Overview

What is a (good) model?

Estimating Model Parameters

Allocating Variation

Confidence Intervals for Regressions
  Parameter Intervals
  Prediction Intervals

Verifying Regression
What Is a (Good) Model?

- For correlated data, model predicts response given an input
- Model should be equation that fits data
- Standard definition of “fits” is least-squares
  - Minimize squared error
  - Keep mean error zero
  - Minimizes variance of errors
What is a (good) model?

Least-Squared Error

- If $\hat{y} = b_0 + b_1x$ then error in estimate for $x_i$ is $e_i = y_i - \hat{y}_i$
- Minimize Sum of Squared Errors (SSE)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- Subject to the constraint

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$
Best regression parameters are

\[
b_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad b_0 = \bar{y} - b_1 \bar{x}
\]

where

\[
\bar{x} = \frac{1}{n} \sum x_i \quad \bar{y} = \frac{1}{n} \sum y_i
\]

Note that book may have errors in these equations!
Parameter Estimation Example

Execution time of a script for various loop counts:

<table>
<thead>
<tr>
<th>Loops</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1.2</td>
<td>1.7</td>
<td>2.5</td>
<td>2.9</td>
<td>3.3</td>
</tr>
</tbody>
</table>

\[ \bar{x} = 6.8, \ y = 2.32, \ \sum xy = 88.54, \ \sum x^2 = 264 \]

\[ b_1 = \frac{88.54 - 5(6.8)(2.32)}{264 - 5(6.8)^2} = 0.29 \]

\[ b_0 = 2.32 - (0.29)(6.8) = 0.35 \]
Estimating Model Parameters

Graph of Parameter Estimation Example
Analysis of Variation (ANOVA):

- If no regression, best guess of $y$ is $\bar{y}$
- Observed values of $y$ differ from $\bar{y}$, giving rise to errors (variance)
- Regression gives better guess, but there are still errors
- We can evaluate quality of regression by allocating sources of errors
The Total Sum of Squares

Without regression, squared error is

\[
\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i^2 - 2y_i\bar{y} + \bar{y}^2) \\
= \left( \sum_{i=1}^{n} y_i^2 \right) - 2\bar{y} \left( \sum_{i=1}^{n} y_i \right) + n\bar{y}^2 \\
= \left( \sum_{i=1}^{n} y_i^2 \right) - 2\bar{y}(n\bar{y}) + n\bar{y}^2 \\
= \left( \sum_{i=1}^{n} y_i^2 \right) - n\bar{y}^2 \\
= \text{SSY} - \text{SS0}
\]
Recall that regression error is

$$\text{SSE} = \sum e_i^2 = \sum (y_i - \bar{y})^2$$

Error without regression is SST (previous slide)

So regression explains $$\text{SSR} = \text{SST} - \text{SSE}$$

Regression quality measured by coefficient of determination

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$
Allocating Variation

Evaluating Coefficient of Determination

- Compute $\text{SST} = (\sum y^2) - n\bar{y}^2$
- Compute $\text{SSE} = \sum y^2 - b_0 \sum y - b_1 \sum xy$
- Compute $R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}$
Example of Coefficient of Determination

For previous regression example:

<table>
<thead>
<tr>
<th>Loops</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1.2</td>
<td>1.7</td>
<td>2.5</td>
<td>2.9</td>
<td>3.3</td>
</tr>
</tbody>
</table>

\[
\sum y = 11.60, \quad \sum y^2 = 29.79, \quad \sum xy = 88.54, \\
\bar{y}^2 = 5(2.32)^2 = 26.9 \\
SSE = 29.79 - (0.35)(11.60) - (0.29)(88.54) = 0.05 \\
SST = 29.79 - 26.9 = 2.89 \\
SSR = 2.89 - 0.05 = 2.84 \\
R^2 = (2.89 - 0.05)/2.89 = 0.98 \]

\[
\]
Variance of errors is SSE divided by degrees of freedom
  - DOF is $n - 2$ because we’ve calculated 2 regression
    parameters from the data
  - So variance (mean squared error, MSE) is $\frac{\text{SSE}}{(n - 2)}$

Standard deviation of errors is square root: $s_e = \sqrt{\frac{\text{SSE}}{n - 2}}$
(minor error in book)
Degrees of freedom always equate:

- SS0 has 1 (computed from $\bar{y}$)
- SST has $n - 1$ (computed from data and $\bar{y}$, which uses up 1)
- SSE has $n - 2$ (needs 2 regression parameters)
- So $\text{SST} = \text{SSY} - \text{SS0} = \text{SSR} + \text{SSE}$
  
  \[
  n - 1 = n - 1 = 1 + (n - 2)
  \]
Allocating Variation

Example of Standard Deviation of Errors

- For regression example, SSE was 0.05, so MSE is $0.05/3 = 0.017$ and $s_e = 0.13$
- Note high quality of our regression:
  - $R^2 = 0.98$
  - $s_e = 0.13$
  - Why such a nice straight-line fit?
Confidence Intervals for Regressions

- Regression is done from a single population sample (size $n$)
  - Different sample might give different results
  - True model is $y = \beta_0 + \beta_1 x$
  - Parameters $b_0$ and $b_1$ are really means taken from a population sample
Calculating Intervals for Regression Parameters

- Standard deviations of parameters:

\[ s_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{x^2}{\sum x^2 - n\bar{x}^2}} \]

\[ s_{b_1} = \frac{s_e}{\sqrt{\sum x^2 - n\bar{x}^2}} \]

- Confidence intervals are \( b_i \pm t_{1-\frac{\alpha}{2};n-2}s_{b_i} \)

- Note that \( t \) has \( n - 2 \) degrees of freedom!
Example of Parameter Confidence Intervals

- Recall $s_e = 0.13$, $n = 5$, $\sum x^2 = 264$, $\bar{x} = 6.8$
- So $s_{b_0} = 0.13 \sqrt{\frac{1}{5} + \frac{(6.8)^2}{264 - 5(6.8)^2}} = 0.16$
  $s_{b_1} = \frac{0.13}{\sqrt{264 - 5(6.8)^2}} = 0.004$
- Using 90% confidence level, $t_{0.95;3} = 2.353$
- Thus, $b_0$ interval is $0.35 \pm 2.353(0.16) = (-0.03, 0.73)$
  - Not significant at 90%
- And $b_1$ is $0.29 \pm 2.353(0.004) = (0.28, 0.30)$
  - Significant at 90% (and would survive even 99.9% test)
Previous confidence intervals are for *parameters*
- How certain can we be that the parameters are correct?

Purpose of regression is *prediction*
- How accurate are the predictions?
- Regression gives mean of predicted response, based on sample we took
Predicting $m$ Samples

- Standard deviation for mean of future sample of $m$ observations at $x_p$ is

$$s_{\hat{y}_{mp}} = se\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x^2 - n\bar{x}^2}}$$

- Note deviation drops as $m \to \infty$
- Variance minimal at $x = \bar{x}$
- Use $t$-quantiles with $n - 2$ DOF for calculating confidence interval
Example of Confidence of Predictions

- Using previous equation, what is predicted time for a single run of 8 loops?
- Time = 0.35 + 0.29(8) = 2.67
- Standard deviation of errors $s_e = 0.13$

\[
s_{\hat{y}_{1,8}} = 0.13 \sqrt{1 + \frac{1}{5} + \frac{(8 - 6.8)^2}{264 - 5(6.8)^2}} = 0.14
\]

- 90% interval is then $2.65 \pm 2.353(0.14) = (2.34, 3.00)$
Prediction Confidence
Regressions are based on assumptions:
- Linear relationship between response $y$ and predictor $x$
  - Or nonlinear relationship used in fitting
- Predictor $x$ nonstochastic and error-free
- Model errors statistically independent
  - With distribution $N(0, c)$ for constant $c$
- If assumptions violated, model misleading or invalid
Testing Linearity

Scatter plot $x$ vs. $y$ to see basic curve type

- **Linear**
- **Piecewise Linear**
- **Outlier**
- **Nonlinear (Power)**
Verifying Regression

Testing Independence of Errors

- Scatter-plot $\varepsilon_i$ versus $\hat{y}_i$
- Should be no visible trend
- Example from our curve fit:

```
0 1 2 3
-0.1
0.0
0.1
0.2
```

2015-06-15
May be useful to plot error residuals versus experiment number
  - In previous example, this gives same plot except for $x$ scaling

No foolproof tests
  - "Independence" test really disproves particular dependence
  - Maybe next test will show different dependence!
Verifying Regression

Testing for Normal Errors

- Prepare quantile-quantile plot of errors
- Example for our regression:
Testing for Constant Standard Deviation

- Tongue-twister: *homoscedasticity*
- Return to independence plot
- Look for trend in spread
- Example:

![Graph with data points]
Regression throws away some information about the data
  To allow more compact summarization
Sometimes vital characteristics are thrown away
  Often, looking at data plots can tell you whether you will have a problem
Example of Misleading Regression

<table>
<thead>
<tr>
<th>I</th>
<th>x</th>
<th>y</th>
<th>II</th>
<th>x</th>
<th>y</th>
<th>III</th>
<th>x</th>
<th>y</th>
<th>IV</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
<td>10.94</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
<td>6.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8.14</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
<td>5.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>8.74</td>
<td>13</td>
<td>12.74</td>
<td>8</td>
<td>7.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>8.77</td>
<td>9</td>
<td>7.11</td>
<td>8</td>
<td>8.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>9.26</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
<td>8.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>8.10</td>
<td>14</td>
<td>8.84</td>
<td>8</td>
<td>7.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6.13</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
<td>5.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>3.10</td>
<td>4</td>
<td>5.39</td>
<td>19</td>
<td>12.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>9.13</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
<td>5.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7.26</td>
<td>7</td>
<td>6.42</td>
<td>8</td>
<td>7.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>4.74</td>
<td>5</td>
<td>5.73</td>
<td>8</td>
<td>6.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2015-06-15

CS147

Verifying Regression

Example of Misleading Regression
What Does Regression Tell Us?

- Exactly the same thing for each data set!
- $n = 11$
- Mean of $y = 7.5$
- $y = 3 + 0.5x$
- Standard error of regression is 0.118
- All the sums of squares are the same
- Correlation coefficient = 0.82
- $R^2 = 0.67$
Now Look at the Data Plots
Now Look at the Data Plots

Verifying Regression

Now Look at the Data Plots
Now Look at the Data Plots
Now Look at the Data Plots
Now Look at the Data Plots

Verifying Regression

Now Look at the Data Plots