

Homework # 6

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

$$\widehat{rdintens} = 2.613 + .00030 sales - .0000000070 sales^2$$

(0.429) (0.00014) (0.0000000037)

$n = 32, R^2 = .1484.$

- At what point does the marginal effect of *sales* on *rdintens* become negative?
- Would you keep the quadratic term in the model? Explain.
- 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 $salesbil^2 = sales^2/(1,000)^2$.]
- For the purpose of reporting the results, which equation do you prefer?

(iii) $\widehat{rdintens} = 2.613 + 0.30 salesbil - 0.007 salesbil^2$

(0.429) (0.14) (0.0037)

$n = 32, R^2 = 0.1484$

(iv) It is easier in (iii) to read due to the lesser zero that it contains, The interpretation of 2 equations is the same once the difference scales are accounted for.

6104640732

Panachai Ngaechaisoenkul

(i) $\frac{\partial rdintens}{\partial sales} = 0.0003 - 0.000000014 sales = 0$

$sales = \frac{0.0003}{0.000000014} = 21,428.57$

∴ The point that marginal effect is negative is 21,428.57

(ii) Yes due to $t_{stat} = -1.89$ which is significant against alternative $H_0: \beta_1 < 0$ at 5% level (CV ≈ -1.70, df = 29)

$$t_{stat} = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{-0.000000007}{0.000000037} = -1.89$$

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

$$\widehat{sleep} = 3,840.83 - .163 totwrk - 11.71 educ - 8.70 age$$

(235.11) (0.018) (5.86) (11.21)

$$+ .128 age^2 + 87.75 male$$

(0.134) (34.33)

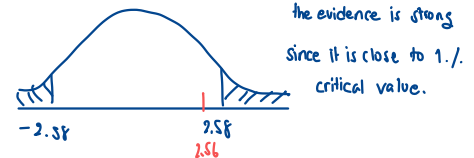
$n = 706, R^2 = .123, \bar{R}^2 = .117.$

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.

- All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?
- Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?
- What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

(i) the coefficient of male is 87.75, within a week man is estimated to sleep more than women.

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



(ii) $t_{totwrk} = \frac{-0.163}{0.018} \approx -0.96$ is statistically significant

the coefficient implies that one more hour of work is 0.163(60) ≈ 9.8 min less sleep.

(iii) the null hypothesis we're testing

$$H_0: \hat{\beta}_3 = \hat{\beta}_4 = 0$$

now run restricted version where age, age^2 are omitted by calculating

$$F = \frac{(R_{ur}^2 - R_r^2)}{(1 - R_{ur}^2)/(n - k - 1)}$$

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?"

- 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?

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

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

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

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

$$\log(\text{wage}) = \beta_0 + \delta_0 \text{light} + \delta_1 \text{moderate} + \delta_2 \text{heavy} + \beta_2 \text{educ} + \beta_3 \text{exper} + \beta_4 \text{female} + u$$

non-user is the omitted category

(iv) The null hypothesis is $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ Perform F test which $q = 3$ $df = n - 6 - 1$

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

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.

$$\Gamma \widehat{\text{score}} = 32.31 + 14.32 \text{ colgpa}$$

(2.00) (0.70)
 $n = 856, R^2 = .329, \bar{R}^2 = .328.$

$$\widehat{\text{score}} = 29.66 + 3.83 \text{ male} + 14.57 \text{ colgpa}$$

(2.04) (0.74) (0.69)
 $n = 856, R^2 = .349, \bar{R}^2 = .348.$

$$\text{WR} \widehat{\text{score}} = 30.36 + 2.47 \text{ male} + 14.33 \text{ colgpa} + 0.479 \text{ male} \cdot \text{colgpa}$$

(2.86) (3.96) (0.98) (1.383)
 $n = 856, R^2 = .349, \bar{R}^2 = .347.$

$$\widehat{\text{score}} = 30.36 + 3.82 \text{ male} + 14.33 \text{ colgpa} + 0.479 \text{ male} \cdot (\text{colgpa} - 2.81)$$

(2.86) (0.74) (0.98) (1.383)
 $n = 856, R^2 = .349, \bar{R}^2 = .347.$

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 the increasing in score by 3.83 when one unit added.

CI at 95% confidence = $3.83 \pm 1.96 (0.74)$

0 is excluded due to interval is between (2.379, 5.2804)

(ii) In Eq (3), we have interaction term among the variables so the estimate on male has a higher SE. Do the F test,

$H_0: \beta_1 = \beta_2 = 0$ in Equation (3)
 $H_a: \text{otherwise}$

$$F = \frac{(0.349 - 0.342) / 2}{\frac{1 - 0.349}{852}} = 13.08$$

} so we reject H_0
 gender differences are significant

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

(iii) Because of Eq (4) variable $\text{male} \cdot (\text{colgpa} - 2.81)$ has been subtract by the mean of *colgpa* (2.81) making it close to 0 and be more precise OLS.

C4. Use the data in GPA2 for this exercise.

i. Consider the equation

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

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?

- β_3 are less than zero due to high school percentile is determined so that the smaller the number the better the student is.
- $\beta_4 > 0$ due to SAT scores only be positive
- others are unclear

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?

```
. reg colgpa hsize hsize^2 hspec sat female athlete
```

Source	SS	df	MS	Number of obs	F(6, 4130)
Model	524.819305	6	87.4698842	=	284.59
Residual	1269.37637	4,130	.307355053	=	0.0000
Total	1794.19567	4,136	.433799728	=	0.2925

colgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hsize	-.0568543	.0163513	-3.48	0.001	-.0889117 -.0247968
hsize^2	.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	.1548914	.0180047	8.60	0.000	.1195926 .1901802
athlete	.1693064	.0423492	4.00	0.000	.0862791 .2523336
_cons	1.241365	.0794923	15.62	0.000	1.085517 1.397212

$$colgpa = 1.241 - 0.569hsize + 0.00468hsize^2 - 0.0132hsperc + 0.00165sat + 0.155female + 0.169athlete$$

$$N = 4,137 \quad R^2 = 0.293$$

An athlete is predicted to have a GPA

≈ 0.169 points higher than non athlete, CETERIS

PARIBUS. Tstat = $\frac{0.169 - 0}{0.042} \approx 4.02$ is 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).

```
. reg colgpa hsize hsize^2 hspec female athlete
```

Source	SS	df	MS	Number of obs	F(5, 4131)
Model	338.217123	5	67.6434247	=	191.92
Residual	1455.97855	4,131	.35245184	=	0.0000
Total	1794.19567	4,136	.433799728	=	0.1885

colgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hsize	-.0534038	.0175092	-3.05	0.002	-.0877313 -.0190763
hsize^2	.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 athlete becomes ≈ 0.0054 which is not significant as part (ii) because we don't control SAT scores

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.

```
. reg colgpa hsize hsize^2 hspec sat femath maleath malenonath
```

Source	SS	df	MS	Number of obs	F(7, 4129)
Model	524.821272	7	74.9744674	=	243.88
Residual	1269.3744	4,129	.307429015	=	0.0000
Total	1794.19567	4,136	.433799728	=	0.2925

colgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hsize	-.0568006	.0163671	-3.47	0.001	-.0888889 -.0247124
hsize^2	.0046699	.0022507	2.07	0.038	.0002573 .0090825
hsperc	-.0132114	.0005733	-23.06	0.000	-.0143349 -.012088
sat	.0016462	.0000669	24.62	0.000	.0015151 .0017773
femath	.1751108	.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	.0755881	18.48	0.000	1.248055 1.544324

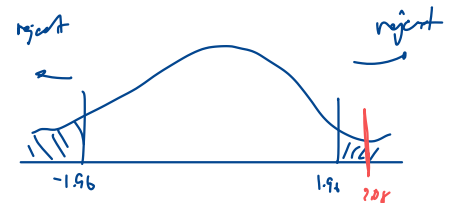
To test the hypothesis we choose

female nonathlete as base group

$$H_0: \delta_i = 0$$

$$t_{0.025, 4129} = 1.96$$

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



Therefore, reject H_0

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

```
. gen femsat=female*sat
```

```
. regress colgpa hsize hsize^2 hspec sat female athlete femsat
```

Source	SS	df	MS	Number of obs	F(7, 4129)
Model	524.867644	7	74.981092	=	243.91
Residual	1269.32803	4129	.307417784	=	0.0000
Total	1794.19567	4136	.433799728	=	0.2925

colgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hsize	-.0569121	.0163537	-3.48	0.001	-.0889741 -.0248501
hsize^2	.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

So, we add *female . sat* to the eqn in (ii) its coefficient is about 0.000051 and

t stat ≈ 0.4 there is little evidence that sat scores differs by gender.