MACHINE LEARNING USING R

Class 8

Regression

- Supervised Learning method
  - Can be used to build Prediction Models
  - Can be used for Hypothesis Testing

- Regression is the relationship between
  - One numeric Dependent Variable or quantitative response, the value to be predicted

    and
  - one or many numeric Independent Variables or predictors

SIMPLE LINEAR REGRESSION

Simple Linear Regression

- Assumption: Dependent variable is continuous
- Value of quantitative Y is predicted on the basis of single predictor variable X.

$$Y \approx \alpha + \beta X$$

- α (the intercept) is the value of Y when X = 0
- β (the slope) is the rise in line with for each increase in X
- Together α and β are called model coefficients or parameters

Simple Linear Regression

- Using the training data we estimate the values for $\hat{\alpha}$ and $\hat{\beta}$ and predict the future value of $\hat{y}$

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

where $X = x$

- Hat symbol, ^ , is used to estimated value for an unknown coefficient or parameter, or to donate predicted value for the response.

ESTIMATING THE COEFFICIENTS

Ordinary least square estimation

- Assume that the data has n datapoints.
- For i = 1,…., n, we predict the value of y_i based on the value of x_i using the following formula:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

- The difference between the response and the predicted value represents the residual e_i:

$$e_i = y_i - \hat{y}_i$$

Ordinary least square estimation

- Residual Sum of Squares (RSS) is defined as:

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

$$RSS = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2$$

$$RSS = (y_1 - \hat{\alpha} - \hat{\beta}x_1)^2 + (y_2 - \hat{\alpha} - \hat{\beta}x_2)^2 + \ldots + (y_n - \hat{\alpha} - \hat{\beta}x_n)^2$$

- The goal of ordinary least square estimation method is to have the minimum value of RSS

Ordinary least square estimation

- Values of $\hat{\alpha}$ and $\hat{\beta}$, or *least squares coefficient estimates* selected to minimize RSS are given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

- where $\bar{x}$ and $\bar{y}$ are simple means:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Ordinary least square estimation

- Based on the data sample, we can calculate the sample mean, $\hat{\mu}$

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

  - where n = number of observations of the sample
- Averaging multiple sample means will give us an estimate of population mean, *μ*
- Variance and Standard Error tells us how different $\hat{\mu}$ and *μ* are
- Of T total observations, if n independent observations have mean $\hat{\mu}$ and standard deviation, σ, then total variance = nσ^2
- Variance of T/n (or sample mean $\hat{\mu}$) is

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

  - where σ is the standard deviation of y_i from Y

Ordinary least square estimation

- Alternatively:

$$\hat{\beta} = \frac{Cov(x,y)}{Var(x)}$$

Covariance is a measure of the joint variability of two random variables.

Random variables whose covariance is zero are called uncorrelated.

Variance is a special case of the covariance in which the two variables are identical.

ASSESSING MODEL ACCURACY

Residual Square Error

- Average amount the response will deviate from the true regression line
- Measure of the lack of fit of the model to the data

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

- Small RSE indicates that model fits data well
- Large RSE indicates that model does not fit data well

R^2 Statistic

- Since RSE is measured in terms of Y, it is not always clear what value of RSE is good.
- R^2 statistic eliminates this by calculating the proportion of variance and takes the value between 0 and 1.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- TSS is the total sum of squares (or variance) $\sum_{i=1}^{n} (y_i - \overline{y})^2$
- R^2 near 0 indicates poor model
- R^2 near 1 indicates good model

Correlation

- Pearson's correlation coefficient
- Measures linear correlation between x and y
- Value between -1 and 1
    - -1 is total negative correlation
    - 0 is no linear correlation
    - 1 is total positive correlation

$$\rho_{x,y} = Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

where

- Cov is the covariance
- σ_x is the standard deviation of x
- σ_y is the standard deviation of y

MULTIPLE LINEAR REGRESSION

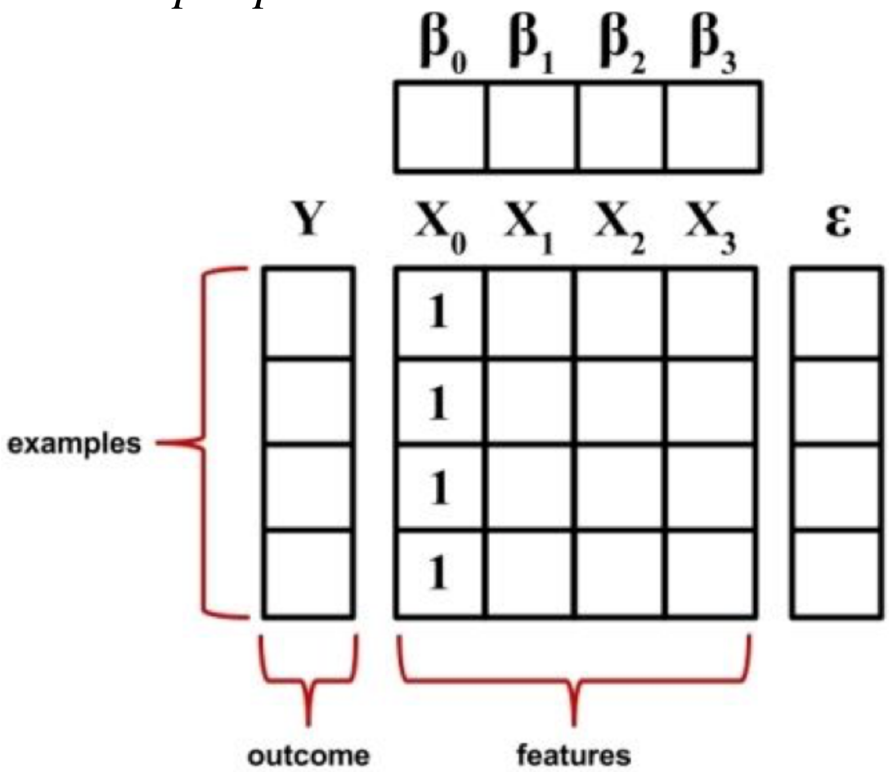
Multiple Linear Regression

- Accommodates multiple predictors by assigning separate slope coefficients

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

$$Y = X\beta + \varepsilon$$

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



Multiple Linear Regression

- Predicted value of y:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

- Residual Sum of Squares (RSS) is defined as:

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

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})^2$$

Multiple Linear Regression: Advantages

- By far the most common approach for modeling numeric data
- Can be adapted to model almost any data
- Provides estimates of the strength and size of the relationships among features and the outcome

Multiple Linear Regression: Disadvantages

- Makes strong assumptions about the data
- The model's form must be specified by the user in advance
- Does not do well with missing data
- Only works with numeric features, so categorical data require extra processing
- Requires some knowledge of statistics to understand the model.

Potential Problems

- Qualitative Data
- Non Linear relationship
- Non-constant variance of error terms
- Outliers
- High-leverage points
- Collinearity

QUALITATIVE PREDICTORS

Predictor with Two Levels

- Create dummy variable that takes two possible values:

$$x_i = \begin{cases} 1 & \text{if } i \text{ th tumor is malignant} \\ 0 & \text{if } i \text{ th tumor is benign} \end{cases}$$

- Use the variable as a predictor:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i \text{ th tumor is malignant} \\ \beta_0 + \varepsilon_i & \text{if } i \text{ th tumor is benign} \end{cases}$$

Predictor with Two Levels

- Create dummy variable that takes two possible values:

$$x_i = \begin{cases} 1 & \text{if } i^{\text{ th}} \text{ tumor is malignant} \\ -1 & \text{if } i^{\text{ th}} \text{ tumor is benign} \end{cases}$$

- Use the variable as a predictor:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i^{\text{ th}} \text{ tumor is malignant} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{if } i^{\text{ th}} \text{ tumor is benign} \end{cases}$$

Predictor with more than Two Levels

- Create dummy variable that takes two possible values:

$$x_{i1} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ flower is setosa} \\ 0 & \text{if } i^{\text{th}} \text{ flower is not setosa} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ flower is versicolor} \\ 0 & \text{if } i^{\text{th}} \text{ flower is not versicolor} \end{cases}$$

- Use the variable as a predictor:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i^{\text{th}} \text{ flower is setosa} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i^{\text{th}} \text{ flower is versicolor} \\ \beta_0 + \varepsilon_i & \text{if } i^{\text{th}} \text{ flower is virginica} \end{cases}$$

Predictor with more than Two Levels

- Number of dummy variable = Number of levels - 1
- Level with no dummy variable is known as the Baseline

NON LINEAR RELATIONSHIP

Non-Linear Relationships

- If the residual plot indicates that there are non-linear associations in the data, use non-linear transformations of the predictors in the regression model.
  - log x
  - $\sqrt{x}$
  - x^2

Polynomial Regression

- Sometimes Predictor variables and Response variables have a non-linear relationship:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$$

$$where \qquad x_2 = x_1^2$$

- The model is still linear since x_1^2 now simply represents x_2

Interaction effects

- Interaction is when two features have a combined effect

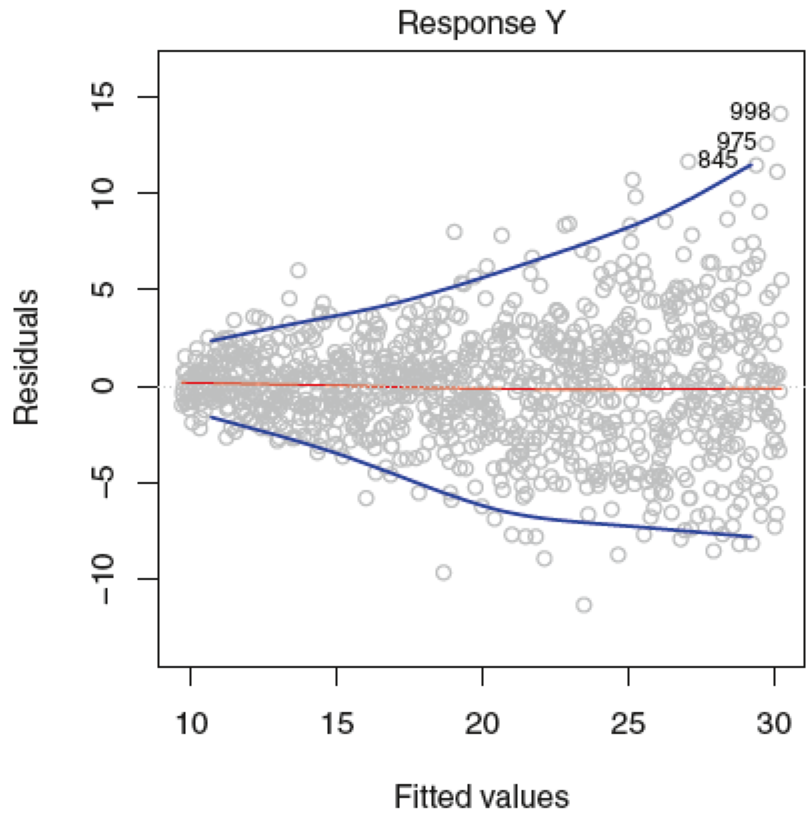
- y~x1*x2
- which translates to y~x1+x2+x1:x2

- Interactions should never be included in a model without also adding each of the interacting variables. If you always create interactions using the * operator, this will not be a problem since R will add the required components for you automatically.

VARIANCE OF ERROR TERMS

Heteroscedasticity

- Linear regression models assume that the error terms have a constant variance, $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
- In real-data, variance of error terms is not constant.
- Heteroscedasticity is the non-constant variance in the errors
- Can be identified by the presence of funnel shape in residual plot
- Solution is to use log Y or $\sqrt{Y}$



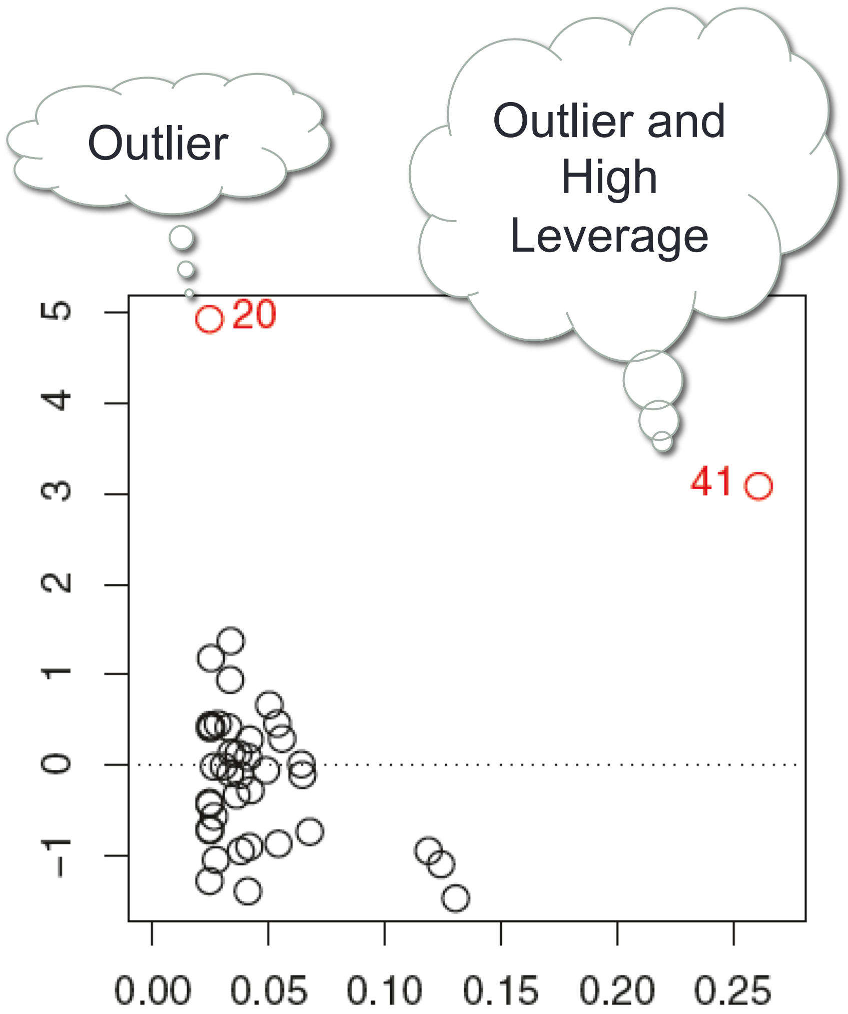
OUTLIERS & HIGH LEVERAGE

Outliers

- Unusual true value of response which is far from the predicted value
- Significantly impact regression estimates
- Should consider if this is due to data collection error or indicates a missing predictor
- Residuals plot can help identify outliers
- One way to programmatically address the issue of Outliers is to use Robust Linear regression with rlm in MASS package.

High Leverage

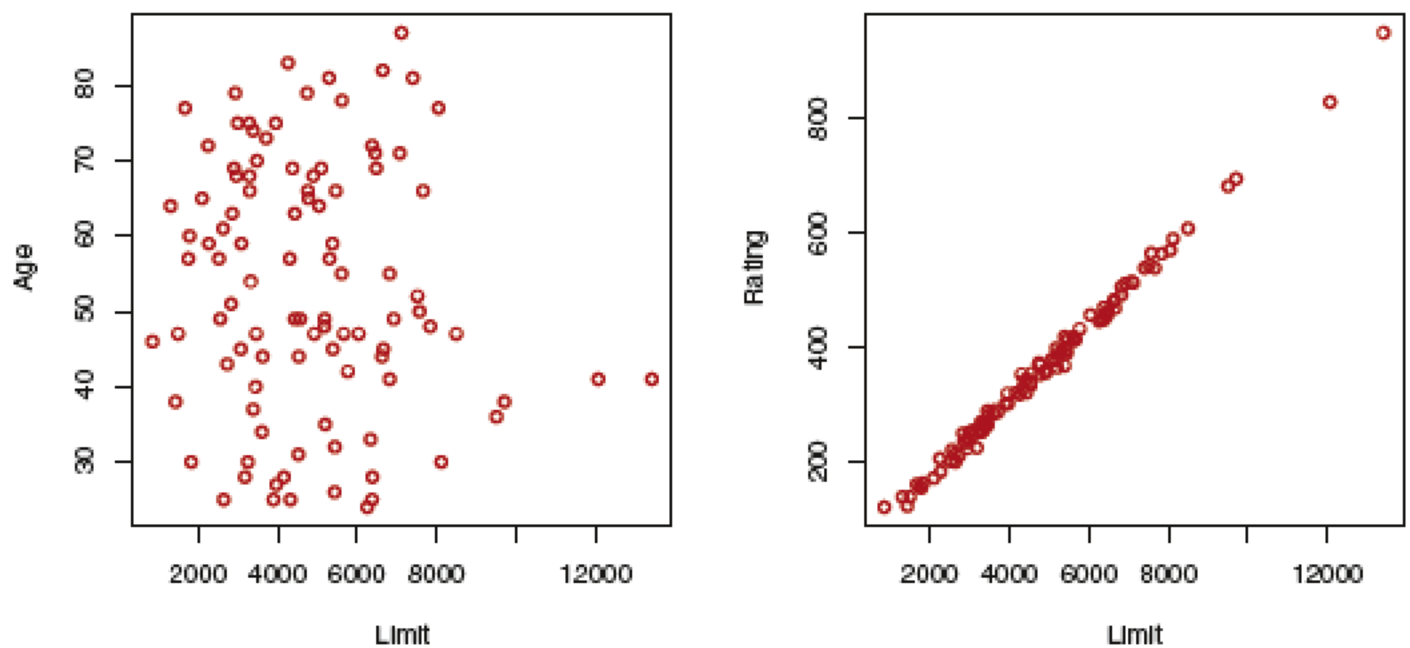
- Unusual value of predictor
- Hard to identify in multiple regression models
- Significantly impact regression estimates
- Can you Leverage Statistics to indicate observations with high leverage



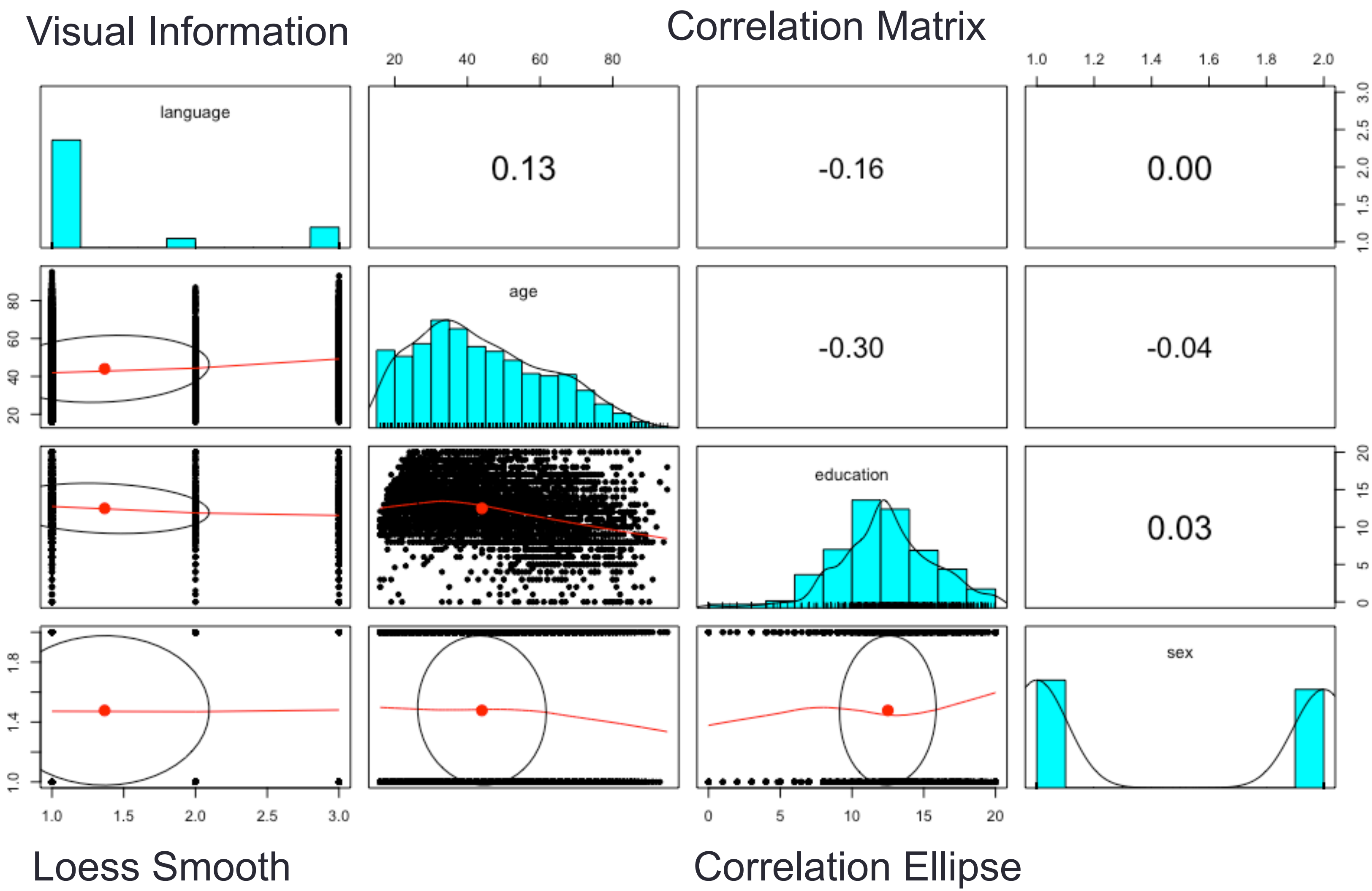
COLLINEARITY

Collinearity

- When two or more predictor variables are closely related to one another.
- When predictor variables are collinear, it is difficult to estimate individual effects.



Visualizing Data



Reading Assignment

- Chapter 6 of Lantz

- Optional Reading:
- http://data.library.virginia.edu/diagnostic-plots/
- https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/

Deep Dive in R

- Simple Linear Regression
  - lm function in stats package

- Robust Linear Regression
  - rlm function in MASS package

- Linear regression on SLID dataset in car package

- Gaussian Model for Generalized Linear Regression
  - glm function in stats package
  - Poisson Model
  - Binomial Model