Introduction to Supervised Learning

Introduction: Some Questions

- Why do we need to understand supervised learning?
- What is Supervised Learning?
- How do we formulate and solve supervised learning problems?

Supervised Learning: An example

- Office arrival time prediction

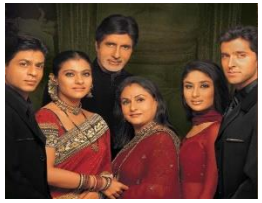
Start time

Journey time

Arrival time

Supervised Learning: An example

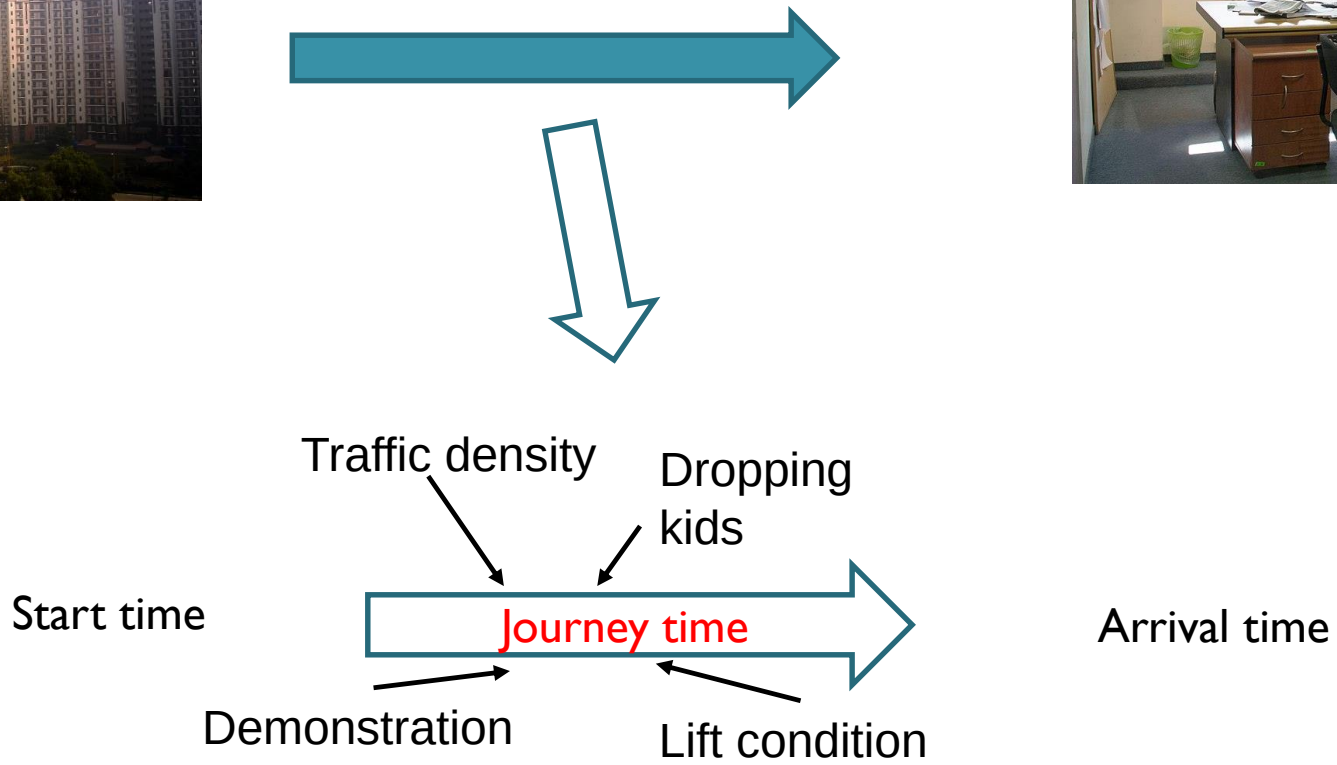
- Office arrival time prediction



How do we predict my office arrival time?

Supervised Learning: An example

- Office arrival time prediction



Supervised Learning: An example

- Office arrival time prediction

Collect Data

Day	Start time	# of demonstration	Dropping time to kids (in min)	Time in lift (in min)	Traffic density (veh/m^2)	Journey time (min)
1	9.26 am	2	10	0.8	0.25	30.2
2	9 am	0	12	1	0.15	29. 6
3	9.25 am	5	9	2.2	0.85	42
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Supervised Learning: An example

- Office arrival time prediction

Collect Data

Day	Start time	# of demonstration	Dropping time to kids (in min)	Time in lift (in min)	Traffic density (veh/m^2)	Journey time (min)
1	9.26 am	2	10	0.8	0.25	30.2
2	9 am	0	12	1	0.15	29. 6
3	9.25 am	5	9	2.2	0.85	42
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Measurements are possible

Our interest

Supervised Learning: An example

- Office arrival time prediction

Day	Start time	# of demonstration	Dropping time to kids (in min)	Time in lift (in min)	Traffic density (veh/m^2)	Journey time (min)
1	9.26 am	2	10	0.8	0.25	30.2
2	9 am	0	12	1	0.15	29. 6
3	9.25 am	5	9	2.2	0.85	42
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Collect Data

?

Journey time = *f*(Traffic density, Dropping time to kids, Time in lift, # of demonstration, Starting time, parameters)

Supervised Learning: An example

- Office arrival time prediction

Collect Data

Day	Start time	# of demonstration	Dropping time to kids (in min)	Time in lift (in min)	Traffic density (veh/m^2)	Journey time (min)	Arrived on time (< 10 am)
1	9.26 am	2	10	0.8	0.25	30.2	Yes
2	9 am	0	12	1	0.15	29. 6	Yes
3	9.25 am	5	9	2.2	0.85	42	No
⋮	⋮	⋮		⋮	⋮	⋮	⋮

Measurements are possible

Our interest

Supervised Learning: An example

- Office arrival time prediction

Day	Start time	# of demonstration	Dropping time to kids (in min)	Time in lift (in min)	Traffic density (veh/m^2)	Journey time (min)	Arrived on time (< 10 am)
1	9.26 am	2	10	0.8	0.25	30.2	Yes
2	9 am	0	12	1	0.15	29. 6	Yes
3	9.25 am	5	9	2.2	0.85	42	No
⋮	⋮	⋮		⋮	⋮	⋮	⋮

Collect Data

?

Arrived on time = *f*(Traffic density, Dropping time to kids, Time in lift, # of demonstration, Starting time, parameters)

Supervised Learning

- **Problem Characteristics**
  - Input variables are measured: Traffic density, # of demonstrations….
  - Output variables are measured: Journey time or arrived on time on time
  - Labelled data: For each observation input and output are available
- **Regression problem**
  - Predicting a continuous quantity output (real number typically): Journey time
- **Classification Problem**
  - Predicting a discrete output (Yes or No): Arrived on time

Supervised Learning

- **Next couple of Classes**

- **How to build regression models?**
  - Predicting a continuous quantity output (real number typically): Journey time

- **How to build classification models?**
  - Predicting a discrete output (Yes or No): Arrived on time

Regression Analysis

Outlines

- Regression
  - Types of regression
  - Fitting a function – Criterion for best fit
  - Least squares
- Correlation vs Regression
- Simple regression
- Multiple regression
- Diagnostics and ANOVA
- Model assessment and validation
- Non-parametric testing

Preliminaries

- n observations for x and y variables (x_i, y_i)
- Sample means $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{y} = \frac{\sum y_i}{n}$$

- Sample variances S_{xx} and S_{yy}

$$S_{xx} = \frac{1}{n}\sum(x_i - \bar{x})^2 \qquad S_{yy} = \frac{1}{n}\sum(y_i - \bar{y})^2$$

- Sample covariance S_{xy}

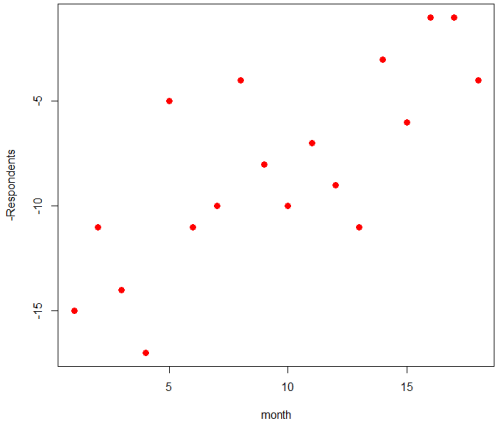
$$S_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$$

Correlation

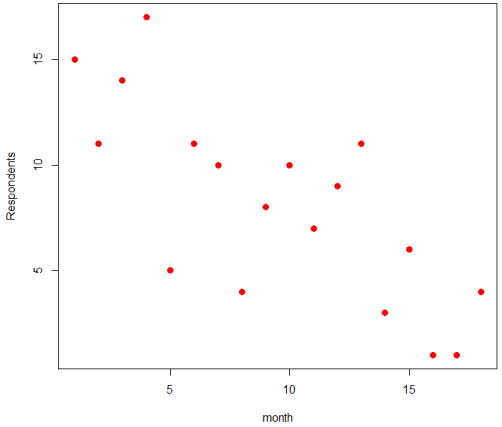
- Correlation: the strength of association between two variables
- Visual representation of correlation:

Scatter grams

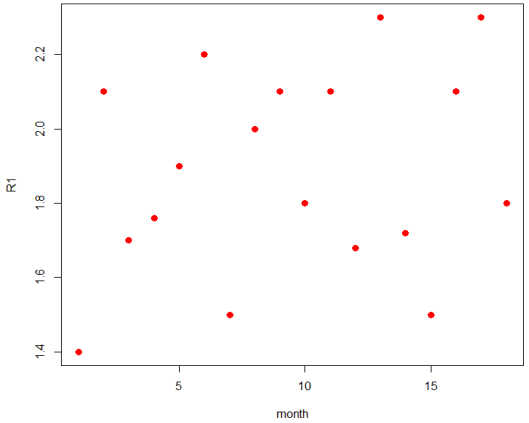
Positive trend



Negative trend



Little or no correlation



Quantitative Metric?

Pearson's Correlation

- n observations for x and y variables (x_i, y_i)
- Pearson's product-moment correlation coefficient (r_{xy})

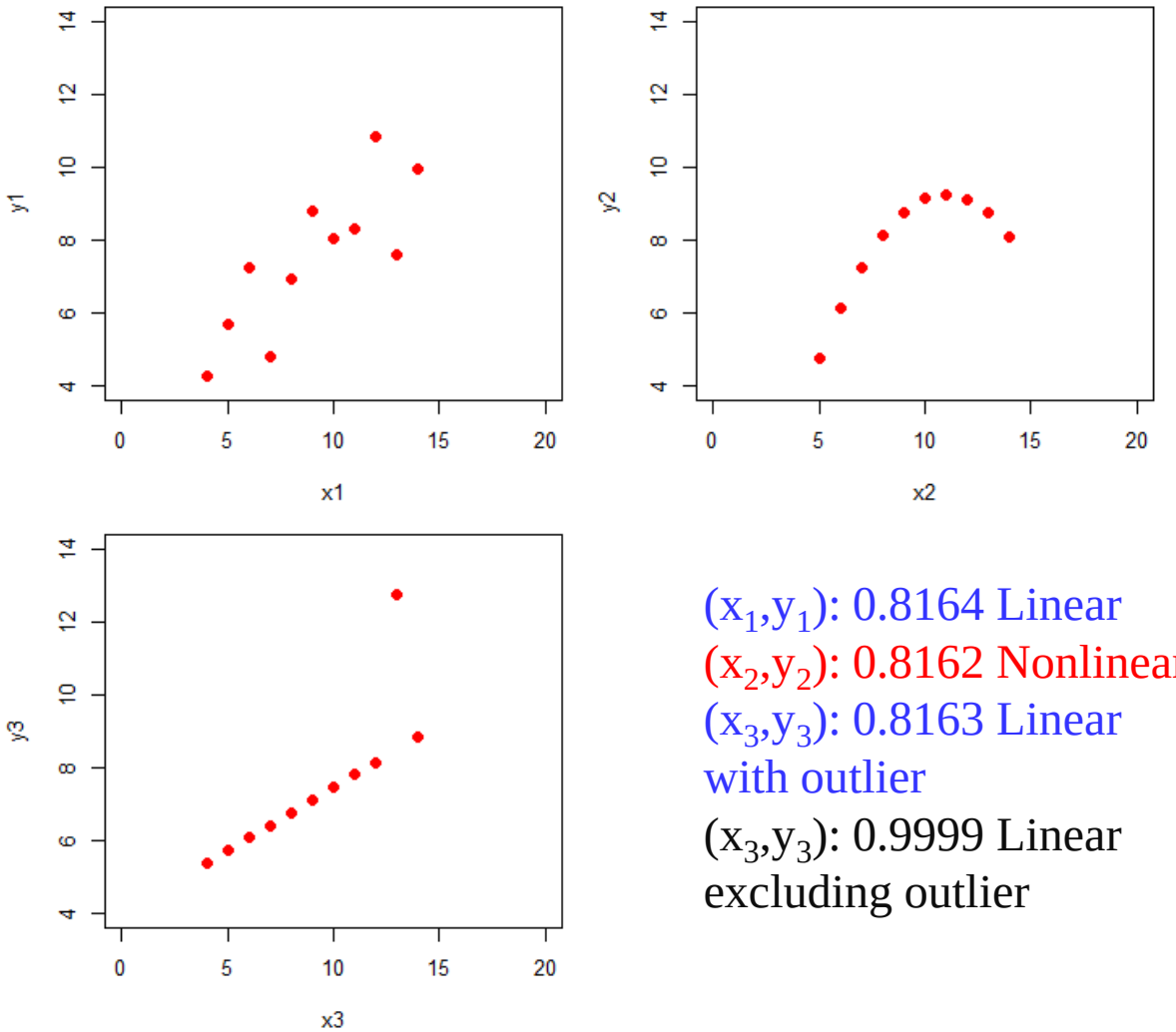
$$r_{xy} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)}\sqrt{(\Sigma y_i^2 - n\bar{y}^2)}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- r_{xy} takes a value between -1 (negative correlation) and 1 (positive correlation)
- r_{xy}=0 means no correlation

Pearson's Correlation (Cont.)

- A measure for the degree of linear dependence between x and y
- Cannot be applied to ordinal variables
- Sample size: Moderate (20-30) for good estimate
- Robustness: Outliers can lead to misleading values

Pearson's Correlation: Anscombe's data



(x_1,y_1): 0.8164 Linear
(x_2,y_2): 0.8162 Nonlinear
(x_3,y_3): 0.8163 Linear with outlier
(x_3,y_3): 0.9999 Linear excluding outlier

LINEAR REGRESSION

Introduction: Regression analysis

- One of the widely used statistical techniques
- Purpose is to build a functional relationship between *dependent variable(s)* and *independent variable(s)*
- Dependent variables also known as *Response variable*, *Regressand*, *Predicted variable* - denoted as variable/s *x*
- Independent variable also known as *Predictor variable*, *Regressor*, *Exploratory variable* - denoted as variable/s *y*
- **Use of regression model:** Given values of independent variable **predict** the values of dependent variable in the future using the regression model (also known as soft sensor)

Introduction: Regression analysis

- Classification of Regression Analysis
  - Univariate vs Multivariate

    Univariate: One dependent and one independent variable

    Multivariate: Multiple independent and multiple dependent variables
  - Linear vs Nonlinear

    Linear: Relationship is linear between dependent and independent variables

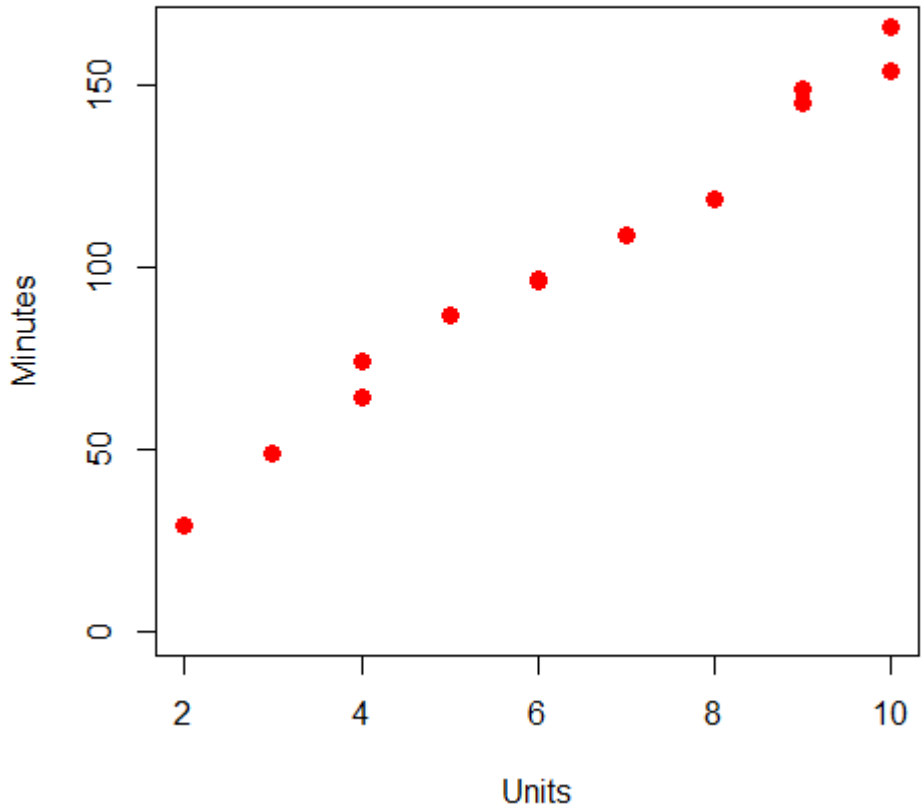
    Nonlinear: Relationship is nonlinear between dependent and independent variables
  - *Simple vs Multiple*

    Simple: One dependent and one independent variable

    Multiple: One dependent and Multiple independent variables

Correlation vs Regression

- **Example**: Length of service calls (y) and number of units repaired (x)



Pearson's correlation co-efficient=0.994
Spearman's rank co-efficient = 0.995

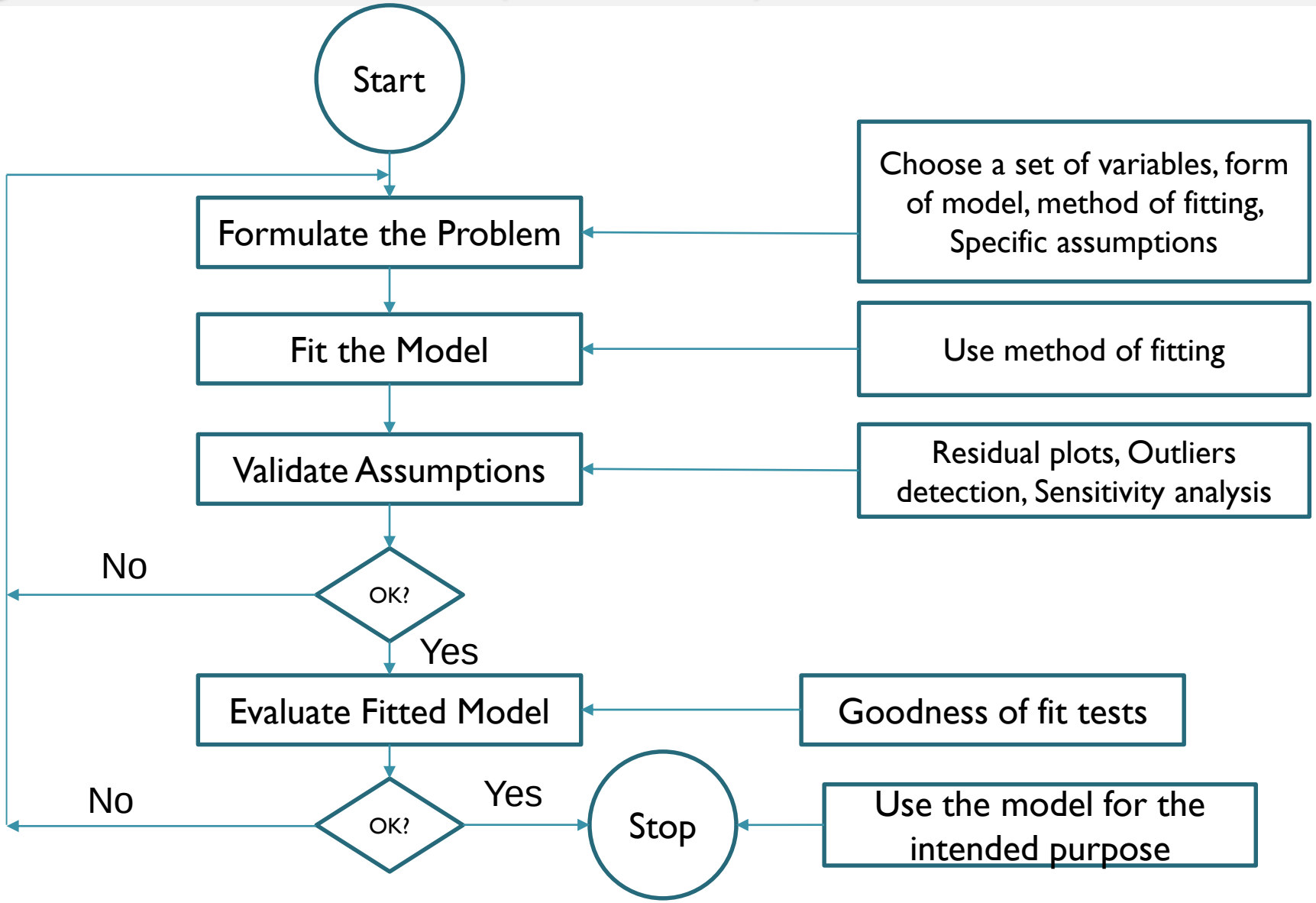
Introduction: Regression analysis

- Is there a relationship between these variables?
- How strong is the relationship?
- How accurately can we estimate these effects?
- How accurately can we predict future?
- Is the relationship linear?

Introduction: Regression analysis

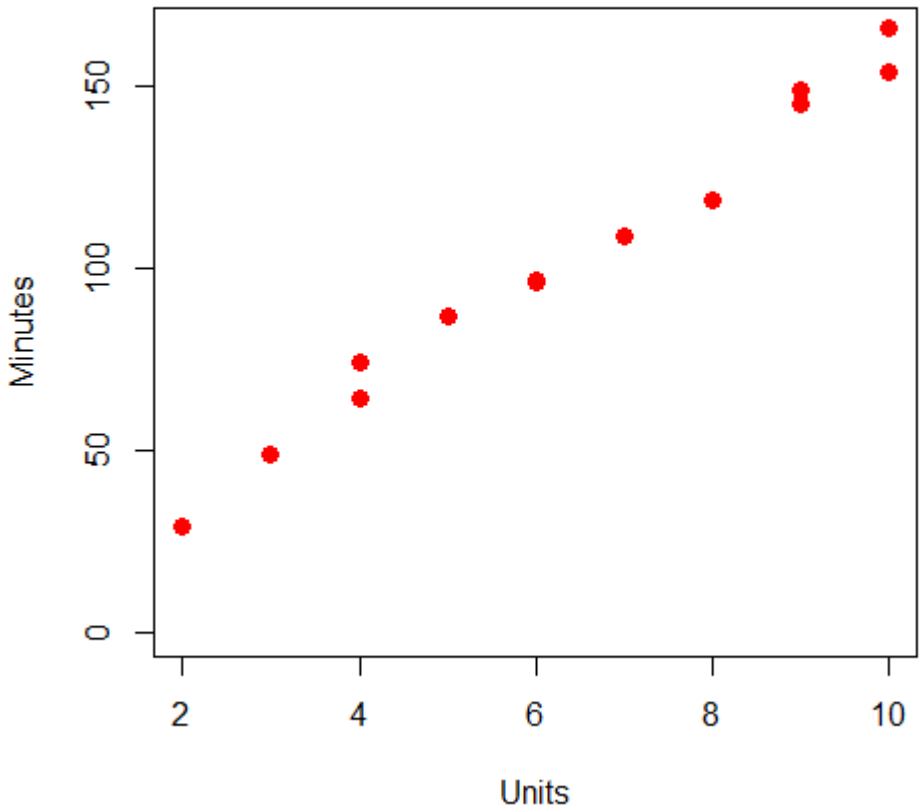
- Regression analysis
  - Linear regression methods
    - Simple linear regression
    - Multiple linear regression
    - Weighted least squares
    - Principal component regression
    - Ridge regression
    - Lasso
    - Partial least squares
  - Nonlinear regression methods
    - Polynomial regression
    - Spline regression

Regression Process (Iterative)



Ordinary Least Squares (OLS)

- **Example**: Length of service calls (y) and number of units repaired (x)



Objective: Find a relationship between y and x.

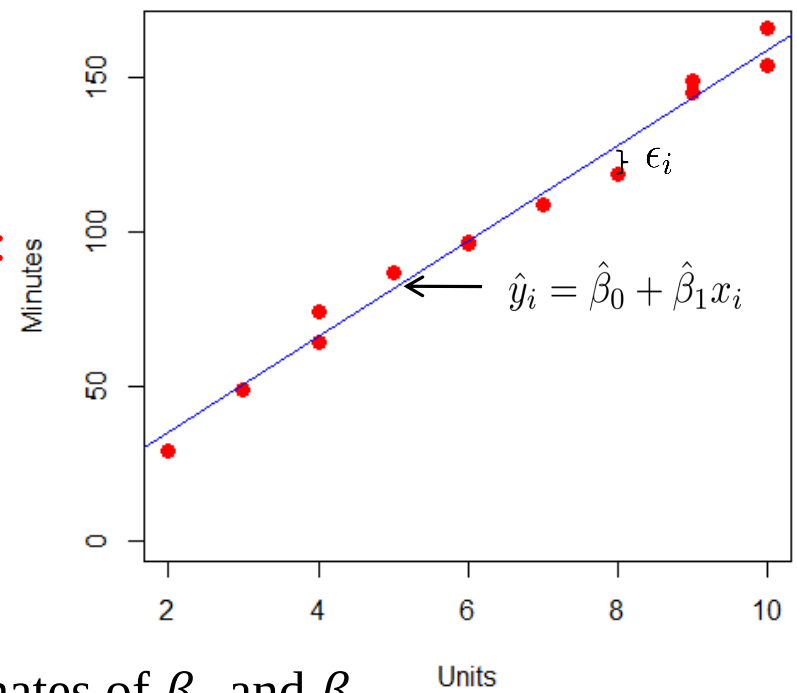
Ordinary Least Squares

- Linear model between y_i and x_i, $i = 1,\ldots, n$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Error in only dependent variable and no error in independent variable:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$



- The sum of squares of errors (SSE)

$$\sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

- The minimization of SSE gives estimates of β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

OLS: Testing Goodness of Model

- Prediction using the regression equation: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- A measure of the variability explained by the regression equation
- R^2 is defined as a measure

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

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

Total variability in y

- R^2 values: Between 0 and 1
  - 0 indicates poor fitting
  - 1 indicates the best fitting
- Adjusted $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2/(n-p-1)}{\sum(y_i - \bar{y})^2/(n-1)}$$

OLS: Assumptions on Error Characteristics

- Assumptions
  - Error distribution: $\epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, 2, \ldots, n$
  - Uncorrelated errors
  - The errors must be uncorrelated to x_i, $i=1,2,\ldots, n$
  - ϵ_i and ϵ_j are also uncorrelated

OLS: Properties of Estimates

- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates are unbiased

$$E[\hat{\beta}_0] = \beta_0, \quad E[\hat{\beta}_1] = \beta_1$$

- Variance of the estimates

$$\mathrm{var}[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}, \quad \mathrm{var}[\hat{\beta}_0] = \sigma^2 \frac{\sum x_i^2}{n\, S_{xx}}$$

- Estimate of σ^2

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\mathrm{SSE}}{n-2}$$

- $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{S_{xx}})$

Simple Linear Regression

- Inferences on $\hat{\beta}_0$ and $\hat{\beta}_1$
- Confidence interval for $\hat{\beta}_1$ at 5% level of significance

$t_{0.025,12}$

$$\beta_1 \in [\hat{\beta}_1 - 2.18\, s_{\hat{\beta}_1}, \hat{\beta}_1 + 2.18\, s_{\hat{\beta}_1}], \quad s_{\hat{\beta}_1} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2) S_{xx}}}$$

- Confidence interval for $\hat{\beta}_0$ at 5% level of significance

$$\beta_0 \in [\hat{\beta}_0 - 2.18\, s_{\hat{\beta}_0}, \hat{\beta}_0 + 2.18\, s_{\hat{\beta}_0}], \quad s_{\hat{\beta}_0} = s_e \sqrt{\frac{\sum x_i^2}{n\, S_{xx}}}$$

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}}$$

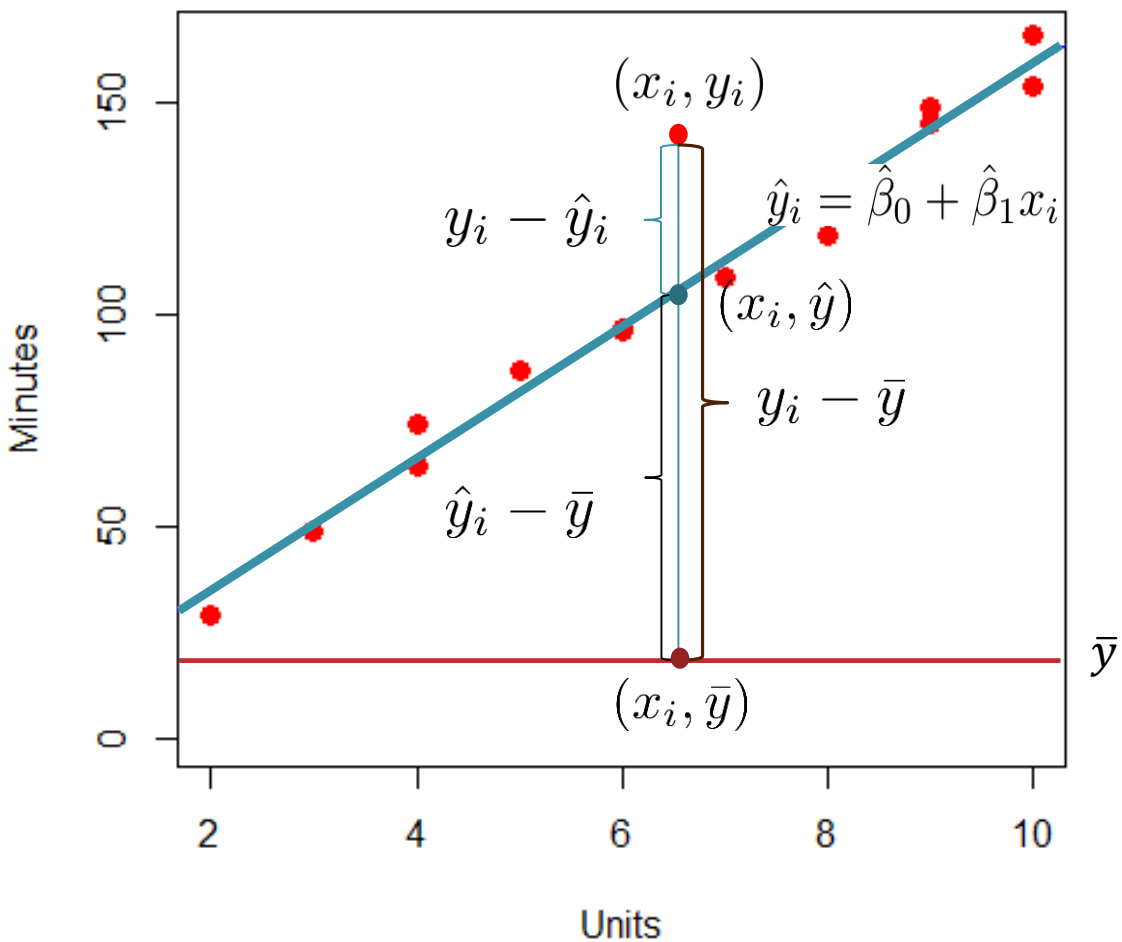
OLS: F-Test

- Fitted model is adequate or can be reduced further?
  - Test significance of individual coefficient $\widehat{\beta}$
  - A test on the full model (FM) vs the reduced model (RM)
- Hypotheses testing

H_0: $\hat{y}_i = \hat{\beta}_0 + \epsilon_i$ ← Reduced Model

H_1: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$ ← Full Model

OLS: Sum Squared Quantities



$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$$

$$\text{SST} = \sum (y_i - \bar{y})^2$$

OLS: F-Test

- F-test for testing two models
- F-statistic

$$F_o = \frac{SST - SSE}{SSE/(n-2)} = \frac{SSR}{SSE/(n-2)}$$

Degrees of freedom

  - Note that the numerator has 1 df

- For α-significance level: Reject H_o if

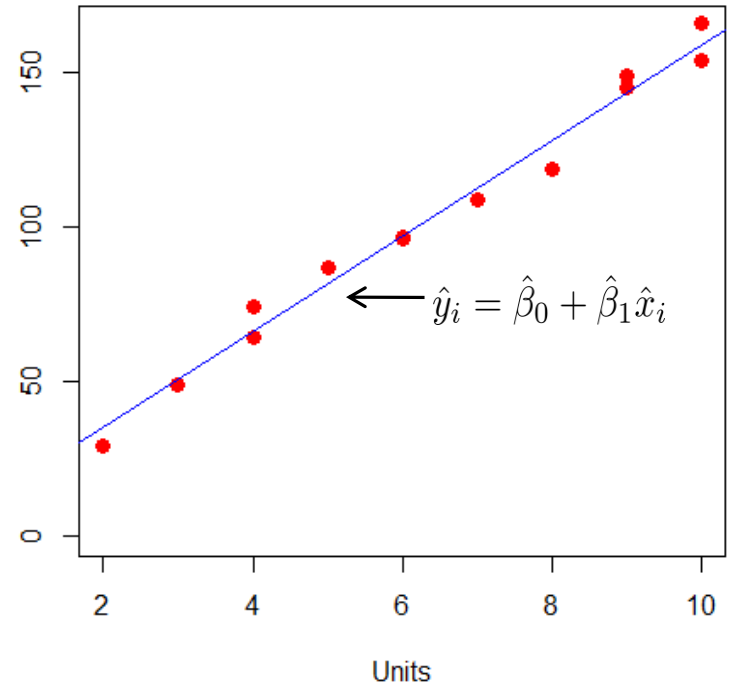
$$F_o \geq F_{(1,n-2;0.05)}$$

where F-statistic for the given dfs from the table

OLS: Example

```
Call:
lm(formula = Minutes ~ Units)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.2318 -3.3415 -0.7143  4.7769  7.8033 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.162      3.355    1.24    0.239    
Units         15.509      0.505   30.71 8.92e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-squared:  0.9874, Adjusted R-squared:  0.9864 
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

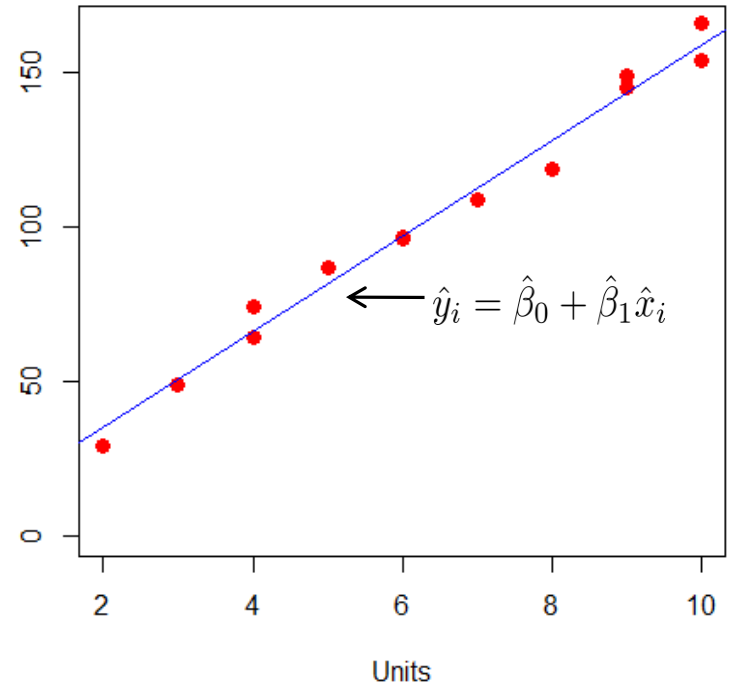
$\hat{\beta}_0$ $s_{\hat{\beta}_0}$ β_1 $s_{\hat{\beta}_1}$



OLS: Example

```
Call:
lm(formula = Minutes ~ Units)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.2318 -3.3415 -0.7143  4.7769  7.8033 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.162      3.355    1.24    0.239    
Units         15.509      0.505   30.71 8.92e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-squared:  0.9874, Adjusted R-squared:  0.9864 
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

$\hat{\beta}_0$ $s_{\hat{\beta}_0}$ $\hat{\beta}_1$ $s_{\hat{\beta}_1}$



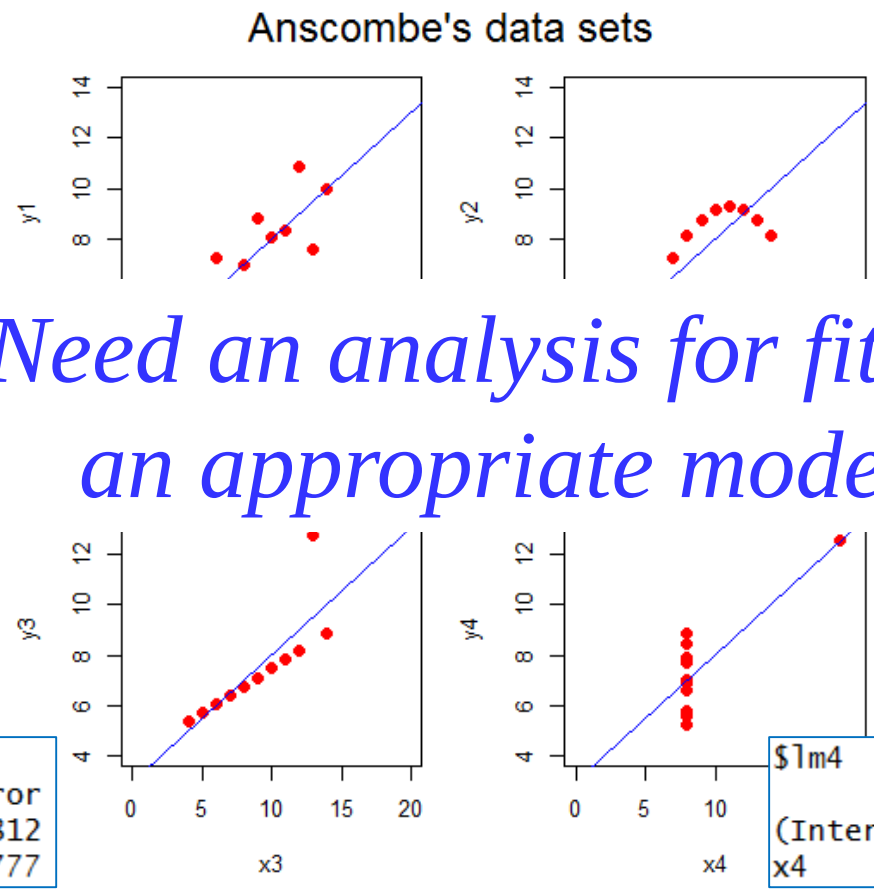

```
                 2.5 % 97.5 %
(Intercept) -3.148 11.472
Units       14.409 16.609
```

$\hat{\beta}_0$ $\hat{\beta}_1$

DIAGNOSTICS

OLS: Anscombe's data

- Got a regression equation from data: Job done?
- Anscombe's (1973) data sets:



Need an analysis for fitting an appropriate model

```
$lm1
            Estimate S
(Intercept) 3.0000909
x1          0.5000909
```

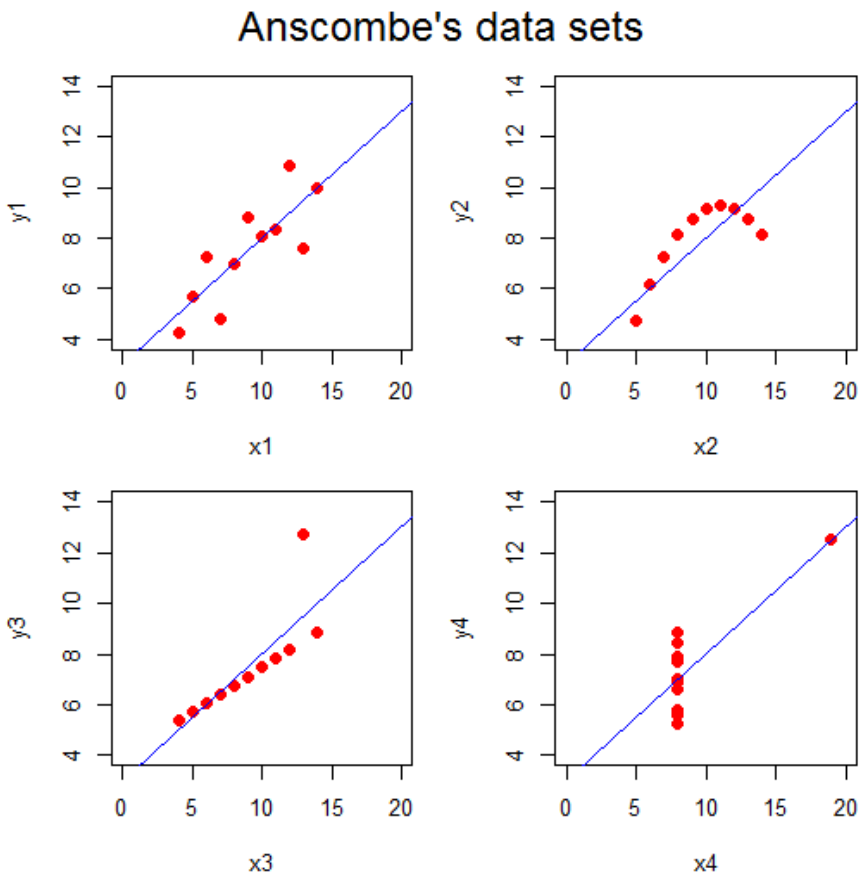
```
mate Std. Error
0909  1.1253024
0000  0.1179637
```

```
$lm3
            Estimate Std. Error
(Intercept) 3.0024545  1.1244812
x3          0.4997273  0.1178777
```

```
$lm4
            Estimate Std. Error
(Intercept) 3.0017273  1.1239211
x4          0.4999091  0.1178189
```

OLS: Diagnostics

- Questions:
  - Do the underlying data satisfy the assumptions?
  - Data are free of outliers?
  - Do some observations exert more influence than others?
  - Can the regression equation be improved by using a nonlinear model?



OLS: Residual plots

- A straightforward method for validation of a model:

Residual plots

- Different types of residual
  - Ordinary least squares residuals

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

  - Variance of e_i is not same for all data points and also correlated
  - $\mathrm{Var}(e_i) = \sigma^2(1 - p_{ii}), \;\; p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$

$$\mathrm{Cov}(e_i e_j) = -\sigma^2(p_{ij}), \;\; p_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum (x_i - \bar{x})^2}$$

OLS: Residual plots

- Standardized residual:

$$z_i = \frac{e_i}{S_e\sqrt{1-p_{ii}}}, \qquad i = 1, 2, \ldots, n$$

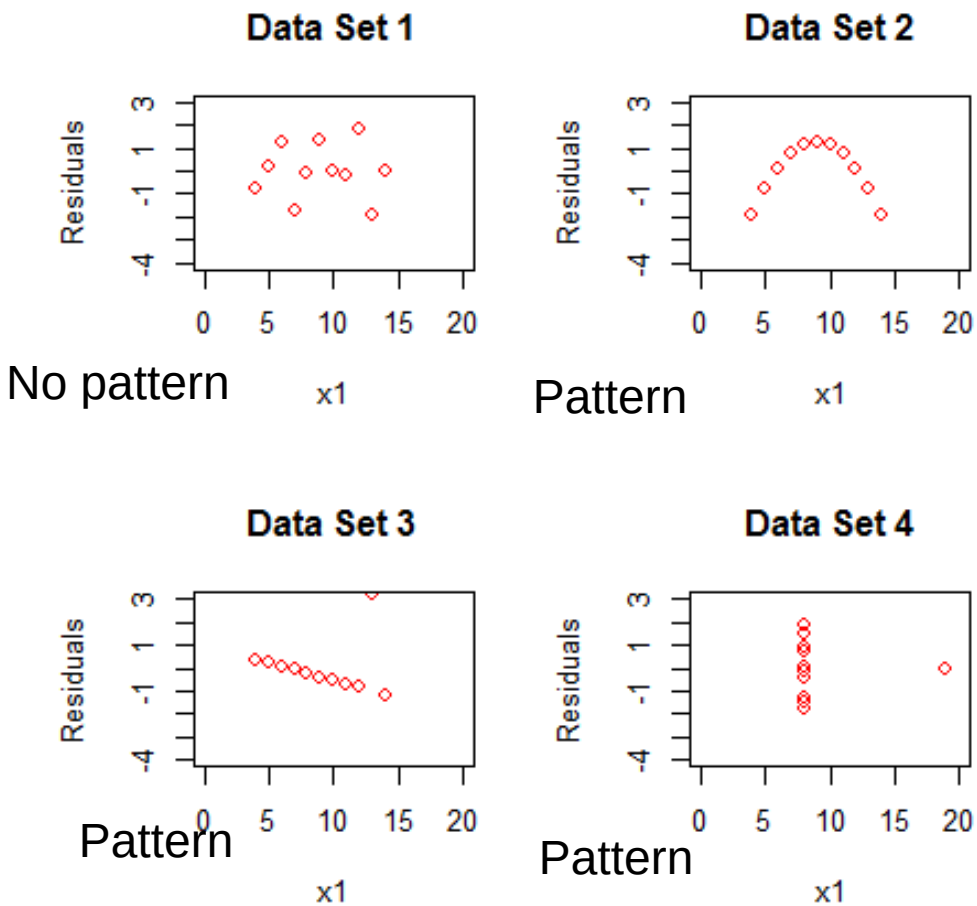
- $z_i \sim \mathcal{N}(0, 1)$

OLS: Residual Plots

- Residual plots: Standardized residuals for assessing
  - Validity of the linear model
  - Normality of the errors
  - Homoscedastic vs heteroscedastic errors
  - Correlated vs uncorrelated errors

OLS: Residual Plots

- Residual plots: OLS residuals for Anscombe's data



- Look for patterns
  - No pattern:
    A valid model
  - Pattern
    Not a valid model
- Shape of Pattern
  Information on the function of *x*

Normal Q-Q Plot

- Plot of sample quantiles against quantiles from normal distribution
- Quantile (or percentile) is the data value below which a certain percentage of data lies
- Quantiles for a standard normal distribution

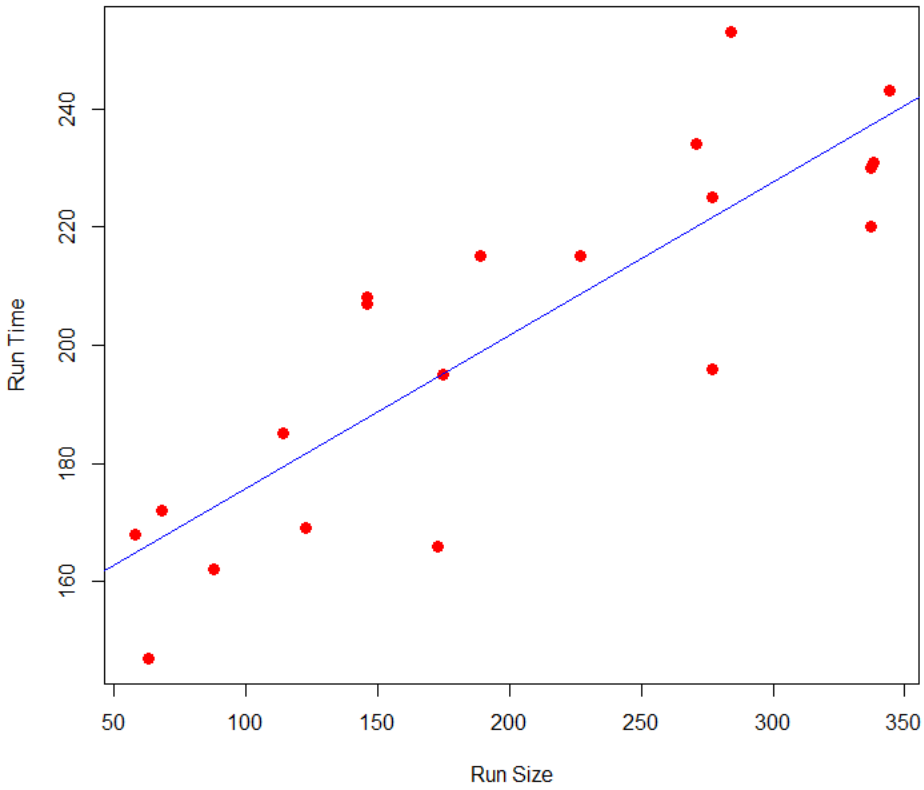
[10% 20% 30% 40% 50% 60% 70% 80% 90%]

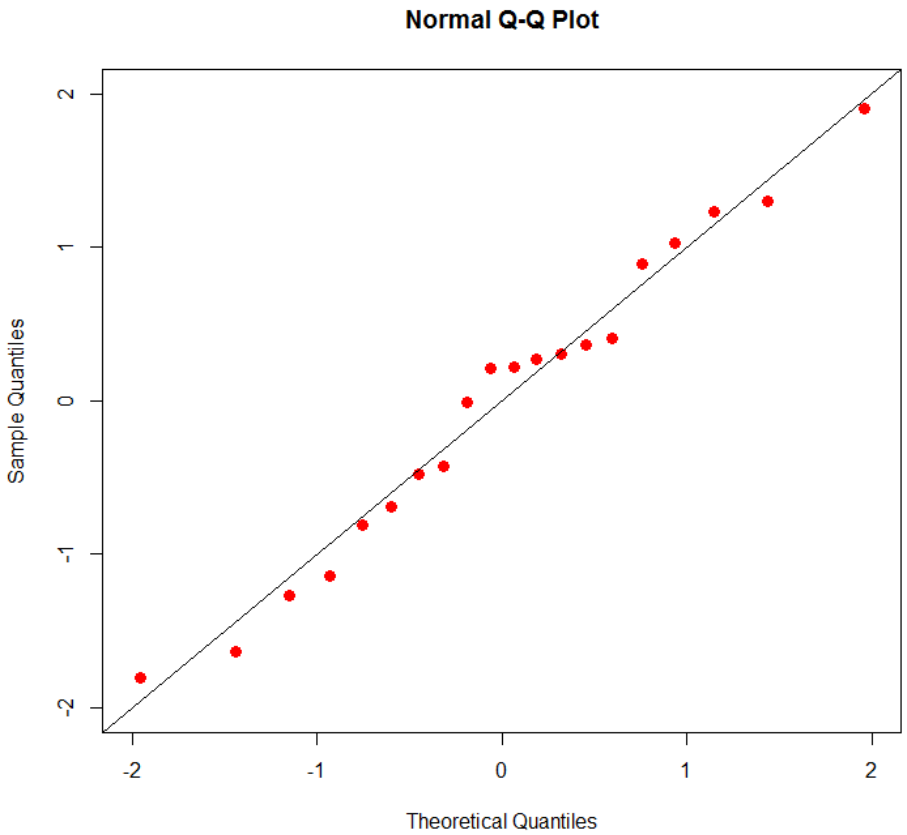
[-1.28 -0.84 -0.52 -0.25 0.0 0.25 0.52 0.841.28]

- Sample quantiles
  - Arrange standardized samples in increasing order
  - Choose number of quantiles = number of samples
  - Find corresponding quantiles of standard normal distribution
  - If data points fall on 45 deg line, then data is from normal distribution
- Function qqnorm(x) in R

OLS: Residual Plots

Normality of the errors : Normal Q-Q plot of standardized residuals: the ordered standardized residuals vs the expected order statistics from a z-distribution

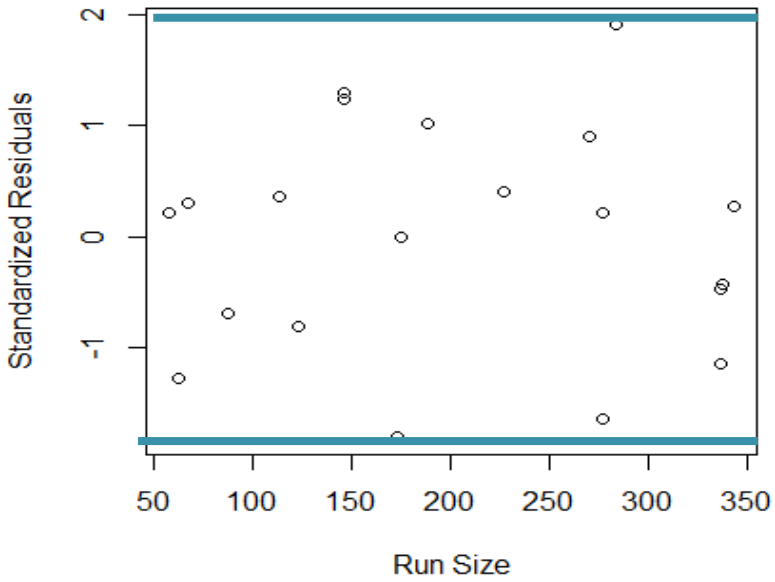




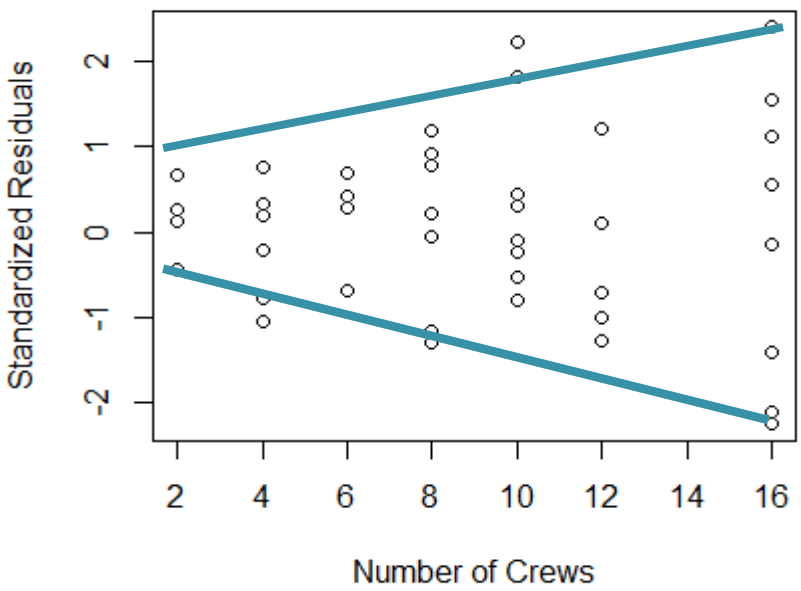
OLS: Residual Plots

Heteroscedastic errors: Standardized residuals vs *x*

Assumptions: Constant variance of errors



Constant variance



Heteroscedastic errors

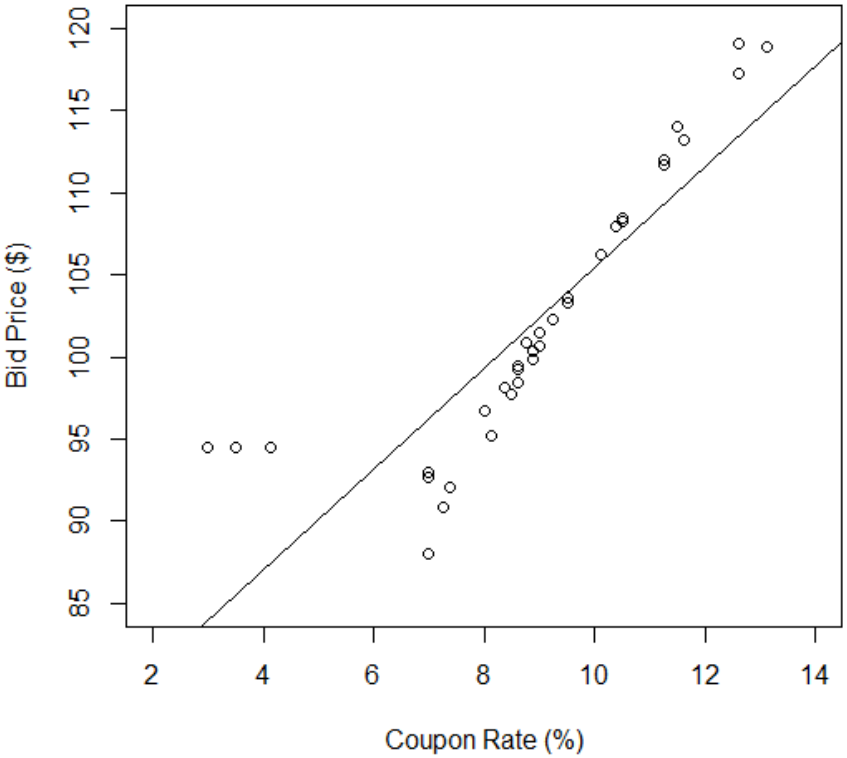
OLS: Residual Plots

- Outliers: Points which do not follow the pattern set by the bulk of the data
- Outliers:
  - For small- and moderate-size data sets
    - The standardized residuals of points which fall out side [-2,2]
  - For large data set
    - The standardized residual of points which fall out side [-4,4]

OLS: Residual Plots

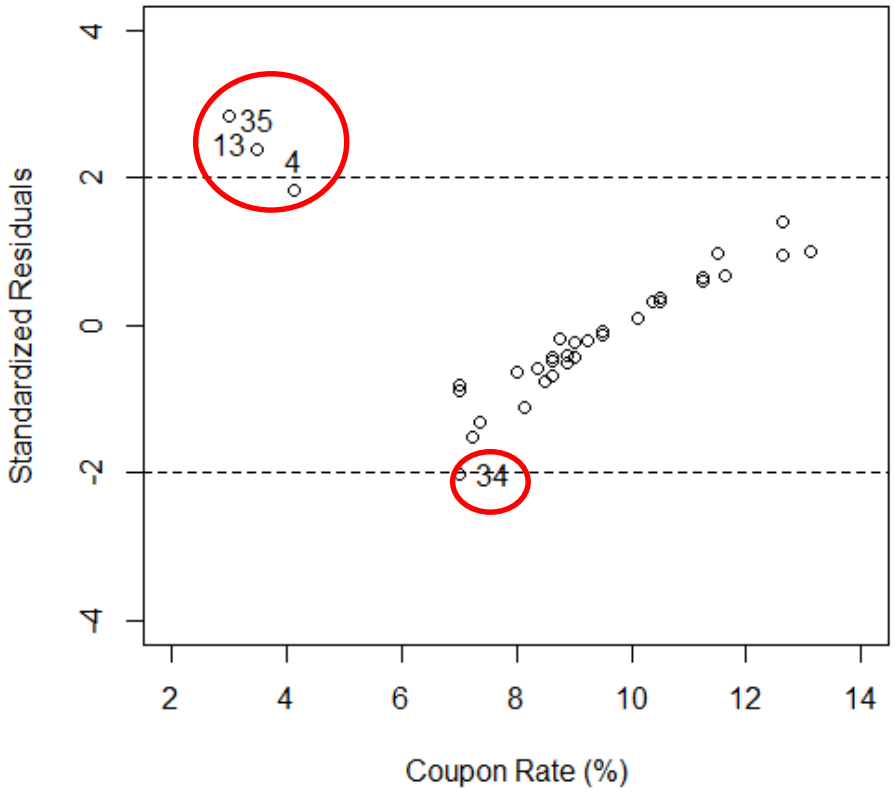
- Outliers: US Bonds example

```
Call:
lm(formula = BidPrice ~ CouponRate)

Residuals:
   Min     1Q Median     3Q    Max 
-8.249 -2.470 -0.838  2.550 10.515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516, Adjusted R-squared:  0.7441 
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```



Simple Linear Regression: Diagnostics

- Outliers: US Bonds example



Simple Linear Regression: Diagnostics

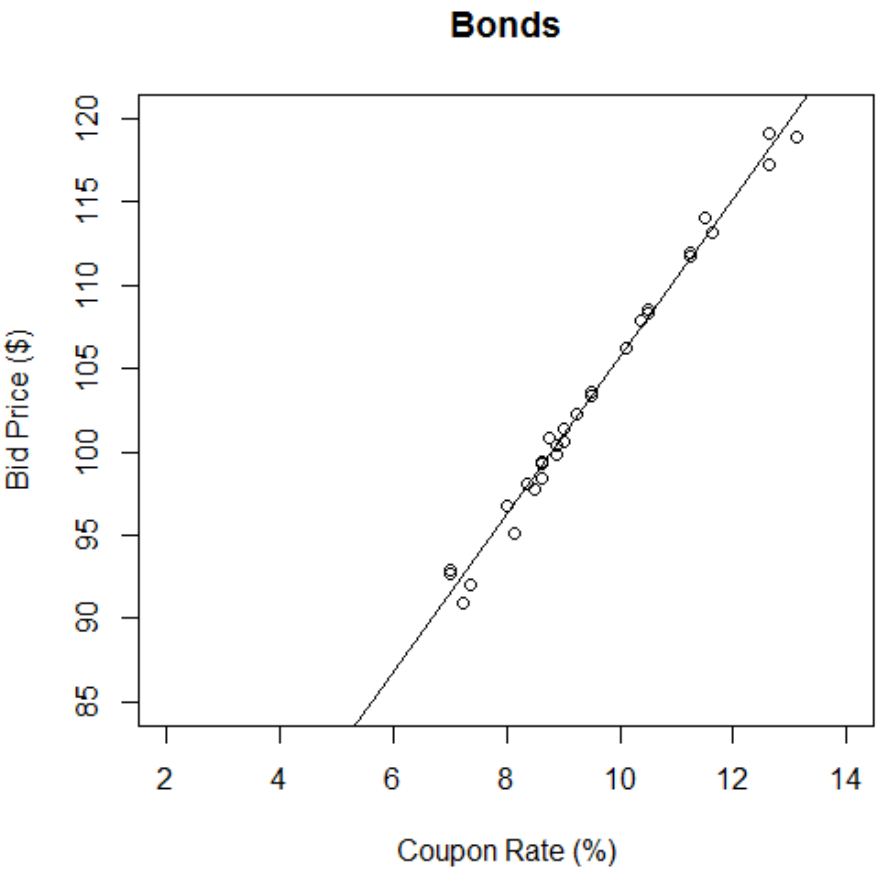
- US Bonds example

With Outliers

```
Call:
lm(formula = BidPrice ~ CouponRate)

Residuals:
   Min     1Q Median     3Q    Max 
-8.249 -2.470 -0.838  2.550 10.515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516, Adjusted R-squared:  0.7441 
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Without Outliers

```
Call:
lm(formula = BidPrice ~ CouponRate, subset = (1:35)[-c(4, 13, 
    35, 34)])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.77099 -0.37978  0.08147  0.44518  1.47401 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 58.28581    0.88993   65.50   <2e-16 ***
CouponRate   4.74002    0.09227   51.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8434 on 29 degrees of freedom
Multiple R-squared:  0.9891, Adjusted R-squared:  0.9888 
F-statistic:  2639 on 1 and 29 DF,  p-value: < 2.2e-16
```



Multiple Linear Regression

- Response variable (y) depends on *p* predictors x_i, $i=1,2,\ldots,p$
- A linear model

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

- For *i*th observation

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \epsilon_i$$

- Objective: Using *n* observations, estimate regression coefficients

Multiple Linear Regression

- Approach similar to simple regression

  Minimize the sum of squares of the errors

- Matrices

$$\mathbf{y} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{1,1} - \bar{x}_1 & x_{2,1} - \bar{x}_2 & \cdots & x_{p,1} - \bar{x}_p \\ x_{1,2} - \bar{x}_1 & x_{2,2} - \bar{x}_2 & \cdots & x_{p,2} - \bar{x}_p \\ \vdots & \vdots & \cdots & \vdots \\ x_{1,n} - \bar{x}_1 & x_{2,n} - \bar{x}_2 & \cdots & x_{p,n} - \bar{x}_p \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

- The linear model in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\epsilon}) = \mathbf{0},\ Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

- SSE

$$S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Multiple Linear Regression

- Minimization of the SSE leads to the normal equations

$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

- Assumption: $(\mathbf{X}^T\mathbf{X})$ is of full rank
- The coefficients vector

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y};\ \beta_0 = \bar{y} - \bar{\mathbf{x}}^T\hat{\boldsymbol{\beta}}$$

- The properties of the estimators

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$
$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE)

Multiple Linear Regression

- Estimate of the error variance

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-p-1}$$

where *(n-p-1)* is the degrees of freedom (df)

- Confidence intervals for β_j, *j=0,1,…,p with α-level*

$$\beta_j \in [\hat{\beta}_j - t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_j), \hat{\beta}_j + t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_j)]$$

$t_{(n-p-1,\, \alpha/2)}$ is the *(1 – α/2)* percentile point of the *t*-distribution with *(n-p-1) df*

$$s.e.(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}}$$

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$$

Multiple Linear Regression

- Multiple correlation coefficient

$$Cor(y, \hat{y}) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}}$$

- The coefficient of determination $R^2 = cor\ (y, \hat{y})$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Adjusted R-squared, R_a^2

$$R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Multiple Linear Regression

- Fitted model is adequate or can be reduced further?
  - Test significance of individual coefficient $\widehat{\beta}$
  - A general unified test on the full model (FM) vs the reduced model (RM)
- Hypothesis testing

  H_0: *Reduced model is adequate*

  H_1: *Full model is adequate*

Multiple Linear Regression

- Testing two models: RM with *k* parameters
- F-statistic

$$F_o = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

Degrees of freedom

- Note that SSE(RM) ≥ SSE(FM)

- For α-significance level: Reject H_o if

$$F_o \geq F_{(p+1-k, n-p-1; \alpha)}$$

where F-statistic for the given dfs from the table

Multiple Linear Regression

Menu pricing in Restaurants of NYC

y : Price of dinner
x_1: Customer rating of the food (Food)
x_2: Customer rating of the décor (Décor)
x_3: Customer rating of the service (Service)
x_4: If the restaurant is east or west (East)

Objective: Build a model

Multiple Linear Regression

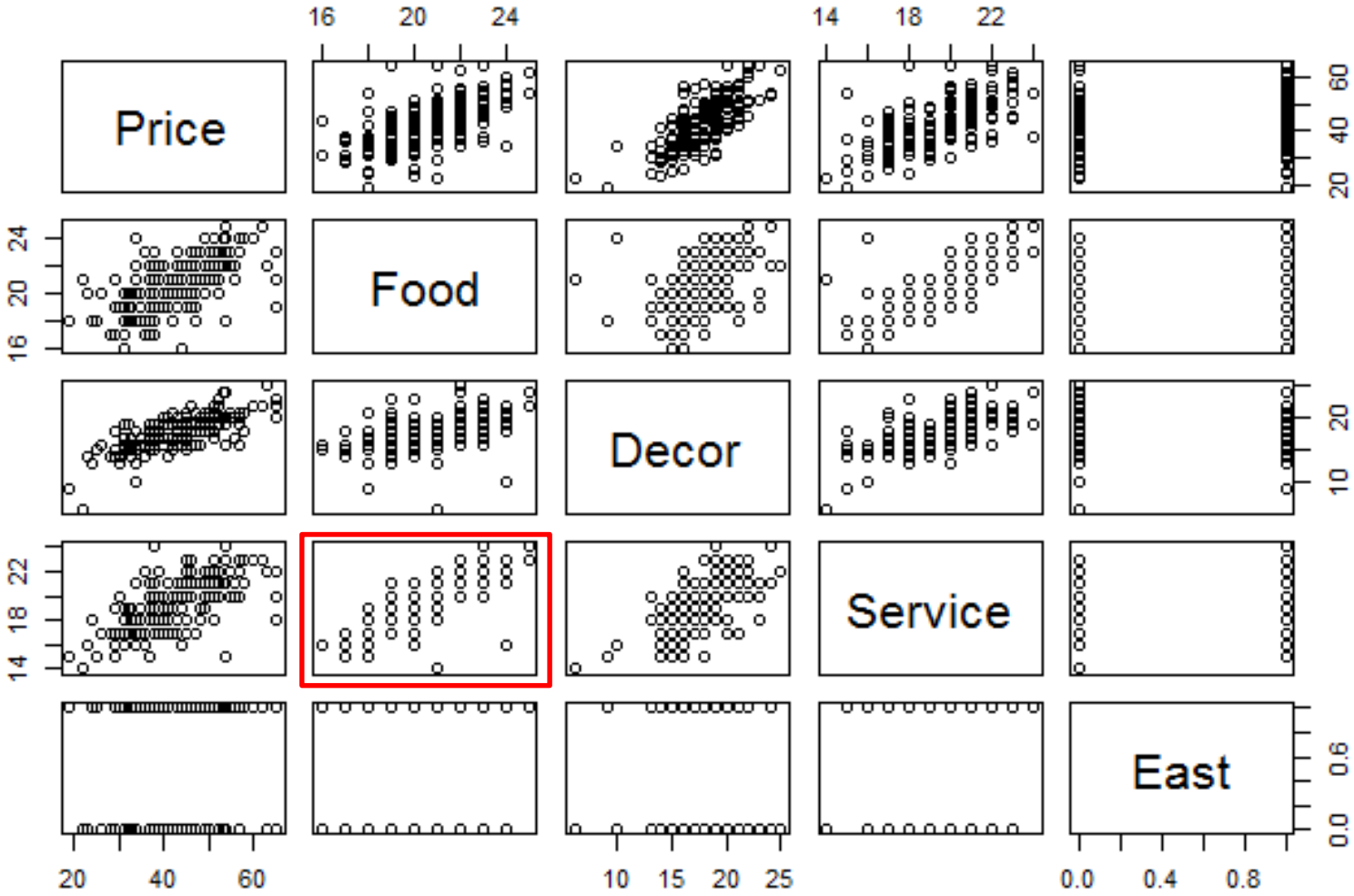
Regression output from R

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.023800   4.708359  -5.102 9.24e-07 ***
Food          1.538120   0.368951   4.169 4.96e-05 ***
Decor         1.910087   0.217005   8.802 1.87e-15 ***
Service      -0.002727   0.396232  -0.007   0.9945
East          2.068050   0.946739   2.184   0.0304 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.738 on 163 degrees of freedom
Multiple R-squared:  0.6279, Adjusted R-squared:  0.6187
F-statistic: 68.76 on 4 and 163 DF,  p-value: < 2.2e-16
```

$$\hat{y}_i = -24.024 + 1.538x_1 + 1.910x_2 - \boxed{0.003x_3} + 2.068x_4$$

Remove x_3

Multiple Linear Regression



Multiple Linear Regression

Regression output from R without Service variable

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.0269     4.6727  -5.142 7.67e-07 ***
Food          1.5363     0.2632   5.838 2.76e-08 ***
Decor         1.9094     0.1900  10.049  < 2e-16 ***
East          2.0670     0.9318   2.218   0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279, Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

$$\hat{y}_i = -24.027 + 1.536x_1 + 1.910x_2 + 2.067x_4$$

Caution: Removing several predictors may lead a dramatic effect on the coefficients in the reduced model

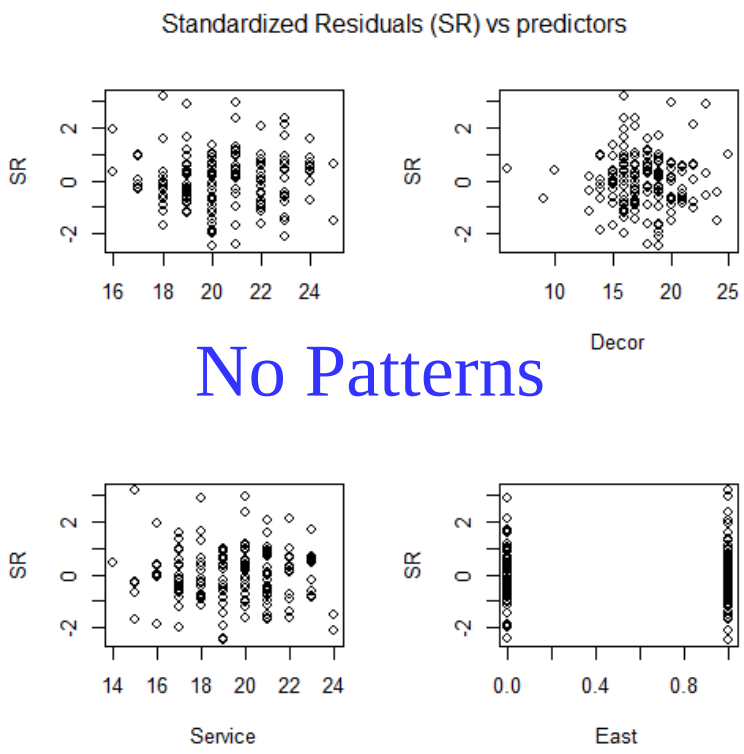
Multiple Linear Regression: Diagnostics

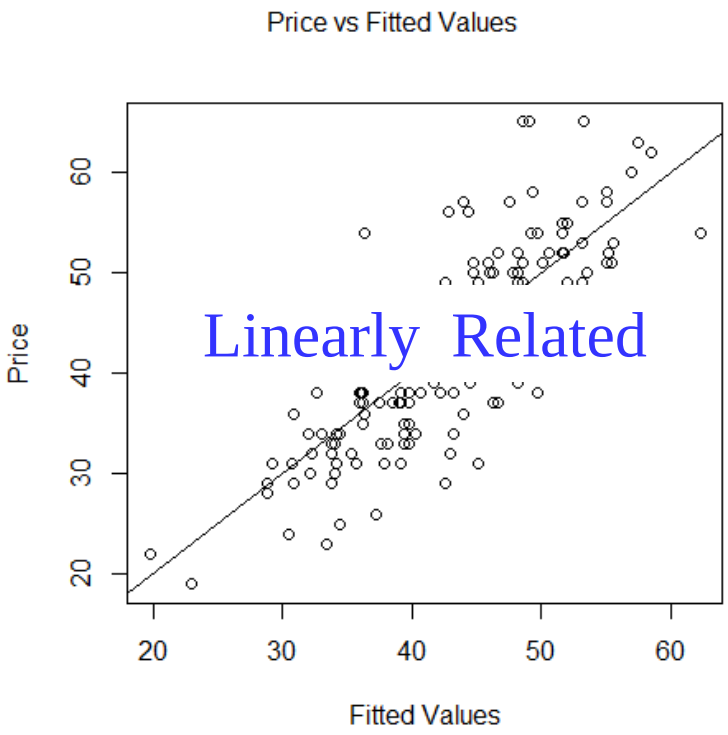
- Residual plots: Standardized residuals for assessing
  - Linear vs nonlinear model
  - Normality of the errors
  - Homoscedastic vs heteroscedastic errors
  - Correlated vs uncorrelated errors
  - Collinearity among variables

Similar to Simple regression

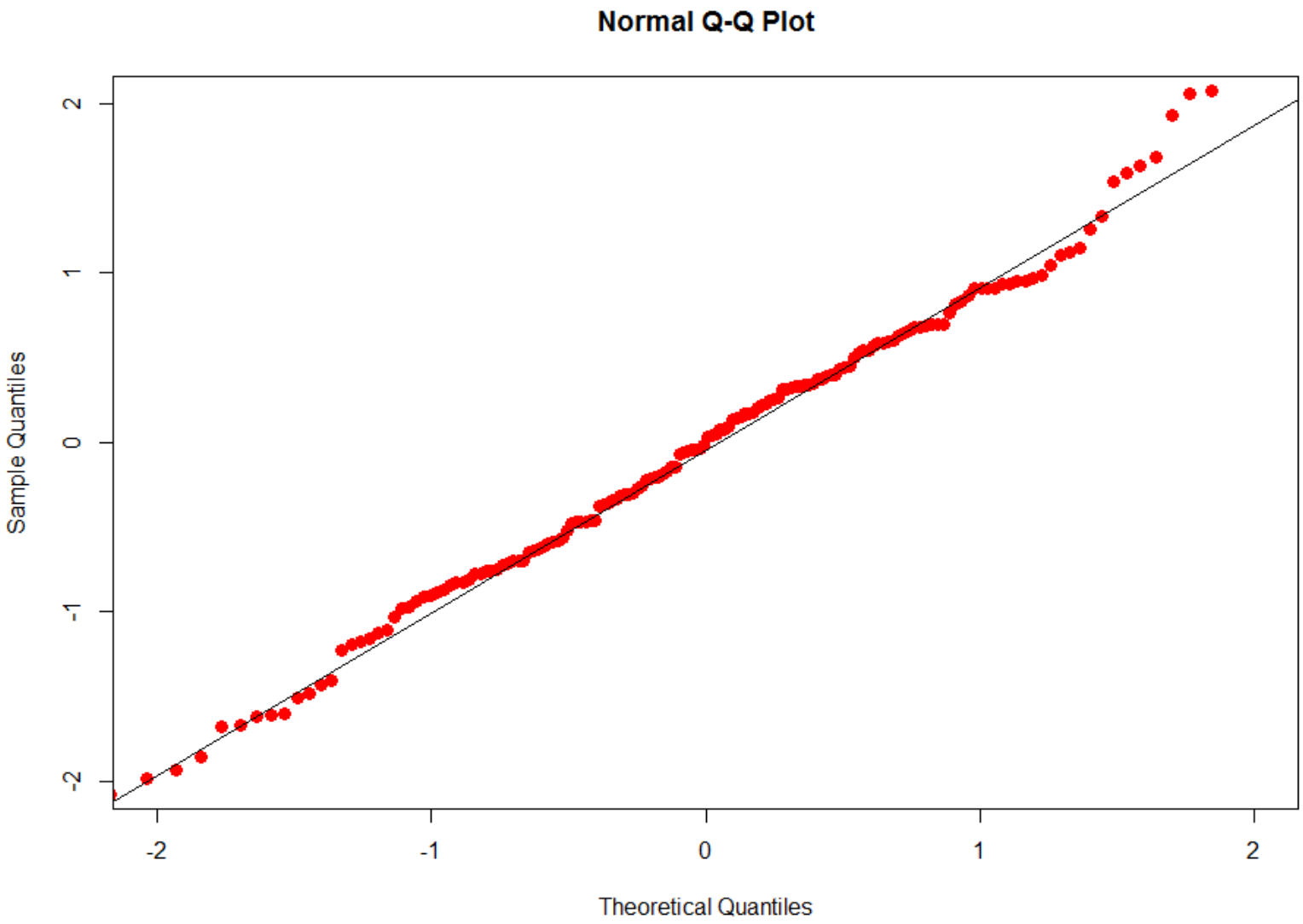
Multiple Linear Regression: Testing for linearity

- Linear vs nonlinear model: Data vs Fitted values

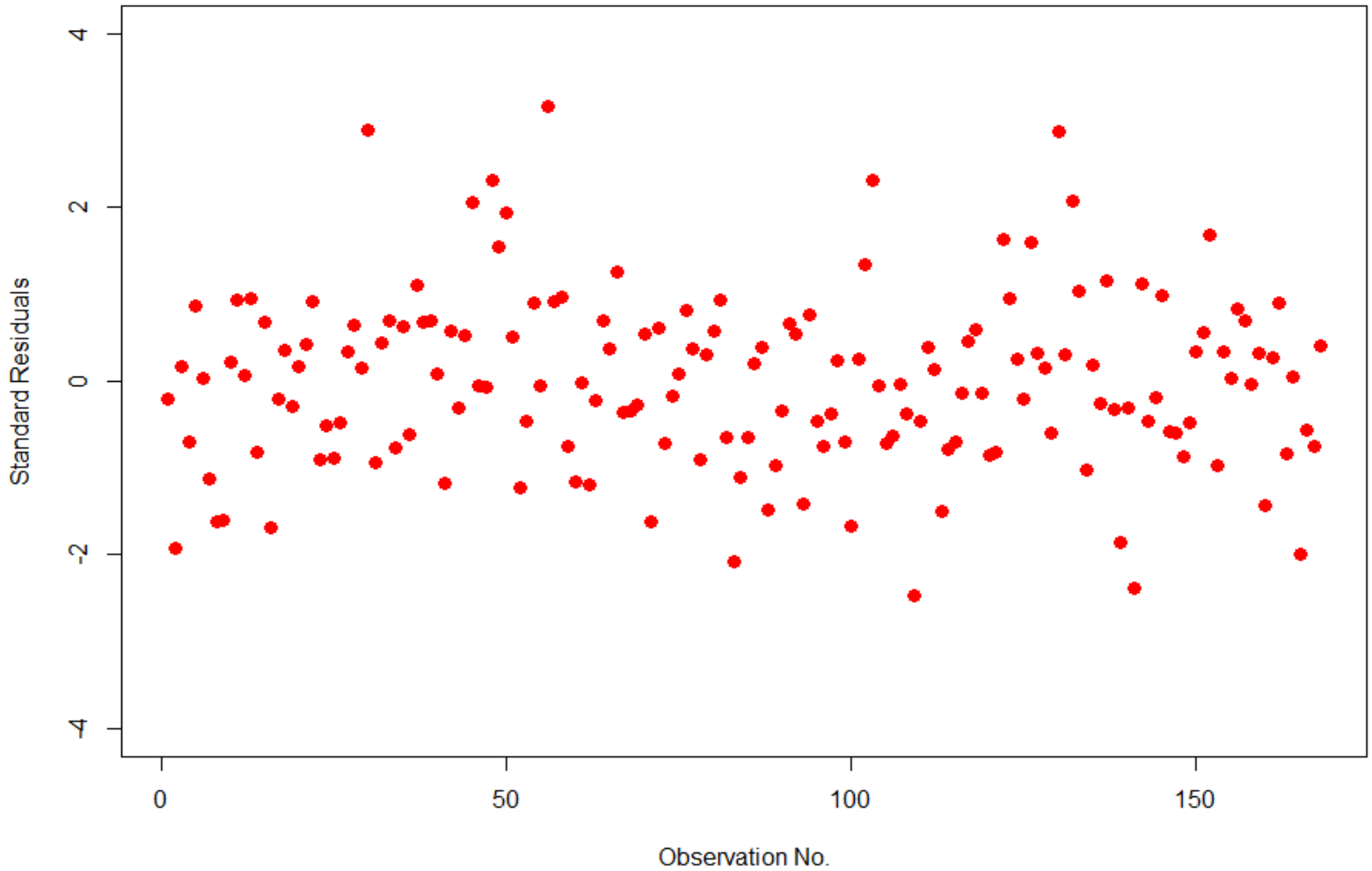




Multiple Linear Regression: Testing for Normality



Multiple Linear Regression: Testing for Outliers



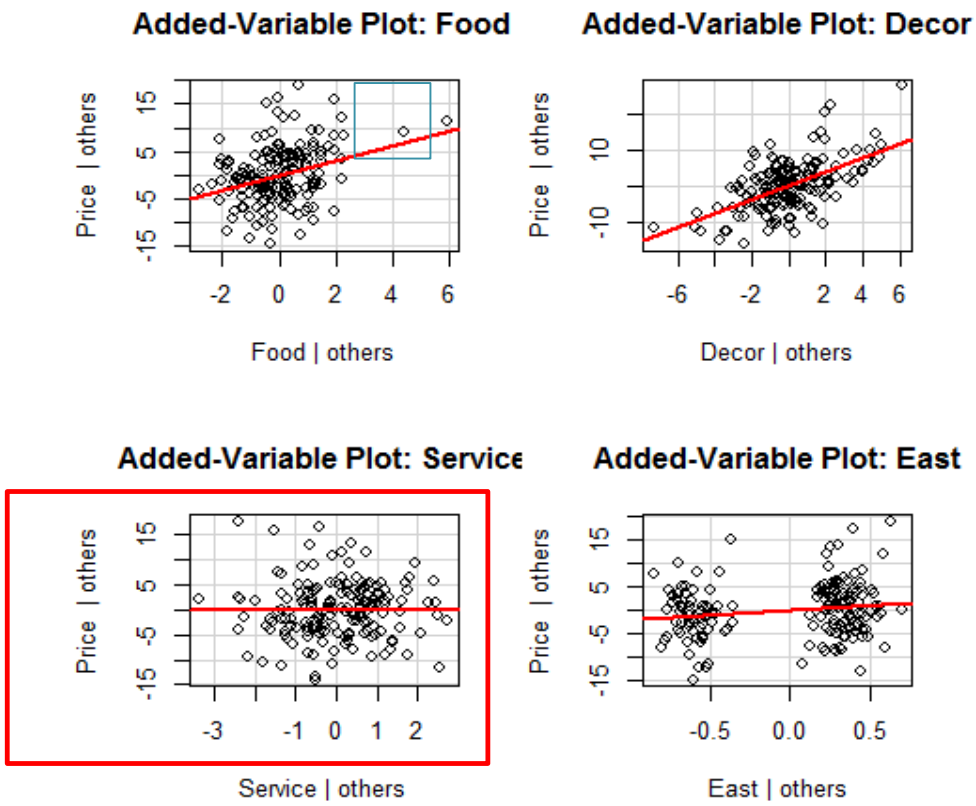
Multiple Linear Regression: Added Variable Plots

- Added variable plots: Plot of x_j *residuals vs y residuals* for each predictor x_j
- *y residuals:* Residual obtained by regressing y with all predictors except x_j
- *x residuals:* Residual obtained by regressing x_j with other predictors
- Linear regression between y residuals and x_j residuals gives $\hat{\beta}_j$
- Importance of predictor variables can be visualized using added variable plots

Multiple Linear Regression: Added Variable Plots

- Added variable plots

  Assessing the effect of each predictor after adjusting the contribution of the other predictors



Multiple Linear Regression: Diagnostics

- Collinearity among variables: Simulated data
- True model with $x_4 \cong 2x_1$

$$y_i = 0.23x_1 + 1.27x_2 + 0.57x_3 + 1.68x_4 + 20$$

- Fitted model using R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.43208    1.60810  13.949   <2e-16 ***
x1           1.25514    3.76156   0.334    0.739
x2           1.26962    0.01860  68.275   <2e-16 ***
x3           0.56911    0.02032  28.012   <2e-16 ***
x4           1.13155    1.88155   0.601    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.951 on 95 degrees of freedom
Multiple R-squared:  0.994, Adjusted R-squared:  0.9938
F-statistic:  3941 on 4 and 95 DF,  p-value: < 2.2e-16
```

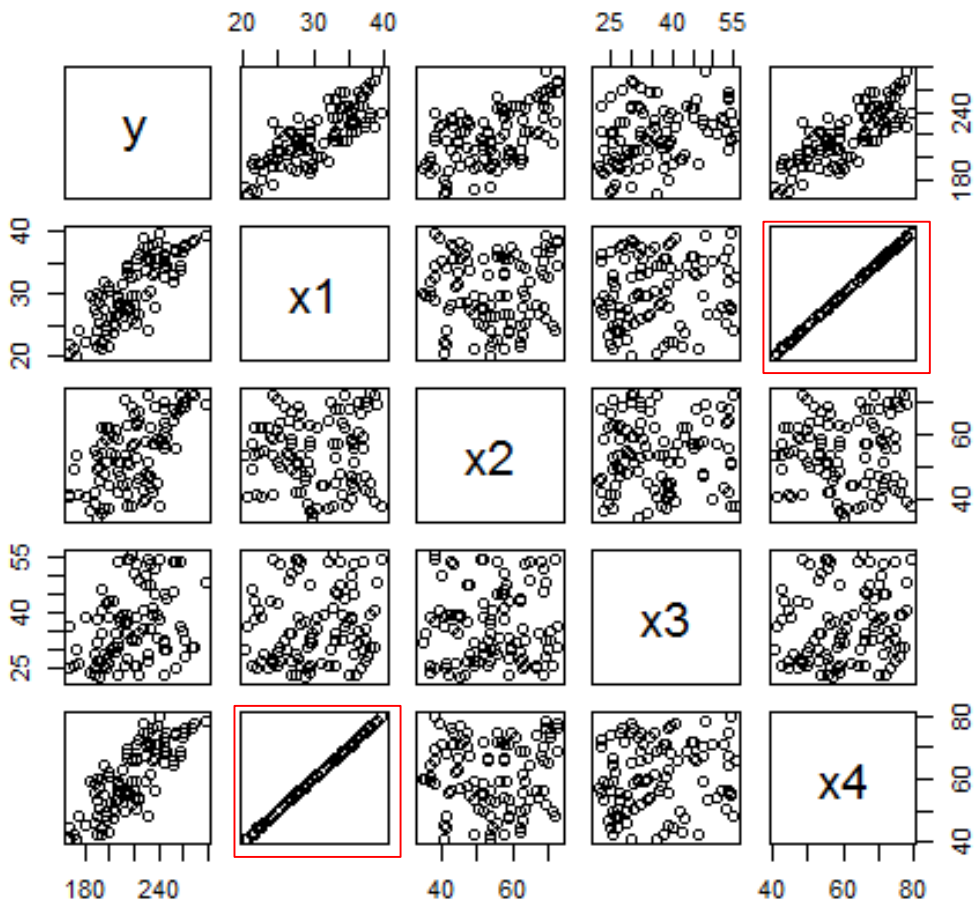
Multiple Linear Regression: Diagnostics

- Collinearity among variables: Simulated data
- True n

$$y_i = ($$

- Fitted

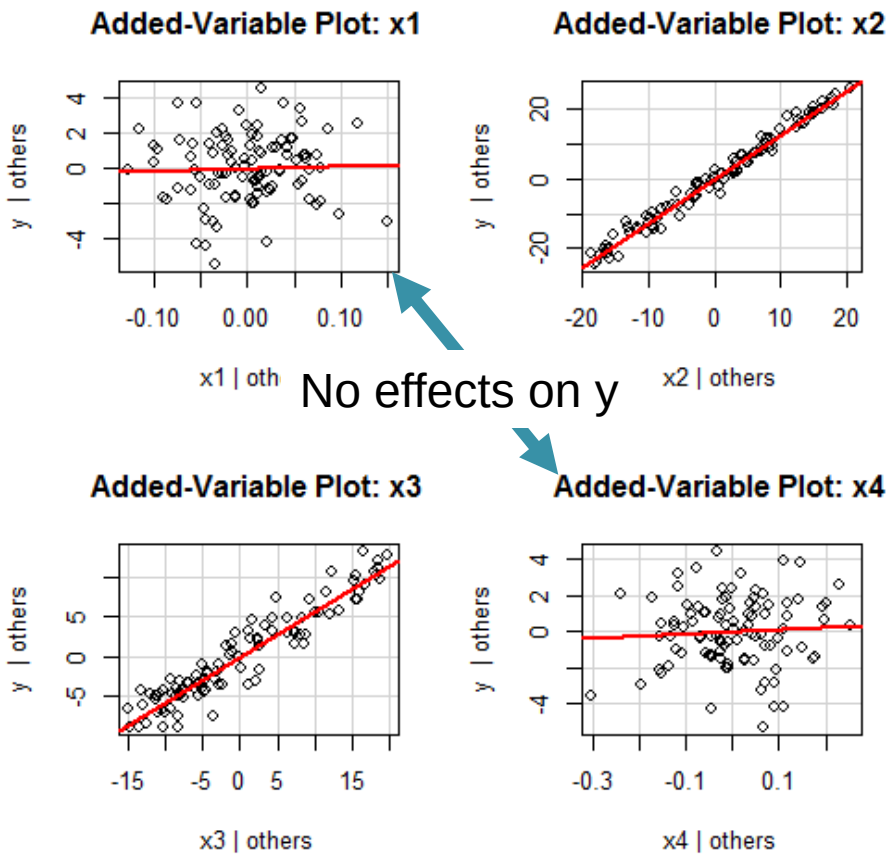
```
Coeffi

(Inter
x1
x2
x3
x4
---
Signif

Residu
Multip
F-stat
```



)

Multiple Linear Regression: Diagnostics

- Collinearity among variables: Simulated data
- Added variable plots



TRANSFORMATIONS: Nonlinear to linear model

OLS: Extensions

- Handling of the following situations:
  - Converting nonlinear models to linear models
  - Weighted least-squares to deal with Heteroscedastic errors

OLS: Transforming nonlinear model to linear model

- Nonlinear relationship known from the physics of the system
  - Bacteria deaths due to x-ray radiation

$$n_t = n_0 e^{\beta_1 t}, \quad t \geq 0$$

where n_t the number of surviving bacteria after exposure time t

- Transformed model by taking logarithms

$$\begin{aligned} \ln n_t &= \ln n_0 + \beta_1 t \\ y &= \beta_0 + \beta_1 t \end{aligned}$$

- The relationship in transformed variables is linear

OLS: Transforming nonlinear model to linear model

- Nonlinear relationship known from the physics of the system
  - Boiling temperature (T_b) and pressure (P) relationship

$$ln(P) = A + \frac{B}{T_b}$$

- Transformed model by defining $y=ln(P)$ and $x = 1/T_b$

$$y = A + B\,x$$

- The relationship in transformed variables is linear

OLS: Transforming nonlinear model to linear model

- Nonlinear relationship known from the physics of the system
  - Boiling temperature (T_b) and pressure (P) relationship:

$$ln(P) = A + \frac{B}{T_b + C}$$

where A, B, and C are unknown parameters

- Transformation to linear model is not always possible

Transformation of Variables: Nonlinearity

- Nonlinear relationship is unknown and has to be derived from the underlying data
- Transformation of
  - Only the response variable
  - Only the predictors or a subset of predictors
  - Both the response and predictor variables

Transformation of Variables: Nonlinearity

- Transformation of
  - Only the response variable
- True regression model between *y* and *x*

$$y = g(\beta_0 + \beta_1 x + \epsilon)$$

- Transforming *y* by g^{-1} (if the inverse exists),

$$g^{-1}(y) = \beta_0 + \beta_1 x + \epsilon$$

- *g(.)* is generally unknown function

Finding g(.)?

Nonlinear transformation of response variable

- Transformation of
  - Only the response variable
- Regression model between *y* and *x*

$$y = g(\beta_0 + \beta_1 x + \epsilon)$$

- Transforming *y* by g^{-1} (if the inverse exists),

$$g^{-1}(y) = \beta_0 + \beta_1 x + \epsilon$$

- *g(.)* is generally unknown function

Determine nonlinear transformation function of response variable

- Two approaches
  - Inverse response plot
  - Choosing a power transformation
- Inverse response plot

  - Fit a linear model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - Inverse response plot: Scatter plot of $\hat{y}_i$ vs y_i
  - Fit a nonlinear model between $\hat{y}_i$ vs y_i

Determine nonlinear transformation function of response variable

- Choosing a power transformation for strictly positive y s

$$\Psi(y,\lambda) = \begin{array}{ll} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln y & \text{if } \lambda = 0 \end{array}$$

- Fit the model of form

$$E(\hat{y}|y) = \alpha_0 + \alpha_1 \, \Psi_s(y,\lambda)$$

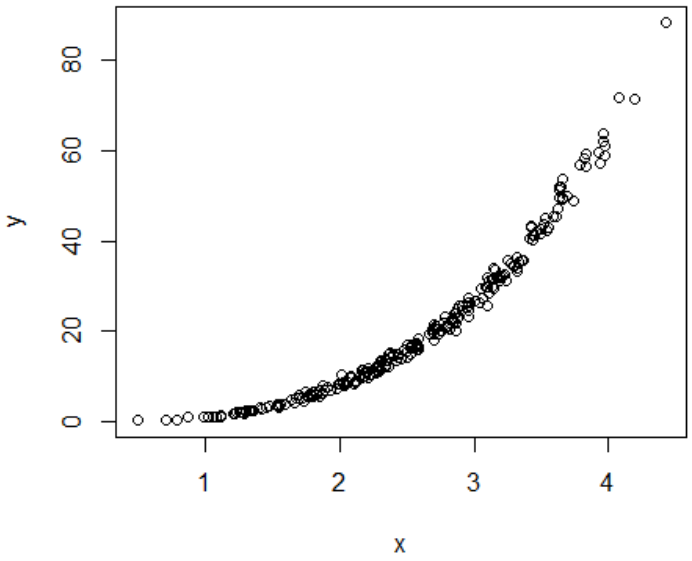
- Note that for a given value of λ, the model is linear in parameter α_0 and α_1
- SSE(λ) vs λ
- Find the value of λ that minimize SSE

Transformation of Variables: Nonlinearity

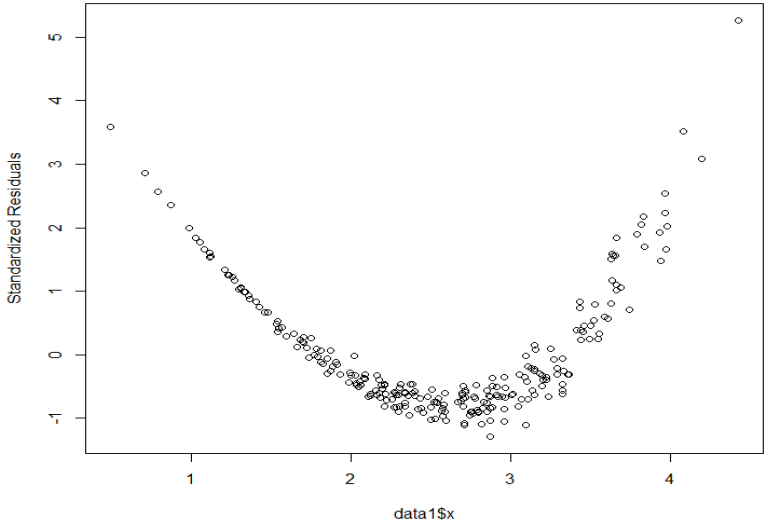
- Example: 250 data points and the model

$$y = (\beta_0 + \beta_1 x + \epsilon)^3$$

Scattered plot



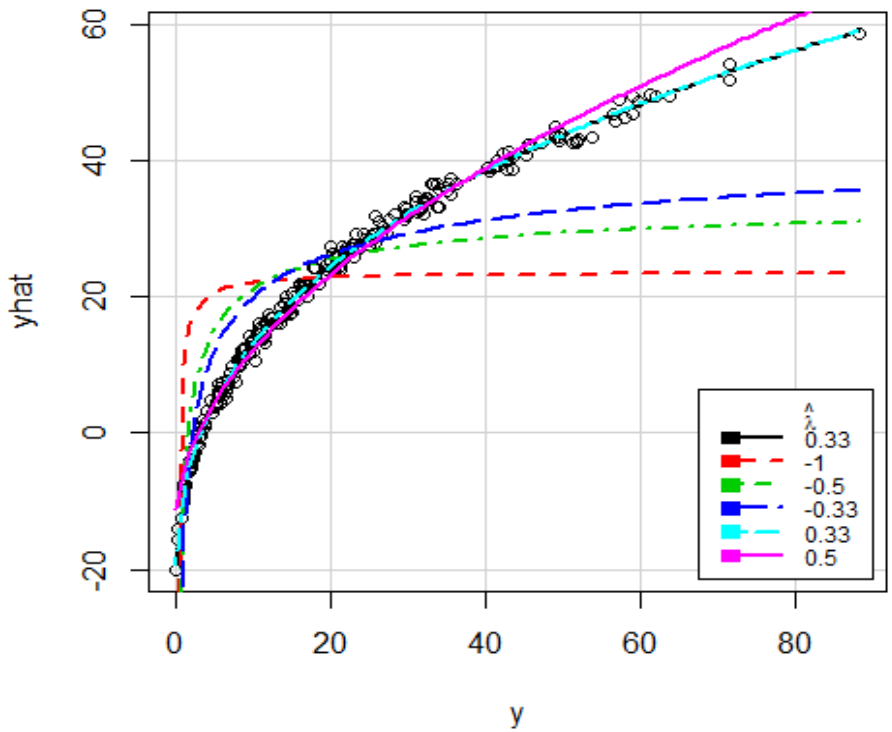
SD vs *x*



Transformation of Variables: Nonlinearity

- Example: 250 data points and the model

$$y = (\beta_0 + \beta_1\, x + \epsilon)^3$$



	lambda	RSS
1	0.3321126	265.8749
2	-1.0000000	46673.8798
3	-0.5000000	24090.7391
4	-0.3333333	15424.6329
5	0.3333333	265.9111
6	0.5000000	880.1573

$g^{-1}(.) = 0.332$

Weighted least squares for heteroscedastic errors

- Weighted least squares (WLS): $Var(\varepsilon_i) = \sigma_i^2$
- WLS estimates of the regression coefficients by minimizing SSE

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{1,i} - \ldots \beta_1 x_{p,i})^2$$

where $w_i = 1/\sigma_i^2$

- If $Var(\varepsilon_i) = \sigma^2$, leads to OLS
- If $w_i = 0$ for some i, then WLS will exclude the ith observation from the estimation process

Weighted least squares for heteroscedastic errors: Solution

- WLS solution for univariate case $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2}, \qquad \hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w$$

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad \bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

- WLS solution for multilinear case for

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

where $\mathbf{W}$=diag([w_1,w_2,…,w_p])

Transformation of Variables: Varying Variance

Crews vs Rooms data

	Case	Crews	Rooms
[1,]	1	16	51
[2,]	2	10	37
[3,]	3	12	37
[4,]	4	16	46
[5,]	5	16	45
[6,]	6	4	11
[7,]	7	2	6
[8,]	8	4	19
[9,]	9	6	29
[10,]	10	2	14
[16,]	16	2	10
[17,]	17	6	19
[18,]	18	10	33
[19,]	19	16	46
[20,]	20	16	69
[21,]	21	10	41
[22,]	22	6	19
[23,]	23	2	6
[24,]	24	6	27
[25,]	25	10	35
[26,]	26	12	55
[27,]	27	4	15

Repetitions:
y measurements
for given *x*

- *Errors*~*Normal*$(0, w_i)$, $y_i \sim Normal(\beta_0 + \beta_1 x_i, w_i)$, $i=1,\ldots 7$
- Computation of $w_i = 1/sd(y_i)^2$
- Apply WLS formulation

```
Call:
lm(formula = Rooms ~ Crews, weights = 1/StdDev^2)

Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-1.43184 -0.82013  0.03909  0.69029  2.01030 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.8095     1.1158   0.725    0.471    
Crews         3.8255     0.1788  21.400   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9648 on 51 degrees of freedom
Multiple R-squared:  0.8998, Adjusted R-squared:  0.8978 
F-statistic:   458 on 1 and 51 DF,  p-value: < 2.2e-16

               2.5 % 97.5 %
   (Intercept) -1.431  3.050
   Crews        3.467  4.184
```

Model Selection and Assessment

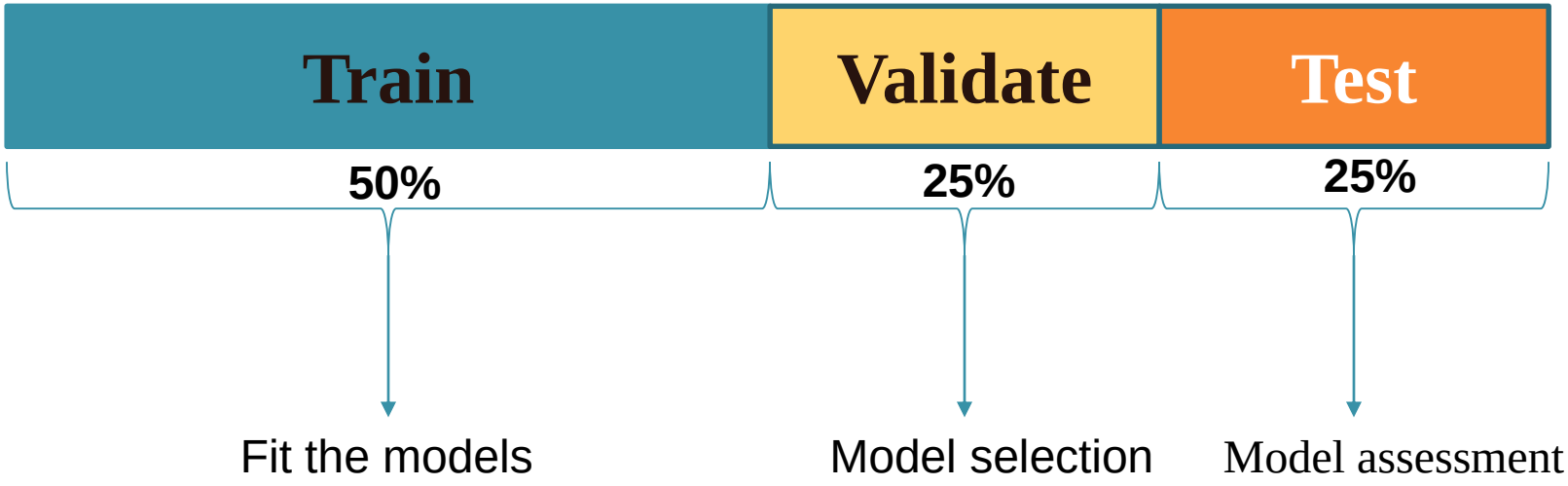
Introduction

- Development of model
  - **Building a model**(s): Verifying all the assumptions
  - Validation of model: Predictive ability of the model
  - Testing the model on new data
- Two goals: Validation of model
  - **Model selection:** Comparing the performance of several models to find the best one
  - **Model assessment:** Assessing the predictive ability of the chosen final model on new data

Introduction

- Model selection is important in multiple linear and nonlinear models
- Data-rich situation: Randomly divide the data in three parts

Ideal Scenario: Data-rich situation

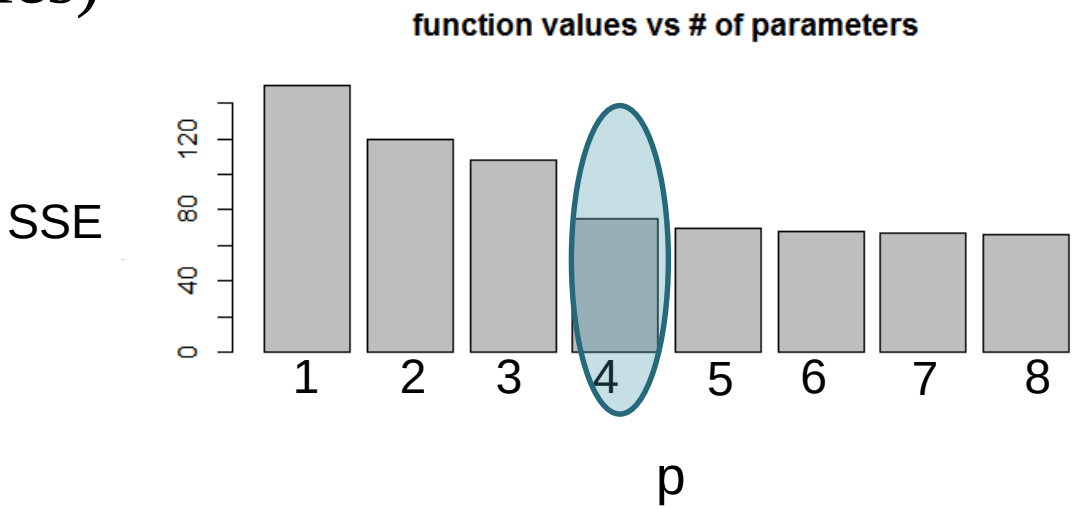


Practice: Limited Amount of Data

Best Model in Practice? Need a Criterion

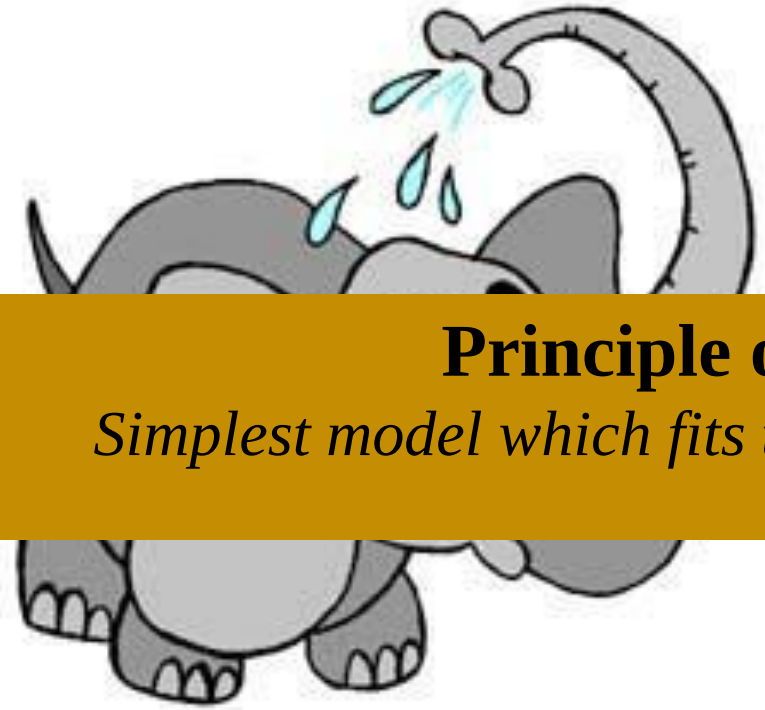
Introduction: Model Complexity

- SSE: $S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
- *SSE* values decreases with increase in parameters (or variables)



- Over fitting as *p* increases
- Under fitting not including parameters
- How to determine # of parameters?

Introduction: Model Complexity



Principle of Parsimony:

Simplest model which fits the data well should be chosen

"Give me three parameters, I can fit an elephant
Give me five and I can include his tail!"
- Fogler and Gurmen (ECRE, 2006)

Introduction: Model Complexity

- Trade off: *SSE* values and model parameters

- Model selection criterion?

  SSE values + model complexity

- SSE values: Assess the quality of the model

- Model complexity: Principle of Parsimony

  penalize complex models = number of parameters increases

Introduction: Model Complexity

- Multiple linear regression: Revisiting

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p,1} & x_{p+1,1} & \cdots & x_{q,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p,2} & x_{p+1,2} & \cdots & x_{q,2} \\ \vdots & \vdots & \vdots & \cdots & \vdots & & & \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p,n} & x_{p+1,n} & \cdots & x_{q,n} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_p & \beta_{p+1} \cdots & \beta_q \end{bmatrix}^T,$$

- The linear model in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\epsilon}) = \mathbf{0},\ Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

- Solution

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

Introduction: Model Complexity

- The full model can be partitioned into *p* variables and *r*=*q*-*p* variables

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\epsilon}$$

- The least square estimates

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\boldsymbol{\beta}}_p^* \\ \hat{\boldsymbol{\beta}}_r^* \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Introduction: Model Complexity

- Multiple linear regression: *p (<q) subset of variables*

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

- The linear model in matrix form

$$\mathbf{y} = \mathbf{X}_p \boldsymbol{\beta}_p + \boldsymbol{\epsilon}$$

- Solution

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

Introduction: Model Complexity

- Expected value and variance of $\hat{\beta}_p$

$$E[\hat{\boldsymbol{\beta}}_p] = \boldsymbol{\beta}_p + \mathbf{A}\boldsymbol{\beta}_r, \quad \mathbf{A} = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{X}_r$$

$$Var[\hat{\boldsymbol{\beta}}_p] = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\sigma^2$$

- Properties of $\hat{\beta}_p$ and $\hat{\beta}_p^*$
  - $\hat{\beta}_p$: Biased Estimates of β_p unless β_r=0 or $\mathbf{X}_p{}^T\mathbf{X}_r = 0$
  - $Var[\hat{\boldsymbol{\beta}}^*] > Var[\hat{\boldsymbol{\beta}}_p]$
  - $\hat{\sigma}_p^2$: Biased estimate of σ^2

Introduction: Model Complexity

- Effect on model miss specification on prediction?
- Prediction of $\hat{y}^*$ corresponding to $\mathbf{x}=[\mathbf{x_p} : \mathbf{x_r}]^T$
  - $\hat{y}^* = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}^*$

$$E[\hat{y}^*] = \mathbf{x}^T\boldsymbol{\beta}$$

$$Var[\hat{y}^*] = \sigma^2(1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x})$$

  - $\hat{y} = \boldsymbol{x}_p^T\widehat{\boldsymbol{\beta}}_p$ with

$$E[\hat{y}] = \mathbf{x}_p^T\,\boldsymbol{\beta}_p + \mathbf{x}_p^T\,\mathbf{A}\boldsymbol{\beta}_r$$

$$Var[\hat{y}] = \sigma^2(1 + \mathbf{x}_p^T(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{x}_p)$$

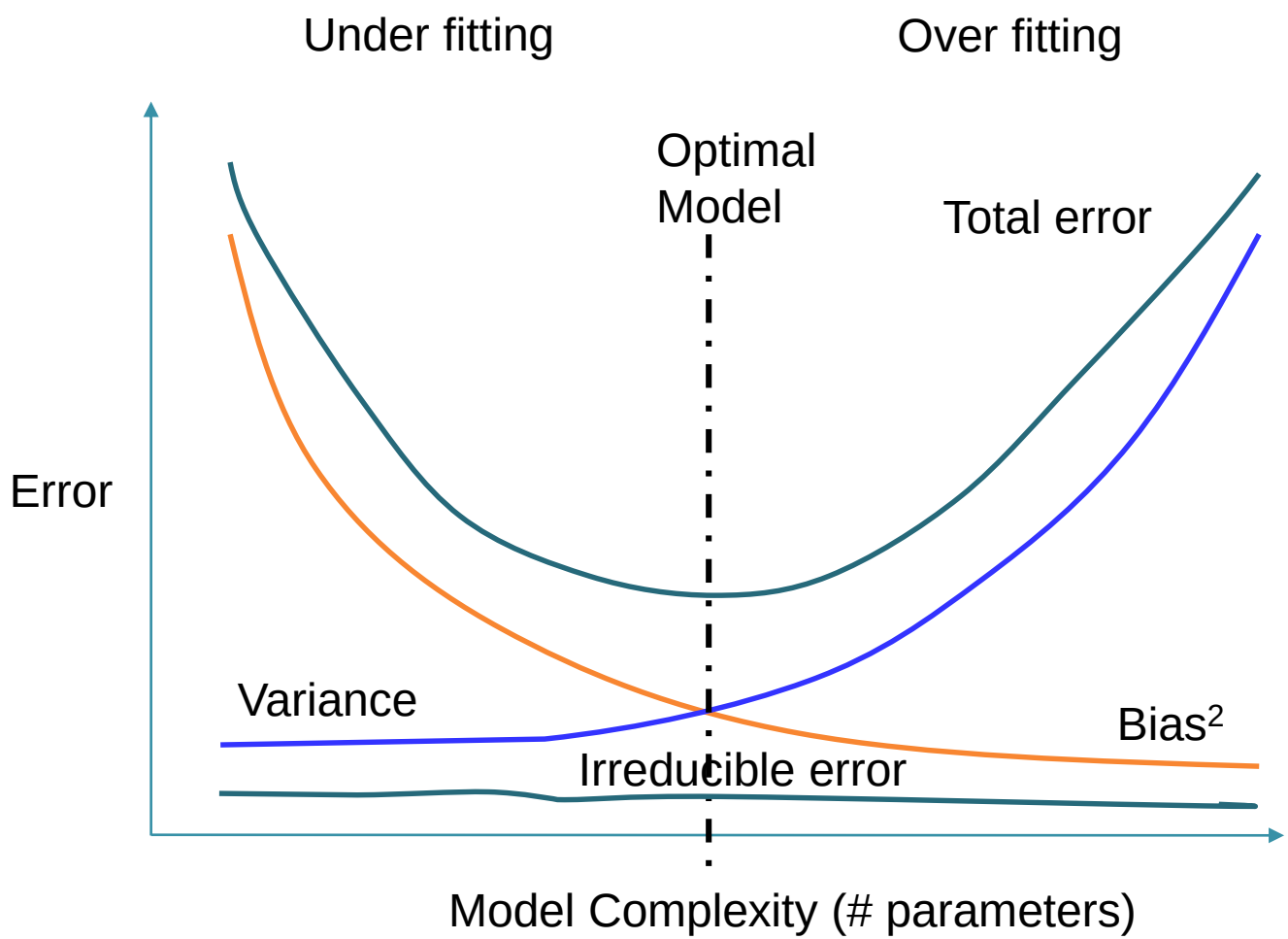
Introduction: Model Complexity

- Properties of $\hat{y}^* = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}^*$ and $\hat{y} = \boldsymbol{x}_p^T\widehat{\boldsymbol{\beta}}_p$
  - $\hat{y}$ is biased unless $\mathbf{X}_p^T\mathbf{X}_r\boldsymbol{\beta}_r = 0$
  - $Var[\hat{y}^*] \geq Var[\hat{y}]$

- Expected prediction error

$$E[(\hat{y} - y)^2] = \underbrace{\sigma^2}_{\text{Irreducible error}} + \underbrace{(\mathbf{x}_p^T Var[\hat{\boldsymbol{\beta}}_p]\mathbf{x}_p)}_{\text{Variance: Expected s.d. of } \boldsymbol{x}_p^T\widehat{\boldsymbol{\beta}}_p \text{ around the true mean}} + \underbrace{(\mathbf{x}_p^T\mathbf{A}\boldsymbol{\beta}_r - \mathbf{x}_r\beta_r)^2}_{\text{Bias}}$$

Bias-Variance Trade-off



Introduction

Model Validation

Analytical Methods
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Mallows Cp

Resampling Methods
- Leave-one-out cross validation
- K-fold cross validation
- Bootstrap

Subset selection
- Best subsets
- Backward elimination
- Forward Selection
- Stepwise regression

Analytical Methods: Mallow C_p

- Mallows C_p:

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + 2\mathbf{p} - n$$

where SSE : Sum of squares errors

σ^2 : An estimate of variance with the full model

p : number of parameters

- Expected value of C_p is p if there is no bias in the fitted equation
- Subsets of variables for $C_p \cong p$ is the desirable subset

Analytical Methods: AIC

- AIC defined as:

$$AIC = -2(SSE_\theta) + 2dim(\theta)$$

Penalty on
# of parameters

where

$\boldsymbol{\theta}$: Parameters

$dim(\boldsymbol{\theta})$: number of parameters

- A good model: A small value of AIC

- Applicability: Likelihood models

Analytical Methods: BIC

- BIC defined as:

$$BIC = -2\,log(L(\hat{\theta})) + log(n)\,dim(\theta)$$

Penalty on Log(L)

where Log(L) : Log-likelihood function value

$\boldsymbol{\theta}$: Parameters

$dim(\boldsymbol{\theta})$: =p, dimension of parameters

n : # of observations

- A good model: A small value of BIC
- $n > e^2$, BIC penalize model having large $dim(\boldsymbol{\theta})$
- Selects simpler models

Example 1

- Linear regression model,

$$y_i = B_0 + B_1 x_{1,i} + B_2 x_{2,i} + \text{error}_i$$

$$B_0 = -560,\ B_1 = 0.08,\ B_2 = 1.56$$

- Given data for *n* instants:

$$y,\ x_1,\ x_2,\ x_3$$

- Objective:
  - Find a relationship and parameters B_0, B_1, B_2

Example 1: Cont…

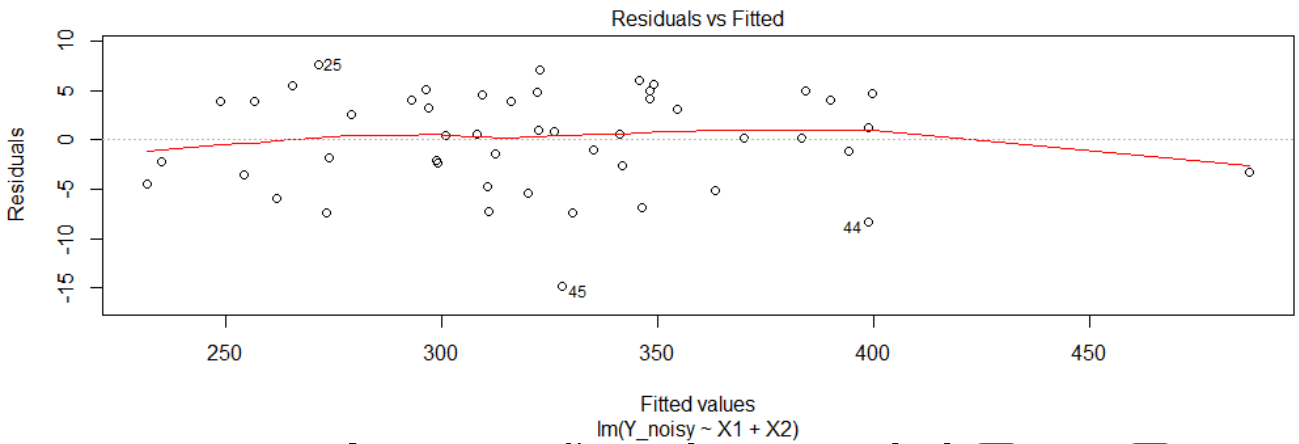
- Fit a linear model with x_1, x_2, x_3
- Fitting objective function

Models $f(x_i, \Theta)$	AIC	BIC
$B_0 + B_2x_1 + B_3x_3$	302.0091	311.4682
$B_0 + B_1x_1 + B_2x_2 + B_3x_3$	303.6010	314.9519
$B_0 + B_1x_1 + B_2x_2$	301.3854	308.9527
$B_0 + B_1x_1$	474.3469	480.0223

Minimum AIC & BIC

Example 1:Cont....

- Residual plot analysis: No patterns, linearity assumption holds



- Parameter estimates for the model $B_0 + B_1x_1+B_2x_2$

$$B_{0,e}= -543.43(-560),$$
$$B_{1,e}= 0.078(0.08),$$
$$B_{2,e}= 1.53(1.56)$$

Example 2

- Nonlinear regression model,

$$y_i = B_0 + B_1 x_{1,i} + B_2 x_{2,i} + B_3 x_{1i}^2 + \text{error}_i$$

$$B_0 = 560,\ B_1 = 3,\ B_2 = 1.56,\ B_3 = 0.08$$

- Given data:

$$y,\ x_1,\ x_2,\ x_3$$

- Objective:
  - Find a relationship and parameters

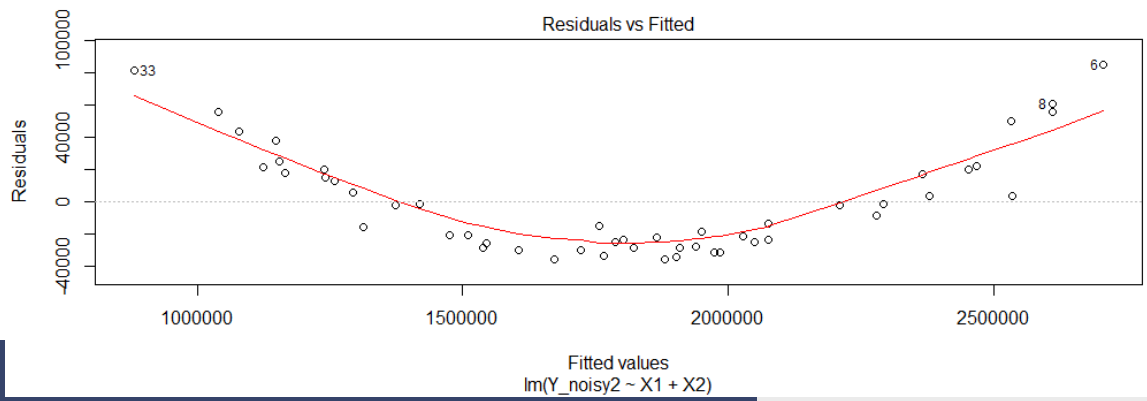
Example 2: Cont…

- Fit linear model with x_1, x_2, x_3

Models $f(x_i, \Theta)$	AIC	BIC
$B_0 + B_1x_1+B_2x_2+B_3x_3$	1188.846	1198.406
$B_0 + B_1x_1+B_3x_3$	1190.131	1197.779
$B_0 + B_1x_1+B_2x_2$	1186.853	1194.501
$B_0 + B_1x_1$	1188.158	1193.894
$B_0+B_2x_2+B_3x_3$	1433.363	1441.011

Minimum AIC & BIC

- Residual plot analysis:



Patterns : Yes
Linearity: Does not hold

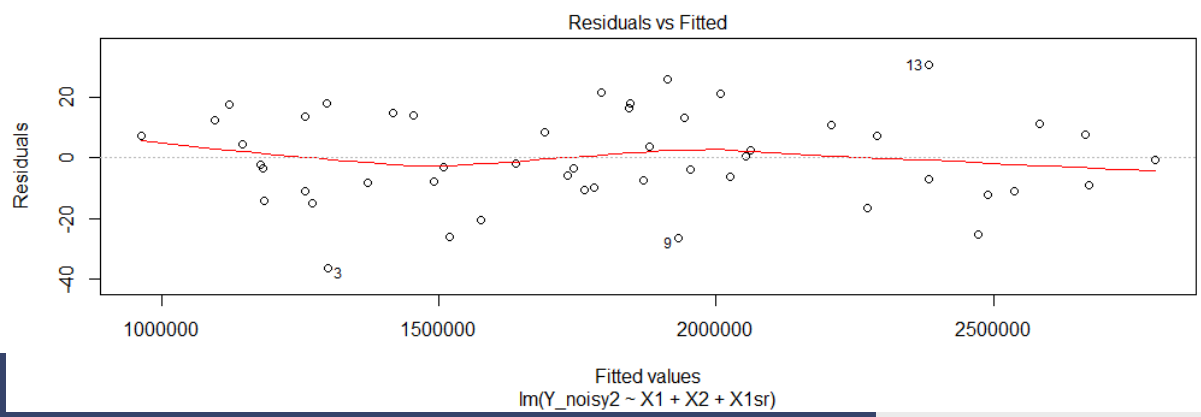
Example 2: Cont…

- Fit nonlinear model with x_1, x_2 and nonlinear terms

Models $f(x_i, \Theta)$	AIC	BIC
$B_0 + B_1x_1+B_2x_2+B_3x_1x_2$	1183.4581	1193.018
$B_0 + B_1x_1+B_2x_2+B_3x_1^2$	420.1689	429.729
$B_0 + B_1x_1+B_2x_2+B_3x_2^2$	1188.2036	1197.764
$B_0 + B_1x_1+B_2x_2+B_3x_1^2+B_4x_2^2$	422.0468	433.519

Minimum AIC & BIC

- Residual plot analysis



Parameter estimates

$B_{0,e}$= 541.43(*560*),
$B_{1,e}$= 3.02(*3*),
$B_{2,e}$= 1.49(*1.56*)
$B_{3,e}$= 0.077(*0.08*)

Subset selection methods

- *Automatically picking variables from a global model?*
- Methods: (i) Best subsets, and (ii) Stepwise
- Best subsets:
  - Builds 2^p where p: # of parameters
  - Best according to some statistics such as AIC, BIC, C_p
- Stepwise:
  - consider a smaller set of models
  - Three approach
    - Backward elimination
    - Forward selection
    - Stepwise regression

Subset selection: Backward Elimination

- Procedure to select variables
  1. Fit the full model with all p independent variables (IV) and
  2. Fit the model by removing each IV and compute AIC values
  3. Find the IV whose removal leads to minimum AIC
  4. Remove the corresponding IV
  5. Fit a new model without this variable and Go to Step 2
  6. If the removal of any variable does not reduce AIC, stop and keep current model

Subset selection: Forward Selection

- Procedure to select variables
  1. Fit all simple regression models
  2. Compute AIC
  3. Consider the IV with the lowest AIC
  4. Fit all two variable models including this variable and Go to Step 2
  5. If Current AIC > previous AIC, stop and keep previous model

Subset selection: Stepwise Regression

- Procedure to select variables
  1. Start like Backward Selection
  2. Find the IV whose removal leads to minimum AIC and remove the IV, and $AIC_{removal}$= minimum AIC for the next iteration
  3. New IV must have $AIC_{enter} \leq AIC_{removal}$ to enter
  4. Re-test all "*old variables*" that have already been entered,
  5. Old variables must have $AIC_{enter} > AIC_{\text{with IV}}$ to stay in model
  6. Continue until no new variables can be entered and no old variables need to be removed

Example 2:Cont…

Term removal(-) or addition(+)	AIC
- x_2^2	278.41
- x_2	278.71
- x_3	278.74
- x_1x_2	278.78
None	280.35
- x_1	431.86
- x_1^2	1040.28

Model: $B_0 + B_1x_1+B_2x_2 +B_3x_3+B_4x_1^2+B_5x_2^2+B_6x_1x_2$

Remove x_2^2

$AIC_{removal}$= 278.41

Term removal(-) or addition(+)	AIC
- x_1x_2	276.79
- x_3	276.93
None	278.41
+x_2^2	280.35
- x_2	281.74
- x_1	432.12
- x_1^2	1039.88

Model: $B_0 + B_1x_1+B_2x_2 +B_3x_3+B_4x_1^2+B_6x_1x_2$

Remove x_1x_2

AIC_{enter} =< $AIC_{removal}$ = 278.41

Add x_2^2 , $AIC_{with\ IV}$ > AIC_{enter}

$AIC_{removal}$= 276.79

Example 2:Cont…

Term removal(-) or addition(+)	AIC
- x_3	276.28
None	276.79
+ x_1x_2	278.41
+ x_2^2	278.78
- x_2	349.09
- x_1	493.16
- x_1^2	1044.95

Model: $B_0 + B_1x_1+B_2x_2 +B_3x_3+B_4x_1^2$

Remove x_3

$AIC_{enter} =< AIC_{removal} = 276.79$

Add x_1x_2 , $AIC_{with\ IV} > AIC_{enter}$

Add x_2^2 , $AIC_{with\ IV} > AIC_{enter}$

Term removal(-) or addition(+)	AIC
None	276.28
+ x_3	276.79
+ x_1x_2	276.93
+ x_2^2	278.15
- x_2	347.16
- x_1	491.32
- x_1^2	1042.96

$AIC_{removal}= 276.28$

Model: $B_0 + B_1x_1+B_2x_2 +B_4x_1^2$

Add x_3 , $AIC_{with\ IV} > AIC_{enter}$

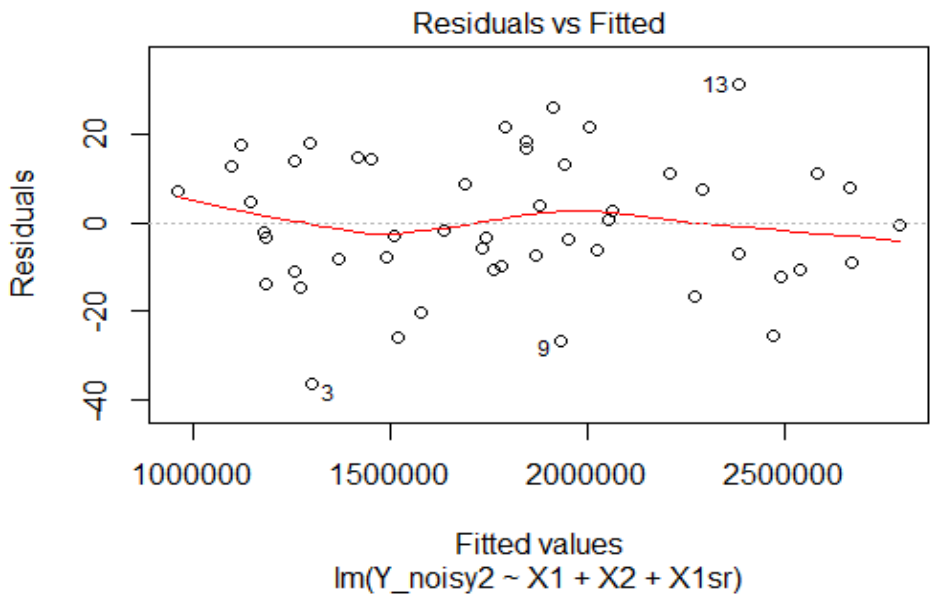
Add x_1x_2 , $AIC_{with\ IV} > AIC_{enter}$

Add x_2^2 , $AIC_{with\ IV} > AIC_{enter}$

No improvement in AIC by adding or removing IVs

Example 2: Cont…

- Identified model: $B_0 + B_1x_1 + B_2x_2 + B_4x_1^2$
- Residual plot analysis



Parameter estimates

$B_{0,e}$= 541.6(*560*),
$B_{1,e}$= 3.02(*3*),
$B_{2,e}$= 1.49(*1.56*)
$B_{3,e}$= 0.08(*0.08*)

Resampling Methods

- Validation of models by repeatedly drawing random samples from a training set
  - K-fold cross validation
  - Bootstrap
- Objective: Predict the performance of model(s) on the test sets using the training sets
- Resampling methods useful for data scarce situations

Resampling Methods

- Consider the following data set
  - Training set: $\{(\boldsymbol{x}_1, y_1);(\boldsymbol{x}_2, y_2);\ldots;(\boldsymbol{x}_n, y_n)\}$
  - Test point: $(\boldsymbol{x}_0, y_0)$ such n_t observations
- Training error rate

$$MSE_{Training} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2$$

Not of our interest for predictive ability of the model

- Test error rates

$$\boxed{MSE_{Test} = \frac{1}{n_t}\sum_{i=1}^{n}(y_{0,i} - \mathbf{x}_{0,i}^T\hat{\boldsymbol{\beta}})^2}$$

Of our interest

Data scarcity: Test data are not available

Resampling Methods

- Expected test error

$$E[(\hat{y}_0 - y_0)^2] = \underbrace{\sigma^2}_{\text{Irreducible error}} + \underbrace{(\mathbf{x_0}_p^T Var[\hat{\boldsymbol{\beta}}_p]\mathbf{x_0}_p)}_{\text{Variance}} + \underbrace{(\mathbf{x_0}_p^T \mathbf{A}\boldsymbol{\beta}_r - \mathbf{x_0}_r\beta_r)^2}_{\text{Bias}}$$

- Interpretation of variance: The amount by which $\widehat{\boldsymbol{\beta}}_p$ would change if we estimate it using different training sets
- Interpretation of Bias: The amount of error introduced by approximating a problem with a simpler model
- *Select the model that achieves low variance and low bias*

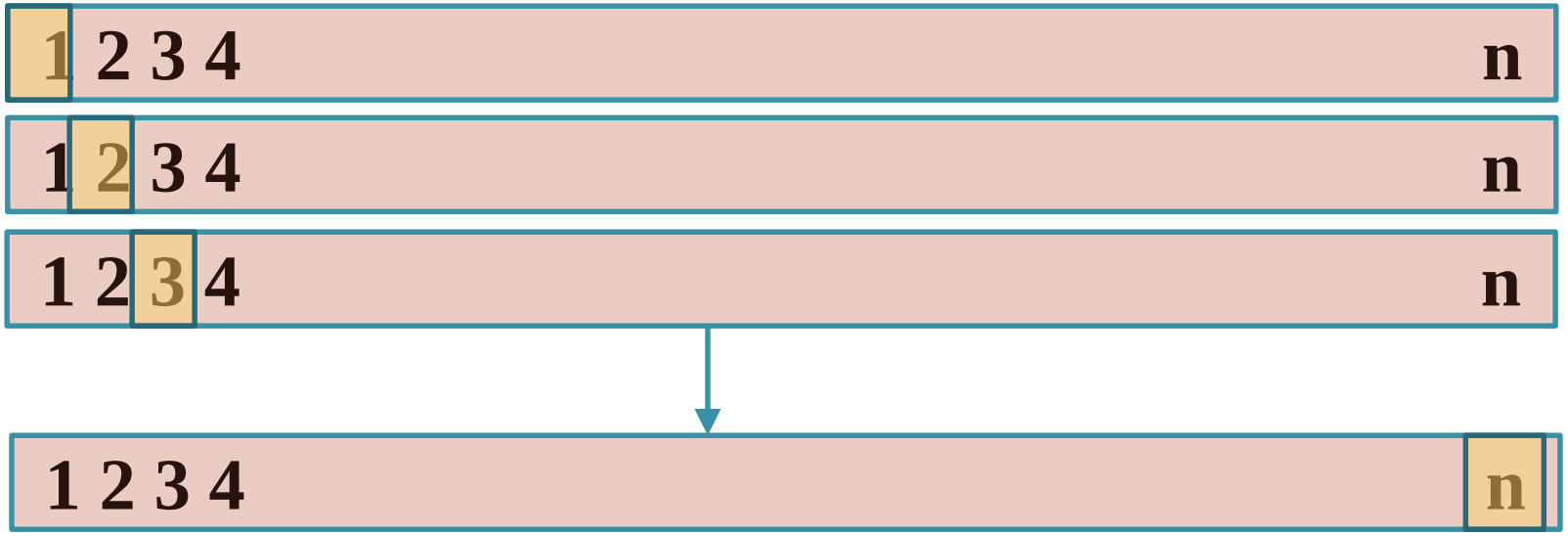
Validation Set Approach

- Enough data: (1) Training set, (2) Validation set, and (3) Test set
- Not enough data: Generate validation sets from a training set
- Validation set approach: Divides (often randomly) the training set into two parts
  - A training set
  - A validation set (or hold-out set)
- Use training set, to fit the model
- Use validation set, to predict validation set errors

  Provides an estimate of test error rates

Leave-one-out-cross-validation (LOOCV)

- Build model using *(n-1)* samples and predict the response (y_i) for *the remaining sample*



$$CV_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(1)})^2$$

LOOCV

- Leave-one-out-cross-validation (LOOCV)
- Advantages
  - Far less bias comparison to the validation set approach

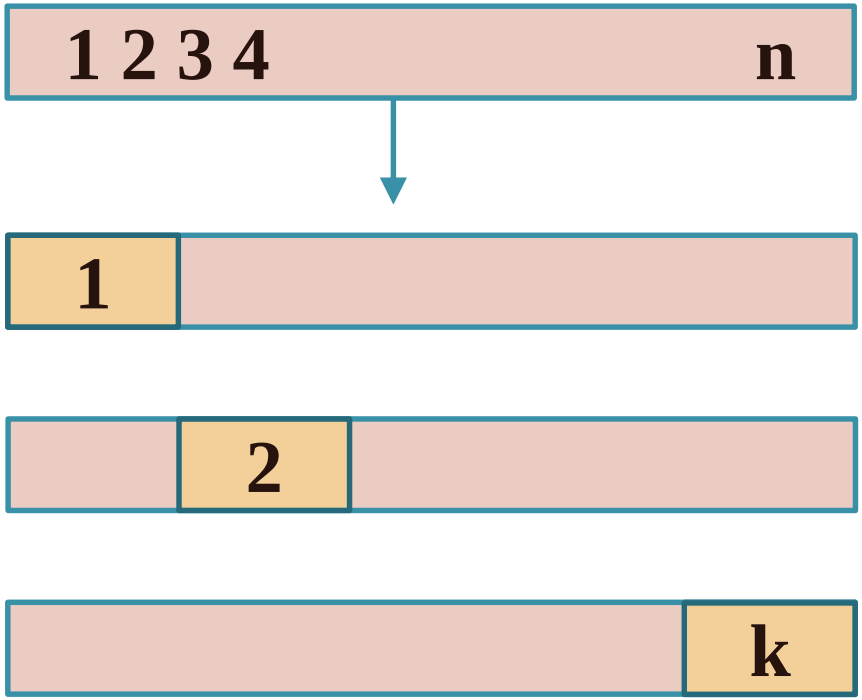
    Training set contains *(n-1)* observations each iteration
  - Yield the same results

    No randomness in the training/validation set splits
  - Does not overestimate the test error rate as much as the validation set approach
- Disadvantages
  - Expensive to implement due to fitting happens *n* times
  - Asymptotical incorrect (n tends to infinity) it does not choose correct model
  - It may select a model of excessive size (more variables) than the optimal model

k-Fold Cross Validation

- Training data into *k* disjoint samples of equal size,

  $Z_1, Z_2 \ldots, Z_k$

- For each validation sample Z_i
  - Use remaining data to fit the model
  - Predict the response for the validation sample Z_i and compute mean square error (MSE_i),
  - Repeat for all *k* samples
  - The k-fold CV

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$