hw5

AUTHOR
Hanbei Xiong

Part A: Problems from Lab 5.

1. For the regression of logdocper on logpopdens bedp1000 hsgrad poverty unemp pcinck, provide an interpretation of each of the regression coefficients.

Answer:

Here is the SAS output for the regression coefficients:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-2.98003	0.31428	-9.48	<.0001	0
logpopdens	1	0.04832	0.01573	3.07	0.0023	1.56293
bedp1000	1	0.15378	0.00854	18.01	<.0001	1.36750
hsgrad	1	0.01711	0.00332	5.16	<.0001	2.53581
poverty	1	0.04025	0.00518	7.76	<.0001	2.73125
unemp	1	-0.02622	0.00815	-3.22	0.0014	1.70162
pcinck	1	0.06515	0.00556	11.72	<.0001	2.38561

The formula with the regression coefficient is:

$$\log(\text{docper}) = -1.57389 + 0.04309\log(\text{popdens}) + 0.16328\text{bedp1000} - 0.0004001\text{hsgrad} + 0.27268\text{bagrad} + 0.01841\text{poverty} - 0.00666\text{unemp} + 0.03425\text{pcinck}$$

We exponentiate both sides, we have:

$$\text{docper} = e^{-1.57389} \times e^{0.04309\log(\text{popdens})} \times e^{0.16328\text{bedp1000}} \times e^{-0.0004001\text{hsgrad}} \times e^{0.27268\text{bagrad}} \times e^{0.01841\text{poverty}} \times e^{-0.00666\text{unemp}} \times e^{0.03425\text{pcinck}}$$

We simplify the formula to:

$$\text{docper} = e^{-1.57389} \times \text{popdens}^{0.04309} \times e^{0.16328\text{bedp1000}} \times e^{-0.0004001\text{hsgrad}} \times e^{0.27268\text{bagrad}} \times e^{0.01841\text{poverty}} \times e^{-0.00666\text{unemp}} \times e^{0.03425\text{pcinck}}$$

logpopdens: For a 1% increase in population density, the expected number of active physician per 1000 people is multiplied by 1.00482 while holding all other predictors constant.

bedp1000: For a 1 unit increase in the number of hospital beds per 1000 people, the expected number of active physician per 1000 people is multiplied by 1.166 while holding all other predictors constant.

hsgrad: For a 1% increase in the percentage of college graduates, the expected number of active physician per 1000 people is multiplied by 1.017 while holding all other predictors constant.

poverty: For a 1% increase in the percentage of people living in poverty, the expected number of active physician per 1000 people is multiplied by 1.041 while holding all other predictors constant.

unemp: For a 1% increase in the percentage of unemployment, the expected number of active physician per 1000 people is multiplied by 0.974 while holding all other predictors constant.

pcinck: For a 1% increase in the per capita income, the expected number of active physician per 1000 people is multiplied by 1.067 while holding all other predictors constant.

2. For the predictor pcinck, verify that the SE of the regression coef is equal to the formula involving R^2_j given in lecture, ie, find each of the quantities in the SE and calculate the SE, verifying it is equal to the SE given in the SAS output for the regression.

Answer:

By regressing pcinck on other predictors, we get the R_j^2

Root MSE	2.64319	**R-Square**	0.5808
Dependent Mean	18.56148	**Adj R-Sq**	0.5760
Coeff Var	14.24016		

The formula is given by:

$$SE = \sqrt{\hat{Var}(\hat{\beta}_j)} = \sqrt{\frac{1}{1-R_j^2}\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_{ij}-\bar{x})^2}} = \sqrt{\frac{1}{1-R_j^2}\frac{\hat{\sigma}^2}{(n-1)s_j^2}}$$

Here are the quantities we have from SAS output:

$R_j^2 = 0.5808$

$\hat{\sigma}^2 = MSE(F) = 0.09370$

$s_j^2 = 4.05919^2 = 16.477$

We can calculate the SE using the formula:

$$SE = \sqrt{\hat{Var}(\hat{\beta}_j)} = \sqrt{\frac{1}{1-R_j^2}\frac{\hat{\sigma}^2}{(n-1)s_j^2}} = \sqrt{\frac{0.09370}{(1-0.5808)\times(440-1)\times 16.477}} \approx 0.00556$$

This matches with the SE of the regression coef given in the SAS output for the regression.

3. 3. In general, would you expect VIFs to increase or decrease as more predictors are added to a model? Provide a justification for your answer. Illustrate by providing some example regressions using the CDI dataset.

Answer:

VIFs would increase as more predictors are added to a model. The VIF is calculated as:

$$VIF_j = \frac{1}{1-R_j^2}$$

where R_j^2 is the R-squared value of the regression of the jth predictor regresses on all the other predictors. As more predictors are added to the model, the R_j^2 value would increase, since more variability of jth predictor would be explained by the additional predictors (It will never decrease). Then, the denominator of the VIF would decrease, which would increase the VIFs for each predictor.

Justification using CDI dataset:

We will start with model which is regressing 'docper100' on 'bedp1000' and 'hsgrad'. Here is the SAS output including VIFs:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-4.94607	0.59853	-8.26	<.0001	0
bedp1000	1	0.55873	0.02574	21.71	<.0001	1.04667
hsgrad	1	0.06485	0.00734	8.83	<.0001	1.04667

We add an addition predictor called 'bagrad' into the model. Here is the SAS output including VIFs:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.36034	0.60070	-0.60	0.5489	0
bedp1000	1	0.51355	0.02176	23.61	<.0001	1.07098
hsgrad	1	-0.02162	0.00878	-2.46	0.0141	2.14167
bagrad	1	0.10845	0.00787	13.78	<.0001	2.05041

We can see the VIF of variable 'bedp1000' increases from 1.04667 to 1.07098, and VIF of variable 'hsgrad' increases from 1.04667 to 2.14167.

We add another predictor 'poverty' into the second model. Here is the SAS output including VIFs:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.63193	0.82505	0.77	0.4441	0
bedp1000	1	0.52561	0.02277	23.08	<.0001	1.17887
hsgrad	1	-0.03282	0.01084	-3.03	0.0026	3.28575
bagrad	1	0.11009	0.00791	13.92	<.0001	2.07962
poverty	1	-0.02321	0.01326	-1.75	0.0809	2.16563

We can see the VIF of variable 'bedp1000' increases from 1.07098 to 1.17887, the VIF of variable 'hsgrad' increases from 2.14167 to 3.28575, and the VIF of variable 'bagrad' increases from 2.05041 to 2.07962.

Hence, the result above validates my answer.

Part B: Least squares using matrix algebra

1. Show that that matrix **I-H** is symmetric and idempotent. **I** is the identity matrix and $H = X(X^TX)^{-1}X^T$

Answer:

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

We can see that
$(I-H)^T = I^T - (X(X^TX)^{-1}X^T)^T = I - X((X^TX)^T)^{-1}X^T = I - X(X^TX)^{-1}X^T = I - H$

Hence, $I - H$ is symmetric.

We can show H is idempotent:

$$\begin{aligned} HH &= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) \\ &= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T = H \end{aligned}$$

Then, we can show $I - H$ is idempotent:

$$(I-H)(I-H) = I^2 - IH - HI + H^2 = I - H - H + HH = I - H - H + H = I - H$$

We finished the proof.

2. Show that the matrix $I - \frac{1}{n}J$ is symmetric and idempotent. **J=11′**, where **1** is a vector of 1′s

Answer:

Given $I - \frac{1}{n}J = I - \frac{1}{n}11'$

We can see that $(I - \frac{1}{n}J)^T = I^T - (\frac{1}{n}11')^T = I - \frac{1}{n}11' = I - \frac{1}{n}J$

Hence, $I - \frac{1}{n}J$ is symmetric.

$JJ = (11')(11') = 1(1'1)'1 = 11' = n11' = nJ$ where n is number of element in vector **1**

Then, we can show $I - \frac{1}{n}J$ is idempotent:

$$\begin{aligned} (I - \frac{1}{n}J)(I - \frac{1}{n}J) &= I^2 - I(\frac{1}{n}J) - (\frac{1}{n}J)I + (\frac{1}{n}J)^2 \\ &= I - \frac{1}{n}J - \frac{1}{n}J + \frac{1}{n^2}JJ \\ &= I - \frac{1}{n}J - \frac{1}{n}J + \frac{1}{n^2}nJ \\ &= I - \frac{1}{n}J - \frac{1}{n}J + \frac{1}{n}J \\ &= I - \frac{1}{n}J \end{aligned}$$

We finished the proof.

3. Consider the model $Y = X\beta + \epsilon$, $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2 I$, Show by using matrix operations:

   a. $E(Y) = X\beta$ and $Var(Y) = \sigma^2 I$
   b. For $\hat{\beta} = (X'X)^{-1}X'Y$, $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$
   c. For $\hat{Y} = HY$ where $H = X(X'X)^{-1}X'$, $E(\hat{Y}) = X\beta$ and $Var(\hat{Y}) = \sigma^2 H$

(a) Answer:

$E(Y) = E(X\beta + \epsilon) = X\beta + E(\epsilon) = X\beta$

Since we have proved $Cov(X\beta, \epsilon) = 0$ in HW2,

$Var(Y) = Var(X\beta + \epsilon) = Var(X\beta) + Var(\epsilon) + 2Cov(X\beta, \epsilon) = 0 + \sigma^2 I + 0 = \sigma^2 I$

(b) Answer:

$$\begin{aligned} E(\hat{\beta}) &= E((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'X\beta \\ &= (X'X)^{-1}(X'X)\beta \\ &= \beta \end{aligned}$$

$$\begin{aligned} Var(\hat{\beta}) &= Var((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'Var(Y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

(c) Answer:

$$E(\hat{Y}) = E(HY) = E(X(X'X)^{-1}X'Y) = X(X'X)^{-1}X'E(Y) = X(X'X)^{-1}X'X\beta = X\beta$$

In the following, we will use the property that H is symmetric and in idempotent.

$$Var(\hat{Y}) = Var(HY) = HVar(Y)H^T = H\sigma^2 IH^T = \sigma^2 HH^T = \sigma^2 H$$

4. A data set contains observations on 2n patients, n females and n males. A dummy variable is created such that $x_i = 0$ for females and $x_i = 1$ for males. The outcome variables are u_i for women and v_i for men. Let $\bar{u}$ be the sample mean for females and $\bar{v}$ be the sample mean for males, and let s_u^2 and s_v^2 be the sample variances for each group. Consider the model $y_i = \beta_0 + \beta_1 x_i + \epsilon$, where y_i equals u_i for women and v_i for men. The data set is sorted such that females are in the first n rows and males are in the next n rows

    a. Write the design matrix X (Show the element of the matrix)
    b. Obtain the $X'X$ matrix, its inverse $(X'X)^{-1}$, and $X'Y$ matrix
    c. By computing $\hat{\beta} = (X'X)^{-1}X'Y$, show that $\begin{bmatrix} \bar{u} \\ \bar{v} - \bar{u} \end{bmatrix}$. How do you interpret $\hat{\beta}_0$ and $\hat{\beta}_1$?
    d. Obtain the hat matrix $H = X(X'X)^{-1}X'$. Express the elements of H using n.
    e. Obtain the vector of fitted values $\hat{Y}$ by computing $\hat{Y} = HY$.
    f. Obtain the SSE by computing $(Y - \hat{Y})'(Y - \hat{Y})$ and the MSE. Express the MSE in terms of n, s_u^2, and s_v^2.
    g. Suppose that you are conducting a two-sample t test to test $H_0 : \mu_{female} = \mu_{male}$, assuming the two groups have equal variance. Obtain an expression for the estimated common variance σ^2 for this t test, expressing it in terms of n, s_u^2, and s_v^2.
    h. How do your estimates in (f) and (g) compare?

(a) Answer:

$X \in \mathbb{R}^{2n \times 2}$

Starting at the nth row, each row changes from (1,0) to (1,1)

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

(b) Answer:

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \cdots & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2n & n \\ n & n \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{2n^2 - n^2} \begin{bmatrix} n & -n \\ -n & 2n \end{bmatrix} = \frac{1}{n^2} \begin{bmatrix} n & -n \\ -n & 2n \end{bmatrix} = \begin{bmatrix} \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{2}{n} \end{bmatrix}$$

$$Y = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \cdots & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 \cdots & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} u_i + v_i \\ \sum_{i=1}^{n} v_i \end{bmatrix}$$

(c) Answer:

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{2}{n} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} u_i + v_i \\ \sum_{i=1}^{n} v_i \end{bmatrix} = \begin{bmatrix} \frac{1}{n}(\sum_{i=1}^{n} u_i + v_i) - \frac{1}{n}(\sum_{i=1}^{n} v_i) \\ -\frac{1}{n}(\sum_{i=1}^{n} u_i + v_i) + \frac{2}{n}(\sum_{i=1}^{n} v_i) \end{bmatrix} = \begin{bmatrix} \bar{u} \\ \bar{v} - \bar{u} \end{bmatrix}$$

$\hat{\beta}_0$ is the intercept of the model which can be interpreted as the sample mean for male.

$\hat{\beta}_1$ is the slope of the model which can be interpreted as the difference between the sample mean for female and sample mean for male.

(d) Answer:

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ &= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{2}{n} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \cdots & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 \cdots & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n} & -\frac{1}{n} \\ \frac{1}{n} & -\frac{1}{n} \\ \vdots & \vdots \\ \frac{1}{n} & -\frac{1}{n} \\ 0 & \frac{1}{n} \\ 0 & \frac{1}{n} \\ \vdots & \vdots \\ 0 & \frac{1}{n} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \cdots & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 \cdots & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & 0 & 0 \cdots & 0 \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & 0 & 0 \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \cdots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & 0 & 0 \cdots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{n} & \frac{1}{n} \cdots & \frac{1}{n} \\ 0 & 0 & \cdots & 0 & \frac{1}{n} & \frac{1}{n} \cdots & \frac{1}{n} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \cdots & \vdots \\ 0 & 0 & \cdots & 0 & \frac{1}{n} & \frac{1}{n} \cdots & \frac{1}{n} \end{bmatrix} \end{aligned}$$

(e) Answer:

$$\hat{Y} = HY = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & 0 & 0 \cdots & 0 \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & 0 & 0 \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \cdots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & 0 & 0 \cdots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{n} & \frac{1}{n} \cdots & \frac{1}{n} \\ 0 & 0 & \cdots & 0 & \frac{1}{n} & \frac{1}{n} \cdots & \frac{1}{n} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \cdots & \vdots \\ 0 & 0 & \cdots & 0 & \frac{1}{n} & \frac{1}{n} \cdots & \frac{1}{n} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n u_i \\ \frac{1}{n}\sum_{i=1}^n u_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n u_i \\ \frac{1}{n}\sum_{i=1}^n v_i \\ \frac{1}{n}\sum_{i=1}^n v_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n v_i \end{bmatrix} = \begin{bmatrix} \bar{u} \\ \bar{u} \\ \vdots \\ \bar{u} \\ \bar{v} \\ \bar{v} \\ \vdots \\ \bar{v} \end{bmatrix}$$

(f) Answer:

$$SSE = (Y-\hat{Y})'(Y-\hat{Y}) = \begin{bmatrix} u_1 - \bar{u} \\ u_2 - \bar{u} \\ \vdots \\ u_n - \bar{u} \\ v_1 - \bar{v} \\ v_2 - \bar{v} \\ \vdots \\ v_n - \bar{v} \end{bmatrix}' \begin{bmatrix} u_1 - \bar{u} \\ u_2 - \bar{u} \\ \vdots \\ u_n - \bar{u} \\ v_1 - \bar{v} \\ v_2 - \bar{v} \\ \vdots \\ v_n - \bar{v} \end{bmatrix} = \sum_{i=1}^n (u_i - \bar{u})^2 + \sum_{i=1}^n (v_i - \bar{v})^2$$

$$MSE = \frac{SSE}{2n-2} = \frac{\sum_{i=1}^n (u_i - \bar{u})^2 + \sum_{i=1}^n (v_i - \bar{v})^2}{2n-2} = \frac{(n-1)s_u^2 + (n-1)s_v^2}{2n-2}$$

(g) Answer:

$\hat{\sigma}^2 = s^2 = \frac{(n-1)s_u^2 + (n-1)s_v^2}{n+n-2} = \frac{(n-1)s_u^2 + (n-1)s_v^2}{2n-2}$

(h) Answer:

They are equivalent.