Regression (cont.)

2.3. The Simple Linear Regression Model

This section contains a distressing number of Greek symbols and numerous equations. For reference, a list of symbols and equations is given in Appendix A.

We have described the $\hat{y}$-line as an estimator of the population average of Y as a function of x. We will denote the corresponding population average as

$$\mu_{Y|x}$$

to emphasize that the population average of Y will now be considered as related to a known value of x.

Recall that we decided for the M&M's to force the estimated population average of Y given knowledge of x to lie on a straight line. However, in general, linear regression is not limited to use of a straight line. We can and do sometimes fit a curve rather than a straight line, but for now we will stick to straight lines for simplicity. (In fact, the term "linear" regression does not refer to fitting a straight line; it refers to fitting a linear *function*. A polynomial function is a linear function, while an exponential function is not.)

Like any straight line, our $\hat{y}$ line will need a y-intercept, denoted b_0, and a slope, denoted b_1. Thus, we can define our y-hat line as:

$$\hat{y} = b_0 + b_1 x.$$

On the other hand, since we are assuming that $\mu_{Y|x}$ also falls on a straight line. Hence, we will need a corresponding y-intercept, denoted β_0, and a slope, denoted β_1. Thus, we can define $\mu_{Y|x}$ by the equation:

$$\mu_{Y|x} = \beta_0 + \beta_1 x.$$

We consider the intercept, b_0, to be an estimate of the population intercept, β_0, and the slope b_1 as an estimate of the population slope, β_1. The symbol β is the Greek letter "beta".

We will seldom know the true value of $\mu_{Y|x}$, β_0 or β_1. We do not know their true values even after we get a sample and form estimates. There are many other possible samples that we could have gotten and each sample would yield a different estimate of the intercept and the slope, producing a different $\hat{y}$-line. Conceptually, the line defined as $\mu_{Y|x}$ is the "true" regression line: it remains unchanged from sample to sample.

We summarize the relationship between Y and x in the following *linear regression model*:

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

This model breaks the random variable Y into two pieces: a non-random piece, $\beta_0 + \beta_1 x$, and a random piece, ε. The random variable ε captures the *unpredictable variation* contained in Y, while the term $\beta_0 + \beta_1 x$ captures the *predictable variation* in Y when x is known. Since we have defined $\mu_{Y|x} = \beta_0 + \beta_1 x$, we can also write the simple linear regression model as:

$$Y = \mu_{Y|x} + \varepsilon.$$

In what follows, we will assume that ε follows a Normal distribution with a mean of zero and with standard deviation defined as σ_ε.

If there is a relationship between the population average of Y and x, then some of the variability of Y, measured by σ_Y, is transferred from being random into a restatement of the population average of Y. The population average of Y now becomes variable rather than fixed, but it is not variable in a random sense: it varies strictly with x. Thus, the variable ε will have a smaller standard deviation than does Y when x is unknown because some of the overall variation in Y has been absorbed into the varying population average.

If we did not have the ε in the model, then the model could not reproduce the scattered nature of most (x,y) plots of interest. The ε supplies the energy which boots the data points off of the line given by $\mu_{Y|x} = \beta_0 + \beta_1 x$. In what direction does ε kick the points? It only kicks up or down, not diagonally. Why only in the vertical direction? Because Y lies on the vertical axis, and ε is an expression of the variability of Y that remains after providing for a varying population average.

Please note that simply because we decide to run a regression analysis on two variables does not mean that there is in fact a relationship between them. If there is actually no relationship between x and Y, then this can be expressed in the linear model by setting

$$\beta_1 = 0.$$

In that case, our linear model is reduced to the statement:

$$Y = \beta_0 + \varepsilon,$$

for which $\beta_0 = \mu_Y$ and ε is a variable with a Normal distribution centered at $\mu_\varepsilon=0$ and with a standard deviation $\sigma_\varepsilon = \sigma_Y$. In that case, Y is left unchanged except for the superficial division of Y into two components: a fixed component equal to the population average and a random component with a Normal distribution centered at 0.

2.4 Using a Computer-generated Regression Analysis

We will now make a study of the Excel regression output. We will focus on learning how to perform the following:

(a) locate on the Excel output the estimated intercept and slope coefficients, b_0 and b_1;
(b) use these to find the y-hat value for a particular value of x;
(c) locate on the output the estimated standard deviation of Y given knowledge of x;
(d) construct an approximate prediction interval for Y given knowledge of x;
(e) measure our net gain from knowledge of x;
(f) pause to contemplate the deeper difficulties of interpreting a regression analysis;
(g) use a specialized t-test to check that x indeed contains information about Y;
(h) locate a 95% confidence interval for the "population" slope, β_1;
(i) understand the derivation of R^2 ("r-squared"), a common regression diagnostic;
(j) find the sample correlation coefficient from R^2 and interpret it.

It is possible to perform regression analysis in several ways in Excel, but the easiest is to use the Regression selection under Data Analysis from Tools. If Data Analysis is not listed, then go to Add-Ons under Tools and click on Analysis ToolPak.

The regression output for the M&M data set is given below in Table 2.4.1. The output has been edited to make it more readable.

Table 2.4.1. OUTPUT FOR REGRESSION OF WEIGHT ON COUNT

Regression Statistics	
Multiple R	0.52
R Square	0.27
Adjusted R Square	0.25
Standard Error	1.64
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Sign F*
Regression	1	38.25	38.25	14.23	0.0006
Residual	38	102.15	2.69		
Total	39	140.4			

	Coef.	*St. Error*	*t Stat*	*P-value*	*Low*	*Upper*

Intercept	34.88	4.35	8.03	0.0000	26.09	43.68
X Variable 1	0.77	0.20	3.77	0.0006	0.36	1.18

When using the Regression command in Excel, it is always a good idea to click on the *Line Fit Plots*. This will instruct Excel to generate a scatterplot of the data with the regression line added to it. Unfortunately, Excel produces a nasty-looking regression plot, so it needs some work it looks presentable.

We will now work, step-by-step, through points (a) through (j) listed above.

(a) To find the values of b_0 and b_1, look at the last two rows under the column labeled "Coef". The value of b_0 is given as 34.88 and the value of b_1 as 0.77. Hence, the equation of the y-hat line is,

$$\hat{y} = 34.88 + 0.77\ x.$$

(b) If we wanted to estimate the population average of Y when X=20, we can now do so by finding the corresponding y-hat value:

$$\hat{y}_{x=20} = 34.88 + 0.77\ x = 34.88 + 0.77\ (20) = 50.28.$$

Sometimes people like to use this estimated average value of Y as a "prediction" of the value that Y will take on. We have previously commented that if we want to try to predict the value of Y for a given value of x, the best thing to do would be to give a prediction *interval* so that our predictions will not always be wrong. To find a prediction interval, we need to know how variable Y is even when x is known. This is measured by using the sample standard deviation of Y given x, denoted $s_{y|x}$.

(c) The Excel output calls $s_{y|x}$ the "Standard Error" and it is found in the fourth row of the first section of the output, marked Regression Statistics. There, we see that $s_{y|x}$ = 1.64. Locating this value is somewhat tricky because there are two other items marked "standard error" in the Excel output: those are associated with the estimated intercept and slope.

> In other statistical software packages, the value of $s_{y|x}$ appears in different locations from that shown here and under different names. In case of confusion, it is useful to know that this value can also be found from a value located in the section denoted ANOVA. If you locate the intersection of the row marked *Residual* and the column marked *MS* in the ANOVA section, you will find the value **2.69**. If we take the square root of 2.69, we get 1.64, which is the value of $s_{y|x}$. What we have found is the sample variance of Y given knowledge of x, $s_{y|x}^2$, which is also called the "mean square residual", or *MS Residual*. Hence, if you are not sure where to find $s_{y|x}$ on a regression output, you can "simply" locate the omnipresent ANOVA section and take the square root of the mean square residual. For those of you who must know, ANOVA stands for "Analysis of Variance" and that is a Whole Other Kettle of Fish, or WOKOF.

(d) We begin the construction of a prediction interval by centering the interval at the corresponding $\hat{y}$ value. For example, if a prediction interval for the net weight of a bag of Peanut M&M's containing 20 M&M’s, then we would center the interval on $\hat{y}_{x=20}$=50.28 grams.

We now need to find a plus/minus distance to move out from 50.28 in hopes of capturing the actual net weight of the bag. As a simple approximation, we can use ±2 $s_{y|x}$, where $s_{y|x}$ measures the vertical variability left over after taking into account the fitted regression line. Hence, our approximate 95% prediction interval for the net weight of a bag containing 20 Peanut M&M’s is

$$\hat{y} \pm 2\ s_{y|x} = 50.28 \pm 2\ (1.64) = 50.28 \pm 3.28 \text{ grams.}$$

This is first approximation is a bit rough, but is quite useful because it allows us to construct a quick 95% prediction interval after only a glance at the software output. Technically, we should use an interval which is looks like a more complicated version of the prediction interval we used previously, but that interval is clumsy to implement unless the software is set up for it and for some reason Excel is not. Therefore, we will swallow our pride and just use

$$\hat{y} \pm 2\ s_{y|x}$$

as our approximate 95% prediction interval.

(e) Has our knowledge that Count=20 gained us any advantage in terms of prediction? The answer is "Yes" and this can be seen by comparing the sample standard deviation of Y with and without knowledge of x. Without knowing the Count, we found a sample standard deviation for Y of s_y=1.90, but knowledge of x reduced this to $s_{y|x}$ = 1.64. Thus, knowing the number of M&M's allows us to reduce our estimated standard deviation of the weight by a factor of

$$(1.90 - 1.64) / 1.90 = 0.137 = 13.7\%.$$

Basically, this means that our prediction interval for the weight of the next bag will be reduced by about 13.7% by knowing the count.

(f) Early in my career as a teacher, I presented a similar M&M regression analysis to a class that contained two remarkable students. If those students were in this class, here is what they would say. The first student would complain that there was something wrong with the estimated intercept of 34.88 grams. He would point out that this suggests that an empty bag would weigh about 34.88 grams, which makes no sense. At that time, I replied with some well-intentioned nonsense because I failed to comprehend the extent of the problem.

The second student would then put forward the claim that something is wrong with the slope of 0.77 grams. That second student is now a professor of economics at Baruch College; I do not know the fate of the first student. The second student would argue as follows. The slope coefficient of 0.77 grams is routinely defined as the change in the estimated population average for a unit increase in x. (Notice that I have not made that particular claim - I learned my lesson from those two.) In this case, this means that we are saying that if we add one Peanut M&M that on average the weight will increase by about 0.77 grams.

This student then did a very original thing. He went back to the original data and found that on average a Peanut M&M weights about 2.4 grams. This can be found from Table 2.1 by seeing that there are about 21 M&M's per bag on average. If we divide the average weight by the average count, we get 51.25 / 21 = 2.4 grams.

He presented these two numbers to me, 0.77 and 2.4, and claimed that they should have been the same. It took me several long minutes to grasp his point. If we randomly pick single M&M's and add them to a bag, we will find (on average, over many repetitions) that the weight increases by 2.4 grams per addition. But, what are we to make of our slope, then? Doesn't our slope assert that the bag will only increase in weight on average by 0.77 grams?

Suddenly, I realized that the first student was right, as well. The slope was far too small and this had raised the intercept to a ridiculous level. I puzzled and struggled over this quandary for weeks, as I recall, before I came to understand it.

The problem is that we had confused what it meant to "increase the number of M&M's." If we are going to add an M&M to a bag, then 2.4 grams is the correct figure. That means that if we take many bags with 20 M&M's and add one more M&M to each, we should **not** use our regression line to estimate that the average weight will increase by a mere 0.77 grams. Clearly, the average will increase by something closer to 2.4 grams.

On the other hand, suppose we are considering two groups of factory-fresh bags: one set contains 20 M&M's and the other contains 21 M&M's. What will be the average difference between their average net weights? 2.4 grams? Or 0.77 grams? The answer will be around 0.77 grams, not 2.4 grams.

Why? I believe that the answer is that the M&M's are put into the bags by volume, not by count. The result is that the number of M&M's in a bag is related to the size of the M&M's. I suspect that what happens is that bags with more M&M's tend to have smaller M&M's, while those with fewer M&M's tend to have larger M&M's. This reduces the slope of the relationship between the count and the weight.

On a deeper level, our problem is that when we go to add an M&M to a bag, we are introducing a new feature to the situation that did not exist for our original data. Previously, we did not act upon the number of M&M's in the bags, we passively *observed* the number of M&M's and the corresponding net weights. The results of our regression analysis should only be applied to the factory-fresh bags because that is all we dealt with before.

The above facts are more difficult to digest than Peanut M&M's, themselves. These complexities made me feel ill for weeks when I first encountered them - particularly because my extensive education in statistics had not covered this "little" problem.

It turns out that this sort of problem often comes up. There is a critical difference between whether the "changes" in the x-value are determined by experiment or merely observed. Regression users who fail to appreciate this difference are an imminent danger to themselves and those around them because they may predict that adding an M&M will on average increase the weight by about 0.77 grams when any fool can see that the answer is 2.4 grams. If we perform regression analysis without comprehending this point, our regression output will be a genuine danger, akin to providing witch-hunters in Salem, Massachusetts in 1650 with a map showing the location of all dwellings that housed an adult female, a black cat, and a broom.

(g) We have previously noted that there is only one case in which we would know the correct value of β_1 and that is when there is no relationship between the population average of Y and X. In that case, it must be that $\beta_1 = 0$.

We have also noted that it is not certain that the number of M&M's should help to predict the weight of a bag since it might be that the bags are filled solely by weight. If that turned out to the case, then it would imply that $\beta_1 = 0$.

Finally, we should be aware that if we try to use knowledge of x to help predict Y, but x is not really related to Y, then we will in turn actually damage our predictions. It is easy to think that regression is a "Can't hurt, might help," sort of situation, but it is not. Garbage-in equals even-more-garbage-out.

For these reasons, all regression software automatically performs a *hypothesis test* to check that x is actually related to the population average of Y. The hypothesis tests compares two hypotheses: the null hypothesis and the alternate hypothesis. Here, the null hypothesis states that there is no relationship between x and the average of Y, which is shortened to the equivalent statement that $\beta_1= 0$. The alternate hypothesis claims that $\beta_1\neq 0$. That's not very specific, is it?

These two hypotheses are often written, then, as

$$H_O\text{:}\ \ \beta_1= 0$$
$$H_A\text{:}\ \ \beta_1\neq 0.$$

Essentially, we will be checking to see if the sample argues strongly against the claim made by the null hypothesis that $\beta_1 = 0$. If it argues strongly against it, then we reject the null hypothesis.

When should we reject the null hypothesis? We reject if b_1 is far from zero! How far is far? We compute the distance by dividing b_1 by the estimated standard deviation of b_1, which is found under the "St. Error" column and in the X-Variable 1 row in Table 2.4.1. It has a value 0.20.

The resulting ratio, 3.77, is called a t-statistic and is located in the next column of X-Variable 1 row. As a rule of thumb, if the t-statistic is less than -2 or greater than +2, we reject the null. More precisely, we would look up the critical value in the t-table for n-2 degrees of freedom. From Excel, the critical value can be found be entering the Tinv() function with arguments 0.05 and 38 into a cell, giving

=Tinv(.05, 38)

which yields an answer of 2.024394. Since **3.77** is greater than 2.024394 we reject the null.

In brief, here is what is happening. There are millions of possible samples containing 40 bags of Peanut M&M's. Each of these samples will yield its own estimated slope, so there is a new population to consider, one consisting of millions of slope estimates. Under a few assumptions, the population average of the slope estimates will equal the true population slope. Hence, if the null hypothesis is true, the population average of the slope estimates will be zero.

In addition, each of these millions of samples will yield its own t-statistic, computed as above. When the null hypothesis is true, these millions of sample t-statistics will have a population distribution given by a t-distribution with 38 degrees of freedom. This will look a lot like a standard Normal distribution (Z), but it will be a little wider in the tails. In particular, 95% of that t-distribution will fall between -2.024394 and 2.024394, rather than -1.96 and 1.96 of the Z distribution.

Now, if our particular sample yields a t-statistic that fall outside that interval, then we will take this as evidence against the null hypothesis. Is this a fool-proof method? No, in fact, it will give us a 5% probability of rejecting a true null hypothesis. This is called the alpha level of the test.

Why did we input the values 0.05 and 38 as arguments to the Tinv function? The 0.05 shows that we are using an alpha level of 5%, which is the probability of rejecting the null hypothesis when it happens to be true. The 38 denotes the degrees of freedom, which in simple regression is (n - 2) = 40 - 2 = 38.

A more modern method of making use of the t-statistic to assess our competing hypotheses is to examine the p-value. This is shown after the t-statistic in the Excel regression output and has a value of 0.0006 in this case. If the p-value is less than 0.05, we will reject the null hypothesis when using an alpha=5% level test. Hence, we would reject the null hypothesis in this case.

What does this p-value mean? In this instance, the p-value gives the probability getting a t-statistic of greater than **3.77** or less than **–3.77** **if** the null hypothesis were true. Since this probability is so small, it seems that the null hypothesis has not give a good accounting for the observed outcome so the null is rejected.

(h) It often happens that we would like an interval estimate of the population slope, β_1. Excel kindly provides a 95% confidence interval for β_1 in the next two columns of the X-Variable 1 row. In Table 2.4.1, we see that this is an interval spanning 0.36 to 1.18. Please note that this interval does not include zero, which is another justification for the rejection of the null hypothesis.

An aside: The original Excel regression output repeated the confidence interval values in an additional pair of columns. I deleted the repetition in Table 2.4.1. Those values would be used to show the endpoints if a confidence level other than 95% was selected.

(i) The derivation of R^2

One of the common goals in a regression analysis is to reduce the uncertainty about future observations of one variable (Y) by making use of information contained in a second variable (X). If we do not know X, then our measure of uncertainty of the Y variable is the sample variance, s_y^2 , or the sample standard deviation, s_y, where

$$s_y^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2 .$$

Alternatively, we could drop the 1/(n-1) term and consider just the sum of squared deviations term, which is denoted as SST, where

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 .$$

The initials SST stand for *sum of squares total.* Like the sample variance, SST is a measure of the variability or uncertainty contained in the Y variable. In the Excel regression output, the SST value is given in the last row (Total) of the SS column of the ANOVA section. Referring back to Table 2.4.1, we see that for the regression of the Weight on the Count for M&M's, SST=140.4.

When we have information about Y from knowing X, we can decompose SST into the sum of two terms: the first being a source of predictable variation and the second a source of unpredictable variation. This decomposition is written:

$$SST = SSR + SSE.$$

SSR is the *sum of squares from regression* and represents the variation in Y which is associated with knowing x,

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \ .$$

The SSR value is given in the Excel output in the Regression row, SS column of the ANOVA section. For the M&M problem, we find that SSR = 38.25.

SSE is the *sum of squares from the residuals* (errors) and represents the uncertainty still present concerning Y even when x is known,

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ .$$

The SSE term is the remaining value in the SS column, in the Residual row. In Table 7.1, we find SSE=102.15.

The SSE term will be relatively small when the y-hat values lie close to the y-values, which takes place when the data lies almost on a straight line. In that case, SSR will almost be as large as SST, so that the ratio of SSR/SST will be close to one. We denote that ratio as R^2 , where

$$R^2 = \text{SSR} / \text{SST}.$$

R-squared is found in the Excel regression output in the second row of the Regression Statistics section. In Table 2.4.1, R-square = 0.27. As a check, we can compute,

$$R^2 = \text{SSR} / \text{SST} = 38.25 / 140.4 = 0.272.$$

R^2 can range between 0 and 1. It is often used as an indication of the importance of the regression analysis, but care must be taken. In some applications in econometrics and in any application in sociology, an r-squared of 0.27 would be considered to be remarkably high, while in other situations this would be disappointingly low.

R-squared is often described as the **coefficient of determination**. This terminology can be misleading because it suggests that a causal or deterministic relationship has been demonstrated between x and Y, whereas what may have been demonstrated is a correlation between the two.

It is only reasonable to describe R-squared as the "coefficient of determination" when the regression data resulted from a controlled, laboratory study. In such a setting, it may make sense to think of the x variable as determining the Y variable (or its average, anyway). For example, the amount of dioxin administered to rats will determine to a large extent the <u>average</u> life expectancy of the rats.

However, you will seldom encounter that sort of data in the ordinary commercial world. When regression is applied to business data, the X and Y variables do not arise from the pristine environment of a laboratory, but from a rough and tumble jungle. In the marketplace, it is easy to find correlations, but rare to find simple, exploitable causality.

Despite these warnings, when your professors ask for the meaning of R-squared, they will expect to be told that it is the coefficient of determination. Just tell them that: they will be happy and will not bother you about it. Be kind and don't bother them, either.

(j) One of the standard measures of relationship between two variables is the sample correlation coefficient, r. This measures the extent to which a plot of the two variables would yield points clustered around a straight line. A plot showing the points clustered tightly around a rising straight line would have r close to +1, while r would be close to -1 if the line was sloped downwards. On the other hand, a plot showing a ball of points would have an r of close to 0.

This r is related to R-squared. In fact, R-squared is the square of r. We can find the value of r from the regression output of Excel by taking the square root of R-squared and attaching the sign of b_1.

Alternatively, r can be found in Excel from the built-in function CORREL.

In terms of the underlying populations, we consider r to be an estimate of the population correlation coefficient, denoted as the Greek rho, *p*.

Exercises 2.4.

Table 2.4.2

Store (No.)	Space (Linear Feet)	Sales ($ Last Month)
1	6	444
2	9	543
3	9	709
4	7	571
5	10	908
6	8	801
7	5	565
8	10	828
9	11	941
10	9	667
11	6	647

1. Use the data in Table 2.4.2 to answer the following questions about the regression of Sales (Y) on Space (X).

(a) Write out the equation of the Regression Model in symbols and then explain it in words.

(b) Regress Sales(Y) on Space(X) using Excel's Data Analysis Regression program. Clean up the output so it is readable.

(c) Prepare a nicely-formatted regression plot of Sales(Y) vs. Space(X). The plot should use points (not bricks) and a regression line.

(d) What was the value of the estimated slope and intercept?

(e) Write out the equation of the y-hat line for this problem.

(f) For a store with 9 feet of linear Space, give a point forecast of Sales.

(g) What was the value of the estimated standard deviation of Y given X?

(h) For a store with 9 feet of linear Space, give an approximate 95% prediction interval for Sales.

(i) In which case would the above forecast be legitimate: (1) a Store which already has 9 linear feet or (2) a Store which had 5 linear feet but has added on 4 more feet to try to increase Sales? Explain!

(j) What conclusion would be reached for the usual regression t-test? Explain how using the t-value or p-value would have led to the same conclusion.

(k) State the 95% confidence interval for the population slope and decide if this value concurs with the result of the t-test in (j).

(l) Give the value of R^2 and show how it can be computed from the values in the ANOVA section.

(m) Compute r, the sample correlation coefficient, from R^2.

Appendix A: Glossary

	X
μ_x	Population average of the X population
σ_x	Population standard deviation of X population
$\bar{x}$	Sample average from the X sample; estimator of μ_x
s_x	Sample standard deviation from the X sample; an estimator of σ_x
	Y
μ_y,	Population average of the Y population
σ_y	Population standard deviation of Y population
$\bar{y}$	Sample average for the Y sample; an estimator of μ_y
s_y	Sample standard deviation from the Y sample; an estimator of σ_y
	Y|x
$\mu_{Y\|x}$	Population average of Y given knowledge of the x value
β_0	The population intercept in the regression model
β_1	The population slope in the regression model
ε	The random error; produces scattering of Y around $\mu_{Y\|x}$.
$\sigma_{Y\|x}$ or σ_ε	Population standard deviation of Y given knowledge that X=x
$\hat{y}$	Estimated value of $\mu_{Y\|x}$
b_0	Estimated value of β_0
b_1	Estimated value of β_1
$s_{y\|x}$	Estimated standard deviation of Y|x; *standard error* in Excel reg. output
s_{b1}	Estimated standard error of b_1;
e_i	The difference between y_i and $\hat{y}_i$.

	Misc.
p	The population correlation coefficient
r	The estimated correlation coefficient
R^2	The square of r; also termed the coefficient of determination
SST	The sum of squared deviations around $\bar{y}$
SSE	The sum of squared deviations around $\hat{y}$
SSR	The sum of squared deviations between $\hat{y}$ and $\bar{y}$

Key Equations

$Y = \beta_0 + \beta_1 x + \varepsilon$	The simple linear regression model
$Y = \mu_{Y\|x} + \varepsilon$	Alternative statement of the linear regression model
$\mu_{Y\|x} = \beta_0 + \beta_1 x$	The true regression line: expresses $\mu_{y\|x}$ as a linear function of x
$\hat{y} = b_0 + b_1 x$	The estimated regression line
$\hat{y} \pm 2 s_{y\|x}$	Approximate 95% prediction interval for Y given x
SST = SSR + SSE	Shows the decomposition of the total variance of Y
R^2 = SSR / SST	Expresses the squared correlation to the ratio of SSR to SST