ANOVA

Dr. Frank Wood

ANOVA

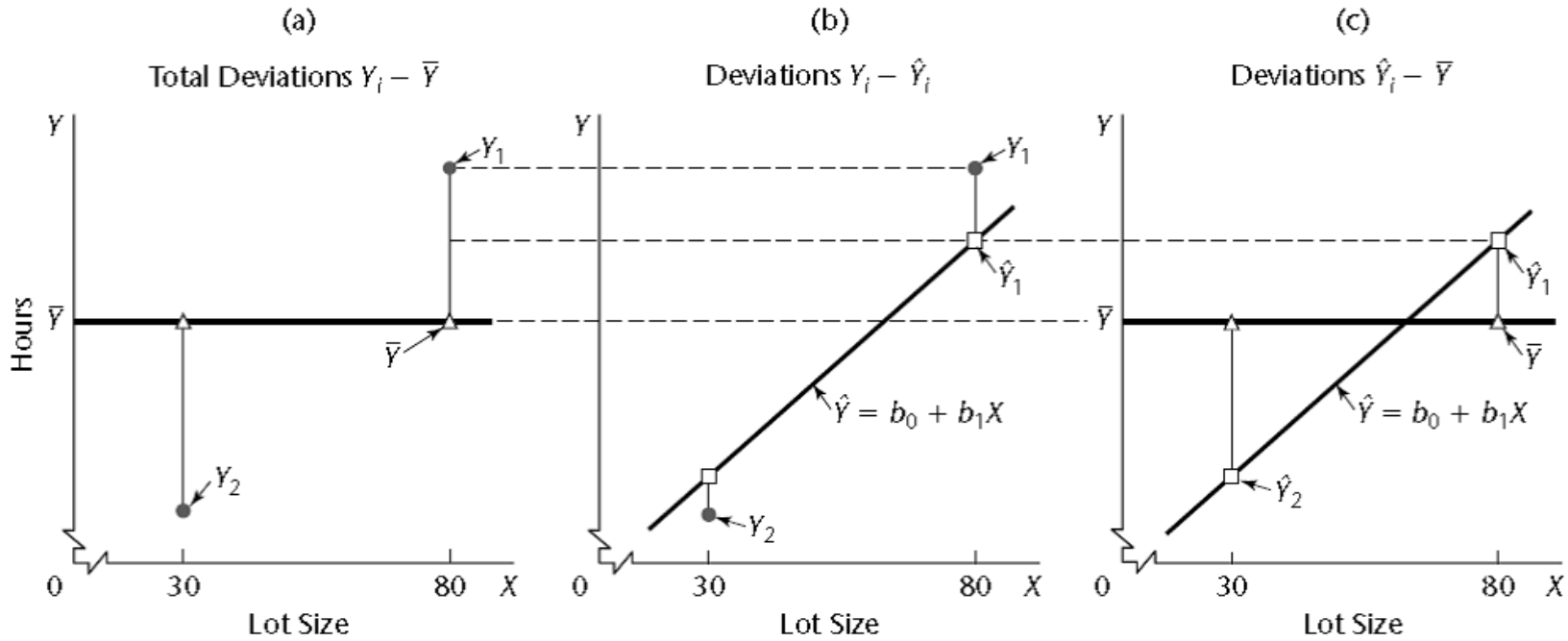
- ANOVA is nothing new but is instead a way of organizing the parts of linear regression so as to make easy inference recipes.
- Will return to ANOVA when discussing multiple regression and other types of linear statistical models.

Partitioning Total Sum of Squares

- "The ANOVA approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y"
- 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

- The measure of total variation is denoted by

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

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

Variation after predictor effect

- 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

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

- SSE denotes error sum of squares

Regression Sum of Squares

- The difference between SSTO and SSE is SSR

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

- SSR stands for regression sum of squares

Partitioning of Sum of Squares

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

Remarkable Property

- The sums of the same deviations squared has the same property!

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

or $SSTO = SSR + SSE$

Remarkable Property

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

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

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. Namely

$$\sum \hat{Y}_i e_i = 0$$

and

$$\sum e_i = 0$$

Remember: Lecture 3

- The i^{th} residual is defined to be

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

- The sum of the residuals is zero:

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

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. The left is given by $\sum e_i = 0$, the right can be expanded to yield the second normal equation.

Breakdown of Degrees of Freedom

- SSTO
  - 1 linear constraint due to the calculation and inclusion of the mean
    - n-1 degrees of freedom
- SSE
  - 2 linear constraints arising from the estimation of β_1 and β_0
    - n-2 degrees of freedom
- SSR
  - Two degrees of freedom in the regression parameters, one is lost due to linear constraint
    - 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 linear regression

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

$E\{MSE\} = \sigma^2$

- Remember the following theorem, presented in an earlier lecture.

  For the normal error regression model, $\frac{SSE}{\sigma^2}$ is distributed as χ^2 with $n-2$ degrees of freedom and is independent of both b_0 and b_1.

  Rewritten this yields

$$SSE/\sigma^2 \sim \chi^2(n-2)$$

- That means that $E\{SSE/\sigma^2\} = n-2$
- And thus that $E\{SSE/(n-2)\} = E\{MSE\} = \sigma^2$

$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

- To begin, we take an alternative but equivalent form for SSR

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

- And note that, by definition of variance we can write

$$\sigma^2\{b_1\} = E\{b_1^2\} - (E\{b_1\})^2$$

$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

- But we know that b_1 is an unbiased estimator of β_1 so $E\{b_1\} = \beta_1$
- We also know (from previous lectures) that

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

- So we can rearrange terms and plug in

$$\begin{aligned} \sigma^2\{b_1\} &= E\{b_1^2\} - (E\{b_1\})^2 \\ E\{b_1^2\} &= \frac{\sigma^2}{\sum (X_i - \bar{X})^2} + \beta_1^2 \end{aligned}$$

$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

- From the previous slide

$$E\{b_1^2\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} + \beta_1^2$$

- Which brings us to our desired result

$$E\{MSR\} = E\{SSR/1\} = E\{b_1^2\} \sum (X_i - \bar{X})^2 = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

Comments and Intuition

- The mean of the sampling distribution of MSE is σ^2 regardless of whether X and Y are linearly related (i.e. whether $\beta_1 = 0$)
- The mean of the sampling distribution of MSR is also σ^2 when $\beta_1 = 0$.
  - When $\beta_1 = 0$ the sampling distributions of MSR and MSE tend to be the same

This intuition leads us to a battery of simple rules for constructing linear regression tests

F Test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$

ANOVA provides a battery of useful tests. For example, ANOVA provides an easy test for

Two-sided test

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

Two-sided t-test statistic from before

$$t^* = \frac{b_1 - 0}{s\{b_1\}}$$

ANOVA test statistic

$$F^* = \frac{MSR}{MSE}$$

The ANOVA framework makes many common linear regression tests into almost "shorthand."

Sampling distribution of F^*

- The sampling distribution of F^* when $H_0 : \beta_1 = 0$ holds can be derived starting from Cochran's theorem

Cochran's theorem

If all n observations Y_i come from the same normal distribution with mean μ and variance σ^2, and SSTO is decomposed into k sums of squares SS_r, each with degrees of freedom df_r, then the SS_r/σ^2 terms are independent χ^2 variables with df_r degrees of freedom if

$$\sum_{r=1}^{k} df_r = n - 1$$

Does this directly apply to the regression case?

The F Test

We have decomposed SSTO into two sums of squares SSR and SSE and their degrees of freedom are additive, hence, by Cochran's theorem: If $\beta_1 = 0$ so that all Y_i have the same mean $\mu = \beta_0$ and the same variance σ^2, SSE/σ^2 and SSR/σ^2 are independent χ^2 variables

F^* Test Statistic

- F^* can be written as follows

$$F^* = \frac{MSR}{MSE} = \frac{\frac{SSR/\sigma^2}{1}}{\frac{SSE/\sigma^2}{n-2}}$$

- But by Cochran' s theorem, we have when H_0 holds

$$F^* \sim \frac{\frac{\chi^2(1)}{1}}{\frac{\chi^2(n-2)}{n-2}}$$

F Distribution

- The F distribution is the ratio of two independent χ^2 random variables.
- The test statistic F^* follows the distribution $F^* \sim F(1, n-2)$

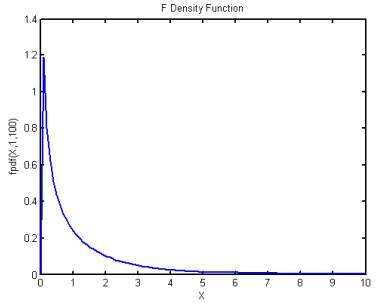
Hypothesis Test Decision Rule

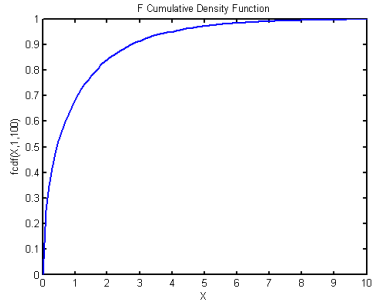
Since F^* is distributed as $F(1, n-2)$ when H_0 holds, the decision rule to follow when the risk of a Type I error is to be controlled at α is:

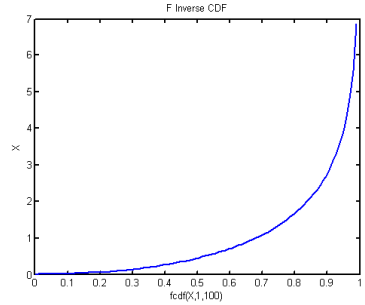
If $F^* \leq F(1-\alpha; 1, n-2)$, conclude H_0
If $F^* > F(1-\alpha; 1, n-2)$, conclude H_a

F distribution

- PDF, CDF, Inverse CDF of F distribution



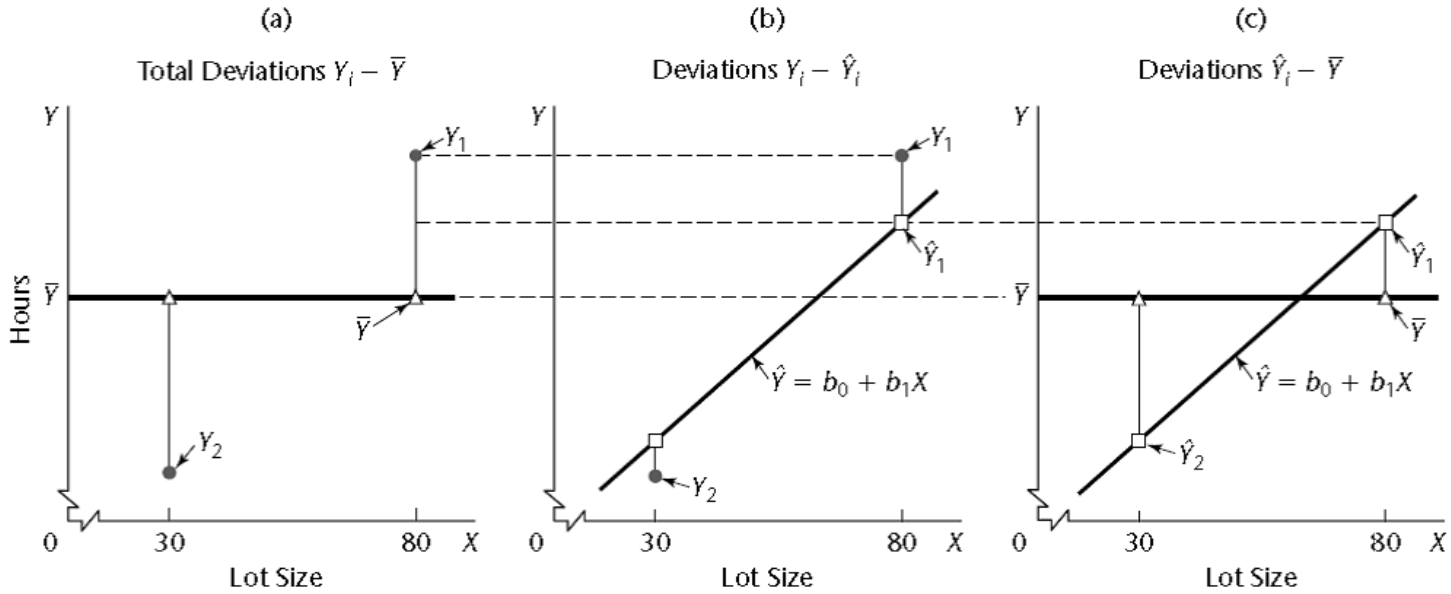




- Note, MSR/MSE must be big in order to reject hypothesis.

Partitioning of Total Deviations

Does this make sense? When is MSR/MSE big?



General Linear Test

- The test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is but a single example of a general test for a linear statistical models.
- The general linear test has three parts
  - Full Model
  - Reduced Model
  - Test Statistic

Full Model Fit

- A full linear model is first fit to the data

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Using this model the error sum of squares is obtained, here for example the simple linear model with non-zero slope is the "full" model

$$SSE(F) = \sum [Y_i - (b_0 + b_1 X_i)]^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE$$

Fit Reduced Model

- One can test the hypothesis that a simpler model is a "better" model via a general linear test (which is really a likelihood ratio test in disguise). For instance, consider a "reduced" model in which the slope is zero (i.e. no relationship between input and output).

  $H_0 : \beta_1 = 0$
  $H_a : \beta_1 \neq 0$

- The model when H_0 holds is called the reduced or restricted model.

$$Y_i = \beta_0 + \epsilon_i$$

- The SSE for the reduced model is obtained

$$SSE(R) = \sum (Y_i - b_0)^2 = \sum (Y_i - \bar{Y})^2 = SSTO$$

Test Statistic

- The idea is to compare the two error sums of squares SSE(F) and SSE(R).
- Because the full model F has more parameters than the reduced model R $SSE(F) \leq SSE(R)$ always
- In the general linear test, the test statistic is

$$F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}}$$

  which follows the F distribution when H_0 holds.
- df_R and df_F are those associated with the reduced and full model error sums of square respectively

R^2

- SSTO measures the variation in the observations Y_i when X is not considered
- SSE measures the variation in the Y_i after a predictor variable X is employed
- A natural measure of the effect of X in reducing variation in Y is to express the reduction in variation ($SSTO - SSE = SSR$) as a proportion of the total variation

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- Note that since $0 \leq SSE \leq SSTO$ then $0 \leq R^2 \leq 1$

Limitations of and misunderstandings about R^2

1. Claim: high R^2 indicates that useful predictions can be made. The prediction interval for a particular input of interest may still be wide even if R^2 is high.
2. Claim: high R^2 means that there is a good linear fit between predictor and output. It can be the case that an approximate (bad) linear fit to a truly curvilinear relationship might result in a high R^2.
3. Claim: low R^2 means that there is no relationship between input and output. Also not true since there can be clear and strong relationships between input and output that are not well explained by a linear functional relationship.