INTRODUCTORY NOTES ON LINEAR REGRESSION

We have data of the form \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\). These will most often be presented to us as two column of a spreadsheet.

As the topic develops we will see both upper case and lower case letters. Where reasonable, we use upper case \(X_i\) and \(Y_i\) to denote random quantities, lower case \(x_i\) and \(y_i\) to denote non-random quantities which are possibly the observed values of random \(X_i\) and \(Y_i\). We are not able to enforce this distinction in a consistent way, so for now do not be overly concerned.

Our first objective will be giving an equation relating \(Y\) and \(x\). This equation will be \(Y = b_0 + b_1 x\). This is called simple linear regression. Here’s it’s regressing \(Y\) on \(X\). The “simple” is used to distinguish this problem from more complicated ones such as \(Y = b_0 + b_1 x + b_2 u + b_3 v + b_4 w\).

The equation is sometimes presented in form \(Y = a + b x\). This is reasonable, but it’s not the notation we use.

The notation \(Y = m x + b\) seems to be universal in high school algebra courses in the United States. This notation does not seem to be used elsewhere, and we won’t use it here.

We should certainly ask: *Why are we doing linear regression?*

Reminder on the equation of a line. If \(Y = b_0 + b_1 x\), then we can interpret \(b_1\) as the increase in \(Y\) for a one unit increase in \(x\); this is the slope. Also in this spirit, \(b_0\) is the value of \(Y\) corresponding to \(x = 0\) (if this concept even makes sense in the context of the actual example); this is the intercept.

The least squares criterion used to find values of \(b_0\) and \(b_1\) (sometimes \(a\) and \(b\)) to minimize \(\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2\).

The minimization process (by calculus or by other tricks) leads to these steps:

\[(1) \quad \text{Find the five sums } \sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} y_i^2, \sum_{i=1}^{n} x_i y_i.\]
(2) Find the five derived quantities \( \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \)

\[
S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}, \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n},
\]

\[
S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}.
\]

(3) Give the slope estimate as \( b_1 = \frac{S_{xy}}{S_{xx}} \) and the intercept estimate as \( b_0 = \bar{y} - b_1 \bar{x} \).

(4) For later use, record \( S_{yy|x} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \).

If you acquire a new value of \( X \), say \( x_{new} \), then the prediction is \( \hat{Y}_{new} = b_0 + b_1 x_{new} \).

So . . . why are we doing this activity of putting a straight line through a cloud of data?

Maybe we want to predict \( Y \) corresponding to a new \( x \).

Maybe we have a special interest in the slope coefficient \( b_1 \) (an elasticity or marginal cost argument).

It’s standard methodology. Moreover, everyone else will do it, so you may as well see the stuff that they are going to find anyway.

Maybe we want to assess the strength of the relationship between the two variables.

Maybe we’re just curious.
There is a concept of randomness here that we’ve not yet pinned down. We’ll do this next. There’s also a concept that there is some underlying truth, and that we are trying to get at the truth with noisy data.

In fact, we say that the truth is \( Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \), where \( Y_i = \beta_0 + \beta_1 x_i \) would represent points on an exact straight line. The subscript \( i \) runs over 1, 2, …, \( n \). The \( \varepsilon_i \) values are statistical noise drawn from some mechanism which produces (over the long haul) averages of 0 and standard deviations of \( \sigma \), which is unknown to us. The challenge for us is that \( \beta_0 \) and \( \beta_1 \) are not known.

The statements that we are making here constitute a model, and we’re not sure if this model is really correct. (We will encounter similar instances of nervousness.)

We sometime say that \( \beta_0 + \beta_1 x_i \) is the signal while \( \varepsilon_i \) is the noise. We get to see \( Y_i = \text{signal} + \text{noise} \).

The model is \( Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \) is the model, thinking of \( i \) as running over 1, 2, …, \( n \). Sometimes people will drop the \( i \) subscript and write simply \( Y = \beta_0 + \beta_1 x + \varepsilon \).

The fitted model, obtained from the data is \( \hat{Y} = b_0 + b_1 x \).

In some books, the symbol \( \hat{\beta}_0 \) is used instead of \( b_0 \), and \( \hat{\beta}_1 \) is used instead of \( b_1 \).

Thus, you’ll see the fitted line as \( \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \). This is often written with \( Y \) or \( y \) rather than \( \hat{Y} \).

Let’s separate known vs unknown, random vs nonrandom:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Observed or unobserved?</th>
<th>Random or Nonrandom?</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_i )</td>
<td>Observed</td>
<td>Nonrandom *</td>
</tr>
<tr>
<td>( Y_i )</td>
<td>Observed</td>
<td>Random</td>
</tr>
<tr>
<td>( \beta_1, \beta_0 )</td>
<td>Unobserved</td>
<td>Nonrandom</td>
</tr>
<tr>
<td>( b_0, b_1 )</td>
<td>Observed</td>
<td>Random</td>
</tr>
<tr>
<td>( \varepsilon_i )</td>
<td>Unobserved</td>
<td>Random</td>
</tr>
</tbody>
</table>

* In virtually every regression situation, the \( x_i \)’s are nonrandom, but we agree to treat them as though we actually planned the investigation around those particular values. We say that we do the work “conditional on the \( x \) values actually obtained.” This verbal sleight of hand lets us treat them as nonrandom.

The table above may be mysterious so far. This will be discussed in class.
The regression calculations for \( b_0 \) and \( b_1 \) estimate the truth, meaning \( \beta_0 \) and \( \beta_1 \).

We can also estimate \( \sigma \), the standard deviation of the noise terms. This estimate will be denoted as \( s \) or possibly \( s_\varepsilon \) or possibly as \( s_{y\mid x} \). The formula for this is

\[
    s_\varepsilon = \sqrt{\frac{1}{n-2} S_{xy|x}}
\]

There are other formulas as well.

This formula applies only for simple linear regression problems, those with one predictor.

Most regressions are summarized in the analysis of variance table.

The quantity \( S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \) measures variation in \( Y \). Indeed we get \( s_y \), the standard deviation of the \( Y \)'s, as \( s_y = \sqrt{\frac{S_{yy}}{n-1}} \).

We use the symbol \( \hat{Y}_i \) to denote the fitted value for point \( i \). This is computed from the estimated intercept and slope as \( \hat{Y}_i = b_0 + b_1 x_i \). Compare this expression to the model equation.

One can show that \[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \] These sums have the names \( SS_{\text{total}} \), \( SS_{\text{regression}} \), and \( SS_{\text{error}} \). They have other names or abbreviations. For instance

\[
    SS_{\text{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \text{ may be written as } SS_{\text{tot}}. 
\]

\[
    SS_{\text{regression}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \text{ may be written as } SS_{\text{reg}}, SS_{\text{fit}}, \text{ or } SS_{\text{model}}. 
\]

\[
    SS_{\text{error}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{ may be written as } SS_{\text{err}}, SS_{\text{residual}}, SS_{\text{resid}}, \text{ or } SS_{\text{res}}. 
\]
This derivation starts from \( SS_{\text{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \).

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 = \sum_{i=1}^{n} ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2
\]

\[
= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
\]

\[
= SS_{\text{error}} + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + SS_{\text{regression}}
\]

Thus, all we have to deal with is the middle term; we need to show that it is equal to zero.

\[
\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))(b_0 + b_1 x_i - \bar{y})
\]

\[
= \sum_{i=1}^{n} (y_i - (\bar{y} - b_0 x_i + b_1 x_i))(b_0 - b_0 x_i + b_1 x_i - \bar{y})
\]

\[
= \sum_{i=1}^{n} (y_i - \bar{y} - b_1 (x_i - \bar{x}))(b_1 (x_i - \bar{x})) = b_1 (S_{xy} - b_1 S_{xx})
\]

However, \( b_1 = \frac{S_{xy}}{S_{xx}} \), so that this is indeed zero!

The more general proof (for two or more predictors) will be done through matrix algebra!

The degrees of freedom accounting is this:

- \( SS_{\text{total}} \) has \( n - 1 \) degrees of freedom
- \( SS_{\text{regression}} \) has \( K \) degrees of freedom (\( K \) is the number of independent variables)
- \( SS_{\text{error}} \) has \( n - 1 - K \) degrees of freedom
Here is how the quantities would be laid out in an analysis of variance table. This is the generic form, with $K$ predictors.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$K$</td>
<td>$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$</td>
<td>$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{K}$</td>
<td>$\frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$</td>
</tr>
<tr>
<td>Error</td>
<td>$n - 1 - K$</td>
<td>$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$</td>
<td>$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - 1 - K}$</td>
<td>[d]</td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$\sum_{i=1}^{n}(y_i - \bar{y})^2$</td>
<td>[c]</td>
<td>[c]</td>
</tr>
</tbody>
</table>

For the case that we’ve got so far, simple regression, we just have $K = 1$. The table is in fact computable by hand, and computing formulas have been inserted.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>$\frac{(S_{xy})^2}{S_{xx}} = \hat{\beta}<em>1 S</em>{xx}$</td>
<td>$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$</td>
<td>$\frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$ [d]</td>
</tr>
<tr>
<td>Error</td>
<td>$n - 2$</td>
<td>$S_{yy} = \frac{(S_{xy})^2}{S_{xx}}$</td>
<td>$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$</td>
<td>[b]</td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ [a]</td>
<td>[c]</td>
<td>[c]</td>
</tr>
</tbody>
</table>

[a] The ratio $\frac{SS_{\text{regression}}}{SS_{\text{total}}}$ is known as $R^2$, the fraction of variation in $y$ that is explained by the regression.
The square root of the mean square error is \( s \) (or \( s_e \)), the estimate of the noise standard deviation. It can be written explicitly in simple regression as

\[
\sqrt{\frac{1}{n-2} \left( S_{xy} - \left( \frac{S_y}{S_x} \right)^2 \right)}.
\]

This box is conventionally left empty. However, if you computed a mean square here, its square root would estimate \( SD(Y) \) as

\[
\sqrt{\frac{1}{n-1} S_{yy}}.
\]

The \( F \) statistic tests the null hypothesis \( H_0: \beta_1 = 0 \).

A measure of quality of the multiple regression is the \( F \) statistic. Formally, this \( F \) statistic tests

\[
H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0
\]

versus

\[
H_1: \text{at least one of } \beta_1, \beta_2, \beta_3, \ldots, \beta_K \text{ is not zero}
\]

Note that \( \beta_0 \) is not involved in this test.

The above is for general regression. For simple regression (\( K = 1 \)), the test is \( H_0: \beta_1 = 0 \) versus \( H_1: \beta_1 \neq 0 \).

Also, note that \( s_e = \sqrt{MS_{Error}} \) is the estimate of \( \sigma_e \). This has many names:

- standard error of estimate
- standard error of regression
- estimated noise standard deviation
- root mean square error (RMS error)
- root mean square residual (RMS residual)

The measure called \( R^2 \) is computed as \( \frac{SS_{Regression}}{SS_{Total}} \). This is often described as the "fraction of the variation in \( Y \) explained by the regression."
You can show, by the way, that

$$\frac{s_e}{s_y} = \sqrt{\frac{n-1}{n-1-K}}(1-R^2)$$

The quantity $R_{adj}^2 = 1 - \frac{n-1}{n-1-K}(1-R^2)$ is called the adjusted $R$-squared. This is supposed to adjust the value of $R^2$ to account for both the sample size and the number of predictors. With a little simple arithmetic,

$$R_{adj}^2 = 1 - \left(\frac{s_e}{s_y}\right)^2$$

The analysis of variance table for simple regression might look like this:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of Freedom</th>
<th>Sum Squares</th>
<th>Mean Squares</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>965.15</td>
<td>965.15</td>
<td>4.1304</td>
</tr>
<tr>
<td>Residual</td>
<td>18</td>
<td>4,206.06</td>
<td>233.67</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>4,871.21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that

The Degrees of Freedom column sums to its total.
The Sum Squares column sums to its total.
Each Mean Square is the corresponding Sum Square divided by its Degrees of Freedom.
The $F$ statistic is the ratio of Mean Square (Regression) / Mean Square (Residual).

There are other useful facts that we can derive from this table:

$s_e = s_{Y|x} =$ standard error of regression $= \sqrt{MS_{Resid}}$.
Here this is $s_e = \sqrt{233.67} \approx 15.29$.

SD$(Y) = s_Y =$ standard deviation of the dependent variable $= \sqrt{\frac{SS_{Total}}{n-1}}$.
Here this is $s_Y = \sqrt{\frac{4,871.21}{19}} \approx \sqrt{256.3795} \approx 16.01$. 

8
\[ R^2 = \text{percent variation in } Y \text{ explained by regression} = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}} . \]

Here this is \( R^2 = \frac{965.15}{4,871.21} \approx 0.1981 = 19.81\% . \)

\[ R^2_{\text{adj}} = \text{adjusted } R^2 = 1 - \left( \frac{s_e}{s_Y} \right)^2 . \] This takes some effort, and neither \( s_e \) nor \( s_Y \) is printed in the analysis of variance table.

Here this is \( R^2_{\text{adj}} = 1 - \left( \frac{15.29}{16.01} \right)^2 \approx 0.0879 = 8.79\% . \)

Since it can happen that \( s_e > s_Y \) (in a really pathetic regression), you will sometimes see \( R^2_{\text{adj}} < 0 . \)