Stat13-lecture 25
regression (continued, SE, t and chi-square)

- Simple linear regression model:
  $$Y = \beta_0 + \beta_1 X + \epsilon$$

- Assumption: \(\epsilon\) is normal with mean 0 variance \(\sigma^2\)

  The fitted line is obtained by minimizing the sum of squared residuals; that is finding \(\beta_0\) and \(\beta_1\) so that
  $$(Y_1 - \beta_0 - \beta_1 X_1)^2 + \ldots + (Y_n - \beta_0 - \beta_1 X_n)^2$$
  is as small as possible

- This method is called least squares method
Least square line is the same as the regression line discussed before

- It follows that estimated slope $\hat{b}_1$ can be computed by
- \[ r \left[ \frac{\text{SD}(Y)}{\text{SD}(X)} \right] = \frac{\text{cov}(X,Y)}{\text{SD}(X)\text{SD}(Y)} \left[ \frac{\text{SD}(Y)}{\text{SD}(X)} \right] \]
- \[ = \frac{\text{cov}(X,Y)}{\text{VAR}(X)} \] (this is the same as equation for $\hat{b}_1$ on page 518)
- The intercept $\hat{b}_0$ is estimated by putting $x=0$ in the regression line; yielding equation on page 518
- Therefore, there is no need to memorize the equation for least square line; computationally it is advantageous to use $\frac{\text{cov}(X,Y)}{\text{VAR}(X)}$ instead of $r \left[ \frac{\text{SD}(Y)}{\text{SD}(X)} \right]$
Finding residuals and estimating the variance of $\sigma^2$

- Residuals = differences between $Y$ and the regression line (the fitted line)
- An unbiased estimate of $\sigma^2$ is $\frac{\text{sum of squared residuals}}{(n-2)}$
- Which divided by $(n-2)$?
- Degree of freedom is $n-2$ because two parameters were estimated
- $\frac{\text{sum of squared residuals}}{\sigma^2}$ follows a chi-square.
Hypothesis testing for slope

• Slope estimate $\hat{\beta}_1$ is random
• It follows a normal distribution with mean equal to the true $\beta_1$ and the variance equal to $\sigma^2 / [n \text{ var}(X)]$

Because $\sigma^2$ is unknown, we have to estimate from the data; the SE (standard error) of the slope estimate is equal to the squared root of the above
t-distribution

• Suppose an estimate hat $\hat{q}$ is normal with variance $c \sigma^2$.
• Suppose $\sigma^2$ is estimated by $s^2$ which is related to a chi-squared distribution.
• Then $(\hat{q} - q) / \sqrt{(c s^2)}$ follows a t-distribution with the degrees of freedom equal to the chi-square degree freedom.
An example

• Determining small quantities of calcium in presence of magnesium is a difficult problem of analytical chemists. One method involves use of alcohol as a solvent.

• The data below show the results when applying to 10 mixtures with known quantities of CaO. The second column gives

• Amount CaO recovered.

• Question of interest: test to see if intercept is 0; test to see if slope is 1.
<table>
<thead>
<tr>
<th>X: CaO present</th>
<th>Y: CaO recovered</th>
<th>Fitted value</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>3.7</td>
<td>3.751</td>
<td>-0.051</td>
</tr>
<tr>
<td>8.0</td>
<td>7.8</td>
<td>7.73</td>
<td>0.070</td>
</tr>
<tr>
<td>12.5</td>
<td>12.1</td>
<td>12.206</td>
<td>-0.106</td>
</tr>
<tr>
<td>16.0</td>
<td>15.6</td>
<td>15.688</td>
<td>-0.088</td>
</tr>
<tr>
<td>20.0</td>
<td>19.8</td>
<td>19.667</td>
<td>0.133</td>
</tr>
<tr>
<td>25.0</td>
<td>24.5</td>
<td>24.641</td>
<td>-0.141</td>
</tr>
<tr>
<td>31.0</td>
<td>31.1</td>
<td>30.609</td>
<td>0.491</td>
</tr>
<tr>
<td>36.0</td>
<td>35.5</td>
<td>35.583</td>
<td>-0.083</td>
</tr>
<tr>
<td>40.0</td>
<td>39.4</td>
<td>39.562</td>
<td>-0.161</td>
</tr>
<tr>
<td>40.0</td>
<td>39.5</td>
<td>39.562</td>
<td>-0.062</td>
</tr>
</tbody>
</table>
Least Squares Estimates:

- Constant: $-0.228090 \pm 0.137840$
- Predictor: $0.994757 \pm 0.005219$

- $R^2$: 0.999780
- $\sigma$: 0.206722
- Number of cases: 10
- Degrees of freedom: 8

\[
\frac{0.22809}{0.1378} = 1.6547
\]

\[
\frac{1 - 0.994757}{5.219485 \times 10^{-3}} = 1.0045052337539044
\]