1. Consider a multivariate regression model of the form \( y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i \). Write the 1st order conditions for the optimization problem where one is interested in minimizing the sum of squared errors \( SSE = \sum_{i=1}^{n} \epsilon_i^2 \).

Suppose in a sample if 25 observations, the following facts are presented about the model above.

\[
\sum_{i=1}^{n} x_{i1}^2 = 40 \quad \sum_{i=1}^{n} x_{i2}^2 = 80 \quad \sum_{i=1}^{n} x_{i1}x_{i2} = 0 \quad \sum_{i=1}^{n} x_{i1} = 0 \quad \sum_{i=1}^{n} x_{i2} = 0 \quad \sum_{i=1}^{n} y_i = 0
\]

\[
\sum_{i=1}^{n} x_{i1}y_i = 120 \quad \sum_{i=1}^{n} x_{i2}y_i = 160
\]

Using the first order conditions (or normal equations) and these facts, provide the estimates for \( \hat{\beta}_0, \hat{\beta}_1 \) and \( \hat{\beta}_2 \).

HINT: Solve for \( \hat{\beta}_0 \) first.

2. Download the data cps87.dta. Generate two new variables. The first is the natural log of weekly earnings. The second is age squared. Next, run a regression of the natural log of weekly earnings on age, age squared and years of education. We can write this model as

\[
\ln(\text{weekly earn}) = \beta_0 + \text{age}\beta_1 + \text{age}^2\beta_2 + \text{educ}\beta_3 + \epsilon_i
\]

Provide a mathematical expression that defined \( \frac{\partial \ln(\text{weekly earn})}{\partial \text{age}} \). Using the results from the regression, what is \( \frac{\partial \ln(\text{weekly earn})}{\partial \text{age}} \) at age 21? Age 35? Age 50?

3. Consider a multivariate regression model of the form \( y_i = \beta_0 + x_{i1}\beta_1 + \epsilon_i \). Suppose the R\(^2\) from this model is R\(_a\). True, False, or Uncertain and explain. The R\(^2\) can never fall below R\(_a\) when additional variables are added to the model? (Think of a special case where someone adds completely irrelevant variables to the model – what will happen to the R\(^2\)?)

4. On the class web page is a STATA data set called house_price.dta. It has data on 114 homes sold in 1998 in a small town in New England. The data set contains information on the sales price of the house (measured on thousands of dollars), the number of bedrooms, bathrooms, other rooms, square feet of living space and age of the home,

Download the data and initially estimate a regression with house prices as the outcome of interest and four covariates: age in years, # number of bedrooms, # of bath rooms, # of other rooms. Call this model 1.

a. Interpret the coefficient on age in years and # of bedrooms by providing a numeric example.

Now, estimate a second model and add to the original regression the square feet of living space. Call this model 2.
b. What happens to the coefficient on \# of rooms, \# of bedrooms and \# of other rooms in this new model compared to the previous one? Why have the coefficients on these three variables changed so dramatically?

c. Interpret the coefficient on square feet of living space.

Now estimate a third model with the same dependent variable but include only two covariates: age in years and square feet.

d. Compare the $R^2$ from this model and that in Model #2. Provide an intuitive explanation for why the difference is so small.

5. On the class web page is a data set law_school_1985.dta that has information about average starting salaries of graduates from 95 law schools in 1985. The data set has four other key variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cost</td>
<td>average annual tuition</td>
</tr>
<tr>
<td>lsat</td>
<td>average lsat of class of 1985</td>
</tr>
<tr>
<td>rank</td>
<td>rank of law school according to national survey, 1=best</td>
</tr>
<tr>
<td>age</td>
<td>age of law school in years</td>
</tr>
</tbody>
</table>

Load the data set into STATA, then construct two new variables:

- `gen lcost=ln(cost)`
- `gen lsalary=ln(salary)`

a) Regress lsalary on four variables: lcost, lsat, rank and age. Interpret the coefficient on lcost – provide a numeric example of the magnitude of the coefficient on this variable?

b) Interpret the coefficient on rank -- provide a numeric example of the magnitude of the coefficient on this variable?

c) Right after the regression statement, generate a new variable that is the predicted residuals from the regression in part a) and call this variable res1. This can be done by using the STATA statement

```
predict res1, residual
``` 

Next, obtain the correlation coefficient between res1 and two of the explanatory variables: lsat and lcost. Do the correlation coefficients between res1 and lcost and res1 and lsat make sense?

You can use the following statement to get the correlation coefficients

```
corr res1 lcost lsat
``` 

d) Next, output the predicted values from the regression and call this variable pred1. This can be done with the following statement

```
predict pred1, xb
``` 

With the predicted values get the correlation coefficient between the actual y and the predicted y

```
corr lsalary pred1
``` 

Square the correlation coefficient and compare to the $R^2$ from the original model. What is the relationship between the correlation coefficient squared and the $R^2$ from the original model?

e) Returning to the model in part a), estimate a similar model but now delete the variable “rank”. What happens to the coefficient on “lsat” when this variable is deleted? Provide an intuitive explanation for why the coefficient on “lsat” changed when “rank” was deleted.

f) Run a regression of lcost (y), on three variables (x’s): lsat, rank and age. Output the residuals from this regression, then regress the variable lnsalary (y) on the residuals of lcost (x). Compare the coefficient on the residual of lcost to the coefficient for lcost in part a).

6. *Return to problem 6 on problem set 3. A pharmaceutical company is investigating the cholesterol lowering benefits of a new drug. In a sample of n subjects the company randomly assigns milligrams of active ingredients (label this as x_{1i}) and the outcome of interest, labeled as y_i, is the change in cholesterol from the start until the end of the trial. Initially, the researchers estimate a model of the form $y_i = \beta_0 + x_{1i} \beta_1 + \epsilon_i$. However, a colleague mentions that as part of the experiment, they also collected detailed data on characteristics of survey participants that predict $y_i$ like their weight at the start of the trial, age, sex, ethnicity/race, plus other variables. The colleague asks whether one should include these covariates (label them as x_{2i}, x_{3i}, …x_{ki}) into the basic regression?

a) By estimating a model of $y_i = \beta_0 + x_{1i} \beta_1 + x_{2i} \beta_2 + \cdots + x_{ki} \beta_k + \epsilon_i$, do you anticipate that the estimate on $\hat{\beta}_1$ will change?

b) In a multivariate model, the estimated variance of $\hat{\beta}_1$ is given as 
$$V(\hat{\beta}_1) = \frac{\sigma^2}{(1 - R_i^2) \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}$$

What is the likely consequence of adding these additional covariates (x_{2i}, x_{3i}, …x_{ki}) to the estimated variance of $\hat{\beta}_1$?

7. *On the next page are the results from two regression models: In model (1), I regress Y on X_1, and note that the standard error on the coefficient on X_1 is very small and the t-statistic on the coefficient on $\hat{\beta}_1$ is over 23. Note that in model (2), when I add X_2 to the model, the standard error on $\hat{\beta}_1$ increases by a factor of 3 and the t-statistic on this parameter falls to 1.39. Using the information given, provide an intuitive explanation for why the standard increases so much on $\hat{\beta}_1$ when X_2 is added to the model. To get full credit, you must provide the proper equation.

8. *Many people get their health insurance through their job and because of high health insurance costs, many employers are considering offering free on-site exercise classes as a way of encouraging healthy behaviors and hopefully reducing medical care costs. The evidence for subsidized exercise classes comes primarily from research in the field of public health. In these models the authors collect data from an employer and estimate a regression of the form $y_i = \beta_0 + x_{1i} \beta_1 + \epsilon_i$, where $y_i$ is annual spending on health care for employee i and $x_{1i}$ is a dummy variable that equals 1 if the person uses the on-site health care services. Call this model (1). Let $\hat{\beta}_1$ be the estimate for $\beta_1$ from model (1) and in this case, the author gets the expected result that $\hat{\beta}_1 < 0$ – people that use on-site exercise classes have lower health care spending. Model (1) has been criticized because it does not control for the fact that the least healthy employees are the ones the least likely to enroll in these classes. Consider a simple extension to the model where the author has detailed data on the health of employees prior to the exercise classes opening. Let $x_{2i}$ be a simple index that equals the number of chronic health conditions a person has (e.g., a person with high blood pressure, obesity, and diabetes has a count of three whereas a healthy person has a count of zero). Now consider estimating model (2) which is of the form $y_i = \beta_0 + x_{1i} \beta_1 + x_{2i} \beta_2 + \epsilon_i$. If Model (2) is the true model, do you anticipate that, the estimate $\hat{\beta}_1$ from model (1), is biased up or down? Explain your answer and to get full credit, you must provide an appropriate equation.
Results for Question 7

Correlation between \( X_1 \) and \( X_2 \)

\[ . \text{ corr } x1 \ x2 \]
\( (\text{obs}=2489) \)

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>0.9994</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Model 1: Regression of \( Y \) on \( X_1 \)

\[ . \text{ reg } y \ x1 \]

\[
\begin{array}{l}
\text{Source} | \text{SS} \ 
\text{df} \  \text{MS} \ 
\text{F( 1, 2487) = 562.63} \\
\text{Model} | 121.044173 | 1 | 121.044173 \\
\text{Residual} | 535.054756 | 2487 | .215140634 \\
\text{Total} | 656.098929 | 2488 | .263705357 \\
\%
\end{array}
\]

\[
\begin{array}{l}
y | \text{ Coef. } \text{ Std. Err.} \text{ t } \text{ P>|t|} \ [95\% \text{ Conf. Interval}] \\
x1 | .0765488 | .0032272 | 23.72 | 0.000 | .0702205 .0828771 \\
_cons | 5.059357 | .0435541 | 116.16 | 0.000 | 4.973951 5.144763 \\
\%
\end{array}
\]

Model 2: Regression of \( Y \) on \( X_1 \) and \( X_2 \)

\[ . \text{ reg } y \ x1 \ x2 \]

\[
\begin{array}{l}
\text{Source} | \text{SS} \ 
\text{df} \  \text{MS} \ 
\text{F( 2, 2486) = 281.41} \\
\text{Model} | 121.115811 | 2 | 60.5579054 \\
\text{Residual} | 534.983118 | 2486 | .215198358 \\
\text{Total} | 656.098929 | 2488 | .263705357 \\
\%
\end{array}
\]

\[
\begin{array}{l}
y | \text{ Coef. } \text{ Std. Err.} \text{ t } \text{ P>|t|} \ [95\% \text{ Conf. Interval}] \\
x1 | .1304861 | .0935397 | 1.39 | 0.163 | -.0529375 .3139098 \\
x2 | -.0539557 | .0935159 | -0.58 | 0.564 | -.2373328 .1294213 \\
_cons | 5.059738 | .0435649 | 116.14 | 0.000 | 4.974311 5.145166 \\
\%
\end{array}
\]
9. Research has shown that students attending higher quality colleges and universities tend to have higher wages after graduation than those attending less selective institutions. Using a nationally representative sample of college graduates aged 30-39, researchers regress the natural log of annual earnings ($y_i$) on the average SAT score from the college the respondent attended ($x_{1i}$) using the simple bivariate regression model $y_i = \beta_0 + x_{1i} \beta_1 + \epsilon_i$. Call this model (1). Let $\hat{\beta}_1$ be the estimate for $\beta_1$ from model (1) and in this case, the author gets the expected result that $\hat{\beta}_1 > 0$ – students that graduated from higher quality schools tends to have higher earnings. Someone criticizes model (1) because it does not control for differences in other characteristics of the students that are likely to be correlated with earnings. For example, the author does not have a measure of academic ability for the student like an SAT score which they argue should be included in the model. Suppose the author considers estimating model (2) which is of the form $y_i = \beta_0 + x_{1i} \beta_1 + x_{2i} \beta_2 + \epsilon_i$ where $x_{2i}$ is the student's own SAT score. If Model (2) is the true model, do you anticipate that $\hat{\beta}_1$, the estimate from model (1), is biased up or down? Explain your answer and to get full credit, you must provide an appropriate equation.

10. A researcher regresses $y$ on $x_1$ and produces the results below. A colleague argues that the model should also include the covariates $x_2$, $x_3$, and $x_4$, which the colleague argues are strong predictors of $y$. Below is a matrix that provides the correlation coefficients for the variables $x_1$, $x_2$, $x_3$ and $x_4$. Given these results, do you expect that adding $x_2$, $x_3$ and $x_4$ to the model will change the results much? Assume your colleague is correct that $x_2$, $x_3$ and $x_4$ are strong predictors of $y$.

Results for Problem 10

```
. reg y x1

Source | SS df MS
---------+--------------------------------------------------
Model | 174.739778 1 174.739778
Residual | 874.36848 3979 .219745785
---------+--------------------------------------------------
Total | 1049.10826 3980 .26359504

Number of obs = 3981
F(  1,  3979) = 795.19
Prob > F = 0.0000
R-squared = 0.1666
Adj R-squared = 0.1664
Root MSE = .46877

------------------------------------------------------------------------------
y | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
x1 | .0739822  .0026236 28.20 0.000 .0688386 .0791259
_cons | 5.105143  .0353055 144.60 0.000 5.035925 5.174362
------------------------------------------------------------------------------

. corr x1 x2 x3 x4
(obs=3981)

|   x1   x2   x3   x4
-------------+------------------
x1 | 1.0000
x2 | 0.0182 1.0000
x3 | 0.0002 0.0061 1.0000
x4 | 0.0025 0.0075 -0.0211 1.0000
```