Linear Regression: Practical Considerations

Introduction to Machine Learning

Matt Magnusson & Marek Petrik

February 7, 2017

Some of the figures in this presentation are taken from ”An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Last Class

1. Simple and multiple linear regression

2. Estimating coefficients (β)

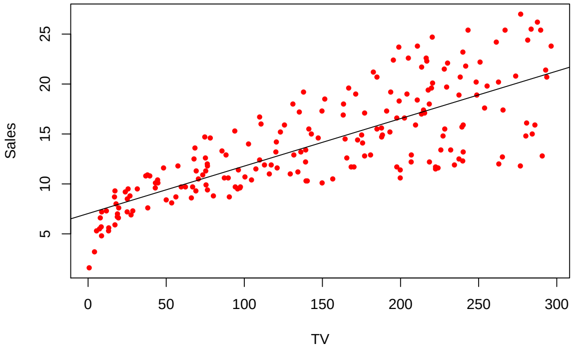
3. R^2 error and correlation coefficient

Simple Linear Regression

- We have only one feature

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

- Example:



$$\text{Sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

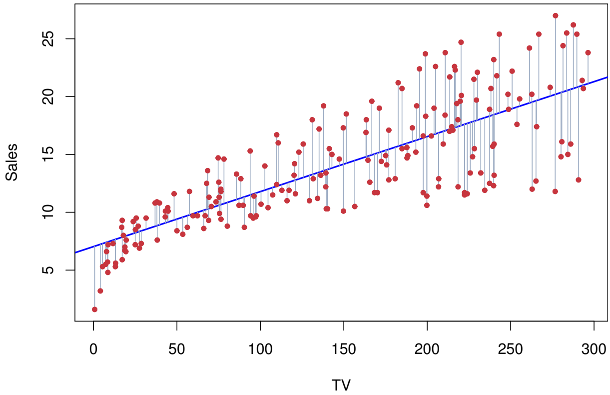
How To Estimate Coefficients

- No line that will have no errors on data x_i
- Prediction:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Errors (y_i are true values):

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



Residual Sum of Squares

- Residual Sum of Squares

$$\text{RSS} = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

- Equivalently:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

R^2 Statistic

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- RSS - residual sum of squares, TSS - total sum of squares
- R^2 measures the goodness of the fit as a proportion
- Proportion of data variance explained by the model
- Extreme values:

  0: Model does not explain data

  1: Model explains data perfectly

Correlation Coefficient

- Measures dependence between two random variables X and Y

$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

- Like R^2 it is between 0,1

  0: Variables are not related

  1: Variables are perfectly related (same)

Correlation Coefficient

- Measures dependence between two random variables X and Y

$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

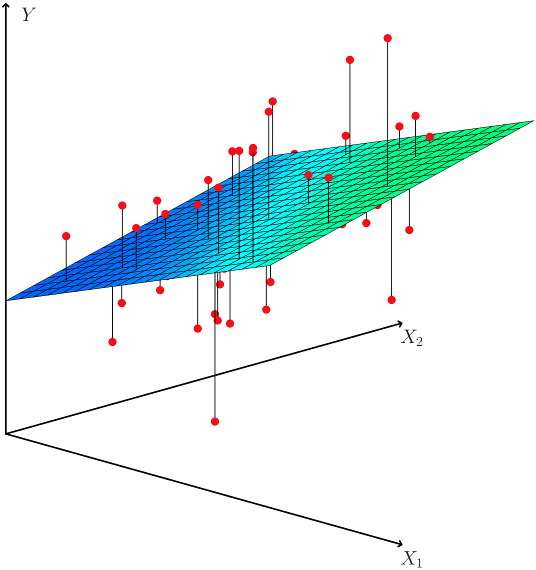
- Like R^2 it is between 0,1

  0: Variables are not related

  1: Variables are perfectly related (same)

- $R^2 = r^2$

Multiple Linear Regression



Estimating Coefficients

- Prediction:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$$

- Errors (y_i are true values):

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

- Residual Sum of Squares

$$\mathrm{RSS} = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

- How to minimize RSS? Linear algebra!

Today: Linear Regression in Practice

1. Inference using linear regression
2. Designing features
3. Possible problems: What can go wrong?
4. Lab!

Multiple Linear Regression

- Usually more than one feature is available

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon$$

- In general:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Inference from Linear Regression

1. Are predictors $X_1, X_2, \ldots, X_p$ really predicting Y?
2. Is only a subset of predictors useful?
3. How well does linear model fit data?
4. What Y should be predict and how accurate is it?

Inference 1

"Are predictors $X_1, X_2, \ldots, X_p$ really predicting Y?"

- Null hypothesis H_0:

$$\text{There is no relationship between } X \text{ and } Y$$
$$\beta_1 = 0$$

- Alternative hypothesis H_1:

$$\text{There is some relationship between } X \text{ and } Y$$
$$\beta_1 \neq 0$$

- Seek to reject hypothesis H_0 with small "probability" (p-value) of making a mistake
- See ISL 3.2.2 on how to compute F-statistic and reject H_0

Inference 2

"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features

Inference 2

"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features
- **RSS always decreases with more features!**

Inference 2

"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features
- **RSS always decreases with more features!**
- Other measures control for number of variables:
  1. Mallows C_p
  2. Akaike information criterion
  3. Bayesian information criterion
  4. Adjusted R^2

Inference 2

"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features
- **RSS always decreases with more features!**
- Other measures control for number of variables:
  1. Mallows C_p
  2. Akaike information criterion
  3. Bayesian information criterion
  4. Adjusted R^2
- Testing all subsets of features is impractical: 2^p options!

Inference 2

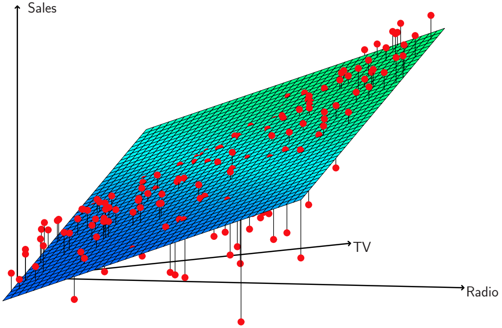
"Is only a subset of predictors useful?"

- Compare prediction accuracy with only a subset of features
- **RSS always decreases with more features!**
- Other measures control for number of variables:
  1. Mallows C_p
  2. Akaike information criterion
  3. Bayesian information criterion
  4. Adjusted R^2
- Testing all subsets of features is impractical: 2^p options!
- More on how to do this later

Inference 3

"How well does linear model fit data?"

- R^2 also always increases with more features (like RSS)
- Is the model linear? Plot it:



- More on this later

Inference 4

"What Y should be predict and how accurate is it?"

- The linear model is used to make predictions:

$$y_{\text{predicted}} = \hat{\beta}_0 + \hat{\beta}_1 \, x_{\text{new}}$$

- Can also predict a confidence interval (based on estimate on ϵ):

Inference 4

"What Y should be predict and how accurate is it?"

- The linear model is used to make predictions:

$$y_{\text{predicted}} = \hat{\beta}_0 + \hat{\beta}_1\, x_{\text{new}}$$

- Can also predict a confidence interval (based on estimate on ϵ):
- **Example**: Spent $\$100\,000$ on TV and $\$20\,000$ on Radio advertising
  - **Confidence interval**: predict $f(X)$ (the average response):

$$f(x) \in [10.985, 11, 528]$$

  - **Prediction interval**: predict $f(X) + \epsilon$ (response + possible noise)

$$f(x) \in [7.930, 14.580]$$

Feature Engineering

What if we have …

1. Qualitative features: (gender, car color, major)
2. Interaction between features: non-additivity
3. Nonlinear relationships

Qualitative Features: 2 Values

- Predict salary as a function of gender
- Feature $\text{gender}_i \in \{\text{male}, \text{female}\}$

Qualitative Features: 2 Values

- Predict salary as a function of gender
- Feature $\text{gender}_i \in \{\text{male}, \text{female}\}$
- Introduce **indicator variable** x_i: (AKA dummy variable, ...)

$$x_i = \begin{cases} 0 & \text{if gender}_i = \text{male} \\ 1 & \text{if gender}_i = \text{female} \end{cases}$$

- Predict salary as:

$$\text{salary} = \beta_0 + \beta_1 \times x_i = \begin{cases} \beta_0 & \text{if gender}_i = \text{male} \\ \beta_0 + \beta_1 & \text{if gender}_i = \text{female} \end{cases}$$

Qualitative Features: 2 Values

- Predict salary as a function of gender
- Feature $\text{gender}_i \in \{\text{male}, \text{female}\}$
- Introduce **indicator variable** x_i: (AKA dummy variable, ...)

$$x_i = \begin{cases} 0 & \text{if } \text{gender}_i = \text{male} \\ 1 & \text{if } \text{gender}_i = \text{female} \end{cases}$$

- Predict salary as:

$$\text{salary} = \beta_0 + \beta_1 \times x_i = \begin{cases} \beta_0 & \text{if } \text{gender}_i = \text{male} \\ \beta_0 + \beta_1 & \text{if } \text{gender}_i = \text{female} \end{cases}$$

- β_1 is the difference between female and male salaries

Qualitative Features: Many Values

- Predict salary as a function of state
- Feature $\text{state}_i \in \{\text{MA}, \text{NH}, \text{ME}\}$
- What about x_i:

$$x_i = \begin{cases} 0 & \text{if } \text{state}_i = \text{MA} \\ 1 & \text{if } \text{state}_i = \text{NH} \\ 2 & \text{if } \text{state}_i = \text{ME} \end{cases}$$

Qualitative Features: Many Values

- Predict salary as a function of state
- Feature $\text{state}_i \in \{\mathrm{MA}, \mathrm{NH}, \mathrm{ME}\}$
- What about x_i:

$$x_i = \begin{cases} 0 & \text{if } \text{state}_i = \mathrm{MA} \\ 1 & \text{if } \text{state}_i = \mathrm{NH} \\ 2 & \text{if } \text{state}_i = \mathrm{ME} \end{cases}$$

- Predict salary as:

$$\text{salary} = \beta_0 + \beta_1 \times x_i = \begin{cases} \beta_0 + \beta_1 & \text{if } \text{state}_i = \mathrm{MA} \\ \beta_0 + \beta_1 & \text{if } \text{state}_i = \mathrm{NH} \\ \beta_0 + 2 \times \beta_1 & \text{if } \text{state}_i = \mathrm{ME} \end{cases}$$

Qualitative Features: Many Values

- Predict salary as a function of state
- Feature $\text{state}_i \in \{\mathrm{MA}, \mathrm{NH}, \mathrm{ME}\}$
- What about x_i:

$$x_i = \begin{cases} 0 & \text{if } \text{state}_i = \mathrm{MA} \\ 1 & \text{if } \text{state}_i = \mathrm{NH} \\ 2 & \text{if } \text{state}_i = \mathrm{ME} \end{cases}$$

- Predict salary as:

$$\text{salary} = \beta_0 + \beta_1 \times x_i = \begin{cases} \beta_0 + \beta_1 & \text{if } \text{state}_i = \mathrm{MA} \\ \beta_0 + \beta_1 & \text{if } \text{state}_i = \mathrm{NH} \\ \beta_0 + 2 \times \beta_1 & \text{if } \text{state}_i = \mathrm{ME} \end{cases}$$

- Does not work: NH salary always average of MA and ME

Qualitative Features: Many Values The Right Way

- Predict salary as a function of state
- Feature $\text{state}_i \in \{\mathrm{MA}, \mathrm{NH}, \mathrm{ME}\}$

Qualitative Features: Many Values The Right Way

- Predict salary as a function of state
- Feature $\text{state}_i \in \{\mathrm{MA}, \mathrm{NH}, \mathrm{ME}\}$
- Introduce 2 **indicator variables** x_i, z_i:

$$x_i = \begin{cases} 0 & \text{if } \text{state}_i = \mathrm{MA} \\ 1 & \text{if } \text{state}_i \neq \mathrm{MA} \end{cases} \qquad z_i = \begin{cases} 0 & \text{if } \text{state}_i = \mathrm{NH} \\ 1 & \text{if } \text{state}_i \neq \mathrm{NH} \end{cases}$$

- Predict salary as:

$$\text{salary} = \beta_0 + \beta_1 \times x_i + \beta_2 \times z_i = \begin{cases} \beta_0 + \beta_1 & \text{if } \text{state}_i = \mathrm{MA} \\ \beta_0 + \beta_2 & \text{if } \text{state}_i = \mathrm{NH} \\ \beta_0 & \text{if } \text{state}_i = \mathrm{ME} \end{cases}$$

Qualitative Features: Many Values The Right Way

- Predict salary as a function of state
- Feature $\text{state}_i \in \{\mathrm{MA}, \mathrm{NH}, \mathrm{ME}\}$
- Introduce 2 **indicator variables** x_i, z_i:

$$x_i = \begin{cases} 0 & \text{if state}_i = \mathrm{MA} \\ 1 & \text{if state}_i \neq \mathrm{MA} \end{cases} \qquad z_i = \begin{cases} 0 & \text{if state}_i = \mathrm{NH} \\ 1 & \text{if state}_i \neq \mathrm{NH} \end{cases}$$

- Predict salary as:

$$\text{salary} = \beta_0 + \beta_1 \times x_i + \beta_2 \times z_i = \begin{cases} \beta_0 + \beta_1 & \text{if state}_i = \mathrm{MA} \\ \beta_0 + \beta_2 & \text{if state}_i = \mathrm{NH} \\ \beta_0 & \text{if state}_i = \mathrm{ME} \end{cases}$$

- Need an indicator variable for ME? Why? hint: linear independence

Removing Additive Assumption

- What is the additive assumption?

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio}$$

- What if TV and radio interact?

Removing Additive Assumption

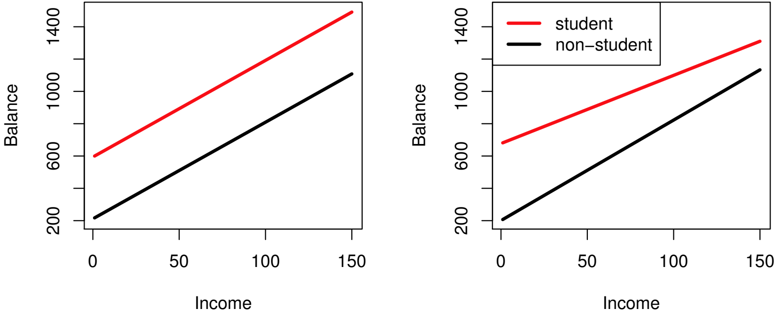
- What is the additive assumption?

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio}$$

- What if TV and radio interact?
- Add new feature:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{TV} \times \text{radio}$$

Example of Interaction

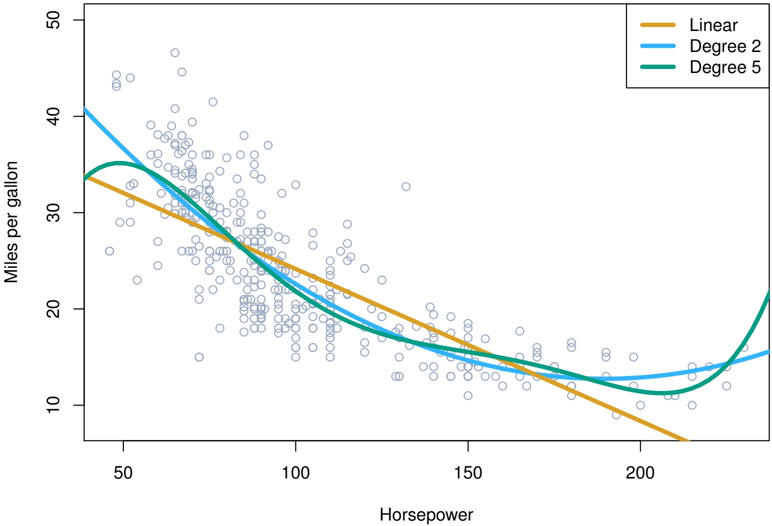


$$\text{balance}_i = \beta_0 + \beta_1 \times \text{income}_i + \beta_2 \times \text{student}_i$$

$$\text{balance}_i = \beta_0 + \beta_1 \times \text{income}_i + \beta_2 \times \text{student}_i + \beta_3 \times \text{student}_i \times \text{income}_i$$

Nonlinear Relationship

Can we use linear regression to fit a nonlinear function?



Nonlinear Relationship

- Linear regression can fit a nonlinear function
- Just introduce new features!
- Linear regression:

$$\textsf{mpg} = \beta_0 + \beta_1 \times \textsf{mpg}$$

- Degree 2 (Quadratic):

$$\textsf{mpg} = \beta_0 + \beta_1 \times \textsf{mpg} + \beta_2 \times \textsf{mpg}^2$$

- Degree k:

$$\textsf{mpg} = \sum_{i=0}^{k} \beta_k \times \textsf{mpg}^k$$

What Can Wrong

Many ways to fail:

1. Response variable is non-linear
2. Errors are correlated
3. Error variance is not constant
4. Outlier data
5. Points with high leverage
6. Features are collinear

What can be done about it?

Response variable is Non-linear

- We can fit a nonlinear model

$$\text{mpg} = \beta_0 + \beta_1 \times \text{mpg} + \beta_2 \times \text{mpg}^2$$

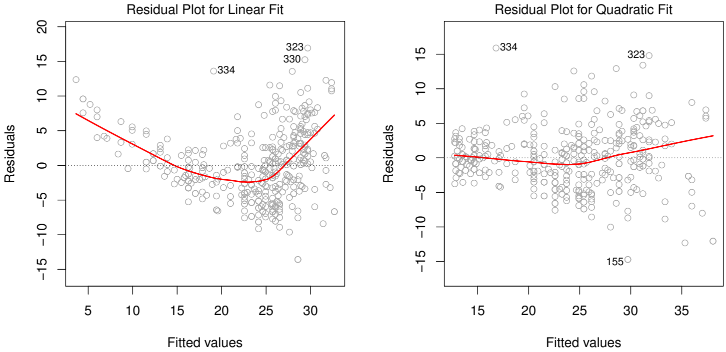
- But how do we know we should?

Response variable is Non-linear

- We can fit a nonlinear model

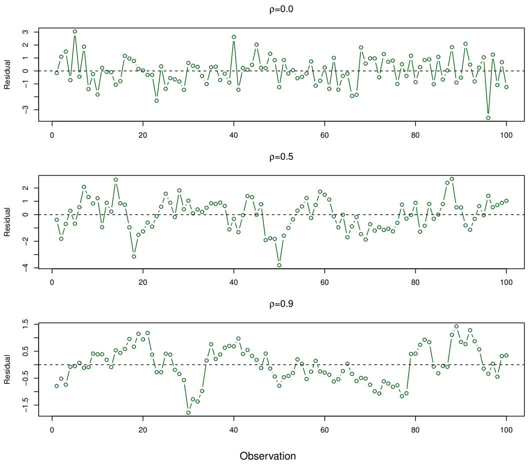
$$\text{mpg} = \beta_0 + \beta_1 \times \text{mpg} + \beta_2 \times \text{mpg}^2$$

- But how do we know we should?
- Residual plot



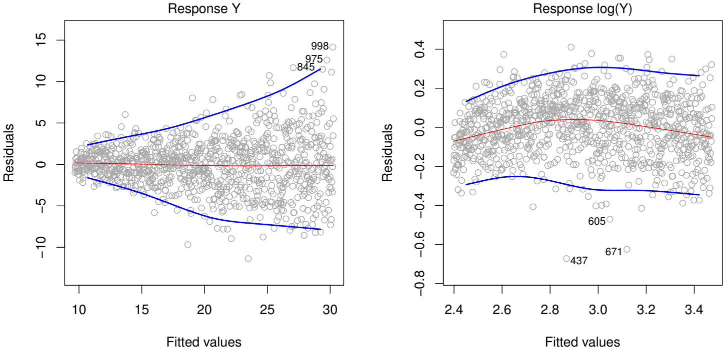
Correlated Errors

- The errors ϵ_i are not independent
- For example, use each data point twice
- No additional information, but error is apparently reduced



Non-constant Variance of Errors

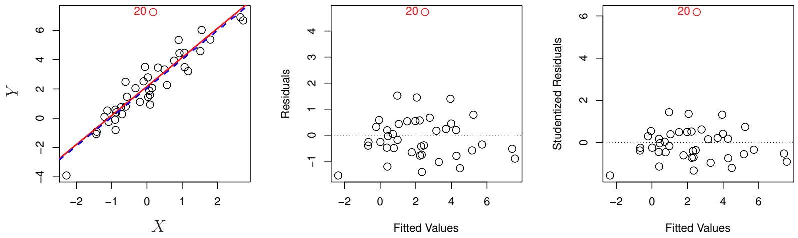
- Errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$
- **Homoscedastic** errors: $\mathrm{Var}[\epsilon_1] = \mathrm{Var}[\epsilon_2] = \ldots = \mathrm{Var}[\epsilon_n]$
- **Heteroscedastic** errors can cause a wrong fit



- **Remedy**: scale response variable Y or use *weighted linear regression*

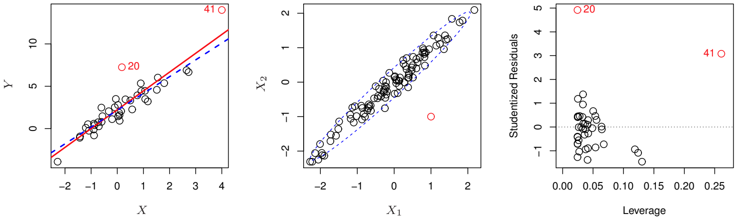
Outlier Data Points

- Data point that is far away from others
- Measurement failure, sensor fails, missing data point
- Can seriously influence prediction quality



Points with High Leverage

- Points with unusual value of x_i
- Single data point can have significant impact on prediction
- R and other packages can compute leverages of data points

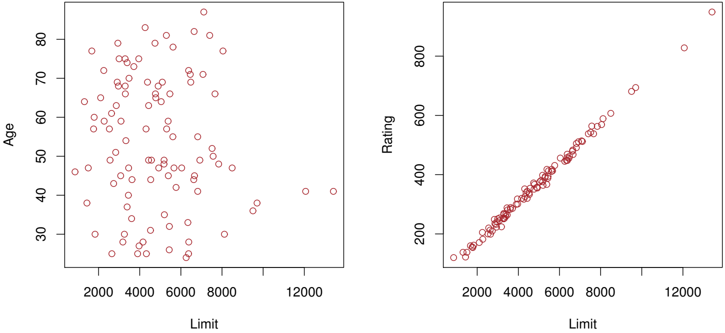


- Good to remove points with high leverage and residual

Collinear Features

- Collinear features can reduce prediction confidence

$$\text{credit} \approx \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{limit}$$



- Detect by computing feature correlations
- Solution: remove collinear feature or combine them

Lab