Regression: Model Fit Measures

1. Coefficient of Multiple Correlation R and Coefficient of Determination R^2

As previously noted one measure of model fit—how well the regression model is able to reproduce the observed scores on the dependent variable Y—is the simple Pearson's correlation between observed Y and predicted Y'.

R = Pearson's correlation, r, between Y and Y'

The closer R is to 1.00 the better the regression model is able to reproduce Y, the closer R is to 0.00, the worse the performance of the model in reproducing Y. While R may be negative, this is not expected or likely; one anticipates R to be positive since the regression model is designed to predict Y as well as is possible given the data.

The coefficient of determination, R^2, is simply R squared:

$R^2 = R \times R$ = proportion of variance in Y predicted (or explained) by regression model

The coefficient of determination may be interpreted as the proportional reduction in error resulting from use of the regression model to predict Y. Another interpretation of the coefficient of determination is explained variance—the proportion of variance in Y explained, or predicted, by the regression model. The complement of this, $1-R^2$, is the amount of variance in Y that is not explained or predicted by the regression model.

$1-R^2$ = proportion of variance in Y not explained by regression model

Recall the student ratings data:

Table 1: Student Ratings and Course Grades Data

Course	Quarter	Year	Student Ratings (mean ratings for course)	Percent A's
EDR852	FALL	1994	3.00	46.00
EDR761	FALL	1994	4.40	47.00
EDR761	FALL	1993	4.40	53.00
EDR751	SUMM	1994	4.50	62.00
EDR751	SUMM	1994	4.90	64.00
EDR761	SPRI	1994	4.40	50.00
EDR751	SPRI	1994	3.70	33.00
EDR751	WINT	1994	3.30	25.00
EDR751	WINT	1994	4.40	53.00
EDR751	FALL	1993	4.80	50.00
EDR751	SUMM	1993	4.80	54.00
EDR751	SUMM	1993	3.80	60.00
EDR751	SPRI	1993	4.60	54.00
EDR761	SPRI	1993	4.10	37.00
EDR751	WINT	1993	4.20	53.00
EDR751	FALL	1992	3.50	41.00
EDR751	FALL	1992	3.80	47.00

SPSS Data File: http://www.bwgriffin.com/gsu/courses/edur8132/notes/student_ratings.sav

1. What is the coefficient of multiple correlation value for the student ratings data; that is, what is the correlation between observed ratings (Y) and predicted ratings (Y')?
2. What is the coefficient of determination value?

2. Residuals and Model Fit: SEE and MSE

Recall that a residual, or error, is the difference between observed Y and predicted Y':

e = Y - Y'

One way to measure model fit is to examine variation in residuals.

From basic statistics note that variance in raw data may be calculated for the population as

$$\sigma^2 = \frac{\sum(Y-\bar{Y})^2}{N}$$

and variance for sample data may be calculated as

$$s^2 = \frac{\sum(Y-\bar{Y})^2}{n-1}$$

The difference between these formula is the degrees of freedom. In the population case the count of all observations is use, N, but in the sample formula degrees of freedom is n − 1 is used (to provide an unbiased estimate of σ^2).

The variance for residuals may also be calculated in the same manner taking into account regression model degrees of freedom:

$$\hat{\sigma}^2 = \frac{\sum(Y-Y')^2}{n-k-1} = \text{MSE}$$

The above produces a variance that as many names:

variance error of residuals, or
variance error of estimate, or
mean squared error (MSE)

and is denoted as $\hat{\sigma}^2$ or *MSE*.

The square root of MSE, $\sqrt{MSE}$, is conceptually the standard deviation of residuals, but since these data are residuals, or errors, $\sqrt{MSE}$ is known as the *standard error of residuals* or *standard error of estimate* and is symbolized as

$$\hat{\sigma} = \sqrt{MSE} = \text{SEE (standard error of estimate)}$$

Note that as SEE, and MSE, become smaller, the fit of the model is better since the residuals are smaller.

3. Adjusted R^2—Incorporating MSE into Standardized Model Fit

Both MSE and SEE are scale dependent—the larger the raw scores, the larger will be MSE and SEE. As a result, use of MSE and SEE as measures of model fit make difficult model fit comparisons across different measures of Y.

One way to incorporate MSE in a standardized solution for model fit is to examine the proportional reduction in error from Y to Y'. Consider the following:

(a) s^2 = variance of Y before prediction explains variation in Y, and

(b) $\hat{\sigma}^2 = \dfrac{\sum (Y - Y')^2}{n - k - 1}$ = MSE = variance in residuals of Y after prediction,

so MSE is the amount of variance in Y that remains after regression. MSE is the amount of unexplained or unpredicted variance in Y; the amount of variation in Y that the regression model does not predict.

The ratio of MSE to s^2 can be used to produce a measure of proportional reduction in error:

Adjusted R^2 = $\dfrac{\sigma_Y^2 - MSE}{\sigma_Y^2} = 1 - \dfrac{MSE}{\sigma_Y^2}$

1. Calculate and show s^2 for Y and MSE for e for the student ratings data
2. Show calculation of adj. R^2 using the above formula