ANOVA

Dr. Frank Wood

ANOVA

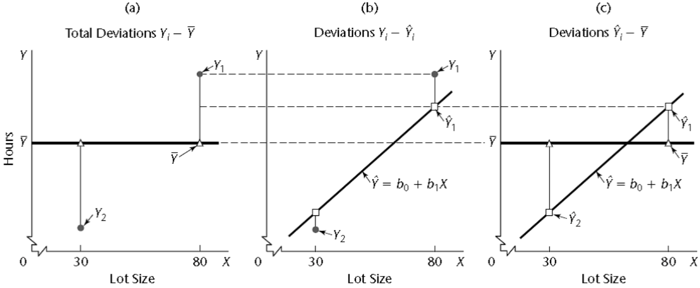
1. ANOVA is nothing new but is instead a way of organizing the parts of linear regression so as to make easy inference recipes.
2. Will return to ANOVA when discussing multiple regression and other types of linear statistical models.

Partitioning Total Sum of Squares

1. "The ANOVA approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y"
2. We start with the observed deviations of Y_i around the observed mean

$$Y_i - \bar{Y}$$

Partitioning of Total Deviations



Measure of Total Variation

1. The measure of total variation is denoted by

$$SSTO = \sum (Y_i - \bar{Y})^2$$

2. SSTO stands for total sum of squares
3. If all $Y_i's$ are the same, SSTO $= 0$
4. The greater the variation of the $Y_i's$ the greater SSTO

Variation after predictor effect

1. The measure of variation of the $Y_i's$ that is still present when the predictor variable X is taken into account is the sum of the squared deviations

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

2. SSE denotes error sum of squares

Regression Sum of Squares

1. The difference between SSTO and SSE is SSR

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2$$

2. SSR stands for regression sum of squares

Partitioning of Sum of Squares

$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Deviation of fitted regression value around mean}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Deviation around fitted}}$$

Remarkable Property

1. The sums of the same deviations squared has the same property!

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2$$

   or $SSTO = SSR + SSE$

2. Proof:

Remarkable Property

Proof: $(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2$

$$
\begin{aligned}
(Y_i - \bar{Y})^2 &= \sum[(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2 \\
&= \sum[(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)] \\
&= \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 + 2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)
\end{aligned}
$$

but

$$\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \sum \hat{Y}_i(Y_i - \hat{Y}_i) - \sum \bar{Y}(Y_i - \hat{Y}_i) = 0$$

By properties previously demonstrated

Remember: Lecture 3

1. The i^{th} residual is defined to be

$$e_i = Y_i - \hat{Y}_i$$

2. The sum of the residuals is zero:

$$\begin{aligned}\sum_i e_i &= \sum (Y_i - b_0 - b_1 X_i) \\ &= \sum Y_i - nb_0 - b_1 \sum X_i \\ &= 0\end{aligned}$$

By first normal equation.

Remember: Lecture 3

The sum of the weighted residuals is zero when the residual in the i^{th} trial is weighted by the fitted value of the response variable for the i^{th} trial

$$\begin{aligned}
\sum_i \hat{Y}_i e_i &= \sum_i (b_0 + b_1 X_i) e_i \\
&= b_0 \sum_i e_i + b_1 \sum_i e_i X_i \\
&= 0
\end{aligned}$$

By previous properties.

Breakdown of Degrees of Freedom

1. SSTO
   1.1 1 linear constraint due to the calculation and inclusion of the mean
      1.1.1 n-1 degrees of freedom
2. SSE
   2.1 2 linear constraints arising from the estimation of β_1 and β_0
      2.1.1 n-2 degrees of freedom
3. SSR
   3.1 Two degrees of freedom in the regression parameters, one is lost due to linear constraint
      3.1.1 1 degree of freedom

Mean Squares

A sum of squares divided by its associated degrees of freedom is called a mean square

The regression mean square is

$$MSR = \frac{SSR}{1} = SSR$$

The error mean square is

$$MSE = \frac{SSE}{n-2}$$

ANOVA table for simple lin. regression

Source of Variation	SS	df	MS	$\mathbb{E}(MS)$
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = SSR/1$	$\sigma^2 + \beta_1^2 \sum(X_i -$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = SSE/(n-2)$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		