Simple Linear Regression

Part 6: ANOVA Table, F and t Tests

STAT 705: Regression and Analysis of Variance

Partitioning the Total Sum of Squares

Start with the Total Sum of Squares : $\text{SSTot} = \text{SS}_{\text{YY}} = \sum\left(Y_i - \bar{Y}\right)^2$

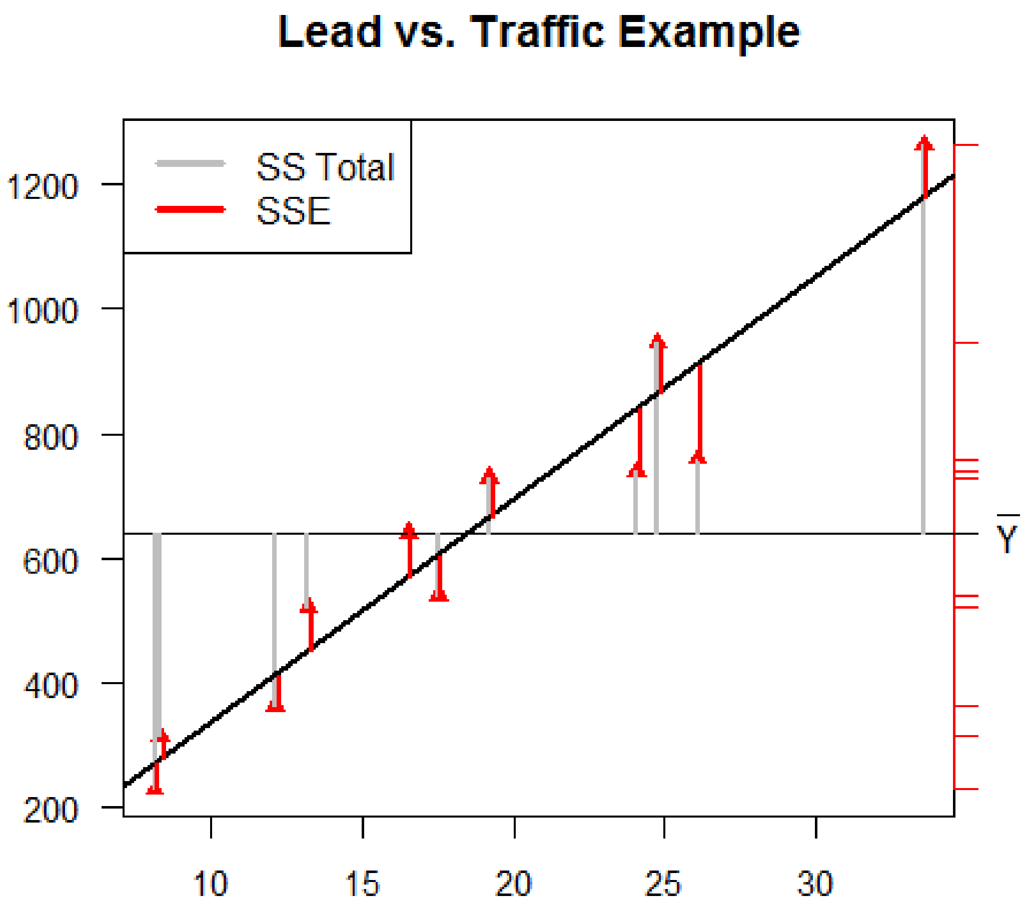
Add and subtract the predicted value $\hat{Y}_i$: $\text{SSTot} = \sum\left(Y_i - \hat{Y} + \hat{Y} - \bar{Y}\right)^2$

Separate the terms : $\text{SSTot} = \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y} - \bar{Y}\right)^2$

Partitioned Sum of Squares : $\text{SSTot} = \text{SSE} + \text{SSreg}$

We use the partitioned sums of squares to help quantify the relationship between X and Y.

Visualize the Partitions



SS Total
(same as SS_{YY})

Regression Line

SSE
deviation from
regression line

SSReg is the difference between SSTotal and SSE.
SSTot = SSReg + SSE

ANOVA Table

ANOVA Table for the Shear Strength vs. Age Example					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1,527,483	1,527,483.000	165.38	<.0001
Error	18	166,255	9236.381		
Corrected Total	19	1,693,738			

ANOVA Table Calculations for Simple Linear Regression					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	k-1	SSReg	SSReg / dfReg	MSReg/MSE	p-value
Error	n-k	SSE	SSE / dfE		
Corrected Total	n-1	SSTot			

Calculations for ANOVA Table

ANOVA Table Calculations for Simple Linear Regression					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	k-1	SSReg	SSReg / dfReg	MSReg/MSE	p-value
Error	n-k	SSE	SSE / dfE		
Corrected Total	n-1	SSTot			

- k = number of parameters in the model
- SAS uses the term ‘Model’ and we use ‘Regression’
- df is degrees of freedom
- p-value comes from a new probability distribution: the F distribution

F Distribution

- The ‘F value’ in the ANOVA table is a statistic
  - it is calculated from the data
  - it has a probability distribution, namely an F distribution
- An F distribution has two parameters
  - numerator df and denominator df
  - for simple linear regression
    - numerator df = dfReg = $k - 1 = 1$ (k is number of parameters)
    - denominator df = dfE = $n - k = n - 2$
  - these two parameters control the shape of the F distribution

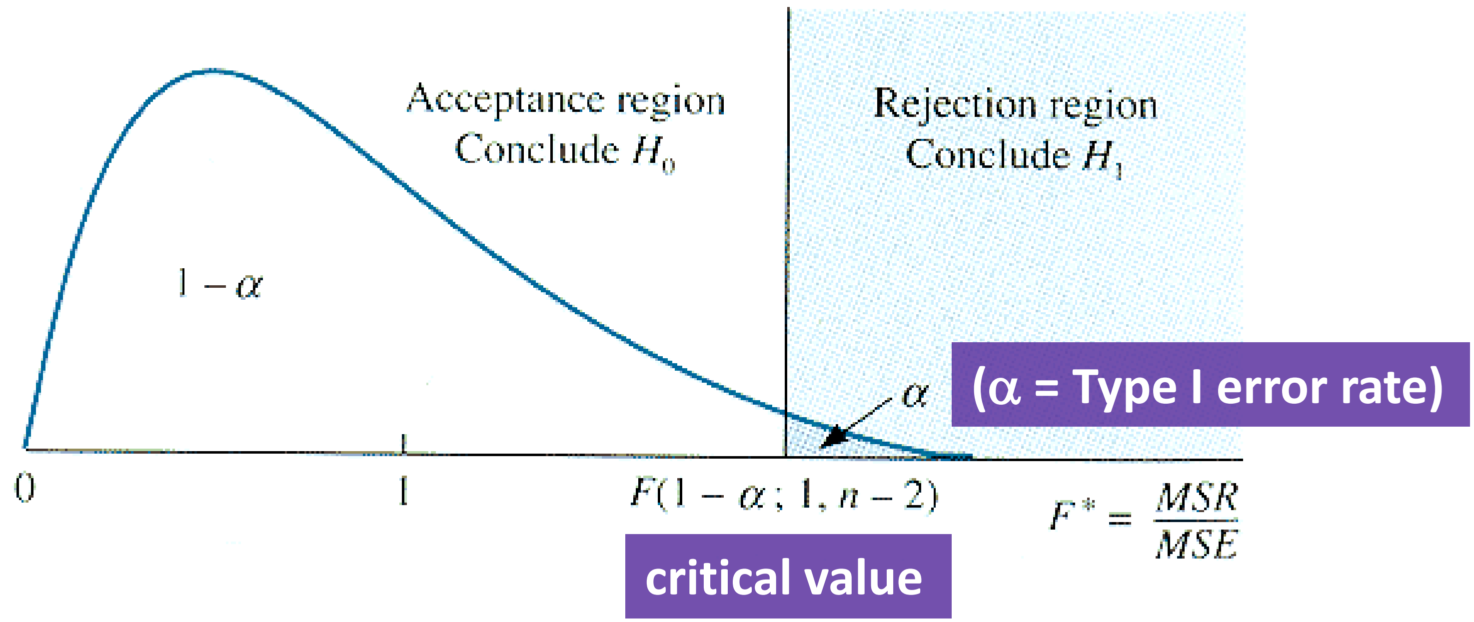
ANOVA F Test

- In the ANOVA table, ‘Pr>F’ provides the p-value for a hypothesis test
- Compare two models
  - Full model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Reduced model: $Y_i = \beta_0 + \varepsilon_i$
- Hypotheses

  H_0: Reduced model adequately fits the data (full model is not needed)

  H_a: The full model is needed to adequately fit the data
- Test statistic is ‘F value’, i.e., F = MSReg/MSE
- Compare this to a critical value from the F distribution

Critical Value from F Distribution



- F tests are always right-tailed, so
  - Rejection region is on the right
  - Right-tailed area is α, the significance level of the test
- Notation for the critical value: $F(1 - \alpha; 1, n - 2)$

Using the F Table

- Table is provided on course website
- For the Shear Strength vs. Age example
  - $n = 20$; $dfReg = 1$; $dfE = 20 - 2 = 18$
- Locate along the top
  - denominator df = dfE = 18
- Locate along the left side
  - numerator df = dfReg = 1
- Select critical value for specified α

Reading the F Table

A Portion of the Provided F table

		Denominator Degrees of Freedom						
df1	area	1	...	12	15	20	24	...
1	0.1	39.86	...	3.18	3.07	2.97	2.93	...
	0.05	161.45	...	4.75	4.54	4.35	4.26	...
	0.03	647.79	...	6.55	6.2	5.87	5.72	...
	0.01	4052.2	...	9.33	8.68	8.1	7.82	...
2	0.1	49.5	...	2.81	2.7	2.59	2.54	...
	0.05	199.5	...	3.89	3.68	3.49	3.4	...
	0.03	799.5	...	5.1	4.77	4.46	4.32	...
	0.01	4999.5	...	6.93	6.36	5.85	5.61	...

Numerator DF

Find numerator df along left side

Find denominator df along top (df = 18 is between 15 and 20)

Select the area (significance level of the test)

Read the critical value at the intersection

For the Shear Strength vs Age example,

- numerator df = 1
- denominator df = 18

Result:
The critical value is between 4.54 and 4.35. With software, we obtain a more accurate value of 4.41.

F-Test and t-test

- Hypotheses for F test:
  - H_0: Model is $Y_i = \beta_0 + \varepsilon_i$
  - H_a: Model is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Hypotheses for t test: H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$
- These two tests are equivalent
- From statistical theory, it can be shown
  - if a random variable T follows a t distribution with df = D
  - then T^2 follows an F distribution with numerator df = 1 and denominator df = D
- So… for simple linear regression, the t test and the F test are performing exactly the same comparison, and should come to exactly the same conclusion

SAS Output for F and t tests

Shear Strength vs. Age example

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	2627.82236	44.18391	59.47	<.0001
Age	1	-37.15359	2.88911	-12.86	<.0001

ANOVA Table					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	1,527,483	1,527,483.000	165.38	<.0001
Error	18	166,255	9236.381		
Corrected Total	19	1,693,738			

For simple linear regression, these two tests are testing the same hypotheses:

- H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$
- (This is why the p-values are the same.)
- Note that $t^2 = F$, i.e. $(-12.86)^2 = 165.38$

Why do We Need Both F and t Tests?

- t tests are used to test ONE parameter in a linear model
- F tests can be used to test many parameters in the linear model
- For simple linear regression, there is only one parameter (not counting the intercept)
- For multiple linear regression, there are many parameters, so the distinction between F and t tests will become important

Some Considerations

- Observational vs. Experimental data
  - Dictates scope of inference
- Interpretation of hypothesis tests
- Implications of failing to reject H_0: slope = 0
- Avoid extrapolation
  - We are always limited by the available data

Observational Data

- Lead concentration vs. Traffic example
  - The number of vehicles at each site was observed
  - No attempt was made to manipulate the number of vehicles
  - The number of vehicles was not randomly assigned to a location
- Shear Strength vs. Age
  - The age of a batch of rocket propellant was observed
  - No attempt was made to control or manipulate the age
  - Age was not randomly assigned to the propellant

Experimental Data

- Sometimes, data are derived from randomized controlled experiments
- Basic types of experiments include Completely Randomized Design, Factorial Design, and Randomized Block Design
  - We will learn more about these later in the course
- In an experiment, the X variable is defined and/or manipulated for the purpose of measuring the effect on Y.
- To assess whether or not a change in X <u>causes</u> a change in Y, the data <u>must</u> come from a randomized experiment

Scope of Inference

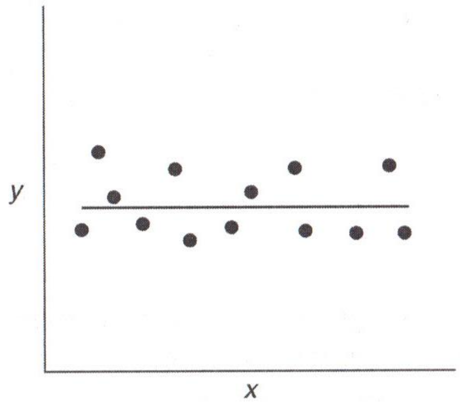
- Observational data
  - Can explore whether or not the variables are associated
  - Can NOT determine a cause-and-effect relationship between the variables
- In the NASA propellant example
  - WE CAN SAY: If the age of the propellant increases by 1 week, then the shear strength decreases, on average, by 37.15 pounds per square inch.
  - WE CANNOT SAY: The increase in age CAUSES the decrease in strength.
  - We only know that the two quantities are associated.

Interpreting a Hypothesis Test

- For hypothesis tests
  - We can NEVER 'prove' H_0 is true
  - We can NEVER 'prove' H_a is true
  - We either reject H_0 or fail to reject H_0 (we do not 'accept H_0')
- Reject H_0 if the sample provides enough evidence for use to be (almost surely) convinced that H_0 is false.
- Fail to reject H_0 if the evidence against H_0 is not convincing

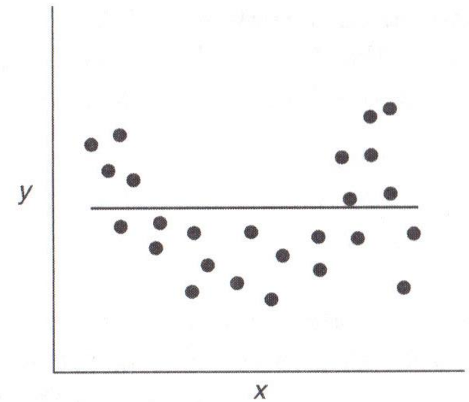
If We Fail to Reject H_0: $\beta_1 = 0$

There may not be a relationship between X and Y



… OR …

The relationship between X and Y may not be linear



There may still be a ‘true’ relationship between X and Y in the population, but the test may not have enough power to detect it in the given dataset. (e.g., The sample may be too small.)

Avoid Extrapolation

- NASA example: Shear Strength = 2627.8 – 37.15 * Age
- For a propellant with Age = 75 weeks
  - Strength = 2627.8 – 37.15*75 = -158.45 psi
- HOWEVER…
  - Available data has Age values from 2 to 24 weeks
  - It is not clear what happens for Age greater than 24
  - The model predicts a negative value for Strength (not possible !)
- Extending inference beyond the scope of the data is called extrapolation, and **<u>is not valid</u>**.

Things You Should Know

- Use SAS to generate the ANOVA table
- Understand the relationship between the values in the ANOVA table
- Write the hypotheses being tested by the F test and the t test
- Find critical values in the F table
- Interpret the results of the F test and t test