Analysis of Variance (ANOVA) for Simple Linear Regression

The variability in Y values can be partitioned into two pieces.

$\sum_{i=1}^n (Y_i - \bar{Y})^2$	$=$	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$+$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
Total Sum of Squares	$=$	Regression Sum of Squares	$+$	Error (or Residual) Sum of Squares
SSTO	$=$	SSREG	$+$	SSE

We can organize the results of a simple linear regression analysis in an ANOVA table.

Source	D.F.	Sum of Squares	Mean Square	F	P-value	
Regression	df_{REG}	$SSREG$	$MSREG$	$\frac{MSREG}{MSE}$	$P(T^2 \geq \frac{MSREG}{MSE})$	$T^2 \sim F(df_{REG}, df_E)$
Error	df_E	SSE	MSE			
Total	df_{TO}	$SSTO$				

Source	D.F.	Sum of Squares	Mean Square	F	P-value	
Regression	1	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{1}$	$\frac{MSREG}{MSE}$	$P(T^2 \geq \frac{MSREG}{MSE})$	$T^2 \sim F(1, n-2)$
Error	$n-2$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$			
Total	$n-1$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$				

The F-statistic $\frac{MSREG}{MSE}$ is used to test

$$H_0 : \mu\{Y|X\} = \beta_0 \quad \text{versus} \quad H_A : \mu\{Y|X\} = \beta_0 + \beta_1 X \text{ for some } \beta_1 \neq 0.$$

or $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ for short. The test is equivalent to the t-test that we learned about previously because

(1) $F = \dfrac{MSREG}{MSE} = \dfrac{\hat{\beta}_1^2}{\left[\text{SE}(\hat{\beta}_1)\right]^2} = t^2$ and (2) $T^2 \sim F$ with 1 and $n-2$ d.f. $\iff T \sim t$ with $n-2$ d.f.

The F-statistic $\frac{MSREG}{MSE}$ is a special case of the F-statistic used to compare full and reduced models.

$$F = \frac{[\text{RSS(red.)} - \text{RSS(full)}]/[\text{df}_{\text{RSS(red.)}} - \text{df}_{\text{RSS(full)}}]}{\text{RSS(full)}/\text{df}_{\text{RSS(full)}}}$$

Recall that our null and alternative hypotheses are

$$H_0 : \mu\{Y|X\} = \beta_0 \quad \text{versus} \quad H_A : \mu\{Y|X\} = \beta_0 + \beta_1 X \text{ for some } \beta_1 \neq 0.$$

The full model corresponds to the situation where β_1 can be any value. The reduced model forces β_1 to be 0, just like H_0. Write down formulas for RSS(red.), RSS(full), $\text{df}_{\text{RSS(red.)}}$, and $\text{df}_{\text{RSS(full)}}$ for the special case of simple linear regression; and show that the resulting reduced vs. full model F-statistic is the same as $F = \frac{MSREG}{MSE}$.

Because $SSTO = SSREG + SSE$, we may write $1 = \frac{SSREG}{SSTO} + \frac{SSE}{SSTO}$.

$\frac{SSE}{SSTO}$ is the proportion of total variation in the Y values that was not explained by the regression of Y on X.

The remaining proportion of variation in the Y values is $1 - \frac{SSE}{SSTO} = \frac{SSREG}{SSTO}$. This quantity – known as the *coefficient of determination* – is the proportion of the variation in the Y values that was explained by the regression of Y on X.

It can be shown that the coefficient of determination is equal to the square of the sample linear correlation coefficient between X and Y.

$$1 - \frac{SSE}{SSTO} = \frac{SSREG}{SSTO} = r^2$$