## Problem Set 1

1. Let $\beta_0, \beta_1, \ldots, \beta_k$ be the OLS estimates from the regression of $y_i$ on $x_{i1}, x_{i2}, \ldots,\ x_{ik},\ i = 1, 2, \ldots n,$ . For nonzero constants $c_1, ..., c_k$, argue that the OLS intercept and slopes from the regression of $c_0 y_i$ on $c_1 x_{i1}, ..., c_k x_{ik}$ $i = 1, 2, ..., n$, are given by $\tilde{\beta}_j = c_0 \tilde{\beta}_0, \tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1$ and so on. [Hint: Use the fact that the $\hat{\beta}_j$ solve the first order conditions in (3.13), and the $\tilde{\beta}_j$ must solve the first order conditions involving the rescaled dependent and independent variables.] (Wooldridge chapter 6, question 2)

2. Using data in RDCHEM, the following equation was estimated by OLS:

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u$$

   where, rdintens is the amount spent on research by a pharma firm as percentage of sales. The R output is shown below:

```
> summary(lm(rdintens ~ sales + salessq, rdchem))

Call:
lm(formula = rdintens ~ sales + salessq, data = rdchem)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1418 -1.3630 -0.2257  1.0688  5.5808 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.613e+00  4.294e-01   6.084 1.27e-06 ***
sales        3.006e-04  1.393e-04   2.158   0.0394 *  
salessq     -6.946e-09  3.726e-09  -1.864   0.0725 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.788 on 29 degrees of freedom
Multiple R-squared:  0.1484,	Adjusted R-squared:  0.08969 
F-statistic: 2.527 on 2 and 29 DF,  p-value: 0.09733
```

(a) At what point does the marginal effect of sales on rdintens become negative?

(b) Would you keep the quardratic term in the model? Explain?

(c) 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 $salesbil^2 = sales^2/(1000)^2$.]

(d) Consider the new model which tries to explain the number of research interns hited

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \beta_3 profits + u$$

Interpret the estimates and explain the following results.

```
> summary(lm(rdintens ~ salesbil + I(salesbil * salesbil) + profits, rdchem))

Call:
lm(formula = rdintens ~ salesbil + I(salesbil * salesbil) + profits,
    data = rdchem)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8845 -1.3763 -0.4077  1.0182  5.7782

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.6372502  0.4413061   5.976 1.95e-06 ***
salesbil               0.2212773  0.2601408   0.851   0.4022
I(salesbil * salesbil) -0.0070063  0.0037868  -1.850   0.0749 .
profits                0.0007573  0.0020852   0.363   0.7192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.815 on 28 degrees of freedom
Multiple R-squared:  0.1524,    Adjusted R-squared:  0.0616
F-statistic: 1.678 on 3 and 28 DF,  p-value: 0.1943
```

3. The following three equations were estimated using 1534 observations in 401K.

$$
\begin{aligned}
prate =& \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 totemp + u \\
prate =& \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 log(totemp) + u \\
prate =& \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 totemp + \beta_4 totemp^2 + u
\end{aligned}
$$

(a) This example is introduced in chapter 3. Interpret the coefficients for the first model.

(b) Which of these models would you prefer and why?

```
> summary(lm(prate ~ mrate + age + totemp, k401k))

Call:
lm(formula = prate ~ mrate + age + totemp, data = k401k)

Residuals:
    Min      1Q  Median      3Q     Max 
-77.698  -8.074   4.716  12.505  30.307 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.029e+01  7.777e-01 103.242  < 2e-16 ***
mrate        5.442e+00  5.244e-01  10.378  < 2e-16 ***
age          2.692e-01  4.514e-02   5.963 3.07e-09 ***
totemp      -1.291e-04  3.666e-05  -3.521 0.000443 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.88 on 1530 degrees of freedom
Multiple R-squared:  0.09954,	Adjusted R-squared:  0.09778 
F-statistic: 56.38 on 3 and 1530 DF,  p-value: < 2.2e-16
```

```
> summary(lm(prate ~ mrate + age + ltotemp, k401k))

Call:
lm(formula = prate ~ mrate + age + ltotemp, data = k401k)

Residuals:
    Min      1Q  Median      3Q     Max 
-74.957  -7.490   4.201  11.354  26.374 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 97.32036    1.94629  50.003  < 2e-16 ***
mrate        5.01554    0.51363   9.765  < 2e-16 ***
age          0.31361    0.04404   7.120 1.65e-12 ***
ltotemp     -2.65631    0.27690  -9.593  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.48 on 1530 degrees of freedom
Multiple R-squared:  0.1437,    Adjusted R-squared:  0.1421 
F-statistic: 85.62 on 3 and 1530 DF,  p-value: < 2.2e-16
```

```
> summary(lm(prate ~ mrate + age + totemp + I(totemp * totemp), k401k))

Call:
lm(formula = prate ~ mrate + age + totemp + I(totemp * totemp),
    data = k401k)

Residuals:
    Min      1Q  Median      3Q     Max
-77.619  -8.099   4.622  12.219  24.721

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.062e+01  7.787e-01 103.529  < 2e-16 ***
mrate               5.342e+00  5.226e-01  10.222  < 2e-16 ***
age                 2.899e-01  4.525e-02   6.406 1.98e-10 ***
totemp             -4.297e-04  8.550e-05  -5.026 5.59e-07 ***
I(totemp * totemp)  3.939e-09  1.013e-09   3.888 0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.81 on 1529 degrees of freedom
Multiple R-squared:  0.1084,    Adjusted R-squared:  0.106
F-statistic: 46.45 on 4 and 1529 DF,  p-value: < 2.2e-16
```

4. Suppose we want to estimate the effects of alcohol consumption ($alcohol$) on college grade point average ($colGPA$). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called $attend$). A standardized test score (say, $SAT$) and high school GPA ($hsGPA$) are also available. (Q8, Chapter 6, JW)

   (a) Should we include $attend$ along with $alcohol$ as explanatory variables in a multiple regression model? (Think about how you would interpret $\beta_{alcohol}$.)

   (b) Should $SAT$ and $hsGPA$ be included as explanatory variables? Explain

5. The following two equations were estimated using the data in MEAPSINGLE. The key explanatory variable is *lexppp*, the log of expenditures per student at the school level.

$$math4 = \beta_0 + \beta_1 lexpp + \beta_2 free + \beta_3 lmedinc + \beta_4 pctsgle + u$$
$$math4 = \beta_0 + \beta_1 lexpp + \beta_2 free + \beta_3 lmedinc + \beta_4 pctsgle + \beta_5 read4$$

   (a) If you are a policy maker trying to estimate the causal effect of per-student spending on math test performance, explain why the first equation is more relevant than the second.

(b) What is the estimated effect of a 10% increase in expenditures per student? Interpret the following results:

```
> summary(lm(math4 ~ lexppp + free + lmedinc + pctsgle, meapsingle))

Call:
lm(formula = math4 ~ lexppp + free + lmedinc + pctsgle, data = meapsingle)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.259  -7.422   1.615   7.274  49.524 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.48949   59.23781   0.413   0.6797    
lexppp       9.00648    4.03530   2.232   0.0266 *  
free        -0.42164    0.07064  -5.969 9.27e-09 ***
lmedinc     -0.75221    5.35816  -0.140   0.8885    
pctsgle     -0.27444    0.16086  -1.706   0.0894 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.59 on 224 degrees of freedom
Multiple R-squared:  0.4716,    Adjusted R-squared:  0.4622 
F-statistic: 49.98 on 4 and 224 DF,  p-value: < 2.2e-16
```

(c) Does adding *read4* to the regression have strange effects on coefficients and statistical significance other than $\beta_{lexppp}$? Why or why not?

```
> summary(lm(math4 ~ lexppp + free + lmedinc + pctsgle + read4, meapsingle))

Call:
lm(formula = math4 ~ lexppp + free + lmedinc + pctsgle + read4,
    data = meapsingle)

Residuals:
     Min       1Q   Median       3Q      Max
-29.5690  -4.6729  -0.0349   4.3644  24.8425

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.37870   41.70293   3.582 0.000419 ***
lexppp        1.93215    2.82480   0.684 0.494688
free         -0.06004    0.05399  -1.112 0.267297
lmedinc     -10.77595    3.75746  -2.868 0.004529 **
pctsgle      -0.39663    0.11143  -3.559 0.000454 ***
read4         0.66656    0.04249  15.687  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.012 on 223 degrees of freedom
Multiple R-squared:  0.7488,    Adjusted R-squared:  0.7432
F-statistic: 132.9 on 5 and 223 DF,  p-value: < 2.2e-16
```

(d) How would you explain to someone with only basic knowledge of regression why, in this case, you prefer the equation with the smaller adjusted R-squared?