Introduction to Regression

Above we considered the relationship between two variables X and Y with the covariance and correlation. Economics frequently tries to explain the relationship between two variables, such as consumption and GDP or age and wealth. Econometrics attempts to explain the relationship between two or more explanatory variables and a dependent variable. However, we do not claim that our models “explain” the dependent variable in terms of an independent variable are any more than when we say that the weatherman explains what will happen to the weather. That is, correlation does not imply causation.

A useful introduction to regression is to consider the following. Suppose you want to predict how much money a student here at UTPB earns in a year. Furthermore, suppose you are given the following data.

<table>
<thead>
<tr>
<th>Annual Earnings</th>
<th>Years Work Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>$20,000</td>
<td>1</td>
</tr>
<tr>
<td>$15,000</td>
<td>2</td>
</tr>
<tr>
<td>$25,000</td>
<td>5</td>
</tr>
<tr>
<td>$35,000</td>
<td>1</td>
</tr>
<tr>
<td>$12,500</td>
<td>15</td>
</tr>
<tr>
<td>$14,750</td>
<td>3</td>
</tr>
<tr>
<td>$17,500</td>
<td>3</td>
</tr>
<tr>
<td>$19,000</td>
<td>4</td>
</tr>
<tr>
<td>$21,000</td>
<td>8</td>
</tr>
<tr>
<td>$45,000</td>
<td>12</td>
</tr>
<tr>
<td>$36,500</td>
<td>8</td>
</tr>
<tr>
<td>$32,000</td>
<td>8</td>
</tr>
<tr>
<td>$18,000</td>
<td>1</td>
</tr>
<tr>
<td>$47,852</td>
<td>15</td>
</tr>
<tr>
<td>$22,000</td>
<td>2</td>
</tr>
<tr>
<td>$25,124</td>
<td>6</td>
</tr>
<tr>
<td>$31,000</td>
<td>15</td>
</tr>
<tr>
<td>$21,250</td>
<td>1</td>
</tr>
<tr>
<td>$14,200</td>
<td>8</td>
</tr>
<tr>
<td>$39,000</td>
<td>8</td>
</tr>
</tbody>
</table>

If you are given not additional information and you want to estimate how much a UTPB student earn, what would be your estimate? Given the above descriptive statistics discussion, you may look at measures of location.
Mean=$25,584
Median=$21,625
Mode=N/A

You may even construct a model to estimate earnings, such as:

\[ \hat{E} = \bar{E} + \epsilon \]

This simple model says your best earnings estimate, \( \hat{E} \), is mean earnings plus a random error term, \( \epsilon \). However, we assume the \( E(\epsilon) = 0 \), hence, your best estimate for any student’s earnings is just mean earnings of all students.

Using mean earnings may be a good estimate of student earnings, but you may be given more information than just earnings. You may be also given years of work experience. This additional information may improve your estimated earnings. For example, if we have two students and one is just out of high school with one year of work experience and the second is returning to finish their degree at night and has 12 years of work experience, who do you think earns more? You may very well predict that the student with 12 years work experience earns more because they have accumulated skills that are rewarded with higher earnings. They’ve been in the labor force long enough to establish labor force connections or have more training. Whatever the reason, we’ll predict the association between work experience and earning is positive.

Recall \( y = mx + b \) from your algebra courses. Statistics measures the relationship as \( y = mx + b + \epsilon \). (Notice, statistics uses \( \alpha \) to represent the y-intercept and \( \beta \) to represent the slope coefficient.) Your math courses suggest such a relationship between \( y \) and \( x \) is deterministic. That is, if I know \( x \), I know \( y \) with certainty. Statistics is less certain of this relationship. We may believe the correct model specification is \( y = mx + b \), but we know we live in a probabilistic world, not a deterministic world. In other words, \( y = mx + b \) may give us an improvement over the mean of \( y \), but \( y = mx + b \) will be measured with error because we will not always be correct. This
is an important distinction because it allows us to estimate the parameters m and b. Perhaps when you first studied mathematics you wondered where the m and b came from and what practical application they had. The typical mathematics response is “They provide a rigorously analytical tool to help us work through the logic of relationships.” However, statistics makes y=mx+b operational. For example, you may be a manager and have to estimate the relationship between sales and advertising. Statistics and regression analysis allow us to estimate the m and b such that knowing our advertising expenses helps us estimate what sales will be given a certain advertising budget.

Making the transition from mathematics to statistics, we do not use the notation y=mx+b. Rather, we use the model

\[ y = \alpha + \beta x + \epsilon \]

This is known as the simple linear regression model. y is called the dependent or endogenous variable. That is, y is dependent on x or y is determined inside the model. x is call the independent or exogenous variable. In other words, x is independent in the model or determined outside the model.

To take full advantage of information in the wage second model, we may construct the following model,

\[ \hat{E} = \alpha + \beta WE + \epsilon \]

where \( \alpha \) represents earnings with no work experience (WE=0), \( \beta \) equals additional earnings for each additional year of work experience (dE/dWE), and \( \epsilon \) represents a random error term. This second model allows us to take advantage of the work experience information where the first model was restricted from using this information.
To formalize the simple linear regression model, there are a series of assumptions that if met produce “good” estimates for $\alpha$ and $\beta$. These assumptions are important in estimating the simple linear regression model.

**Assumptions of the Classical Linear Regression Model**

A1. The error terms, $\varepsilon$, are normally distributed.

A2. The error has an expected (average) value of zero, i.e. $E(\varepsilon)=0$.

A3. The error terms have a common variance, $\text{var}(\varepsilon)=\sigma^2$ for all error terms.

A4. Each error term is independent from all other error terms, $\text{cov}(\varepsilon_i, \varepsilon_j)=0$

A5. Regressors and error terms are independent of each other, $\text{cov}(x_i, \varepsilon_i)=0$

These assumptions will hold not only for the simple linear regression model but the multivariate regression model to be considered later. A few comments on the assumptions may be helpful. A1 says the errors are distributed randomly. A2 says predicted $y$ will not be biased because of the error terms. A3 says the variance of the error terms will not change as $x$ changes. This is termed homoskedasticity. A4 says the error terms are not correlated with other error terms. This correlation of errors between surrounding errors is known as autocorrelation. A5 says that knowing information about $\varepsilon$ gives us no information about $x$. If such is not the case, our independent variables are measured with error and we will have to find alternative explanatory variables if we hope to correctly model $y$. While these assumptions may appear trivial, they form the basis of estimating the regression model. If all of the assumptions are satisfied, the estimates for the (simple) classical linear regression model are the minimum variance of all unbiased estimates. That is, we can not obtain better estimates of $\alpha$ and $\beta$.

Now that we’ve considered the assumptions of the classical linear regression model (CLRM), let’s consider two means of estimating $\alpha$ and $\beta$. First, we can use CLRM restrictions to estimate $\alpha$ and $\beta$. This is known as method of moments estimation. Second, we can minimize
the sum of squared errors. This is known as ordinary least squares estimation. As you will see, the latter is merely a special case of the former.

**Method of Moments**

The CLRM makes two important *population* assumptions.

A2. $E(\varepsilon_i) = 0$

A5. $cov(x_i, \varepsilon_i) = 0$

These are known as moment restrictions. The method of moments estimator simply sets the population moment equal to the sample moment and solves for the estimated parameters.

<table>
<thead>
<tr>
<th>Population Moment Assumption</th>
<th>Sample Counterpart</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(\varepsilon) = 0$</td>
<td>$\frac{1}{n} \sum \varepsilon_i = 0 \Rightarrow \sum \varepsilon_i = 0$</td>
</tr>
<tr>
<td>$cov(x_i, \varepsilon) = 0$</td>
<td>$\frac{1}{n} \sum x_i \varepsilon_i = 0 \Rightarrow \sum x_i \varepsilon_i = 0$</td>
</tr>
</tbody>
</table>

We know that

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \varepsilon_i \Rightarrow \varepsilon_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

Hence

$$\sum \varepsilon_i = 0 \text{ or } \sum (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\sum x_i \varepsilon_i = 0 \text{ or } \sum x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

We can solve for the y intercept by dividing the first equation by n,

$$\sum y_i = n \hat{\alpha} + \hat{\beta} \sum x_i \Rightarrow \bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$$

Solving for $\hat{\alpha}$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$
We can solve for the slope coefficient by expanding the summation operators through and rearranging results in

\[
\sum y_i = n\alpha + \hat{\beta} \sum x_i, \quad \sum \alpha = n\alpha \\
\sum x_i y_i = \hat{\alpha} \sum x_i + \hat{\beta} \sum x_i^2
\]

Multiplying the first equation by the sum of \( x_i \) and the second equation by \( n \) yields

\[
\sum x_i \sum y_i = n \sum x_i \hat{\alpha} + (\sum x_i)^2 \hat{\beta} \\
n \sum x_i \sum y_i = n \sum x_i \hat{\alpha} + n \sum x_i^2 \hat{\beta}
\]

Subtracting the first equation from the section yields

\[
\sum x_i y_i - \sum x_i y_i = [(\sum x_i)^2 + n \sum x_i^2] \hat{\beta}
\]

\textit{Solving for} \( \hat{\beta} \)

\[
\Rightarrow \hat{\beta} = \frac{n \sum x_i y_i - \sum x_i y_i}{(\sum x_i)^2 + n \sum x_i^2}
\]

These are the method of moment estimators of \( \alpha \) and \( \beta \).

\textit{Ordinary least squares}

We just derived parameter estimates for \( \alpha \) and \( \beta \) by restricting the population to sample moments and solving for sample parameters. Ordinary least squares (OLS) is a special case of MM and produces the same parameter estimates by minimizing the sum of squared errors (SSE). In the process of minimizing the SSE, we obtain the moment restrictions employed in MM. These are known as the normal equations. Hence, OLS is a special case of MM.

First, we define the SSE.
Now, we’re going to minimize SSE with respect to $\alpha$ and $\beta$.

\[
\frac{\partial SSE}{\partial \alpha} = -2 \sum (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0
\]

\[
\frac{\partial SSE}{\partial \beta} = -2x_i \sum (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0
\]

Dividing both equations by $-2$ produces

\[
\sum \varepsilon = 0
\]

\[
\sum x\varepsilon = 0
\]

These are known as the OLS normal equations. Solutions for $\alpha$ and $\beta$ using the normal equations yield OLS estimates. You may also recognize these normal equations as the moment restrictions used in the derivation of the MM estimates. Hence, from this point we use the same solution process to derive OLS estimates in the simple CLRM.

Multiplying the first equation by the sum of $x_i$ and the second equation by $n$ yields

\[
\sum x_i \sum y_i = n \sum x_i \hat{\alpha} + \left(\sum x_i\right)^2 \hat{\beta}
\]

\[
n \sum x_i \sum y_i = n \sum x_i \hat{\alpha} + n \sum x_i^2 \hat{\beta}
\]

Subtracting the first equation from the section yields
Dividing the first equation by $n$

$$\sum x_i y_i - \sum x_i y_i = \left[ (\sum x_i)^2 + n \sum x_i^2 \right] \hat{\beta}$$

Solving for $\hat{\beta}$

$$\Rightarrow \hat{\beta} = \frac{n \sum x_i y_i - \sum x_i y_i}{(\sum x_i)^2 + n \sum x_i^2}$$

These are the sum of squared estimators of $\alpha$ and $\beta$. When we estimate the above earnings model, these estimates of $\alpha$ and $\beta$ are what the computer uses.

**Goodness of Fit and Hypothesis Tests**

Recall, we wanted to know which model we should use when predicting student earnings. First, we can use

$$\hat{E} = \bar{E} + \epsilon$$

Or we can use

$$\hat{E} = \alpha + \beta WE + \epsilon$$

We may want to use the simple linear regression model because it uses more information, therefore, may yield better estimates of the earnings data generating process. Now let’s consider statistics that have been constructed which help us select a better estimator. Suppose we wanted
to compare the mean of $y$ as a predictor versus the simple regression model as a predictor of $y$.

The following graph helps explain the concept.
Consider observation Y. It is easy to see that Y-Hat is a better predictor of Y than Y-bar. An important measure of how well the estimated regression line is $R^2$. $R^2$ measures the variation in Y that is explained by variations in X. A more cryptic, yet useful definition of $R^2$ is the percent of Y that is explained by the model $\alpha + \beta x$. However, before we calculate $R^2$, a few definitions are necessary.

$R^2$ is calculated as follows:

$y - \bar{y}$ equals total variation
1) Add Zero: $y_i - \bar{y} = (y_i - \hat{y}) + (\hat{y} - \bar{y})$
2) Square and Sum: $\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$
3) Rename: $SST = SSE + SSR$

where $SST = \sum (y_i - \bar{y})^2$

$SSE = \sum (y_i - \hat{y})^2$
$SSR = \sum (\hat{y} - \bar{y})^2$

We use the third equation to calculate $R^2$. $R^2$ is the ratio of explained variation by the regression to total variation in y.

$$\frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSR}{SST} \Rightarrow R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

So long as $SSE < SST$, $R^2$ is positive, i.e., $R^2$ has a lower bound of 0. We also know that $SST \geq SSR$, therefore, $R^2$ has an upper bound of 1. The closer $R^2$ is to one, the greater explanatory
power of the equation. Alternatively, the closer $R^2$ is to 0, the less explanatory power of the model.

*The t-statistic—Testing a single parameter*

After $\alpha$ and $\beta$ are estimated, we must consider if they are “good” estimates. We think in terms of the statistical significance of $x$ in explaining $y$. This relies on hypothesis testing you learned in your principles course. To test a hypothesis, we must first understand the meaning of a hypothesis. We start with a proposition such as mean UTPB student’s earnings is $25,584$. This hypothesis may or may not be true. We call the hypothesis the null hypothesis and denote it as $H_0$. If we reject the null hypothesis, we opt for an alternative hypothesis, $H_1$. Corresponding to the two realizations and the two conclusions drawn, we have the following four possibilities:

<table>
<thead>
<tr>
<th>Reality</th>
<th>Test Result</th>
<th>$H_0$ is true</th>
<th>$H_0$ is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressor significant (Reject $H_0$)</td>
<td>Type I error</td>
<td>Correct Conclusion</td>
<td></td>
</tr>
<tr>
<td>Regressor Not significant (Fail to Reject $H_0$)</td>
<td>Correct Conclusion</td>
<td>Type II error</td>
<td></td>
</tr>
</tbody>
</table>

We must also define what we mean by significant and insignificant.

*Significant:* Sample variation is unlikely to explain the difference between the null hypothesis and sample values, i.e., the independent variable matters in the explanation of the dependent variable.

*Insignificant:* Sampling variation is likely to explain the difference between the null hypothesis and sample values, i.e., the independent variable does not matter in the explanation of the dependent variable.

We formalize our hypothesis by attaching probability to our making errors:

Type I error: $\alpha = \text{Prob}(\text{rejecting } H_0 \mid H_0 \text{ is true})$

Type II error: $\beta = \text{Prob}(\text{not rejecting } H_0 \mid H_0 \text{ is not True})$
In this class we will only be concerned with $\alpha$. That is, when $H_0$ is true, we don’t want to reject it.

To test the null hypothesis, we must evaluate a test statistic to determine if the null hypothesis is a likely event. For example, our null hypothesis may state

$$H_0 : \hat{\beta} = \beta_0$$
$$H_1 : \hat{\beta} \neq \beta_0$$

To determine if $H_0$ is or is not supported by available evidence, we calculate a test-statistic to determine the likelihood of an event. This t-statistic is defined as

$$t_{n-k} = \frac{\hat{\beta} - \beta_0}{s_\beta}$$

where $s_\beta$ is the standard deviation of the estimated slope coefficient. $t_{n-k}$ measures the standardized distance between the estimated and hypothesized values. In other words, $t_{n-k}$ measures the number of standard deviations the estimate is from the hypothesized value. When there is a large difference between the estimated coefficient and the hypothesized value corrected for the standard deviation of the estimate, we don’t believe $\beta_0$ is likely. So we reject the null

$$H_0 : \hat{\beta} = \beta_0$$

hypothesis that

and fail to reject the alternative hypothesis that

$$H_1 : \hat{\beta} \neq \beta_0$$

So, now we must define by what we mean by large. Remember, our test statistic is a random variable and has its own distribution. This distribution is called the t-distribution. If the
t-statistic based on the regression estimate is greater than the critical value base upon the theoretical distribution, we reject the hypothesis. If the t-statistic is less than the theoretical critical value, we fail to reject the hypothesis. We’ll discuss critical values in greater depth during class.

The t-statistic with regression estimates may seem counter-intuitive. Given the regression equation

\[ y_i = \alpha + \beta x_i \]

our null hypothesis is stated as \( \beta = 0 \). This means that \( x \) is not significant in the explanation of \( y \). Hence, rejecting the null hypothesis that \( \beta = 0 \) means that we can’t say the influence of \( x \) on \( y \) is statistically meaningless, i.e., \( x \) matters in the explanation of \( y \). If we fail to reject the null hypothesis, this means \( x \) does not explain \( y \).

Example:

Let \( \hat{\beta} = 6 \)

\( s_\beta = 2 \)

Critical Value = 2

\[ \Rightarrow t_1 = 3 \]

\( \therefore \) \( x \) is statistically significant.

The F-statistic—Testing the equation

We’ve considered the t-statistic, a test for individual parameters. We can also test the significance of the equation. The null hypothesis in this case is \( \alpha = \beta = 0 \). If the null hypothesis is true, the model has little power in explaining \( y \). If the null hypothesis is false, the model is significant in explaining \( y \). Where the t-statistic follows a t-distribution, the F-statistic follows an F-distribution. The F-statistic for the simple linear regression model is

\[ F_{1,N-2} = \frac{\text{explained variation}}{\text{Unexplained Variation}} = \frac{RRS / 1}{ESS / (N - 2)} \]
The subscripts on F denote the degrees of freedom in the numerator and denominator, 1 and N-2.

After finding the F critical value in the appendix (page 608-609), we compare the F-statistic to the theoretical critical value. If the F-statistic is greater than the critical value, the null hypothesis \( \alpha = \beta = 0 \) is rejected. The equation explains \( y \). If the F-statistic is less than the critical value, we can’t reject the null. The equation does not explain \( y \). We now consider an example to solidify the concepts for the simple linear regression model.

**Example**

Recall our UTPB data. We now use Excel to illustrate a simple linear regression model.

This is the only time we use Excel. Remaining illustrations and assignments use Shazam.

<table>
<thead>
<tr>
<th>SUMMARY OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression Statistics</td>
</tr>
<tr>
<td>Multiple R 0.930401</td>
</tr>
<tr>
<td>R Square 0.865646</td>
</tr>
<tr>
<td>Adjusted R Square 0.858182</td>
</tr>
<tr>
<td>Standard Error 3970.343</td>
</tr>
<tr>
<td>Observations 20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
<th>Lower 95.0%</th>
<th>Upper 95.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15102.52</td>
<td>1317.358</td>
<td>11.46424</td>
<td>1.05E-09</td>
<td>12334.85</td>
<td>17870.18</td>
<td>12334.85</td>
</tr>
</tbody>
</table>

The estimate regression model is
\[ \hat{E} = 15,103 + 2,438WE \]

The interpretation of the model is that when a student has no work experience, her entry-level wage is $15,103. For each additional year of work experience, the average student earns $2,438. To test the individual significance of work experience on earnings, observe that the t-statistic is 10.77. This is greater than the t-distributions critical value 2.093. Hence, the individual influence of work experience is statistically significant. By comparing the calculated F-statistic with the critical value from the F-distribution, we see that the calculated F-statistic 115.98 is greater than the critical value of 8.18. Hence, the model is statistically significant in explaining earnings.

The estimated R\(^2\) indicates that variation in work-experience explains 93% of the variation in earnings. An additional issue is to check the normal equations. In our example,

\[
\frac{1}{n} \sum \varepsilon_i = 0 \quad \text{Example} : \quad 1.09^{-11}
\]

\[
\frac{1}{n} \sum x_i \varepsilon_i = 0 \quad \text{Example} : \quad -1.09^{-10}
\]

So, our normal equations hold. Both are very close to zero.

I like to visualize the relationship between the dependent and independent variable, so I like to graph the results. I recommend you always graph your data. Figure 4 shows the fitted line between earnings and work experience in a scatter plot.

![Scatter plot of earnings and work experience](scatter_plot.png)
The graph helps develop intuition. It also may help to determine if there are any observations that deserve particular attention as outliers. We can learn a lot about our estimates by plotting the model’s errors. Figure 5 shows the relationship between the errors and work experience.

![Figure 5: Errors on Work Experience]

Notice how the model’s errors get smaller as work experience increases. This violates assumption A3, meaning we will have to correct for hetroskedasticity.
Homework 2

Note: This assignment must be done with Shazam for full credit.

1. Use file EX32 on the PR data disk to calculate the mean and standard deviation of Personal Consumption Expenditures (GC).

2. Use file EX32 to regress Personal Consumption Expenditures (GC) on Aggregate Disposable Income (GYD). Answer the following:
   A. What is the estimated simple linear regression model of GC regressed on GYD?
   B. Is GYD statistically significant in explaining GC?
   C. Is your estimated equation statistically significant in the explanation of GC?
   D. What percent of the variation in GC can be explained by the variation in GYD?
   E. Plot GC on GYD.
   F. Plot the error terms on GYD. What do you notice about the pattern of the error terms?
   G. Test the normal equations.

3. Use file EX33 to regress Personal Consumption Expenditure on Autos (GCDAN) on National Income (GWY). Answer the following:
   A. What is the estimated simple linear regression model of GCDAN regressed on GWY?
   B. Is GWY statistically significant in the explanation of GCDAN?
   C. Is your estimated equation statistically significant in the explanation of GCDAN?
   D. What percent of the variation in GCDAN can be explained by the variation in GWY?
   E. Plot GCDAN on GWY.
   F. Plot the error terms on GWY. What do you notice about the pattern of the error terms?
   G. Test the normal equations.
Shazam Overview

This section provides a brief overview to Shazam. Some important commands are:

1. Opening and reading a file in Shazam
   
   Data file:  gdp.dat

   *Enter Shazam*

   File 4 a:gdp.dat

   READ (4) variable names/ Options

   *Commands*

   STOP

2. Setting a range

   SMPL beg end

3. Descriptive statistics

   STAT variables

4. Save screen output to a disk

   FILE SCREEN filename

5. Create a new variable

   GENR newvar=equation

6. Plot dependent variable on independent variables

   PLOT depvars indep/options

7. Ordinary least squares regression

   OLS y x1 x2 . ./options

7. Autocorrelation

   AUTO y x1 x2 . ./options

8. Least absolute deviation regression

   ROBUST y x1 x2 . ./options

9. Maximum likelihood estimation
MLE y x1 x2 . /options

10. Qualitative response models

   PROBIT y x1 x2 . /options
   LOGIT y x1 x2 . /options

11. Non-linear models

   NL # of equations/NCOEF= options
   EQ equation
   COEF coef value coef

12. Two-stage least squares

   2SLS lhsvars rhsvars (indepvars)/ options
The Multiple Regression Model

Now we know the basics of regression, we can expand on our knowledge to construct more realistic models. For example, work experience may be a good predictor for earnings, but it certainly isn’t the only predictor. Hence, we want to expand our model to include multiple explanatory factors. This is precisely what multi-variate regression does, expands our model from a single explanatory variable into multiple variables. The model specification is

\[ y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} + \varepsilon \]

Note the coefficients are also linear in \( x_i \). (\( \beta_i \) does not change when \( x_i \) changes.) Like our simple linear regression, the multivariate model has a set of assumptions that when met yield the “best” estimates of the model.

**Assumptions of the Multivariate Classical Linear Regression Model**

A1. The error terms, \( \varepsilon \), are normally distributed.

A2. The error has an expected (average) value of zero, i.e. \( E(\varepsilon) = 0 \).

A3. The error terms have a common variance, \( \text{var}(\varepsilon) = \sigma^2 \) for all error terms.

A4. Each error term is independent of all other error terms, \( \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \).

A5. Regressors and error terms are independent of each other, \( \text{cov}(x_i, \varepsilon_i) = 0 \).

A6. There is no relationship between explanatory variables, \( \text{cov}(x_i, x_j) = 0 \).

These assumptions are the same as those underlying the simple CLRM, except A6. A6 just says the explanatory variables are independent.

To derive the multivariate CLRM, we can use scalar algebra. Define the model as

\[ y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon \]

We now seek to minimize the sum of squares.
\[ SSE = \sum (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \]
\[ \Rightarrow \frac{\partial SSE}{\partial \beta_i} = -2x_i \sum (y_i - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0 \]
\[ A) \sum x_{i1}y_i = \beta_1 \sum x_{i1}^2 + \beta_2 \sum x_{i1} x_{i2} \]
\[ \Rightarrow \frac{\partial SSE}{\partial \beta_2} = -2x_2 \sum (y_i - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0 \]
\[ B) \sum x_{i2}y_i = \beta_1 \sum x_{i1} x_{i2} + \beta_2 \sum x_{i2}^2 \]

Multiply A by
\[ \sum x_{i2}^2 \]

And B by
\[ \sum x_{i1} x_{i2} \]

Now, subtract B' from A'.
\[ \Rightarrow \sum x_{i2}y_i \sum x_{i2}^2 - \sum x_{i2}y_i \sum x_{i1}x_{i2} = \hat{\beta}_1 \left[ \sum x_{i1}^2 \sum x_{i2}^2 - \left( \sum x_{i1}x_{i2} \right)^2 \right] \]
\[ \Rightarrow \hat{\beta}_1 = \frac{\sum x_{i1}y_i \sum x_{i2}^2 - \sum x_{i2}y_i \sum x_{i1}x_{i2}}{\sum x_{i1}^2 \sum x_{i2}^2 - \left( \sum x_{i1}x_{i2} \right)^2} \]

Similarly,
\[ \Rightarrow \hat{\beta}_2 = \frac{\sum x_{i2}y_i \sum x_{i1}^2 - \sum x_{i1}y_i \sum x_{i1}x_{i2}}{\sum x_{i1}^2 \sum x_{i2}^2 - \left( \sum x_{i1}x_{i2} \right)^2} \]

Now we can reverse back to find the intercept.
\[ \hat{\alpha} = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \]
however, such is rather cumbersome and unnecessary. Instead, we opt for linear algebra where multiple variables are included in a simple matrix form. You will see that this matrix representation is easier to manage. The simple CLRM also a special case.

Let’s make some necessary definitions. Let

\[
y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 & x_2 & \cdots & x_m \\ 1 & x_1 & x_2 & \cdots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdot & \cdot & \cdots & \cdot \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}
\]

We can now define our model as

\[ y = X\beta + \epsilon \]

We use this model to minimize the sum of squares to obtain parameter estimates.

1) Define error vector: \( \epsilon = y - X\beta \)

2) Minimize the SSE: \( SSE = (y - X\beta)^T (y - X\beta) = \)
\[
\epsilon^T \epsilon = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta
\]
\[
\frac{\partial SSE}{\partial \beta^T} = -2X^T y + 2X^T X\beta = 0 \quad \text{(First Order Condition)}
\]
\[
\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y
\]

Notice the first order condition produces the normal equations.

\[ X^T \epsilon \]

This includes the normal equations from the simple linear regression model. You may also want to think about how the method of moment’s moment restrictions are used to obtain parameter estimates.

**Estimator Properties in Matrix Form**

1) Unbiased
A) \( y = X\beta + \varepsilon \)

B) \( \hat{\beta} = \left(X^TX\right)^{-1}X^Ty \)

C) \( \hat{\beta} = \left(X^TX\right)^{-1}X^T(X\beta + \varepsilon) = \left(X^TX\right)^{-1}X^TX\beta + \left(X^TX\right)^{-1}X^T\varepsilon \)

D) Assumption A2

\( E(\varepsilon) = 0 \)

E) \( E(\hat{\beta}) = \beta \)

\( \therefore \) Unbiased

2) Minimum variance

Let \( A = c \)

\( \Rightarrow A\varepsilon = (X^TX)^{-1}X^T\varepsilon = (X^TX)^{-1}X^T(y - X\beta) = (X^TX)^{-1}X^T\varepsilon - \beta = \left(\hat{\beta} - \beta\right) \)

\[ V(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = E[(A\varepsilon)(A\varepsilon)^T] = E(A\varepsilon^TA^T) = A\varepsilon^TA^T = A(\sigma^2I)A^T = \sigma^2AA^T \]

However,

\( A = \left(X^TX\right)^{-1}X^T \)

\( \Rightarrow AA^T = \left[(X^TX)^{-1}X^T\right]\left[(X^TX)^{-1}X^T\right] = \left[(X^TX)^{-1}X^T\right]\left[X\left(X^TX\right)^{-1}\right] = \left(X^TX\right)^{-1} \)

\( \therefore V(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = \sigma^2\left(X^TX\right)^{-1} \)

We know that OLS is linear and unbiased. To show that OLS is minimum variance of linear, unbiased estimators, we use the Gauss-Markov Theorem.

Gauss-Markov Theorem-Given A2-A5, OLS estimators are most efficient (minimum variance) in the sense that they have minimum variance of all linear unbiased estimators.

These estimators are also known as Best Linear Unbiased Estimators (BLUE). If A1-A5 hold, then OLS will be Minimum Variance of All Unbiased Estimators.
Back to BLUE. To show BLUE, we must show that any other linear unbiased estimator \( \mathbf{b} \) has greater variance than \( \mathbf{b}_{\text{OLS}} \).

\[
\mathbf{b} - \beta = (A + C)\epsilon
\]

\[
V(\mathbf{b}) = E[(\mathbf{b} - \beta)(\mathbf{b} - \beta)^T] = E[(A + C)\epsilon (A + C)\epsilon^T] = E[(A + C)\epsilon \epsilon^T (A + C)^T] = (A + C)E(\epsilon \epsilon^T)(A + C)^T = \sigma^2 (A + C)(A + C)^T
\]

However,

\[
(A + C)(A + C)^T = AA^T + CA^T + AC^T + CC^T = (X^TX)^{-1}X^TX(X^TX)^{-1} + CX(X^TX)^{-1} + (X^TX)^{-1}X^TC^T + CC^T
\]

Recall, \( CX = X^TC^T \)

\[
\Rightarrow (A + C)(A + C)^T = (X^TX)^{-1} + CC^T
\]

\[
\therefore V(\mathbf{b}) = \sigma^2 [(X^TX)^{-1} + CC^T] = V(\hat{\beta}) + \sigma^2 CC^T
\]

Since \( CC^T \) is a positive semidefinite matrix

\[
\therefore V(\beta_{\text{OLS}}) \leq V(\mathbf{b}), \quad \text{Only equal if } C = 0.
\]

Now let's look at the \( V(\mathbf{b}) \):

\[
\beta_{\text{OLS}} = (X^TX)^{-1}X^Ty = Ay
\]

\[
b = (A + C)y = Ay + Cy = \hat{\beta}y + Cy = (A + C)X\beta = (A + C)X\hat{\beta} + (A + C)\epsilon
\]

If \( \mathbf{b} \) is unbiased,

\[
E(b) = (X^TX)^{-1}X^TX\beta + CX\beta = (I + CX)\beta = \beta
\]

If and only if \( CX = 0 \)
Collinearity

With the additional explanatory variables in the multivariate regression model, a new problem arises. What happens if the explanatory variables are correlated? Remember how we interpreted $\beta$?

$$
\beta = \frac{\partial y}{\partial x_i} \bigg|_{x_{i-1}=0}
$$

In other words, $\beta$ is the change in the dependent variable for a change in an independent variable, *holding all other independent variables constant*. There was no relationship between independent variables, so a change in $x_1$ left $x_2$ constant. However, multicollinearity says let’s change $x_1$ but $x_2$ changes at the same time because they are correlated. Our traditional measure of $\beta$ is out the window because we aren’t holding all other variables constant. So, a significant problem with multi-collinearity is that we don’t know what $\beta$ means.

How do we know if multi-collinearity exists? The following indicate when collinearity may exist between regressors (the Xs):

1. In a typically model that explains the dependent variable well, we expect there to be a high $R^2$ and significant t-statistics. However, when multi-collinearity exists, we may see a high $R^2$ but few, if any, significant t-statistics. This may imply that there is a high $F$-statistic but no independently significant explanatory variables. In other words, there is a high degree of explanatory power for the model, but independent variables don’t matter.

   What non-sense!! Again, multi-collinearity makes it difficult to interpret the model.

2. Relatively high simple correlations between explanatory variables *may* indicate multi-collinearity. However, conclusions concerning multi-collinearity based solely on simple correlation between independent variables may be misleading.

3. Estimators may be very sensitive to the addition or deletion of a few observations or the deletion of an apparently insignificant variable.
There are different ways of dealing with multi-collinearity. When appropriate, we may opt to make a ratio of the two independent variables that are collinear. For example, capital and labor may be highly correlated in a production function. One way around this is to divide capital by labor to express production as a function of the capital-labor ratio. A second way to deal with collinearity is to obtain additional observations. However, multi-collinearity is a problem with the data. So, we have to get better data or hope to deal with it by using a transformation.

In sum, multicollinearity is not a violation of the basic assumptions concerning the CLRM. OLS is still BLUE when assumptions A1-A5 are satisfied. The problem may be that estimators are too imprecise to yield meaningful results.

Elasticity

From your principle’s course, you learned about elasticity. Elasticity measures the responsiveness of the change in a dependent variable when an independent variable changes. For example, how does quantity respond to a change in price? It can be either responsive or non-responsive. To measure elasticity, we seek a unitless measure of responsiveness since the choice of base will alter the measure of elasticity. Our measure of elasticity is

\[ \Sigma = \frac{\% \Delta q}{\% \Delta p} = \frac{dq}{dp} \frac{p}{q} = \frac{dq}{dp} \frac{p}{q} \]

Alternatively, elasticity can be expressed

\[ \Sigma = \frac{\% \Delta q}{\% \Delta p} = \frac{\ln q}{\ln p} = \frac{d \ln q}{dp} \]

We can use our regression coefficients to obtain elasticity at the means of the dependent and independent variables.
Elasticity can also be measured at specific prices and quantities.

\[
\Sigma = \frac{dq}{dp} \frac{p}{q}
\]

\[\text{However, } \frac{dq}{dp} = \beta
\]

\[\therefore \Sigma = \beta \frac{p}{q}
\]

Elasticity can also be measured at specific prices and quantities.

\[
\Sigma = \beta \frac{p_i}{q_i}
\]

**Example**

We now return to our previous earnings example. However, we’ll add an additional regressor, education.

<table>
<thead>
<tr>
<th>Annual Earnings</th>
<th>Years Work Experience</th>
<th>Planned College Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>$20,000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$15,000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$25,000</td>
<td>2</td>
<td>.25</td>
</tr>
<tr>
<td>$35,000</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>$12,500</td>
<td>1</td>
<td>.25</td>
</tr>
<tr>
<td>$14,750</td>
<td>1.5</td>
<td>.5</td>
</tr>
<tr>
<td>$17,500</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$19,000</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>$21,000</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>$45,000</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>$36,500</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>$32,000</td>
<td>8</td>
<td>3.5</td>
</tr>
<tr>
<td>$18,000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$47,852</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>$22,000</td>
<td>2</td>
<td>.5</td>
</tr>
<tr>
<td>$25,124</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$31,000</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>$21,250</td>
<td>1.5</td>
<td>.5</td>
</tr>
<tr>
<td>$14,200</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$39,000</td>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>
Throughout the remainder of the course, we’ll use the statistical software *Shazam*.

Shazam is a command driven system that is very user friendly. The dependent variable earnings was regressed on the independent variables work experience and planned college education. The following is the Shazam regression output:

```
TYPE COMMAND
:_UNIT 4 IS NOW ASSIGNED TO: a:utpb.dat

TYPE COMMAND
:_
...SAMPLE RANGE IS NOW SET TO: 1 20

REQUIRED MEMORY IS PAR= 2 CURRENT PAR= 500
OLS ESTIMATION
20 OBSERVATIONS DEPENDENT VARIABLE = EARN
...NOTE..SAMPLE RANGE SET TO: 1, 20

R-SQUARE = 0.8665 R-SQUARE ADJUSTED = 0.8508
VARIANCE OF THE ESTIMATE-SIGMA**2 = 0.16581E+08
STANDARD ERROR OF THE ESTIMATE-SIGMA = 4072.0
SUM OF SQUARED ERRORS-SSE= 0.28188E+09
MEAN OF DEPENDENT VARIABLE = 25584.
LOG OF THE LIKELIHOOD FUNCTION = -192.991

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>ESTIMATED COEFFICIENT</th>
<th>STANDARD ERROR</th>
<th>T-RATIO</th>
<th>17 DF</th>
<th>P-VALUE</th>
<th>CORR. COEFFICIENT</th>
<th>PARTIAL STANDARDIZED COEFFICIENT</th>
<th>ELASTICITY AT MEANS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WE</td>
<td>2654.8</td>
<td>688.8</td>
<td>3.854</td>
<td>0.001</td>
<td>0.683</td>
<td>1.0133</td>
<td>0.4462</td>
<td></td>
</tr>
<tr>
<td>PED</td>
<td>-505.57</td>
<td>1509.</td>
<td>-0.3351</td>
<td>0.742</td>
<td>0.081</td>
<td>-0.0881</td>
<td>-0.0346</td>
<td></td>
</tr>
<tr>
<td>CONSTANT</td>
<td>15053.</td>
<td>1359.</td>
<td>11.07</td>
<td>0.000</td>
<td>0.937</td>
<td>0.0000</td>
<td>0.5884</td>
<td></td>
</tr>
</tbody>
</table>
```

**Interpretation**

The estimated regression model is:

\[ \hat{E} = 15,053 + 2,656WE - 506PED \]

Interpreting the earnings model is straight-forward. Start with the t-statistics. The t-statistic for the constant demonstrates that when work experience and planned college education are jointly zero a worker, on average, earns $15,053. Each additional year of work experience adds an additional $2,656 and work experience continues to be statistically significant. However, the t-statistic for planned college education has a negative sign and is statistically insignificant. In
other words, planned education doesn't matter in explaining wages. This is an interesting result because variables that are statistically insignificant give us important information. In the finance literature, we say the dog that doesn’t bark tells us as much as the dog that does. In other word, the variable that is insignificant may be as important as the variable that is significant.

Now let’s look at the $R^2$. We would expect that inclusion of an additional regressor would increase $R^2$ because it generally increases the sum of squared residuals. However, in this case, the $R^2$ declines (.93 to .87). In chapter 7, we’ll learn that inclusion of an irrelevant variable, such as planned college education, causes us to lose a degree of freedom and a loss in efficiency. However, the estimates will still be consistent and unbiased.

Shazam also produces some additional statistics. The fourth column displays the *p*-value. The *p*-value gives us the probability that the null hypothesis is true. Recall the null is that the coefficient equals zero. We already know that work experience is significant and planned college education is insignificant (the *t*-statistics told us). Hence, we already know that the probability that the work experience regressor is zero is low and the probability that planned college education equals zero is high. So, these conclusions are confirmed by the *t*-statistics. In fact, there will always be an inverse relationship between *t*-statistics and *p*-values. High *t*-statistics imply low *p*-values, and low *t*-statistics imply high *p*-values.

The fifth column displays what is called the partial correlation coefficient. The partial correlation coefficient between earnings and work experience or between earnings and planned college education measures the effect of work experience on earnings net of the influence of planned college education. Likewise, the partial correlation coefficient for planned college education measures the relationship between planned college education and earnings in isolation from work experience. As you might expect, partial correlation coefficients are bounded between one and negative one where values closer to one or negative one indicate that there is a strong relationship between earnings and the variable after the effect of all other variables have been accounted for. A frequent use of partial correlation coefficients is in stepwise regression.
Stepwise regression adds variables to a model to maximize what is called adjusted $R^2$. While stepwise regression may seem to make sense (ceteris peribus, we’d like a high $R^2$), it has little value in economics because we construct a model theoretically and support it with the evidence. Stepwise regression gives a high $R^2$, but is divorced from theory, the antithesis of what we are trying to do in econometrics.

The sixth column gives us the standardized coefficient. Standardized coefficients describe the relative importance of the independent variable in a regression model. Large standardized coefficients suggest a variable is important, while small, standardized coefficients suggest a variable is not important. We already know that work experience is important while planned college education is not important. This is supported by the standardized coefficients. Finally, the last column measures the elasticity at the means.
Homework 3

1. Using matrix notation, demonstrate that given \( y = X\beta + \varepsilon \)

\[
\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y
\]

2. Using matrix notation, demonstrate that given \( y = X\beta + \varepsilon \)

\[
R^2 = \frac{\hat{\beta}^T X^T X \beta}{Y^T Y}
\]

3. Demonstrate that if \( E(\varepsilon) = 0 \), OLS is unbiased.  *Hint:* Use matrix notation.

4. Demonstrate that the OLS error term can be written as

\[
e = (I - (X^T X)^{-1} X^T) y = My
\]

5. Demonstrate that \( M \) is symmetric and idempotent.

6. Use data set EX42 to answer the following:
   A. What is the estimated regression for the model

   \[
   \hat{IP} = \alpha + \hat{\beta}_1 FYGN3 + \hat{\beta}_2 FM2 + \hat{\beta}_3 PW + \varepsilon
   \]
   
   B. Which variables are significant in explaining industrial production, IP?
   
   C. Is the model significant in explaining IP?
   
   D. What is the variation in IP that can be explained by variation in the model?

7. Use data set EX45 to answer the following:
   A. What is the estimated regression equation for the model

   \[
   RTDR = \alpha + \hat{\beta}_1 IVRDR + \hat{\beta}_2 FYCP + \hat{\beta}_3 LEH + \hat{\beta}_4 PUCD + \varepsilon
   \]
B. Which variables are significant in explaining retail sales, RTDR?

C. Is the model significant in explaining RTDR?

D. What is the variation in RTDR that can be explained by variation in the model?