

Regression 2: the Output

Dr Tom Ilvento

Department of Food and Resource Economics



Overview

- In this lecture we will examine the typical output of regression.
- I will use Excel's output, but the results will be the same in JMP, SAS, Minitab, or any other program.
- We will stop short of the formal inference, but you will see
 - The ANOVA Table
 - An F-test
 - t-tests
 - Standard Errors (SE)
 - p-values and confidence intervals

2

Let's look closely at the Excel Output for the Regression of Catalog Sales on Salary

SUMMARY OUTPUT of SALES Regressed on SALARY

Regression Statistics	
Multiple R	0.700
R Square	0.489
Adjusted R Square	0.489
Standard Error	687.068
Observations	1000

ANOVA					
	df	SS	MS	F	Sig F
Regression	1	451624335.68	451624335.68	956.71	0.000
Residual	998	471117860.07	472061.98		
Total	999	922742195.74			

	Coef.	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-15.332	45.374	-0.338	0.736	-104.373	73.708
SALARY	0.021961	0.000710	30.931	0.000	0.021	0.023

3

Regression Statistics

- **Multiple R** – in a bivariate regression, this is the absolute value of the correlation coefficient $|r|$. In a multivariate regression it is the square root of R^2
- **R-Square** – A measure of association that gives us an indication of the linear fit of the model. R-square ranges from 0 (nothing explained by the model) to 1 (a perfect fit).
- **Adjusted R-Square** – R-square will always increase as you add independent variables to a model. To account for this, the adjusted R-square modifies R^2 to account for the number of independent variables in the model.

Regression Statistics	
Multiple R	0.700
R Square	0.489
Adjusted R Square	0.489
Standard Error	687.068
Observations	1000

4

Regression Statistics

- **Standard Error** – The standard error of the model - the square root of the MSE.
- This is an overall standard error of the model and is used in calculating the standard error of the coefficients in the model.
- The standard error is the square root of the MSE, which will be discussed in a later section.
- **Observations** – the number of observations in the data – always check this!

Regression Statistics	
Multiple R	0.700
R Square	0.489
Adjusted R Square	0.489
Standard Error	687.068
Observations	1000

5

The Regression ANOVA Table

ANOVA	df	SS	MS	F	Sig F
Regression	1	451624335.68	451624335.68	956.71	0.000
Residual	998	471117860.07	472061.98		
Total	999	922742195.74			

- Excel uses different terms for the components of the ANOVA Table
- However, there is a direct connect to what we learned in the previous section
- Regression SS = **SSR** Regression MS = **MSR**
- Residual SS = **SSE** Residual MS = **MSE**
- Total SS = **SST**

6

The Sums of Squares

- **Total Sum of Squares for Y**
- Since we are fitting a model to the data, it is easier to express the Sums of Squares
- We decompose the Total Sum of Squares Total for Y into a part due to
- **Regression (SSR)** – think of this as explained **451624335.68**
 - **k d.f.** where k is the number of independent variables
- **Residual (SSE or error)** – think of this as unexplained **471117860.07**
 - **n-k-1 d.f.** based on fitting k+1 estimated parameters to the models - the coefficients and the intercept

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

7

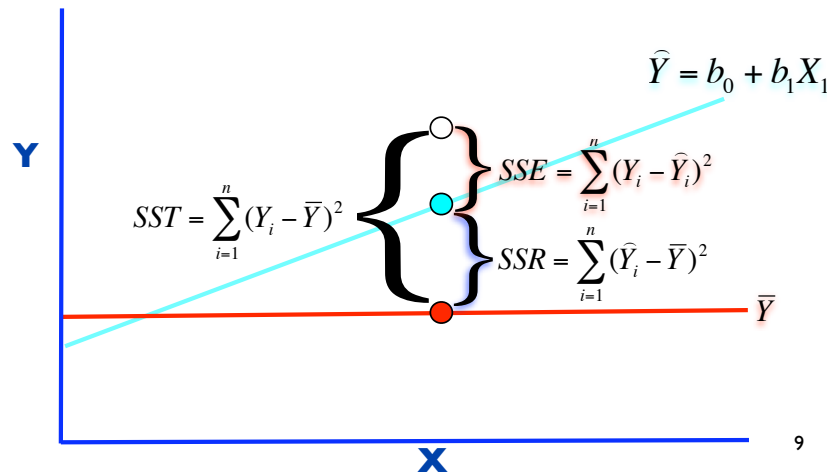
The Regression ANOVA Table

ANOVA	df	SS	MS	F	Sig F
Regression	1	451624335.68	451624335.68	956.71	0.000
Residual	998	471117860.07	472061.98		
Total	999	922742195.74			

- **Regression Sum of Squares (SSR)** - The sum of squares due to the fit of the model. The df for regression is equal to the number of independent variables in the model and is denoted by k.
- The **Mean Square due to Regression (MSR)** in the next column is equal to the SSR divided by the df.
- **Residual or Sum of Squares Error (SSE)** - this is the part of the Total Sum of Squares that is unexplained by the model. The df for the SSE is equal the sample size (n) minus 1 minus the degrees of freedom for regression: n - 1 - k.
- The **Mean Square Error (MSE)** - is equal to the SSE divided by its df. The MSE is the pooled variance of the model.

8

Decomposing the Sums of Squares



9

A Note About R-square

- $R^2 = SSR/SST$
- The Sum of Squares due to Regression divided by the Total Sum of Squares for Y (SST)
- R^2 represents what part of the total variability in Y is “explained” by knowing something about the independent variable(s)
- $R^2 = 1 - SSE/SST$
- Shows the linear “fit of the model”
- Ranges from 0 to 1
 - $R^2 = 0$ implies no linear relationship
 - $R^2 = 1$ perfect linear relationship
- What is a high R^2 depends upon the data you are working with

10

Degrees of Freedom

- Overall, the degrees of freedom are $n-1$
- Think of k as the number of independent variables in the model
- The degrees of freedom for **Regression** is k
 - In our example, $k = 1$ because we only have Salary as an independent variables
 - So, **d.f. Regression = 1**
- The degrees of freedom for **Residual** is $n-k-1$
 - The sample size minus the number of parameters estimated by the model (intercept and slope coefficients)
 - In our example, d.f. **Residual = 1000 - 1 - 1 = 998**
- The **d.f. Regression + d.f. Residual = d.f. Total**
 - In our example, **1 + 998 = 999**

11

Mean Squares

- We divide the Sums of Squares by their respective degrees of freedom to the Mean Squares

- **MS Regression = MSR = SSR/(k)**
 - **451,624,335.68/1 = 451,624,335.68**

$$MSR = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{k}$$

- **MS Residual = MSE = SSE/(n-k-1)**
 - **471,117,860.07/(1000-2) = 472,061.98**

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n - k - 1)}$$

- Think of these as “average squared deviations” or variances

12

Root Mean Square Error

- The **Root Mean Square Error** is the Square Root of the MSE
- Excel calls this the **Standard Error** under Regression Statistics
- It is the Standard Error for the model
- As with any standard error, it is based on a sampling distribution of estimating the regression on many samples of size n
- The **Root Mean Square Error** factors into the standard errors of the regression coefficients

$$\sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n - k - 1)}} = \sqrt{472061.98} = 687.068$$

Regression Statistics	
Multiple R	0.700
R Square	0.489
Adjusted R Square	0.489
Standard Error	687.068
Observations	1000

13

The Regression ANOVA Table

ANOVA	df	SS	MS	F	Sig F
Regression	1	451624335.68	451624335.68	956.71	0.000
Residual	998	471117860.07	472061.98		
Total	999	922742195.74			

- F** - The F-value is the ratio of two variances. In this case it is the ratio of The Mean Square due to Regression divided by the Mean Square Error (MSR/MSE).
- The F-distribution is a probability distribution with two separate degrees of freedom (one for MSR and one for MSE).
- A ratio of one (or close to one based on a sample) would imply that the model was a poor fit and there is no relationship of any of the independent variables with the dependent variable.
- Significance F** - The significance level associated with the F-value is the p-value (chance of being wrong) to reject a null hypothesis that the model is a poor fit (all the coefficients for the independent variables are equal to zero).
- Generally we are looking for a significance level less than .05 (p-value).

14

The F-test

- F-Test**: a very general test that none of the independent variables are significantly different from zero
- It is the ratio of the MS due to Regression divided by the MS due to Residual
- If the two MS's are equal to each other, the ratio should be about or near one
- If there is only one independent variable, the F-Test equals the square of the t-test, i.e., $F^* = t^{*2}$
- The general null and alternative hypothesis for the F-test is
 - $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
 - H_a : at least one $\beta_k \neq 0$
- Focus on the Significance F (p-value) < .05**

15

The Last Part of the Output: the Coefficients

	Coef.	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-15.332	45.374	-0.338	0.736	-104.373	73.708
SALARY	0.021961	0.000710	30.931	0.000	0.021	0.023

- Coefficients** that we estimate
 - The Intercept of the line $\hat{Y} = -15.33 + .022(\text{Salary})$
 - The slope coefficient of each independent variable
- The **Standard Error** of each coefficient
- The **t-statistic or t*** $t^* = (\hat{\beta} - 0) / s_{\hat{\beta}}$
 - Based on a Null Hypothesis that the slope coefficient is equal to zero
 - Our estimate is divided by Standard Error
- The **p-value** associated with t*
 - Probability of finding a value of t or greater given a Null Hypothesis that the coefficient is equal to zero for a **two-tailed test**
- Upper and lower Confidence Intervals** for the coefficients

16

The Regression Coefficient Confidence Interval

- The confidence interval for a coefficient is similar to the C.I. for the mean
- It is the estimate, plus or minus a component that is a function of a t-value and a standard error of the estimate
- It places a Bound of Error around our estimate
- Excel gives the upper and lower bound, based on a t-value

$$\hat{\beta} \pm t_{\alpha/2, (n-k)d.f.} (SE)$$

17

The meaning of the coefficients

• Intercept or estimated β_0

- The value of the Dependent variable if all independent variables equal zero
- When using dummy variables, the intercept is the mean of the reference category
- If a customer has no salary, the sales are: \$-15.33
- We need the intercept in our model, but its interpretation may be outside the range of our data.

• Slope or estimated β_1

- The change in Y for a unit change in X
- For every dollar increase in Salary, sales increase by \$.022
- For every \$1,000 increase in Salary, Sales increase by \$22

18

Symmetry of Regression Coefficients

• A correlation coefficient is a symmetrical measure of association

- The correlation between Y and X is the same as the correlation between X and Y
- The order doesn't matter and neither is established as the dependent or the independent variable

• A regression coefficient is not symmetrical!

- The slope and intercept resulting from a regression of Y on X
- Is not the same as a regression of X on Y

19

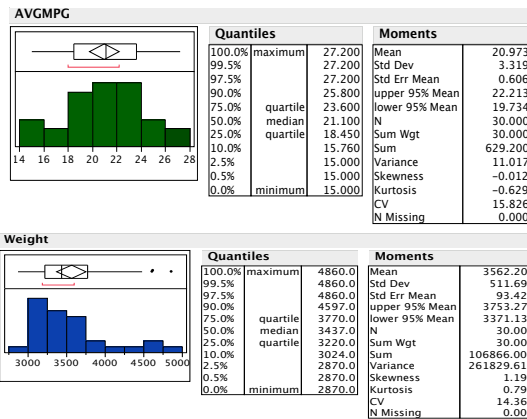
Run the Regression and see

- SALES = -15.33 + .022(SALARY)
- SALARY = 28,986.27 + 22.29(SALES)
 - The correlation between them is .700, regardless of which is first or second.
 - R-Square also remains the same in the two models.
 - But the coefficients change
- In regression, it does matter which is the dependent variable and which is the independent variable
- We typically say we "regress Y on a set of X independent variables"

20

Example: Mpg as a function of Vehicle Weight

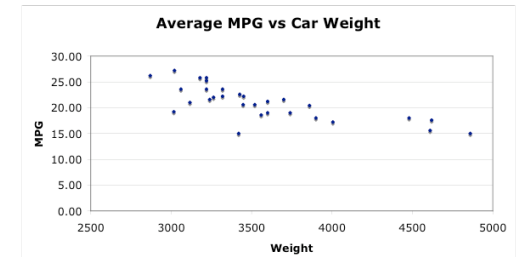
- I have a small sample of vehicles from the late 1990s - 30 cars and trucks
- Our dependent variable is AVGMPG - the average miles per gallon the vehicle gets while driving on the highway and city
- Our independent variable is the weight of the vehicle in pounds.
- It is important in regression to get a sense of the distribution, center, and spread of each variable in the model



21

Vehicle example

- The Covariance between these two variables is -1231.515
- The correlation is -.750, a much easier measure to interpret
- We can see a moderately strong negative relationship between the two variables
- The next step is to regress AVGMPG on vehicle WEIGHT



22

Regression of AVGMPG on WEIGHT

- The regression result is given below
- R^2 for the model is .563: 56.3% of the variability in AVGMPG is explained by the vehicle WEIGHT
- Note that this is the r^2
- The equation generated is **est AVGMPG = 38.3059 - .0049*WEIGHT**

Regression of AVGMPG on WEIGHT

Regression Statistics	
Multiple R	0.750
R Square	0.563
Adjusted R Square	0.547
Standard Error	2.234
Observations	30

ANOVA					
	df	SS	MS	F	Sig F
Regression	1	179.765	179.765	36.021	0.000
Residual	28	139.734	4.990		
Total	29	319.499			

	Coefficients	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	38.3059	2.9166	13.1339	0.0000	32.3316	44.2802
Weight	-0.0049	0.0008	-6.0018	0.0000	-0.0065	-0.0032

23

- When WEIGHT = zero, estimated AVGMPG = 38.3059
- The coefficient for WEIGHT is small (-.0049), mostly because the average WEIGHT is large (3562)
- When WEIGHT =
 - 3000 est AVGMPG = $38.3059 - .0049(3000) = 23.61$
 - 3500 est AVGMPG = $38.3059 - .0049(3500) = 21.16$
 - 4000 est AVGMPG = $38.3059 - .0049(4000) = 18.71$

Regression of AVGMPG on WEIGHT

Regression Statistics	
Multiple R	0.750
R Square	0.563
Adjusted R Square	0.547
Standard Error	2.234
Observations	30

ANOVA					
	df	SS	MS	F	Sig F
Regression	1	179.765	179.765	36.021	0.000
Residual	28	139.734	4.990		
Total	29	319.499			

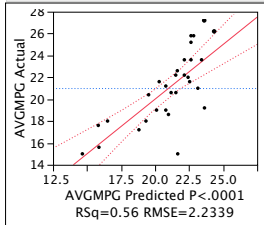
	Coefficients	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	38.3059	2.9166	13.1339	0.0000	32.3316	44.2802
Weight	-0.0049	0.0008	-6.0018	0.0000	-0.0065	-0.0032

24

JMP Results

Whole Model

Actual by Predicted Plot



Summary of Fit

RSquare	0.562647
RSquare Adj	0.547027
Root Mean Square Error	2.233941
Mean of Response	20.97333
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	179.76491	179.765	36.0215
Error	28	139.73376	4.990	Prob > F
C. Total	29	319.49867		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	38.30588	2.916555	13.13	<.0001*
Weight	-0.004866	0.000811	-6.00	<.0001*

25

Summary

- We walked through typical output in Excel and JMP
- I want you to be able to look at the output and see the relevant information to help make a decision
- The next lecture will focus on the inferential aspects of regression
 - The Overall F-test
 - The individual t-tests for the regression coefficients
- I also want to point out the strategy I want you to employ
 - Investigate the individual variables first
 - Look at the scatterplot
 - Run and interpret a model

26