**28-5**

This set of questions is about interpreting the result of a regression analysis.

With a regression analysis, we generate a bunch of statistics.

Statistics that have to do with the equation of a line: $\hat{y}$=a+bx [[1]](#footnote-1)

slope coefficient b – says how much y changes with one unit of x

intercept a – this is the value of $\hat{y}$ when x=0

to get y, add this amount after multiplying b and x together

 Note that $\hat{y}$ is different from y. y are the values used together with x to build the regression line. They are part of the data. $\hat{y}$ are calculated using the parameters a and b and the x-values. They are the ‘predicted’ values.

**Example**

 If you eat 1 green apple you may get a stomach ache for 10 min, 2 apples for 15 min, and 3 apples gives a stomach ache for 20 min.

It is likely that the equation that predicts a stomach ache depending on the number of green apples eaten will be $\hat{y}$=a+bx. $\hat{y}$=5+5x. Note that sometimes the intercept is nonsense. Eating 0 green apples will not give give you a stomach ache for 5 minutes.

This equation could be used to predict how long your stomach ache will be if you eat 11 apples.

$\hat{y}$=5+5(11) = 5+55= 60 minutes

You might also get an R2 value, such as .7, .13, etc (R2 is between 0 and 1, since R is between -1 an 1).

In the apple example, R2 = 1, since the x values predict the y values perfectly. That usually does not happen, in fact, that is why we have regression analysis, because we have to find a line that fits data with randomness.

R2 is the is the proportion of the variability in the y-variable that is explained by the least-squares line with the x-variable. That may not seem helpful. Let ‘s try to expand on the idea. There is variability in the y values. How much of that is random and how much is due to another factor we don’t know, we just have a bunch of y values. Let’s now introduce x values that are associated with y values in some way. Let’s say some change in x is associated with a change in y, and we can build a regression line.

R2 tells how much of that original variation is due to its association with x via the regression line.

**Problem 6**

Use excel to create the scatterplot. Open the excel data file (or type in the data, it is not very much) and then use Insert->Scatter then pick the scatter without the lines. This last part is important

Equations for regression parameters a and b are on page 623

Generate $\hat{y}$=a+bx using calculated a an b

Refer to the apple example above and the equation for a line to answer the other questions

6.i is R2

**Problem 7 - Residuals**

1. e = yi-$\hat{y}$
2. use excel for the graph

**Problem 8 Influential points**

Without recalculating, use =SLOPE(), =INTERCEPT(), and =RSQ() in excel (each in their own cell – follow instructions in excel), then points can be inserted or removed from the group easily.

It is important to use

**Problem 10**

Again, use equations on page 623 for a.

B is important as it shows that we can use the regression line to *predict* values

D is R-squared again.

**Problem 11**

Use excel for the scatterplot. If you want to use it for dotplots, watch this youtube video <http://www.youtube.com/watch?v=N7HHmTpccZI>

Again, use equations from 28-3

1. This the same as a more familiar equation for a line, y=mx+b, where m=b and b=a [↑](#footnote-ref-1)