(\theta)$; 3. Variance $V(\theta) = \mu'(\theta) = K''(\theta)$.	1. MLE: $\hat{\mu} = \bar{Y}$; 2. Deviance extends Euclidean distance: 	1. Fisher information: $i^{(n)} = nV(\theta), i^{(n)}(\mu) = n/V(\mu)$; 2. Central limit theorem: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, 1/V(\theta))$. 3. LRT is uniformly most powerful.
Chapter 4 Generalized linear model	1. $Y_i \mid \mathbf{X}_i \sim e^{\{y_i \theta_i - K(\theta_i)\} / \sigma_i^2} f_0(y_i; \sigma_i^2)$; 2. Linkage $g(\mu_i) = \eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$; 3. Canonical link: $\eta_i = \theta_i$; 4. $\sigma_i^2 = \underbrace{w_i}_{\text{known}} \underbrace{\sigma^2}_{\text{over/under-dispersion}}$	1. MLE: $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} D_+(\mathbf{Y}, \boldsymbol{\mu}(\boldsymbol{\beta}))$. 2. Score equation: $\sum_{i=1}^n \frac{(Y_i - \mu_i) \mathbf{X}_i}{\text{Var}(Y_i) g'(\mu_i)} = \mathbf{0}$; (Canonical form: $\mathbf{X}^T \{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\} = \mathbf{0}$.) 3. Deviance additivity (for canonical link): $D_+(\mathbf{Y}, \hat{\boldsymbol{\mu}}_0) = D_+(\mathbf{Y}, \hat{\boldsymbol{\mu}}) + D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_0)$.	1. Fisher information: $\mathbf{I}(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \mathbf{X}^T \mathbf{W} \mathbf{X} / \sigma^2 & 0 \\ 0 & * \end{pmatrix}$, where $\mathbf{W} = \text{diag}(\mathbf{w}_i / [V(\mu_i) \{g'(\mu_i)\}^2])$; 2. Asymptotic distribution: $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1})$; 3. $\hat{\sigma}^2 = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 / \{(n - p) V(\hat{\mu}_i)\}$; 4. Nested models: Analysis of deviance/LRT $D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_0) \xrightarrow{d} \chi_{p-p_0}^2$.

Table 5.1 A concise summary of this course.

- Replace residual with one of the following (and standardize by dividing $\sqrt{1 - H_{ii}}$ as before):

- Pearson's residual:

$$\frac{Y_i \hat{\mu}_i}{\hat{\sigma} \sqrt{V(\hat{\mu}_i) / w_i}};$$

- Deviance residual:

$$\text{sign}(Y_i - \hat{\mu}_i) \sqrt{D(Y_i, \hat{\mu}_i)} / \sqrt{\hat{\sigma}^2 / w_i}.$$

Model selection

Why model selection?

- Select a simpler and more interpretable model.
- Better bias-variance tradeoff.

See Table 5.2 for a summary of model selection criteria.

Linear model	Generalized linear model
1. $C_p = \ \mathbf{Y} - \hat{\boldsymbol{\mu}}\ ^2 + 2 \cdot \text{df} \cdot \sigma^2$ is an unbiased estimator of mean squared prediction error.	Unavailable
2. Cross-validation: $\text{CV} = \sum_{i=1}^n (Y_i - \hat{\mu}_{-i})^2 = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 / (1 - H_{ii})^2$.	Replace squared error with deviance.
3. $\text{AIC} = -2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + 2\text{df}$.	Same.
4. $\text{BIC} = -2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + \text{df} \cdot \log n$.	Same.

Table 5.2 Criteria for model selection.

Algorithmically, model selection or regularization can be achieved by the best subset method, greedy (forward or backward) stepwise selection method, or adding penalty terms to the likelihood function. Ignoring model selection may lead to biased post-selection inference.

Asymptotic theory for likelihood inference

- The log-likelihood function is given by $l(\boldsymbol{\beta}; \mathbf{Y}) = \sum_{i=1}^n \log f(Y_i | \mathbf{X}_i; \boldsymbol{\beta}_i)$.
- Score function: $U(\boldsymbol{\beta}; \mathbf{Y}) = \nabla l(\boldsymbol{\beta}; \mathbf{Y})$.
- Fisher information: $\mathbf{I}(\boldsymbol{\beta}) = \text{Var}_{\boldsymbol{\beta}}(U(\boldsymbol{\beta}; \mathbf{Y})) = \mathbb{E}_{\boldsymbol{\beta}}\{-\nabla^2 l(\boldsymbol{\beta}; \mathbf{Y})\}$.

- MLE: $\hat{\beta} = \arg \max_{\beta} l(\beta; \mathbf{Y})$.
- Under regularity conditions, $\hat{\beta} \sim N(\beta, \mathbf{I}(\beta)^{-1})$. This leads to asymptotic confidence intervals and hypothesis tests.
- LRT: Suppose $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ and interested in testing $H_0 : \beta_1 = \mathbf{0}$ vs. $H_1 : \beta_1 \neq \mathbf{0}$.

Wilks' theorem:

$$2 \left\{ \sup_{H_0 \cup H_1} l(\beta; \mathbf{Y}) - \sup_{H_0} l(\beta; \mathbf{Y}) \right\} \xrightarrow{d} \chi_{\dim(\beta_1)}^2.$$

Other topics

- Heteroskedasticity-robust standard error for linear models.
- Bias-variance tradeoff in linear models.
- Simpson's paradox.
- Box-Cox transformation.
- Computation for GLMs: Newton-Raphson, Fisher scoring, and iteratively reweighted least squares.
- Binomial regression: common link functions and latent variable interpretations.
- Poisson log-linear regression, multinomial model, and the Poisson trick.
- Contingency tables: parametrizing Poisson regression for independence testing.

5.2 Look forward

- Mixed effect models: Assume some elements of (high-dimensional) β are random. (Part III, *Statistical Learning in Practice*.)
- Generalized additive models and kernel regression: Replace $\mathbf{X}_i^T \beta$ by some basis function expansion. (Part III, *Modern Statistical Methods*.)
- Trees, random forests, and boosting: Replace $\mathbf{X}_i^T \beta$ by a (complicated) step function. (Part II, *Mathematics of Machine Learning*; Part III, *Statistical Learning in Practice*.)
- Neural networks: Replace $\mathbf{X}_i^T \beta$ by a composition of GLMs. (Part II, *Mathematics of Machine Learning*; Part III, *Statistical Learning in Practice*.)
- Regularization. (Part III, *Modern Statistical Methods*.)
- Graphical models. (Part III, *Bayesian Statistics, Causal Inference*.)
- Distinguishing correlation from causation. (Part III, *Causal Inference*.)