Regression Analysis

BU609-3

Objectives

- Understand reasoning behind regression models
- Be able to use Excel to run multiple regressions models
- Understand and interpret computer output
- Understand limitations of regression

Some Business Applications

- How do employee wages relate to experience, education, gender, … ?
- Does a stock's current price depend on past values, as well as the current and past market index value?
- To what extent do current sales levels depend on current & past advertising levels, advertising levels of competitors, past sales levels, and the general market level?
- How does the unit cost of producing an item depend on the total quantity produced?

Simple Linear Regression

Fitting a straight line to sample data

Introduction

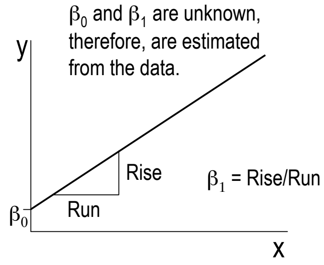
- Modeling technique for relationships
  - **can't show causality, just a relationship**
- Technique also used to predict value of one variable (dependent variable, y) based on value of other variables (independent variables $x_1, x_2, \ldots x_k$)
- Scatterplot is always best way to start

First Order Linear Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y = dependent variable
- x = independent variable
- β_0 = y-intercept
- β_1 = slope of the line
- ε = error variable

  "everything else"



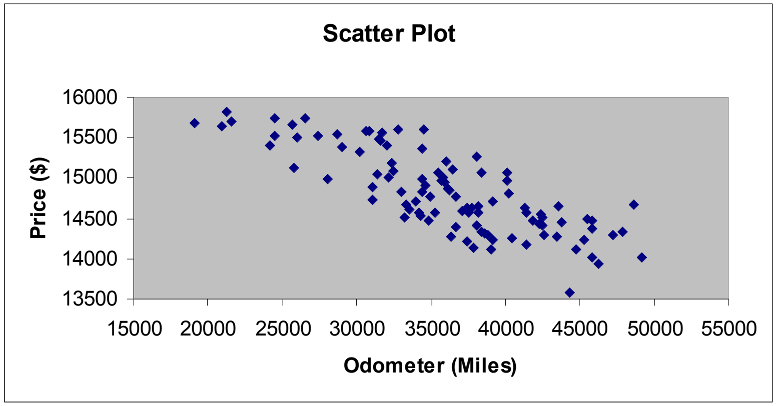
β_0 and β_1 are unknown, therefore, are estimated from the data.

Estimating Model Coefficients

- Estimates are determined by
  - drawing a sample from population of interest,
  - producing a straight line that cuts into data points, and which minimizes sum of squared vertical differences between all points and the line
- Resulting equation is

$$\hat{y} = b_0 + b_1 x$$

Scatterplot - Used Car Price vs Odometer



Always a great way to start looking at your data

Simple Linear Regression Line

- Tools > Data analysis > Regression > [Shade y range & x range] > OK

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted F	0.6466
Standard E	303.1
Observatio	100

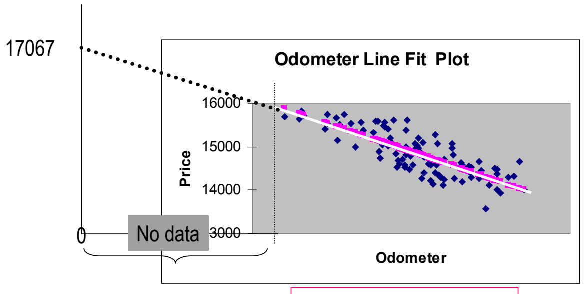
$$\hat{y} = 17,067 - .0623x$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	16734111	16734111	182.11	0.0000
Residual	98	9005450	91892		
Total	99	25739561			

	Coefficients	*Standard Errc*	*t Stat*	*P-value*
Intercept	17067	169	100.97	0.0000
Odometer	-0.0623	0.0046	-13.49	0.0000

Interpreting Results



$$\hat{y} = 17,067 \boxed{-} .0623x$$

The intercept is b_0 = $17067.

This is the slope of the line.
For each additional mile on the odometer,
the price decreases by an average of $0.0623

Do not interpret the intercept as the
“Price of cars that have not been driven”

Error Variable: Required Conditions

- Error ε is a critical part of regression model
- Four requirements involving distribution of ε ***must*** be satisfied:
  - Probability distribution of ε is Normal
  - Mean of ε is zero: E(ε) = 0
  - Std dev of ε is σ_ε for **all** values of x
    - » Minor deviation not a problem
  - Set of errors associated with different values of y are all independent
    - » Usually only a time series problem

Assessing Regression Model

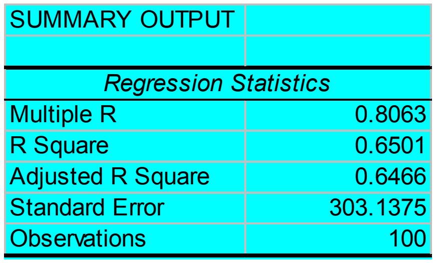
- Least squares method will produce a regression line whether or not there is a linear relationship between x and y
- Consequently, it is important to assess how well linear model fits data
- Several methods are used to assess model:
  - Testing and/or estimating coefficients
  - Using various descriptive measurements

Standard Error of Estimate

- Assumption: Mean error is equal to zero
  - Error = residual = observed - predicted
- If σ_ε is small, errors tend to be close to zero (close to mean error); model fits data well
- So, can use σ_ε as a measure of suitability of using linear model
- Unbiased estimator of σ_ε^2 is given by s_ε^2 (Standard Error in output)

Car Example (cont.)

- What does standard error of estimate for previous example say about model fit?



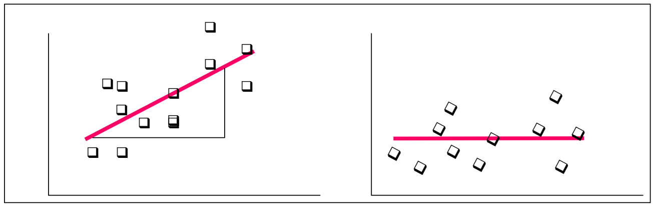
SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted R Square	0.6466
Standard Error	303.1375
Observations	100

Hard to assess model based on s_ε even when compared with mean value of y

$s_\varepsilon = 303.14, \ \bar{y} = 14{,}822.82$

Testing the slope

- When no linear relationship exists between two variables, regression line should be horizontal



Linear relationship. Different inputs (x) yield different outputs (y).

The slope is not equal to zero

No linear relationship. Different inputs (x) yield the same output (y).

The slope is equal to zero

(cont.)

- Can draw inference about β_1 from b_1 by testing

  H_0: $\beta_1 = 0$

  H_a: $\beta_1 \neq 0$ (or < 0, or > 0)

  - Test statistic is

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \quad \text{where} \quad s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

Standard error of b_1

  - If error variable is normally distributed, statistic is Student t dist. with d.f. = n-2
  - p-value given in Excel output

Car Example (cont.)

- Solution
  - From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	17,066.77	169.02	100.97	7.2785E-101
Odometer	-0.0623	0.0046	-13.49	4.44346E-24

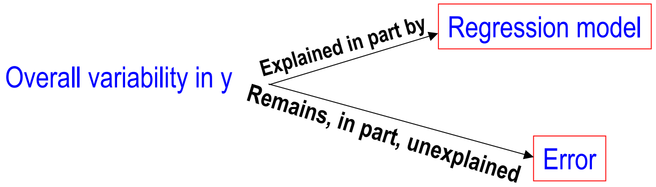
Overwhelming evidence to infer odometer reading affects auction selling price

Correlation Coefficient (R)

- Correlation Coefficient R gives **strength of association** between x & y (or model & y); also gives **direction of association**
  - Regression s/w only shows + value!!
  - Scatterplot or Correlation Analysis gives sign
- Recall correlation coeff range: $-1 \leq r \leq 1$
  - If r = -1 (negative association) or r = +1 (positive association) every point falls on the regression line
  - If r = 0 there is no linear pattern
- To test for linear relationship between two variables, test for $\beta_1 = 0$

Coeff. of Determination (R^2)

- Amt of variation explained by model



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	16,734,110.88	16,734,110.88	182.11	4.44346E-24
Residual	98	9,005,449.88	91,892.35		
Total	99	25,739,560.76			

Adj. R^2: adjusts for d.f.; penalizes unnecessarily complex model

Testing Full Model (ANOVA)

H_0: All $\beta = 0$;

H_a: at least one $\beta \neq 0$

Test Statistic: F statistic used

p-value: given in computer output (significance)

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	16,734,110.88	16,734,110.88	182.11	4.44346E-24
Residual	98	9,005,449.88	91,892.35		
Total	99	25,739,560.76			

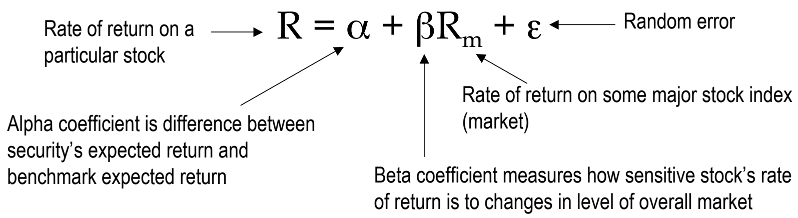
Using Regression Equation

- Before using regression model, need to assess how well it fits data
- If satisfied with how well model fits data, can use it to make predictions for y
- Illustration
  - Predict the selling price of a three-year-old Ford with 40,000 miles on the odometer (prev. example)

$$\hat{y} = 17{,}067 - .0623x = 17{,}067 - .0623(40{,}000) = 14{,}575$$

Finance Application: Market Model

- One important application of linear regression is the ***market model***
- Assume rate of return on a stock (R) is linearly related to rate of return on overall market



$$R = \alpha + \beta R_m + \varepsilon$$

Rate of return on a particular stock
Random error
Rate of return on some major stock index (market)
Alpha coefficient is difference between security's expected return and benchmark expected return
Beta coefficient measures how sensitive stock's rate of return is to changes in level of overall market

Market Model Alpha & Beta

- Alpha measures how well security performed on a risk-adjusted basis
  - >0: security did better than benchmark
  - <0: security did worse than benchmark
- Beta is a measure of sensitivity of security return to market
  - >1.0: aggressive security
  - <1.0: defensive security
- Can consider market index & one stock, or market index & portfolio

Market Model & Risk Analysis

- Market model provides useful insights into analyzing risk-return characteristics of a portfolio. From the market model, can determine the alpha, beta, and residual risk:
  - Alpha: measure of how large (small) "abnormal" return is
  - Beta: measure of how large market risk is (**market-related** or **systematic risk**)
  - Coeff. of Determination: measures proportion of total risk that is market related; remainder is **firm-specific (nonsystematic)**
  - Residual risk (epsilon): risk unrelated to market

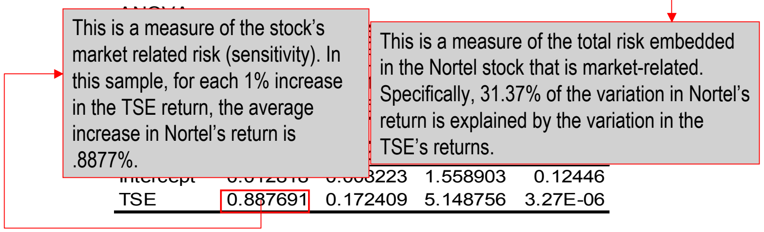
Example: Market Model

- Estimate the market model for Nortel, a stock traded on the Toronto Stock Exchange
- Data consisted of monthly percentage return for Nortel and monthly percentage return for all the stocks

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.560079
R Square	0.313688
Adjusted F	0.301855
Standard E	0.063123
Observatio	60

ANOVA



This is a measure of the stock's market related risk (sensitivity). In this sample, for each 1% increase in the TSE return, the average increase in Nortel's return is .8877%.

This is a measure of the total risk embedded in the Nortel stock that is market-related. Specifically, 31.37% of the variation in Nortel's return is explained by the variation in the TSE's returns.

Intercept	0.012818	0.008223	1.558903	0.12446
TSE	0.887691	0.172409	5.148756	3.27E-06

Example: Nortel & Royal Bank

- We'll use our data from 609-1
  - TSX Data Regression
- Which stock is more sensitive to changes in the market index?
  - Compare betas
- Which stock has the larger firm-specific risk?
  - Compare complement of R^2 (Coeff. of Det.)

(NB: Excel doesn't like missing data)

Regression Diagnostics

- Recall conditions required for validity of regression analysis:
  - error variable is normally distributed with mean zero
  - error variance is constant for all values of x
  - errors are independent of each other (watch with **time series**)
- How can we diagnose violations of these conditions?

Residual Analysis

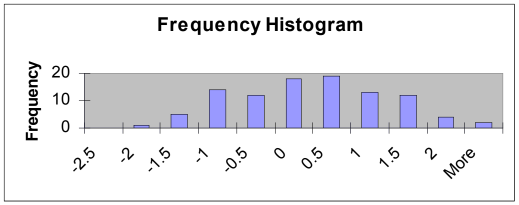
- Examining residuals (or standardized residuals), can identify violations of required conditions
  - Residual = error = observed - predicted
  - Standardized residual = Residual / Std Error
  - Testing Normality requirement
    - » Use Excel to obtain standardized residual histogram (all within +/- 3 std dev)
    - » Examine histogram and look for a bell shape with mean close to zero

(cont.)

RESIDUAL OUTPUT

Standardized residual i = Residual i / Standard error

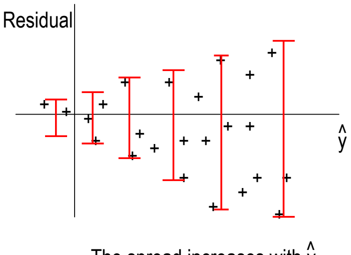
Observation	*Predicted Price*	*Residuals*	*Standard Residuals*
1	14736.915	-100.9149985	-0.334595895
2	14277.64993	-155.6499296	-0.516076186
3	14210.66079	-194.6607914	-0.645421421
4	15143.5858	446.4141955	1.480140312
5	15091.05386	476.946143	1.58137268
6	14947.41668	-229.4166814	-0.760658782



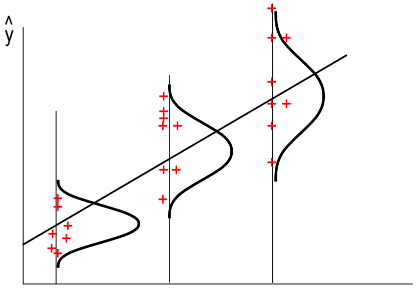
Can also apply Lilliefors test or χ^2 test of Normality

Heteroscedasticity

- When requirement of constant variance is violated, have heteroscedasticity
  - test by plotting residuals vs predicted y

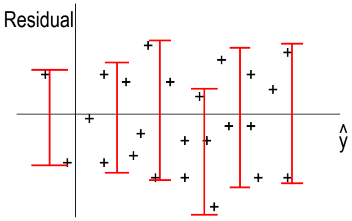


The spread increases with $\hat{y}$

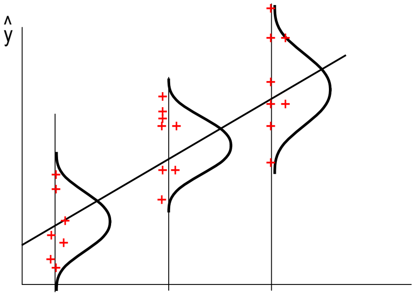


Homoscedasticity

- When requirement of constant variance is not violated, have homoscedasticity



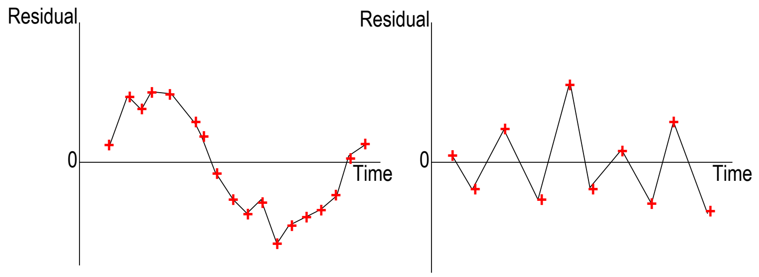
Spread of data points does not change much



Nonindepend. of error variables

- Examining residuals over time, no pattern should be observed if errors independent
- When pattern is detected, errors said to be **autocorrelated**
- Autocorrelation can be detected by graphing residuals against time
- *Time series* if data collected over time
  - may be better to use time series analysis
  - this is for our Forecasting class

Patterns in the appearance of the residuals over time indicates that autocorrelation exists.



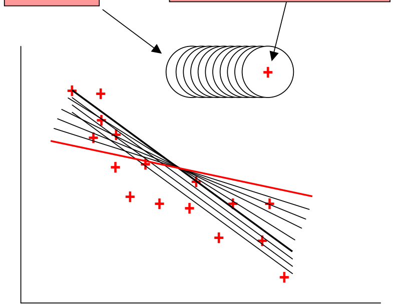
Note the runs of positive residuals, replaced by runs of negative residuals

Note the oscillating behavior of the residuals around zero.

Outliers

- Outlier is an observation that is unusually small or large
- Several possibilities need to be investigated when an outlier is observed:
  - There was an error in recording value
  - Point does not belong in sample
  - Observation is valid
- Identify outliers from the scatter diagram
- Customary to suspect an observation is an outlier if its |standard residual| > 2

(cont.)



An outlier

An influential observation

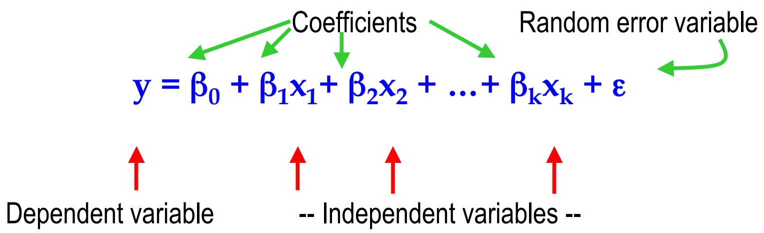
… but, some outliers may be very influential

The outlier causes a shift in the regression line

Multiple Linear Regression

Model & Required Conditions

- Allow for **k** independent variables to potentially be related to a single dependent variable



Coefficients

Random error variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Dependent variable

-- Independent variables --

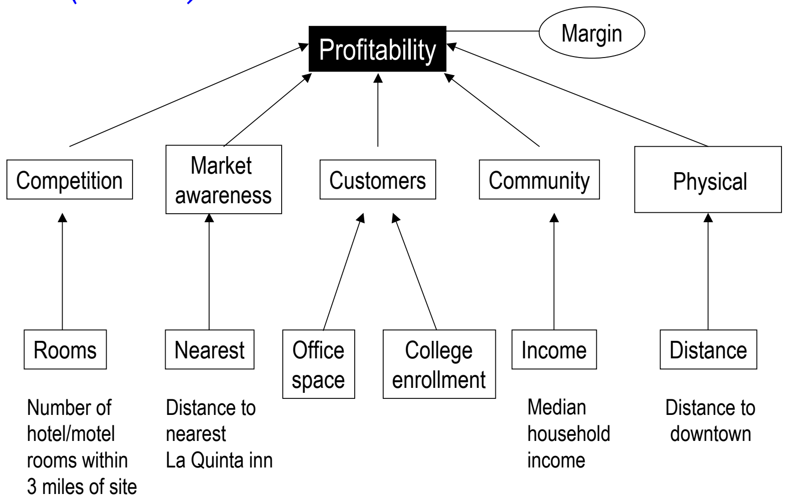
Estimating Coeffs. & Assessing Model

- Procedure
  - Obtain model coefficients and statistics using statistical computer software
  - Diagnose violations of required conditions. Try to remedy problems when identified
  - Assess model fit and usefulness using the model statistics
  - If model passes assessment tests, use it to interpret coefficients and generate predictions

Example

- La Quinta Motor Inns is planning expansion
  - Management wishes to predict which sites are likely to be profitable
  - Several areas where predictors of profitability can be identified are:
    - » Competition
    - » Market awareness
    - » Demand generators
    - » Demographics
    - » Physical quality

(cont.)



Profitability

Margin

Competition

Market awareness

Customers

Community

Physical

Rooms

Nearest

Office space

College enrollment

Income

Distance

Number of hotel/motel rooms within 3 miles of site

Distance to nearest La Quinta inn

Median household income

Distance to downtown

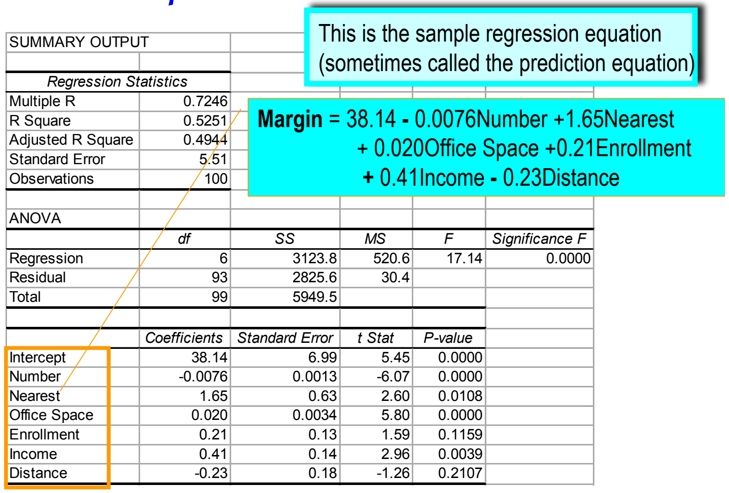
(cont.)

Data was collected from randomly selected 100 inns that belong to La Quinta, and ran for the following suggested model:

Margin = β_0 + β_1Number + β_2Nearest + β_3Office + β_4 College + β_5Income + β_6Distance + ε

Margin	Number	Nearest	Office Space	Enrollment	Income	Distance
55.5	3203	4.2	549	8	37	2.7
33.8	2810	2.8	496	17.5	35	14.4
49	2890	2.4	254	20	35	2.6
31.9	3422	3.3	434	15.5	38	12.1
57.4	2687	0.9	678	15.5	42	6.9

Excel Output



This is the sample regression equation (sometimes called the prediction equation)

Margin = 38.14 - 0.0076Number +1.65Nearest
+ 0.020Office Space +0.21Enrollment
+ 0.41Income - 0.23Distance

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7246
R Square	0.5251
Adjusted R Square	0.4944
Standard Error	5.51
Observations	100

ANOVA

	df	SS	MS	F	Significance F
Regression	6	3123.8	520.6	17.14	0.0000
Residual	93	2825.6	30.4		
Total	99	5949.5			

	Coefficients	Standard Error	t Stat	P-value
Intercept	38.14	6.99	5.45	0.0000
Number	-0.0076	0.0013	-6.07	0.0000
Nearest	1.65	0.63	2.60	0.0108
Office Space	0.020	0.0034	5.80	0.0000
Enrollment	0.21	0.13	1.59	0.1159
Income	0.41	0.14	2.96	0.0039
Distance	-0.23	0.18	-1.26	0.2107

Assessing & Using the Model

- Coefficient of determination (R^2)
- Linear relationship: ANOVA (all β = 0 ?)
- Testing coefficients (each β = 0 ?)
- Standard error of estimate
- Interpreting coefficients
- Using linear regression equation
  - predicting
  - explaining

Multicollinearity, Example

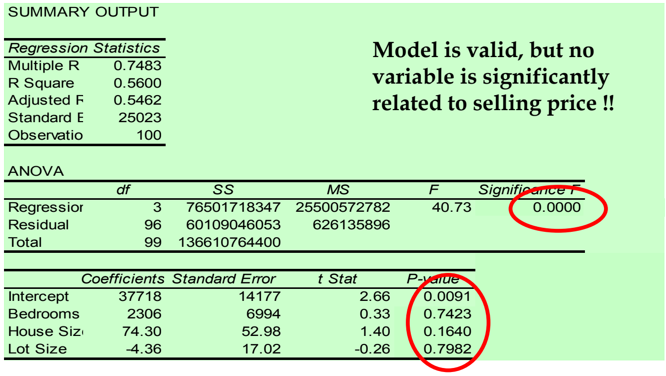
- Real estate agent believes that house selling price can be predicted using house size, number of bedrooms, and lot size
- Random sample of 100 houses was drawn and data recorded

Price	Bedrooms	H Size	Lot Size
124100	3	1290	3900
218300	4	2080	6600
117800	3	1250	3750
.	.	.	.
.	.	.	.

- Analyze relationship among four variables

Solution

- The proposed model is
  $\textbf{PRICE} = \beta_0 + \beta_1\textbf{BEDROOMS} + \beta_2\textbf{H-SIZE} + \beta_3\textbf{LOTSIZE} + \varepsilon$



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7483
R Square	0.5600
Adjusted F	0.5462
Standard E	25023
Observatio	100

Model is valid, but no variable is significantly related to selling price !!

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	3	76501718347	25500572782	40.73	0.0000
Residual	96	60109046053	626135896		
Total	99	136610764400			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	37718	14177	2.66	0.0091
Bedrooms	2306	6994	0.33	0.7423
House Siz	74.30	52.98	1.40	0.1640
Lot Size	-4.36	17.02	-0.26	0.7982

- Investigating each independent variable alone, it is found that each is strongly related to selling price (correlation analysis)
- **Multicollinearity** is source of problem

	Price	*Bedrooms*	*H Size*	*Lot Size*
Price	1			
Bedrooms	0.645411	1		
H Size	0.747762	0.846454	1	
Lot Size	0.740874	0.83743	0.993615	1

- Correlation Table shows each independent variable is also correlated with the others!
  - Could have anticipated this

Multicollinearity

- Two or more independent variables in the model are linearly related to each other
  - check by regressing each X on all the other X's (VIF)

$$\text{VIF} = 1 \;/\; (1 - R^2)$$

- Causes two problems:
  - t statistics appear to be too small (insignificant p)
  - β coefficients cannot be interpreted as "slopes"

Regression Diagnostics (Review)

Required conditions for model assessment to apply must be checked

Is error variable normally distributed?	Use Normality plot or histogram of residuals
Is error variance constant?	Plot std residuals versus y-hat
Are errors independent?	Plot std residuals versus time periods
Can identify outliers?	Scatterplot, residual analysis
Is multicollinearity a problem?	Use VIF values (issue if any VIF > 5)

Remedying Violations of Required Conditions

- **Nonnormality** or **heteroscedasticity** can usually be remedied by using transformations on y variable
- Transformations can improve linear relationship between dependent variable and independent variables
- Many computer software systems allow us to make transformations easily

Brief List of Transformations

- normalize the y appropriately: y′ = y / size_factor
- y′ = log y (for y > 0)
  - When s_ε increases with y, or
  - When error distribution is positively skewed
- $y' = y^2$
  - When the s_ε^2 is proportional to E(y), or
  - When error distribution is negatively skewed
- $y' = y^{1/2}$ (for y > 0)
  - When s_ε^2 is proportional to E(y)
- y′ = 1/y
  - When s_ε^2 increases significantly as y increases beyond some value

Cautions – Interpreting R^2

- R^2 does *NOT* tell whether:
  - independent variables are true cause of changes in dependent variable;
  - an important independent variable was left out (omitted-variable bias);
  - correct/best regression equation used;
  - most appropriate set of independent variables chosen
  - multi-collinearity is present in data
  - model might be improved by using transformed versions of existing set of independent variables

Model Building - Introduction

- Regression analysis is one of the most commonly used techniques in statistics
- Considered powerful because:
  - Can cover variety of mathematical models
    - » linear relationships
    - » non - linear relationships
    - » qualitative variables
  - Provides efficient methods for model building, to select best fitting set of variables

Polynomial Models

- Independent variables may appear as functions of a number of predictor variables
  - Polynomial models of order p with one predictor variable: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + \varepsilon$
  - Polynomial models with two predictor variables

    For example:

    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

    Interaction term

    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

Indicator (Dummy) Variables

- For categorical (qualitative) variables
- 0 - 1 value
- If n categories, need n - 1 indicator variables
  - Gender: 2 categories, 1 variable (0, 1)
  - Education: 5 categories, 4 variables
    - » 0, 0, 0, 0; 1, 0, 0, 0; 0, 1, 0, 0; 0, 0, 1, 0; 0, 0, 0, 1
- Requires larger sample size

Developing a model

- Much better to have a **logical** model in mind, rather than to just start working with a pile of independent variables
- Identify **dependent variable**; clearly define it
- Identify **potential predictors**
  - remember **multicollinearity** problem
  - consider **cost** of gathering & processing data
  - be parsimonious
- Rule of thumb: ≥ 8 observations for every independent variable in model; if violated, adjusted r^2 value will be significantly less
  - Green's rule of thumb is 50 + 8 * # ind. variables

Developing model (cont.)

- Identify several possible models
  - scatterplot of variables can help
  - if uncertain, start with 1st order and 2nd order models, with and without interaction
  - try other relationships (transformations) if polynomial models fail to provide a good fit
- Consider stepwise regression: introduce independent variables one-at-a-time, based on their contribution to current model (reduces multicollinearity)
  - SAS, SPSS provide this

Summary of Regression Issues

- Data quality
  - outliers (influential observations)
  - missing data and/or variables
- Relationship
  - Linear? Nonlinear?
  - Choice of independent variable (*cause?*) & dependent (*effect?*)
- Developing potential model
  - use graphical tools & descriptive statistics (scatterplot, correlation analysis)

(cont.)

- Model assumptions
  - Met? If not, how serious?
  - Residual analysis/plots
  - transformation of variable values
- Interpretation
  - business issues
  - answer questions
  - extrapolation danger
- Prediction & prediction intervals
  - individual response
  - mean response

*Additional Topics**

- Prediction intervals
- Heteroskedasticity: types, tests, correcting for (we've covered some of this)
- Serial correlation: Durbin-Watson test, detecting serial correlation