Chapter 4: Simple Linear Regression

4.1 Linear regression and prediction

Linear regression uses the relationship between distributions of scores in making predictions. If there is a relationship between two distributions, it is possible to predict a person's score in one distribution on the basis of their score in the other distribution (e.g., using a score on an aptitude test to predict actual job performance). Simple regression refers to the situation where there are only two distributions of scores, X and Y. By convention, X is the predictor variable, and Y the criterion (or predicted) variable. Multiple regression, which we will address later, refers to situations with a single criterion (Y), but more than one predictor (X_1, X_2, etc).

4.2 Definitions

a) A **scatterplot** is a graph of paired X and Y values

b) A **linear relationship** is one in which the relationship between X and Y can best be represented by a straight line.

c) A **curvilinear relationship** is one in which the relationship between X and Y can best be represented by a curved line.

d) A **perfect relationship** exists when all of the points in the scatter plot fall exactly on the line (or curve). An **imperfect relationship** is one in which there is a relationship, but not all points fall on the line (or curve).

e) A **positive relationship** exists when Y increases as X increases (i.e., when the slope is positive).

f) A **negative relationship** exists when Y decreases as X increases (i.e., when the slope is negative).

4.3 Equation for a straight line

The equation for a straight line is usually written as:

$$Y = bX + a \tag{4.1}$$

where b = slope of the line
$= (Y_2 - Y_1) / (X_2 - X_1)$
= "the rise" divided by "the run"

and a = the Y-intercept
= the value of Y when $X = 0$

Perhaps an example will help to clarify what this means. Imagine that you have decided to start working out at a gym. The annual membership fee is £25, and in addition to that, you must pay £2 every time you go to the gym.[1] If we let X = the number of times you go to the gym, and Y = the total cost, we would find that:

$$Y = 2X + 25 \tag{4.2}$$

The Y-intercept is 25. That is, if you never go to the gym ($X = 0$), your total cost is £25. And the slope of the line (b) is 2: Every time you go to the gym, it costs you another £2. Putting this another way, every time there is an increase of 1 on the X-axis, there is an increase of 2 on the Y-axis.

4.4 The best fitting regression line: Least squares criterion

When the relationship between X and Y is perfect, one can use the method shown above for calculating the slope (b), and then calculate the Y-intercept (a) by substituting known values for X and Y and solving for a. The method shown above does not work, however, when the relationship between X and Y is imperfect. Therefore, we need to calculate the slope (b) and Y-intercept (a) by some other means. The most frequently used method is the **least squares** method. And the formula for the least squares regression line is:

$$Y' = b_Y X + a_Y \tag{4.3}$$

Note that some textbooks use $\hat{Y}$ (pronounced "Y-hat") in place of Y'. And some books use b_0 in place of a_Y. Y', (pronounced "Y-prime") is the predicted value of Y.

4.4.1 How to calculate the slope

There will be a few formulae in this section, but many of them should already be familiar. Please bear with me. The most common version of the formula for the slope constant in least squares regression is:

$$b_Y = \frac{SP}{SS_X} \tag{4.4}$$

The "Y" subscript on b_Y indicates that this is the slope for the *regression of Y on X*—that is for the equation that allows prediction of Y-values from X-values.

In case anyone has forgotten, *SS* is shorthand for *the sum of the squared deviations about the mean*. The "X" subscript indicates that it is the "sum of squares" for the X-scores that we need. The **conceptual formula** for SS_X is:

[1] This method of charging for gym memberships is common in the UK. I was living there when I wrote the first version of these notes—hence the reference to Pounds Sterling rather than dollars.

$$SS_X = \sum (X - \overline{X})^2 = \sum (X - \overline{X})(X - \overline{X}) \quad (4.5)$$

SP stands for "sum of products". Its conceptual formula is shown in equation 4.6. The **product** for each (*X,Y*) pair is obtained by multiplying the $(X - \overline{X})$ difference by the $(Y - \overline{Y})$ difference. *SP* = the **sum** of these **products**.

$$SP = \sum (X - \overline{X})(Y - \overline{Y}) \quad (4.6)$$

4.4.2 How to calculate the Y intercept (a_Y)

It is a fact that **the least squares regression line passes through the point** $(\overline{X}, \overline{Y})$**.** Therefore, we can substitute these two means into the equation for a straight line, and solve for *a*:

$$\overline{Y} = b_Y \overline{X} + a_Y \quad (4.7)$$

Subtracting $b_Y \overline{X}$ from both sides, we get:

$$a_Y = \overline{Y} - b_Y \overline{X} \quad (4.8)$$

4.5 Why this method is called the *least squares* method

This method of finding the best fitting regression line is called the least squares method because it *minimizes the sum of the squared errors in prediction*. In other words, if we let *Y'* = the predicted value of *Y*, then

$$\sum (Y - Y')^2 = \text{ a minimum} \quad (4.9)$$

The quantity (*Y* - *Y'*) is the amount of error in prediction. It is the difference between the actual Y score and the predicted Y score. We square this term before summing, because **the sum of the errors in prediction equals zero** (just like the sum of the deviations about the mean equals zero).

Note that because the least squares regression line minimizes the sum of the squared errors, it gives greater accuracy in prediction than any other possible regression line.

4.5.1. Proof that $b_Y = SP/SS_X$

The following is from David W. Stockburger's online textbook, *Introductory Statistics: Concepts, Models, and Applications* (http://www.psychstat.smsu.edu/introbook/sbk16.htm).

The problem is presented to the mathematician as follows: "The values of a and b in the linear model $Y'_i = a + b X_i$ are to be found which minimize the algebraic expression

$$\sum_{i=1}^{N} (Y_i - Y_i')^2$$

The mathematician begins as follows:

$\sum(Y_i - Y'_i)^2$ is the expression to be minimized

$\sum(Y_i - (a + bX_i))^2$ substituting $a + bX$ for Y'

$\sum(Y_i - a - bX_i)^2$ deleting the innermost parentheses

$\sum(Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX)$ squaring the expression

$\sum Y^2 + \sum a^2 + \sum b^2X^2 - 2\sum aY - 2\sum bXY + 2\sum abX$ taking the summation sign inside

Now comes the hard part that requires knowledge of calculus. At this point even the mathematically sophisticated student will be asked to "believe." What the mathematician does is take the first-order partial derivative of the last form of the preceding expression with respect to b, set it equal to zero, and solve for the value of b. This is the method that mathematicians use to solve for minimum and maximum values. Completing this task, the result becomes:

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

Notice the following about Stockburger's expression for the regression coefficient *b*:

$$N\sum XY - \sum X \sum Y = \sum XY - \frac{\sum X \sum Y}{N} = \text{ a computational formula for } SP$$

$$N\sum X^2 - \left(\sum X\right)^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{N} = \text{ a computation formula for } SS_X$$

Therefore, as noted earlier in this chapter:

$$b = \frac{SP}{SS_X} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

4.6 The regression of *X* on *Y*

What we have been talking about so far concerns the prediction of Y given a value of X. The regression line we constructed when attempting to predict Y from X is called the **regression of Y on X**. It may sometimes be the case, however, that we wish to predict X given Y. In order to do so, we must construct a new (different) regression line that minimizes the sum of the squared errors in predicting X given Y. That is to say that errors in prediction are measured **horizontally** (left to right) rather than **vertically** in this situation. This new regression line is called the **regression of X on Y**.

Note that it is always the regression of the **predicted variable** on the **known variable**. Thus, if X is known and you are trying to predict Y, you have the regression of Y on X. If Y is known and you are trying to predict X, you have the regression of X on Y. If height is known and you are trying to predict weight, then you have the regression of weight on height.

The linear equation for the regression of X on Y is given by:

$$X' = b_X Y + a_X \tag{4.10}$$

where X' = predicted value of X
b_X = slope
a_X = X intercept

The slope (b_X) can be computed with either of the following formulae:

$$b_X = \frac{SP}{SS_Y} \tag{4.11}$$

And because the regression line passes through the point $(\overline{X}, \overline{Y})$, the X-intercept can be computed as follows:

$$a_X = \overline{X} - b_X \overline{Y} \tag{4.12}$$

Note that all we have done is swap around the X's and Y's in the formulae shown previously.

4.7 The standard error of estimate

When you are looking at a single distribution of scores, the **variance** and **standard deviation** are measures that indicate the degree to which scores are spread out around the mean. In the case of bivariate distributions and linear regression, it would be useful to have similar measures that indicate the amount of spread around the regression line. There are in fact two such measures: The **variance error of estimate**, and the **standard error of estimate**. The latter is

the square root of the former. The standard error of estimate is to the least squares regression line what the standard deviation is to the mean of a distribution. (I will say more about this later.)

4.7.1 Conceptual Formula

The conceptual formula for the standard deviation of a distribution of *Y*-scores is given by:

$$s_Y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{n-1}} \tag{4.13}$$

The conceptual formula for the *standard error of estimate when estimating Y given X* is very similar. See if you can spot the two differences in equation (4.14):

$$s_{Y|X} = \sqrt{\frac{\sum (Y - Y')^2}{n-2}} \tag{4.14}$$

The first difference from equation (4.13) is that the deviations in the numerator are around the least squares regression line (i.e., around the predicted values of Y) rather than around the mean. And the second difference is that the denominator is *n-2* rather than *n-1*. In other words, whereas *df* = *n-1* for the variance and standard deviation of a single distribution of scores, *df* = *n-2* for the standard error of estimate for a bivariate distribution. (More generally, as we will see later, *df* = *n-p-1*, where *p* = the number of predictor variables.)

4.7.3 The standard error of estimate when predicting X given Y

If you wish to predict X from Y, as we saw in section 4.6, you must use a least squares regression line that is different from the one used to predict Y from X. If follows from this that the standard error of estimate when predicting X given Y will also be different from the standard error when prediction Y given X. The conceptual and computational formulae are given below:

$$s_{X|Y} = \sqrt{\frac{\sum (X - X')^2}{n-2}} \qquad \{ \text{conceptual formula} \} \tag{4.15}$$

4.8 Miscellaneous points about linear regression

Linear regression is used to predict a Y score from a score on X. Bear in mind the following:

1) The relationship between X and Y must be **linear**. If the relationship is not linear, prediction will not be very accurate.

2) Normally, we are not interested in predicting Y scores that are already known. We derive our regression equation with sample data that consists of paired X and Y scores, but use the equation to predict Y scores when only X values are given. Because we use data collected from a sample to make these predictions, it is vital to have a **representative** sample when deriving a regression equation.

3) A regression equation is properly used only for the range of the variables on which it was based. We do not know whether the relationship between X and Y continues to be linear beyond the range of sample values.

4) Prediction is most accurate if the data have the property of **homoscedasticity**—i.e., if the variability of the Y scores is constant at all points along the regression line.

5) When X and Y are both normally distributed and the number of paired scores is large, the data in a bivariate frequency distribution often produce a so-called **bivariate normal** distribution. When you have such a distribution, the **standard error of estimate** can be used in the same way we used the standard deviation of a normal distribution. That is, we could say that *about 68% of the scores in the scatterplot fall within 1 standard error of the regression line; and about 95% of the scores fall within 2 standard errors of the regression line*.

4.9 Linear regression and Pearson *r*

Like linear regression, correlation is concerned with the relationship between two (or more) variables. Regression is typically concerned with using the relationship for prediction. Correlation, on the other hand, is concerned with finding out whether there is a relationship, and if there is, determining its strength and direction.

If two variables are correlated, then one variable **may** be the cause of the other. If two variables are not correlated, there is no causal relationship between the two--at least not a **linear** causal relationship.

A **correlation coefficient** is a number that expresses the magnitude and direction of the relationship between two (or more) variables. The **Pearson product moment correlation coefficient (Pearson *r*)** is a correlation coefficient that ranges from - 1.00 to 1.00 .

a) if $r = 1.00$ -- perfect positive linear relationship

b) if r = -1.00 -- perfect negative linear relationship

c) if r = 0.00 -- no linear relationship

4.9.1 Pearson r: Conceptual Formula

Conceptually, Pearson r is **the slope of the least-squares regression line when linear regression is carried out on z-scores** rather than on raw scores. You will recall that a z-score indicates the position of a score relative to the mean of the distribution, in standard deviation units. In other words, a z-score of 2.5 indicates that the raw score is 2.5 standard deviations greater than the mean.

Let us suppose that we have two sets of scores, X and Y. If the z-score corresponding to X_1 is equal to the z-score corresponding to Y_1, then we know that X_1 and Y_1 occupy the same relative positions within their respective distributions. In other words, **Pearson r is a measure of the extent to which paired scores occupy the same or opposite relative positions within their respective distributions**.

4.9.2 Pearson r from Raw Scores

In your reading, you may come across various computational formulae for Pearson r. Two of the most common computational formulae are shown below. Unlike most computational formulae, the second of these does actually provide some insight into the meaning of Pearson r. It indicates that Pearson r is to a scatterplot as a z-score is to a univariate distribution of scores. That is, Pearson r is a measure of the covariance of the X and Y scores, but is measured in standard score units rather than in raw score units.

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} \tag{4.16}$$

4.10 The coefficient of determination

If we square Pearson r, we obtain

r^2 = proportion of the variability of Y accounted for by the linear relationship between X and Y

Note that "the proportion of the variability of Y accounted for by the linear relationship between X and Y" is usually abbreviated to "**the variability of Y accounted for by X**". Because r^2 equals the proportion of the variability of Y accounted for by X, it is called the **coefficient of determination**.

In order to understand what is meant by "proportion of variability of Y accounted for by X", let us consider a concrete example where X is a score on a test of spelling competence, and Y is a score on a test of writing ability. The data for N=6 subjects are shown below.

Case Summaries[a]

Subject		Spelling score (X)	Writing score (Y)
1		20	20
2		30	50
3		45	35
4		60	60
5		78	45
6 (Mary)		88	**90**
Total	N	6	6
	Mean	53.50	50.00
	Std. Deviation	26.76	23.87
	Variance	715.900	570.000

a. Limited to first 100 cases.

Let us imagine for just a moment that we do not know Mary's Y-score (writing ability), and wish to predict it from her X-score. If there were no linear relationship between X and Y, then the best prediction we could make would be the mean of the Y scores, or 50 in this case. However, because there is a linear relationship between X and Y, we can do better than that by using the regression line to make our prediction. As can be seen in Figure 4.1, Mary's **predicted** Y-score is about 75.

Mary's **actual** Y-score is 90. Therefore, the deviation of her actual Y-score from the mean of Y is 90 - 50 = 40.

$$(Y_i - \overline{Y}) = 90 - 50 = 40 \qquad (4.17)$$

Note that if you were to take the same kind of $(Y_i - \overline{Y})$ deviation score for each of the six people, square them, and add them up, you would have the sum of the squared deviations about the mean, or SS_Y.

Returning to Mary's case, note that her deviation (from the mean of Y) score can be broken down, or partitioned into two components. The first component is the **deviation of her predicted score from the mean of Y**. As noted earlier, Mary's predicted score is 75, so:

$$(Y' - \overline{Y}) = 75 - 50 = 25 \qquad (4.18)$$

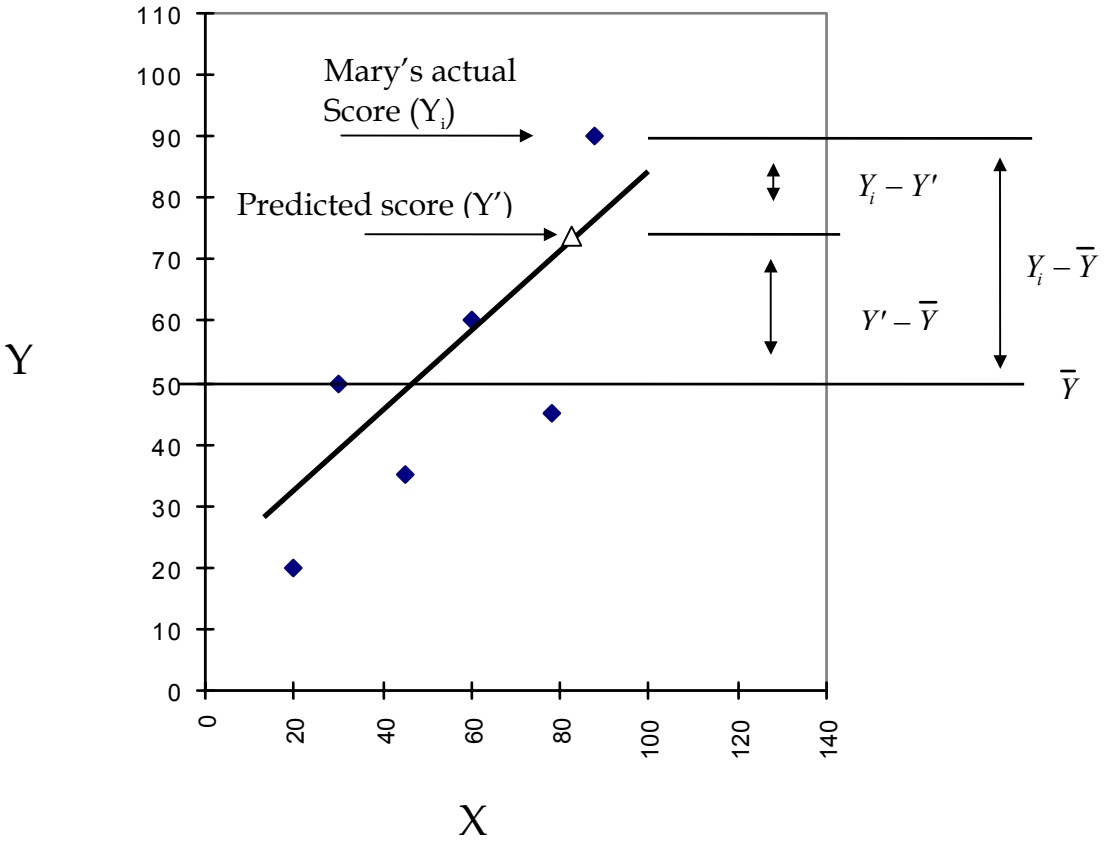


Figure 4.1 Relationship between spelling competence (X) and writing ability (Y).[2]

The second component of the deviation of Mary's actual score from the mean of Y is the **deviation of her actual score from her predicted score**:

$$(Y_i - Y') = 90 - 75 = 15 \tag{4.19}$$

The important thing to note is that these components are additive. That is, when you look at an individual case, the following will be true:

$$(Y_i - \overline{Y}) = (Y' - \overline{Y}) + (Yi - Y') \tag{4.20}$$

Finally, if you were compute these components for each of the X-Y pairs and **square them**, you would find that the following is also true:

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(Y' - \overline{Y}\right)^2 + \sum\left(Y_i - Y'\right)^2 \tag{4.21}$$

[2] Figure is based on Figure 7.6 from Pagano (1990).

You should recognize the left-hand portion of equation (4.21) as SS_Y. The first term on the right-hand side is that portion of SS_Y that can be accounted for (or explained) by the linear relationship between X and Y. It is sometimes symbolized as $SS_{Y'}$ or $SS_{Y|X}$, or $SS_{regression}$. And the final term is that portion of SS_Y that cannot be explained by the linear relationship between X and Y. It is also know as SS_{error}, $SS_{Y-Y'}$, or $SS_{residual}$.

And so, equation (4.21) can also be expressed as follows:

$$SS_Y = SS_{Y'} + SS_{error} \tag{4.22}$$

Note, however, that many textbook authors (including Norman & Streiner) refer to $SS_{Y'}$ as $SS_{regression}$. Also, SS_{error} is often referred to as $SS_{residual}$ (to reflect that fact that it is the portion of SS_Y that is left over after we've explained as much as we can by virtue of the linear relationship between X and Y). So equation (4.21) may also be expressed as:

$$SS_Y = SS_{regression} + SS_{residual} \tag{4.23}$$

Conclusion

$$r^2 = \text{ proportion of variability of Y accounted for by X}$$

$$= \frac{\text{variability of Y accounted for by X}}{\text{total variability of Y}} \tag{4.24}$$

$$= \frac{SS_{regression}}{SS_Y} = 1 - \frac{SS_{residual}}{SS_Y}$$

4.11 Significance testing in linear regression

We have just seen that SS_Y can be partitioned into $SS_{regression}$ and $SS_{residual}$. This is very similar to the partitioning of SS_{Total} in one-way ANOVA. (In that case, one partitions SS_{Total} into $SS_{Between-groups}$ and $SS_{Within-groups}$.) In linear and multiple regression, there are p degrees of freedom associated with $SS_{regression}$, where p = the number of predictor variables. So in the case of *simple* linear regression (i.e., only one predictor variable), $df_{regression} = 1$.

It follows from this that $MS_{regression} = SS_{regression}$, because $MS_{regression} = SS_{regression} / df_{regression}$.

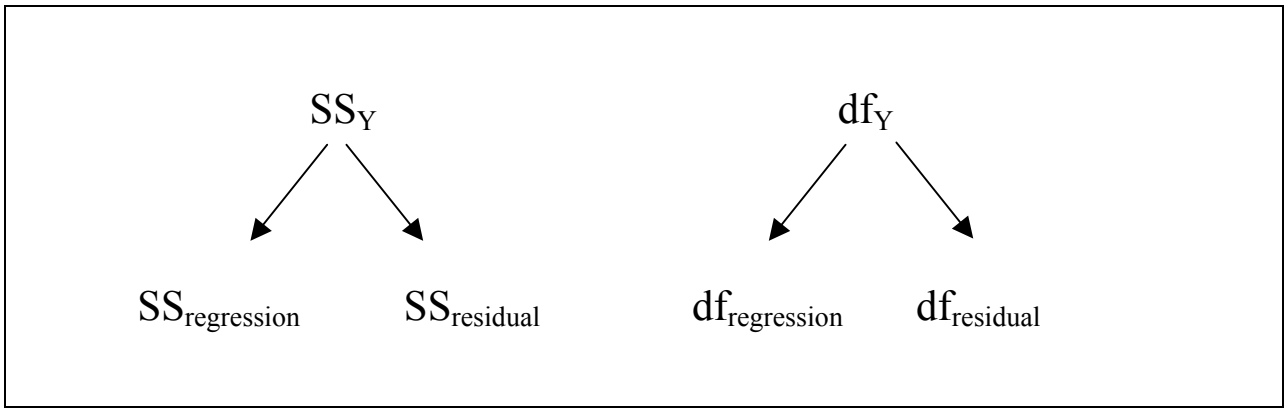


Figure 4.2: Partitioning diagram for ANOVA of simple linear regression.

There *are N-p-1* degrees of freedom associated with $SS_{residual}$. (N = the total number of subjects, p = the number of predictor variables.) Thus,

$$MS_{residual} = \frac{SS_{residual}}{(N - p - 1)} \tag{4.25}$$

And finally,

$$F_{(p,N-p-1)} = \frac{MS_{regression}}{MS_{residual}} \tag{4.26}$$

This F-value can be used to test the **significance of r^2** (or R^2 in multiple regression). The null hypothesis for this test states that **none** of the variability of Y is accounted for by a linear relationship between X and Y. Note that as usual, we are using values we have sampled to make an inference about a population. So a more accurate statement of the null hypothesis would be that *in the population from which we have sampled our X and Y scores, there is no linear relationship between X and Y.*

4.12 The standard error of estimate revisited

Earlier, I gave conceptual and computational formulae for **standard error of estimate** in linear regression--see equations (4.14) and **Error! Reference source not found.**. Many textbooks also offer the following formula:

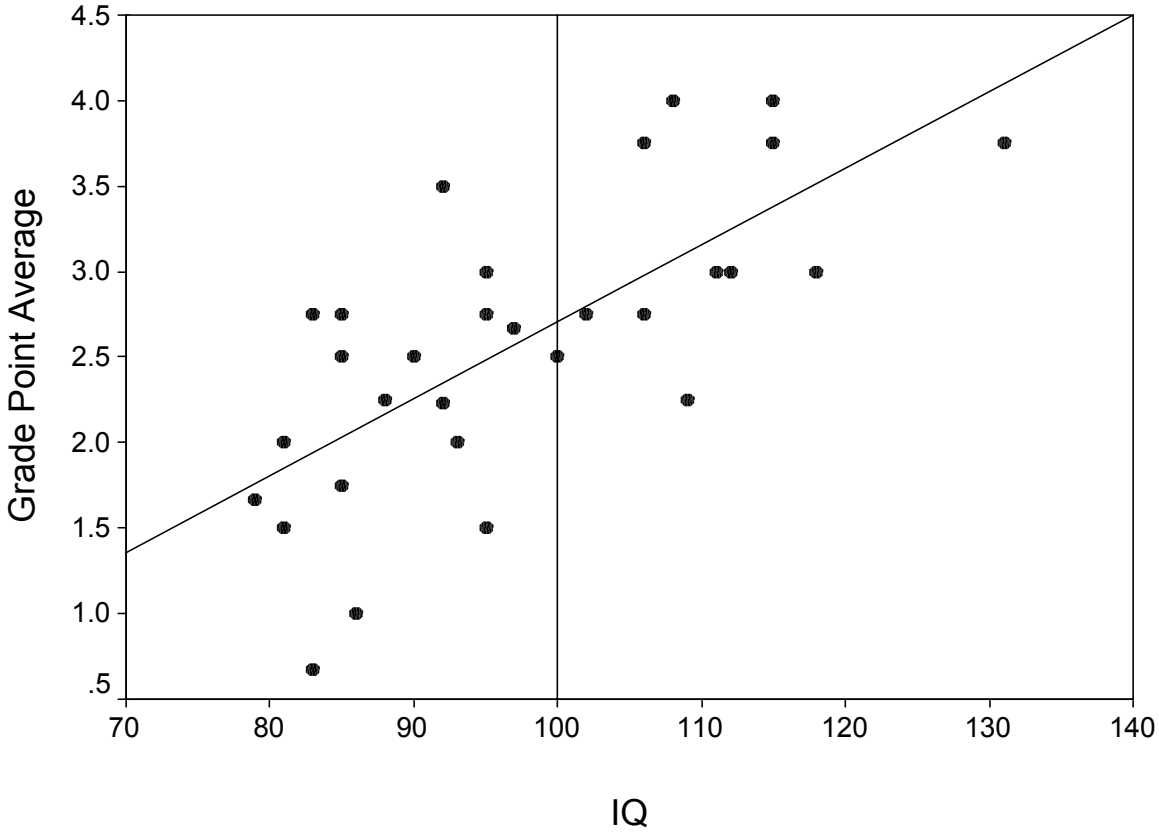
$$s_{Y|X} = s_Y \sqrt{\frac{(1 - r^2)(n - 1)}{(n - 2)}} \tag{4.27}$$

Other texts give this formula:

$$s_{Y|X} = s_Y \sqrt{(1 - r^2)} \tag{4.28}$$

Equation (4.28) can be used in place of equation (4.27) **when the sample size is large**. When that is the case, the ratio of (n-1) / (n-2) will be virtually equal to 1.

Figure 4.3: GPA as a Function of IQ



4.14 Pearson r and restriction of the range of X

Restriction of the range of X-values results in attenuation of the correlation between X and Y (i.e., the absolute value of *r* will be closer to 0). This can be illustrated with a simple example. Figure 4.3 shows the relationship between IQ and grade point average (GPA) for a sample of 30 students. The Pearson r for these data is .702, so roughly 49% of the variance of the GPA scores is accounted for by the linear relationship between IQ and GPA (i.e., $r^2 = .492$).

But now suppose that you were to divide these 30 people into two groups on the basis of IQ. One group has IQ < 100, and the other has IQ greater than or equal to 100. I did this, and computed the correlations for the two groups separately. Here's what I obtained, plus the results for the whole sample:

Group	Pearson r	r^2	p
IQ < 100	.441	.194	.067
IQ = 100 or more	.461	.213	.131
All students	.702	.492	< .001

Clearly, the linear relationship is much weaker when I restrict the range on X. The proportion of variance accounted for by the linear relationship drops from almost 50% with the whole sample to somewhere around 20% when I restrict the range of IQ values.

Why is it important to know that restriction of the range of X attenuates the correlation between X and Y? One reason is that this phenomenon occurs pretty routinely in a lot of settings. For example, you might expect to see a reasonably strong, positive correlation between high school grades and university grades. However, only those students with good high school grades typically get into university. Therefore, you are probably have restricted range on X, and the correlation may be quite a bit weaker than you anticipated.

4.15 A final word of caution

Consider the following example: You compute the correlation between amount of study time (X) and performance on an exam (Y). Your intuition probably tells you that there should be a positive relationship between these variables. Furthermore, you probably suspect that increasing study time **causes** performance on the exam to improve.

For the sake of argument, let us assume that this causal relationship does exist (i.e., increasing study time does in fact lead to improved performance on the exam). Despite the existence of this causal relationship, it is entirely possible that your calculated value of *r* could indicate **no correlation**, or even **a negative correlation**.

"HOW IS THIS POSSIBLE?", you ask, your arms gesticulating wildly. Note that amount of study time is not the only variable that determines how well someone does on an exam. Many other factors affect one's performance, not the least of which is intelligence (whatever that is).

The students in your sample could be divided into groups based on some measure of intelligence (e.g., Bright, Average, Poor). Let us suppose that the brightest students spent the least time studying, and that the poorest students spent the most time studying. That is, the amount of study time is negatively correlated with intelligence. If this were so, and if mean performance was the same for all groups, you would probably calculate $r = 0$ approximately.

Now imagine the same situation, but with the mean test scores ordered Bright > Average > Poor (which could well happen). Now you may actually come up with a negative value or *r*, which would appear to suggest that the less time spent studying, the better one will do on the exam!

In order to get a clear picture of what's going on here, you would have to take into account not only the amount of study time, but all other relevant variables as well. You could "take them into account" by either *controlling them* (i.e., holding them constant, or equating them across groups); or by *including them as a predictor variables*, and using **multiple regression/correlation** procedures, which are discussed in another chapter.

4.16 Linear regression and correlation using SPSS

To access the linear regression dialog box in SPSS, click on **Analyze→Regression→Linear**. Then move your independent (or predictor) and dependent variables into the appropriate boxes. Various options are available by clicking on the **Statistics**, **Plots**, **Save**, and **Options** buttons.

To generate a correlation matrix, click on **Analyze→Correlate→Bivariate**.

To generate a scatterplot, click on **Graphs→Scatter**. After you have created your scatterplot, you may wish to add a regression line. To do so, double-click on the scatterplot in the SPSS Output window to open the Chart Editor. Then click on **Chart→Options**, and put a check in the "Fit line" box. Several lines (or curves) are available, but the least squares regression line is the default.

TIP:	If you don't understand something in an SPSS dialog box, **right-click** it with the mouse to produce a brief explanation.

An SPSS syntax file demonstrating simple linear regression and correlation can be downloaded from my homepage here:

http://www.angelfire.com/wv/bwhomedir/spss/linreg.txt

Review Questions

1. The most commonly used regression line is the *least squares* regression line. Why it is called the *least squares* regression line?

2. Why is it necessary to use one regression line for prediction Y from X, and another line for prediction of X from Y?

3. Do the regression of Y on X and the regression of X on Y ever produce the same regression line? If so, under what circumstances?

4. What is *homoscedasticity*?

5. What is a bivariate normal distribution?

6. Assume that X and Y are both normally distributed, and that you have carried out the regression of Y on X. If the number of X-Y pairs is quite large, approximately what percentage of the points in the scatter plot fall within (±):

   a) one standard error of the regression line?
   b) two standard errors of the regression line?
   c) 2.5 standard errors of the regression line?

7. What is covariance? How does it relate to Pearson *r*?

8. What is another name for the covariance of a variable *with itself*?

9. Why is r^2 called the coefficient of determination?

10. Draw partitioning diagrams (i.e., partitioning of *SS*) for:

    a) the one-way ANOVA (if this has already been taught)
    b) simple linear regression

11. In linear regression, how many degrees of freedom are associated with:

    a) $SS_{regression}$?
    b) $SS_{residual}$?

12. How can one test the significance of r^2? What is the null hypothesis for this test?

13. How is r^2 related to the standard error of estimate?

Appendix 1: Mean and Individual Prediction Intervals

DISCLAIMER: The material in this appendix is quite advanced, and is for your information only. It will **not** be covered on the final exam.

Two kinds of *prediction intervals* are available in SPSS scatterplots, **mean** and **individual** prediction intervals. These are specific and extreme cases of a family of possible prediction intervals which are generated using a standard error computed as follows:

$$s_{pred} = SE(Y_{new}) = \sqrt{MS_{error} \cdot \left(\frac{1}{m} + \frac{1}{n} + \frac{(X_{new} - \overline{X})^2}{SS_X} \right)}$$

where m = the number of new cases
n = the number of X-Y pairs used to create the regression model

If you have the X-value for one new case, you would substitute m = 1 into the equation, and obtain the standard error that is used in computing the **individual prediction interval**.

Substituting m = ∞ (∞ is the symbol for infinity), on the other hand, you would obtain the standard error used in calculating the **mean prediction interval**.

Finally, note that the curvature of the prediction intervals (i.e., they get wider as you get further away from the mean of X) is due to the $(X_{new} - \overline{X})^2$ in the numerator of the formula.

Appendix 2: Herman Rubin on Linear Regression and Normality

The following is from a post to usenet newsgroup sci.stat.edu:

From: hrubin@odds.stat.purdue.edu (Herman Rubin)
Subject: Assumptions for regression Date: 1999/05/29
Newsgroups: sci.stat.edu

In article <7imc1rrqh1@usenet01.srv.cis.pitt.edu>,
Richard F Ulrich <wpilib+@pitt.edu> wrote:
>G. Anthony Reina (reina@nsi.edu) wrote:
>: I have a regression that seems to fit the data well. The residuals don't
>: appear to have any trends so I think what's left is just random (normal
>: distribution) error.

>: Is there a way that I can quantitatively test the residuals to see if
>: they are truly in a normal disribution?

> - I have read 5 responses and no one has mentioned "independence."
>Dependency is more subtle, and (maybe) more damning to a simple model.

>Residuals should not be correlated with X or with the sequence of
>sampling. Et cetera.

It cannot be overemphasized that **normality is NOT necessary for the validity of regression.** What has just been said is what is most important, that the disturbances (actual deviations from the "true" regression expression) must be uncorrelated with the "explanatory" variables.

The precise probabilities of various tests do depend on normality, some more than others. But regression has rather good robustness, by which I mean that the properties of the procedure do not depend much on those assumptions which one does not wish (or need) to make.

--
This address is for information only. I do not claim that these views are those of the Statistics Department or of Purdue University.
Herman Rubin, Dept. of Statistics, Purdue Univ., West Lafayette IN47907-1399
hrubin@stat.purdue.edu Phone: (765)494-6054 FAX: (765)494-0558

The "precise probabilities of various tests" Professor Rubin refers to are *p*-values.

Appendix 3: Variance, covariance, and standard scores

You may recall that the **variance** of a set of scores is equal to the sum of squared deviations about the mean divided by the degrees of freedom:

$$s^2 = \frac{\sum (X - \overline{X})^2}{n-1} = \frac{SS_X}{df} \tag{4.29}$$

If you have 2 sets of scores (*X* and *Y*), it is possible to calculate the **covariance** of the two sets of scores as follows:

$$COV_{XY} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n-1} = \frac{SP}{df} \tag{4.30}$$

The covariance provides information similar to that given by Pearson *r*, the (linear) correlation coefficient. And in fact, one version of the formula for calculating *r* is as follows:

$$r_{XY} = \frac{COV_{XY}}{s_X s_Y} \tag{4.31}$$

From this formula, it should be clear that the difference between covariance and correlation is one of scaling: The covariance is in raw score (i.e., original scale) units, whereas the Pearson correlation is in standardized (z-score) units.

Just as an interesting side issue, note that the covariance of a set of scores with itself is equal to the variance of those same scores. (Hey, don't laugh! You never know when a titbit like that will come in handy--e.g., when making small talk at a party.)

The following equation was presented earlier in this chapter:

$$b_X = \frac{SP}{SS_Y} \tag{4.32}$$

Note that if you took the right-hand portion of equation 4.35, and divided both numerator and denominator by *n-1*, the equation for the slope could be rewritten as:

$$b_Y = \frac{COV_{XY}}{s_X^2} \tag{4.33}$$

Appendix 4: Pearson r and regression coefficients

The relationship between the slope constants of linear regression and Pearson r may be useful in some situations:

$$b_Y = r_{XY}\left(\frac{s_Y}{s_X}\right) \tag{4.34}$$

$$b_X = r_{XY}\left(\frac{s_X}{s_Y}\right) \tag{4.35}$$

Also note that the product of the regression coefficients from the regression of Y-on-X and the regression of X-on-Y equals r^2:

$$b_Y b_X = r\left(\frac{s_Y}{s_X}\right) r\left(\frac{s_X}{s_Y}\right) = r^2\left(\frac{s_Y s_X}{s_X s_Y}\right) = r^2 \tag{4.36}$$