Lecture 17 – Relationships in Regression Leading to the *F-Ratio*

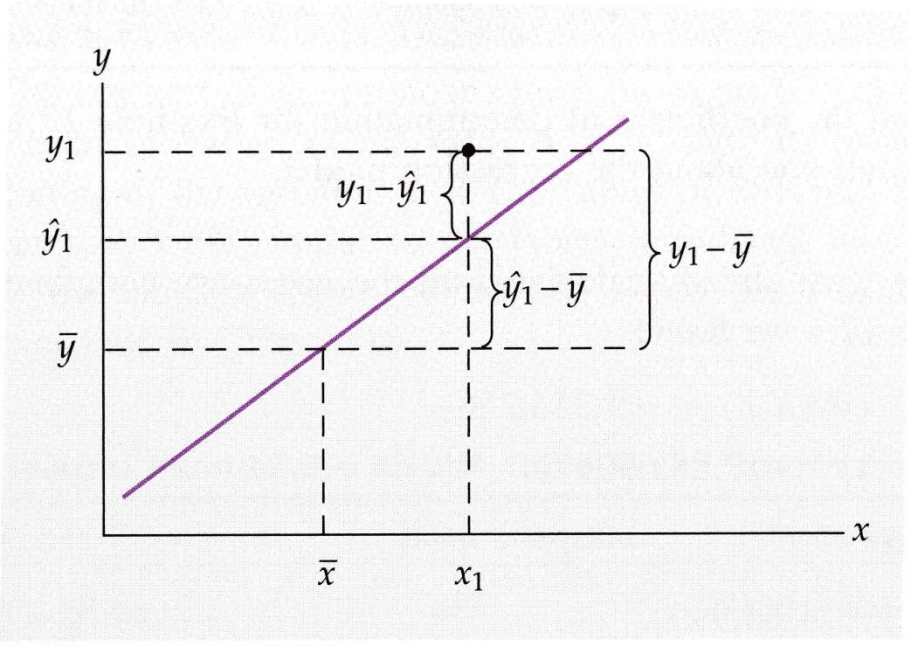
Most hypothesis tests of significance in regression involve the *F*-Distribution. While you don't need to thoroughly understand the *F*-distribution to perform a regression analysis of data, a cursory understanding is essential to understanding how regression evaluates competing models in multiple regression.

Full vs Reduced Models – Simple Linear Regression

Much of this lecture is concerned with single variable (simple linear) regression. The hope is that our familiarity with simple regression will allow us to focus on the goal of the lecture: to investigate the role of the *F*-distribution in regression. However, the F-distribution plays an even bigger role in multiple regression and analysis of variance.

We wish to compare the regression model, regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$, to a simpler model without X. For the purposes of this discussion, we will call the former the Full model, and the latter the Reduced model. This terminology will later be extended to multiple regression for the purpose of comparing different models. The table below summarizes some terminology and expressions for these models.

Terminology	Full Model in Simple Regression	Reduced Model
Random Model	$\boldsymbol{Y} = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$	$\boldsymbol{Y} = \beta_0 + \boldsymbol{\varepsilon}$
Fitted Model	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$y = \bar{y}$
Prediction Errors	Residuals: $y_i - \hat{y}_i$	Deviations: $y_i - \bar{y}$
Sum of Squares, *SS*	$SSE = \sum (y_i - \hat{y}_i)^2$	$SST = \sum (y_i - \bar{y})^2$
Degrees of Freedom, *df*	$df_{SSE} = n - 2$	$df_{SST} = n - 1$



The Residual $e_1 = y_1 - \hat{y}_1$, and Deviation $y_1 - \bar{y}$, for the first observation (x_1, y_1)

Example 1: The small set of four observations below will demonstrate the definitions. The fitted full and reduced models, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and $y = \bar{y}$, are used to compute the residuals and deviations from which the sums of squares *SSE* and *SST* are calculated.

x	y	Reduced Model	Deviations	Full Model	Residuals	Full - Reduced
		y-bar	y - ybar	y^ = 8 - x	y - y^	y^ - ybar
1	8	5	3	7	1	2
2	4	5	-1	6	-2	1
4	6	5	1	4	2	-1
5	2	5	-3	3	-1	-2

For the example above:

- $SST = \sum (y_i - \bar{y})^2 = 3^2 + (-1)^2 + 1^2 + (-3)^2 = 20$ (Total variation of ***Y*** about the reduced model)
- $SSE = \sum (y_i - \hat{y}_i)^2 = 1^2 + (-2)^2 + 2^2 + (-1)^2 = 10$ (Residual variation of ***Y*** about the full model)
- $SSR = \sum (\hat{y}_i - \bar{y})^2 = 2^2 + 1^2 + (-1)^2 + (-2)^2 = 10$ (Reduced (explained) variation of full model)

Discussion: It is no coincidence that $SST = SSE + SSR$. The variation *SSR* is the difference between the total variation *SST* of ***Y*** about the reduced model and the residual variation *SSE* of ***Y*** about the regression model. This is why *SSR* is interpreted as the variation of ***Y*** explained by the regression.

The Mean Squares in Multiple Regression

A sum of squares divided by its degrees of freedom represents a variance called a mean square. We've already encountered a mean square in regression, the error mean square *MSE*. The table below summarizes relationships between sums of squares, degrees of freedom, and mean squares in a multiple regression model with ***k*** independent (predictor) variables.

One dependent variable:
k Independent variables:
Number of observations: *n*

Analysis of Variance

Source (of Variation)	**SS**	***Df***	**Mean Square**	**F-Ratio**
Regression (Model)	*SSR*	*k*	*MSR = SSR/k*	*F = MSR/MSE*
Error (Residuals)	*SSE*	*n – k* - 1	*MSE = SSE/(n – k – 1)*	
Total	*SST*	*n* - 1		

The F-Ratio in Regression

If the assumptions of the regression model are satisfied, then the ratio $\frac{MSR}{MSE}$ follows an *F*-distribution with degrees of freedom *k* and *n* - *k*-1. (**Note:** The *F-Ratio* has two degrees of freedom, one for the numerator mean square and one for the denominator mean square.) Heuristically, *MSR* represents an explained variance while *MSE* represents the unexplained variance of the error variable in the model. A useful model explains more variation in ***Y*** than it leaves unexplained, so we look for a large *F-Ratio*.

In multiple regression the test for the model has the form,

$H_0 : \beta_1 = 0, \cdots, \beta_k = 0$, i.e., all slope parameters are zero

H_A : at least one slope parameter is not zero

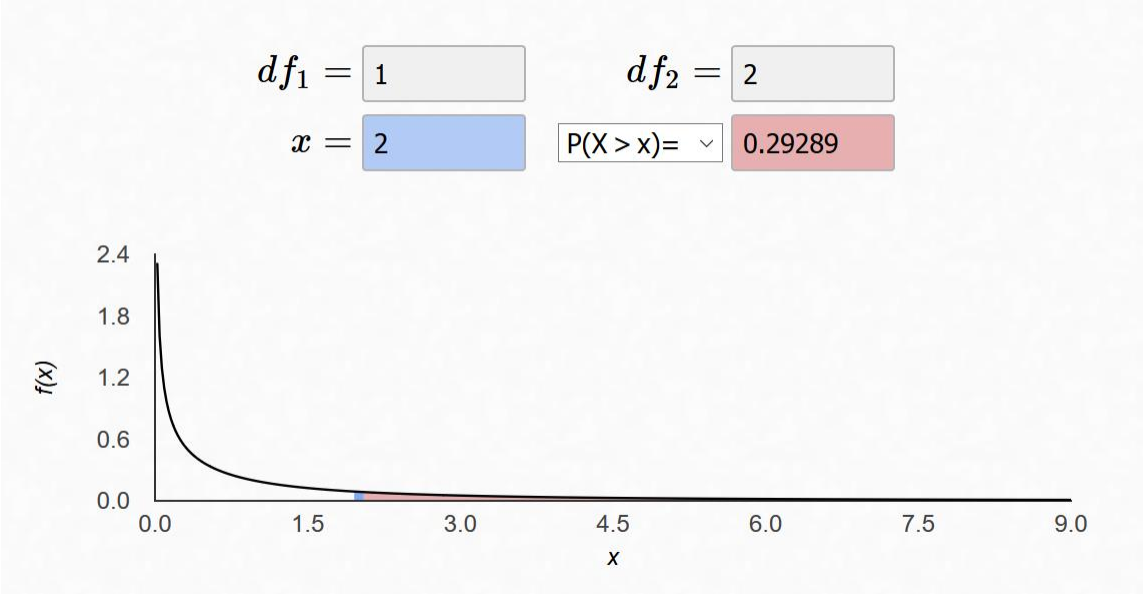
The test statistic is the *F-Ratio*.

Assuming model assumptions are satisfied, the *P*–value for the test is the probability $P\left(F_{k,n-k-1} \geq MSR/MSE\right)$, so it is a right-tailed test in *F*.

Example 1: Returning to the simple regression of ***Y*** on ***X*** begun previously, $n = 4$ and $k = 1$ for the (full) regression model. Inserting the values obtained previously,

Source	SS	*Df*	Mean Square	F-Ratio
Regression	*SSR* =10	$k = 1$	$MSR = 10/1 = 10$	$F = MSR/MSE = 10/5 = 2$
Error	$SSE = 10$	$n - k - 1 = 2$	$MSE = SSE/2 = 5$	
Total	$SST = 20$	$n - 1 = 3$		

Finally, the *P*-value for the model is $P\left(F_{k,n-k-1} \geq MSR/MSE\right) = P\left(F_{1,2} \geq 2\right) = 0.2929$, where either an online applet, such as the one below, or a graphics calculator such as the TI-84 is used to evaluate the *P*-value.



the *F*-Distribution for Example 1 with the area corresponding to the *P*-value shaded

Sadly, the relatively large *P*-value suggests that the regression model we have painstakingly constructed is not statistically significant. The full model $\boldsymbol{Y} = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$ may not be an improvement over the reduced model $\boldsymbol{Y} = \beta_0 + \boldsymbol{\varepsilon}$.

Of course, we will typically do all of the above in a dedicated statistical package. Below is the *Analysis of Variance* table for this example in Statgraphics.

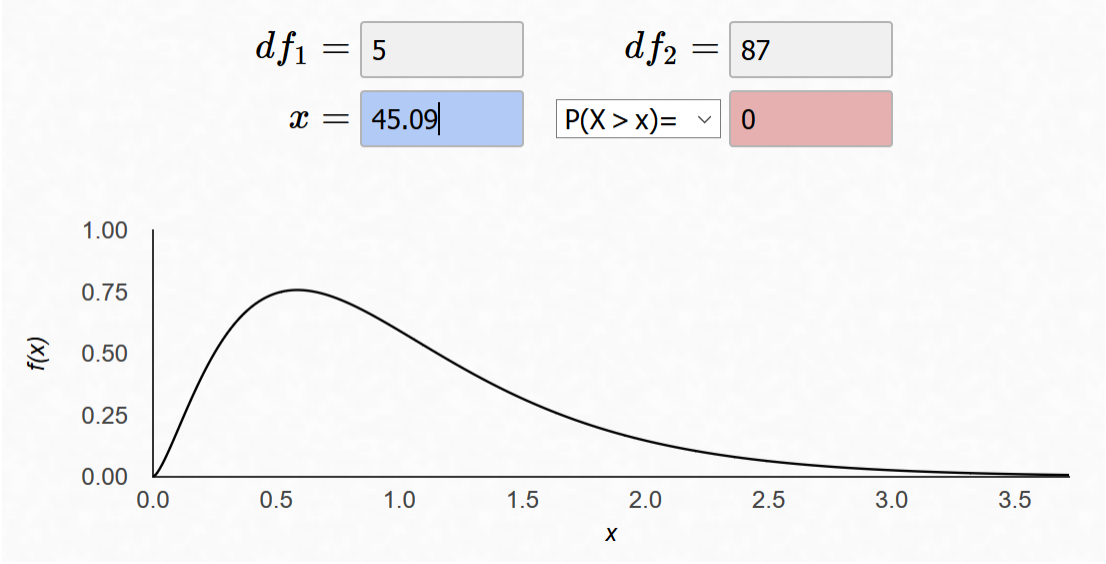
Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	10.0	1	10.0	2.00	0.2929
Residual	10.0	2	5.0		
Total (Corr.)	20.0	3			

Example 2**:** Using the 93cars dataset from homework 5, the *Analysis of Variance* table for regressing *MPG Highway* on the five predictors *Engine Size*, *Horsepower*, *Fueltank*, *Wheelbase*, and *Weight* appears below. You should verify that the values given in the table for degrees of freedom, mean squares, the *F-Ratio*, and the *P*-value are correct.

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1887.12	5	377.424	45.09	0.0000
Residual	728.193	87	8.37003		
Total (Corr.)	2615.31	92			



the *F*-Distribution for Example 2. The *P*-value is essentially zero.

Finally, it should be noted that a large *F*-Ratio and small *P*-value for the model does *not* imply that all of the variables in the model are significant. Individual *t*-tests of the predictors and further analysis of the model must still be performed, including the residuals diagnostics. As the table below shows, the variables *Engine Size* and *Horsepower* are probably correlated, making neither of them appear significant to *MPG Highway*.

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	30.7234	7.70917	3.9853	0.0001
Engine Size	0.925742	0.576821	1.6049	0.1121
Horsepower	0.00914446	0.0104104	0.8784	0.3821
Fueltank	-0.46557	0.20849	-2.23306	0.0281
Wheelbase	0.326317	0.103934	3.13965	0.0023
Weight	-0.0102779	0.00188816	-5.44337	0.0000