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 nxp 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^{th} diagonal element of $X(X'X)^{-1}X'$, $s_{(i)}$ is the estimate of σ without the i^{th} case, t_i is the externally studentized residual and r_i is the i^{th} internally studentized residual.

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

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

where ϕ_i is the vector with all its components equal to 0 except for the ith element, which is equal to 1. Derive a test whether the i^{th} case is outlying and show that this test statistic is the i^{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^{th} internally studentized residual. [Hint: see Exercise 10a of Seber and Lee, page 270]

4 Suppose we have a regression model is $E(y)= \beta_0 +\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ and the standard assumptions hold. An added variable plot is constructed by first regressing y on $x_1, .., x_k$ and regressing x_k on $x_1..,x_{k-1}$, and then regressing the first set of residuals e_1 on the second set of residuals e_2. y on $x_1,\ldots,x_{k-1}$?

(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 $x_1, x_2, ..,x_k$.

(ii) What is correlation e_1 and e_2 and describes its relationship to the partial correlation between y and x_k controlling for $x_1, x_2, ..,x_{k-1}$.

5 The Cook's distance of the i^{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.

56
60

Q1: Define $P_X = X(X^TX)^{-1}X^T$, $Q_X = I - P_X$

$\phi_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i^{th}$ position, $S^2 = \frac{1}{n-p} Y^T Q_X Y$

$h_{ii} = (P_X)_{ii}$, q_i: i^{th} column of Q_X.

$e_i = Y^T q_i = q_i^T Y$.

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

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

where $RSS_{H_0} = Y^T Q_X Y$ and

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

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,

$$RSS_{H_0} - RSS = Y^T \frac{Q_X \phi_i \phi_i^T Q_X}{\phi_i^T Q_X \phi_i} Y = \frac{e_i^2}{1-h_{ii}}$$

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

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

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

where t_i is the external studentized residual.

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

Clearly explain why $F \sim F(.,.)$, it is the test statistic for a Gen'l F-test in a linear model.

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

$t_i = \frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}}$,

$$\frac{r_i^2}{t_i^2} = \frac{S_{(i)}^2}{S^2} \overset{(\Delta)}{=} \frac{\frac{n-p}{n-p-1} S^2 - \frac{1}{n-p-1} \frac{e_i^2}{1-h_{ii}}}{S^2} = \frac{n-p}{n-p-1} - \frac{1}{n-p-1} \frac{e_i^2}{1-h_{ii}} \frac{1}{S^2}.$$

where (Δ) 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. $\hat{\beta}$ be the LSE.

then by 250A:

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

Then:

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

Also note that

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

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

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

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

? want t_i as a function of r_i

$t_i = \frac{r_i \sqrt{n-p-1}}{\sqrt{n-p-r_i^2}}$

Q3: Using the notations in Q1, we have

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

But $Y^T Q_X Y = Y^T \underbrace{\left(Q_X - \frac{Q_X \phi_i \phi_i^T Q_X}{\phi_i^T Q_X \phi_i}\right)}_{A_2} Y + Y^T \underbrace{\frac{Q_X \phi_i \phi_i^T Q_X}{\phi_i^T Q_X \phi_i}}_{A_1} Y$ — 1

We have $A_1^T = A_1$, $A_1^2 = A_1$
$A_2^T = A_2$, $A_2^2 = A_2$
$A_1 A_2 = 0$.

} Have to show, Sorry.

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

$\text{rank}(A_2) = \text{Tr}(A_2) = \text{Tr}(Q_X) - \text{Tr}(A_1)$
$= n-p-1$

Thus, $Y^T A_1 Y \perp\!\!\!\perp Y^T A_2 Y$ &
$Y^T A_1 Y \sim \chi_1^2$ or $Ga(\frac{1}{2}, 2)$
$Y^T A_2 Y \sim \chi_{n-p-1}^2$ or $Ga(\frac{n-p-1}{2}, 2)$

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

c)

Q4: Let $X_{(k)} = [1\ X_1 \cdots X_{k-1}]$.

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

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

Then $\underset{n\times 1}{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

$= X_{(k)} \beta_{(j)} + \beta_k X_k$

where $\beta_{(j)}^T = [1, \cdots, \beta_k]$.

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

$= X_{(k)} \left(\beta_{(j)} + (X_{(k)}^T X_{(k)})^{-1} X_{(k)}^T X_k \beta_k\right)$
$+ Q_{(j)} X_k \beta_k$

$\Rightarrow \hat{\beta}_k \overset{(*)}{=} \frac{X_k^T Q_{(j)} y}{X_k^T Q_{(j)} X_k}$ ... ①

(*) comes from the fact that

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

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:

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

which is exactly ①. □

(ii): $r_{yx_k|x_{(k)}} = \dfrac{\langle y^{\perp}, x_k^{\perp} \rangle}{\|y^{\perp}\|_2 \|x_k^{\perp}\|_2}$

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

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

$\Rightarrow r^2_{yx_k|x_{(k)}} = \dfrac{(y^T Q_{(k)} x_k)^2}{y^T Q_{(k)} y \cdot x_k^T Q_{(k)} x_k}$

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

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

$= y^T Q_{(k)} x_k .$

$\|e_1\|_2^2 = y^T Q_{(k)} y, \quad \|e_2\|_2^2 = x_k^T Q_{(k)} x_k$

$\Rightarrow Cor^2(e_1, e_2) = r^2_{yx_k|x_{(k)}}$ ✓

or $Cor(e_1, e_2) = r_{yx_k|x_{(k)}}$.

Q5: $D_i = \dfrac{(\hat{Y} - \hat{Y}_{(i)})^T (\hat{Y} - \hat{Y}_{(i)})}{p S^2}$

$= \dfrac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p S^2}$

① $= \dfrac{e_i^2 \cdot x_i^T (X^T X)^{-1} x_i}{S^2 (1 - h_{ii})^2 \cdot p}$

② $= \dfrac{r_i^2}{p} \dfrac{h_{ii}}{1 - h_{ii}}$ ✓

① is due to

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

② comes from

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

□

Q6: please see attached file.

Next time
Combine pdfs
when submitting
Can embed pdf
pages of LaTeX
into Rmd.

Biostat 250B HW2 Q6

Elvis Cui

January 28, 2021

1 Linear Model

The model is

$$\text{arsnails} = \beta_0 + \beta_1 * \text{age} + \beta_2 * \text{drinkuse} + \beta_3 * \text{cookuse} + \beta_4 * \text{arswater}$$

The fitted values and t-statistics are

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.452972   0.418132   1.083    0.296
age         -0.001290   0.003444  -0.374    0.713
sexMale     -0.145038   0.107448  -1.350    0.197
drinkuse    -0.011719   0.047010  -0.249    0.807
cookuse     -0.027471   0.082861  -0.332    0.745
arswater    13.195586   1.639792   8.047 8.01e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2302 on 15 degrees of freedom
Multiple R-squared:  0.8323,    Adjusted R-squared:  0.7764
F-statistic: 14.89 on 5 and 15 DF,  p-value: 2.339e-05
```

Figure 1: Fitted values

The figure on studentized residuals vs. leverage is

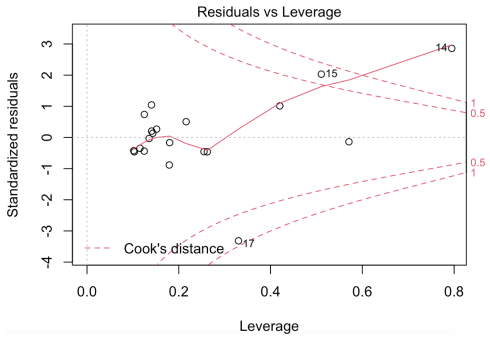

Figure 2: Studentized residuals

The Cook's distances are

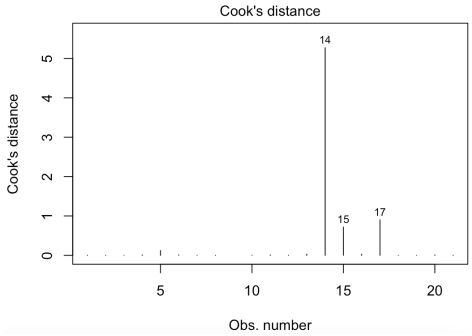

Figure 3: Cook's distance

The diagnostics on heteroscedasticity is

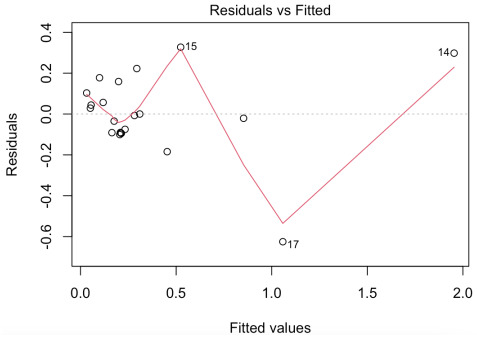

Figure 4: Fitted values

The mean-shift model outputs 2 outliers:

```
> outlierTest(mod)
    rstudent unadjusted p-value Bonferroni p
17 -6.209763         2.2762e-05   0.00045525
14  4.092296         1.0985e-03   0.02197100
```

Figure 5: Mean-shift model

The codes are

```r
setwd("~/Desktop/UCLA_Study/Bio250B/HW2")
rm(list=ls())

library(haven)
library(broom)
library(tidyverse)
theme_set(theme_classic())
data <- read_dta("arsenic.dta")

mod <- lm(arsnails~., data=data)
summary(mod)
plot(mod)

model.diag.metrics <- augment(mod)

# Cook's distance
plot(mod, 4)
# Residuals vs Leverage
plot(mod, 5)
model.diag.metrics %>%
  top_n(3, wt = .cooksd)

library(car)
outlierTest(mod)

```

Figure 6: R code