Math 3339

Section 27204
MWF 10-11:00am AAAud 2

Bekki George
bekki@math.uh.edu
639 PGH

Office Hours:
M & Th noon – 1:00 pm & T 1:00 – 2:00 pm
and by appointment

The *F* test for Significance of Regression

The F distribution with v_1 and v_2 degrees of freedom is found by $F = \dfrac{\chi_1^2 / v_1}{\chi_2^2 / v_2}$

χ_1^2 has v_1 degrees of freedom and χ_2^2 has v_2 degrees of freedom
For a regression model with 2 parameters (let p represent the number of parameters for the general formula and n is the number of values), $v_1 = p - 1 = 2 - 1 = 1$ and $v_2 = n - p = n - 2$

For a regression model with 2 parameters, the F test statistic can also be calculated by

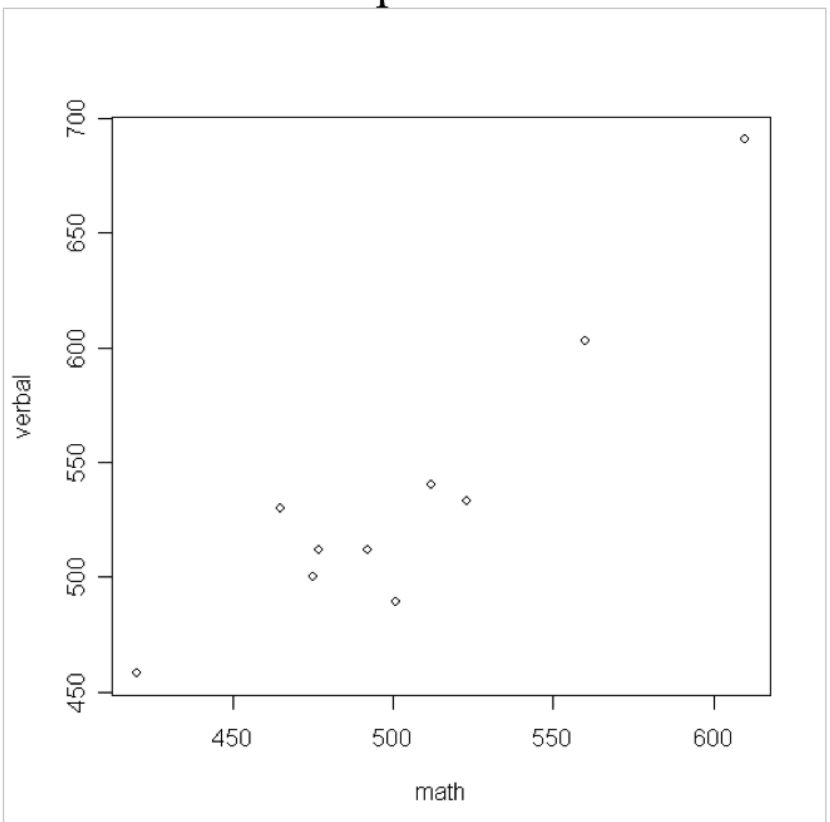
$$F = \frac{MS(regr)}{MS(resid)}$$

This is also equal to the square of the t statistic (t^2) on the test for $H_0 : \beta_1 = 0 \; vs. H_a : \beta_1 \neq 0$.

Example: The table below displays the performance of 10 randomly selected students on the SAT Verbal and SAT Math tests taken last year.

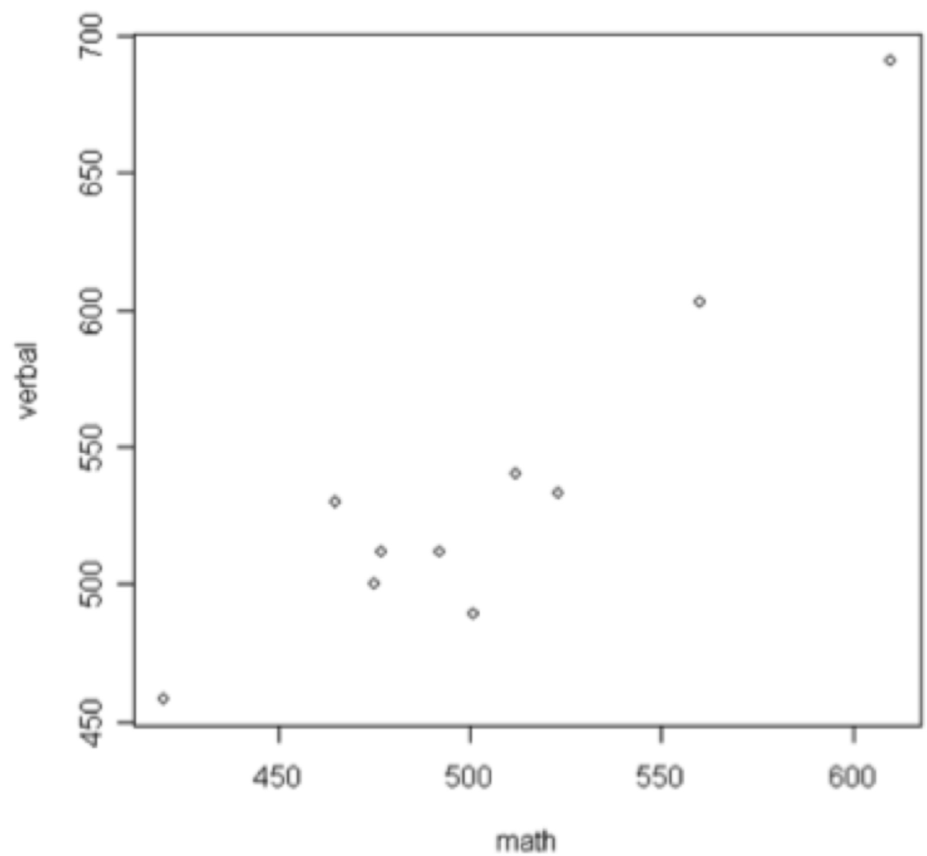
Student	1	2	3	4	5	6	7	8	9	10
Math	475	512	492	465	523	560	610	477	501	420
Verbal	500	540	512	530	533	603	691	512	489	458

Here is the scatter plot:



Popper 32

1. What can be said about this scatter plot?
   a. There is a strong negative linear relationship
   b. There is a weak negative linear relationship
   c. There is a strong positive linear relationship
   d. There is a weak positive linear relationship
   e. None of these



Here is the computer output:

```
> math=c(475,512,492,465,523,560,610,477,501,420)
> verbal=c(500,540,512,530,533,603,691,512,489,458)
> summary(lm(verbal~math))

Call:
lm(formula = verbal ~ math)

Residuals:
    Min      1Q  Median      3Q     Max
-44.904 -10.271  -1.517  14.915  37.796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -46.4249    84.8013  -0.547 0.599002
math          1.1583     0.1676   6.912 0.000123 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 26.55 on 8 degrees of freedom
Multiple R-Squared: 0.8566      Adjusted R-squared: 0.8386
F-statistic: 47.77 on 1 and 8 DF,  p-value: 0.0001231
```

coeff.

(x)

v_1 v_2

$r = \sqrt{.8566}$

Popper

2. Calculate the least-squares regression line for this data.
   a. $\hat{y} = -46.425 + 1.158x$
   b. $\hat{y} = 1.158 - 46.425x$
   c. $\hat{y} = -46.425 + .9255x$
   d. $\hat{y} = 84.8013 + 1.158x$
   e. none of these

3. What is the value of *r*?
   a. .8566
   b. .9255
   c. 1.158
   d. .1676
   e. none of these

Example. The file fire_theft.csv contains the following data: the number of fires per 1000 housing units and the number of thefts per 1000 population within the same Zip code in the Chicago metro area. (Reference: U.S. Commission on Civil Rights) The data can be found here:
http://www.math.uh.edu/~bekki/3339/notes/fire_theft.csv

Test whether there is a significant relationship between fire and thefts in that zip code.

Non-Linear Methods
Many times a scatter-plot reveals a curved pattern instead of a linear pattern.

We can **transform** the data by changing the scale of the measurement that was used when the data was collected. In order to find a good model we may need to transform our *x* value or our *y* value or both.

Let's exam this data to see if the LSRL is a good fit.

Year	1790	1800	1810	1820	1830	1840	1850	1860	1870	1880
People per square mile	4.5	6.1	4.3	5.5	7.4	9.8	7.9	10.6	10.09	14.2
Year	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980 ← 190
People per square mile	17.8	21.5	26	29.9	34.7	37.2	42.6	50.6	57.5	64

Here we will let 1790 be year 0, 1800 will be year 10, …

$$\log \hat{y} = 1.379 + .015x$$

$$\hat{y} = e^{1.379 + .015x}$$

pred 1990

$$\hat{y}(200) = e^{1.379 + .015(200)} = 79.73$$

Popper $4-6 = A$