Fox Module 10 Advanced multiple regression

REGRESSION ANALYSIS SUM OF SQUARES AND R^2 *PRACTICE PROBLEMS*

(The attached PDF file has better formatting.)

Know the three types of sums of squares: total, residual, and regression.

- Ordinary least squares estimators minimize the sums of squared residuals.
- The estimator for σ is a sum of squares adjusted for degrees of freedom.
- The R^2 is a ratio of two sums of squares.
- The adjusted (corrected) R^2 adjusts this ratio for degrees of freedom.
- The *F*-statistic is a similar ratio, also adjusted for degrees of freedom.

Most regression concepts are based on sums of squares. Standardized coefficients and generalized linear models adjust the sum of squares for the conditional distributions of the explanatory and response variables. GLMs use maximum likelihood estimation, which is similar to (not identical to) minimizing a normalized sum of squares.

Final exam problems are of two types.

- Quantitative problems compute the various sums of squares, R^2, adjusted R^2, analysis of variance, *F*-statistic, and similar items.
- Qualitative problems ask how these items change with units of measurement, number of observations, and displacement.

** Exercise 10.1: R^2

Ten pairs of observations (X_i, Y_i) are fit to the model $Y_i = \alpha + \beta \times X_i + \epsilon_i$, where ε_i are independent, normally distributed random variables with mean 0 and variance σ^2.

$\sum x_i = 50$
$\sum x^2_i = 1{,}050$
$\sum y_i = 60$
$\sum y^2_i = 3{,}560$
$\sum x_i\, y_i = 1{,}260$

A. What is TSS, the total sum of squares?
B. What is RegSS, the regression sum of squares?
C. What is R^2, the coefficient of determination?
D. What is the correlation of X and Y?
E. What is RSS, the residual sum of squares?
F. What is the (omnibus) F-value for this regression?

Part A: $TSS = \sum(y_i - \overline{y})^2 = 3{,}560 - 60^2 / 10 = 3{,}200$

Part B: $RegSS = [\, \sum(x_i - \overline{x})(y_i - \overline{y})\,]^2 / \sum(x_i - \overline{x})^2 = 960^2 / 800 = 1{,}152$

$\sum(x_i - \overline{x})^2 = 1{,}050 - 50^2 / 10 = 800$
$\sum(y_i - \overline{y})^2 = 3{,}560 - 60^2 / 10 = 3{,}200$
$\sum(x_i - \overline{x})(y_i - \overline{y}) = 1{,}260 - 50 \times 60 / 10 = 960$

Jacob: What is the rationale for this formula?

Rachel: The regression sum of squares RegSS =

$\sum(\, \hat{y}_i - \overline{y})^2 = \sum [\, (\alpha + \beta x_i) - (\alpha + \beta \overline{x})\,]^2 = \beta^2 \sum\, (x_i - \overline{x})^2$

$\beta = \sum(x_i - \overline{x})(y_i - \overline{y}) / \sum(x_i - \overline{x})^2$, so $RegSS = [\, \sum(x_i - \overline{x})(y_i - \overline{y})\,]^2 / \sum(x_i - \overline{x})^2$

Part C: $R^2 = 1{,}152 / 3{,}200 = 0.360 = 36\%$

Part D: The correlation of X and Y is

$(\sum x_i\, y_i - N \sum x_i \sum y_i) / [\, (\sum x^2_i - N \sum x_i) \times (\sum y^2_i - N \sum y_i)\,]^{0.5}$

$= (1{,}260 - 50 \times 60 / 10) / [\, (1{,}050 - 50^2 / 10) \times (3{,}560 - 60^2 / 10)\,]^{½} = 0.600$

Using deviations from the means, the correlation is

$\sum(x_i - \overline{x})(y_i - \overline{y}) / (\sum(x_i - \overline{x})^2 \times \sum(y_i - \overline{y})^2\,)^{0.5} = = 960 / (800 \times 3{,}200)^{½} = 0.600$

Note: R^2 is the square of the correlation $= 0.600^2 = 0.360$

Part E: The residual sum of squares RSS = TSS – RegSS = 3,200 – 1,152 = 2,048

Part F: The omnibus *F*-value is (RegSS / *k*) / (RSS / N-*k*-1) = 1,152 / (2,048 / (10 - 1 - 1) = 4.500