**Math 114 ACTIVITY 4:** Coefficient of Determination - how good is the "best" line?

**Why**

The correlation coefficient measures the strength (and direction) of linear relation between two variables (measured on the same subjects) and the regression line gives us the best *line* for predicting the response from the explanatory variable - but how good is the line (the "best" may not be very good)? The *coefficient of determination* gives us a handle on how well the line predicts the response and also indicates why we use "smallest sum of squares of residuals" as the criterion for the "best" line.

**LEARNING OBJECTIVES**

1. Understand the coefficient of determination $R^2$.
2. Use the coefficient of determination to judge how well the regression line fits the data. s
3. Continue developing skills in working with teams.

**CITERIA**

1. Success in working as a team and in fulfilling the team roles.
2. Success in involving all members of the team in the conversation.
3. Success in completing the exercises.

**RESOURCES**

1. The course syllabus
2. The team role desk markers (handed out in class for use during the semester)
3. Your text - especially section 4.3
4. Your calculator
5. 40 minutes

**PLAN**

1. Select roles, if you have not already done so, and decide how you will carry out steps 2 and 3 (5 minutes)
2. Work through the exercises given here - be sure everyone understands all results & procedures(30 minutes)
3. Assess the team's work and roles performances and prepare the Reflector's and Recorder's reports including team grade (5 minutes).

**DISCUSSION**

When we calculate a regression line, we are trying to predict a variable $Y$ from a variable $X$. The most important feature of this is that we relate the changes in $Y$-values to the changes in the corresponding $X$-values (using the slope of the line).

Our principal measure of the "usefulness" of the regression line is called the *coefficient of determination* [symbol $r^2$, though many people – including your text and Minitab – also use $R^2$, for reasons that you would see in a second statistics course]. This is the $r^2$ shown on your calculator (and the $R-sq$ shown by Minitab) when you calculate a regression line. Here's what it measures:

Remember that our usual measure of the variability in the $Y$ values is the "odd average of deviations" known as standard deviation $s_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$. For our purpose it's easier to look at the *total* variation, rather than the average – this is given by the *sum of squares of the deviations* $SS_y = \sum(y_i-\bar{y})^2$ [which we can calculate by taking $(n-1)s_y^2$]. This variation divides into two parts – the "explained" variation ($SS_R = \sum(\hat{y}_i-\bar{y})^2$ (change in y-values that comes from following the line) and the "error" variation ($SS_E = \sum(y_i-\hat{y}_i)^2$) – which we see is the sum of the squares of the residuals [see the diagrams on p.210 in your text]. That is:
**Total sum of squares (of deviations) = Explained sum of squares + sum of the squares of the residuals.**

[Note – this is why we use the sum of *squares of* the residuals as our measure of "goodness or badness" for the line – it corresponds to the sum of squares of the deviations].

If our points fit a line exactly, then the residuals will all be 0 and all the variation in $Y$ will be "explained" by the line. If the points don't give any line, then the "explained" sum of squares will be 0 and all the variation in $Y$ will be in the residuals.

The value of $r^2$ is the fraction of total sum of squares that is in the "explained" variation. That is:
$r^2 = \frac{\text{Explained sum of squares}}{\text{total sum of squares (of the deviations) of } Y} = \frac{SS_R}{SS_y}$

The value is usually converted to a percentage: if $r^2 = .67$ we read it as 67% and say "67% of the variation in $Y$ is explained (or predicted) by the regression with respect to $X$"

**EXAMPLE**

For our example in class - predicting gas usage based on average temperature, we have the line $\hat{y} = 1425 - 19.9x$; we found the sum of the squares of the residuals is 12966. The mean of the $y$'s is 615 and the total sum of squares (of deviations) for the $Y$ values is 375600 [Note - this is *not* $\sum y^2$ – it's the sum of the squares of the *deviations*. To get it from the calculator output we take $(n-1)s_y^2$ - rounding the result]. The "explained sum of squares" is $375,600 - 12,966 = 362,634$, so the coefficient of determination is $\frac{362634}{375600} = .966$ — about 97% of the variation in gas usage is explained by [the linear equation using] temperature variation. That means about 3% is random variation which is "unexplained" [related to other variables].

**EXERCISE**

1. Calculate $r^2$ from its definition as the coefficient of determination: For the following set of points:

| X | Y |
|---|---|
| 1 | 3 |
| 2 | 4 |
| 3 | 6 |
| 4 | 7 |

   (a) Calculate the regression line for predicting $y$ from $x$

   (b) Calculate the sum of squares (of the deviations) for $y$ [The "total sum of squares" $SS_y$] [the easiest way to do this (assuming you have entered the values in your calculator) is to get $s_y$ from the calculator and take $(n-1)s_y^2$ (round to one place). Or you could calculate $\bar{y}$ and the deviations $(y_i - \bar{y})$ and add the squares.]

   (c) Calculate the predicted value $(\hat{y}_i)$ and residual $(y_i - \hat{y}_i)$ for each data point, square the residuals and add to get the sum of the squares of the residuals. [The "sum of the squares of the errors" $SS_E$]

   (d) Calculate the mean and sum of squares (of deviations) of the predicted $(\hat{y})$ values (round to two places).[This is the "explained" sum of squares – the variation corresponding to following the line $SS_R$]

   (e) The sums from (c) and (d) should add up to the sum from (b) — the total variation in $Y$ – measured by the sum of squares – is the sum of the "explained" part and the "error" part.

   (f) Divide the "explained" sum of squares by the total sum of squares - the result should match the $r^2$ given by your calculator at step (a).

2. Interpreting the coefficient of determination for comparison of regression lines [Use the calculator to get $r^2$]:
Eight people took a two-part test (similar to SAT) intended to predict their first-year GPA. The table below shows the scores on the two parts and the first-year GPA for the eight students.

| Part A | Part B | GPA |
|---|---|---|
| 50 | 45 | 2.9 |
| 70 | 55 | 3.1 |
| 65 | 65 | 3.3 |
| 35 | 40 | 2.7 |
| 45 | 50 | 2.3 |
| 70 | 75 | 3.8 |
| 75 | 70 | 3.6 |
| 55 | 60 | 3.3 |

   (a) Make a scatter plot and find the regression line for predicting first-year GPA based on Part A of the test (keep track of $r$ and $r^2$).

   (b) Give the predicted GPA (based on Part A) and the residual for person number 2.

(c) Make a scatter plot and find the regression line for predicting first-year GPA based on Part B of the test (keep track of $r$ and $r^2$)

(d) Give the predicted GPA (based on Part B) and the residual for person number 2.

(e) Based on the $r^2$ values, which part of the test (A or B) gave better predictions for these eight students? Do your scatter plots agree?

**READING ASSIGNMENT** (in preparation for next class)
No new reading assignment for Monday - review chapters 1-4, prepare questions for Mondays class preparing for test Wedensday

**SKILL EXERCISES:**(hand in - individually - at next class meeting Remember Minitab and calculator) Sullivan p.212 #3, 7, 9, 10