Solutions to Extra Practice Questions - Week 2

Statistics and Econometrics

Question 1

Using the data set ceosal1.RData to answer the following questions. Consider an equation to explain salaries of CEOs in terms of annual firm sales, return on equity (*roe*, in percentage form), and return on the firm's stock (*ros*, in percentage form):

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u$$

1. In terms of the model parameters, state the null hypothesis that, after controlling for *sales* and *roe*, *ros* has no effect on CEO salary. State the alternative that better stock market performance increases a CEO's salary.
2. Estimate the model and report your results. By what percentage is *salary* predicted to increase if *ros* increases by 50 basis points (i.e., *ros* increases by 50)? Does *ros* have a practically large effect on *salary*?
3. Test the null hypothesis that *ros* has no effect on *salary* against the alternative that *ros* has a positive effect. Carry out the test at the 10% significance level (please show clearly the test statistic and the critical value used in your testing).
4. Would you include *ros* in a final model explaining CEO compensation in terms of firm performance? Explain.

Solutions

1. $H_0 : \beta_3 = 0; H_1 : \beta_3 > 0.$
2.

```r
load("ceosal1.RData")
fitted.salary <- lm(log(salary) ~ log(sales) + roe + ros, data = data)
```

The estimated equation is

$$\widehat{\log(salary)} = \underset{(.32)}{4.31} + \underset{(.035)}{.280} \log(sales) + \underset{(.0041)}{.0174} roe + \underset{(.00054)}{.00024} ros,$$

$n = 209, R^2 = .283$. The proportionate effect on *salary* is $.00024 \times 50 = .012$. To obtain the percentage effect, we multiply this by 100: 1.2%. Therefore, 50 basis points increase in *ros* is predicted to increase salary by only 1.2%. Practically speaking, this is a very small effect for such a large change in *ros*.

3. For $n = 209$, we can use the standard normal distribution. The 10% critical value for a one-tailed test is 1.28. The t statistic on *ros* is $.00024/.00054 \approx .44$, which is well below the critical value. Therefore, we fail to reject H_0 at the 10% significance level.

4. Based on this sample, the estimated *ros* coefficient appears to be different from zero only because of sampling variation. On the other hand, including *ros* may not be causing any harm; it depends on how correlated it is with the other independent variables (although these are very significant even with *ros* in the equation).

Question 2

Use the data set lawsch85.RData to answer the following questions. Consider an equation to explain the median starting salary for new law school graduates

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u,$$

where $LSAT$ is the median LSAT score for the graduating class, GPA is the median college GPA for the class, $libvol$ is the number of volumes in the law school library, $cost$ is the annual cost of attending law school, and $rank$ is a law school ranking (with $rank = 1$ being the best).

1. Estimate the model. State and test the null hypothesis that the rank of law schools has no causal effect on median starting salary (please show clearly the test statistic and the critical value used in your testing).
2. Are features of students - namely, $LSAT$ and GPA - individually or jointly significant for explaining $salary$?
3. Test whether the size of the class ($clsize$) or the size of the faculty ($faculty$) needs to be added to this equation: carry out a single test for joint significance of the two variables.

Solutions

1.

```
load("lawsch85.RData")
fitted.salary <- lm(log(salary) ~ LSAT + GPA + log(libvol) + log(cost) + rank, data = data)
```

The estimated equation is

$$\widehat{\log(salary)} = \underset{(.533)}{8.343} + \underset{(.004)}{.005}\, LSAT + \underset{(.090)}{.248}\, GPA + \underset{(.033)}{.095}\, \log(libvol) + \underset{(.032)}{.038}\, \log(cost) - \underset{(.0003)}{.0033} rank,$$

$n = 136, R^2 = .842$. The hypothesis that $rank$ has no effect on $\log(salary)$ is $H_0 : \beta_{rank} = 0$. The t statistic on $rank$ is $-.0033/.0003 \approx -11$, which is very significant (critical value for 1% significance level against a two-sided alternative is 2.576). If $rank$ decreases by 10 (which is a move up for a law school), median starting salary is predicted to increase by about 3.3%.

2.

```
data.sub <- data %>% select(lsalary, LSAT, GPA, llibvol, lcost, rank) %>% na.omit
salary.ur <- lm(lsalary ~ LSAT + GPA + llibvol + lcost + rank, data = data.sub)
salary.r <- lm(lsalary ~ llibvol + lcost + rank, data = data.sub)

# calculate F statistic
F.stat <- (summary(salary.ur)$r.squared - summary(salary.r)$r.squared)/2 /
  ((1 - summary(salary.ur)$r.squared)/salary.ur$df.residual)

# p value for the F test
pf(F.stat, 2, salary.ur$df.residual, lower.tail = FALSE)
```

```
## [1] 9.518119e-05
```

```
# using the built-in function
linearHypothesis(salary.ur, c("LSAT = 0", "GPA = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## LSAT = 0
```

```
## GPA = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ LSAT + GPA + llibvol + lcost + rank
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    132 1.8942
## 2    130 1.6427  2   0.25151 9.9518 9.518e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$LSAT$ is not statistically significant (t statistic ≈ 1.17, p-value $\approx .244$) but GPA is very significant (t statistic ≈ 2.75, p-value $\approx .007$). The test for joint significance is moot given that GPA is so significant, but for completeness, the F statistic is about 9.95 (with 2 and 130 *df*) and p-value $\approx .0001$.

3.

```r
data.sub2 <- data %>% select(lsalary, LSAT, GPA, llibvol, lcost, rank, clsize, faculty) %>% na.omit
salary.ur2 <- lm(lsalary ~ LSAT + GPA + llibvol + lcost + rank + clsize + faculty, data = data.sub2)
salary.r2 <- lm(lsalary ~ LSAT + GPA + llibvol + lcost + rank, data = data.sub2)

# calculate F statistic
F.stat2 <- (summary(salary.ur2)$r.squared - summary(salary.r2)$r.squared)/2 /
  ((1 - summary(salary.ur2)$r.squared)/salary.ur2$df.residual)

# p value for the F test
pf(F.stat2, 2, salary.ur2$df.residual, lower.tail = FALSE)
```

```
## [1] 0.3901833
```

```r
# using the built-in function
linearHypothesis(salary.ur2, c("clsize = 0", "faculty = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## clsize = 0
## faculty = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ LSAT + GPA + llibvol + lcost + rank + clsize + faculty
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    125 1.5974
## 2    123 1.5732  2  0.024259 0.9484 0.3902
```

When we add $clsize$ and $faculty$ to the regression we lose five observations. The test of their joint significance (with 2 and 123 *df*) gives $F \approx .95$ and p-value $\approx .39$. So these two variables are not jointly significant unless we use a very large significance level.

Question 3

Use the data in discrim.RData to answer this question. These are ZIP code-level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of African Americans.

1. Consider a model to explain the price of soda in its log form, $\log(psoda)$, in terms of the proportion of the population that is African American and log of median income:

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

Estimate the model and report your results. If $prpblck$ increases by .20, what is the estimated percentage change in $psoda$?
2. Compare the estimate from part 1 with the simple regression estimate from $\log(psoda)$ on $prpblck$. Is the discrimination effect larger or smaller when you control for income? Explain.
3. Now add the variable $prppov$ to the regression in part 1 and report your results. What happens to $\hat{\beta}_{prpblck}$? Is $\hat{\beta}_{prpblck}$ statistically different from zero at the 5% level against a two-sided alternative? What about at the 1% level? (please show clearly the test statistic and the critical value used in your testing)
4. Find the correlation between $\log(income)$ and $prppov$. Is it roughly what you expected?
5. Now add the variable $\log(hseval)$ to the regression in part 3. Interpret its coefficient and report the two-sided p-value for $H_0 : \beta_{\log(hseval)} = 0$.
6. In the regression in part 5, what happens to the individual statistical significance of $\log(income)$ and $prppov$? Are these variables jointly significant? (Report p-value) What do you make of your answers?
7. Given the results of the previous regressions, which one would you report as most reliable in determining whether the racial makeup of a zip code influences local fast-food prices?

Solutions

1.

```r
load("discrim.RData")
fitted.price <- lm(log(psoda) ~ prpblck + log(income), data = data)
```

The estimated equation is:

$$\widehat{\log(psoda)} = -\underset{(.179)}{.794} + \underset{(.026)}{.122}\, prpblck + \underset{(.017)}{.077} \log(income),$$

where $n = 401$, $R^2 = .068$. If $prpblck$ increases by .20, $\log(psoda)$ is estimated to increase by $.20 \times .122 = .0244$, or about 2.44 percent.

2.

```r
fitted.simple <- lm(log(psoda) ~ prpblck, data = data)
stargazer(fitted.simple, header = FALSE, type = 'latex', title = "Question 3.2")
```

```r
cor(data$lincome, data$prpblck, use = "complete.obs")
```

[1] -0.4966359

The simple regression estimate on $prpblck$ is .062, so the simple regression estimate is actually lower. This is because $prpblck$ and $\log(income)$ are negatively correlated ($-.50$) and $\log(income)$ has a positive coefficient in the multiple regression.

3.

```r
fitted.price2 <- lm(log(psoda) ~ prpblck + log(income) + prppov, data = data)
stargazer(fitted.price2, header = FALSE, type = 'latex', title = "Question 3.3")
```

$\hat{\beta}_{prpblck}$ falls to about .073 when $prppov$ is added to the regression.

For $n = 401$, we can use the standard normal distribution. The 5% critical value for a two-tailed test is 1.96, and the 1% critical value for a two-tailed test is 2.57. The test statistic is $0.073/0.031 \approx 2.373$, which is greater than 1.96 but less than 2.57. So that we can reject H_0 at the 5% level but not at the 1% level.

Table 1: Question 3.2

	Dependent variable:
	log(psoda)
prpblck	0.062***
	(0.023)
Constant	0.033***
	(0.005)
Observations	401
R^2	0.018
Adjusted R^2	0.016
Residual Std. Error	0.084 (df = 399)
F Statistic	7.451*** (df = 1; 399)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Question 3.3

	Dependent variable:
	log(psoda)
prpblck	0.073**
	(0.031)
log(income)	0.137***
	(0.027)
prppov	0.380***
	(0.133)
Constant	−1.463***
	(0.294)
Observations	401
R^2	0.087
Adjusted R^2	0.080
Residual Std. Error	0.081 (df = 397)
F Statistic	12.604*** (df = 3; 397)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.

```r
cor(data$lincome, data$prppov, use = "complete.obs")
```

```
## [1] -0.838467
```

The correlation is about -.84 , which makes sense because poverty rates are determined by income (but not directly in terms of median income).

5.

```r
fitted.price3 <- lm(log(psoda) ~ prpblck + log(income) + prppov + log(hseval), data)
stargazer(fitted.price3, header = FALSE, type = 'latex', title = "Question 3.5", no.space = TRUE)
```

Table 3: Question 3.5

	Dependent variable:
	log(psoda)
prpblck	0.098***
	(0.029)
log(income)	−0.053
	(0.038)
prppov	0.052
	(0.134)
log(hseval)	0.121***
	(0.018)
Constant	−0.842***
	(0.292)
Observations	401
R^2	0.184
Adjusted R^2	0.176
Residual Std. Error	0.077 (df = 396)
F Statistic	22.313*** (df = 4; 396)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient on log($hseval$) indicates that one percent increase in housing value, holding the other variables fixed, increases the predicted price by about .12 percent. The two-sided p-value is approximately zero.

6.

```r
linearHypothesis(fitted.price3, c("log(income) = 0", "prppov = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(income) = 0
## prppov = 0
##
## Model 1: restricted model
## Model 2: log(psoda) ~ prpblck + log(income) + prppov + log(hseval)
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    398 2.3911
## 2    396 2.3493  2  0.041797 3.5227 0.03045 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding log(*hseval*) makes log(*income*) and *prppov* individually insignificant (at even the 15% significance level against a two-sided alternative for log(*income*), and *prppov* does not have a t statistic even close to one in absolute value). Nevertheless, they are jointly significant at the 5% level because the outcome of the $F_{2,396}$ statistic is about 3.52 with p-value = .030. All of the control variables - log(*income*), *prppov*, and log(*hseval*) - are highly correlated, so it is not surprising that some are individually insignificant.

7. Because the regression in part 5 contains the most controls, log(*hseval*) is individually significant, and log(*income*) and *prppov* are jointly significant, part 5 seems the most reliable. It holds fixed three measure of income and affluence. Therefore, a reasonable estimate is that if the proportion of African Americans increases by .10, *psoda* is estimated to increase by 1%, other factors held fixed.