I. What does it all mean?

A) Notice that so far all we’ve done is math.

1) One can calculate the Least Squares Regression Line for anything, regardless of any assumptions.

2) But, if we then try to figure out if it means anything (in other words, are our slope and/or intercept significant?) we do need to worry about our assumptions.

We'll say more about the assumptions a bit later, but they're really important in regression.

B) Determining significance for slope.

1) In general, this is the one we’re interested in. The intercept, too, can be “tested”, but your book does not even present this.

2) Notice that most of our tests (at least those with normal assumptions) use the standard error of the quantity we’re testing in the denominator. We’re doing the same thing here.

3) The standard error of \( b_1 \) is given as follows:

\[
SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)}}}
\]

********

Once again the 4th edition does this differently and gives the following equation:

\[
SE_{b_1} = \frac{s_e}{s_e \sqrt{n-1}}
\]

This time, though they're almost identical, as the 4th edition uses \( s_e = s_{Y|X} \). Now note that for the denominator:

\[
s_e \sqrt{n-1} = \sqrt{SS_x} \sqrt{n-1} = \sqrt{SS_x}
\]

still, I think the first expression is probably a bit easier to work with, so we'll stick with the 2nd and 3rd edition presentation.

********

what is \( s_{Y|X} (= s_e) \) ?? Basically the standard deviation of the residuals.

a) Sum of squares for residuals:
\[ SS_{\text{resid}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

where \( \hat{y}_i \) is the predicted value for \( y_i \)

i) the predicted value for \( y_i \) is gotten from our regression equation.

ii) Note how this is similar to any sum of squares. We just use \( \hat{y} \) instead of \( \bar{y} \).

b) The residual standard deviation is then given by:

\[ s_{Y|X} = s_{\text{resid}} = s_e = \sqrt{\frac{SS_{\text{resid}}}{n - 2}} \]

(note three symbols all meaning the same thing)

i) The \( Y|X \) notation basically means “\( Y \) given \( X \)”, or in other words, we get the standard deviation of the \( Y \)'s given \( X \) (which in this case is the same as getting the residuals).

c) Before we get confused, let’s rewrite our standard error:

\[
SE_{b_1} = \sqrt{\frac{SS_r}{n - 2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}
\]

d) Yuck! Fortunately, software often does most of the work for us.

e) and we’re not done yet!!

4) So now we have the standard error of \( b_1 \). Next we figure out if \( b_1 \) means anything. For this we’ll use a regular t-test.

\[ t_s = t^* = \frac{b_1}{SE_{b_1}} \]

5) Compare to \( t_{\text{table}} \) with \( df = n - 2 \). The rest you know.

a) incidentally, one sided tests are the same as always:

- make sure you use the correct row of your t-table
- verify your data agree with \( H_1 \)
- make sure you use the correct decision rule.

6) Comments:
a) in general, we test if \( b_1 = 0 \) (our \( H_0 \)), but we can just as easily test if \( b_1 = w \), where \( w \) is anything at all.

b) this tells us if our slope is significantly different from 0.

c) our \( H_0: \beta_1 = 0 \).

d) our \( H_1: \beta_1 \neq 0 \) (or \( H_1: \beta_1 < 0 \), \( H_1: \beta_1 > 0 \))

7) Enough math & theory, here’s an example (exercise 12.6, p. 538) \( \{12.2.6, p491 \text{ and } 12.3.4, p.503 \text{ (note that due to the different presentation in the 4th edition, they give you different values; we’ll use the values given in the 2nd and 3rd edition below!)} \}:

a) Set-up: Laetisaric acid is thought to control fungus in crops. To test this, fungus was exposed to various concentrations of Laetisaric acid. Note that \( n = 12 \), though each concentration is duplicated.

b) The book provides us with the following:

\[
\begin{align*}
SS_x &= 1,303 \\
SS_y &= 677.349 \\
SS_{cp} &= -927.75 \\
SS_r &= 16.7812
\end{align*}
\]

c) You know how to get \( SS_x, SS_y, \) and \( SS_{cp} \), but let’s do a couple of steps for \( SS_r \):

\( Y = 31.8298 - 0.71201 X \)

(keep all significant digits until you report your results - particularly important here)

e) Here are a few data points (only giving 3 rows out of the original 12):

<table>
<thead>
<tr>
<th>X (acid conc.) (micro-g/ml)</th>
<th>Y (growth) (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>33.3</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>28.0</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>30</td>
<td>11.7</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

f) to get our \( \hat{Y} \)'s, we plug each \( x_i \) into our equation above.
for \( x = 0 \), we have:

\[
\hat{y}_i = 31.8298 - 0.71201 (0) = 31.8298
\]

for \( x = 6 \), we have:

\[
\hat{y}_i = 31.8298 - 0.71201 (6) = 27.55774
\]

and then for \( x = 30 \), we get a \( \hat{y} \) of 10.4695.

g) now we (finally!) get our residuals:

for \( x = 0 \), we have \( \hat{y}_i = 31.8298 \), \( y_i = 33.3 \)

and our residual is \( 33.3 - 31.8298 = 1.4702 \)

for \( x = 6 \), we have \( 28.0 - 27.55774 = 0.44226 \)

for \( x = 30 \), we have \( 11.7 - 10.4695 = 1.2305 \)

h) of course, we’d need to do this for all 12 data points.

i) to get our \( SS_r \), we’d square each of these quantities (i.e., each residual), then add them up. That would give us 16.7812 (\( = SS_r \)). Of course, it's nice that the book just gives this to us.

j) Back to our hypothesis test. As usual, we set up our hypotheses:

\[ H_0: \beta_1 = 0 \quad \text{(in words, our acid does not affect fungus growth)} \]

\[ H_1: \beta_1 < 0 \quad \text{(in words, increasing acid causes decreased fungus growth)} \]

k) the book says to use \( \alpha = .05 \), so we’ll do that.

l) Since our \( b_1 < 0 \) (see equation above), we can proceed and calculate our \( t^* \):

\[
t^* = \frac{b_1}{SE_{b_1}} = \frac{-0.712}{\sqrt{\frac{SS_r}{n-2}}} = \frac{-0.712}{\sqrt{\frac{16.7812}{10}}} = \frac{-0.712}{3.588} = -19.84
\]

m) without even looking, you should know by now that it’s significant (it's an absurdly large \( t^* \)). But let’s do our comparison:

Our tabulated \( t \), with \( d.f. = n - 2 = 10 \) and \( \alpha = .05 \) is 1.812, and our comparison \( (t^* \geq t_{\text{table}}) \) indicates that we reject.

R tells us that our \( p \)-value is about 1.2E-9, so we are highly confident in our result.
n) So we conclude that increasing acid levels inhibit fungus growth.

o) Incidentally, here’s what it all looks like, together with a fitted regression line:

![Regression Line](image)

C) Now, our assumptions:

1) Remember, we didn’t have any for our Least squares line. But once we started a test, there’s at least four:

   i) $X$ and $Y$ have a linear (straight) relationship. This is important. If they are not linear then \textit{NOTHING} else works or is valid. You need to check this before you do anything else.

      - use residual plots.

   ii) For each level of $X$, the residuals are normally distributed (see fig. 12.9, p. 517, [12.9, p. 543], [12.4.1, p. 507]).

      - to verify, do a q-q plot of the residuals.

   iii) Each residual is independent of every other residual.

      - you’ll just have to live with this one. (If you’ve heard of something called “Time series”, it deals with this).

      - sometimes a residual plot will show obvious problems.

   iv) The variance of the residuals is constant. For example, for low values of $X$, the
variance of the residuals is the same as at high values.

- use residual plots.

v) Our assumption of randomness is still there.

D) Ouch. That’s a lot of assumptions. We will not learn what to do if they fail. There are techniques (e.g. rank regression) that avoid a lot of these. Basically, if it appears that you have a serious problem with the assumptions, TALK TO A STATISTICIAN!!

E) Residual plots (or how to know when you need to talk to a statistician).

1) Easy to understand. You plot the first residual (the residual that goes with the first \( X \)), then the second, the third, etc.

In other words, put your \( X \)-values on the \( x \)-axis, and plot each residual in the appropriate place on the \( y \)-axis (this axis is now NOT the \( y \)-variable we used in our calculations).

(Your text (all editions) uses the residual on the \( y \)-axis and \( \hat{y} \) on the \( x \)-axis; there are actually several different ways to do residual plots, but we'll stick with the method presented here. It doesn't make that much difference).

2) Often done by software, but here’s a simple example using previous exercise.

a) we already figured out how to get our residuals (see above, it’s just our \((y_i - \hat{y}_i)\)).

b) if we get all the residuals for this exercise, we have:

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.47021*</td>
</tr>
<tr>
<td>0</td>
<td>-0.82979</td>
</tr>
<tr>
<td>3</td>
<td>0.10624</td>
</tr>
<tr>
<td>3</td>
<td>-1.89376</td>
</tr>
<tr>
<td>6</td>
<td>0.44227*</td>
</tr>
<tr>
<td>6</td>
<td>1.44227</td>
</tr>
<tr>
<td>10</td>
<td>0.79032</td>
</tr>
<tr>
<td>10</td>
<td>-0.90968</td>
</tr>
<tr>
<td>20</td>
<td>0.71042</td>
</tr>
<tr>
<td>20</td>
<td>-2.08958</td>
</tr>
<tr>
<td>30</td>
<td>1.23053*</td>
</tr>
<tr>
<td>30</td>
<td>-0.46947</td>
</tr>
</tbody>
</table>

*(these numbers are those for which we calculated the residuals above)

c) many computer packages will use “standardized residuals” instead (they'll convert the residuals to \( z \) scores). That's fine and will give you a very similar plot.

d) do the plot (\( x_i \) on \( x \) axis vs. residual on \( y \) axis):
3) The above plot looks fine. Now let’s illustrate a couple of example of when things are screwed up (see web page).

a) curve
b) funnel
c) curve & funnel together

4) If you see any of these, this usually means that you need to do some kind of transformation. This is an advanced topic, and you’ll need to talk to a statistician who can help you.

a) in particular, a curve means your results are **GARBAGE** until you straighten out the curve. Don’t even try to do anything with a curve.

b) a funnel? well, this means your results aren’t as good as they could be, but you’re not necessarily just producing garbage.

5) When you look at someone else’s regression analysis or when you do your own, the residual plot should always be the first thing you look at.

II. $R^2$ (not the software...).

1) $R^2$ is also known as the Coefficient of Determination. Simply, this is the square of our correlation coefficient. So you already know how to calculate this. But here’s another way:

$$R^2 = \frac{SS_{reg}}{SS_y}$$

this might look a lot easier than the formula you got for the correlation coefficient, but here’s $SS_{reg}$:

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

which is rather messy to calculate (important: $SS_r = SS_{resid}$ is not equal to $SS_{reg}$)
a) let’s dissect this a little bit. $SS_y$ is the sum of squares for $y$. In other words, this is as “messy” as $y$ can get; we take the total “spread” in $y$, regardless of what $x$ is doing.

b) $SS_{reg}$ takes the variation and adjusts it for each level of $x$. For each level of $x$, we have a “spread” of $y$-values.

c) hopefully it’s obvious that if we adjust the spread of $y$ for $x$, this is smaller than the total spread of the $y$-values, UNLESS, the adjustment of $y$ by $x$ isn’t doing anything.

If the adjustment by $x$ is irrelevant, than the numerator will be close to 0 (because $\hat{y}_i$ isn’t going up and down with $x$, it’s staying close to $\bar{y}$ (if our slope = 0, then $\hat{y}_i = \bar{y}$, and we get an $R^2$ of 0).

d) If the adjustment by $x$ is perfect, than the numerator = denominator and we get an $R^2$ of 1.

Suppose we had four points perfectly on a line. Our denominator is simply the usual $SS_y$. For each $\hat{y}_i$, this will be exactly equal to $y_i$, so our numerator and denominator are identical.

e) So what does our $R^2$ tell us? The proportion of variability that is explained by $x$.

So, for example, if $R^2 = 0.95$ would tell us that 95% of the variation in $y$ can be explained by $x$. This is good (very good).

On the other hand, if $R^2 = 0.12$ would tell us that only 12% of the variation in $y$ can be explained by $x$. This is not so good.

f) Bottom line - you know how to calculate $R^2$. Either use the formula above or just square your correlation coefficient. You should always report your $R^2$ when you do a regression.

- even if you have a highly significant regression, if your $R^2$ is low, you might need to look for a better explanation as to what’s happening to your $y$. On the other hand, if you have a good $R^2$, then you’ve probably explained a lot about $y$ (but ONLY if your regression is significant!).

IV. Concluding remarks on regression:

A) Nothing is valid if your linear model isn’t working (i.e., curves). For example, don’t even think about reporting $R^2$ if your residual plots are awful.

B) Watch out for outliers. We haven’t talked a lot about this, but outliers can have a very bad impact on your regression. Look at your fitted line plot, as well as your residual plot.

- outliers are points far away from other points.

- never (never!) just remove outliers without investigating the cause of the outlier.

C) Regression is a complicated topic. It includes such topics as:

- multiple regression/polynomial regression/multivariate regression
- in fact, most of the statistical procedures we've learned in class can be “modeled” as a regression problem.

D) Taking a regression class is well worth the time and effort. You arguably learn more about applied statistics from a regression class than any other statistics class because you learn a lot about how statistics work (you should brush up on your matrix (or linear) algebra, though).