17. 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.

\[ E(Y|X) = \alpha + \beta X \]

*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 $X$:

$$E(Y|x) = \alpha + \beta x$$

The parameter $\beta$ is the slope of the true regression line, and can be interpreted as the population mean change in $y$ for a unit change in $x$.

The parameter $\alpha$ 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, $\alpha$ would represent the mean baseline sales level without any advertisements.

• Unfortunately, the notation $\alpha$ is traditionally used for both the significance level of a hypothesis test and for the intercept of the true regression line.

In practice, $\alpha$ and $\beta$ 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?)

In an advertising application, for example, $\beta$ might represent the mean increase in sales of Pepsi for each additional 30 second ad.

In the Salary vs. Height example, $\beta$ is the mean increase in monthly salary for each additional inch of height.

Instead, we assume that the true regression line is observed with error, i.e., that $y_i$ is given by

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

for $i = 1, \ldots, n$. Here, $\varepsilon_i$ 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 $\varepsilon$ 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, \( E(y|x) = \alpha + \beta x \).

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 \( E(Y|X = x) = -900 + 100x \). 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 ½ 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 Least Squares Estimators

Given our data on \( x \) and \( y \), we want to estimate the unknown intercept and slope \( \alpha \) and \( \beta \) 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 \( a + bx \) and pick the line which comes "closest" to the data. This procedure is called "fitting a line to data".

See website which allows you to do this interactively, at


To objectively fit a line to data, we need a way to measure the "distance" between a line \( (a + bx) \) and our data.

We use the sum of squared vertical prediction errors,

\[
f(a, b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2.
\]

The smaller \( f(a, b) \) is, the closer the line \( a + bx \) is to our data. Thus the best fitting line is obtained by minimizing \( f(a, b) \).

The solution is denoted by \( \hat{\alpha}, \hat{\beta} \), and may be found by calculus methods.
• The resulting line $\hat{y} = \hat{\alpha} + \hat{\beta} x$ is called the 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 fitted line, and the regression line, but do not confuse the regression line $\hat{y} = \hat{\alpha} + \hat{\beta} x$ with the true regression line $\mathbb{E}(y|x) = \alpha + \beta x$.

It can be shown that the sample statistics $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of the population parameters $\alpha$ and $\beta$. That is,

$$E(\hat{\alpha}) = \alpha, \quad E(\hat{\beta}) = \beta.$$

• Give interpretations of $\hat{\alpha}$ and $\hat{\beta}$ 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 $\hat{\alpha}$ and $\hat{\beta}$ from Minitab.

\textbf{Eg:} For the $y =$ Salary, $x =$ Height data, Minitab gives

$\hat{\alpha} = -902$ and $\hat{\beta} = 100.4$.

Thus, the fitted model is

$$\hat{y} = -902 + 100.4 x .$$

• Give an interpretation of this fitted model.

\textbf{Regression Analysis: Salary versus Height}

\begin{tabular}{lrrrr}
\hline
\textbf{Term} & \textbf{Coef} & \textbf{SE Coef} & \textbf{T-Value} & \textbf{P-Value} \\
\hline
Constant & -902 & 837 & -1.08 & 0.290 \\
Height & 100.4 & 12.0 & 8.35 & 0.000 \\
\hline
\end{tabular}

\textbf{Regression Equation}

Salary = -902 + 100.4 Height
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%