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

6104641268

Lalita T.

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

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

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

$$sales = \frac{0.0003}{0.000000014}$$

$$sales = 21{,}428.57$$

$\therefore$ the marginal effect of sales on rdintens become negative when sales = 21,428.57

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

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

This is significant against alternative $H_0: \beta < 0$ at 5% level of sig ($CV \approx 1.7$, d.f. 29)

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

$$\widehat{rdintens} = 2.613 + \underbrace{0.3}_{0.0003 \times 1{,}000}\, salesbil - \underbrace{0.007}_{0.000000007 \times (1000)^2}\, salesbil^2$$

$$(0.429) \quad (0.00014) \quad (0.0000000037)$$

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

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

The equation in (iii) interprets clear mathematic explanation more than the original equation with different scale of sales (measure in billion of dollars)

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

$$\widehat{sleep} = 3{,}840.83 - \underset{\beta_1}{.163}\, totwrk - \underset{\beta_2}{11.71}\, educ - \underset{\beta_3}{8.70}\, age$$
$$(235.11)\quad(.018)\qquad(5.86)\qquad(11.21)$$
$$+ \underset{\beta_4}{.128}\, age^2 + \underset{\beta_5}{87.75}\, male$$
$$(.134)\qquad(34.33)$$
$$n = 706,\ R^2 = .123,\ \bar{R}^2 = .117.$$

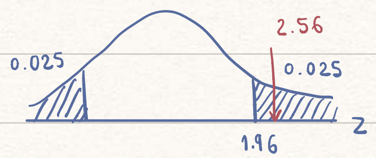
The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

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

i) The coefficient on men is 87.75

$$t_{male} = \frac{\hat{\beta}_5 - 0}{s.e\ \hat{\beta}_5} = \frac{87.75}{34.33} \approx 2.56$$

$n = 706$, df. = 0.05



∴ 2.56 falls in to the rejection region ,so we reject H_0 at 5% of significant.

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

ii) $t_{totwork} = \frac{\hat{\beta}_1 - 0}{s.e\ \hat{\beta}_1} = -\frac{0.163}{0.181} \approx -9.06$ falls into the rejection region, so we reject $H_0 : \hat{\beta}_1 \neq 0$ at 5% level of significant, therefore there is a significantly significant tradeoff between working and sleeping

Coefficient β_1 = -0.163 means that 1 minute spent working will decrease your sleep for 0.163 minute per week

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?

test hull hypothesis where $H_0 = \beta_4 = \beta_5 = 0$ and run the restricted version of regression where age, age² are omitted by calculating

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

number of H_0

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%."

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

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?

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

to test $H_0 : \beta_5 \neq 0$

$H_a : \beta_5 = 0$

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.

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

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

non-user is the omitted category.

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.

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

Perform F test which $q = 3$ and $d.f. = n-k-1 = n-6-1$

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

(v) The responses may not be accurate report because the respondents may aware of the legal repurcussion 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.

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

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

ur $$\widehat{score} = \underset{(2.86)}{30.36} + \underset{(3.96)}{\overset{\beta_1}{2.47}}\, male + \underset{(0.98)}{14.33}\, colgpa + \underset{(1.383)}{\overset{\beta_3}{0.479}}\, male \cdot colgpa$$
$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .347.$$

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

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?

(i) When the score increase by 3.83, 1 more male is added

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

zero is excluded because the interval is between $(2.379, 5.2804)$

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.]

(ii) In equation 3, we have an interaction term amoung variable, so the estimate on male has higher standard error.

Compute the F-test

$H_0: \beta_1 = \beta_3 = 0$ in equation 3

$H_a:$ otherwise

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

We reject H_0 and gender differences are significant

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?

(iii) Because in equation 4 variable $male \times (colgpa - 2.81)$ has been subtract 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

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

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?

(i) - β_3 is definitely less than zero because the smaller number school graduating class, the better the student's score.

- $\beta_4 > 0$ because SAT score cannot be negative

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 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
------------------------------------------------------------------------------
```

$$\widehat{colgpa} = \underset{(0.079)}{1.241} - \underset{(0.0164)}{0.569}\, hsize + \underset{(0.00225)}{0.00468}\, hsize^2 - \underset{(0.0006)}{0.0132}\, hsperc + \underset{(0.00007)}{0.00165}\, sat$$

$$+ \underset{(0.018)}{0.155}\, female + \underset{(0.042)}{0.169}\, athlete$$

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

- An althlete is predicted to have a GPA ≈ 0.169 point higher than non-athlete ceteris paribus. The t state $= \frac{0.169 - 0}{0.042} \approx 4.02$ 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 hsizesq hsperc female athlete
```

Source	SS	df	MS
Model	338.217123	5	67.6434247
Residual	1455.97855	4,131	.35245184
Total	1794.19567	4,136	.433799728

Number of obs	=	4,137
F(5, 4131)	=	191.92
Prob > F	=	0.0000
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 coefficient on athlete becomes ≈ 0.0054 which is not as significant as part (ii) because we donot 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.

To test the hypothesis, we choose female nonalthlete as a basegroup

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

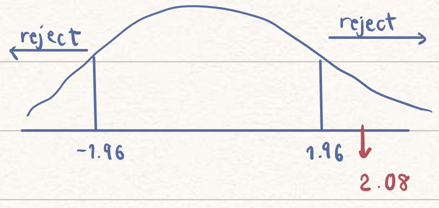
Source	SS	df	MS
Model	524.821272	7	74.9744674
Residual	1269.3744	4,129	.307429015
Total	1794.19567	4,136	.433799728

Number of obs	=	4,137
F(7, 4129)	=	243.88
Prob > F	=	0.0000
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
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

$H_0 : \delta_1 = 0$

$H_a :$ otherwise



$t_{0.25,\,4129} = 1.96$

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

$\therefore$ We reject $H_0 : \delta_1 = 0$ at 5% level of significant

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

```
. 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
------------------------------------------------------------------------------
```

We add female · sat on the equation in (ii), its coefficient is about 0.000051 and t state ≈ 0.4, there is a little evidence that sat score differs by gender.