Chapter 8: Simple Linear Regression

Shiwen Shen

University of South Carolina

2017 Summer

Introduction

- A problem that arises in engineering, economics, medicine, and other areas is that of investigating the relationship between two (or more) variables. In such settings, the goal is to model a continuous random variable Y as a function of one (or more) independent variables, say, $x_1, x_2, ..., x_p$. Mathematically, we can express this model as

$$Y = g(x_1, x_2, ..., x_p) + \epsilon$$

  where g is called a **regression model**.
- ϵ is the random error, which indicates that the relationship bettwen Y and $x_1, x_2, ..., x_p$ through g is not deterministic.
- ϵ is where the variability comes from.
- ϵ is the reason why regression models are treated as statistical models.

Linear Regression Model

- Let's consider the model

$$Y = g(x_1, x_2, ..., x_k) + \epsilon$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

  where g is a linear function of $\beta_0, \beta_1, ..., \beta_p$. We call this a **linear regression model**.
- Y is called **response/dependent variable**. (random, observed)
- $x_1, x_2, ..., x_p$ are called **explanatory/independent variables**. (fixed, observed)
- $\beta_0, \beta_1, ..., \beta_p$ are called **regression parameters**. (fixed, unknown/unobserved)
- ϵ is the **random error** term. (random, unknown/unobserved)

Simple Linear Regression

- In the case of only one explantory/independent variable, x, the linear regression model becomes

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

  which is called **Simple Linear Regression Model**.
- Note that $g(x) = \beta_0 + \beta_1 x$

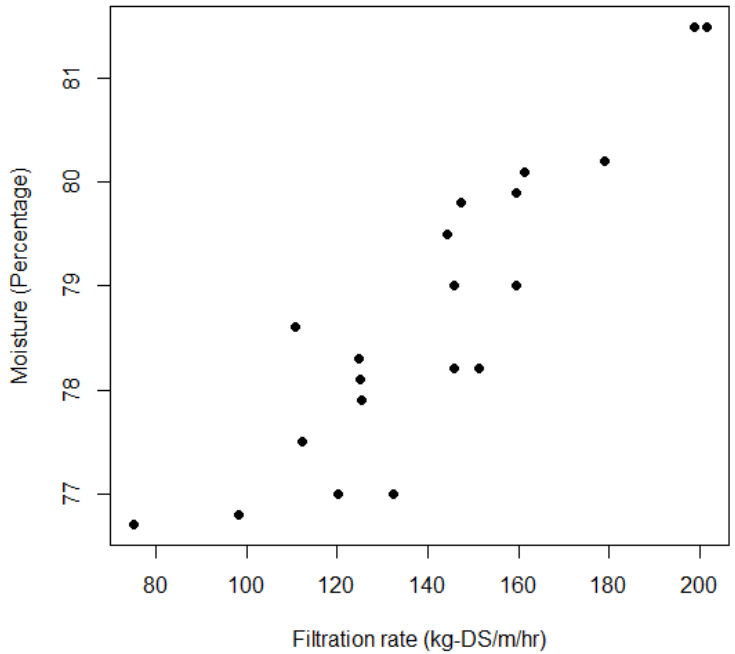
An Motivational Example

- As part of a waste removal project, a new compression machine for processing sewage sludge is being studied. In particular, engineers are interested in the following variables:

  Y = moisture control of compressed pellets (measured as a percent)

  x = machine filtration rate (kg-DS/m/hr).

- Engineers collect $n = 20$ observations of (x, Y).



An Motivational Example

- No simple curve passed exactly through all the points.
- All the points scattered randomly around a straight line.
- It is reasonable to assume that the mean of the random variable Y is related to x by the following straight-line relationship:

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

- **Regression coefficients:** β_0 (intercept), β_1 (slope)
- Naturally, a statistical model is

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

- We assume that $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$

Properties of Simple Linear Regression

- β_0 quantifies the mean of Y when $x = 0$.
- β_1 quantifies the change in $E(Y)$ brought about by a one-unit change in x
- For the model $Y = \beta_0 + \beta_1 x + \epsilon$, we have

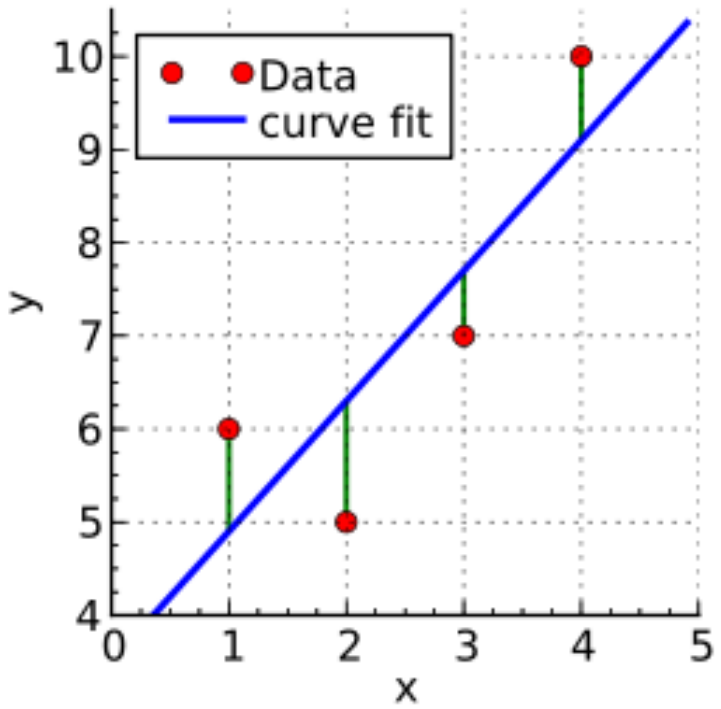
$$E(Y) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x,$$

and

$$\mathrm{Var}\,(Y) = \mathrm{Var}\,(\beta_0 + \beta_1 x + \epsilon) = \mathrm{Var}\,(\epsilon) = \sigma^2.$$

How to Find the Regression Line?

- When we want to use simple linear regression model (a stright line) to fit the data, we want to find the line which is the closest to the observations points.
- What is the meaning of closest?



- Closest means smallest sum of squared distances (green line segaments).

Least Squares Method

- The method which finds the stright line whose sumation of squared distances to observation points are smallest is called the **Method of Least Squares (LS)**.
- Least squares says to choose the values β_0 and β_1 that minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

- Recall that we can minimize or maximize a multivariable function by taking the derivatives with respect to each arguments and set them to 0. So, taking partial derivative of $Q(\beta_0, \beta_1)$, we obtain

$$\begin{aligned} \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \stackrel{\text{set}}{=} 0 \\ \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i \stackrel{\text{set}}{=} 0 \end{aligned}$$

Least Squares Estimators

- Solve above system of equations yields the **least squares estimators**

$$\begin{aligned}\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{SS_{xy}}{SS_{xx}}.\end{aligned}$$

- $SS_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})$ is the Sum of Cross-deviations of Y and x.
- $SS_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$ is Sum of Squared deviations of x.
- $\bar{x} = \frac{1}{n}\sum_{i=1^n} x_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1^n} Y_i$
- Therefore, the estimator of Y (given x) is

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

Fit with R

- In real life, it is rarely necessary to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ by hand.
- Let's see how to use R to fit a regression model in the waste removal project example

```
#enter the data
filtration.rate=c(125.3,98.2,201.4,147.3,145.9,124.7,112.2,120.2,161.2,178.9,
                  159.5,145.8,75.1,151.4,144.2,125.0,198.8,132.5,159.6,110.7)
moisture=c(77.9,76.8,81.5,79.8,78.2,78.3,77.5,77.0,80.1,80.2,79.9,
                79.0,76.7,78.2,79.5,78.1,81.5,77.0,79.0,78.6)

# Fit the model
fit = lm(moisture~filtration.rate)
fit
Call:
lm(formula = moisture ~ filtration.rate)
Coefficients:
    (Intercept)  filtration.rate
       72.95855          0.04103
```

Solution of LSE

- From the output, we see that the least squares estimates are $\hat{\beta}_0 = 72.959$, and $\hat{\beta}_1 = 0.041$.
- Therefore, the equation of the least squares line that relates moisture percentage Y to the filtration rate x is

$$\widehat{Y} = 72.959 + 0.041x.$$

That is to say an estimate of expected moisture is given by

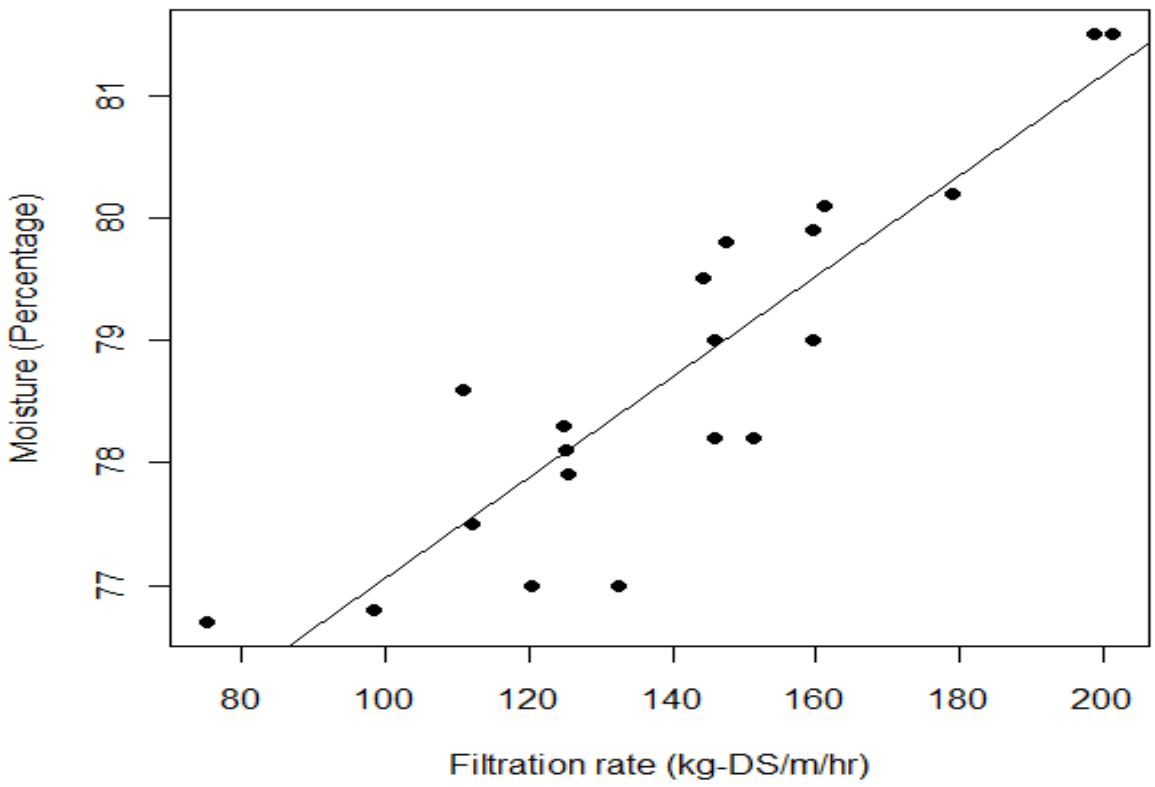
$$\widehat{\text{Moisture}} = 72.959 + 0.041 \times \text{Filtration rate}.$$

- The least squares line is also called *prediction equation*. We can predict the mean response $E(Y)$ for any value of x. For example, when the filtration rate is $x = 150\text{kg}\cdot\text{ DS/m/hr}$, we would predict the mean moisture percentage to be

$$\widehat{Y}(150) = 72.959 + 0.041(150) = 79.109.$$

Scatter Plot with Least Squares Line

```
plot(filtration.rate,moisture,xlab = "Filtration rate (kg-DS/m/hr)",
      ylab = "Moisture (Percentage)",pch=16)
abline(fit)
```



Model Assumptions

- The simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, ..., n$$

  We assume the error term ϵ_i follows
  - $E(\epsilon_i) = 0$, for $i = 1, 2, \ldots, n$
  - $\mathrm{Var}(\epsilon_i) = \sigma^2$, for $i = 1, 2, \ldots, n$, i.e., the variance is constant
  - the random variable ϵ_i are independent
  - the random variable ϵ_i are normally distributed
- Those assumptions of the error terms can be summarized as

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2),$$

  where $i.i.d.$ stands for **independent and identically distributed**.

Model Assumptions

- Under the assumption

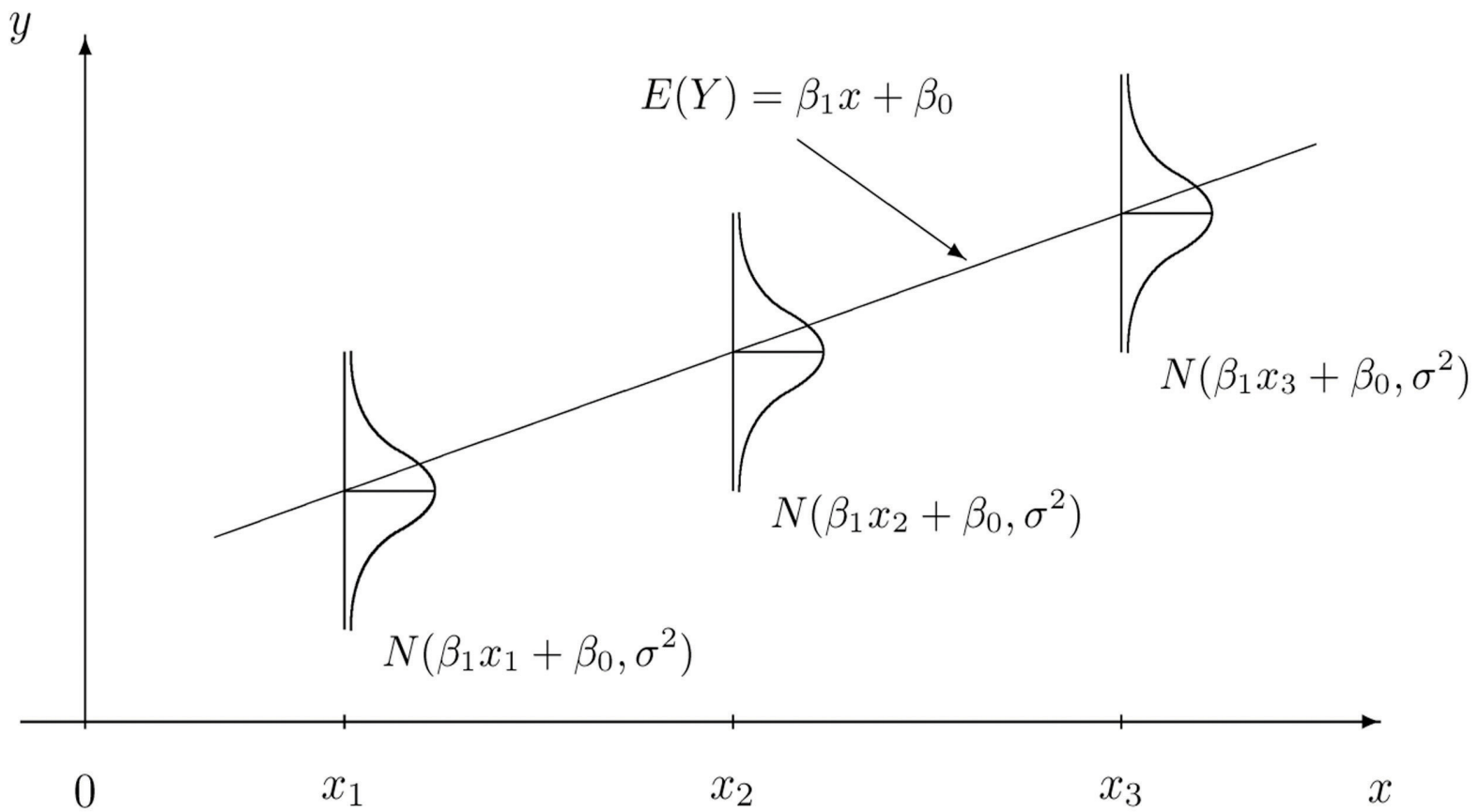
$$\epsilon_1, \epsilon_2, \ldots, \epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

  it follows that

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

- In this normal distribution, we have three unknown but fixed parameters to estimate, namely, β_0, β_1, and σ^2.

Pictorial Illustration of Model Assumptions



Estimating σ^2

- We can use least squares method to estimate β_0 and β_1.
- The residuals

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

are used to obtain an estimator of σ^2. The sum of squares of the residuals, often called the **error sum of squares**, is

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

- **Fact:** $E(SSE) = (n-2)\sigma^2$.
- Using fact, therefore, an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

$\hat{\sigma}^2$ is also called mean squared error (**MSE**).

Calculating $\hat{\sigma}^2$ in R

In R, `predict(fit)` gives the predicted value at each x_i, namely, $\hat{Y}(x_1), \hat{Y}(x_2), \ldots, \hat{Y}_{x_n}$.

```
> fit = lm(moisture~filtration.rate)
> fitted.values = predict(fit)
> residuals = moisture-fitted.values
> # Calculate MSE
> sum(residuals^2)/18
[1] 0.4426659
```

We have $\hat{\sigma}^2 = MSE = 0.443$.

Properties of Least Squares Estimators

- Recall that

$$\begin{aligned}\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{SS_{xy}}{SS_{xx}}.\end{aligned}$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of Y_i, so they are random variables and have their **sampling distributions.**
- It can be shown that

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \left(\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}\right)\sigma^2\right) \text{ and } \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SS_{xx}}\right)$$

- Note that both $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased.

Properties of Least Squares Estimators

- Since σ^2 is unknown, the **estimated standard error** of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\widehat{se(\hat{\beta}_0)} = \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}\right)\hat{\sigma}^2}$$

$$\widehat{se(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}}$$

  where $\hat{\sigma}^2 = \frac{SSE}{n-2}$.

- Given $\hat{\beta}_0$ and $\hat{\beta}_1$ are both normal and the value of standard errors can be estimated, we are able to conduct hypothesis tests (as well as find confidence intervals) on them.

Hypothesis Tests in Simple Linear Regression

- An important part of assessing the adequacy of a linear regression model is testing statistical hypotheses about the model parameters and constructing certain confidence intervals.
- In practice, inference for the slope parameter β_1 is of primary interest because of its connection to the independent variable x in the model.
- Inference for β_0 is less meaningful, unless one is explicitly interested in the mean of Y when $x = 0$. We will focus on inference on β_1.
- Under the model assumptions, the following sampling distribution arises:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\widehat{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / SS_{xx}}} \sim t(n-2)$$

Confidence Interval of $\hat{\beta}_1$

- The sampling distribution of $\hat{\beta}_1$ leads to the following $(1-\alpha)100\%$ confidence interval of β_1:

$$\underbrace{\hat{\beta}_1}_{\text{Point Estimate}} \pm \underbrace{t_{\alpha/2,n-2}}_{\text{Quantile}} \underbrace{\sqrt{\hat{\sigma}^2/SS_{xx}}}_{\text{standard error}}$$

- Note that this is two-sided confidence interval, which corresponds to the test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.
  - If '0'" is covered by this interval, we fail to reject H_0 at significance level of α. **This suggests that Y and x are not linearly related.**
  - If '0'" is not covered by this interval, we reject H_0 at significance level of α. **This suggests that Y and x are linearly related.**

Hypothesis Test for β_1

- Suppose we want to test β_1 equals to a certain value, say $\beta_{1,0}$, that is our interest is to test

$$H_0 : \beta_1 = \beta_{1,0} \text{ versus } H_a : \beta_1 \neq \beta_{1,0}$$

  where $\beta_{1,0}$ is often set to 0 (why?)
- The test statistic under the null is

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / SS_{xx}}} \sim t(n-2).$$

- The p-value of the test is $2P(T_{n-2} < -|t_0|)$, you can use R to find this probability. Remember that smaller p-value provide stronger evidence against H_0
- Let us look at removal project example.

Removal Project Example

We wish to test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.

```
fit = lm(moisture~filtration.rate)
summary(fit)

Call:
lm(formula = moisture ~ filtration.rate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39552 -0.27694  0.03548  0.42913  1.09901

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     72.958547   0.697528 104.596  < 2e-16 ***
filtration.rate  0.041034   0.004837   8.484 1.05e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6653 on 18 degrees of freedom
Multiple R-squared: 0.7999,     Adjusted R-squared: 0.7888
F-statistic: 71.97 on 1 and 18 DF,  p-value: 1.052e-07
```

Note that the residual standard error is $\sqrt{\hat{\sigma}^2} = \sqrt{MSE} = 0.6653$.

Removal Project Example

- R result is

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     72.958547   0.697528 104.596  < 2e-16 ***
filtration.rate  0.041034   0.004837   8.484 1.05e-07 ***
```

- For β_1, we have $\hat{\beta}_1 = 0.0410$, $se(\hat{\beta}_1) = 0.0048$, and by t-table $t_{18,0.025} = 2.1009$
- The 95% confidence interval is

$$\hat{\beta}_1 \pm t_{18,0.025} se(\hat{\beta}_1) = 0.0410 \pm 2.1009(0.0048) = (0.0309, 0.0511)$$

- The p-value of t test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ is less than $1.05 \times 10^{-7} \approx 0$.
- What is your conclusion?

Confidence and prediction intervals for a given $x = x_0$

- Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

- We are interested in using the fitted model to learn about the response variable Y at a certain setting for the independent variable, say, $x = x_0$.
- Two potential goals:
  - Estimating the **mean response** of Y. This value, $E(Y|x_0)$, is the mean of the following probability distribution

$$\mathcal{N}(\beta_0 + \beta_1 x_0, \sigma^2)$$

  - Predicting a **new response** Y, denoted by $Y^*(x_0)$. This value is one new outcome from

$$\mathcal{N}(\beta_0 + \beta_1 x_0, \sigma^2)$$

Confidence and prediction intervals for a given $x = x_0$

- Two potential goals:
  - Estimating the **mean response** of Y.
  - Predicting a **new response** Y.
- **Difference:** In the first problem, we are estimating the mean of a distribution. In the second problem, we are predicting the value of a new response from this distribution. The second problem is more difficult than the first one.
- **Goals:** We would like to create $100(1-\alpha)\%$ intervals for the mean $E(Y|x_0)$ and for the new value $Y^*(x_0)$.
- The former is called a **confidence interval** and the latter is called a **prediction interval**.

Confidence and prediction intervals for a given $x = x_0$

- A $100(1-\alpha)\%$ **confidence interval** for the mean $E(Y|x_0)$ is

$$\widehat{Y}(x_0) \pm t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]}$$

- A $100(1-\alpha)\%$ **prediction interval** for the new value $Y^*(x_0)$ is

$$\widehat{Y}(x_0) \pm t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]}$$

Confidence and prediction intervals for a given $x = x_0$

- Note that the prediction interval is wider than the confidence interval! The extra "1" in the prediction interval's standard error arises from the additional uncertainty associated with predicting a new response from the distribution.
- The length of the interval is smallest when $x_0 = \overline{x}$ and will get larger the farther x_0 is from $\overline{x}$ in either direction.
- **Warning:** It can be very dangerous to estimate $E(Y|x_0)$ or predict $Y^*(x_0)$ based on the fit of the model for values of x_0 outside the range of x values used in the experiment/study. This is called extrapolation.

Removal Project Example

In the removal Project example, suppose that we are interested in estimating $E(Y|x_0)$ and predicting a new $Y^*(x_0)$ when the filtration rate is $x_0 = 150$.

- $E(Y|x_0)$ denotes the mean moisture percentage for compressed pellets when the machine filtration rate is $x_0 = 150$.
- $Y^*(x_0)$ denotes a possible value of Y for a single run of the machine when the filtration rate is set at $x_0 = 150$.

Removal Project Example

- Confidence interval:

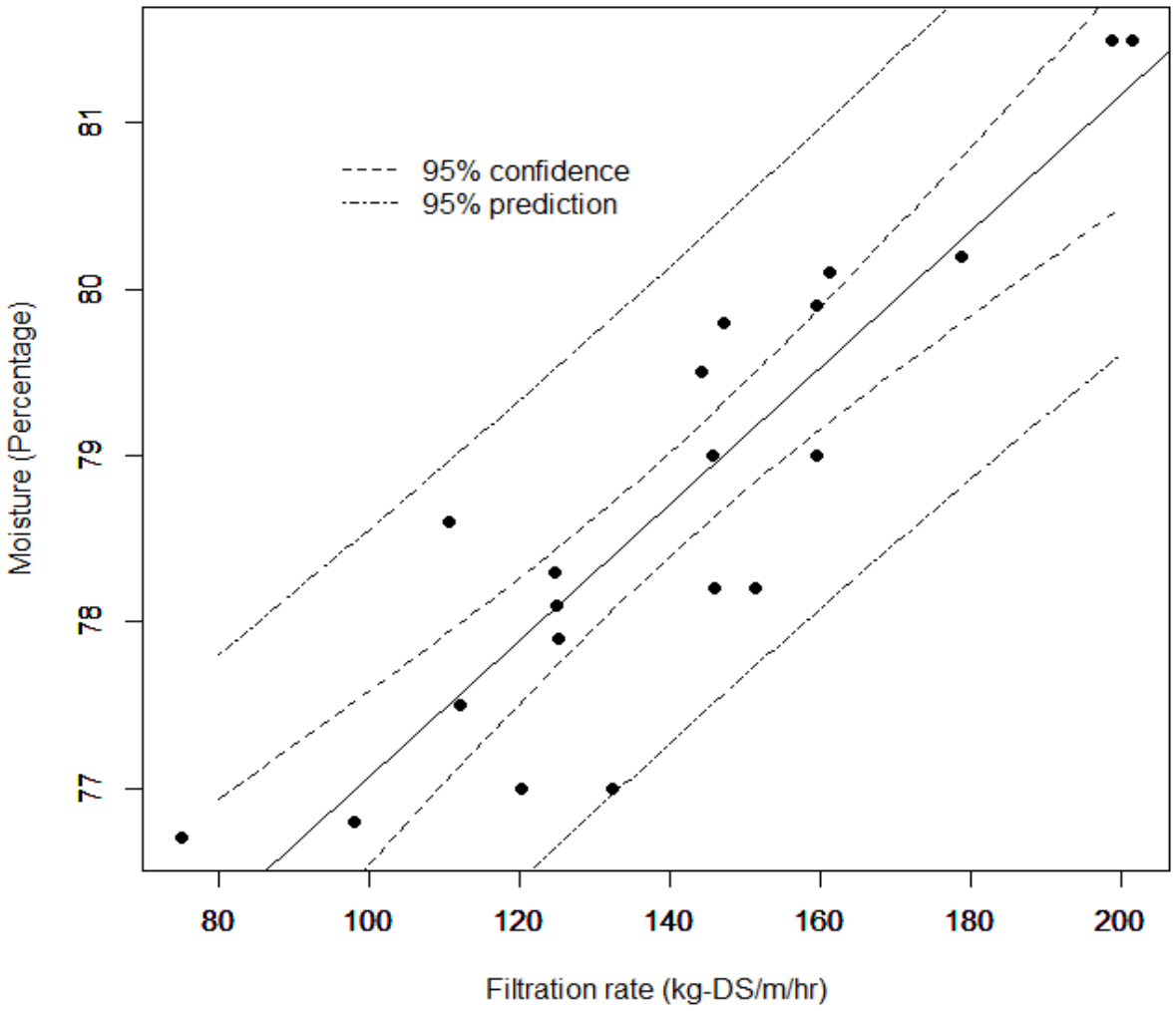
```
>  predict(fit,data.frame(filtration.rate=150),level=0.95,interval="confidence")
       fit      lwr      upr
1 79.11361 78.78765 79.43958
```

- Prediction interval:

```
>  predict(fit,data.frame(filtration.rate=150),level=0.95,interval="prediction")
       fit     lwr      upr
1 79.11361 77.6783 80.54893
```

- Interpretation
  - A 95% confidence interval for $E(Y|x_0 = 150)$ is $(78.79, 79.44)$. When the filtration rate is $x_0 = 150$ kg-DS/m/hr, we are 95% confident that the **mean** moisture percentage is between 78.79 and 79.44 percent.
  - A 95 percent prediction interval for $Y^*(x_0 = 150)$ is $(77.68, 80.55)$. When the filtration rate is $x_0 = 150$ kg-DS/m/hr, we are 95% confident that the moisture percentage for a **single run** of the experiment will be between 77.68 and 80.55 percent.

Confidence Interval v.s. Prediction Interval



Model and Assumptions Checking: Three problems

1. How good the regression line is?
2. Is the error term ϵ really normally distributed?
3. Is the assumption that variances of $\epsilon_1, \epsilon_2, .., \epsilon_n$ are the same true?

Problem 1: How good the regression line is?

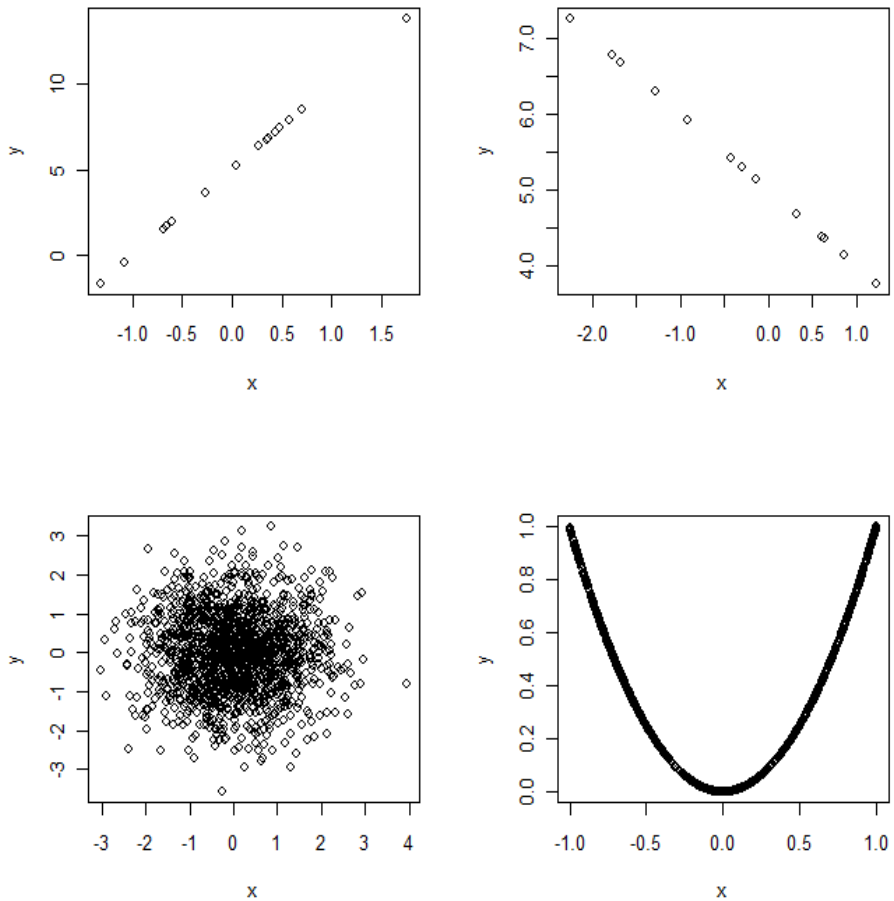
- In simple liear regression, this problem can be transferred to how strong Y and x are linearly correlated. If they have a close linear correlation, the model should work very well. If not, we get a bad model.
- **sample Coefficient of Correlation** is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}.$$

- $r \in [-1, 1]$.

Coefficient of Correlation

The plot in the top left corner has $r = 1$; the plot in the top right corner has $r = -1$; the plot in the bottom left and right corner have $r \approx 0$;



Coefficient of Determination

- **Coefficient of Determination**, denoted by r^2, measures the contribution of x in the predicting of y. (Sometimes it is called **Regression R-square**.)
- Define

$$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2, SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- If x makes no contribution to prediction of y, then $\beta_1 = 0$. In this case,

$$Y = \beta_0 + \epsilon.$$

  It can be shown that $\hat{Y}_i = \hat{\beta}_0 = \overline{Y}$, and $SSE = SSTO$.
- If x contribute to prediction of Y_i, then we expect $SSE << SSTO$. In other words, the independent variable x "explain" significant amount of variability among data.

Coefficient of Determination

- Intuitively, $SSTO$ is total sample variation around $\overline{Y}$, and SSE is unexplained sample variability after fitting regression line.
- Coefficient of determination is defined as

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{\text{Total Variability - Unexplained Variability}}{\text{Total Variability}}$$

  which can be understood as the proportion of total sample variance explained by linear relationship.
- In simple linear regression, the coefficient of determination **equals to** the squared sample coefficient of correlation between x and Y.

Removal Project Example

We can use command `cor` to calculate sample coefficient of correlation. The coefficient of determination r^2 is called `Multiple R-squared` in the summary of simple linear regression.

```
> fit<-lm(moisture~filtration.rate)
> summary(fit)

Residual standard error: 0.6653 on 18 degrees of freedom
Multiple R-squared: 0.7999,     Adjusted R-squared: 0.7888
F-statistic: 71.97 on 1 and 18 DF,  p-value: 1.052e-07

> r <- cor(filtration.rate,moisture)
> r^2
[1] 0.7999401
```

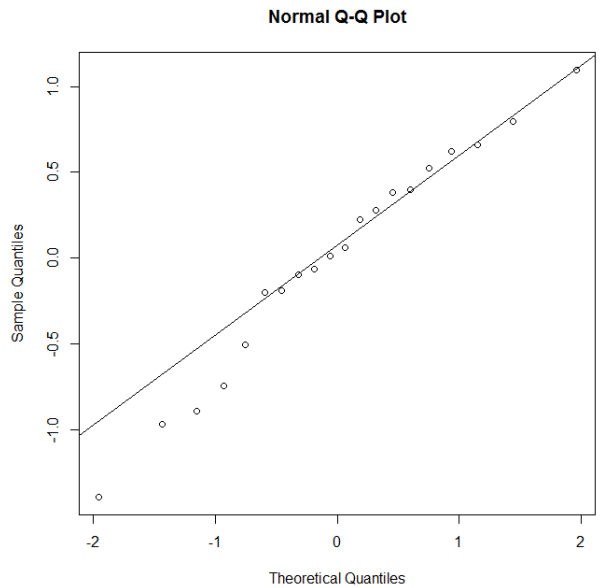
Problem 2: Is the Error Term Really Normally Distributed?

- The **residuals** from a regression model are $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n.$, where y_i is an actual observation and $\hat{y}_i$ is the corresponding fitted value from the regression model.
- Analysis of the residuals is frequently helpful in checking the assumption that the errors are approximately normally distributed with constant variance, and in determining whether additional terms in the model would be useful.
- As an approximate check of normality, we can use apply the fat pencil test to the normal qq plot of residuals.

Normal qq plot for Removal Project Example

- Normal qq plot for removal project

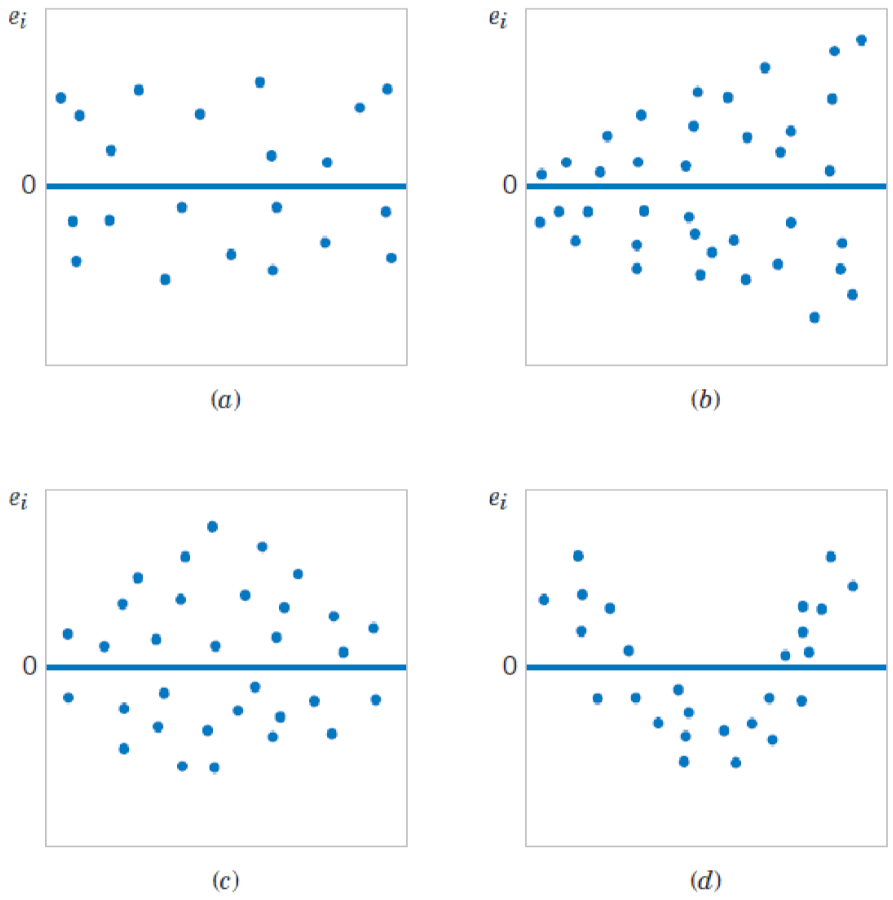
```
resid <- residuals(fit)
qqnorm(resid)
qqline(resid)
```



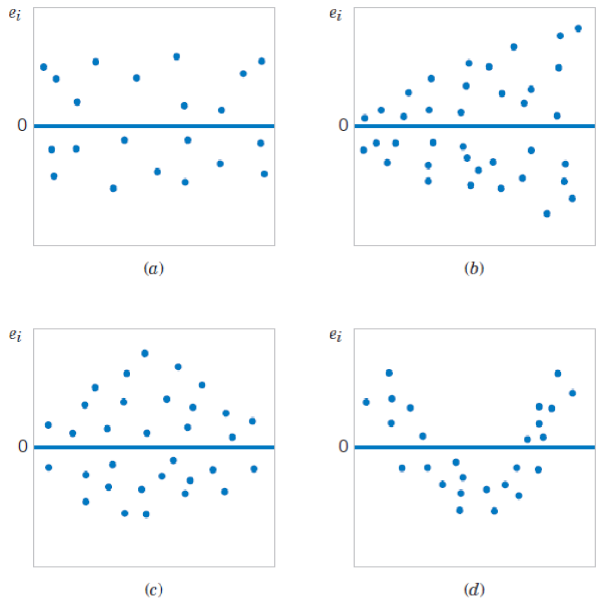
- What is your conclusion?

Problem 3: Are the variances of $\epsilon_1, \epsilon_2, .., \epsilon_n$ Equal?

We use residual plot to check this assumption. Residual plot is simply the scatterplot of residuals e_i's and predicted values. For example,



Residual Plots

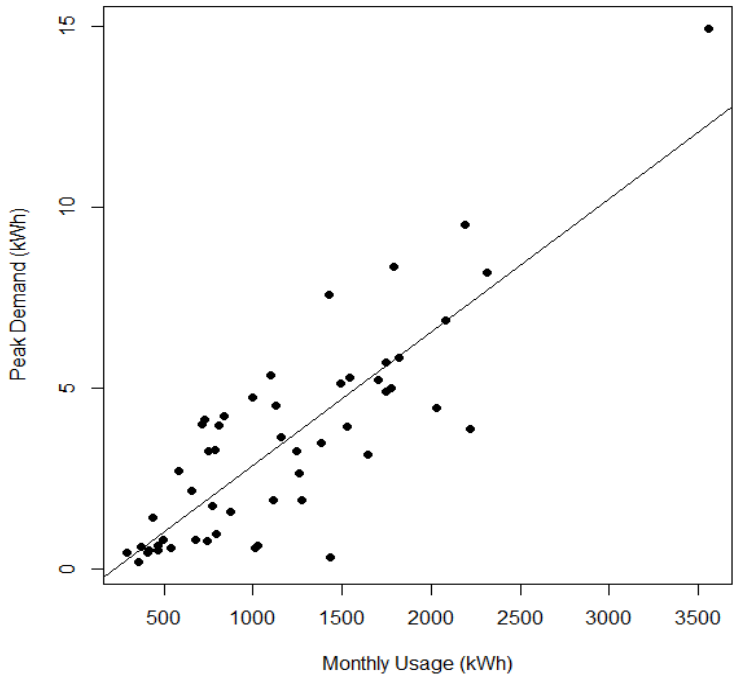


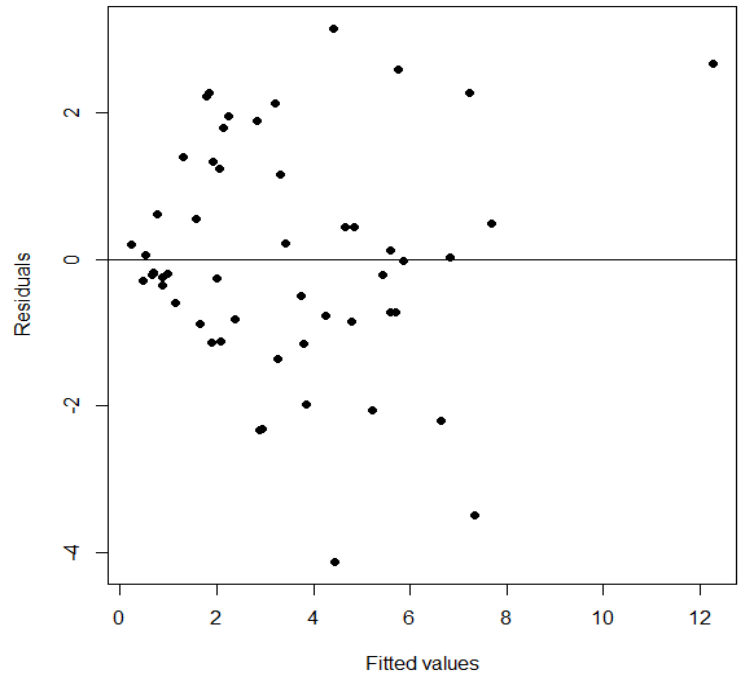
- Pattern (a) represents the ideal situation.
- Pattern (b) represents the cases where the variance of the observations may be increasing with the magnitude of y_i or x_i. Pattern (b) and (c) represents the unequal variance cases.
- Pattern (d) indicates the linear relationship between $E(Y_i)$ and x_i is not proper. We need to add higher order term, which requires multiple linear regression.

Example: Electricity Consumption

An electric company is interested in modeling peak hour electricity demand (Y) as a function of total monthly energy usage (x). This is important for planning purposes because the generating system must be large enough to meet the maximum demand imposed by customers.

```
electricity <- read.table(file="D:/electricity.txt",head=TRUE)
# Define variables
monthly.usage = electricity[,1]
peak.demand = electricity[,2]
# Fit the model
fit = lm(peak.demand ~ monthly.usage)
summary(fit)

# Plots were constructed separately
# Scatterplot
plot(monthly.usage,peak.demand,xlab = "Monthly Usage (kWh)",
      ylab = "Peak Demand (kWh)", pch=16)
abline(fit)
# Residual plot
plot(fitted(fit),residuals(fit),pch=16,
      xlab="Fitted values",ylab="Residuals")
abline(h=0)
```



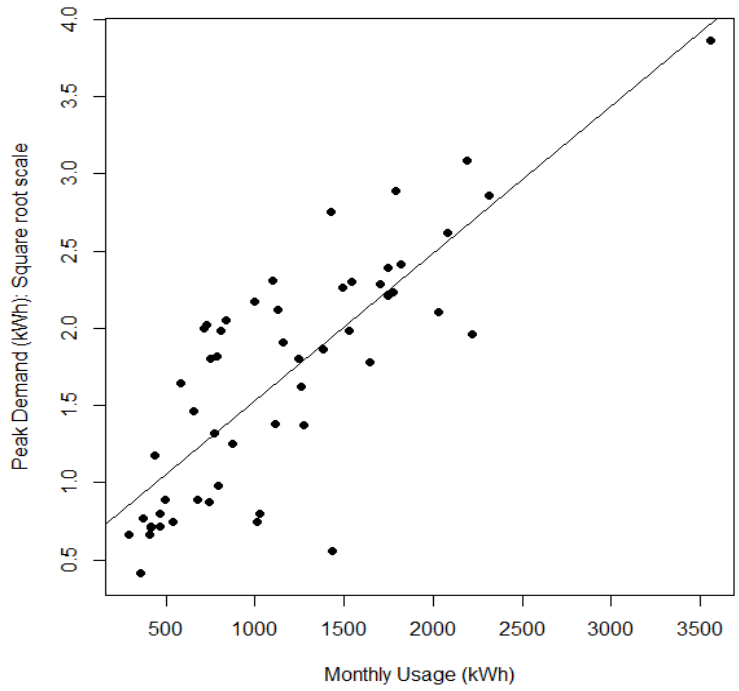


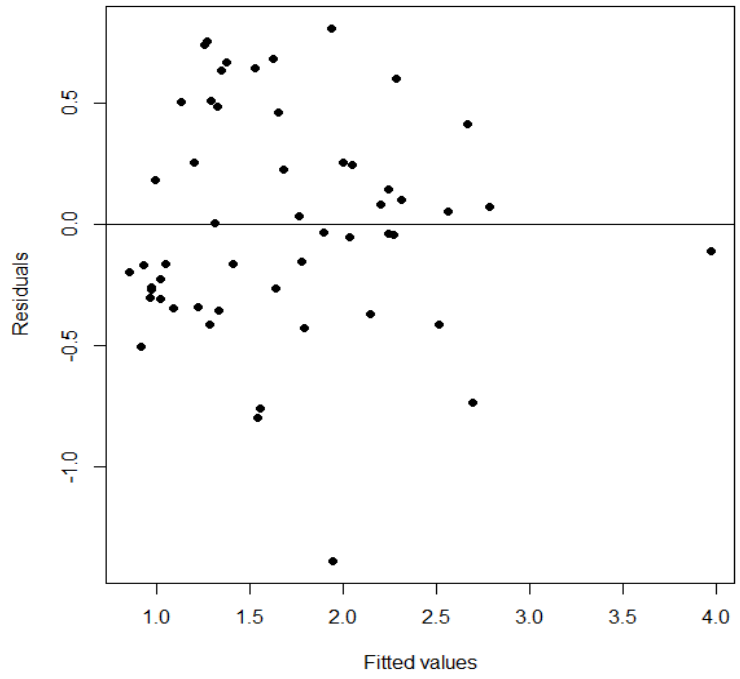
- The residual plot shows clearly a "megaphone" shape, which indicates that the equal variance assumption is violated.
- Widely used variance-stabilizing transformations include the use of $\sqrt{y}$, $\log y$, or $1/y$ as the response.
- Let us try $\sqrt{y}$ as the response.

Transforming the Response

You can use `sqrt(peak.demand)` in `R` to transform the response variable directly.

```
# Fit the transformed model
fit.2 <- lm(sqrt(peak.demand) ~ monthly.usage)

# Plots were constructed separately
# Scatterplot
plot(monthly.usage,sqrt(peak.demand),xlab = "Monthly Usage (kWh)",
      ylab = "Peak Demand (kWh): Square root scale", pch=16)
abline(fit.2)
# Residual plot
plot(fitted(fit.2),residuals(fit.2),pch=16,
      xlab="Fitted values",ylab="Residuals")
abline(h=0)
```





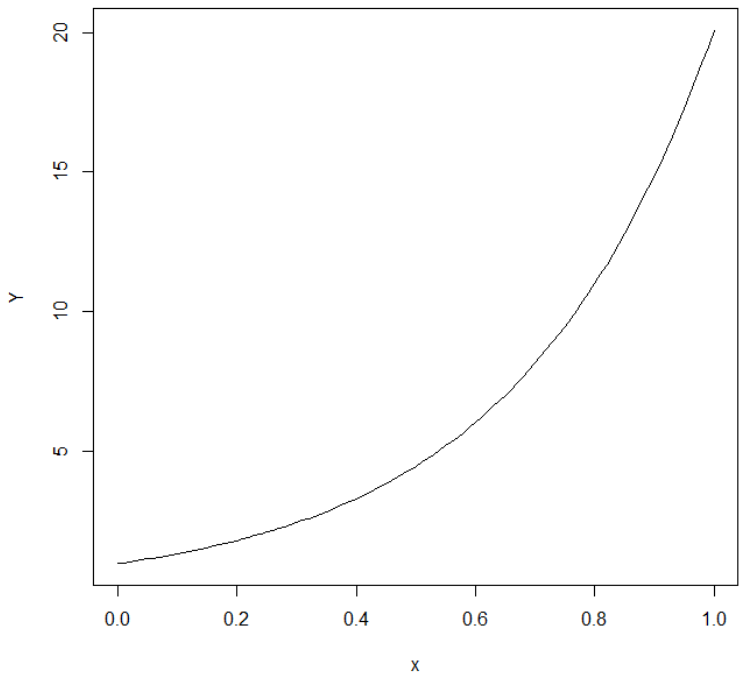
- The residual plot looks much better.
- Model interpretation: $\sqrt{Y_i} = \beta_0 + \beta_1 x_i + \epsilon_i$.

Transformation of Variables

- Variable transformation can be used in multiple situations (not only in variance-stabilizing) to make the simple linear regression available.
- For example,

$$Y = \beta_0 e^{\beta_1 x} \epsilon$$

- Clearly, Y and x are not linearly related. Let's check its shape:



Transformation of Variables

- If we take a logarithmic ransformation to the both side of the equation of $Y = \beta_0 e^{\beta_1 x} \epsilon$, then

$$\log(Y) = \log(\beta_0 e^{\beta_1 x} \epsilon) = \log(\beta_0) + \beta_1 x + \log(\epsilon)$$

- If we denote $Y^* = \log(Y)$, $\beta_0^* = \log(\beta_0)$ and $\epsilon^* = \log(\epsilon)$, the equation becomes

$$Y^* = \beta_0^* + \beta_1 x + \epsilon^*$$

- Now, the relation between Y^* and x is linear! A nonlinear function, which can be expressed as linear function (straight line) by using a suitable transformation, is called **intrinsically linear**.

Tranfer Functions to Linear

1. $Y^2 = \beta_0 + \beta_1 x + \epsilon$
2. $Y = \sqrt{\beta_0 + \frac{\beta_1}{x} + \epsilon}$
3. $Y = \frac{1}{e^{\beta_0 + \beta_1 x + \epsilon}}$
4. $Y = \frac{x}{\beta_0 x + \beta_1 + x\epsilon}$
5. $Y = \beta_0 x^{\beta_1} \epsilon$
6. $Y = \beta_0 \beta_1^x \epsilon$

Transformation Example

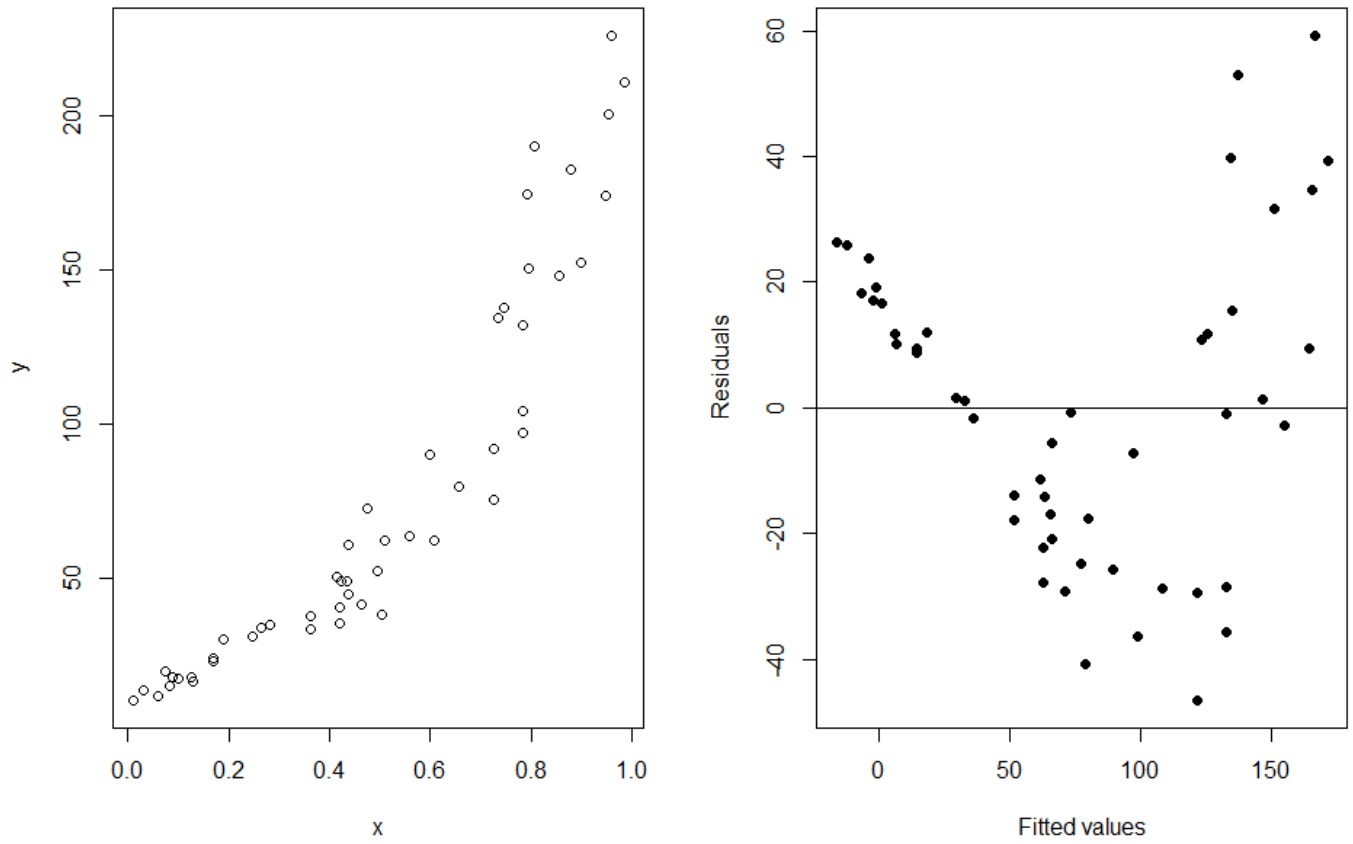
Suppose we are given a dataset and we want to investigate the relation between two variables, X and Y. We fit the data with simple linear regression with `lm()` and R gives

```
> fit1 <- lm(y~x)
> summary(fit1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -17.92       7.01  -2.556   0.0138 *
x             192.52      12.22  15.760   <2e-16 ***

Residual standard error: 25.13 on 48 degrees of freedom
Multiple R-squared:  0.8381,    Adjusted R-squared:  0.8347
F-statistic: 248.4 on 1 and 48 DF,  p-value: < 2.2e-16
```

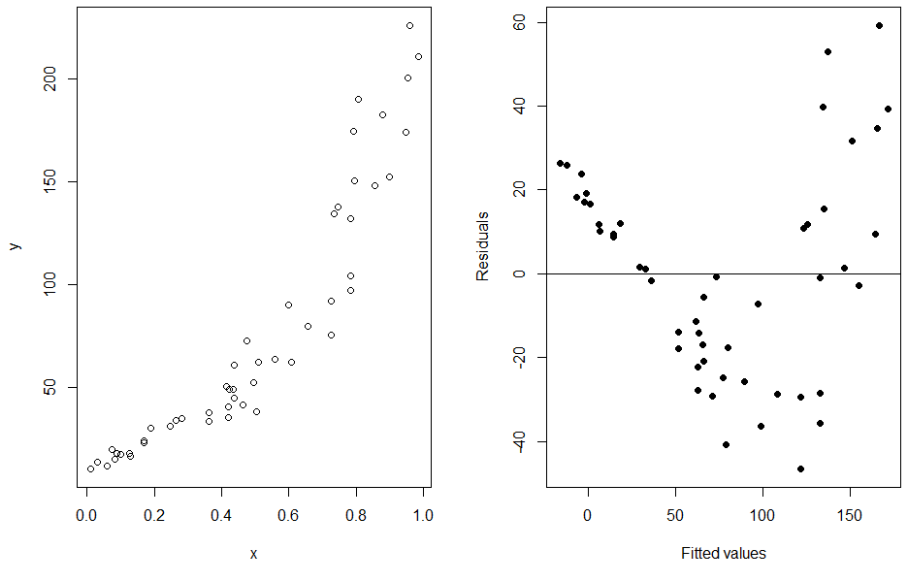
Are you happy with the results?

Transformation Example

Let's check the scatter plot (left) and residual plot (right)



Transformation Example



- The scatter plot (left) shows that X and Y are not linearly related!
- The residual plot shows that the variance of the residual changes with different $\hat{y}$, indicating that the equal variance assumption is violated.
- Even thoguth we are happy with the regression model, it turns out to be a bad one.

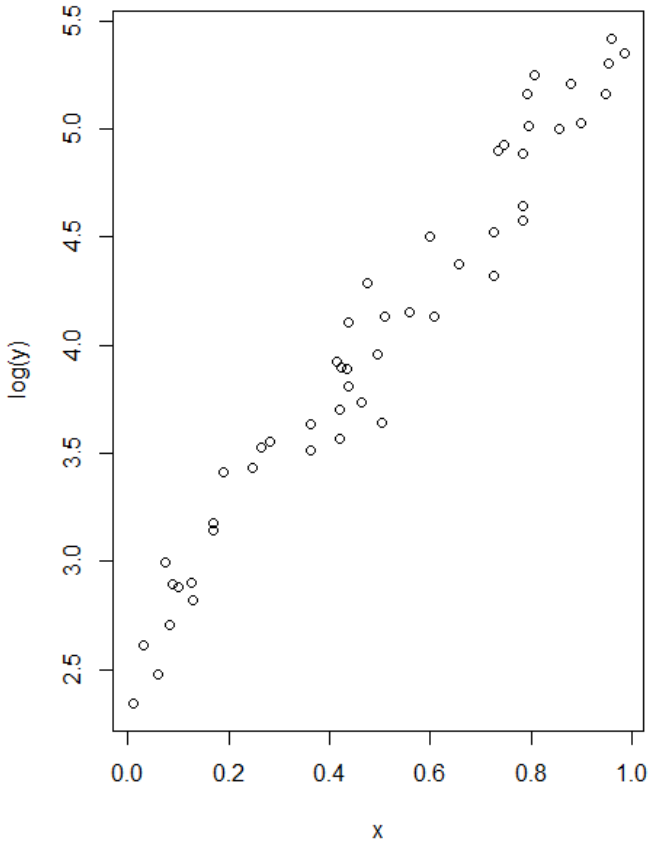
Transformation Example

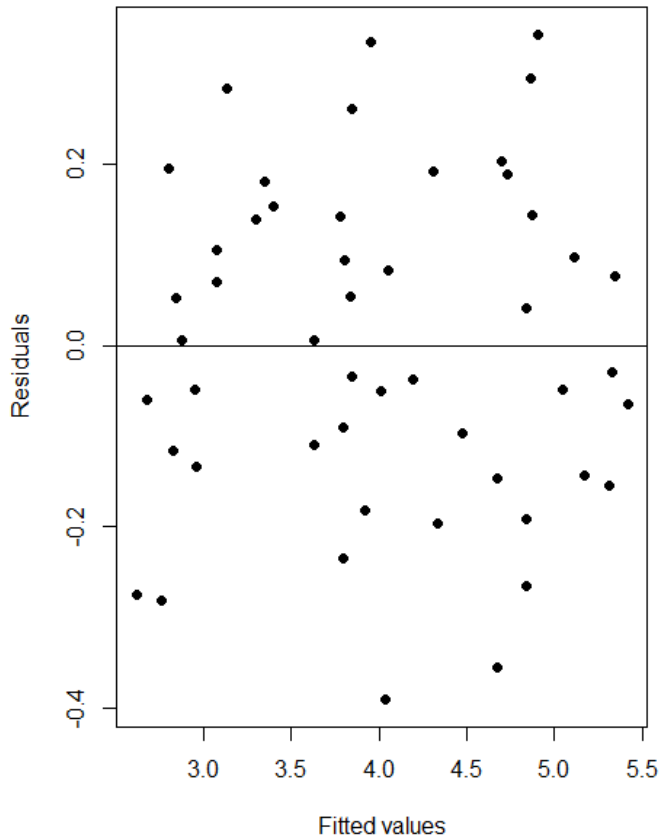
- Let's try the log transformation to the Y, which means we denote $Y^* = \log(Y)$ and fit the model in R using the new Y^*.

```
> fit2 <- lm(log(y) ~ x)
> summary(fit2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.58767    0.05092   50.82   <2e-16 ***
x            2.87523    0.08873   32.40   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1825 on 48 degrees of freedom
Multiple R-squared:  0.9563,    Adjusted R-squared:  0.9554
F-statistic:  1050 on 1 and 48 DF,  p-value: < 2.2e-16
```

Transformation Example

Let's check the scatter plot of X v.s.log(Y) (left) and residual plot (right)





Now, are you truly happy?

Box-Cox Transformation

- You might ask "What is the best transformation we should use?" In the last example, log transformation seems to work well, but is there a better one?
- In 1964, Box and Cox suggested the following transformations

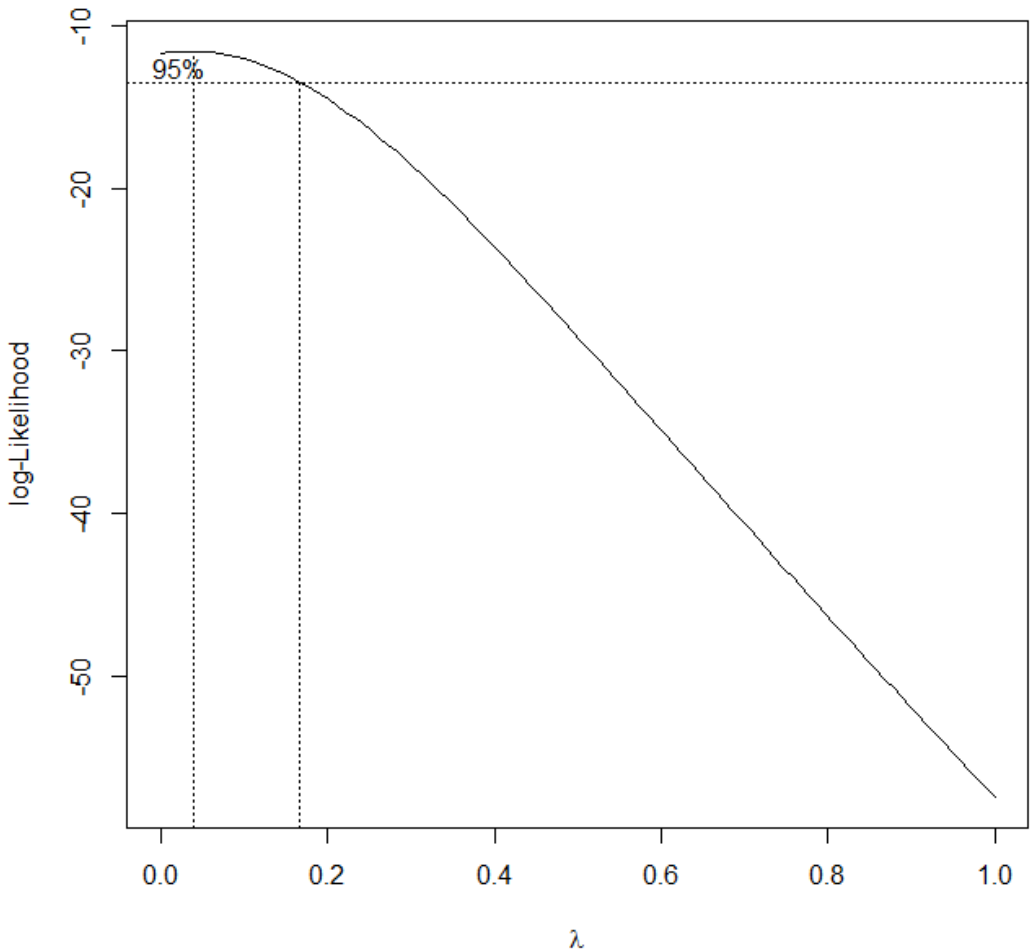
$$\text{BoxCox}(Y) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}$$

- When $\lambda \approx 0$, $\frac{Y^\lambda - 1}{\lambda} \approx \log(Y)$.
- Box-Cox transformation depends on the parameter λ, which is unknown. How to estimate the best λ?

Find Best λ

Let's use the previous dataset to learn how to find the best λ in R.

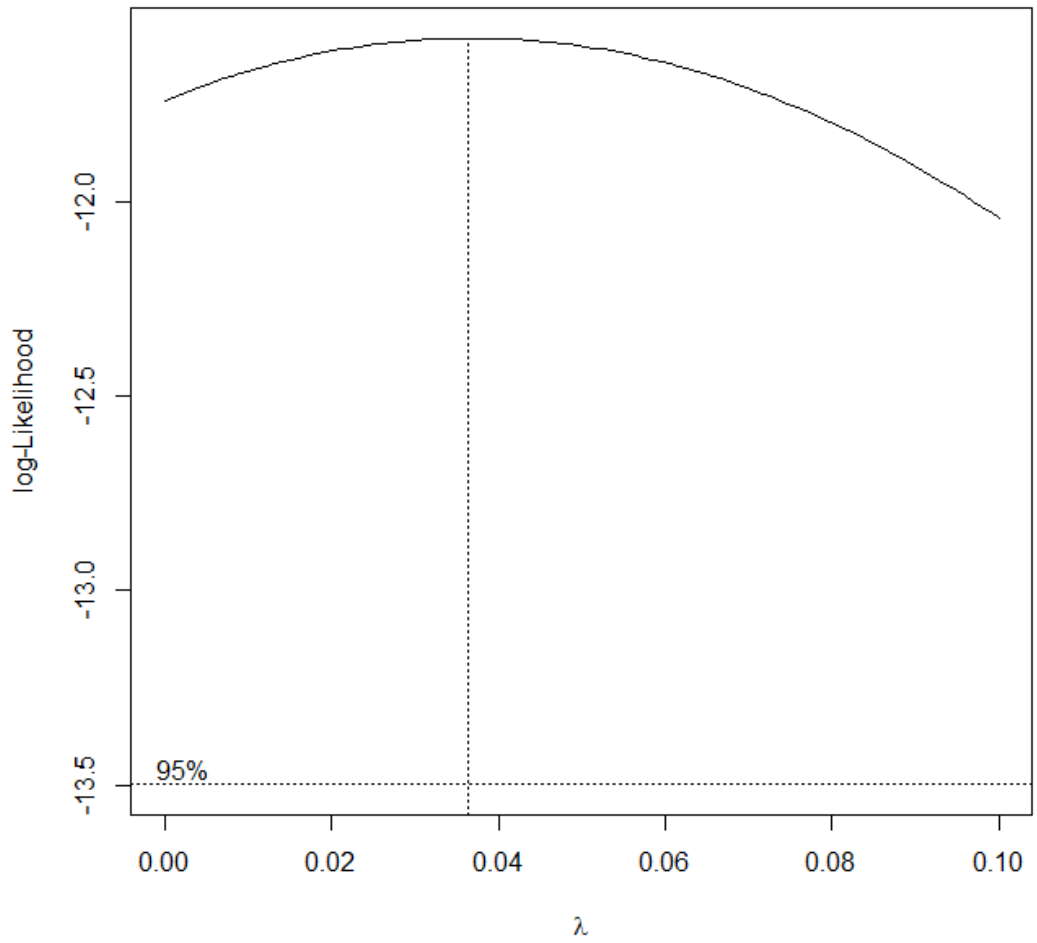
```
library(MASS)  # We need to MASS package first
boxcox(lm(y~x), lambda=seq(0,1,by=0.1))
```



Find Best λ

It seems the best λ is in $[0, 0.1]$. Let's adjust the range of λ in R.

```
boxcox(lm(y~x), lambda=seq(0,0.1,by=0.01))
```



Find Best Transformation

- R suggests us the following transformation

$$\text{BoxCox}(Y) = \frac{Y^{0.04} - 1}{0.04}$$

- Remark: $\lambda = 0.04$ is very close to 0, and when $\lambda \approx 0$, the log transformation $\log(Y)$ is the best.
- What we have done in the previous example is actually not far away from the best!

Logistic Regression

- So far, we have learnt how to find a good simply linear regression model to fit the data, whose response variable Y is quantitative, e.g. continuous numbers.
- However, in reality, we might be interested in the case that the response variable Y is binary, e.g. success and failure, 0 and 1.
- For example, we want to predict whether it will rain tomorrow, or whether the O-ring will fail in the space shuttle launch, etc.

Bernoulli Response Variable

- In chapter 2, we have learnt the Bernoulli trials, which meet the following properties:
  1. each trial results a "success" or "failure"
  2. trials are independent
  3. P(success) is the same for each trial, denoted as p, $0 < p < 1$.
- In logistic regression, we have each response variable Y_i is a Bernoulli random variable, with $E(Y_i) = P(Y_i = 1) \equiv p_i$.

Failure of Interpretation

- If we still use the normal simple linar regression method to regress Y, the results might not be interpretable. For example, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $E(\epsilon_i) = 0$ gives

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

  where $E(Y_i) = P(Y_i = 1) = p_i$.
- It is possible that the fitted value, say $\hat{\beta}_0 + \hat{\beta}_1 x_i$, is greater than 1 or less than 0.
- By the Kolmogorov Axioms (Chapter 2 page 17), we cannot have probability greater than 1 or less than 0! Therefore, we are not able to interpret the regression result properly.

Logit Function

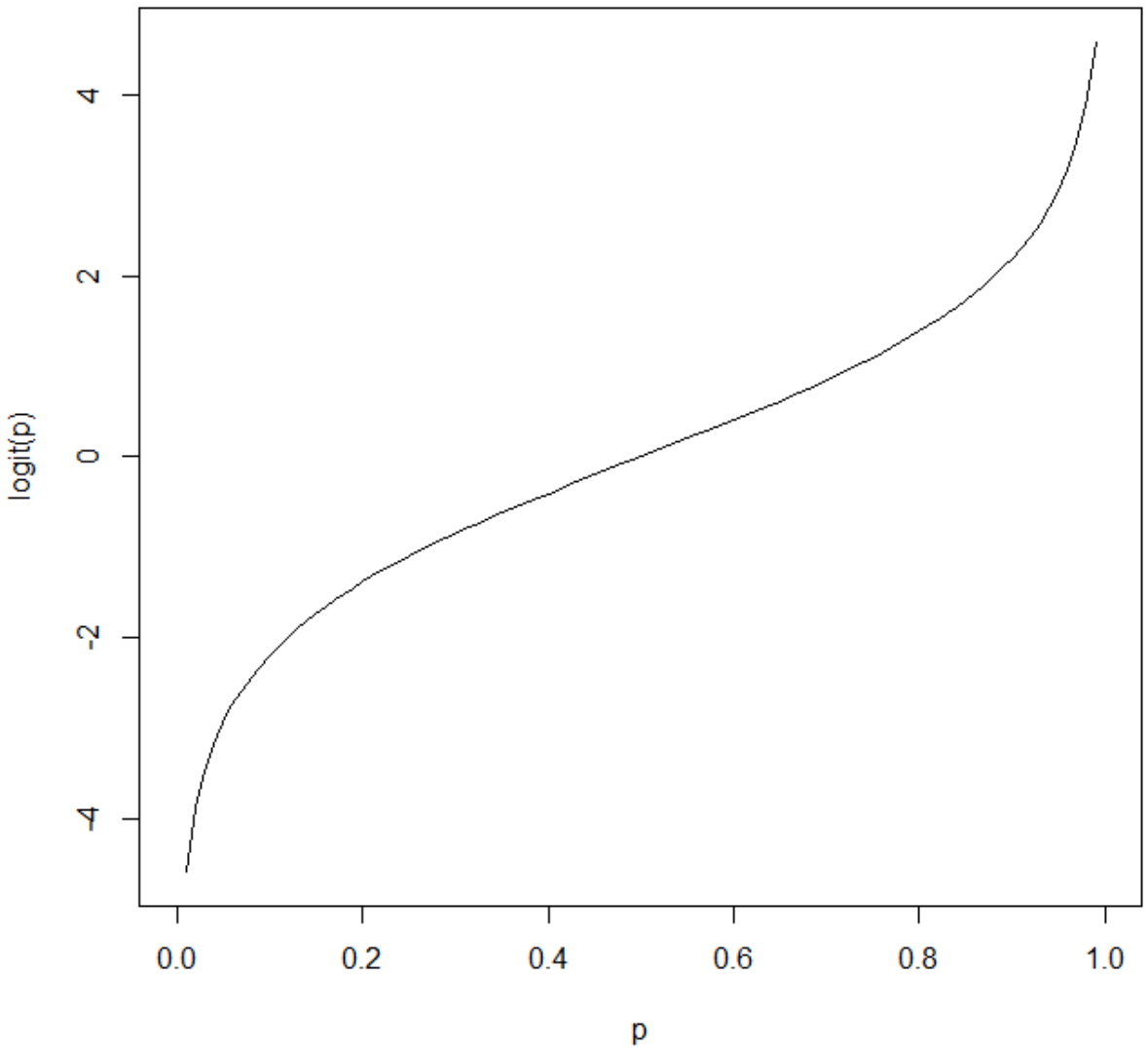
- **Logit function** is frequently used in mathematics and statistics. It is defined as

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \qquad 0 < p < 1$$

- The reason it is popular is that it can transfer a random variable from $(0, 1)$ to the entire real line.

- Note that when p is close to 0, $\text{logit}(p)$ is close to $-\infty$, and when p is close to 1, $\text{logit}(p)$ is close to ∞.

Shape of the Logit Function



Logistic Regression Model

- Using the property of the logit function, the logistic regression model is that

$$\log\left(\frac{E(Y_i)}{1-E(Y_i)}\right) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

- Solving for p_i, this gives

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$$

- Remark: logistic regression is NOT a model simply transfer Y_i with the logit function. The logit transformation is conducted with respect to $E(Y_i)$.

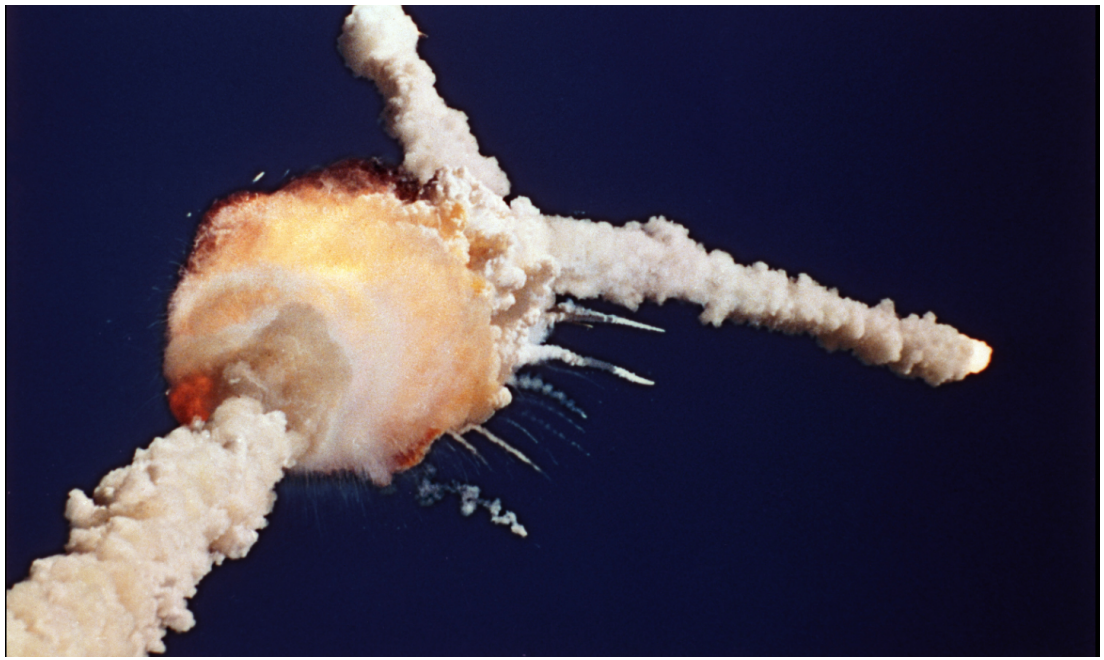
Reasons to Use Logistic Regression

Logistic regression is one the most commonly used tools for applied statistics and discrete data analysis. There are basically four reasons for this.

1. Tradition (David Cox 1958, the same Cox in CoxBox).
2. $\frac{p}{1-p}$ is called **odds**, so that the logit function, $\log\left(\frac{p}{1-p}\right)$, is the **log odds**, which plays an important role in the analysis of contingency tables.
3. It 's closely related to "exponential family" distributions, which is massively used in many contexts, e.g. engineering, physics, etc.
4. It often works surprisingly well!

Example: O-Ring Failure

The **Space Shuttle Challenger disaster** occurred on January 28, 1986, when the NASA space shuttle orbiter *Challenger* broke apart 73 seconds into its flight, leading to the deaths of its seven crew members. Disintegration of the vehicle began aftter an **O-Ring** seal in its right solid rocket booster failed at liftoff.



Example: O-Ring Failure

O-Ring seal failed becuase the launch temperature is lower than expected. Therefore, it is critical to carefully test the reliability of O-Ring under different circumstance. Here we have 24 data points, including the lauching temperature and whether at least one O-Ring failure has occured.

Table : My caption

O-Ring Failure	Temperature
1	52
1	56
1	57
0	63
0	66
...	...
0	81

Example: O-Ring Failure

```
# input data
oring <- c(1,1,1,0,0,0,0,0,0,0,0,1,1,1,0,0,0,1,0,0,0,0,0,0)
temperature <- c(53, 56, 57, 63, 66, 67, 67, 67, 68, 69, 70,
  70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 80, 81)

# fit the logistic regression model
fit <- glm(oring ~ temperature, family=binomial)
summary(fit)

# R output
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.87535    5.70291   1.907   0.0565 .
temperature -0.17132    0.08344  -2.053   0.0400 *
```

Example: O-Ring Failure

- The fitted logistic regression model is

$$\log\left(\frac{p_i}{1-p_i}\right) = 10.875 - 0.171x_i$$

- It is equivalent to

$$E(Y_i) = p_i = \frac{1}{1+e^{-(10.875-0.171x_i)}}$$

- Remark: in testing $H_0 : \beta_1 = 0$ v.s. $H_a : \beta_1 \neq 0$, the p-value is 0.04, indicating that the linear relationship between $\log\left(\frac{E(Y_i)}{1-E(Y_i)}\right)$ and x is significant.

Example: O-Ring Failure

- The actual temperature at the *Challenger* launch was 31 F.

$$p_i = \frac{1}{1 + e^{-(10.875 - 0.171(31))}} = 0.996$$

- The probability that at least one O-Ring failure is 99.6%! It is almost certainly going to happen!
- The **Odds Ratio** is $e^{\hat{\beta}_1} = e^{-0.171} = 0.843$, so every 1 degree increase in temperature reduces the odds of failure by 0.843.
- It is interesting to note that all of these data were available **prior** to launch. However, engineers were unable to effectively analyze the data and use them to provide a convincing argument against launching *Challenger* to NASA managers.