Regression using R

A CSCAR Workshop

Corey Powell

June 14, 2017

Acknowledgements

These materials are adapted from the Fall 2013 regression workshop taught by Kathy Welch and Missy Plegue.

Workshop Goals

1. To review the theory and practice of regression

2. To get experience performing regression analyses in R

Outline

- Simple Regression
- Diagnostics
- Categorical Predictors
- ANCOVA (Interactions)
- Multiple Regression
- Model Selection

What is Regression?

A technique for learning about the relationship between independent variables, X, and a dependent variable Y.

$$X_1, X_2, ..., X_p \rightarrow Y \tag{1}$$

Terminology

- X: independent variable, covariate, predictor
- Y: dependent variable, response, outcome

Questions Answered by Regression

What factors influence student achievement?

How effective are various treatments for depression?

Does income depend on gender?

Simple Linear Regression

Simple linear regression explores the relationship between a **single predictor**, X, and a response variable Y.

Example

The relationship between height and wingspan

Height and Wingspan Data

height	wingspan
70.2	69.9
66.1	69.6
68.9	70.6
65.8	70.4
63.2	68.3
69.7	71.9

The Simple Linear Regression Model

Review: Statistics and Parameters

Statistics

- A **statistic** is a summary measure of a sample
- Examples
  1. Sample Mean ($\bar{X}$)
  2. Sample Standard Deviation (s)

Parameters

- A **parameter** is a characteristic of a population
- Examples
  1. Population Mean (μ)
  2. Population Standard Deviation (σ)

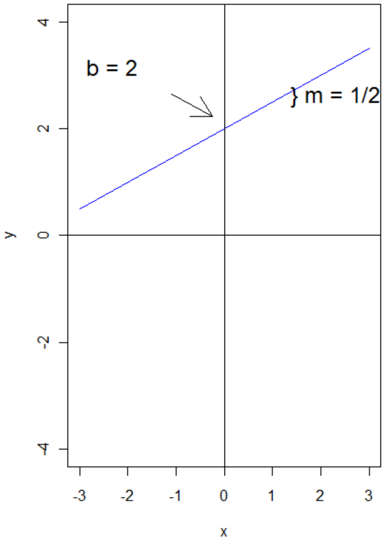
Review of Lines

Common notation:

$$y = mx + b \tag{2}$$

- m is the slope
- b is the y-intercept

$y = .5x + 2$



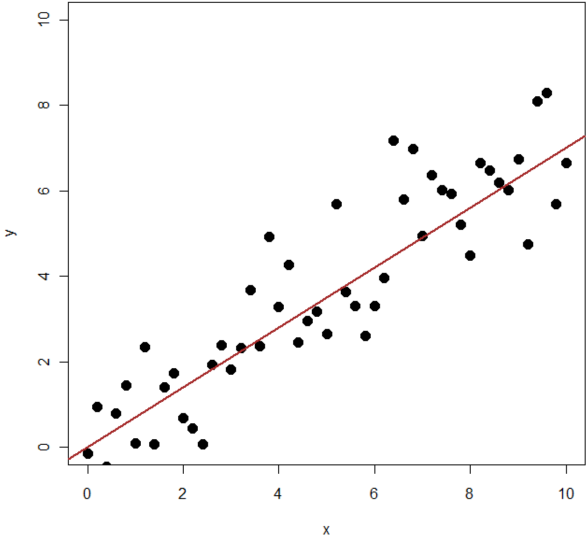
Simple Linear Regression Model

- Assumes a linear relationship between X and the expected value of Y

$$E[Y] = \beta_0 + \beta_1 X \tag{3}$$

- E stands for "expected"
- β_0 is the intercept
- β_1 is the slope, or "effect" of X

$E[Y]$ and Y versus X



Two Equations

Expected

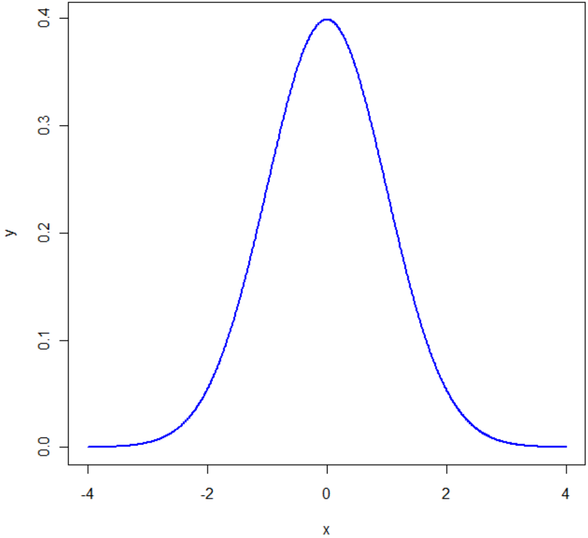
$$E[Y] = \beta_0 + \beta_1 X \tag{4}$$

Individual

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{5}$$

- ϵ is the **error**
- ϵ is normally distributed with mean 0 and variance σ^2
- σ^2 is the error variance

Normal Distribution

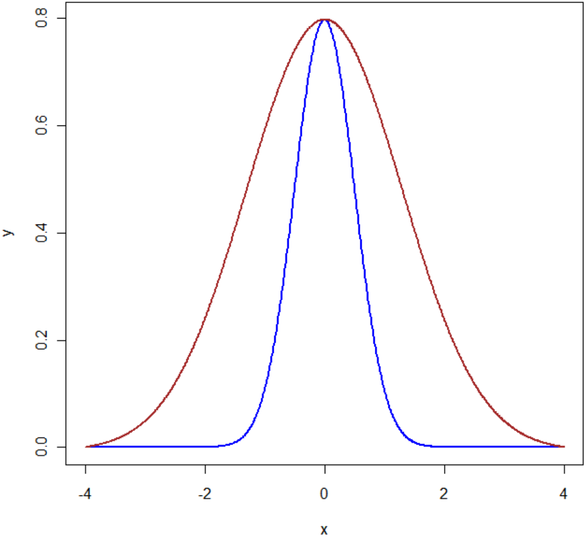


Normal Distribution

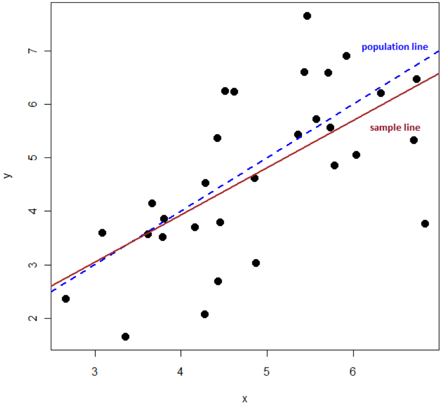
Two Components

- μ denotes the center of the distribution
- σ denotes the standard deviation of the distribution
- σ^2 denotes the variance of the distribution

Same Mean, Different σ^2



Population and Sample Lines

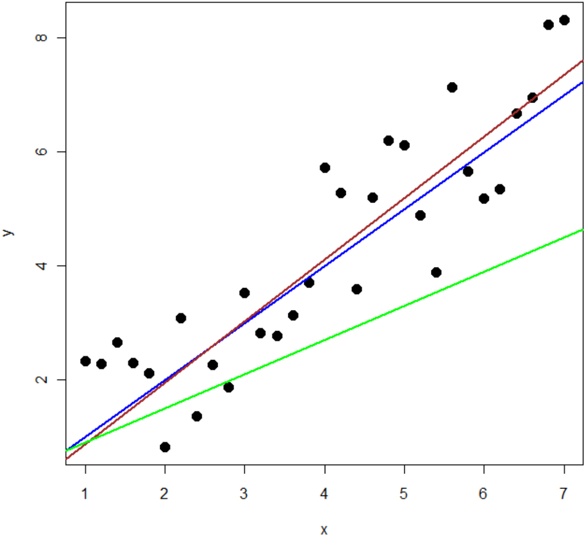


Regression Parameters and Statistics

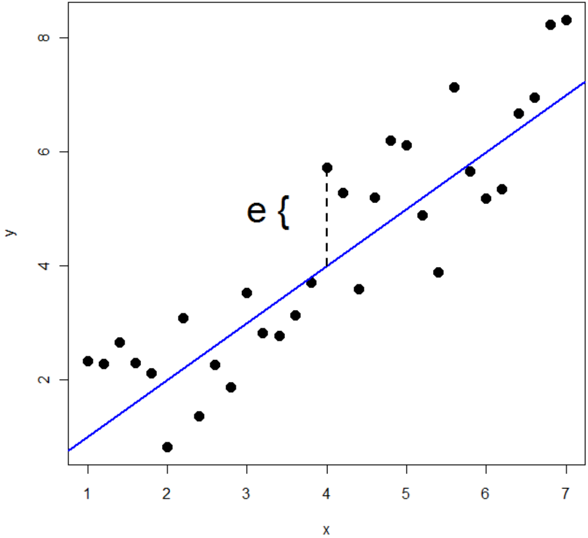
Parameter	Statistic	Alternate Notation
β_0	$\hat{\beta}_0$	b_0
β_1	$\hat{\beta}_1$	b_1
σ^2	$\hat{\sigma}^2$	s^2

- $\hat{\beta}_0$ – estimated intercept
- $\hat{\beta}_1$ – estimated slope
- $\hat{\sigma}^2$ – estimated error variance
- n – sample size

Which Line Is Best?



Residuals



Residuals and Least Squares

Residuals

- Definition: the vertical distance between a point and the line
- Each point has a residual
- The residual of the i^{th} person is denoted e_i

Method of Least Squares

- The **Least Squares regression line** is the line that minimizes the sum of the squared residuals (RSS)

$$\Sigma e_i^2 \tag{6}$$

Least Squares Equations

Slope

$$\hat{\beta}_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \quad (7)$$

Intercept

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (8)$$

Residual Variance

$$\hat{\sigma}^2 = \frac{\Sigma e_i^2}{n - 2} \quad (9)$$

Population and Sample Lines

Means

- Population

$$E[Y] = \beta_0 + \beta_1 X \tag{10}$$

- Sample

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{11}$$

Individual

- Population

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{12}$$

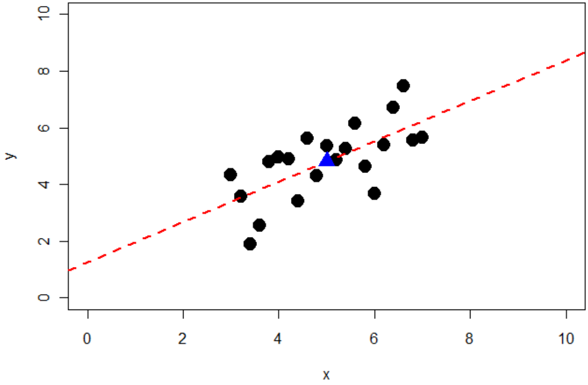
- Sample

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e \tag{13}$$

Properties of Least Squares Regression Line

- Residuals sum to zero: $\Sigma e_i = 0$
- Line passes through the middle $(\bar{X},\ \bar{Y})$ of the data
- Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased **if** model is correct

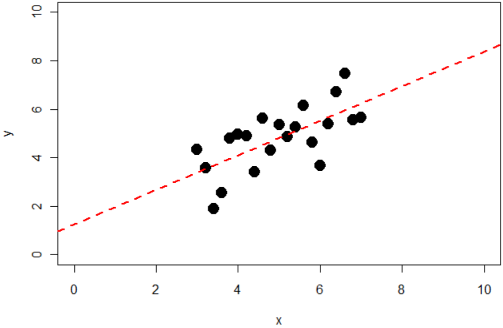
Least Squares Fit



Prediction

Goal of prediction is to predict the Y value for a **new** observation

Fitted Line: $\hat{Y} = 0.7 + 1.3X$



- For every unit increase in X, $\hat{Y}$ increases by 1.3

Prediction

Least Squares Line

$$\hat{Y} = 0.7 + 1.3X \tag{14}$$

Prediction at $X = 4$

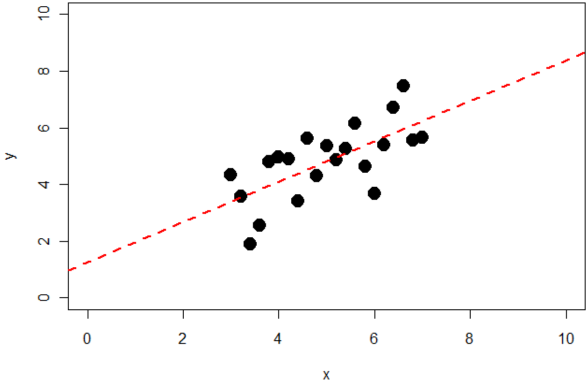
$$\hat{Y} = 0.7 + 1.3(4) = 5.9 \tag{15}$$

Prediction at $X = 5$

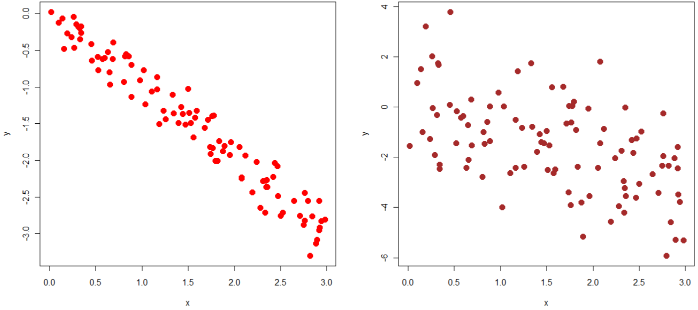
$$\hat{Y} = 0.7 + 1.3(5) = 7.2 \tag{16}$$

$$7.2 - 5.9 = 1.3 \tag{17}$$

Avoid Extrapolation



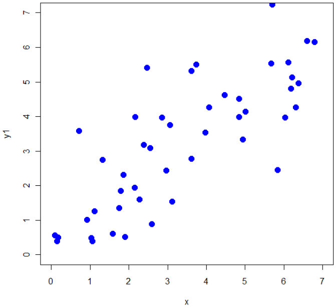
Strength of Relationship

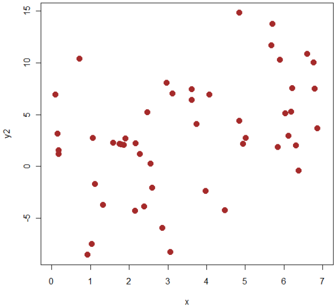


Correlation

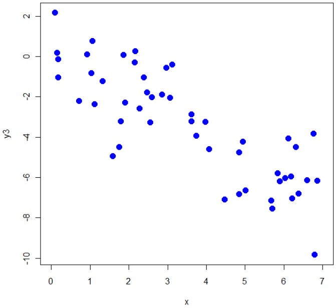
- The **correlation coefficient**, r, measures the strength of the **linear** relationship between Y and X
- r is between -1 and 1
- $r =$ -1 indicates an exact negative linear relationship
- $r = 1$ indicates an exact positive linear relationship
- $r = 0$ indicates no linear relationship
- The regression line slope $\hat{\beta}_1$ is related to r through the equation $\hat{\beta}_1 = r * sd_P(Y)/sd_P(X)$.

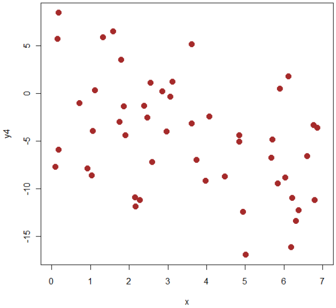
Correlation: $r = 0.8$, $r = 0.4$



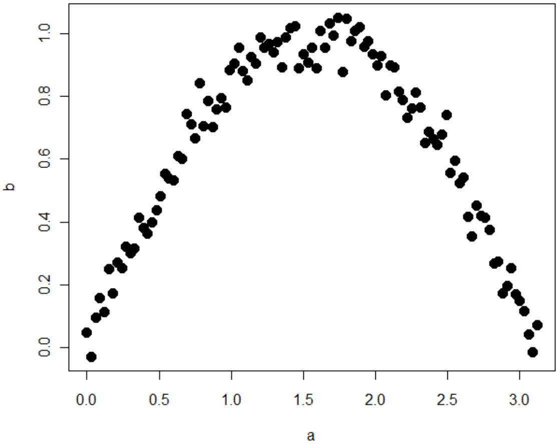


Correlation: r = -0.8, r = -0.4





Correlation: $r \approx 0$



Common Question

How much of the variation in Y is explained by X?

R^2

- The **squared correlation coefficient** (r^2 or R^2) is the proportion of variation in Y accounted for by the linear relationship with X

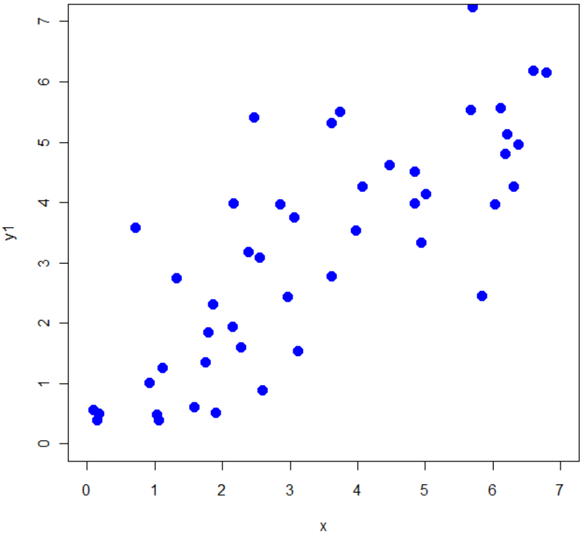
$$R^2 = 1 - \sum e_i^2 / \sum (Y_i - \bar{Y})^2 = 1 - SSError/SSTotal \quad (18)$$

- R^2 is between 0 and 1
- R^2 is a commonly reported measure of model fit
- $SSError = (1 - R^2)SSTotal$

Terminology

- R^2: Coefficient of Determination

Proportion of Variation Explained = 0.64



Proportion of Variation Explained

- $R^2 = .64$ indicates 64% of the variation in Y is accounted for by the line

Summary of Simple Linear Regression

- A simple linear regression assumes that $Y = \beta_0 + \beta_1 X + \epsilon$
- The **Least Squares** method estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$
- The least squares equation, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, can be used for prediction
- The correlation measures the strength of the **linear** relationship

Inference

- We have discussed how to estimate regression coefficients.
- How precise are these estimates?
- **Statistical Inference** is the process of drawing conclusions from data

Two Types of Inference

1. Confidence Intervals
2. Hypothesis Tests

Standard Error

- All estimates have variability associated with them
- The **standard error** of an estimate gives an idea of how much the statistic would vary from sample to sample

Standard Error of $\hat{\beta}_1$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\Sigma(X_i - \bar{X})^2}} = \frac{\hat{\sigma}}{\sqrt{(n-1)\mathrm{Var}(X)}} \tag{19}$$

Confidence Intervals Overview

- Confidence intervals give a range of reasonable values for a parameter
- Example: Instead of saying only that $\hat{\beta}_1 = 4$, we could say that a 95% confidence interval for β_1 is (3.2, 4.8)

Confidence Interval for β_1

- Take best guess, and go up and down "a few" std. errors

$$\hat{\beta}_1 \pm t^\star se(\hat{\beta}_1) \tag{20}$$

$t^\star$ depends on

1. Confidence Level (90%, 95%, etc)
2. Sample Size

Confidence Intervals and Hypothesis Tests

- Confidence intervals give range of reasonable values for β_1
- Hypothesis tests help us decide if a particular value (usually 0) is reasonable
- Hypothesis tests use **test statistics** to make a decision
- A **test statistic** is a summary of the data that helps us make a decision

Hypothesis Tests

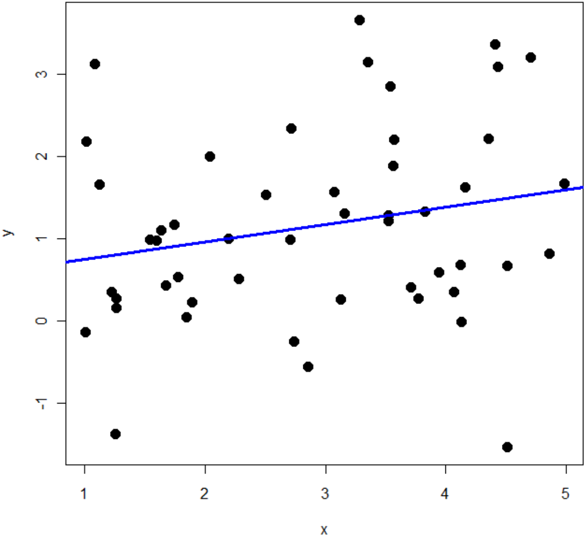
- How do we know if X has **any** effect on Y?

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{21}$$

- If $\beta_1 = 0$, then X has no effect on Y

- A common hypothesis test is $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

Intuition of Hypothesis Test



t-test for Regression Coefficient

Hypotheses

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$

Find Test Statistic

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \tag{22}$$

P-value

- The **P-value** is the probability of getting such an extreme result if H_0 is true.

Distributions

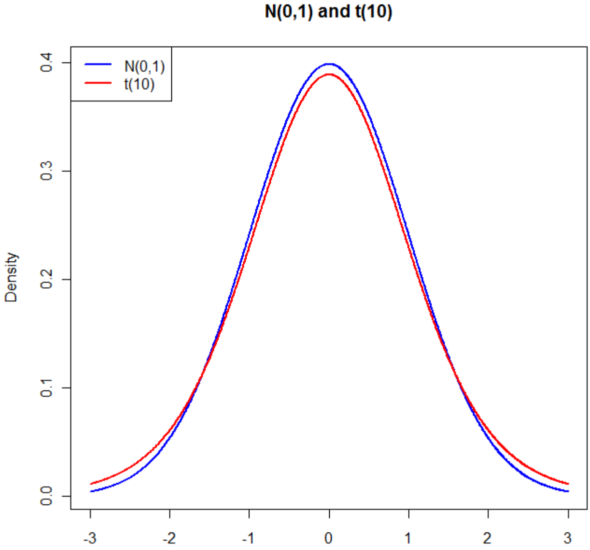
Definition

- A **distribution** describes the possible values of a random quantity.

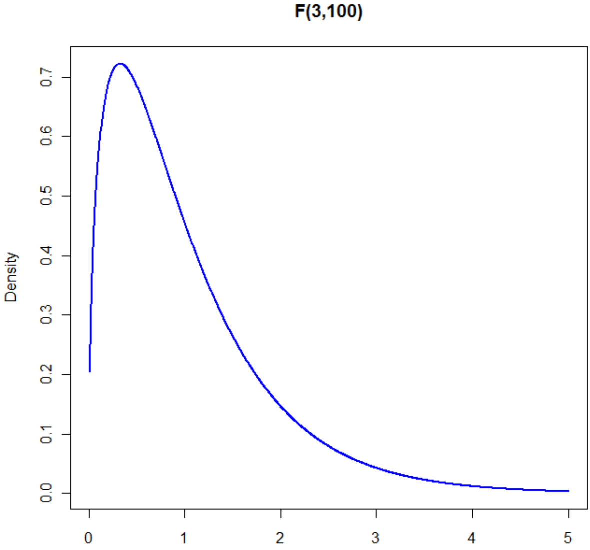
Regression Distributions

- Normal(μ, σ)
- t(df)
- F(df_1, df_2)

Standard Normal and t Distributions



F($df_1 = 3$, $df_2 = 100$)



Distribution of t Statistic

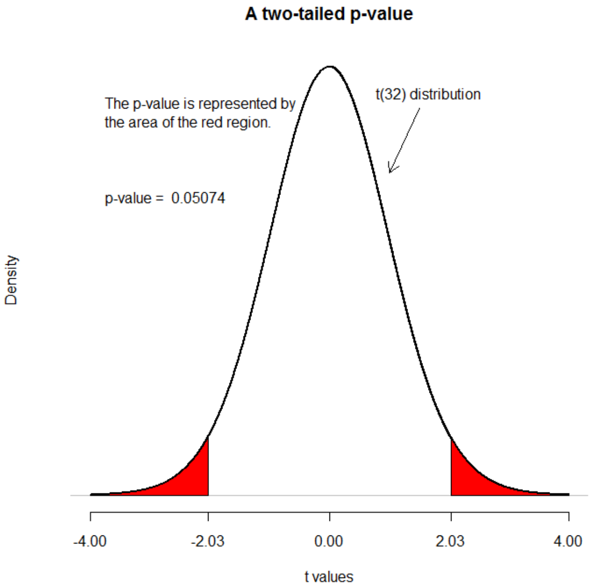
Hypotheses

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$

Distribution of Test Statistic

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t(df = n - 2) \tag{23}$$

P-value


A two-tailed p-value
The p-value is represented by
the area of the red region.
p-value = 0.05074
t(32) distribution
Density
-4.00
-2.03
0.00
2.03
4.00
t values

Results

	Est	Std Err	95% CI	t stat	P-val
Intercept	0.54	0.43	(-0.33, 1.40)	1.25	0.22
Predictor	0.21	0.14	(-0.06, 0.49)	1.55	0.13

$$R^2 = 0.19,\ \hat{\sigma} = 0.35$$

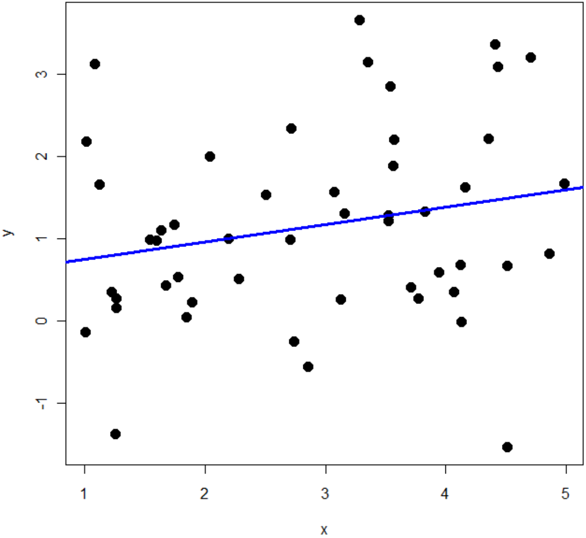
- Always report confidence intervals in addition to P-values!

Overall Test

Question

- Does the model explain more variation in Y than would be expected by chance?

Intuition of Overall Test



Decomposition of Variance

Total Variation

$$\mathrm{SSTotal} = \sum (Y_i - \bar{Y})^2$$

Variation Explained by Model

$$\mathrm{SSReg} = \sum (\hat{Y}_i - \bar{Y})^2$$

Variation Due to Error

$$\mathrm{SSError} = \sum (Y_i - \hat{Y}_i)^2$$

Decomposition

$$\text{SSTotal} = \text{SSReg} + \text{SSError}$$

F Test

Hypotheses

- H_0: Model explains no variation in Y
- H_a: Model explains some variation in Y

Test Statistic

$$F = \frac{\text{SSReg}/1}{\text{SSError}/(n-2)}$$

Analysis of Variance (ANOVA) table

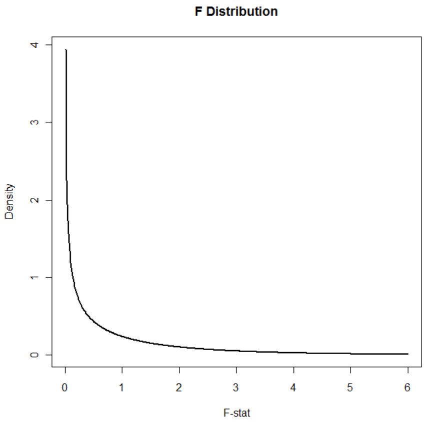
	Df	Sum Sq	Mean Sq	F value	P-val
Regression	1	32.27	32.27	5.34	0.0264
Error	38	229.60	6.04		

Conclusion

- Since $p \leq 0.05$, conclude that the model explains some variation in Y.

F Distribution

- In Simple Linear Regression $F \sim F(1, n-2)$



Simple Regression Results

Table: Coefficients

	Est	Std Err	95% CI	t stat	P-val
Intercept	2.85	0.39	(2.05, 3.64)	7.26	0.000
Predictor	1.27	0.55	(0.16, 2.39)	2.31	0.026

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	P-val
Regression	1	32.27	32.27	5.34	0.0264
Error	38	229.60	6.04		

Confidence Bands For Regression Line

Two Lines

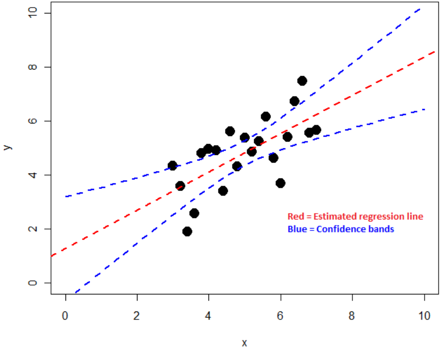
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{24}$$

$$E[Y] = \beta_0 + \beta_1 X \tag{25}$$

Question

- How close is the fitted line to the true line?

Confidence Bands for Regression Line



Confidence Bands for Regression Line

Narrow Intervals

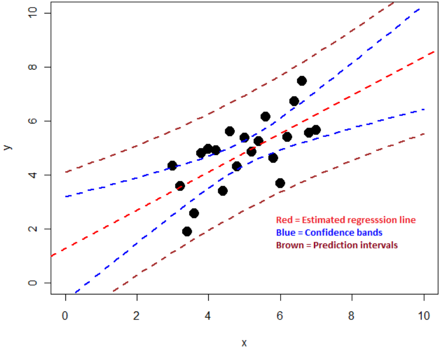
1. X close to $\bar{X}$
2. Large n
3. Small $\hat{\sigma}$

Formula

$$\text{Best Guess} \pm \text{(A few)*std. errors}$$

$$\hat{y} \pm t^{\star}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}} \tag{26}$$

Prediction Intervals



Prediction Intervals

Narrow Prediction Intervals

1. X close to $\bar{X}$
2. Large n
3. Small $\hat{\sigma}$
4. Prediction intervals are wider than confidence bands for population line.

Formula

Best Guess $\pm$ (A few)*std. errors

$$\hat{y} \pm t^{\star}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}} \tag{27}$$

Summary of Regression Inferences

- Estimated regression coefficients have a sampling error.
- Confidence intervals and hypothesis tests are tools for making inferences in the presence of sampling error.
- Confidence bands quantify uncertainty in the estimated regression line.
- Prediction intervals quantify uncertainty in the estimated value of y for a given value of x.

Computer Lab #1

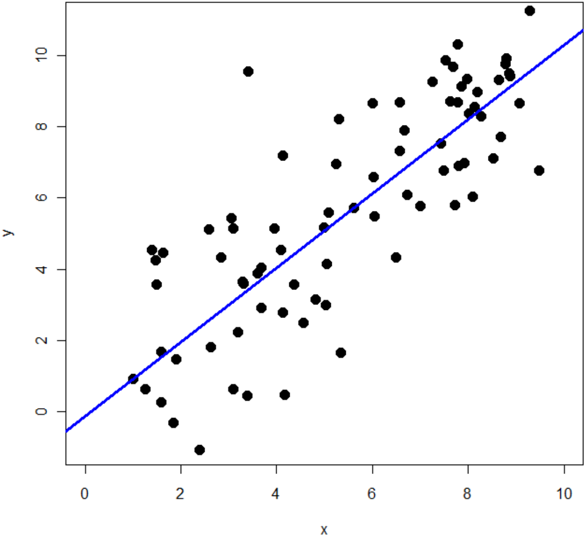
Simple Regression

Diagnostics

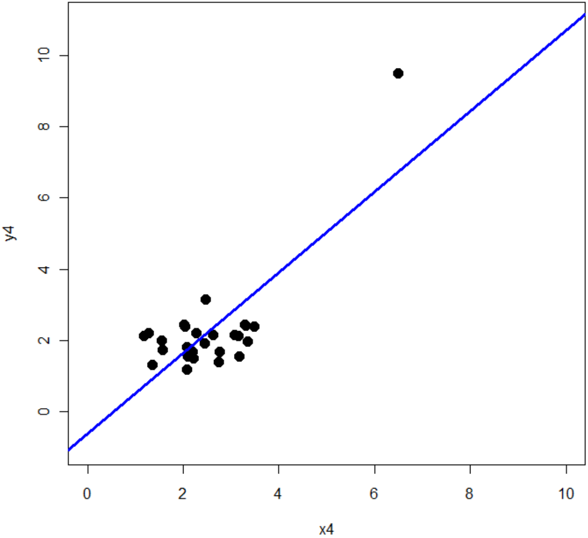
Are the results meaningful and appropriate?

Look at the following 3 graphs: Are they the same?

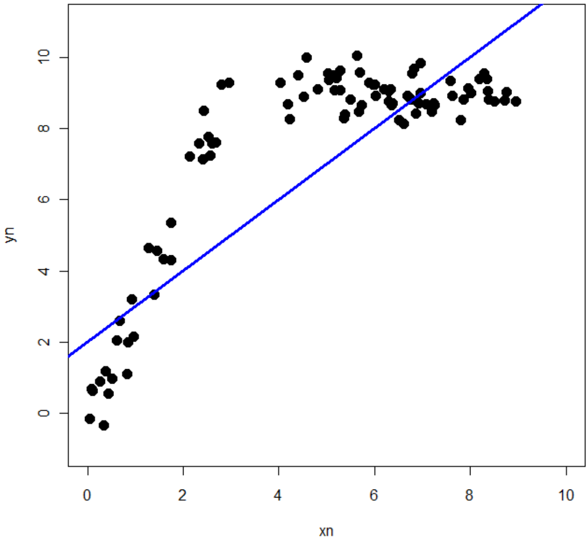
$\hat{\beta}_1 = 1,\ r^2 = 0.8$



$$\hat{\beta}_1 = 1,\ r^2 = 0.8$$



$\hat{\beta}_1 = 1,\ r^2 = 0.8$



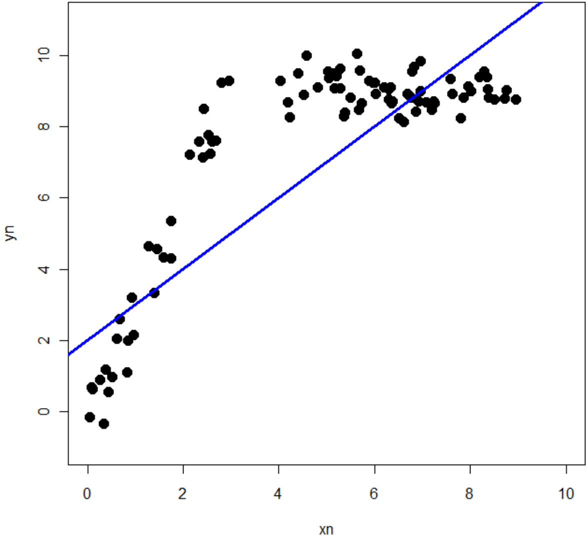
Regression Assumptions

- Sample is representative of population
- The relationship between X and $E[Y]$ is linear
- The errors are independent
- The errors have constant variance
- The errors are normally distributed
  - Not important with large sample sizes!

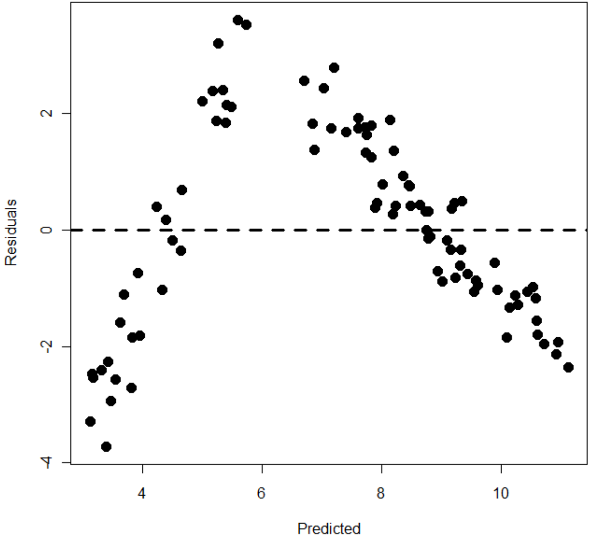
Checking Linearity Assumption

- Scatter plot of Y versus X
- **Residual plots**: residuals (e) versus predicted values ($\hat{Y}$)
- Residuals should be randomly scattered around 0

Checking Form: Scatter Plot of Y versus X



Checking Form: Residual Plots



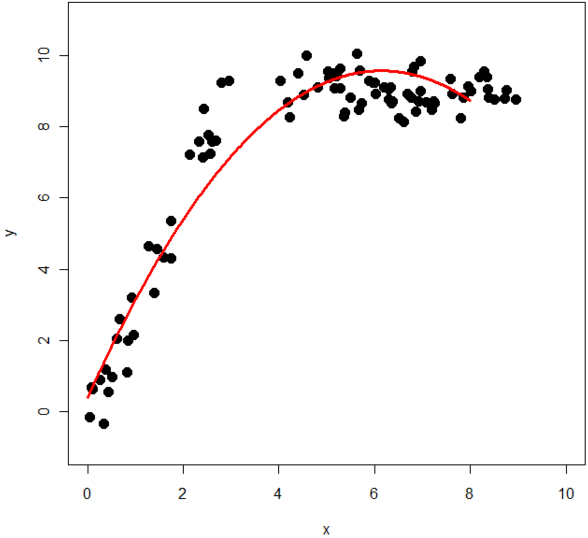
Remedial Measures

What if the diagnostic plots show linearity is violated?

One solution: model the relationship quadratically

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \tag{28}$$

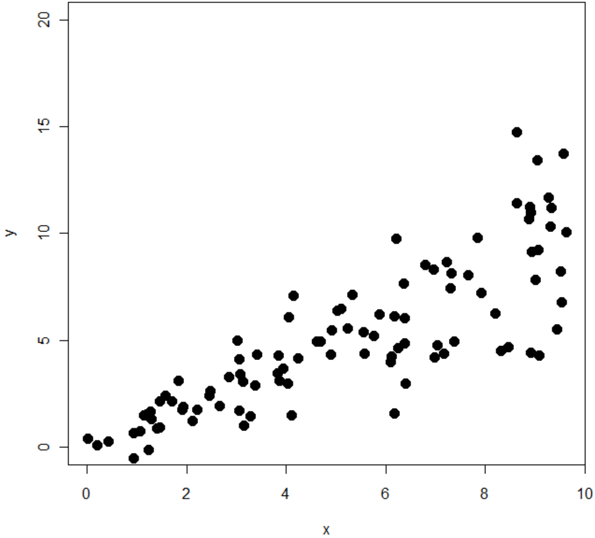
A quadratic fit



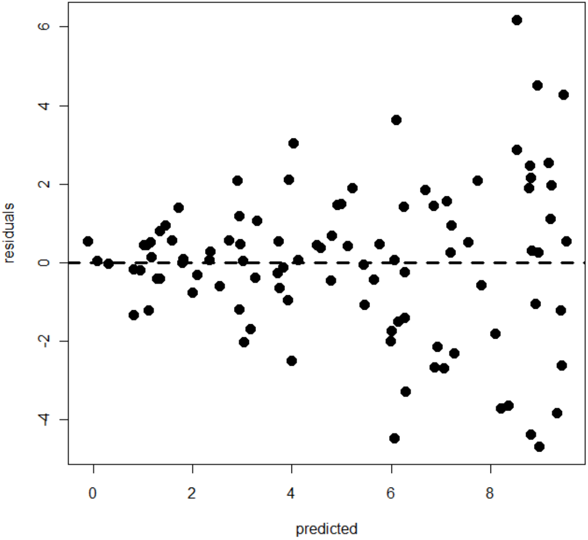
Checking Constant Variance

Is the assumption of constant variance met?

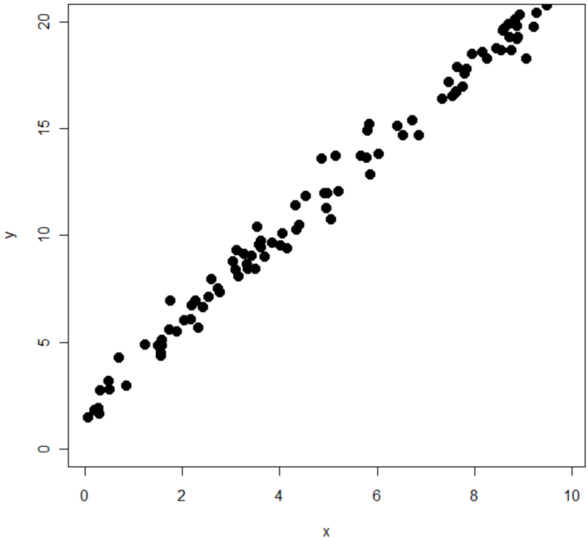
Scatterplot #1 – Constant Variance?



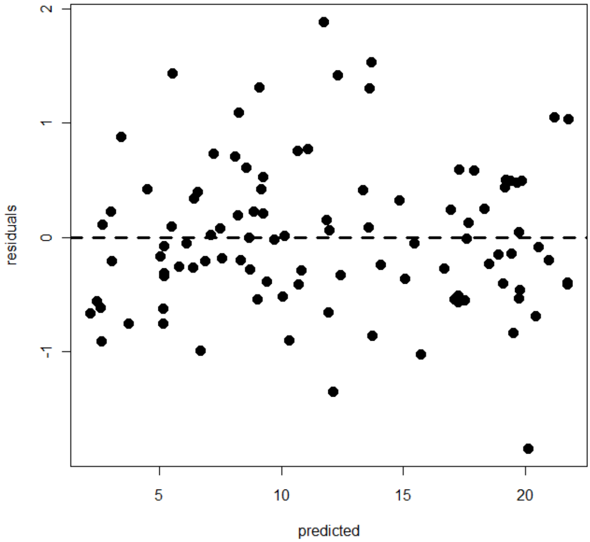
Residual plot #1 – Constant Variance?



Scatterplot #2 – Constant Variance?



Residual plot #2 – Constant Variance?



Heteroscedasticity

- **Homoscedasticity** is when errors have same variance.
- **Heteroscedasticity** is when errors have different variance.
- A common example of heteroscedasticity is when there is a **mean-variance relationship.**
- Heteroscedasticity threatens the accuracy of inferences.

Transformations

Replaces variable with some function of that variable

Examples

1. $Y \to \sqrt{Y}$
2. $X \to \log(X)$

A tranformation may help with:

1. Heteroscedasticity
2. Lack of fit to a straight line

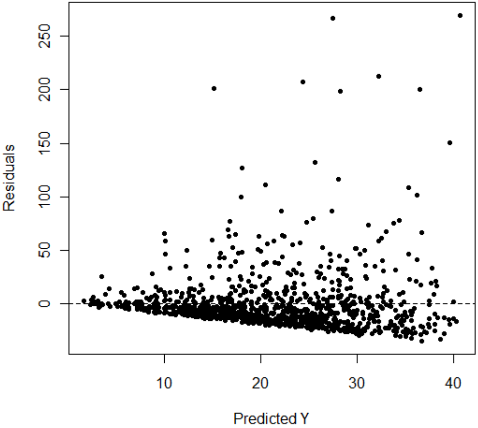
Transformation Example

	Y	X_1
1	14.9	3.8
2	34.2	4.5
3	4.0	3.4
4	94.9	3.1
5	28.0	2.6
6	2.7	2.8

Original Model

$$\hat{Y} = 24.5 + 7.72X_1 \tag{29}$$

Residuals versus Predicted

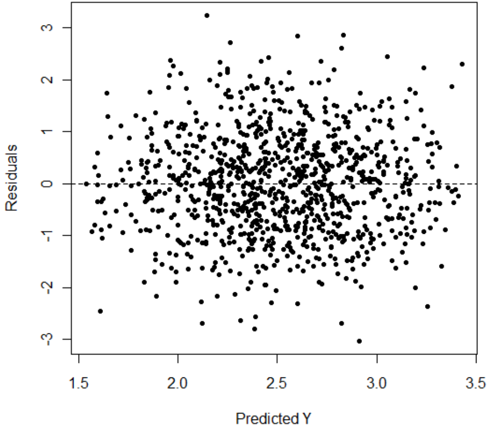


Transformed Model

- Use $\log(Y)$ as outcome, and regress against X_1.

$$\log(\hat{Y}) = 2.65 + 0.33X_1 \tag{30}$$

Residuals versus Predicted



Back Transforming

$$\log(\hat{Y}) = 2.65 + 0.33X_1 \tag{31}$$
$$\hat{Y} = e^{2.65}e^{.33X_1} \tag{32}$$

Interpretation

- For every unit increase in X_1, $\hat{Y}$ increases by factor of $e^{.33} = 1.39$ or (39%).

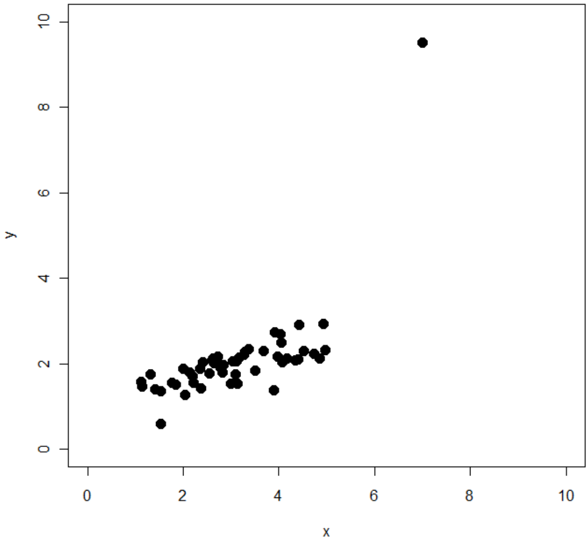
Which Transformation?

$$\sqrt{Y} \text{ or } \log(Y) \text{ or } Y^3 \text{ or } Y^2 \text{ or } Y^{-1}$$

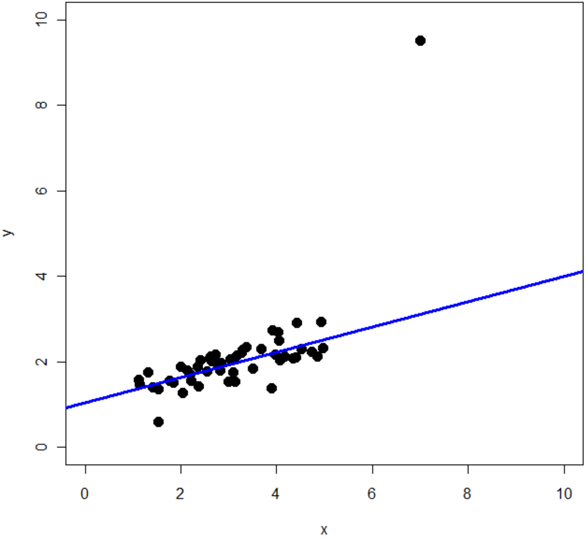
Considerations

1. Quality of Fit
2. Interpretability

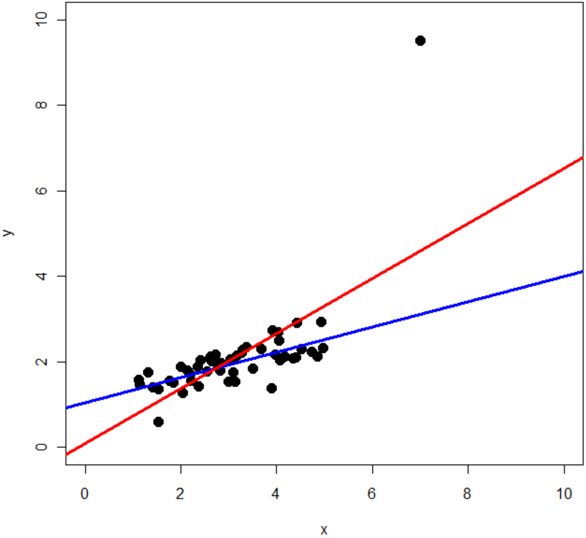
Outliers



Outliers



Outliers



Outlier Effects

- Model 1: Without Outlier
- Model 2: With Outlier

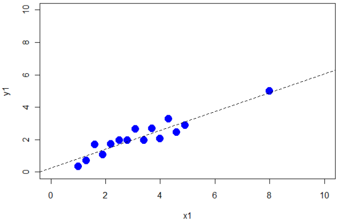
	M1	M2
$\hat{\beta}_0$	0.91	0.00
$\hat{\beta}_1$	0.32	0.67
r^2	0.64	0.50
$\hat{\sigma}$	0.26	0.83

Leverage and Influence

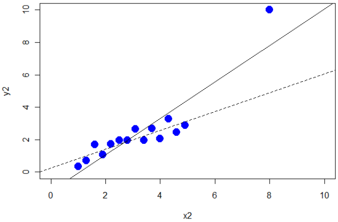
- **Leverage** measures how far each point is from $\bar{X}$.
- **Influence** measures how each point changes the fitted line.
- High Influence $\Longrightarrow$ High Leverage
- High Leverage $\Longrightarrow$ High Influence ?

Leverage and Influence

High Leverage



High Influence



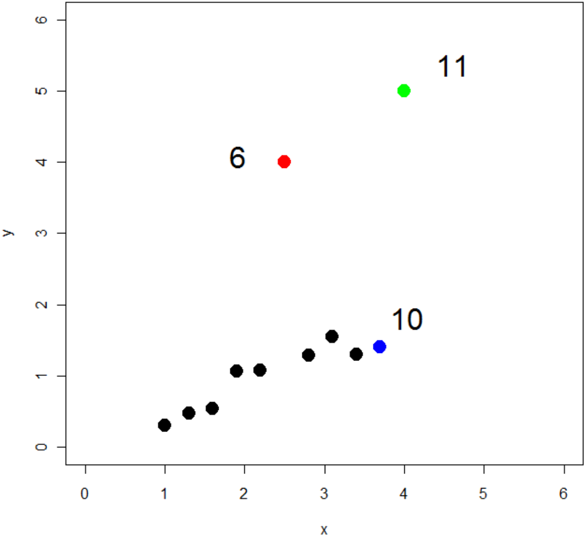
Cook's Distance

- Each point $i = 1, ..., n$ has a Cook's Distance.
- Measures impact of deleting each point
- "Large" values of Cook's Distance warrant investigation.
- Rules of thumb for "large"
    1. $D_i > 1$
    2. $D_i > 4/n$

DFFITS and DFBETA

- **DFFITS**: effect of deleting each observation on fitted values
  1. Rule of thumb for "large": $|DFFITS| > 2\sqrt{p/n}$, where p is the number of parameters in the model.
- **DFBETA**: effect of deleting each observation on coefficient estimates
  1. Rules of thumb for "large": $|DFBETA| > 1$ or $|DFBETA| > 2/\sqrt{n}$.
- Cook's, DFFITS, and DFBETA are **Leave-One-Out** diagnostics.

Example of Influence Diagnostics



Example of Influence Diagnostics

	Obs	Lev	DFBETA	DFFITS	Cook
	1	0.32	0.01	0.01	0.00
	2	0.24	0.08	0.08	0.00
	3	0.17	-0.04	-0.04	0.00
	4	0.13	-0.16	-0.20	0.02
	5	0.10	-0.01	-0.02	0.00
$\star$	6	0.09	0.30	0.84	0.21
	7	0.10	-0.01	-0.14	0.01
	8	0.13	0.06	-0.30	0.05
	9	0.17	0.13	-0.35	0.06
$\star$	10	0.24	0.38	-0.74	0.25
$\star$	11	0.32	-0.93	1.54	0.82

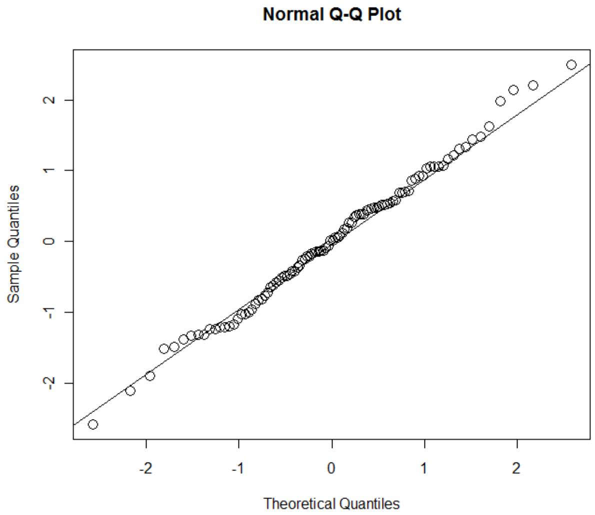
What to do with influential points?

1. Investigate!
2. Consider robust methods.
3. Do **not** remove without consideration.

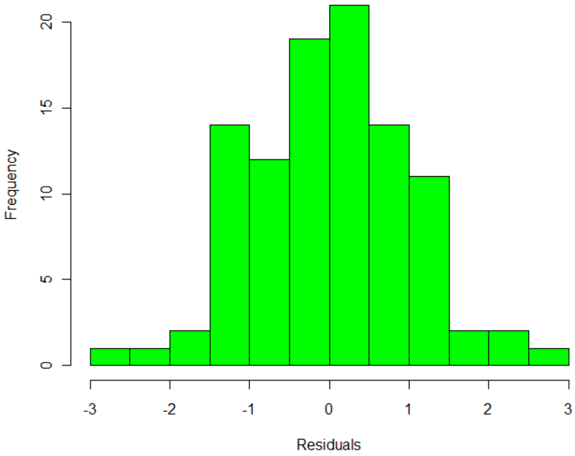
Error Distribution

- Regression assumes errors are normally distributed
- Assess with QQ plot and histogram

QQ plot of residuals



Histogram of residuals



The Normality Assumption

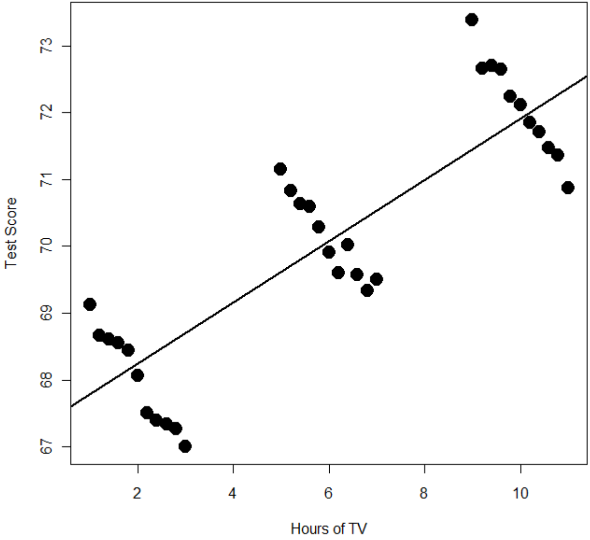
- The assumption is **not** that Y is normal
- The assumption is that Y varies from its mean normally (i.e. the errors are normal)
- Because of the **Central Limit Theorem**, error normality is **not** crucial in large samples

Omitted Variables and Confounding

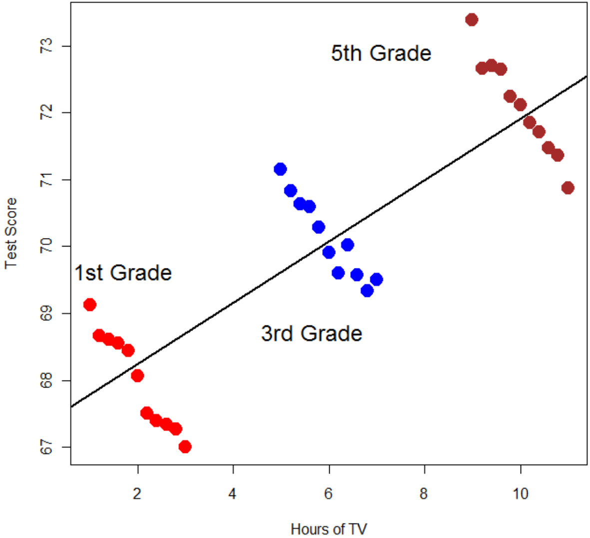
Omitted variables can obscure the relationship between X and Y

Reading Score vs Hours of TV

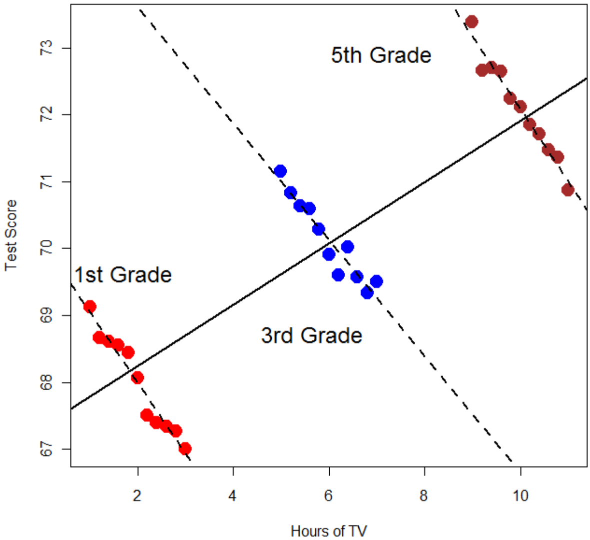
Reading Score vs Hours TV



Reading Score vs Hours TV



Reading Score vs Hours TV



Confounders

Definition

- A **confounder** is a variable correlated with both X and Y

Potential Problems

- Reverse direction of a relationship
- Create false appearance of a relationship
- Create inaccurate estimates

Correlation $\not\Rightarrow$ Causation

- Regression results are **associative** rather than **causal.**

Motivation

- How does a new medical treatment affect patient outcomes?

Methods for Causal Inference

- Statistical Adjustment
- Matching
- Propensity Scores

Categorical Predictors

Motivating Example

Suppose you are interested in understanding how average horsepower differs between three types of cars: Japanese American, and European.

Dummy Variables

Common Method

- If a categorical variable has k levels, use $k-1$ dummy variables.
- Let one category (Japanese) be the **reference category.**
- Compare other categories to reference category.
- If country $=$ American, then $X_1 = 1$, else $X_1 = 0$.
- If country $=$ European, then $X_2 = 1$, else $X_2 = 0$.

Dummy Variables

Level	X_1	X_2
Japanese	0	0
American	1	0
European	0	1

Model

Average Horsepower by Country

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{33}$$

- $X_1 = 1$ if car is American
- $X_2 = 1$ if car is European

Parameter Interpretations

- β_0: Average horsepower for Japanese cars
- $\beta_0 + \beta_1$: Average horsepower for American cars
- $\beta_0 + \beta_2$: Average horsepower for European cars

Hypothesis Testing

Model

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{34}$$

$H_0 : \beta_1 = 0$

- Do Japanese and American cars have same mean horsepower?

$H_0 : \beta_2 = 0$

- Do Japanese and European cars have same mean horsepower?

ANOVA Output

	Estimate	Std. Error	t value	p-value
(Intercept)	80.51	0.85	95.11	< 0.001
car="American"	29.25	1.20	24.44	< 0.001
car="European"	1.84	1.20	1.54	0.130

Overall Test

- The dummy variable approach compares each category to a reference level
- What about an overall test for significance of country?

Hypotheses

- $H_0 : \mu_J = \mu_A = \mu_E$
- H_a: at least one μ_j is different

Alternative Expression

- $H_0 : \beta_1 = \beta_2 = 0$
- H_a: $\beta_1 \neq 0$ or $\beta_2 \neq 0$

Overall F Test

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	P-val
country	2	17445.63	8722.82	343.30	$< .001$
Residuals	87	2210.55	25.41		

Conclusion

- There is strong evidence that horsepower depends on country.

Summary: Categorical Predictors

- Categorical predictors can be included via dummy variables.
- An F test is an overall test.

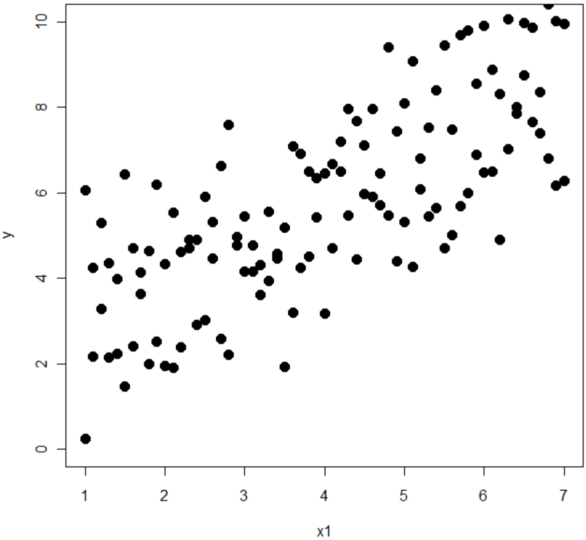
Analysis of Covariance (ANCOVA)

- Includes both categorical and continuous predictors
- Combination of regression and ANOVA

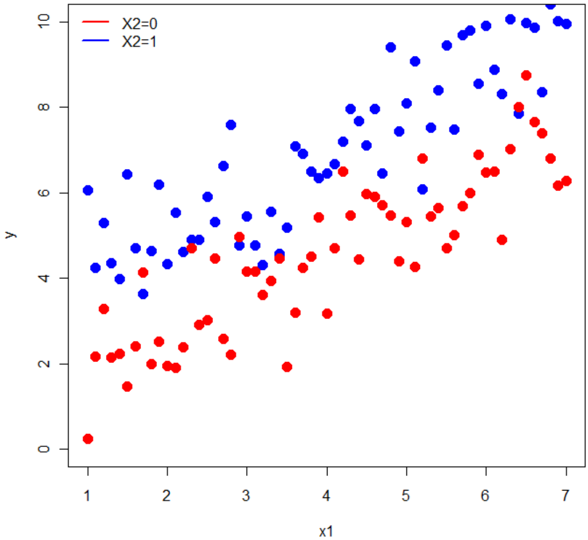
Example

- Consider the case of one continuous predictor, X_1, and one binary predictor, X_2

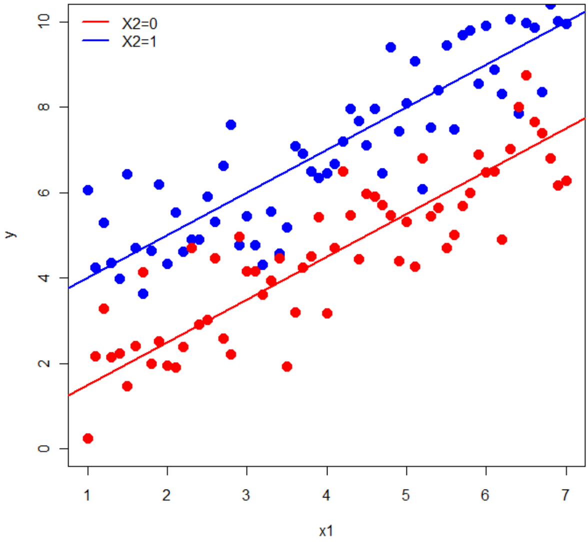
Y versus X_1



X_1 continuous, X_2 binary



X_1 continuous, X_2 binary



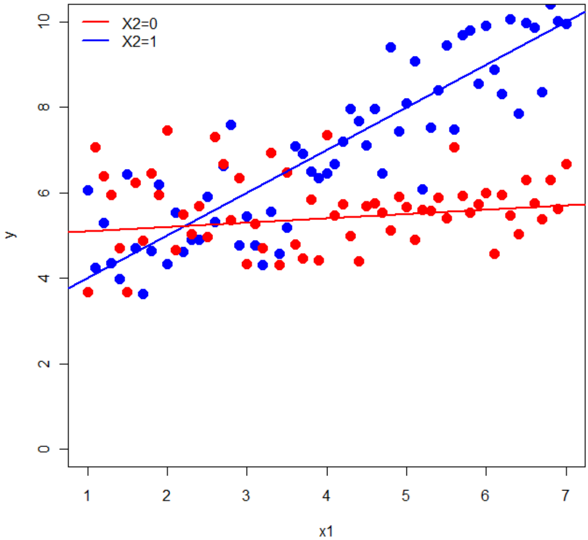
ANCOVA Model

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{35}$$

- X_1 is continuous.
- X_2 is a 0/1 dummy variable.

- β_1 is the slope of the line.
- β_2 is the vertical distance between the two lines.

What if the lines are not parallel?



ANCOVA Interaction Model

Effect of X_1 depends on X_2

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{36}$$

For $X_2 = 0$,the effect of X_1 is β_1.

For $X_2 = 1$, the effect of X_1 is $\beta_1 + \beta_3$.

Question: What is the interpretation of β_2?

Computer Lab #2

Diagnostics, categorical variables, and ANCOVA (interactions)

Simple and Multiple Regression

Simple Regression Model

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{37}$$

Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{38}$$

Motivating Example

What factors influence student achievement?

Goals

1. Learn about relationship between $X_1, ..., X_p$ and Y
2. Estimate β_i with $\hat{\beta}_i$
3. Construct confidence intervals for β_i
4. Determine strength of relationships

Sample Data Set

Y	X1	X2	X3
8.18	9.08	3.81	11.96
7.02	10.69	5.22	11.28
9.52	10.71	3.68	13.08
8.54	8.69	2.58	11.96
7.84	7.26	2.17	12.33
0.95	10.61	3.33	13.08

Summarize Each Variable Separately

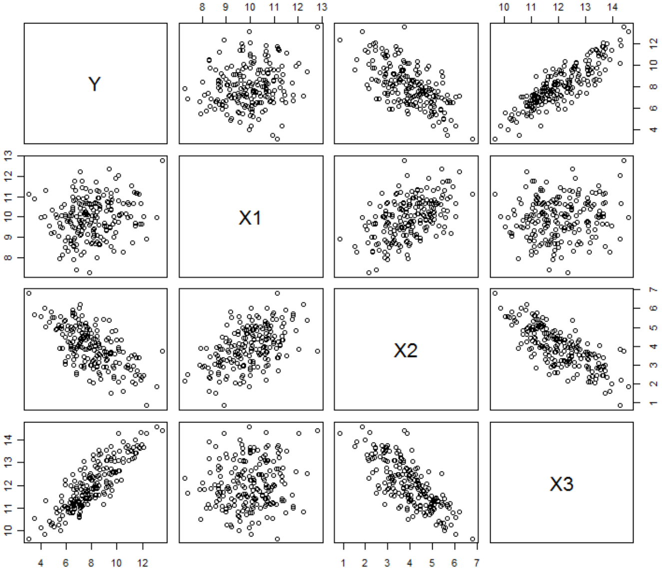
Graphical Summaries

- Histogram
- Boxplot

Numerical Summaries

- Mean
- Quartiles
- Minimum
- Maximum
- Standard Deviation

Scatterplot Matrix



Correlation Matrix

	Y	X1	X2	X3
Y	1.00	0.16	-0.64	0.86
X1	0.16	1.00	0.50	0.11
X2	-0.64	0.50	1.00	-0.75
X3	0.86	0.11	-0.75	1.00

Method: Least Squares

Sum of Squared Residuals

$$\Sigma e_i^2 = \Sigma(Y_i - \hat{Y}_i)^2 \tag{39}$$

Predicted Values

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p \tag{40}$$

Interpretations

- $\hat{\beta}_i$ is the predicted change in Y for every one unit increase in X_i, holding all other variables constant

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p \tag{41}$$

Example Interpretation

Y	X1	X2	X3
8.18	9.08	3.81	11.96
7.02	10.69	5.22	11.28
9.52	10.71	3.68	13.08
8.54	8.69	2.58	11.96
7.84	7.26	2.17	12.33
0.95	10.61	3.33	13.08

$$\hat{Y} = -7.02 + 0.51X_1 - 0.62X_2 + 1.04X_3 \tag{42}$$

- The predicted value of Y increases by 0.51 for every unit increase in X_1, holding all other variables constant.

Prediction

- Use fitted line to predict Y for new observations.
- Individual with $X_1 = 10$, $X_2 = 4$, $X_3 = 11$

$$\hat{Y} = -7.02 + 0.51X_1 - 0.62X_2 + 1.04X_3 \tag{43}$$

$$\hat{Y} = -7.02 + 0.51(10) - 0.62(4) + 1.04(11) = 6.93 \tag{44}$$

Results

1. Coefficient estimates $\hat{\beta}_0$, $\hat{\beta}_1$, ... ,$\hat{\beta}_p$
2. Standard errors of coefficients
3. Confidence intervals for population coefficients
4. Results from $H_0 : \beta_i = 0$ for $i > 0$

Sample Output

$$\hat{Y} = -6.15 + 0.54X_1 - 0.71X_2 + 0.96X_3 \tag{45}$$

	Est	SE	Lower CI	Upper CI	t value	p-value
Int	-6.15	1.87	-9.84	-2.47	-3.29	< 0.001
X1	0.54	0.16	0.23	0.85	3.40	0.002
X2	-0.71	0.21	-1.13	-0.29	-3.33	0.003
X3	0.96	0.20	0.58	1.35	4.94	< 0.001

Multiple Testing

- A Type I Error is $P(\text{reject } H_0 \text{ if } H_0 \text{ is true})$.
- Every test has an α (often 5%) chance of a Type I error.
- For a **single** test, there is a 5% chance of a Type I error.
- For **ten** independent tests, there is a $1 - (1 - 0.05)^{10} \approx 40\%$ chance of at least one Type I error.

Overall Test

- T-tests consider each predictor separately.
- Multiple Testing Issue
- F test is an "overall" test

F-test for Regression

- H_0: model has no predictive power for Y
- $\beta_1 = \beta_2 = ... = \beta_p = 0$

- H_a: model has some predictive power
- At least one non-intercept $\beta_i \neq 0$

F-test for Regression

- $p = (\#\text{ of predictors})$
- n = sample size
- Test statistic has a $F(p, n - p - 1)$ distribution

Example

- $p = 3,\ n = 100$
- $F(3, 96) = 2.31$, p-val $= 0.081$

Conclusion

- Not enough information (at $\alpha = 0.05$ level) to conclude that model has any predictive ability

Multiple R^2

- Proportion of variation in Y explained by model

- If $R^2 = 0.74$, then 74% of the variation in Y is explained by the model.

- $R^2 = \text{cor}(\hat{Y}, Y)^2$

Multiple Regression Diagnostics

- Form of Model
- Error Distribution
- Confounding
- Correlation among predictors

Independent Predictors

- If predictors are **independent**, then multiple regression and simple regressions yield **same** estimated coefficients.

Multiple Regression

$$\hat{Y} = 0.14 + 0.48X_1 - 1.02X_2 \tag{46}$$

Simple Regressions

$$\hat{Y} = -5.65 + 0.48X_1 \tag{47}$$

$$\hat{Y} = 4.96 - 1.02X_2 \tag{48}$$

Correlated Predictors

- If predictors are **correlated**, then multiple regression and simple regressions can yield **very different** results.

Multiple Regression $\quad \mathrm{cor}(X_1, X_2) = .8$

$$\hat{Y} = -0.05 + 0.52X_1 + 1.03X_2 \tag{49}$$

Simple Regressions

$$\hat{Y} = 4.00 - 0.50X_1 \tag{50}$$

$$\hat{Y} = 3.08 + 0.68X_2 \tag{51}$$

Multicollinearity

- **Multicollinearity** is when two or more predictors are highly correlated.

Consequences

- Limits ability to estimate effects of individual predictors
- Coefficient estimates have high variability.
- Can still use model to make predictions

Detecting Multicollinearity

1. Correlation Matrix

	Y	X1	X2	X3
Y	1.00	0.16	-0.64	0.86
X1	0.16	1.00	0.50	0.11
X2	-0.64	0.50	1.00	-0.75
X3	0.86	0.11	-0.75	1.00

Variance Inflation Factors

- Pairwise Correlations do not fully capture multicollinearity.

- **Variance Inflation Factors (VIF)** are a useful tool for quantifying collinearity in a data set.

- Each regression coefficient has a VIF.

Calculating VIF for X_j

1. Temporarily treat X_j as response.
2. Regress X_j against all other predictors.
3. R_j^2 is the multiple R^2 for this regression.
4. High R_j^2 means X_j is affected by multicollinearity.

Variation in $\hat{\beta}_j$

No Collinearity

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{(n-1)\mathrm{Var}(X_j)} \tag{52}$$

Collinearity

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{(n-1)\mathrm{Var}(X_j)} * \frac{1}{1-R_j^2} \tag{53}$$

VIF

$$\frac{1}{1-R_j^2} \tag{54}$$

VIF Example

Predictor	Estimate	VIF
X_1	$\hat{\beta}_1$	8.08
X_2	$\hat{\beta}_2$	6.24
X_3	$\hat{\beta}_3$	3.42

- Multicollinearity increases $\text{Var}(\hat{\beta}_1)$ by factor of 8.08 (808%).
- The standard error of $\hat{\beta}_1$ increases by factor of $\sqrt{8.08} = 2.84$.

Large VIFs

How large is too large?

- Various cutoffs have been proposed:
  1. VIF > 5
  2. VIF > 10.
- The choice depends on the goals of a particular analysis.

How to fix collinearity?

- Remove variables from model
- Ridge Regression
- More under Model Selection

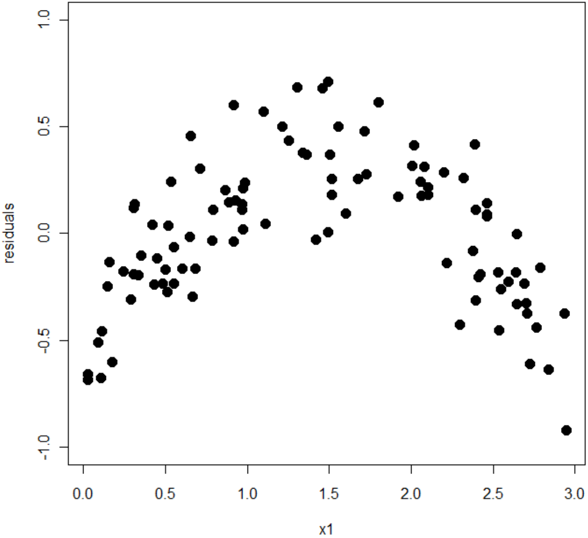
Residual Diagnostics

1. Residual Plots: e_i versus $\hat{Y}_i$

2. QQ Plot of Residuals

3. Plot residuals versus each predictor

Residual versus individual predictor (X_1)



Added Variable Plots

- A scatter plot between Y and X shows only the marginal relationship.

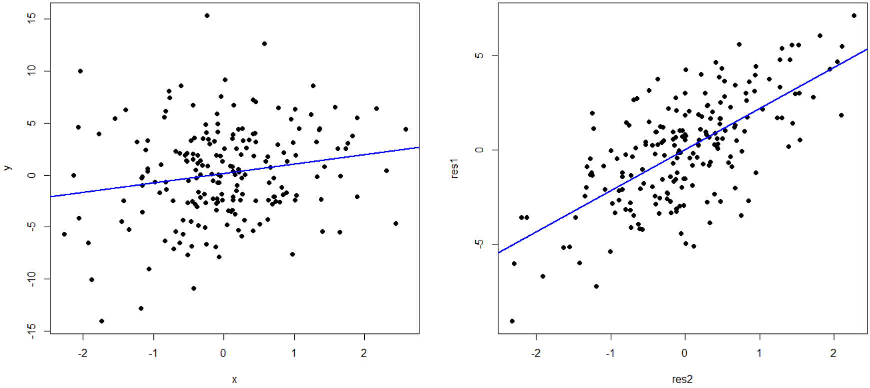
- An **added variable plot** shows relationships after adjusting for other predictors.
- One use of this plot is to assess confounding.

Added Variable Plot for X_1

Steps

1. Regress Y on X_2 and X_3.
2. Compute residuals from #1 (res1).
3. Regress X_1 on X_2 and X_3.
4. Compute residuals from #3 (res2).
5. Plot res1 versus res2.

Scatter and Added Variable Plots



Higher Order Terms

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_1^2 + \hat{\beta}_3 X_2 \quad (55)$$

Effect of X_1

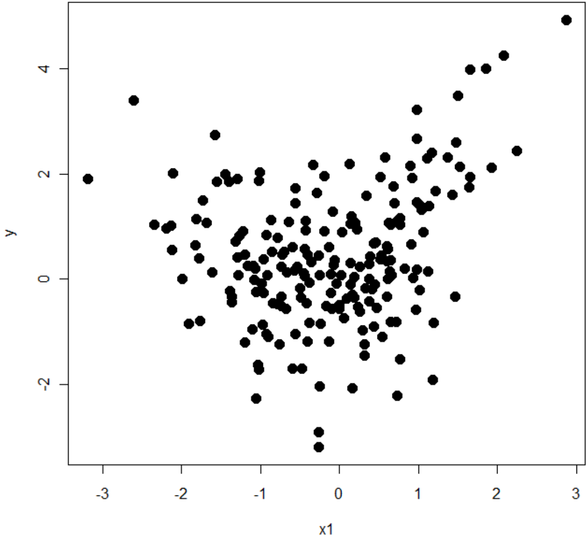
- The effect of X_1 is no longer constant.
- $\hat{\beta}_1$ should **not** be interpreted in isolation.
- For each unit increase in X_1, the predicted value of Y increases by $\hat{\beta}_1 + 2\hat{\beta}_2 X_1 + \hat{\beta}_2$.

Interactions

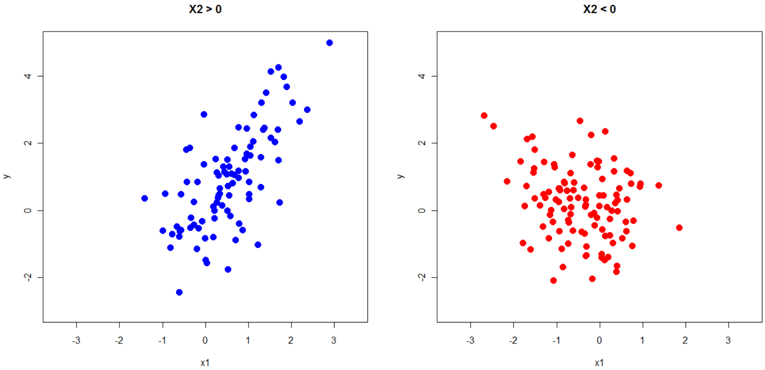
- The model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ assumes that the effect of X_1 does **not** depend on X_2.

- The following 3 graphs plot X_1 versus Y for:
  1. all values of X_2,
  2. only points where $X_2 > 0$, and
  3. only points where $X_2 < 0$.

Y versus X_1



Y versus X_1



Interaction Model

- Interactions are usually modeled with a multiplicative term.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 \qquad (56)$$

Effect of X_1

- The effect of X_1 depends on X_2
- For each unit increase in X_1, the predicted value of Y increases by $\hat{\beta}_1 + \hat{\beta}_3 X_2$.

Diagnostics

- The sample is representative of the population.
- The relationship between X and $E[Y]$ is linear.
- The errors are independent.
- The errors have constant variance.
- The errors are normally distributed.
  - Not important with large sample sizes!
- Correlation among predictors
- Interactions

Model Selection

Which Model Is Best?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_1 X_2 \tag{57}$$

OR

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 \tag{58}$$

Two Considerations

1. Which predictors to include
2. Form (X_1 or X_1^2 or $X_1 X_2$)

Model Selection

1. "Underfitting"

2. "Overfitting"

3. Model selection criteria

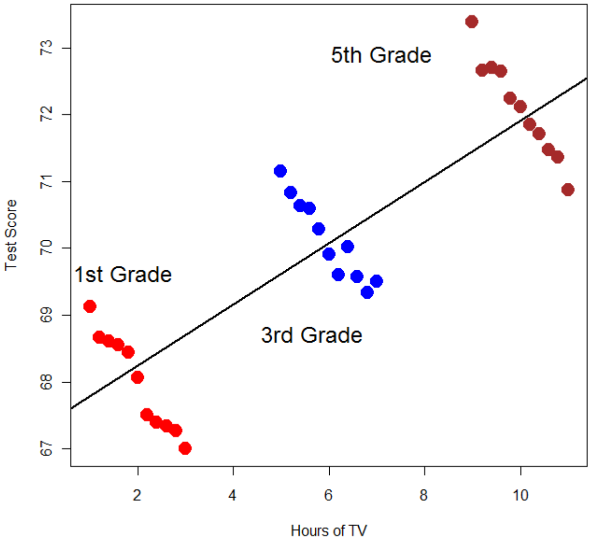
“Underfitting”

1. Omitted confounders can obscure relationships of interest.

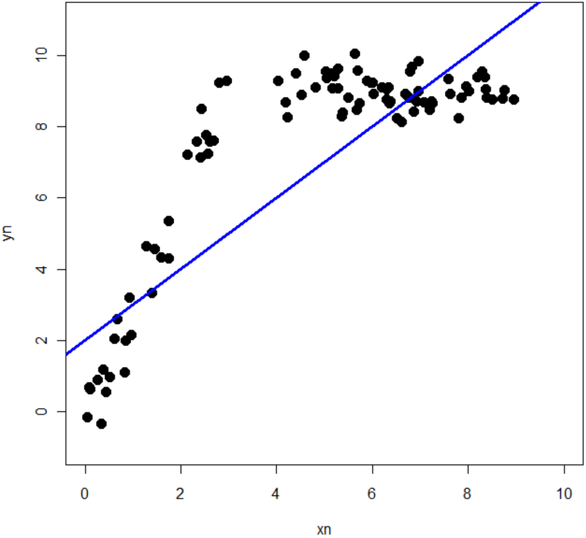
2. If assumptions of **linearity** and **additivity (no interaction)** do not hold, the true relationship may be missed.

3. Next two slides show effect of 1) omitting a confounder, and 2) falsely assuming linearity.

Omitting A Confounder



Falsely Assuming Linearity



How To Avoid Underfitting

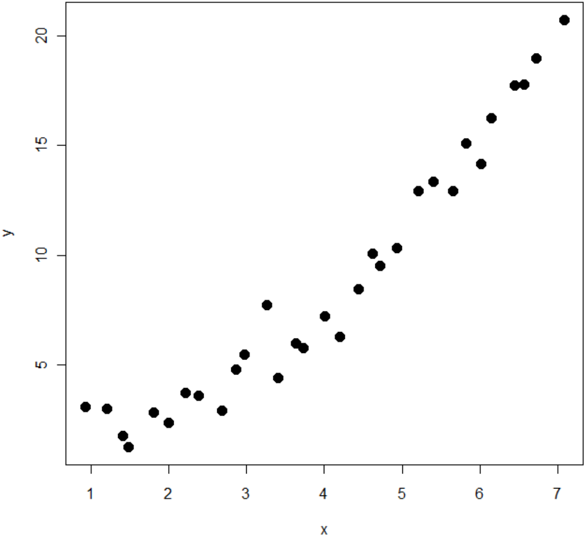
Confounding

- Do **not** rely only on bivariate relationships.
- Measure potential confounders and include in model.

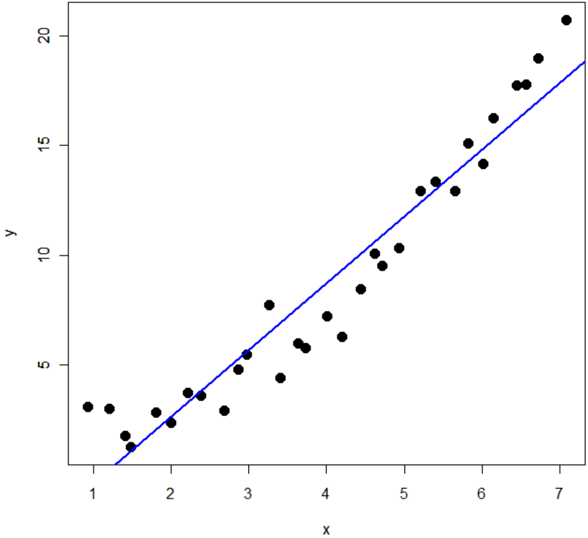
Complexity

- Test for interactions and non-linearities.

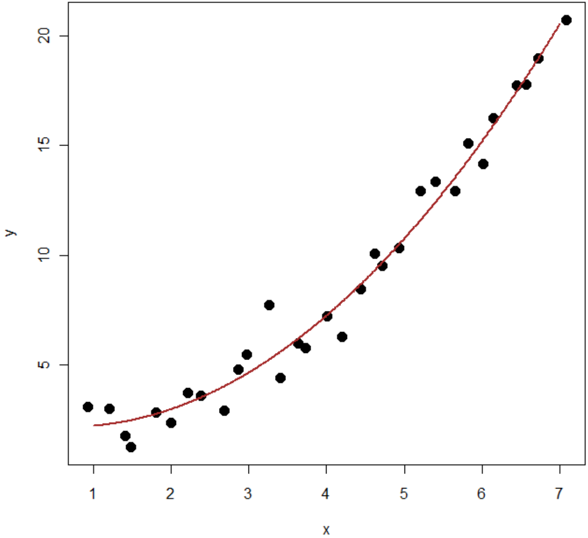
"Overfitting"



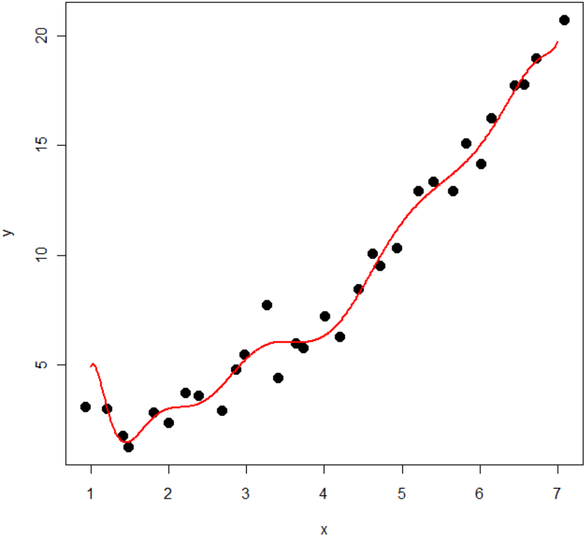
"Overfitting"



“Overfitting”



"Overfitting"



"Overfitting"

- More complex models will always "fit" the data better.
- **Overfitting** occurs when model fits random fluctuations in data.
- Overfit models may perform well on the **training** (original) data, but may perform very poorly on **test** (new) data.
- Need to balance quality of fit and **parsimony** (simplicity)

Avoid Overfitting

Rule Of Thumb

- Rule of Thumb: no more than $n/10$ or $n/20$ parameters
- If $n = 67$, no more than $3 - 6$ parameters

Adjusted R^2

- Adding a variable always increases R^2.
- The adjusted R^2 "adjusts" for model complexity.
- Adding an additional variable can decrease the adjusted R^2.
- Adjusted R^2 is a criteria for comparing two potential models.

Model Selection Criteria

1. $R^2 = \mathrm{cor}(Y, \hat{Y})^2$
2. Adjusted R^2
3. Akaike Information Criterion (AIC)
4. Bayesian Information Criterion (BIC)

Nested Models

- Two models are **nested** if predictors in one model are subset of predictors in other.

Nested Models

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
2. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Non-Nested Models

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
2. $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$

F Test for Nested Models

- If two models are **nested**, larger model will “fit” better.
- Does the improvement in fit justify additional complexity?
- Two nested models can be compared with an F test.

F Test for Nested Models

Table: Example ($n = 100$)

	Small Model	Large Model
p	3	5
Σe^2	130	120

$$F = \frac{\frac{130-120}{5-3}}{\frac{120}{100-5-1}} = 3.92 \tag{59}$$

p-val $= 0.023 \Rightarrow$ choose larger model

Criterion-Based Approach

- Choose a set of potential models that are meaningful
- Choose a final model using a criterion such as AIC

Remember:

- Don't assume linearity and additivity (no interactions)
- Avoid overfitting
- Look at model diagnostics

Model Selection

Two people modeling the same data set will usually have different final models.

"All models are wrong, but some are useful."

– George Box

Computer Lab #3

Multiple regression and model selection