Statistical Modelling—Additional Information

Qingyuan Zhao

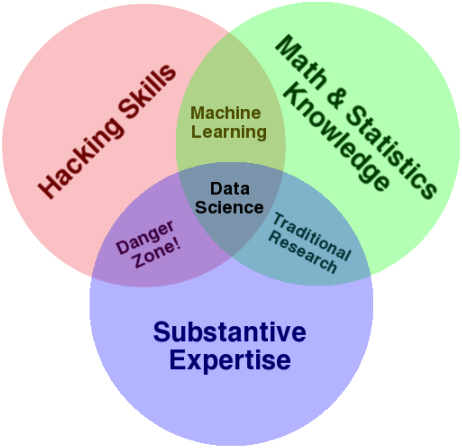
December 4, 2024

Outline

Logistics

- Most information about the course can be found on this webpage (not Moodle): `https://www.statslab.cam.ac.uk/~qz280/teaching/modelling-2024/`.
- Prerequisite: *IB Statistics*.
- This course complements *Principles of Statistics* by providing a more applied/computational perspective.
- If you have any need reasonable adjustments, please let me know.

Data science Venn diagram



(credit: Drew Conway)

Practical sessions are ESSENTIAL

It is nearly impossible to learn (applied) statistics well without hands-on experience of data analysis. So please come to and engage in the practical sessions (led by Louis Christie).

- Please read over the sheet and try the R code before each practical.
- First 10 minutes will discuss questions about the sheet.
- Rest of the session will go over the exercises (in small groups).
- Make sure R and R Studio are correctly installed before the first session (`https://posit.co/download/rstudio-desktop/`).

Big picture

- Statistics is not mathematics but uses mathematics (deduction) to justify induction:

  In statistical inference, as distinct from mathematical inference, there is a world of difference between the two statements "p is true" and "p is known to be true".

- Models guide the cycle of scientific research: conjecture $\rightarrow$ design $\rightarrow$ experiment $\rightarrow$ analysis $\rightarrow$ conjecture $\rightarrow \cdots$.
- Statistical models have different levels:
  - (conditional) expectation;
  - (conditional) distribution;
  - causal.
- Why regression?
- Why (generalized) linear models?

Lecture 1

- Notation.
- Normal linear model: Definition.
- Normal linear model: Maximum likelihood estimator.
- Geometry of ordinary least squares.
- Partial regression.

Lecture 2

- Exact inference for normal linear model.
- Demo of `lm` and partial regression.

```
?lm
?LifeCycleSavings
fit1 <- lm(sr ~ pop15 + pop75 + ddpi, data = LifeCycleSavings)
summary(fit1)
attach(LifeCycleSavings)
fit2 <- lm(residuals(lm(sr ~ pop15 + pop75)) ~
           residuals(lm(ddpi ~ pop15 + pop75)))
detach(LifeCycleSavings)
summary(fit2)
```

Lecture 3

- Generalized least squares.
- Heteroscedasticity and sandwich variance.
- Projection interpretation of regression coefficients.
- Demo of projection interpretation.

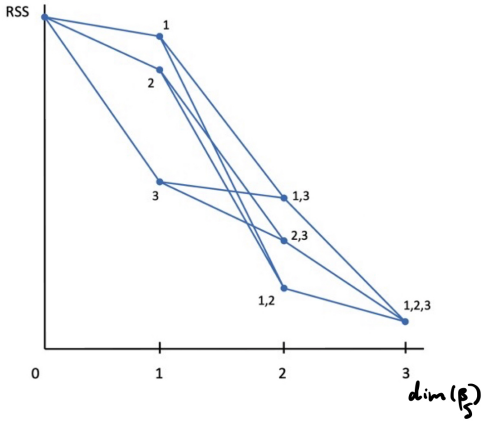
```
n <- 100
X1 <- rnorm(n, 1, 1)
Y1 <- X1^2 + X1 + rnorm(n)
(fit1 <- lm(Y1 ~ X1))
X2 <- rnorm(n, -1, 1)
Y2 <- X2^2 + X2 + rnorm(n)
(fit2 <- lm(Y2 ~ X2))
plot(X1, Y1, col = "red", xlim = c(-4, 4))
abline(fit1$coefficients, col = "red")
points(X2, Y2, col = "blue")
abline(fit2$coefficients, col = "blue")
```

- Diagnostics for normal linear model.
- Bias-variance decomposition.
- Demo of `plot.lm`.

```
plot(fit1)
X3 <- rnorm(n, -1, 1)
Y3 <- X3 + rnorm(n) * (X3 + 0.5)
(fit3 <- lm(Y3 ~ X3))
plot(fit3)
```

Lecture 5

- Quantitative criteria for model selection: C_p, cross-validation, AIC, BIC.
- Algorithms for model selection: best subset and forward/backward stepwise.



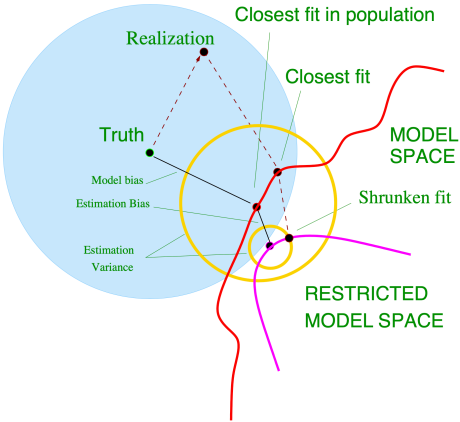
Linear models: Review

- Three assumptions in normal linear model:
  1. Linear conditional expectation;
  2. Independent noise;
  3. Normal and homoscedastic noise.
- MLE for β is obtained by least squares; fitted value $\hat{\boldsymbol{\mu}}$ is obtained by projection; partial regression/Gram-Schmidt.
- Exact inference: t-test, F-test, confidence set.
- Deviations from the normal linear model:
  1. Generalized least squares.
  2. Heteroscedasticity-robust inference.
  3. Projection interpretation when model is mis-specified.
  4. Bias-variance trade-off.

Linear models: Review (cont.)

- Diagnostics for linear model: R^2, leverage, studentized residual, different plots (residual vs. fitted, Q-Q, scale-location, residual vs. leverage), Cook's distance.
- Criteria for model selection:
  1. $C_p = \|Y - \hat{\mu}\|^2 + 2p\sigma^2$ [$p = \dim(\Theta)$ is model complexity].
  2. LOO-CV $= \sum_{i=1}^n (Y_i - \hat{\mu}_{(-i)})^2 = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2/(1 - P_{ii})^2$ [can be extended to general prediction problems].
  3. AIC $= -2\sum_{i=1}^n \log f(Y_i; \hat{\theta}) + 2\dim(\Theta)$ [works for general models with a likelihood].
  4. AIC $= -2\sum_{i=1}^n \log f(Y_i; \hat{\theta}) + \dim(\Theta) \cdot \log n$ [works for general models with a likelihood].
- Algorithms for model selection: best subset; forward/backward stepwise.

Schematic of the bias-variance trade-off



Lecture 6

- *Yule-Simpson paradox.
- *Regularization.
- Box-Cox transformation.
- Demo of model selection and Box-Cox transformation.

```
library(MASS)
X <- runif(100, min = 0, max = 2); mu <- sin(X * pi);
Y <- mu + 0.3 * rnorm(100)
lower <- Y ~ 1; upper <- Y ~ X + I(X^2) + I(X^3) + I(X^4)
fit.forward <- stepAIC(lm(lower), list(lower = lower, upper = upper),
                       direction = "forward")
fit.backward <- stepAIC(lm(upper), list(lower = lower, upper = upper),
                       direction = "backward")
Y <- exp(rnorm(50))
boxcox(Y ~ 1)
```

Lecture 7

- Definition of exponential family.
- Cumulants.
- Mean-value parametrization.

Lecture 8

- Asymptotic inference: Fisher's "master theorem" for MLE; delta method; Wilks' theorem.
- Application to one-parameter exponential family.
- Posterior distribution.
- *Empirical Bayes, Stein's estimator.

Shrinkage and empirical Bayes

- Tweedie's formula: If $Y \sim \mathrm{N}(\mu, \sigma^2)$ and $\mu \sim \pi$, then

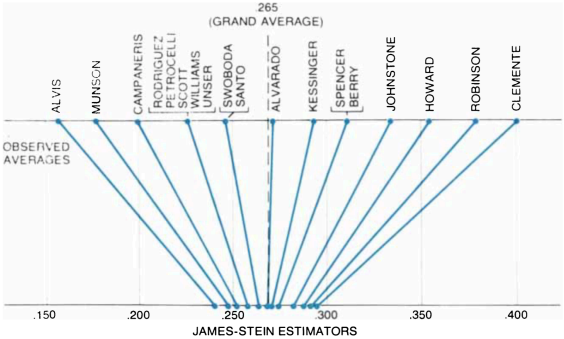
$$E(\mu \mid Y) = Y + \sigma^2 \cdot \frac{f'(Y)}{f(Y)},$$

  where $f(y) = \int f(y \mid mu)\pi(\mu)d\mu$ is the marginal density function of Y.
- Empirical Bayes: Given independent observations $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, p$, we can estimate $f(y)$ and plug it into Tweedie's formula.
- When the prior is $\mu_i \sim \mathrm{N}(\eta, \tau^2)$, this leads to Stein's estimator

$$\hat{\mu}_{\text{Stein}} = \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2}\right)(\mathbf{Y} - \bar{Y}\mathbf{1}) + \bar{Y}\mathbf{1} \quad \left[= \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{Y}\|^2}\right)\mathbf{Y} \text{ if } \bar{Y} = 0\right].$$

- This 'dominates' the MLE $\hat{\mu}_{\text{MLE}} = Y$ in a strong sense (*Principles of Statistics*).
- It is similar to regularization: $\hat{\mu}_\lambda = \arg\min_{\boldsymbol{\mu}} \|\mathbf{Y} - \boldsymbol{\mu}\|^2 + \lambda\|\boldsymbol{\mu}\|^2 = \mathbf{Y}/(1+\lambda)$.

Demonstration of shrinkage estimator



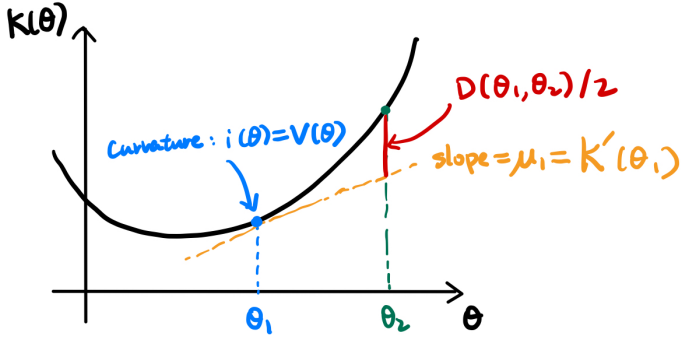
JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Bradley Efron and Carl Morris. "Stein's paradox in statistics." Scientific American 236.5 (1977): 119-127. `https://www.jstor.org/stable/24954030`.

Lecture 9

- *Demonstration of shrinkage estimator.
- Deviance and its geometric interpretation.
- Deviance residual.

Connection between deviance and convex analysis



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

Lecture 10

- Canonical form GLM and MLE.
- Analysis of deviance.

Lecture 11

- Overdispersion due to clustering.
- Exponential dispersion family: definition and examples.
- General form of GLM (with linkage and overdispersion).
- Estimation and asymptotic inference.

Lecture 12

- Asymptotic inference for general GLMs.
- Numerical computation: Newton-Raphson; Fisher scoring; IRLS.

Lecture 13

- GLM diagnostics and model selection.
- Binomial GLM.
- Latent variable interpretation.
- Case-control sampling.

GLM diagnostics

- Key idea: let $\boldsymbol{Z}^{(t)} = \boldsymbol{\eta}^{(t)} + \boldsymbol{R}^{(t)}$ be the "pseudo-response" and $\hat{\boldsymbol{Z}} = \lim_{t\to\infty} \boldsymbol{Z}^{(t)} = \hat{\boldsymbol{\eta}} + \hat{\boldsymbol{R}}$ where $R_i = (Y_i - \hat{\mu}_i)g'(\hat{\mu}_i)$.
- $\hat{\boldsymbol{\beta}}$ can be obtained from weighted least squares: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^T\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{Z}}$.
- We can then treat $\hat{\boldsymbol{W}}^{1/2}\hat{\boldsymbol{Z}}, \hat{\boldsymbol{W}}^{1/2}\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{W}}^{1/2}\hat{\boldsymbol{R}}$ as, respectively, the "adjusted" responses, fitted values, and residuals.
- Leverage of an observation can be defined as the corresponding diagonal entry of the "adjusted hat matrix" $\boldsymbol{H} = \hat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^T\boldsymbol{X}^T\hat{\boldsymbol{W}}^{1/2}$.
- Pearson and deviance residuals are defined as

$$R_{P,i} = \frac{Y_i - \hat{\mu}_i}{(\hat{\sigma}^2/w_i)V(\hat{\mu}_i)}, \quad R_{D,i} = \operatorname{sign}(Y_i - \hat{\mu}_i)\sqrt{\frac{D(Y_i, \hat{\mu}_i)}{\hat{\sigma}^2/w_i}}.$$

- Like in diagnostics for linear models, they are often standardized by dividing by $\sqrt{1 - H_{ii}}$.
- Cook's distance can be similarly extended to GLMs: $D_i = \frac{1}{p}\frac{H_{ii}}{1-H_{ii}}\tilde{R}_{P,i}$.

Lecture 14

- Variance stabilizing transformation.
- Poisson GLM.
- Deviance and Pearson's χ^2.
- Demonstration of Poisson GLM.

```
Day <- seq(1, 12);
Count <- rpois(length(Day), 5 * exp(0.2 * Day))
data <- data.frame(Day, Count)
rownames(data) <- NULL
rbind(head(data, 2), tail(data, 2))
fit <- glm(Count ~ Day, family = poisson, data)
summary(fit)
plot(Count ~ Day, data)
lines(data$Day, predict(fit, data, type = "response"))
```

Lecture 15

- Multinomial data and the Poisson trick.
- Surrogate Poisson model for 2-way contingency tables.

Lecture 16

- Surrogate Poisson model for 3-way contingency tables.
- *Undirected graphical models.
- Review and look forward.