Stat 135, Fall 2006
HOMEWORK 9 (due Friday 11/17)

You have two weeks to do these problems. If you decide to take a week off and just do them in the second week, you will not find me helpful. On Tuesday 11/14 I will help you with at most three problems, and on Thursday 11/16, I will help you with at most one part of one problem.

1. 12.26. **2.** 12.27. **3.** 12.28.

4. Let X and Y be two square-integrable random variables. That means $E(X^2)$ and $E(Y^2)$ are both finite. Define the *correlation* between X and Y to be

$$R(X,Y) \quad = \quad E\Big[\Big(\frac{X-E(X)}{SD(X)}\Big)\Big(\frac{Y-E(Y)}{SD(Y)}\Big)\Big]$$

Let X^* be X in standard units, so that $X^* = (X-E(X))/SD(X)$. Let Y^* be Y in standard units.

a) Show that $R(X,Y) = R(Y,X) = R(X^*,Y^*)$. This means that the correlation between two variables does not depend on the order of the variables nor on the units in which the variables were measured.

b) Show that $-1 \le R(X,Y) \le 1$.

[Hint: The random variables $(X^*+Y^*)^2$ and $(X^*-Y^*)^2$ are non-negative, therefore so are their expectations.]

In what follows, you have two observations on each of n individuals. The observations are $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$. Define $\bar{x}$ and $\bar{y}$ as usual, σ_x^2 and σ_y^2 as usual with n in the denominator, s_x^2 and s_y^2 as usual with $n-1$ in the denominator, and the correlation

$$r \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{x_i-\bar{x}}{\sigma_x}\Big)\Big(\frac{y_i-\bar{y}}{\sigma_y}\Big)$$

Notice the change in notation: everything defined in this paragraph pertains to lists of numbers, not random variables.

5. Some texts define r as follows:

$$r \quad = \quad \frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{x_i-\bar{x}}{s_x}\Big)\Big(\frac{y_i-\bar{y}}{s_y}\Big)$$

Show that the two definitions are equivalent. This form is hard to use in derivations. I recommend that you stick to our original form of the definition.

6. A more comprehensible form of the equation of the regression line for estimating y based on x. In class we derived expressions for the slope and the intercept of the regression line.

a) Show that the slope of the regression line is

$$\hat{b} \;=\; r\frac{\sigma_y}{\sigma_x} \;=\; r\frac{s_y}{s_x}$$

b) Show that the intercept of the regression line is $\hat{a} = \bar{y} - \hat{b}\bar{x}$, and that the regression line passes through the *point of averages* $(\bar{x}, \bar{y})$.

c) Show that the equation of the regression line can be written as

$$\Big(\frac{\hat{y} - \bar{y}}{\sigma_y}\Big) \;=\; r\Big(\frac{x - \bar{x}}{\sigma_x}\Big)$$

In other words, the estimate of y (in y-standard units) equals r times the given x (in x-standard units). This is the most easily remembered form of the equation.

d) The "regression effect". Consider a student who scored 1.5 SDs above average on the Stat 135 midterm. Will the regression estimate of the student's final exam score be 1.5 SDs above average, more than 1.5 SDs above average, or fewer than 1.5 SDs above average?

Now consider a student who scored 1.5 SDs below average on the midterm. Will the regression estimate of the student's final exam score be 1.5 SDs below average, more than 1.5 SDs below average, or fewer than 1.5 SDs below average?

Justify your answers.

In what follows, the result of any problem may prove very useful in solving subsequent problems.

7. The fitted values. The estimated values $\hat{y}$ are often called the *fitted* values because they are obtained by fitting the regression model to the data. Use the fact that $\hat{y}$ is a linear function of x to show that

a) $\bar{\hat{y}} \;=\; \bar{y}$. That is, the average of the fitted values equals the average of the original values.

b) $\sigma_{\hat{y}}^2 \;=\; r^2\sigma_y^2$. Check that this gives sensible answers when $r = 0$ and when $r = 1$.

8. The residuals. For each $i = 1, 2, 3, \ldots, n$, define $\hat{e}_i = \hat{y}_i - y_i$ to be the ith *residual*, that is, the error in the regression estimate of y_i.

a) Show that $\bar{\hat{e}} = 0$.

b) Show that $\sigma_{\hat{e}}^2 = (1 - r^2)\sigma_y^2$. This implies that the larger r gets, the less overall error there is in the regression. Check that the answer is sensible when $r = 0$ and when $r = 1$.

9. Decomposing the sum of squares.

a) Show that $\sigma_y^2 = \sigma_{\hat{e}}^2 + \sigma_{\hat{y}}^2$.

b) Show that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad = \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \; + \; \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

That is, $SS_{total} = SS_{error} + SS_{model}$.

10. The residuals and x.

a) Show that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \quad = \quad 0$$

b) Show that the residuals and x are uncorrelated. This explains why the "residual plot" (a plot of the residuals versus x) should show no trend upwards or downwards.