Basic Linear Regression

Dave Goldsman

Georgia Institute of Technology, Atlanta, GA, USA

1/7/15
Simple Linear Regression Model

Basic Properties

Confidence Intervals and Inference for $\beta_0$ and $\beta_1$
Suppose we have a data set with the following paired observations:

\[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\]

Example:

\[x_i = \text{height of person } i\]
\[y_i = \text{weight of person } i\]

Can we make a model expressing \(y_i\) as a function of \(x_i\)?
Estimate \( y_i \) for fixed \( x_i \). Let’s model this with the simple linear regression equation,

\[
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,
\]

where \( \beta_0 \) and \( \beta_1 \) are unknown constants and the error terms are usually assumed to be

\[
\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2) \quad \Rightarrow \quad y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).
\]
Simple Linear Regression Model

$y = \beta_0 + \beta_1 x$

with “high” $\sigma^2$

$y = \beta_0 + \beta_1 x$

with “low” $\sigma^2$
Warning! Look at data before you fit a line to it:

doesn’t look very linear!
<table>
<thead>
<tr>
<th></th>
<th>$x_i$</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>4.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Feb</td>
<td>3.6</td>
<td>2.3</td>
</tr>
<tr>
<td>Mar</td>
<td>4.3</td>
<td>2.5</td>
</tr>
<tr>
<td>Apr</td>
<td>5.1</td>
<td>2.8</td>
</tr>
<tr>
<td>May</td>
<td>5.6</td>
<td>3.0</td>
</tr>
<tr>
<td>Jun</td>
<td>5.0</td>
<td>3.1</td>
</tr>
<tr>
<td>Jul</td>
<td>5.3</td>
<td>3.2</td>
</tr>
<tr>
<td>Aug</td>
<td>5.8</td>
<td>3.5</td>
</tr>
<tr>
<td>Sep</td>
<td>4.7</td>
<td>3.0</td>
</tr>
<tr>
<td>Oct</td>
<td>5.6</td>
<td>3.3</td>
</tr>
<tr>
<td>Nov</td>
<td>4.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Dec</td>
<td>4.2</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Great... but how do you fit the line?
Fit the regression line \( y = \beta_0 + \beta_1 x \) to the data 

\[(x_1, y_1), \ldots, (x_n, y_n)\]

by finding the “best” match between the line and the data. The “best” choice of \( \beta_0, \beta_1 \) will be chosen to minimize

\[
Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_{i=1}^{n} \varepsilon_i^2.
\]
This is called the *least squares* fit. Let’s solve...

\[
\frac{\partial Q}{\partial \beta_0} = -2 \sum (y_i - (\beta_0 + \beta_1 x_i)) = 0
\]

\[
\frac{\partial Q}{\partial \beta_1} = -2 \sum x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0
\]

\[
\sum y_i = n \beta_0 + \beta_1 \sum x_i
\]

\[
\sum x_i y_i = \sum x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0
\]

After a little algebra, get

\[
\hat{\beta}_1 = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \text{ where } \bar{y} \equiv \frac{1}{n} \sum y_i \text{ and } \bar{x} \equiv \frac{1}{n} \sum x_i.
\]
Let’s introduce some more notation:

\[ S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \]

\[ S_{xy} = \sum (x_i - \bar{x})y_i = \sum (x_i - \bar{x})(y_i - \bar{y}) \]

\[ = \sum x_i y_i - n\bar{x}\bar{y} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \]

These are called *sums of squares*. 
Then, after a little more algebra, we can write

\[
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}
\]

Fact: If the \( \varepsilon_i \)'s are iid \( N(0, \sigma^2) \), it can be shown that \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are the maximum likelihood estimators for \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), respectively. (See any text for easy proof).

Anyhow, the fitted regression line is:

\[
\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.
\]
Fix a specific value of the explanatory variable $x^*$, the equation gives a fitted value $\hat{y}|x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ for the dependent variable $y$. 

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]
Notation Summary: For actual data points $x_i$, the fitted values are $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

observed values: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Example: Suppose

\[ n = 12, \quad \sum_{i=1}^{12} x_i = 58.62, \quad \sum y_i = 34.15, \]

\[ \sum x_i^2 = 291.231, \quad \sum y_i^2 = 98.697, \quad \sum x_i y_i = 169.253 \]

These give \( \hat{\beta}_0 = 0.4090 \) and \( \hat{\beta}_1 = 0.49883 \), and so the fitted regression line is

\[ \hat{y} = 0.409 + 0.499x. \]

For example, \( \hat{y}|_{5.5} = 3.1535. \)
Outline

1. Simple Linear Regression Model
2. Basic Properties
3. Confidence Intervals and Inference for $\beta_0$ and $\beta_1$
Since the \( y_i \)'s are independent with \( y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \) (and the \( x_i \)'s are constants), we have

\[
E[\hat{\beta}_1] = \frac{1}{S_{xx}} E[S_{xy}] = \frac{1}{S_{xx}} \sum (x_i - \bar{x}) E[y_i]
\]

\[
= \frac{1}{S_{xx}} \sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i)
\]

\[
= \frac{1}{S_{xx}} \left[ \beta_0 \sum (x_i - \bar{x}) + \beta_1 \sum (x_i - \bar{x}) x_i \right]
\]

\[
= \frac{\beta_1}{S_{xx}} \sum (x_i^2 - x_i \bar{x}) = \frac{\beta_1}{S_{xx}} \left( \sum x_i^2 - n \bar{x}^2 \right)
\]

\[
= \beta_1
\]

Thus, \( \hat{\beta}_1 \) is an unbiased estimator of \( \beta_1 \).
Further, since $\hat{\beta}_1$ is a linear combination of independent normals, $\hat{\beta}_1$ is itself normal. We can also derive

$$
\text{Var}(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \text{Var}(S_{xy}) = \frac{1}{S_{xx}^2} \sum (x_i - \bar{x})^2 \text{Var}(y_i) = \frac{\sigma^2}{S_{xx}}.
$$

Thus, $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$. 
While we’re at it, we can do the same kind of thing with the intercept parameter, $\beta_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$  

We have

$$E[\hat{\beta}_0] = E[\bar{y}] - \bar{x} E[\hat{\beta}_1] = \beta_0 + \beta_1 \bar{x} - \bar{x} \beta_1 = \beta_0,$$

so that $\hat{\beta}_0$ is unbiased for $\beta_0$.

Similar to before, since $\hat{\beta}_0$ is a linear combination of independent normals, it is also normal. Finally,

$$\text{Var}(\hat{\beta}_0) = \frac{\sum x_i^2}{nS_{xx}} \sigma^2.$$
Proof:

\[
\text{Cov}(\bar{y}, \hat{\beta}_1) = \frac{1}{S_{xx}} \text{Cov}(\bar{y}, \sum (x_i - \bar{x})y_i)
\]

\[
= \frac{\sum (x_i - \bar{x})}{S_{xx}} \text{Cov}(\bar{y}, y_i)
\]

\[
= \frac{\sum (x_i - \bar{x}) \sigma^2}{S_{xx} n} = 0
\]

\[\Rightarrow \text{Var}(\hat{\beta}_0) = \text{Var}(\bar{y} - \hat{\beta}_1 \bar{x})\]

\[
= \text{Var}(\bar{y}) + \bar{x}^2 \text{Var}\hat{\beta}_1 - 2\bar{x} \text{Cov}(\bar{y}, \hat{\beta}_1)
\]

\[
= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}}
\]

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

Thus, \(\hat{\beta}_0 \sim N(\beta_0, \frac{\sum x_i^2}{nS_{xx}} \sigma^2).\)
Now let’s estimate the error variation $\sigma^2$ by considering the deviations between $y_i$ and $\hat{y}_i$, i.e., the sum of squared errors,

$$SSE \equiv \sum (y_i - \hat{y}_i)^2$$

$$= \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

$$= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i.$$

Turns out that a good estimator for $\sigma^2$ is

$$\hat{\sigma}^2 \equiv \frac{SSE}{n - 2} \sim \frac{\sigma^2 \chi^2(n - 2)}{n - 2}.$$
Outline

1. Simple Linear Regression Model
2. Basic Properties
3. Confidence Intervals and Inference for $\beta_0$ and $\beta_1$
Back to $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / S_{xx}) \ldots$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0, 1)$$

In addition, it turns out:

1. $\hat{\sigma}^2 = \frac{SSE}{n-2} \sim \frac{\sigma^2 \chi^2(n-2)}{n-2}$;
2. $\hat{\sigma}^2$ is independent of $\hat{\beta}_1$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}} \sim \frac{N(0, 1)}{\sqrt{\hat{\sigma}^2 / \sigma^2}} \sim t(n - 2)$$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} / \sqrt{S_{xx}}} \sim t(n - 2).$$
Confidence Intervals and Inference for $\beta_0$ and $\beta_1$

\[ t(n - 2) \]

\[ 1 - \alpha \]

\[ -t_{\alpha/2, n-2} \quad \text{and} \quad t_{\alpha/2, n-2} \]
Confidence Intervals and Inference for $\beta_0$ and $\beta_1$

2-sided Confidence Intervals for $\beta_1$:

$$1 - \alpha = P\left(-t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \leq t_{\alpha/2, n-2}\right)$$

$$= P\left(\hat{\beta}_1 - t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)$$

1-sided CI’s for $\beta_1$:

$$\beta_1 \in (-\infty, \hat{\beta}_1 + t_{\alpha, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}})$$

$$\beta_1 \in (\hat{\beta}_1 - t_{\alpha, n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \infty)$$

Can also do CI’s for $\beta_0$ as well as hypothesis testing.