

Ti Sahornvace 6104640658

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

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

(.429) (.00014) (.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: $\text{salesbil} = \text{sales}/1,000$. 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 $\text{salesbil}^2 = \text{sales}^2/(1,000)^2$.]
- For the purpose of reporting the results, which equation do you prefer?

$$i) \frac{\partial \widehat{rdintens}}{\partial \text{sales}} = .0003 - .000000014 \text{ sales} = 0$$

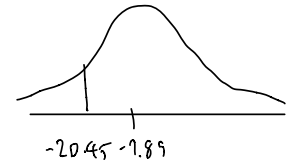
$$\text{sales} = \frac{.0003}{.000000014}$$

$$\text{sales} = 21,428.57 \quad \#$$

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

so, at 95% CI, $df = 29$, $t_{crit} = 2.045$ & $|-1.89| < 2.045$

$\therefore$ No, because quadratic term has no significant



$$iii) \widehat{rdintens} = 2.613 + 0.30 \text{ salesbil} - 0.0070 \text{ salesbil}^2$$

(.429) (.014) (.0037)

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

iv) The equation from iii because it is easier to read as it contains fewer zeros.

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

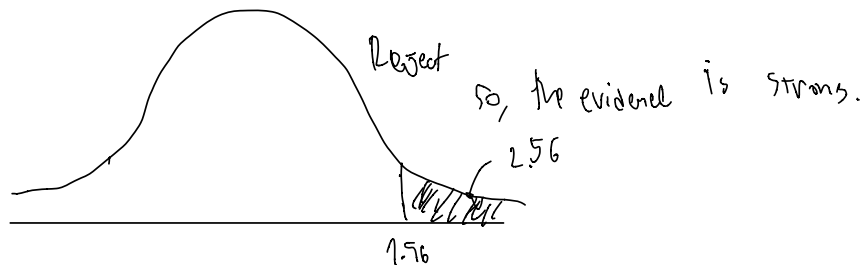
$$\begin{aligned}\widehat{sleep} = & 3,840.83 - .163 \text{ totwrk} - 11.71 \text{ educ} - 8.70 \text{ age} \\ & (235.11) \quad (.018) \quad (5.86) \quad (11.21) \\ & + .128 \text{ age}^2 + 87.75 \text{ male} \\ & (.134) \quad (34.33) \\ n = & 706, R^2 = .123, \bar{R}^2 = .117.\end{aligned}$$

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 on male is 87.75 so, men is estimated to sleep more per week compare to woman.

$$t_{\text{male}} = \frac{87.75}{34.33} \approx 2.56 \quad \text{at } 95\% \text{ CI } t_{\text{crit}} = 1.96$$



$$ii) t_{\text{totwrk}} = \frac{-1.63}{2.914} \approx -0.56 \quad \text{is statistically significant}$$

the coefficient implies that one more hour of work is .163(60) $\approx$ 9.8 minute less sleep.

iii) the null hypothesis is we're testing $H_0: \beta_3 = \beta_4 = 0$

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

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 if 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?

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

$$\text{To test } H_0: \beta_5 = 0$$

$$H_a: \beta_5 \neq 0$$

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

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

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

perform F test which $Q=3$ $df = k - b - 1$

v) respondents may not accurately report their marijuana

usage out of fear of suspicion 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.

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

(2.00) (0.70)

$n = 856, R^2 = .329, \bar{R}^2 = .328.$

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

(2.04) (0.74) (0.69)

$n = 856, R^2 = .349, \bar{R}^2 = .348.$

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

(2.86) (3.96) (0.98) (1.383)

$n = 856, R^2 = .349, \bar{R}^2 = .347.$

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

(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 increase in score by 3.83 when 1 more male is added
 95% confidence interval = $3.83 \pm 1.96(0.74)$
 0 is excluded because the interval is
 b/w (2.379, 5.2804)

ii In equation 3 we have an interaction term among the variables so the estimate on male has a higher s.e.

F-testing

$$H_0: \beta_1 = \beta_2 = 0 \text{ in Equation 3}$$

$$H_1: \text{otherwise}$$

$$F = \frac{(0.349 - 0.329)/2}{\frac{0.049}{852}} = 13.09$$

$$F_{(.05, 852)} = 3.006$$

So, we reject H_0

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

C4. Use the data in GPA2 for this exercise.

i. Consider the equation

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

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.

- i) . β_3 are definitely less than zero b/c high school percentile is defined so that the smaller the number the better the student do.
- . $\beta_4 > 0$ because SAT scores cannot be negative
 - . others coefficients are unclear.

ii)

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

Source	SS	df	MS	Number of obs	=	4,137
Model	524.819305	6	87.4698842	F(6, 4130)	=	284.59
Residual	1269.37637	4,130	.307355053	Prob > F	=	0.0000
Total	1794.19567	4,136	.433799728	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
hsize^2	.0046754	.0022494	2.08	0.038	.0002654 .0090854
hsperc	-.0132126	.0005708	-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

$$\hat{\text{colgpa}} = 1.241 - 0.0569 \text{hsize} + .00467 \text{hsize}^2 - .0132 \text{hsperc} + .00165 \text{sat}$$

(0.079) (0.0164) (0.00225) (0.0006) (0.00007)

$$+ 0.155 \text{female} + 0.169 \text{athlete}$$

(0.016) (0.042)

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

- An athlete is predicted to have a GPA ≈ 0.169 points higher than non athlete ceteris paribus. the t stat $\frac{0.169-0}{0.042} \approx 4.02$ is very significant.

(ii)

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

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

colgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hsize	-.0534038	.0175092	-3.05	0.002	-.0877313 -.0190763
hsizeeq	.0053228	.0024086	2.21	0.027	.0004607 .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 as significant as part (ii) b/c we don't control SAT scores.

(iv)

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

Source	SS	df	MS	Number of obs	=	4,137
Model	524.821272	7	74.9744674	F(7, 4129)	=	245.88
Residual	1269.3744	4,129	.307429015	Prob > F	=	0.0000
Total	1794.19567	4,136	.433799728	R-squared	=	0.2925
				Adj R-squared	=	0.2913
				Root MSE	=	.55446

colgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hsize	-.0568006	.0163671	-3.47	0.001	-.0888889 -.0247124
hsizeeq	.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

to test the hypothesis we choose female non athlete as a baseline

$$H_0: \beta_1 = 0$$

$$t_{.025, 4129} = 1.96$$

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

$\therefore$ So, we reject H_0

