Regression II - hypothesis testing

Now that we know how to calculate our least squares line, we need to figure out if the line means anything (in the *statistical* sense).

Remember that we are using b_1 and b_0 to estimate the slope (β_1) and the intercept (β_0).

Since both β_1 and β_0 are population parameters, we could do a hypothesis test for either of these.

In practice, we usually are only interested in β_1.

In particular, we want to test to see if $\beta_1 = 0$, because that would tell us whether or not there is a relationship between x and y.

Testing the hypothesis that $\beta_1 = 0$.

We will use a t-test to do this.

Whenever we do any kind of t-test (e.g., two sample, paired, etc.), we take whatever we're interested in (for example, $\bar{y}$ or $(\bar{y}_1 - \bar{y}_2)$) and divide this by the **S**tandard **E**rror (SE) of this quantity. For example, we do:

$$t^* = \frac{\bar{y} - \mu}{SE_{\bar{y}}} = \frac{\bar{y} - \mu}{s/\sqrt{n}} \qquad \text{or} \qquad t^* = \frac{(\bar{y}_1 - \bar{y}_2)}{SE_{\bar{y}_1 - \bar{y}_2}} = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

So in this case, we're interested in the SE for b_1.

In other words, we need to get the SE of b_1 to calculate our t^*:

$$t^* = \frac{b_1}{SE_{b_1}}$$

The SE of b_1 is give as follows:

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

Okay, so that looks complicated. Let's take it apart slowly:

The denominator is simply $\sqrt{SS_x}$; you should know this by now.

The numerator is $s_{residuals} = \sqrt{\frac{SS_{residuals}}{n-2}}$. In other words, it's the standard deviation of the residuals (although we use $n-2$ in the denominator).

(Note that some people like to use different notation here: $s_{residuals} = s_r = s_e = s_{y|x}$).

So, we can rewrite our expression for SE_{b_1} as follows:

$$SE_{b_1} = \frac{\sqrt{\frac{SS_{residuals}}{n-2}}}{\sqrt{SS_x}} = \frac{s_{residuals}}{\sqrt{SS_x}}$$

Before we go on, we need to mention that some texts once again take a slightly different approach. They define SE_{b_1} as follows:

$$SE_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}$$

So we note the following for the denominator:

$$s_x\sqrt{n-1} = \frac{\sqrt{SS_x}}{\sqrt{n-1}} \times \sqrt{n-1} = \sqrt{SS_x}$$

We will stick with the what we showed above (it's the more traditional way of doing things). Note also that for problems (i.e., homework or exams), you'll be given SS_r, not $s_r (= s_e = s_{y|x})$.

So, finally, we know how to calculate t^*, and we're ready to figure out how to do a hypothesis test:

1. Set up your hypotheses:

   $H_0 : \beta_1 = 0$
   $H_1 : \beta_1 \neq 0$

   (Of course we could do $<$ or $>$ for H_1).

2. Pick your value for α as usual.

3. Verify your assumptions (more on this soon).
   (Don't forget to make sure your data agree with H_1 if you're doing a one sided test).

4. Calculate t^* as described above.

5. Compare $|t^*|$ with t_{table} using $n-2$ degrees of freedom.

6. If $|t^*| \geq t_{\text{table}}$, we reject the null hypothesis and say the slope is significantly different from zero.

Enough theory. Let's continue with our soil respiration example. Remember we got the following results:

$$SS_x = 42 \quad SS_y = 0.1334 \quad SS_{cp} = 2.26 \quad \bar{x} = 20.5 \quad \bar{y} = 0.475$$

To this we now have to add SS_r, our Sum of Squares for the residuals. Let's see how to get this:

$$\hat{y} = -0.62809 + 0.05381x$$

(If you don't remember how we got this, review the previous set of notes). So here's our data, together with the columns we need in order to calculate SS_r (see below):
(X = temperature (Celsius), Y = NCER (µmol/m^2)

X	Y	$\hat{Y}$	r	r^2
17	0.31	0.2866667	0.023333333	5.444444e-04
18	0.27	0.3404762	-0.070476190	4.966893e-03
19	0.45	0.3942857	0.055714286	3.104082e-03
20	0.43	0.4480952	-0.018095238	3.274376e-04
21	0.51	0.5019048	0.008095238	6.553288e-05
22	0.55	0.5557143	-0.005714286	3.265306e-05
23	0.65	0.6095238	0.040476190	1.638322e-03
24	0.63	0.6633333	-0.033333333	1.111111e-03
Sum:				**0.01179048**

We know where the X and Y columns come from - that's just our original data. Let's figure out the rest of the columns.

Our column for $\hat{Y}$:

Our $\hat{Y}$ is the estimated value for Y at each value of X. For example, since $x_1 = 17$ then we have:

$$\hat{y}_1 = -0.62809 + 0.05381(17) = 0.2866667$$

Which is the value you see in the table for $\hat{Y}$ (in the first numerical row of the table).

Similarly, we could calculate the estimated value for Y for $x_3 = 3$:

$$\hat{y}_3 = -0.62809 + 0.05381(19) = 0.3942857$$

And again, this value is given in the table above.

We need to carry out this calculation eight times to get all the $\hat{Y}$ values to fill in our table.

Our column for r:

> This one's a little easier. We need the actual residual, which is simply $r = y - \hat{y}$. So for $x_1 = 17$ and $y_1 = 0.31$ we have $\hat{y}_1 = 0.2866667$ (see the table), and we get:
>
> $$r_1 = 0.31 - 0.2866667 = 0.023333333$$
>
> And again this value is given in the first numerical row in the table above.
>
> Similarly, we can do:
>
> $$r_3 = 0.45 - 0.3942857 = 0.055714286$$
>
> And so on.

Finally, our column for r^2:

> This is almost trivial, since it's literally r^2 (take r and square it). The values for the two rows we've used in the text just above are also, of course, in the table.

So where is our SS_r? This is the sum of the squared residuals. In other words, we add up the last column (as should be obvious from the table above).

So now we know how to calculate SS_r. Doing our hypothesis test is then pretty straight forward:

Set up our hypotheses:

$H_0 : \beta_1 = 0$ (in words, temperature does not affect soil respiration)
$H_1 : \beta_1 > 0$ (in words, increasing temperature increases soil respiration)

Let's stick with $\alpha = 0.05$.

Since our $b_1 > 0$ (see the slope for our equation above), we can proceed and calculate our t^*:

$$t^* = \frac{b_1}{SE_{b_1}} = \frac{0.05381}{\frac{\sqrt{\frac{SS_r}{n-2}}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = \frac{0.05381}{\frac{\sqrt{\frac{0.01179048}{8-2}}}{\sqrt{42}}} = \frac{0.05381}{.00684015} = 7.867$$

Without even looking, you should know by now that $|t^*| = 7.867$ will be significant (it's a large value for t^*).

But, to do it right, $t_{\text{table}} = t_{0.05,8} = 1.860$ (one sided), so we reject the null hypothesis.

Incidentally, R tells us that the p-value $= 0.000223$, so we're highly confident about our result.

We conclude that increasing temperature increases the level of soil respiration.

So we've done our first hypothesis test for regression. It's important to note that we really haven't checked our assumptions yet (other than verifying that our data agree with H_1).

It turns out that checking our assumptions is **very** important in regression. If you don't do this, you could be generating nothing but garbage!

We'll take a look at our assumptions in the next set of notes.