6104640260

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?

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

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 $salesbil^2 = sales^2/(1{,}000)^2$.]

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

3. i) $\frac{d\,rdintens}{d\,sales} = 0.00030 - 0.00000014\,sales$

$0 = 0.00030 - 0.00000014\,sales$

$-0.00030 = -0.00000014\,sales$

$21428.57 = sales$

Marginal effect of sales on rdintens starts to become negative when sales = 21428.57

ii) $t = \frac{\hat{\beta}_2 - a_j}{s.e.\ \hat{\beta}_2}$

$= \frac{-0.0000000070 - 0}{0.0000000057}$

$= -1.89$

In this case, $|t| < 1.96$ which means that $sales^2$ is not a significant variable at 5% level of significance. We can choose to not keep the quadratic term.

iii) $\widehat{rdintens} = 2.613 + 0.0003\,sales - 0.000000007\,sales^2$

$salesbil = \frac{sales}{1000}$	$salesbil^2 = \frac{sales^2}{1000^2}$
$salesbil(1000) = sales$	$salesbil^2(1000) = sales^2$

$\widehat{rdintens} = 2.613 + 0.0003\,salesbil(1000) - 0.000000007\,salesbil^2(1000)^2$

$= 2.613 + 0.3\,salesbil - 0.007\,salesbil^2$

$s.e.(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{\sum (X_{ij} - \bar{X}_j)^2 (1 - R_j^2)}} = \sqrt{\frac{\sigma^2}{\sum (SST)(1 - R^2)}}$

Changing from sales to salesbil will have no effect on R^2 but will have an effect on the SST_{sales}, SST_{sales^2} and standard errors.

iv) The equation in part iii because it is easier to read with fewer zeros.

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

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

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.

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

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

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?

1) i. In this case, there is a strong evidence that men sleep more than women because the coefficient on male is 87.75. Therefore, being a male increases expected total sleeping minutes by 87.75

ii) The coefficient is −0.163 (trade off) meaning that when total weekly minutes spent working increases by 1, the expected sleeping time decreases by 0.163

The t-statistic on totwork is $\left|\frac{-0.163}{0.018}\right| = 9.06$ which is greater than 1.96, totwork has a significant impact on $\widehat{sleep}$ at 5% level

iii) We need to test that coefficients on age and age2 are jointly zero using F-test. $H_0: \beta_{age} = \beta_{age^2} = 0$

$$F = \frac{R^2_{ur} - R^2_r}{q} \Big/ \frac{1 - R^2_{ur}}{n-k-1}$$

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?"

   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) $\log(wage) = \beta_0 + \beta_1 education + \beta_2 experience + \delta_3 female + \beta_4 marijuana\_usage + u$

ii) $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \delta_3 female + \beta_4 Marijuana\_usage + \beta_5 female.marijuana\_usage + u$

Test hypothesis $H_0: \beta_5 = 0$ against $H_1: \beta_5 \neq 0$

iii) $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \delta_3 female + \delta_4 light + \delta_5 moderate + \delta_6 heavy + u$

Non-user is the omitted variable

iv) $H_0: \delta_4 = \delta_5 = \delta_6 = 0$

Using F-test $F = \dfrac{R^2_{ur} - R^2_r}{q} \Big/ \dfrac{1 - R^2_{ur}}{n-k-1}$ with $q = 3$ and $k = 6$

→ df for numerator $N_1 = 3$

df for denominator $N_2 = n-k-1 = n-6-1 = n-7$

We would be obtaining a critical value from the $F_{q,n-7}$ distribution

v) There could be some important variables omitted which determine marijuana usage and wage

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.

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

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

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

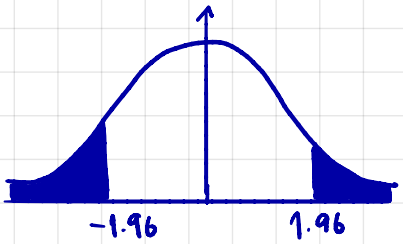
$$\widehat{score} = 30.36 + 3.82\, male + 14.33\, colgpa + 0.479\, male \cdot (colgpa - 2.81)$$
$$(2.86) \quad (0.74) \quad (0.98) \quad (1.383)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .347.$$

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) The coefficient on male is 3.83 which means that the expected difference in class score between male and female is 3.83



The 95% confidence interval for β_{male}

$= [\hat{\beta}_j - 1.96\, s.e.(\hat{\beta}_j),\ \hat{\beta}_j + 1.96\, s.e.(\hat{\beta}_j)]$

$= 3.83 - 1.96(0.74),\ 3.83 + 1.96(0.74)$

$\approx 2.3796,\ 5.2804$

The confidence interval excludes 0.

ii) $H_0: \beta_{male} = 0$, $H_0: \beta_{male.colgpa} = 0$

$$F = \frac{R^2_{ur} - R^2_r}{q} \Big/ \frac{1 - R^2_{ur}}{n-k-1} = \frac{0.348 - 0.329}{2} \Big/ \frac{1 - 0.348}{856 - 3 - 1}$$

$= 13.09$

$F_{2,854} = 3.01$, since $13.09 > 3.01$, we can reject the H_0 that there is no gender difference in score at 5% sig level

C4. Use the data in GPA2 for this exercise.

i. Consider the equation

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

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) I expect the coefficients on hsperc to be negative while the coefficients on sat, female to be positive. I'm unsure about the rest

ii)

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

Source	SS	df	MS		
Model	524.819305	6	87.4698842	Number of obs	4,137
Residual	1269.37637	4,130	.307355053	F(6, 4130)	284.59
				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
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

- The expected GPA differentials between athletes and non atheletes is 0.1693. The t-statistic is greater than 1.96 which shows that athlete or the difference between the GPA of the athlete and non athlete is statistically significant at 5% level of significance
- Also, the p-value of the coefficient is 0.000 which is less than the critical p-value of 0.05 at 5% level of significance.

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
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
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 estimate effect of being an athlete is given by the coefficient of athlete which is 0.005449.

Now, the expected difference in college GPA between athletes and non-athletes is only 0.005 and the t-statistics shows that it is not a statistically significant variable.

iv)

```
gen male = female==0

gen nonathlete = athlete == 0

gen maleathlete = male*athlete

gen malenonathlete = male*nonathlete
```

```
. reg colgpa hsize hsizesq hsperc sat femaleathlete maleathlete malenonathlete
```

Source	SS	df	MS	Number of obs	=	4,137
				F(7, 4129)	=	243.88
Model	524.821272	7	74.9744674	Prob > F	=	0.0000
Residual	1269.3744	4,129	.307429015	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
femaleathlete	.1751106	.0840258	2.08	0.037	.0103748	.3398464
maleathlete	.0128034	.0487395	0.26	0.793	-.0827523	.1083591
malenonathlete	-.1546151	.0183122	-8.44	0.000	-.1905168	-.1187133
_cons	1.39619	.0755581	18.48	0.000	1.248055	1.544324

In this case, the coefficient on female athlete is 0.175 which means that the expected difference of female athletes and female non-athletes is 0.175. The p-value of coefficient is 0.037 which is less than the critical p-value of 0.05 at 5% level of significance. Therefore, there is a significant difference in college GPA between female athletes and female non-athletes.

v)

```
. gen femalesat = female * sat

. reg colgpa hsize hsizesq hsperc sat femaleathlete maleathlete malenonathlete femalesat
```

Source	SS	df	MS	Number of obs	=	4,137
				F(8, 4128)	=	213.37
Model	524.873728	8	65.609216	Prob > F	=	0.0000
Residual	1269.32195	4,128	.307490781	R-squared	=	0.2925
				Adj R-squared	=	0.2912
Total	1794.19567	4,136	.433799728	Root MSE	=	.55452

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0568198	.0163688	-3.47	0.001	-.0889114	-.0247282
hsizesq	.0046773	.002251	2.08	0.038	.0002641	.0090904
hsperc	-.0132236	.0005738	-23.04	0.000	-.0143487	-.0120986
sat	.001624	.0000858	18.93	0.000	.0014558	.0017922
femaleathlete	.1779989	.0843247	2.11	0.035	.0126771	.3433207
maleathlete	.0652958	.1361172	0.48	0.631	-.2015674	.3321589
malenonathlete	-.0990198	.1358427	-0.73	0.466	-.3653447	.1673051
femalesat	.0000539	.0001306	0.41	0.680	-.0002021	.00031
_cons	1.364334	.1079746	12.64	0.000	1.152646	1.576023

- The p-value of coefficient femalesat is 0.680 which is more than the critical p-value of 0.05 at 5% level of significance. Therefore, there is no statistically significant difference of effect of sat on colgpa by gender.