Biostat 250B HW1

Elvis Cui
Han Cui
Dept. of Biostat
UCLA

Biostat 250B Homework 1 due 1/18/2021

Reading material: Seber and Lee's Text, Chapter 4, section 4.4 and two papers posted on the class website entitled Reading Material 1 and Reading Material 2.

1 Recall that when we want to test whether the p-1 covariates are useful in a linear model with intercept, we showed in class that under the null hypothesis that all coefficients of the p-1 covariates are equal to 0, we have $ER^2=(p-1)/(n-1)$, where n is the sample size and R^2 is the square of the multiple correlation coefficient. This result was arrived at using the fact that if F is a F-variate with numerator degree of freedom a and denominator degree of freedom b, then (i) $\{aF/b\}/(1+aF/b)$ is a Beta distributed with parameters a/2, b/2 and (ii) the expectation of this distribution is a/(a+b). Verify the latter two assertations.

Questions 2 - 3 concern pitfalls of the multiple correlation coefficient and Questions 4 - 5 review basic concepts of standardized regression coefficients and its properties along with its relationship with partial correlation with a covariate when there are two covariates in the model. Question 6 relates the partial correlation coefficient of the response with a covariate in a multiple regression model to the test statistic for testing if the coefficient of the covariate in the model is 0.

2 Review the paper entitled Reading Material 1, use the two data sets in the paper and see if you can confirm the reported results in Tables 1 and 2.

3 Write a ¾ page of your commentary on the paper entitled Reading Material 2 for your supervisor, who is not well trained in statistics, has an advanced degree in Public Health and wanted to know the main gist of the paper.

4 Refer to the Lecture Notes by Ernesto Amaral posted on the website. Use the data set on p.6 and verify all numerical results and the formula shown on p. 30. The last formula is for a linear regression on two explanatory variables with an intercept and expresses the square of the multiple correlation coefficient in terms of an ordinary correlation and a partial correlation coefficient.

5 Do Problem 2 in Ex. 4c on p.113 of Seber and Lee's text.

6 Consider the multiple regression of Y on x_1, …, , x_{k-1}, x_k. Express the partial correlation coefficient of Y and x with the linear effects of x_1,…, x_{k-1} removed in terms of the test statistic for testing whether the coefficient of x_k in the multiple regression model is 0.

Suggested further reading: Reading Material 3 concerns use of square of the multiple correlation coefficient in nonlinear models and potential pitfalls. Review this paper when you find time to do so.

Q1: Let $F \sim F_{a,b}$. Show

(a) $(aF/b)/(1+aF/b) \sim \text{Beta}(\frac{a}{2}, \frac{b}{2})$

(b) $\mathbb{E}\,\text{Beta}(\frac{a}{2}, \frac{b}{2}) = a/(a+b)$

Pf: (a): By definition F, we write

$$F = \frac{\chi_a^2}{\chi_b^2}\frac{b}{a}, \quad \chi_a^2 \perp\!\!\!\perp \chi_b^2$$

where χ_a^2, χ_b^2 are χ^2-variables with df $= a$ & b.

Since $\mathbb{P}(\{F \le 0\}) = 0$, we have

$$\frac{aF/b}{1+aF/b} = \frac{F}{F+\frac{b}{a}} = \frac{1}{1+\frac{b}{a}\frac{1}{F}}$$

$$= \frac{1}{1+\frac{\chi_b^2}{\chi_a^2}} = \frac{\chi_a^2}{\chi_a^2+\chi_b^2} \quad (\Delta)$$

Further, note $\chi_a^2 \sim \text{Gamma}(\frac{a}{2}, 2)$

So (Δ) is the definition of a Beta variable:

$$(\Delta) \sim \text{Beta}(\frac{a}{2}, \frac{b}{2})$$

(b): Let $B \sim \text{Beta}(\frac{a}{2}, \frac{b}{2})$, then

$$\mathbb{E}B = \int_{\mathbb{R}^+} b\, f_B(b)\, db$$

$$\overset{(*)}{=} \int_{\mathbb{R}^+} b\, b^{\frac{a}{2}-1}(1-b)^{\frac{b}{2}-1} B(\tfrac{a}{2},\tfrac{b}{2})^{-1}\, db$$

$$= \int_{\mathbb{R}^+} b^{\frac{a+2}{2}-1}(1-b)^{\frac{b}{2}-1} B(\tfrac{a}{2},\tfrac{b}{2})^{-1}\, db$$

$$= B(\tfrac{a+2}{2}, \tfrac{b}{2}) \,/\, B(\tfrac{a}{2}, \tfrac{b}{2})$$

where $B(\cdot,\cdot)$ is the Beta function, i.e.

$$B(\tfrac{a}{2}, \tfrac{b}{2}) = \frac{\Gamma(\frac{a}{2})\Gamma(\frac{b}{2})}{\Gamma(\frac{a+b}{2})}$$

and $(*)$ follows from the pdf of a Beta RV.

Thus,

$$\mathbb{E}B = \frac{\Gamma(\frac{a}{2}+1)/\Gamma(\frac{a+b}{2}+1)}{\Gamma(\frac{a}{2})/\Gamma(\frac{a+b}{2})}$$

$$\overset{(\circ)}{=} \frac{a}{a+b}$$

where $(\circ)$ follows from the recursion formula

$$\Gamma(x+1) = x\Gamma(x).$$

Q2 : See Codes & Results below.

```r
R_square <- function(y, y_hat) {
  R1 <- 1 - sum((y - y_hat) ** 2) / sum((y - mean(y)) ** 2)
  R2 <- sum((y_hat - mean(y)) ** 2) / sum((y - mean(y)) ** 2)
  R3 <- sum((y_hat - mean(y_hat)) ** 2) / sum((y - mean(y)) ** 2)
  e <- y - y_hat
  R4 <- 1 - sum((e - mean(e)) ** 2) / sum((y - mean(y)) ** 2)
  R6 <- cor(y, y_hat) ** 2
  R7 <- 1 - sum(e ** 2) / sum(y ** 2)
  R8 <- sum(y_hat ** 2) / sum(y ** 2)

  output <- tibble("R1" = R1,
                   "R2" = R2,
                   "R3" = R3,
                   "R4" = R4,
                   "R6" = R6,
                   "R7" = R7,
                   "R8" = R8)
  output
}

mod1 <- lm(y~x, data = data1)
mod2 <- lm(y~0 + x, data = data1)
mod3 <- lm(log(y)~log(x), data = data1)

R_square(data1$y, mod1$fitted.values)
R_square(data1$y, mod2$fitted.values)
R_square(data1$y, exp(mod3$fitted.values))
```

```
> R_square(data1$y, mod1$fitted.values)
# A tibble: 1 x 7
     R1    R2    R3    R4    R6    R7    R8
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.981 0.981 0.981 0.981 0.981 0.997 0.997
> R_square(data1$y, mod2$fitted.values)
# A tibble: 1 x 7
     R1    R2    R3    R4    R6    R7    R8
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.978  1.08  1.08 0.978 0.981 0.996 0.996
> R_square(data1$y, exp(mod3$fitted.values))
# A tibble: 1 x 7
     R1    R2    R3    R4    R6    R7    R8
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.978  1.10  1.10 0.978 0.981 0.996  1.02
```

Q4 : See results below.

```r
# Q4 # formula on page 30

y <- c(1,2,3,5,3,1,5,0,6,3,7,4)
x <- c(1,1,1,1,2,2,3,3,4,4,5,5)
z <- c(12,14,16,16,18,16,12,12,10,12,10,16)

mod_q4 <- lm(y~x + z)
summary(mod_q4) # 0.2539
ry1 <- cor(y, x) ** 2
library(ppcor)
ry2.1 <- pcor(t(rbind(y,x,z)))$estimate[1,3]

ry1 + ry2.1 ** 2 * (1 - ry1) # 0.2538
```

So 0.2539 ≈ 0.2538

↑ definition. ↑ formula.

Q3: Commentary

1 Commentary

Recall the classical linear regression model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \mathbf{W})$$

As practitioners, we would like to estimate β based on observed data and address the goodness-of-fit based on some measurements such as **coefficient of determinants**.

There are 2 common ways to estimate β:

- Ordinary Least Squares (OLS): $\widehat{\beta}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
- Weighted Least Squares (WLS): $\widehat{\beta}_{WLS} = (\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{y}$.

For the second one, usually we write it in the **transformed** version, i.e.,

$$\widehat{\beta}_{WLS} = (\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_*^T\mathbf{y}_*$$

where

$$\mathbf{X}_* = \mathbf{W}^{-1/2}\mathbf{X},\ \mathbf{y}_* = \mathbf{W}^{-1/2}\mathbf{y}$$

Then, the coefficient of determination can be calcualted in three ways:

- $$R^2_{OLS} = 1 - \frac{\|\mathbf{y} - \mathbf{X}\widehat{\beta}_{OLS}\|_2^2}{\mathbf{y^T y} - n\bar{y}^2}$$
- $$R^2_{WLS} = 1 - \frac{\|\mathbf{y}_* - \mathbf{X}_*\widehat{\beta}_{WLS}\|_2^2}{\mathbf{y}_*^\mathbf{T}\mathbf{y}_* - n\bar{y}_*^2}$$
- $$\text{pseudo } R^2_{WLS} = 1 - \frac{\|\mathbf{y} - \mathbf{X}\widehat{\beta}_{WLS}\|_2^2}{\mathbf{y^T y} - n\bar{y}^2}$$

The first two are more intuitive than the last one. While usually the second one (R^2_{WLS}) is higher than the first one (R^2_{OLS}), the third one (pseudo R^2_{WLS}) is always less than the first one. Therefore, "sole reliance on the **coefficient of determination** may fail to reveal important data characteristics and model inadequacies".

Q5: General full rank LRM:

R^2 & F for testing $H: \beta_j = 0, j \neq 0$ are indept of the units in which Y_i & x_{ij} are measured.

Sol. $F = \frac{(RSS)_0 - RSS}{RSS} \frac{n-k}{k}$, $= |\{j : \beta_j = 0, j \neq 0\}|$

$R^2 = 1 - \frac{RSS}{TSS}$, $TSS = (RSS)_0$.

$(RSS)_0 = y^T(I - \frac{11^T}{n})y$, $RSS = y^T(I-P)y$

where $y = (Y_1, \cdots, Y_n)$ & $P = X(X^TX)^{-1}X^T$

① if $y \leftarrow ay$, $a \neq 0$.

then $(RSS)_0 - RSS = a^2 y^T(P - \frac{11^T}{n})y$.

$RSS = a^2 y^T(I-P)y$.

So $[(RSS)_0 - RSS]/RSS$ and RSS_0/RSS stay the same since a^2 cancels out.

② if $X \leftarrow aX$, $a \neq 0$.

then $P = aX(a^2X^TX)^{-1}aX^T$

$= X(X^TX)^{-1}X^T$

is unchanged.

So $((RSS)_0 - RSS)/RSS$ is the same

To sum up, since R^2 & F are functions of RSS_0/RSS and $(RSS_0 - RSS)/RSS$, so both of them stay the same.

Q6: Let $y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$

Show $r_{yx_k \cdot x_1 \cdots x_{k-1}}$ is a function of the test statistic for $H_0: \beta_k = 0$.

Sol. Define the following:

$X = [x_1 \cdots x_{k-1}]$

$\tilde{X} = [x_1 \cdots x_k]$

$P_{x_k^\perp} = \frac{x_k^\perp (x_k^\perp)^T}{\|x_k^\perp\|_2^2}$

$P_X = X(X^TX)^{-1}X$; $P_{\tilde{X}} = \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T$

$y^\perp = (I - P_X)y$; $\hat{y} = P_X y$

$x_k^\perp = (I - P_X)x_k$; $\hat{x}_k = P_X x_k$

- $a = y^T P_{x_k^\perp} y$; $b = y^T(I - P_X)y$

Then we have:

① $r^2_{yx_k \cdot x_1 \cdots x_{k-1}} = \frac{\langle y^\perp, x_k^\perp \rangle^2}{\|y^\perp\|_2^2 \, \|x_k^\perp\|_2^2}$

$= \frac{y^\perp P_{x_k^\perp} y^\perp}{y^T(I-P_X)y}$

$= \frac{y^T P_{x_k^\perp} y - \hat{y} P_{x_k^\perp} \hat{y}}{y^T(I-P_X)y}$

$\overset{(*)}{=} a/b$ (1)

(*) is because $\hat{y} \in V(X)$ & $P_{x_k^\perp}\hat{y} \in V^\perp(X)$

② Test stat for $\beta_k = 0$ is

$F = (n-k)\, y^T(P_{\tilde{X}} - P_X)y \,/\, y^T(I - P_{\tilde{X}})y$

$\overset{(\Delta)}{=} (n-k)\, y^T P_{x_k^\perp} y / y^T(I - P_{\tilde{X}})y$

$= (n-k)\, a / [y^T(I-P_X)y - y^T(P_{\tilde{X}} - P_X)y]$

$\overset{(\Delta)}{=} (n-k)\, a/(b-a)$ (2)

(Δ) are due to $(P_{\tilde{X}} - P_X)y \in V(\tilde{X})$ & $\notin V(X)$

Combine (1) & (2):

$r^2 = F/(n-k+F)$.