Regression, Part I

I. Difference from correlation.

A) Correlation describes the relationship between two variables, where neither is “independent” or a “predictor”.

- In correlation, it would be irrelevant if we changed the axes on our graph.

- This is NOT true for regression.

B) In regression, we often have a variable that is independent, and another that “depends” on this variable.

- For example, temperature (independent) and activity level (dependent) in cold blooded animals. Activity level obviously depends on temperature (and NOT vice-versa).

- Often we also use this “independent” variable to predict the “dependent” variable.

- We're also interested in testing for significance, though in this case we look at a line, not just the relationship between the variables.

II. Basic idea:

A) There is a possible relationship between the variables under consideration. This relationship is modeled using an equation. In the simplest case, an equation for a line.

1) This raises the question: is the line significant? This is one place where statistics come in.

2) There are actually many different ways of estimating the line, but we'll only learn the most common method.

III. Fitting the line.

A) There are several ways to do this, but the most popular (and arguably best) is least squares.:

B) Illustrate. See fig. 17., p. 331

C) The basic idea is that we add up all the residuals, and then rotate the line until the sum of squares for these distances is minimized.

1) Residual - the vertical distance from a given point to the line going through the points (at least in regression).

D) To do this, you need calculus - (if you've had calculus: you differentiate, set the result to 0, solve, and then make sure it’s actually a minimum)

1) If you do all the calculus, you wind up with our estimates for the equation of a line (see equations 17.4, 17.7 and 17.8):

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

where b_0 is the intercept and b_1 is the slope. So basically you'll wind up with the equation of a line:

$$\hat{Y} = b_0 + b_1 X$$

where $\hat{Y}$ is the value of Y predicted for the particular X.

Comment: your text uses a and α instead of b_0 and β_0 (and, equivalently, uses b and β for b_1 and β_1). For a number of reasons, the notation given here is better, particularly if you start doing multiple regression and using matrix algebra).

Also, the numerator in the equation for b_1 is often called SS_{cp} (Sum of cross products, the "SS" is to indicate it's a bit like a Sum of Squares).

2) Recap: You calculate b_0 and b_1, then put everything into the form for the equation of a line.

3) Note:

b_0 *estimates* β_0

b_1 *estimates* β_1

As usual, β_0 and β_1 are unknown. Since we don't know what they are, we estimate them with b_1 and b_0.

Side comment: another way of looking at things:

$$Y_i = b_0 + b_1 X_i + e_i$$

we often use this because this tells us what each y_i is equal to. The e_i*'s* are the "deviation" or "residual" between our equation of a line and the actual value of y. The earlier equation gives us a line, this gives us an exact relationship between x and y.

- Incidentally, minimizing the sum of the squares of the e_i*'s* will give us our least squares.

4) The next step is to figure out if β_0 and β_1 are "significant", that is, do they mean anything. As usual, we use our estimates.

a) this is similar to what we've done previously: we hypothesize some value for β_1

- although we can do the same thing for β_0, this isn't done too often (though we do occasionally get confidence intervals for β_0).

- β_1 is almost always tested for equivalence to 0, since this indicates "no slope" => so no effect of x on y.

- (Incidentally, if we do test β_0, we're hardly ever interested in $\beta_0 = 0$. But, as mentioned, we won't cover this).

IV. But before we do that, let's do an example. We'll use the exercise 17.1 in your textbook:

The problem investigates oxygen consumption in birds, as it varies with temperature. 8 temperatures (in °C) were selected, and the oxygen consumption (in ml/g/hr) was recorded:

Temp:	-18	-15	-10	-5	0	5	10	19
O_2 cons.:	5.2	4.7	4.5	3.6	3.4	3.1	2.7	1.8

To get our equation of a line, we need to calculate two quantities: SS_x and SS_{cp}.

You know how to do SS_x, so we'll just give it: 1135.5

Here's how to do SS_{cp}:

(-18 - (-1.75)) (5.2 - 3.625) = -25.59375
...
etc. ...
...
(10 - (-1.75))(4.7 - 3.625) = -14.24375

Sum = -99.65

(obviously you need to add all 8 products)

So now we have everything we need to get our b_1:

$$\frac{-99.65}{1135.5} = -.0877587$$

And now we can use this to calculate b_0:

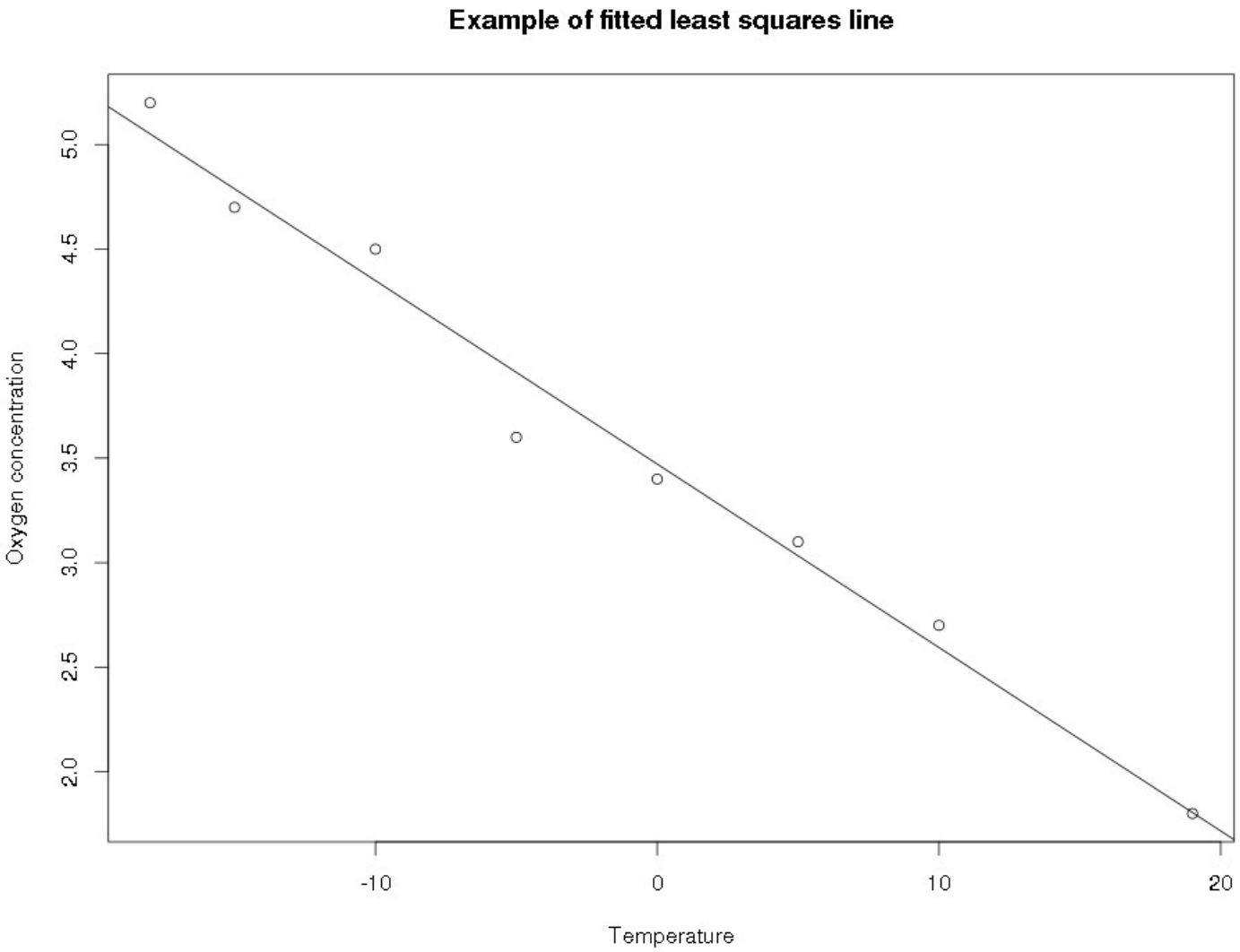
$$3.625 - (-0.0877587 \times -1.75) = 3.471422$$

So our equation for our line becomes:

$$\hat{Y} = 3.471422 - 0.0877587\, X$$

This gives us a "best fit" line through the points, using our "least squares" criterion.

And we can graph all of this (R instructions will come later):



V. Establishing significance - does the least squares line mean anything?

As it turns out, there are two ways of doing this: we can use SS's and use an F-test, or we can use a t-test.

Both will give you exactly the same answer (reject/"fail to reject").

We'll use the *t*-test approach:

It's a bit easier to understand, and directly uses the quantities that we've already calculated.

You can also do one sided tests - they're much easier here.

There's nothing wrong with the F-test, and if you really get into regression, you'll find lots of uses for it.

Your book covers both tests.

t-test for regression:

The t-test for regression is given by the following formula:

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

where

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

We know what all of these things are, except SS_r:

SS_r, or the sum of squares residuals is what we get when we take each value for y, subtract of the predicted value for that particular y, i.e., $\hat{y}$, square that, and then add the next squared difference and so on.

We'll see how this works below.

Once have t^*, we compare this to t_{table} with $df = n - 2$. The rest you know.

VI. Continuing with our example:

We want to test whether or not the slope (β_1) is significantly different from 0.

Let's set up your hypotheses:

H_0: $\beta_1 = 0$
H_1: $\beta_1 < 0$ (why? because as temperature decreases, we would expect our birds to expend more energy keeping warm, so O_2 consumption would go down with temperature)

Let's use $\alpha = 0.05$.

We need to calculate SS_r (= Sum of squares for residuals):

This is a real pain, but here goes:

For i = 1 (i.e., our first observation):

$\hat{Y}$ = 3.471422 - 0.0877587 (-18) = 5.051

For i = 2:

$\hat{Y}$ = 3.471422 - 0.0877587 (-15) = 4.788

For i = 3:

$$\hat{Y} = 3.471422 - 0.0877587\ (-10) = 4.349$$

etc. *for all 8 values*.

Now we use these numbers to get our SS_r as follows:

$$(5.2 - 5.051)^2 + (4.7 - 4.788)^2 + (4.5 - 4.349)^2 + \ldots = 0.1698$$

(yes, this is a real pain).

Finally, we can calculate t^*:

$$t^* = -\frac{0.0877587}{\frac{\sqrt{\frac{0.1698}{6}}}{\sqrt{1135.5}}} = 17.5789$$

The last step is to compare this to t_{table} with $df = n - 2$ (one sided):

$t_{0.05,6} = 1.943$, and since our t^* is bigger, we reject H_0.

VII. How to do regression in R:

Read your data in as separate columns. For example:

```
temp <- scan(nline = 1)
-18  -15  -10  -5   0    5    10   19

o2 <- scan(nline = 1)
5.2  4.7  4.5  3.6  3.4  3.1  2.7  1.8
```

(incidentally, the "`nline = 1`" statement tells "`scan`" that all the data are in one line; you probably don't need that statement).

Now you do:

```
rtemp <- lm(o2 ~ temp)
summary(rtemp)
```

(Note that R defaults to a t-test here)

And you'll get back:

```
Call:
lm(formula = o2 ~ temp)

     Residuals:
Min       1Q   Median       3Q      Max
-0.31022 -0.07552  0.03168  0.11685  0.15099
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.471422   0.060123   57.74 1.81e-09 ***
temp        -0.087759   0.004993  -17.58 2.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1682 on 6 degrees of freedom
Multiple R-squared: 0.9809,   Adjusted R-squared: 0.9778
F-statistic: 308.9 on 1 and 6 DF,  p-value: 2.177e-06
```

So what does it all mean? Look at the bits in bold above:

Note that the first "row" above is labeled "(Intercept)". It has an estimate of 3.471422. That is your intercept (= b_0).

The second row is labeled "temp" (for temperature). It has an estimate of -0.087759. This is your slope (= b_1).

The t^* value is given in this row as -17.58. R calculates the probability for us as 2.18e-06.

Notice that R also gives you an F* (last row - note the p-value is almost identical) if you really want it.

If you want to plot this, do a regular scatter plot, something like:

```
plot(temp,o2,ylab = "Oxygen concentration",xlab =
 "Temperature",main = "Example of fitted least squares line")
```

(all on one line, of course).

Now to add the least squares line, do:

```
abline(rtemp)
```

This only works after you've done "`rtemp <- lm(..etc.)`"

VIII. Checking your regression.

In regression it's very (very!) important that you check your assumptions. There are several:

i) *X* and *Y* have a linear (straight) relationship. This is important. If they are not linear then *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

- 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.

Aasumption (i) is so important that if you violate it, STOP. Don't do anything else, anything else you do is GARBAGE!!

So how do you check for this?

In R it's not too difficult. Let's use the above example and do our checks:

Residual plots:

```
plot(o2,rtemp$residuals,ylab = "Residuals",main = "Residual
                                  plot for temperature data")
abline(0,0)
```

Q-Q plots:

```
qqnorm(rtemp$residuals)
qqline(rtemp$residuals)
```

See your text on page 360 (fig. 17.12).

You should always (ALWAYS) check your assumptions for regression.

It's really (really) easy to generate garbage.

IX. Final comments:

1) NOTHING is valid if you violate the linearity assumption.

2) Do not use your regression equation to “reverse” predict. It's not straight forward.

Regression minimizes the “vertical” distances, not the horizontal distances.

But see 17.6 in your text if you're really interested.

3) If we have time or interest, we can explore some other regression related topics:

- multiple regression

- ANCOVA

- logistic regression

4) All of the parametric tests discussed so far (t-tests, ANOVA (even nested designs), etc.) can be modeled as a regression problem.