250B Linear Statistical Models B Review

Tomoki Okuno

Summer 2023

Contents

1 Filler's Theorem

- Suppose the joint distribution of *uncorrelated* U and V

$$\begin{pmatrix} U \\ V \end{pmatrix} \quad \sim \quad N_2\left(\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \sigma^2 \begin{pmatrix} a_1^2 & 0 \\ 0 & a_2^2 \end{pmatrix}\right)$$

Further, let s^2 be an estimate of σ^2 with m degrees of freedom, namely, $ms^2/\sigma^2 \sim \chi^2_m$ and assume

$$g = \frac{t^2_{m,\alpha/2} s^2 a_2^2}{V^2} < 1$$

then a $100(1-\alpha)\%$ CI for $\theta = \mu/\nu$, ratio of the expected values of U and V is

$$\theta = \frac{\mu}{\nu} \quad \in \quad \frac{1}{1-g}\left(\frac{U}{V} \pm \frac{t_{m,\alpha/2} s}{V}\sqrt{(1-g)a_1^2 + \frac{U^2}{V^2}a_2^2}\right).$$

Note that the requirement of $g < 1$ is equivalent to requiring V is significantly different from 0.

Proof: Let $W = U - \theta V$, then $W \sim N(0, \sigma^2(a_1^2 + \theta^2 a_2^2))$. Hence

$$t = \frac{W/(\sigma\sqrt{a_1^2 + \theta^2 a_2^2})}{\sqrt{s^2/\sigma^2}} = \frac{U - \theta V}{s\sqrt{a_1^2 + \theta^2 a_2^2}} \quad \sim \quad t_m,$$
$$\Rightarrow (U - \theta V)^2 = t^2_{m,\alpha/2} s^2 (a_1^2 + \theta^2 a_2^2)$$

The CI can be obtained solving this w.r.t θ. This is the case where two variables are uncorrelated.

2 Simultaneous Interval Estimation

- We begin with the full-rank model $Y = X\beta + \epsilon$ with $X : n \times p$. Then one way to construct $100(1-\alpha)\%$ CI for k linear combination $a_j'\beta$ $(j = 1, \ldots, k)$ is

$$\begin{aligned} a_j'\beta \quad &\in \quad a_j'\widehat{\beta} \pm t^{\alpha/2}_{n-p} S\sqrt{a_j'(X'X)^{-1}a_j} \\ &\in \quad a_j'\widehat{\beta} \pm t^{\alpha/2}_{n-p}\widehat{\sigma}_{a_j'\widehat{\beta}} \end{aligned}$$

But this has a multiple comparison problem.

- Suppose that E_j $j = 1, \ldots, k$ is the event that j-th statement is correct, and let $Pr(E_j) = 1 - \alpha_j$ and let δ be the probability of getting at least one statement wrong. Then we have

$$1 - \delta = Pr\left(\bigcap_{j=1}^k E_j\right) = 1 - Pr\left(\overline{\bigcap_{j=1}^k E_j}\right) = 1 - Pr\left(\bigcup_{j=1}^k \overline{E_j}\right) \geq 1 - \sum_{j=1}^k Pr(\overline{E_j}) = 1 - \sum_{j=1}^k \alpha_j.$$

This is called Bonferroni inequality, which is not as crude as one might expect with $k \leq 5$ and small α. $\gamma = \sum_{j=1}^k \alpha_j$ is the expected number of incorrect statements if k is finite. *Proof*: Define $I_j = 1$ if E_j is incorrect with $Pr(\overline{E}_j) = \alpha_j$. Then $I_j \sim$ Bernoulli(α_j), so that $E(\sum_j I_j) = \sum_j E(I_j) = \sum_j \alpha_j = \gamma$.

- If the dependence between the events E_j is small, we can write

$$\begin{aligned} Pr\left(\bigcap_{j=1}^k E_j\right) &= Pr(E_1)Pr(E_2 \mid E_1)\cdots Pr(E_k \mid E_1, \ldots E_{k-1}) \\ &\approx Pr(E_1)Pr(E_2)\cdots Pr(E_k) \\ &= \prod_{j=1}^k (1 - \alpha_k) \geq 1 - \sum_{j=1}^k \alpha_j \end{aligned}$$

which provides better bounds than the previous one.

Proof: Set $\alpha_i = \alpha$ for simplicity. Let $f(\alpha) = (1-\alpha)^k - 1 - k\alpha$. Then $f(\alpha) \geq f(0) = 0$

- **Bonferroni** t-interval: If an individual significant level for each k is common ($\alpha_j = \alpha$), and we use this as $\alpha' = \alpha/k$ instead of α, then we have

$$Pr\left(\bigcap_{j=1}^{k} E_j\right) \geq 1 - k\alpha = 1 - \alpha'.$$

This is the Bonferroni correction for the multiple comparisons. Then their CIs become

$$a_j'\beta \quad \in \quad a_j'\widehat{\beta} \pm t_{n-p}^{\alpha/2k} S\sqrt{a_j'(X'X)^{-1}a_j}$$

- **Scheffe's** interval. We assume WLS that the first d vectors of the set $\{a_1, \ldots, a_k\}$ are linearly independent, and the remaining vectors are linearly dependent on the d vectors. Consider $A\beta$, where

$$A = \begin{pmatrix} a_1' \\ \vdots \\ a_d' \end{pmatrix} \quad \Rightarrow \quad A\beta = \begin{pmatrix} a_1'\beta \\ \vdots \\ a_d'\beta \end{pmatrix}.$$

Using the F-statistic we have in 250A, and setting $b = A(\widehat{\beta} - \beta)$ and $L = A(X'X)^{-1}A'$ yields

$$\begin{aligned} 1 - \alpha &= Pr(F \leq F_{d,n-p}^{\alpha}) \\ &= Pr(b'L^{-1}b \leq dS^2 F_{d,n-p}^{\alpha}) \\ &= Pr\left[\max_{h \neq 0}\left\{\frac{(h'b)^2}{h'L^{-1}h}\right\} \leq dS^2 F_{d,n-p}^{\alpha}\right] \\ &= Pr\left[\frac{(h'b)^2}{h'L^{-1}h} \leq dS^2 F_{d,n-p}^{\alpha},\ \forall h \neq 0\right] \\ &= Pr\left(\frac{|h'A(\widehat{\beta} - \beta)|}{S\sqrt{h'L^{-1}h}} \leq \sqrt{dF_{d,n-p}^{\alpha}},\ \forall h \neq 0\right), \end{aligned}$$

which gives CIs for $h'A\beta$ like this:

$$\begin{aligned} h'A\beta \ &\in \ h'A\widehat{\beta} \pm \sqrt{dF_{d,n-p}^{\alpha}}\, S\sqrt{h'L^{-1}h}, \\ &\in \ h'A\widehat{\beta} \pm \sqrt{dF_{d,n-p}^{\alpha}}\, \widehat{\sigma}_{h'A\widehat{\beta}} \end{aligned}$$

since $\text{var}(h'A\widehat{\beta}) = \sigma^2 h'L^{-1}h$. In particular, setting $h = c_j$, we get

$$a_j'\beta \ \in \ a_j'\widehat{\beta} \pm \sqrt{dF_{d,n-p}^{\alpha}}\, \widehat{\sigma}_{a_j'\widehat{\beta}}, \quad j = 1, \ldots, d.$$

In addition, an interval for every $a_j'\beta$ ($j = d+1, \ldots k$) is also included in this set, owing to the linear dependence of the a_j on the other a_j ($j = 1, \ldots, d$). For example, if $a_{d+1} = h_1 a_1 + \cdots + h_d a_d$, then $a_{d+1}'\beta = \sum_{j=1}^{d} h_i a_i'\beta = h'A\beta$. Thus, the above CI also applies to the other a_j, $j = d+1, \ldots, k$.

- Suppose that we wish to test $H : \beta_1 = \cdots = \beta_{d+1}$, which can be written in the form $\gamma_i = \beta_i - \beta_{d+1}$ ($i = 1, \ldots, d$) so that Sheffe's method will provide confidence intervals for all linear combination

$$\sum_{i=1}^{d} h_i\gamma_i = \sum_{i=1}^{d} h_i\beta_i + \left(-\sum_{i=1}^{d} h_i\right)\beta_{d+1} = \sum_{i=1}^{d+1} c_i\beta_i, \quad c_{d+1} = -\sum_{i=1}^{d} h_i,$$

where $\sum_{i=1}^{d+1} c_i = 0$; thus every linear combination of the γ_i is a **contrast** in the β_i ($i = 1, \ldots, d+1$).

- Once we have estimated β from n observation Y, we can use predictor $\widehat{Y} = x'\beta$. If we are interested in a particular value of $x = x_0$, using $\widehat{Y}_0 = x_0'\widehat{\beta}$, then the interval is $\widehat{Y}_0 \pm t_{n-p}^{\alpha/2} S\sqrt{x_0'(X'X)^{-1}x_0}$. If we are interested in all values of $x = (1, x_1, \ldots, x_{p-1})$, then Sheffe's method provides simultaneous intervals

$$x'\beta \ \in \ x'\widehat{\beta} \pm (pF_{p,n-p}^{\alpha})^{1/2} S\sqrt{x'(X'X)^{-1}x}$$

with an exact overall probability of $1-\alpha$.

- Example: Let $p = 2$, that is, $\widehat{Y}_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $(i = 1, \ldots, n)$, where $\epsilon_i \sim N(0, \sigma^2)$. Then a set of multiple $100(1-\alpha)\%$ CIs for all linear combinations $a_0\beta_0 + a_1\beta_1$ $(a_0 \neq 0, a_1 \neq 0)$ is

$$a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1 \pm (2F_{2,n-2}^{\alpha})\widehat{v}^{1/2},$$

where $\widehat{v} = S^2\left[a_0(\sum_i x_i^2/n) - 2a_0a_1\overline{x} + a_1^2\right]/\sum_i(x_i - \overline{x})^2$.

- In practice, we are generally more interested in predicting the value, Y_0, say, of the random variable Y, where $Y_0 = x_0'\beta + \epsilon_0$. If we assume $\epsilon_0 \sim N(0, \sigma^2)$ and ϵ_0 is independent of Y. Then

$$\begin{aligned} &E(\widehat{Y}_0 - Y_0) = 0, \quad \text{var}(\widehat{Y}_0 - Y_0) = \sigma^2[x_0'(X'X)^{-1}x_0 + 1] \\ &\Rightarrow \quad \widehat{Y}_0 - Y_0 \quad \sim \quad N(0, \sigma^2(x_0'(X'X)^{-1}x_0 + 1)). \end{aligned}$$

Thus, the interval estimate is $\widehat{Y}_0 \pm t_{n-p}^{\alpha/2} S\sqrt{x_0'(X'X)^{-1}x_0 + 1}$. Similarly, simultaneous prediction intervals for k future values of Y is given by

$$x'\beta \ \in \ x'\widehat{\beta} \pm [(p+k)F_{p+k,n-p}^{\alpha}]^{1/2} S\sqrt{x'(X'X)^{-1}x + 1}.$$

3 Straight Line Regression

- Again, consider the simplest regression model, $\widehat{Y}_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $(i = 1, \ldots, n)$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
- Important properties are

$$\text{SSE} = \|Y - \widehat{Y}\| = \sum_i [Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})]^2 = \text{TSS} - \widehat{\beta}_1^2 \sum_i (x_i - \overline{x})^2$$

$$\text{cov}(\overline{Y}, \widehat{\beta}_1) = \frac{1}{n}\,\text{cov}(\sum_i Y_i, \sum_j c_j Y_j) = \frac{1}{n}\sum_i c_i\,\text{Var}(Y_i) = \frac{\sigma^2}{n}\sum_i c_i = 0.$$

- Since $\text{Var}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$, the simultaneous CIs for β_0 and β_1 are, respectively,

$$\widehat{\beta}_0 \pm \lambda S\sqrt{\frac{\sum_i x_i^2}{n\sum_i(x_i - \overline{x})}}, \quad \widehat{\beta}_1 \pm \lambda S\sqrt{\frac{1}{\sum_i(x_i - \overline{x})}},$$

where $\lambda = t_{n-2}^{\alpha/4}$ (Bonferroni) or $\lambda = (2F_{2,n-2}^{\alpha})^{1/2}$ (Sheffe's method).

- Confidence interval for the x-intercept: Since $E(Y) = \beta_0 + x\beta_1 = 0$ the x-intercept can be written as $\phi = -\beta_0/\beta_1$. Let $U = \widehat{\beta}_0 + \phi\widehat{\beta}_1$. As both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased and normal, $U \sim N(0, V)$, where

$$V = \text{var}(\widehat{\beta}_0 + \phi\widehat{\beta}_1) = \text{var}\left[\overline{Y} + \widehat{\beta}_1(\phi - \overline{x})\right] = \sigma^2\left[\frac{1}{n} + \frac{(\phi - \overline{x})^2}{\sum_i(x_i - \overline{x})^2}\right] = \sigma^2 w.$$

Hence, we have t-statistic

$$T = \frac{U/(\sigma\sqrt{w})}{S/\sigma} = \frac{\overline{Y} + \widehat{\beta}_1(\phi - \overline{x})}{S\sqrt{w}} \sim t_{n-2}.$$

so that a $100(1-\alpha)\%$ CI for ϕ is given by $T^2 \le (t_{n-2}^{\alpha/2})^2 = F_{1,n-2}^{\alpha}$. Hence, the interval reduces $d_1 + \overline{x} \le \phi \le d_2 + \overline{x}$, where d_1 and d_2 are the roots of the quadratic

$$\begin{aligned}
(\overline{Y} + \widehat{\beta}_1 d)^2 &= F_{1,n-2}^{\alpha} S^2 \left[\frac{1}{n} + \frac{d^2}{\sum_i (x_i - \overline{x})^2}\right] \\
\Rightarrow \overline{Y}^2 + 2\overline{Y}\widehat{\beta}_1 d + \widehat{\beta}_1^2 d^2 &= \frac{F_{1,n-2}^{\alpha} S^2}{n} + \frac{F_{1,n-2}^{\alpha} S^2 d^2}{\sum_i (x_i - \overline{x})^2} \\
\Rightarrow d^2 \left[\widehat{\beta}_1^2 - \frac{F_{1,n-2}^{\alpha} S^2}{\sum_i (x_i - \overline{x})^2}\right] + 2\overline{Y}\widehat{\beta}_1 d + \left(\overline{Y}^2 - \frac{F_{1,n-2}^{\alpha} S^2}{n}\right) &= 0.
\end{aligned}$$

- Prediction intervals for **mean response** $E(Y_0)$: a set of k multiple $100(1-\alpha)\%$ prediction intervals is

$$\widehat{Y}_0 \pm \lambda S \sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}},$$

  where $\lambda = t_{n-2}^{\alpha/(2k)}$ (Bonferroni) or $\lambda = (2F_{2,n-2}^{\alpha})^{1/2}$ (Sheffe's method) since still $p = 2$.

- **Prediction intervals for response** Y_0: a set of k multiple $100(1-\alpha)\%$ prediction intervals for the random variable $Y_0^{(j)} = \beta_0 + \beta_1 x_0^{(j)} + \epsilon_0^{(j)}$ $(j = 1, \ldots, k)$ is

$$\widehat{Y}_0 \pm \lambda S \sqrt{1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}},$$

  where $\lambda = t_{n-2}^{\alpha/(2k)}$ (Bonferroni) or $\lambda = (kF_{k,n-2}^{\alpha})^{1/2}$ (Sheffe's method) since still $p = 2$.

- Inverse prediction (Calibration). A natural estimate of x_0 (which is also MLE) is found by solving the fitted equation $Y_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x$, namely,

$$\widehat{x}_0 = \frac{Y_0 - \widehat{\beta}_0}{\widehat{\beta}_1} = \overline{x} + \frac{Y_0 - \overline{Y}}{\widehat{\beta}_1}.$$

  A CI for x_0 can be constructed using the filler theorem. Suppose

$$x_0 = \overline{x} + \frac{Y_0 - \overline{Y}}{\beta_1}, \quad U = (Y_0 - \overline{Y}) - (x_0 - \overline{x})\widehat{\beta}_1.$$

  Then $E(U) = 0$ and since $\mathrm{cov}(\overline{Y}, \widehat{\beta}_1) = 0$,

$$\mathrm{var}(U) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right] = \sigma_U^2$$

  so that the t-statistic is given by

$$T = \frac{U/\sigma_U}{S/\sigma} = \frac{(Y_0 - \overline{Y}) - (x_0 - \overline{x})\widehat{\beta}_1}{S\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right]^{1/2}} \quad \sim \quad t_{n-2}.$$

  Hence, we have the interval $d_1 \le x_0 - \overline{x} \le d_2 \Rightarrow d_1 + \overline{x} \le x_0 \le d_2 + \overline{x}$, where d_1 and d_2 are roots of

$$[(Y_0 - \overline{Y}) - d\widehat{\beta}_1]^2 = (t_{n-2}^{\alpha/2})^2 S \left[1 + \frac{1}{n} + \frac{d^2}{\sum_i (x_i - \overline{x})^2}\right].$$

- Exercise 6a.4. Using the notation of the above, show that when $\overline{x} = 0$, $(\widehat{x}_0 - \widetilde{x}_0)/\widehat{x}_0 = 1 - r^2$, where $\widetilde{x}_0$ (inverse estimate) is obtained by regressing x on Y and r is the correlation coefficient of the pairs (x_i, Y_i). This implies that when $r^2 \to 1$, $\widehat{x}_0 \approx \widetilde{x}_0$, i.e., little difference between the two estimates.

Solution: $\widehat{x}_0 = (Y_0 - \overline{Y})/\widehat{\beta}_1$ and $\widetilde{x}_0 = \widehat{a}_0 + \widehat{a}_1 Y_0 = \widehat{a}_1(Y_0 - \overline{Y})$ as $\overline{x} = 0$. Hence,

$$\frac{\widehat{x}_0 - \widetilde{x}_0}{\widehat{x}_0} = \frac{1/\widehat{\beta}_1 - \widehat{a}_1}{1/\widehat{\beta}_1} = 1 - \widehat{a}_1\widehat{\beta}_1 = 1 - \frac{[\sum_i (Y_i - \overline{Y})(x_i - \overline{x})]^2}{\sum_i (Y_i - \overline{Y})^2 \sum_i (x_i - \overline{x})^2} = 1 - r^2.$$

- Miscellaneous Exercise 6.1. Let F is the F-statistic for testing $H : \beta_1 = 0$ for a straight line. Prove

$$\widetilde{x}_0 - \overline{x} = \frac{F}{F + (n-2)}(\widehat{x}_0 - \overline{x})$$

Solution: $\widetilde{x}_0 - \overline{x} = (\widehat{a}_0 - \widehat{a}_1\overline{Y}) - \overline{x} = \widehat{a}_1(Y_0 - \overline{Y}) = \widehat{a}_1\widehat{b}_1(\widehat{x}_0 - \overline{x}) = r^2(\widehat{x}_0 - \overline{x})$.

$$F = \frac{(\text{SSE}_H - \text{SSE})/1}{\text{SSE}/(n-2)} = \frac{\text{TSS} - \text{SSE}}{\text{SSE}/(n-2)} = \frac{1 - (1-r^2)}{(1-r^2)/(n-2)} = \frac{(n-2)r^2}{1-r^2} \;\Rightarrow\; r^2 = \frac{F}{F + (n-2)}.$$

- Replicated prediction. Suppose we have m replications Y_{0j} $(j = 1, \ldots, m; m > 1)$ with sample mean $\overline{Y}_0$, at the unknown value $x = x_0$. Let $U = (\overline{Y}_0 - \overline{Y}) - (x_0 - \overline{x})\widehat{\beta}_1$, then it variance is

$$\text{var}(U) = \sigma^2 \left[\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right] = \sigma_U^2$$

If the two estimated variances are

$$V_1 = \sum_{i=1}^{n} \left[Y_i - \overline{Y} - \widehat{\beta}_1(x_i - \overline{x})\right]^2 = \text{SSE}, \quad V_2 = \sum_{j=1}^{m} (Y_{0j} - \overline{Y}_0^2),$$

then U, V_1, and V_2 are mutually independent and hence

$$\frac{(n - 2 + m - 1)\widehat{\sigma}^2}{\sigma^2} = \frac{V_1 + V_2}{\sigma^2} \sim \chi^2_{m+n-3}.$$

Therefore, the t-statistic is given by

$$T = \frac{U/\sigma_U}{S/\sigma} = \frac{(\overline{Y}_0 - \overline{Y}) - (x_0 - \overline{x})\widehat{\beta}_1}{S\left[\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right]^{1/2}} \quad \sim \quad t_{n+m-3}$$

and solve $T^2 = F^{\alpha}_{1,n+m-3}$ w.r.t $d = x_0 - \overline{x}$ to obtain the CI for x_0.

4 Comparing straight lines

- Suppose that we with to compare K regression lines

$$Y = \alpha_k + \beta_k x + \epsilon, \quad k = 1, \ldots, K,$$

where $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2$ (same for each line). If we are given n_k pairs of observations (x_{ki}, Y_{ki}) $(i = 1, \ldots, n_k)$ on the k-th line and let $N = \sum_{k=1}^{K} n_k$, then we have the model $Y_{ki} = \alpha_k + \beta_k x_{ki} + \epsilon_{ki}$, $i = 1, \ldots, n_k$, where $\epsilon_{ki} \overset{iid}{\sim} N(0, \sigma^2)$ or $\epsilon \sim N_N(0, \sigma^2 I_N)$. Then we have

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{K1} \\ \vdots \\ Y_{Kn_K} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 & x_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 & x_{1n_1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & x_{K1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & x_{Kn_K} \end{pmatrix}}_{N \times 2K} \underbrace{\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}}_{2K} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. Thus we can test any hypothesis of $H : A\gamma = c$.

- Test for *parallelism*. Suppose that we wish to test whether the K lines are parallel; then $H_1 : \beta_1 = \cdots = \beta_K (= \beta)$ or $H_1 : \boldsymbol{\beta} = \mathbf{1}_K \beta$; in the matrix form this is

$$\underbrace{\begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}}_{(K-1)\times 2K} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = A_1 \gamma = 0,$$

with $\text{rank}(A_1) = K - 1$, so that the F-statistics is given by

$$F = \frac{(\text{SSE}_{H_1} - \text{SSE})/(K-1)}{\text{SSE}/(N-2K)}.$$

To obtain SSE_{H_1}, when H_1 true, the design matrix, X_1 say, is obtained by simply adding together the last K columns of X. That is, the model reduces to

$$\mathbf{Y} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 & x_{11} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & x_{1n_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & x_{K1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & x_{Kn_K} \end{pmatrix}}_{N\times(K+1)} \underbrace{\begin{pmatrix} \boldsymbol{\alpha} \\ \beta \end{pmatrix}}_{K+1} + \boldsymbol{\epsilon} = \mathbf{X}_1 \boldsymbol{\gamma}_1 + \boldsymbol{\epsilon}.$$

- Test for *coincidence*. To test whether the K lines are coincident, we consider $H_2 : \alpha_1 = \cdots = \alpha_K (= \alpha)$ and $\beta_1 = \cdots = \beta_K (= \beta)$ or $H_2 : \boldsymbol{\alpha} = \mathbf{1}_K \alpha$ and $\boldsymbol{\beta} = \mathbf{1}_K \beta$; in the matrix form this is

$$\underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & -1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}}_{(2K-2)\times 2K} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = A_2 \gamma = 0,$$

with $\text{rank}(A_2) = 2K - 2$, so that the F-statistics is given by

$$F = \frac{(\text{SSE}_{H_2} - \text{SSE})/(2K-2)}{\text{SSE}/(N-2K)}.$$

To compute SSE_{H_2}, the design matrix for H_2, X_2, is obtained by adding together the first K columns of X and then the last K columns, i.e., the model reduces to

$$\mathbf{Y} = \underbrace{\begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{Kn_K} \end{pmatrix}}_{N\times 2} \underbrace{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}}_{2} + \boldsymbol{\epsilon} = \mathbf{X}_2 \boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}.$$

In practice, we would probably test for parallelism first and then, if H_1 is not rejected, test for H_2 (given that H_1 is true) using

$$F = \frac{(\text{SSE}_{H_2} - \text{SSE}_{H_1})/(K-1)}{\text{SSE}_{H_1}/(N-K-1)}.$$

If this is also not significant, then we can check this nested procedure using the F-statistic for testing H_2 (without restriction of H_1) as a final test statistic.

- Test for *concurrence* with x-coordinate known. To test H_3 that all the lines meet at a point on the y-axis ($x = 0$), that is, $H_3 : \alpha_1 = \cdots = \alpha_K (= \alpha)$, the F statistic is

$$F = \frac{(\text{SSE}_{H_3} - \text{SSE})/(K-1)}{\text{SSE}/(N-2K)},$$

which is the same as testing H_1. The design matrix X_3 is obtained by adding together the first K columns of X. An estimate α, i.e., the y coordinate of the point of occurrence is obtained automatically when the reduced model by H_3 is fitted.

- Test for the intersection. Consider two straight lines. We wish to derive an F-statistic for testing the hypothesis that their lines intersect at the point (a, b). Under this hypothesis, H, we have two models

$$\mathbb{E}(Y_{ki}) = b + \beta_k(x_{ki} - a), \quad k = 1, 2, \quad i = 1, \ldots, n_k,$$

so that least squares estimate of β_1 and β_2 and ESS_H can be obtained by minimizing

$$f(\beta_1, \beta_2) = \sum_{k=1}^{2} \sum_{i=1}^{n_i} [Y_{ki} - b - \beta_k(x_{ki} - a)]^2.$$

$\partial f / \partial \beta_k = 0$ $(k = 1, 2)$ give us

$$\widehat{\beta}_k = \frac{\sum_{i=1}^{n_1}(Y_{ki} - b)(x_{ki} - a)}{\sum_{i=1}^{n_k}(x_{ki} - a)^2}, \quad k = 1, 2.$$

The F statistic is given by

$$F = \frac{(\text{ESS}_H - \text{ESS})/2}{\text{ESS}/(n_1 + n_2 - 4)}.$$

- Test for the horizontal distance. Again, consider two straight lines to obtain an estimate and a CI for the horizontal distance between two *parallel* lines. Suppose two models with the same slope parameter:

$$E(Y_{ki}) = \alpha_k + \beta x_{ki}, \quad k = 1, 2, \quad i = 1, \ldots, n_k$$

and then the point estimate is given by

$$d = \frac{\alpha_2 - \alpha_1}{\beta} \quad \Rightarrow \quad \widehat{d} = \frac{\widehat{\alpha}_2 - \widehat{\alpha}_1}{\widehat{\beta}}.$$

The least squares estimates of α_1, α_2, and β can be obtained by minimizing

$$f(\alpha_1, \alpha_2, \beta) = \sum_{k=1}^{2} \sum_{i=1}^{n_i} (Y_{ki} - \alpha_k - \beta x_{ki})^2.$$

The three first derivatives equal to zero yield

$$\widehat{\alpha}_k = \overline{Y}_{k\cdot} - \widehat{\beta}\overline{x}_{k\cdot}, \quad k = 1, 2, \quad \widehat{\beta} = \frac{\sum_{k=1}^{2}\sum_{i=1}^{n_i} Y_{ki}(x_{ki} - \overline{x}_{k\cdot})^2}{\sum_{k=1}^{2}\sum_{i=1}^{n_i}(x_{ki} - \overline{x}_{k\cdot})^2}.$$

Suppose $U = \widehat{\alpha}_2 - \widehat{\alpha}_1 - d\widehat{\beta}$. Then $E(U) = 0$ and

$$\begin{aligned}\text{var}(U) = \sigma_U^2 &= \text{var}[\overline{Y}_{2\cdot} - \overline{Y}_{1\cdot} - \widehat{\beta}(\overline{x}_{2\cdot} - \overline{x}_{1\cdot} + d)] \\ &= \sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\overline{x}_{2\cdot} - \overline{x}_{1\cdot} + d)^2}{\sum_{k=1}^{2}\sum_{i=1}^{n_i}(x_{ki} - \overline{x}_{k\cdot})^2}\right], \quad \because \text{cov}(\overline{Y}_{k\cdot}, \widehat{\beta}) = 0 \\ &= \sigma^2 w\end{aligned}$$

Hence, the confidence limits for d are roots of $U^2 - F^{\alpha}_{1,n_1+n_2-3} S^2 w = 0$, where

$$S^2 = \frac{\text{SSE}_H}{n_1 + n_2 - 3} = \frac{\sum_{k=1}^{2}\sum_{i=1}^{n_i}(Y_{ki} - \widehat{\alpha}_k - \widehat{\beta} x_{ki})^2}{n_1 + n_2 - 3}.$$

5 Two phase regression

- Consider a linear two-phase model, where it undergoes a change in slope at $x = \gamma$. Assuming the change is continuous, we have the model

$$E(y) = \begin{cases} \alpha_1 + \beta_1 x, & x < \tau \\ \alpha_2 + \beta_2 x, & x \geq \tau \end{cases},$$

where continuity requires that $\alpha_1 + \beta_1\tau = \alpha_2 + \beta_2\tau$ $(= \theta)$. Note that γ may be (1) known, (2) unknown but know to line between two observed values of x, and (3) completely unknown. Some say γ the changeover point and θ the changeover value.

- (1) γ (Changeover Point) is known. Given n_1 obs on the first line and n_2 on the second, the two models can be written as

$$\begin{aligned} Y_{1i} &= \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i}, \quad i = 1, \ldots, n1, \\ Y_{2i} &= \alpha_1 + \beta_1\gamma + \beta_2(x_{2i} - \gamma) + \epsilon_{2i}, \quad i = 1, \ldots, n_2, \end{aligned}$$

where $x_{11} < x_{12} < x_{1n_1} < \gamma < x_{21} < x_{22} < x_{2n_2}$. Using $\mathbf{Y}_1, \mathbf{x}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{Y}_2, \mathbf{x}_2 \in \mathbb{R}^{n_2}$, we can write

$$\mathbf{Y} = \begin{pmatrix}\mathbf{Y}_1 \\ \mathbf{Y}_2\end{pmatrix} = \underbrace{\begin{pmatrix}\mathbf{1}_{n_1} & \mathbf{x}_1 & \mathbf{0} \\ \mathbf{1}_{n_2} & \gamma 1_{n_2} & \mathbf{x}_2 - \gamma 1_{n_2}\end{pmatrix}}_{(n_1+n_2)\times 3} \begin{pmatrix}\alpha_1 \\ \beta_1 \\ \beta_2\end{pmatrix} + \begin{pmatrix}\boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2\end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Given a value of γ, we can find the least squares estimate $\widehat{\boldsymbol{\beta}} = (\widehat{\alpha}_1, \widehat{\beta}_1, \widehat{\beta}_2)'$. Also, the estimate $\widehat{\theta} = \widehat{\alpha}_1 + \widehat{\beta}_1\gamma = (1, \gamma, 0)\widehat{\boldsymbol{\beta}} = \mathbf{a}'\widehat{\boldsymbol{\beta}}$, so that we can construct a t-CI for $a'\widehat{\boldsymbol{\beta}}$ in the usual manner.

If we want to test $H : \gamma = c$, where $x_{1n_1} < c < x_{21}$, then testing H is equivalent to testing for concurrence at $x = c$ when $K = 2$.

- (2) Unknown Changeover Point but known to the interval with it. In practice γ will be unknown but suppose it is known that $x_{1n_1} < \gamma < x_{21}$, then γ can be estimated by

$$\gamma = -\frac{\alpha_1 - \alpha_2}{\beta_1 - \beta_2} \quad \Rightarrow \quad \widehat{\gamma} = -\frac{\widehat{\alpha}_1 - \widehat{\alpha}_2}{\widehat{\beta}_1 - \widehat{\beta}_2},$$

where $\widehat{\alpha}_k$ and $\widehat{\beta}_k$ are usual least squares estimates. Since the radio of two correlated normal variables, we can use Filler's method for finding a CI for γ as follows. Suppose $U = (\widehat{\alpha}_1 - \widehat{\alpha}_2) + \gamma(\widehat{\beta}_1 - \widehat{\beta}_2)$. Then $\mathbb{E}(U) = (\alpha_1 - \alpha_2) + \gamma(\beta_1 - \beta_2) = 0$ and

$$\begin{aligned}\text{var}(U) &= \text{var}[(\widehat{\alpha}_1 - \gamma\widehat{\beta}_1) - (\widehat{\alpha}_2 - \tau\widehat{\beta}_2)] \\ &= \text{var}[\overline{Y}_1 - \widehat{\beta}_1(\gamma - \overline{x}_1)] + \text{var}[\overline{Y}_2 - \widehat{\beta}_1(\gamma - \overline{x}_2)] \\ &= \sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\gamma - \overline{x}_1)^2}{\sum_{i=1}^{n_1}(x_{1i} - \overline{x}_1)^2} + \frac{(\gamma - \overline{x}_2)^2}{\sum_{i=1}^{n_2}(x_{2i} - \overline{x}_2)^2}\right] = \sigma^2 w\end{aligned}$$

as $\overline{Y}_k \perp\!\!\!\perp \widehat{\beta}_1$ $(k = 1, 2)$. Thus, $100(1-\alpha)\%$ CI for τ is given by the roots of $U^2 = F^{\alpha}_{1,n-4} S^2 w$, i.e.,

$$\left[(\widehat{\alpha}_1 - \widehat{\alpha}_2) + \gamma(\widehat{\beta}_1 - \widehat{\beta}_2)\right]^2 - F^{\alpha}_{1,n-4} S^2 \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\gamma - \overline{x}_1)^2}{\sum_{i=1}^{n_1}(x_{1i} - \overline{x}_1)^2} + \frac{(\gamma - \overline{x}_2)^2}{\sum_{i=1}^{n_2}(x_{2i} - \overline{x}_2)^2}\right] = 0,$$

where $S^2 = (\|\mathbf{Y}_1 - \widehat{\mathbf{Y}}_1\|^2 + \|\mathbf{Y}_2 - \widehat{\mathbf{Y}}_2\|^2)/(n_1 + n_2 - 4)$.

- (3) γ is completely unknown. If $\widehat{\gamma}$ does not lie in the interval (x_{1n_1}, x_{21}), then the experimenter must decide whether to attribute this to sampling errors or to an incorrect assumption about the position of γ. When the position of γ is unknown, the problem becomes much more difficult, as it is now nonlinear.

6 Multiple Correlation coefficient

- Consider $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$. If the model is adequate, then $\mathrm{Corr}(Y, \widehat{Y})$ should be high:

$$r = \mathrm{Corr}(Y, \widehat{Y}) = \frac{\sum_i (Y_i - \overline{Y})(\widehat{Y}_i - \overline{\widehat{Y}})}{\sqrt{\sum_i (Y_i - \overline{Y})^2 \sum_i (\widehat{Y}_i - \overline{\widehat{Y}})^2}}.$$

Note that $\overline{\widehat{Y}} = 1_n' \widehat{Y}/n = 1_n' PY/n = 1_n' Y/n = \overline{Y}$ if the intercept is included, i.e., $1_n \in C(X)$. Then

$$\sum_i (Y_i - \overline{Y})(\widehat{Y}_i - \overline{\widehat{Y}}) = \sum_i (Y_i - \overline{Y})(\widehat{Y}_i - \overline{Y}) = \sum_i (Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y})(\widehat{Y}_i - \overline{Y}) = \sum_i (\widehat{Y}_i - \overline{Y})^2$$

since $\sum_i (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) = Y'(I - P)(P - 11'/n)Y = 0$. Hence,

$$r = \frac{\sum_i (\widehat{Y}_i - \overline{Y})^2}{\sqrt{\sum_i (Y_i - \overline{Y})^2 \sum_i (\widehat{Y}_i - \overline{Y})^2}} = \sqrt{\frac{\sum_i (\widehat{Y}_i - \overline{Y})^2}{\sum_i (Y_i - \overline{Y})^2}} = \sqrt{\frac{\mathrm{SSReg}}{\mathrm{TSS}}} = \sqrt{R^2}.$$

- Suppose $E(y) = \beta_0 + \sum_{i=1}^{p-1} \beta_i x_i$. Then, the F-statistic for testing $H_0 : \beta_i = 0$ $(i = 1, \ldots, p-1)$ is

$$F = \frac{(\mathrm{SSE}_{H_0} - \mathrm{SSE})/(p-1)}{\mathrm{SSE}/(n-p)}.$$

Note that $\mathrm{SSE}_{H_0} = \min \sum_i (Y_i - \beta_0)^2 = \sum_i (Y_i - \overline{Y})^2 = \mathrm{TSS}$, so that

$$F = \frac{(\mathrm{TSS} - \mathrm{SSE})/(p-1)}{\mathrm{SSE}/(n-p)} = \frac{n-p}{p-1} \frac{1 - \mathrm{SSE}/\mathrm{TSS}}{\mathrm{SSE}/\mathrm{TSS}} = \frac{n-p}{p-1} \frac{1 - (1 - R^2)}{1 - R^2} = \frac{n-p}{p-1} \frac{R^2}{1 - R^2} \sim F_{p-1,n-p}$$

We reject H_0 if $F > F^{\alpha}_{p-1,n-p}$.

- Show that R^2 has a beta distribution and hence $E(R^2) = p - 1/(n-1)$. See HW.
- Change the notation to $E(Y) = \sum_{j=1}^{k} x_j \beta_j = \theta$. Define linear subsets

$$V_k = \mathcal{L}(x_1, \ldots, x_k), \quad V_{k-1} = \mathcal{L}(x_1, \ldots, x_{k-1})$$

and $\widehat{Y}_k = P_{C(X)} Y = P(Y \mid V_k)$ and $\widehat{Y}_{k-1} = P(Y \mid V_{k-1})$. Further let $x_k^{\perp} = x_k - P(x_k \mid V_{k-1}) = x_k - \widehat{x}_k$. $x_k^{\perp}$, called a signal of x_k, accounts for the linear relationship between y and x_k beyond these already explained by $x_1, \ldots, x_{k-1}$.

Note that $\widehat{x}_k \in V_{k-1}$ and so $x_k^{\perp} \in V_{k-1}^{\perp}$. Thus, we have

$$\|x_k^{\perp}\|^2 = (x_k^{\perp}, x_k^{\perp}) = (x_k - \widehat{x}_k, x_k^{\perp}) = (x_k, x_k^{\perp}) - (\widehat{x}_k, x_k^{\perp}) = (x_k, x_k^{\perp}),$$

$$(\theta, x_k^{\perp}) = \left(\sum_{j=1}^{k-1} x_j \beta_j, x_k^{\perp}\right) = \beta_k (x_k, x_k^{\perp}) = \beta_k \|x_k^{\perp}\|^2 \quad \because x_j \perp x_k^{\perp},\ j = 1, \ldots, k-1.$$

Hence, we have

$$\beta_k = \frac{(\theta, x_k^\perp)}{\|x_k^\perp\|^2} \quad \Rightarrow \quad \widehat{\beta}_k = \frac{(\widehat{Y}, x_k^\perp)}{\|x_k^\perp\|^2} = \frac{(\widehat{Y} - Y + Y, x_k^\perp)}{\|x_k^\perp\|^2} = \frac{(Y, x_k^\perp)}{\|x_k^\perp\|^2}$$

since $\widehat{Y} - Y \in V_k^\perp \Rightarrow (\widehat{Y} - Y, x_k^\perp) = (\widehat{Y} - Y, x_k - \widehat{x}_k) = (\widehat{Y} - Y, x_k) - (\widehat{Y} - Y, \widehat{x}_k) = 0$.

- $\widehat{Y}_k = \widehat{Y}_{k-1} + \widehat{\beta}_k x_k^\perp$ and $P_{C(X)} = P_1 + Q_1 X_2 (X_2' Q_1 X_2)^{-1} X_2' Q_1$, where $P_1 = P_{C(X_1)}$, and $Q_1 = Q_{C(X_1)}$.

  Proof: If X petitions $X = (X_1 \mid X_2)$, then we have $E(y) = X_1\beta_1 + X_2\beta_2$. Then we have

$$\begin{aligned}
\widehat{\beta}_2 &= (X_2' Q_1 X_2)^{-1} X_2' Q_1 y, \\
\widehat{\beta}_1 &= (X_1' X_1)^{-1} X_1' (y - X_2 \widehat{\beta}_2) = (X_1' X_1)^{-1} X_1' [y - X_2 (X_2' Q_1 X_2)^{-1} X_2' Q_1 y], \\
\Rightarrow P_{C(X)} y &= X_1 \widehat{\beta}_1 + X_2 \widehat{\beta}_2 \\
&= P_1 [y - X_2 (X_2' Q_1 X_2)^{-1} X_2' Q_1 y] + X_2 (X_2' Q_1 X_2)^{-1} X_2' Q_1 y \\
&= P_1 y + Q_1 X_2 (X_2' Q_1 X_2)^{-1} X_2' Q_1 y \\
&= [P_1 + Q_1 X_2 (X_2' Q_1 X_2)^{-1} X_2' Q_1] y, \ \forall y.
\end{aligned}$$

  Hence, $P_{C(X)} = P_1 + Q_1 X_2 (X_2' Q_1 X_2)^{-1} X_2' Q_1$. Using this, we have

$$\widehat{y}_k = \widehat{y}_{k-1} + Q_1 x_k (x_k' Q_1 x_k)^{-1} x_k' Q_1 y = \widehat{y}_{k-1} + \frac{x_k^\perp (x_k^\perp, y)}{\|x_k^\perp\|^2} = \widehat{y}_{k-1} + \widehat{\beta}_k x_k^\perp.$$

- We know $\mathrm{Var}(\widehat{\beta}) = \sigma^2 (X'X)^{-1}$, that is, $\mathrm{cov}(\widehat{\beta}_i, \widehat{\beta}_j) = \sigma^2 (X'X)^{-1}_{ij}$. But using the above notation,

$$\mathrm{cov}(\widehat{\beta}_i, \widehat{\beta}_j) = \mathrm{cov}\left(\frac{(Y, x_i^\perp)}{\|x_i^\perp\|^2}, \frac{(Y, x_j^\perp)}{\|x_j^\perp\|^2} \right) = \sigma^2 \frac{(x_i^\perp, x_j^\perp)}{\|x_i^\perp\|^2 \|x_j^\perp\|^2}, \quad \mathrm{var}(\widehat{\beta}_i) = \frac{\sigma^2}{\|x_i^\perp\|^2},$$

  Compared with the well-known result, we have

$$(X'X)^{-1}_{ij} = \frac{(x_i^\perp, x_j^\perp)}{\|x_i^\perp\|^2 \|x_j^\perp\|^2}.$$

  Also, we observe that if $x_i^\perp$ is nearly zero, then $\mathrm{var}(\widehat{\beta}_i)$ is inflated. Since $x_i^\perp = x_k - P(x_k \mid V_{k-1})$, if $x_k^\perp \approx 0$, then $x_k \approx P(x_k \mid V_{k-1})$, meaning that x_k can be written as nearly linear combination of $x_1, \ldots, x_{k-1}$ (Multicolinearlity).

- Compare $H_0 : \beta_k = 0$ vs H_A. That is, the reduced model does not include x_k. We have

$$\mathrm{SSE}_{k-1} = \|y - \widehat{y}_{k-1}\|^2 = \|y\|^2 - \|\widehat{y}_{k-1}\|^2, \quad \mathrm{SSE}_k = \|y - \widehat{y}_k\|^2 = \|y\|^2 - \|\widehat{y}_k\|^2$$

  because $(y, \widehat{y}_i) = (y, P(y_i \mid V_i)) = (P(y_i \mid V_i), P(y_i \mid V_i)) = \|\widehat{y}_i\|^2$, $i = k-1, k$. Hence, the F-statistic for testing H_0 is

$$F = t^2 = \frac{(\mathrm{SSE}_{k-1} - \mathrm{SSE}_k)/1}{\mathrm{SSE}_k/(n-k)} = \frac{\|\widehat{y}_k\|^2 - \|\widehat{y}_{k-1}\|^2}{\mathrm{SSE}_k/(n-k)} = \frac{\widehat{\beta}_k^2 \|\widehat{x}_k^\perp\|^2}{\mathrm{SSE}_k/(n-k)} = \frac{\widehat{\beta}_k^2}{\widehat{\sigma}^2/\|\widehat{x}_k^\perp\|^2}$$

  since $\|\widehat{y}_k\|^2 = \|\widehat{y}_{k-1}\|^2 + \widehat{\beta}_k^2 \|x_k^\perp\|^2$ by $(\widehat{y}_{k-1}, x_k^\perp) = 0$. Further, define R_k^2 and R_{k-1}^2 as

$$R_i^2 = 1 - \frac{\mathrm{SSE}_i}{\mathrm{TSS}}, \quad i = k-1, k.$$

  and let $d = t^2/(n-k)$, then

$$d = \frac{\mathrm{SSE}_{k-1} - \mathrm{SSE}_k}{\mathrm{SSE}_k} \quad \Rightarrow \quad \mathrm{SSE}_{k-1} = (1+d)\mathrm{SSE}_k$$

so that

$$\frac{R_k^2 - R_{k-1}^2}{1 - R_{k-1}^2} = \frac{\text{SSE}_{k-1} - \text{SSE}_k}{\text{SSE}_{k-1}} = \frac{d}{1+d},$$

which can be interpreted as the proportional reduction in SSE. We see that when $d \to 0$ ($t \to 0$), then H is not rejected, as $R_k^2 \to R_{k-1}^2$. Conversely, when d is large, H is rejected and R_k is closed to 1.

- If r is the partial correlation of Y and x_k with the effects of $x_1, x_2, \ldots, x_{k-1}$ removed, i.e.,

$$\begin{aligned} r &= \frac{(Y^\perp, x_k^\perp)}{\|Y^\perp\|\|x_k^\perp\|}, \quad \text{where} \quad Y^\perp = Y - \widehat{Y}_{k-1} \in V_{k-1}^\perp, \quad x_k^\perp \in V_{k-1}^\perp \\ &= \frac{(Y, x_k^\perp)}{\|Y - \widehat{Y}_{k-1}\|\|x_k^\perp\|} \quad \because (Y^\perp, x_k^\perp) = (Y, x_k^\perp) - (\widehat{Y}_{k-1}, x_k^\perp) = (Y, x_k^\perp), \end{aligned}$$

then

$$r^2 = \frac{(Y, x_k^\perp)^2 / \|x_k^\perp\|^2}{\|Y - \widehat{Y}_{k-1}\|^2} = \frac{\text{ESS}_{k-1} - \text{ESS}_k}{\text{ESS}_{k-1}} = \frac{R_k^2 - R_{k-1}^2}{1 - R_{k-1}^2} = \frac{d}{1+d}.$$

7 Partial Correlation Coefficient (PCC)

- Let $V = \mathcal{L}(x_1, \ldots, x_k)$. If $v_1 \notin V$ and $v_2 \notin V$, the partial correlation coefficient (PCC) for v_1 and v_2 with $x_1, \ldots, x_k$ removed is given by

$$r_{v_1, v_2 \cdot x_1, \ldots, x_k} = \frac{(v_1 - P(v_1 \mid V), v_2 - P(v_2 \mid V))}{\|v_1 - P(v_1 \mid V)\|\|v_2 - P(v_2 \mid V)\|} = \frac{(v_1^\perp, v_2^\perp)}{\|v_1^\perp\|\|v_2^\perp\|},$$

which is undefined otherwise ($v_1 \in V$ or $v_2 \in V$) since $v_i^\perp$ becomes 0.

  - Note that PCC is independent of the lengths of v_1 and v_2.
  - We say the notation $r_{v_1, v_2 \cdot x_1, \ldots, x_k}$ is of order k.

- Let I and J are two sets of indices such that $J \subset I$. Specifically, WLOG. suppose $I = \{3, 4, \ldots, k\}$ and $J = \{4, \ldots, k\}$ and define V_I and V_J. Then

$$r_{12,I} = \frac{r_{12,J} - r_{13,J}\, r_{23,J}}{\sqrt{(1 - r_{13,J}^2)(1 - r_{23,J}^2)}}.$$

Proof: Use the fact that $(x_3^\perp, x_i) = (x_3^\perp, x_i^\perp)$ and

$$P(x_i \mid V_I) = P(x_i \mid V_J) + \frac{Q_3 x_3 x_3' Q_3}{x_3' Q_3 x_3} x_i$$

- True or False. The value of the partial correlation of two regressors in a standard linear regression model is not larger than their simple correlation value. *Answer*: False; since correlations can be negative. It will be true if you compare the absolute value of their correlations.

8 Polynomial Regression

- Orthogonal Polynomial. Consider the model

$$Y_i = \gamma_0 \phi_0(x_i) + \gamma_1 \phi_1(x_i) + \cdots + \gamma_k \phi_k(x_i) + \epsilon_i,$$

where $\phi_r(x_i)$ is an rth degree polynomial in x_i $(r = 0, 1, \ldots, k)$ with orthogonal constraint

$$\sum_{i=1}^{n} \phi_r(x_i)\phi_s(x_i) = 0, \quad r \neq s.$$

Then $Y = X\gamma + \epsilon$, where

$$X = \underbrace{\begin{pmatrix} \phi_0(x_1) & \cdots & \phi_k(x_1) \\ \vdots & \ddots & \vdots \\ \phi_0(x_n) & \cdots & \phi_k(x_n) \end{pmatrix}}_{n \times (k+1)},$$

so that from $\widehat{\gamma} = (X'X)^{-1}X'Y$, we have

$$\widehat{\gamma}_r = \frac{\sum_{i=1}^{n} \phi_r(x_i)Y_i}{\sum_{i=1}^{n} \phi_r^2(x_i)}, \quad r = 0, 1, \ldots, k,$$

which holds for *all* k. The orthogonal structure of X implies that the least squares estimate of γ_r is independent of the degree k of the polynomial. Since $\phi_0(x_i)$ is a polynomial of degree zero, we can set $\phi_0(x) \equiv 1$ and obtain $\widehat{\gamma}_0 = \overline{Y}$. Using this, the residual sum of squares is given by

$$\begin{aligned} \text{SSE}_{k+1} &= (Y - X\widehat{\gamma})'(Y - X\widehat{\gamma}) \\ &= Y'Y - 2Y'X\widehat{\gamma} + \widehat{\gamma}'X'X\widehat{\gamma} \\ &= Y'Y - \widehat{\gamma}'X'X\widehat{\gamma} \quad \because X'X\widehat{\gamma} = X'Y \\ &= \sum_i Y_i^2 - \sum_{r=0}^{k} \left[\sum_{i=1}^{n} \phi_r^2(x_i)\right] \widehat{\gamma}_r^2 \\ &= \sum_i (Y_i - \overline{Y})^2 - \sum_{r=1}^{k} \left[\sum_{i=1}^{n} \phi_r^2(x_i)\right] \widehat{\gamma}_r^2. \end{aligned}$$

If we wish to test $H : \gamma_k = 0$, then

$$\begin{aligned} \text{SSE}_k &= \sum_i (Y_i - \overline{Y})^2 - \sum_{r=1}^{k-1} \left[\sum_{i=1}^{n} \phi_r^2(x_i)\right] \widehat{\gamma}_r^2 \\ &= \text{SSE}_{k+1} + \left[\sum_{i=1}^{n} \phi_k^2(x_i)\right] \widehat{\gamma}_k^2 \end{aligned}$$

and the F-statistic is

$$F = \frac{(\text{SSE}_k - \text{SSE}_{k+1})/1}{\text{SSE}_{k+1}/(n-k-1)} = \frac{\left[\sum_{i=1}^{n} \phi_k^2(x_i)\right] \widehat{\gamma}_k^2}{\text{SSE}_{k+1}/(n-k-1)}.$$

The problem is how to choose the degree, K.

- *Equally Spaced* x- values. Suppose that x-values are equally spaced so that they can be transformed to

$$x_i = i - \frac{n+1}{2}, \quad i = 1, \ldots, n.$$

Then we have the system of orthogonal polynomials. Suppose that $n = 3$ (up to 2 degrees). Then $x_i = -1, 0, 1$, $\phi_0(x) = 1$, $\phi_1(x) = \lambda_1 x = x$, $\phi_2(x) = \lambda_2(x^2 - 2/3) = 3x^2 - 2$, where λ_1 and λ_2 are chosen so that the values of $\phi_r(x_i)$ are all integers, and the fitted polynomial is

$$\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2(3x^2 - 2).$$

- *Fractional* Polynomials. This was proposed by Royston and Altman (See an attached paper). A reasonably adequate set of powers to consider are

$$\left\{-2, -1, \frac{1}{2}, 0, \frac{1}{2}, 1, 2 \ldots, \max(3, m)\right\}$$

See HW2 Question 4 for more detail.

- *Piecewise* Polynomial. Sometimes a polynomial fit is unsatisfactory even when orthogonal polynomials up to, say, 20 are fitted. One symptom is the failure of SSE_k to settle down to constant values as k increases. Another symptom is the behavior of the residuals by a residual plot, where $r_i = Y_i - \widehat{Y}_i$ versus x_i will continue to exhibit a systematic pattern instead of a random. An alternative approach to the problem is to divide up the range of x into segments and fit a low-degree polynomial in each segment. The most useful method employs the theory of spline functions.

  - *Spline* Functions. We define the spline function $s(x)$ of order M (degree $M-1$) with knots $\xi_1, \ldots, \xi_K$ ($\xi_1 < \cdots < \xi_K$) and having domain $[a, b]$ ($-\infty \le a < \xi_1,\ \xi_K < b \le \infty$) to be a function with the following properties:

    1. In each of the intervals, $s(x)$ is a polynomial of degree $M-1$ at most.

    2. $s(x)$ and its derivatives up to order $(M-2)$ are continuous.

    The cubic spline ($M = 4$) is a satisfactory function for fitting data in most situations, and second-derivative continuity is usually adequate.

  - *Smoothing Splines.* We want to avoid the problem of knot selection completely by using a maximal set of knots. The first idea would be to use the least squares criterion of minimizing $\text{SSE}(f) = \sum_i [Y_i - f(x_i)]^2$. However, this could lead to overfitting; the piecewise linear graph would not be smooth, as its derivative do not exist. What we would like to do is impose a penalty function that measures the degree of roughness. Since we would not want our measure to be affected by the addition of a constant or a linear function, we could utilize the second derivative f''. Combining the two ideas of least squares and roughness leads to the criterion of finding f which minimizes

$$\begin{aligned}\text{SSE}(f) &= \sum_i [Y_i - f(x_i)]^2 + \lambda \int_a^b [f''(t)]^2 dt \\ &= \text{Closeness of fit} + \text{Smoothness},\end{aligned}$$

    which is called the penalized residual sum of squares. If λ (smoothing parameter) is zero, then f is just the ESS. which leads to a wiggly or rough graph. If $\lambda = \infty$, $f'' = 0$ and we get the least squares fit of a straight line. As λ ranges from 0 to ∞, f can very from very rough to very smooth.

9 Hypothesis testing with less-than-full-rank X

- Suppose $Y = \theta + \epsilon$, where $\theta \in \Omega$ (an r-dimensional subspace of $\mathbb{R}^n$), and wish to test $H : \theta \in \omega$, where ω is an $(r-q)$-dimensional subspace of Ω. When H is true and $\epsilon \sim N_n(0, \sigma^2 I_n)$, then the F-statistic is

$$\begin{aligned}F = \frac{(\text{SSE}_H - \text{SSE})/q}{\text{SSE}/(n-r)} &= \frac{[Y'(I_n - P_\omega)Y - Y'(I_n - P_\Omega)Y)/q}{Y'(I_n - P_\Omega)Y/(n-r)} \\ &= \frac{Y'(P_\Omega - P_\omega)Y/q}{Y'(I_n - P_\Omega)Y/(n-r)} \\ &= \frac{\epsilon'(P_\Omega - P_\omega)\epsilon/q}{\epsilon'(I_n - P_\Omega)\epsilon/(n-r)} \ \sim \ F_{q,n-r}\end{aligned}$$

since

  - $P_\Omega \theta = P_\omega \theta = \theta$ under $H \Rightarrow (P_\Omega - P_\omega)Y = (P_\Omega - P_\omega)(\theta + \epsilon) = (P_\Omega - P_\omega)\epsilon$ and $(I_n - P_\Omega)Y = (I_n - P_\Omega)\epsilon$.

– $P_\Omega - P_\omega \perp\!\!\!\perp I_n - P_\Omega$ and $\epsilon \sim N_n(0, \sigma^2 I_n) \Rightarrow \epsilon'(P_\Omega - P_\omega)\epsilon$ and $\epsilon'(I_n - P_\Omega)\epsilon$ are independent χ^2 distribution by Craig's theorem.

- In one-way ANOVA:

$$\underbrace{\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix}}_{N\times 1} = \underbrace{\begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1_{n_k} & 0 & 0 & \cdots & 1_{n_k} \end{pmatrix}}_{N\times(k+1)} \begin{pmatrix} \alpha \\ \tau_1 \\ \vdots \\ \tau_k \end{pmatrix} = X\beta$$

  If $\sum_{i=1}^k c_i\tau_i$ is estimable, then we only focus on estimable functions. Thus, we want to identify them.

- Let $Y = X\beta + \epsilon$ and the error has the usual assumptions and $\text{rank}(X) = r(\leq p)$. Find H such that $H\beta = 0$, where $H \in R^{(p-r)\times p}$. H is called *identifiable constraints*. Suppose that rows of H and rows of X are all linearly independent, Let $G = (X' \mid H')' : (n + p - r) \times p$ then $G'G = X'X + H'H$.

  (i) $G'G$ is invertible. *Proof*: Since $\text{rank}(G) = p$, G has full column rank, i.e., $G'G$ is nonsingular.

  (ii) $X'X(G'G)^{-1}H' = 0$. *Proof*: $G'G(G'G)^{-1}H' = H' \Rightarrow X'X(G'G)^{-1}H' = H'(I - H(G'G)^{-1}H') \Rightarrow X'X(G'G)^{-1}H'a = H'(I - H(G'G)^{-1}H')a = 0$, $\forall a$ by assumption $(C(X') \cap C(H') = \{0\})$. Hence $X'X(G'G)^{-1}H' = O$.

  (iii) $H(G'G)^{-1}H' = I$. By (ii), $H'(I - H(G'G)^{-1}H')a = O$, $\forall a$. Since H has full row rank, $(I - H(G'G)^{-1}H')a = 0 \Rightarrow H(G'G)^{-1}H' = I$.

  (iv) $(G'G)^{-1}$ is a g-inverse of $X'X$, i.e., $(G'G)^{-1} = (X'X)^-$. *Proof*: $X'X(G'G)^{-1}X'X = X'X(G'G)^{-1}(G'G - H'H) = X'X - X'X(G'G)^{-1}H'H = X'X$ by (i).

  (v) $E(\widehat{\beta}) = \beta$. *Proof*: $E(\widehat{\beta}) = (X'X)^-X'X\beta = (G'G)^{-1}(G'G - H'H)\beta = \beta$ since $H\beta = 0$.

- Thus, even if X has less than full rank, the model with constraints $H\beta = 0$:

$$\begin{pmatrix} \theta \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ H \end{pmatrix}\beta = G\beta$$

  has a unique $\widehat{\beta}$, which is unbiased for β because G is nonsingular.

10 ANOVA in Seber's textbook

- When the regressors are qualitative so that indicator variables are involved (taking values 0 or 1), then we refer to the model as an ANOVA model.

- In ANOVA models, if we begin with $E(Y) = X\beta = \theta$, where $\theta \in \Omega$, then we find a matrix X such that this space is $\mathcal{C}(X)$. Clearly, such a representation is not unique, as $\theta = X\beta = XB^{-1}B\theta = X_B\gamma$, where B is a $p \times p$ nonsingular matrix. The choice of X_B depends on which linear combination of the β's we are interested in. Typically, we are interested in just the individual β_j or in *contrasts* $\sum_i c_j\beta_j$, where $\sum_j c_j = 0$, for example $\beta_1 - \beta_2$.

- Consider $Y_{ij} = \mu_i + \epsilon_{ij},\ i = 1, \ldots, a,\ j = 1, \ldots, n_i$, where the error term are normal. Let $n = \sum_{i=1}^a n_i$ If we wish to test $H : \mu_1 = \mu_2 = \cdots = \mu_I\ (= \mu$, say), then $\theta = X_H\mu = 1_n\mu$. Hence the F-statistic is

$$F = \frac{(\text{SSE}_H - \text{SSE})/(a-1)}{\text{SSE}/(n-a)} = \frac{\sum_i n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2/(a-1)}{\sum_i\sum_j(Y_{ij} - \overline{Y}_{i.})^2/(n-a)} = \frac{S_H^2}{S^2} \sim F_{a-1,n-a},$$

when H is true, where $a-1$ means the difference in ranks or # of parameters between X and X_H and

$$\text{SSE} = \sum_i \sum_j (Y_{ij} - \overline{Y}_{i.})^2, \quad \text{SSE}_H = \sum_i \sum_j (Y_{ij} - \overline{Y}_{..})^2,$$

as $\widehat{\mu}_i = \overline{Y}_{i.}$ on Ω and $\widehat{\mu}_i = \overline{Y}_{..}$ under H. We can also express $F = S_H^2/S^2$.

- The hypothesis H can also be expressed in the form $H : \mu_i - \mu_a = 0$, $i = 1, \ldots, a-1$, i.e., we have $k-1$ contrasts $\alpha_i = \mu_i - \mu_a$. Putting $\mu_k = \mu$, we have $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, or $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where $\boldsymbol{\gamma} = (\mu, \alpha_1, \ldots, \alpha_{a-1})$ and $\mathbf{X}_1$ is of rank a with $H : \alpha_i = \cdots = \alpha_{a-1} = 0$.
- Other reparameterization are to define $\mu = \sum_i \mu_i/k$ or $\mu = \sum_i n_i\mu_i/n$.
- Exercise. Find $E[\sum_i \sum_j (\overline{Y}_{i.} - \overline{Y}_{..})^2]$ and $E[\sum_i \sum_j (Y_{ij} - \overline{Y}_{i.})^2]$.

  Solution: $Q_1 = \sum_i \sum_j (\overline{Y}_{i.} - \overline{Y}_{..})^2 = Y'AY$, where

$$A = \text{diag}\left(\frac{1_{n_1}1'_{n_1}}{n_1}, \ldots, \frac{1_{n_a}1'_{n_a}}{n_k}\right) - \frac{1_n 1'_n}{n}$$

  is symmetric and idempotent with $\text{rank}(A) = \text{tr}(A) = k-1$. Hence,

$$\begin{aligned} E(Q_1) &= \sigma^2(a-1) + [Y'AY]_{Y_{ij}=\mu_i} \\ &= \sigma^2(a-1) + \sum_i \sum_j \left(\mu_i - \frac{\sum_i n_i\mu_i}{n}\right)^2. \end{aligned}$$

  Likewise, $Q_2 = \sum_i \sum_j (Y_{ij} - \overline{Y}_{i.})^2 = Y'BY$, where

$$B = I_n - \text{diag}\left(\frac{1_{n_1}1'_{n_1}}{n_1}, \ldots, \frac{1_{n_a}1'_{n_a}}{n_k}\right)$$

  is symmetric and idempotent with $\text{rank}(B) = \text{tr}(B) = n-a$. Hence,

$$E(Q_2) = \sigma^2(n-a) + [Y'BY]_{Y_{ij}=\mu_i} = \sigma^2(n-a).$$

- **Confidence Intervals**. Consider one contrast $\theta = \sum_i c_i\mu_i$. Since $\widehat{\mu}_i = \overline{Y}_{i.}$ and $\text{var}(\widehat{\theta}) = \sigma^2 \sum_i c_i^2/n_i$, a $100(1-\alpha)\%$ CI for θ is given by

$$\sum_i c_i\overline{Y}_{i.} \pm t_{n-a}^{\alpha/2} S \left(\sum_i c_i^2/n_i\right)^{1/2}, \quad \sum_i c_i = 0$$

  where $S^2 = \text{SSE}/(n-a)$.

  - If we are interested in k contrasts, the Bonferroni intervals are given using $t_{n-a}^{\alpha/(2k)}$.
  - If we are interested in all possible contrasts, then Sheffe's method is appropriate. Since H can be expressed as $\phi_i = \mu_i - \mu_a = 0$ $(i = 1, \ldots, a-1)$, we know that set of all possible linear combinations of ϕ_i is the same as the set of all possible contrasts in the μ_i . Hence, $c'\mu$ lies in the interval

$$\sum_i c_i\overline{Y}_{i.} \pm [(a-1)F_{a-1,n-a}^{\alpha}]^{1/2} S \left(\sum_i c_i^2/n_i\right)^{1/2}, \quad \forall c \text{ s.t. } 1'c = 0.$$

    with an overall probability of $1-\alpha$. H is rejected iff at least oneof these CIs does not contain 0.
  - *Tukey Method*. Some of the above CIs can be quite wide. If we are interested in just the set of pairwise contrasts $\mu_r - \mu_s$ for all r and s $(r \neq s)$, then the Tukey intervals are given by

$$\overline{Y}_{r.} - \overline{Y}_{s.} \pm q_{a,n-a}^{\alpha} S \sqrt{\frac{1}{2}\left(\frac{1}{n_r} + \frac{1}{n_s}\right)},$$

    where $q_{a,n-a}^{\alpha}$ is the upper α quantile of the *Studentized range distribution*.

– If the sample size are all equal ($n_i = n$), then the simultaneous CIs for all contrasts $\sum_{i=1}^{a} c_i\mu_i$ is

$$\sum_i c_i \overline{Y}_{i.} \pm q^{\alpha}_{a,n-a} \frac{S}{\sqrt{n}} \sum_i \frac{|c_i|}{2}$$

- So far, we assume that σ^2 is the same for all the populations. Sheffe concluded that any heteroscedasticity of variance does not affect the F-test if the design is approximately balanced. However, the confidence intervals are not so robust against variance differences.
- Miscellaneous Exercise 8.2. Let $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ $(i = 1, \ldots, a; j = 1, \ldots, n)$, where $\sum_i d_i\alpha_i = 0$ $(\sum_i d_i \neq 0)$ and $E(\epsilon_{ij}) = 0$. Find the LS estimate of μ and α_i.

  Solution. Using the Lagrange multiplier, we define

$$\begin{aligned} f(\mu, \alpha_i) &= \sum_i \sum_j (Y_{ij} - \mu - \alpha_i)^2 + \lambda \sum_i d_i\alpha_i \\ \Rightarrow \quad \partial f/\partial\mu &= -2 \sum_i \sum_j (Y_{ij} - \mu - \alpha_i) \\ \Rightarrow \quad \partial f/\partial\alpha_k &= -2 \sum_j (Y_{ij} - \mu - \alpha_k) + \lambda d_k. \end{aligned}$$

  Both first derivatives zero leas to

$$\sum_i \lambda d_i = \lambda \sum_i d_i = 2 \sum_i \sum_j (Y_{ij} - \widehat{\mu} - \widehat{\alpha}_i) = 0 \quad \Rightarrow \quad \lambda = 0.$$

  Then $\sum_j Y_{ij} - b\widehat{\mu} - b\widehat{\alpha}_i = 0 \;\Rightarrow\; \widehat{\alpha}_i = \overline{Y}_{i.} - \widehat{\mu}$ and by $\sum_i d_i\widehat{\alpha}_i = 0$, $\widehat{\mu} = \sum d_i\overline{Y}_{i.} / \sum d_i$.
- Consider a two-way ANOVA model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

  for $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, n_{ij}$. Let $n = \sum_{ij} n_{ij}$. On the whole space, the least square estimate of μ_{ij} can be estimated by minimizing $\sum_i \sum_j \sum_k (Y_{ijk} - \mu_{ij})^2$, which leads to $\widehat{\mu}_{ij} = \overline{Y}_{ij.}$

  Then F-statistic for testing $H : \mu_{ij} = \mu$ for all i, j is

$$F = \frac{(\text{SSE}_H - \text{SSE})/(a-1)}{\text{SSE}/(n-a)} = \frac{\sum_i \sum_j n_{ij}(\overline{Y}_{ij.} - \overline{Y}_{...})^2/(ab-1)}{\sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij.})^2/(n-ab)} = \frac{S_H^2}{S^2} \;\sim\; F_{ab-1,n-ab},$$

  when H is true, as $\mu_{ij} = \overline{Y}_{...}$ under H.
- Missing Observations. Since unbalanced designs are problematical, a sensible approach might be to use a balanced design with some missing observations. The idea would be to put in suitable estimates of the missing values.
  – Exercise.1 Let

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \epsilon = X\beta + \epsilon,$$

  where X has full rank. Suppose that the obs Y_2 are missing. For the data observed, the LS estimate of β is $\widehat{\beta} = (X_1'X_1)^{-1}X_1'Y_1$. Let $\widehat{Y}_2 = X_2\widehat{\beta}$. Show that $\widehat{\beta}$ can be obtained by applying least squares to the observed data augmented by $\widehat{Y}_2$ and the regression matrix X.

  Solution: We have $X_1'X_1\widehat{\beta} = X_1'Y_1$ and $X_2'X_2\widehat{\beta} = X_2'\widehat{Y}_2$. Adding them provides

$$(X_1'X_1 + X_2'X_2)\widehat{\beta} = X_1'Y_1 + X_2'\widehat{Y}_2 \quad \Rightarrow \quad X'X\widehat{\beta} = X' \begin{pmatrix} Y_1 \\ \widehat{Y}_2 \end{pmatrix}.$$

– Exercise 2. Suppose a one-way classification $E(Y_{ij} = \mu_i)$ $(i = 1, \ldots, a; j = 1, \ldots, n)$, and that Y_{ab} is missing. Prove that the appropriate estimate of Y_{ab} is the mean of the remaining observations of the mean μ_a.

Solution. From the above perspective,

$$\widehat{Y}_{ab} = \overline{Y}_{a.} = \frac{\sum_{j=1}^{a-1} Y_{aj} + \widehat{Y}_{ab}}{n} \quad \Rightarrow \quad \widehat{Y}_{ab} = \frac{\sum_j Y_{a.}}{n-1}.$$

- ANCOVA. A general ANCOVA model takes the form

$$G : E[Y] = X\beta + Z\gamma = (X, Z)\begin{pmatrix}\beta \\ \gamma\end{pmatrix} = W\delta.$$

Exercise 8.3. Let $Y_{ijk} = \mu_{ij} + \gamma_{ij} z_{ijk} + \epsilon_{ijk}$, where $i = 1, \ldots, a$; $j = 1, \ldots, b$; $k = 1, \ldots n$; and the $\epsilon_{ijk} \sim N(0, \sigma^2)$. Obtain a test statistic for testing $H : \gamma_{ij} = \gamma$ for all i, j. *Solution*: Below.

$$F = \frac{(\text{SSE}_H - \text{SSE})/(ab-1)}{\text{SSE}/[ab(n-2)]}.$$

11 ANOVA in class

- Confidence intervals for σ^2, σ_τ^2, and ICC in a one-way random effect model. Assume $\tau_i \sim N(0, \sigma_\tau^2)$
  - For σ^2, since $\text{SSE}/\sigma^2 \sim \chi^2_{k(n-1)}$, $100(1-\alpha)\%$ C.I for σ^2 is given by

$$\chi^2_{k(n-1),\alpha/2} \le \frac{\text{SSE}}{\sigma^2} \le \chi^2_{k(n-1),1-\alpha/2} \quad \Rightarrow \quad \frac{\text{SSE}}{\chi^2_{k(n-1),1-\alpha/2}} \le \sigma^2 \le \frac{\text{SSE}}{\chi^2_{k(n-1),\alpha/2}}.$$

  - For the variance component σ_τ^2, although we have

$$E(\text{SSTr}) = (k-1)(n\sigma_\tau^2 + \sigma^2) \quad \Rightarrow \quad E(\text{MSTr}) = n\sigma_\tau^2 + \sigma^2$$

and hence

$$E\left(\frac{\text{MSTr} - \text{MSE}}{n}\right) = \sigma_\tau^2,$$

MSTr − MSE does not have an exact distribution. Thus, we *cannot* construct a CI for σ_τ^2 unless the bootstrapping is used.

  - For the intraclass correlation coefficient (ICC), we first need to verify

$$\frac{\text{SSTr}}{n\sigma_\tau^2 + \sigma^2} \sim \chi^2_{k-1}.$$

from $\overline{y}_{i.} \sim N(\mu, \sigma_\tau^2 + \sigma^2/n)$, $i = 1, \ldots, k$. Thus, we obtain the F statistic with σ^2 and σ_τ^2:

$$F = \frac{\text{SSTr}/(k-1)}{n\sigma_\tau^2 + \sigma^2} \Big/ \frac{\text{SSE}/(nk-k)}{\sigma^2} = \frac{\text{MSTr}}{\text{MSE}} \frac{\sigma^2}{n\sigma_\tau^2 + \sigma^2} \sim F_{k-1,k(n-1)}.$$

Use the fact to construct a 95% C.I. for ICC as follows:

$$F_{k-1,k(n-1),\alpha/2} \le \frac{\text{MSTr}}{\text{MSE}} \frac{\sigma^2}{n\sigma_\tau^2 + \sigma^2} \le F_{k-1,k(n-1),1-\alpha/2}$$

$$\Rightarrow \quad \frac{1}{n}\left(\frac{\text{MSTr}}{\text{MSE}} \frac{1}{F_{k-1,k(n-1),1-\alpha/2}} - 1\right) \le \frac{\sigma_\tau^2}{\sigma^2} \le \frac{1}{n}\left(\frac{\text{MSTr}}{\text{MSE}} \frac{1}{F_{k-1,k(n-1),\alpha/2}} - 1\right)$$

Let $\ell = \text{LHS}$ and $u = \text{RHS}$,

$$\ell \le \frac{\sigma_\tau^2}{\sigma^2} \le u \;\Rightarrow\; \frac{1}{u} \le \frac{\sigma^2}{\sigma_\tau^2} \le \frac{1}{\ell} \;\Rightarrow\; \frac{1+u}{u} \le \frac{\sigma^2 + \sigma_\tau^2}{\sigma_\tau^2} \le \frac{1+\ell}{\ell} \;\Rightarrow\; \frac{\ell}{1+\ell} \le \text{ICC} \le \frac{u}{1+u}.$$

- Test statistic for a one-way **unbalanced** ANOVA: $y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. Let

$$N = \sum_{i=1}^{k} n_i, \quad \overline{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}, \quad \overline{y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{N}, \quad \overline{\mu} = \frac{\sum_{i=1}^{k} n_i \mu_i}{N},$$

then SSTr can be written in a quadratic form as follows.

$$\text{SSTr} = \sum_{i=1}^{k} n_i \overline{y}_i^2 - N\overline{y}^2 = \sum_{i=1}^{k} \frac{(\sum_j y_{ij})^2}{n_i} - \frac{(\sum_i \sum_j y_{ij})^2}{N} = \mathbf{y}'\mathbf{A}\mathbf{y},$$

where $\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{k1}, \ldots, y_{kn_1})' \in \mathbb{R}^N$

$$\mathbf{A} = \text{diag}\left(\frac{\mathbf{1}_{n_1}\mathbf{1}'_{n_1}}{n_1}, \ldots, \frac{\mathbf{1}_{n_k}\mathbf{1}'_{n_k}}{n_k}\right) - \frac{\mathbf{1}_N\mathbf{1}'_N}{N} \in \mathbb{R}^{N\times N}.$$

Let $\boldsymbol{\mu} = (\mu_1\mathbf{1}'_{n1}, \ldots, \mu_k\mathbf{1}'_{nk})' \in \mathbb{R}^N$, then since $y_{ij} \sim N(\mu_i, \sigma^2) \Rightarrow y_{ij}/\sigma \sim N(\mu_i/\sigma, 1)$,

$$\widetilde{\mathbf{y}} = \frac{\mathbf{y}}{\sigma} \sim N(\widetilde{\boldsymbol{\mu}}, \mathbf{I}_N), \quad \text{where} \quad \widetilde{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}}{\sigma}.$$

Furthermore, $\mathbf{A}$ is clearly symmetric and idempotent because

$$\mathbf{A}^2 = \text{diag}\left(\frac{\mathbf{1}_{n_1}\mathbf{1}'_{n_1}}{n_1}, \ldots, \frac{\mathbf{1}_{n_k}\mathbf{1}'_{n_k}}{n_k}\right) - 2\frac{\mathbf{1}_N\mathbf{1}'_N}{N} + \frac{\mathbf{1}_N\mathbf{1}'_N}{N} = \mathbf{A}.$$

Hence,

$$\frac{\text{SSTr}}{\sigma^2} = \widetilde{\mathbf{y}}'\mathbf{A}\widetilde{\mathbf{y}} \sim \chi^2_r(\delta),$$

where

$$r = \text{tr}(\mathbf{A}) = \text{tr}\left[\text{diag}\left(\frac{\mathbf{1}_{n_1}\mathbf{1}'_{n_1}}{n_1}, \ldots, \frac{\mathbf{1}_{n_k}\mathbf{1}'_{n_k}}{n_k}\right)\right] - \text{tr}\left(\frac{\mathbf{1}_N\mathbf{1}'_N}{N}\right) = k - 1$$

$$\delta = \widetilde{\boldsymbol{\mu}}'\mathbf{A}\widetilde{\boldsymbol{\mu}} = \widetilde{\mathbf{y}}'\mathbf{A}\widetilde{\mathbf{y}}\big|_{\mathbf{y}=\boldsymbol{\mu}} = \frac{\text{SSTr}}{\sigma^2}\bigg|_{\mathbf{y}=\boldsymbol{\mu}} = \frac{\sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2}{\sigma^2}.$$

Next, for the denominator of the test statistic,

$$\frac{\text{SSE}}{\sigma^2} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2}{\sigma^2} \sim \chi^2_{N-k}(0).$$

Hence, the test statistic is given by

$$\frac{\frac{\text{SSTr}}{\sigma^2}/(k-1)}{\frac{\text{SSE}}{\sigma^2}/(N-k)} = \frac{\text{MSTr}}{\text{MSE}} \sim F_{k-1,N-k}\left(\frac{\sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2}{\sigma^2}\right).$$

Note that $\delta = 0$ if $n_i = n$ (balanced model) or under $H_0 : \mu_i = \mu$, $\forall i$ (μ_i is fixed).

- The **power** of the test for $H_0 : \mu_i = \mu$, $\forall i$ vs $H_1 :$ not (fixed effects model).

$$\alpha = P\left(\frac{\text{MSTr}}{\text{MSE}} > F_{k-1,N-k,1-\alpha} \,\middle|\, H_0 : \delta = 0\right), \quad \text{where} \quad \frac{\text{MSTr}}{\text{MSE}} \sim F_{k-1,N-k}(0).$$

where $F_{k-1,N-k,1-\alpha}$ is an upper critical value of a central F distribution with $k-1$ and $N-k$ degrees of freedom at α. Then the power $1-\beta$ is expressed as

$$1 - \beta = P\left(\frac{\text{MSTr}}{\text{MSE}} > F_{k-1,N-k,1-\alpha} \,\middle|\, H_1 : \delta \neq 0\right), \quad \text{where} \quad \frac{\text{MSTr}}{\text{MSE}} \sim F_{k-1,N-k}(\delta).$$

- The **power** of the test for $H_0 : \sigma_\tau^2 = 0$ vs $H_1 : \sigma_\tau^2 = \sigma^2$ (balanced mixed effects model). We have

$$\frac{\text{SSTr}}{n\sigma_\tau^2 + \sigma^2} \sim \chi^2_{k-1}, \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2_{(n-1)k} \quad \Rightarrow \quad F = \frac{\text{MSTr}}{\text{MSE}} \frac{\sigma^2}{n\sigma_\tau^2 + \sigma^2} \sim F_{k-1,(n-1)k}.$$

Let $V = \sigma_\tau^2/\sigma^2$ (called the variance ratio). We can rewrite this as

$$\frac{\text{MSTr}}{\text{MSE}} \sim (1 + nV)\, F_{k-1,(n-1)k},$$

where $V = \sigma_\tau^2/\sigma^2$ is called the variance ratio. Under $H_0 : \sigma_\tau^2 = 0$, or equivalently $V = 0$, we have

$$\frac{\text{MSTr}}{\text{MSE}} \sim F_{k-1,(n-1)k} \quad \Rightarrow \quad \alpha = P\left(\frac{\text{MSTr}}{\text{MSE}} > F_{k-1,(n-1)k,1-\alpha} \;\middle|\; H_0\right),$$

On the other hand, under $H_a : \sigma_\tau^2 = \sigma^2$, or equivalently $V = 1$,

$$\frac{\text{MSTr}}{\text{MSE}} \sim \frac{1}{n+1} F_{k-1,(n-1)k} \quad \Rightarrow \quad \frac{1}{n+1}\frac{\text{MSTr}}{\text{MSE}} \sim F_{k-1,(n-1)k}.$$

It follows that the power is

$$\begin{aligned} 1 - \beta &= P\left(\frac{\text{MSTr}}{\text{MSE}} > F_{k-1,(n-1)k,1-\alpha} \;\middle|\; H_1\right) \\ &= P\left(\frac{1}{n+1}\frac{\text{MSTr}}{\text{MSE}} > \frac{1}{n+1} F_{k-1,(n-1)k,1-\alpha} \;\middle|\; H_1\right). \end{aligned}$$

- **Studentized range interval**. Studentized range distribution is defined below. Let $Z_1, \ldots, Z_k$ be iid $Z(0,1)$ and let $U \sim \chi^2_m(0)$ and suppose $U \perp\!\!\!\perp Z_i$ $(i = 1, \ldots, k)$, then

$$\max_{1 \le i \ne j \le k} \frac{|Z_i - Z_k|}{\sqrt{U/m}} \quad \sim \quad q_{k,m}.$$

Consider one-way balanced ANOVA. Since $\overline{y}_i \sim N(\alpha_i, \sigma^2/n)$,

$$\max_{1 \le i \ne j \le k} \frac{\sqrt{n}|\overline{y}_i - \overline{y}_j - (\alpha_i - \alpha_j)|/\sigma}{S/\sigma} = \max_{1 \le i \ne j \le k} \frac{\sqrt{n}|\overline{y}_i - \overline{y}_j - (\alpha_i - \alpha_j)|}{S} \sim q_{k,k(n-1)}.$$

Thus, the test statistic for testing $H : \alpha_1 = \cdots = \alpha_k$ is

$$M = \max_{1 \le i \ne j \le k} \frac{\sqrt{n}|\overline{y}_i - \overline{y}_j|}{S} \sim q_{k,k(n-1)}.$$

For $100(1-\alpha)\%$ confidence interval for $\alpha_i - \alpha_j$,

$$\begin{aligned} 1 - \alpha &= Pr\left(\max_{1 \le i \ne j \le k} \frac{\sqrt{n}|\overline{y}_i - \overline{y}_j - (\alpha_i - \alpha_j)|}{S} \le q^\alpha_{k,k(n-1)}\right) \\ &= Pr\left(|\overline{y}_i - \overline{y}_j - (\alpha_i - \alpha_j)| \le q^\alpha_{k,k(n-1)} \frac{S}{\sqrt{n}},\ 1 \le i \ne j \le k\right) \\ &= Pr\left(\overline{y}_i - \overline{y}_j - q^\alpha_{k,k(n-1)} \frac{S}{\sqrt{n}} \le \alpha_i - \alpha_j \le \overline{y}_i - \overline{y}_j + q^\alpha_{k,k(n-1)} \frac{S}{\sqrt{n}},\ 1 \le i \ne j \le k\right) \end{aligned}$$

leads to

$$\alpha_i - \alpha_j \in \overline{y}_i - \overline{y}_j \pm q^\alpha_{k,k(n-1)} \frac{S}{\sqrt{n}}, \quad 1 \le i \ne j \le k.$$

- Lemma. Let $a_1, \dots a_k$ be numbers. Then

$$|a_i - a_j| \le b, \ \forall i, j \ \Leftrightarrow \ \left|\sum_{i=1}^{k} c_i a_i\right| \le \frac{b \sum_i |c_i|}{2} \quad \text{s.t.} \ \sum_i c_i = 0.$$

Proof: $\Rightarrow$ is hard. $\Leftarrow$ is easy, just take $c_i = 1, c_j = -1$,

- Using this lemma, we obtain the general CI for $\sum_{i=1}^{k} c_i \alpha_i$:

$$\begin{aligned} 1 - \alpha &= Pr\left(\max_{1 \le i \ne j \le k} \frac{\sqrt{n}|\overline{y}_i - \overline{y}_j - (\alpha_i - \alpha_j)|}{S} \le q^{\alpha}_{k,k(n-1)}\right) \\ &= Pr\left(|(\overline{y}_i - \alpha_i) - (\overline{y}_j - \alpha_j)| \le q^{\alpha}_{k,k(n-1)} \frac{S}{\sqrt{n}}, \quad \forall i, j\right) \\ &= Pr\left(\left|\sum_{i=1}^{k} (\overline{y}_i - \alpha_i)\right| \le \frac{\sum_i |c_i|}{2} q^{\alpha}_{k,k(n-1)} \frac{S}{\sqrt{n}}, \quad \text{s.t.} \ \sum_i c_i = 0\right) \end{aligned}$$

so that

$$\sum_{i=1}^{k} c_i \alpha_i \in \sum_{i=1}^{k} c_i \overline{y}_i \pm q^{\alpha}_{k,k(n-1)} \frac{S}{\sqrt{n}} \frac{\sum_i |c_i|}{2}, \quad \text{s.t.} \ \sum_{i=1}^{k} c_i = 0.$$

- (HW4): We consider a four-way mixed ANOVA:

$$\begin{aligned} y_{ijklm} = \mu &+ \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} \\ &+ (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ijl} + (\alpha\gamma\delta)_{ikl} + (\beta\gamma\delta)_{jkl} + (\alpha\beta\gamma\delta)_{ijkl} + \epsilon_{(ijkl)m}, \end{aligned}$$

for $i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, c, \quad l = 1, \dots, d, \quad m = 1, \dots, n$. Further, let

$$a' = a - 1, \quad b' = b - 1, \quad c' = c - 1, \quad d' = d - 1, \quad n' = n - 1$$

The important point is that fixed effects appear in its own term only, while random effects are present in all terms with their factors.

12 Factors A, B, C are fixed and Factor D is random.

Model Term	Factor	Expected Mean Squares
α_i	A, main fixed effect	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + cn\sigma^2_{\alpha\beta\delta} + bn\sigma^2_{\alpha\gamma\delta} + bcn\sigma^2_{\alpha\delta} + \frac{bcdn\sum_i \alpha_i^2}{a'}$
β_j	B, main fixed effect	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + cn\sigma^2_{\alpha\beta\delta} + an\sigma^2_{\beta\gamma\delta} + acn\sigma^2_{\beta\delta} + \frac{acdn\sum_j \beta_j^2}{b'}$
γ_k	C, main fixed effect	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + bn\sigma^2_{\alpha\gamma\delta} + an\sigma^2_{\beta\gamma\delta} + abn\sigma^2_{\gamma\delta} + \frac{abdn\sum_k \gamma_k^2}{c'}$
δ_l	D, main random effect	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + cn\sigma^2_{\alpha\beta\delta} + bn\sigma^2_{\alpha\gamma\delta} + an\sigma^2_{\beta\gamma\delta} + bcn\sigma^2_{\alpha\delta} + acn\sigma^2_{\beta\delta} + abn\sigma^2_{\gamma\delta} + abcn\sigma^2_{\delta}$
$(\alpha\beta)_{ij}$	AB, 2-factor fixed interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + cn\sigma^2_{\alpha\beta\delta} + \frac{cdn\sum_{i,j}(\alpha\beta)_{ij}^2}{a'b'}$
$(\alpha\gamma)_{ik}$	AC, 2-factor fixed interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + bn\sigma^2_{\alpha\gamma\delta} + \frac{bdn\sum_{i,k}(\alpha\gamma)_{ik}^2}{a'c'}$
$(\alpha\delta)_{il}$	AD, 2-factor random interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + cn\sigma^2_{\alpha\beta\delta} + bn\sigma^2_{\alpha\gamma\delta} + bcn\sigma^2_{\alpha\delta}$
$(\beta\gamma)_{jk}$	BC, 2-factor fixed interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + an\sigma^2_{\beta\gamma\delta} + \frac{adn\sum_{j,k}(\beta\gamma)_{jk}^2}{b'c'}$
$(\beta\delta)_{jl}$	BD, 2-factor random interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + cn\sigma^2_{\alpha\beta\delta} + an\sigma^2_{\beta\gamma\delta} + acn\sigma^2_{\beta\delta}$
$(\gamma\delta)_{kl}$	CD, 2-factor random interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + bn\sigma^2_{\alpha\gamma\delta} + an\sigma^2_{\beta\gamma\delta} + abn\sigma^2_{\gamma\delta}$
$(\alpha\beta\gamma)_{ijk}$	ABC, 3-factor fixed interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + \frac{dn\sum_{i,j,k}(\alpha\beta\gamma)_{ijk}^2}{a'b'c'}$
$(\alpha\beta\delta)_{ijl}$	ABD, 3-factor random interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + cn\sigma^2_{\alpha\beta\delta}$
$(\alpha\gamma\delta)_{ikl}$	ACD, 3-factor random interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + bn\sigma^2_{\alpha\gamma\delta}$
$(\beta\gamma\delta)_{jkl}$	BCD, 3-factor random interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta} + an\sigma^2_{\beta\gamma\delta}$
$(\alpha\beta\gamma\delta)_{ijkl}$	ABCD, 4-factor random interaction	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma\delta}$
$\epsilon_{(ijkl)m}$	Error	σ^2

H_0	Test Statistic	Exact/Approx.	Distribution
$\alpha_i = 0, \forall i$	MSA / MSAD	Exact	$F_{a',\ a'd'}$
$\beta_j = 0, \forall j$	MSB / MSBD	Exact	$F_{b',\ b'd'}$
$\gamma_k = 0, \forall k$	MSC / MSCD	Exact	$F_{c',\ c'd'}$
$\sigma^2_\delta = 0$	$\frac{\text{MSD}}{\text{MSAD + MSBD + MSCD - MSABD - MSACD - MSBCD + MSABCD}}$ or $\frac{\text{MSD + MSABD + MSACD + MSBCD}}{\text{MSAD + MSBD + MSCD + MSABCD}}$	Approx.	$F_{*,*}$
$(\alpha\beta)_{ij} = 0, \forall i, j$	MSAB / MSABD	Exact	$F_{a'b',\ a'b'd'}$
$(\alpha\gamma)_{ik} = 0, \forall i, k$	MSAC / MSACD	Exact	$F_{a'c',\ a'c'd'}$
$\sigma^2_{\alpha\delta} = 0$	$\frac{\text{MSAD}}{\text{MSABD + MSACD - MSABCD}}$ or $\frac{\text{MSAD + MSABCD}}{\text{MSABD + MSACD}}$	Approx.	$F_{*,*}$
$(\beta\gamma)_{jk} = 0, \forall j, k$	MSBC / MSBCD	Exact	$F_{b'c',\ b'c'd'}$
$\sigma^2_{\beta\delta} = 0$	$\frac{\text{MSBD}}{\text{MSABD + MSBCD - MSABCD}}$ or $\frac{\text{MSBD + MSABCD}}{\text{MSABD + MSBCD}}$	Approx.	$F_{*,*}$
$\sigma^2_{\gamma\delta} = 0$	$\frac{\text{MSCD}}{\text{MSACD + MSBCD - MSABCD}}$ or $\frac{\text{MSCD + MSABCD}}{\text{MSACD + MSBCD}}$	Approx.	$F_{*,*}$
$(\alpha\beta\gamma)_{ijk} = 0, \forall i, j, k$	MSABC / MSABCD	Exact	$F_{a'b'c',\ a'b'c'd'}$
$\sigma^2_{\alpha\beta\delta} = 0$	MSABD / MSABCD	Exact	$F_{a'b'd',\ a'b'c'd'}$
$\sigma^2_{\alpha\gamma\delta} = 0$	MSACD / MSABCD	Exact	$F_{a'c'd',\ a'b'c'd'}$
$\sigma^2_{\beta\gamma\delta} = 0$	MSBCD / MSABCD	Exact	$F_{b'c'd',\ a'b'c'd'}$
$\sigma^2_{\alpha\beta\gamma\delta} = 0$	MSABCD / MSE	Exact	$F_{a'b'c'd',\ n'abcd}$

13 Cochran's Theorem

- In more than 2-way ANOVA, some F-statistics have no exact F distribution, as both denominator and numerator have a linear combination of independent χ^2 distribution. Then Cochran's theorem can be applied to obtain their asymptotic distributions.
- Suppose $X_i \sim \chi^2_{r_i}$ and $X_i \perp X_j$, $1 \le i \ne j \le k$. Let $X = \sum_{i=1}^k a_i X_i$, $a_i \ge 0$ and define $X \sim c\chi^2_r(0)$.
- $E(X_i) = r_i$ and $E(X) = cr \Rightarrow \sum_{i=1}^k a_i r_i = cr$.
- $\text{var}(X_i) = 2r_i$ and $\text{Var}(X) = c^2(2r) \Rightarrow \sum_{i=1}^k 2a_i^2 r_i = 2c^2 r$. Hence,

$$\sum_{i=1}^k a_i^2 r_i = c^2 r = c(cr) = c\sum_{i=1}^k a_i r_i \quad \Rightarrow \quad c = \frac{\sum_{i=1}^k a_i^2 r_i}{\sum_{i=1}^k a_i r_i}, \quad r = \frac{\sum_{i=1}^k a_i r_i}{c} = \frac{\left(\sum_{i=1}^k a_i r_i\right)^2}{\sum_{i=1}^k a_i^2 r_i}.$$

- Apply this theorem to the ANOVA setting:

$$\frac{\text{SSX}_i}{\sigma_i^2} = \frac{r_i \text{MSX}_i}{\sigma_i^2} = X_i \ \sim \ \chi^2_{r_i}.$$

We are interested in $X = \sum_{i=1}^k \text{MSX}_i$.

$$X = \sum_{i=1}^k \frac{\sigma_i^2}{r_i}\frac{r_i \text{MSX}_i}{\sigma_i^2} = \sum_{i=1}^k a_i X_i,$$

so that

$$c = \frac{\sum_{i=1}^k a_i^2 r_i}{\sum_{i=1}^k a_i r_i} = \frac{\sum_{i=1}^k \sigma_i^4/r_i}{\sum_{i=1}^k \sigma^2} \quad \Rightarrow \quad \widehat{c} = \frac{\sum_{i=1}^k \text{MSX}_i^2/r_i}{\sum_{i=1}^k \text{MSX}_i},$$
$$r = \frac{\left(\sum_{i=1}^k a_i r_i\right)^2}{\sum_{i=1}^k a_i^2 r_i} = \frac{\left(\sum_{i=1}^k \sigma^2\right)^2}{\sum_{i=1}^k \sigma_i^4/r_i} \quad \Rightarrow \quad \widehat{r} = \frac{\left(\sum_{i=1}^k \text{MSX}_i\right)^2}{\sum_{i=1}^k \text{MSX}_i^2/r_i}.$$

14 Bias

- Here, suppose $\epsilon \sim N_n(0, \sigma^2 I_n)$.
- **Underfitting**. Consider

$$\begin{aligned} \text{True model } (G): &\quad E(Y) = X\beta + Z\gamma = W\delta, \\ \text{Working model } (R): &\quad E(Y) = X\beta. \end{aligned}$$

Then the least square estimate is $\widehat{\beta} = (X'X)^{-1}X'Y$ based on the postulated model.

(1) $\widehat{\beta}$ is biased. $E(\widehat{\beta}) = (X'X)^{-1}X'E(Y) = \beta + (X'X)^{-1}X'Z\gamma \ne \beta$ unless $X'Z = O$ (i.e., the columns of Z are orthogonal to the columns of X).

(2) $\text{Var}(\widehat{\beta})$ is correct (same as the postulated model). $\text{Var}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$, as Z is also fixed.

(3) S^2 is an overestimate of σ^2 (biased).

$$\mathbb{E}[(n-p)S^2] = \mathbb{E}[Y'(I_n - P_X)Y] = \sigma^2(n-p) + E(Y')Q_X E(Y) = \sigma^2(n-p) + \gamma' Z' Q_X Z\gamma$$
$$\Rightarrow \quad E[S^2] = \sigma^2 + \frac{\gamma' Z' Q_X Z\gamma}{n-p} > \sigma^2 \quad \because Q_X \succeq O, \ \gamma \ne 0, \ Z\gamma \notin C(X)$$

noting that if $Z\gamma \in C(X)$, then we write $Z\gamma = Xa$ and hence $Q_X Z\gamma = Q_X Xa = 0$.

(4) Fitted model (not prediction) and the *observed* residual vector are biased.

$$E(\widehat{Y}) = P_X E(Y) = P(X\beta + Z\gamma) = X\beta + PZ\gamma \neq X\beta + Z\gamma \quad \because Z\gamma \notin C(X)$$
$$E(e) = (I - P_X)E(Y) = (I - P_X)Z\gamma \neq 0 \quad \because Z\gamma \notin C(X).$$

(5) $\mathrm{Var}(e)$ is unchanged. $\mathrm{Var}(e) = \sigma^2(I - P_X)$, which is the same if the postulated model is used.

(6) The prediction mean is biased. Suppose that we wish to predict Y at $w_0 = (x_0', z_0')'$, where only x_0 is observed.

$$E(\widehat{Y}_{0R}) = x_0' E(\widehat{\beta}) = x_0'\beta + x_0'(X'X)^{-1}X'Z\gamma,$$
$$E(\widehat{Y}_{0G}) = w_0' E(\widehat{\delta}) = x_0'\beta + z_0'\gamma.$$

(7) The prediction variance is smaller than (or equal to) the true prediction variance.

$$\begin{aligned}
\mathrm{var}(\widehat{Y}_{0G}) &= \sigma^2 w_0' \,\mathrm{Var}(\widehat{\delta}) w_0 \\
&= \sigma^2 (x_0', z_0') \begin{pmatrix} (X'X)^{-1} + LML' & -LM \\ -ML' & M \end{pmatrix} \begin{pmatrix} x_0 \\ z_0 \end{pmatrix} \\
&= \sigma^2 [x_0'(X'X)^{-1}x_0 + (L'x_0 - z_0)'M(L'x_0 - z_0)] \\
&\geq \mathrm{var}(\widehat{Y}_{0R}) \quad \because M \succ O
\end{aligned}$$

with the equality iff $z_0 = L'x_0$, which could happen.

- **Overfitting**. Consider

$$\begin{aligned}
\text{True model } (R): &\quad E(Y) = X\beta, \\
\text{Working model } (G): &\quad E(Y) = X\beta + Z\gamma = W\delta,
\end{aligned}$$

Then the LS estimate is

$$\widehat{\delta} = \begin{pmatrix} \widehat{\beta}_G \\ \widehat{\gamma}_G \end{pmatrix} = \begin{pmatrix} (X'X)^{-1}X(y - Z\widehat{\gamma}_G) \\ (Z'Q_X Z)^{-1}Z'Q_X Y \end{pmatrix}$$

based on the postulated model.

(1) $\widehat{\beta}_G$ is unbiased. Since $E(\widehat{\gamma}_G) = (Z'Q_X Z)^{-1}Z'Q_X E(Y) = (Z'Q_X Z)^{-1}Z'Q_X X\beta = 0$,

$$E(\widehat{\beta}_G) = (X'X)^{-1}X'[E(Y) - 0] = (X'X)^{-1}X'X\beta = \beta.$$

(2) $\mathrm{Var}(\widehat{\beta}_G)$ is inflated.

$$\mathrm{Var}(\widehat{\beta}_G) = \sigma^2[(X'X)^{-1} + LML'] \succeq \mathrm{Var}(\widehat{\beta}) \quad \because M \succ O,$$

so that $\mathrm{var}(\widehat{\beta}_{Gi}) = \sigma^2(X'X)^{-1}_{ii} + \sigma^2(LML')_{ii} > \mathrm{var}(\widehat{\beta}_i)$.

(3) S^2 is an unbiased estimate of σ^2.

$$\mathbb{E}[(n-p)S^2] = \mathbb{E}[Y'(I_n - P_X)Y] = \sigma^2(n-p) + \beta'X'Q_W X\beta = \sigma^2(n-p) \;\Rightarrow\; E[S^2] = \sigma^2$$

since $Q_W W = (I - P_W)(X \mid Z) = O \Rightarrow Q_W X = O$.

(4) Fitted model (not prediction) and the *observed* residual vector are unbiased.

$$E(\widehat{Y}) = E(X\widehat{\beta}_G + Z\widehat{\gamma}_G) = XE(\widehat{\beta}_G) = X\beta,$$
$$E(e) = E[Y - (X\widehat{\beta}_G + Z\widehat{\gamma}_G)] = X\beta - X\beta = 0.$$

(5) $\mathrm{Var}(e)$ is attenuated. Compare P_W and P_X.

$$\begin{aligned} P_W = W(W'W)^{-1}W' &= (X \mid Z)\begin{pmatrix} (X'X)^{-1} + LML' & -LM \\ -ML' & M \end{pmatrix}\begin{pmatrix} X' \\ Z' \end{pmatrix} \\ &= P_X + (XL - Z)M(XL - Z)' \succeq P_{X_1} \end{aligned}$$

so that apparent $\mathrm{Var}(e) = \sigma^2(I_n - P_W) \preceq \sigma^2(I_n - P_X) = \text{true } \mathrm{Var}(e)$, which means overfitting.

(6) The prediction mean is unbiased. Suppose that we predict Y at $w_0 = (x_0', z_0')'$, which is observed.

$$E(\widehat{Y}_{0G}) = E(w_0'\widehat{\delta}) = x_0'E(\widehat{\beta}_G) + z_0'E(\widehat{\gamma}_G) = x_0'\beta = E(\widehat{Y}_{0R}).$$

(7) The prediction variance is larger than (or equal to) the true prediction variance.

$$\begin{aligned} \mathrm{var}(\widehat{Y}_{0G}) &= \sigma^2 w_0' \,\mathrm{Var}(\widehat{\delta}) w_0 \\ &= \sigma^2(x_0', z_0')\begin{pmatrix} (X'X)^{-1} + LML' & -LM \\ -ML' & M \end{pmatrix}\begin{pmatrix} x_0 \\ z_0 \end{pmatrix} \\ &= \sigma^2[x_0'(X'X)^{-1}x_0 + (L'x_0 - z_0)'M(L'x_0 - z_0)] \\ &\geq \mathrm{var}(\widehat{Y}_{0R}) \quad \because M \succ O \end{aligned}$$

with the equality iff $z_0 = L'x_0$, which could happen.

- In summary,
  - Underfitting gives a biased estimate of β and σ^2, but correct variance of estimate of β, while overfitting gives an unbiased estimate of β and σ^2, but inflated variance of estimate of β.
  - Underfitting leads to a biased residual vector but its variance is unchanged, while overfitting gives an unbiased residual vector, but its variance is *attenuated* (i.e., overfitting).
  - Underfitting introduces bias but reduces the variance of $\widehat{Y}_0$, while overfitting provides unbiasedness but increases the variance of $\widehat{Y}_0$.

	$\widehat{\beta}$ and S^2	$\mathrm{Var}(\widehat{\beta})$	e	$\mathrm{Var}(e)$	$\widehat{Y}_0$	$\mathrm{Var}(\widehat{Y}_0)$ vs truth
Underfitting	Biased	Correct	Biased	Correct	Biased	Smaller (better)
Overfitting	Unbiased	Inflated	Unbiased	Attenuated	Unbiased	Larger (worse)

15 Residuals and Hat matrix diagonals

- We begin with the usual model $Y = X\beta + \epsilon$, where $X \in \mathbb{R}^{n\times p}$ of rank p and $\mathrm{Var}(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I_n$.
- The major tools for diagnosing model faults are the residuals. Here the projection matrix is often denoted by H rather than P.
- Let $H = (h_{ij})$ and $h_{ii} = h_i$. Important properties are:
  - $1/n \leq h_i \leq 1$. First, show $h_i \geq 1/n$. WLOG, consider $Y = X\beta + \epsilon$, where $X = (1\ X_c)$, i.e., with intercept and after centering. Then

$$H = X(X'X)^{-1}X' = \frac{11'}{n} + X_c C^{-1} X_c' \quad \Rightarrow \quad h_i = c_i' H c_i = \frac{1}{n} + (x_i - \overline{x})'C^{-1}(x_i - \overline{x}) \geq \frac{1}{n},$$

  which implies that h_i is a measure of how *outlying* the ith data point is, at least as far as the predictors are concerned. Also, $H = H^2$ gives $h_i = h_i^2 + \sum_{k\neq 1} h_{ki}^2 \geq h_i^2 \Rightarrow h_i(1-h_i) \geq 0 \Rightarrow h_i \leq 1$.
  - $-0.5 \leq h_{ij} \leq 0.5$ $(i \neq j)$. From $h_i = h_i^2 + \sum_{k\neq 1} h_{ki}^2$,

$$\sum_{k\neq 1} h_{ki}^2 = h_i(1 - h_i) \leq \frac{1}{4} \quad \Rightarrow \quad h_{ki}^2 \leq 0.25.$$

– The average hat matrix diagonal is

$$\overline{h} = \frac{\sum_{i=1}^{n} h_i}{n} = \frac{\text{tr}(H)}{n} = \frac{p}{n}$$

- The residual is given by $e = Y - X\widehat{\beta} = (I-H)Y = (I-H)\epsilon$, so that $E(e) = 0$ and $\text{Var}(e) = \sigma^2(I_n - H)$, so that $\text{Var}(e_i) = \sigma^2(1-h_i)$.
- The fitted values is $\widehat{Y} = X\widehat{\beta}$ so that $E(\widehat{Y}) = X\beta$ and $\text{Var}(\widehat{Y}) = \sigma^2 X(X'X)^{-1}X' = \sigma^2 H \preceq \text{Var}(\epsilon)$.
- Moreover, $\text{Cov}(e, \widehat{Y}) = \text{Cov}[(I-H)Y, HY] = 0$ as $H^2 = H$.
- Let $H = (h_{ij})$ and $h_{ii} = h_i$, then we have $\text{Var}(e_i) = \sigma^2(1-h_i)$. That is, the model is correct, the variances of the residuals depend on the diagonal of H. Hence, the residuals are sometimes scaled.
  - Internally studentized residual:

$$r_i = \frac{e_i}{S\sqrt{1-h_i}} \quad \Rightarrow \quad r_i^2 = \frac{e_i^2}{S^2(1-h_i)}, \quad \sim \quad (n-p) \times \text{Beta}\left(\frac{1}{2}, \frac{n-p-1}{2}\right),$$

    where $S^2 = Y'(I-H)Y/(n-p) = e'e/(n-p)$ is the usual estimate of σ^2.
  - Externally studentized residual:

$$t_i = \frac{e_i}{S_{(i)}\sqrt{1-h_i}} \sim t_{n-p-1}, \quad \Rightarrow \quad t_i^2 = \frac{e_i^2}{S_{(i)}^2(1-h_i)} \sim F_{1,n-p-1},$$

    where $S_{(i)}$ is calculated from the $n-1$ data points that remain after deleting the ith case.
- Show $B = r_i^2/(n-p) \sim \text{Beta}(1/2, (n-p-1)/2)$.

  Proof: Since $e_i = c_i'(I-H)Y = c_i'(I-H)\epsilon$,

$$B = \frac{\epsilon'(I-H)c_ic_i'(I-H)\epsilon}{(n-p)S^2(1-h_i)} = \frac{\epsilon'Q\epsilon}{\epsilon'(I-H)\epsilon} = \frac{\epsilon'Q\epsilon}{\epsilon'(I-H-Q)\epsilon + \epsilon'Q\epsilon},$$

  where $\epsilon \sim N_n(0, \sigma^2 I_n)$ and $Q = (I-H)c_ic_i'(I-H)/(1-h_i)$ is a projection matrix as $c_i'(I-H)c_i = 1-h_i$. Note that $Q(I-H) = Q \neq O$ since $QH = O$ and hence the numerator and the denominator are *not* independent. However, $I-H-Q$ is another projection matrix and $Q(I-H-Q) = O$, meaning that $\epsilon'Q\epsilon \perp\!\!\!\perp \epsilon'(I-H-Q)\epsilon$ by Craig's theorem.

  Finally, since $\text{rank}(Q) = \text{tr}(Q) = 1$,

$$B = \frac{\chi_1^2}{\chi_{n-p-1}^2 + \chi_1^2} = \frac{\chi_1^2/\chi_{n-p-1}^2}{1 + \chi_1^2/\chi_{n-p-1}^2} = \frac{F}{n-p-1+F} \sim \text{Beta}\left(\frac{1}{2}, \frac{n-p-1}{2}\right).$$

  from the following important relationships:

$$F \sim F_{\alpha,\beta} \quad \Rightarrow \quad B = \frac{\alpha F}{\beta + \alpha F} = \frac{F}{\beta/\alpha + F} \sim \text{Beta}\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

$$B \sim \text{Beta}\left(\frac{\alpha}{2}, \frac{\beta}{2}\right) \quad \Rightarrow \quad F = \frac{\beta B}{\alpha(1-B)} = \frac{(\beta/\alpha)B}{1-B} \sim F_{\alpha,\beta}.$$

- Let $\widehat{\beta}_{(i)}$ denote the LSE of β without the ith case. Then

$$\widehat{\beta} - \widehat{\beta}_{(i)} = \frac{(X'X)^{-1}x_ie_i}{1-h_i}.$$

Proof: Since $X'_{(i)}X_{(i)} = X'X - x_i x_i'$ (outer product),

$$(X'_{(i)}X_{(i)})^{-1} = (X'X - x_i x_i')^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i}.$$

Further, since $X'_{(i)}Y_{(i)} = X'Y - x_i Y_i$,

$$(X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)} = \left[(X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_i}\right](X'Y - x_i Y_i)$$

$$\Rightarrow \quad \widehat{\beta}_{(i)} = \widehat{\beta} - \frac{(X'X)x_i(Y_i - X_i'\widehat{\beta})}{1-h_i}.$$

We see that the change is proportional to the residual but greatly inflated if h_i is close to 1.

- Using the above, we have a relationship between two estimates of σ^2.

$$(n-p-1)S^2_{(i)} = \sum_{k\neq i}[Y_k - x_k'\widehat{\beta}_{(i)}]^2 = \sum_{k\neq i}\left[e_k + \frac{h_{ki}e_i}{1-h_i}\right]^2 = \sum_{k=1}^{n}\left[e_k + \frac{h_{ki}e_i}{1-h_i}\right]^2 - \left(e_i + \frac{h_i e_i}{1-h_i}\right)^2.$$

Since $He = H(I-H)Y = 0$ and $\sum_k h_{ki}^2 = h_i$,

$$(n-p-1)S^2_{(i)} = (n-p)S^2 + \frac{h_i e_i^2}{(1-h_i)^2} - \frac{e_i^2}{(1-h_i)^2}$$

$$= (n-p)S^2 - \frac{e_i^2}{1-h_i},$$

i.e., $\text{SSE}_{(i)} = \text{SSE} - e_i^2/(1-h_i)$.

- Hence, from $e_i^2/(1-h_i) = r_i^2 S^2$, we have

$$t_i^2 = \frac{e_i^2}{S^2_{(i)}(1-h_i)} = \frac{r_i^2(n-p-1)}{n-p-r_i^2} = \frac{B}{1-B}(n-p-1) \;\sim\; F_{1,n-p-1}.$$

- By the way, the i-th leverage value without i-th case is given by

$$h_i(i) = x_i(X'_{(i)}X_{(i)})^{-1}x_i = h_i + \frac{h_i^2}{1-h_i} = \frac{h_i}{1-h_i}.$$

- HW3 Question 3. Let c_i be the i-th coordinate of the standard basis and you want to test whether the ith case is problematic in the model $y = X\beta + c_i\delta + \epsilon$, where $\epsilon \sim N_n(0, \sigma^2 I_n)$. Show that the test statistic for testing $H_0 : \delta = 0$ vs $H_A : \delta \neq 0$ is t_i^2.

Solution: Suppose $X \in \mathbb{R}^{n\times p}$ with the rank of p. Let SSE_{H_0} and SSE_{H_a} be SSE under H_0 and H_a, respectively. Then $\text{SSE}_{H_0} = (n-p)s^2$. where s^2 is the estimated error variance under $H_0 : E(y) = X\beta$. We can estimate δ and β_{H_a} by considering this model as a G model

$$\widehat{\delta} = (c_i'Q_X c_i)^{-1}c_i'Q_X y = [c_i'(I-H)c_i)^{-1}]c_i'e = \frac{e_i}{1-h_i},$$

$$\widehat{\beta}_{H_a} = \widehat{\beta} - (X'X)^{-1}X'c_i\widehat{\delta},$$

where $\widehat{\beta}$ is the LS estimate under H_0. Hence,

$$\text{SSE}_{H_0} - \text{SSE}_{H_a} = \|(X\widehat{\beta}_{H_a} + c_i\widehat{\delta}) - X\widehat{\beta}\|^2 = \widehat{\delta}^2\|(I-H)c_i\|^2 = \frac{e_i^2}{1-h_i},$$

$$\Rightarrow\; \text{SSE}_{H_a} = \text{SSE}_{H_0} - \frac{e_i^2}{1-h_i} = (n-p-1)s^2_{(i)},$$

so that the test statistic can be written as

$$F = \frac{(\text{SSE}_{H_0} - \text{SSE}_{H_a})/1}{\text{SSE}_{H_a}/[n-(p+1)]} = \frac{e_i^2}{s^2_{(i)}(1-h_i)} = t_i^2.$$

16 Type of outliers

- There are two kinds of outliers in regression data. First, there may be a big difference between $\mathbf{x}_i$ and the center of the x-data, which is called a high-leverage point. Second, there may be a large difference between Y_i and the mean $x_i'\widehat{\beta}$, we called such a point an outlier.
- Outliers that are not high-leverage points do not have a strong influence on $\widehat{Y}$ unless they are very large. The situation is very different when outliers are also high-leverage points.
- Since $\widehat{Y}_i = x_i'\widehat{\beta} + e_i$, it is reasonable to treat e_i as estimate of the error ϵ_i and examine the residuals for extremes. We can use the raw residuals e_i, r_i, or t_i, which are adjusted to be identically distributed. Since $t_i \sim t_{n-p-1}$ in the absence of outliers, a reasonable distribution of "large" is $|t_i| > 2$.
- The diagnostic approach described above works well if the data point in question does not have high leverage. If it does, we cannot expect the corresponding residual to reveal the presence of an outlier.
- Suppose that the ith response is recorded as $Y_i - \Delta_i$ rather than Y_i, so that $Y_i = x_i'\beta + \Delta_i + \epsilon_i$, which we can interpret as the error being changed by an amount Δ_i. Let Δ be the vector with ith element Δ_i and the rest zero. Then

$$E(e) = (I - H)E(Y) = (I - H)\Delta \quad \Rightarrow \quad E(e_i) = (1 - h_i)\Delta_i,$$

  which means that if the x_i is close to $\overline{x}$ so that h_i is small, we can expect the residual to reflect the outlier quite well. Since $\overline{h} = p/n$, a reasonable definition of a high-leverage point is $h_i > 2p/n$.

17 Leave-One-Out Case Diagnostics

- Case diagnostics focus on identifying outliers, which are not always bad.
- Change in fitted values:

$$\text{DFFITS}_i = x_i'\widehat{\beta} - x_i'\widehat{\beta}_{(i)} = \frac{x_i(X'X)^{-1}x_i e_i}{1 - h_i} = \frac{h_i e_i}{1 - h_i},$$
$$\text{DFFITSS}_i = \frac{x_i'\widehat{\beta} - x_i'\widehat{\beta}_{(i)}}{S_{(i)}h_i^{1/2}} = \frac{e_i h_i^{1/2}}{S_{(i)}(1 - h_i)} = t_i\left(\frac{h_i}{1 - h_i}\right)^{1/2}.$$

- Covariance ratio:

$$\text{COVRATIO} = \frac{|S_{(i)}^2(X_{(i)}'X_{(i)})^{-1}|}{|S^2(X'X)^{-1}|} = \left(\frac{S_{(i)}^2}{S^2}\right)^p \frac{|X'X|}{|X_{(i)}'X_{(i)}|} = \left(\frac{n - p - 1}{n - p} + \frac{t_i^2}{n - p}\right)^{-p} \frac{1}{1 - h_i}$$

  since $|X_{(i)}'X_{(i)}| = |X'X - x_i x_i'| = |X'X||I_p - (X'X)^{-1}x_i x_i'| = |X'X|(1 - x_i'(X'X)^{-1}x_i)$ and

$$\frac{(n - p)S^2}{(n - p)S_{(i)}^2} = \frac{(n - p - 1)S_{(i)}^2 + \frac{e_i^2}{1 - h_i}}{(n - p)S_{(i)}^2} = \frac{n - p - 1}{n - p} + \frac{t_i^2}{n - p}.$$

- Cook's Distance: Analogy to F-test statistics,

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})'X'X(\widehat{\beta}_{(i)} - \widehat{\beta})}{pS^2} = \frac{e_i^2 h_i}{pS^2(1 - h_i)^2} = r_i^2\frac{h_i}{p(1 - h_i)}.$$

  Thus, a point will have a large Cook's D if it has a large r_i or is a high-leverage point (close to 1).
- Volume of confidence ellipsoid

$$\frac{(\widehat{\beta}_{(i)} - \widehat{\beta})'X_{(i)}'X_{(i)}(\widehat{\beta}_{(i)} - \widehat{\beta})}{pS_{(i)}^2}$$

- Andrews and Pregibon Statistic. Consider the augmented matrix $X_A = (X, Y)$.

$$AP_i = \frac{|Z'_{(i)}Z_{(i)}|}{|Z'Z|} = \frac{\text{SSE}_{(i)}}{\text{SSE}} \frac{|X'_{(i)}X_{(i)}|}{|X'X|} = \left[1 - \frac{e_i^2}{\text{SSE}(1-h_i)}\right](1-h_i) = \left(1 - \frac{r_i^2}{n-p}\right)(1-h_i)$$

since, by $|A| = |A_{11}|A_{22.1}| = |A_{22}||A_{11.2}|$,

$$|Z'Z| = |X'X||Y'Y - Y'X(X'X)^{-1}X'Y| = |X'X||Y'(I_n - H)Y| = |X'X| \times \text{SSE}.$$

Note that since $r_i^2/(n-p)$ has a Beta$(1/2, (n-p-1)/2)$, the first factor of this statistic is Beta$((n-p-1)/2, 1/2)$. The AP statistic can be interpreted as a hat matrix diagonal. Let $h_{i,A}$ be the ith diagonal element of the hat matrix based on Z, then $AP_i = 1 - h_{i,Z}$.

18 Added variable plot

- Consider the usual regression: $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1}x_{k-1} + \beta_k x_k$.
- If we have n observation, we can estimate $\widehat{\beta}_k$ by setting

$$E(y) = (1_n \; x_2 \cdots x_{k-1}) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{k-1} \end{pmatrix} + x_k\beta_k = X\beta + x_k\beta_k$$

to obtain $\widehat{\beta}_k = (x_k'Qx_k)^{-1}x_k'Qy$, where $Q = I_n - P$. P projects a vector onto $\mathcal{L}(x_1, \ldots, x_{k-1}) = V$.

- How can we obtain $\widehat{\beta}_k$ *graphically*?
  1. Regress y on $x_1, \ldots, x_{k-1}$ and keep residuals as $e_1 = Qy = y_k^{\perp}$, where $1_n'e_1 = \bar{e}_1 = 0$.
  2. Regress x_k on $x_1, \ldots, x_{k-1}$ and keep residuals as $e_2 = Qx_k = x_k^{\perp}$, where $1_n'e_2 = \bar{e}_2 = 0$.
  3. Regress e_1 on e_2, i.e., $e_1 = \gamma_0 + \gamma_1 e_2$. Then the least squares estimates $\widehat{\gamma}_0 = \bar{e}_1 - \widehat{\gamma}_1\bar{e}_2 = 0$ and

$$\widehat{\gamma}_1 = \frac{\sum_{i=1}^n e_{2i}e_{1i}}{\sum_{i=1}^n e_{2i}^2} = \frac{e_2'e_1}{\|e_2\|^2} = \frac{(Qx_k)'(Qy)}{\|Qx_k\|^2} = \frac{x_k'Qy}{x_k'Qx_k} = \frac{(y, x_k^{\perp})}{\|x_k^{\perp}\|^2} = \widehat{\beta}_k.$$

- Added-variable plots are plots of r_1 against r_2, which is helpful for identifying influential observations; suggesting whether the relationship is monotone or not as well as checking for nonlinearity.

19 Test for Outliers

- Suppose that we wish to test if k obs contain outliers, assuming that the remaining $n-k$ cases are clean. Arrange the data so that the clean obs come first, followed by the k possibly outlying obs. We will use the *outlier shift* model.

$$Y = X\beta + Z\gamma + \epsilon, \quad Z = \begin{pmatrix} 0 \\ I_k \end{pmatrix},$$

where γ is a k-vector containing the shifts for the possibly outlying obs. Let H be a hat matrix under $H : \gamma = 0$, i.e., $H = P_X = X(X'X)^{-1}X'$ and let

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix},$$

where H_{11} is $n-k$ by $n-k$, and $(I-H)Y = e = (e_1', e_2')'$. Then the LSE of γ is

$$\widehat{\gamma} = [Z'(I-H)Z]^{-1}Z'(I-H)Y = (I_k - H_{22})^{-1}e_2.$$

Further, the numerator of the F-test for testing $H : \gamma = 0$ is

$$\widehat{\gamma}' Z'(I - H)Y = e_2'(I_k - H_{22})^{-1} e_2,$$

so that the F-statistic is given by

$$F = \frac{e_2'(I_k - H_{22})^{-1} e_2/k}{[\text{SSE}_H - e_2'(I_k - H_{22})^{-1} e_2]/(n - p - k)}.$$

When $k = 1$, we can test if the ith observation is an outlier. Then

$$F = \frac{e_i^2/(1 - h_i)}{[\text{SSE}_H - e_i^2/(1 - h_i)]/(n - p - 1)} = \frac{(n - p - 1)e_i^2}{(1 - h_i)\text{SSE}_H - e_i^2} = t_i^2 \ \sim \ F_{1,n-p-1}$$

under H, i.e., when the ith observation is *not* an outlier; a large F suggests otherwise as e_i^2 is large.

20 Remedies for Collinearity

- Centering and scaling do not cure collinearity, just leads to simple formulas.
  - Scaling merely changes the units of measurement and leads to variability on a different scale.
  - Centering merely reparameterizes the regression surface, while the estimated surface (i.e., the set of fitted values) remains the same.
- There are three ways to overcome the effects of collinearity.
  1. Collect fresh data to repair the deficiencies in the regression matrix. But this is not always possible.
  2. Discard variables until the remaining set is not collinear. Yet, the variable deleted may well have a relationship with the outcome.
  3. Abandon the use of least squares and use a *biased estimation method* such as **ridge regression** and **principal component regression**.

21 Shrinkage estimators

- As mentioned above, we may wish to abandon the use of least squares. Instead, we shall use various *shrinkage* estimates where the least squares estimates of $\boldsymbol{\beta}$ are shrunk toward zero.
- Define $\text{MSE}(\widehat{\beta}) = E[(\widehat{\beta} - \beta)'(\widehat{\beta} - \beta)] = E[\|\widehat{\beta} - \beta\|]^2$. If $\widehat{\beta}$ is unbiased, then $\text{MSE}(\widehat{\beta}) = \text{Var}(\widehat{\beta})$.

  Consider a *shrinkage* estimate $\widetilde{\beta} = \widehat{\beta}/a$ $(a > 1)$. $\widetilde{\beta}$ is biased as $E(\widetilde{\beta}) = \beta/a < \beta$ and hence

$$\begin{aligned}
\text{MSE}(\widetilde{\beta}) &= E[\|\widetilde{\beta} - E(\widetilde{\beta}) + E(\widetilde{\beta}) - \beta\|^2] \\
&= E[\|\widetilde{\beta} - E(\widetilde{\beta})\|^2] + E[\|E(\widetilde{\beta}) - \beta)\|^2] \\
&= \text{tr}[\text{Var}(\widetilde{\beta})] + \text{Bias}^2
\end{aligned}$$

  Suppose $\widehat{\beta} \sim N(1, 1)$ and hence $\widetilde{\beta} \sim N(1/a, 1/a^2)$. Then

$$\text{MSE}(\widetilde{\beta}) = \frac{1}{a^2} + \left(\frac{1}{a} - 1\right)^2 \equiv g(a).$$

  Note when $a = 1$ (unbiased), $g(1) = 1$. Taking $g'(a)$ and $g''(a) > 0$ shows $\text{MSE}(\widetilde{\beta})$ is minimized at $a = 2$; $g(2) = 0.5 < g(1)$. That is, this $\widetilde{\beta}$ is biased but more efficient than $\widehat{\beta}$.

- **Stein Shrinkage**. Consider $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \sigma^2\mathbf{I}_p)$, where $p > 2$. Then the LS estimate $\widehat{\boldsymbol{\mu}} = \mathbf{Y}$ is the minimum variance unbiased estimate. However, its squared length $\|\mathbf{Y}\|^2 = \mathbf{Y}'\mathbf{Y}$ tends to be too large;

$$E(\|\widehat{\boldsymbol{\mu}}\|^2) = E\|\mathbf{Y}\|^2 = \widehat{\boldsymbol{\mu}}'\widehat{\boldsymbol{\mu}} + \text{tr}(\sigma^2\mathbf{I}_p) = \|\widehat{\boldsymbol{\mu}}\|^2 + p\sigma^2 \geq \|\widehat{\boldsymbol{\mu}}\|^2.$$

  This suggests "shrinking" the elements of $\mathbf{Y}$, and considering an estimate of the form $\widetilde{\boldsymbol{\mu}} = c\mathbf{Y}$, where $0 < c < 1$. $\widetilde{\boldsymbol{\mu}}$ is biased, but it is possible to choose c so that $\widetilde{\boldsymbol{\mu}}$ has a smaller MSE than $\widehat{\boldsymbol{\mu}} = \mathbf{Y}$ ($c = 1$).

$$\begin{aligned} E\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 &= E\|c\mathbf{Y} - \boldsymbol{\mu}\|^2 \\ &= c^2E\|\mathbf{Y}\|^2 - 2cE(\mathbf{Y}'\boldsymbol{\mu}) + \|\boldsymbol{\mu}\|^2 \\ &= c^2(\|\boldsymbol{\mu}\|^2 + p\sigma^2) - 2c\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|^2 \equiv g(c). \end{aligned}$$

  Since $g'(c) = 2c(\|\boldsymbol{\mu}\|^2 + p\sigma^2) - 2\|\boldsymbol{\mu}\|^2$ and $g''(c) = 2(\|\boldsymbol{\mu}\|^2 + n\sigma^2) > 0$, $g(c)$ is minimized at

$$c^* = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + p\sigma^2} = 1 - \frac{n\sigma^2}{\|\boldsymbol{\mu}\|^2 + p\sigma^2} \ (< 1) \quad \Rightarrow \quad \widetilde{\boldsymbol{\mu}} = \left(1 - \frac{p\sigma^2}{\|\boldsymbol{\mu}\|^2 + p\sigma^2}\right)\mathbf{Y}.$$

  Unfortunately, this is not a practical estimate, since $\|\boldsymbol{\mu}\|$, we are trying to estimate, and σ^2 are unknown. In practice, we use the **Stein's estimator**

$$\widetilde{\boldsymbol{\mu}} = \left(1 - \frac{c\widehat{\sigma}^2}{\|\mathbf{Y}\|^2}\right)\mathbf{Y},$$

  where c is some constant, as we know $E\|\mathbf{Y}\|^2 = \|\widehat{\boldsymbol{\mu}}\|^2 + p\sigma^2$. Further, James and Stein showed that of all estimates of the form $(1 - b/\|\mathbf{Y}\|)\mathbf{Y}$, the best of choice is $b = (p-2)\widehat{\sigma}^2$, i.e.,

$$\widetilde{\boldsymbol{\mu}} = \left(1 - \frac{(p-2)\widehat{\sigma}^2}{\|\mathbf{Y}\|^2}\right)\mathbf{Y},$$

  which can be improved even further, though.

- This James-Stein shrinkage estimate also has a nice Bayesian interpretation. Suppose $\mu_i \overset{iid}{\sim} N(0, \sigma_0^2)$ and $Y_i \mid \mu_i \overset{indep}{\sim} N(\mu_i, \sigma^2)$, then

$$\boldsymbol{\mu} \mid \mathbf{Y} \ \sim \ N((1-\omega)\mathbf{Y}, \ \omega\sigma_0^2\mathbf{I}_p), \quad \omega = \frac{\sigma^2}{\sigma^2 + \sigma_0^2},$$

  meaning that the Bayes estimate of $\boldsymbol{\mu}$ is the posterior mean $(1-\omega)\mathbf{Y}$.

- Assuming that σ^2 is known, this Bayesian approach requires that we specify a value for σ_0^2. An alternative approach is to consider the marginal distribution of $\mathbf{Y}$ as $\mathbf{Y} \sim N_p(\mathbf{0}, (\sigma^2 + \sigma_0^2)\mathbf{I}_p)$, as

$$p(y_i) = \int_{-\infty}^{\infty} p(y_i \mid \mu_i)p(\mu_i)d\mu_i = \cdots = \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_0^2)}} \exp\left[-\frac{y_i^2}{2(\sigma^2 + \sigma_0^2)}\right],$$

  so that we have

$$\frac{\mathbf{Y}'\mathbf{Y}}{\sigma^2 + \sigma_0^2} \sim \chi_p^2 \quad \Rightarrow \quad E\left(\frac{\sigma^2 + \sigma_0^2}{\|\mathbf{Y}\|^2}\right) = \frac{1}{p-2}.$$

  It follows that $(p-2)/\|\mathbf{Y}\|^2$ is an unbiased estimate of $1/(\sigma^2 + \sigma_0^2)$, leading to estimate $\widetilde{\boldsymbol{\mu}}$. This type of estimate, based on the posterior mean but using the marginal distribution, is called an *empirical Bayes estimate*.

- In the case of regression with p *orthonormal* predictors and no constant term, we can apply the James-Stein estimate directly by setting $\mathbf{Y} = \widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\mu} = \boldsymbol{\beta}$. In the orthonormal case $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2\mathbf{I}_p)$, so that the Stein's estimator becomes

$$\widetilde{\boldsymbol{\beta}} = \left(1 - \frac{c\widehat{\sigma}^2}{\|\boldsymbol{\beta}\|^2}\right)\boldsymbol{\beta},$$

  where c may be $p-2$. This has the smallest MSE.

- Assume $Y \mid \beta \sim N_n(X\beta, \sigma^2 I_n)$ and the prior $\beta \sim N_p(m, \sigma^2 V)$, then the posterior mean of β is

$$m^* = E(\beta \mid Y) = (V^{-1} + X'X)^{-1}(X'Y + V^{-1}m).$$

If we take $V = r^{-1}I_p$ and $m = 0$, then the Bayes estimate is $\widehat{\beta}_r = (X'X + rI_p)^{-1}X'y$, which is the same form as the ridge estimate. If $V = r^{-1}(X'X)^{-1}$ and $m = 0$, then

$$\widetilde{\beta} = (1+r)^{-1}(X'X)^{-1}X'Y = \frac{\widehat{\beta}}{1+r},$$

which is often called the *James-Stein regression* estimate.

22 Ridge regression

- Ridge regression has a means not only for improving the estimation of β when the predictors are highly correlated, but also for improving the accuracy of prediction. The ridge estimate is given by

$$\widehat{\beta}_r = (X'X + rI_p)^{-1}X'y = (X'X + rI_p)^{-1}X'X\widehat{\beta}, \quad r \geq 0.$$

- $\widehat{\beta}_r$ is biased; $E(\widehat{\beta}_r) = (X'X + rI_p)^{-1}X'X\beta \neq \beta$ unless $r = 0$.

- Let $\lambda_1, \ldots, \lambda_p$ be eigenvalues of $X'X$, then by spectral decomposition, $X'X = T\Lambda T'$, where $TT' = T'T = I_p$. Further let $T = (t_1, \ldots, t_p)$. Then,

$$\widehat{\beta}_r = T(\Lambda + rI_p)^{-1}T'T\Lambda T'\widehat{\beta} = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + r} t_i t_i'\widehat{\beta} \quad < \quad \sum_{i=1}^{p} t_i t_i'\widehat{\beta} = \widehat{\beta},$$

$$\begin{aligned}\operatorname{Var}(\widehat{\beta}_r) &= \sum_{i=1}^{p}\sum_{j=1}^{p} \operatorname{Cov}\left(\frac{\lambda_i}{\lambda_i + r} t_i t_i'\widehat{\beta}, \frac{\lambda_j}{\lambda_j + r} t_j t_j'\widehat{\beta}\right) = \sum_{i=1}^{p} \left(\frac{\lambda_i}{\lambda_i + r}\right)^2 t_i t_i' \operatorname{Var}(\widehat{\beta}) \\ &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + r)^2} t_i t_i' \quad \preceq \quad \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} t_i t_i' = \operatorname{Var}(\widehat{\beta}).\end{aligned}$$

since $t_i't_j = \delta_{ij}$ and $\operatorname{Var}(\widehat{\beta}) = \sigma^2(X'X)^{-1} = \sigma^2 T\Lambda^{-1}T'$. Check if $r = 0$ for verification.

- Compare the two estimates, taking trace and determinant. Note $\operatorname{tr}(t_i t_i') = \operatorname{tr}(t_i' t_i) = 1$ and $|T||T'| = 1$.

$$\operatorname{tr}[\operatorname{Var}(\widehat{\beta}_r)] - \operatorname{tr}[\operatorname{Var}(\widehat{\beta})] = \sigma^2 \sum_{j=1}^{p} \left[\frac{\lambda_i}{(\lambda_i + r)^2} - \frac{1}{\lambda_i}\right] = -\sigma^2 \sum_{j=1}^{p} \frac{r(r + 2\lambda_i)}{(\lambda_i + r)^2 \lambda_i} \leq 0$$

$$\frac{|\operatorname{Var}(\widehat{\beta}_r)|}{|\operatorname{Var}(\widehat{\beta})|} = \frac{\prod_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + r)^2}}{\prod_{i=1}^{p} \frac{1}{\lambda_i}} = \prod_{i=1}^{p} \left(\frac{\lambda_i}{\lambda_i + r}\right)^2 \leq 1.$$

Thus, $\operatorname{tr}[\operatorname{Var}(\widehat{\beta}_r)] \leq \operatorname{tr}[\operatorname{Var}(\widehat{\beta})]$ and $\det[\operatorname{Var}(\widehat{\beta}_r)] \leq \det[\operatorname{Var}(\widehat{\beta})]$, i.e., The trace and the determinant of the ridge estimate are smaller than LS estimate.

- What about the difference in trace of MSE? MSE consists of variance and bias. The bias of $\widehat{\beta}_r$ is

$$E(\widehat{\beta}_r) - \beta = \sum_{i=1}^{p} \left(\frac{\lambda_i}{\lambda_i + r} - 1\right) t_i t_i'\beta = -\sum_{i=1}^{p} \left(\frac{r}{\lambda_i + r}\right) t_i t_i'\beta \ < \ \mathbf{0}$$

and $\widehat{\beta}$ is unbiased, so that

$$\begin{aligned}M(r) &= \operatorname{tr}[\operatorname{MSE}(\widehat{\beta}_r) - \operatorname{MSE}(\widehat{\beta})] \\ &= \operatorname{tr}[\operatorname{Var}(\widehat{\beta}_r)] - \operatorname{tr}[\operatorname{Var}(\widehat{\beta})] + \operatorname{tr}[(\text{Bias of } \widehat{\beta}_r)(\text{Bias of } \widehat{\beta}_r)'] \\ &= -\sum_{j=1}^{p} \frac{\sigma^2 r(r + 2\lambda_i)}{(\lambda_i + r)^2 \lambda_i} + \sum_{i=1}^{p} \frac{r^2 (t_i'\beta)^2}{(\lambda_i + r)^2} = \sum_{i=1}^{p} \frac{[\lambda_i (t_i'\beta)^2 - \sigma^2] r^2 - 2\sigma^2 \lambda_i r}{(\lambda_i + r)^2 \lambda_i}.\end{aligned}$$

By a complicated calculation, the derivative is

$$M'(r) = \sum_{j=1}^{p} \frac{2\lambda_i[r(t_i'\beta)^2 - \sigma^2]}{(\lambda_i + r)^3},$$

which is zero if $r^* = \sigma^2/(t_i'\beta)^2$. This implies that there exists some $r > 0$ such that $M(r) = 0$. In other words, we can find some $r > 0$ such that $\text{tr}[\text{MSE}(\widehat{\beta}_r)] < \text{tr}[\text{MSE}(\widehat{\beta})]$.

- Also, we can compare the two mean squares errors themselves. Then for some r, $\text{MSE}(\widehat{\beta}_r) \preceq \text{MSE}(\widehat{\beta})$. Let $G = (X'X + rI_p)^{-1}$ ($G' = G$), then we have $\widehat{\beta}_r = GX'y$ and

$$E(\widehat{\beta}_r) = GX'X\beta = G(X'X + rI_p)\beta - rG\beta = \beta - rG\beta.$$

Thus, bias of $\widehat{\beta}_r$ is $E(\widehat{\beta}_r) - \beta = -rG\beta$. Hence,

$$\begin{aligned}
\text{MSE}(\widehat{\beta}_r) &= \text{Var}(\widehat{\beta}_r) + (\text{Bias})(\text{Bias})' = \sigma^2 GX'XG + r^2 G\beta\beta'G = G(\sigma^2 X'X + r^2\beta\beta')G, \\
\text{MSE}(\widehat{\beta}) &= \text{Var}(\widehat{\beta}) = \sigma^2 G[G^{-1}(X'X)^{-1}G^{-1}]G = G[\sigma^2 X'X + 2\sigma^2 rI_p + \sigma^2 r^2 (X'X)^{-1}]G,
\end{aligned}$$

so that

$$\begin{aligned}
\Delta = \text{MSE}(\widehat{\beta}) - \text{MSE}(\widehat{\beta}_r) &= G[\sigma^2(2rI_p + r^2(X'X)^{-1}) - r^2\beta\beta']G \\
&= rG[\sigma^2(2I_p + r(X'X)^{-1}) - r\beta\beta']G.
\end{aligned}$$

Since $G \succeq O$, $\Delta \succeq O$ if and only if $\sigma^2(2I_p + r(X'X)^{-1}) - r\beta\beta' \succeq O$. Then the sufficient (not necessary) condition is

$$2\sigma^2 I_p - r\beta\beta' \succeq O \;\Leftrightarrow\; I_p - \sqrt{\frac{r}{2\sigma^2}}\beta\left(\sqrt{\frac{r}{2\sigma^2}}\beta\right)' \succeq O \;\Leftrightarrow\; 1 - \frac{r}{2\sigma^2}\|\beta\|^2 \geq 0.$$

Hence, $\text{MSE}(\widehat{\beta}) \succeq \text{MSE}(\widehat{\beta}_r)$ at least when $0 < r \leq 2\sigma^2/\|\beta\|^2$ as $\|\beta\|^2 > 0$.

- Lemma: $\mathbf{I}_n - \mathbf{a}\mathbf{a}' \succeq \mathbf{O} \;\Leftrightarrow\; 1 - \mathbf{a}'\mathbf{a} \geq 0$.

($\Rightarrow$) If $\mathbf{I} - \mathbf{a}\mathbf{a}' \succeq \mathbf{O}$, then $\mathbf{x}'(\mathbf{I} - \mathbf{a}\mathbf{a}')\mathbf{x} = \|\mathbf{x}\|^2 - (\mathbf{a}'\mathbf{x})^2 \geq 0 \;\Rightarrow\; (\mathbf{a}'\mathbf{x})^2 \leq \|\mathbf{x}\|^2$. But, Chebyshev's inequality says $(\mathbf{a}'\mathbf{x})^2 \leq \|\mathbf{x}\|^2\|\mathbf{a}\|^2$, implying that $\mathbf{a}'\mathbf{a} \leq 1$ is necessary.

($\Leftarrow$) If $1 - \mathbf{a}'\mathbf{a} \geq 0$, then $\mathbf{x}'(\mathbf{I} - \mathbf{a}\mathbf{a}')\mathbf{x} = \|\mathbf{x}\|^2 - (\mathbf{a}'\mathbf{x})^2 \geq \|\mathbf{x}\|^2(1 - \mathbf{a}'\mathbf{a}) \geq 0$, meaning that $\mathbf{I} - \mathbf{a}\mathbf{a}' \succeq \mathbf{O}$.

23 Principal Component (PC) regression

- Principal component analysis (PCA) is a dimension reduction technique that finds the most informative linear combinations of the p random variables $x \in \mathbb{R}^p$ by transforming it into a new coordinate system where the principal components capture the most variation in the data.
- That is, let $\text{Cov}(x) = \Sigma$, then we want to find α_1 that maximizes

$$\text{Cov}(\alpha_1' x) = \alpha_1'\Sigma\alpha_1 \quad \text{subject to} \quad \alpha_1'\alpha_1 = 1.$$

Given a predictor matrix $X = (x_1, \ldots, x_n)' \in \mathbb{R}^{n\times p}$ that is column-normalized in advance. Normalizing X (at least centering it) is essential because PCA is sensitive to data centering. Then we use X to estimate the covariance as $\widehat{\Sigma} = X'X/(n-1)$. Hence, we practically consider $\max_{\alpha_1'\alpha_1=1} \alpha_1' X'X\alpha_1$. Then α_1 is the largest eigenvector of $X'X$.

24 Ridge vs PC regression using SVD

- Consider a standard linear model: $Y = X\beta = \epsilon$.
- By the singular value decomposition of X, we have $X = U\Sigma V'$, where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p)$, assuming $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_p \geq 0$, and $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ are both orthogonal sets of vectors denoting the left and right singular vectors of X, respectively.
- In this setting, the fitted values for ridge regression is given by

$$\begin{aligned}
\widehat{Y}_{\text{ridge}} &= X\widehat{\beta}_\lambda \\
&= X(X'X + \lambda I_p)^{-1}X'Y \\
&= U\Sigma V'(V\Sigma^2 V' + \lambda I_p)^{-1}V\Sigma U'Y \quad \because U'U = I_p \\
&= U\Sigma V'[V(\Sigma^2 + \lambda I_p)V']^{-1}V\Sigma U'Y \quad \because VV' = I_p \\
&= U\Sigma(\Sigma^2 + \lambda I_p)^{-1}\Sigma U'Y \\
&= \sum_{j=1}^{p} u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j'Y.
\end{aligned}$$

- In terms of PCA regression, we can write $E(Y) = X\beta = \beta_0 + X(VV')\beta = (XV)(V'\beta) = W\gamma$, where the columns of $W = XV$ are called the (full) principal scores. We can choose the value of $k \in \{1, \ldots, p\}$ so that $W_k \in \mathbb{R}^{n\times k}$ and $\gamma \in \mathbb{R}^k$. Then since $X = U_k\Sigma_k V_k'$ so that $W_k = XV_k = U_k\Sigma_k$, we have

$$\widehat{Y}_{\text{PCA}} = W_k(W_k'W_k)^{-1}W_k'Y = U_k\Sigma_k(\Sigma_k U_k'U_k\Sigma_k)^{-1}\Sigma_k U_k'Y = U_kU_k'Y = \sum_{j=1}^{k} u_j u_j'Y.$$

- We see that PCA sets the k largest singular values to be 1 and the remaining smaller ones to be 0. On the other hand, for Ridge regression, λ shrinks the effect of each principal component by a factor or $\sigma_j^2/(\sigma_j^2 + \lambda)$; in particular, it has a smaller impact on larger singular values but a bigger impact on smaller singular values, meaning that ridge regression is interpreted as a soft PCA regression method.

25 Wald/Score/LR test

- Wald test. Let $\widehat{\theta}$ be the MLE of $\theta \in \mathbb{R}^p$. Then $\widehat{\theta} \sim N_p(\theta, I^{-1}(\theta))$. $\text{Var}(\widehat{\theta}) = I^{-1}(\theta) \Rightarrow \widehat{\text{Var}}(\widehat{\theta}) = I^{-1}(\widehat{\theta})$. Under $H_0 : \theta = \theta_0$, the Wald test statistic is given by

$$X_W = (\widehat{\theta} - \theta_0)'I(\theta_0)(\widehat{\theta} - \theta_0) \quad \text{or} \quad (\widehat{\theta} - \theta_0)'I(\widehat{\theta})(\widehat{\theta} - \theta_0) \quad \sim \quad \chi_p^2(0).$$

- Score test. Define the *score function* u such that

$$u = u(y \mid \theta) = \frac{\partial \ln f(y \mid \theta)}{\partial \theta}, \quad E(u) = 0.$$

  Since $\text{Var}(u) = E(uu') = I(\theta)$, we have $u \sim N_p(0, I(\theta))$. Under $H_0 : \theta = \theta_0$, the score test statistic is

$$X_S = u'(\theta_0)I^{-1}(\theta_0)u(\theta_0) \quad \text{or} \quad u'(\theta_0)I^{-1}(\widehat{\theta})u(\theta_0) \quad \sim \quad \chi_p^2(0),$$

  where there are no other nuisance parameters (for the first definition).
- Likelihood Ratio test. Let Ω be the whole parameter space and

$$\begin{aligned}
\widehat{\theta} &= \arg_{\theta\in\Omega} \max \ell(\mathbf{y} \mid \theta), \\
\widehat{\theta}_0 &= \arg_{\theta\in H_0} \max \ell(\mathbf{y} \mid \theta).
\end{aligned}$$

Then LR test statistic is $\Lambda < c$, where

$$\Lambda = \frac{L(\widehat{\theta}_0 \mid \mathbf{y})}{L(\widehat{\theta} \mid \mathbf{y})}.$$

Further, asymptotically,

$$-2 \ln \Lambda = 2[\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0)] \xrightarrow{D} \chi^2_p(0).$$

- The likelihood ratio statistic is invariant under parameterization (e.g., log or exp). However, the calculation of a likelihood ratio test requires fitting two (smaller and larger) models compared to only one model for the Wald test (larger model) and sometimes no model at all for the score test.
- The Wald statistic is not invariant to the parameterization chosen. The Wald statistic uses the MLE but not the value of the maximized likelihood.
- The three test statistics are asymptotically equivalent and are therefore expected to give similar results in large samples. Their small-sample properties are not known, but some simulation studies suggest that the LRT test may be better than its competitors in small samples.

26 Detecting Nonconstant Variance

- We consider $Y_i = x_i'\beta + \epsilon_i$, where

$$\text{var}[\epsilon_i] = \sigma_i^2 = w(z_i, \boldsymbol{\lambda}) \equiv w_i,$$

where σ_i^2 depends either on the mean $E[Y_i] = x_i'\beta$ or on vector of (possibly additional) known predictors z_i. Also, w is a variance function with the property that for some $\boldsymbol{\lambda}_0$, $w(z, \boldsymbol{\lambda}_0)$ does not depend on $\mathbf{z}$.

- Then although the least squares estimate of β is still unbiased, it may not be efficient if $\sigma_i^2 \neq \sigma^2$ (non constant variance). We need to test $H_0 : \sigma_i^2 = \sigma^2$ or equivalently, $H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}_0$.
- A popular choice is $w(z, \boldsymbol{\lambda}) = \exp(z'\boldsymbol{\lambda})$, where $z = (1, z_1, \ldots, z_k)$. In this case, $\boldsymbol{\lambda}_0 = (\lambda_0, 0 \ldots, 0)'$ and, for this value, $\text{var}[Y_i] = \exp(z'\boldsymbol{\lambda}_0) = e^{\lambda_0} = \sigma^2$.
- The residuals e from a least squares fit contain information about the variances. If $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, then $\text{Var}(e) = (I_n - H)\,\text{Var}(Y) = (I_n - H)\Sigma(I_n - H)$, so that

$$\text{var}[e_i] = E[e_i^2] = (1 - h_i)^2\sigma_i^2 + \sum_{k \neq i} h_{ki}^2 \sigma_k^2.$$

- $b_i = e_i^2/(1 - h_i)$ are useful when making various kinds of graphic displays. When H is true, then

$$E[b_i] = (1 - h_i)\sigma^2 + \sum_{k \neq i} \frac{h_{ki}^2}{1 - h_i}\sigma^2 = \sigma^2$$

since $h_i(1-h_i) = \sum_{k \neq i} h_{ki}^2$. Thus, when the variances are constant, the b_i's have constant expectations.

- Even if the variances are unequal, the fitted values $\widehat{Y}_i = x_i'\widehat{\beta}$ still have expectation $E[Y_i]$. However, obs with large means often have large variances, too. Thus, plotting the b_i's (or equivalently, r_i^2) against $\widehat{Y}_i$ should result in a wedge-shaped display if the variances increase with the means. Alternatively, e_i's (raw residuals) against $\widehat{Y}_i$ have a fan-shaped pattern, indicating variances increasing with the means.
- Inference. The log-likelihood function is given by

$$\begin{aligned}\ell(\beta, \lambda) &= c - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta) \\ &= c - \frac{1}{2}\sum_{i=1}^{n} \log w_i - \frac{1}{2}\sum_{i=1}^{n} \frac{(Y_i - x_i'\beta)^2}{w_i},\end{aligned}$$

so that the score functions are

$$\frac{\partial \ell}{\partial \beta} = X'\Sigma^{-1}(Y - X\beta) = X'\Sigma^{-1}\epsilon,$$
$$\frac{\partial \ell}{\partial \lambda} = -\frac{1}{2}\left[\sum_{i=1}^{n}\left\{\frac{1}{w_i} - \frac{(Y_i - x_i'\beta)^2}{w_i^2}\right\}\frac{\partial w_i}{\partial \lambda}\right].$$

The MLEs can be obtained by the following algorithm.

1. Put $\lambda = \lambda_0$.
2. Compute $\widehat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$.
3. Solve $\partial \ell(\widehat{\beta}, \lambda)/\partial\lambda = 0$ for λ.
4. Repeat steps 2 and 3 until converge.

- In the special case $w_i = \exp[z_i'\boldsymbol{\lambda}]$, step 3 can be implemented as a least squares calculation. Then

$$\frac{\partial \ell}{\partial \lambda} = -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{w_i} - \frac{\epsilon_i^2}{w_i^2}\right)z_i w_i = -\frac{1}{2}\sum_{i=1}^{n}\left(1 - \frac{\epsilon_i^2}{w_i}\right)z_i = \frac{1}{2}Z'(d - 1_n),$$

where $d_i = \epsilon_i^2/w_i \sim \chi_1^2$, $d = (d_1, \ldots, d_n)'$, $d_i \perp d_j$ $(i \neq j)$, and Z is the matrix whose ith row is z_i. Then the information matrix is given by

$$I(\beta, \lambda) = \begin{pmatrix} \mathrm{Var}\left(\frac{\partial \ell}{\partial \beta}\right) & \mathrm{Cov}\left(\frac{\partial \ell}{\partial \beta}, \frac{\partial \ell}{\partial \lambda'}\right) \\ \mathrm{Cov}\left(\frac{\partial \ell}{\partial \lambda}, \frac{\partial \ell}{\partial \beta'}\right) & \mathrm{Var}\left(\frac{\partial \ell}{\partial \lambda}\right) \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}X & O \\ O' & \frac{1}{2}Z'Z \end{pmatrix}.$$

since $\mathrm{cov}(\epsilon_i, d_j) = 0$ for all i, j. To solve the likelihood equations we can use Fisher scoring. The updating equations are

$$\widehat{\beta}_{(m+1)} = (X'\widehat{\Sigma}_{(m)}^{-1}X)^{-1}X'\widehat{\Sigma}_{(m)}^{-1}Y$$
$$\widehat{\lambda}_{(m+1)} = \widehat{\lambda}_{(m)} + (Z'Z)^{-1}Z'(d - 1_n)$$
$$\widehat{\Sigma}_{(m+1)} = \mathrm{diag}\left[w(z_1, \widehat{\lambda}_{(m)}), \ldots, w(z_n, \widehat{\lambda}_{(m)})\right].$$

We note that $Z'Z\widehat{\lambda}_{(m+1)} = Z'(d - 1_n + Z\widehat{\lambda}_{(m)})$ are the normal equations for a formal regression of $d - 1_n + Z\lambda_{(m)}$ on Z.

- We can test $H_0 : \lambda = \lambda_0$ $(\Sigma = \sigma^2 I_n)$ using a likelihood ratio test or score test. The score test has the advantage that we do not have to calculate the unrestricted MLEs. The score test statistic is given by

$$\left(\frac{\partial \ell}{\partial \lambda}\Big|_{H_0}\right)' I(H_0)_{\lambda\lambda}^{-1}\left(\frac{\partial \ell}{\partial \lambda}\Big|_{H_0}\right) = \frac{1}{2}(u - 1_n)'D(D'D)^{-1}D'(u - 1_n).$$

Proof (250B HW4). we have the score functions (the score vector)

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}, \quad \frac{\partial \ell}{\partial \boldsymbol{\lambda}} = -\frac{1}{2}\sum_{i=1}^{n}\left(1 - \frac{\epsilon_i^2}{w_i}\right)\frac{1}{w_i}\frac{dw_i}{d\boldsymbol{\lambda}}.$$

Under $H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}_0$, since $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ and $w_i = \sigma^2$, the score functions are

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}}\Big|_{H_0} = \frac{\mathbf{X}'\boldsymbol{\epsilon}}{\sigma^2}, \quad \frac{\partial \ell}{\partial \boldsymbol{\lambda}}\Big|_{H_0} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{\epsilon_i^2}{\sigma^2} - 1\right)\frac{dw_i}{d\boldsymbol{\lambda}} = \frac{1}{2\sigma^2}\mathbf{D}'(\mathbf{d} - \mathbf{1}_n),$$

where $\mathbf{D}$ is the matrix with i-th row $(\partial w_i/\partial \boldsymbol{\lambda})'$ and $\mathbf{d} = (d_1, \ldots, d_n)$, where $d_i = \epsilon_i^2/\sigma^2 \overset{indep}{\sim} \chi_1^2$. Then their variances and covariances are, respectively,

$$\begin{aligned}
\text{Var}\left(\left.\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right|_{H_0}\right) &= \text{Var}\left(\frac{\mathbf{X}'\boldsymbol{\epsilon}}{\sigma^2}\right) = \frac{\mathbf{X}'\text{Var}(\boldsymbol{\epsilon})\mathbf{X}}{\sigma^4} = \frac{\mathbf{X}'\mathbf{X}}{\sigma^2}, \\
\text{Var}\left(\left.\frac{\partial \ell}{\partial \boldsymbol{\lambda}}\right|_{H_0}\right) &= \text{Var}\left(\frac{1}{2\sigma^2}\mathbf{D}'(\mathbf{d} - \mathbf{1}_n)\right) = \frac{\mathbf{D}'\text{Var}(\mathbf{d})\mathbf{D}}{4\sigma^4} = \frac{\mathbf{D}'\mathbf{D}}{2\sigma^4}, \\
\text{Cov}\left(\left.\frac{\partial \ell}{\partial \boldsymbol{\beta}}, \frac{\partial \ell}{\partial \boldsymbol{\lambda}}\right|_{H_0}\right) &= \text{Cov}\left(\frac{\mathbf{X}'\boldsymbol{\epsilon}}{\sigma^2}, \frac{1}{2\sigma^2}\mathbf{D}'(\mathbf{d} - \mathbf{1}_n)\right) = \frac{\mathbf{X}'\text{Cov}(\boldsymbol{\epsilon}, \mathbf{d})\mathbf{D}}{2\sigma^4} = \mathbf{0}
\end{aligned}$$

because $\text{var}(\epsilon_i^3) = 0$ and $\text{cov}(\epsilon_i, \epsilon_j^2) = 0,\ i \neq j$. Hence, the information matrix under H_0 is

$$\mathbf{I}(H_0) = \begin{pmatrix} \mathbf{X}'\mathbf{X}/\sigma^2 & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2}\mathbf{D}'\mathbf{D}/\sigma^4 \end{pmatrix} \Rightarrow \mathbf{I}^{-1}(H_0) = \begin{pmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & 2\sigma^4(\mathbf{D}'\mathbf{D})^{-1} \end{pmatrix}.$$

Therefore, the score test statistic is

$$\begin{aligned}
\left(\left.\frac{\partial \ell}{\partial \boldsymbol{\lambda}}\right|_{H_0}\right)' \mathbf{I}^{-1}(H_0)_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \left(\left.\frac{\partial \ell}{\partial \boldsymbol{\lambda}}\right|_{H_0}\right) &= \frac{1}{2\sigma^2}(\mathbf{d} - \mathbf{1}_n)'\mathbf{D}[2\sigma^4(\mathbf{D}'\mathbf{D})^{-1}]\frac{1}{2\sigma^2}\mathbf{D}'(\mathbf{d} - \mathbf{1}_n) \\
&= \frac{1}{2}(\mathbf{d} - \mathbf{1}_n)'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{d} - \mathbf{1}_n).
\end{aligned}$$

In practice, since ϵ and σ^2 are unknown, we replace them with e_i and $\widehat{\sigma}^2$ (MLE of σ^2), respectively. Further, let $\mathbf{u} = \widehat{\mathbf{d}} = (u_1, \ldots, u_n)$, where $u_i = \widehat{d}_i = e_i^2/\widehat{\sigma}^2$, then the score test statistic is replaced by

$$\frac{1}{2}(\mathbf{u} - \mathbf{1}_n)'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{u} - \mathbf{1}_n).$$

27 Lack of Fit test

- The lack-of-fit test compares a simple regression $E(y_{ij}) = x_i'\beta = \eta_i$ (reduced model) with one-way ANOVA $y_{ij} \sim N(\mu_i, \sigma^2)$ (full model), **assuming the linearity**. A small p-value indicates a lack-of-fit, i.e., means are not linear.
- Needs multiple observations at various predictors:

$$\begin{aligned}
&n_1 \text{ obs. at } x_{1.}' = (x_{11}, \ldots, x_{1p}), \\
&n_2 \text{ obs. at } x_{2.}' = (x_{21}, \ldots, x_{2p}), \\
&\qquad\vdots \\
&n_g \text{ obs. at } x_{g.}' = (x_{g1}, \ldots, x_{gp}).
\end{aligned}$$

Let $n = \sum_{i=1}^g n_i$. The model is given by

$$E\begin{pmatrix} y_{11} \\ \vdots \\ y_{n_1 1} \\ \vdots \\ y_{1g} \\ \vdots \\ y_{n_p g} \end{pmatrix} = \begin{pmatrix} x_{1.}' \\ \vdots \\ x_{1.}' \\ \vdots \\ x_{g.}' \\ \vdots \\ x_{g.}' \end{pmatrix} \beta$$

- The number of parameters in the full model is g, which is greater than p, that in the reduced model.
- Wish to test $H_0 : E(y) = X\beta = \eta$ (linear) vs. $H_1 : E(y) = \gamma \neq \eta$ (nonlinear).
- Let $\widehat{\eta} = Py$ and then $\widehat{\eta}_{tr} = \widehat{\eta}_{.r}$ for $\forall t = 1, \ldots, n_r$.
- Let $Pr = r_0$ (wrong place). Then $r - r_0$ is called a mean model residual vector and the lack-of-fit parameter is defined as $\Lambda^2 = (r - r_0)'(r - r_0)$. Then $H_0 : E(y) = \eta$ is equivalent to $\Lambda^2 = 0$.
- We can split $y'y$ into

$$\begin{aligned} y'y &= \text{SSReg} + \text{SSE} = y'Py + y'(I - P)y \\ &= \text{SSReg} + \text{SSLoF} + \text{SSPE (pure error)} \\ &= \widehat{\eta}'\widehat{\eta} + y'(I - P - U)y + y'Uy \\ &= \sum_{r=1}^{g}\sum_{t=1}^{n_r}\widehat{\eta}_{.r}^2 + \sum_{r=1}^{g}\sum_{t=1}^{n_r}(\overline{y}_{.r} - \widehat{\eta}_{.r})^2 + \sum_{r=1}^{g}\sum_{t=1}^{n_r}(y_{tr} - \overline{y}_{.r})^2, \\ &= \sum_{r=1}^{g} n_r\widehat{\eta}_{.r}^2 + \sum_{r=1}^{g} n_r(\overline{y}_{.r} - \widehat{\eta}_{.r})^2 + \sum_{r=1}^{g}\sum_{t=1}^{n_r}(y_{tr} - \overline{y}_{.r})^2. \end{aligned}$$

  where

$$U = U' = U^2 = \underbrace{\begin{pmatrix} I_{n_1} - \frac{11'}{n_1} & & O \\ & \ddots & \\ O & & I_{n_g} - \frac{11'}{n_g} \end{pmatrix}}_{n \times n}, \quad \text{rank}(U) = \text{tr}(U) = n - g$$

  and $\text{rank}(I - P - U) = n - p - (n - g) = g - p$. Note that $UP = O$ since

$$UX = \begin{pmatrix} I_{n_1} - \frac{11'}{n_1} & & O \\ & \ddots & \\ O & & I_{n_g} - \frac{11'}{n_g} \end{pmatrix} \underbrace{\begin{pmatrix} x'_{1.} \\ \vdots \\ x'_{1.} \\ \vdots \\ x'_{g.} \\ \vdots \\ x'_{g.} \end{pmatrix}}_{n \times p} = O.$$

  Then $U(I - P - U) = O \Rightarrow$ SSLoF $\perp\!\!\!\perp$ SSPE by Craig's theorem. The F-statistic for testing H_0 is

$$F = \frac{\text{MSLoF}}{\text{MSPE}} \sim F_{g-p,n-g}.$$

- If H_0 is rejected (F is high), we conclude that the regression (reduced) model is not adequate.
- Corresponding expected mean squares are, under H_1,

$$\begin{aligned} E(\text{MSReg}) &= E\left(\frac{y'Py}{p}\right) = \sigma^2 + \frac{\gamma_0'\gamma_0}{p}, \\ E(\text{MSLoF}) &= E\left[\frac{y'(I - P - U)y}{g - p}\right] = \frac{\Lambda^2}{g - p} + \sigma^2, \\ E(\text{MSPE}) &= E\left(\frac{y'Uy}{n - g}\right) = \sigma^2, \\ E(\text{MSTO}) &= E\left(\frac{y'y}{n}\right) = \sigma^2 + \frac{\gamma'\gamma}{n}. \end{aligned}$$

- In practice, run a regression to obtain

$$\text{SSE} = \sum_r \sum_t (y_{tr} - \widehat{\eta}_{\cdot r})^2, \quad \text{SSPE} = \sum_{r=1}^{g} \sum_{t=1}^{n_r} (y_{tr} - \overline{y}_{\cdot r})^2$$

  then calculate SSLoF = SSE − SSPE.

28 Linear Mixed Effects Model

- Consider a linear mixed effects model

$$y = X\beta + Zu + \epsilon \in \mathbb{R}^n, \quad u \sim N_q(0, D), \quad \epsilon \sim N_n(0, R), \quad u \perp\!\!\!\perp \epsilon.$$

- Since $y \mid u \sim N_n(X\beta + Zu, R)$, the joint density of Y and u is given by $f(y \mid u)f(u)$. Hence,

$$\ell(\beta, u) = C - \frac{1}{2}[(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'D^{-1}u],$$
$$\frac{\partial \ell}{\partial \beta} = X'R^{-1}(y - X\beta - Zu), \quad \frac{\partial \ell}{\partial u} = Z'R^{-1}(y - X\beta - Zu) - D^{-1}u$$

  so that we obtain the Henderson normal equations:

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{pmatrix}.$$

- The second equation follows

$$\widetilde{u} = (Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}(y - X\widetilde{\beta})$$

- Let $V = \text{Var}(y) = ZDZ' + R$, then, by Sherman-Morrison formula,

$$V^{-1} = (R + ZDZ')^{-1} = R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})ZR^{-1}.$$

- Substituting $\widetilde{u}$ into the first equation,

$$\begin{aligned} &X'R^{-1}X\widetilde{\beta} + X'R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}(y - X\widetilde{\beta}) = X'R^{-1}y \\ &\Rightarrow X'R^{-1}X\widetilde{\beta} + X'(R^{-1} - V^{-1})(y - X\widetilde{\beta}) = X'R^{-1}y \\ &\Rightarrow X'V^{-1}X\widetilde{\beta} = X'V^{-1}y \quad \Rightarrow \quad \widetilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y. \end{aligned}$$

- By standard algebra,

$$(Z'R^{-1}Z + D^{-1})DZ' = Z'R^{-1}ZDZ' + Z' = Z'R^{-1}(ZDZ' + R) = Z'R^{-1}V$$

  so that $\widetilde{u}$ can be expressed as follows:

$$\widetilde{u} = (Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}VV^{-1}(y - X\widetilde{\beta}) = DZ'V^{-1}(y - X\widetilde{\beta}),$$

  which is called **BLUP** (best linear unbiased prediction) of u. Let $A = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$,

$$\widetilde{u} = DZ'V^{-1}[I - X(X'V^{-1}X)^{-1}X'V^{-1}]y = DZ'Ay.$$

- An alternative approach to get $\widetilde{u}$ is to use the Lagrange multiplier. We seek $C \in \mathbb{R}^{q\times n}$ such that $E(Cy) = E(u) = 0$. $E(Cy) = CX\beta$ follows $CX = 0$ is a necessary condition. Under this restriction, we want to choose C that minimizes

$$\operatorname{Cov}(Cy - u) = \operatorname{Cov}(Cy) - 2\operatorname{Cov}(Cy, u) + \operatorname{Cov}(u) = CVC' - 2CZD + D,$$

leading to the Lagrange function and the derivatives w.r.t C and L

$$\begin{aligned} r(C, L) &= p'(CVC' - 2CZD)p - \operatorname{tr}(CXL), \quad \forall p, \\ \frac{\partial r}{\partial C} &= 2pp'(CV - D'Z') - L'X' = 0, \quad \frac{\partial r}{\partial L} = C'X' = 0. \end{aligned}$$

Then combining them yields

$$\begin{aligned} 2pp'CV = 2pp'D'Z' + L'X' \;&\Rightarrow\; 2pp'C = 2pp'D'Z'V^{-1} + L'X'V^{-1} \\ &\Rightarrow\; 2pp'CX = 2pp'D'Z'V^{-1}X + L'X'V^{-1}X = 0 \\ &\Rightarrow\; L' = -2pp'D'Z'V^{-1}X(X'V^{-1}X)^{-1} \end{aligned}$$

Substituting L' into the first derivative gives

$$\begin{aligned} &2pp'(CV - D'Z') + 2pp'D'Z'V^{-1}X(X'V^{-1}X)^{-1}X' = 0 \\ &\Rightarrow\; 2pp'CV = 2pp'D'Z'[I - V^{-1}X(X'V^{-1}X)^{-1}X'], \quad \forall p \\ &\Rightarrow\; C = D'Z'V^{-1}[I - X(X'V^{-1}X)^{-1}X'V^{-1}] \\ &\Rightarrow\; Cy = D'Z'V^{-1}(y - X\widetilde{\beta}). \end{aligned}$$

- Compute some variances and covariances. Again, $u \perp\!\!\!\perp \epsilon$.
  1. $\operatorname{Var}(y) = ZDZ' + R = V$ and $\operatorname{Var}(u) = D$.
  2. $\operatorname{Cov}(y, u) = \operatorname{Cov}(Zu + \epsilon, u) = ZD$.
  3. $\operatorname{Cov}(\widetilde{\beta}) = (X'V^{-1}X)^{-1}$.
  4. $\operatorname{Cov}(\widetilde{u}) = DZ'AVAZD = DZ'AZD$ as $AVA = A$ (just calculation).
  5. $\operatorname{Cov}(\widetilde{\beta}, \widetilde{u}) = (X'V^{-1}X)^{-1}X'V^{-1}\operatorname{Cov}(y)A'ZD = (X'V^{-1}X)^{-1}X'A'ZD = O$ as $X'A' = O$.
  6. $\operatorname{Cov}(\widetilde{\beta}, u) = (X'V^{-1}X)^{-1}X'V^{-1}\operatorname{Cov}(y, u) = (X'V^{-1}X)^{-1}X'V^{-1}ZD$.
  7. $\operatorname{Cov}(\widetilde{u}, u) = DZ'A\operatorname{Cov}(y, u) = DZ'AZD = \operatorname{Cov}(\widetilde{u})$.
  8. $\operatorname{Cov}(\widetilde{u} - u) = \operatorname{Cov}(\widetilde{u}) - 2\operatorname{Cov}(\widetilde{u}, u) + \operatorname{Cov}(u) = D - DZ'AZD$.
- $\widetilde{u}$ can be interpreted as the estimate of the mean for u given y.

$$\begin{aligned} \begin{pmatrix} u \\ y \end{pmatrix} \sim N_{q+n}\left(\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} D & DZ' \\ ZD & V \end{pmatrix}\right) \;&\Rightarrow\; u \mid y \sim N_q(DZ'V^{-1}(y - X\beta), D - DZ'V^{-1}ZD)) \\ &\Rightarrow\; E(u \mid y) = DZ'V^{-1}(y - X\beta) \\ &\Rightarrow\; \widehat{E(u \mid y)} = DZ'V^{-1}(y - X\widetilde{\beta}) = \widetilde{u}. \end{aligned}$$

- Consider a 1-way random effect ANOVA $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, \ldots, a$ and $j = 1, \ldots, n$. Let $N = na$. We can write a matrix form

$$y = 1_N\mu + Zu + \epsilon, \quad Z = \underbrace{\begin{pmatrix} 1_n & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1_n \end{pmatrix}}_{N\times a}, \quad u = \begin{pmatrix} \tau_1 \\ \vdots \\ \tau_a \end{pmatrix},$$

where $D = \sigma_\tau^2 I_a$ and $R = \sigma^2 I_N$. Then

$$V = ZDZ' + R = \sigma_\tau^2 ZZ' + \sigma^2 I_N = \begin{pmatrix} \sigma_\tau^2 J_n + \sigma^2 I_n & \cdots & O \\ \vdots & \ddots & O \\ O & \cdots & \sigma_\tau^2 J_n + \sigma^2 I_n \end{pmatrix} : N \times N.$$

We can obtain $\widetilde{\mu} = (1_N' V^{-1} 1_N)^{-1} 1_N' V^{-1} y = \overline{y}$ and

$$\widetilde{u} = DZ'V^{-1}(y - 1_N\widetilde{\mu}) = \frac{2\sigma_\tau^2}{\sigma^2 + 2\sigma_\tau^2} \begin{pmatrix} \overline{y}_{1\cdot} - \overline{y} \\ \vdots \\ \overline{y}_{a\cdot} - \overline{y} \end{pmatrix} : a \times 1.$$

- Use linear mixed effects model to enlarge *longitudinal data*. Consider a linear model for subject i,

$$E(y_i) = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{n_i 1} \end{pmatrix} = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{n_i 1} \end{pmatrix} \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} = Z_i \beta_i.$$

Suppose there are 3 groups (L_i: low dose, H_i: high dose, C_i: control) and assume

$$\beta_i = \underbrace{k_i}_{2\times 4} \underbrace{\beta}_{4\times 1} + \underbrace{b_i}_{2\times 1} = \begin{cases} \beta_0 + b_{10} \\ \beta_0 + \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{20} \end{cases},$$

where β is a fixed effect and $b_i = (b_{10}, b_{20})'$ is a random effect between subjects (independent of i),

$$k_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & L_i & H_i & C_i \end{pmatrix}.$$

Then $y_i = Z_i(k_i\beta + b_i) + \epsilon = Z_i k_i \beta + Z_i b_i + \epsilon$ is a form of $X\beta + Zu + \epsilon$.

29 Bayesian Estimation

- We begin with $f(y)$, the probability density of Y, which we assume to be multivariate normal, say $Y \sim N_n(\beta, \sigma^2 I_n)$, and we now wish to incorporate *prior* knowledge about $\boldsymbol{\theta}$ with density $f(\boldsymbol{\theta})$. Our aim is to make inferences on $f(\theta \mid Y = y)$, the *posterior* density of θ.
- By Bayes theorem, $f(\theta \mid y) = cf(y \mid \theta) f(\theta) \propto f(y \mid \theta) f(\theta)$.
- Usual assumptions of $\theta = (\beta, \sigma)$ is that they have independent prior distributions: $f(\theta) = f_1(\beta) f_2(\sigma)$. Frequently, one uses the *noninformative prior* in which β and $\log\sigma$ are assumed to be locally uniform and $\sigma > 0$. This means that $f_1(\beta) \propto 1$ and $f_2(\sigma) \propto \sigma^{-1}$. *Proof* of the latter: $g(\log\sigma) \propto 1$. Let $v = \sigma$ ($\log\sigma = \log v$), then $f_2(\sigma) = g(v)\partial \log\sigma / \partial v = \sigma^{-1}$. These priors are described as *improper*, as their integrals are technically infinite.
- Using these, we obtain

$$f(\beta, \sigma \mid y) \propto (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right),$$

- Integrating out σ, we have

$$\begin{aligned} f(\beta \mid y) &\propto \int_0^\infty (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right) d\sigma \\ &\propto (\|y - X\beta\|^2)^{-n/2} \\ &= [(n-p)s^2 + (\beta - \widehat{\beta})' X' X (\beta - \widehat{\beta})]^{-n/2} \\ &\propto \left[1 + \frac{(\beta - \widehat{\beta})' X' X (\beta - \widehat{\beta})}{(n-p)s^2}\right]^{-n/2}, \end{aligned}$$

meaning that $\beta \mid y \sim t_p(n-p,\ \widehat{\beta},\ s^2(X'X)^{-1})$, i.e., p-dimensional multivariate t-distribution.

- The least squares estimate of β can be obtained as the mean (or the mode) of the posterior distribution, that is, $\widehat{\beta} = E(\beta \mid Y = y)$.
- (HW5) The conditional posterior density $f(\beta \mid y, \sigma)$ is multivariate normal since

$$f(\beta \mid y, \sigma) \propto f(y \mid \beta, \sigma) f(\beta) = f(y \mid \beta, \sigma) \propto \exp\left[-\frac{(\beta - \widehat{\beta})'(X'X)(\beta - \widehat{\beta})}{2\sigma^2}\right],$$

which means that $\beta \mid y, \sigma \sim N_p(\widehat{\beta}, \sigma^2(X'X)^{-1})$.

- Suppose Z has a density proportional to

$$z^{n/2-1} \exp(-nz/2)$$

and conditional on Z, $X \mid Z = z$ is multivariate normal with mean 0 and covariance $(1/z)I_k$. Determine the density of X, its mean and its covariance matrix.

Solution We notice $Z \sim \text{Gamma}(n/2, n/2)$. Since $\mathbf{X}|Z = z \sim N_k(\mathbf{0}, (1/z)\mathbf{I}_k)$, the conditional density is

$$f(\mathbf{x}|z) = \frac{1}{\sqrt{(2\pi)^k |(1/z)\mathbf{I}_k|}} \exp\left(-\frac{1}{2}\mathbf{x}'[(1/z)\mathbf{I}_k]^{-1}\mathbf{x}\right) \propto z^{k/2} \exp(-z\mathbf{x}'\mathbf{x}/2), \quad \mathbf{x} \in \mathbb{R}^k$$

Hence,

$$f_{\mathbf{X}}(\mathbf{x}) = \int_0^\infty f(\mathbf{x}, z)dz = \int_0^\infty f(\mathbf{x}|z)f(z)dz \propto \int_0^\infty z^{(n+k)/2-1} \exp\left[-\frac{(n + \mathbf{x}'\mathbf{x})z}{2}\right] dz.$$

We see that a gamma density with positive parameters $(n+k)/2$ and $(n+\mathbf{x}'\mathbf{x})/2$ for z, that is,

$$\frac{(\frac{n+\mathbf{x}'\mathbf{x}}{2})^{(n+k)/2}}{\Gamma(\frac{n+k}{2})} \int_0^\infty z^{(n+k)/2-1} \exp(-(n + \mathbf{x}'\mathbf{x})z/2)dz = 1.$$

Using this, we obtain

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &\propto \int_0^\infty z^{(n+k)/2-1} \exp(-(n + \mathbf{x}'\mathbf{x})z/2)dz \\ &\propto \left(\frac{n + \mathbf{x}'\mathbf{x}}{2}\right)^{-(n+k)/2} \propto \left(1 + \frac{\mathbf{x}'\mathbf{x}}{n}\right)^{-(n+k)/2}, \end{aligned}$$

which is, when properly normalized, $\mathbf{X} \sim t_k(n, \mathbf{0}, \mathbf{I}_k)$. Use the iterative expectation and variance to get the mean and variance of $\mathbf{X}$. Again, since $\mathbf{X}|Z = z \sim N_k(\mathbf{0}, (1/z)\mathbf{I}_k)$, we have

$$E(\mathbf{X} \mid Z) = \mathbf{0}, \quad \text{Var}(\mathbf{X} \mid Z) = \mathbf{I}_k/Z.$$

Mean: $E(\mathbf{X}) = E(E(\mathbf{X} \mid Z)) = E(\mathbf{0}) = \mathbf{0}$ for $n > 1$; otherwise undefined.

Covariance matrix: $\text{Var}(\mathbf{X}) = E(\text{Var}(\mathbf{X} \mid Z)) + \text{Var}(E(\mathbf{X} \mid Z)) = E(\mathbf{I}_k/Z) + \text{Var}(\mathbf{0}) = E(1/Z)\mathbf{I}_k$. Here, we know that if $Z \sim \text{Gamma}(n/2, n/2)$, then $X = 1/Z \sim \text{InvGamma}(n/2, n/2)$. Thus,

$$E(1/Z) = \frac{n/2}{n/2 - 1} = \frac{n}{n-2} \quad \Rightarrow \quad \text{Var}(\mathbf{X}) = \frac{n}{n-2}\mathbf{I}_k, \quad n > 2,$$

otherwise undefined.

30 Miscellaneous Exercises

- Let $Y_1, \ldots, Y_n$ be independent random variables such that $E(Y_i \mid X = x_i) = \beta_1 x_i$ and $\text{var}(Y_i \mid X = x_i) = \sigma^2 w_i^{-1}$ for $i = 1, \ldots, n$. If the conditional distribution of Y given x is Poisson, and $w_i^{-1} = x_i$, show that the MLE is the same as WLS.

  Solution: By assumption, we have

  $$f(y \mid x) = e^{-a_x} a_x^y / y!,$$

  where the parameter a_x is a function of x. Since $E(Y \mid X = x) = a_x$, we have $a_{x_i} = \beta_1 x_i$, so that we have $f(y_i \mid x_i) = e^{-\beta_1 x_i}(\beta_1 x_i)^{y_i}/y_i!$. Hence, the likelihood and the log-likelihood are

  $$L(\beta_1) = C + e^{-\beta_1 \sum_i x_i} + \prod_i \beta_1^{y_i} x_i^{y_i}$$
  $$\Rightarrow \ell(\beta_1) = C' - \beta_1 \sum_i x_i + \log \beta_1 \sum_i y_i$$

  Solving the $\partial \ell / \partial \beta_1 = 0$ yields $\widehat{\beta}_{1,\text{MLE}} = \sum_i Y_i / \sum_i x_i$. If $w_i^{-1} = x_i$, the WLS estimate is

  $$\widehat{\beta}_{1,\text{WLS}} = \frac{\sum_i w_i Y_i (x_i - \overline{x})}{\sum_i w_i (x_i - \overline{x})^2} = \frac{\sum_i w_i Y_i x_i}{\sum_i w_i x_i^2} = \frac{\sum_i Y_i}{\sum_i x_i} = \widehat{\beta}_{1,\text{MLE}}.$$

- Suppose that the regression curve $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$ has a local maximum at $x = x_m$, where x_m is near the origin. If Y is observed at n points x_i in $[-a, a]$, $\overline{x} = 0$, and the usual normality assumptions hold, outline a method for finding a CI for x_m.

  Solution: Solving $dY/dx = 0$ gives $x_m = -\beta_1/(2\beta_2)$. Let $U = \widehat{\beta}_1 + 2x_m \widehat{\beta}_2$. Then $E(U) = 0$ since $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are unbiased and $\sigma_U^2 = \text{var}(U) = \text{var}(\widehat{\beta}_1) + 4x_m \,\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2) + 4x_m^2 \,\text{var}(\widehat{\beta}_2)$, where $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X'X)^{-1}$. Thus, we have $T^2 = (U^2/\sigma_U^2)/(S^2/\sigma^2) \sim t_{n-3}^2 = F_{1,n-3}$. The confidence limits are root of $T^2 = F_{1,n-3}^{\alpha}$ w.r.t x_m.

- Suppose that $Y = (Y_1, \ldots, Y_n)'$, where $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$. Find the restricted likelihood for $Q'Y$ and show that the REML estimate of σ^2, which is unbiased, unlike MLE.

  Solution: Note that Q is a n by $n-p = n-1$ matrix whose columns form an orthogonal basis for $C(X)^{\perp}$, the orthogonal complement of $C(X)$. Thus $Q'Q = I_{n-1}$ and $QQ' = I_n - X(X'X)^{-1}X' = I_n - 1_n 1_n'/n$. Since $Y \sim N_n(\mu 1_n, \sigma^2 I_n)$, $Q'Y \sim N_{n-1}(0, \sigma^2 I_{n-1})$ as $1_n \in C(X)$. Hence,

  $$\ell_R(\sigma^2) = C - \frac{n-1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|Q'Y\|^2 \;\Rightarrow\; \widehat{\sigma}_R^2 = \frac{Y'QQ'Y}{n-1} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}.$$