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

pavita kriathkungwakai
6104640674

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

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?

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

when the marginal effect of sales become negative

$$\frac{\partial \widehat{rdintens}}{\partial sales} < 0$$

$$0.0003 - 0.000000014\, sales < 0$$

$$0.000000014\, sales > 0.0003$$
$$sales > 21{,}428.5714$$

∴ At sales equal or lower than 21,428.74 the marginal effect of sales become negative

(ii) from this equation it is quadratic. So, we test if $\beta_2 < 0$ we will keep the quadratic term

$t_{stat} = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{-0.000000007}{0.0000000037} = -1.89$, which is significant at the 5% level of significant

(iii) $\widehat{rdintens} = 2.613 + 0.3\, salesbil - 0.007\, salesbil^2$
$(0.429)\quad (0.14)\qquad (0.0037)$
$n = 32$, $R^2 = 0.1484$

(iv) Actually, 2 equations are identical, but equation in (iii) is easier to read due to fewer zeroes to the right of decimal.

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?

(i) The coefficient of male is 87.75 male so, man is estimated to sleep more per week comparable to women.

$t_{stat} = \frac{87.75}{34.33} \approx 2.56$, which is closing to 1% critical value (2.58) means that the evidence is strong.

(ii) $t_{totwork} = \frac{-0.163}{0.018} \approx -9.06$ » statistic significant

The coefficient implies that to spend one more hour of working he has to trade-off $-(0.163)(60) = -9.78$ minutes of sleeping.

(iii) the null hypothesis we're testing

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

Now run restricted version of the regression where age, age^2 are omitted by calculating

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

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

(ii) $\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \delta_1 gender + \delta_2 gender \cdot usage + u$

and to test whether there are different in the effect of drug usage for men and women

$H_0: \delta_2 = 0$

$H_a: \delta_2 \neq 0$

(iii) $\log(wage) = \beta_0 + \beta_1 light + \beta_2 moderate + \beta_3 heavy + \beta_4 edu + \beta_5 exper + \beta_6 gender + u$

(iv) the null hypothesis is

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

we can test the null hypothesis by using F-test

The numerator degrees of freedom is $q = 3$

The denominator degrees of freedom is $n-6-1$

(v) Sometimes the data may be not accurate. Because people might not be truthful about their marijuana 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.

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

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

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

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

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) from equation 2 $\widehat{score} = 29.66 + 3.83 male + 14.57\, colgpa$
$(2.04) \quad (0.74) \quad (0.69)$

the coefficient of male is 3.83 so, increasing one more male score will increase by 3.83

$t_{stat} = \frac{3.83}{0.74} \approx 5.18$, which is strongly significant.

(ii) unrestricted model $\widehat{score} = 29.66 + 3.83 male + 14.57\, colga$
$(2.04) \quad (0.74) \quad (0.69)$

restricted model $\widehat{score} = 32.31 + 14.32\, colgpa$
$(2.00) \quad (0.7)$

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)}$$
$$= \frac{(0.348 - 0.328)/2}{(1-0.348)/(856-3-1)}$$
$$= \frac{0.02}{0.00076} \approx 26.3$$

, which is very significant. we can implies that male has an effect on score

(iii) Because in equation 4 variable male (colga-2.81) has subtracted by the mean of colga (2.81) making it closer to 0 and more precious OLS

C4. Use the data in GPA2 for this exercise.

i. Consider the equation

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

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 denifinitely less than zero because highschool percentile is defined so that the smaller the number the better the student do.

• $\beta_4 > 0$ because SAT scores cannot be negative

• other coefficient are unclear.

(ii)

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

Source	SS	df	MS	Number of obs	=	4,137
				F(6, 4130)	=	284.59
Model	524.819305	6	87.4698842	Prob > F	=	0.0000
Residual	1269.37637	4,130	.307355053	R-squared	=	0.2925
				Adj R-squared	=	0.2915
Total	1794.19567	4,136	.433799728	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

$$\widehat{colgpa} = 1.241 - 0.569 hsize + 0.00468\, hsize^2 - 0.0132\, hspec + 0.00165\, SAT + 0.155\, female + 0.169\, athlete$$
$$(0.079)\quad (0.0164)\quad (0.00225)\quad (0.0006)\quad (0.00007)\quad (0.018)\quad (0.042)$$

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

• An athlete is predicted to have a GPA ≈ 1.69 points higher than non athlete CETERIS PARIBUS. The $t_{stat} = \frac{0.169 - 0}{0.042} \approx 4.02$ is very significant.

(iii)

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

Source	SS	df	MS	Number of obs	=	4,137
				F(5, 4131)	=	191.92
Model	338.217123	5	67.6434247	Prob > F	=	0.0000
Residual	1455.97855	4,131	.35245184	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 becomes ≈ 0.0054 which is not as significant as part (ii) because we don't control SAT scores.

(iv)

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

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
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 nonathlete as a basegroup.

So, we add female sat to the equation in (ii) its coefficient is about 0.000051 and $t_{stat} \approx 0.4$ there is little evidence that SAT scores differs by gender.

$H_0: \delta_1 = 0$

$t_{0.025,4129} = 1.96$

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

So, we reject H_0