Statistical Modelling with Data

May 23 – June 02, 2023

Instructor: Qing (Leah) Li, Ph.D. Candidate at Cumming School of Medicine

qing.li2@ucalgary.ca

Thank you Dr. Thuntida Ngamkham for contributing the contents

Thank you Dr. Qingrun Zhang and Dr. Quan Long for contributing some slides

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression, Forward selection and Backward Elimination
  - Lecture 5: Model selection: Evaluate the reliability of the model chosen
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, equal variance assumptions and normality assumptions.
  - Lecture 7: Multiple regression diagnostics: identify multicollinearity and outliers and data transformation.
- Topic 5: Transfer learning
  - Lecture 8: Deep learning basics
  - Lecture 9: Transfer-learning (Bonus): standing on the shoulders of giants.

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression, Forward selection and Backward Elimination
  - Lecture 5: Model selection: Evaluate the reliability of the model chosen
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, equal variance assumptions, and normality assumptions.
  - Lecture 7: Multiple regression diagnostics: identify multicollinearity and outliers and data transformation.
- Topic 5: Transfer learning
  - Lecture 8: Deep learning basics
  - Lecture 9: Deep learning advances: Transfer-learning (Bonus).

Statistical Modelling with Data

Learning Outcomes: At the end of the course, participants will be able to

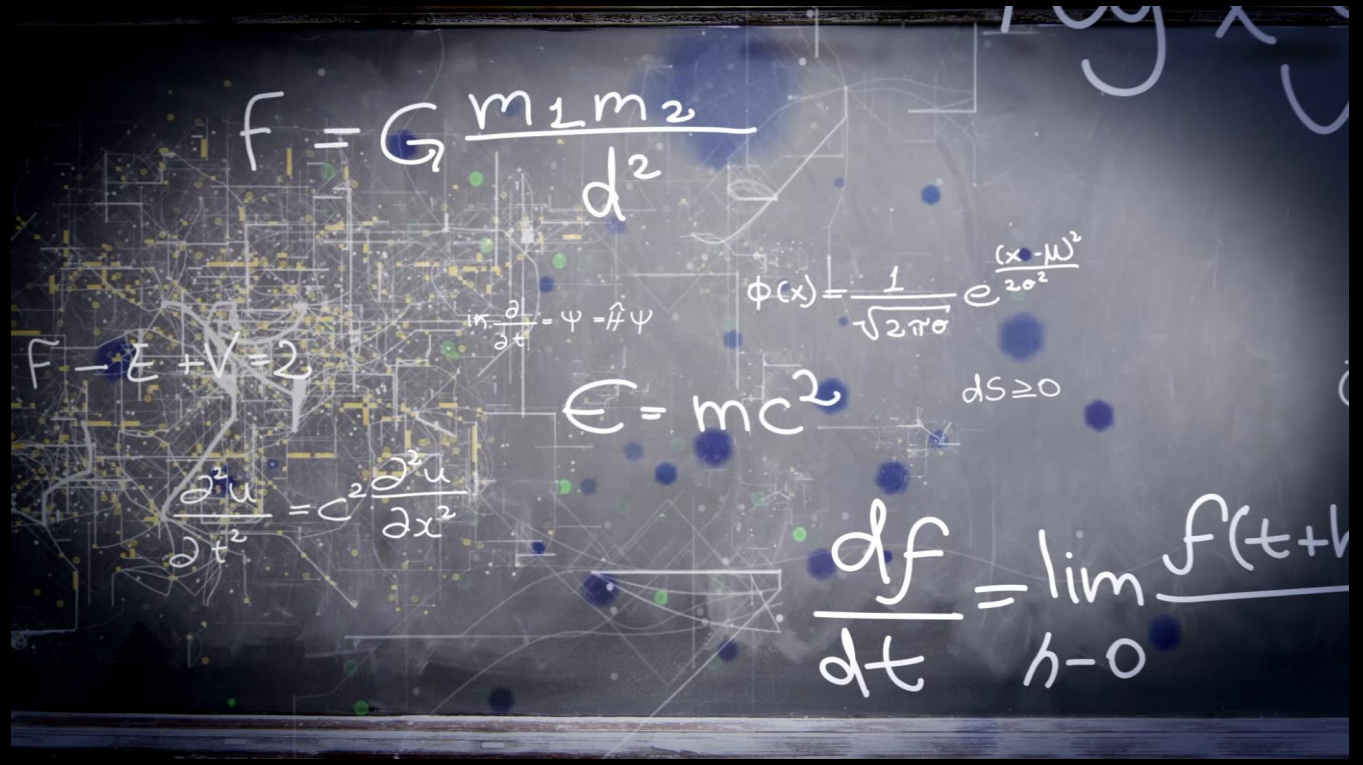
1. Model the multiple linear relationships between a response variable (Y) and all explanatory variables (both categorical and numerical variables) with interaction terms. Interpret model parameter estimates, construct confidence intervals for regression coefficients, evaluate model fits, and visualize correlations between a response variable (Y) and all explanatory variables (X) by graphs (scatter plot, residual plot) to assess model validity.
2. Predict the response variable at a certain level of the explanatory variables once the fit model exists.
3. Implement R-software and analyze statistical results for biomedical and other data.

- **Evaluations**
  1. Assignments will be posted on Slack (our communication tool with students).
  2. Students must attend 70% (6/9) of the sessions in order to receive the certificate and are encouraged to work on the assignments progressively throughout the course as the relevant material is covered.

Statistical Modelling with Data

- Supportive materials
  - Lectures slides (2023)
  - R code scripts (2023)
  - PDF (dated 2022)
  - Two Assignments (dated 2022)
- Slack channels
    - Recoding videos
    - Exercises
    - Course-documents

Lecture 7: Multiple regression diagnostics: identify multicollinearity and outliers and data transformation



Plots

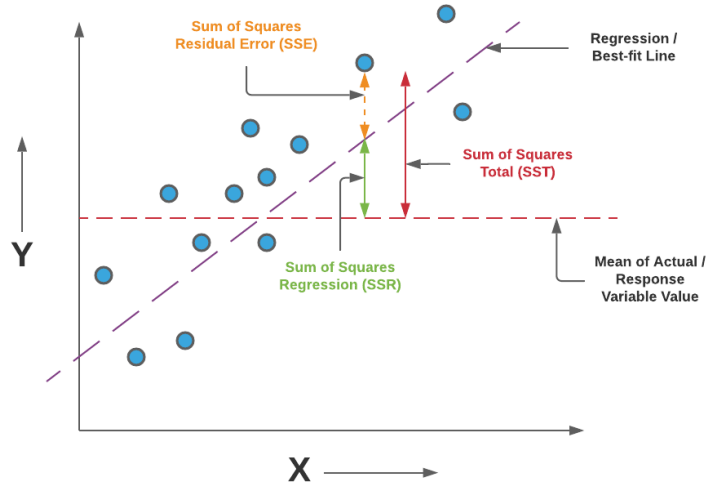
- R: plot()
- Pros:
  - Easy to write and learn by heart
- Cons:
  - Does not incorporate advanced functions such as fitting curves to points
- Multiple figures in one panel
  - par(mfrow=c(nrow, ncol))

- ggplot2 package: ggplot()
- Pros:
  - Multiple choices, high-level functions
- Cons:
  - It takes a while to understand and learn the functions by heart
- Multiple figures in one panel
  - library(gridExtra)
  - grid.arrange(p1,p2, nrow = 3,top = “Title").

Assumptions

response, coefficients of correlation, predictors

$$g(y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \mathrm{g}(y_i) = y_i$$



1. Linearity Assumption

   Relationships between all predictors and the response are linear

2. Independence Assumption

   Independence of observations

3. Equal Variance Assumption

   Error term (Residuals) has equal variance given any values of independent variables

4. Normality Assumption

   Error term (Residuals) is normally distributed

5. Multicollinearity

   No perfect collinearity and non-zero variance of independent variables

6. Outlier

   Error terms (Residuals) has expected value of zero given the values of independent variables

5. Multicollinearity

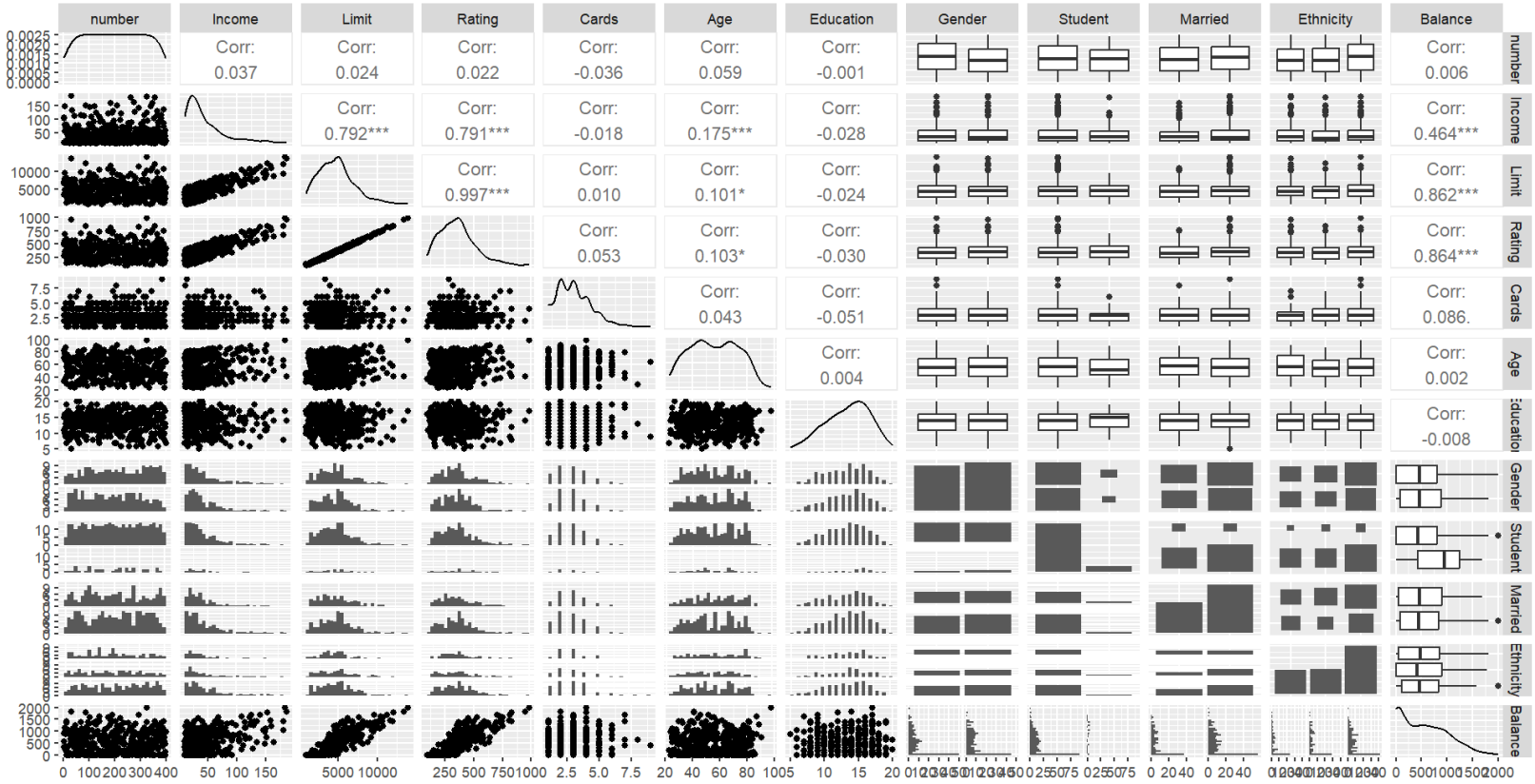
Often, two or more of the independent variables used in the model for E(Y) provide redundant information. That is, the independent variables will be correlated with each other. **When the independent variables are (linearly) correlated, we say that multicollinearity exists.** In practice, it is common to observe correlations among the independent variables. However, a few problems arise when serious multicollinearity is present in the regression analysis.

What Problems Do Multicollinearity Cause?

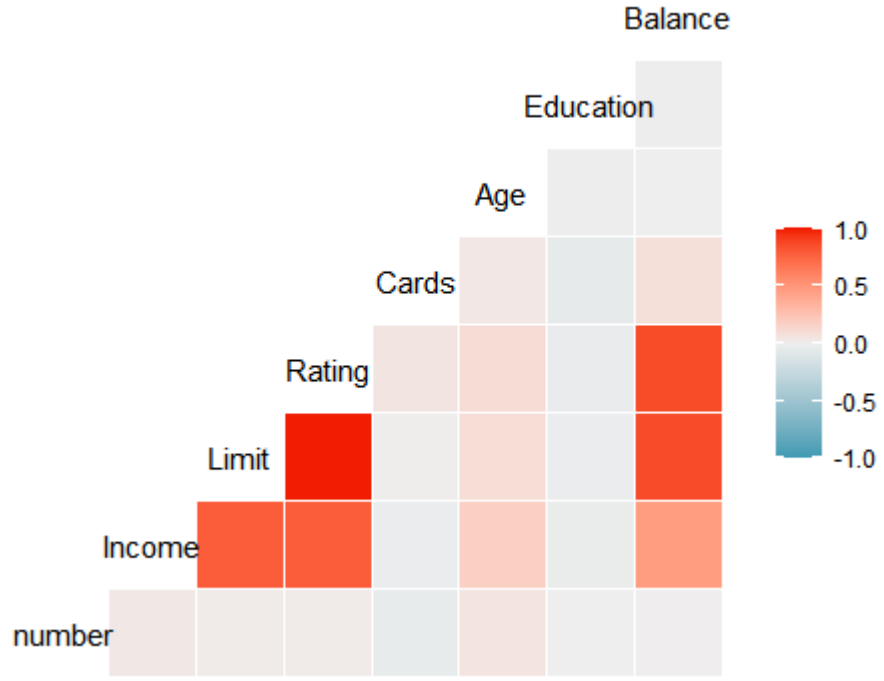
- The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.

- Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model. You might not be able to trust the p-values to identify independent variables that are statistically significant.

An example from credit data



ggpairs(credit)



ggcorr(credit)

Testing for Multicollinearity with Variance Inflation Factors (VIF)

VIF identifies correlation between independent variables and the strength of that correlation. It can be computed using the formula

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors. If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.

Statistical software calculates a VIF for each independent variable. Value of VIFs start at 1 and have no upper limit and can be interpreted as following;

*VIFs=1 indicates that there is no collinearity between this independent variable and any others.

*1<= VIFs <=5 suggest that there is a moderate collinearity, but it is not severe enough to warrant corrective measures.

VIFs > 5 or 10 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

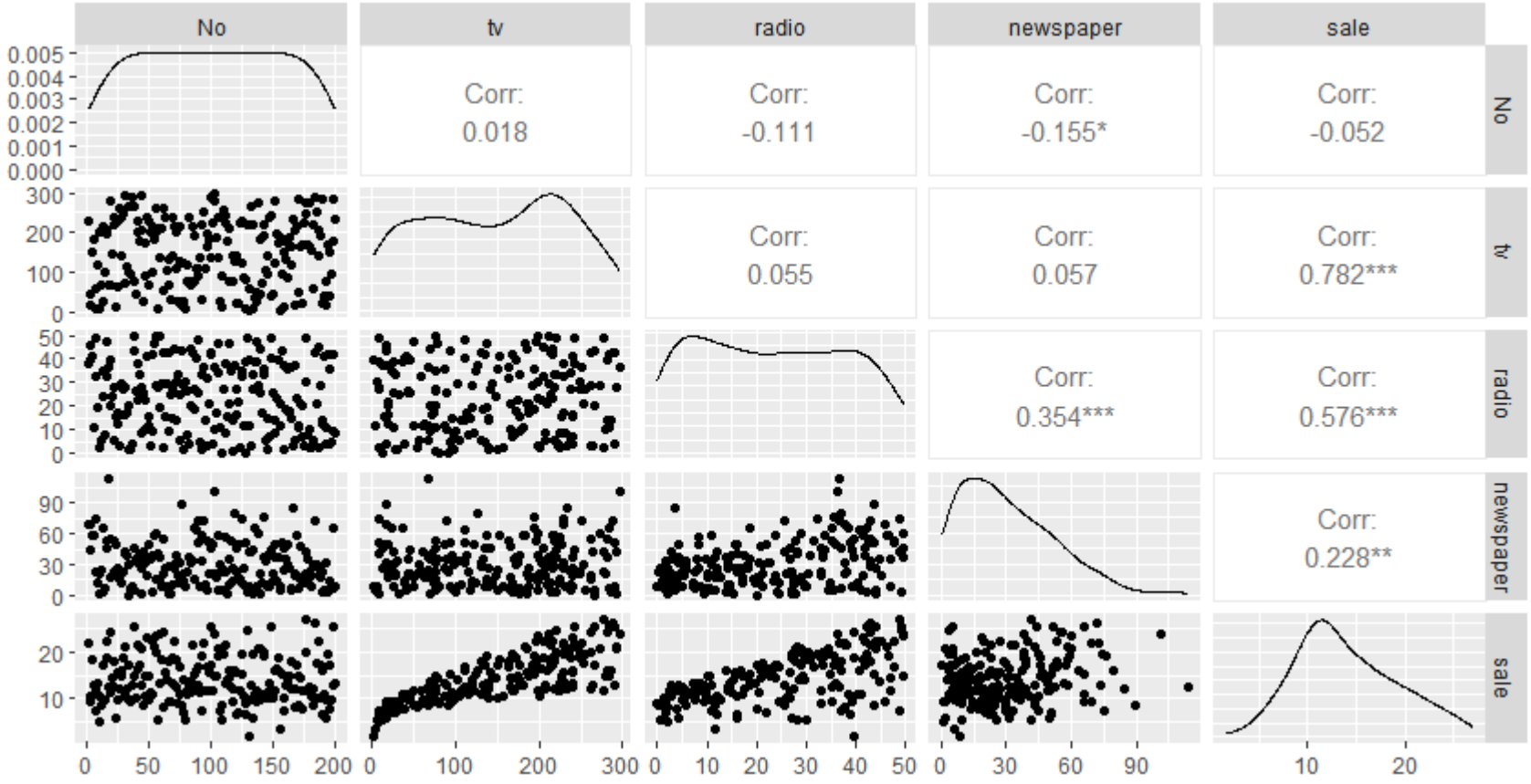
We use VIFs to identify correlations between variables and determine the strength of the relationships.

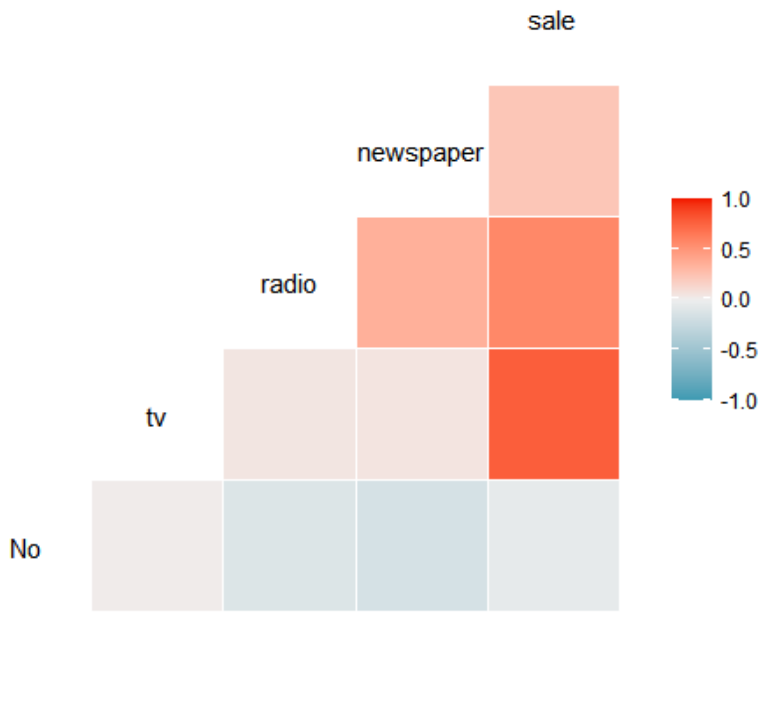
Attention: *When the high VIFs are caused by the inclusion of powers or products of other variables, you can Safely Ignore Multicollinearity*

```
library(mctest)
>>imcdiag(model,
method="VIF")

library(car)
>>VIF(model)
```

An example from Advertising data





An example from Advertising data

```
> firstordermodel<-lm(sale~tv+radio, data=Advertising)
>
> #install.packages("mctest")
> library(mctest)
> imcdiag(firstordermodel, method="VIF")

Call:
imcdiag(mod = firstordermodel, method = "VIF")


 VIF Multicollinearity Diagnostics

       VIF detection
tv    1.003         0
radio 1.003         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test
```

```
> library(car)
> vif(firstordermodel)
       tv     radio
1.003013 1.003013
```

From the output, you can see that the $VIF_{TV} = VIF_{Radio} = 1.003013$, which suggests that there is no correlation between these predictors.

Solutions for Multicollinearity

- **The first solution** is to drop one of the problematic variables from the regression model. This can usually be done without much compromise to the regression model, since the presence of multicollinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables.
- **The second solution** is to combine the collinear variables together into a single predictor. For instance, we might take the average of standardized versions of Limit and Rating in order to create a new variable that measures credit worthiness.

In class Practice Problem 19

Use the CREDIT.CSV data.

From the credit card example, check whether **multicollinearity** exists or not.

Lecture_7.R

In class Practice Problem 19 Answers

```
> first_order=lm(Balance ~ Income + Rating + Age + Limit + Cards + factor(Student), data=credit)
> imcdiag(first_order, method="VIF")

Call:
imcdiag(mod = first_order, method = "VIF")


 VIF Multicollinearity Diagnostics

                       VIF detection
Income               2.7769         0
Rating             230.8695         1
Age                  1.0397         0
Limit              229.2385         1
Cards                1.4390         0
factor(Student)Yes   1.0091         0

Multicollinearity may be due to Rating Limit regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

====================================
```

```
> firstorder1 <- lm(Balance ~ Income + Rating + Age + Cards + factor(Student), data=credit)
> imcdiag(firstorder1, method="VIF")

Call:
imcdiag(mod = firstorder1, method = "VIF")


 VIF Multicollinearity Diagnostics

                     VIF detection
Income             2.7760         0
Rating             2.7226         0
Age                1.0396         0
Cards              1.0161         0
factor(Student)Yes 1.0029         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test
```

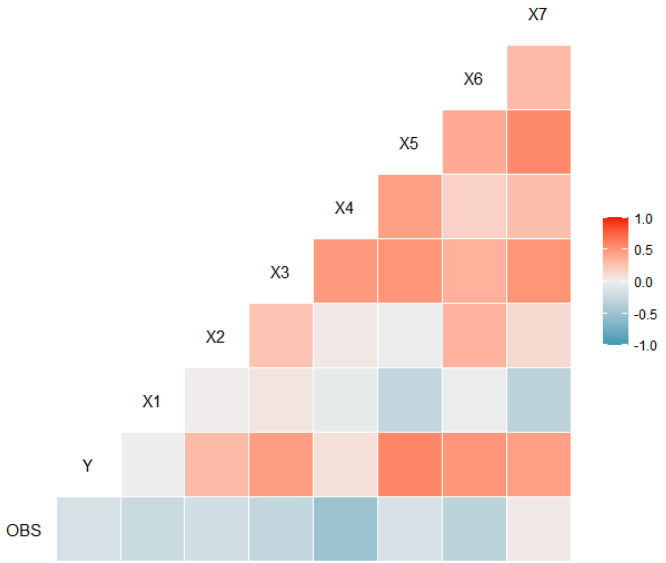
Lecture_7.R

In class Practice Problem 20

From the clerical staff work hours, check whether **multicollinearity** exists in the optimal models or not. If you detect a trend, how would you like to transform the predictors in the model?

Lecture_7.R

In class Practice Problem 20 Answers



optimal_model_approach2=lm(Y ~ X2+X4+X5+I(X2^2), data=workhours)

```
> imcdiag(optimal_model_approach2,method="VIF")

Call:
imcdiag(mod = optimal_model_approach2, method = "VIF")


 VIF Multicollinearity Diagnostics

           VIF detection
X2      14.4682         1
X4       1.2493         0
X5       1.2478         0
I(X2^2) 14.4871         1

Multicollinearity may be due to X2 I(X2^2) regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

==================================
```

optimal_model_approach3 <- lm(Y~(X1+X2+X3+X4+X5+X6+X1:X6+X2:X6), data=workhours)

```
> imcdiag(optimal_model_approach3,method="VIF")

Call:
imcdiag(mod = optimal_model_approach3, method = "VIF")


 VIF Multicollinearity Diagnostics

          VIF detection
X1     32.1883         1
X2     28.3857         1
X3      2.0138         0
X4      1.4124         0
X5      2.1160         0
X6     20.4776         1
X1:X6 41.9543         1
X2:X6 49.1724         1

Multicollinearity may be due to X1 X2 X6 X1:X6 X2:X6 regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Lecture_7.R

6. Outlier (The effect on individual cases)

- Outlier: An outlying case is defined as a particular observation (Y,$X1$,$X2$,...Xp) that differs from the majority of the cases in the data set. Potentially caused by a different mechanism (something might wrong)
- Possible causes of outliers
  - Error in recording data (e.g. 28 to 82)
  - Measurement process/tool problem
  - Failure of the experimental process
- For non-robust statistics, outlier can ruin the analysis. (e.g. mean)

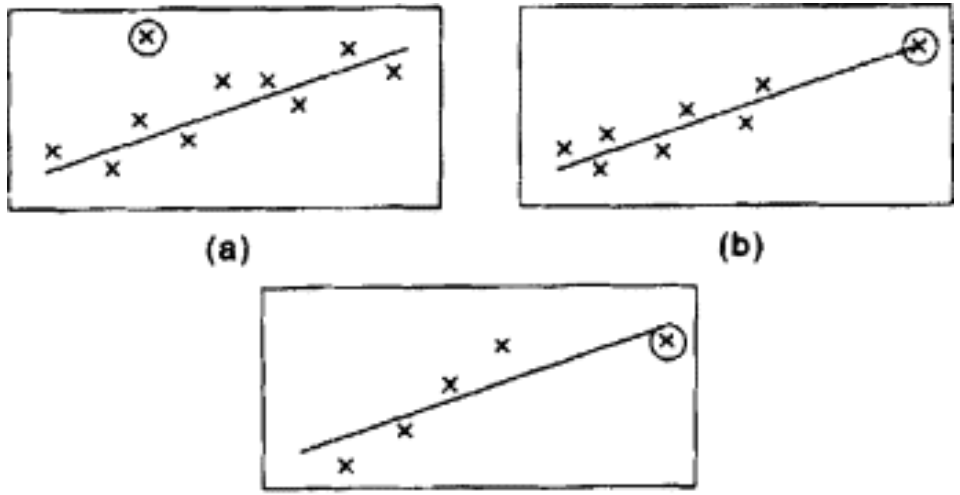
Several key concepts

Outliers: a point that falls far from the other data points.

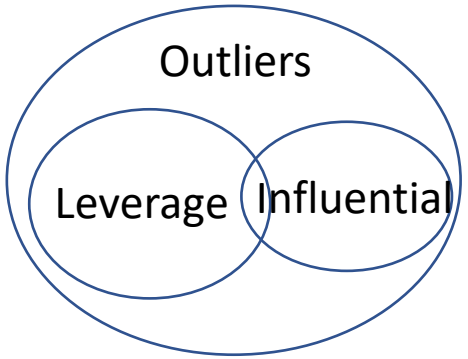
Leverage: leverage is a measure of how far away the independent variable values of an observation are from those of the other observations. High-leverage points, if any, are outliers with respect to the independent variables.

Influential: If the coefficients of predictors change dramatically when a point is removed, that point is influential. How influential a point is, is a combination of how much leverage it has and how extreme it is in the y-direction.

Good leverage

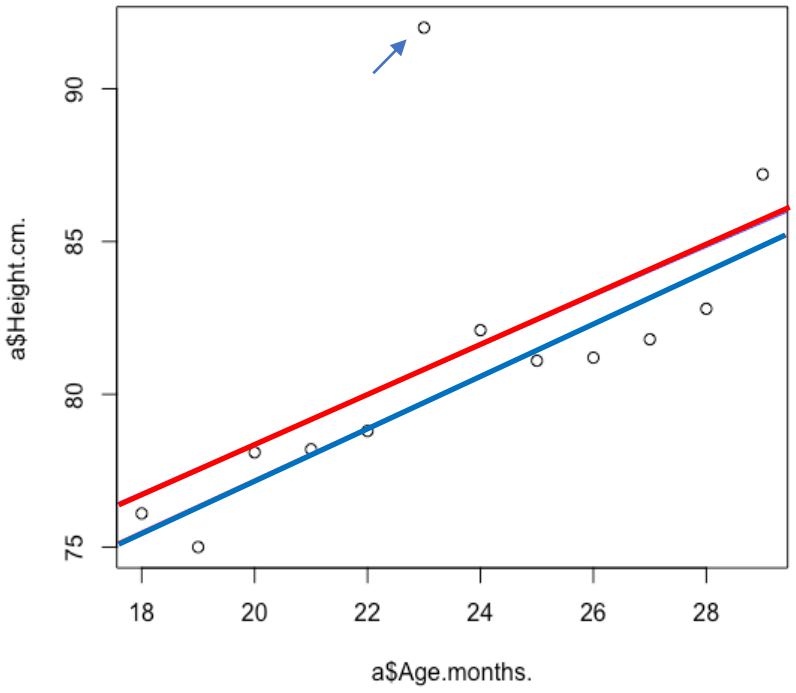


Bad leverage

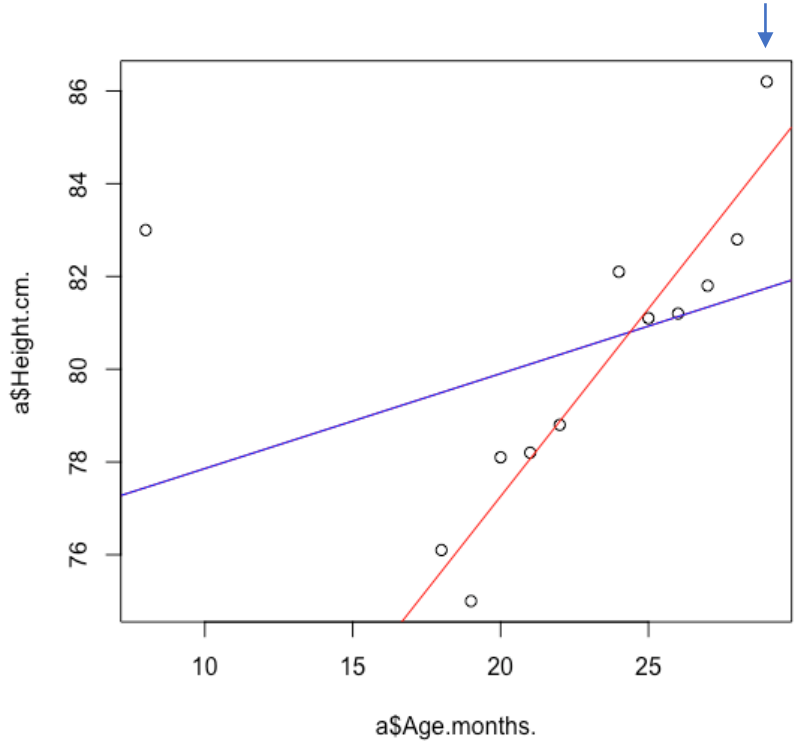


Leverage & influence

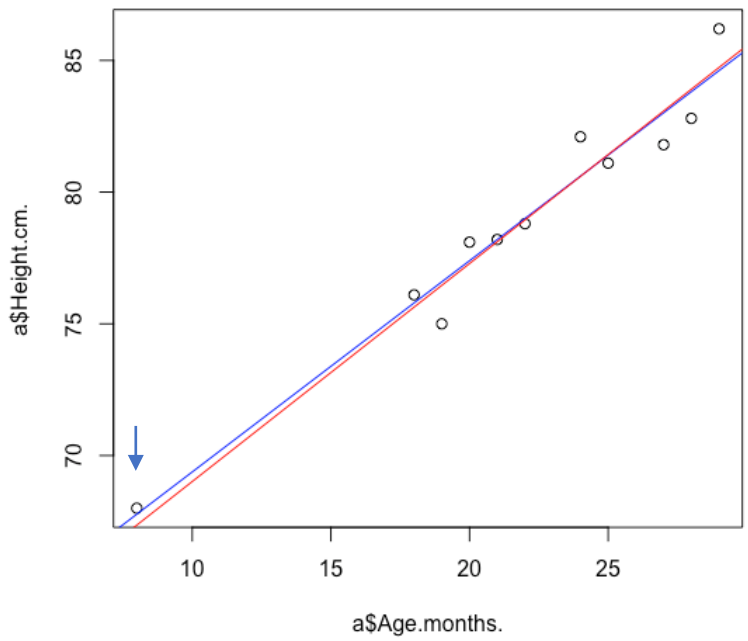
Low leverage, some Influential



High leverage, high Influential



High leverage, low Influential



- What to do with influential point?

  Make sure the observation is correctly recorded. Can we get more observations near that value of X? Do the conclusions change?

Numeric measures for leverage

Leverage values: $h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}}$

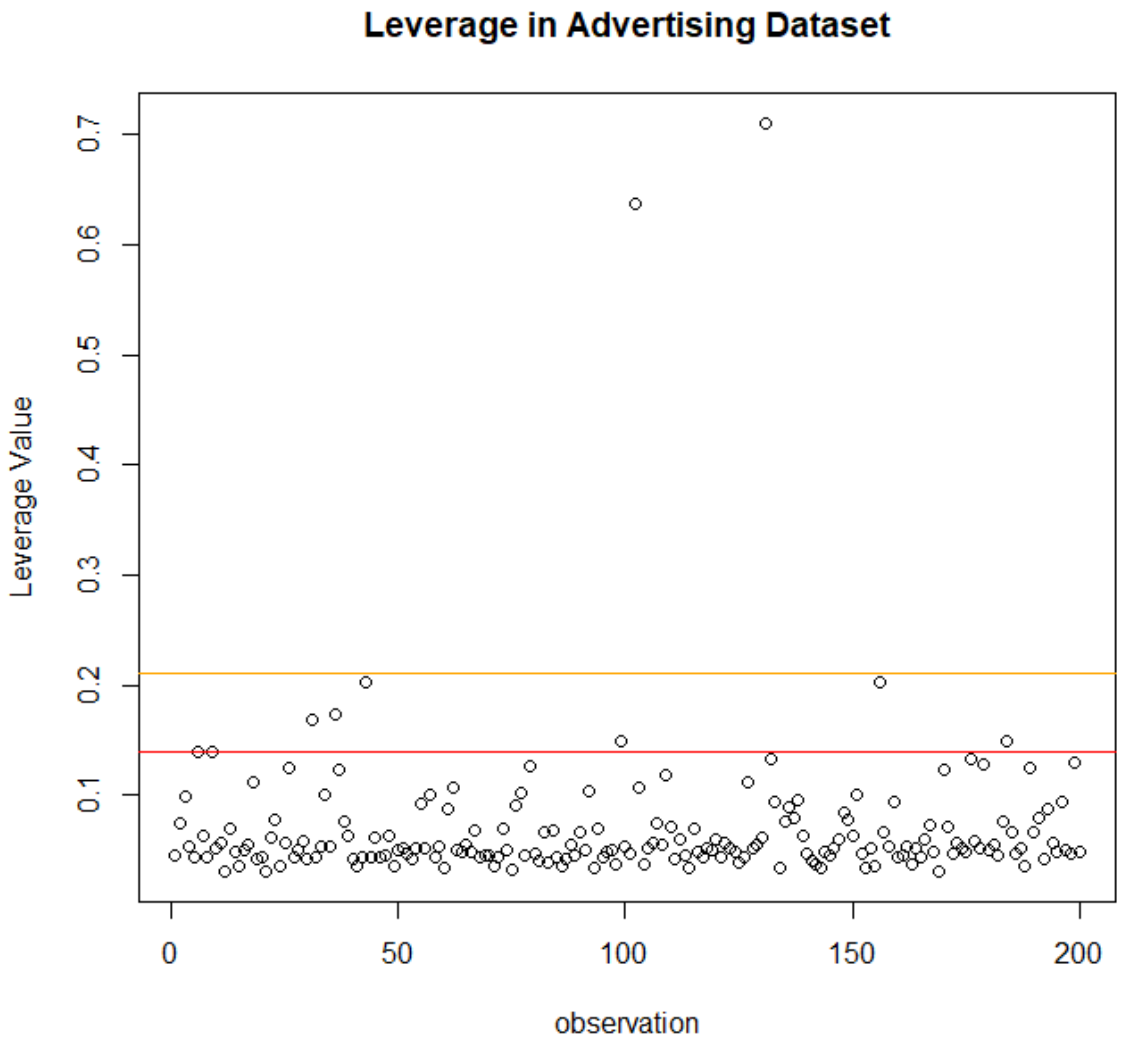
- Leverage values for multiple regression models are extremely difficult to calculate without the aid of a computer
- A good rule of thumb to identify an observation y_i as influential if its leverage value h_i is

$$h_i > \frac{2p}{n}$$

p = the number of predictors

n= the number of the sample size

Numeric measures for leverage




```
lev=hatvalues(morepower)
p = length(coef(morepower))
n = nrow(Advertising)
outlier2p = lev[lev>(2*p/n)]
outlier3p = lev[lev>(3*p/n)]

plot(rownames(Advertising),lev, main =
"Leverage in Advertising Dataset",
xlab="observation", ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
abline(h = 3 *p/n, lty = 1)
```

Effects of influential points

- Not all outliers are influential in linear regression analysis (whatever outliers mean). Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn't be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don't really matter; they are not influential.
- On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don't get along with the trend in the majority of the cases.

Numeric measure for influence

- A measure of the overall influence an outlying observation has on the estimated coefficients was proposed by R. D. Cook (1979). Cook's Distance:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)} \right)^2}{ps^2}$$

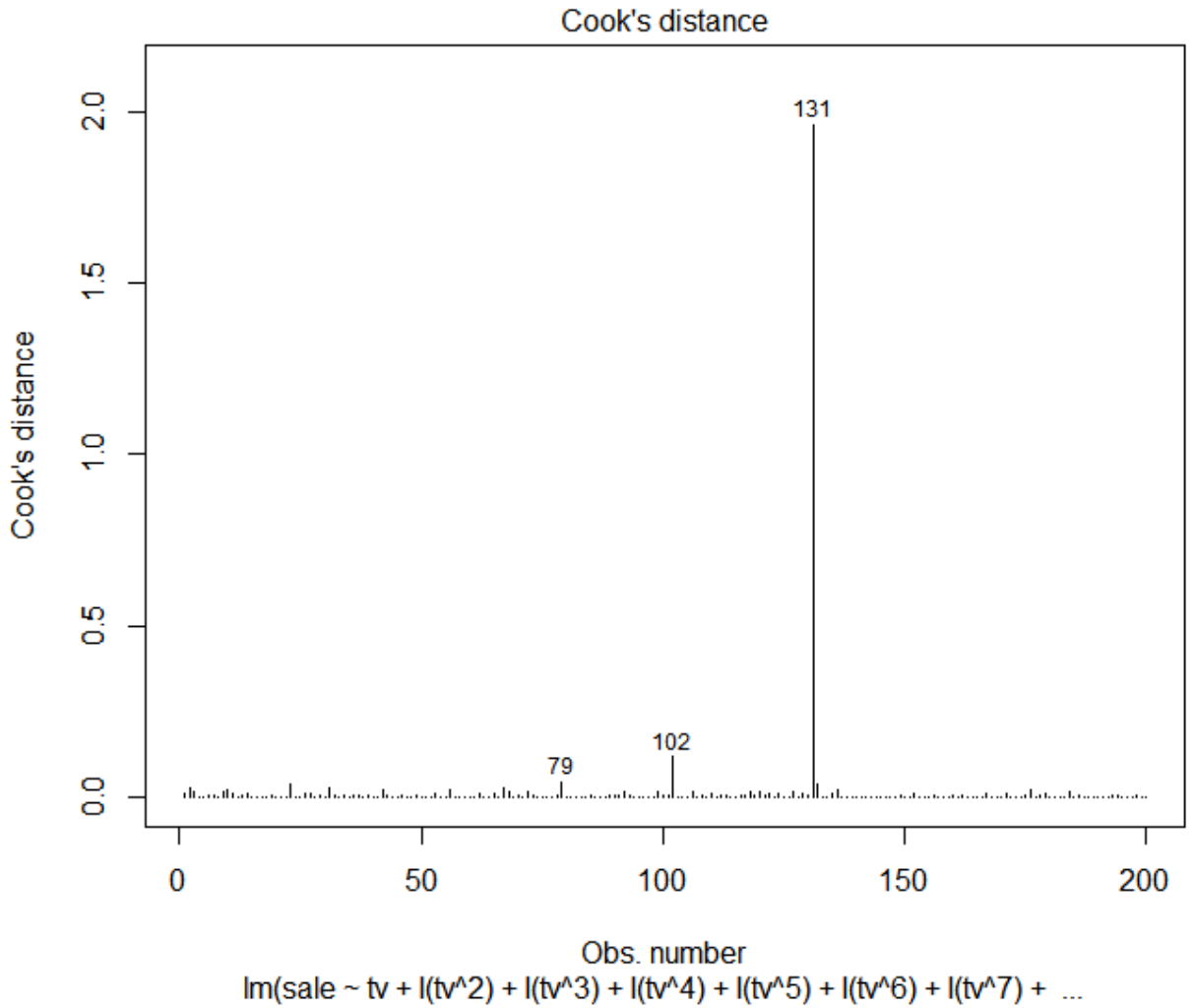
where $\hat{y}_{j(i)}$ is the fitted response value obtained when excluding i,

S is the mean square error of the regression model. p is the number of the parameters in the model

Cook's Distance

- The Cook's distance D_i measures the effect of deleting a given observation.
- A large value of D_i indicates that the observed Y_i value has strong influence on the estimated coefficients
- A general rule of thumb is that observations with a Cook's D of more than 3 times the mean, μ, is a possible outlier.
- An alternative interpretation is to investigate any point over $4/n$, where n is the number of observations
- Any point with large D_i. The consensus seems to be that a D_i value of more that 1 indicates an influential value, but you may want to look at values above 0.5.

Identification of influential points for Advertising data

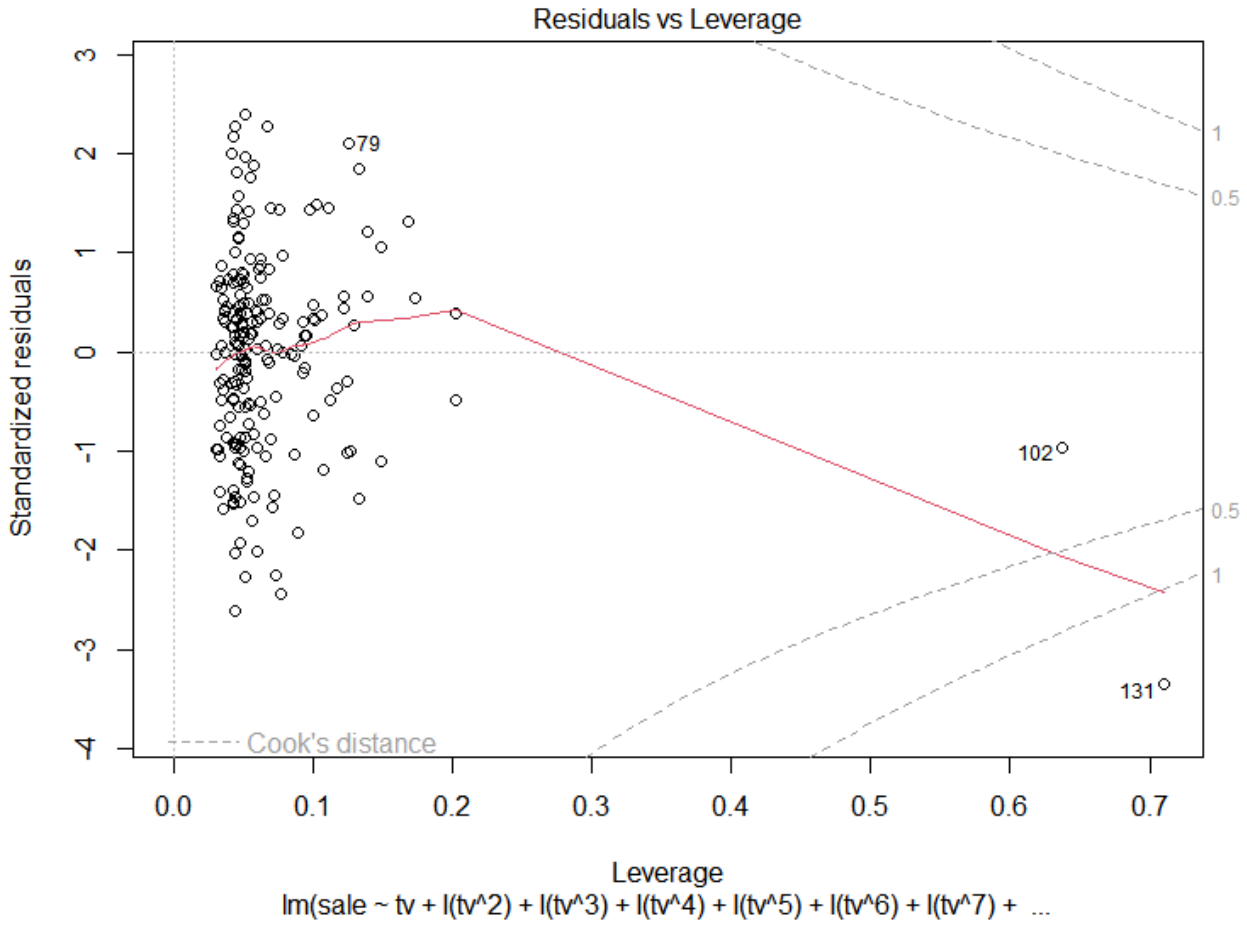


plot(highorder11, which=4)

Which()

- plot(model, which=n)
- Which selects which plot (n) to be displayed:
  1. A plot of residuals against fitted values
  2. A normal Q-Q plot
  3. A Scale-Location plot of sqrt(| residuals |) against fitted values
  4. A plot of Cook's distances versus row labels
  5. A plot of residuals against leverages
  6. A plot of Cook's distances against leverage/(1-leverage)

Identification of high influential and high leverage points for Advertising data



Residual VS leverage plot helps the identification of influential points

```
Advertising=read.table("Advertising.txt", header = TRUE, sep ="\t" )

Highorder11=lm(sale~tv+I(tv^2)+I(tv^3)+I(tv^4)+I(tv^5)+I(tv^6)+I(tv^7)+I(tv^8)+I(tv^9)+I(tv^10)+I(tv^11)+radio+tv*radio, data=Advertising)

plot(highorder11,which=5)
```

What to do with outlier

- For a good understanding of the regression model, if we have some outliers or influential points, we may want to
- See what happens when we exclude these from the model as an outlier has occurred due to an error in data collection or recoding, then the solution is to simply remove the observation.
- Investigate these cases separately as it may happen that we mistyped.

In class Practice Problem 21

From the clerical staff work hours, using residual plots to conduct a residual analysis of the data. Check any potential outliers.

optimal_model_approach2=lm(Y ~ X2+X4+X5+I(X2^2), data=workhours)

optimal_model_approach3 <- lm(Y~(X1+X2+X3+X4+X5+X6+X1:X6+X2:X6), data=workhours)

Lecture_7.R

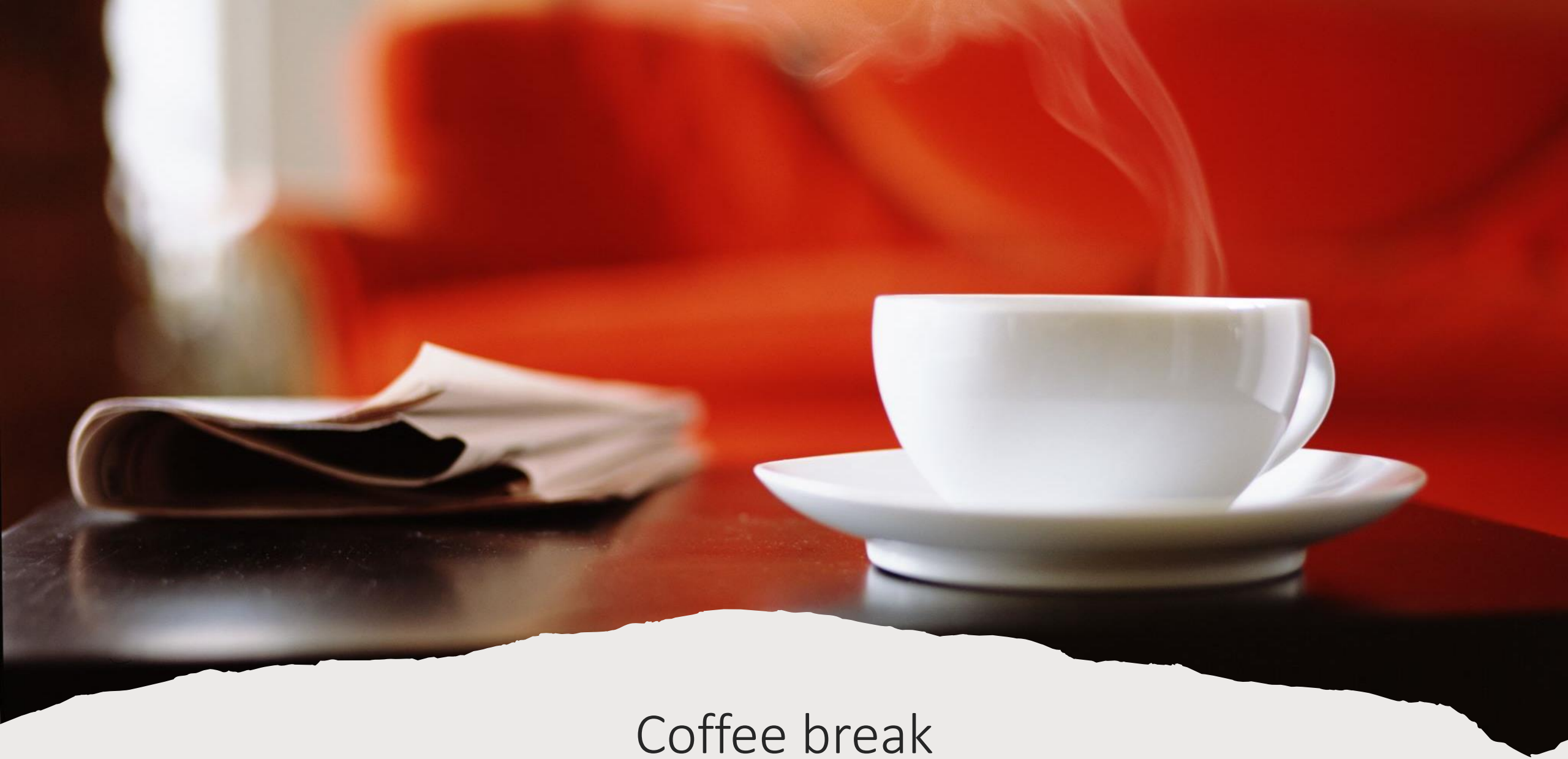
In class Practice Problem 22

Use the CREDIT.CSV data.

Check assumptions for the model below to predict executive salary (Y)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_1 + \beta_6 X_5 + \beta_7 X_3 * X_4 + \epsilon$$

Lecture_7.R



Coffee break

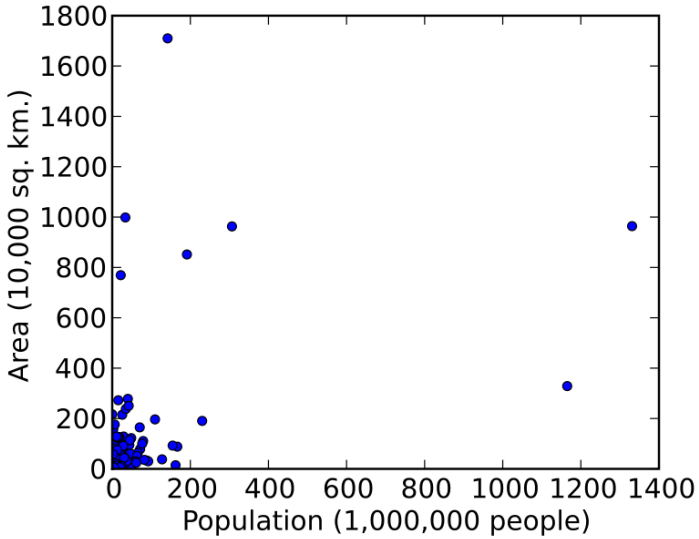
Transformation

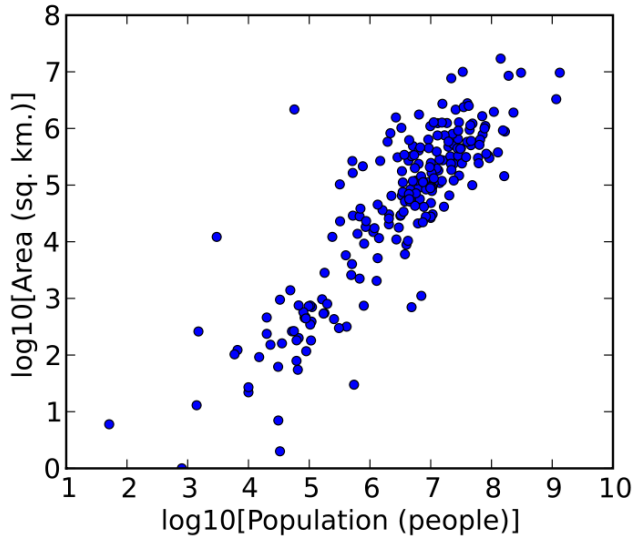
What is data transformation

- Data transformation is the process of changing the format, structure, or value of data
- Transformation is used to transform nonnormal data to normal data
- Transformation is done by using many functions such as square root, logarithm, power, reciprocal or arcsine, etc.

Motivation of transformation

- One of the most common assumptions for statistical analyses is that of normality (normality of error terms)
- To make interpterion of the data easier
- To make appearance of the graphs easier to understand
- A scatterplot in which the areas of the sovereign states and dependent territories in the world are plotted on the vertical axis against their populations on the horizontal axis. The upper plot uses raw data. In the lower plot, both the area and population data have been transformed using the logarithm function.





https://en.wikipedia.org/wiki/Data_transformation_%28statistics%29

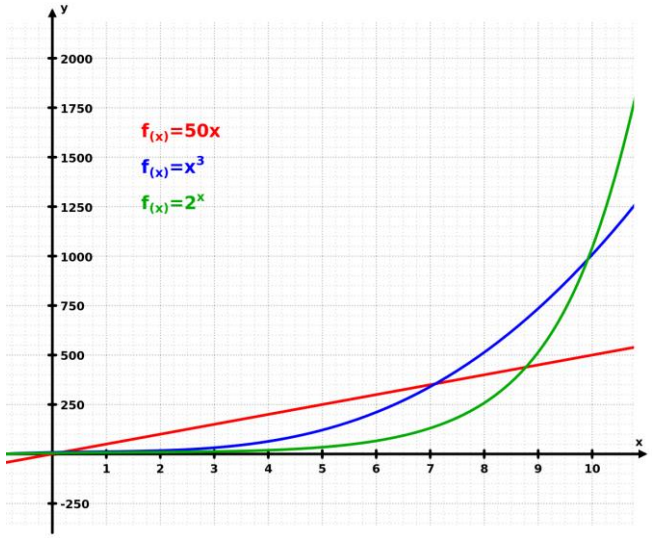
Transformations

- Logarithm transformation: If y goes exponentially or as a power of x, transform by taking the log
- Square root transformation: If data exhibiting a Poisson distribution (variance =mean), transform by taking the square root.
- Box-Cox transformation: a family of power transformation.

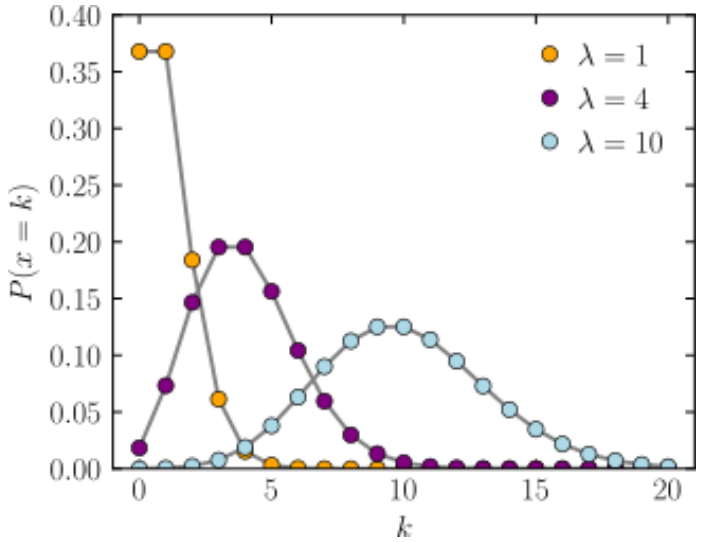
$$Y^{(\lambda)} = Y^{\lambda}$$

  - where λ is a parameter to be determined using the data, vairies from -5 to 5

Exponential OR Power of x



Poisson distribution



Box-Cox Transformation

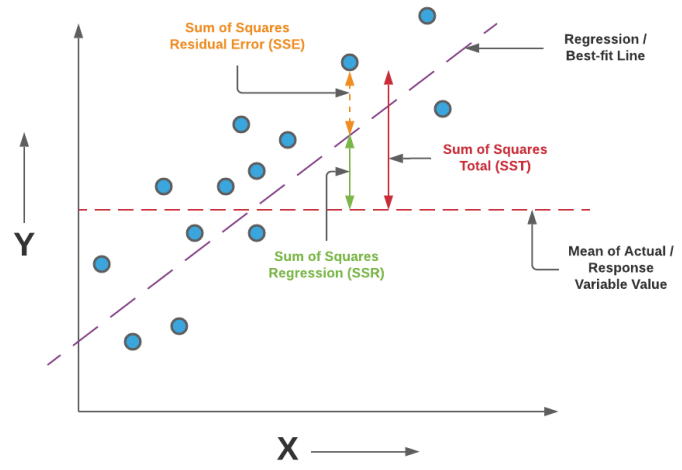
$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, \lambda \neq 0 \\ log_e Y, \lambda = 0 \end{cases}$$

$$Y_i^{(\lambda)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon_i$$

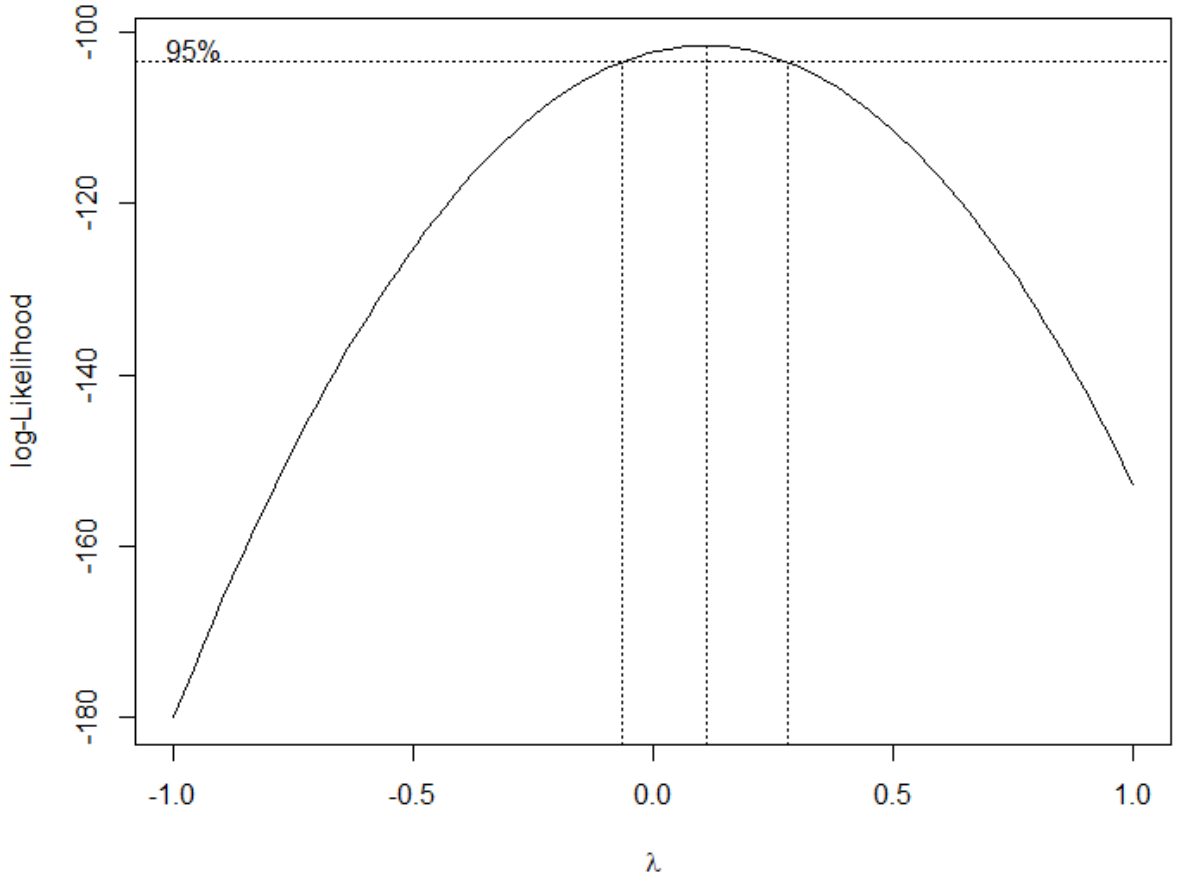
$$\lambda = 2, Y' = Y^2$$
$$\lambda = 0.5, Y' = \sqrt{Y}$$
$$\lambda = 0, Y' = log_e Y \text{ (by definition)}$$
$$\lambda = -0.5, Y' = 1/\sqrt{Y}$$
$$\lambda = -1.0, Y' = 1/Y$$

Note! the transformation for 0 is $\log_e(Y)$.

- Transformations for Nonnormality and Heteroscedasticity
- Strictly **positive response variables**
- Use maximum likelihood estimation (MLE) to find the best fit λ and betas
- Or repeat the ordinary least square (OLS) for different value of and find the one with minimum sum of squares of residual errors (SSE).
- Round the best fit to the nearest multiple of 0.5, then repeat the regression



Best lambda for Box-Cox Transformation




```
library(MASS)

bc=boxcox(reg,lambda=seq(-1,1))

bestlambda=bc$x[which(bc$y==
max(bc$y))]

bestlambda
```

Example 1 for Box-Cox transformation

A practical question for the head of a university could be how study fees (stfees) raise the universities net assets (nassets).A simple linear regression could help to explain the relation between these two variables.

Hints: Use library(Ecdat) to load University dataset

Example 1 for Box-Cox transformation

Check assumptions	Results
Linearity	Failed
Independence	Pass
Equal variance (no hetero)	P=0.001 Failed
Normality	P=6e-8 Failed
Multicollinearity	NA
Outliers	One point

Check assumptions (after Box-Cox transformation)	Results
Linearity	Looks OK
Independence	Pass
Equal variance (no hetero)	P=0.796 Pass
Normality	P=0.013 Pass
Multicollinearity	NA
Outliers	None

Example 2 for Box-Cox transformation

Now we will use **the gala dataset** as an example of using the Box-Cox method to justify a transformation for Multiple Linear Regression. There are 30 Galapagos islands and 7 variables in the dataset. The relationship between the number of plant species and several geographic variables is of interest. We fit an additive multiple regression model with Species as the response and most of the other variables as predictors. *The dataset gala* contains the following variables:

- Species:the number of plant species found on the island
- Endemics:the number of endemic species
- Area:the area of the island (*km*2)
- Elevation:the highest elevation of the island (m)
- Nearest:the distance from the nearest island (km)
- Scruz:the distance from Santa Cruz island (km)
- Adjacent:the area of the adjacent island (square km)

Hints: use library(faraway) to load the dataset

Example 2 for Box-Cox transformation

Check assumptions	Results
Linearity	?
Independence	Pass
Equal variance (no hetero)	P=0.02146 Failed
Normality	P=0.005706 Failed
Multicollinearity	Pass
Outliers	OK

Check assumptions (after Box-Cox transformation)	Results
Linearity	?
Independence	Pass
Equal variance (no hetero)	P=0.1836 Pass
Normality	P=0.1724 Pass
Multicollinearity	Pass
Outliers	Remove some data points

Johnson Transformation

- The Johnson Transformation is a mathematical transformation used to create new variables from existing variables. It can be used to linearize nonlinear relationships and to create normally distributed variables from non-normal ones.
- The Johnson Transformation relies on the fact that any function of a normal random variable is also normal. Suppose we have a random variable X that is not normally distributed. We can create a new random variable Y=g(X) where g() is a function that transforms X into a new random variable Y that is normally distributed. If we can find such a function g(), then we can use Y in place of X in any statistical analysis that assumes normality.
- There are several cases where the Johnson Transformation can be used, but the most common case is when X is skewed to the left or right. In these cases, the transformation can be used to create normally distributed variables from skew variables.
- The advantage of using the Johnson Transformation
  - it preserves many of the features of the original skewed distribution.
  - it's not subject to positive data as it is available for any data values including negative values.
  - It's very powerful for determining an appropriate distribution and can by applied to cases which Box-Cox transformation does not fit.

Four types of Johnson Transforms

- Four types: linear, cumulative probability, uniform, logit.
- The linear transformation is the most common and simplest form of the Johnson Transformation. It transforms data by adding or subtracting a constant and then multiplying or dividing by another constant. This transformation creates a new variable that is linearly related to the original variable.
- Cumulative probability transformations are used when you want to preserve information about how likely it is for values to fall above or below certain thresholds. These transformations are especially useful for extreme value analysis.
- Uniform transformations are used when you want to create variables that are uniformly distributed between 0 and 1. These transformed variables can then be used in Monte Carlo simulations.
- Logit transformations are used when you want to create dummy variables from categorical data. Dummy variables are commonly used in regression analysis.

R function to conduct transformation

yeo.johnson {VGAM}

Yeo-Johnson Transformation

Description

Computes the Yeo-Johnson transformation, which is a normalizing transformation.

Usage

```
yeo.johnson(y, lambda, derivative = 0,
            epsilon = sqrt(.Machine$double.eps), inverse = FALSE)
```

Arguments

`y`
Numeric, a vector or matrix.

transfo {cellWise} R Documentation

Robustly fit the Box-Cox or Yeo-Johnson transformation

Description

This function uses reweighted maximum likelihood to robustly fit the Box-Cox or Yeo-Johnson transformation to each variable in a dataset. Note that this function first calls `checkDataSet` to ensure that the variables to be transformed are not too discrete.

Usage

```
transfo(X, type = "YJ", robust = TRUE,
        standardize = TRUE,
                    prestandardize = NULL,
        quant = 0.99, nbsteps = 2, checkPars = list())
```

Arguments

`X`
A data matrix of dimensions n x d. Its columns are the variables to be transformed.

`type`
The type of transformation to be fit. Should be one of:

- `"BC"`: Box-Cox power transformation. Only works for strictly positive variables. If this type is given but a variable is not strictly positive, the function stops with a message about that variable.
- `"YJ"` Yeo-Johnson power transformation. The data may have positive as well as negative values.

When to use Data Transformation

- 1. When the data is non-normal
- One reason to use data transformation is when the data is non-normal. Non-normal data is data that does not follow a normal distribution, which is a symmetrical bell-shaped curve. Many statistical tests assume that the data is normal, so transforming non-normal data can help to make the results of these tests more accurate.
- 2. When the data is skewed
- Another reason to use data transformation is when the data is skewed. Skewed data is data that is not evenly distributed, with most of the values clustered around one side of the distribution. Skewed data can often be transformed into a more normal distribution, which can be helpful for statistical testing.
- 3. When the data has outliers
- Outliers are extreme values that are significantly different from the rest of the data. When outliers are present, they can often have a significant impact on the results of statistical tests. Therefore, it can be helpful to transform the data in order to reduce the impact of outliers on the results.

When to use Data Transformation

- 4. When you want to compare two or more groups of data
- Another reason to use transformation is when you want to compare two or more groups of data. Some statistical tests require that the groups being compared have equal variance, which is a measure of how spread out the values are within a group. Transforming the data can often help to equalize the variance between groups, which can make comparisons more accurate.
- 5. When you want to stabilize variance over time
- If you have data that was collected over time, you may want to use transformation in order to stabilize variance over time. Variance stabilization helps to ensure that any changes in the mean are not due simply to changes in variance over time, which can make interpretation of results more difficult

Summary of model diagnostics

	Assumptions	Descriptions	Measurements	Potential actions
1	Linearity Assumption	Relationships between all predictors and the response are linear	1. residuals – fitted y values plot	Add high order terms or transform data
2	Independence Assumption	Independence of observations	1. Residuals vs predictors	Use other models than MLR
3	Equal Variance Assumption	Error term has equal variance given any values of independent variables	1. Scale-Location plot. [square root of standardized residuals – fitted y values plot] 2. Breusch-Pagan test (H0: no hetero)	Add high order terms or transform data
4	Normality Assumption	Error term is normally distributed	1. Histogram 2. Q-Q plot 3. Kolmogorov-Smirnov test 4. Shapiro-Wilk test	Add high order terms or transform data
5	Multicollinearity	No perfect collinearity and non-zero variance of independent variables	1. Variance Inflation Factors (VIF)	- Drop one of the problematic variables - Combine colinear predictors to form a new predictor
6	Outlier	Far away from other data points	1. Leverage values 2. Cook's distance	- Exclude them to see the effects on regression models - Check data points

Thank you

- Questions OR Comments?

- Slack channel: section2-course-documents
- Email: qing.li2@uclagary.ca