34 Lecture 34: April 27

Last time

- Theoretical background of linear model

Today

- Course evaluation (12/17)
- Multivariate Normal and Cochran's theorem
- Bootstrap
- Logistic Regression (JF Chapter 14)

Additional reference

"Essential Statistical Inference Theory and Methods" by Dr. Dennis D. Boos and Dr. L. A. Stefanski.
Dr. Hua Zhou's Computational Statistics notes.

Normal distribution in scaler case

- A random variable Z has a standard normal distribution, denoted $Z \sim \mathcal{N}(0,1)$, if

$$F_Z(t) = \Pr(Z \leqslant t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz,$$

  or equivalently Z has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

  or equivalently, Z has moment generating function (mgf)

$$m_Z(t) = \mathrm{E}(e^{tZ}) = e^{t^2/2}, \quad -\infty < z < \infty$$

- Non-standard normal random variable
  - Definition 1: A random variable X has normal distribution with mean μ and variance σ^2, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if

$$X = \mu + \sigma Z$$

    where $Z \sim \mathcal{N}(0,1)$
  - Definition 2: $X \sim \mathcal{N}(\mu, \sigma^2)$ if

$$m_X(t) = \mathrm{E}(e^{tX}) = e^{t\mu + \sigma^2 t^2/2}, \quad -\infty < t < \infty$$

  - In both definitions, $\sigma^2 = 0$ is allowed. If $\sigma^2 > 0$, it has a density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

Multivariate normal distribution

- The standard multivariate normal is a vector of independent standard normals, denoted $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. The joint density is

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}} e^{-\sum_{i=1}^p z_i^2/2}.$$

  The mgf is

$$m_{\mathbf{Z}}(\mathbf{t}) = \prod_{i=1}^p m_{Z_i}(t_1) = \prod_{i=1}^p e^{t_i^2/2} = e^{\mathbf{t}^T\mathbf{t}/2}.$$

- Consider the affine transformation $\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. $\mathbf{X}$ has mean and variance

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}, \quad \mathrm{Var}(\mathbf{X}) = \mathbf{A}\mathbf{A}^T$$

  and the moment generating function is

$$m_{\mathbf{X}}(\mathbf{t}) = \mathrm{E}(e^{\mathbf{t}^T(\boldsymbol{\mu}+\mathbf{A}\mathbf{Z})}) = e^{\mathbf{t}^T\boldsymbol{\mu}}\mathrm{E}(e^{\mathbf{t}^T\mathbf{A}\mathbf{Z}}) = e^{\mathbf{t}^T\boldsymbol{\mu}+\mathbf{t}^T\mathbf{A}\mathbf{A}^T\mathbf{t}/2}.$$

- $\mathbf{X} \in \mathbb{R}^p$ has a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\mathbf{V} \in \mathbb{R}^{p\times p}, \mathbf{V} \succeq_{p.s.d.} \mathbf{0}$, denoted $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, if its mgf takes the form

$$m_{\mathbf{X}}(\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu}+\mathbf{t}^T\mathbf{V}^T\mathbf{t}/2}, \quad \mathbf{t} \in \mathbb{R}^p$$

- if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{V}$ is non-singular, then
  - $\mathbf{V} = \mathbf{A}\mathbf{A}^T$ for some non-singular $\mathbf{A}$
  - $\mathbf{A}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$
  - The density of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{V}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}.$$

- (Any affine transform of normal is normal) If $\mathbf{X} \in \mathbb{R}^p, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$, where $\mathbf{a} \in \mathbb{R}^q$ and $\mathbf{B} \in \mathbb{R}^{q\times p}$, then $\mathbf{Y} \sim \mathcal{N}(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{V}\mathbf{B}^T)$.
- (Marginal of normal is normal) If $\mathbf{X} \in \mathbb{R}^p, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, then any subvector of $\mathbf{X}$ is normal too.
- A convenient fact about normal random variables/vectors is that zero correlation/covariance implies independence.
  If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and is partitioned as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_m \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \dots & \mathbf{V}_{1m} \\ \vdots & & \vdots \\ \mathbf{V}_{m1} & \dots & \mathbf{V}_{mm} \end{bmatrix}$$

  then $\mathbf{X}_1, \dots, \mathbf{X}_m$ are jointly independent if and only if $\mathbf{V}_{ij} = \mathbf{0}$ for all $i \neq j$.
  Proof:

Independence and Cochran's theorem

- (Independence between two linear forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{Y}_1 = \mathbf{a}_1 + \mathbf{B}_1\mathbf{X}$ and $\mathbf{Y}_2 = \mathbf{a}_2 + \mathbf{B}_2\mathbf{X}$. Then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if $\mathbf{B}_1\mathbf{V}\mathbf{B}_2^T = 0$.
Proof:

- Consider the normal linear model $\mathbf{y} \sim \mathcal{N}(\mathbf{Xb}, \sigma^2\mathbf{I}_n)$
  - Using $\mathbf{A} = (1/\sigma^2)(\mathbf{I} - \mathbf{P_X})$, we have
  $$SSE/\sigma^2 = ||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2 = \mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2_{n-r},$$
  where $r = \text{rank}(\mathbf{X})$. Note the noncentrality parameter is
  $$\phi = \frac{1}{2}(\mathbf{Xb})^T(1/\sigma^2)(\mathbf{I} - \mathbf{P_X})(\mathbf{Xb}) = 0 \text{ for all } \mathbf{b}.$$
  - Using $\mathbf{A} = (1/\sigma^2)\mathbf{P_X}$, we have
  $$SSR/\sigma^2 = ||\hat{\mathbf{y}}||_2^2/\sigma^2 = \mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2_r(\phi),$$
  with the noncentrality parameter
  $$\phi = \frac{1}{2}(\mathbf{Xb})^T(1/\sigma^2)\mathbf{P_X}(\mathbf{Xb}) = \frac{1}{2\sigma^2}||\mathbf{Xb}||_2^2.$$
  - The joint distribution of $\hat{\mathbf{y}}$ and $\hat{\boldsymbol{\epsilon}}$ is
  $$\begin{bmatrix} \hat{\mathbf{y}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{P_X} \\ \mathbf{I}_n - \mathbf{P_X} \end{bmatrix}\mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{Xb} \\ \mathbf{0}_n \end{bmatrix}, \begin{bmatrix} \sigma^2\mathbf{P_X} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I} - \mathbf{P_X}) \end{bmatrix}\right).$$
  So $\hat{\mathbf{y}}$ is independent of $\hat{\boldsymbol{\epsilon}}$. Thus $||\hat{\mathbf{y}}||_2^2/\sigma^2$ is independent of $||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2$ and
  $$F = \frac{||\hat{\mathbf{y}}||_2^2/\sigma^2/r}{||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2/(n-r)} \sim F_{r,n-r}(\frac{1}{2\sigma^2}||\mathbf{Xb}||_\mathbf{2}^\mathbf{2}).$$

- (Independence between linear and quadratic forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$. Let $\mathbf{A}$ be symmetric with rank s. Then $\mathbf{BX}$ and $\mathbf{X}^T\mathbf{AX}$ are independent if $\mathbf{BVA} = \mathbf{0}$.
Proof:

- (Independence between two quadratic forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{A}$ be symmetric with rank r, and $\mathbf{B}$ be symmetric with rank s. If $\mathbf{BVA} = \mathbf{0}$, then $\mathbf{X}^T\mathbf{AX}$ and $\mathbf{X}^T\mathbf{BX}$ are independent.
Proof:

- (Cochran's theorem) Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ and $\mathbf{A}_i$, $i = 1, \ldots, k$ be symmetric idempotent matrix with rank s_i. If $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n$, then $(1/\sigma^2)\mathbf{y}^T\mathbf{A}_i\mathbf{y}$ are independent $\chi^2_{s_i}(\phi_i)$, with $\phi_i = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}$ and $\sum_{i=1}^k s_i = n$.
Proof:

- Application to the one-way ANOVA: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$. We have the classical ANOVA table

Source	df	Projection	SS	Noncentrality
Mean	1	$\mathbf{P_1}$	$SSM = n\bar{y}^2$	$\frac{1}{2\sigma^2} n(\mu + \bar{\alpha})^2$
Group	$a-1$	$\mathbf{P_X} - \mathbf{P_1}$	$SSA = \sum_{i=1}^a n_i \bar{y}_i^2 - n\bar{y}^2$	$\frac{1}{2\sigma^2} \sum_{i=1}^a n_i(\alpha_i - \bar{\alpha})^2$
Error	$n-a$	$\mathbf{I} - \mathbf{P_X}$	$SSE = \sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	0
Total	n	$\mathbf{I}$	$SST = \sum_i \sum_j y_{ij}^2$	$\frac{1}{\sigma^2} \sum_{i=1}^a n_i(\mu + \alpha_i)^2$

Bootstrap

We follow JF Chapter 21 to discuss the version of nonparametric bootstrap here. The term *bootstrapping*, coined by Efron (1979), refers to using the sample to learn about the sampling distribution of a statistic without reference to external assumptions – as in "pulling oneself up by one's bootstraps."

Bootstrapping offers a number of advantages:

- The bootstrap is quite general, although there are some cases in which it fails.
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small.
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically.
- It is relatively simple to apply the bootstrap to complex data collection plans.

Bootstrap standard errors

For simplicity, we start with an iid sample $Y_1, \ldots, Y_n$ with each Y_i having distribution function F, and a real parameter θ is estimated by $\hat{\theta}$. When necessary, we think of $\hat{\theta}$ as a function of the sample, $\hat{\theta}(Y_1, \ldots, Y_n)$. The variance of $\hat{\theta}$ is then

$$\mathrm{Var}_F(\hat{\theta}) = \int \left\{ \hat{\theta}(y_1, \ldots, y_n) - \mathrm{E}_F(\hat{\theta}) \right\}^2 dF(y_1) \ldots dF(y_n),$$

where

$$\mathrm{E}_F(\hat{\theta}) = \int \hat{\theta}(y_1, \ldots, y_n) dF(y_1) \ldots dF(y_n).$$

The nonparametric bootstrap estimate of $\mathrm{Var}(\hat{\theta})$ is just to replace F by the empirical distribution function $F_n(y) = n^{-1} \sum_{i=1}^n I(Y_i \leqslant y)$:

$$\mathrm{Var}_{F_n}(\hat{\theta}) = \int \left\{ \hat{\theta}(y_1, \ldots, y_n) - \mathrm{E}_{F_n}(\hat{\theta}) \right\}^2 dF_n(y_1) \ldots dF_n(y_n),$$

Please refer to Chapter 11 of Boos and Stefanski for a complete discussion.

A practical bootstrapping procedure follows:

1. Create r number of bootstrap replications or pseudo-replicates – that is, for each bootstrap sample (replicate) $b = 1, \ldots, r$, we randomly draw n observations $\{Y^*_{b_1}, Y^*_{b_2}, \ldots, Y^*_{b_n}\}$ with replacement from the original sample $\{Y_1, Y_2, \ldots, Y_n\}$.
2. Obtain an estimate $\hat{\theta}^*_b$ of each bootstrap sample.
3. Use the distribution of $\hat{\theta}^*_b$ to estimate properties of the sampling distribution of $\hat{\theta}$.For example, the sample standard deviation of $\hat{\theta}^*_b$ gives the bootstrap standard error estimates of $\widehat{SE}^*(\hat{\theta})$.

Bootstrap example

We use the example in JF 21.1 for illustration. Imagine that we sample (fake) ten working, married couples, determining in each case the husband's and wife's income, as recorded in the table (JF table 21.3) below.

Observation	husband's Income	Wife's Income	Difference Y_i
1	34	28	6
2	24	27	-3
3	50	45	5
4	54	51	3
5	34	28	6
6	29	19	10
7	31	20	11
8	32	40	-8
9	40	33	7
10	34	25	9

A point estimate of this population mean difference μ is the sample mean,

$$\bar{Y} = \frac{\sum Y_i}{n} = 4.6$$

Elementary statistical theory tells us that the standard deviation of the sampling distribution of sample means is $\text{SD}(\bar{Y}) = \sigma/\sqrt{n}$, where σ is the population standard deviation of Y. Because we do not know σ in most real applications, the usual estimator of σ is the sample

standard deviation

$$\hat{S} = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}$$

and we obtain the 95% confidence interval by

$$\bar{Y} \pm t_{n-1,0.025}\frac{\hat{S}}{\sqrt{n}}$$

In the present case, $\hat{S} = 5.948$, $\widehat{SE}(\bar{Y}) = 5.948/\sqrt{10} = 1.881$, and $t_{9,0.025} = 2.262$. The 95% confidence interval for the population mean μ is therefore

$$4.6 \pm 2.262 \times 1.881 = 4.6 \pm 4.255$$

or equivalently,

$$0.345 < \mu < 8.855$$

To illustrate the bootstrap procedure,

1. We can draw $r = 2000$ bootstrap samples (using a computer), each of size $n = 10$, from the original data given in table 21.3.
2. We then calculate the mean $\bar{Y}_b^*$, with $b = 1, \ldots, r$ for each bootstrap sample.
3. The bootstrap estimate of the standard error is then given by $\widehat{SE}^*(\bar{Y}^*) = \sqrt{\frac{\sum_{b=1}^{r}\left(\bar{Y}_b^* - \bar{\bar{Y}}^*\right)^2}{r-1}}$

From the 2000 replicates that Dr. Fox drew, he obtained $\bar{\bar{Y}}^* = 4.693$ and $\widehat{SE}(\bar{Y}^*) = 1.750$. Both are quite close to the theoretical values (read JF 21.1 for a discussion over $\sqrt{n/n-1}$ for the differences in calculating the standard errors, which is often negligible, especially when n is large).

Now, we can get a bootstrap estimate for the $100(1-\alpha)\%$ confidence interval by using the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the bootstrap sampling distribution of $\hat{\theta}_b^*$ which means

1. We order $\hat{\theta}_b^*$ such that $\hat{\theta}_{(1)}^* \leqslant \hat{\theta}_{(2)}^* \leqslant \cdots \leqslant \hat{\theta}_{(r)}^*$.
2. Find the two quantiles $\hat{\theta}_{(lower)}^* = \hat{\theta}_{(\alpha/2\times r)}^*$ and $\hat{\theta}_{(upper)}^* = \hat{\theta}_{((1-\alpha/2)\times r)}^*$
3. Construct the confidence interval by $(\hat{\theta}_{(lower)}^*, \hat{\theta}_{(upper)}^*)$.

In this case,

$$\begin{aligned}
\text{lower} &= 2000(0.05/2) = 50 \\
\text{upper} &= 2000(1 - 0.05/2) = 1950 \\
\bar{Y}_{(50)}^* &= 0.7 \\
\bar{Y}_{(1950)}^* &= 7.8 \\
0.7 &< \mu < 7.8
\end{aligned}$$

Bias-corrected bootstrap intervals

We introduce the bias-corrected version of the above bootstrap intervals through two "correction factors" Z and A defined below:.

1. Calculate
$$Z \equiv \Phi^{-1}\left[\frac{\sum_{b=1}^{r} I(\hat{\theta}_b^* < \hat{\theta})}{r}\right]$$
where $\Phi^{-1}(\cdot)$ is the inverse of the standard-normal distribution and $\sum_{b=1}^{r} I(\hat{\theta}_b^* < \hat{\theta})/r$ is the proportion of bootstrap replicates below the estimate $\hat{\theta}$. If the bootstrap sampling distribution is symmetric and if $\hat{\theta}$ is unbiased, then this proportion will be close to 0.5, and the "correction factor" Z will be close to 0.

2. Let $\hat{\theta}_{(-i)}$ represent the value of $\hat{\theta}$ produced when ith observation is deleted from the sample (known as the jackknife values of $\hat{\theta}$). There are n of these quantities. Let $\bar{\theta} = \sum \hat{\theta}_{(-i)}/n$. Then calculate
$$A \equiv \frac{\sum_{i=1}^{n}(\bar{\theta} - \hat{\theta}_{(-i)})^3}{6\left[\sum_{i=1}^{n}(\bar{\theta} - \hat{\theta}_{(-i)})^2\right]^{3/2}}$$
With the correction factors Z and A, compute
$$A_1 \equiv \Phi\left[Z + \frac{Z - z_{\alpha/2}}{1 - A(Z - z_{\alpha/2})}\right]$$
$$A_2 \equiv \Phi\left[Z + \frac{Z + z_{\alpha/2}}{1 - A(Z + z_{\alpha/2})}\right]$$
And the corrected interval is
$$\hat{\theta}^*_{(lower)} < \theta < \hat{\theta}^*_{(upper)}$$
where lower$^* = rA_1$ and upper$^* = rA_2$ (rounding or interpolating as required).

When the correction factors Z and A are both 0, $A_1 = \Phi(-z_{\alpha/2}) = \alpha/2$ and $A_2 = \Phi(z_{\alpha/2}) = 1 - \alpha/2$.

For the 2000 bootstrap samples that Dr. Fox drew, there are 926 bootstrapped means below $\bar{Y} = 4.6$, and so $Z = \Phi^{-1}(926/2000) = -0.09288$. The $\bar{Y}_{(-i)}$ are $4.444, 5.444, \ldots, 4.111$. And $A = -0.05630$. Using the correction factors Z and A,

$$\begin{aligned} A_1 &= \Phi\left[-0.09288 + \frac{-0.09288 - 1.96}{1 - [-0.05630(-0.09288 - 1.96)]}\right] \\ &= \Phi(-2.414) = 0.007889 \\ A_2 &= \Phi\left[-0.09288 + \frac{-0.09288 + 1.96}{1 - [-0.05630(-0.09288 + 1.96)]}\right] \\ &= \Phi(1.597) = 0.9449 \end{aligned}$$

Multiplying by r, we have $2000 \times 0.007889 \approx 16$ and $2000 \times 0.9449 \approx 1890$, from which

$$\bar{Y}^*_{(16)} < \mu < \bar{Y}^*_{(1890)}$$
$$-0.4 < \mu < 7.3$$

Logistic regression

So far, we only considered cases where the response variable is continuous. Logistic regression belongs in the family of Generalized Linear Model that can be used for analyzing binary responses.

Motivation Let p be the probability of a specific outcome. We are interested in how this probability is affected by the explanatory variables. A naive approach could be:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Problem p must be between 0 and 1.

Solution Model log odds of p (i.e. logit of p) which are defined as

$$\text{odds} = \frac{p}{1-p} \in [0, \infty)$$
$$\text{logit} = \log(\frac{p}{1-p}) \in (-\infty, \infty)$$

This forms the logistic regression

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Note that

1. Increase in log odds $\iff$ increase in p.
   Decrease in log odds $\iff$ decrease in p.
2. No ϵ in logistic regression because we observe a binary outcome y_i, not p itself.

The density

$$\begin{aligned} f(y_i|p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= e^{y_i \log(p_i) + (1-y_i)\log(1-p_i)} \\ &= e^{y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i)} \end{aligned}$$

where

$$\text{E}(y_i) = p_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \quad \text{(mean function, inverse link function)}$$
$$\mathbf{x}_i^T \boldsymbol{\beta} = \log(\frac{p_i}{1-p_i}) \quad \text{(logit link function)}$$

We obtain parameter estimates by maximum likelihood. Read page 131 - page 133 of Dr. Hua Zhou's Computational Statistics notes (link) for algorithms to find these MLE (maximum likelihood estimates).