Multiple Regression

Dr Tom Ilvento
Department of Food and Resource Economics
University of Delaware

Overview

- This lecture will expand our discussion on Regression by allowing more than one independent variable on the right hand side of the equation.
- Multiple Regression fits a model to a single dependent variable (Y) that is a function of more than one independent variable
- This allow for a richer, more in-depth model
    - We can test the effect of multiple variables on the dependent variable at the same time
    - While controlling for the effect of other variables
- This makes regression a very powerful tool, and also a tool open for exploitation
- We will also introduce **Standardized Coefficients**

Multiple Regression

- What makes regression really powerful is the ability to estimate models with many independent variables
- In this case we still estimate a linear equation which can be used for prediction.
- For a case with three independent variables we estimate:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3}$$

- Much of the output will look familiar
    - R^2
    - ANOVA Table
    - F-test
    - Coefficients and t-tests
- But now we will have several coefficients; one for each independent variable in the model
- The degrees of freedom for Regression will change accordingly

Multiple Regression

- In Multiple Regression, the interpretation of each coefficient is somewhat different
    - The slope coefficient for X_1 is now the change in Y for a unit change in X_1 holding all other independent variables constant.
    - We take into account the other independent variables when estimating the impact of X_1
    - By incorporating the covariance of X_1 with the other independent variables
- How this is done is more complicated: computed with simultaneous equations via Matrix algebra

Formulas for regression coefficients for two independent variables

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

- In terms of the equations
  - **V** stands for the variance
  - **C** stands for the covariance
- The Least Squares estimates are a function of
  - the variances of the independent variables,
  - the covariances of each independent variables with Y,
  - and covariances of the independent variables with each other

$$b_1 = \frac{(V_2C_{Y1} - C_{12}C_{Y2})}{(V_1V_2 - C_{12}^2)}$$

$$b_2 = \frac{(V_1C_{Y2} - C_{12}C_{Y1})}{(V_1V_2 - C_{12}^2)}$$

$$b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

Multiple Regression

- The ability to estimate the affect of an independent variable (X1) independent of the other independent variables in the model is a very powerful and compelling feature of regression
- It allows for **"statistical control"** as opposed to control via an experimental design
  - **Multiple regression estimates the unique effect** of each independent variable **by accounting for covariances** between independent variables
  - Hence it's popularity in the social sciences, medicine, nutrition, and business
- Compared to the bivariate regression, controlling for the other independent variables may:
  - **Increase** the strength of the relationship between an independent variable (X) and the dependent variable (Y)
  - **Decrease** the strength of the relationship
  - **Reverse** the sign (e.g., from positive to negative)
  - Or, **leave it relatively unchanged**

Collinearity in Regression

- In fact, if X_1 is uncorrelated with the other independent variables in the model, i.e., it is independent of the other Xs in the model,
- then the bivariate regression estimate of the β_1 will equal the multivariate regression estimate of β'_1
- If there is high correlation between X_1 and the other independent variables we will have a problem
  - **Collinearity** when X_1 highly correlated with one other independent variable
  - **Multi-collinearity** when X_1 is highly correlated with a set of independent variables
- Too much collinearity means we can't estimate the affect of X_1 very well
- **Extreme collinearity means the regression can't be estimated at all!**

Collinearity in Regression with Dummy Variables

- This is why we can't have all the levels represented in a model when dealing with dummy variables
- For example, if we have three levels of a categorical variable, we said we could represent this with 2 dummy variables
- The third level is referred to as the "reference" level or category and is captured in the intercept.
- **The reference level has a perfect linear relationship with the other two dummy variables and must be left out of the model**
- If we included all three dummy variables in the model, the software will warn you there is a problem

Collinearity and Standard Errors in Multiple Regression

- In bivariate regression, we established that the standard error for β_1 is a function of:
  - The Root Mean Squared Error for the model
  - The Sum of Squares for X

$$\text{Std Error for } b_1 = \frac{Root\ MSE}{\sqrt{SS_X}}$$

- **In multiple regression, the standard error is also a function of the covariance between the independent variables**
- We take into account how the independent variables are related to each other
  - If the correlation between independent variables is large, the standard errors will be inflated
  - The estimates of standard errors will no longer have minimum variance

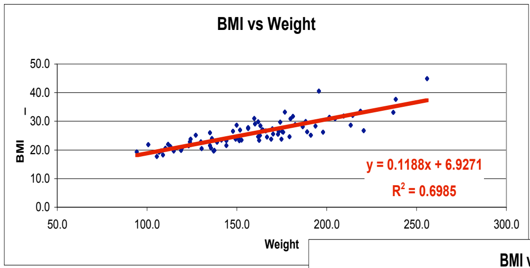
$$\text{var}(b_1) = \frac{\sigma^2}{n}\left[\frac{1/V_1}{1-r_{12}^2}\right]$$

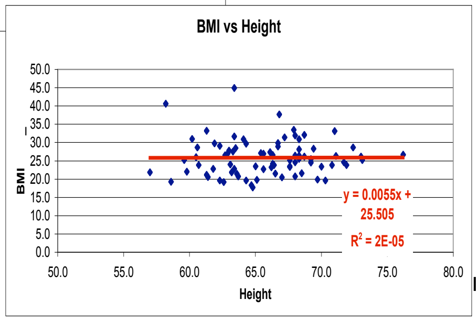
A quick example of a Multiple Regression

- **BMI** stand for Body Mass Index.
- It is a function of a person's weight (w in kg) and height (h in cm).
- In fact, the exact relationship is expressed as:
- I have a small data set with the subjects BMI, height and weight
- We will regress BMI on WEIGHT and HEIGHT of 80 random subjects

$$BMI = \frac{w}{h^2}$$

Look at the Scatterplots





- **The relationship of BMI and WEIGHT is strong and positive**
- **The relationship of BMI and HIEGHT is near zero**

Look what happens in Multiple Regression

- This is output from Excel
- Most things look the same: the only real difference is now we have a estimated coefficient for WEIGHT and HEIGHT
- Notice that:
  - R^2 is much larger than that for WEIGHT alone (.987)
  - The coefficient for WEIGHT is positive, significant and larger than the bivariate relationship
  - The coefficient for HEIGHT is negative and significant

Regression of BMI on WEIGHT and HEIGHT

Regression Statistics	
Multiple R	0.994
R Square	0.987
Adjusted R Square	0.987
Standard Error	0.563
Observations	80

ANOVA

	df	SS	MS	F	Sig F
Regression	2	1907.887	953.943	3009.595	0.000
Residual	77	24.406	0.317		
Total	79	1932.293			

	Coef	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	53.012	1.138	46.582	0.000	50.746	55.278
WEIGHT	0.370	0.005	77.583	0.000	0.361	0.380
HEIGHT	-32.309	0.770	-41.943	0.000	-33.843	-30.775

$$est\ \ BMI = 53.169 + .168 * \text{WEIGHT} - .823 * \text{HEIGHT}$$

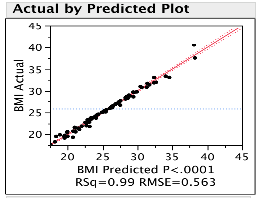
When we control for WEIGHT in the model, HIEGHT becomes negative and significant

Some thoughts on our model

- I actually prefer the JMP output: it gives me more information
- Notice the plot on top:
  - this shows the predicted values from the model in a Scatterplot of the actual values
  - This is a picture of R^2 = .987
- This example shows what can happen in a multiple regression versus bivariate regressions
  - Both coefficients are significant in the model
  - The coefficient for WEIGHT was strengthened from .119 to .370
  - The coefficient for HEIGHT changed sign; now it is negative

Regression of BMI on WEIGHT and HEIGHT

Actual by Predicted Plot



Summary of Fit

RSquare	0.987
RSquare Adj	0.987
Root Mean Square Error	0.563
Mean of Response	25.818
Observations (or Sum Wgts)	80.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1907.8869	953.943	3009.595
Error	77	24.4065	0.317	Prob > F
C. Total	79	1932.2934		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	53.012	1.138	46.58	<.0001*
WEIGHT	0.370	0.005	77.58	<.0001*
HEIGHT	-32.309	0.770	-41.94	<.0001*

Some thoughts on the model

- There is one more, very important point about this model
- Which speaks to a very important point about Multiple Regression
- **THE MODEL IS WRONG!!!!**
- Remember, I gave you the actual form of BMI

$$BMI = \frac{w}{h^2}$$

  - It is **deterministic**
  - It is **not a linear function**
- **The fact that I get a high R^2 and significant coefficients in the model does not make my model good, correct, or preferred!**
- The is no substitute for knowledge and background information from the researcher

Regression is a tool that can be easily misused

Requirements and Assumptions of Regression

- **Requirements of Regression**
  1. Y is measured as a continuous level variable – not a dichotomy or ordinal
     - The independent variables can be continuous, dichotomies, or ordinal
  2. The independent variables are not highly correlated with each other
  3. The number of independent variables is 1 less than n (preferably n is far greater than the number of independent variables)
  4. Same number of cases for each variable – any missing values for any variable in the regression removes that case from the analysis
- **Assumptions about the Error Term**
  - Mean of Probability Distribution of the Error term is zero
  - Probability Distribution of Error Has Constant Variance = 2
  - Probability Distribution of Error is Normal
  - Errors Are Independent – they are uncorrelated with each other

A Multivariate Example: The Value of Apartment Buildings

- This is a small data set of attributes of apartment buildings in a mid-sized city- a random sample of 25 apartments
- The sale price of the apartment building (**PRICE**) is seen as a function of
  - The number of apartments in the building **#APTS +**
  - The age of the apartment building **AGE -**
  - The lot size that the building is on **LOTSIZE +**
  - The number of parking spaces **PARKING +**
  - The total area is square footage **AREA +**
- A model of PRICE based on attributes would be useful for Real Estate appraisal purposes

MNApts.xls or MNApts.JMP are on the website

- Our Strategy for the analysis
  - Generate descriptive statistics
  - Examine the correlation matrix
  - Examine scatter plots of Price with key variables
  - Then begin building multivariate regression models
- Because PRICE is such a large number (several hundred thousand dollars), I will re-express it per $1,000.
  - This will not change any of the essential results - scatterplots, correlations, and regression results will be the same
  - But the Sums of Squares will not be so huge

Descriptive Statistics

	PRICE	#APTS	AGE	LOTSIZE	PARKING	AREA
Mean	290.57	12.16	52.92	8554.12	2.52	11423.40
Standard Error	42.31	2.52	5.18	839.86	0.99	2003.87
Median	268.00	8.00	62.00	7425.00	0.00	7881.00
Mode	#N/A	4.00	82.00	#N/A	0.00	#N/A
Standard Deviation	211.53	12.58	25.89	4199.30	4.93	10019.35
Sample Variance	44744.58	158.31	670.49	17634110.11	24.34	100387322.33
Coef. Variation	72.8%	103.5%	48.9%	49.1%	195.8%	87.7%
Kurtosis	2.80	10.04	-1.40	2.33	6.28	2.19
Skewness	1.61	2.84	-0.48	1.52	2.44	1.71
Range	870.70	58.00	72.00	16635.00	20.00	36408.00
Minimum	79.30	4.00	10.00	4365.00	0.00	3040.00
Maximum	950.00	62.00	82.00	21000.00	20.00	39448.00
Sum	7264.34	304.00	1323.00	213853.00	63.00	285585.00
Count	25	25	25	25	25	25

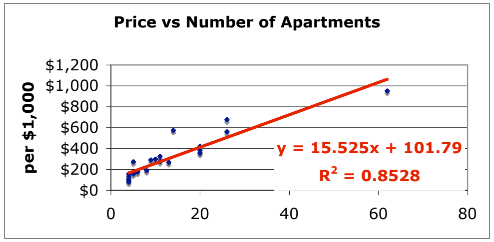
- The mean PRICE is 290.57 ($290,570), slightly higher than the median
  - The CV for PRICE is large - 72.8% - which reflects a lot of variability in the price of the apartments
- The means of the independent variables are
  - 12.16 for #APTS
  - 52.92 years for AGE
  - 8554.12 sq ft for LOTSIZE
  - 2.52 parking spaces for PARKING (some with no parking)
  - 11,423.40 sq ft for AREA

The Correlation Matrix

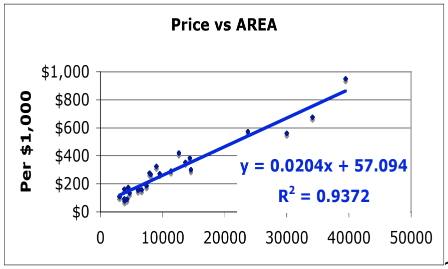
	PRICE	#APTS	AGE	LOTSIZE	PARKING	AREA
PRICE	1.000					
#APTS	0.923	1.000				
AGE	-0.114	-0.014	1.000			
LOTSIZE	0.742	0.800	-0.191	1.000		
PARKING	0.225	0.224	-0.363	0.167	1.000	
AREA	0.968	0.878	0.027	0.673	0.089	1.000

- High correlations of Price with
  - **No. Apts (.923),**
  - **Lot Size (.742),**
  - **Area (.968)**
- **Age is negatively related to Price (-.114)**
- High correlations between independent variables for
  - **No. Apts with Lot Size (.80) and Area (.878)**
  - **Lot Size and Area (.673)**

Scatterplots

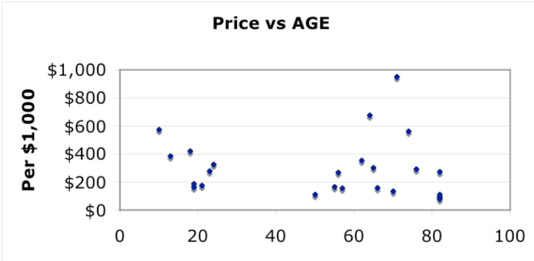


- PRICE vs #APTS shows a strong, positive linear relationship: each apartment brings $15.525 dollars in value per $1,000, or $15,525.00
- PRICE va AREA shows a strong, positive, linear relationship. Each square foot brings .0204 in price per $1,000, or $20.40



Scatterplot of PRICE and AGE

- The relationship of PRICE with AGE is weak
- It is hard to see that it is in fact, negative!
- Based on the regression line, for each year of age the apartment loses $.9353 in value per $1,000 or $935.30



Look at the Regression results of PRICE on AGE

Regression of PRICE on AGE

Regression Statistics	
Multiple R	0.114
R Square	0.013
Adjusted R Square	-0.030
Standard Error	214.658
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	14076.534	14076.534	0.305	0.586
Residual	23	1059793.413	46077.974		
Total	24	1073869.947			

	Coef	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	340.069	99.309	3.424	0.002	134.633	545.505
AGE	-0.935	1.692	-0.553	0.586	-4.436	2.565

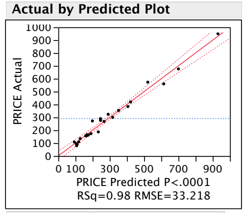
- R^2 is very low (**.013**)
- The F-test is not significant: **F* = .305 and the p value = .586**
- The t-test for AGE is not significant: **t* = -.553, p =.586**
- **Even if we decided the test for AGE was a one-tailed test, the p-value would be .293 (1/2 of .586)**
- Even though our regression equation predicts a negative relationship between PRICE and AGE, we can't tell if it is any different from zero

Look at the Multiple Regression

- I used JMP to estimate the Regression of PRICE on #APTS, AREA, LOTSIZE, PARKING, and AGE
- R^2 is quite high, .980; **98% of the variability in PRICE is explained by our model**
- The Sums of Squares are quite large (it was good idea to use per $1,000)
- LOTSIZE and PARKING are not significant at α = .05 for a two-tailed test, but other coefficients are significant
- **In the Multiple Regression**
  - The coefficient for #APTS dropped considerably compared with the bivariate regression (4.140 vs 15.525)
  - The coefficient for AREA dropped a little (.016 vs .020)
  - The coefficient for AGE is now significant

Response PRICE

Actual by Predicted Plot



Summary of Fit

RSquare	0.980
RSquare Adj	0.975
Root Mean Square Error	33.218
Mean of Response	290.574
Observations (or Sum Wgts)	25.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	1052904.6	210581	190.841
Error	19	20965.3	1103	**Prob > F**
C. Total	24	1073869.9		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	92.789	28.688	3.23	0.0044*
#APTS	4.140	1.490	2.78	0.0120*
AREA	0.016	0.001	10.63	<.0001*
LOTSIZE	0.001	0.003	0.34	0.7411
PARKING	2.696	1.577	1.71	0.1036
AGE	−0.853	0.299	−2.86	0.0101*

Regression Results

est. PRICE = 92.789 + 4.140*#APTS +.016*AREA + .001*LOTSIZE + 2.696*PARKING - .853*AGE

- All coefficients are in the expected direction
- If we want to find the predicted value of an apartment building with
  - 20 apartments
  - 22,000 sq ft of area
  - 2,000 sq ft of lot size
  - 20 parking spaces
  - 50 years old

est. PRICE = 92.789 + 4.140*20 +.016*22,000 + .001*2,000 + 2.696*20 - .853*50

- **est PRICE = $540.859 per $1,000 or $540,859**

The Hypothesis Test for AGE

- **Ho:** Ho: $\beta_{AGE} = 0$
- **Ha:** Ha: $\beta_{AGE} < 0$
- **Assumptions** Equal variances, normal distribution
- **Test Statistic** t* = (-.853-0)/.299 = -2.86 p = .005
- **Conclusion:** p = .005 we can reject at α = .01
  Reject Ho: $\beta_{AGE} = 0$

In the Multiple Regression, when we control for other variables in the model, the coefficient for AGE is now significant

Controlling for other variables in the model helped us to better estimate the unique effect of AGE on PRICE

How can we tell which coefficient is most important in the model?

- The coefficients reflect the metric of each independent variable
- The are also referred to as unstandardized coefficients
- This makes it difficult to determine which variable has the most influence
- We can create a **Standardized Coefficient**, **b'**, which will be standardized between -1 and 1
- b' is equal to the unstandardized coefficient times the ratio of the standard deviation of x to the standard deviation of y

$$b_i' = b_i * \frac{s_{x_i}}{s_y}$$

Standardized Coefficients

$$b_i' = b_i * \frac{s_{x_i}}{s_y}$$

- Standardized coefficients transform the coefficients as if each variable has a mean of zero and a standard deviation of 1 (like a z-score)
- The interpretation for b' is how many standard deviations predicted Y changes, with a 1 standard deviation change in X (holding all other variables constant)
- The are analogous to a correlation coefficient – the theoretical range is –1 to 1
- We can compare the strength of the relationship by looking at the relative size of the standardized coefficients

Standardized Coefficients

Variable	Coef	Std Dev	Sx/Sy	Std Coef
PRICE		211.529		
#APTS	4.140	12.582	0.059	**0.246**
AREA	0.016	10019.347	47.366	**0.736**
LOTSIZE	0.001	4199.299	19.852	**0.019**
PARKING	2.696	4.934	0.023	**0.063**
AGE	-0.853	25.894	0.122	**-0.104**

- **The variable with the most impact on PRICE is AREA (.736), followed by #APTS (.246) and then much lower, AGE (-.104)**
- Be careful with standardized coefficients!
- They should not be used with predictions
- They are sample specific – when we want to make an inference via a significance test, use the unstandardized coefficients

Software and Standardized Coefficients

- JMP will generate the Standardized Coefficients if you request it

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	92.789	28.688	3.23	0.0044*	0.000
#APTS	4.140	1.490	2.78	0.0120*	0.246
AREA	0.016	0.001	10.63	<.0001*	0.736
LOTSIZE	0.001	0.003	0.34	0.7411	0.019
PARKING	2.696	1.577	1.71	0.1036	0.063
AGE	-0.853	0.299	-2.86	0.0101*	-0.104

Summary

- Now we have more than one independent variable in the model: Multiple Regression stands for multiple independent variables
- Many aspects of the output remains the same and should be familiar to you
- But the estimates and standard errors of the coefficients take into account what is in the model
- The interpretation of each regression coefficient for each variable is its effect on the dependent variable, ***holding constant all other variables in the model***
- Controlling for other variables in the model can change things!