Introduction to Regression

Dr Tom Ilvento
Department of Food and Resource Economics

Overview

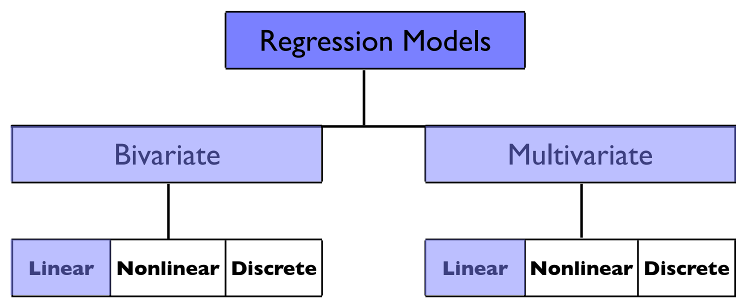
- The last part of the course will focus on Regression Analysis
- This is one of the more powerful statistical techniques
  - Provides estimates from **a model**
  - Allows for inference and testing hypotheses
  - Extends our abilities from ANOVA
  - Enable us to test theories
- We will start will simple, bivariate models: **Y is a function of a single X variable**
- And then move toward the complex, multivariate models: **Y is a function of a set of X variables**

Regression

- We are looking at the relationship between two or more variables
  - One is called the **Dependent Variable** (Y), which is to modeled or predicted
  - The others are called **Independent Variables** (X or a set of Xs), which are used to ***explain, estimate, or predict*** **Y**
- In a bivariate (two variable) case, one way to express the relationship is in terms of covariance and correlation:
  - Expressed as a linear measures of association
  - Symmetric measures
- Regression is an extension of correlation/covariance
  - Still linear
  - No longer symmetric
  - Covariance is the basic building block of regression

Regression Models

- Regression models represent an assortment of models with assumptions about the dependent variable - continuous or discrete - and the form of the relationship with independent variables - linear or nonlinear
- We will focus on the following



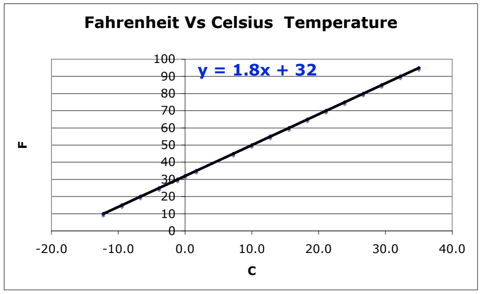
Fahrenheit versus Celsius

- When I go to Europe I have to deal with temperatures in Celsius
- How do I convert from C to F?
- ***A friend once told me a quick "rule of thumb" was to double C and add 30***
- I used my calculator to make a small data set of values
- And I used it in a regression

F	C
10	-12.22
15	-9.44
20	-6.67
25	-3.89
30	-1.11
32	0.00
35	1.67
45	7.22
50	10.00
55	12.78
60	15.56
65	18.33
70	21.11
75	23.89
80	26.67
85	29.44
90	32.22
95	35.00

Fitting a Line to the data

- The relationship between F and C is perfect, r = 1
- It is a deterministic function
- I will run a regression of F on C and see what equation I get.
- Regression will generate a "best fitting line" to the data
- In Excel I will use
  - **Tools, Data Analysis**
  - **Regression**



Regression of F on C

- This is the regression result from Excel
- The estimated equation is
  - **F = 32 + 1.8 C**

SUMMARY OUTPUT

Regression Statistics	
Multiple R	1
R Square	1
Adjusted R Square	1
Standard Error	7.0966E-05
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	12372.94	12372.94	2456783461497.83	0.00
Residual	16	0.00	0.00		
Total	17	12372.94			

	Coefficients	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	32.0	0.0	1519489.6	0.0	32.0	32.0
C	1.8	0.0	1567413.0	0.0	1.8	1.8

Requirements of Regression

- We specify one variable as **Dependent**
  - Usually represented as **Y**
  - It must be measured as a continuous variable – not a dichotomy or ordinal
- The dependent variable is thought to be a function of one or more **Independent** variables
  - Usually represented as **X**
  - Can be continuous, dichotomies, or ordinal
- Regression is limited to **Linear Relationships** in the parameters in the form of:
  - $Y = b_0 + b_1X_1 + b_2X_2 + \ldots b_kX_k$
  - We will will have **k** independent variables

Nonlinear relationships that can be represented by regression

- It is possible to represent a nonlinear relationship with a linear approach, such as a Polynomial or Log function
- Log function – take the log of both sides
  - **$Y = aX^b$**
  - **Ln(Y) = Ln(a)+ b*Ln(X)**
- Polynomial of the kth order
  - **$Y = b_0 + b_1X + b_2X^2 + b_3X^3 + \ldots b_kX^k$**
- It is not terribly restrictive to be limited to linear relationships

The equation of a Line

- I suspect you have seen the equation of a line written as
  - Y = mX + b
  - Where m is the slope and b is the intercept
- We specify a dependent variable Y, and independent variable X
- We will use the form $Y = b_0 + b_1X_1$
- Note: in multiple regression there may be more than one X: $Y = b_0 + b_1X_1 + b_2X_2$
- When referring to the population I will use Greek terms:

$$Y = \beta_0 + \beta_1 X_1$$

Equation of a Line

- **Y = 5 + .5X**
  - X=0 then Y=5 The intercept
  - X=10 then Y=10
  - X=20 then Y=15
  - X=30 then Y=20
- The slope shows how much Y changes for a unit change in X: **Y changes .5 for each 1 unit change in X**
- This is a **deterministic model** - there is an exact relationship between the two variables

In reality, we often have a random component

- A Probabilistic Model has a deterministic component and a random error component, denoted as e_i or ε_i

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_{i1}$$

- Our Expectation of Y is the deterministic component

$$E(Y_i) = \beta_0 + \beta_1 X_{i1}$$

The error term in our model

- The error component is very important
  - Observed in population/sample $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$
  - Predicted from model $\hat{Y}_i = \beta_o + \beta_1 X_{i1}$
  - The difference between what we predict and what we observe $\varepsilon_i = \hat{Y}_i - Y_i$

Have we seen the error term before?

- Consider the following model using the mean

$Y_i = \mu + \varepsilon_i$ A simple model based on the mean

$\varepsilon_i = Y_i - \mu$ Deviations about the mean

$\Sigma\varepsilon_i^2 = \Sigma(Y_i - \mu)^2$ Sum of Squared Deviations

$\Sigma\varepsilon_i^2 / n = \Sigma(Y_i - \mu)^2 / n$ Mean Squared Deviations

$\Sigma\varepsilon_i^2 / n = \sigma^2$ Population Variance

The error term in Regression is important!

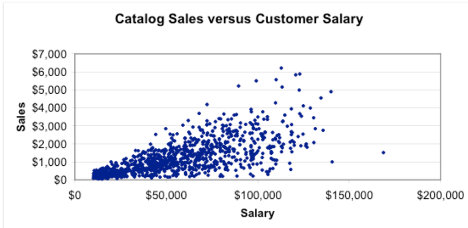
- The error term in regression is a measure of the:
  - **Variance of the Model**
  - **Standard Deviation of the Model**
  - And ultimately contributes to the estimate of the **Standard Error** for our coefficients
- We will assume equal variances for Y (dependent variable) across each level of X (independent variable)
- In essence we will **pool the measure of the variance** in regression
- This is called, **Homoscadasticity**

How do we fit a line to our data?

- We will use the property of **Least Squares**
- We will find estimates for β_0 and β_1 that will minimize the squared deviations about the fitted line
- First an example, and then the details
  - A catalog sales company which sells electronic equipment wants to improve its marketing campaign
  - They collect data on a random sample of 1,000 customers
  - The main variable of interest is the amount of sales (in dollars) in the previous year.

Catalog Sales data

- There is an Excel file (Catalogs.xls) and a JMP file (Catalogs.jmp)
- Y is the Dependent Variable: **SALES**
- X is the Independent Variable: **SALARY**
- The correlation between SALES and SALARY is **.700**
- Look at a Scatter Plot
- Excel will add a trendline and an equation and R-square which is based on regression



Estimated Regression of Sales on Salary

SALES = -15.332 + .022(SALARY)

- If SALARY = 0
  - SALES = -15.332 + .022(0)
  - SALES = **-15.332**
- A unit change in SALARY ($1) results in a **.022** change in SALES
  - **This is better expressed as: $1,000 change in SALARY results in Sales of $22.00**
- Our prediction of SALES for a household with a SALARY of $50,000 is:
  - **SALES = -15.332 + .022($50,000)**
  - **SALES = $1,084.67**
- **I will refer to this as solving the equation for a person with a salary of $50,000**

How to do this in Excel

- Organize data in columns
  - One column contains Y (dependent)
  - Remaining Columns contain contiguous Xs (independent)
- **TOOLS Data Analysis Regression**
  - Specify Y variable
  - Specify X variables – need to be contiguous columns (for more Xs in model, columns must be next to each other)
  - Remember to specify if first row has labels
  - Specify Output
- I modify the output
  - How many decimal places are showing (3 to 4)
  - Change Headings to make them fit
  - Bold Headers

Excel output

- **The correlation and R-square**
- **The ANOVA Table**
- **The estimated coefficients**

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

Output from JMP

Response SALES

Whole Model

Summary of Fit

RSquare	0.489434
RSquare Adj	0.488923
Root Mean Square Error	687.0649
Mean of Response	1216.77
Observations (or Sum Wgts)	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	451615197	451615197	956.6940
Error	998	471114029	472058.14	Prob > F
C. Total	999	922729225		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	634	345072568	544278	1.5718
Pure Error	364	126041460	346268	Prob > F
Total Error	998	471114029		<.0001*
				Max RSq
				0.8634

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-15.31783	45.37416	-0.34	0.7357
SALARY	0.0219608	0.00071	30.93	<.0001*

A few points about our model

- It is possible to predict outside the range of the data
  - When Salary = 0: SALES = -15.332 + .022($0) = **-$15.33**
  - When Salary = 1,000,000 SALES = -15.332 + .022($1,000,000) = **$21,985**
- **The model parameters should be interpreted only within the sampled range of the independent variables**
- The prediction part of our model is deterministic, but we know we will have some error – our prediction won't match the data exactly
  - We are fitting a model to the data
  - ***"All models are wrong, some models are useful"*** George Box
- We will have the ability to test coefficients and construct confidence intervals - there is a known sampling distribution for regression coefficients

How to generate a "Best Fitting Line"

- We will use the property of Least Squares
- We will find estimates for β_0 (intercept) and β_1 (slope) that will minimize the squared deviations about the fitted line
- 'Best Fit' means the **Difference Between Actual Y Values & Predicted Y Values Are a Minimum**
- Least Squares generates a set of coefficients that minimizes the Sum of the Squared Errors (SSE)

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \widehat{Y})^2 = \sum_{i=1}^{n}\widehat{\varepsilon}^2 = \text{minimum}$$

Bi-variate Regression Formulas for estimates of β_0 and β_1

- **I will tend to use b_0 and b_1 for the estimated values**

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1}$$

- **The slope coefficient is based on the covariance of Y and X, adjusted for the variability in X**

$$\widehat{\beta}_1 = \frac{SS_{XY}}{SS_X}$$

$$where\ SS_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}$$

$$SS_X = \sum (X_i - \overline{X})^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}$$

- **The Intercept is based on the estimate of b_1 and the means of the other variables**

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

Summary

- Regression is a strategy to model the relationship of a set of independent variables (Xs) on a dependent variable (Y).
- **We say we "regress" Y on X or as set of Xs.**
- Regression estimates a best fitting line to the data by minimizing squared deviations about that line.
- It is a natural extension of much of what we have covered before, especially ANOVA.
- We will cover the regression output, the ANOVA table, understanding the regression coefficients, inference in regression, and multiple regression.