18. SIMPLE LINEAR REGRESSION III

**Fitted Values and Residuals**

To each observed \( x_i \), there corresponds a \( y \)-value on the fitted line,

\[
y_i = \hat{\alpha} + \hat{\beta} x_i.
\]

The \( \hat{y}_i \) are called **fitted values**.

They are the values of \( y \) which would be predicted by the estimated linear regression model, at the observed values of \( x \).

**Eg:** For the Beer example, the fitted Calories/12 oz are

\[
y = 38.1 + 22.02 x,
\]

where \( x \) is the % Alcohol.

**Regression Analysis: Calories/12 oz versus % Alcohol**

Model Summary

\[
\begin{array}{cccc}
S & R-sq & R-sq(adj) & R-sq(pred) \\
20.8238 & 48.01\% & 47.49\% & 37.56\%
\end{array}
\]

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>38.1</td>
<td>11.1</td>
<td>3.42</td>
<td>0.001</td>
</tr>
<tr>
<td>% Alcohol</td>
<td>22.02</td>
<td>2.30</td>
<td>9.56</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Regression Equation

Calories/12 oz = 38.1 + 22.02 % Alcohol

O’Doul’s has a % Alcohol of 0.4. Therefore, the corresponding fitted value is

\[
\hat{y}=38.1+(22.02)(0.4)=46.908.
\]

Note that this "prediction" is smaller than the actual Calories for this beer, \( y = 70 \).
In general, the observed data values $y_i$ will be different from the fitted values $\hat{y}_i$.

The differences between the data and fitted values are called the **residuals**,

$$e_i = y_i - \hat{y}_i.$$

**Eg:** In the Beer example, the residual corresponding to O'Doul's is $70 - 46.908 = 23.092$.

We can think of the residuals $e_i$ as proxies for the underlying errors $\varepsilon_i$.

A given residual will be positive if the data point lies above the fitted line, and negative if the point is below the fitted line.

Because of the way the least squares estimators are determined, the residuals will always sum to zero.

- The residuals are very important, since they tell us how well the least squares line fits the data.

To assess the quality of the fit, we can plot the residuals against the $x_i$'s, or against the fitted values.

If the model is adequate, then these plots should show no systematic patterns or structure. If they do show structure, then the model may need to be modified.

Residual plots should be a routine part of any regression analysis.

They often reveal unexpected features in the data, such as nonlinear relationships (parabola), non-constant variance (wedge), and outliers, which might never have been discovered otherwise.
For the Beer example, the plot of residuals versus the %Alcohol shows the biggest outlier, Sam Adams Light, though the residual for O’Douł’s does not look that bad. The actual errors for these two points may be much larger, though, since they may have dragged up the fitted line, causing the appearance of a linear pattern with positive slope for the other residuals.

After we remove these two outliers, the fitted slope increases from 22.02 to 35.14, and the residual plot seems reasonably free of patterns.

In particular, the variability seems not to change radically with $x$.

Thus, the cleaned data set does not strongly contradict any of the assumptions underlying the linear regression model.

Next, we assess the variability of $y$ for a given $x$. A helpful example for this is Salary vs. Height.
A numerical measure of the closeness of the least squares line to the data is provided by the sum of squared residuals,
\[ \text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]
If all the data points lie exactly on a line, then SSE will be zero. (Why?)
SSE stands for "Sum of Squares for Error".
But SSE is not equal to the sum of squares of the actual errors, \( \sum_{i=1}^{n} e_i^2 \). (Why?)
Nevertheless, we can use SSE to estimate the variance \( \sigma^2 \) of the actual errors \( \varepsilon \).

An unbiased estimator of \( \sigma^2 \) is given by
\[ s^2 = \frac{\text{SSE}}{n-2} = \text{Resid. Mean Square}. \quad (1) \]
Since we are estimating two parameters, we have \( n - 2 \) "degrees of freedom".
We can think of \( s^2 \) as an "average squared prediction error", since SSE is the sum of the squared deviations between the actual and predicted values (\( y \) and \( \hat{y} \)).
You might be tempted to use \( s^2 \) or SSE to compare different models fitted to a given set of \( y \)'s.
This is actually a very bad way to pick a model!
You need to first adjust for the fact that different models may have different numbers of explanatory variables.
(We will describe how to do this later if time permits).

Eg: For the Salary vs. Height data, \( s^2 = 37134 \). Thus, we can estimate the standard deviation of the salary for a given height to be \( s = 192.702 \) Month. This tells us about the natural variability which would be expected in salary for a given fixed height.

**Regression Analysis: Salary versus Height**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F-Value</th>
<th>P-Value</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>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>

Model Summary

- \( S = 192.702 \)
- \( R^2 = 71.36\% \)
- \( R^2(\text{adj}) = 70.34\% \)
- \( R^2(\text{pred}) = 65.65\% \)

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-902</td>
<td>837</td>
<td>-1.08</td>
<td>0.290</td>
</tr>
<tr>
<td>Height</td>
<td>100.4</td>
<td>12.0</td>
<td>8.35</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Don't confuse \( s^2 \) with the sample variance of a batch of numbers,
\[ \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2. \quad (2) \]
We also used the name \( s^2 \) for (2), but it's not the same as (1).
There is a connection, though, since we can think of (2) as an "average squared prediction error" based on a simpler predictor of \( y \): Just use \( \bar{y} \) to predict all the \( y \)'s, and forget about the \( x \)'s.
Testing Whether $\beta$ is Zero: Is the Straight-Line Model Useful for Predicting $y$ from $x$?

A basic principle of model building is that we should always use the simplest model which adequately describes the data.

• Why do we think that the Earth goes around the Sun?

• What is the best predictor for the number of tournaments that Serena Williams will win next year?

Accordingly, unless there is strong evidence of a linear relationship between $x$ and $y$, it is better to adhere to the null hypothesis of no linear relationship, $H_0: \beta = 0$.

This produces a much simpler model, $y_i = \alpha + \epsilon_i$, in other words, the $y$'s are a batch of data with a normal distribution (mean $\alpha$, variance $\sigma^2$).

In fact, this is the same model used earlier in the course to make inferences about the population mean $\mu$ based on the sample mean and standard deviation.

• So if $\beta = 0$, the best predictor of a future $y$ value is just the average of the observed $y$ values. It is only if $\beta \neq 0$ that the values of the "explanatory variable" $x$ can improve our ability to predict a future $y$. If $\beta = 0$, the use of $x$ will usually degrade the quality of the forecast!

Of course, we can use a scatterplot of $y$ versus $x$ as well as a residual plot to help us to decide whether a linear relationship exists, but it would be nice to have a more objective tool.

• Why can't we just check whether $\hat{\beta} = 0$?

Because even if $\beta = 0$, the estimator $\hat{\beta}$ is a random variable, and will not in general be exactly zero.

So $\hat{\beta}$ has a sampling distribution. This distribution is normal, with mean $\beta$. The standard error of $\hat{\beta}$ can be estimated from the data.

We will call this $S_{\hat{\beta}}$.

Minitab calls it “SE Coef”, that is, the standard error of the estimated coefficient.

For the Stock Market returns, $S_{\hat{\beta}}$ is 0.00928.

Regression Analysis: Today versus Yesterday

Model Summary

<table>
<thead>
<tr>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.971546</td>
<td>0.56%</td>
<td>0.55%</td>
<td>0.46%</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.01654</td>
<td>0.00905</td>
<td>1.83</td>
<td>0.067</td>
</tr>
<tr>
<td>Yesterday</td>
<td>0.07486</td>
<td>0.00928</td>
<td>8.06</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Regression Equation

Today = 0.01654 + 0.07486 Yesterday
Now, we can test $H_0: \beta = 0$ versus $H_1: \beta \neq 0$ by converting $\hat{\beta}$ into its own $t$-score,

$$t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}}.$$ 

The result is given by Minitab as "T".

For the Stock Market data,

$$t = \frac{0.07486}{0.00928} = 8.07.$$ 

- If $\beta = 0$, then $t$ will have a $t$ distribution with $n - 2$ degrees of freedom.

The $p$-value corresponding to a 2-tailed test is given by Minitab as "P-Value".

For the Stock Market data, we get $p = 0.000$. Conclusion: Strong evidence of a linear relationship between today's return and yesterday's return! (But the returns are still quite hard to predict.)

For the Salary vs. Height example, the $t$-statistic for $\beta$ is 8.35, yielding $p = 0.000$ (rounded to nearest 1/1000). Apparently, height matters.

For the Rotten Tomatoes Audience Score vs. Critics Score, the $p$-value for $\beta$ is 0.004, so there is strong evidence of a linear relationship. But we might be interested in a different null hypothesis than $\beta = 0$.

Regression Analysis: Rotten Tomatoes Audience Score versus Critics Score

Model Summary

<table>
<thead>
<tr>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.30058</td>
<td>65.64%</td>
<td>61.34%</td>
<td>41.23%</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>34.0</td>
<td>11.3</td>
<td>3.01</td>
<td>0.017</td>
</tr>
<tr>
<td>Critics Score</td>
<td>0.588</td>
<td>0.151</td>
<td>3.91</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Regression Equation

Audience Score = 34.0 + 0.588 Critics Score
For the RT example, a 95% confidence interval for β is given by
\[
\hat{\beta} \pm t_{0.025} s_{\hat{\beta}},
\]
where from Table 6, \(t_{0.025} = 2.306\) (df=8). From the Minitab output, we obtain the interval 0.588 ± (2.306) (0.151), which reduces to (0.240, 0.936).

The interval does not contain 1.0, so it is not plausible that \(\beta = 1\).

What would be the practical importance of this?

To test \(H_0 : \beta = 1\) versus \(H_1 : \beta \neq 1\) at level 0.05, we can use the test statistic
\[
t = \frac{\hat{\beta} - 1}{s_{\hat{\beta}}} = \frac{-0.412}{0.151} = -2.73,
\]
which is less than the critical value of -2.306, so we reject \(H_0\).

Clearly, the \(p\)-value is less than 0.05, but we cannot compute it without additional tables.

(Can get tail areas for a \(t\)-distribution in Minitab, but we won’t).

In any case, we have statistical evidence to suggest that \(\beta\) is different from 1.

In this example, it was more reasonable to test if \(\beta = 1\) than to test if \(\beta = 0\). Keep in mind, though, that Minitab bases its "T" and "P" on a null hypothesis that the true parameter values are zero. If you are interested in a different null, you must compute the test statistic manually, as above.

- We can get confidence interval and hypothesis test for the intercept \(\alpha\) in a similar way. Use the standard errors, \(t\)-statistics and \(p\)-values for \(\hat{\alpha}\), as given by Minitab.
**Eg:** July Mean Temperatures in UK versus the year.

**UK July Mean Temperatures (Degrees C)**

\[ \text{JUL} = -4.025 + 0.009518 \text{ Year} \]

Regression Analysis: JUL versus Year

Model Summary

<table>
<thead>
<tr>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.01247</td>
<td>6.11%</td>
<td>5.08%</td>
<td>1.63%</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.02</td>
<td>7.67</td>
<td>-0.53</td>
<td>0.601</td>
</tr>
<tr>
<td>Year</td>
<td>0.00952</td>
<td>0.00391</td>
<td>2.43</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Regression Equation

\[ \text{JUL} = -4.02 + 0.00952 \text{ Year} \]

**Eg:** Returns on two stocks.

**Hershey vs. Apple Returns, Monthly**

\[ \text{HersheyRet} = 0.01066 - 0.09251 \text{ AppleRet} \]

Regression Analysis: HersheyRet versus AppleRet

Model Summary

<table>
<thead>
<tr>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0661965</td>
<td>4.28%</td>
<td>3.42%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.01066</td>
<td>0.00633</td>
<td>1.68</td>
<td>0.095</td>
</tr>
<tr>
<td>AppleRet</td>
<td>-0.0925</td>
<td>0.0415</td>
<td>-2.23</td>
<td>0.028</td>
</tr>
</tbody>
</table>

Regression Equation

\[ \text{HersheyRet} = 0.01066 - 0.0925 \text{ AppleRet} \]
**Eg: Business Week Rankings for UG Business Schools**

**Regression Analysis: BW 2007 Rank versus BW 2006 Rank**

Rows unused 42

**Model Summary**

<table>
<thead>
<tr>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.70014</td>
<td>84.81%</td>
<td>84.50%</td>
<td>83.53%</td>
</tr>
</tbody>
</table>

**Coefficients**

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.12</td>
<td>2.67</td>
<td>-1.17</td>
<td>0.248</td>
</tr>
<tr>
<td>BW 2006 Rank</td>
<td>1.3182</td>
<td>0.0797</td>
<td>16.54</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Regression Equation**

BW 2007 Rank = -3.12 + 1.3182 BW 2006 Rank

**Eg: Median Starting Salaries for Undergrad Business Students**

**Med. Starting Salary vs BW 2007 Rank**

**Med. Starting Salary**

- 60000
- 55000
- 50000
- 45000
- 40000
- 35000

**BW 2007 Rank**

- 0
- 10
- 20
- 30
- 40
- 50
- 60
Regression Analysis: Med. Starting Salary versus BW 2007 Rank

Model Summary
- S: 3627.10
- R-sq: 58.73%
- R-sq(adj): 58.26%
- R-sq(pred): 56.69%

Coefficients
- Constant: 53137, SE: 785, T-Value: 67.69, P-Value: 0.000
- BW 2007 Rank: -160.0, SE: 14.3, T-Value: -11.19, P-Value: 0.000

Regression Equation
Med. Starting Salary = 53137 - 160.0 BW 2007 Rank

Eg: Salary vs. Height.

Monthly Salary (Dollars) vs. Height (Inches)

Do you see how the two outliers on the right may have dragged down the line?

A demo of the effect of outliers on the fitted line is at http://www.amstat.org/publications/jse/v6n3/applets/Regression.html