STA 4033
Mathematical Statistics
with Computer Applications
Chapter 7∗

Brett Presnell†
Department of Statistics‡
University of Florida§

10 June 2002

Contents

7 Simple Linear Regression 2
  7.1 A Statistical Model ........................................... 2
  7.2 Least Squares Estimators ....................................... 5
  7.3 Formulas for the LSEs .......................................... 6
  7.4 Standard Errors and Estimating the Variance ............... 6
  7.5 Formal Inference ................................................ 9
    7.5.1 Inference About the Slope .............................. 10
    7.5.2 Inference About the Intercept ......................... 10
    7.5.3 Inference About the Mean .............................. 11
  7.6 Prediction Intervals ........................................... 11

∗Copyright ©2002 by Brett Presnell
†http://www.stat.ufl.edu/presnell/
‡http://www.stat.ufl.edu/
§http://www.ufl.edu/
7 Simple Linear Regression

Suppose that we want to find the best fitting line to a set of points \((x_1, y_1), \ldots, (x_n, y_n)\), such as the heights and weights of 22 male college students (see Figure 1).

- If all the points lie on a single straight line this is easy, but real data is never like that.
- So, what does “best fitting” mean?
- One answer is “least squares”:
  
  Choose the slope \(b\) and the intercept \(a\) to minimize the sum of squares of the vertical distances between the points and the line.

  - Let \(\hat{y}_i = a + bx_i\), so \((x_i, \hat{y}_i)\) is the point on the line with \(x\)-coordinate \(x_i\).
  - Let \(e_i = y_i - \hat{y}_i\).
  - Then we choose \(a\) and \(b\) to minimize
    
    \[
    \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2
    \]

    - It is a simple calculus problem to find \(a\) and \(b\) (differentiate with respect to \(a\) and \(b\), set the two derivatives equal to 0, and solve for \(a\) and \(b\)).

See Figure 2.

7.1 A Statistical Model

Take the \(x\) values as given. A statistical model for the dependence of \(Y\) on \(x\) is

\[
Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where

- \(\varepsilon_1, \ldots, \varepsilon_n\) are independent random variables, with
- \(E(\varepsilon_i) = 0\).
Figure 1: Heights and weights of 22 male students (Summer 2001).
Male Student Heights and Weights

Figure 2: Heights and weights of 22 male students (Summer 2001). Least squares line and residuals.
The random errors $\varepsilon_1, \ldots, \varepsilon_n$ model the failure of the observations to fall exactly on a straight line.

Note: another way of expressing the model is to say that $Y_1, \ldots, Y_n$ are independent random variables with

$$E(Y_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$ 

So the model says that the mean of $Y$ is a linear function of $x$.

**Interpretation of $\beta_1$:** The slope, $\beta_1$, is the change in the mean of $Y$ associated with a unit increase in $x$:

- If $x$ increases by 1, then $E(Y)$ changes by $\beta_1$.
- If $x$ changes by $\Delta x$, then $E(Y)$ changes by $\beta_1 \times \Delta x$.

### 7.2 Least Squares Estimators

Given data $(x_1,y_1), \ldots, (x_n,y_n)$, the *least squares estimators* (LSEs) of $\beta_0$ and $\beta_1$ are

- denoted $\hat{\beta}_0$ and $\hat{\beta}_1$
- chosen to minimize the "sum of squared residuals"

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

- where the *ith residual* is $e_i = y_i - \hat{y}_i$
- and the *ith fitted value* is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- The line with equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

is called the *fitted regression line*.

Assuming that the model is correct, the LSEs are unbiased, i.e.,

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad E(\hat{\beta}_1) = \beta_1$$
**Example.** For the height/weight example (male students, Summer 2001), the least squares estimates are

\[
\hat{\beta}_0 = -82.74 \quad \text{and} \quad \hat{\beta}_1 = 3.56
\]

Thus fitted line is

\[
\hat{y} = -82.74 + 3.56x
\]

Thus we estimate that on the average, weight increases by 3.56 pounds for every 1 inch increase in height.

The data together with the fitted values and residuals are given in Table 1.

See R transcript.

### 7.3 Formulas for the LSEs

Here are the formulas for the LSEs in simple linear regression.

\[
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]

where \(S_{xx}, S_{yy}, \) and \(S_{xy}\) represent the *corrected sums of squares and cross-products*,

\[
S_{xx} = \sum_{i=1}^{n} (x - \bar{x})^2 \quad S_{yy} = \sum_{i=1}^{n} (y - \bar{y})^2 \quad S_{xy} = \sum_{i=1}^{n} (x - \bar{x})(y - \bar{y})
\]

### 7.4 Standard Errors and Estimating the Variance

In order to carry out more formal statistical inference, we must add some additional assumptions to the model. First we add the assumption of homogeneity of variances (equal variances):

\[
Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,
\]

where

- \(\epsilon_1, \ldots, \epsilon_n\) are independent random variables, with
- \(E(\epsilon_i) = 0\) for all \(i = 1, \ldots, n\), and
- \(\text{Var}(\epsilon_i) = \sigma^2\) for all \(i = 1, \ldots, n\).
Table 1: Data, fitted values, and residuals for height/weight example (male students, Summer 2001).

<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
<th>Fitted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>66.0</td>
<td>165</td>
<td>152.2</td>
<td>12.8</td>
</tr>
<tr>
<td>71.5</td>
<td>198</td>
<td>171.8</td>
<td>26.2</td>
</tr>
<tr>
<td>70.0</td>
<td>170</td>
<td>166.4</td>
<td>3.6</td>
</tr>
<tr>
<td>66.0</td>
<td>135</td>
<td>152.2</td>
<td>−17.2</td>
</tr>
<tr>
<td>71.0</td>
<td>125</td>
<td>170.0</td>
<td>−45.0</td>
</tr>
<tr>
<td>69.0</td>
<td>150</td>
<td>162.9</td>
<td>−12.9</td>
</tr>
<tr>
<td>72.0</td>
<td>190</td>
<td>173.6</td>
<td>16.4</td>
</tr>
<tr>
<td>63.0</td>
<td>128</td>
<td>141.5</td>
<td>−13.5</td>
</tr>
<tr>
<td>71.0</td>
<td>150</td>
<td>170.0</td>
<td>−20.0</td>
</tr>
<tr>
<td>70.0</td>
<td>170</td>
<td>166.4</td>
<td>3.6</td>
</tr>
<tr>
<td>72.0</td>
<td>195</td>
<td>173.6</td>
<td>21.4</td>
</tr>
<tr>
<td>73.0</td>
<td>225</td>
<td>177.1</td>
<td>47.9</td>
</tr>
<tr>
<td>68.0</td>
<td>150</td>
<td>159.3</td>
<td>−9.3</td>
</tr>
<tr>
<td>72.0</td>
<td>155</td>
<td>173.6</td>
<td>−18.6</td>
</tr>
<tr>
<td>67.0</td>
<td>165</td>
<td>155.8</td>
<td>9.2</td>
</tr>
<tr>
<td>73.0</td>
<td>183</td>
<td>177.1</td>
<td>5.9</td>
</tr>
<tr>
<td>64.0</td>
<td>125</td>
<td>145.1</td>
<td>−20.1</td>
</tr>
<tr>
<td>61.0</td>
<td>187</td>
<td>134.4</td>
<td>52.6</td>
</tr>
<tr>
<td>75.0</td>
<td>200</td>
<td>184.2</td>
<td>15.8</td>
</tr>
<tr>
<td>71.5</td>
<td>140</td>
<td>171.8</td>
<td>−31.8</td>
</tr>
<tr>
<td>68.0</td>
<td>155</td>
<td>159.3</td>
<td>−4.3</td>
</tr>
<tr>
<td>69.0</td>
<td>140</td>
<td>162.9</td>
<td>−22.9</td>
</tr>
</tbody>
</table>
With this model it can be shown that
\[
\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad \text{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{nS_{xx}}, \quad \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}\sigma^2}{S_{xx}}
\]

**Recall:** For random variables \(X\) and \(Y\),
\[
\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]
\]

Also, if \(X\) and \(Y\) are independent, then \(\text{Cov}(X, Y) = 0\).

**Recall:** For any two random variables \(X\) and \(Y\) and constants \(a\) and \(b\),
\[
E(aX + bY) = aE(X) + bE(Y)
\]

and
\[
\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y).
\]

It follows that with the model above,
\[
E(\hat{Y}|x) = E(\hat{\beta}_0 + \hat{\beta}_1 x) = \hat{\beta}_0 + \hat{\beta}_1 x = E(Y|x)
\]

(so \(\hat{Y}\) is an unbiased estimator of the mean of \(Y\) at \(x\)) and
\[
\text{Var}(\hat{Y}|x) = 1^2 \text{Var}(\hat{\beta}_0) + x^2 \text{Var}(\hat{\beta}_1) + 2(1)(x) \text{Cov}(\hat{\beta}_0, \hat{\beta}_1)
\]
\[
= \frac{\sigma^2 \sum x_i^2}{nS_{xx}} + x^2 \frac{\sigma^2}{S_{xx}} + 2x\frac{-\bar{x}\sigma^2}{S_{xx}}
\]
\[
= \ldots \text{algebra} \ldots
\]
\[
= \sigma^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]
\]

Define the **sum of squares for error** to be
\[
\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]

It can be shown that
\[
E(\text{SSE}) = (n - 2)\sigma^2.
\]

Thus an unbiased estimator of \(\sigma^2\) is the **mean squared error**, denoted by various authors as MSE or \(S^2\) or \(\hat{\sigma}^2\), and given by
\[
\hat{\sigma}^2 = \frac{\text{SSE}}{n - 2}
\]
Note: We say that the SSE (or $\hat{\sigma}^2$) has $n - 2$ degrees of freedom. Roughly speaking, we lose two degrees of freedom because we must first estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ in order to calculate SSE.

These results are summarized in Table 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unbiased Estimator</th>
<th>Std Error of Estimator</th>
<th>Estimated Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>$\hat{\beta}_0$</td>
<td>$\sigma \sqrt{\frac{\sum x_i^2}{nS_{xx}}}$</td>
<td>$\hat{\sigma} \sqrt{\frac{\sum x_i^2}{nS_{xx}}}$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$\hat{\beta}_1$</td>
<td>$\sigma \sqrt{\frac{1}{S_{xx}}}$</td>
<td>$\hat{\sigma} \sqrt{\frac{1}{S_{xx}}}$</td>
</tr>
</tbody>
</table>

$E(Y|X) = \beta_0 + \beta_1 x$ \quad $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ \quad $\sigma \sqrt{\frac{1}{n} + \frac{(x-x)^2}{S_{xx}}}$ \quad $\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x-x)^2}{S_{xx}}}$

$\sigma^2 = \frac{\text{SSE}}{n-2}$

Table 2: Estimators in simple linear regression and their standard errors.

7.5 Formal Inference

In order to carry out formal inference (confidence intervals and tests of hypotheses), we need to add one more assumption (normality) to our regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where

- $\epsilon_1, \ldots, \epsilon_n$ are independent random variables, with
- $E(\epsilon_i) = 0$ for all $i = 1, \ldots, n,$
- $\text{Var}(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n,$ and
- $\epsilon_1, \ldots, \epsilon_n$ are normally distributed.

Note: The results that follow hold (approximately) even without the normality assumption, as long as the sample size is large.
7.5.1 Inference About the Slope

Often in a simple linear regression model, interest centers on the relationship between the predictor $x$ and the response $Y$ as reflected in the slope $\beta_1$.

Under the model above, a $(1 - \alpha)100\%$ CI for $\beta_1$ is given by

$$\hat{\beta}_1 \pm t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{S_{xx}}}$$

where $t_{\alpha/2}$ is the upper $(\alpha/2)$th quantile of a $t$ distribution with $n - 2$ degrees of freedom.

Similarly, to test $H_0 : \beta_1 = 0$ against any of the alternatives $H_a : \beta_1 > 0$, $H_a : \beta_1 < 0$, or $H_0 : \beta_1 \neq 0$, we refer the test statistic

$$T = \frac{\hat{\beta}_1}{\hat{\sigma} \sqrt{\frac{1}{S_{xx}}}}$$

to a $t$ distribution with $n - 2$ degrees of freedom.

7.5.2 Inference About the Intercept

Generally the parameter $\beta_0$ is not of much direct interest. Nevertheless, confidence intervals and test of hypotheses about $\beta_0$ can be constructed in the obvious way.

Under the model above, a $(1 - \alpha)100\%$ CI for $\beta_0$ is given by

$$\hat{\beta}_0 \pm t_{\alpha/2} \hat{\sigma} \sqrt{\frac{\sum x_i^2}{nS_{xx}}}$$

where $t_{\alpha/2}$ is the upper $(\alpha/2)$th quantile of a $t$ distribution with $n - 2$ degrees of freedom.

Similarly, to test $H_0 : \beta_0 = 0$ against any of the alternatives $H_a : \beta_0 > 0$, $H_a : \beta_0 < 0$, or $H_0 : \beta_0 \neq 0$, we refer the test statistic

$$T = \frac{\hat{\beta}_0}{\hat{\sigma} \sqrt{\frac{\sum x_i^2}{nS_{xx}}}}$$

to a $t$ distribution with $n - 2$ degrees of freedom.
7.5.3 Inference About the Mean

It is often of interest to estimate the mean response $Y$ for a given value of the predictor $x$.

Under the model above, a $(1 - \alpha)100\%$ CI for $E(Y|x) = \beta_0 + \beta_1 x$ is given by

$$
\hat{y} \pm t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}
$$

where $t_{\alpha/2}$ is the upper $(\alpha/2)$th quantile of a $t$ distribution with $n - 2$ degrees of freedom.

7.6 Prediction Intervals

If we are predicting a new (independent) $Y$ given the corresponding $x$ value, then we must take into account the variability of $Y$ as well as the variability in our best prediction, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$, of the value of $Y$. Since the new observation is statistically independent of our original data,

$$
\text{Var}(Y - \hat{Y}) = \text{Var}(Y) + \text{Var}(\hat{Y}) = \sigma^2 + \sigma^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]
$$

$$
= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]
$$

A $(1 - \alpha) \times 100\%$ prediction interval for $Y$ is thus given by

$$
\hat{y} \pm t_{\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}
$$

where $t_{\alpha/2}$ is the upper $(\alpha/2)$th quantile of a $t$ distribution with $n - 2$ degrees of freedom.

Note that this prediction interval is usually much wider than a confidence interval for the mean of $Y$ at the given value of $x$.

See R transcript.