Lecture 22

Part 5 Linear Regression

Simple Linear Regression

Simple Linear Regression

- Simple linear regression is used to investigate the relationship between two variables, denoted X and Y.
- X is called the **independent**, **explanatory**, **predictor** or **regressor** variable.
- Y is called the **dependent** or **response** variable.

Simple Linear Regression

- Simple linear regression can be used to:

1. Determine whether there is a linear relationship between X and Y (does the value of X have any effect on the value of Y?).
2. Determine the nature of the linear relationship between X and Y (as X changes, how does Y change?).
3. Predict the value of Y from a value of X.

Simple Linear Regression

- X and Y are both usually continuous variables.
- However, we will consider situations in multiple regression that involve categorical independent variables.
- Note: Multiple regression (more on this next topic) refers to when there is more than one independent variable, i.e., $X_1, X_2, X_3, \ldots$

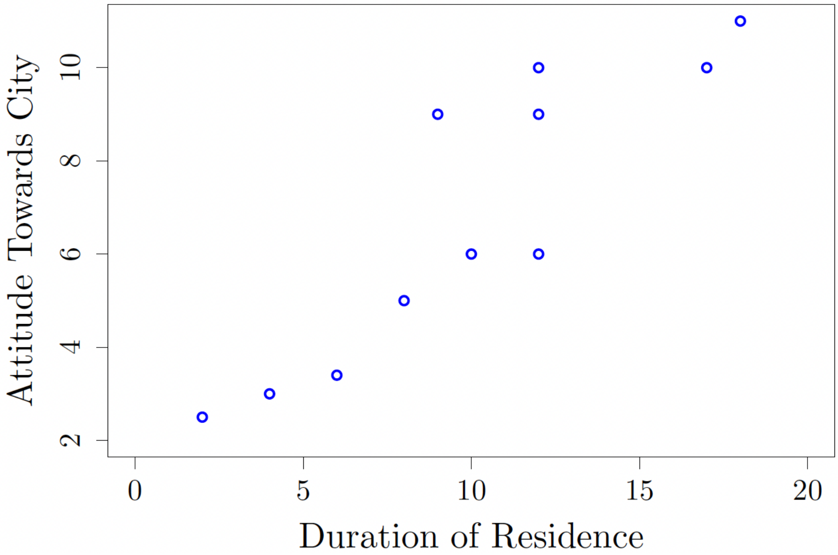
Example

- Suppose we want to use simple linear regression to see whether the time a person has lived in a city affects their attitude toward that city in a linear manner.
- Since we want to see how attitude is affected by duration of residence, we would set:
  - Duration of residence as the independent variable X.
  - Attitude toward the city as the dependent variable Y.

Example

- Let our sample data be denoted by the pairs of values $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
- Now, the first step in conducting a simple linear regression analysis is to construct a scatter plot so we can "eyeball" the data.
- Plot the independent variable X on the x-axis and dependent variable Y on the y-axis.

Scatter Plot



Simple Linear Regression Model

- The **simple linear regression model** assumes that the relationship between the dependent and independent variables is given by a straight line:

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

- β_0 is the y-intercept of the line.
- β_1 is the slope of the line.
- ϵ is called the error variable.

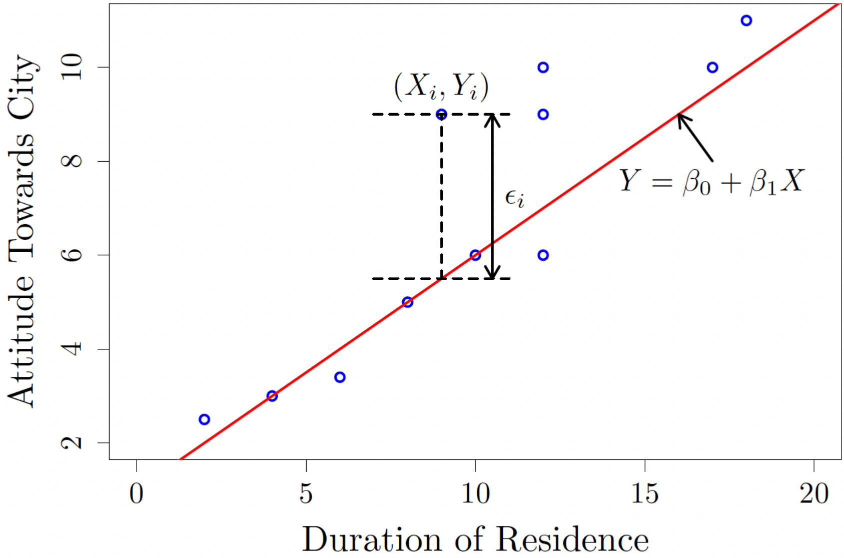
Simple Linear Regression Model

- In our city example, if we took the ith person from our sample with duration of residence X_i and attitude Y_i, then the simple linear regression model states that:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- That is, their attitude Y_i is equal to $\beta_0 + \beta_1 X_i$ plus some amount ϵ_i.
- The amount ϵ_i signifies a random error component that can be either positive, negative or zero.

Simple Linear Regression Model



Model Assumptions

- As was the case with ANOVA, the simple linear regression model makes a number of assumptions and these are stated in terms of the error variable ϵ.
- The assumptions are that the errors:
  - Are normally distributed.
  - Have mean equal to 0.
  - Have constant variance denoted by σ_ϵ^2, regardless of the value of X.
  - Are independent.
- Shorthand we write $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.
  - *iid* stands for "independently and identically distributed".

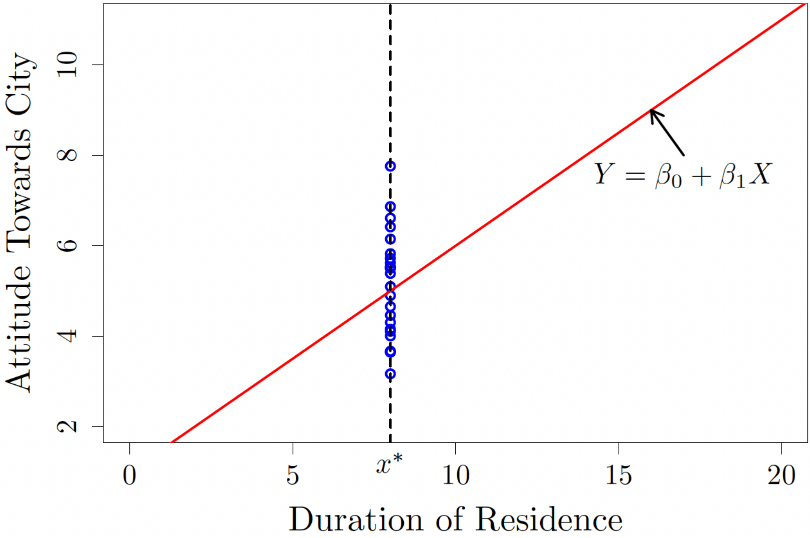
The Error Variable ϵ

- Why is there an error variable?
- Suppose you take two observations from the population with the same value of X.
- Will they have the same value of Y?
- Probably not. Why?
- Because of the inherent variability in the population.

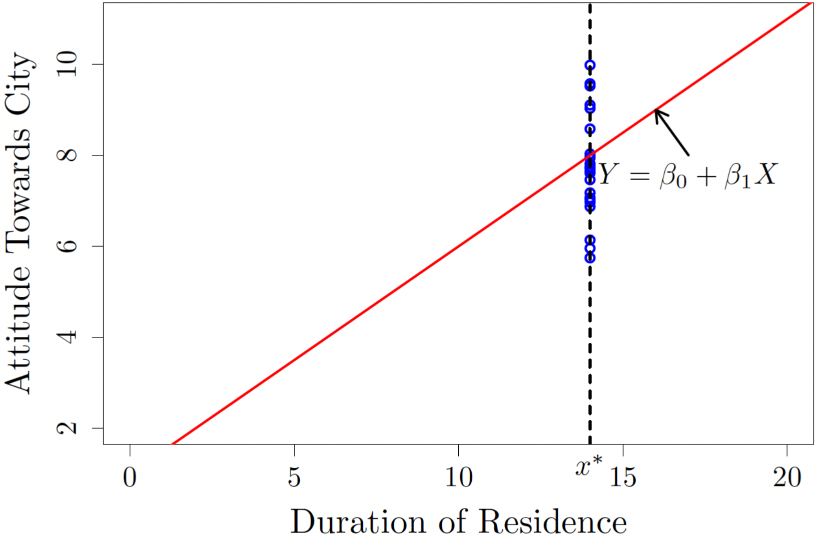
The Error Variable ϵ

- The error variable represents this inherent variability that exists in the population.
- In fact, if we look at *all* the observations in the population that have a particular value of $X = x^*$, due to this inherent variability, there will be a corresponding distribution of Y values.

The Error Variable ϵ



The Error Variable ϵ



The Error Variable ϵ

- Based on the model assumptions, the distribution of the Y values given $X = x^*$ is actually normal with mean and variance given by:

$$\begin{aligned} E(Y) &= E(\beta_0 + \beta_1 x^* + \epsilon) \\ &= \beta_0 + \beta_1 x^* + E(\epsilon) \\ &= \beta_0 + \beta_1 x^* \end{aligned}$$

$$\begin{aligned} V(Y) &= V(\beta_0 + \beta_1 x^* + \epsilon) \\ &= V(\epsilon) \\ &= \sigma_\epsilon^2 \end{aligned}$$

Simple Linear Regression Model

- Therefore, another way to state the simple linear regression model is that the dependent variable Y is normally distributed with mean equal to:

$$E(Y) = \beta_0 + \beta_1 X$$

and variance equal to:

$$V(Y) = \sigma_\epsilon^2$$

Parameter Estimation

- The intercept β_0 and slope β_1 of the simple linear regression model are unknown *population parameters*.
- Therefore, to actually *fit* the model, we need to *estimate* β_0 and β_1.
- How do we do that? By using our sample data, $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

Parameter Estimation

- In other words, based on the observations in our sample, we are trying to find the best estimate of the straight line given by our model:

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

- Suppose we have somehow obtained estimates for β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.

Parameter Estimation

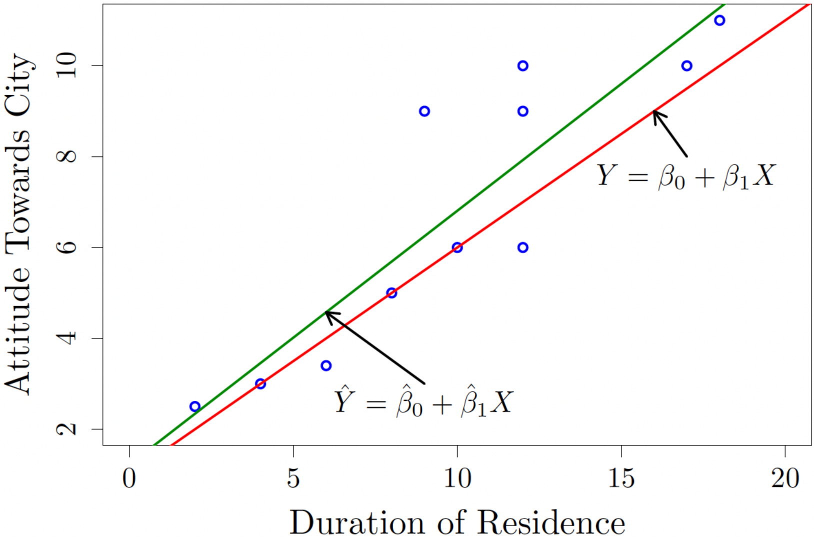
- Then our **estimated** or **fitted regression line** is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

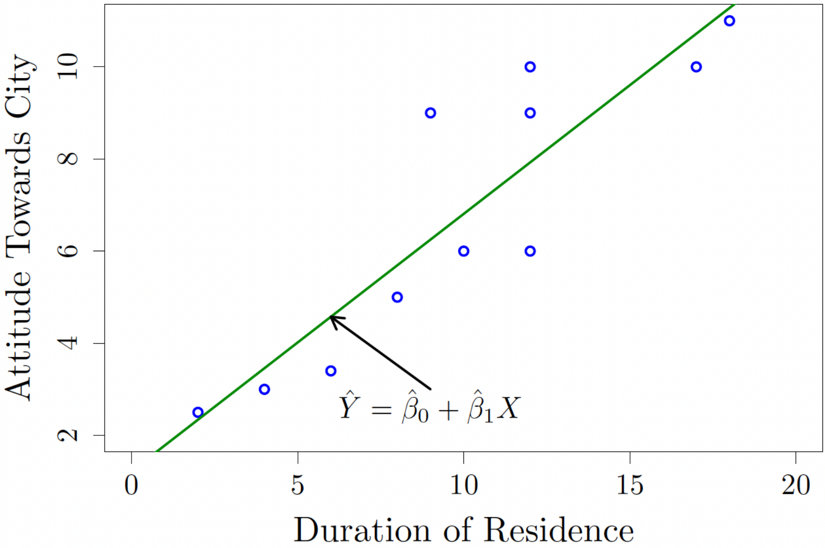
- So for each observation in our sample, we can calculate its **fitted value**:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Parameter Estimation



Parameter Estimation



Parameter Estimation

- Recall that, based on the simple linear regression model, the Y_i value for each observation in our sample can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We can do a similar thing based on...

Parameter Estimation

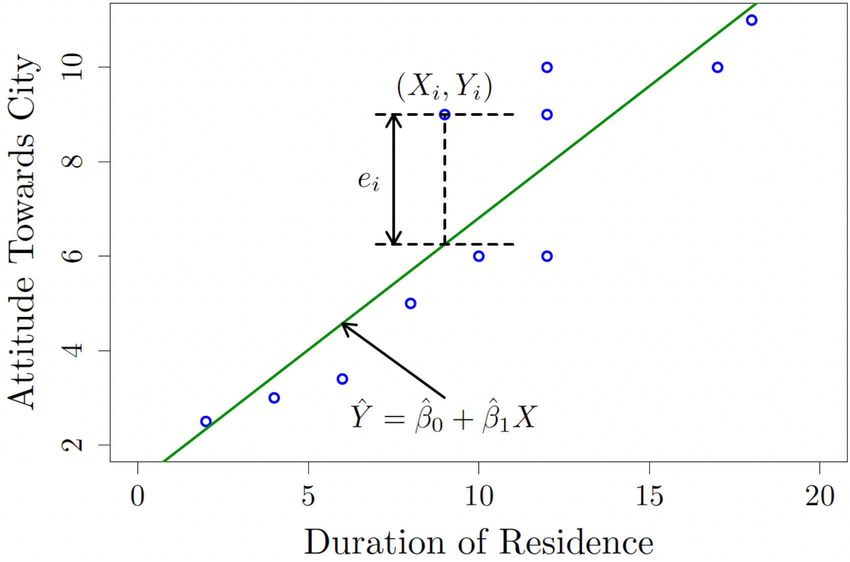
- ...our *fitted regression line*:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

- The term e_i is called the **residual** of the ith observation and is just equal to:

$$\begin{aligned} e_i &= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= Y_i - \hat{Y}_i \end{aligned}$$

Parameter Estimation



Parameter Estimation

- We would like our estimated regression line to be as close as possible to the observations in our sample.
- That is, we want the observed values Y_i to be as close as possible to the fitted values $\hat{Y}_i$.
- In other words, we want to make the residuals e_i as small as we can.

Parameter Estimation

- We estimate β_0 and β_1 using the **method of least squares**.
- That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen as the values that minimize the **sum of squared residuals**:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2$$

Parameter Estimation

- Using calculus, we can show that the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals are given by:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

where s_X^2 is the sample variance of X and s_{XY} is the sample covariance between X and Y.

Parameter Estimation

- It turns out that these estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are actually unbiased. That is:

$$E(\hat{\beta}_0) = \beta_0, \qquad E(\hat{\beta}_1) = \beta_1$$

- Although we have the formulae to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, in practice we usually use a statistical software package (R, Minitab, Excel, etc) to obtain them, rather than doing it by hand.

R Output

```
Call:
lm(formula = attitude ~ duration, data = city.dat)

Residuals:
Min       1Q  Median      3Q      Max
-1.9262 -0.7640 -0.4579  0.6165  2.7494

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.2237     1.0531    1.162 0.275114
duration       0.5585     0.0952    5.867 0.000239

Residual standard error: 1.493 on 9 degrees of freedom
Multiple R-squared: 0.7927,    Adjusted R-squared: 0.7697
F-statistic: 34.42 on 1 and 9 DF,  p-value: 0.0002387
```

Assessing the Model

- So we have defined the simple linear regression model and we know how to fit (or estimate) the model.
- The next important step is to *assess* our simple linear regression model.
- In other words, we want to determine whether or not the model is any good.

Assessing the Model

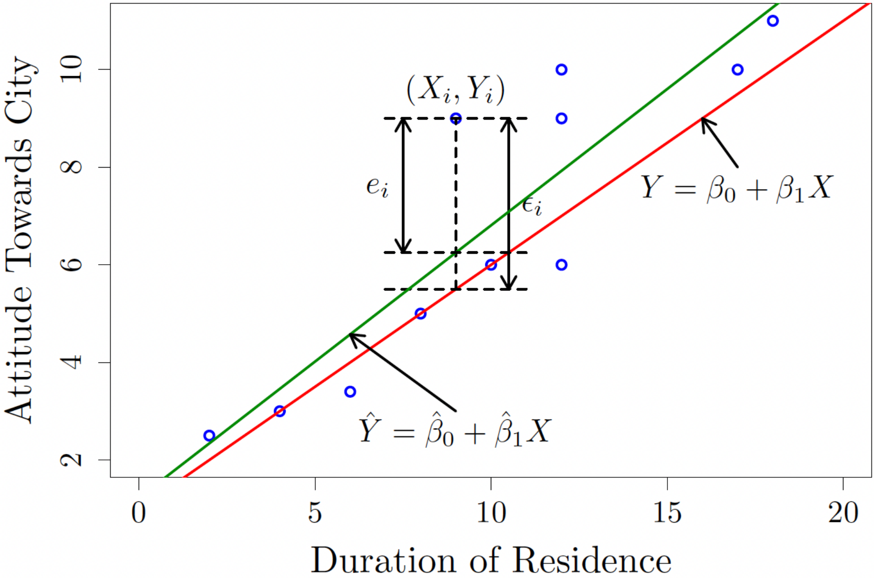
- There are a number of things we can do in order to assess our simple linear regression model:

1. Check to see if the model assumptions hold.
2. Test the overall significance of the model.
3. Estimate σ_ϵ^2, the variance of the error variable.
4. Calculate R^2, the proportion of variation in Y explained by the model.

1. Checking the Model Assumptions

- The first thing we should do after fitting the model is to check if the model assumptions hold.
- If they do not hold, it means that a simple linear regression model is not appropriate for our data.
- Recall that the model assumptions were stated in terms of the error variable, namely, $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.
- We don't know the true errors ϵ_i, but we do know the residuals $e_i = Y_i - \hat{Y}_i$, which estimate the true errors.

1. Checking the Model Assumptions



1. Checking the Model Assumptions

- So we should check to see if the residuals e_i satisfy the model assumptions:
  (a) Are they normally distributed?
  (b) Do they have mean 0 and constant variance?
  (c) Are they independent?

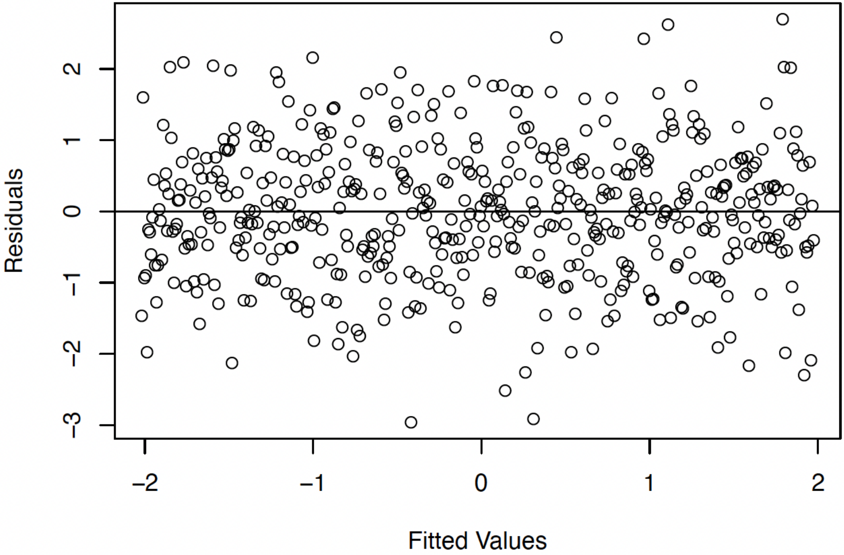
(a) To check the normality of the residuals, we can generate:
  - A histogram of the residuals, which should look like a normal distribution (bell-shaped and symmetric).
  - A normal probability plot of the residuals, which should be linear.

1. Checking the Model Assumptions

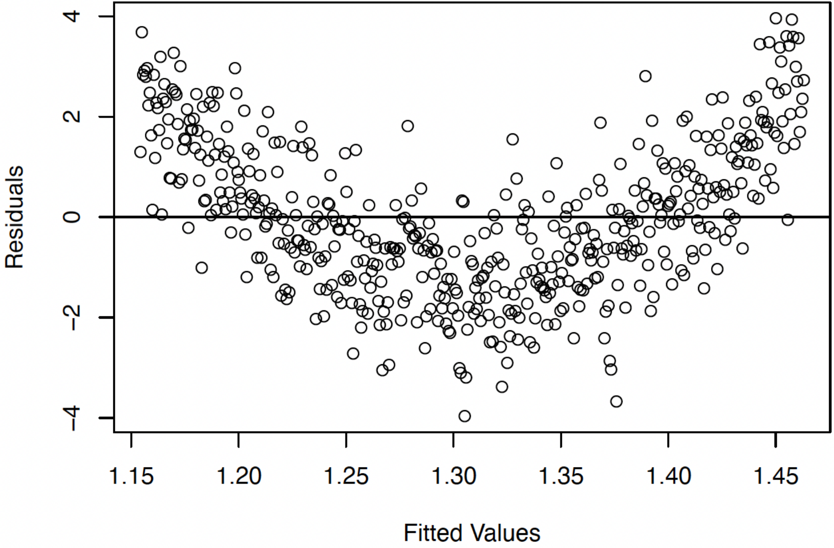
(b) To check that the residuals have mean 0 and constant variance, we can examine scatter plots of the residuals against the X values or fitted values.

- These *residual plots* should look like a random scatter of points about 0 with no obvious patterns or trends.
- If there are clear patterns or trends, we might need to transform the data.

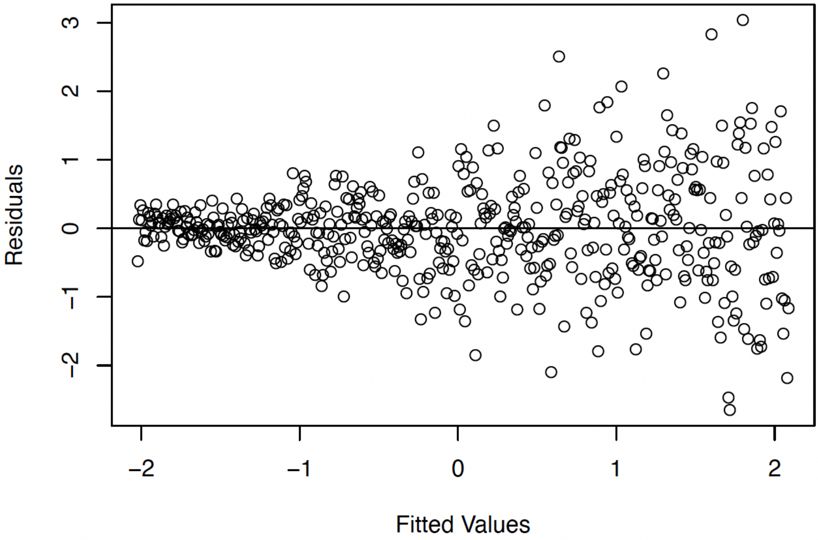
A Good Residual Plot



A Bad Residual Plot



Another Bad Residual Plot



1. Checking the Model Assumptions

(c) Certain plots can be used for checking the independence of the residuals, but if the sample is collected properly (i.e., randomly) this hopefully shouldn't be a major problem.

- Can plot the residuals against the order in which the observations were collected to see if there is any time correlation between the residuals.