4.1 Fitting a Line by Least Squares Regression

Scatterplots

- Plot of $Y$ vs $X$, two quantitative variables, measured on the same individual
- $X$=explanatory variable, $Y$=response variable

- Interpreting scatterplots
  Direction (positive, negative)

Form (linear, clusters)

Strength (strong or weak or so-so)
Example: Interpret the following scatterplot.
X=percent taking SAT  Y=median SAT math score for each state

Correlation

- A numerical measure of the strength and direction of linear association between two quantitative variables.
- Why? Because it is very hard to determine the strength of the linear association by eye.
Facts about Correlation (r)

- Correlation does not distinguish between explanatory and response variables.
- r is always between -1 and +1 (no units).
- Interpretation: positive/negative, strong/weak.
- r measures only the strength of the linear relationship between x and y.
- Outliers can have a strong effect on r.
- Some examples:

Formula: Correlation Coefficient

Note: The formula for the correlation coefficient is mostly for illustration purposes. On the exam and quizzes, you will never need to compute r. For homework problems, you can use Minitab, or your calculator (if it does 2-variable statistics).
Least Squares Regression

- A method for finding the "best-fitting" line through a set of \((x, y)\) points.
- Does require an explanatory / response variable setting.
- "Least Squares" regression finds the line with the smallest possible sum of squared vertical distances to all the points.

Formulas:

- Formal Model: \( Y = \beta_0 + \beta_1 x + \epsilon \) where \( \beta_0 \) and \( \beta_1 \) are unknown and \( \epsilon \sim \mathcal{N}(0, \sigma^2) \).
- Fit Model  \( \hat{y} = b_0 + b_1 x \)

(These slope and y intercept formulas differ from the book.)
Coefficient of Determination
$R^2 = \text{proportion of the variability in y that is explained by x}$

Formula for $R^2$

Example: The number of pounds of steam used per month by a chemical plant is thought to be related to the average ambient temperature (in F degrees) for that month. The past year’s usage and temperature are shown in the following table.

<table>
<thead>
<tr>
<th>temperature</th>
<th>usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>185.79</td>
</tr>
<tr>
<td>24</td>
<td>210.00</td>
</tr>
<tr>
<td>32</td>
<td>283.00</td>
</tr>
<tr>
<td>47</td>
<td>414.04</td>
</tr>
<tr>
<td>50</td>
<td>454.58</td>
</tr>
<tr>
<td>59</td>
<td>529.03</td>
</tr>
<tr>
<td>68</td>
<td>620.00</td>
</tr>
<tr>
<td>74</td>
<td>670.25</td>
</tr>
<tr>
<td>62</td>
<td>572.03</td>
</tr>
<tr>
<td>50</td>
<td>452.93</td>
</tr>
<tr>
<td>41</td>
<td>369.95</td>
</tr>
<tr>
<td>30</td>
<td>273.98</td>
</tr>
</tbody>
</table>
Descriptive Statistics: Temperature, Usage/1000

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperat</td>
<td>12</td>
<td>46.50</td>
<td>48.50</td>
<td>46.30</td>
<td>17.34</td>
<td>5.01</td>
</tr>
<tr>
<td>Usage/10</td>
<td>12</td>
<td>419.6</td>
<td>433.5</td>
<td>418.0</td>
<td>160.0</td>
<td>46.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperat</td>
<td>21.00</td>
<td>74.00</td>
<td>30.50</td>
<td>61.25</td>
</tr>
<tr>
<td>Usage/10</td>
<td>185.8</td>
<td>670.3</td>
<td>276.2</td>
<td>561.3</td>
</tr>
</tbody>
</table>

Correlations: Temperature, Usage/1000

Pearson correlation of Temperature and Usage/1000 = 0.999
P-Value = 0.000

1.) Find the slope and interpret.

2.) Find the y intercept and interpret.
3.) Predict the amount of steam used when x=30.

4.) What is the value of $R^2$? Interpret.

5.) With such a high value of $R^2$, can you say that x causes y?
6.) Identify the important parts in the computer output.

Regression Analysis: Usage/1000 versus Temperature

The regression equation is
Usage/1000 = -9.01 + 9.22 Temperature

Predictor    Coef     SE Coef     T      P
Constant   -9.014      4.762     -1.89  0.088
Temperat  9.21819    0.09644      95.58  0.000

S = 5.548    R-Sq = 99.9%    R-Sq(adj) = 99.9%

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>281182</td>
<td>281182</td>
<td>9135.83</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>10</td>
<td>308</td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>281490</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Residual Analysis

- Helps check how well the regression line fits the data
- Magnifies any patterns present - points out any problems.

- **Residual Plot**
  - plot of the **residuals** vs explanatory variable
  - **residuals** = difference between the observed and predicted values of y

- residuals always sum to zero

Examples:
- Original Plot: Y vs X
- Residual Plot: resids vs X
Residual Analysis

- Check the residual plot.
- Check either a histogram or a normal probability plot to check to see if the errors could be Normal. (The residuals should be Normally distributed if it is just random error.)

Example: The number of pounds of steam used per month by a chemical plant is thought to be related to the average ambient temperature (in F degrees) for that month. The past year’s usage and temperature are shown in the following table.

<table>
<thead>
<tr>
<th>temperature</th>
<th>usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>185.79</td>
</tr>
<tr>
<td>24</td>
<td>214.47</td>
</tr>
<tr>
<td>32</td>
<td>28803</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

7.) Find the residual for the first data point when x=21.

8.) Comment on the residual plot below.
9.) Comment on the histogram of the residuals below.

```
Histogram of the Residuals
(response is Usage/10)
```

10.) Comment on the normal probability plot below.

```
Normal Probability Plot of the Residuals
(response is Usage/10)
```
11.) Suppose that 13\textsuperscript{th} point was added to the plot. X= 0 and Y=700. Did this point affect anything?
Getting Ready for Inference

F distribution

To be able to some of the inferences in this section, it is necessary to learn another probability distribution, the F distribution.

\[
f(x) = \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right) \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} x^{\frac{v_1}{2} - 1}}{\Gamma\left(\frac{v_1}{2}\right) \Gamma\left(\frac{v_2}{2}\right) \left(1 + \frac{v_1 x}{v_2}\right)^{\frac{v_1 + v_2}{2}}}
\]

This distribution is often used to compare two variances.

If \( x \) is said to have an F distribution, it can be denote as \( F(\nu_1, \nu_2) \).

The cdf is denoted as \( Q_{\nu_1, \nu_2}(p) = \frac{1}{Q_{\nu_2, \nu_1}(1-p)} \)

Use the table in your packet to find the following quartiles.

1. Find \( f(3,5) \) at the 90\(^{th} \) quartile.

2. Find \( f(24, 26) \) at the 95\(^{th} \) quartile.

3. Find \( f(3,4) \) at the 5\(^{th} \) quartile.
9.1 Inference Methods Related to Least Squares Fitting of a Line (Simple Linear Regression)

9.1.1 The Simple Linear Regression Model and Corresponding Variance Estimate

Graphical Representation of the Simple Linear Model

Formal Model: \( Y = \beta_0 + \beta_1 x + \epsilon \) where \( \beta_0 \) and \( \beta_1 \) are unknown and \( \epsilon \sim \text{N}(0, s^2) \).

\[
s_{LF}^2 = \frac{1}{n-2} \sum (y - \hat{y})^2
\]

is used to estimate \( s^2 \)

\( s_{LF}^2 \) is called the line-fitting sample variance.
**Inferences for Slope**

- $\beta_1$ is the true rate of change of average response with respect to $x$
- $b_1$ estimates the true population slope $\beta_1$
- Test: We often want to see if $\beta_1$ is significantly different from 0. If $\beta_1$ is different from zero, then $x$ is a good predictor of $y$.

Examples where slope is a good predictor:

Example where slope is not a good predictor:

How do you test this?

- CI for $\beta_1$
- Significance test for $\beta_1$
- ANOVA F test
Confidence Intervals and Hypothesis Test for Slope, $\beta_1$

- Because $b_1$ is a statistic, every sample will have a different $b_1$.
- The sampling distribution of $b_1$ is $N\left(\beta_1, \frac{\sigma^2}{\sum (x - \bar{x})^2}\right)$
- Problem: We don’t know $\sigma^2$ (the variance of the errors), so we estimate it with $s_{LF}^2$ (the line fitting sample variance.)
- Since $s_{LF}$ is an estimate of $s$, we will be using the $t$ with $n-2$ degrees of freedom.
- $\sqrt{\frac{s_{LF}^2}{\sum (x - \bar{x})^2}} = s_h$ from the Minitab output

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Statistic</th>
<th>Std. Error</th>
<th>Sampling Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sqrt{\frac{s_{LF}^2}{\sum (x - \bar{x})^2}} = s_h$</td>
<td>t(n-2)</td>
<td></td>
</tr>
</tbody>
</table>

Confidence Interval for $\beta_1$

Hypothesis Test for $\beta_1$

Ho: ______________  Ha: ______________

Test Statistic:

**Note:** You can also test that the Ho: $\beta_1 = # \text{ versus Ha: } \beta_1 \neq #$, this test to see if $y$ changes by $#$ for every one unit increase in $x$. 

What is that ANOVA output?

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>SSM</td>
<td>SSM / 1</td>
<td>MSM / MSE</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>n-2</td>
<td>SSE</td>
<td>SSE / (n-2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SST</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2
\]

Notes:
- MSE – variability of points around line
- \( s_{\hat{y}} \) is the estimator of \( \sigma \) (constant variance)
- \( s_{LF}^2 = \text{MSE} = \text{estimate of } \sigma^2 \)
- \( s_{LF} = s = \sqrt{\text{MSE}} \)
- \( R^2 = \frac{SS \text{ RegModel}}{SS_{\text{Total}}} \)

F Test for Slope, Ho: \( \beta_1 = 0 \) versus Ho: \( \beta_1 \neq 0 \)

Test Statistic: \( F = \frac{MSR}{MSE} \)

Sampling Distribution: F(1, n-2).
Example: Below is the regression output of the relationship between Sarah’s age and Sarah’s height.

<table>
<thead>
<tr>
<th>Age (mo.)</th>
<th>36</th>
<th>48</th>
<th>51</th>
<th>54</th>
<th>57</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (cm)</td>
<td>86</td>
<td>90</td>
<td>91</td>
<td>93</td>
<td>94</td>
<td>95</td>
</tr>
</tbody>
</table>

Regression Analysis

The regression equation is

\[ \text{height} = 72.0 + 0.383 \text{ age} \]

Predictor Coef       StDev          T        P
Constant  71.950       1.053      68.33    0.000
age      0.38333     0.02041      18.78    0.000

S = 0.3873    R-Sq = 98.9%     R-Sq(adj) = 98.6%

Analysis of Variance

Source  DF    SS          MS         F        P
Regress 1   52.900     52.900    352.67    0.000
Res Err 4    0.600      0.150
Total   5   53.500

In Minitab:    Go to Stat – Regression – Fitted Line Plot

a.) Test the hypothesis that Ho: \( \beta_1 = 0 \) versus Ha: \( \beta_1 \neq 0 \) using the t distribution.

b.) Test the hypothesis that Ho: \( \beta_1 = 0 \) versus Ha: \( \beta_1 \neq 0 \) using the F distribution.
c.) Find a 99% confidence interval for Sarah’s growth rate. Interpret.

d.) The normal growth rate for girls is 6 cm per year. Is Sarah growing Normally?

In Simple Linear Regression, these two tests are the same.

<table>
<thead>
<tr>
<th>t test for $\beta_1$</th>
<th>ANOVA F Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Statistic = 18.78</td>
<td>Test Statistic = $(18.78)^2$</td>
</tr>
<tr>
<td>p-value = 0.000</td>
<td>p-value = 0.000</td>
</tr>
</tbody>
</table>
12.4 Prediction Intervals

One of the main reasons for doing a simple linear regression analysis is because it allows us to do predictions. However, you can only do predictions for points within the set of data. Fitting a least squares regression line to a point that is outside the range of data is called extrapolation.

There are two types of intervals for $y$

- Predicting the average $y$ at a value of $x$
- Predicting the next $y$ at a value of $x$

Predicting the average $y$ at a value of $x$

- The distribution of $\hat{y} \sim N(\hat{Y}, \sigma^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2} \right))$

- However, we don’t know $s^2$ (the variance of the errors), so, we estimate it with $s_{LF}^2$.

- Since $s_{LF}$ is an estimate of $s$, we will be using the $t$ with $n-2$ degrees of freedom.

\[
s_{LF} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = s_{FIT}
\]

- $s_{FIT}$ is on the Minitab output for every value of $x$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Statistic</th>
<th>Std. Error</th>
<th>Sampling Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$s_{LF} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = s_{FIT}$</td>
<td>t(n-2)</td>
<td></td>
</tr>
</tbody>
</table>

Confidence interval for the mean value of $y$ at $x$
Hypothesis test for the mean value of $y$ at $x$

Predicting the next value of $y$ at a new value of $x$

- The distribution of $Y_{new} \sim N\left(\hat{\mu}_x, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}\right)\right)$

- However, we don’t know $\sigma^2$ (the variance of the errors), so, we estimate it with $s_{LF}^2$.

- Since $s_{LF}$ is an estimate of $\sigma$, we will be using the $t$ with $n-2$ degrees of freedom.

- $s_{LF} \sqrt{\frac{1 + \frac{1}{n} + \sum (x - \bar{x})^2}{\sum (x - \bar{x})^2}} = \sqrt{s_{FIT}^2 + s_{LF}^2}$

- $s_{FIT}$ and $s_{LF}$ are on the Minitab output for every value of $x$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Statistic</th>
<th>Std. Error</th>
<th>Sampling Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$s_{LF} \sqrt{\frac{1 + \frac{1}{n} + \sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = \sqrt{s_{FIT}^2 + s_{LF}^2}$</td>
<td>t(n-2)</td>
<td></td>
</tr>
</tbody>
</table>

Confidence Interval for the next $y$ at a certain $x$

Hypothesis Test for the next $y$ at a certain $x$
Notes:

- Prediction interval is always wider than CI
- Both intervals are centered around $\hat{y} = b_o + b_x x^*$
- Std errors (and intervals) will be smaller when $x^*$ is closer to $\bar{X}$, bigger when $x^*$ is far from $\bar{X}$.

Warning: Be very careful when making multiple intervals. Using the equations for confidence intervals and prediction intervals separately for several values of $x$ yield a set of CI for which the joint or simultaneous confidence level is guaranteed to be at least $100(1-k\alpha)%$.

Simultaneous Two Sided Confidence Limits for all mean of $y$ at $x$.

If you want to make a confidence interval for the mean of $y$ at every possible value of $x$, use the following equation.

$$\hat{y} \pm 2 f s_{LF} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$  

where $f$ is $F(2, n-2)$.

The F distribution has two degrees of freedom. The f critical values can be found in table B.6A.

Note: $s_{LF}$ is not under the square root sign.

Example: Suppose that we want to make a confidence interval of the average value of $y$ at $x$ and to predict the value of $y$ for the next $x$.

New Example: $Y =$ the yield of grain $g/m-row$  $X =$ the distance up slope on a sloping watershed $m$

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>45</th>
<th>50</th>
<th>70</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
<th>160</th>
<th>170</th>
<th>190</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>500</td>
<td>590</td>
<td>410</td>
<td>470</td>
<td>450</td>
<td>480</td>
<td>510</td>
<td>450</td>
<td>360</td>
<td>400</td>
<td>300</td>
<td>410</td>
<td>280</td>
<td>350</td>
</tr>
</tbody>
</table>
Regression Analysis

The regression equation is
\[ y = 515 - 1.06 \, x \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>StDev</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>515.45</td>
<td>25.14</td>
<td>20.50</td>
<td>0.000</td>
</tr>
<tr>
<td>x</td>
<td>-1.0601</td>
<td>0.2414</td>
<td>-4.39</td>
<td>0.001</td>
</tr>
</tbody>
</table>

\[ S = 54.80 \quad R-Sq = 61.6\% \quad R-Sq(adj) = 58.4\% \]

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>Regress</td>
<td>1</td>
<td>57906</td>
<td>57906</td>
<td>19.28</td>
<td>0.001</td>
</tr>
<tr>
<td>Residual</td>
<td>12</td>
<td>36037</td>
<td>3003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>93943</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted Values

<table>
<thead>
<tr>
<th>Fit</th>
<th>StDev Fit</th>
<th>95.0% CI</th>
<th>95.0% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>480.5</td>
<td>19.2</td>
<td>(438.6, 522.4)</td>
<td>(353.9, 607.0)</td>
</tr>
</tbody>
</table>

1. What is the average yield of the grain at 33 meters up the watershed?

2. What is the 95% confidence interval for the average yield at 33 meters up the watershed?
3. What is the prediction interval for the next year’s yield at 33 meters up the watershed?

4. Find the 95% simultaneous confidence bounds at x=33.