Suchaya Chuldilok
6104641060

Homework 6

Chapter 6

3. Using the data in RDCHEM, the following equation was obtained by OLS:

$$\widehat{rdintens} = 2.613 + .00030\ sales - .0000000070\ sales^2$$
$$(.429) \quad (.00014) \qquad (.0000000037)$$
$$n = 32, R^2 = .1484.$$

i. At what point does the marginal effect of *sales* on *rdintens* become negative?

$$\frac{\partial rdintens}{\partial sales} = 0.003 - 0.000000014\ sales = 0$$

$$sales = \frac{0.003}{0.000000014}$$

$$sales = 21,428.57$$

∴ The point marginal effect become negative is sales = 21,428.57

ii. Would you keep the quadratic term in the model? Explain.

Yes, because $t_{stat} = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{-0.000000007}{0.0000000037} = -1.89$

the tstat is less than zero so, it is significant.

iii. Define *salesbil* as sales measured in billions of dollars: $salesbil = sales/1{,}000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the *R*-squared. [*Hint*: Note that $salesbil^2 = sales^2/(1{,}000)^2$.]

$$\widehat{ardintens} = 2.613 + 0.30\ salesbil - 0.0070\ salesbil^2$$
$$(0.429) \quad (0.14) \qquad (0.0037)$$

$$n = 32,\ R^2 = 0.1484$$

iv. For the purpose of reporting the results, which equation do you prefer?

The equation from question (iii) is easier to read because it contains fewer zeros, the interpretation of 2 equations is the same once the different scales are accounted for.

Suchaya Chuldilok
6104641060

Chapter 7

1. Using the data in SLEEP75 (see also Problem 3 in Chapter 3), we obtain the estimated equation

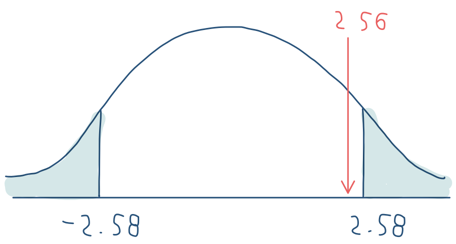
$$\widehat{sleep} = \underset{(235.11)}{3{,}840.83} - \underset{(.018)}{.163}\, totwrk - \underset{(5.86)}{11.71}\, educ - \underset{(11.21)}{8.70}\, age$$
$$+ \underset{(.134)}{.128}\, age^2 + \underset{(34.33)}{87.75}\, male$$
$$n = 706,\ R^2 = .123,\ \overline{R}^2 = .117.$$

The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

i. All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?

the coefficient on male is 87.75 so, men is estimated to sleep more per week compare to women.

$$t_{male} = \frac{87.75}{34.33} \approx 2.56$$



close to 1% critical value so, the evidence is strong.

ii. Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

$$t_{totwork} = \frac{-1.63}{0.018} \approx -9.06$$ is statiscally significant

the coefficient implies that one more hour of work is $(0.163)(60) \approx 9.8$ minutes less sleep.

iii. What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

the null hypothesis we are testing $H_0: \hat{\beta}_3 = \hat{\beta}_4 = 0$
run restricted version of the regression where age, age^2 are omitted by calculating

$$F = \frac{(R^2_{ur} - R^2_{r})}{(1 - R^2_{ur}) / (n-k-1)}$$

8. Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: "On how many separate occasions last month did you smoke marijuana?"

Suchaya Chuldilok
6104641060

i. Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by *x*%."

ii. Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

iii. Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.

iv. Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

v. What are some potential problems with drawing causal inference using the survey data that you collected?

(i) $\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 female$

(ii) $\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 female + \beta_5 usage \cdot female$

To test $H_0 : \beta_5 = 0$
$H_a : \beta_5 \neq 0$

(iii) Assuming that there is no interaction between sex and usage

$\log(wage) = \beta_0 + \beta_1 light + \beta_2 moderate + \beta_3 heavy + \beta_2 educ + \beta_3 exper + \beta_4 female + u$

non-user is the omitted category

(iv) The null hypothesis is $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

peform F test which $q = 3$ df. $= n - 6 - 1$

(v.) Respondents may not accurately report their marijuana usage out of fear of legal repurcussion or there maybe omitted variables which determine wage and usage.

Suchaya Chuldilok
6104641060

11. The following equations were estimated using the data in ECONMATH, with standard errors reported under coefficients. The average class score, measured as a percentage, is about 72.2; exactly 50% of the students are male; and the average of *colgpa* (grade point average at the start of the term) is about 2.81.

$$\widehat{score} = \underset{(2.00)}{32.31} + \underset{(0.70)}{14.32}\, colgpa$$
$$n = 856, R^2 = .329, \bar{R}^2 = .328.$$

$$\widehat{score} = \underset{(2.04)}{29.66} + \underset{(0.74)}{3.83}\, male + \underset{(0.69)}{14.57}\, colgpa$$
$$n = 856, R^2 = .349, \bar{R}^2 = .348.$$

$$\widehat{score} = \underset{(2.86)}{30.36} + \underset{(3.96)}{2.47}\, male + \underset{(0.98)}{14.33}\, colgpa + \underset{(1.383)}{0.479}\, male \cdot colgpa$$
$$n = 856, R^2 = .349, \bar{R}^2 = .347.$$

$$\widehat{score} = \underset{(2.86)}{30.36} + \underset{(0.74)}{3.82}\, male + \underset{(0.98)}{14.33}\, colgpa + \underset{(1.383)}{0.479}\, male \cdot (colgpa - 2.81)$$
$$n = 856, R^2 = .349, \bar{R}^2 = .347.$$

i. Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero?

ii. In the second equation, how come the estimate on *male* is so imprecise? Should we now conclude that there are no gender differences in *score* after controlling for *colgpa*? [*Hint*: You might want to compute an *F* statistic for the null hypothesis that there is no gender difference in the model with the interaction.]

iii. Compared with the third equation, how come the coefficient on *male* in the last equation is so much closer to that in the second equation and just as precisely estimated?

(i) interpretation : as 1 male increase, the score will increase by 3.83

CI at 95% confidence = $3.83 \pm 1.96(0.74)$
0 is excluded because the interval is btw 2.379 and 5.2804

(ii) From equation 3, we have interaction term among the variables so the estimate on male has a higher s.e.

F-test ; $H_0 : \beta_1 = \beta_3 = 0$ in equation 3

H_1 : otherwise

$$F = \frac{(0.349 - 0.329)/2}{\frac{1 - 0.344}{852}} = 13.08$$

$F_{(0.05, 852)} = 3.006$

Reject Ho because gender differences are significant.

(iii) because in the equation 4 variable male* (colgpa - 2.81) has been subtracted and the result is closer to zero and more precise OLS.

C4. Use the data in GPA2 for this exercise.

Suchaya Chuldilok
6104641060

i. Consider the equation

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

where *colgpa* is cumulative college grade point average; *hsize* is size of high school graduating class, in hundreds; *hsperc* is academic percentile in graduating class; *sat* is combined SAT score; *female* is a binary gender variable; and *athlete* is a binary variable, which is one for student-athletes. What are your expectations for the coefficients in this equation? Which ones are you unsure about?

ii. Estimate the equation in part (i) and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?

iii. Drop *sat* from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).

iv. In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no ceteris paribus difference between women athletes and women nonathletes.

v. Does the effect of *sat* on *colgpa* differ by gender? Justify your answer.

(i) >> β_3 are definitely less than zero because high school percentile is defined so that the smaller the number the better the student do.

>> β_4 > 0 because SAT scores cannot be negative.

>> others coefficients are unclear.

(ii)

```
. reg colgpa hsize hsizesq hsperc sat female athlete
```

Source	SS	df	MS
Model	524.819305	6	87.4698842
Residual	1269.37637	4,130	.307355053
Total	1794.19567	4,136	.433799728

Number of obs =	4,137
F(6, 4130) =	284.59
Prob > F =	0.0000
R-squared =	0.2925
Adj R-squared =	0.2915
Root MSE =	.5544

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0568543	.0163513	-3.48	0.001	-.0889117	-.0247968
hsizesq	.0046754	.0022494	2.08	0.038	.0002654	.0090854
hsperc	-.0132126	.0005728	-23.07	0.000	-.0143355	-.0120896
sat	.0016464	.0000668	24.64	0.000	.0015154	.0017774
female	.1548814	.0180047	8.60	0.000	.1195826	.1901802
athlete	.1693064	.0423492	4.00	0.000	.0862791	.2523336
_cons	1.241365	.0794923	15.62	0.000	1.085517	1.397212

$$\widehat{colgpa} = \underset{(0.079)}{1.241} - \underset{(0.0164)}{0.569}hsize + \underset{(0.00225)}{0.00468}hsize^2 - \underset{(0.0006)}{0.132}hsperc + \underset{(0.00007)}{0.00165}sat + \underset{(0.018)}{1.55}female + \underset{(0.042)}{1.69}athlete$$

$n = 4{,}137 \quad R^2 = 0.293$

The tstat $= \frac{169 - 0}{0.042} \approx 4.02$ which is significant

Suchaya Chuldilok
6104641060

(iii)

```
. reg colgpa hsize hsizesq hsperc female athlete
```

Source	SS	df	MS		
Model	338.217123	5	67.6434247	Number of obs =	4,137
Residual	1455.97855	4,131	.35245184	F(5, 4131) =	191.92
				Prob > F =	0.0000
				R-squared =	0.1885
				Adj R-squared =	0.1875
Total	1794.19567	4,136	.433799728	Root MSE =	.59368

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0534038	.0175092	-3.05	0.002	-.0877313	-.0190763
hsizesq	.0053228	.0024086	2.21	0.027	.0006007	.010045
hsperc	-.0171365	.0005892	-29.09	0.000	-.0182916	-.0159814
female	.0581231	.0188162	3.09	0.002	.0212333	.095013
athlete	.0054487	.0447871	0.12	0.903	-.0823582	.0932556
_cons	3.047698	.0329148	92.59	0.000	2.983167	3.112229

The coefficient on athele is approximately 0.0054 in this case, which is not that significant when compare to question (ii) because the SAT score is not being controlled.

(iv) Choosing female nonathlete as a baseline

```
. reg colgpa hsize hsizesq hsperc sat femath maleath malenonath
```

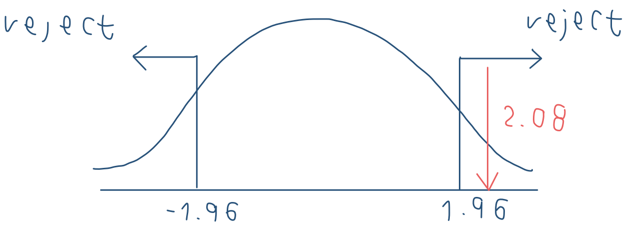
Source	SS	df	MS		
Model	524.821272	7	74.9744674	Number of obs =	4,137
Residual	1269.3744	4,129	.307429015	F(7, 4129) =	243.88
				Prob > F =	0.0000
				R-squared =	0.2925
				Adj R-squared =	0.2913
Total	1794.19567	4,136	.433799728	Root MSE =	.55446

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0568006	.0163671	-3.47	0.001	-.0888889	-.0247124
hsizesq	.0046699	.0022507	2.07	0.038	.0002573	.0090825
hsperc	-.0132114	.000573	-23.06	0.000	-.0143349	-.012088
sat	.0016462	.0000669	24.62	0.000	.0015151	.0017773
femath	.1751106	.0840258	2.08	0.037	.0103748	.3398464
maleath	.0128034	.0487395	0.26	0.793	-.0827523	.1083591
malenonath	-.1546151	.0183122	-8.44	0.000	-.1905168	-.1187133
_cons	1.39619	.0755581	18.48	0.000	1.248055	1.544324

$H_0: \delta_1 = 0$

$t_{0.025, 4129} = 1.96$

$t_{cal} = \frac{0.175}{0.084} = 2.08$



$\therefore$ Reject H_0

(v)

Suchaya Chuldilok
6104641060

```
. gen femsat=female*sat

. regress  colgpa hsize hsizesq hsperc sat female athlete  femsat

      Source |       SS       df       MS              Number of obs =    4137
-------------+------------------------------           F(  7,  4129) =  243.91
       Model |  524.867644     7  74.981092            Prob > F      =  0.0000
    Residual |  1269.32803  4129  .307417784           R-squared     =  0.2925
-------------+------------------------------           Adj R-squared =  0.2913
       Total |  1794.19567  4136  .433799728           Root MSE      =  .55445

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsize |  -.0569121   .0163537    -3.48   0.001    -.0889741   -.0248501
     hsizesq |   .0046864   .0022498     2.08   0.037     .0002757    .0090972
      hsperc |   -.013225   .0005737   -23.05   0.000    -.0143497   -.0121003
         sat |   .0016255   .0000852    19.09   0.000     .0014585    .0017924
      female |   .1023066   .1338023     0.76   0.445    -.1600179    .3646311
     athlete |   .1677568   .0425334     3.94   0.000     .0843684    .2511452
      femsat |   .0000512   .0001291     0.40   0.692     -.000202    .0003044
       _cons |   1.263743   .0974952    12.96   0.000       1.0726    1.454887
------------------------------------------------------------------------------
```

Add female·sat to the equation in question (ii), the coefficient is 0.000051 and t stat ≈ 0.4 It can be concluded that SAT scores differs by gender.