Chapter 5

Review and look forward (not covered this year)

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 ith observation: H_{ii}, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ 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:

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

- 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 $\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}$ obtained from iteratively reweighted least squares.

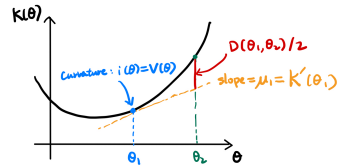
	Distribution of Y	MLE & Geometry	Statistical inference
Chapter 2 Linear model	1. Normal LM: $\boldsymbol{Y} \mid \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$; 2. Nonparametric relaxation: $Y_i = g(\boldsymbol{X}_i) + \epsilon_i,\ \epsilon_i \perp\!\!\!\perp \boldsymbol{X}_i$.	1. Euclidean: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$; 2. Nested projections: $\boldsymbol{X} = (\boldsymbol{X}_0\ \boldsymbol{X}_1) \Rightarrow \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}_0} = \boldsymbol{P}_{\boldsymbol{X}_0}$; 3. Partial regression: $\hat{\beta}_j = \texttt{lm}(\texttt{resid}(Y \sim \boldsymbol{X}_{-j}) \sim \texttt{resid}(X_j \sim \boldsymbol{X}_{-j}))$.	1. $\hat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)$; 2. With normality, use pivor $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\hat{\sigma}$; 3. Without normality, establish $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d}$ 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:  $\frac{D(\theta_1,\theta_2)}{2} = K(\theta_2) - K(\theta_1) - (\theta_2 - \theta_1)\mu_1$	1. Fisher information: $i^{(n)} = nV(\theta),\ i^{(n)}(\mu) = n/V(\mu)$; 2. Central limit theorem: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0, 1/V(\theta))$. 3. LRT is uniformly most powerful.
Chapter 4 Generalized linear model	1. $Y_i \mid \boldsymbol{X}_i \sim e^{\{y_i\theta_i - K(\theta_i)\}/\sigma_i^2} f_0(y_i;\sigma_i^2)$; 2. Linkage $g(\mu_i) = \eta_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$; 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: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} D_+(\boldsymbol{Y}, \boldsymbol{\mu}(\boldsymbol{\beta}))$. 2. Score equation: $\sum_{i=1}^n \frac{(Y_i - \mu_i)\boldsymbol{X}_i}{\mathrm{Var}(Y_i)g'(\mu_i)} = 0$; (Canonical form: $\boldsymbol{X}^T\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\} = \boldsymbol{0}$.) 3. Deviance additivity (for canonical link): $D_+(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}_0) = D_+(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}) + D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_0)$.	1. Fisher information: $\boldsymbol{I}(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}/\sigma^2 & 0 \\ 0 & * \end{pmatrix}$, where $\boldsymbol{W} = \mathrm{diag}(w_i V(\mu_i)\{g'(\mu_i)\}^2)^{-1}$; 2. Asymptotic distribution: $\hat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1})$; 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 $D_+(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}) - D_+(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}_0) \xrightarrow{d} \chi^2_{p-p_0}$.

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:
$$\frac{Y_i\hat{\mu}_i}{\hat{\sigma}\sqrt{w_iV(\hat{\mu}_i)}};$$
  - Deviance residual:
$$\text{sign}(Y_i-\hat{\mu}_i)\sqrt{D(Y_i,\hat{\mu}_i)}.$$

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. $C_p = \|\boldsymbol{Y}-\hat{\boldsymbol{\mu}}\|^2 + 2\cdot\text{df}\cdot\sigma^2$ is an unbiased estimator of mean squared prediction error.	Unavailable
2. Cross-validation: $\text{CV} = \sum_{i=1}^n (Y_i - \hat{\mu}_{-i})^2 = \sum_{i=1}^n \frac{(Y_i-\hat{\mu}_i)^2}{(1-H_{ii})^2}$.	Replace squared error with deviance.
3. $\text{AIC} = -2l(\hat{\boldsymbol{\beta}},\hat{\sigma}^2) + 2\text{df}$.	Same.
4. $\text{BIC} = -2l(\hat{\boldsymbol{\beta}},\hat{\sigma}^2) + \text{df}\log n$.	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 $l(\boldsymbol{\beta};\boldsymbol{Y}) = \sum_{i=1}^n \log f(Y_i \mid \boldsymbol{X}_i;\boldsymbol{\beta}_i)$.
- Score function: $U(\boldsymbol{\beta};\boldsymbol{Y}) = \nabla l(\boldsymbol{\beta};\boldsymbol{Y})$.
- Fisher information: $\boldsymbol{I}(\boldsymbol{\beta}) = \text{Var}_{\boldsymbol{\beta}}(U(\boldsymbol{\beta};\boldsymbol{Y})) = \mathbb{E}_{\boldsymbol{\beta}}\{-\nabla^2 l(\boldsymbol{\beta};\boldsymbol{Y})\}$.

- MLE: $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}; \boldsymbol{Y})$.
- Under regularity conditions, $\hat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{I}(\boldsymbol{\beta})^{-1})$. This leads to asymptotic confidence intervals and hypothesis tests.
- LRT: Suppose $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}$ and interested in testing $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ vs. $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$. Wilks' theorem:

$$2\left\{\sup_{H_0 \cup H_1} l(\boldsymbol{\beta}; \boldsymbol{Y}) - \sup_{H_0} l(\boldsymbol{\beta}; \boldsymbol{Y})\right\} \xrightarrow{d} \chi^2_{\dim(\boldsymbol{\theta}_1)}.$$

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) $\boldsymbol{\beta}$ are random. (Part III, *Statistical Learning in Practice*.)
- Generalized additive models and kernel regression: Replace $\boldsymbol{X}_i^T \boldsymbol{\beta}$ by some basis function expansion. (Part III, *Modern Statistical Methods*.)
- Trees, random forests, and boosting: Replace $\boldsymbol{X}_i^T \boldsymbol{\beta}$ by a (complicated) step function. (Part II, *Mathematics of Machine Learning*; Part III, *Statistical Learning in Practice*.)
- Neural networks: Replace $\boldsymbol{X}_i^T \boldsymbol{\beta}$ 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*.)