Homework 2

Yunlin Zhang

2015/10/09

1 Problem 1

1.1

For OLS, we can derive the following properties:

$Cov[X, Y] = \hat{\beta}_1/\hat{\beta}_1 Cov[X, Y]$
$\quad = Cov[\hat{\beta}_1 X, Y]/be\hat{t}a_1$
$\quad = Cov[\hat{\beta}_1 X + \hat{\beta}_0, Y]/\hat{\beta}_1$
$\quad = Cov[\hat{Y}, Y]/\hat{\beta}_1$
$\sigma[X] = \hat{\beta}_1/\hat{\beta}_1 \sigma[X]$
$\quad = \sigma[\hat{\beta}_1 X]/\hat{\beta}_1$
$\quad = \sigma[\hat{\beta}_1 X + \hat{\beta}_0]/\hat{\beta}_1$
$\quad = \sigma[\hat{Y}]/\hat{\beta}_1$
$\Rightarrow \rho_{xy} = Cov[X, Y]/\sigma[X]\sigma[Y] = Cov[\hat{Y}, Y]/\sigma[\hat{Y}]\sigma[Y]$
$\quad = \sum(\hat{y}_i - \overline{y})(y_i - \overline{y})/\sqrt{\sum(\hat{y}_i - \overline{y})^2}\sqrt{\sum(y_i - \overline{y})^2}$
$\quad = \sum(\hat{y}_i - \overline{y})(y_i - \hat{y}_i + \hat{y}_i - \overline{y})/\sqrt{\sum(\hat{y}_i - \overline{y})^2}\sqrt{\sum(y_i - \overline{y})^2}$
$\quad = \sum[(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})(\hat{y}_i - \overline{y})]/\sqrt{\sum(\hat{y}_i - \overline{y})^2}\sqrt{\sum(y_i - \overline{y})^2}$
$\quad = \sum[(\hat{y}_i - \overline{y})\hat{u}_i + (\hat{y}_i - \overline{y})^2]/\sqrt{\sum(\hat{y}_i - \overline{y})^2}\sqrt{\sum(y_i - \overline{y})^2}$
$\quad = \sum[(\hat{y}_i - \overline{y})^2]/\sqrt{\sum(\hat{y}_i - \overline{y})^2}\sqrt{\sum(y_i - \overline{y})^2} \qquad \because \sum(\hat{y}_i - \overline{y})\hat{u}_i = 0$
$\quad = \sqrt{\sum[(\hat{y}_i - \overline{y})^2]}/\sqrt{\sum(y_i - \overline{y})^2}$
$\quad = \sqrt{SSE/SST}$
$\quad = \sqrt{R^2}$
$\Rightarrow \rho_{xy}^2 = R^2$ for single variate OLS

1.2

In general, $R^2 \neq \rho_{xy}^2$ in the multivariate case. Observe that in the proof for the single variable case, a transformation is made from X to $\hat{Y}$. However, in the multivariable case, the covariance terms would not reduce to 0 after making the transformation, and therefore the two values would not converge.

2 Problem 2

2.1

For a general linear model of the form $y_i = \sum_{j=0}^{k} \beta_j x_{ij} + u_i$

If MLR.1-MLR.6 are satisfied

$\Rightarrow \hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2/[SST_j(1-R_j^2)])$

For simplicity let $SST_j' = SST_j(1-R_j^2)$

$se[\hat{\beta}_j] = \sqrt{s^2/SST_j'}$

$s^2 = \frac{1}{n-k-1}\sum \hat{u}_i^2$

By MLR.6 $\hat{u} \sim \mathcal{N}(0,\sigma^2) \Rightarrow \hat{u}/\sigma \sim \mathcal{N}(0,1) \Rightarrow \sum \hat{u_i}^2 \sim \sigma^2 \mathcal{X}^2$

$\Rightarrow s^2(n-k-1)/\sigma^2 \sim \mathcal{X}_{n-k-1}^2$

$\Rightarrow (\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) = (\hat{\beta}_j - \beta_j)/\sqrt{s^2/SST_j'}$

$= (\hat{\beta}_j - \beta_j)/\sqrt{\sigma^2/SST_j'}/\sqrt{s^2/\sigma^2}$ $\because$ multiplying by 1

$= (\hat{\beta}_j - \beta_j)/\sqrt{\sigma^2/SST_j'}/\sqrt{s^2(n-k-1)/[\sigma^2(n-k-1)]}$ $\because$multiplying by 1

$\sim \mathcal{N}(0,1)/\sqrt{\mathcal{X}_{n-k-1}^2/(n-k-1)}$

$\sim t_{n-k-1}$

2.2

Based on results from previous problem, $s^2(df)/\sigma^2 \sim \mathcal{X}_{df}^2$

Also in general, $SSR = (df)s^2 \sim \sigma^2 \mathcal{X}_{df}^2$

Since unrestricted model has $n-k-1$ degrees of freedom$\Rightarrow$ restricted model has $n-k-1-q$ degrees of freedom

This implied: $(SSR_r - SSR_{ur})/q \sim \sigma^2(\mathcal{X}_{n-k-1-q}^2 - \mathcal{X}_{n-k-1}^2)/q$

$\sim \sigma^2(\mathcal{X}_q^2/q)$

And

$SSR_{ur}/(n-k-1) \sim \sigma^2\mathcal{X}_{n-k-q}^2$

The σ^2 in top and bottom will cancel

$\Rightarrow [(SSR_{ur} - SSR_r)/q]/[SSR_{ur}/(n-k-1)] \sim (\mathcal{X}_q^2/q)/(\mathcal{X}_{n-k-1}^2/(n-k-1))$

$\sim F_{q,n-k-1}$

3 Problem 3

$\tilde{\beta}_1 = \sum \hat{r}_{i1}y_i / \sum \hat{r}_{i1}^2$ Plugging in $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$

Working with the numerator:

$\sum \hat{r}_{i1}y_i = \sum \hat{r}_{i1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i)$

But $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = \beta_1 \hat{r}_{i1}$

$\Rightarrow \sum \hat{r}_{i1}y_i = \beta_1 \sum \hat{r}_{i1}^2 + \beta_3 \sum \hat{r}_{i1}x_{i3}$

$\Rightarrow E[\tilde{\beta}_1 = E[(\beta_1 \sum \hat{r}_{i1}^2 + \beta_3 \sum \hat{r}_{i1}x_{i3}) / \sum \hat{r}_{i1}^2]$

$= \beta_1 + \beta_3 \ \sum \hat{r}_{i1}x_{i3} / \sum \hat{r}_{i1}^2$

4 Problem 4

$\log wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$

4.1

Another year of general work experience has as much effect on $\log wage$ as another year of tenure with current employer

$H_0 : \hat{\beta}_2 = \hat{\beta}_3$

However, this can also be written as:

$H_0 : \theta = \hat{\beta}_2 - \hat{\beta}_3 = 0$

And the model equation will be transformed to:

$\log(wage) = \beta_0 + \beta_1 educ + \theta exper + \beta_2(exper + tenure)$

4.2

Calculated t-statistic and p-value for θ:

$t_\theta = 0.412$

$p_\theta = 0.680$

Based on the t-test, θ is only non-zero with a 68% significance. This means that we cannot reject the null hypothesis that θ is 0, and therefore we cannot reject that the effect of tenure at current position and experience at previous position are the same on current (log)wages.

5 Problem 5

$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + \beta_3 prppov + u$

5.1

Data is cleaned by removing empty entries and 0 entries for income and psoda. Calculated t-statistics and p-value for β_1:

$t_{\beta_1} = 2.373$

$p_{\beta_1} = 0.018 = 1.8\%$

Based on the t-test, the null hypothesis would be rejected at 5% significance level, while it would not be rejected at 1% significance level.

5.2

Calculation is carried out in Excel $Corr[\log(income), prppov] = -0.840$

These two variables are pretty highly correlated, which is also visible based on a scatter plot between them.

p-value for prppov = 0.0044

and

p-value for log(income) = 4.8e-7

With very high confidence we can reject the null hypothesis, therefore these two variables are significant in the regression model.

5.3

p-value for log(hseval) = 2.67e-11
With very high confidence we can reject the null hypothesis, and therefore this parameter is significant.

5.4

p-value for log(income) became 0.159
and
p-value for prppov became 0.699
Therefore individually, the became non-significant.

For testing join significance:

$H_0 : \beta_2 = \beta_3 = 0$

Running F test by restricting these two variables: $SSR_r = 0.450982$
From part iii, $SSR_{ur} = 0.0443098$
Running an F-test: $q = 2$, $n - k - 1 = 401 - 4 - 1 = 396$
$F = [(SSR_r - SSR_{ur})/q]/[SSR_{ur}/(n - k - 1)] = 1857$
At 95% confidence/5% significance, $F_{2,396} = 3$ therefore we reject the null hypothesis, and these variables are significant jointly.

5.5

Based on the tests in the previous sections, the most reliable model is from part iii, where
$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + \beta_3 prppov + \beta_4 \log(hseval) + u$

6 Problem 6

The second model is preferable since it has a higher $\overline{R}^2$ value than that of the other two models. Without seeing the data, basing just on the three models, it looks like a linear response on totemp is too strong and had to be corrected by a smaller quadratic term, which makes the log model work better in the regime that the sampled data is in.