27. SIMPLE LINEAR REGRESSION II

The Model

In linear regression analysis, we assume that the relationship between X and Y is linear. This does not mean, however, that Y can be perfectly predicted from X. In real applications there will almost always be some random variability which blurs any underlying systematic relationship which might exist.

We formalize these ideas by writing down a model which specifies exactly how the systematic and random components come together to produce our data.

Note that a model is simply a set of assumptions about how the world works. Of course, all models are wrong, but they can help us in understanding our data, and (if judiciously selected) may serve as useful approximations to the truth.
We start by assuming that for each value of $X$, the corresponding value of $Y$ is *random*, and has a normal distribution.

$$E(Y|X) = \beta_0 + \beta_1 X$$
Eg: Consider the heights in inches (X) of male MBA graduates from the University of Pittsburgh and their monthly salary (Y, Dollars) for their first post-MBA job.
Even if we just consider graduates who are all of the same height (say 6 feet), their salaries will fluctuate, and therefore are not perfectly predictable. Perhaps the mean salary for 6-foot-tall graduates is $6300/Month. Clearly, the mean salary for 5 ½ - foot-tall graduates is less than it is for 6-foot-tall graduates.

Based on our data, we are going to try to estimate the mean value of $y$ for a given value of $x$, denoted by $E(Y|X)$.

(Can you suggest a simple way to do this?)
In general, \( E(Y|X) \) will depend on \( X \).

Viewed as a function of \( X \), \( E(Y|X) \) is called the true regression of \( Y \) on \( X \).

In linear regression analysis, we assume that the true regression function \( E(Y|X) \) is a linear function of \( X \):

\[
E(y|x) = \beta_0 + \beta_1 x
\]

The parameter \( \beta_1 \) is the slope of the true regression line, and can be interpreted as the population mean change in \( y \) for a unit change in \( x \).
In an advertising application, for example, $\beta_1$ might represent the mean increase in sales of Godiva chocolates for each additional ad shown.

In the Salary vs. Height example, $\beta_1$ is the mean increase in monthly salary for each additional inch of height.
The parameter $\beta_0$ is the intercept of the true regression line and can be interpreted as the mean value of $Y$ when $X$ is zero.

For this interpretation to apply in practice, however, it is necessary to have data for $X$ near zero.

In the advertising example, $\beta_0$ would represent the mean baseline sales level without any advertisements.

It is almost always a good idea to include an intercept in the model, even if $\beta_0$ does not have a natural interpretation.

In practice, $\beta_0$ and $\beta_1$ will be unknown parameters, which must be estimated from our data. The reason is that our observed data will not lie exactly on the true regression line. (Why?)
Instead, we assume that the true regression line is observed with error, i.e., that \( y_i \) is given by

\[
y_i = \beta_0 + \beta_1 x_i + \epsilon_i,
\]

for \( i = 1, \ldots, n \), where \( \epsilon_i \) is a normally distributed random variable (called the **error**) with mean 0 and variance \( \sigma^2 \) which does not depend on \( x \).

We also assume that the values of \( \epsilon \) associated with any two values of \( y \) are independent.

Thus, the error at one point does not affect the error at any other point.
If (1) holds, then the values of $y$ will be randomly scattered about the true regression line, and the mean value of $y$ for a given $x$ will be this true regression line, $E(y|x) = \beta_0 + \beta_1 x$.

The error term $\varepsilon$ accounts for all of the variables -- measurable and un-measurable -- that are not part of the model.

For the heights of graduates and their salaries, suppose that $E(Y|X = x) = -900 + 100 x$. Then the salary of a 6-foot-tall graduate will differ from the mean value of $6300$/Month by an amount equal to the error term $\varepsilon$, which lumps together a variety of effects such as the graduate’s skills, desire to work hard, etc.

Furthermore, if the model (1) holds, then the salaries for graduates who are $5 \frac{1}{2}$ feet tall will fluctuate about their own mean ($5700$/Month) with the same level of fluctuation as before.
The variance of the error, $\sigma^2$, measures how close the points are to the true line, in terms of expected squared vertical deviation.

Under the model (1) which has normal errors, there is a 95% probability that a $y$ value will fall within $\pm 2 \sigma$ from the true line (measured vertically).

So $\sigma$ measures the "thickness" of the band of points as they scatter about the true line.
The Regression Effect

Sir Francis Galton (1822-1911) studied a data set of 1,078 heights of fathers and sons. His data set is on the next page. Two lines are superimposed: The dashed line is a 45-degree line, shifted up by one inch since on the average the sons were one inch taller than the fathers. The solid line is one that better fits the data set. (It’s the least squares line, to be described soon.)

Galton observed that tall fathers tend to have tall sons but the sons are not, on average, as tall as the fathers. Also, short fathers have short sons who, however, are not as short on average as their fathers. Galton called this effect “regression to the mean”.

In other words, the son’s height tends to be closer to the overall mean height than the father’s height was.

Nowadays, the term “regression” is used more generally in statistics to refer to the process of fitting a line to data.
Heights of Fathers and Sons
But the regression effect noticed by Galton is a very widespread phenomenon. Here are some examples:

(1) In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test -- and the top group will on average fall back.

(2) Consider a manager who rewards workers for good performance and punishes workers for bad performance. Subsequently, the performance of the rewarded workers declines, while that of the punished workers improves. The manager concludes that punishment is good, while rewards just make workers lazy. But in fact, the workers’ subsequent performance may be just another example of the regression effect: A worker’s performance next year will on average be closer to the mean than it was this year.
The Least Squares Estimators

Given our data on $x$ and $y$, we want to estimate the unknown intercept and slope $\beta_0$ and $\beta_1$ of the true regression line.

One approach is to find the line which best fits the data, in some sense.

In principle, we could try all possible lines $b_0 + b_1 x$ and pick the line which comes "closest" to the data. This procedure is called "fitting a line to data".

[R Demo: LeastSquaresFit]
To objectively fit a line to data, we need a way to measure the "distance" between a line \((b_0 + b_1 x)\) and our data.

We use the sum of squared vertical prediction errors,

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

The smaller \(f(b_0, b_1)\) is, the closer the line \(b_0 + b_1 x\) is to our data. Thus the best fitting line is obtained by minimizing \(f(b_0, b_1)\).

The solution is denoted by \(\hat{\beta}_0, \hat{\beta}_1\), and may be found by calculus methods.
• The resulting line \( y = \hat{\beta}_0 + \hat{\beta}_1 x \) is called the \textbf{least squares line}, since it makes the sum of squared vertical prediction errors as small as possible.

The least squares line is also referred to as the \textbf{fitted line}, and the \textbf{regression line}, but do not confuse the regression line \( \hat{\beta}_0 + \hat{\beta}_1 x \) with the \textit{true} regression line \( \beta_0 + \beta_1 x \).

Formulas for \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are available.

You do not need to learn these formulas, however, since regression calculations are best left to a computer.
It can be shown that the sample statistics \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are unbiased estimators of the population parameters \( \beta_0 \) and \( \beta_1 \). That is,

\[
E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1.
\]

- Give interpretations of \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) in the Salary vs. Height example, and the Beer example (Calories / 12 oz serving vs. %Alcohol).

Instead of using complicated formulas, you can obtain the least squares estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) from Minitab.
Eg: For the $y =$ Salary, $x =$ Height data, Minitab gives 

$$\hat{\beta}_0 = -902.2 \text{ and } \hat{\beta}_1 = 100.36.$$ 

Thus, the fitted model is 

$$\hat{y} = -902.2 + 100.36 x.$$ 

• Give an interpretation of this fitted model.
Regression Analysis: Salary versus Height

The regression equation is
Salary = - 902 + 100 Height

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-902.2</td>
<td>837.0</td>
<td>-1.08</td>
<td>0.290</td>
</tr>
<tr>
<td>Height</td>
<td>100.36</td>
<td>12.02</td>
<td>8.35</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 192.702    R-Sq = 71.4%    R-Sq(adj) = 70.3%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2590433</td>
<td>2590433</td>
<td>69.76</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>28</td>
<td>1039754</td>
<td>37134</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>3630187</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fitted Line Plot for Salary vs. Height

Salary = -902.2 + 100.4 Height

S 192.702
R-Sq 71.4%
R-Sq(adj) 70.3%
Under repeated sampling, \( \hat{\beta}_0, \hat{\beta}_1 \)
as well as the least squares line have sampling distributions.

[R Demo: RegressionSampling]

This allows us to use the data to draw conclusions (learn; make inferences) about the true intercept and slope \( \beta_0, \beta_1 \).