Multiple Linear Regression Basics

Laura E. Barnes
&
Julianne Quinn

Agenda

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Review of EISE and Data Visualization
- Regression Overview
- Assumptions for Regression
- Parameter Estimation
- A Regression Example in R

Review of EISE

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Evidence Informed Systems Engineering
  - Problem Description
  - Evidence-Informed Approach
  - Evidence
  - Recommendation
- Evidence-Informed Approach
  - Hypothesis
  - Visualization or Graphical Analysis
  - Models and Analysis

Review of Data Visualization

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

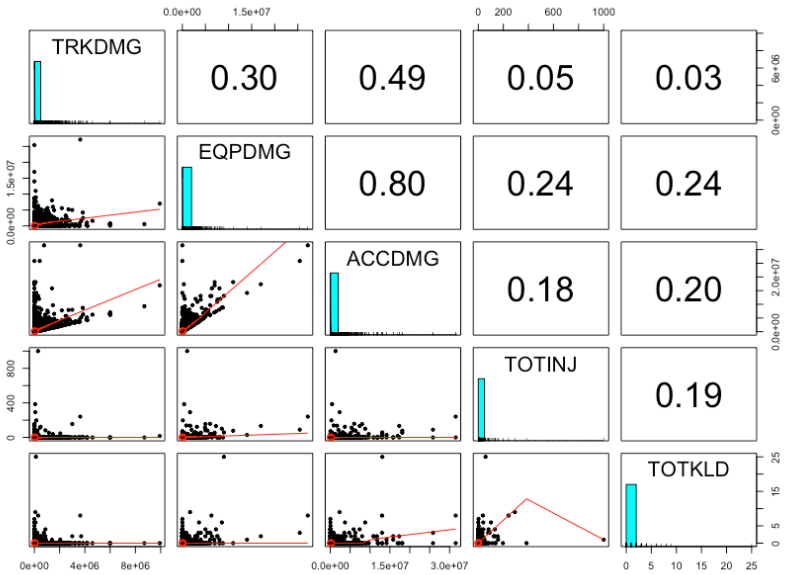
- Univariate Observation & Visualization
  - Histograms
  - Bar Plots
  - Density Plots
  - Box Plots
  - QQ Plots
- Multivariate Observation & Visualization
  - Scatter Plots
  - Scatter Plot Matrices
  - Categorical Variable Plots
  - Plots of Principal Components

Are Data Visualization and Simple Statistics Good Enough?

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Simple statistics are not
  - Sufficient for most engineering problems adjusting for confounding variables.
  - Multiplicity: even low probability events can show significance if we do enough tests.
- Regression and ANOVA provide analytical tools for understanding, prediction, and control in engineering problems.

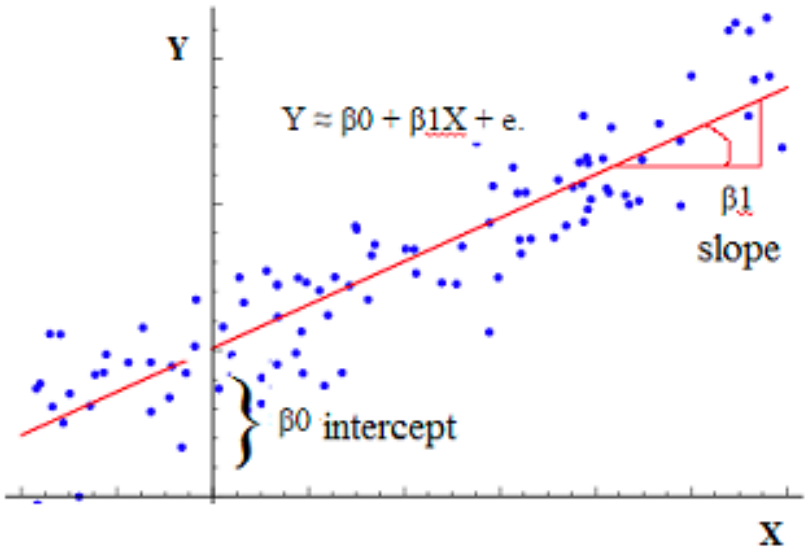


Univariate Linear Regression

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Univariate linear regression reveals the relationship between two variables
- Origin of the name "Regression"?
  - Francis Galton pioneer of statistics.



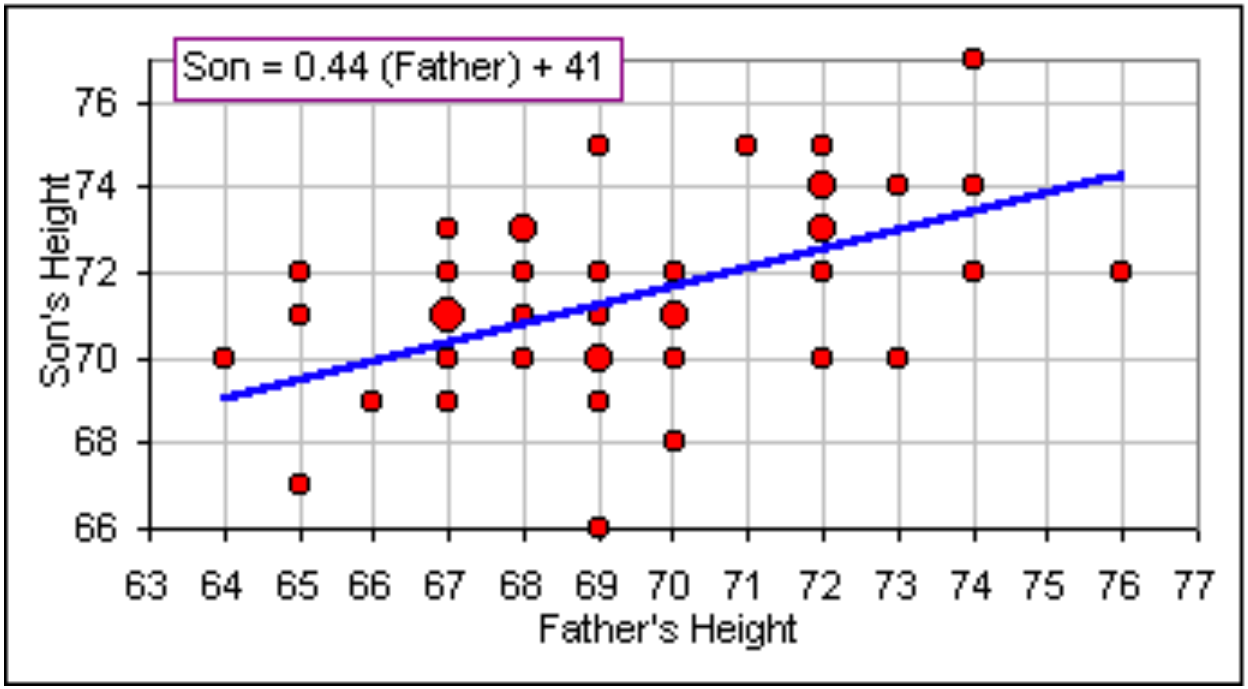
https://datacadamia.com/data_mining/simple_regression

Regression to the Mean

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- The original paper by Galton, which regressed sons heights on the heights of fathers, exposed a common fallacy: Regression to the Mean

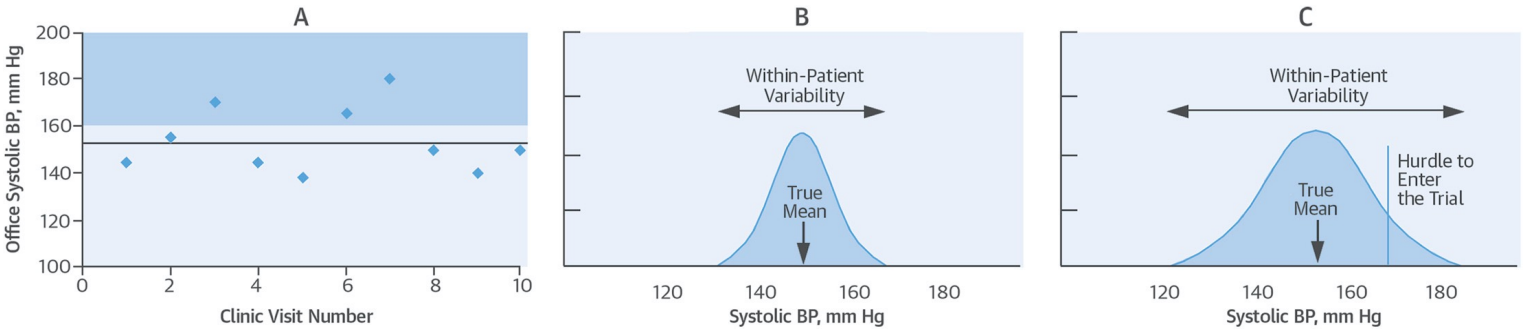


Regression to the Mean

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Clinical trial data



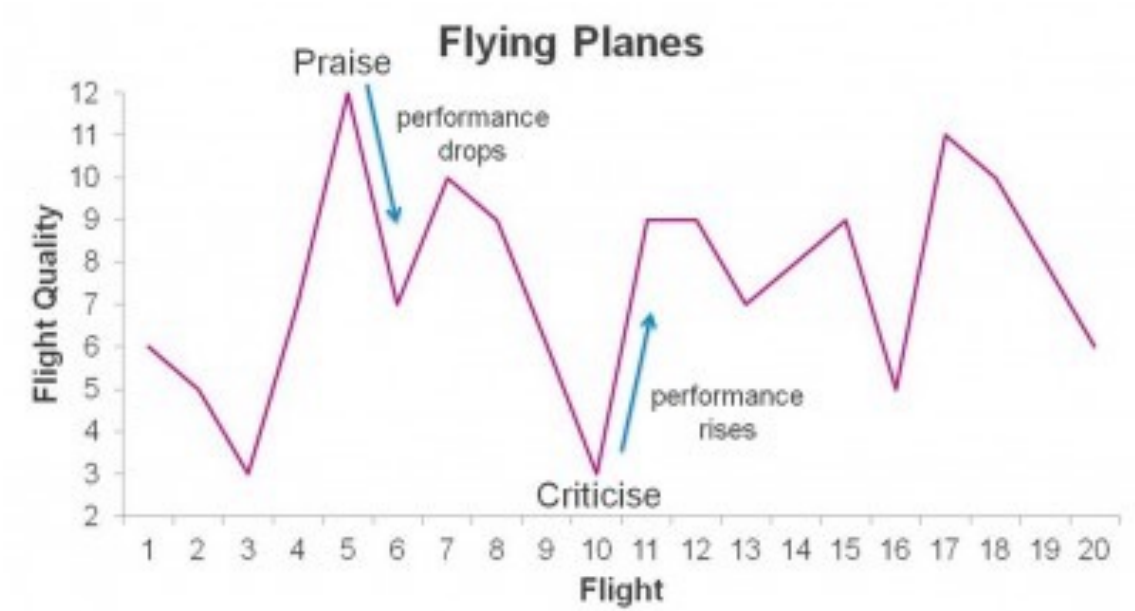
https://www.sciencedirect.com/science/article/pii/S0735109716351099

Regression to the Mean

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Another example: Israeli Air Force - Daniel Kahneman



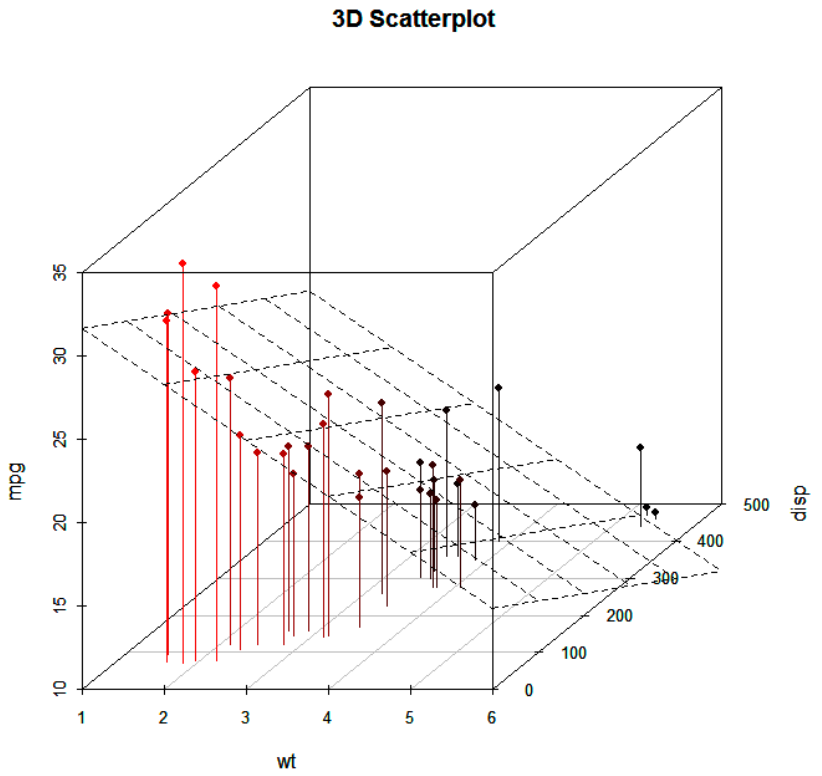
https://www.squawkpoint.com/2013/01/regression-to-the-mean/

Multiple Linear Regression Summary

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Multiple Regression: A method for measuring and modeling the relationship between sets of variables.



Multiple Linear Regression Summary

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Examples:
  - Relationship between SAT score and high school grades, gender, preparation courses, …
  - Relationship between number of crimes and incomes, population, police, …
  - Relationship between salary and years experience, gender, age, …
- Relationship does not imply causation.

Linear Regression Models

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Regression models are one type of mathematical model. Models allow us to focus attention on the key elements that describe or predict a system's performance.
- Types of mathematical models:
  - Functional: $y = f(x)$
  - Stochastic: $y = f(x) + \epsilon$ where ϵ is a random variable.
- Linear regression uses stochastic models with two components:
  - Deterministic: $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and
  - Stochastic: ϵ
- Matrix notation: $y = f(X) + \epsilon = X\beta + \epsilon$
- If we have n observations, what are dimensions of y, X, and ϵ?

Terminology of Linear Regression Models

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Linear Regression Model
  - $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$
- y: response variable, predicted variable, regressand, dependent variable, outcome variable
- x: explanatory variable, predictor variable, regressor, independent variable, input variable
- β_0: intercept
- $\beta_i (i = 1, \cdots, k)$: regression coefficients, effects
- ϵ: residual, error term, noise

Metric Goal for Regression

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Regression is the solution to an optimization problem.
- Find a linear fit to the data that minimizes the sum of squared errors.
- Why do we use the metric sum of squared errors?

Assumptions for Optimization

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- The data for the input variables or predictors, $x_1, \cdots, x_k$, are known.
- The predictors are linearly independent.
- The response variable, y, is quantitative.

Assumptions for Inference

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- For a sample size of n, the distribution of the $\epsilon_i, i = 1, \cdots, n$ are independent, identical and Gaussian with
  - $\mathrm{E}(\epsilon_\mathrm{i}) = 0$, and
  - $\mathrm{Var}(\epsilon_\mathrm{i}) = \sigma^2$
- What's the distribution of Y_i?
  - The above assumptions imply that Y_i also have Gaussian distributions with $E(Y_i) = X_i\beta$ and $\mathrm{Var(Y)} = \sigma^2$.
- Hence, Y is multivariate Gaussian with $E(Y) = X\beta$ and $\mathrm{Var(Y)} = \sigma^2$

Least Squares Estimate

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- We find optimal estimates for the coefficients where the criterion is least squares.
- The optimization problem is:

$$\text{minimize} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} - y_i)^2$$

Or

$$\text{minimize}(X\beta - y)^T(X\beta - y)$$

- What is the least squares estimate?

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Estimation of Variance

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- The estimate for σ^2 where k is the number of predictors or input variables:

$$\sigma^2 = \frac{(X\hat{\beta}-y)^T(X\hat{\beta}-y)}{n-k-1} = \frac{y^T(I-H)y}{n-k-1}$$

$H = X(X^TX)^{-1}X^T$ is the hat matrix:

$$\hat{y} = X\hat{\beta} = Hy$$

- So,

$$\sigma^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k-1}$$

Sum of Squares Decomposition

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- The sum of squares has a convenient decomposition:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2$$

Total S.S. = Residual S.S. +Model S.S.

- This decomposition is important and useful. Why?

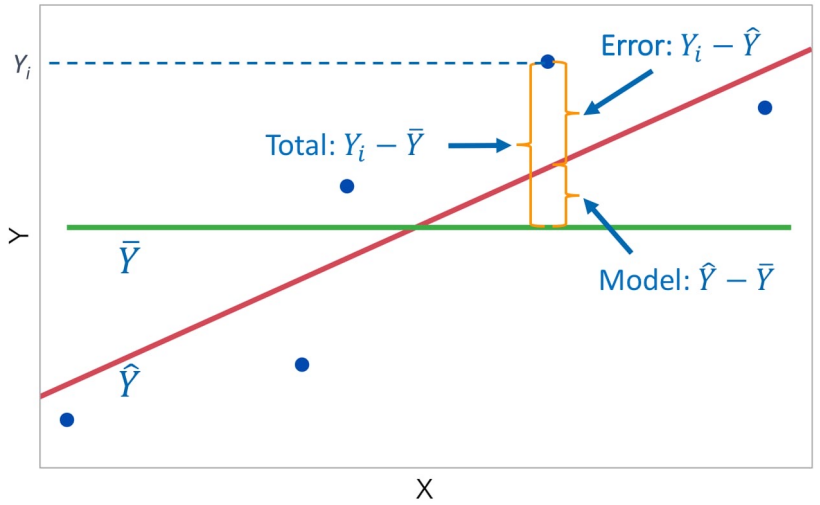
Sum of Squares Decomposition

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- TSS is the sum of squared errors we would get by predicting the mean Y value (green line) for all observations
- Our regression model (red line) reduces this to the residual sum of squares (RSS)
- The difference between these is how much variability our model explained (MSS)

$$\text{TSS} = \text{RSS} + \text{MSS}$$



https://www.jmp.com/en_gb/statistics-knowledge-portal/what-is-regression/interpreting-regression-results.html

ANOVA Table and F Test

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Model Utility Test- Tests whether there are any useful linear relationships between predictors and predictands and the response.
  - $H_0: \beta_1 = \cdots = \beta_k = 0$
  - $H_A: \beta_i \neq 0, i \in \{1, \cdots, k\}$

ANOVA Table and F Test

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- ANOVA table and F test

Source	Sum of Squares	d.f.	Mean Square
Model (MSS)	$\sum_{i=1}^{n}(\hat{y}_i - \overline{y_i})^2$	k	$\frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y_i})^2}{k}$
Residual (RSS)	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-k-1$	$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$
Total (TSS)	$\sum_{i=1}^{n}(y_i - \overline{y})^2$	$n-1$	Sample Variance

Source	F	Pr(F)
Model Utility	$\frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{k}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}$	$F_{(k,n-k-1)}$

F test

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- F-statistic with k and $n-k-1$ degrees of freedom is the ratio of how much variability our model explained to how much remains

$$F = \frac{MSS/k}{RSS/n-k-1} = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{k}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}$$

- The larger the F-statistic, the more useful the model.

ANOVA Table and F Test

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

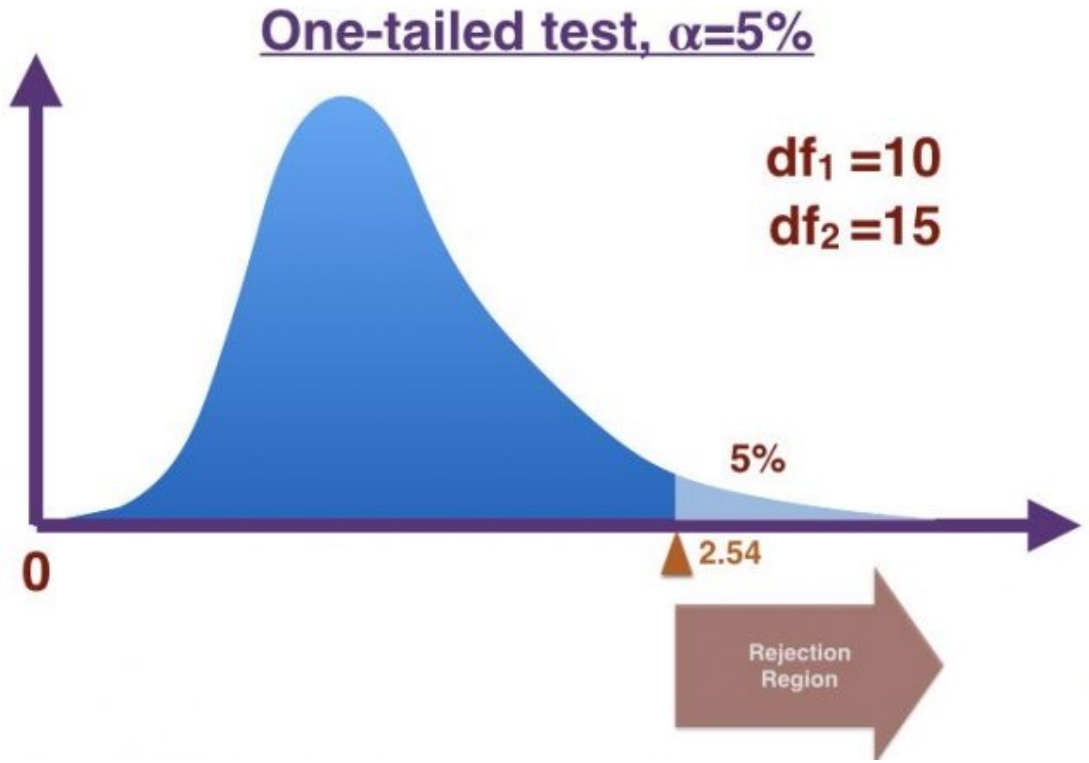
- Model Utility Test
  - $H_0: \beta_1 = \cdots = \beta_k = 0$
  - $H_A: \beta_i \neq 0, i \in \{1, \cdots, k\}$
- Under our null hypothesis, the model does not explain much, as all the coefficients are 0.
- If MSS>RSS, F is large and we reject that null hypothesis in favor of the alternative that the model has some predictive power.

Example F Test

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- We reject H0 if our F statistic is so large the probability of seeing it if H0 is true (p-value) is < α, usually 0.05. In this example, that is if F>2.54. R will report the exact p-value of the observed F statistic.



t-tests

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Hypotheses:
  - $H_0: \beta_i = 0$
  - $H_A: \beta_i \neq 0$
- T-statistic with n - k - 1 d.f.:

$$t = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

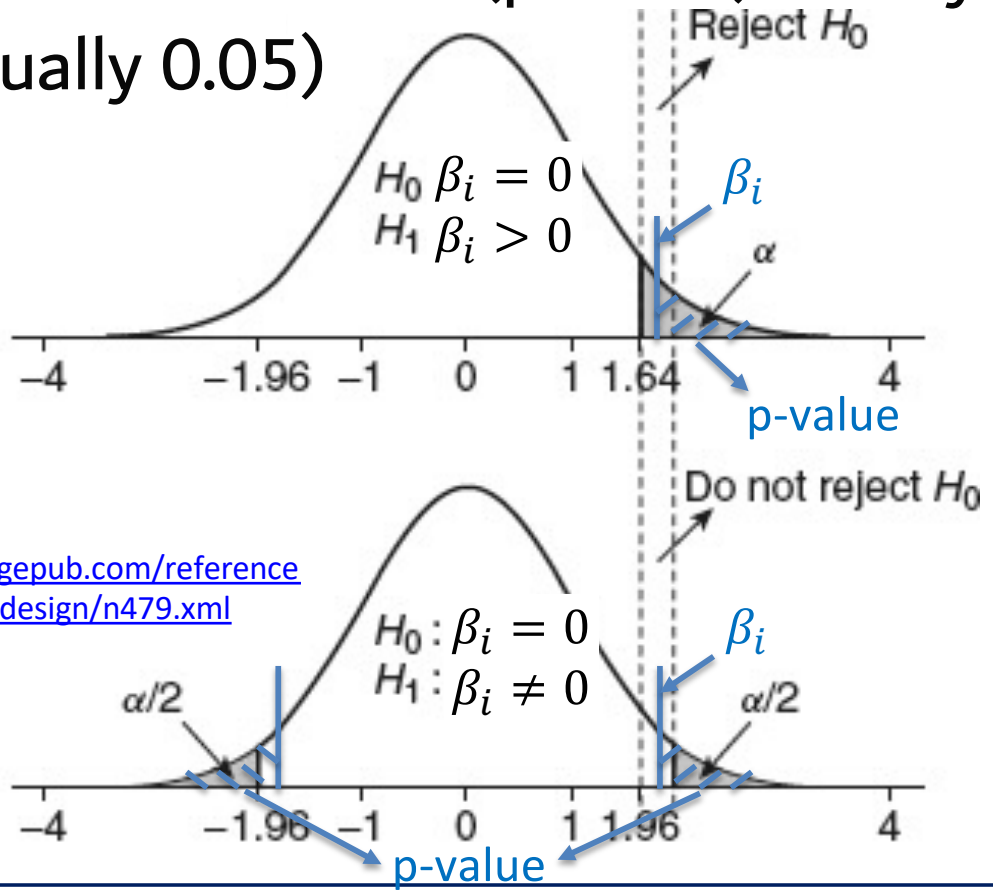
- Check whether the particular X is useful given the presence of other variables.

t-tests

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- We reject H0 that $\beta_i = 0$ if we observe a high or low value whose probability of being observed if H0 is true (p-value) is very low (<α, usually 0.05)



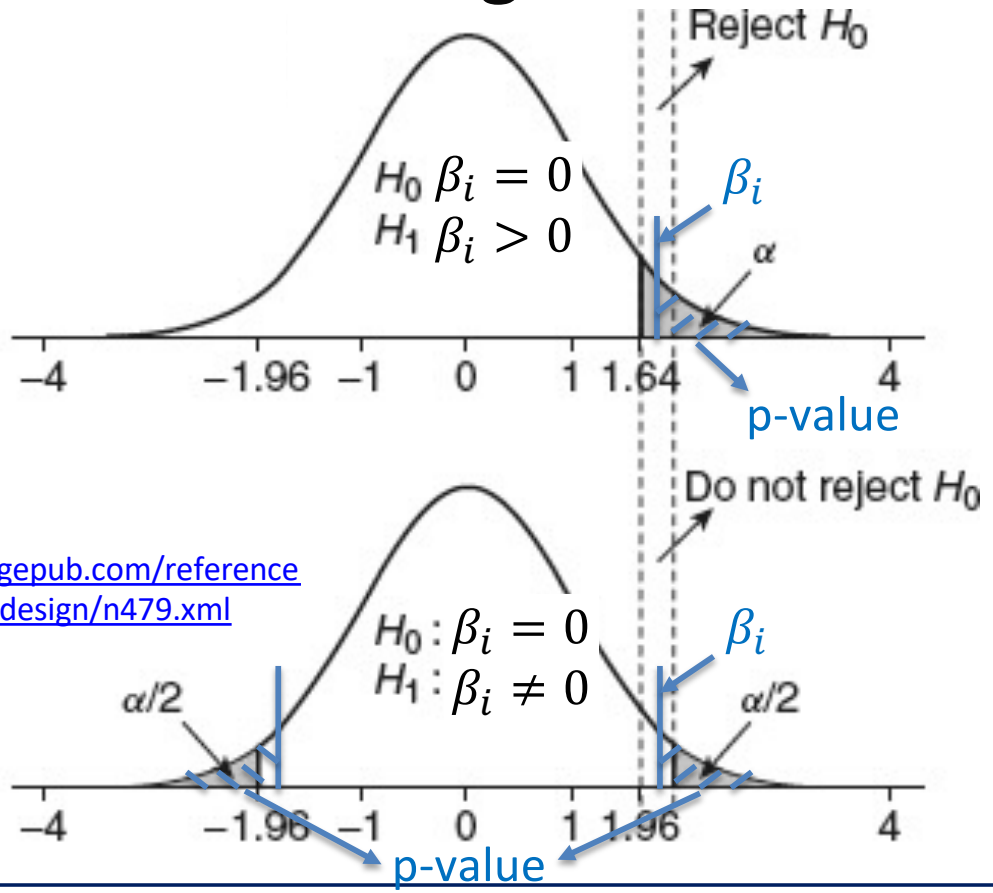
https://methods.sagepub.com/reference/encyc-of-research-design/n479.xml

t-tests

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- It is easier to reject for a 1-sided test (Ha: β_i>0 or β_i<0), as seen below
- R assumes we are using a 2-sided test



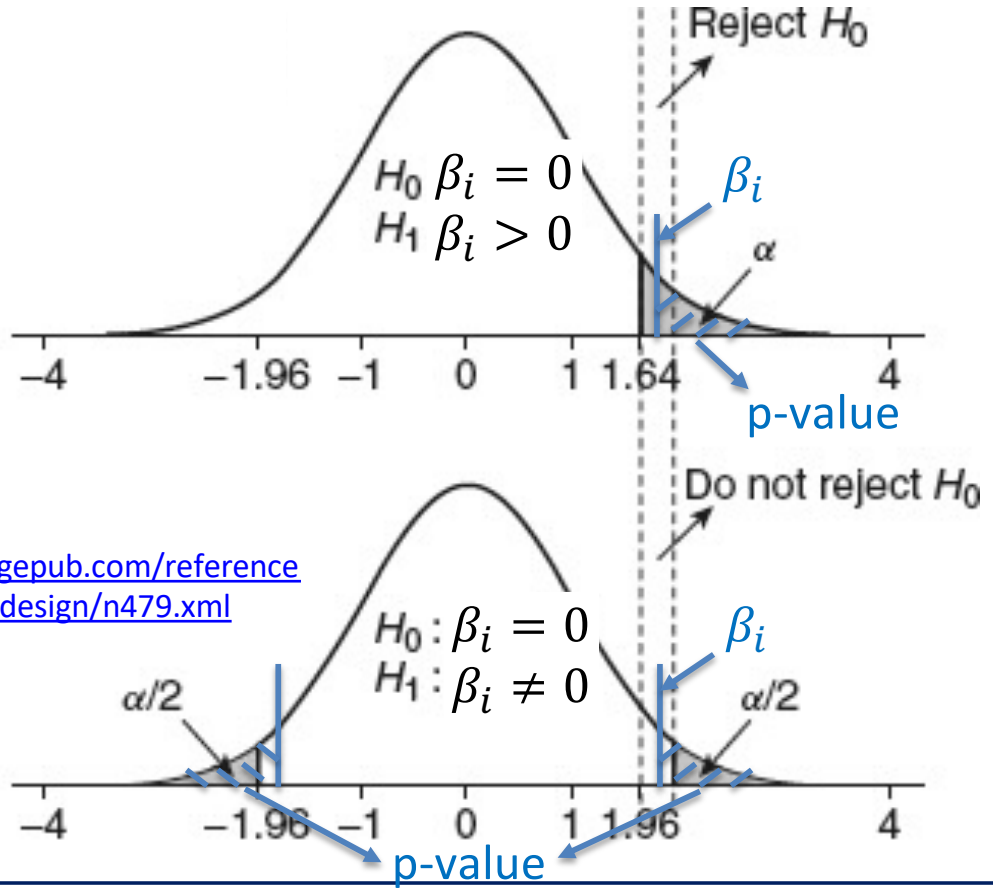
https://methods.sagepub.com/reference/encyc-of-research-design/n479.xml

t-tests

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- If you have reason to believe β_i>0 or β_i<0, you can do a 1-sided test and divide the p-value reported by R by 2



https://methods.sagepub.com/reference/encyc-of-research-design/n479.xml

Example in R

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

```
#Linear regression model with 3 predictors
xdmgnd.lm3<-lm(ACCDMG~TEMP+TRNSPD+CARS,data=xdmgnd)
summary(xdmgnd.lm3)
```

```
> summary(xdmgnd.lm3)

Call:
lm(formula = ACCDMG ~ TEMP + TRNSPD + CARS, data = xdmgnd)

Residuals:
     Min       1Q   Median       3Q      Max
-2069963  -368104  -203101    45608 31180112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 353863.79   40504.14   8.736  < 2e-16 ***
TEMP            68.27     584.60   0.117  0.90704
TRNSPD       16299.41     777.63  20.960  < 2e-16 ***
CARS          2774.64    1051.98   2.638  0.00837 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1184000 on 7250 degrees of freedom
Multiple R-squared:  0.05872,   Adjusted R-squared:  0.05833
F-statistic: 150.7 on 3 and 7250 DF,  p-value: < 2.2e-16
```

Additional Resources and References

Agenda

- Review
- Regression Overview
- Regression Assumptions
- Parameter Estimation
- R Example

- Chapter 3- Tibshirani et. Al. “An Introduction to Statistical Learning”, 2021

Multiple Linear Regression Metrics and Variable Selection

Laura E. Barnes
&
Julianne Quinn

Agenda

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Review of Multiple Linear Regression
- Measurement of Model Performance
- Variable Selection for Multiple Linear Regression
- Model Comparison

Review of Regression

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Multiple Regression: A method for measuring and modeling the relationship between sets of variables.
- Linear regression uses stochastic models with two components:
  - Functional: $y = f(x)$
  - Stochastic: ϵ
  - Matrix notation: $y = f(X) + \epsilon = X\beta + \epsilon$
- Regression is the solution to an optimization problem. We find a linear fit to the data that minimizes the sum of squared errors.

Assumptions for Multiple Linear Regression

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Assumptions for Optimization
  - The data for the input variables or predictors, $x_1, \cdots, x_p$ are known.
  - The predictors are linearly independent.
  - The response variable, y, is quantitative.
- Assumptions for Inference
  - For a sample size of n, the distribution of the $\epsilon_i, i = 1, \cdots, n$ are independent, identical and Gaussian with
  - $E(\epsilon_i) = 0$, and
  - $Var(\epsilon_i) = \sigma^2$

Least Squares Estimates

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- We find optimal estimates for the coefficients β where the criterion is least squares. The optimization problem is:

$$\text{minimize}(X\beta - y)^T(X\beta - y)$$

- The least squares estimate:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- The estimate for σ^2 where k is the number of predictors or input variables:

$$\hat{\sigma}^2 = \frac{(X\hat{\beta} - y)^T(X\hat{\beta} - y)}{n - k - 1} = \frac{y^T(I - H)y}{n - k - 1}$$

- $H = X(X^T X)^{-1} X^T$ is the hat matrix: $\hat{y} = X\hat{\beta} = Hy$

Sum of Squares Decomposition, F-Test, t-tests

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- The sum of squares has a convenient decomposition:
  - Total S.S. = Residual S.S + Model S.S
- Model Utility Test:
  - $H_0: \beta_1 = \cdots = \beta_p = 0$
  - $H_A: \beta_i \neq 0, i \in \{1, \cdots, p\}$
- T-tests:
  - $H_0: \beta_i = 0$
  - $H_A: \beta_i \neq 0$

Interpret Coefficients

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

```
#Linear regression model with 3 predictors
xdmgnd.lm3<-lm(ACCDMG~TEMP+TRNSPD+CARS,data=xdmgnd)
summary(xdmgnd.lm3)
```

```
> summary(xdmgnd.lm3)

Call:
lm(formula = ACCDMG ~ TEMP + TRNSPD + CARS, data = xdmgnd)

Residuals:
     Min       1Q   Median       3Q      Max
-2069963  -368104  -203101    45608 31180112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 353863.79   40504.14   8.736  < 2e-16 ***
TEMP            68.27     584.60   0.117  0.90704
TRNSPD       16299.41     777.63  20.960  < 2e-16 ***
CARS          2774.64    1051.98   2.638  0.00837 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1184000 on 7250 degrees of freedom
Multiple R-squared:  0.05872,   Adjusted R-squared:  0.05833
F-statistic: 150.7 on 3 and 7250 DF,  p-value: < 2.2e-16
```

Interpret Coefficients

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- The t-test tells us if the predictor is significant, but what about the coefficient itself, β_i?
- β_i tells us the change in Y for a unit change in X_i

```
> summary(xdmgnd.lm3)

Call:
lm(formula = ACCDMG ~ TEMP + TRNSPD + CARS, data = xdmgnd)

Residuals:
     Min       1Q   Median       3Q      Max
-2069963  -368104  -203101    45608 31180112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 353863.79   40504.14   8.736  < 2e-16 ***
TEMP            68.27     584.60   0.117  0.90704
TRNSPD       16299.41     777.63  20.960  < 2e-16 ***
CARS          2774.64    1051.98   2.638  0.00837 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1184000 on 7250 degrees of freedom
Multiple R-squared:  0.05872,   Adjusted R-squared:  0.05833
F-statistic: 150.7 on 3 and 7250 DF,  p-value: < 2.2e-16
```

Steps to Build Multiple Linear Regression Models

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Suppose we have an interesting problem like reducing severity of rail accidents. Also, we have data related to this problem.
- What will you do to solve this problem?

Evidence-Informed Problem Solving

1. Background, goals, sources of evidence
2. Hypothesis(es)
3. Visualization and Graphical Analysis
4. Build multiple regression models:
   - How well do the models perform?
   - What explanatory variables should we use?
   - …

Multiple Coefficient of Determination: R^2

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Sum of squares decomposition:

  Total S.S. = Residual S.S. + Model S.S.

- Coefficient of determination:

$$R^2 = 1 - \frac{\text{Residual S.S}}{\text{Total S.S}} = \frac{\text{Model S.S}}{\text{Total S.S}}$$

- What's the range of R^2 ?

  $[0,1]$

- What does $R^2 = 1$ mean?

  A perfect fit to data, but not necessarily a perfect model.

- Will R^2 decrease if we add variables to a regression model?

  No

Multiple Coefficient of Determination: R^2

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- The range of R^2 is from 0 to 1. A value near 0 indicates little linear association between the independent variables and the dependent variable.
- A value near 1 means a strong association.
- R^2 cannot go down when another predictor is added to the model.
- R^2 can almost always be made close to 1 by using a model with k predictors where k is very close to n.

Criterion Based Assessments

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- What criteria can we use to compare performance? Our goal is to predict accurately on new data.
- Why don't we use R^2?
- Most criteria reward good fits and penalize size (complexity).

Common Criterion-Based Methods

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

$$R^2_{adj} = 1 - \frac{Residual\ M.S.}{Total\ M.S.}$$

$$= 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

- Adjusted R^2 penalizes the addition of extraneous predictors
- Adjusted R^2 is smaller than R^2.

Common Criterion-Based Methods

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Akaike Information Criterion (AIC):

$$\mathrm{AIC} = \mathrm{nlog}(\mathrm{R.S.S}/\mathrm{n}) + 2\mathrm{p}$$

- Bayesian Information Criterion (BIC):

$$\mathrm{BIC} = \mathrm{nlog}(\mathrm{R.S.S}/\mathrm{n}) + \mathrm{plog}(\mathrm{n})$$

Common Criterion-Based Methods

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- The preferred model gives the smaller value of AIC and BIC.
- In most cases, BIC penalizes the complexity more strongly than AIC does.
- You may see different values for AIC or BIC in different functions. This results from the use of different definitions of the terms and different logs. All that matters is order and consistency.

Example: Rail Accidents

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

```
#Build linear regression models in R: lm
xdmgnd.lm1<-lm(ACCDMG ~ TEMP + TRNSPD + CARS + HEADEND1,data=xdmgnd)

xdmgnd.lm2<-lm(ACCDMG ~ TEMP + TRNSPD + CARS, data=xdmgnd)
```

```
> summary(xdmgnd.lm1)$adj.r.squared
[1] 0.06710172
> AIC(xdmgnd.lm1)
[1] 223200.5
> BIC(xdmgnd.lm1)
[1] 223241.8
>
>
> summary(xdmgnd.lm2)$adj.r.squared
[1] 0.06086843
> AIC(xdmgnd.lm2)
[1] 223247.8
> BIC(xdmgnd.lm2)
[1] 223282.2
```

Importance of Variable Selection

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Occam's Razor
- Extra terms can add noise to the predictions. More explanatory variables is not necessarily better.
- Correlation among the variables creates instabilities in the model. Small changes have big effects that are not wanted. This is called multicollinearity.
- Leaving out variables can also cause inaccurate understanding and predictions.
- Getting the data to add predictors can be costly.

Complexity of Variable Selection

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- If the number of available explanatory variables is very small, such as 2-3, we might build models with all combinations of explanatory variables.
- What if we have k explanatory variables? We will have $2^k - 1$ main effects models.
- We should consider more than the main effects models as we will see later. In addition, we will add qualitative variables and consider transformations of some or all of the variables.
- What is the next step?

Automated Selection

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Automated procedures have been developed to choose variables.
- Since the number of models to consider is so large, none of the procedures are optimal on any particular criteria. However, they can provide a convenient approach to narrowing your search and providing some insight.
- Three simple automated selection techniques:
  - Forward selection
  - Backward selection
  - Stepwise regression
- Regression models with automated variable selection is a hot topic of recent research.

Automated Selection Techniques

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- All the automated techniques require some criterion:
  - We used to use F tests and the significance level on F tests.
  - Most modern approaches use AIC or BIC or some other version of them.
- Forward selection: starts with the mean (intercept) model and adds terms one at a time. It picks the term that does the best on the criterion. It stops when the threshold is reached.
- Backward selection: starts with the complete (full main effects) model. It then eliminates terms one at a time and stops when a threshold is met.

Stepwise Regression

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Stepwise regression:
  - A sequential process that adds or drops one new variable to the model at each step.
  - It always selects the variable that provides the greatest improvement to the selection criterion given the variables already in the model.
  - Stepwise stops when no variable can be added or dropped according to a threshold.
- Stepwise works best when started by backward selection (i.e., start with a large model).
- Stepwise is computationally expensive. For large variable sets expect a wait.
- In R stepwise defaults to using AIC.

Stepwise Regression Example

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

```
> xdmgnd.lm1.step<-step(xdmgnd.lm1, trace=T)
Start:  AIC=202615.3
ACCDMG ~ TEMP + TRNSPD + CARS + HEADEND1

            Df  Sum of Sq        RSS    AIC
- TEMP       1 9.0130e+09 9.8198e+15 202613
<none>                    9.8198e+15 202615
- CARS       1 1.8065e+13 9.8378e+15 202627
- HEADEND1   1 6.6976e+13 9.8868e+15 202663
- TRNSPD     1 6.8428e+14 1.0504e+16 203102

Step:  AIC=202613.3
ACCDMG ~ TRNSPD + CARS + HEADEND1

            Df  Sum of Sq        RSS    AIC
<none>                    9.8198e+15 202613
- CARS       1 1.8061e+13 9.8379e+15 202625
- HEADEND1   1 6.6967e+13 9.8868e+15 202661
- TRNSPD     1 6.8566e+14 1.0505e+16 203101
>
```

Stepwise Regression Example

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

```
> summary(xdmgnd.lm1.step)

Call:
lm(formula = ACCDMG ~ TRNSPD + CARS + HEADEND1, data = xdmgnd)

Residuals:
     Min       1Q   Median       3Q      Max
-2281311  -357072  -199270    63906 31049534

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 489220.0    28687.5  17.053  < 2e-16 ***
TRNSPD       17545.7      779.9  22.498  < 2e-16 ***
CARS          3810.5     1043.6   3.651 0.000263 ***
HEADEND1    -71997.8    10240.0  -7.031 2.24e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1164000 on 7249 degrees of freedom
Multiple R-squared:  0.06762,   Adjusted R-squared:  0.06723
F-statistic: 175.2 on 3 and 7249 DF,  p-value: < 2.2e-16
```

Partial F Tests

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- How can we tell whether the reduction in sum of squares by an inclusive model (one that includes all the variables of the smaller model and more) is statistically significant?
  - Partial F tests
- Suppose the smaller model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \epsilon$$

and the larger model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \beta_{i+1} x_{i+1} + \cdots + \beta_k x_k + \epsilon$$

- We use the F statistic to test:
  - $H_0: \beta_{i+1} = \cdots = \beta_k = 0$
  - $H_A: \beta_j \neq 0, j \in \{i+1, \cdots, k\}$

Partial F Tests

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- In a Partial F test, instead of our test statistic comparing MSS and RSS , we compare the RSS from the nested models.
- Let Model 1 be nested within Model 2, so Model 2 has all the same predictors as Model 1 and q more.
- To test if at least one of the additional predictors in Model 2 adds predictive value, we compare RSS_1 and RSS_2 where the smaller model has:

$$F = \frac{(RSS_1 - RSS_2)/q}{RSS_2/(n-k-1)}$$

- Once again, a large value of F indicates $RSS_2 << RSS_1$, so our larger model adds great predictive value and we reject H0 that none of the additional predictors have any significance

Partial F Tests

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Partial F test example: full model vs. stepwise model:

```
> anova(xdmgnd.lm1,xdmgnd.lm2)
Analysis of Variance Table

Model 1: ACCDMG ~ TEMP + TRNSPD + CARS + HEADEND1
Model 2: ACCDMG ~ TEMP + TRNSPD + CARS
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1   7248 9.8198e+15
2   7249 9.8868e+15 -1 -6.6976e+13 49.435 2.237e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(xdmgnd.lm1,xdmgnd.lm1.step)
Analysis of Variance Table

Model 1: ACCDMG ~ TEMP + TRNSPD + CARS + HEADEND1
Model 2: ACCDMG ~ TRNSPD + CARS + HEADEND1
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1   7248 9.8198e+15
2   7249 9.8198e+15 -1 -9.013e+09 0.0067  0.935
```

- Do we reject the null hypothesis here?

  Yes to first test, No to second test

- So which model should we pick?

  Keep Headend1, Drop Temp

Partial F Tests

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Partial F test example: full model vs. stepwise model:

```
> anova(xdmgnd.lm1,xdmgnd.lm2)
Analysis of Variance Table

Model 1: ACCDMG ~ TEMP + TRNSPD + CARS + HEADEND1
Model 2: ACCDMG ~ TEMP + TRNSPD + CARS
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1   7248 9.8198e+15
2   7249 9.8868e+15 -1 -6.6976e+13 49.435 2.237e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(xdmgnd.lm1,xdmgnd.lm1.step)
Analysis of Variance Table

Model 1: ACCDMG ~ TEMP + TRNSPD + CARS + HEADEND1
Model 2: ACCDMG ~ TRNSPD + CARS + HEADEND1
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1   7248 9.8198e+15
2   7249 9.8198e+15 -1 -9.013e+09 0.0067  0.935
```

- Partial F tests provide a convenient, parametric method to judge between models.
- Also allows us to assess all of the variables.

Test Sets

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Partial F test is a good approach, but what if we want to compare non-nested models?
- Recall our goal for model selection. In using test sets we sample a subset (without replacement) of the data and do not use it to build the model. We use this subset for testing.
- A common choice is to sample 1/3 of the data for testing (called the test set or out-of-sample set) and 2/3 for model building (called the training set or the sample).
- Check that your test set is representative.
- What criterion do we use to choose among the models? Predicted Mean Squared Error (PMSE)

Example: Rail Accidents Test Sets Results

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

```
source("TestSet.R")

##set test sets size:
test.size<-1/3

##generate training sets and test sets from original data:
xdmgnd.data<-test.set(xdmgnd,test.size)


##Build models with training set:
xdmgnd.lm1.train<-lm(ACCDMG~TEMP+TRNSPD+CARS+HEADEND1,data=xdmgnd.data$train)

xdmgnd.lm2.train<-lm(ACCDMG~TEMP+TRNSPD+CARS,data=xdmgnd.data$train)


##Recall that we need to measure predicted MSE.
##First, how to predict with lm models:
xdmgnd.lm1.pred<-predict(xdmgnd.lm1.train,newdata=xdmgnd.data$test)

xdmgnd.lm2.pred<-predict(xdmgnd.lm2.train,newdata=xdmgnd.data$test)


##Next, compute PMSE:
pmse.xdmgnd.lm1<-mse(xdmgnd.lm1.pred,xdmgnd.data$test$ACCDMG)

pmse.xdmgnd.lm2<-mse(xdmgnd.lm2.pred,xdmgnd.data$test$ACCDMG)
```

```
> pmse.xdmgnd.lm1
                [,1]
[1,] 1.344837e+12
> pmse.xdmgnd.lm2
                [,1]
[1,] 1.355913e+12
>
```

Cross-Validation

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Test sets method requires enough data.
- If we don't have enough data, we can use cross validation.
- In cross validation we start by sampling without replacement to get k parts, called folds.
- We build a model on k - 1 of the folds and test on the remaining fold.
- We repeat this using each fold as a test set for the other k - 1 folds.

https://www.wikiwand.com/en/Cross-validation_(statistics)

Cross-Validation

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- If we have n data points how much of it do we use for training?

  n

- How much do we use for testing?

  n

- Models are compared based on the averages of PMSE.

https://www.wikiwand.com/en/Cross-validation_(statistics)

Example: Rail Accidents CV Results

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

```
# Cross-Validation

##Need the boot library
library(boot)

xdmgnd.lm1.cv<-glm(ACCDMG~TEMP+TRNSPD+CARS+HEADEND1,data=xdmgnd)

xdmgnd.lm2.cv<-glm(ACCDMG~TEMP+TRNSPD+CARS,data=xdmgnd)

##Cross-validation error

xdmgnd.lm1.err<-cv.glm(xdmgnd,xdmgnd.lm1.cv,K=10)

xdmgnd.lm2.err<-cv.glm(xdmgnd,xdmgnd.lm2.cv,K=10)


xdmgnd.lm1.err$delta

xdmgnd.lm2.err$delta
```

```
> xdmgnd.lm1.err$delta
[1] 1.357638e+12 1.357442e+12
>
> xdmgnd.lm2.err$delta
[1] 1.366671e+12 1.366485e+12
>
```

Additional Resources and References

Agenda

- Review
- Metrics
- Variable Selection
- Model Comparison

- Chapters 2, 3 & 5- Tibshirani et. Al. “An Introduction to Statistical Learning”, 2021

Multiple Linear Regression Diagnostics and Transformations

Laura E. Barnes
&
Julianne Quinn

Agenda

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Review of Multiple Regression
- Model Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Transformation of Response Variables and Predictors

Review of Regression

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Multiple Regression: A method for measuring and modeling the relationship between sets of variables.
- Multiple Linear Regression:

$$y = f(X) + \epsilon = X\beta + \epsilon$$

- Regression is the solution to an optimization problem. We find a linear fit to the data that minimizes the sum of squared errors.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{\sigma}^2 = \frac{(X\hat{\beta} - y)^T (X\hat{\beta} - y)}{n - k - 1} = \frac{y^T (I - H) y}{n - k - 1}$$

Sum of Squares Decomposition, F-Test, t-tests

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- The sum of squares has a convenient decomposition:
  - Total S.S. = Residual S.S + Model S.S
- Model Utility Test:
  - $H_0: \beta_1 = \cdots = \beta_p = 0$
  - $H_A: \beta_i \neq 0, i \in \{1, \cdots, p\}$
- T-tests:
  - $H_0: \beta_i = 0$
  - $H_A: \beta_i \neq 0$

Metrics to Evaluate Models

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Multiple Coefficient of Determination: R^2
- Adjusted-R^2
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

Variable Selection

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Importance of variable selection
- Automated Selection:
  - Forward selection
  - Backward selection
  - Stepwise regression
- Partial F Test
  - $H_0: \beta_{i+1} = \cdots = \beta_k = 0$
  - $H_A: \beta_j \neq 0, j \in \{i+1, \cdots, k\}$

Test Sets

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- In using test sets we sample a subset (without replacement) of the data and do not use it to build the model. We use this subset for testing.
- What criterion do we use to choose among the models? Predicted Mean Squared Error (PMSE)
- Because test sets are generated by sampling, you will see different PMSE on different runs. Therefore, it is better to generate paired groups of PMSE and then use statistical tests to see whether the difference between models is significant.

Cross Validation

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- In cross validation we start by sampling without replacement to get k parts, called folds.
- We build a model on k - 1 of the folds and test on the remaining fold. We repeat this using each fold as a test set for the other k - 1 folds.
- We use all the data points for training and testing.

Regression Assumptions

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Recall our goal: determine if there is a relationship between the variables.
- Regression allows us to test hypotheses about relationships between the predictor variables and the response variable. What tests do we use?
- Regression allows for association tests while controlling for the values of other variables in the equation (e.g., ACCDMG vs. TONS and TRNSPD)
- To gain these powerful results regression requires certain assumptions:
  - Independent, identically distributed Gaussian errors with zero mean and constant variance;
  - Linearly independent predictor variables; and
  - Correct linear model for the response.

Diagnostic Plots

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- We can use diagnostic plots to exam how well those assumptions are satisfied.
- Four common diagnostic plots:
  - Residuals vs. fitted
  - Scale-Location plot of square root of absolute standardized residuals vs. fitted
  - QQ plot of standardized residuals
  - Residual-Leverage plot

Residual vs. Fitted

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

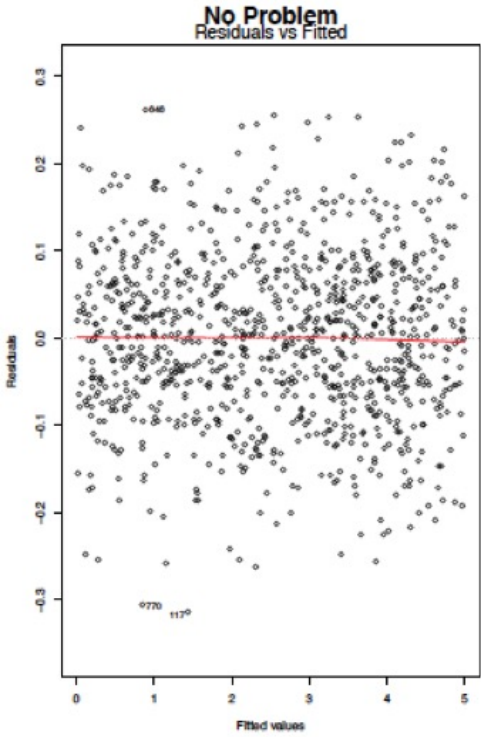
- The most important and useful plot.
- You look for
  - Absence of patterns in a symmetric display around zero.
  - Constant variance.
- Problem if you see:
  - A relationship between the residuals and the fitted values. This indicates a lack of fit for the model.
  - Changing variance which means you need to transform your response or explicitly account for the variance in the model (e.g. with a GLM, next unit).
  - Extreme, influential, or outlying points. Investigate these and make a decision on their influence. Could be lack of fit, heteroscedasticity, both, or neither.

Examples of Residual vs. Fitted

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

No Problem



- Relationship is nonlinear, should try transforming a predictor
- Can also explore Residuals vs. Predictors plots to see if they can explain remaining patterns

Predictor Transformations

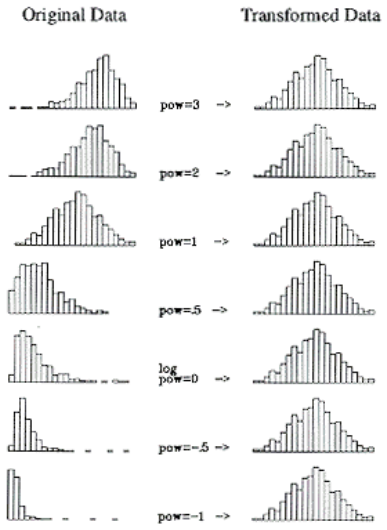
Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- When lack of fit is detected, consider transformations of the predictors.
- Predictors need to cover the range of desired values.
- Log and other transformations can help reduce the effects of skewness.
- Squaring a predictor can also capture nonlinear relationships.
- What kinds of transformations can we do on predictors? Any

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.



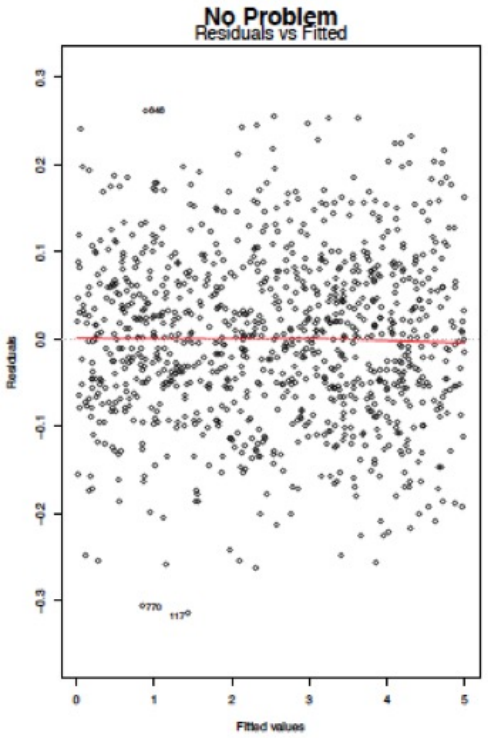
https://nielsen.sites.oasis.unc.edu/soci709/m3/m3.html

Examples of Residual vs. Fitted

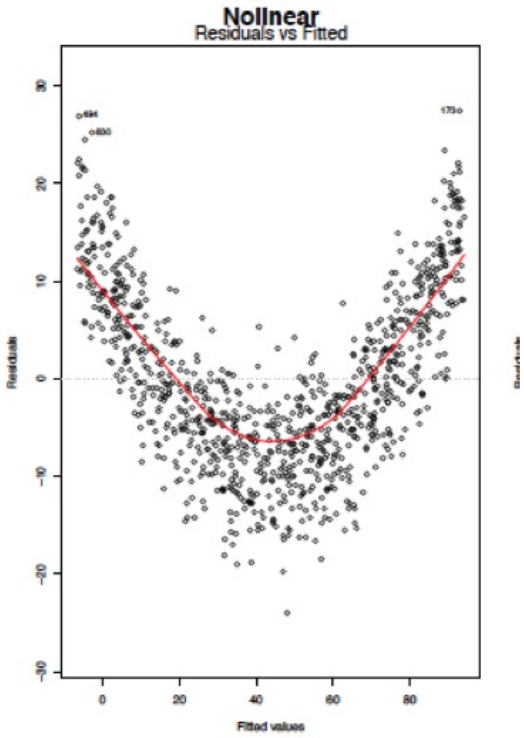
Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

No Problem



Lack of Fit,
Non-constant mean



- Relationship is nonlinear, should try transforming a predictor
- Can also explore Residuals vs. Predictors plots to see if they can explain remaining patterns

Testing for constant variance

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- We can test for constant variance using the Breusch-Pagan test

  H0: Constant variance (homoscedastic)

  Ha: Non-constant variance (heteroscedastic)

- If the p-value of the test <0.05, we reject H0, meaning the assumption of constant variance (homoscedasticity) is violated

Testing for constant variance

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

```
> xdmgnd.lm2<-lm(ACCDMG~TEMP+TRNSPD,data=xdmgnd)
> library(olsrr)
```

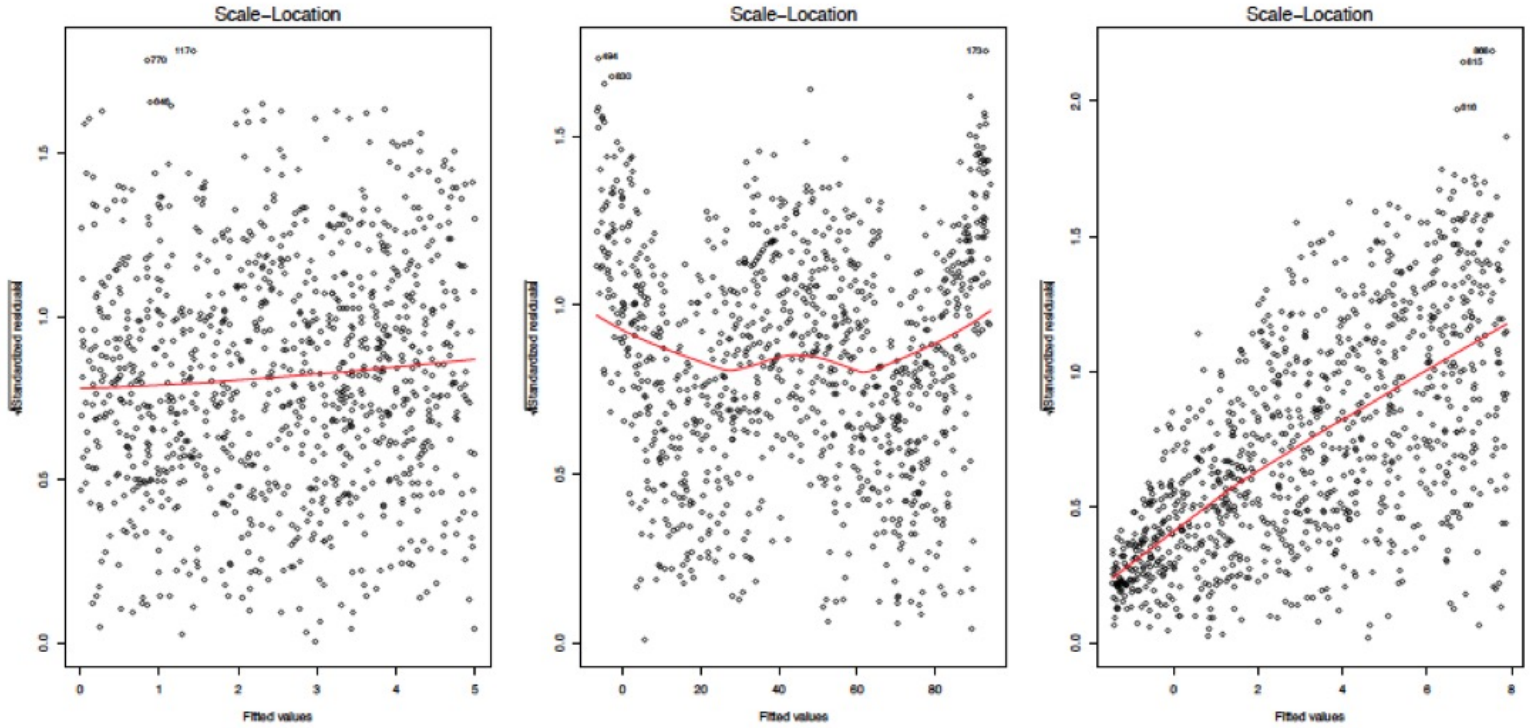
```
> ols_test_breusch_pagan(xdmgnd.lm2)

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant
 Ha: the variance is not constant

                 Data
 ---------------------------------
 Response : ACCDMG
 Variables: fitted values of ACCDMG

         Test Summary
 ----------------------------
 DF              =    1
 Chi2            =    2876.5594
 Prob > Chi2     =    0.0000
```

Scale-Location Plot

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Similar to residual vs. fitted, so look for the same things.
- S-L plots are less affected by skew.



QQ Plot of Standardized residuals

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

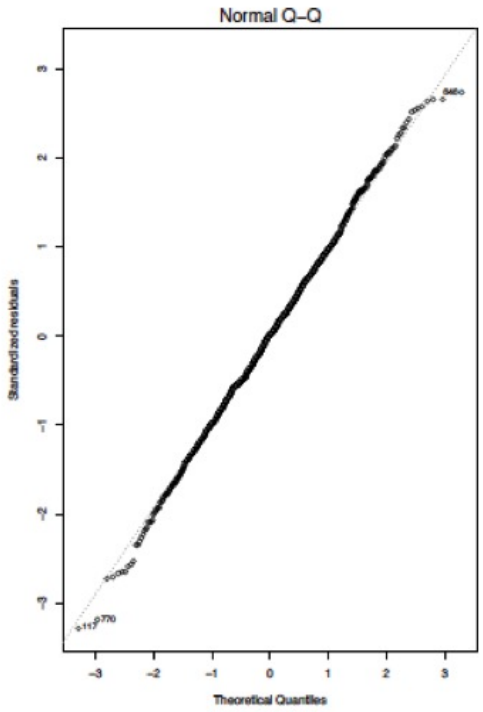
- What is QQ plot?
- A QQ plot can diagnose failure to meet the Gaussian assumption for the error term.
- Residuals are the estimated errors.
- Failure is indicated by points far from the 1:1 line.

QQ Plot Examples

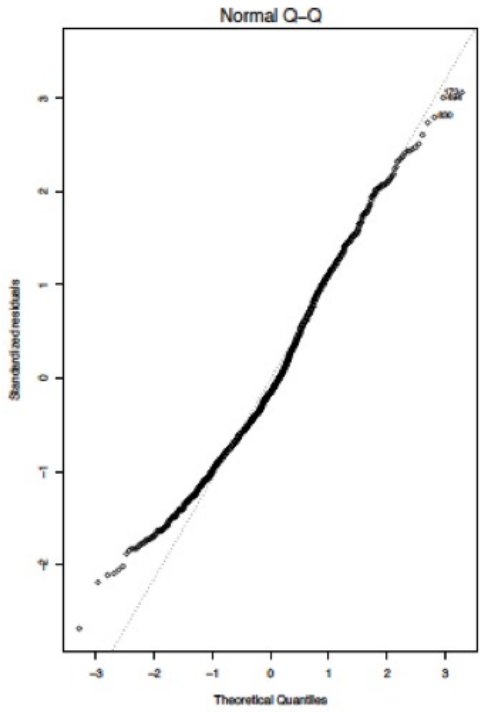
Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

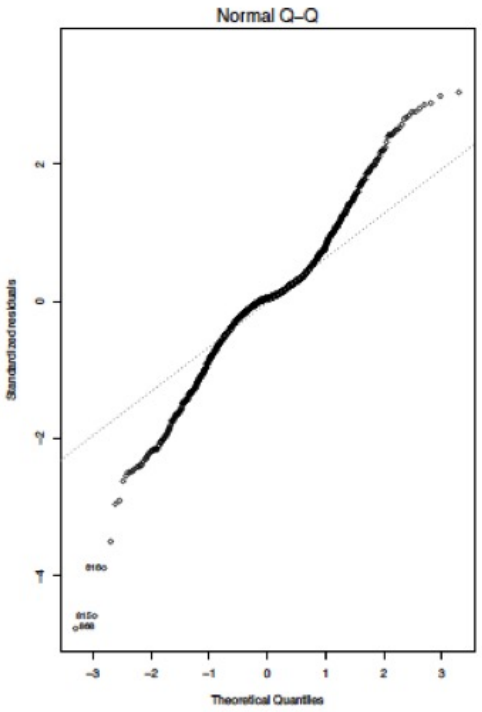
Good fit



Lower tail is non-Gaussian



Whole distribution is non-Gaussian



We can try transforming the response variable to see if these improve

Response Transformation: Box-Cox

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- When lack of Gaussian errors or heteroscedasticity are detected consider transformation of the response. This can be helped with Box-Cox plots.
- Box-Cox plots allow us to consider monotonic power transformations of the form:

$$t_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, \lambda \neq 0 \\ \log(y), \qquad \lambda = 0 \end{cases}$$

- The plot shows the 95% confidence interval for the λ that maximizes the log-likelihood that the errors are Gaussian.
- Look for the integers that this interval brackets. If it is around 1, what do we do?

  No transformation.
- If it is around 0, what do we do?

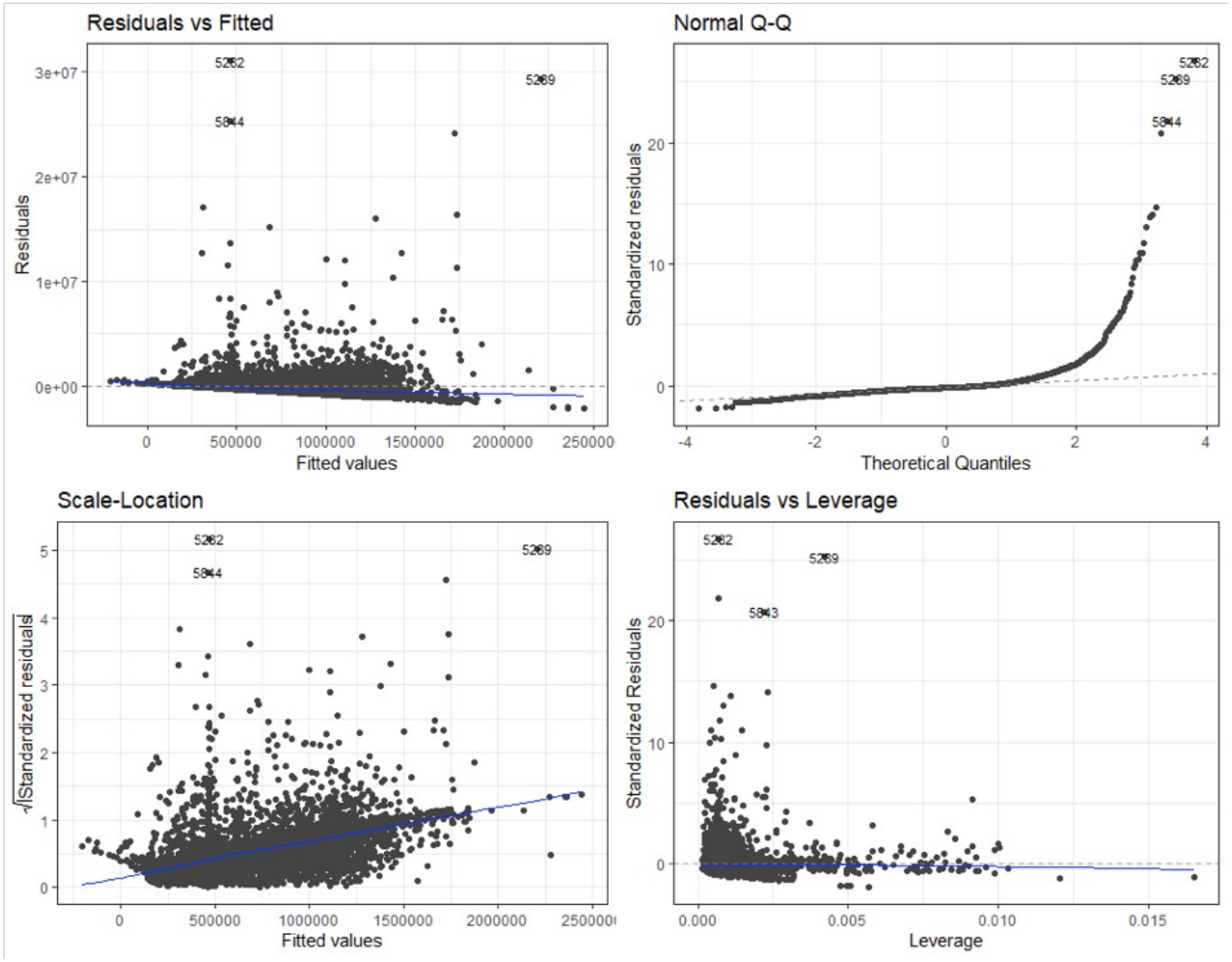
  Log transformation.

Diagnostic Plot Examples

xdmgnd.lm1<-lm(ACCDMG ~ TEMP + TRNSPD + TONS + CARS+ HEADEND1,data=xdmgnd)

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

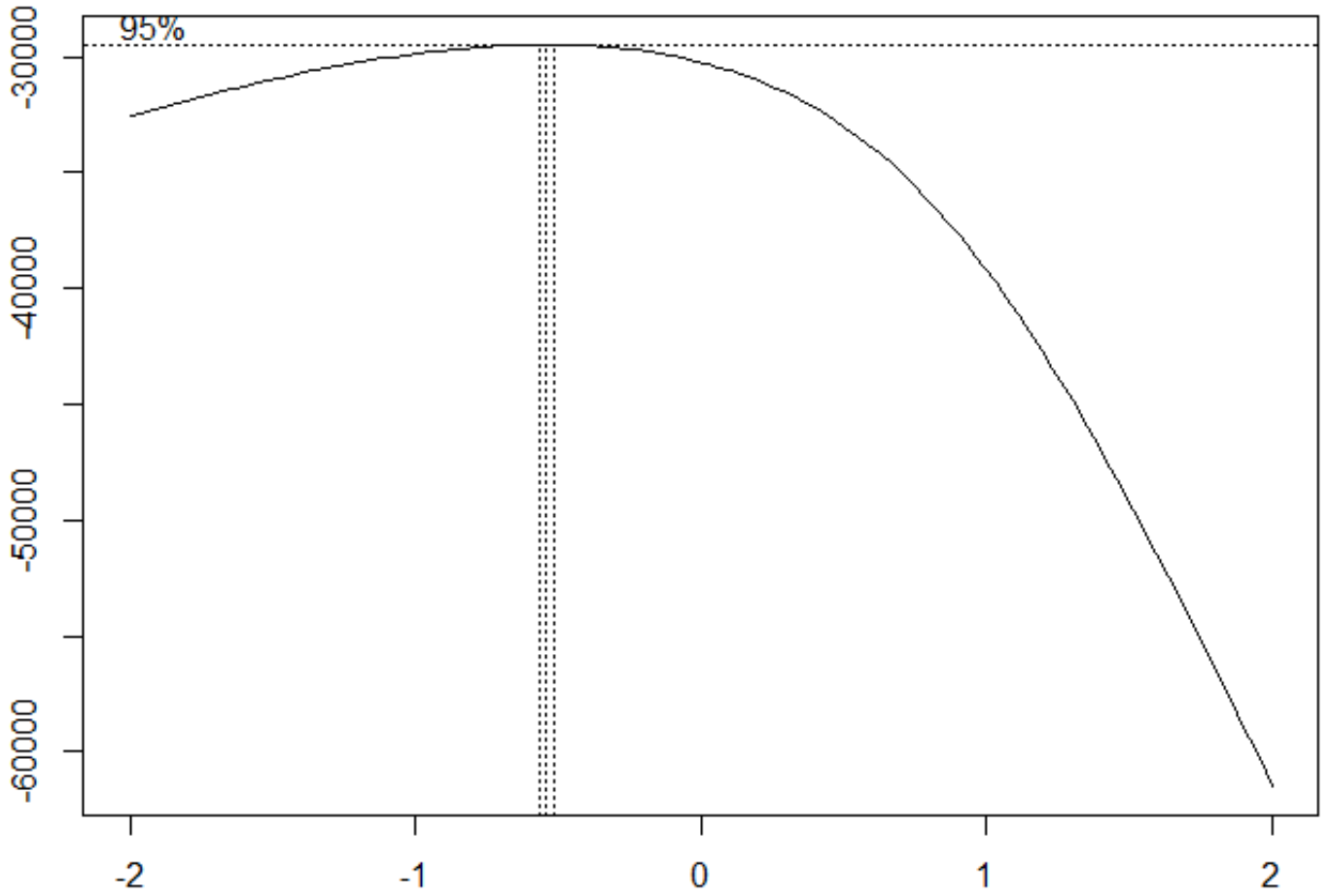


Example Box-Cox Plot

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

```
boxcox(xdmgnd.lm1, plotit=T, lambda=seq(-2,2,by=0.5))
```

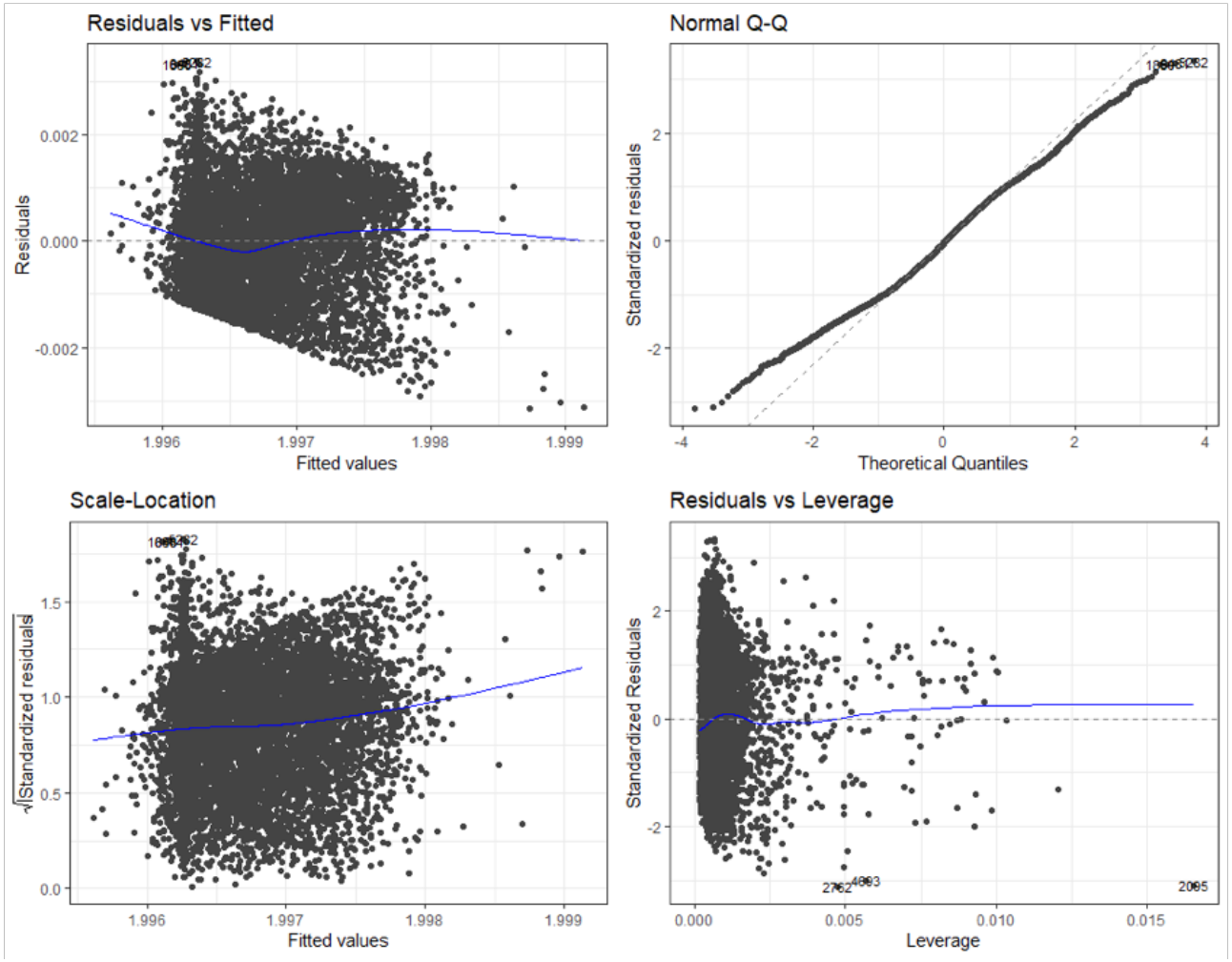


Example of Transformed Diagnostics

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

xdmgnd.lm2<-lm((ACCDMG^L-1)/L ~ TEMP + TRNSPD + TONS + CARS+ HEADEND1,data=xdmgnd)



Influential Points

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

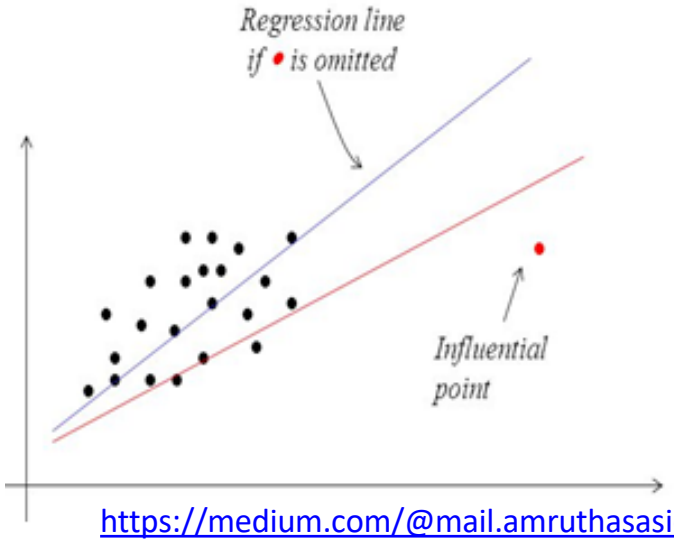
- Recall that $\hat{\beta} = (X^T X)^{-1} X^T y$
- Some data points might influence $\hat{\beta}$ more than the others.
- Influential points can change the values of parameter estimates and other results by their removal.
- Outliers do not fit the model well.
- Observations can be either, neither, or both.
- Influential observations can come from a lack of fit, heteroscedasticity, data entry errors, etc.
- To discover influential points and outliers we need a measures of influence: combine leverage and standardized residuals.

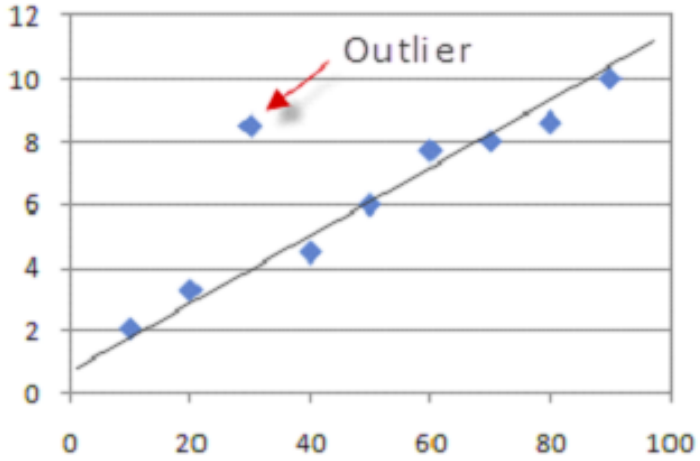
Influential point vs. outlier

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Left: influential point. Its removal drastically changes the regression line.
- Right: outlier. It does not fit the model well, but it doesn't influence it.





https://medium.com/@mail.amruthasasidharan/influential-points-vs-outliers-2c70c7b8d676

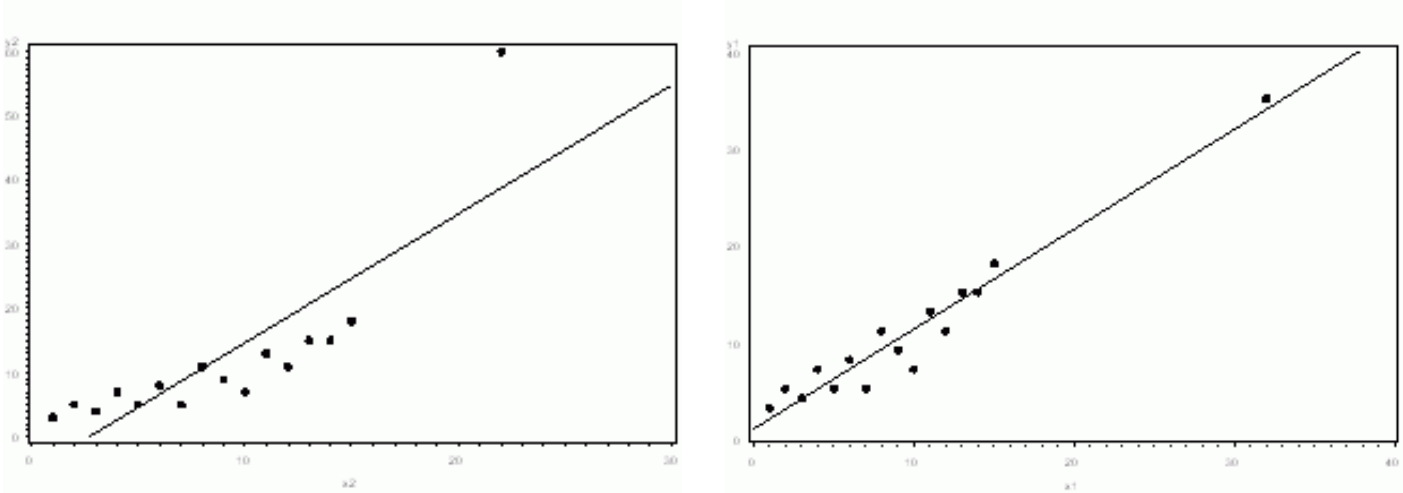
- The left point is influential while the right is not because it has more leverage being far from the mean of the data.

Leverage vs. influence

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Left: the extreme point has high leverage and it is an outlier that doesn't fit the model well (has a high residual).
- Right: the extreme point has high leverage, meaning it has high potential to influence the data. But it fits the model well, so it is not influential.



- The left point is influential, moving the regression line, while the right is not. This can be diagnosed because it has high leverage and a high residual.

https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

Cook's Distance

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Potentially influential points have high leverage. Observations with high leverage are far away from the average of predictor values.
- Leverage are the diagonal entries in H or $H_{ii} = h_i$
- Cook's distance attempts to identify influential observations using leverage.
- Cook's distance is calculated from the leverage and standardized residual of an observation:

$$C_i = r_i^2 \frac{h_i}{p(1-h_i)}$$

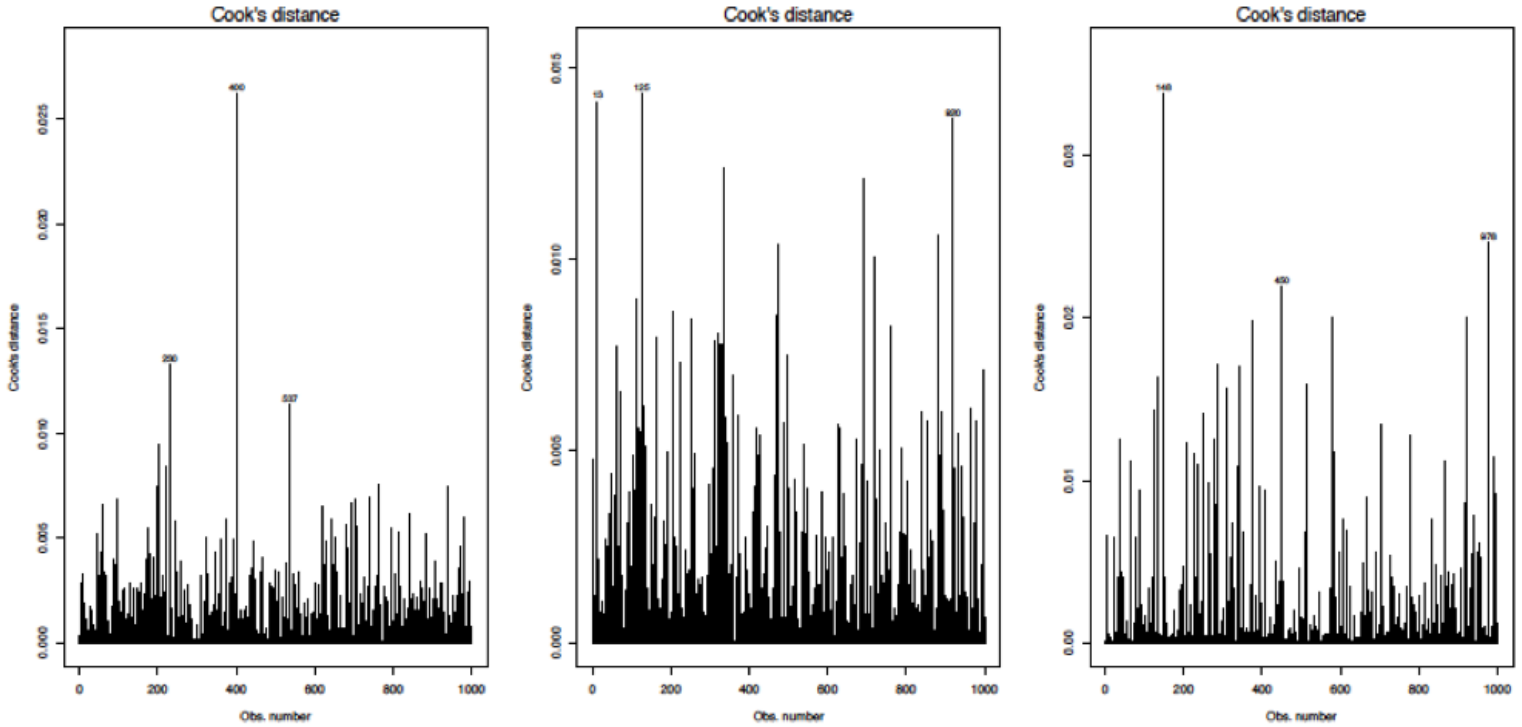
- where r_i is the standardized residual, and p is the number of parameters.
- What does a large Cook's distance mean?

Point has high influence and should likely be removed. Investigate if C_i>1 or even 0.5.

Cook's Distance Plots Examples

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

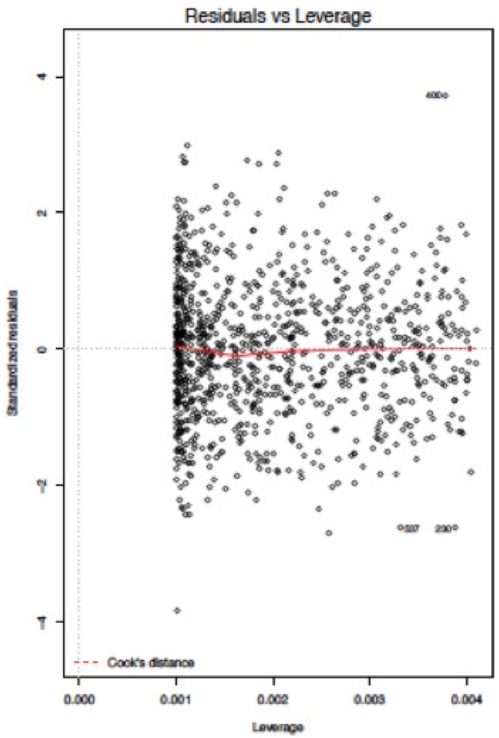


No points here have a Cook's distance > 0.05, so no points are influential and need to be removed.

Residual vs. Leverage Plot Examples

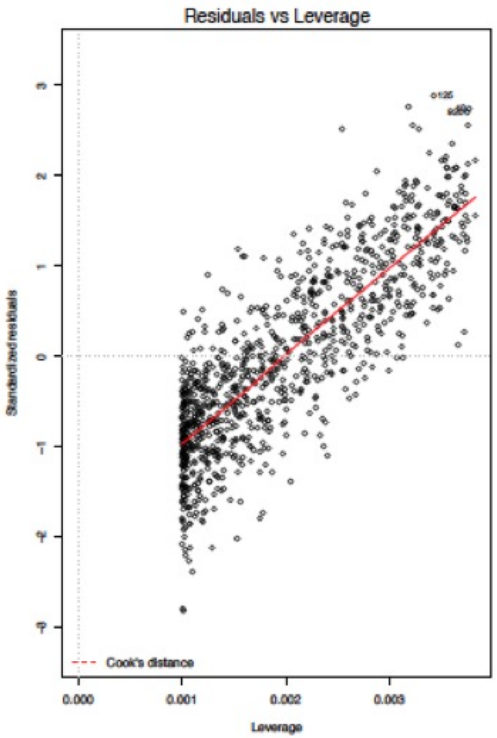
Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary



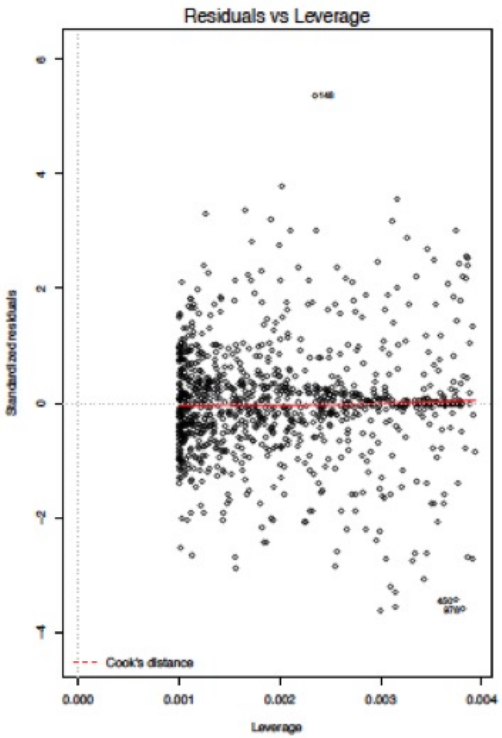
No Problem

Points with high leverage (far away from the mean) have no higher residuals than other points



Very Concerning

Points with high leverage (far away from the mean) are predicted very poorly and are likely influential



Somewhat Concerning

Points with high leverage (far away from the mean) are more likely to have low residuals and may be driving the model fit

Diagnosis from Regression Results

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- We can also diagnose problems in the regression by analyzing the results.
  - Coefficient values
  - Correlations
  - Performance tests

Diagnosis of Multicollinearity: Coefficients

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Multicollinearity means the assumption of linearly independent predictors is violated.
- If two or more predictors are highly correlated, it is hard to parse out whether the response is influenced by one or another moving with it.
- With correlated predictors in the model, the regression estimation will "split the difference" between them. This could result in very different coefficient estimates on these predictors depending on whether or not other correlated predictors are also in the model.

Diagnosis of Multicollinearity: VIF

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Multicollinearity is apparent in the wild swings in coefficients (e.g., flipped signs) based on other independent predictors in model.
- This means the variance of our coefficient estimates, $s(\beta)^2$, is high.
- We can use the Variance Inflation Factor (VIF) to see how much the variance of our coefficient estimate is inflated by other predictors in the model.
- The VIF of the i^{th} predictor coefficient is:

$$VIF_i = \frac{1}{1 - R_i^2}$$

  where R_i^2 is the coefficient of determination of a model predicting X_i as a function of all other predictors.
- If VIF>5, we should consider removing correlated predictors, and definitely if VIF>10.

Diagnosis of Multicollinearity: VIF

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Squaring predictors can capture nonlinear relationships, but increase collinearity

```
> library(car)
> vif(xdmgnd.lm2)
     TRNSPD I(TRNSPD^2)
   7.189156    7.189156
> vif(xdmgnd.lm3)
     TRNSPD I(TRNSPD^2)     HIGHSPD
   7.976421    7.737656    3.555189
> vif(xdmgnd.lm4)
      TRNSPD  I(TRNSPD^2)      HIGHSPD I(HIGHSPD^2)
    25.12493     27.81672     28.03560     30.83330
```

- "Centering" variables by subtracting their mean can sometimes overcome this:

```
> vif(xdmgnd.lm2a)
  I(TRNSPD - mean(TRNSPD)) I((TRNSPD - mean(TRNSPD))^2)
                  1.526767                    1.526767
```

Simpson's Paradox Example

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Simpson's Paradox is a phenomenon in which a relationship appears within several groups of data, but not when all the data is combined together.
- One of the best-known examples of Simpson's paradox: apparent gender bias among graduate school admissions to University of California, Berkeley in 1973

Gender	Applicants	Admitted
All	12,763	41%
Men	8442	44%
Women	4321	35%

Simpson's Paradox Example

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- A closer look . . .

Table 1: Data From Six Largest Departments of 1973 Berkeley Discrimination Case

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Source: Bickel, Hammel, and O'Connell (1975); table accessed via Wikipedia at https://en.wikipedia.org/wiki/Simpson%27s_paradox.

Simpson's Paradox Example

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Israeli data on vaccine effectiveness

Nearly 60% of hospitalized COVID-19 patients in Israel fully vaccinated, data shows

Erica Carbajal - Thursday, August 19th, 2021 Print I Email

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**

https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated

Simpson's Paradox Example

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Digging deeper …

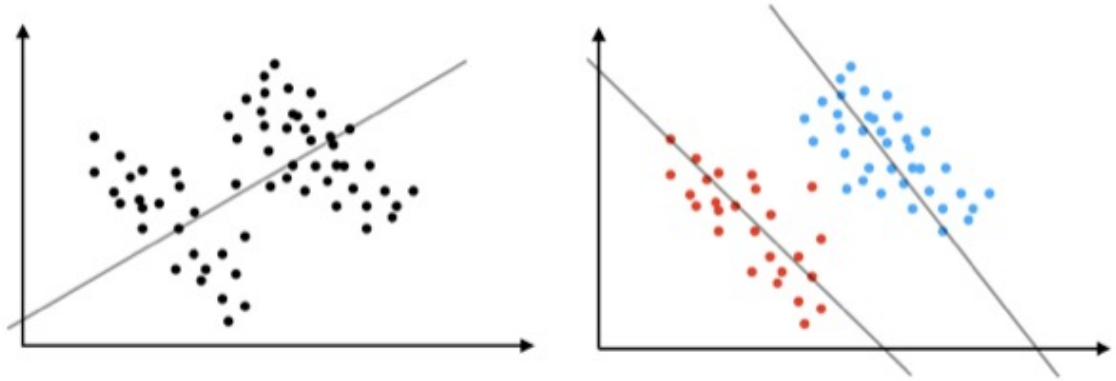
Age	Population (%)		Severe cases/100k		Severe Case Risk	Efficacy
	% Not Vax	% Fully Vax	Not Vax	Fully Vax	Ratio w/ 30-39 UnVax	vs. severe disease
12-15	62.1%	29.9%	0.30	0.00	1/20x	**100%**
16-19	21.9%	73.5%	1.60	0.00	1/4x	**100%**
20-29	20.5%	76.2%	1.50	0.00	1/4x	**100%**
30-39	16.2%	80.9%	6.20	0.20	1	**96.8%**
40-49	13.2%	84.4%	16.50	1.00	2.7x	**93.9%**
50-59	10.0%	88.0%	40.20	2.90	6.5x	**92.8%**
60-69	8.8%	89.8%	76.60	8.70	12.4x	**88.7%**
70-79	4.2%	94.6%	190.10	19.80	30.7x	**89.6%**
80-89	5.6%	92.6%	252.30	47.90	40.7x	**81.1%**
90+	6.1%	90.5%	510.9	38.60	82.4x	**92.4%**

Diagnosing Simpson's Paradox

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Coefficients show contradictory patterns.
- Adding variables or replacing variables causes major changes to coefficient values.



A visual example: the overall trend reverses when data is grouped by some colour-represented category.

https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated

Summary: When Tests Fail

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Transform the response if errors are non-Gaussian or heteroscedastic;
- Transform the predictors if lack of fit indicates non-linear relationships are present;
- Add new variables (i.e., new models) for Simpson's paradox and lack of fit;
- Remove, combine or center variables for multicollinearity;
- Use principal components regression for multicollinearity
- Additional methods (not covered in this course):
  - Ridge regression for multicollinearity
  - Robust methods for regression.

Additional Resources and References

Agenda

- Review
- Diagnostics
  - Graphical Diagnostics
  - Analytical Diagnostics
- Summary

- Chapter 3 & 6- Tibshirani et. Al. “An Introduction to Statistical Learning”, 2021
- Variable Selection Methods-
  - https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html
  - https://www.r-bloggers.com/2010/05/variable-selection-using-automatic-methods/

Multiple Linear Regression
Transformations, Qualitative Models & ANCOVA

Laura E. Barnes
&
Julianne Quinn

Agenda

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Review of Multiple Regression
- Transformation of Response and Predictors
- Qualitative Models
- ANCOVA

Review of Regression

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Multiple Regression: A method for measuring and modeling the relationship between sets of variables.
- Multiple Linear Regression:
  - $y = f(X) + \epsilon = X\beta + \epsilon$
- Regression is the solution to an optimization problem. We find a linear fit to the data that minimizes the sum of squared errors.
- Model assumptions
- Metrics of models: R^2, Adjusted-R^2, AIC, BIC
- Model comparison: Partial F test, test sets and cross-validation

Model Diagnostics

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Graphical Diagnostics:
  - Residuals vs. fitted
  - Scale-Location plot of square root of absolute standardized residuals vs. fitted
  - QQ plot of standardized residuals
  - Residual-Leverage plot
- Analytical Diagnostics:
  - We can also diagnose problems in the regression by analyzing the results: coefficient values, performance tests.
- Multicollinearity is apparent in the flipped signs for coefficients and for large changes in coefficient values with small changes to values in the observations.
- Diagnosing Simpson's Paradox with coefficient values: coefficients show contradictory patterns; adding variables or replacing variables causes major changes to coefficient values.

Influential Points and Outliers

Agenda

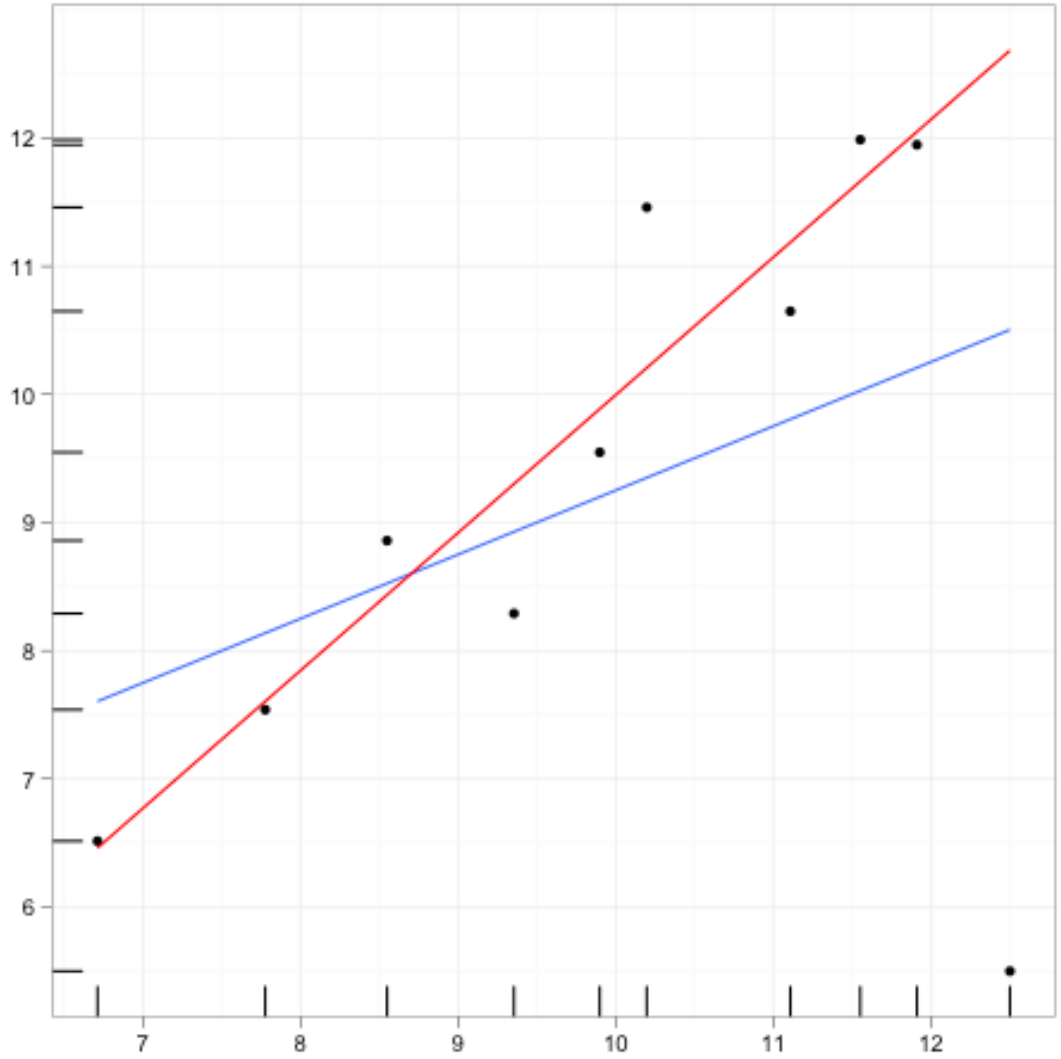
- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Influential observations, outliers and high leverage points:
  - Influential points can change the values of parameter estimates and other results by their removal.
  - Outliers do not fit the model well.
  - Observations with high leverage are far away from the average of predictor values.
  - High leverage observations are potential influential points.

Influential Points and Outliers

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

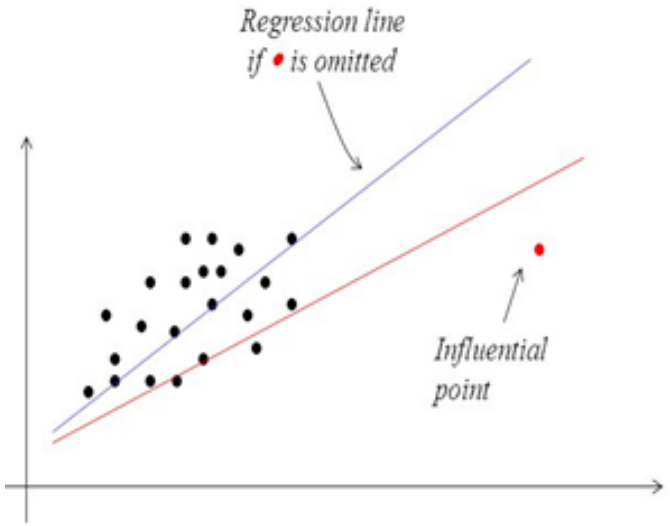


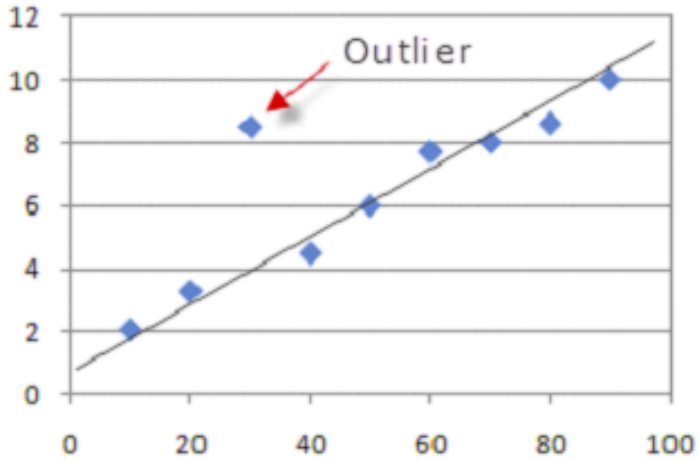
https://stats.stackexchange.com/questions/65912/precise-meaning-of-and-comparison-between-influential-point-high-leverage-point

Influential Points & Outliers

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview



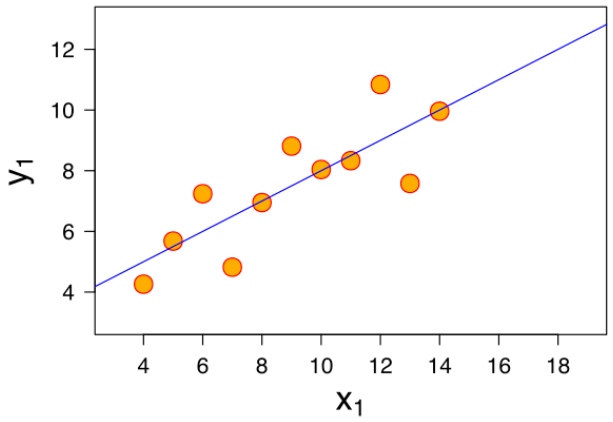


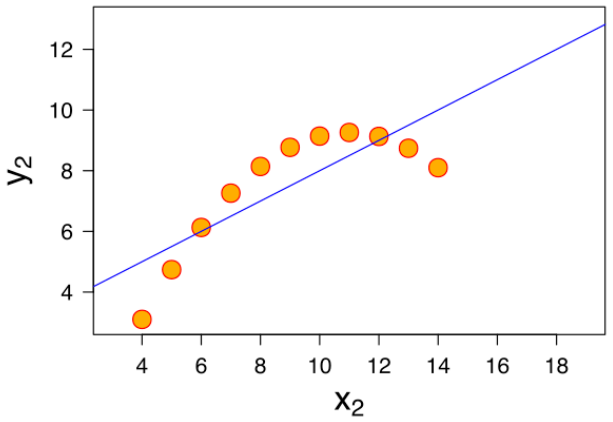
https://medium.com/@mail.amruthasasidharan/influential-points-vs-outliers-2c70c7b8d676

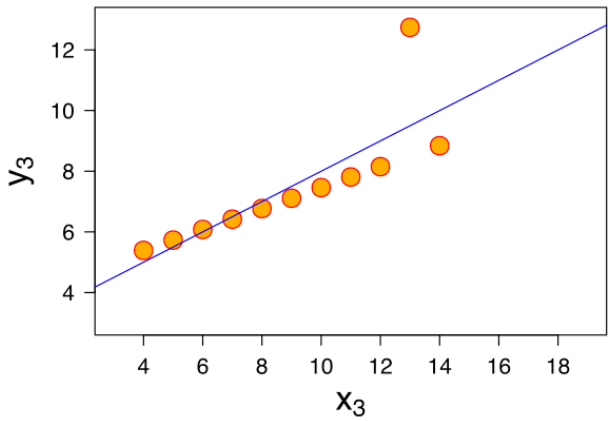
Influential Points Example

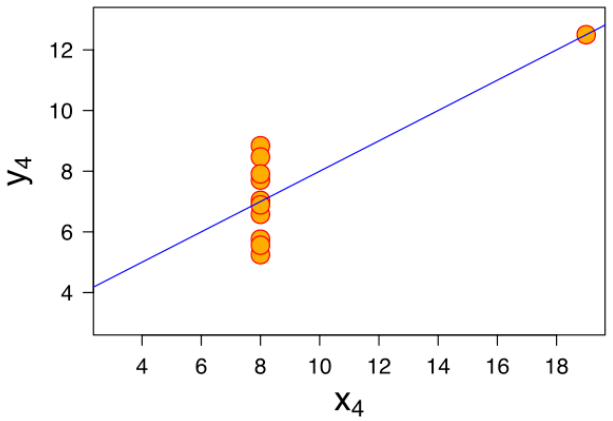
Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview









Leverage and Cook's Distance

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Cook's distance attempts to identify influential observations using leverage, *h*.
- Cook's distance is calculated from the leverage and standardized residual of an observation:

$$C_i = r_i^2 \frac{h_i}{p(1-h_i)}$$

- where r_i is the standardized residual, and p is the number of parameters including β_0.
- What does a large Cook's distance mean?

When Tests Fail

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Transform the response: Box-Cox plot
- Transform the predictors;
- Add new variables (i.e., new models) for Simpson's paradox and lack of fit;
- Remove, combine or center variables for multicollinearity;
- Use principal components regression or ridge regression for multicollinearity; and
- Use robust methods for regression.

Response Transformation: Box-Cox

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- When lack of Gaussian errors or lack of fit are detected consider transformation of the response. This can be helped with Box-Cox plots.
- Box-Cox plots allow us to consider power transformations of the form:

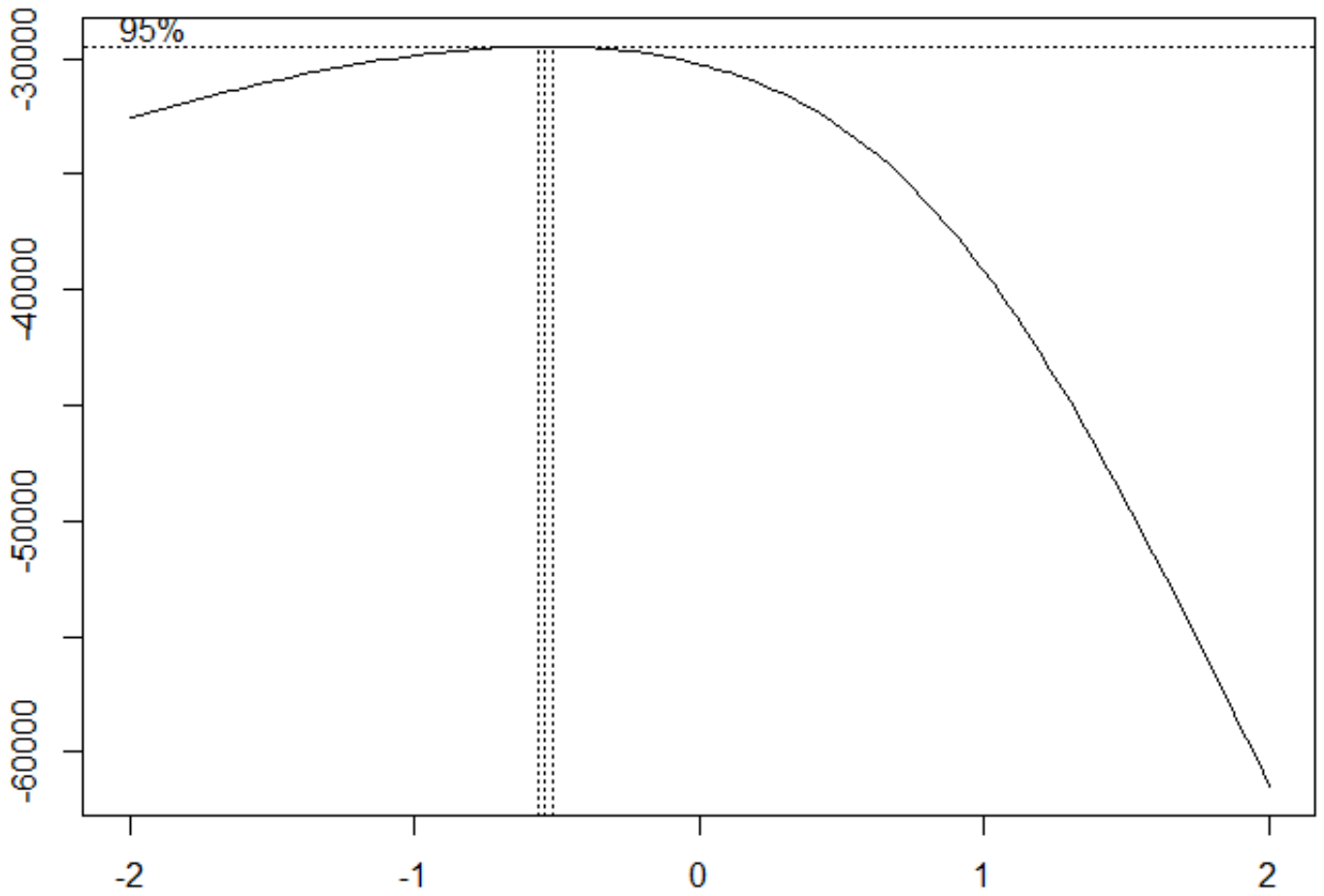
$$t_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

- The plot shows the 95% confidence interval for λ.
- Look for the integers that this interval brackets. If it is around 1 what do we do? No transformation.
- If it is around zero what do we do? Log transformation.

Example Box-Cox Plot

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

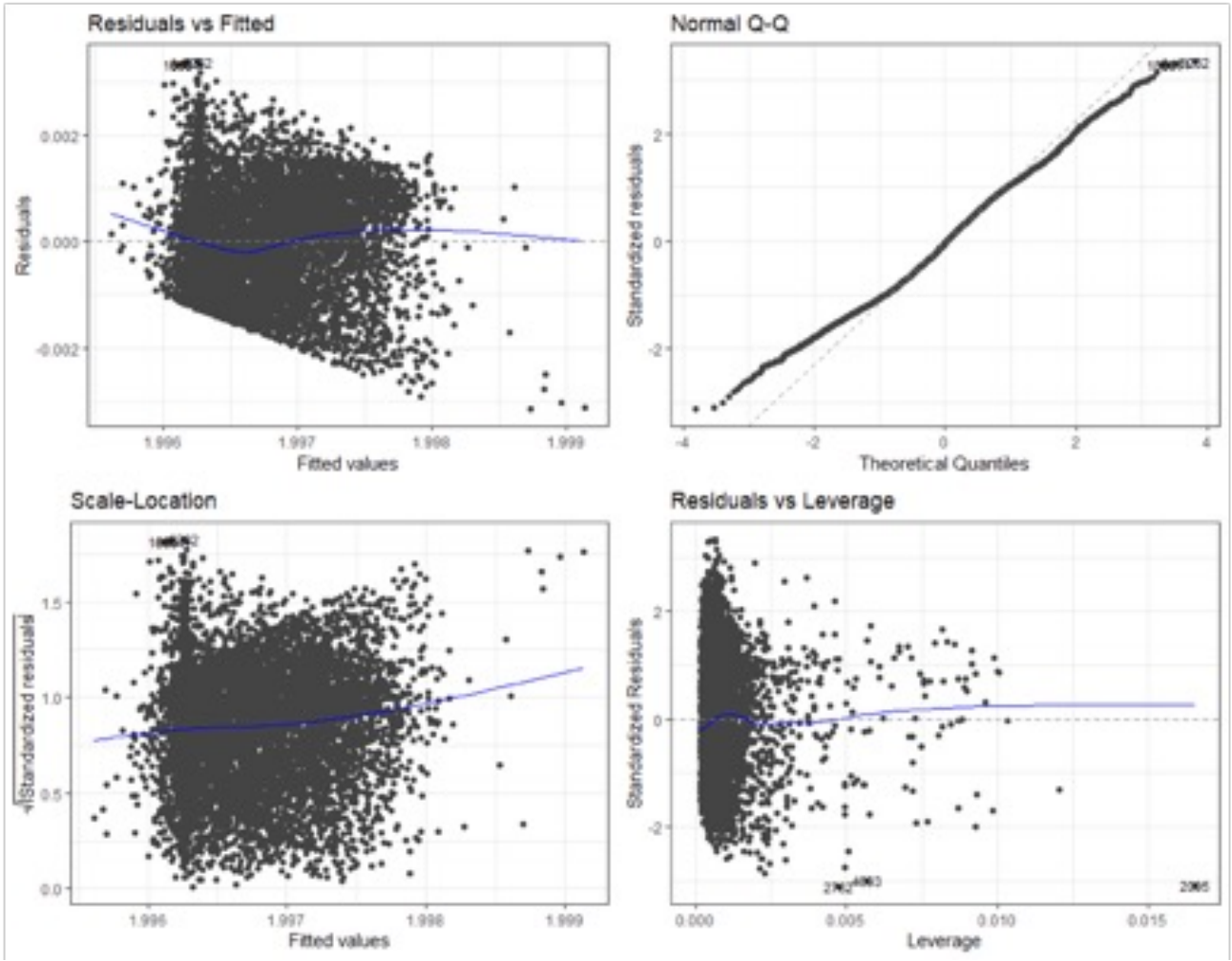


Example Diagnostics with Transformation

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

xdmgnd.lm2<-lm((ACCDMG^L-1)/L ~ TEMP + TRNSPD + TONS + CARS+ HEADEND1,data=xdmgnd)



Quantitative Regression Models

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- We begin with linear models but the world for our analyses has many nonlinearities.
- Least square regression requires that the models are linear in parameters, but not in variables. So, we can still model nonlinear processes with least squares regression.
- We apply nonlinear transformations to the predictor variables as we have used them for the response variable.
- Transformations of the predictors provide for model fit and reduce model bias.

First Order Models

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

- In first order models, all the exponents are 1 on the x variables (predictors)

First Order Models

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Consider the prediction of task performance of a person as a function of sleep.
- Let,

$$\mathrm{Y} = \mathrm{Task\ Performance}$$
$$\mathrm{X} = \mathrm{Sleep}$$

- The first order model to show this relationship:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Draw the first order model. Do you think this is correct?

Second Order Models

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1^2 + \cdots + \beta_k x_k + \beta_k x_k^2 + \epsilon$$

- Write a second order model.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Draw the second order model. Which of these models is more correct?

Review

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- What kind of model is this one?

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- What kind of model is this one?

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Are either of these models nonlinear in parameters?

  No

- Nonlinear in variables?

  Yes

More Than One Predictor

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Now, add caffeine intake as a predictor to the task performance example:

$$\mathrm{Y} = \text{Task Performance}$$
$$\mathrm{X}_1 = \text{Sleep}$$
$$\mathrm{X}_2 = \text{Caffeine Intake}$$

- A linear (main effects) model with the two predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Interactions

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Suppose we think that the effect of sleep on task performance depends on caffeine intake.
- This implies that these variables interact in their relationship with the response.
- A model with main effects and interaction terms for the task performance problem:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Is this model nonlinear in variables?

  Yes
- In parameters?

  No

Interaction Terms

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

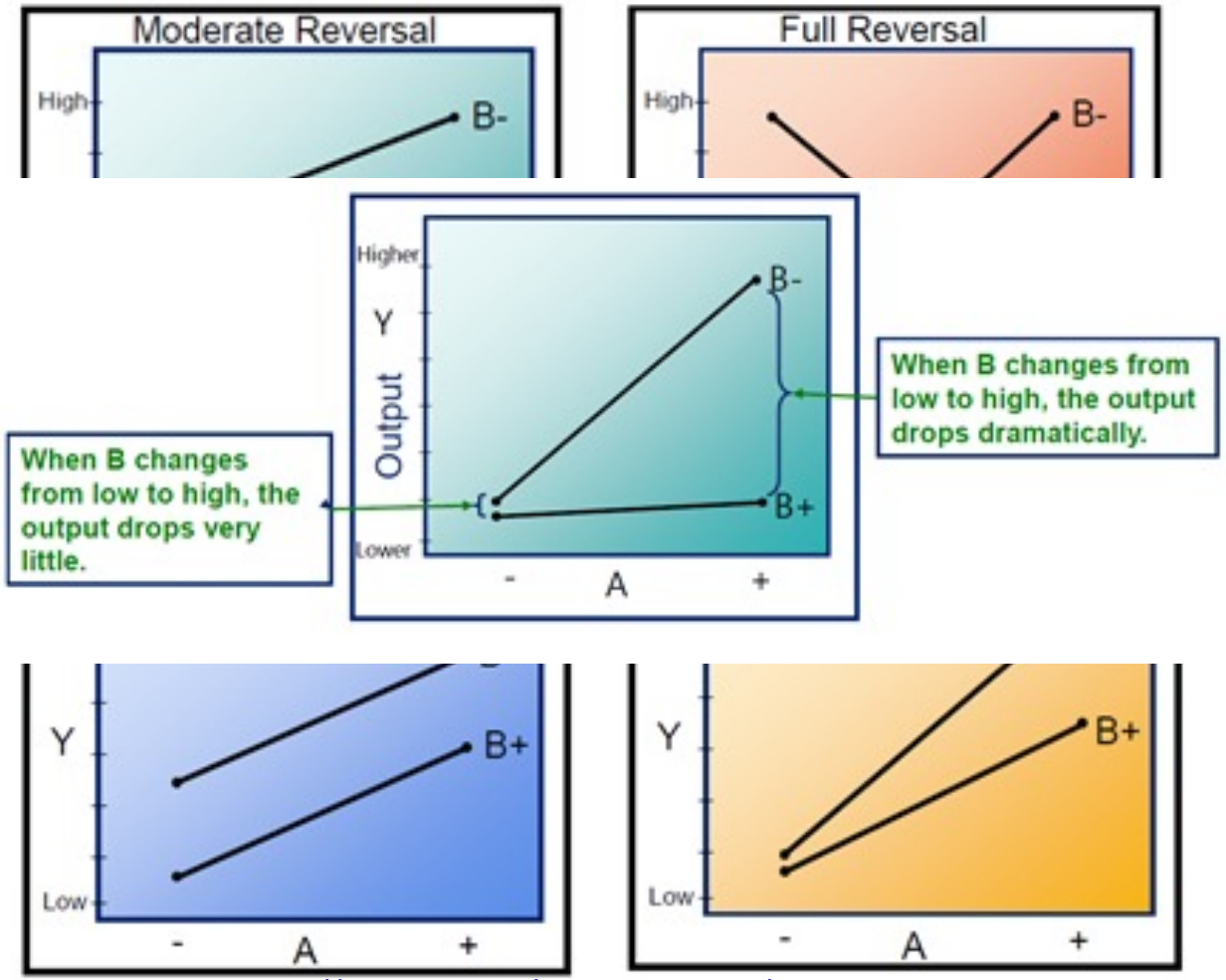
- An interaction term means the effect of a variable on the response depends on the values of the other variables in the interaction.
- For the task performance example an interaction between sleep and caffeine intake means that the effect of sleep on task performance depends on caffeine intake. It also means that the effect of caffeine intake on task performance depends on sleep.
- Can you develop an example of a possible interaction for your project?

Interaction Plots

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Different examples of interactions:



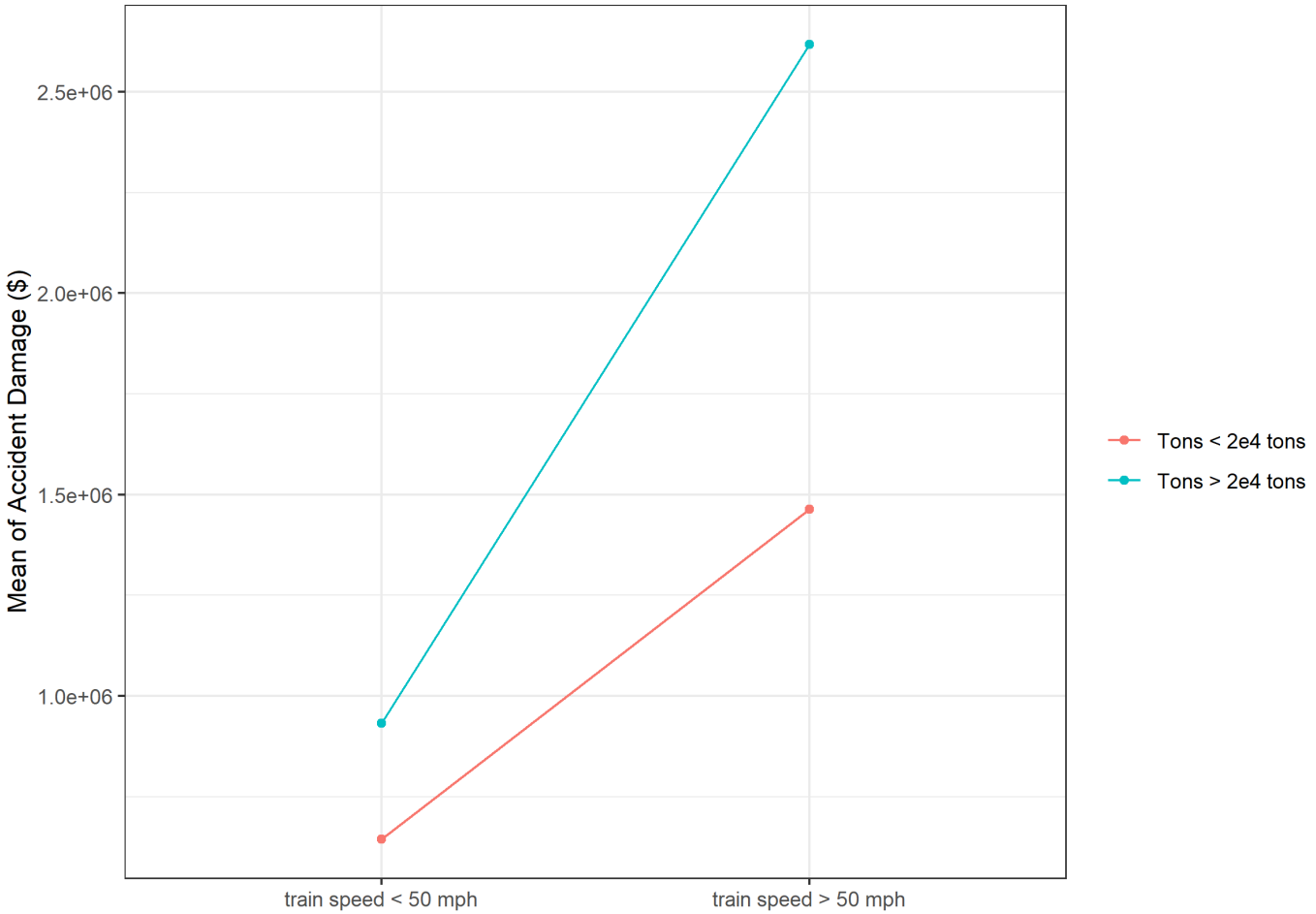
https://medium.com/analytics-vidhya/the-significance-of-interaction-plots-in-statistics-6f2d3a6f77a3

Interaction Plot

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

Interaction between Train Speed and Train Weight in Influencing Accident Damage



Complete Second Order Model

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Complete second order models include all the first and second order predictors (including interaction terms).
- Write a complete second order model of exam performance versus sleep.
- Write a complete second order model of exam performance versus sleep and caffeine intake.
- Suppose we add a third predictor variable, hours studied. How many terms (coefficients or parameters) do we have in a complete second order model with 3 variables?

  9 coefficients and 10 parameters

Complexity

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- In general, for a complete second order model with $p, p > 1$ variables, we have $2p + \binom{p}{2}$ coefficients.
- As we increase the number of variables the number of terms is growing as p^2.
- Choose second order variables with caution. Occam's Razor!
- We may want to consider transformations other than polynomials and interactions.

Qualitative Regression Models

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Qualitative predictors or categorical predictors require coding to use in regression models.
- The different values of the qualitative variable are called levels. Suppose we have a qualitative variable color with three values: red, white and blue. This variable has 3 levels.
- The only nonlinear relationship between qualitative variables and the response is with interactions. Why don't we use squared terms?

Treatment Contrasts

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- A convenient method to represent categorical variables by numbers is with treatment contrasts or dummy variables.
- The default coding for qualitative variables in R is with treatment contrasts or dummy variables.
- If the qualitative variable has m levels then it can be encoded with m – 1 dummy variables. Each dummy variable represents a level of the qualitative variable.

Treatment Contrasts

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Each dummy variable takes on one of two values: 1 if the value of the qualitative variable equals the level represented by the dummy variable and 0 otherwise.
- One level for the qualitative variable is represented by all 0 values for the dummy variables. As we will see the choice of which level is represented this way is important.
- R will create the m – 1 dummy variables automatically using lexicographic selection.

Example Qualitative Regression Model

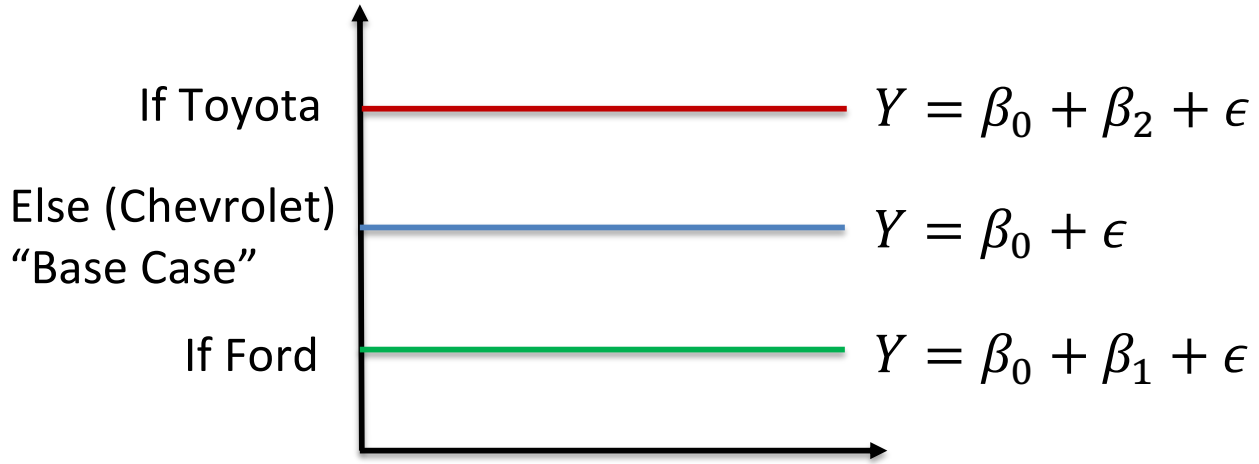
Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Consider MPG of different cars. Suppose we have cars in three brands: Toyota, Ford and Chevrolet.
- Brand is a qualitative variable that may be related to MPG.
- Code the brand variable with 2 dummy variables as follows:

$$X_1(y) = \begin{cases} 1 & \text{if Ford} \\ 0 & \text{else} \end{cases} \qquad X_2(y) = \begin{cases} 1 & \text{if Toyota} \\ 0 & \text{else} \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$



Example Qualitative Regression Model

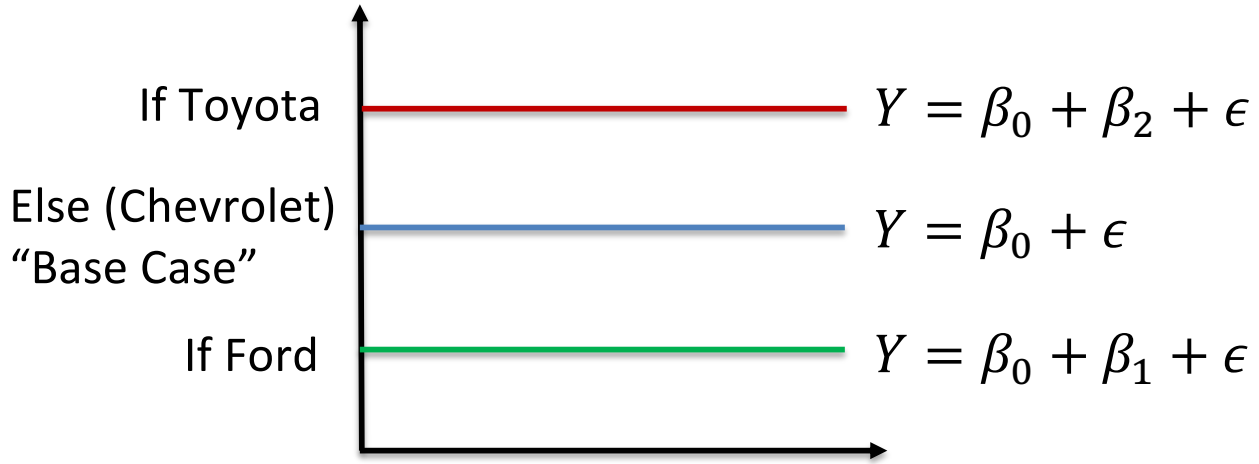
Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

So β_0 is the intercept under the base case (Chevrolet)
β_1 tells us how the intercept changes for Ford relative to the base case
β_2 tells us how the intercept changes for Toyota relative to the base case

$$X_1(y) = \begin{cases} 1 & \text{if Ford} \\ 0 & \text{else} \end{cases} \qquad X_2(y) = \begin{cases} 1 & \text{if Toyota} \\ 0 & \text{else} \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$



Example Qualitative Regression Model

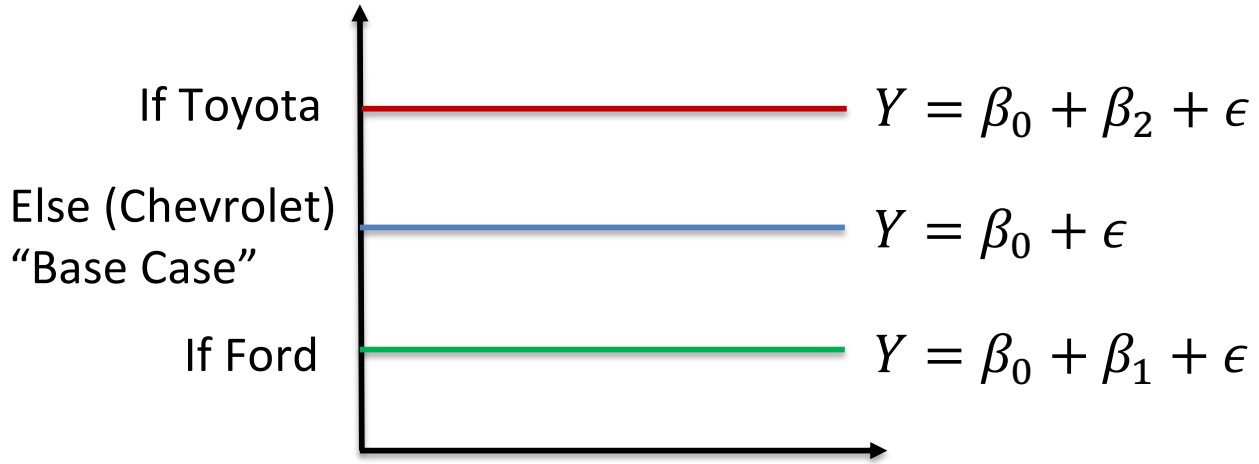
Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

What would be the encoding if we added another brand (e.g. Honda) to the model?

$$X_1(y) = \begin{cases} 1 & \text{if Ford} \\ 0 & \text{else} \end{cases} \qquad X_2(y) = \begin{cases} 1 & \text{if Toyota} \\ 0 & \text{else} \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$



Regression Results of MPG

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model.cause: *mpg ~ brand*
- Result:

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	21.9	1.18	18.5	1.04e-29
2	brandford	-2.53	1.57	-1.62	1.10e- 1
3	brandtoyota	3.02	1.48	2.04	4.44e- 2

Tests of Understanding

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- The main effects model with MPG (Y) as a function of brand:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Which value is the base case? Is this the value that would be chosen by R?
- For your project, how would you code the variable CAUSE?
- CAUSE has 5 levels: E, H, M, S, T.
- Write a linear model with ACCDMG as a function of CAUSE (with five levels)

Regression Results - Train Accidents

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model.cause: *ACCDMG ~ Cause*
- Result:

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	13.0	0.0225	579.	0
2	Cause(H) Train operation - Human Factors	-0.167	0.0306	-5.46	0.0000000490
3	Cause(M) Miscellaneous Causes Not Otherwise Listed	0.0688	0.0322	2.13	0.0328
4	Cause(S) Signal and Communication	-0.392	0.111	-3.55	0.000393
5	Cause(T) Track, Roadbed and Structures	0.00591	0.0271	0.218	0.827

Examples of Qualitative Regression Model: 2 Variables

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- We now add another categorical variable Cylinder to the MPG problem. Cylinder has two levels: more than 5 cylinders or not.
- For cylinder, define the dummy variable:

$$X_3(y) = \begin{cases} 1 & \text{if more than 5 cylinders} \\ 0 & \text{else} \end{cases}$$

Tests of Understanding

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Write a linear, main effects model with MPG (Y) as a function of brand and cylinder.
- Write a main effects plus interaction model with MPG (Y) as a function of brand and cylinder.
- Write a complete second order model with MPG (Y) as a function of brand and cylinder.

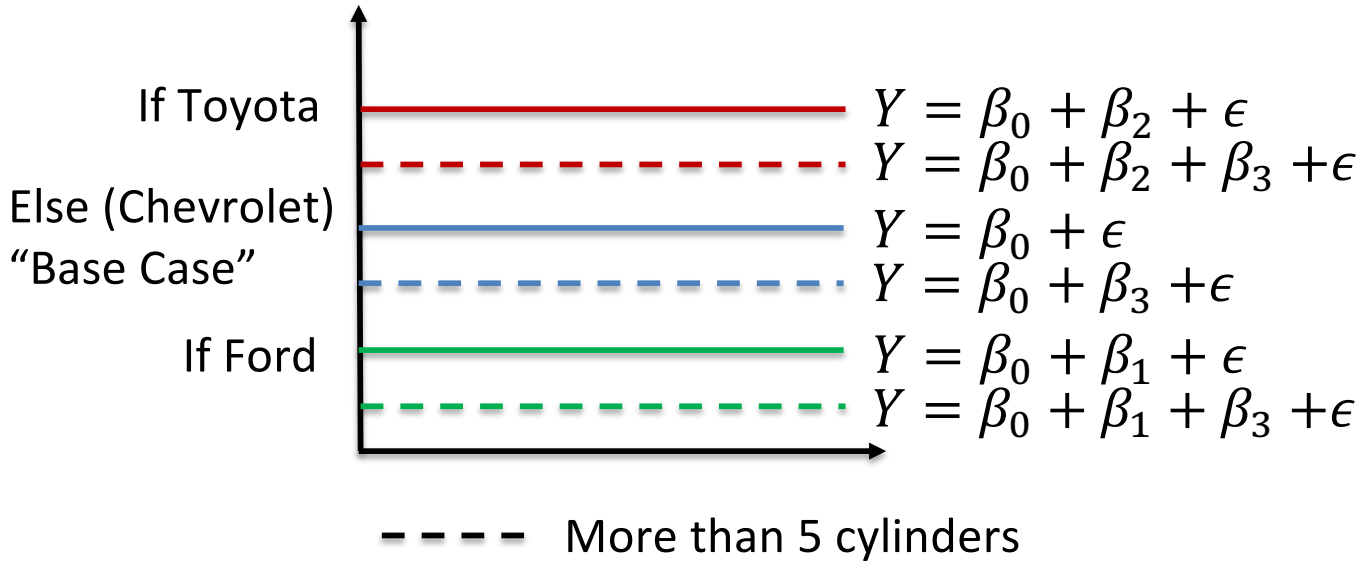
Tests of Understanding

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Write a linear, main effects model with MPG (Y) as a function of brand and cylinder.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$



β_3 tells us how the intercept changes with >5 cylinders, which here is assumed to be the same across all brands

Regression Results of MPG

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model.cause: *mpg ~ brand + cylinder*
- Result:

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	28.2	1.61	17.5	4.99e-28
2	brandford	-1.79	1.37	-1.31	1.95e- 1
3	brandtoyota	0.00754	1.41	0.00536	9.96e- 1
4	greaterthan5TRUE	-7.10	1.39	-5.09	2.60e- 6

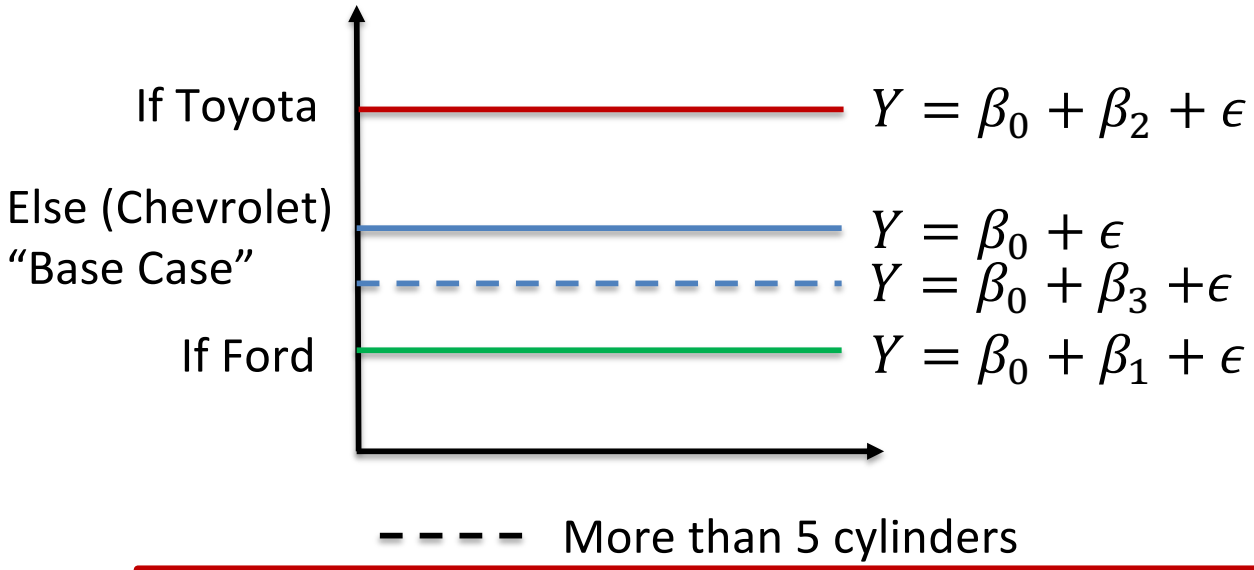
Tests of Understanding

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Write a main effects plus interaction model with MPG (Y) as a function of brand and cylinder.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon$$



β_3 tells us how the intercept changes with >5 cylinders, FOR THE BASE CASE ONLY

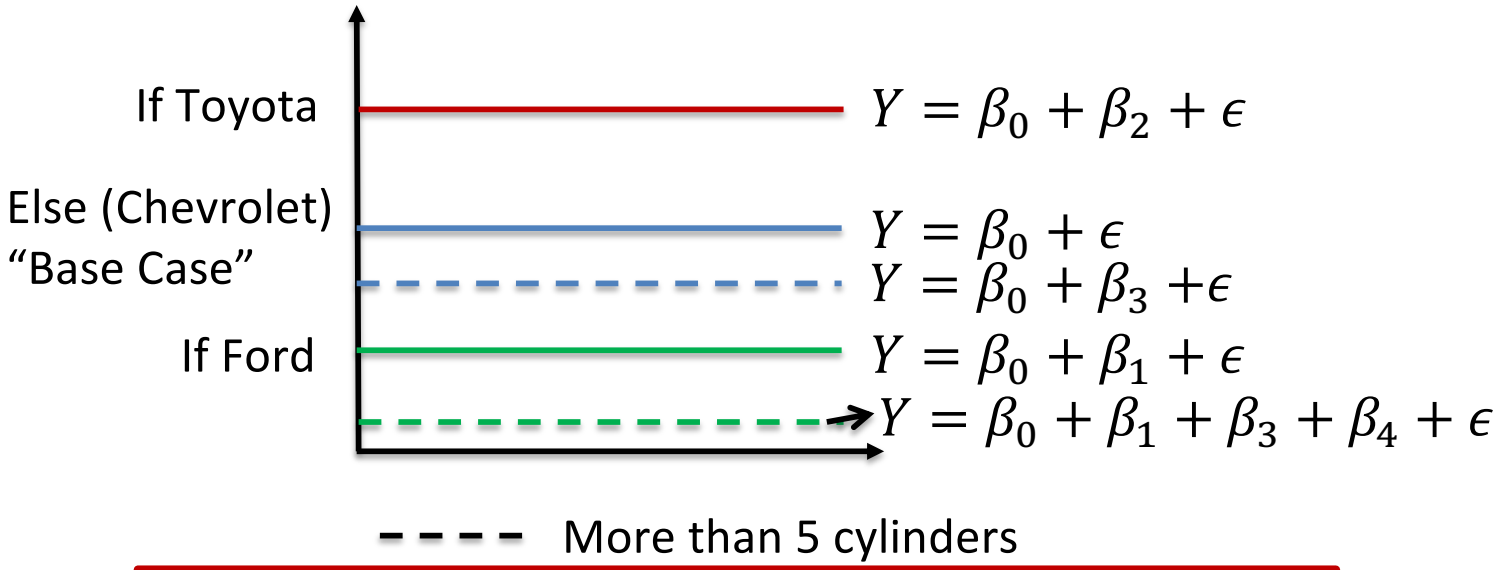
Tests of Understanding

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Write a main effects plus interaction model with MPG (Y) as a function of brand and cylinder.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon$$



β_4 tells us how the intercept changes with >5 cylinders for a Ford compared to how it changes for a Chevy

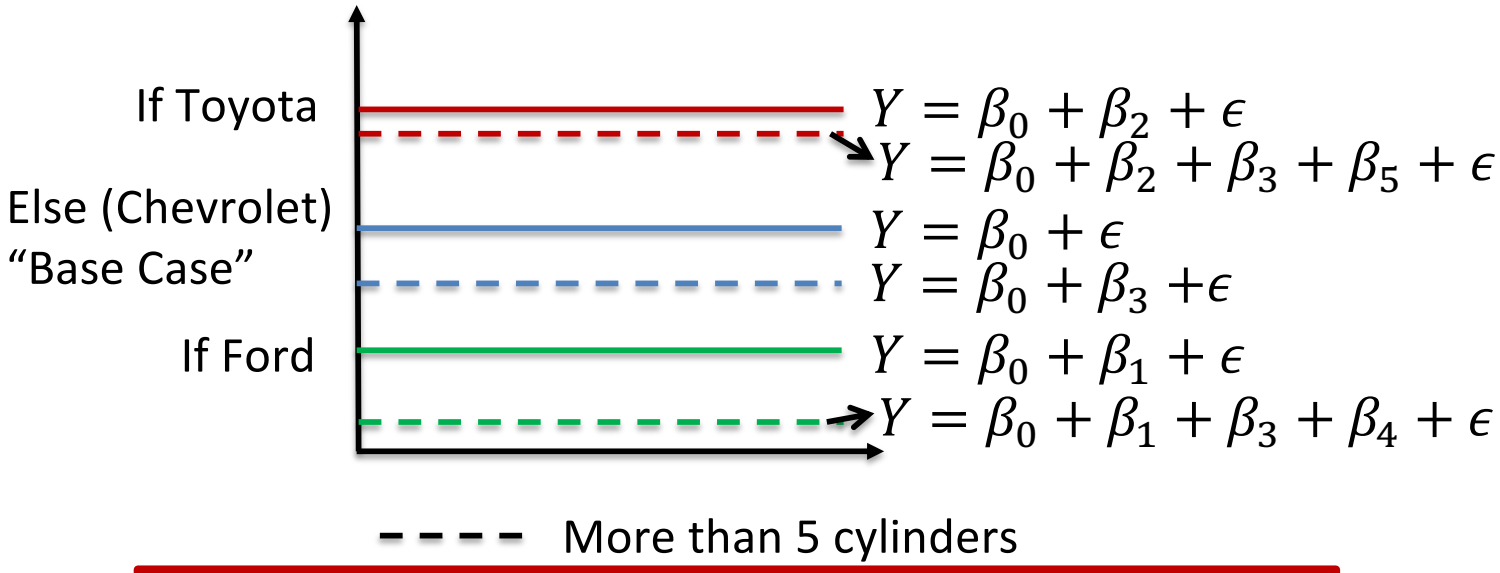
Tests of Understanding

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Write a main effects plus interaction model with MPG (Y) as a function of brand and cylinder.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon$$



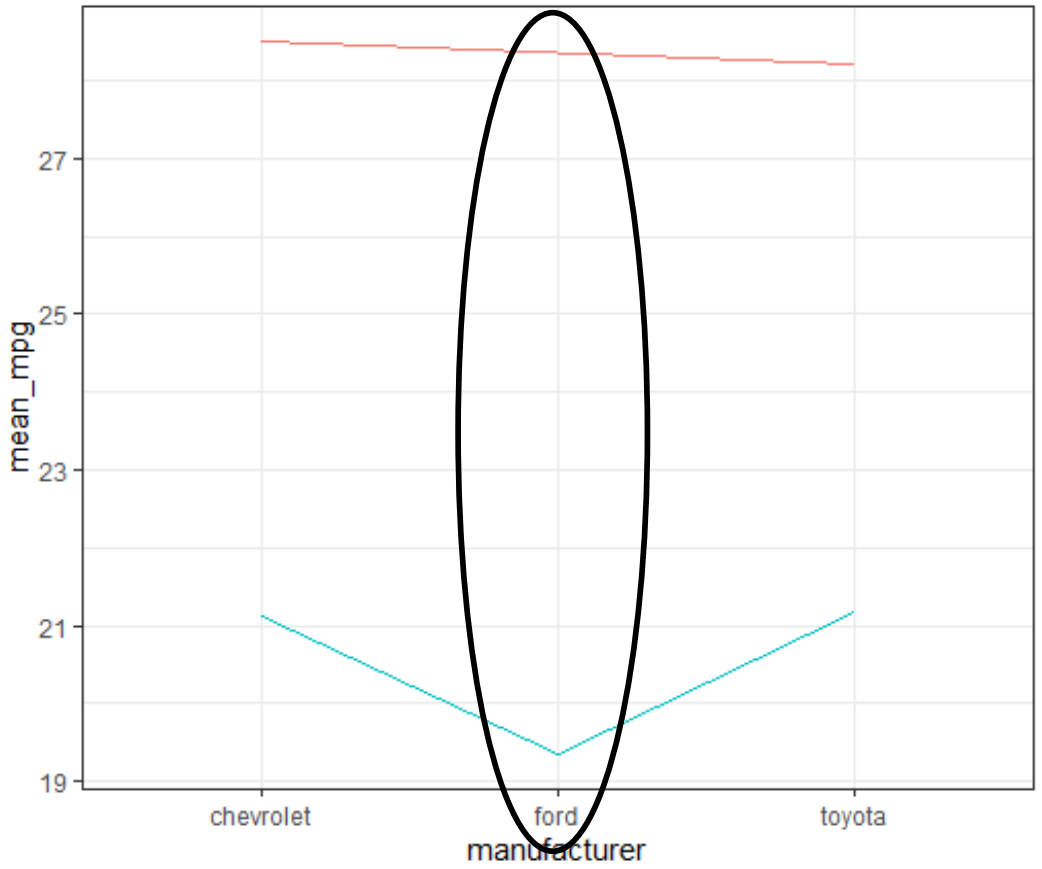
β_5 tells us how the intercept changes with >5 cylinders for a Toyota compared to how it changes for a Chevy

Interaction Plot of MPG

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- We can see if the effect of having >5 cylinders differs by brand from an interaction plot



The difference is greater for a Ford than for a Chevy or Toyota

Analysis of Variance

Agenda

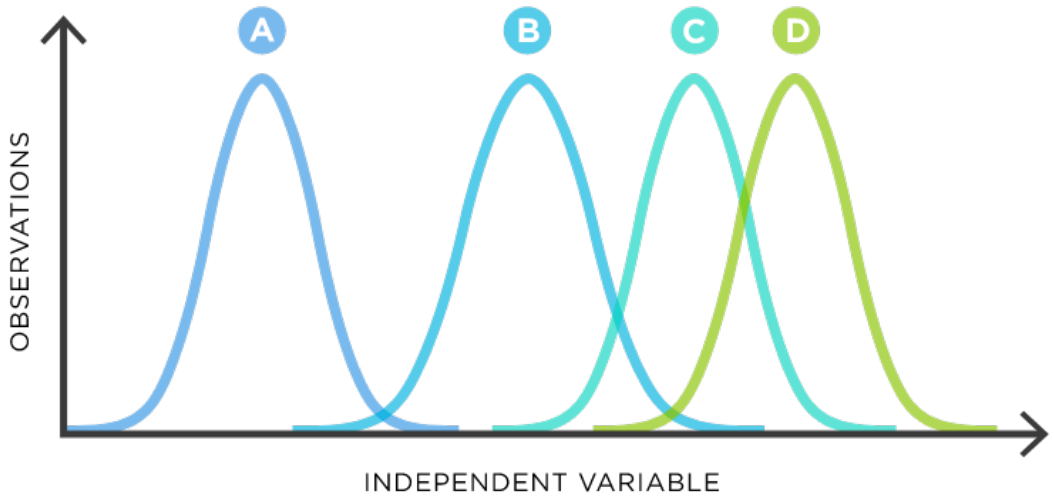
- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- ANOVA provides a method for multiple comparisons of means.
- A one-way ANOVA considers one predictor variable at multiple levels.
- ANOVA treats all predictors as qualitative variables or factors. So it converts quantitative variables to qualitative variables.
- This means it does not require the linear independence assumption but it does reduce the interpretability of the results. Both regression and ANOVA produce identical results for qualitative variables.
- Since regression gives us more information we will use it with both observational and experimental data.

ANOVA vs. Regression

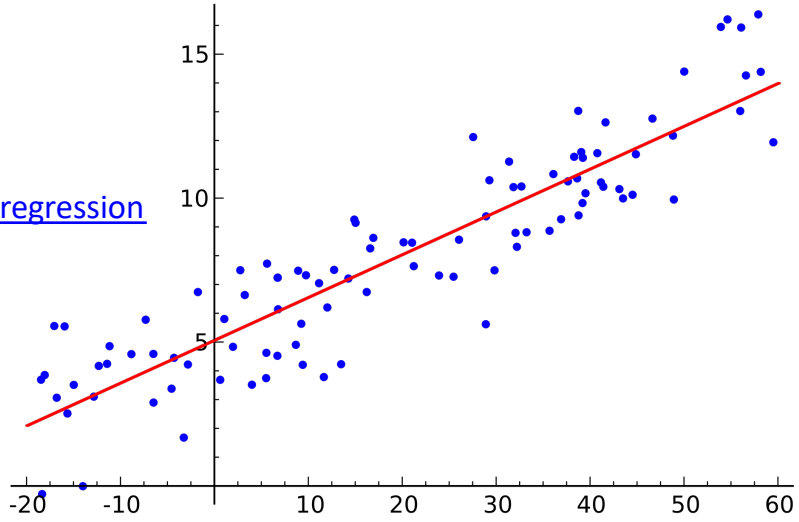
Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview



https://www.tibco.com/reference-center/what-is-analysis-of-variance-anova

https://en.wikipedia.org/wiki/Linear_regression



Analysis of Covariance

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

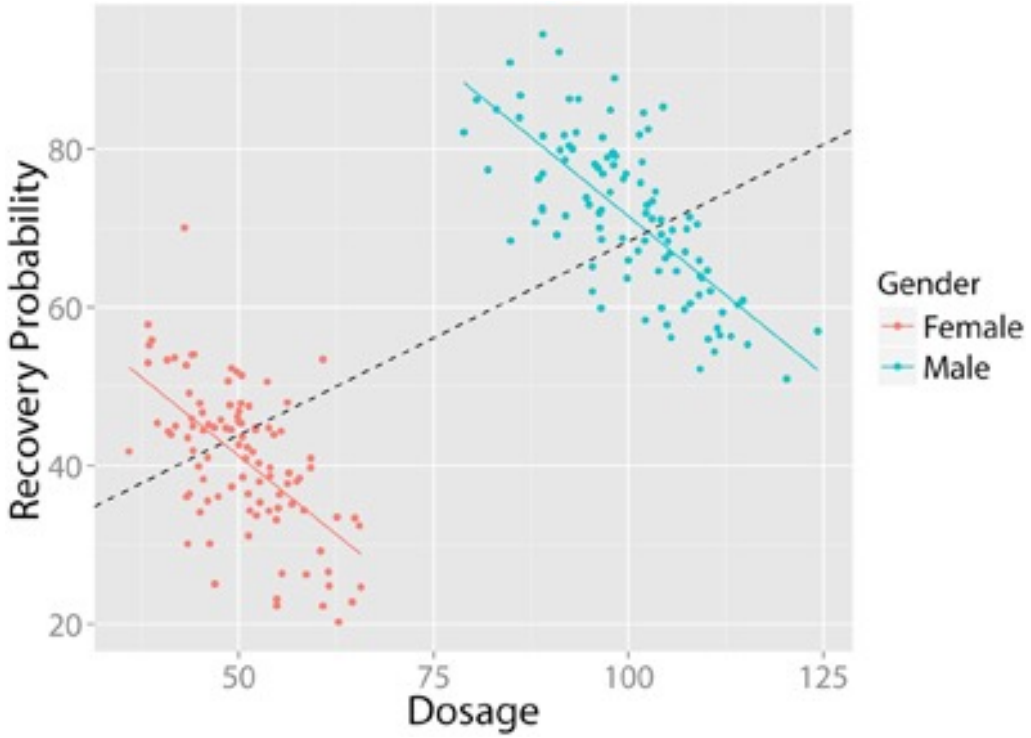
- Recall that regression allows for association tests while controlling for the values of other variables in the equation.
- ANCOVA combines qualitative and quantitative predictors or explanatory variables.
- Why would we want to use ANCOVA?

ANCOVA Example

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Patients who received a certain drug dosage and their probability of recovery. Did it work? Simpson's paradox strikes again!



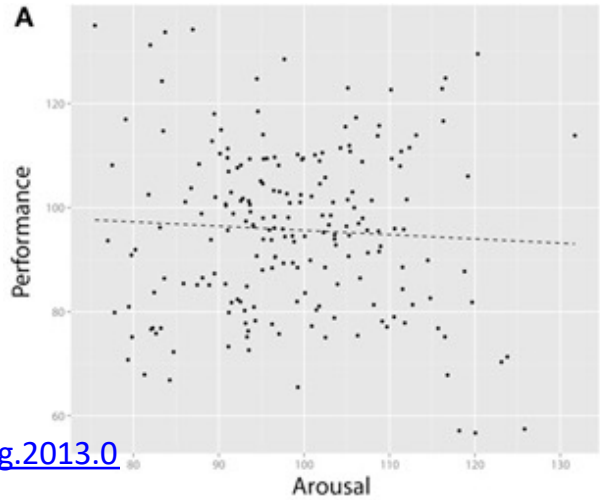
https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00513/full

ANCOVA Example

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Athletic performance vs. arousal



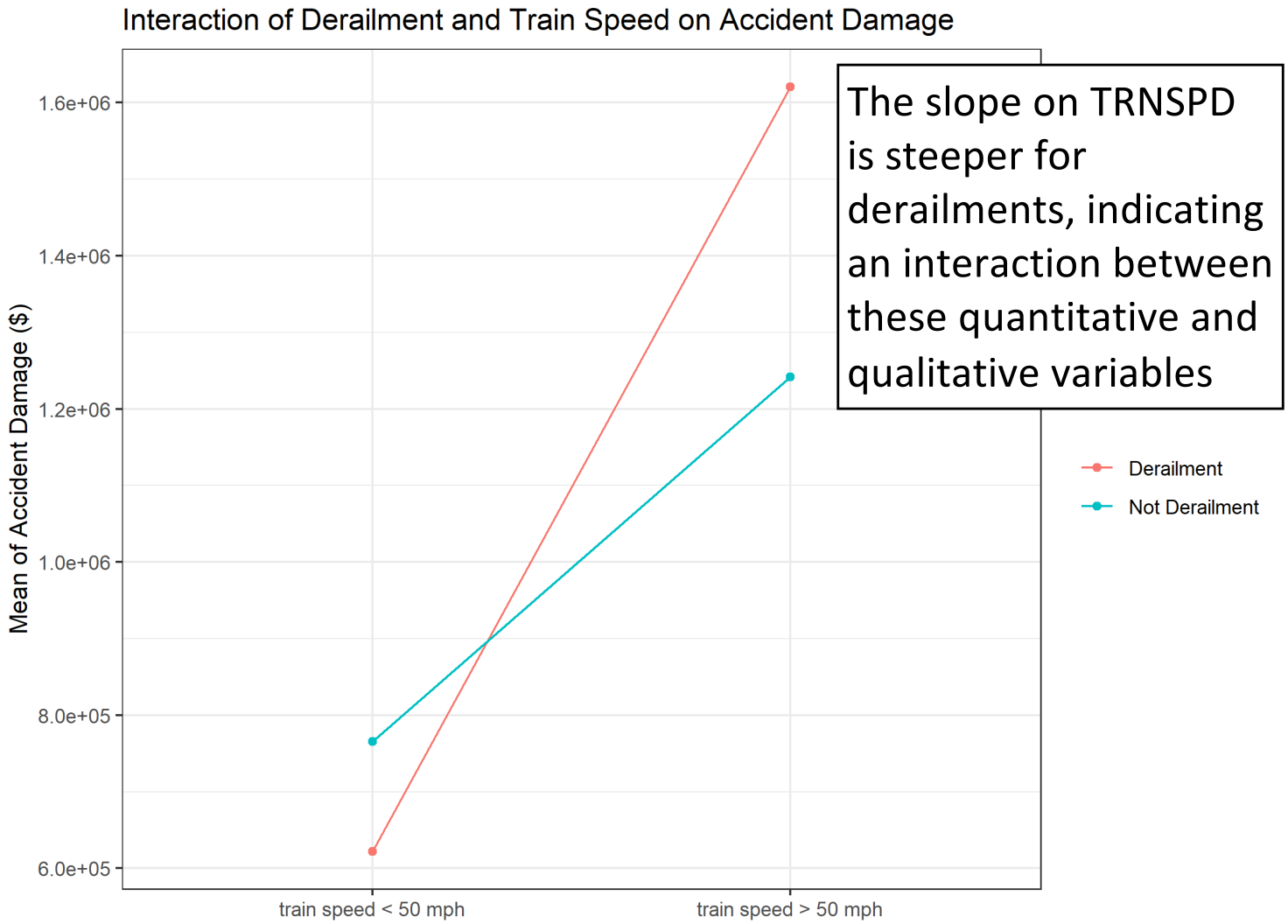
https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00513/full

- No apparent trend until variables on play styles is created.

Using the train dataset

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview



The slope on TRNSPD is steeper for derailments, indicating an interaction between these quantitative and qualitative variables

Quantitative Variable Added

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Recall the train accident problem. We built a model: ACCDMG ~ CAUSE (five levels: E, H, M, S, T)
- Train speed (TRNSPD) may predict damages and we want to add this quantitative variable to the model. This is the essence of ANCOVA.
- Write the main effects models: ACCDMG ~ CAUSE and ACCDMG ~ CAUSE + TRNSPD and the interaction model.
- How can we compare them?
  - In general we test interactions using the Partial F test.
  - Regardless of the encoding we test a qualitative variable using the Partial F test.

Quantitative Variable Added

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Look at the main effects and the interaction results that follow. Use Partial F tests to make a recommendation.
- Look at the t-tests. Do we ever remove a variable in an interaction term from the main effects part of the model?

Main Effects Results ANCOVA

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model.cause: log(*ACCDMG* + 1) ~ *CAUSE*
- Result:

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.1220 -0.6564 -0.1945  0.4872  4.4042

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    13.029499   0.022487 579.416  < 2e-16 ***
Cause(H) Train operation - Human Factors       -0.167001   0.030582  -5.461  4.9e-08 ***
Cause(M) Miscellaneous Causes Not Otherwise Listed  0.068765   0.032210   2.135 0.032804 *
Cause(S) Signal and Communication              -0.391914   0.110520  -3.546 0.000393 ***
Cause(T) Track, Roadbed and Structures          0.005906   0.027063   0.218 0.827251
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8241 on 7249 degrees of freedom
Multiple R-squared:  0.01097,   Adjusted R-squared:  0.01043
F-statistic:  20.1 on 4 and 7249 DF,  p-value: < 2.2e-16
```

Main Effects Results ANCOVA

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model.cause + trnspd: log(*ACCDMG* + 1) ~ *CAUSE* + *TRNSPD*
- Result:

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.4189 -0.5489 -0.1543  0.4473  4.6206

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   12.5789606  0.0246717 509.854  < 2e-16 ***
Cause(H) Train operation - Human Factors       0.0671701  0.0291899   2.301 0.021412 *
Cause(M) Miscellaneous Causes Not Otherwise Listed  0.0711896  0.0298879   2.382 0.017250 *
Cause(S) Signal and Communication             -0.1053513  0.1028932  -1.024 0.305921
Cause(T) Track, Roadbed and Structures         0.0928901  0.0252405   3.680 0.000235 ***
TRNSPD                                         0.0178435  0.0005214  34.224  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7647 on 7248 degrees of freedom
Multiple R-squared:  0.1486,    Adjusted R-squared:  0.148
F-statistic: 252.9 on 5 and 7248 DF,  p-value: < 2.2e-16
```

Interaction Results ANCOVA

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model.interaction: log(*ACCDMG* + 1) ~ *CAUSE* + *TRNSPD* + *CAUSE* : *TRNSPD*
- Results:

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.2273 -0.5374 -0.1439  0.4214  4.7154

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        12.731595   0.036384 349.922  < 2e-16 ***
Cause(H) Train operation - Human Factors           -0.180311   0.044111  -4.088 4.40e-05 ***
Cause(M) Miscellaneous Causes Not Otherwise Listed  0.167421   0.048111   3.480 0.000505 ***
Cause(S) Signal and Communication                  -0.280175   0.138565  -2.022 0.043216 *
Cause(T) Track, Roadbed and Structures             -0.225087   0.042553  -5.290 1.26e-07 ***
TRNSPD                                              0.011798   0.001190   9.911  < 2e-16 ***
Cause(H) Train operation - Human Factors:TRNSPD     0.013867   0.001794   7.729 1.23e-14 ***
Cause(M) Miscellaneous Causes Not Otherwise Listed:TRNSPD -0.003865   0.001512  -2.555 0.010633 *
Cause(S) Signal and Communication:TRNSPD            0.008460   0.009892   0.855 0.392484
Cause(T) Track, Roadbed and Structures:TRNSPD       0.014160   0.001462   9.688  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7513 on 7244 degrees of freedom
Multiple R-squared:  0.1785,    Adjusted R-squared:  0.1775
F-statistic: 174.9 on 9 and 7244 DF,  p-value: < 2.2e-16
```

Partial F Tests ANCOVA

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model.cause vs. Model.cause+trnspd
- Analysis of Variance Table
  - Model 1: log(*ACCDMG* + 1) ~ *CAUSE*
  - Model 2: log(*ACCDMG* + 1) ~ *CAUSE* + *TRNSPD*

```
Analysis of Variance Table

Model 1: log_accdmg_pls1 ~ Cause
Model 2: log_accdmg_pls1 ~ Cause + TRNSPD
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   7249 4923.0
2   7248 4238.1  1    684.87 1171.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Partial F Tests ANCOVA

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model.cause+trnspd vs. Model.interaction
- Analysis of Variance Table
  - Model 1: log(*ACCDMG* + 1) ~ *CAUSE* + *TRNSPD*
  - Model 2: log(*ACCDMG* + 1) ~ *CAUSE* + *TRNSPD* + *CAUSE* : *TRNSPD*

```
Analysis of Variance Table

Model 1: log_accdmg_pls1 ~ Cause + TRNSPD
Model 2: log_accdmg_pls1 ~ Cause + TRNSPD + Cause * TRNSPD
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   7248 4238.1
2   7244 4089.0  4    149.15 66.057 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Multiple Linear Regression

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Concept of multiple linear regression
- Least squares estimate
- Model assumptions
- Variable selections
- Model diagnostics
- Nonlinear variables
- Qualitative variables
- ANOVA and ANCOVA

How to Build Multiple Linear Regression Models?

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Building multiple linear regression model is not as simple as typing **lm(r ~.,data=xdmgnd)** in R!
- Begin with graphical analysis;
- Select variables (both quantitative and qualitative variables) for multiple linear regression models;
- Measure the performance of models: F tests, Adj-R^2, AIC, BIC, etc.;
- Diagnose models: graphical and analytical;
- Adjust models: transformations, higher order models, variable selection, PC regression, etc.;
- Repeat the above steps as necessary;
- Get several alternative models, select the best model(s) for recommendation.

Model Selection

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Model selection means choosing the model or models to use as the basis for our analysis and ultimately our recommendations.
- Model selection consists of choosing variables, transformations, and combinations among the variables and levels in qualitative variables.
- Model selection is important, but why?

Reasons for Model Selection

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Occam's Razor.
- Extra terms can add noise to the predictions. More data is not necessarily better.
- Multicollinearity.
- Leaving out variables causes inaccurate understanding and predictions. This is Simpson's paradox.
- It can cost more to get data on more variables.
- We have to make recommendations. If models give competing answers, we need to pick from among these.

Approaches to Model Selection

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- t-tests. This is not a good approach because of multiplicity.
- Partial F test results. This is a good approach, but not for non-nested models.
- Criterion based also good but with limits.
- Automated selection, forward, backward, and stepwise. Quick and dirty.
- Principal components can provides variable extraction versus selection.

Approaches to Model Selection

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Test sets when we have enough data.
- Cross-validation when we don't have enough data.
- Choose models based on diagnostics.
- Bootstrapping when nonparametric methods are needed.
- Model selection is hard! Focus on the problem (not the mechanics). This is why good systems engineers are in such demand.

Additional Resources and References

Agenda

- Review
- Transformations
- Qualitative Models
- Analysis of Covariance
- Overview

- Chapter 2, 3, 5 & 6- Tibshirani et. Al. “An Introduction to Statistical Learning”, 2021.