HOMEWORK 6

Pattarit Yamploy
6104640393

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

df. = 29

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

i.) $\frac{\partial\, rd\ intens}{\partial\, sales} = 0.0003 - 0.00000014\ sales = 0$

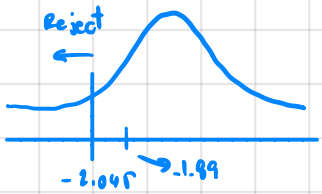
$sales = \frac{0.0003}{0.00000014}$

$sales = 21{,}428.57 \rightarrow$ Negative at 21,428.57

ii) Yes, because $t_{stat} = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{-0.0000000007}{0.0000000037} = -1.89$

So, at 95% conf. interval, df. = 29, $t_{cv} = 2.045 \rightarrow |-1.89| < 2.045$

∴ No, because quadratic term has no significant impact.



iii) $\widehat{rd\ intens} = \underset{(0.429)}{2.613} + 0.030\ salesbil - \underset{(0.0037)}{0.0040}\ salesbil^2$

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

iv.) The equation from (iii) is easier to read because of fewer zeros in the equation so, these two equations are the same but the scale is different.

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?

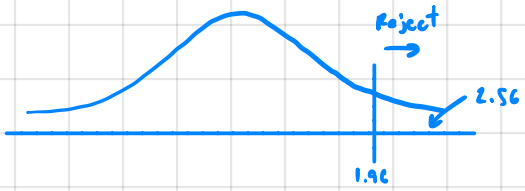
   ii. Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

   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 on male is 87.75 so, man is estimated to sleep more per week compared to women.

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

at 95% conf. interval $t_{crit} = 1.96$



∴ t_{male} is fall in the rejection area $2.56 > 1.96$ which means it has significant impact on sleep time so, it is strong evidence.

ii) $t_{totwork} = -\frac{0.163}{0.018} \approx -9.06$ is statistically significant.

The coefficient implies that one more hour of work (60 mins) is $0.163 \approx 9.8$ minutes less sleep.

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

$H_1 : \hat{\beta}_3 \neq \hat{\beta}_4 \neq 0$

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

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

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

Test $H_0: \beta_5 = 0$

$H_a: \beta_5 \neq 0$

iii) Assuming that there's no interact 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) The null hypothesis is $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

Perform F-test which $q = 3$ $df. = n - 6 - 1$

v.) Respondants may not accurately report their marijuana usage out of fear of legal repurcussion or there may be 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.

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

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

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

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

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.) We can interprete from the coefficient on male that the increase in score by 3.83 when 1 more male is added.

confidence interval at 95% = $3.83 \pm 1.96(0.74)$ ∴ The interval is between $(2.379, 5.2804)$ so, 0 is excluded.

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

Do the F-test, $H_0: \beta_1 = \beta_2 = 0$ in Equ 3

H_1 : otherwise

$$F = \frac{(0.349 - 0.329)/2}{(1 - 0.349)/853} = 13.08$$

∴ Reject H_0 So, the impact of gender on colgpa would be difference.

$F_{0.05, 853} = 3.006$

iii) b/c in equation 4 variable male * (colgpa - 2.81) has been subtract by the mean of colgpa (2.81) making it closer to zero and more precise OLS.

C.4 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?

The two signs that are clear are $\beta_3 < 0$ because hsperc is defined so that the smaller the number the better the student and $\beta_4 > 0$ because SAT score cannot be less than zero. Other coefficients 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 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
				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

$\widehat{colgpa} = \underset{(0.079)}{1.241} - \underset{(0.016)}{0.0569}\, hsize + \underset{(0.0022)}{0.0046}\, hsize^2 - \underset{(0.0006)}{0.0132}\, hsperc + \underset{(0.00007)}{0.0016}\, sat + \underset{(0.018)}{0.155}\, female + \underset{(0.42)}{0.169}\, athlete$

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

∴ Holding other factors fixed, an athlete is predicted to have a GPA about 0.169 points higher than a nonathlete. The $t_{stat} = 0.169/0.042 = 4.02$ which is very significant.

iii. Drop *sat* from the model and re-estimate 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	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

When drop sat from the model, we can see that the coefficient on athletes become about 0.0054, which is not statistically significant.

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 hsizesq hsperc sat femath maleath malenonath
```

Source	SS	df	MS		
Model	524.821272	7	74.9744674	Number of obs	= 4,137
Residual	1269.3744	4,129	.307429015	F(7, 4129)	= 243.88
				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
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

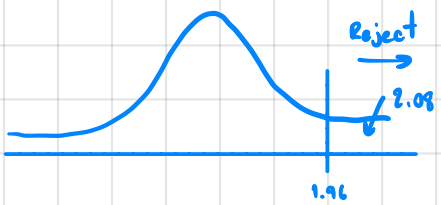
To test the hypothesis, we should choose one of there as a base group.
In this, case we choose female non athletes as a base group.

H_0 : $\delta_0 = 0$

H_1 : $\delta_0 \neq 0$

$t_{critical} = 1.96$

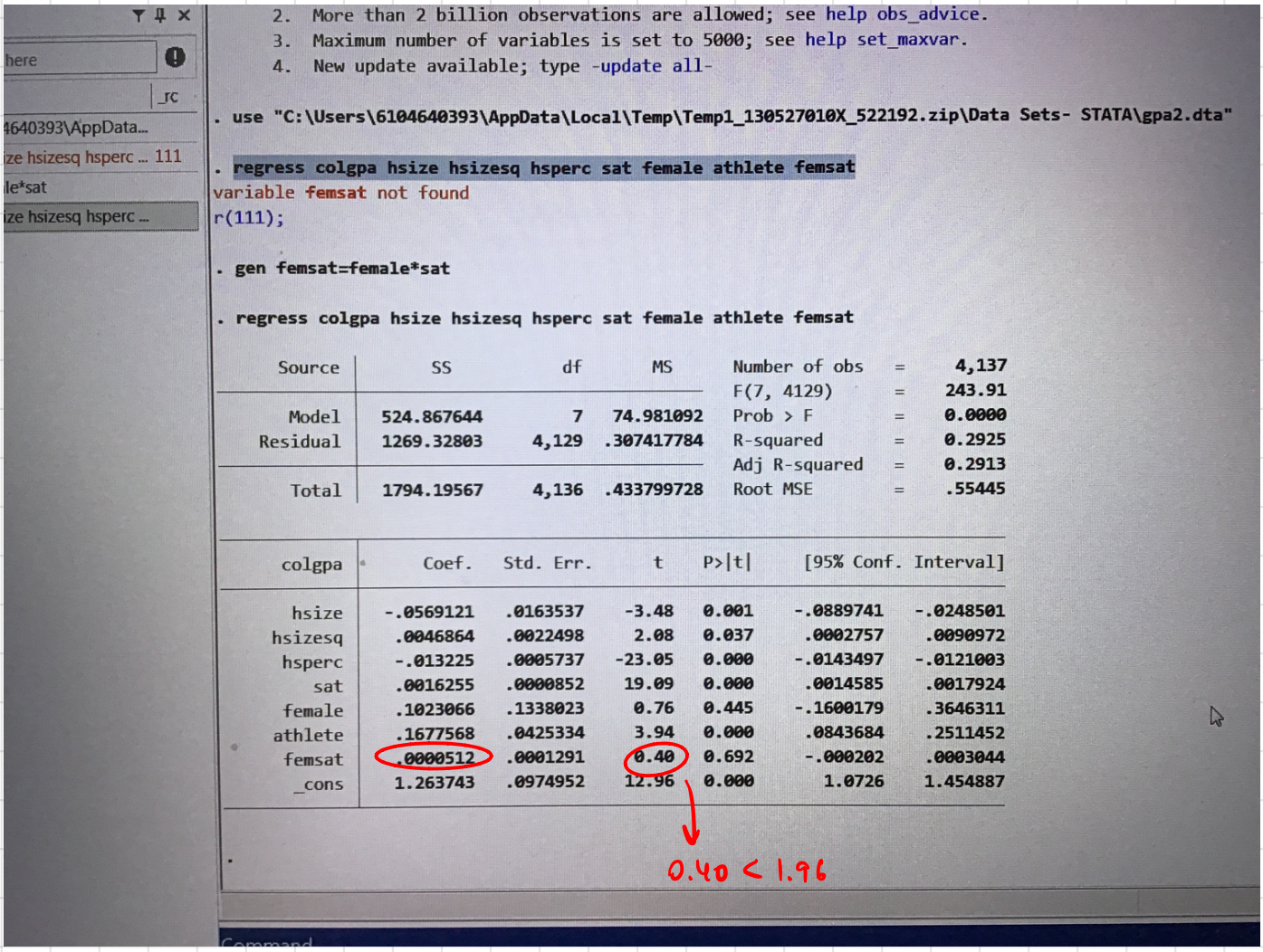
$t_{stat} = \frac{0.175}{0.084} \approx 2.09$



∴ Reject H_0 so, there is diff. btw womenathletes and women-nonathletes.

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

Whether we add the interaction female · sat to the equation in part (ii) or part (iv.), the outcome is partially the same. For example, when female · sat is added to the equation in part (ii), it coefficient is about 0.000051 and its t-statistic is about 0.40. There is very little evidence that the effect of sat differ by gender.



```
      2.  More than 2 billion observations are allowed; see help obs_advice.
      3.  Maximum number of variables is set to 5000; see help set_maxvar.
      4.  New update available; type -update all-

. use "C:\Users\6104640393\AppData\Local\Temp\Temp1_130527010X_522192.zip\Data Sets- STATA\gpa2.dta"

. regress colgpa hsize hsizesq hsperc sat female athlete femsat
variable femsat not found
r(111);

. gen femsat=female*sat

. regress colgpa hsize hsizesq hsperc sat female athlete femsat

      Source |       SS       df       MS              Number of obs =     4,137
-------------+------------------------------           F(7, 4129)    =    243.91
       Model |  524.867644     7  74.981092            Prob > F      =    0.0000
    Residual |  1269.32803 4,129  .307417784           R-squared     =    0.2925
-------------+------------------------------           Adj R-squared =    0.2913
       Total |  1794.19567 4,136  .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
------------------------------------------------------------------------------

.
```

0.40 < 1.96