Introduction
Outline

1. Website (https://academic.macewan.ca/burok/index.html)
2. Office hours (online, in person)
3. Homework/Labwork – Crowdmark (practice assignment 0)
4. email for contacting me outside of lecture, lab, and office hours
5. text book (recommended): Introduction to Linear Regression Analysis, Montgomery, Peck, Vining (any edition), is available online in the MacEwan library
6. major project in the lab
7. Install R and RStudio on your computer
8. Weekly email (as long as we are online)

Prerequisites

- Assumption, you can interpret statistics:
  - What is inferential statistics?
  - What does it mean to be 95% confident?
  - What is measured by the P-value?
  - Rational in testing
  - Why can we not avoid allowing errors in inferential statistics?
  - Why do we need the assumption to be met?
  - How to conduct a test? (6 steps)

  On my website for STAT 252 there are detailed notes on these concepts ((Introduction and Review), which will probably beneficial to review
- Assumption, you remember the basics in probability theory
  - Normal distribution
  - Conditional distribution
- Assumption, you know the basics from Linear Algebra
  - Matrix Algebra (addition, scalar multiplication, multiplication, etc.)
  - Solving a system of linear equations
  - Properties of the inverse of a matrix

- In this course we add to your statistical "toolbox".

  Regression is one of the most common and versatile tools in applied statistics.

  After you have completed the course you will know how to identify scenarios when regression can be used, how to check model assumptions, and address possible violations. You will have experienced the application of regression to a larger scientific question, and how to build a model fitting data of your choice.

What are your expectations?

Course contents

- The model
- How are estimators chosen? Tests
- Finding and dealing with model inadequacies
- More models
- Regression analysis of categorical data
- Using R in the process (in class and in the lab).

1 Introduction

1.1 Basics

- Population versus sample
- Estimation (point, interval)
- Testing
- Linear Algebra

1.2 Models

"Essentially, all models are wrong, but some are useful. " (George Box, 1978)

In statistics we use models to describe the population and then we use sample data, to estimate parameters in the model, and conduct tests relating to these parameters.
Regression is a statistical tool for analyzing models which relate response variables (variables we want to learn about), and independent(explanatory) variables, these might or might not be related to the response variable.
Regression is one of the most commonly used tools in applied statistics and is applied in many diverse fields like psychology, engineering, management, biology, medicine, etc.
Examples:

- house prices in dependency on size, location number of bed rooms, lot size, etc.
- time to delivery in dependency on distance, number of left turns, number of traffic lights
- decrease in blood pressure in dependency on medication, dose of medication, initial blood pressure, diabetes (yes,no), etc.
- degree of procrastination in dependency on sleep quality, emotional well-being, social well being, and psychological well being, and number of assignments
- environmental impact/cost of anesthetic depending on agent(gas) used, Fresh Gas Flow, and duration of surgery

1.3 An applied take

In this section the main concepts are introduced, which will later discussed in more general scenarios. The simple linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

In which:

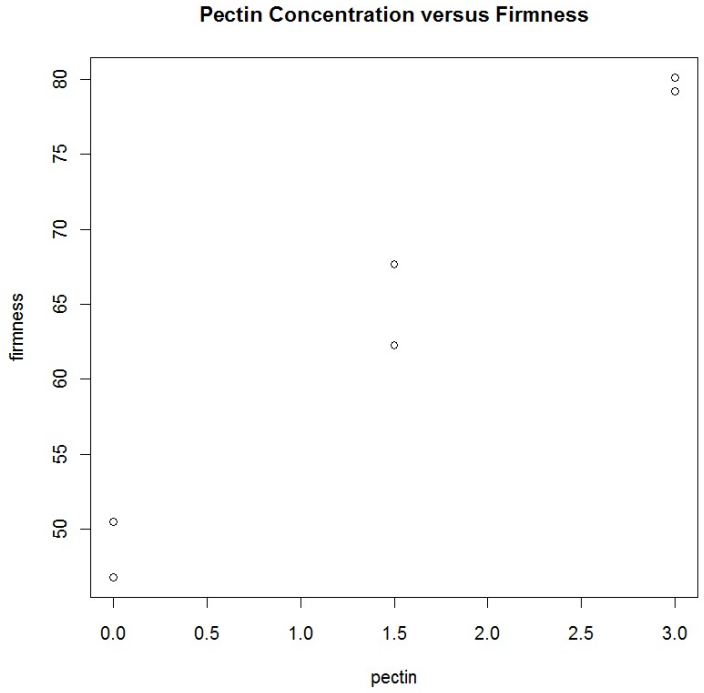
- y is the response variable
- x is the only explanatory variable
- β_0 is the intercept of the population regression line
- β_1 is the slope if the population regression line
- ε is the error, explaining the deviation of the measurements in the population from the regression line

When using inferential statistics ε is assumed to be normally distributed with mean zero and standard deviation σ.

Example 1.1.
From Ott, Longnecker: An Introduction to Statistical Methods and Data Analysis (1977)
A food processor conducted an experiment to study the relationship between concentration of pectin on the firmness of canned sweet potatoes. For each pectin concentration of 0%, 1.5%, and 3% they chose randomly 2 cans which after 30 days of canning at 25 degrees the firmness of the potatoes were measured.
The measurements are in the following table:

pectin	firmness
0	50.5
0	46.8
1.5	62.3
1.5	67.7
3.0	80.1
3.0	79.2

A scatter plot of the data shows how changes in the pectin concentration affect the firmness of the sweet potatoes.



R-code for creating the scatterplot:

```
pectin<-c(0,0,1.5,1.5,3,3)
firmness<-c(50.5,46.8,62.3,67.7,80.1,79.2)
(Pectin.data<-data.frame(pectin,firmness))
with(Pectin.data,plot(pectin,firmness, main="Pectin Concentration versus Firmness"))
```

Let Y be the firmness, and x the pectin concentration then motivated by the diagram we propose a model for the population which claims that Y is a linear function in x + error:

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{1.1}$$

with the error, ε, being normally distributed with mean 0 and standard deviation σ, regardless of the value of x.
Later we will see how we can use the data to gain insight into the population by estimating the parameters in the equation, and conducting tests.

Definition 1.1.
Equation (1.1) together with the assumption that the error is normally distributed with mean 0 and common variance σ^2 defines the *Simple Linear Regression Model*.

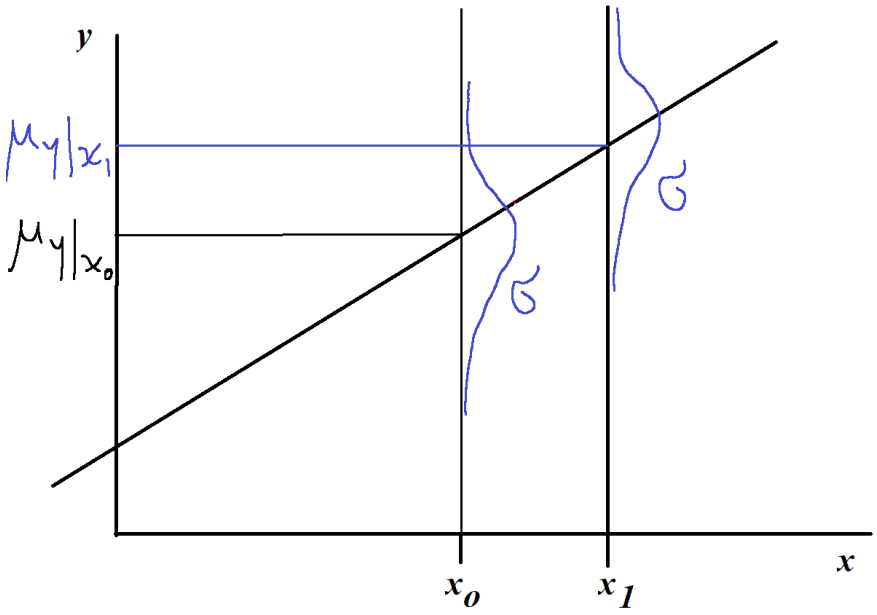
Conclusion:
The model implies that for values of x the conditional expectation of Y given x is

$$\mu_{Y|x} = E(Y|x) = E(\beta_0 + \beta_1 x + \varepsilon) = E(\beta_0 + \beta_1 x) + E(\varepsilon) = \beta_0 + \beta_1 x \tag{1.2}$$

since $\beta_0 + \beta_1 x$ is deterministic and $E(\varepsilon) = 0$.
The conditional variance is

$$Var(Y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = Var(\varepsilon) = \sigma^2 \tag{1.3}$$

Equations (1.2) and (1.3) indicate that a line describes the *mean* of the response variable, Y, in dependency on the regressor, and that the measurements are scattered around the line according to a normal distribution with the *same* variance for all x.



Procedure:
After proposing a model, the model parameters will estimated using sample data, usually followed by statistical tests regarding the values of the parameters or usefulness of the model. The model assumptions have to hold and will have to be confirmed for the tests to be valid. If violated the model must be modified, and a new analysis started, until a satisfactory model has been found.
The final model can then be interpreted to give insight into the population, and if fitting well enough be used for predicting future measurements.

Continue Example 2.2.1. For illustrating the estimation process use the example above. The R-code for obtaining and printing the estimates for the parameters of the simple linear regression model proposed are:

```
model<-lm(firmness~pectin, data=Pectin.data)
model
```

producing the following output:

```
Call:
lm(formula = firmness ~ pectin, data = Pectin.data)

Coefficients:
(Intercept)       pectin
      48.93        10.33
```

Indicating that the "best" estimate for the intercept β_0 is $\hat{\beta}_0 = 48.93$, saying that the estimated mean firmness after 30 days when using 0 pectin is 48.93 MPa, and the "best" estimate for the slope β_1 is $\hat{\beta}_1 = 10.33$ indicating, that for every extra percent of pectin added to the potatoes the firmness increases in average by 10.33 MPa.
The estimated regression line is $\hat{y} = 48.93 + 10.33x$

At this point an analysis of model fit should follow. How this is done will be introduced in the following chapters.

2 Simple Linear Regression

2.1 The model

Definition 2.1.
A random variable Y fits a *Simple Linear Regression Model*, iff there exist $\beta_0, \beta_1 \in \mathbb{R}$ so that for all $x \in \mathbb{R}$

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Corollary:
If Y fits a *Simple Linear Regression Model*, then for a fixed value of $x \in \mathbb{R}$ the conditional expectation of Y given x equals

$$E(Y|x) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x \tag{2.2}$$

and the conditional variance of Y given x is

$$Var(Y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = Var(\varepsilon) = \sigma^2 \tag{2.3}$$

β_0 and β_1 are called the regression coefficients, and are the parameters of the model.

This implies that the mean of Y depends on x in a linear fashion and the variance of Y is independent from the value of x, which is called homoscedasticity.

2.2 Estimation

This section demonstrates how sample data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ can be used to estimate the values of β_0 and β_1 of the simple linear regression model, and properties of the estimators are discussed.
In this context we view the sample data $y_1, \ldots, y_n$ as n measurements from the population described by the linear regression model, or as n realizations of random variables $Y_1, \ldots, Y_n$ fitting the linear regression model.

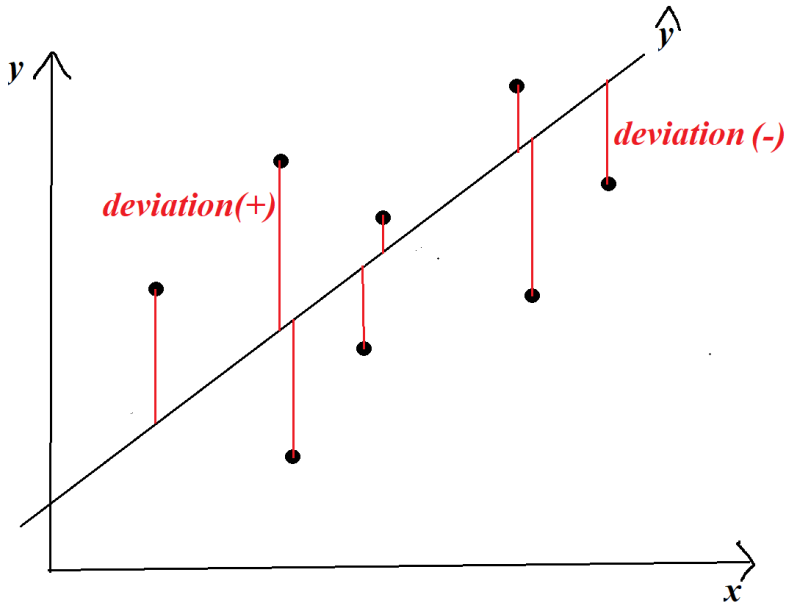
The least-squares estimators for the regression coefficients will be introduced and only at a later time consider Maximum Likelihood estimators.

2.2.1 Point estimators

The least-squares estimators have the property that the total squared vertical distances of the measurements to the least squares line are minimal, i.e. the function

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{2.4}$$

assumes its minimum for the least-square estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.



Theorem 2.1.
The least square estimators for the simple linear regression model are

$$\hat{\beta}_1 = \frac{SS_{xY}}{SS_{xx}}, \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{2.5}$$

with

$$SS_{xx} = \left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right), \quad \text{and} \quad SS_{xY} = \left(\sum_{i=1}^{n} x_i Y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n} \right)$$

Proof:

Assume data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
The least square estimators are the solution of the system of equations obtained by putting the partial derivatives of S to zero:

$$\frac{\partial S}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

and

$$\frac{\partial S}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Simplification of the first equation:

$$\begin{aligned} 0 &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \end{aligned}$$

which is equivalent to:

$$\hat{\beta}_0 = \left(\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \right) /n \tag{2.6}$$

which gives $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, as claimed.
Simplification of the second equation:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i \\ &= \sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{aligned} \tag{2.7}$$

Plucking (2.6) into (2.7) gives

$$\begin{aligned} 0 &= \sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} + \hat{\beta}_1 \frac{(\sum_{i=1}^{n} x_i)^2}{n} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &\Leftrightarrow \\ \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right) &= SS_{xy} \\ &\Leftrightarrow \\ \hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} \end{aligned}$$

Therefore for given data points the least square solution is $\hat{\beta}_1 = \frac{SS_{xY}}{SS_{xx}}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$, providing the estimators as claimed above.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least-squares estimators of β_0 and β_1 and

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.8}$$

is called the least-squares regression line.

Continue Example (Pectin-firmness data)
To calculate the least squares line describing the relationship between firmness and pectin concentration first find

$$\sum_{i=1}^{n} x_i = 9, \quad \sum_{i=1}^{n} y_i = 386.6, \quad \sum_{i=1}^{n} x_i^2 = 22.5, \quad \sum_{i=1}^{n} y_i^2 = 25,893.72, \quad \sum_{i=1}^{n} x_i y_i = 672.9$$

From these:

$$SS_{xx} = 9, \quad SS_{yy} = 983.79, \quad SS_{xy} = 93$$

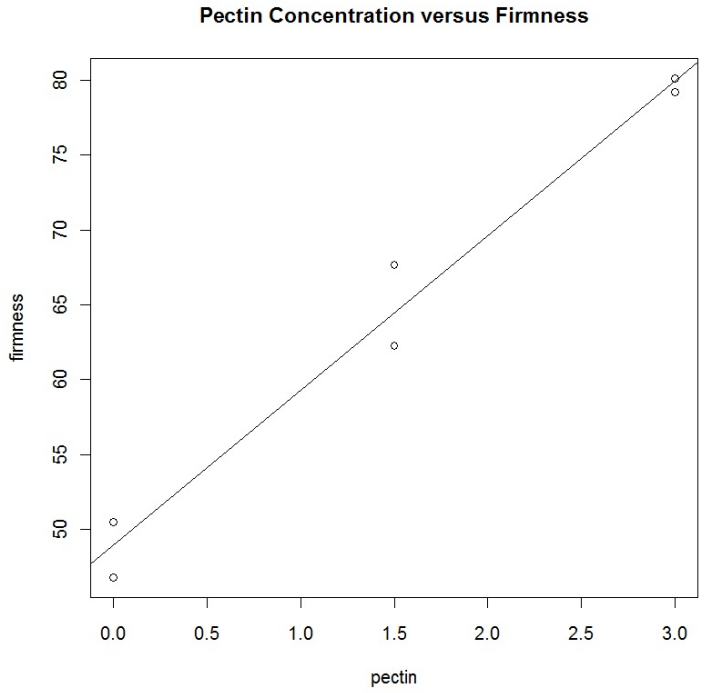
which gives

$$\hat{\beta}_1 = 93/9 = 10.333, \text{ and, } \hat{\beta}_0 = 386.6/6 - 10.333(9/6) = 48.93$$

The least squares regression line is

$$\hat{y} = 48.93 + 10.333x$$

With the interpretation, that when the pectin concentration is 0% then the mean firmness of the potatoes is estimated to be 48.93, and for every extra percent of pectin the firmness increases on average by 10.333.




```
#R code for obtaining estimators and creating plot
pectin<-c(0,0,1.5,1.5,3,3)
firmness<-c(50.5,46.8,62.3, 67.7, 80.1, 79.2)

Pectin.data<-data.frame(pectin,firmness)
model<-lm(firmness~pectin, data=Pectin.data)
summary(model)

with(Pectin.data,plot(pectin,firmness, main="Pectin Concentration versus Firmness"))
abline(model)
```

Properties of the Least Square Estimators

- The estimators are linear in the random variables $Y_i, 1 \leq i \leq n$:

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i, \quad \text{with } c_i = \frac{x_i - \bar{x}}{SS_{xx}}, 1 \leq i \leq n,$$

and

$$\hat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i, \quad \text{with } d_i = 1/n - c_i \bar{x}, 1 \leq i \leq n,$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators for β_0 and β_1, if the model assumptions are met.

  Proof:

$$E(\hat{\beta}_1) = E(\sum_{i=1}^{n} c_i Y_i) = \sum_{i=1}^{n} c_i E(Y_i) = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1$$

  since $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$ (Proof!!!).

  Therefore $\hat{\beta}_1$ is an unbiased estimator for β_1.

  Similar

$$E(\hat{\beta}_0) = \beta_0 \sum_{i=1}^{n} d_i + \beta_1 \sum_{i=1}^{n} d_i x_i = \beta_0$$

  since $\sum_{i=1}^{n} d_i = 1$ and $\sum_{i=1}^{n} d_i x_i = 0$ (Proof!!!).

-

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SS_{xx}}, \text{and} \quad \text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right)$$

  Proof:

$$\text{Var}(\hat{\beta}_1) = \text{Var}\left(\sum_{i=1}^{n} c_i Y_i \right) = \sum_{i=1}^{n} c_i^2 \text{Var}(Y_i) = \sum_{i=1}^{n} c_i^2 \sigma^2 = \frac{\sigma^2}{SS_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sigma^2}{SS_{xx}}$$

  and

$$\text{Var}(\hat{\beta}_0) = \text{Var}\left(\sum_{i=1}^{n} d_i Y_i \right) = \sum_{i=1}^{n} d_i^2 \text{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{x} \right) = \sigma^2 \left(\frac{1}{n} + \frac{1}{n} \bar{x} \sum_{i=1}^{n} c_i + \bar{x}^2 \sum_{i=1}^{n} c_i^2 \right)$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right)$$

- Given the last property the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are, respectively,

$$SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{SS_{xx}}}, \text{and} \quad SE(\hat{\beta}_0) = \sigma \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}} \right)}$$

- **Gauss-Markov-Theorem** (Proof later with multiple regression):
  $\hat{\beta}_0$ and $\hat{\beta}_1$ are the best linear unbiased estimators (BLUEs) for β_0 and β_1 in the linear regression model, when $Y_1, \ldots, Y_n$ are uncorrelated.

  Where "best" means, they are the estimators with the smallest possible variance. Why is this a good thing?

Definition 2.2.
For sample data $(x_1, y_1), \ldots, (x_n, y_n)$ the i^{th} residual is defined as

$$\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Properties of the residuals:

- $\sum_{i=1}^{n} e_i = 0$
- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$
- The centroid $(\bar{x}, \bar{y})$ falls onto the least squares regression line $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
- $\sum_{i=1}^{n} x_i e_i = 0$
- $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

The regression model includes one more parameter, the common variance σ^2

Theorem 2.2.
Let $SS_{Res} = \sum_{i=1}^{n} e_i^2$ then

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

with $SS_{Res} = \sum_{i=1}^{n} e_i^2$ is an unbiased estimator for σ^2.

Proof:
Later with Linear Algebra for Multiple Linear Regression.

Comment:

$$SS_{Res} = SS_{yy} - \hat{\beta}_1 SS_{xy} \tag{2.9}$$

(to see that this is true use $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$)

- $MS_{Res} = \hat{\sigma}^2$ is also the called the Residual Mean square.
- $\hat{\sigma}$ is called the (estimated) standard error of regression, and
- the (estimated) standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SS_{xx}}}, \text{and} \quad se(\hat{\beta}_0) = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}$$

Observe the difference between $SE(\hat{\beta}_0)$ and $se(\hat{\beta}_0)$.

Continue Example 2.2.1.(Pectin-firmness data)
We already found $SS_{yy} = 983.79, SS_{xy} = 93, \hat{\beta}_1 = 10.333,$ and, $\hat{\beta}_0 = 48.93$ Therefore

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2} = \frac{983.79 - (10.333)93}{6-2} = 5.71$$

and $\hat{\sigma} = 2.39$. We estimate that the measurement are spread with a standard deviation of 2.39 around the least squares regression line.

```
# R code producing output including the result

pectin<-c(0,0,1.5,1.5,3,3)
firmness<-c(50.5,46.8,62.3, 67.7, 80.1, 79.2)
Pectin.data<-data.frame(pectin,firmness)

model<-lm(firmness~pectin, data=Pectin.data)
summary(model)
```

As usual point estimators have the critical shortcoming, that the probability to give the true value is 0.
Therefore we usually request interval estimates or confidence intervals.

2.2.2 Confidence intervals for estimating slope, intercept, and variance

Read "On Distributions" before continuing

The confidence intervals are based on the distribution of a standardized scores for the least-square estimates

Theorem 2.3.

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{SS_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} \quad \text{and} \quad t = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_0}{\mathrm{se}(\hat{\beta}_0)}$$

are t-distributed with $df = n - 2$.

Proof idea:
$Y_1, \ldots, Y_n$ are according to definition normally distributed, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations in the Y_i, therefore $\hat{\beta}_0$ and $\hat{\beta}_1$ normally distributed with means β_0 and β_1, respectively, and variances

$$\frac{\sigma^2}{SS_{xx}}, \quad \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)$$

respectively.
Then

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SS_{xx}}} \quad \text{and} \quad Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}}$$

are standard normally distributed.
Also it can be shown that $V = SS_{Res}/\sigma^2 \sim \chi^2(n-2)$.
One can prove that

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SS_{xx}}} \quad \text{and} \quad V = \frac{SS_{Res}}{\sigma^2}$$

and

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}} \quad \text{and} \quad V = \frac{SS_{Res}}{\sigma^2}$$

are independent, which we will do in the more general context of multiple regression.
Then

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

is t-distributed with $df = n - 2$,
and

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}} \times \sqrt{\frac{\sigma^2(n-2)}{SS_{Res}}} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}}$$

is t-distributed with $df = n - 2$.

Corollary:
Let t_p^{df} be the p critical value of a t-distribution with df degrees of freedom, then according to theorem 2.3:

(a) A $(1-\alpha) \times 100\%$ confidence interval for β_0 is given by

$$\hat{\beta}_0 \pm t_{\alpha/2}^{n-2} \text{se}(\hat{\beta}_0)$$

(b) A $(1-\alpha) \times 100\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2}^{n-2} \text{se}(\hat{\beta}_1)$$

Continue Example 2.2.1.(Pectin-firmness data)
95% confidence intervals require the 0.025 critical value of the t-distribution with 4 degrees of freedom, which is $t^* = 2.776$. The standard errors are

$$\text{se}(\hat{\beta}_0) = 2.39\sqrt{\left(\frac{1}{6} + \frac{1.5^2}{9}\right)} = 1.5409 \quad \text{se}(\hat{\beta}_1) = \frac{2.39}{\sqrt{9}} = 0.7957$$

Therefore 95% confidence intervals for intercept and slope are given by

$$50.43 \pm 2.776(1.5409) \text{ and } 10.33 \pm 2.776(0.7957)$$

We are 95% confident that the true intercept falls between 44.7 and 53.2, and that the true slope falls between 8.12 and 12.54.

Interpretation:
We are 95% confident that the mean firmness falls between 44.7 and 53.2 when using no pectin, and that the firmness increases between 8.1 and 12.5, when increasing pectin by 1%.

```
# R code producing confidence intervals for parameters

pectin<-c(0,0,1.5,1.5,3,3)
firmness<-c(50.5,46.8,62.3, 67.7, 80.1, 79.2)
Pectin.data<-data.frame(pectin,firmness)

model<-lm(firmness~pectin, data=Pectin.data)
confint(model)
```

Theorem 2.4.
If the model assumptions are met, then

$$\chi^2 = \frac{(n-2)MS_{Res}}{\sigma^2}$$

is χ^2 distributed with $df = n - 2$

Proof:

It is

$$\frac{(n-2)MS_{Res}}{\sigma^2} = \frac{SS_{Res}}{\sigma^2}$$

which is (see prove above) χ^2 distributed with $df = n-2$, which proves the claim.

Corollary:
Let χ^k_p be the p critical value of a χ^2-distribution with $df = k$, then

$$\left[\frac{(n-2)MS_{Res}}{\chi^{n-2}_{\alpha/2}}, \frac{(n-2)MS_{Res}}{\chi^{n-2}_{1-\alpha/2}}\right]$$

is a $(1-\alpha)\times 100\%$ confidence interval for σ^2.
Proof:
According to theorem 2.4

$$P\left(\chi^{n-2}_{1-\alpha/2} \leq \frac{(n-2)MS_{Res}}{\sigma^2} \leq \chi^{n-2}_{\alpha/2}\right) = 1-\alpha$$

which is equivalent to

$$P\left(\frac{1}{\chi^{n-2}_{1-\alpha/2}} \geq \frac{\sigma^2}{(n-2)MS_{Res}} \geq \frac{1}{\chi^{n-2}_{\alpha/2}}\right) = 1-\alpha$$

which is equivalent to

$$P\left(\frac{(n-2)MS_{Res}}{\chi^{n-2}_{1-\alpha/2}} \geq \sigma^2 \geq \frac{(n-2)MS_{Res}}{\chi^{n-2}_{\alpha/2}}\right) = 1-\alpha$$

which concludes the proof.

Continue Example 2.2.1.(Pectin-firmness data)

A 95% confidence interval for σ^2 is given by

$$\left[\frac{4(5.71)}{11.143}\ ,\ \frac{4(5.71)}{0.484}\right] \longleftrightarrow [2.05\ ,\ 47.19]$$

which gives a 95% confidence interval for σ: $[\sqrt{2.05}, \sqrt{47.19}] = [1.43\ ,\ 6.87]$.

```
# R code for calculating confidence interval for the error variance

# lower bound:
sum(model$residuals^2)/qchisq(0.975,df=4)
# upper_bound
sum(model$residuals^2)/qchisq(0.025,df=4)
```

2.3 Tests concerning the parameters

Also based on theorem 2.3 statistical tests for the model parameters can be derived. The test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is of particular importance, is called model utility test, since with this test we can show the usefulness of the model. If $\beta_1 = 0$ x is not linearly related with Y and the model is not of further interest.

t-test for the slope in the simple linear regression model

1. Hypotheses: Parameter of interest β_1.

test type	hypotheses
upper tail	$H_0 : \beta_1 \leq \beta_{10}$ vs. $H_a : \beta_1 > \beta_{10}$
lower tail	$H_0 : \beta_1 \geq \beta_{10}$ vs. $H_a : \beta_1 < \beta_{10}$
two tail	$H_0 : \beta_1 = \beta_{10}$ vs. $H_a : \beta_1 \neq \beta_{10}$

Choose α.

2. Assumptions: Random sample, regression model is appropriate

3. Test statistic:
$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\text{se}(\hat{\beta}_1)}, \quad df = n - 2$$

4. P-value:

test type	P-value
upper tail	$P(t > t_0)$
lower tail	$P(t < t_0)$
two tail	$2P(t > abs(t_0))$

5. Decision: If $P - value < \alpha$ reject H_0, otherwise do not reject H_0.

6. Context:

Continue Example 2.2.1(Pectin-firmness data)
Model utility test for the simple linear regression model relating the firmness of sweet potatoes with the percentage of pectin added.

1. Hypotheses: Parameter of interest β_1. $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, $\beta_{10} = 0$. $\alpha = 0.05$.

2. Assumptions: Random sample (the description suggests it is a random sample), regression model is appropriate, for checking this assumption we will learn how to do a residual analysis.

3. Test statistic:
$$t_0 = \frac{10.33 - 0}{0.7957} = 12.99, \quad df = 4$$

4. P-value: This is a two-tailed test, therefore P-value= $2P(t > abs(t_0))$, or 2 times the area to the right of $t_0 = 12.99$. According to the t-distribution table, the area is smaller than 0.005.
$P - value < 2(0.005) = 0.01$.

5. Decision: $P - value < 0.01 < 0.05$, reject H_0.

6. Context: At significance level of 5% the data provide sufficient evidence that the mean firmness of the sweet potatoes depends on the percentage of pectin used in the canning process.

```
 \begin{verbatim}
# R code producing test results, but you have to interpret the output!

pectin<-c(0,0,1.5,1.5,3,3)
firmness<-c(50.5,46.8,62.3, 67.7, 80.1, 79.2)
Pectin.data<-data.frame(pectin,firmness)

model<-lm(firmness~pectin, data=Pectin.data)
summary(model)
```

Usually of less interest:
t-test for the intercept in the simple linear regression model

1. Hypotheses: Parameter of interest β_0.

test type	hypotheses
upper tail	$H_0 : \beta_0 \leq \beta_{00}$ vs. $H_a : \beta_0 > \beta_{00}$
lower tail	$H_0 : \beta_0 \geq \beta_{00}$ vs. $H_a : \beta_0 < \beta_{00}$
two tail	$H_0 : \beta_0 = \beta_{00}$ vs. $H_a : \beta_0 \neq \beta_{00}$

Choose α.

2. Assumptions: Random sample, regression model is appropriate

3. Test statistic:
$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\text{se}(\hat{\beta}_0)}, \quad df = n - 2$$

4. P-value:

test type	P-value
upper tail	$P(t > t_0)$
lower tail	$P(t < t_0)$
two tail	$2P(t > abs(t_0))$

5. Decision: If $P - value < \alpha$ reject H_0, otherwise do not reject H_0.

6. Context:

Continue Example 2.2.1.(Pectin-firmness data)
Test if the intercept is different from 50.

1. Hypotheses: Parameter of interest β_0. $H_0 : \beta_0 = 50$ vs. $H_a : \beta_0 \neq 50$, $\beta_{10} = 50$. $\alpha = 0.05$.

2. Assumptions: Random sample (the description suggests it is a random sample), regression model is appropriate, for checking this assumption we will learn how to do a residual analysis.

3. Test statistic:
$$t_0 = \frac{50.43 - 50}{1.5409} = 0.279, \quad df = 4$$

4. P-value: This is a two-tailed test, therefore P-value= $2P(t > abs(t_0))$, or 2 times the area to the right of $t_0 = 0.279$. According to the t-distribution table, the area is larger than 0.10.
$P - value > 2(0.10) = 0.20.$

5. Decision: $P - value > 0.20 > 0.05 = \alpha$, do not reject H_0.

6. Context: At significance level of 5% the data do not provide sufficient evidence that the mean firmness of the sweet potatoes when using no pectin is different from 50.

Analysis of Variance:
The model utility test can be also approached through an analysis of variance (ANOVA) in the response variable. Here we are analysing (taking apart) the sample variance of the response variable observations, $y_1, \ldots, y_n$.
When conducting an ANOVA the different sources of the variance in the response variable are considered. In regression analysis the potential sources are the regression, i.e. the dependency of the response variable on the predictor, and the error (residual), i.e. the fact that for fixed values of the predictor variable the response varies (for example not all people of the height(predictor) have the weight(response)).
It all starts with observing
$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \tag{2.10}$$
From here we get:
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.11}$$

Proof:
When squaring equation 2.10 and totalling for all observations we get

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\hat{y}_i(y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i)) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\underbrace{\sum_{i=1}^{n}\hat{y}_i e_i}_{=0} - 2\bar{y}\underbrace{\sum_{i=1}^{n} e_i}_{=0} + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2
\end{aligned}$$

The left side of 2.11 is the total sum of squares, SS_T, measuring the total variation present in the measurements for the response variable y.

The first sum on the right side, SS_R, measures how much of the variation can be accounted for through the regression model, and the second sum, SS_{Res}, is the residual variation which can not be accounted for through the regression model.
We get

$$SS_T = SS_R + SS_{Res} \tag{2.12}$$

Observe, that $SS_T = SS_{yy}$, and that therefore $SS_R = \hat{\beta}_1 SS_{xy}$ (see equation 2.9).
A similar equation holds for the degrees of freedom

$$df_T = df_R + df_{Res} \tag{2.13}$$

with $df_T = n - 1$, $df_R = 1$, and $df_{Res} = n - 2$
In order to assess, if the regression model is explaining any of the variation in y, i.e. if $\beta_1 \neq 0$, use the F-score

$$F = \frac{MS_R}{MS_{Res}} = \frac{SS_R/df_R}{SS_{Res}/df_{Res}}$$

as we will later see F is F-distributed. In the context of multiple regression the equivalent test will be more relevant than here.

```
# R code for calculating the sum of squares

anova(model)
```

2.4 Estimation of the mean response

In some applications, beyond investigating the relationship, one aims to use the data and model to estimate the mean of the response for given values of the predictor.
A typical example would be models for the real estate market where property prices are estimated based on size, living space, age, number of bedrooms, etc.

A point estimator for the mean response given the predictor is x_0 can be found from the least squares line

$$\widehat{E(Y|x_0)} = \hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Properties of the estimator $\hat{\mu}_{Y|x_0}$:

- linear in $Y_1, \ldots, Y_n$, because $\hat{\beta}_0$ and $\hat{\beta}_1$ linear in $Y_1, \ldots, Y_n$,
- normally distributed because it is linear in normally distributed random variables,
- with mean $E(\hat{\mu}_{Y|x_0}) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_0 = \beta_0 + \beta_1 x_0 = \mu_{Y|x_0} = E(Y|x_0)$,
- and with variance

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{Y|x_0}) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \mathrm{Var}(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0) \\
&= \mathrm{Var}(\bar{Y} + \hat{\beta}_1 (x_0 - \bar{x})) \\
&=^* \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{SS_{xx}} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right).
\end{aligned}$$

  Using at $=^*$ that $Cov(\bar{Y}, \hat{\beta}_1) = 0$.

Thus, under the assumption of the simple linear regression model

$$t = \frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}\right)}}$$

is t-distributed with $df = n - 2$.

Corollary:
A $(1 - \alpha) \times 100\%$ confidence interval for $E(Y|x_0)$ is given by

$$\hat{\mu}_{Y|x_0} \pm t_{\alpha/2}^{n-2} \, \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \right)}$$

Continue Example 2.2.1 (Pectin-firmness data)
Estimate the mean firmness of potatoes for $x_0 = 2\%$ of pectin using a 95% confidence interval.

$$[48.93 + 10.333(2)] \pm 2.776(2.39)\sqrt{\left(\frac{1}{6} + \frac{(2-1.5)^2}{9}\right)}$$

gives 69.596 ± 2.938, gives $[66.658, 72.534]$. We are 95% confident that the mean firmness of potatoes falls between 66.7 and 72.5, when using 2% of pectin.

```
#R code for calculating confidence interval for mean response
new<-data.frame(pectin=2,firmness=0)
predict(model,new,interval="confidence",level=0.95)
```

2.5 Prediction of new observations

In other applications, one aims to use the data and model to predict future values of the response, i.e. predicting values of y when $x = x_0$. Since a normal error is assumed, the range of possible values of y would have to be $-\infty$ to ∞ for all x_0. Therefore the goal is to predict the middle $(1-\alpha) \times 100\%$ of y values when $x = x_0$.

Mathematically this calls for an interval so that the probability for $Y_0 = Y|x_0$ to fall within equals $1 - \alpha$.
$\hat{y}_0 = \hat{Y}|x_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is an unbiased point estimator for a future value $Y_0 = Y|x_0$. and

$$\Psi = Y_0 - \hat{y}_0$$

is normally distributed with mean zero and variance

$$\text{Var}(\Psi) = Var(Y_0 - \hat{y}_0) = Var(Y_0) + Var(\hat{y}_0) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}\right)$$

We can add the variances because Y_0 and $\hat{y}_0$ are independent.
Therefore

$$t = \frac{Y_0 - \hat{y}_0}{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}\right)}$$

is t-distributed with $df = n - 2$.

Corollary:
A $(1-\alpha) \times 100\%$ prediction interval for $Y|x_0$ is given by

$$[\hat{\beta}_0 + \hat{\beta}_1 x_0] \pm t_{\alpha/2}^{n-2}\ \hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}\right)}$$

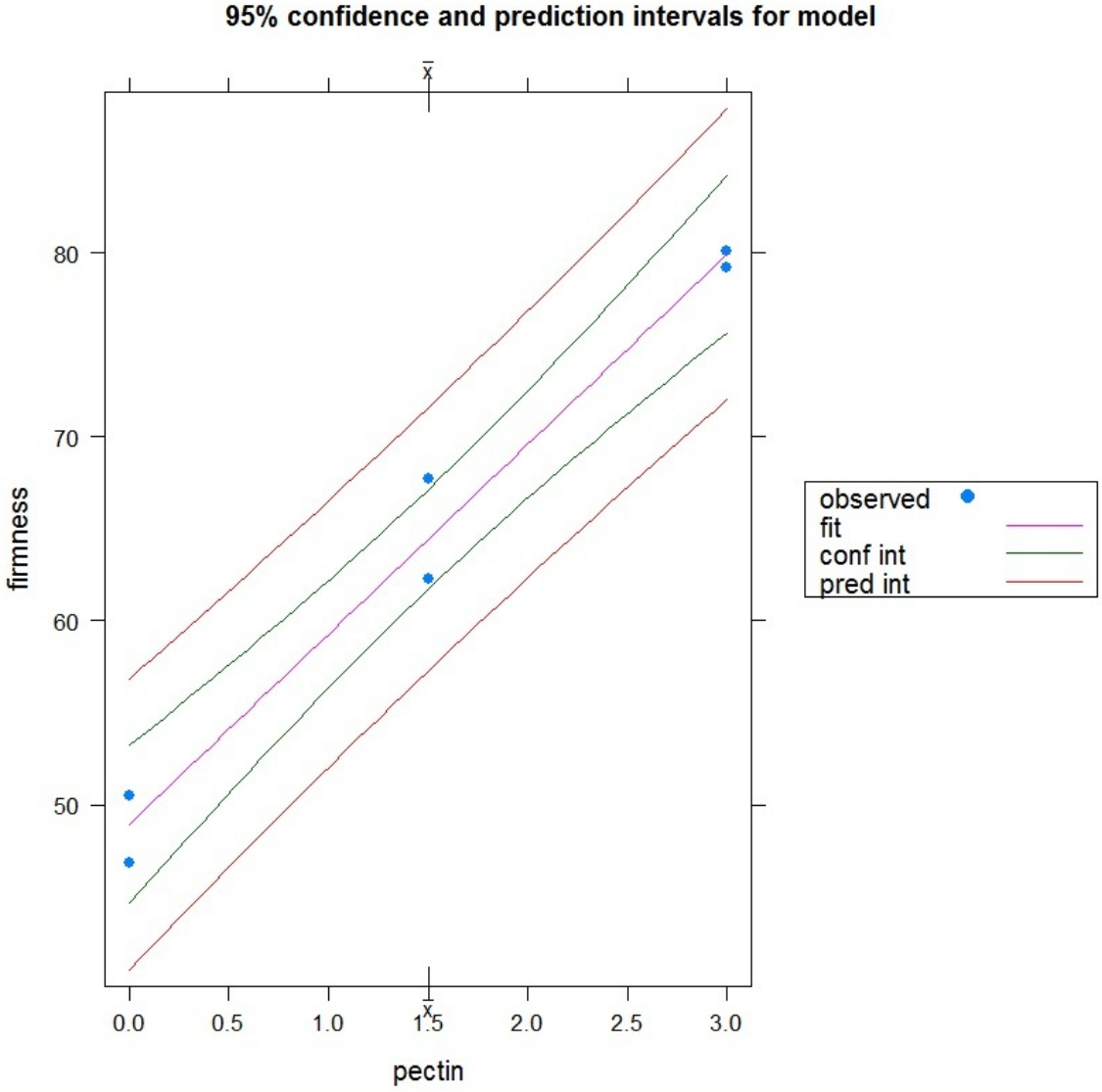
Proof:

$$P\left(\hat{y}_0 - t_{\alpha/2}^{n-2}\,\hat{\sigma}\sqrt{\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}\right)} \le Y_0 \le \hat{y}_0 + t_{\alpha/2}^{n-2}\,\hat{\sigma}\sqrt{\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}\right)}\right) = 1-\alpha$$

Continue Example 2.2.1. (Pectin-firmness data)
Predict the firmness of potatoes for $x_0 = 2\%$ of pectin using a 95% prediction interval.

$$[48.93 + 10.333(2)] \pm 2.776(2.39)\sqrt{\left(1+\frac{1}{6}+\frac{(2-1.5)^2}{9}\right)}$$

gives 69.596 ± 7.251, gives $[62.345, 76.847]$. We predict that 95% of potatoes have a firmness between 62.3 and 76.8, when using 2% of pectin.



```
#R code for calculating prediction interval and creating plot
new<-data.frame(pectin=2,firmness=0)
predict(model,new,interval="prediction",level=0.95)

library(HH)
ci.plot(model)
```

2.6 Model fit

One way to measure model fit is through the Coefficient of Determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

The SS_{Res} is the amount in the total Sum of Squares in y, which remains unexplained by the model. Therefore the R^2 is the proportion of the variance in y which can be explained by the model.
For the example $R^2 = 0.9768$. It seems as if the firmness is almost totally dependent on the pectin concentration.
Caution:
If the there is little variation in the predictor/regressor variable, i.e. $\beta_1^2 SS_{xx}/(n-1)$ is small relative to σ^2, then R^2 tends to be small. this can be observed from a result by Hahn(1973), who showed that

$$E(R^2) \approx \frac{\beta_1^2 SS_{xx}/n - 1}{\beta_1^2 SS_{xx}/(n-1) + \sigma^2}$$

R^2 does not indicate if the model is appropriate, but if it is a useful model how much of the variation is captured by it.

```
#R code for creating output including R^2
summary(model)
```

2.7 Potential Problems with Regression

1. non linear
2. outliers.
3. violation of homoscedasticity assumption.
4. multicollinearity

3 Maximum Likelihood

In the last chapter estimation by least-squares was introduced. In this chapter Maximum-Likelihood estimators for the simple linear regression model will be introduced and discussed.

Read "On maximum likelihood estimation"

To determine the Maximum Likelihood (ML) estimator first obtain the likelihood or log likelihood function.
Assuming that the error for the n data points, (x_i, y_i) , $1 \leq i \leq n$ is NID(0,σ^2) (NID means: normal, identical, and independent), the density function of the normal distribution is needed.

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Since the mean of y_i is $\mu_{y_i} = \beta_0 + \beta_1 x_i$ and variance $\sigma^2_{y_i} = \sigma^2$ the likelihood for point each (x_i, y_i), $1 \leq i \leq n$ is

$$f(y_i, \mu_{y_i}, \sigma^2_{y_i}) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

Since we have to find the likelihood for the entire sample which is the likelihood for independent measurements (x_1, y_1) and (x_2, y_2) and $\ldots$ and (x_n, y_n).
(Probabilities for independent events multiply)

$$\begin{aligned} L((y_i), (x_i), \beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n} -(y_i - \beta_0 - \beta_1 x_i)^2\right) \end{aligned}$$

For this function we now have to find the values for β_0, β_1 and σ^2 for which L is the largest. Those are then called the maximum likelihood estimators $\tilde{\beta}_0, \tilde{\beta}_1$ and $\tilde{\sigma}^2$.
This is much easier when we first use the logarithm of the function for the log likelihood:

$$lnL((y_i), (x_i), \beta_0, \beta_1, \sigma^2) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

The ML estimates are the roots of the system of partial derivatives of the log likelihood function.

$$\frac{\partial ln(L)}{\partial \beta_0} = -\left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = \frac{n}{\sigma^2}\left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i\right)$$

and

$$\frac{\partial ln(L)}{\partial \beta_1} = -\left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = \frac{n}{\sigma^2}\left(\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2\right)$$

and

$$\frac{\partial ln(L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \left(\frac{1}{2\sigma^4}\right)\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Claim: The solution to the system

$$\frac{\partial ln(L)}{\partial \beta_0} = \frac{\partial ln(L)}{\partial \beta_1} = \frac{\partial ln(L)}{\partial \sigma^2} = 0$$

is given by

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum x_i^2}, \quad \tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}, \text{ and } \tilde{\sigma^2} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

Proof: Homework

Comment: Interestingly the ML estimator is the same as the Least-Squares estimator