

Regression Inference

Dr Tom Ilvento

Department of Food and Resource Economics



Overview

- We will look at the main components of inference in the regression model
 - F-test for the overall model
 - t-tests for individual coefficients
- Most of these ideas should seem familiar to you
 - We have an estimate
 - And a sampling distribution of the estimator
 - Which gives us a standard error
- For Regression I will also show Confidence Intervals around predictions of Y

2

Regression Model Assumptions

- **Mean of Probability Distribution of Error Is 0**
 - The error terms are centered around zero
 - No bias in our estimates
- **Prob. Distribution of Error has Constant Variance = σ^2**
 - We will pool the variance to make our estimate of sigma
- **Probability Distribution of Error is Normally distributed**
 - So we can use the normal distribution to make inferences and determine probabilities
- **Errors Are Independent – uncorrelated with each other**
 - Error terms are not related to each other or to another variable, such as time
- **For each assumption there is a way to test for it and potential solutions if there are problems with the assumption**

3

Inference in Regression

- We have two main inferential tests with regression
 - **F-test for the overall model**
 - **t-test for individual coefficients**
- We can also generate a confidence interval for our coefficients – Excel will give us this without even asking
- We will focus on general conclusions by looking at p-values
- There are other, more sophisticated tests in regression
 - We will only look at the main ones
- I would recommend a more advanced course in regression

4

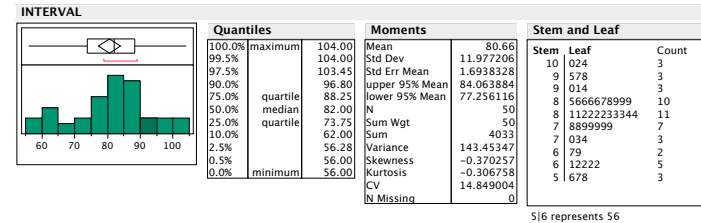
Data Example: Old Faithful eruptions

- Old Faithful is a geyser located in Yellowstone National Park
- It was so named because it erupts on a regular basis. Eruptions can shoot 3,700 to 8,400 gallons of boiling water to a height of 106–185 feet lasting from 1.5 to 5 minutes.
- The average height of an eruption is 145 feet
- Eruptions often occur about 90 minutes apart, but this interval can range from 45 to 125 minutes on occasion.
- A model was posed by Harry Woodward that expressed the interval between eruptions as a function of the duration of the previous eruption
- The duration is timed from the first heavy surge which lifts water skyward at the start of the eruption until the last small splash above the cone at the very end.

5

Old Faithful Data

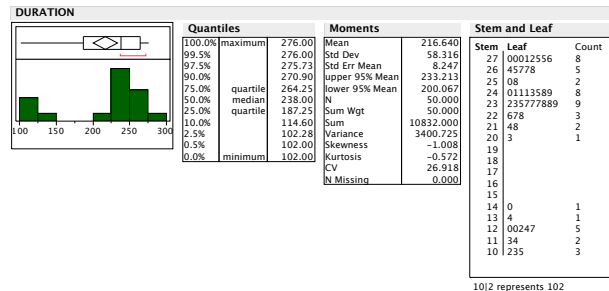
- The Dependent Variable: Interval** (in minutes)
 - The mean is 80.66 minutes and close to the median
 - The distribution is something like symmetrical, mound-shaped
 - Mean and Median are close to each other
 - CV is 14.85



6

Old Faithful Data

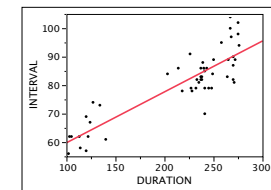
- The Independent Variable: Duration** of the eruption (seconds)
- This is bimodal with a mean of 216
- The value at the first decile is 114.6
- The median is 238.00



7

Regression of Interval on Duration

- The correlation between INTERVAL and DURATION is .870
- The Scatterplot shows a strong positive relationship
- The Regression Model has $R^2 = .757$, not a perfect fit!
- The estimated equation is:
 - est INTERVAL = 41.9450 + .1787 DURATION**
- When DURATION = 0, INTERVAL = 41.95
- For each second increase in the duration, the INTERVAL increases by .179 minutes



Linear Fit

INTERVAL = 41.944986 + 0.1787067*DURATION

Summary of Fit

RSquare	0.75708
RSquare Adj	0.75202
Root Mean Square Error	5.964365
Mean of Response	80.66
Observations (or Sum Wgts)	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5321.6847	5321.68	149.5962
Error	48	1707.5353	35.57	
C. Total	49	7029.2200		

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob > t
Intercept	41.944986	3.275789	12.80	<.0001*
DURATION	0.1787067	0.014611	12.23	<.0001*

8

F-test of the model: a first step

- F-Test
- very general test that none of the independent variables are significantly different from zero
- It is a test of the ratio of two variances to determine if the variances are equal to each other – we are looking for a value near 1 to say the ratios are equal
- If there is only one independent variable, the F-Test = (t-test)² i.e., $F = t^2$
- The null and alternative hypothesis for the F-test are
 - Ho: $\beta_1 = \beta_2 = \dots \beta_k = 0$** Nothing is going on in the model
 - Ha: at least one not equal to zero** Something is going on in the model

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

9

The F-test for the Old Faithful Data

ANOVA					
	df	SS	MS	F	Sig F
Regression	1	5321.685	5321.685	149.596	0.000
Residual	48	1707.535	35.574		
Total	49	7029.220			

- General Test
 - Ho: $\beta_1 = \beta_2 = \dots \beta_k = 0$
 - Ha: at least one $\beta \neq 0$
 - This is always a two-tailed test
- For Old Faithful data
 - Ho: $\beta_1 = 0$**
 - Ha: $\beta_1 \neq 0$**
 - Assumptions: Interval is approximately normally distributed; pooled measure of variance across all levels of Duration**
 - F* = 149.596 p = .000**
 - F_{critical} is based on $\alpha = .05$ and 1 and 48 d.f. = 4.043**
 - Conclusion: reject Ho: $\beta_1 = 0$**

10

Next Step: examine the coefficients in the model

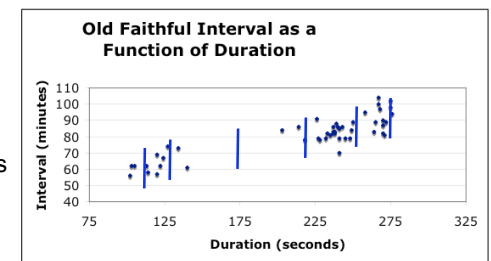
	Coef	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	41.945	3.276	12.805	0.000	35.359	48.531
DURATION	0.179	0.015	12.231	0.000	0.149	0.208

- For the intercept and each independent variable in the model, the output will supply an estimated coefficient
 - Each has a standard Error
 - t Stat or Test Statistic or t*
 - p-value
- The test is based on whether the coefficient is equal to zero
 - For the intercept, does the function go through the origin?
 - For the independent variable, a coefficient = zero means there is no relationship

11

The Standard Error

- We said the error term of our model is related to the variance and the standard error
- And that we assumed constant error variance across all levels of the independent variable X
- So the Standard Error of the Model is given as
 - You can find it in the regression statistics
 - The Standard Error for the model will factor into the standard error of the coefficients



$$s = \sqrt{\frac{SSE}{(n - k - 1)}} = \text{Root MSE}$$

Regression Statistics	
Multiple R	0.870
R Square	0.757
Adjusted R Square	0.752
Standard Error	5.964
Observations	50

12

The Standard Error of the Regression Coefficient

- It is based on the Root MSE and the total sum of squares for the independent variable (the variability of X)

$$\text{Standard Error for } \hat{\beta}_1 = \frac{\text{Root MSE}}{\sqrt{SS_X}}$$
- The numerator is the **Standard Error for the model**
 - aka the **Root Mean Squared Error** (MSE)
- And the denominator is based on the **Total Sum of Squares for X**
 - More spread in X,
 - the better we can estimate the standard error, and the smaller the standard error
- The **sample size is also a factor – degrees of freedom**
- In multivariate regression, the covariance among the independent variables also factors into the standard error for coefficients**

13

Calculations for the Standard Error of the coefficient from Old Faithful Example

- $SS_X = 166,635.5201$
- $\text{SQRT}(SS_X) = 408.2101$
- Std Error for Model = 5.964

$$\text{Standard Error for } \hat{\beta}_1 = \frac{\text{Root MSE}}{\sqrt{SS_X}} = \frac{5.964}{408.2101} = .0146$$

	Coef	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	41.9450	3.2758	12.8045	0.0000	35.3586	48.5314
DURATION	0.1787	0.0146	12.2310	0.0000	0.1493	0.2081

14

The Hypothesis Test for a Regression Coefficient

- We ask the question, ***“Is there a Linear Relationship Between X & Y?”***
- The test involves testing the sample estimate of the slope coefficient, b_1
- The Hypothesis Test
 - Ho: $\beta_1 = 0$ (No Linear Relationship)**
 - Ha: $\beta_1 \neq 0$ (Linear Relationship)**
- The theoretical basis of the test is the Sampling Distribution of the slope coefficient
 - It is based on repeated samples of Y and X of size n
 - And estimating the coefficients for each sample

15

The Hypothesis Test for a Regression Coefficient

- Ho:** Ho: $\beta_1 = 0$
- Ha:** Ha: $\beta_1 \neq 0$
- Assumptions** Equal variances, normal distribution
- Test Statistic** $t^* = 12.231$ $p < .001$
- Rejection Region** $t_{.05/2, 48} = 2.011$
- Conclusion:** $t^* > t_{.05, 48}$
or $p < .001$
Reject Ho: $\beta_1 = 0$

Now we can say that the duration of Old Faithful's eruption has a significant affect on the interval between eruptions

Our best estimate of this relationship is that one second of DURATION results in an increase of .1787 minutes in the INTERVAL

16

The t-test from Regression Output

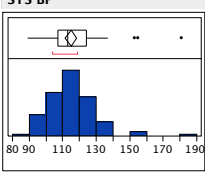
	Coef	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	41.9450	3.2758	12.8045	0.0000	35.3586	48.5314
DURATION	0.1787	0.0146	12.2310	0.0000	0.1493	0.2081

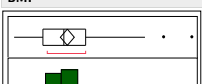
- $t^* = (.1787-0)/.0146 = 12.2310$
- The p-value is based on looking up t^* in the t-distribution
- The t-test in most regression output is a **two-tailed test**
- **Confidence interval of a regression coefficient**
 - $b_1 \pm t_{n-2, d.f.} \cdot S_{b1}$
 - The degrees of freedom can be thought of as $n-k-1$
- Excel gives the upper and lower part – 95% C.I. is the default

17

A Problem for you to try

- The following data is for 80 subjects in a health study.
- We will focus on systolic blood pressure (**SYS BP**) regressed on Body Mass Index (**BMI**).
- BMI is a measure of body fat that is a function of weight and height.

SYS BP		Quantiles		Moments		
		100.0%	maximum	181.00	Mean	114.850
		99.5%		181.00	Std Dev	14.671
		97.5%		154.95	Std Err Mean	1.640
		90.0%		131.00	upper 95% Mean	118.115
		75.0%	quartile	124.00	lower 95% Mean	111.585
		50.0%	median	113.00	N	80.000
		25.0%	quartile	107.00	Sum Wgt	80.000
		10.0%		97.10	Sum	9188.000
		2.5%		92.03	Variance	215.243
		0.5%		89.00	Skewness	1.451
		0.0%	minimum	89.00	Kurtosis	4.786
					CV	12.774
					N Missing	0.000

BMI		Quantiles		Moments		
		100.0%	maximum	44.900	Mean	25.869
		99.5%		44.900	Std Dev	4.959
		97.5%		40.528	Std Err Mean	0.554
		90.0%		31.880	upper 95% Mean	26.972
		75.0%	quartile	28.650	lower 95% Mean	24.765
		50.0%	median	25.350	N	80.000
		25.0%	quartile	22.175	Sum Wgt	80.000
		10.0%		19.810	Sum	2069.500
		2.5%		18.323	Variance	24.594
		0.5%		17.700	Skewness	1.114
		0.0%	minimum	17.700	Kurtosis	2.295
					CV	19.171
					N Missing	0.000

18

JMP Results

- Can you see a few outliers?
- **Conduct the F-test**
- **Conduct the hypothesis test that the coefficient for BMI is equal to zero**
- **Solve the equation for a value BMI=25**

Bivariate Fit of SYS BP By BMI				
Linear Fit				
Linear Fit				
SYS BP = 75.718579 + 1.5126908*BMI				
Summary of Fit				
RSquare		0.261455		
RSquare Adj		0.251987		
Root Mean Square Error		12.68876		
Mean of Response		114.85		
Observations (or Sum Wgts)		80		
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	4445.836	4445.84	27.6131
Error	78	12558.364	161.00	Prob > F
C. Total	79	17004.200		<.0001*
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	75.718579	7.580698	9.99	<.0001*
BMI	1.5126908	0.287868	5.25	<.0001*

19

Results from F and t-tests

- **F-test**
 - H_0 : None of the variables are related to SYS BP
 - H_a : At least one of the variables is related to SYS BP
 - $F^* = 27.613$
 - $p < .001$
 - Reject H_0 :
- **t-test**
 - $H_0: \beta_1 = 0$
 - $H_a: \beta_1 \neq 0$
 - $t^* = 5.255$
 - $p < .001$
 - Reject $H_0: \beta_1 = 0$
- **Since this is a bi-variate regression, the two tests completely agree with each other**

20

Estimate of Sys BP when BMI = 25

- Solve the equation for BMI=25
 - Est Y = 75.710 + 1.513(BMI)
 - Est Y = 75.710 + 1.513(25)
 - Est Y = 75.710 + 37.825
 - Est Y = 113.535
- **Our estimate of Sys BP = 113.535 when BMI = 25**

21

Other Inferences – Prediction of the Mean of Y for a given level of X

- Confidence interval for the mean of Y: The mean of Y for each level of X

$$\hat{Y}_i \pm t_{n-2, d.f.} S_{XY} \sqrt{h}$$

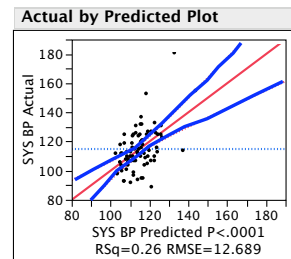
- This is a prediction of Y for a given level of X, where the prediction is really thought of as the average of Y at a given level of X
- This depends upon
 - the sample size (n)
 - The Standard Error for the model (S_{XY}) which is the Root MSE
 - The variability in X (SS_X)

22

Other Inferences: Prediction of Y for a given level of X

- Confidence interval for a prediction of Y: a new prediction of Y for a given level of X
- This is a prediction of Y for a given level of X, where the prediction is really thought of as a new single value
- This depends upon
 - the sample size (n)
 - The Standard Error for the model (S_{XY})
 - The variability in X (SS_X)

$$\hat{Y}_i \pm t_{n-2, d.f.} S_{XY} \sqrt{1+h}$$



23

Summary

- We discussed basic inferential tests in Regression - the F and t-tests
- Regression also has a known sampling distribution for the estimates of the intercept and slope coefficients
- The sampling distribution, like that for the mean is based on estimates from repeated samples.
- And it follows a normal distribution (or the t-distribution)
- Tests are based on a pooled estimate of the variance – MSE and the Standard Error for the model
- I also showed two other confidence intervals
 - For the Mean of Y for a given X
 - For a new prediction of Y for a level of X

24