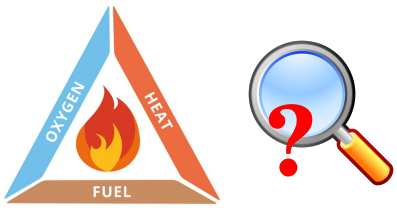
Root Cause Analysis and Linear Regression



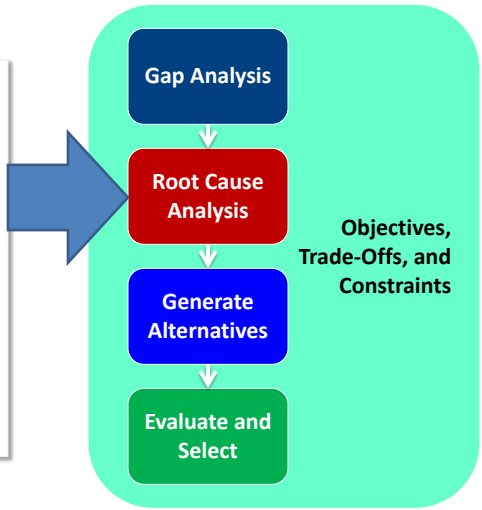
From BUS 462 (Business Analytics) at SFU
Amin Milani Fard - Fall 2018

Some slides are from Michael Brydon

1

Root Cause Analysis

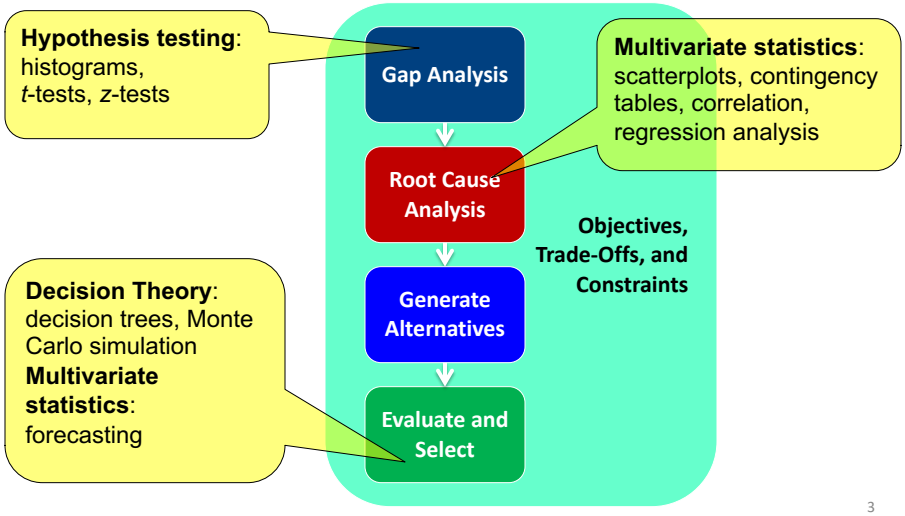
- We have confirmed that a gap exists
- What factor or combination of factors is the "root cause" of the problem?



2

2

Recall Our Model of Decision Making



3

3

Some Root Cause Analysis Techniques

LOGICAL:

- Kepner-Tregoe troubleshooting
  - comparables
  - what is/what is not
  - goal: hypotheses + refuted hypotheses

THEORY DRIVEN:

- Scientific method
  - a priori beliefs and theory
  - influence diagrams
  - hypotheses to test using data

DATA-DRIVEN:

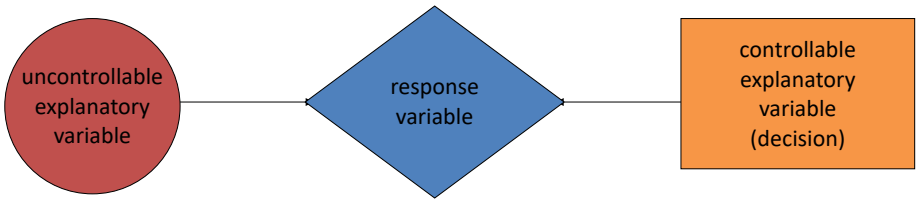
- No a priori model
- Conventional statistical analysis
  - visualization/graphing
  - regression
- Data mining
  - classification trees
  - stepwise regression
  - others (neural networks)

4

4

Influence Diagrams

- Shows how different elements of the problem are related to each other
- Systematic and standardized
- Supports computation of optimal decisions

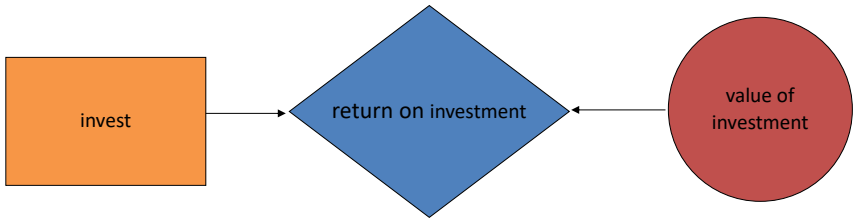


5

5

Influence Diagrams

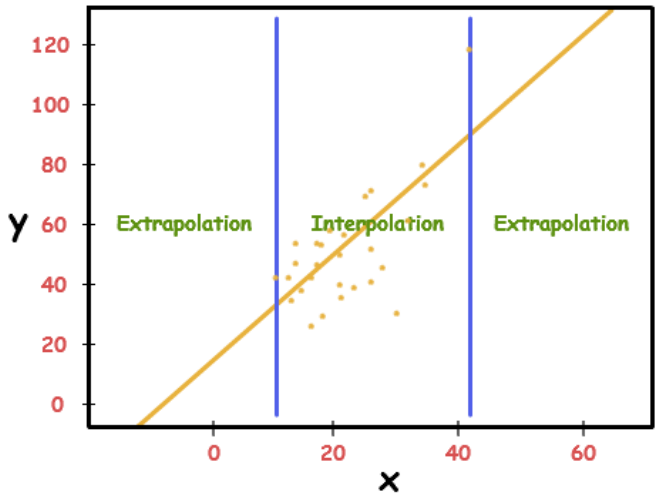
- Similar to Bayesian networks
- Arcs indicate dependence
- Absence of arcs indicate independence



6

6

Extrapolation and Interpolation



7

7

Explanatory vs. Predictive

Explanatory Modeling

- How is X related to Y?
- Sample sizes are small.
- Include few variables.
- To assess, use *p*-values and confidence intervals.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} \ldots + \hat{\beta}_k X_{ki}$$

Predictive Modeling

- If you know X, can you predict Y?
- Sample sizes are large.
- Include many variables.
- To assess, validate with holdout sample data.

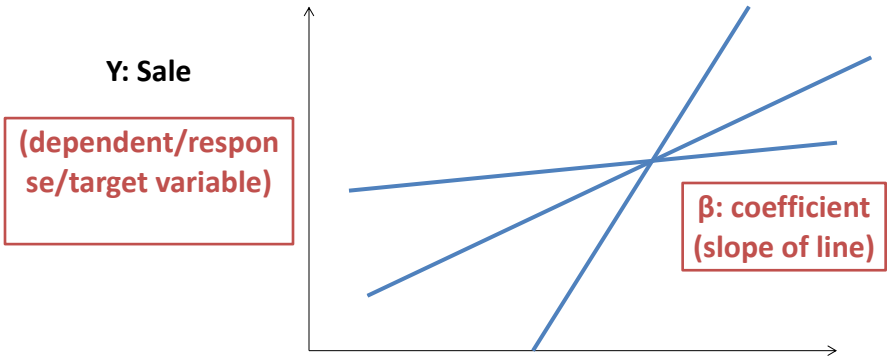
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} \ldots + \hat{\beta}_k X_{ki}$$

Source: SAS Enterprise Guide: Anova, Regression, and Logistic Regression

8

8

Single Regression: Graphical Interpretation



9

9

(Linear) Regression Equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- Predicted response variable (or dependent variable)
- Intercept
- Coefficient or slope
- Explanatory variable

- Slope, rise over run: a point estimate for every unit increase in x we should expect y increase.
- Intercept is more useful for predictive model

10

10

(Linear) Regression Equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- Predicted response variable (or dependent variable)
- Intercept
- Coefficient or slope
- Explanatory variable

Epsilon is a random variable showing error (residuals).

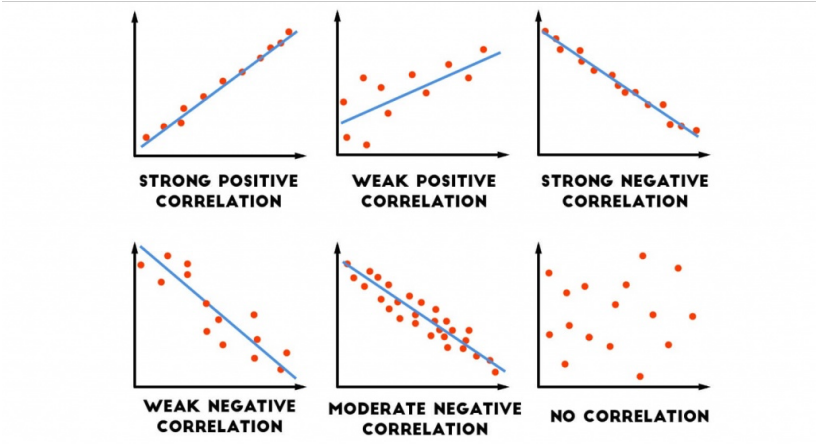
The best-fitting line for the observed data is a line that minimizes the sum of squared residual (epsilon) (differences between actual and predicted values of the dependent variable y)

11

11

Correlation

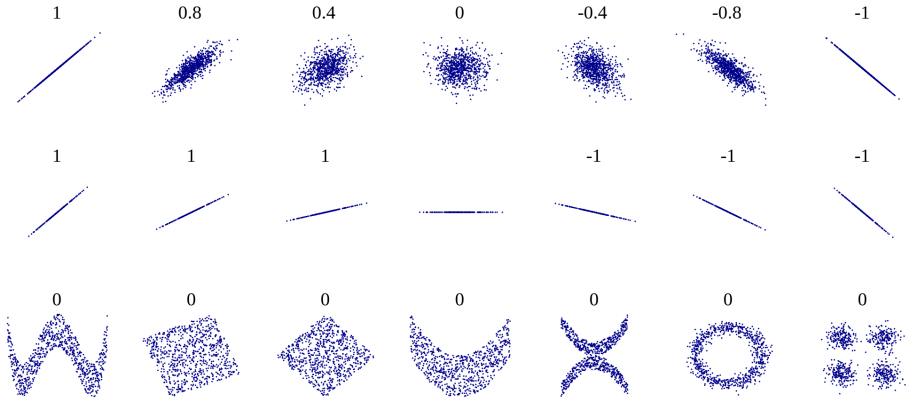
The correlation of two random variables is the strength of the linear relationship between them.



12

12

Correlation



Source: Wikipedia "correlation"

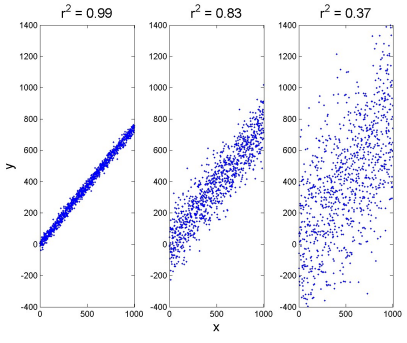
13

13

(Linear) Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- R^2: Coefficient of determination (explained variance), is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
- R^2 tells us how well the fitted data account for the raw data



14

14

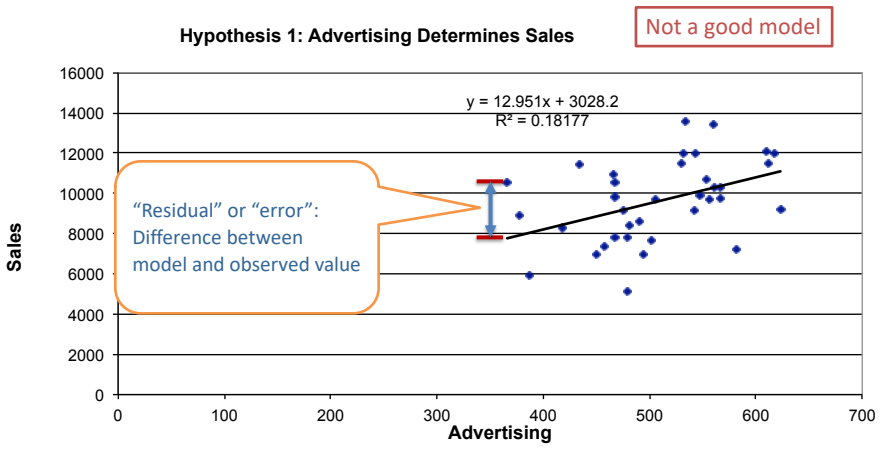
Confidence

- The *p*-value gives the probability that the coefficient (slope of the line) is zero.
- The null hypothesis for liner regression is that the regression model does not fit the data better than the base line model with slope 0.

15

15

Root Cause of Sales?



16

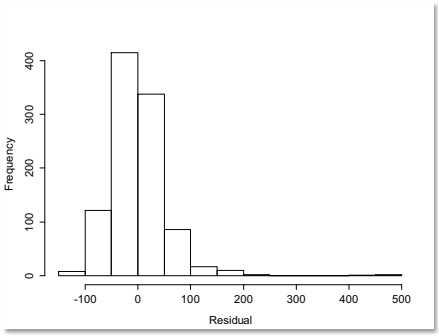
16

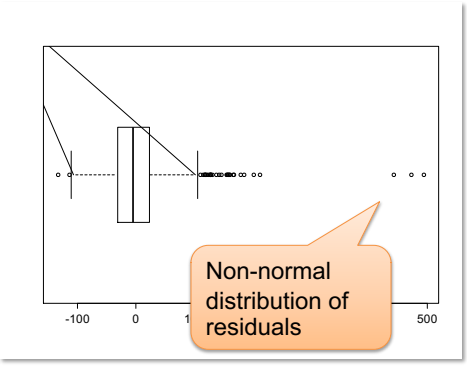
Normality of the Residuals

- An important assumption of the regression technique is that the "residuals" (the difference between observed values of the response variable and the values predicted by the regression model) are approximately normally distributed.
- The output from the regression analysis becomes questionable if these assumptions are not satisfied.

17

Check the Residuals

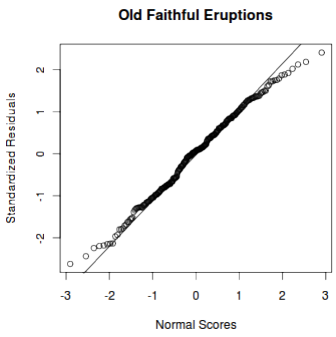




18

Checking Normality of Errors

- If our error distribution is ***not*** normal, it will compromise our estimates for our coefficients and the calculation of statistical significance.
- One of the best ways to test for the normality of distributed errors is to look at the *normal probability plot* of the residuals (Q-Q plot). If the distribution is normal, the points on the plot will fall close to the diagonal line. A bowed pattern or an S-shape pattern indicate typical problems (e.g., highly skewed errors or highly kurtotic residuals)



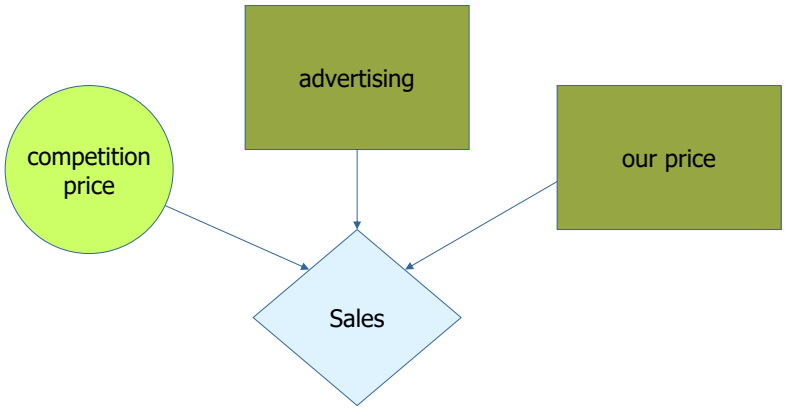
Example normal probability plot of residuals

Example from: http://www.r-tutor.com/elementary-statistics/simple-linear-regression/normal-probability-plot-residuals

19

19

Influence Diagram: A Theory of Sales



Note the difference between **controllable** and **uncontrollable** inputs

Better model requires business knowledge

20

20

Collect Data and Test the Theory …

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.976305882
R Square	0.953173175
Adjusted R Square	0.94878316
Standard Error	459.0979388
Observations	36

ANOVA

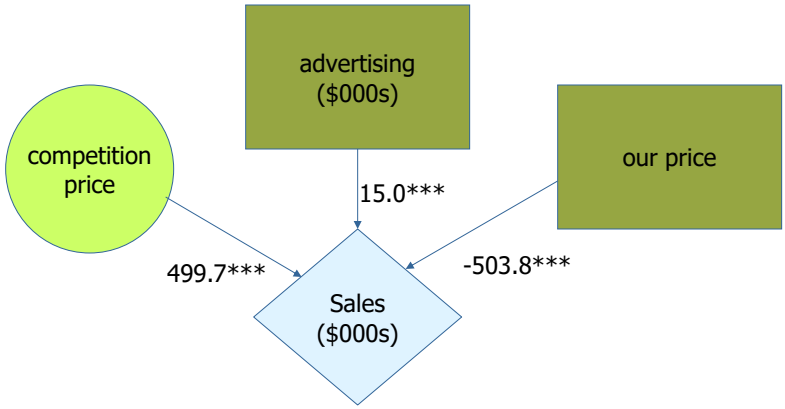
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	137289637.6	45763213	217.123	2.41332E-21
Residual	32	6744669.357	210770.9		
Total	35	144034307			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2199.342251	3839.735609	0.572785	0.570794	-5621.943207	10020.62771
Advertising ($1000s)	15.04660288	1.172569433	12.83216	3.67E-14	12.65815711	17.43504864
Price ($)	-503.7640378	28.3435642	-17.7735	3.84E-18	-561.4979885	-446.0300871
Competition Price ($)	499.6712512	30.55929246	16.35088	4.29E-17	437.4240097	561.9184927

21

21

The Causes of Sales



22

22

Multiple Regression

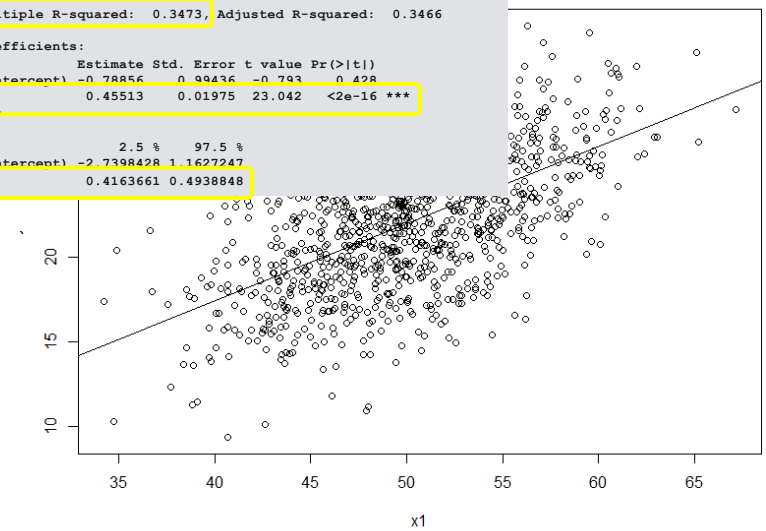
	x1	x2	y
1	50.20952	3.65678616	18.490011
2	43.87222	7.30814474	20.955372
3	46.41969	7.34204440	20.854209
4	49.80087	9.84246352	20.134128
5	55.08988	6.52755044	27.752809
6	48.74429	15.66753462	21.897068
7	43.98581	13.33706312	13.507983
8	44.05346	3.88454236	19.859177
9	50.16832	6.29607297	22.347720
10	48.34303	12.39558697	16.933327
11	49.14650	14.67294304	23.508966
12	49.80759	5.45849924	28.156904
13	44.45976	9.22811751	20.189028
14	49.66238	17.18127061	19.448855
15	50.03527	7.46259532	22.186256
16	48.29303	11.50504225	17.865755
17	45.06410	12.76367725	20.326278
18	51.15800	12.24336042	16.936480

Consider a set of data:

- Explanatory variables x1 and x2
- Response variable y

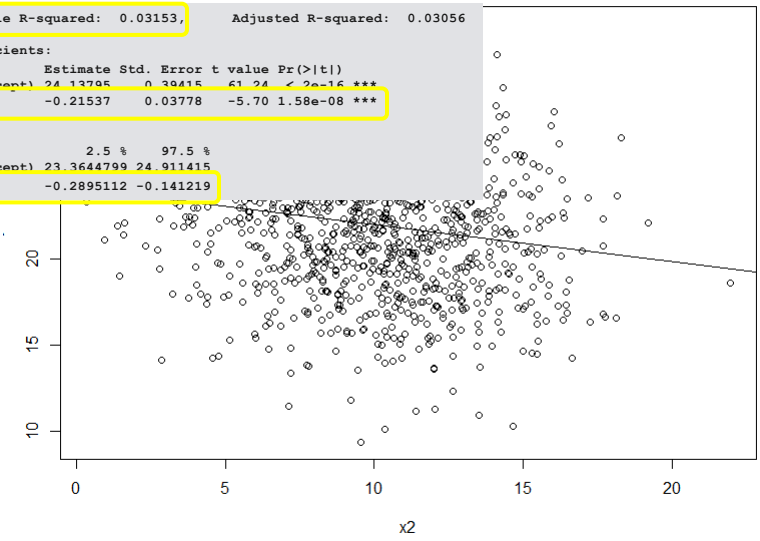
Root Cause Analysis 1

```
Multiple R-squared:  0.3473,  Adjusted R-squared:  0.3466

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.78856    0.99436  -0.793    0.428
x1           0.45513    0.01975  23.042   <2e-16 ***
---

                2.5 %    97.5 %
(Intercept) -2.7398428 1.1627247
x1           0.4163661 0.4938848
```



Root Cause Analysis 2

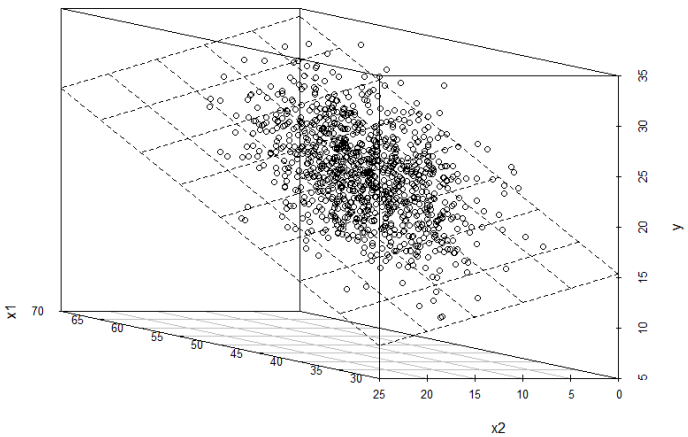
```
Multiple R-squared:  0.03153,     Adjusted R-squared:  0.03056
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.13795    0.39415   61.24  < 2e-16 ***
x2          -0.21537    0.03778   -5.70 1.58e-08 ***

                2.5 %    97.5 %
(Intercept) 23.3644799 24.911415
x2          -0.2895112 -0.141219
```



25

25

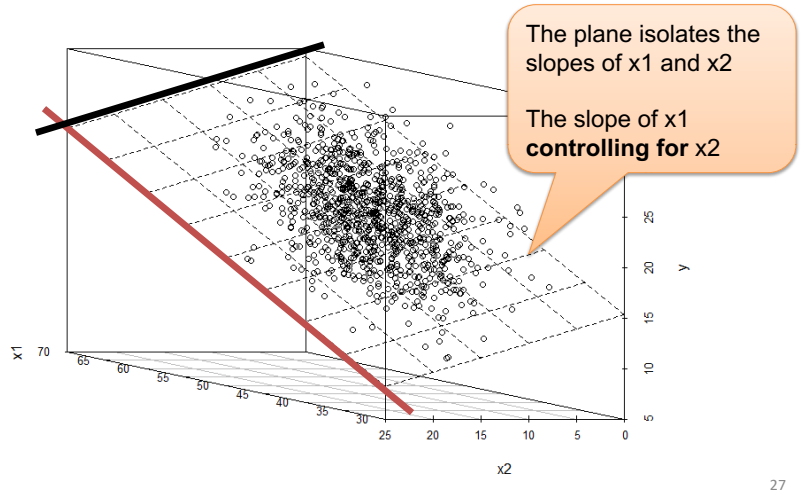
Multiple Regression: x1 and x2 together



26

26

Multiple Regression: x1 and x2 together



27

27

Multiple Regression Results

```
Multiple R-squared:  0.402,   Adjusted R-squared:  0.4008

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.18089    0.97429   1.212    0.226
x1           0.47218    0.01900  24.854   <2e-16 ***
x2          -0.28506    0.02984  -9.554   <2e-16 ***
---

                 2.5 %     97.5 %
(Intercept) -0.7309938  3.0927794
x1           0.4349027  0.5094668
x2          -0.3436119 -0.2265097
```

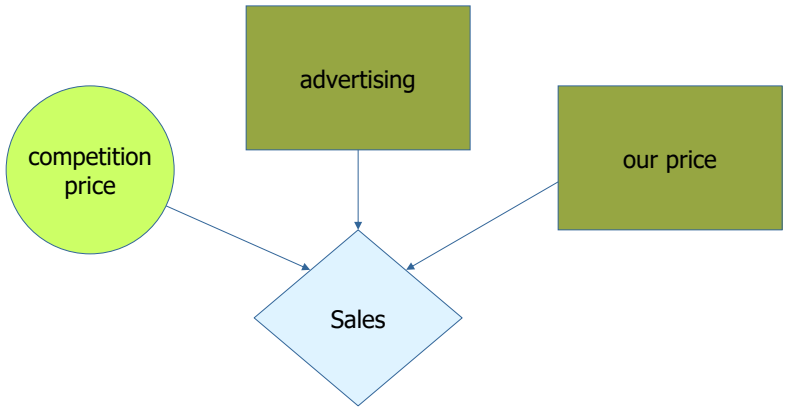
Actual function to generate points:

```
y = 0.5*x1 -0.292*x2 + rnorm(1000, 0,3)
```

28

28

Influence Diagram: A Theory of Sales



Note the difference between **controllable** and **uncontrollable** inputs

Better model requires business knowledge

29

29

Collect Data and Test the Theory …

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.976305882
R Square	0.953173175
Adjusted R Square	0.94878316
Standard Error	459.0979388
Observations	36

ANOVA

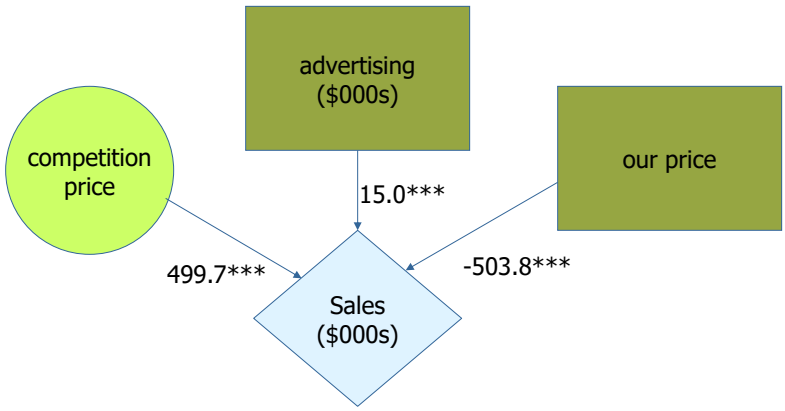
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	137289637.6	45763213	217.123	2.41332E-21
Residual	32	6744669.357	210770.9		
Total	35	144034307			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2199.342251	3839.735609	0.572783	0.570794	-5621.943207	10020.62771
Advertising ($1000s)	15.04660288	1.172569433	12.83216	3.67E-14	12.65815711	17.43504864
Price ($)	-503.7640378	28.3435642	-17.7735	3.84E-18	-561.4979885	-446.0300871
Competition Price ($)	499.6712512	30.55929246	16.35088	4.29E-17	437.4240097	561.9184927

30

30

The Causes of Sales



31

31

Linear Regression Assumptions

- linearity of the relationship between dependent and independent variables
- independence of the errors (no serial/auto correlation)
- homoscedasticity (constant variance) of the errors
- normality of the residuals

The output from the regression analysis becomes questionable if these assumptions are not satisfied.

32

32

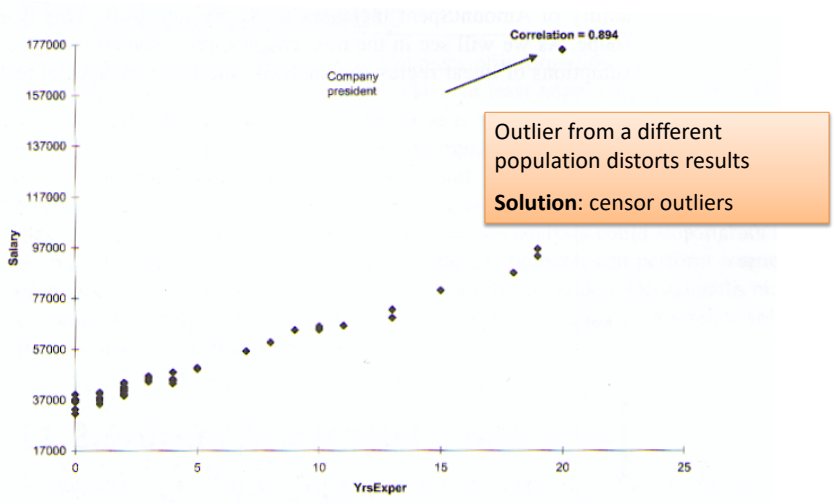
Linear Regression Gotchas

- Outliers
- Categorical explanatory variables
- Transformation
  - non-normal distribution of residuals
  - non-linear relationship between variables
  - heteroskedasticity (non-constant error variance)
- Multicolinearity (colinearity between explanatory variables)

33

33

Checking the Data: Outliers Detection

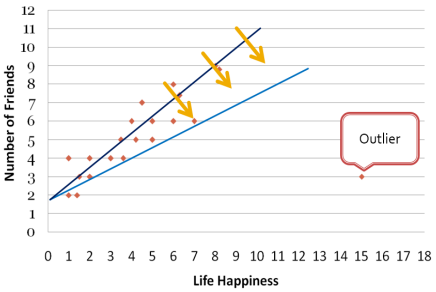


34

34

Problem: Outliers

- A few outliers can drastically shif our predicted line.
- Look for cases that are far outsic the norm (**very large residuals)**.
- We may also be interested in hov much ***influence*** our outliers actually have on the model.



35

35

Categorical Variables

- The charge is that its female employees receive substantially smaller salaries than the firm's male employees
- Is gender bias the root cause of pay variance?

	A	B	C	D	E	F	G	H	I
1	**Bank salary data**								
2									
3	Employee	EducLev	JobGrade	YrHired	YrBorn	Gender	YrsPrior	PCJob	Salary
4	1	3	1	92	69	Male	1	No	32
5	2	1	1	81	57	Female	1	No	39.1
6	3	1	1	83	60	Female	0	No	33.2
7	4	2	1	87	55	Female	7	No	30.6
8	5	3	1	92	67	Male	0	No	29
9	6	3	1	92	71	Female	0	No	30.5
10	7	3	1	91	68	Female	0	No	30
11	8	3	1	87	62	Male	2	No	27
12	9	1	1	91	33	Female	0	No	34
13	10	3	1	86	64	Female	0	No	29.5
209	206	5	6	63	33	Male	0	No	88
210	207	5	6	60	36	Male	0	No	94
211	208	5	6	62	33	Female	0	No	30

Example from: Albright, Winston, Zappe, "Data Analysis and Decision Making"

36

36

Question 1: Is there a gap?

- Hypothesis test in R (open source statistics package)

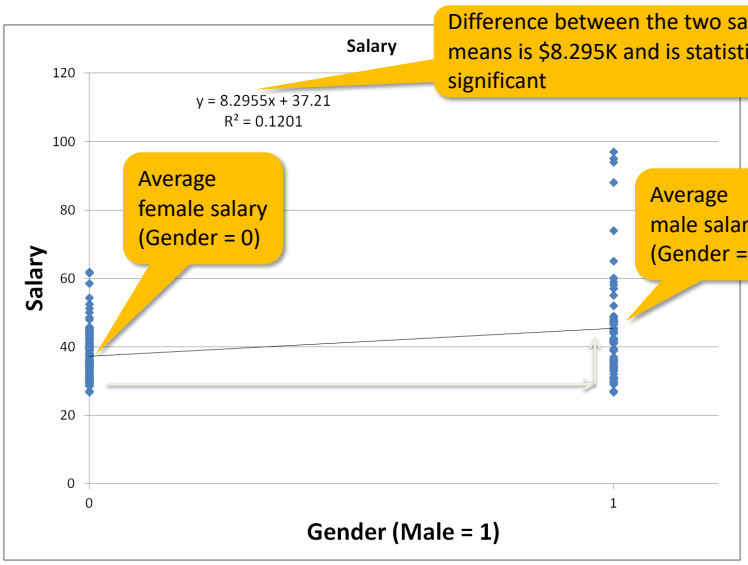
```
t.test(Salary~Gender, alternative='two.sided', conf.level=.95, var.equal=TRUE, data=df)

     Two Sample t-test

data:  Salary by Gender
t = -5.3024, df = 206, p-value = 2.935e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.379984  -5.211041
sample estimates:
mean in group Female   mean in group Male
            37.20993             45.50544
```

http://www.r-project.org/

37

37

T-Test = Regression



38

38

Categorical Variables

- Use “dummy” (or “indicator”) variables to encode categories as a family of zero-one variables

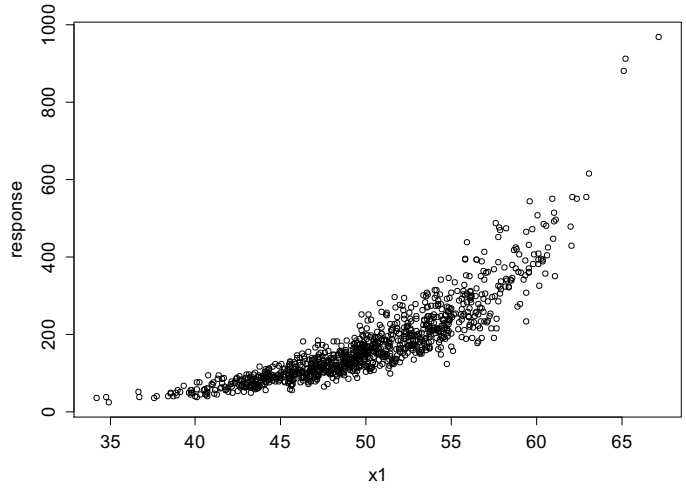
Color
red
green
green
blue
red
blue
red
green
red
red

Red	Green	Blue
1	0	0
0	1	0
0	1	0
0	0	1
1	0	0
0	0	1
1	0	0
0	1	0
1	0	0
1	0	0

39

39

What is Wrong?

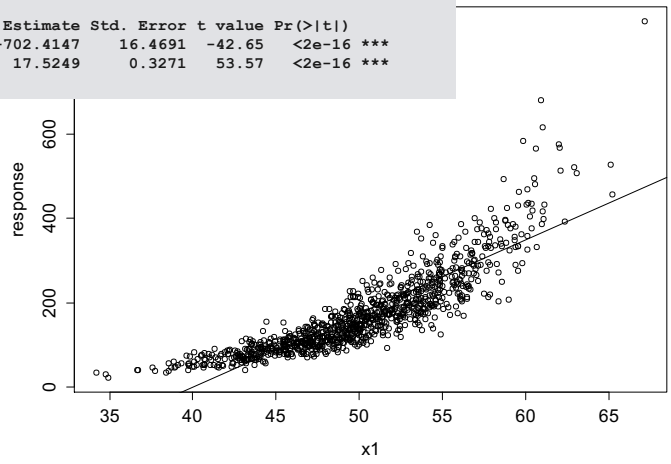


40

40

Can Draw a Line

```
Multiple R-squared:  0.742,   Adjusted R-squared:  0.7417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -702.4147    16.4691  -42.65   <2e-16 ***
x1            17.5249     0.3271   53.57   <2e-16 ***
---
```



41

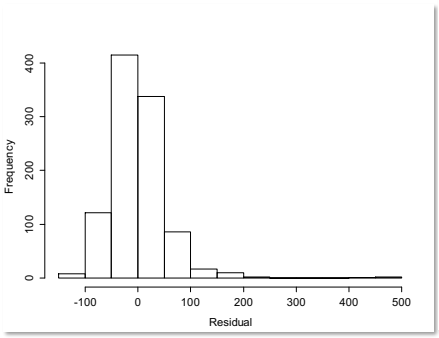
41

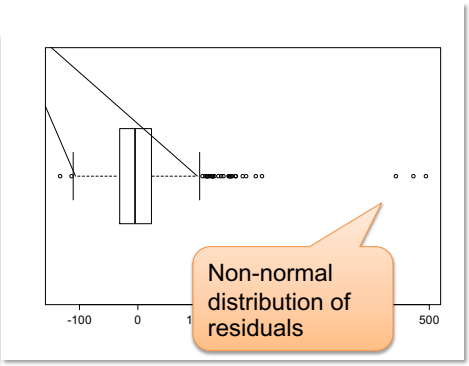
Normality of the Residuals

- An important assumption of the regression technique is that the "residuals" (the difference between observed values of the response variable and the values predicted by the regression model) are approximately normally distributed.
- The output from the regression analysis becomes questionable if these assumptions are not satisfied.

42

Check the Residuals

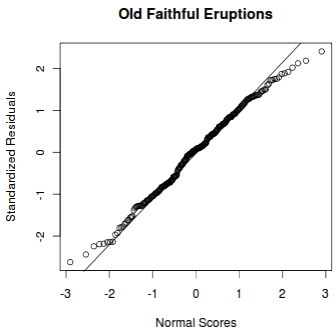




43

Checking Normality of Errors

- If our error distribution is ***not*** normal, it will compromise our estimates for our coefficients and the calculation of statistical significance.
- One of the best ways to test for the normality of distributed errors is to look at the *normal probability plot* of the residuals (Q-Q plot). If the distribution is normal, the points on the plot will fall close to the diagonal line. A bowed pattern or an S-shape pattern indicate typical problems (e.g., highly skewed errors or highly kurtotic residuals)



Example normal probability plot of residuals

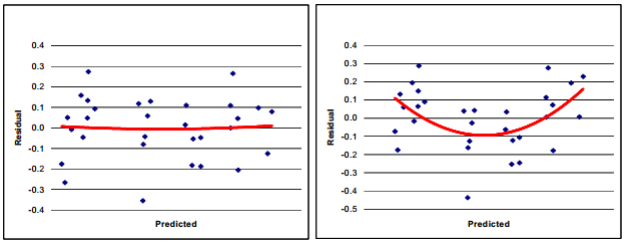
Example from: http://www.r-tutor.com/elementary-statistics/simple-linear-regression/normal-probability-plot-residuals

44

44

Problem : Checking Linearity

- We should check to make sure that our data is actually linearly related.
- We can examine the ***residuals versus predicted values plot***



Example 1: Normal

Example 2: Non-Linear
(notice the bowed shape)

45

45

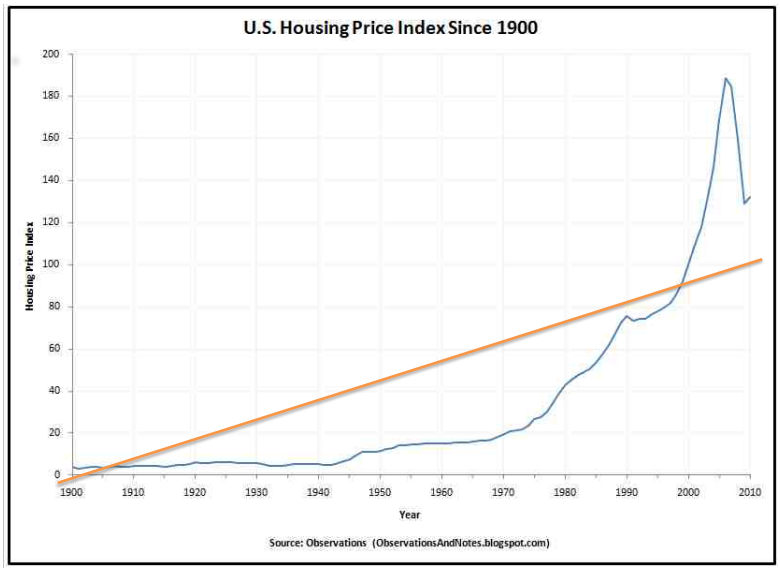
Fixing Linearity problems

- One way to address a non-linear relationship is to transform the variable.
- We can use transformations to shift a highly skewed distribution so that we can better fit a linear relationship between the independent and dependent variables.

46

46

Non-Linear Growth



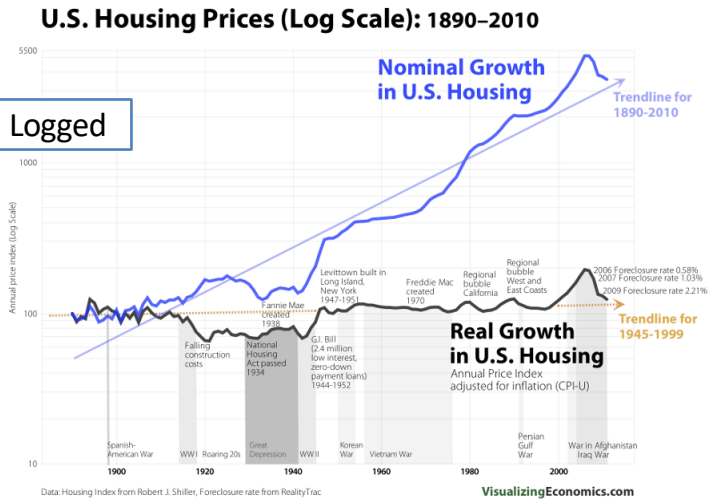
47

47

Data Transformation

Same Data Logged

If we plot the response variable on a log scale, we can fit a line.



48

48

Trick: Transform the Response Variable

- Original linear model:

$$Y = \beta_o + \beta_1 X$$

- Transformed linear model:

$$\log_e(Y) = \beta_o + \beta_1 X$$

Meaning:

$$Y = e^{\beta_0} e^{\beta_1 X}$$

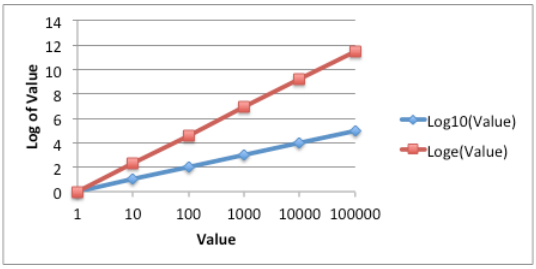
49

49

Logarithm Example

Value	Log_{10}(Value)	Log_e(Value)
1	0	0.00
10	1	2.30
100	2	4.61
1000	3	6.91
10000	4	9.21
100000	5	11.51

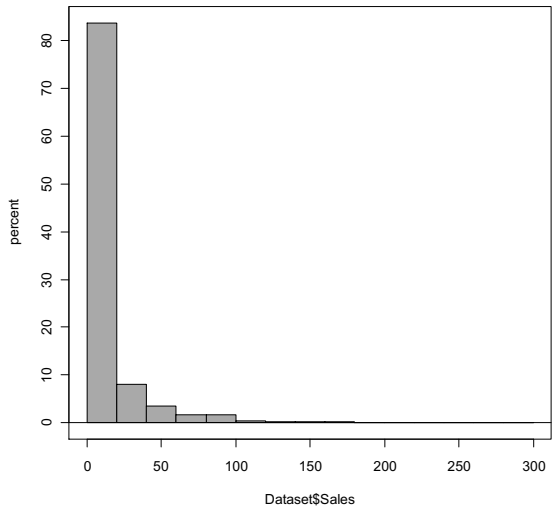
The log of a value grows mush slower than the value.



50

50

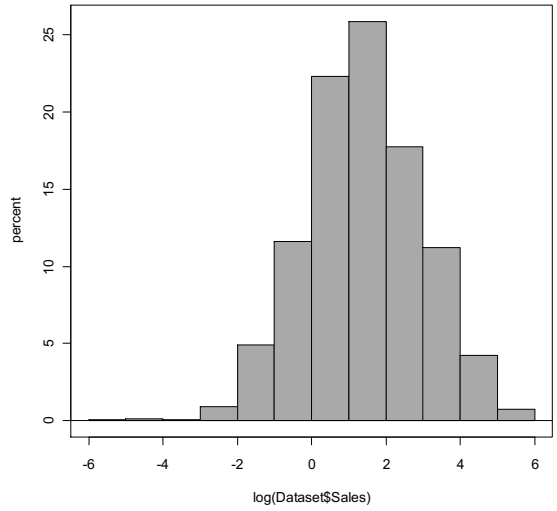
The Long Tail/Power Law Distribution



51

51

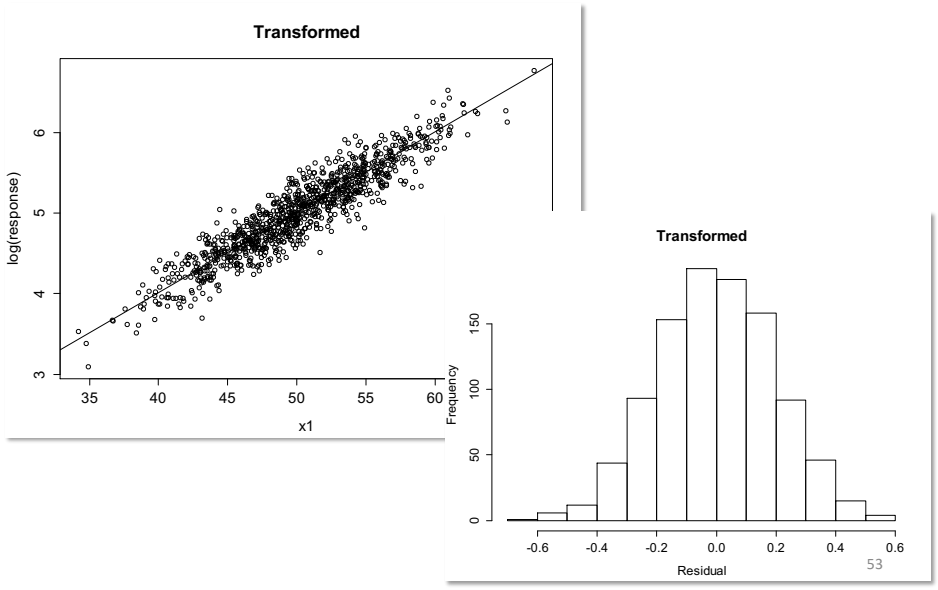
Distribution after Log Transformation



52

52

Transform→Log(y)



53

Interpretation of $Log_e(Y)$ Results

Root MSE	1.27346	R-Square	0.2787
Dependent Mean	1.11104	Adj R-Sq	0.2784
Coeff Var	114.61925		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.08407	0.15276	-20.19	<.0001
Review Score	1	0.05977	0.00214	27.95	<.0001

A one-unit increase in ReviewScore results, on average, in a 5.98% increase in LTD_Sales

54

54

Other Transformations

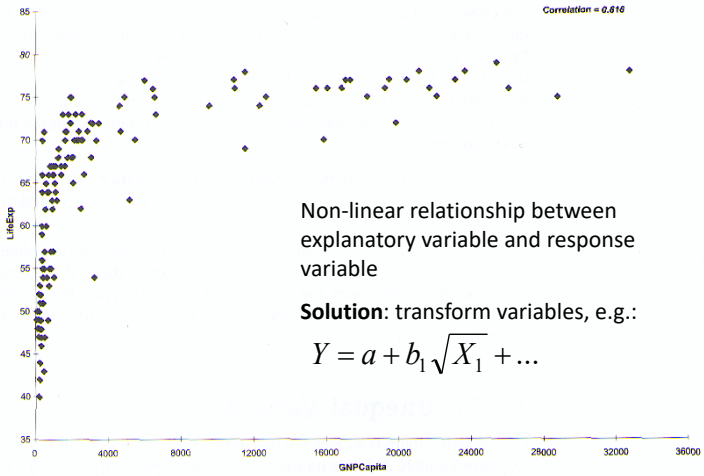


FIGURE 11.8 Scatterplot of Life Expectancy versus GNP per Capita

Non-linear relationship between explanatory variable and response variable

Solution: transform variables, e.g.:

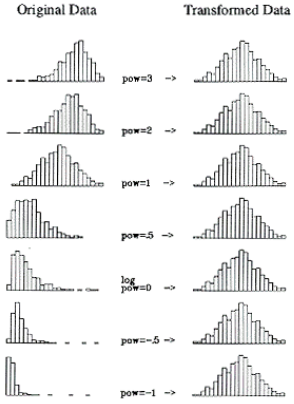
$$Y = a + b_1\sqrt{X_1} + \ldots$$

55

55

Figure 22.3
Ladder of Power Transformations

Power P	SYSTAT BASIC	Name	Notes
P	Y^P	power	DOWN: shorten upper tail.
⋮	⋮	⋮	⋮
3	Y^3	Cube	Not commonly used.
2	Y^2	Square	The highest commonly used power.
—> 1	Y^1	Original data	No transformation.
1/2	Y^(1/2)	Square root	Commonly used for counts.
"0"	LOG(Y)	Logarithm	Commonly used for financial data.
-1/2	-1/Y^(1/2)	Reciprocal root	The minus sign preserves order.
-1	-1/Y	Reciprocal	Lowest commonly used power.
-2	-1/Y^2	Reciprocal square	
⋮	⋮	⋮	⋮
- P	-1/Y^P	Reciprocal power	UP: shorten lower tail.



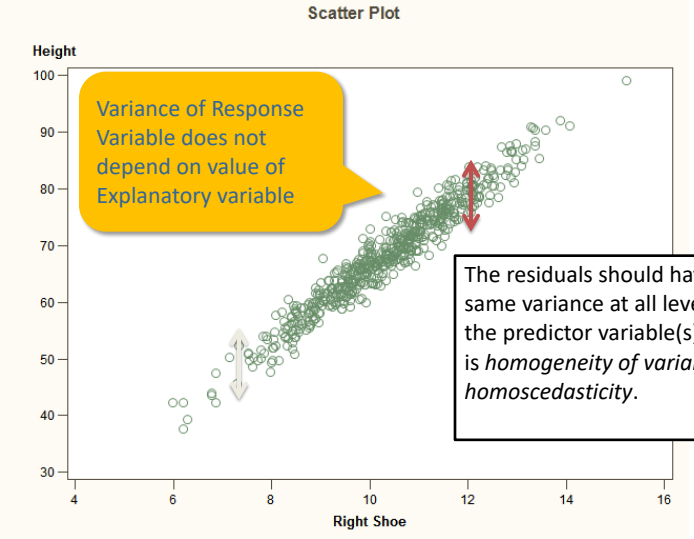
Ladder of Powers

- Helps to correct for skew and non-symmetric distributions.
- Commonly used when the dependent variable in a regression is highly skewed.
- Power transformations will not 'fix' distributions that are bimodal, uniform, etc.

56

56

Homoscedasticity
(Constant Conditional Variance)

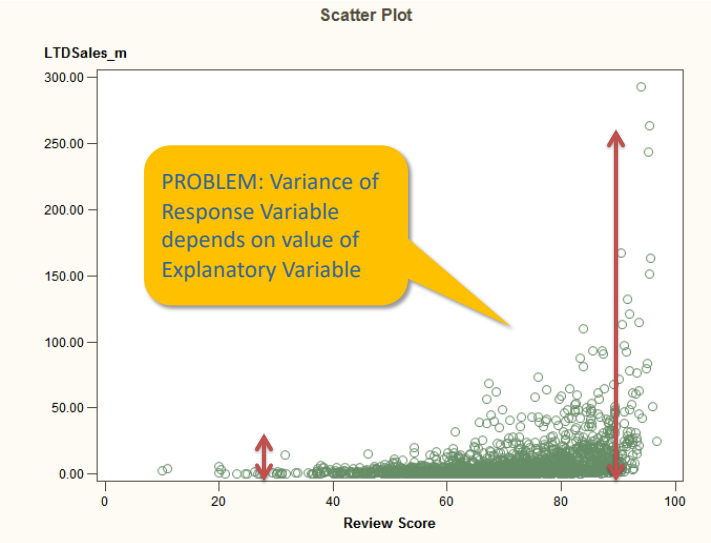


The residuals should have the same variance at all levels of the predictor variable(s). This is *homogeneity of variance*, or *homoscedasticity*.

57

57

Heteroscedasticity
(Non-Constant Conditional Variance)



58

58

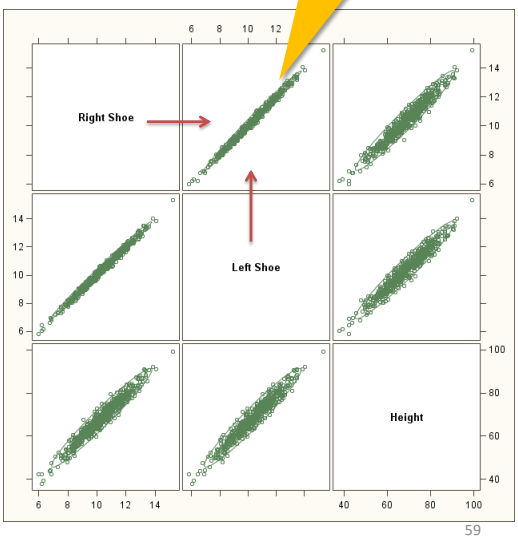
Colinearity

Right shoe size is correlated with left shoe size

Two explanatory variables are correlated

Example:

- Explain **height** as a function of **left shoe size** and **right shoe size**
- Scatter Plot matrix in SAS

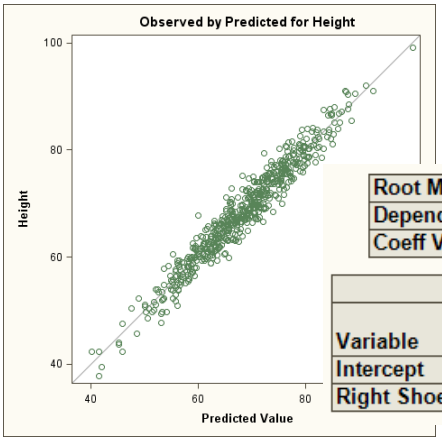


High correlations between independent variables can lead to large errors in our model.

59

59

Height vs **Right** Shoe Size



Root MSE	2.20871	R-Square	0.9444
Dependent Mean	68.40286	Adj R-Sq	0.9443
Coeff Var	3.22897		

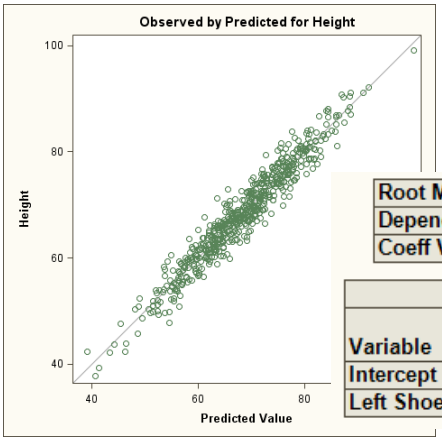
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.24764	0.67967	1.84	0.0669
Right Shoe	1	6.49106	0.06510	99.71	<.0001

60

60

Height vs **Left** Shoe Size



Root MSE	2.36790	R-Square	0.9361
Dependent Mean	68.40286	Adj R-Sq	0.9360
Coeff Var	3.46170		

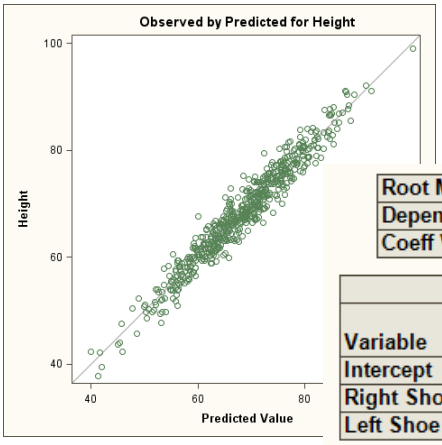
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.71393	0.72683	2.36	0.0187
Left Shoe	1	6.44773	0.06963	92.59	<.0001

61

61

Height vs **Right** & **Left** Shoe Size



Root MSE	2.20857	R-Square	0.9445
Dependent Mean	68.40286	Adj R-Sq	0.9443
Coeff Var	3.22876		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.22803	0.67989	1.81	0.0714
Right Shoe	1	5.84992	0.62201	9.40	<.0001
Left Shoe	1	0.64321	0.62059	1.04	0.3004

62

62

Regression Diagnostics

- Heteroskedasticity: Do the errors have a constant variance?
- Linearity: Is there a linear relationship between the dependent and independent variables?
- Normality: The error term should be normally distributed. (We can use the central limit theorem for large samples)
- Outliers: Do we have a few influential cases that may need to be dropped from the analysis?
- Multicollinearity: If we have multiple predictor variables, are they too related to each other?

63

63

R, R^2, and Adjusted R^2

- *R* is the correlation between the predicted values and the observed values of Y. R^2 (the coefficient of determination) is the square of this coefficient and indicates the percentage of variation explained by your regression line out of the total variation.
- R^2 tends to increase as you include additional predictors in the model. Thus, one can artificially get higher R^2 by increasing the number of variables. The adjusted R^2 penalizes the model for having more variables in the model. When you compare models with their complexity, you should then rely on Adjusted R^2.

64

64

The problem with the R^2 for multiple regression

- Sometimes for model refinement we take insignificant variables out (e.g. backward elimination) or due to colinearity of predictor variables. This results in the R^2 (measure of goodness of fit) to go down slightly. We can't get a more accurate model by removing a variable fro the model.

- R^2 will always increase by adding predictor variables to the model, even if those predictor variables don't significantly contribute to the model. The Adjusted-R^2 takes into consideration the sample size and the number of predictor variables. Its value can increase with fewer variables or smaller sample sizes.

65

65