Lecture 3

Simple Linear Regression

We've spent our first couple of lectures talking about the relationship between variables in different ways using correlation, linearity and other factors. Now, we want to begin to develop models of the data. What if we have an observation of one of the variables not in our original data set? How do we use the information on the relationship between the variables to predict what the observation of the other variable is most likely to be? Or provide a range of such values? To begin to do this, we need to generate a model equation. We'll start with the simplest case of a linear equation that depends on only one other variable. We'll look at this in some detail in the coming weeks and then extend to other scenarios: multiple variables, non-linear models, categorical variables, and so on.

Our linear model is going to take the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

We will estimate this model by finding the descriptive linear model that best fits the data:

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

The "hat" notation is to indicate that this is an estimate. You can see that we follow the Greek letter for the parameter and the English/Latin letter for the descriptive statistic or the estimate, i.e. $b_i = \hat{\beta}_i$, the coefficients in the model are estimates for the model parameters. The ϵ in the inferential model is the error. It is assumed to have a mean of zero and a constant standard deviation. We'll discuss below how we estimate the standard deviation of these errors.

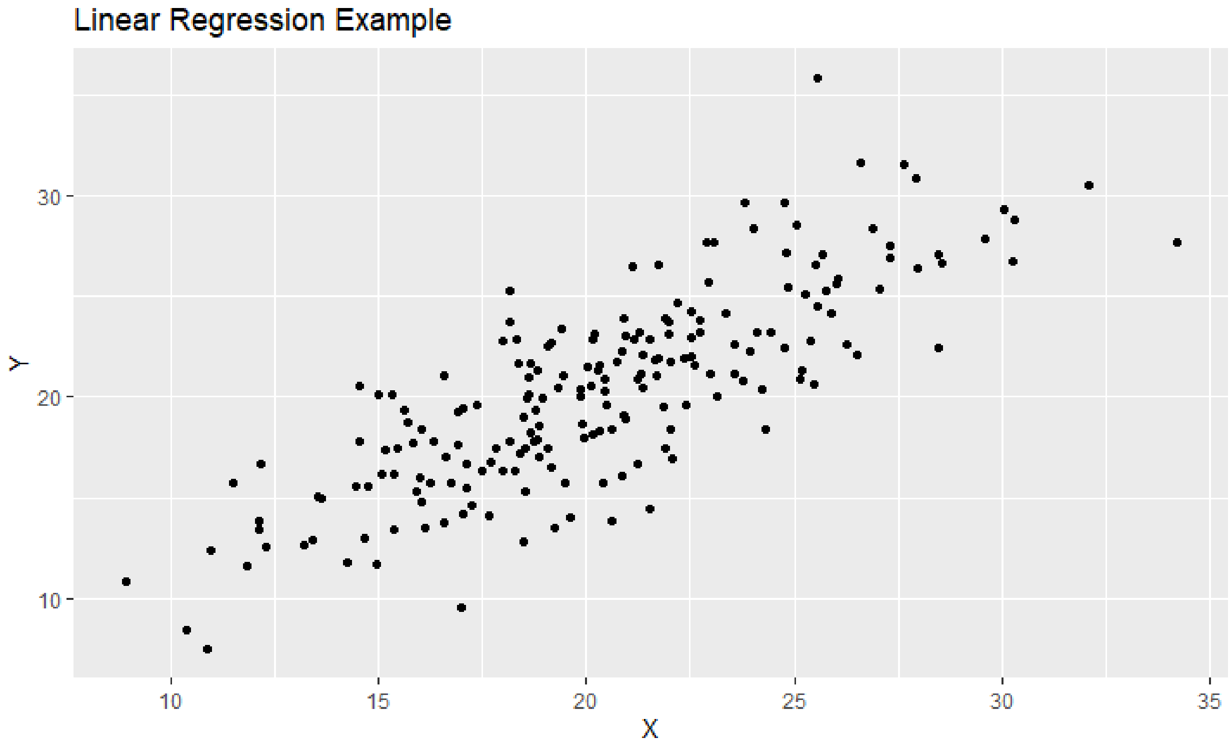
A note on terminology. The x variable in the model is the "input" variable. It may be referred to as the independent variable, the explanatory variable or the predictor variable. In practice, this variable may occur first in time, it may be presumed to be causal, or it may just be easier to measure. While causation may be a factor in our models, it is not required. You've often heard the phrase "correlation does not equal causation". Causation has to be established through experiment or other means, but while we can use causation to identify our explanatory variable if it is available, we do not depend on it. You may also hear it referred to as a "proxy" for the y variable.

The y variable may be referred to as the dependent variable, the response variable or the prediction. This may occur later in time than the explanatory variable, it may be caused by the independent variable, or it may simply be more difficult to measure. We can think of it in function terms as the output variable.

When we plot these variables on a scatterplot, the order does not matter to detect correlation, but it does matter when we build a model. The independent variable should go on the horizontal axis, and the dependent variable, the predicted variable, goes on the vertical axis.

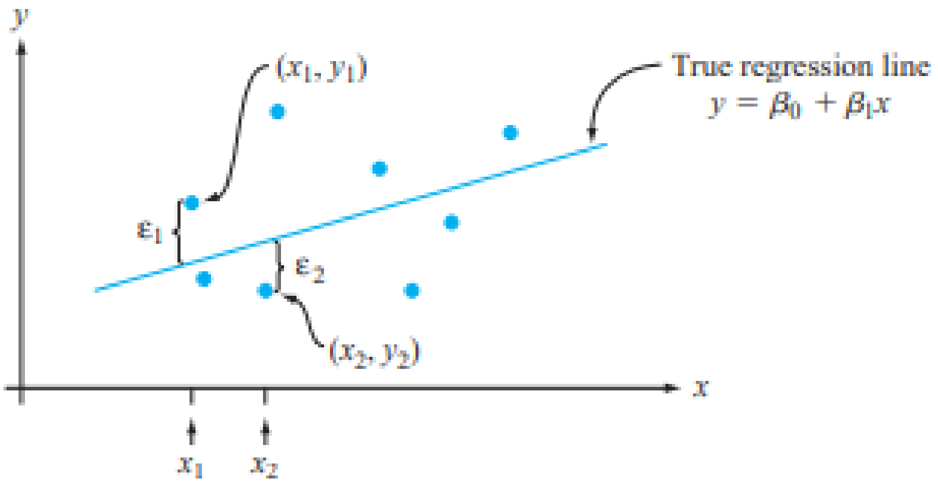
To indicate observational pairs, we subscript the variables. One pair of observations is (x_i, y_i). When creating our scatterplots, we generally are concerned about the relationship of the variables to each other, so we should remove any white space from the graph so that the observed values take up the entire graph. Unlike bar graphs, it's not necessary to start at zero.

Let's look at one of our simulated examples and construct a model of the variables.



We can see that the data represented by X and Y have a positive linear relationship with a relatively strong correlation. If we try to draw a curve through the middle of data, it will be a straight line with a positive slope.

If we think of the straight line $y = \beta_0 + \beta_1 x$ as the true regression line, then the ϵ's are the errors or deviations from that line.



These errors are commonly referred to as residuals. We can estimate their values from the estimation of the line we generate. In other words,

$$\epsilon = y_i - \hat{y}_i$$

Where $\hat{y}_i = b_0 + b_1 x_i$ from our estimated regression equation. Our goal is to find values for b_0 and b_1 that make the squares of the ϵ's as small as possible. This is the source of the name least squares regression.

Using calculus, we can minimize the sum of the squares $\sum \epsilon^2$, by replacing our expressions for epsilon in terms of our equations and x_i and y_i, and then set the derivative with respect to the two variables b_0 and b_1 equal to zero. Then we solve the system to obtain formulas for b_0 and b_1 that produce the minimum. (The critical point must be a minimum since there can't be a maximum.)

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \big(y_i - (b_0 + b_1 x_i)\big)^2 = f(b_0, b_1)$$

$$\frac{\partial f}{\partial b_0} = \sum_{i=1}^{n} 2(\, y_i - (b_0 + b_1 x_i))(-1) = 0$$

$$\frac{\partial f}{\partial b_1} = \sum_{i=1}^{n} 2(\, y_i - (b_0 + b_1 x_i))(-x_i) = 0$$

We can rearrange these to obtain.

$$n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Solving for b_1 and b_0, we get

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

It turns out the calculus solution and the linear algebra solution produce the same results. Consider that we are trying to find the best-fit equation $y = b_0 + b_1 x$ to a set of observations. Let's say the pairs $\{(1,5), (2,6), (3,8), (4,11)\}$. We set up a set of linear equations by replacing x and y in the equation with b_0 and b_1.

$$b_0 + b_1(1) = 5$$
$$b_0 + b_1(2) = 6$$
$$b_0 + b_1(3) = 8$$
$$b_0 + b_1(4) = 11$$

We can't solve this system exactly, but we can convert it into a matrix form with A being the coefficient matrix and Y being the constant vector, and B being the unknown coefficients.

$$AB = Y$$

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, B = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, Y = \begin{bmatrix} 5 \\ 6 \\ 8 \\ 11 \end{bmatrix}$$

To solve this for the best estimate we multiply the equation by A^T. (The reasoning for this is deep in linear algebra, so for now we will just worry about the procedural side.)

$$A^T AB = A^T Y$$

It turns out that in most cases, $A^T A$ will not only be a square matrix but also one that is invertible, so that we can solve for it.

$$B = (A^T A)^{-1} A^T Y$$

Is the solution to the system in matrix form. It turns out that B is a vector of our estimates for b_0 and b_1 and if we break out the formulas for the components from the matrix operations, we obtain the same formulas for the coefficients that we did above from calculus.

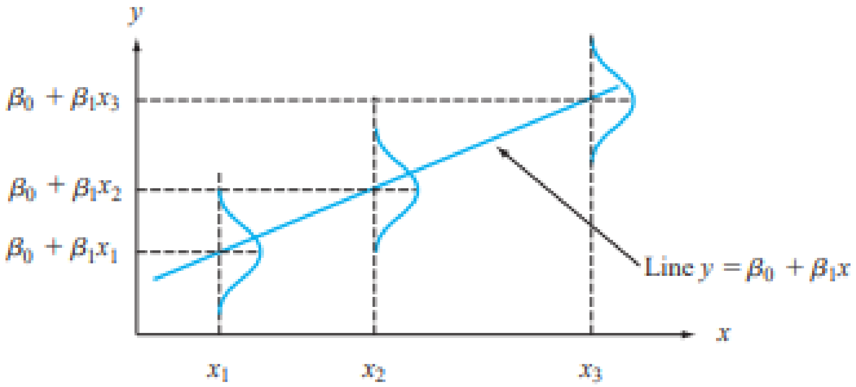
The linear algebra approach has significant advantages. For one thing, we can use the same procedure for almost all systems of equations, and linear algebra can construct workarounds for systems that don't meet the usual requirements for our matrices (for instance, if $A^T A$ turns out not to be invertible). The calculus approach will require separate equations for every possible type of solution with multiple variables, nonlinear approaches, etc. The linear algebra approach is much more compact, and computers work with arrays rather easily, so the linear algebra approach is very easy to encode. It's far more likely that your computer is using the linear algebra approach than going through the raw summation formulas.

There are sources in the reference list that go deeper into the linear algebra approach to finding these coefficients.

If we solve the small example above for the best-fit line we get

$$y = 2x + 2.5$$

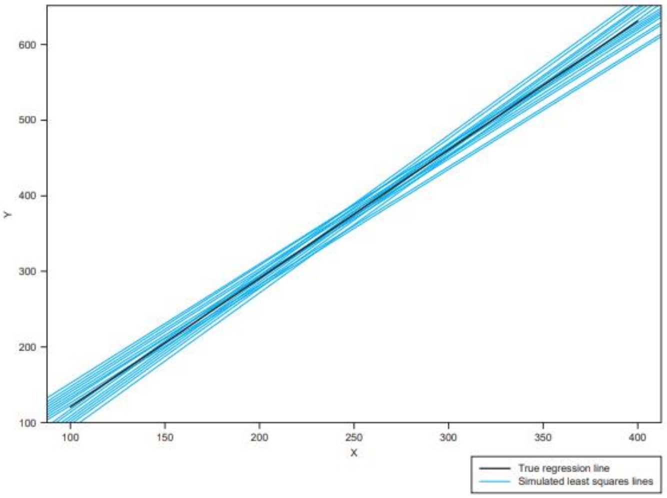
How do we interpret this equation? Essentially, this equation is an estimate of the population true regression line. The prediction from the equation can be best viewed as an estimate of the mean observed y-value. If we were to observe y at the same value of x several times, we'd expect the mean of those observations to agree with our prediction. Those observations will be normally distributed around the mean value that sits on the estimated regression line.



The meaning of the y-intercept of the equation is the mean value of y when $x = 0$, however, it will depend on the data as to whether or not the model is even meaningful at $x = 0$. For instance, if x is years, no model is likely to work going back in time over 2000 years. In some applications, this value may be negative, in which case, it may not have any physical meaning, for instance, if you are measuring heights. So, interpret with care.

We also want to be able to interpret the slope. We should do this in context. In general, the slope is a rate. It's the rate that y changes relative to x. For instance, if x changes by 1, then the y value changes (on average) by the value of the slope. So, perhaps our measurements are the height of a tree in each year after planting. Then we would say that the tree grows on average 2.5 feet per year.

Because the predictions are the expected means of outcomes, we can construct prediction intervals around the mean, similar to confidence intervals, but we will use the standard error (standard deviation) of the residuals to construct it. We can estimate it with just that, but a more sophisticated construction includes the distance from the mean of the data, where are estimates are the most accurate. As you move towards the edge of the data, the potential variability increases as the slope of the line is itself an estimate and the swing of the slope changes the most at the ends.



We could construct a bootstrap simulation of our regression data and predict the regression line from each version. We'd get similar, but not identical results. The biggest differences would be on the endpoints.

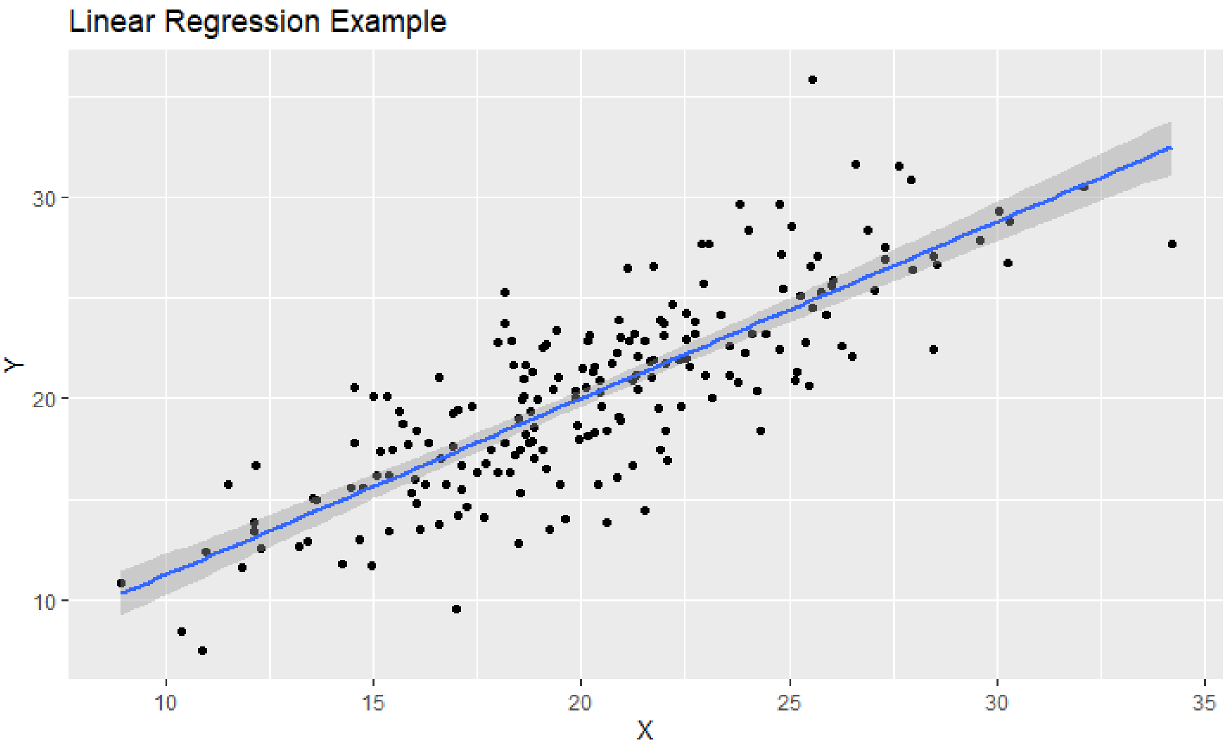
The exact formula we can use for the confidence interval depends on the standard error of the residuals in place of the standard error of the sampling distribution.

$$s_{\hat{Y}} = s_{\hat{\beta}_0+\hat{\beta}_1 x^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

In most cases, the second term under the radical is small, and so sometimes we can ignore it, and just use the standard error of the residuals divided by the square root of n, where n is the number of observations used to obtain the mean. But, as you get further from the mean of x, this value gets bigger and increases rapidly outside the range of the data. This is one of the reasons why it is best to avoid extrapolation (predicting values outside the range of the original data), and because we can't be sure that the trend continues. (Interpolation is when we predict values inside the range of the data.)

When building your confidence interval with the t-distribution, the degrees of freedom is $n - 2$; it's 2 because we've estimated two parameters, b_0 and b_1.

We can plot both the line and the prediction margins in the graph of our data.



You can see how they get wider on the ends.

One difficulty with the way that we calculate these estimates is how much of this data is outside the confidence bounds. Thus, how we interpret these error bounds is important. This confidence interval is smaller than the prediction interval.

$$\begin{aligned}
&\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \\
&= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} \\
&= \hat{y} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{Y}}}
\end{aligned}$$

The prediction interval has an extra +1 under the radical so that it is more similar to the residual error (and therefore wider), and not narrower for a prediction made repeatedly at a single observation.

As we saw with correlation, we can conduct hypothesis testing on our model, both of the model broadly and the model parameters, especially slope. Testing the slope in a simple linear model is essentially equivalent to testing the model overall, and testing the correlation. Generally, we test the slope compared to zero, to determine if the model has any value at all (no relationship is equivalent to zero correlation).

$$H_0\colon \beta_1 = 0$$
$$H_a\colon \beta_1 \neq 0$$

Are the default tests, although, we can test against specific parameter values, for example, if we wanted to slope to be larger than 1, we can modify our procedure accordingly.

To conduct the test, we need and standard error value for the slope parameter specifically.

$$s_{\beta_1} = \frac{s}{\sqrt{S_{xx}}}$$

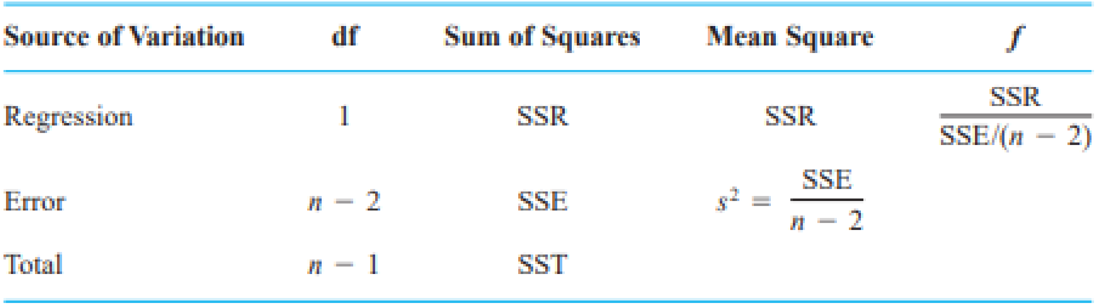
(Recall that S_{xx} = $\sum(x_i - \bar{x})^2$ or another equivalent formulation.)

Our test statistic then becomes

$$T = \frac{\beta_1 - \beta_{10}}{s_{\beta_1}}$$

We use $n - 2$ degrees of freedom to find the P-value. We can use this same standard error to construct a confidence interval on the slope.

Regression and ANOVA are closely related ideas and we'll say more about this in later lectures. For now, we'll just note that ANOVA tables are frequently included in regression analysis output. They should be interpreted as a test of the overall model. For now, this equivalent to testing the slope parameter or correlation, but as we extend to multivariable problems, it will take on more of the character of traditional ANOVA, only saying that at least one of the model parameters is meaningful.



Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	$n-2$	SSE	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

In the next lecture we'll look more closely at the residuals and model diagnostics, checking our assumptions and other tools to assess the quality of our models.

References:

1. http://www.statpower.net/Content/313/Lecture%20Notes/MatrixStatistics.pdf
2. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
3. https://godatadriven.com/blog/the-linear-algebra-behind-linear-regression/
4. https://towardsdatascience.com/building-linear-regression-least-squares-with-linear-algebra-2adf071dd5dd
5. https://machinelearningmastery.com/gentle-introduction-linear-algebra/
6. https://hefferon.net/linearalgebra/