

Biostat 250 A

---

Hw 2

Elvis Cui

---

Han Cui

---

Dept. of Biostat

---

UCLA

---



**Reading material:** (a) Seber and Lee's text, Sections 10.1 and 10.2 of Chapter 10 and (b) Chapter 7 of Reference 2 on the recommended list.

For performing regression diagnostics analysis, check out the IDRE website at UCLA (Regression with Stata Chapter 2 – Regression Diagnostics).

The model of interest is the standard linear model with  $E(y) = X\beta$  where  $X$  is a  $n \times p$  matrix and  $X$  has full column rank. The notation and terms in this homework set are the same used in class; for example,  $h_{ii}$  is the  $i^{\text{th}}$  diagonal element of  $X(X'X)^{-1}X'$ ,  $s_{(i)}$  is the estimate of  $\sigma$  without the  $i^{\text{th}}$  case,  $t_i$  is the externally studentized residual and  $r_i$  is the  $i^{\text{th}}$  internally studentized residual.

- 1 (Mean shift model) Suppose you want to test whether the  $i^{\text{th}}$  case has outlying  $x$ -values using the mean-shift mode given by

$$E y = X\beta + \Theta\phi_i$$

where  $\phi_i$  is the vector with all its components equal to 0 except for the  $i^{\text{th}}$  element, which is equal to 1. Derive a test whether the  $i^{\text{th}}$  case is outlying and show that this test statistic is the  $i^{\text{th}}$  externally studentized statistic and its square has a F-distribution with 1 and  $n-p-1$  degrees of freedom.

- 2 Establish an algebraic relationship between internally and externally studentized residuals.
- 3 Show that the distribution of  $r_i^2/(n-p)$  is Beta( $1/2, (n-p-1)/2$ ) if  $r_i$  is the  $i^{\text{th}}$  internally studentized residual. [Hint: see Exercise 10a of Seber and Lee, page 270]
- 4 Suppose we have a regression model is  $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$  and the standard assumptions hold. An added variable plot is constructed by first regressing  $y$  on  $x_1, \dots, x_k$  and regressing  $x_k$  on  $x_1, \dots, x_{k-1}$ , and then regressing the first set of residuals  $e_1$  on the second set of residuals  $e_2$ . *Y on  $x_1, \dots, x_{k-1}$ ?*
  - (i) What are the fitted regression coefficients when you regress  $e_1$  on  $e_2$  and relate your answers to those fitted coefficients from regressing  $y$  on  $x_1, x_2, \dots, x_k$ .
  - (ii) What is correlation  $e_1$  and  $e_2$  and describes its relationship to the partial correlation between  $y$  and  $x_k$  controlling for  $x_1, x_2, \dots, x_{k-1}$ .
- 5 The Cook's distance of the  $i^{\text{th}}$  case is defined by  $(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)}) / ps^2$ . Express it in terms of  $p$ ,  $h_{ii}$  and  $r_i$ .
- 6 The enclosed small data set comes from a pilot study to assess the use of toenail arsenic concentrations as an indicator of ingestion of arsenic-containing water. Twenty-one participants were interviewed regarding use of their private (unregulated) wells for drinking and cooking, and each provided a sample of water and toenail clippings. Trace concentrations of arsenic were detected in 15 of the 21 well-water samples and in all toenail clipping samples. Using the variable "arsnails" as the outcome variable, perform regression diagnostics on the data using methods or measures discussed in class to identify all potentially problematic cases in the data, and also existence of heteroscedasticity. When appropriate, identify outliers using the mean shift model. Summarize your results clearly so that a public health officer understands your write-up.

Q1: Define  $P_X = X(X^T X)^{-1} X^T$ ,  $Q_X = I - P_X$   
 $\phi_i = \begin{pmatrix} 0 \\ \vdots \\ i \\ \vdots \end{pmatrix}$   $\leftarrow i^{th}$  position,  $S^2 = \frac{1}{n-p} Y^T Q_X Y$   
 $h_{ii} = (P_X)_{ii}$ ,  $g_i = i^{th}$  column of  $Q_X$ .  
 $e_i = Y^T g_i = g_i^T Y$ .

To test  $H_0: \theta = 0$ , define

$$F = \frac{\text{RSS}_{H_0} - \text{RSS}}{\text{RSS}} \frac{n-p-1}{1}$$

where  $\text{RSS}_{H_0} = Y^T Q_X Y$  and

$$\text{RSS} = Y^T [Q_X - \frac{Q_X \phi_i \phi_i^T Q_X}{\phi_i^T Q_X \phi_i}] Y$$

The second equality comes from the fact that

$$P_{C(X|\phi_i)} = Q_X \phi_i (\phi_i^T Q_X \phi_i)^{-1} \phi_i^T Q_X$$

Thus,

$$\begin{aligned} \text{RSS}_{H_0} - \text{RSS} &= Y^T \frac{Q_X \phi_i \phi_i^T Q_X}{\phi_i^T Q_X \phi_i} Y \\ &= \frac{e_i^2}{1-h_{ii}} \end{aligned}$$

$$\begin{aligned} \text{RSS} &= Y^T Q_X Y - \frac{e_i^2}{1-h_{ii}} = (n-p)S^2 - \frac{e_i^2}{1-h_{ii}} \\ &= (n-p-1)S_{(i)}^2 \end{aligned}$$

where  $S_{(i)}$  is the estimate of  $S$  without the  $i^{th}$  case.

$$\Rightarrow F = \frac{e_i^2}{S_{(i)}^2} \frac{1}{1-h_{ii}} = t_i^2$$

Where  $t_i$  is the external studentized residual.

Under  $H_0$ ,  $F \sim F_{1, n-p-1}(0)$ .

Q2:  $r_i = \frac{e_i}{S_{(i)} \sqrt{h_{ii}}}$ ,  $S^2 = \frac{1}{n-p} Y^T Q_X Y$

$$t_i = \frac{e_i}{S_{(i)} \sqrt{h_{ii}}},$$

$$\frac{t_i^2}{S^2} = \frac{S_{(i)}^2}{S^2} \stackrel{(*)}{=} \frac{\frac{n-p}{n-p-1} S^2 - \frac{1}{n-p-1} \frac{e_i^2}{1-h_{ii}}}{S^2}$$

$$= \frac{n-p}{n-p-1} - \frac{1}{n-p-1} \frac{e_i^2}{1-h_{ii}} \frac{1}{S^2}.$$

where  $(*)$  comes from:

$$(n-p-1) S_{(i)}^2 = (n-p) S^2 - \frac{e_i^2}{1-h_{ii}} \quad (*)$$

Proof of  $(*)$ : Let  $\hat{\beta}(i)$  be the LSE without the  $i^{th}$  case.  $\beta$  be the LSE.

then by 250A:

$$\hat{\beta} - \hat{\beta}(i) = \frac{(X^T X)^{-1} x_i e_i}{1-h_{ii}}$$

Then:

$$\begin{aligned} (n-p-1) S_{(i)}^2 &= \sum_{j \neq i} (y_j - x_j^T \hat{\beta}(i))^2 \\ &= \sum_{j \neq i} (y_j - x_j^T [\beta - \frac{x_i^T x_j e_i}{1-h_{ii}}])^2 \\ &= \sum_{j \neq i} (e_j + \frac{x_j^T (X^T X)^{-1} x_i e_i}{1-h_{ii}})^2 \\ &= \sum_{j=1}^n (e_j + \frac{h_{ji} e_i}{1-h_{ii}})^2 - \frac{e_i^2}{(1-h_{ii})^2} \end{aligned}$$

Also note that

$$P_X Q_X Y = 0 \Rightarrow \sum_{j=1}^n h_{ji} e_j = 0$$

$$P_X P_X = P_X \Rightarrow \sum_{i=1}^n h_{ji}^2 = h_{ii}$$

$$P_X Q_X = 0 \Rightarrow \sum_{j=1}^n e_j = 0$$

$$\Rightarrow (n-p) S_{(i)}^2 = (n-p) S^2 - \frac{e_i^2}{(1-h_{ii})^2} \quad \square$$

Q<sub>3</sub>: Using the notations in Q1,  
we have

$$\frac{r_i^2}{n-p} = \frac{Y^T \frac{Q_x \phi_i \phi_i^T Q_x}{\phi_i^T Q_x \phi_i} Y}{Y^T Q_x Y}$$

$$\text{But } Y^T Q_x Y = Y^T \left( Q_x \frac{Q_x \phi_i \phi_i^T Q_x}{\phi_i^T Q_x \phi_i} \right) Y + Y^T \frac{Q_x \phi_i \phi_i^T Q_x}{\phi_i^T Q_x \phi_i} Y$$

$$\text{We have } A_1^T = A_1, A_1^2 = A_1$$

$$A_2^T = A_2, A_2^2 = A_2$$

$$A_1 A_2 = 0.$$

$$\text{rank}(A_1) = \text{Tr}(A_1) = 1$$

$$\begin{aligned} \text{rank}(A_2) &= \text{Tr}(A_2) = \text{Tr}(Q_x) - \text{Tr}(A_1) \\ &= n-p-1 \end{aligned}$$

Thus,  $Y^T A_1 Y \perp\!\!\!\perp Y^T A_2 Y$  &

$$Y^T A_1 Y \sim \chi_1^2 \text{ or } G_a(\frac{1}{2}, 2)$$

$$Y^T A_2 Y \sim \chi_{n-p-1}^2 \text{ or } G_a(\frac{n-p-1}{2}, 2)$$

$$\Rightarrow \frac{r_i^2}{n-p} = \frac{Y^T A_1 Y}{Y^T (A_1 + A_2) Y} \sim \text{Beta}(\frac{1}{2}, \frac{n-p-1}{2})$$

(i) Q<sub>4</sub>: Let  $X_{(k)} = [1 \ x_1 \ \dots \ x_{k-1}]$ .

$$P_{(k)} = X_{(k)} (X_{(k)}^T X_{(k)})^{-1} X_{(k)}^T$$

$$Q_{(k)} = I - P_{(k)}$$

$$\begin{aligned} \text{Then } Y &= \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \\ &= X_{(k)} \beta_{(k)} + \beta_k x_k \end{aligned}$$

$$\text{where } \beta_{(k)}^T = [1, \dots, \beta_k].$$

$$\Rightarrow Y = X_{(k)} \beta_{(k)} + \beta_k Q_{(k)} x_k + \beta_k P_{(k)} x_k$$

$$= X_{(k)} (\beta_{(k)} + (X_{(k)}^T X_{(k)})^{-1} X_{(k)}^T x_k \beta_k)$$

$$+ Q_{(k)} x_k \beta_k$$

$$\Rightarrow \hat{\beta}_k^{(*)} = \frac{x_k^T Q_{(k)} Y}{x_k^T Q_{(k)} x_k} \quad \dots \textcircled{1}$$

(\*) comes from the fact that

$$\langle Q_{(k)} x_k, X_{(k)} m \rangle = 0 \quad \forall m \in \mathbb{R}^k$$

On the other hand,

$$e_1 = Q_{(k)} Y, e_2 = Q_{(k)} x_k$$

So reg coef of  $e_1$  on  $e_2$  is:

$$\frac{\langle e_1, e_2 \rangle}{\|e_2\|_2^2} = \frac{x_k^T Q_{(k)} Y}{x_k^T Q_{(k)} x_k}$$

which is exactly  $\textcircled{1}$ .  $\square$ .

$$(ii): r_{YX_k | X_{(k)}} = \frac{\langle Y^\perp, X_k^\perp \rangle}{\|Y^\perp\|_2 \|X_k^\perp\|_2}$$

$$\text{where } Y^\perp = Q_{(k)} Y$$

$$X_k^\perp = Q_{(k)} X_k$$

$$\Rightarrow r_{YX_k | X_{(k)}}^2 = \frac{(Y^\top Q_{(k)} X_k)^2}{Y^\top Q_{(k)} Y \cdot X_k^\top Q_{(k)} X_k}$$

$$\text{Cor}(e_1, e_2) = \frac{\text{Cov}(e_1, e_2)}{\|e_1\|_2^2 \|e_2\|_2^2},$$

$$\begin{aligned} \text{Cov}(e_1, e_2) &= \text{Cov}(Q_{(k)} Y, Q_{(k)} X_k) \\ &= Y^\top Q_{(k)} X_k. \end{aligned}$$

$$\|e_1\|_2^2 = Y^\top Q_{(k)} Y, \|e_2\|_2^2 = X_k^\top Q_{(k)} X_k$$

$$\Rightarrow \text{Cor}(e_1, e_2) = r_{YX_k | X_{(k)}}^2$$

$$\text{Or } \text{Cor}(e_1, e_2) = r_{YX_k | X_{(k)}}$$

$$\begin{aligned} Q5: D_i &= \frac{(\hat{Y} - \hat{Y}_{(i)})^\top (\hat{Y} - \hat{Y}_{(i)})}{PS^2} \\ &= \frac{(\hat{\beta}_{(i)} - \hat{\beta})^\top X_i^\top X (\hat{\beta}_{(i)} - \hat{\beta})}{PS^2} \end{aligned}$$

$$\begin{aligned} \textcircled{1} &= \frac{e_i^2 \cdot X_i^\top (X^\top X)^{-1} X_i}{S^2 (1-h_{ii})^2 \cdot P} \\ &= \frac{r_i^2}{P} \frac{h_{ii}}{1-h_{ii}} \end{aligned}$$

① is due to

$$\hat{\beta}_{(i)} - \hat{\beta} = (X^\top X)^{-1} X_i e_i$$

② comes from

$$r_i = \frac{e_i}{S \sqrt{1-h_{ii}}}$$

□

Q6: please see attached file.