

## Regression IV: Multiple Regression

Tom Ilvento  
FREC 408

### 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:

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

### Multiple Regression

- But 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
  - Computed with simultaneous equations via Matrix algebra – better left to computers

### Multiple Regression

- The ability to estimate the affect of an independent variable ( $X_1$ ) **independent** of the other independent variables in the model is a very powerful and compelling feature of regression
- It allows to use "statistical control" as opposed to control via an experimental design
- Hence it's popularity in the social sciences, medicine, nutrition

### Multiple Regression

- 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)
  - Leave it relatively unchanged

### Multiple Regression

- If  $X_1$  is uncorrelated with the other independent variables in the model, i.e., it is independent of the other  $X$ s in the model,
- then the bivariate regression estimate of the  $\beta_1$  will equal the multivariate regression estimate of  $\beta_1$

## Multiple Regression

- 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

- 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

## Requirements of Regression

- Y is measured as a continuous level variable – not a dichotomy or ordinal
  - The independent variables can be continuous, dichotomies, or ordinal
- The independent variables are not highly correlated with each other
- The number of independent variables is 1 less than n (preferably n is far greater than the number of independent variables)
- Same number of cases for each variable – any missing values for any variable in the regression removes that case from the analysis

## Linear Regression Assumptions about the error term

- Mean of Probability Distribution of the Error term is zero
- Probability Distribution of Error Has Constant Variance =  $\sigma^2$
- Probability Distribution of Error is Normal
- Errors Are Independent – they are uncorrelated with each other

## A Multivariate Example

- Apartment sales in Minnesota
- The value of the apartment building (price) is seen as a function of
  - The number of apartments in the building
  - The age of the apartment building
  - The lot size that the building is on
  - The number of parking spaces
  - The total area is square footage
  - Condition of the building
- We have a random sample of 25 apartment buildings to estimate a model

## MN Apartment Sales Example - with five independent variables

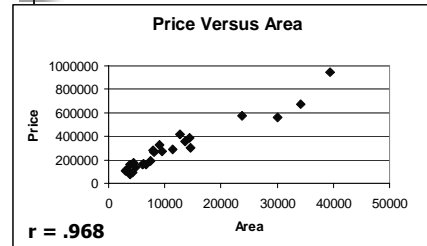
	Price	No. Apts	Age	Lot Size	Parking	Area
Mean	290573.52	12.16	52.92	8554.12	2.52	11423.40
Standard Error	42305.83	2.52	5.18	839.86	0.99	2003.87
Median	268000.00	8.00	62.00	7425.00	0.00	7881.00
Mode	#N/A	4.00	82.00	#N/A	0.00	#N/A
Standard Deviation	211529.15	12.58	25.89	4199.30	4.93	10019.35
Sample Variance	44744581138.09	158.31	670.49	17634110.11	24.34	100387322.33
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	870700.00	58.00	72.00	18635.00	20.00	36408.00
Minimum	79300.00	4.00	10.00	4365.00	0.00	3040.00
Maximum	950000.00	62.00	82.00	21000.00	20.00	39448.00
Sum	7264338.00	304.00	1323.00	213853.00	63.00	285585.00
Count	25.00	25.00	25.00	25.00	25.00	25.00
Confidence Level(95.0%)	87314.92	5.19	10.69	1733.38	2.04	4135.78

## MN Apartment Sales Example

### Correlation Matrix

	Price	No. Apts	Age	Lot Size	Parking	Area
Price	1.000					
No. Apts	0.923	1.000				
Age	-0.114	-0.014	1.000			
Lot Size	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

## MN Apartment Sales Example



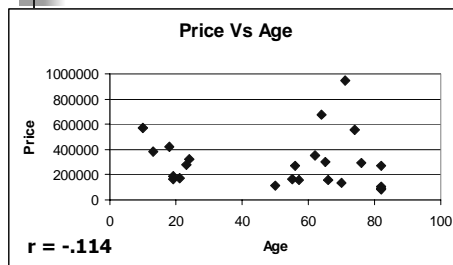
## MN Apartment Sales Example

- Bi-variate regression of PRICE on AREA

	Coefficients	Standard Error	t Stat	P-value
Intercept	57093.707	16612.173	3.437	0.002
Area	20.439	1.103	18.532	0.000

- $R^2 = .937 = r^2 = (.968)^2$
- $\hat{Y} = \$57,093.71 + \$20.44(\text{AREA})$

## MN Apartment Sales Example



## MN Apartment Sales Example

- Bi-variate regression of PRICE on AGE

	Coefficients	Standard Error	t Stat	P-value
Intercept	340068.92	99308.86	3.42	0.00
Age	-935.29	1692.17	-0.55	0.59

- $R^2 = .01 = r^2 = (-.114)^2$
- $\hat{Y} = \$340,068.92 - \$935.29(\text{AGE})$

## Hypothesis Test for a slope coefficient for AGE,

- Null hypothesis:  $H_0: \beta_{\text{AGE}} = 0$
- Alternative hypothesis:  $H_a: \beta_{\text{AGE}} \neq 0$  **two-tailed test**
- Assumptions: Large sample, normal
- Test Statistic:  $t^* = (-935.29 - 0)/1692.17$
- Calculation:  $t^* = -.55$
- P-value:  $P = .59$
- Conclusion: Cannot Reject  $H_0: \beta_{\text{AGE}} = 0$

While the model shows AGE has a negative influence on value (each year of age decreases value by  $-\$935$ ), I can't be sure that the real value is different from zero.

## Multiple Regression Output with five independent variables

SUMMARY OUTPUT					
<b>Regression Statistics</b>					
Multiple R	0.990				
R Square	0.980				
Adjusted R Square	0.975				
Standard Error	33217.938				
Observations	25.000				
<b>ANOVA</b>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	5.000	1052904750916.020	210580950183.205	190.842	0.000
Residual	19.000	20965196398.216	1103431389.380		
Total	24.000	1073869947314.240			

## MN Apartment Sales Example

- Multivariate Regression – many independent variables

	<i>Coef</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	92787.87	28688.252	3.234	0.004
# Apts	4140.42	1489.789	2.779	0.012
Age	-853.18	298.766	-2.856	0.010
Lot Size	0.96	2.869	0.335	0.741
Parking	2695.87	1576.742	1.710	0.104
Area	15.54	1.462	10.631	0.000

$$R^2 = .98$$

## MN Apartment Sales

- Predicted Regression Equation

$$\hat{Y} = \$92,788 + \$4,140(\text{NUMBER}) - \$853(\text{AGE}) + \$1(\text{LOTSIZE}) + \$2,696(\text{PARK}) + \$16(\text{AREA})$$

## MN Apartment Sales

- We want to find the predicted value of an apartment building with
  - 20 apartments
  - 50 years old
  - 2,000 sq ft of lot size
  - 20 parking spaces
  - 22,000 sq ft of area

## MN Apartment Sales

$$\begin{aligned} \text{Predicted Value} &= \$92,788 + 4,140(20) \\ &\quad - \$853(50) + \$1(2000) + \$2,696(20) \\ &\quad + \$16(22,000) \\ &= \mathbf{\$540,858} \end{aligned}$$

## MN Apartment Sales Example

- Focus on the coefficient for AGE

	<i>Coef</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	92787.87	28688.252	3.234	0.004
# Apts	4140.42	1489.789	2.779	0.012
<b>Age</b>	<b>-853.18</b>	<b>298.766</b>	<b>-2.856</b>	<b>0.010</b>
Lot Size	0.96	2.869	0.335	0.741
Parking	2695.87	1576.742	1.710	0.104
Area	15.54	1.462	10.631	0.000

$$R^2 = .98$$

### Hypothesis Test for a slope coefficient for AGE,

- Null hypothesis    ■  $H_0: \beta_2 = 0$
- Alternative        ■  $H_a: \beta_2 \neq 0$     two-tailed test
- Assumptions      ■ Large sample, normal
- Test Statistic    ■  $t^* = (-853.184 - 0)/298.766$
- Calculation       ■  $t^* = -2.856$
- P-value            ■  $P = .01$
- Conclusion        ■ Reject  $H_0: \beta_2 = 0$

Age of the apartment has a negative influence on its value. For each year of age the value decreases by -\$853

### MN Apartment Sales Example (continued)

- We can also include dummy variables into the multiple regression model
- Let's see how our model changes when we add dummy variables

### New Correlation Matrix

	Price	# Apts	Age	Lot Size	Park Sps	Area	Excellent	Good
Price	1.000							
# Apts	0.923	1.000						
Age	-0.114	-0.014	1.000					
Kot Size	0.742	0.800	-0.191	1.000				
Park Sps	0.225	0.224	-0.363	0.167	1.000			
Area	0.968	0.878	0.027	0.673	0.089	1.000		
Excellent	0.162	-0.083	0.153	-0.202	-0.177	0.199	1.000	
Good	0.082	0.175	-0.447	0.273	0.112	0.020	-0.634	1.000

- Note the following:
- Excellent and Good are positively related to Price
  - Good is negatively related to Age????

### MN Apartment Sales

Regression Statistics	
Multiple R	0.994
R Square	0.988
Adjusted R Square	0.982
Standard Error	28027.252
Observations	25

- With two new variables added into the model,  $R^2$  changes a bit, increasing from 0.98 to 0.988, very close to 1.

Overall, the exploratory variables help explain the variability in Y. Our model is very good.

### New Predicted Regression Equation:

$$\hat{Y} = \$72,827 + \$5,644(\# \text{ APTS}) - \$838(\text{AGE}) + \$2(\text{LOTSIZE}) + \$2,919(\text{PARKING}) + \$12(\text{AREA}) + \$56,022(\text{Excellent}) + \$7,134(\text{Good})$$

Compare to the old five variable model

$$\hat{Y} = \$92,788 + \$4,140(\# \text{ APTS}) - \$853(\text{AGE}) + \$1(\text{LOTSIZE}) + \$2,696(\text{PARKING}) + \$16(\text{AREA})$$

### Regression Output: ANOVA Table

ANOVA	df	SS	MS	F	Sig F
Regression	7	1060515990828.65	151502284404.09	192.87	0.00
Residual	17	13353956485.59	78526852.09		
Total	24	1073869947314.24			

### F Test for the estimated model

- Null hypothesis     ■  $H_0: \beta_1 = \beta_2 = \beta_3 \dots = \beta_k = 0$
- Alternative         ■  $H_a$ : At least one coefficient differs from 0
- Assumptions       ■ Equal variances, normal distribution
- Test Statistic     ■  $F^* = 192.867$
- Rejection Region ■  $F_{.05, 7, 17 \text{ d.f.}} = 2.61$
- Conclusion         ■  $F^* > F$
- Reject  $H_0: \beta_1 = \beta_2 = \beta_3 \dots = \beta_k = 0$

### The t-tests

	Coeff	Std Error	t Stat	P-value
Intercept	72827.75	31914.969	2.282	0.036
# Apts	5644.48	1351.794	4.176	0.001
Age	-838.54	291.773	-2.874	0.011
Lot Size	2.38	2.466	0.967	0.347
Park Sps	2919.17	1366.992	2.135	0.048
Area	12.98	1.484	8.750	0.000
Excellent	56022.90	20892.059	2.682	0.016
Good	7134.14	17315.472	0.412	0.685

The t-tests for *Excellent* and *Good* represent a test if each are significantly different from Fair

### Hypothesis Test for a slope coefficient for Excellent

- Null hypothesis     ■  $H_0: \beta_1 = 0$
- Alternative         ■  $H_a: \beta_1 \neq 0$      two-tailed test
- Assumptions       ■ small sample, normal
- Test Statistic     ■  $t^* = (56022)/20892$
- Calculation        ■  $t^* = 2.682$
- P-value             ■  $P = .016$
- Conclusion         ■ Reject  $H_0: \beta_1 = 0$

### Comparing models

- Recall in the previous model which only includes two dummy variables *Excellent* and *Good* as exploratory variables,
- We couldn't reject  $H_0: \beta_1 = 0$

	Coeff	Std Error	t Stat	P-value
Intercept	176940.00	94618.185	1.870	0.075
Excellent	173310.00	128113.628	1.353	0.190
Good	128641.29	110226.850	1.167	0.256

### MN Apartment Sales

- We want to find the predicted value of an apartment building with
  - 20 apartments
  - 50 years old
  - 2,000 sq ft of lot size
  - 20 parking spaces
  - 22,000 sq ft of area
  - In Excellent Condition

### Solve the Equation

$$\begin{aligned} \text{Predicted Value} &= \$72,828 + \$5644(20) \\ &\quad - \$839(50) + \$2(2000) + \$2,919(20) \\ &\quad + \$13(22,000) + \$56,023(1) \\ &= \mathbf{\$548,161} \end{aligned}$$

Predicted in Fair condition : **\$492,138**

**Our old model predicted \$540,858**

### Comparing models

- When more exploratory variables were involved in the model, our conclusions changed.
- In our new model which includes all variables, we were able to reject  $H_0: \beta_1 = 0$  for Excellent
- However, we don't have the same conclusion for Good
- Most of the coefficients changed in the new model, though most only slightly

### What is the meaning of the dummy variables and the dummy variable tests?

- The coefficients
  - If the building is considered in **Excellent** condition, there is an increase of \$56,023 over a building that is considered **Fair**
  - If the building is considered in **Good** condition, there is an \$7,134 premium over a building in **Fair** condition
  - However, I can't be sure if the premium for **Good** is really different from zero
- The tests represent whether **Excellent** and **Good** are different from **Fair**

### Comparing models - conclusion

- Models with different amounts of exploratory variables can affect our conclusions like "rejecting  $H_0: \beta_k = 0$  or not"
- But  $R^2$  will always go up when more exploratory variables are included in the model
- That is why we use adjusted  $R^2$  as a way to compare models

### The downside of statistical control

- Our model should be based on sound theoretical beliefs about the variables that influence the dependent variable
- Other variables, left out of the model, could also influence the dependent variable
- Left out variables might change the results
- One can never be sure!