STAT 308 – Chapter 5

Background Information

Suppose we have n observations from two random variables X and Y (i.e. we have pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$). We believe that we can quantify the variation of the **dependent variable** Y by our knowledge of the **independent variable** X.

Important Definitions

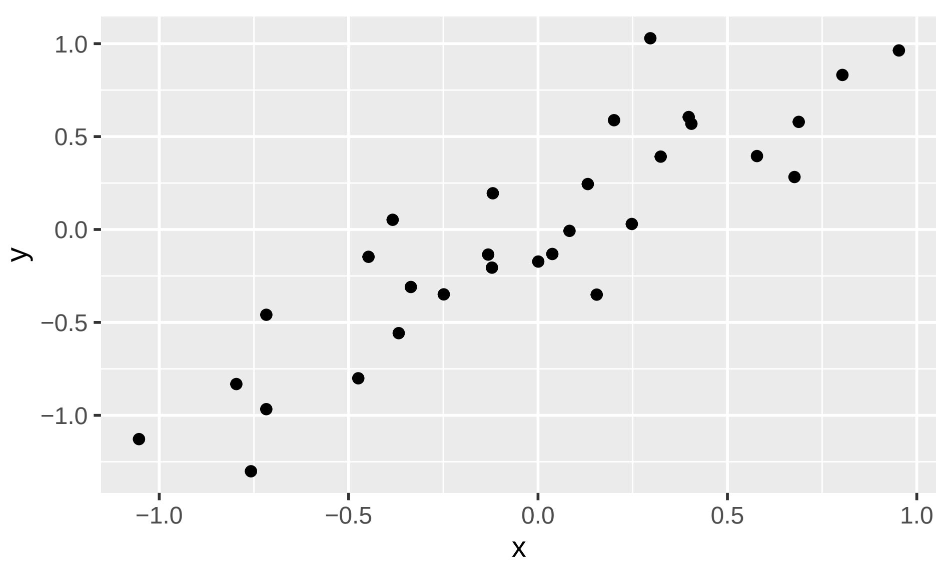
Scatter Diagram (Plot): A plot of the observations from the independent variable $(x_1, \ldots, x_n)$ against the observations from the dependent variable $(y_1, \ldots, y_n)$.

Information we can gather from scatterplot

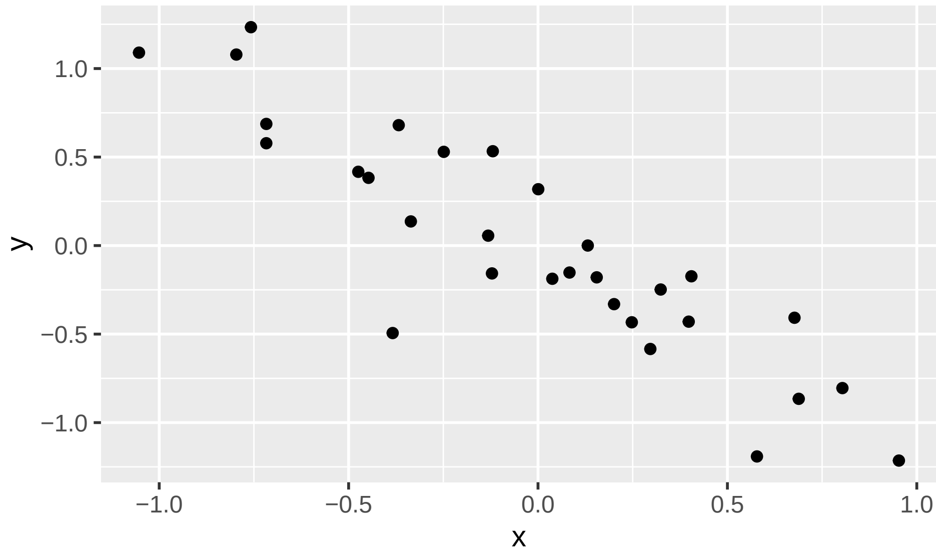
```
## Warning: package 'fGarch' was built under R version 4.2.1

## NOTE: Packages 'fBasics', 'timeDate', and 'timeSeries' are no longer
## attached to the search() path when 'fGarch' is attached.
##
## If needed attach them yourself in your R script by e.g.,
##         require("timeSeries")
```

Direction:

- Positive - observed y tend to get larger as x increases

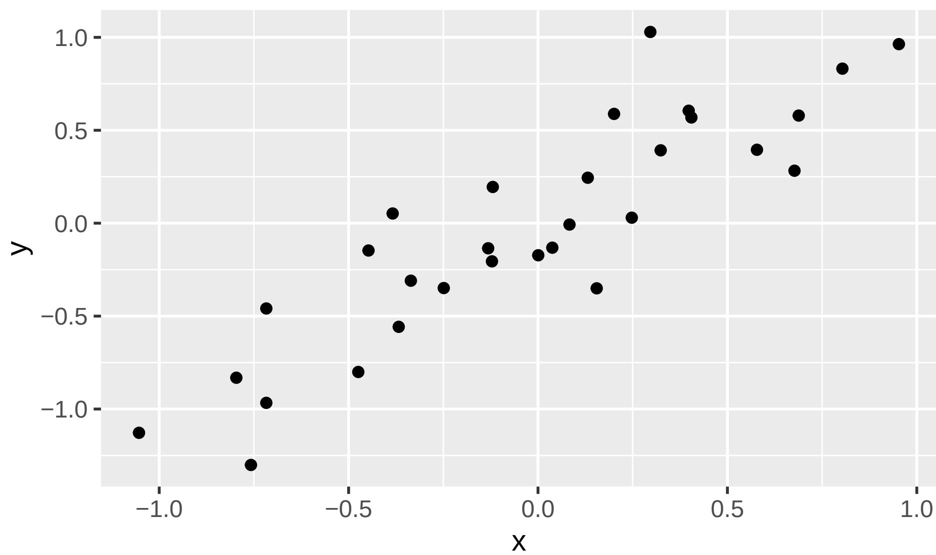


- Negative - observed y tend to get smaller as x increases

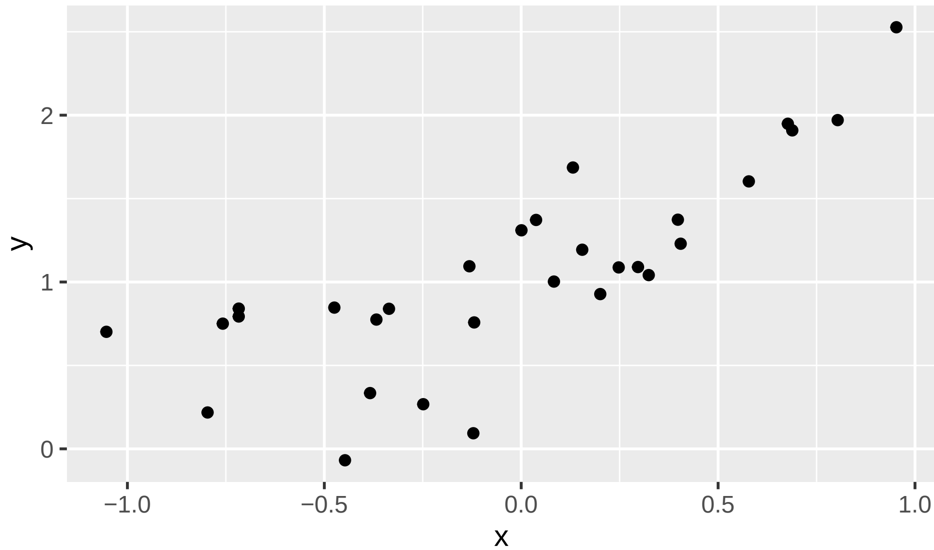


Form:

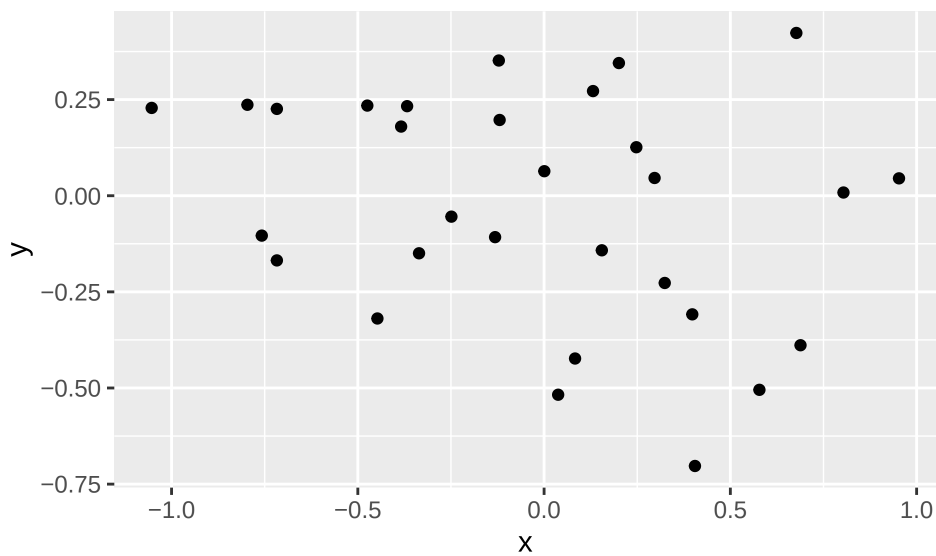
- Linear - Can reasonably make out a straight line pattern from the data



- Non-linear - Can reasonably make out a pattern that is non-linear (parabolic, exponential, logaritmic, etc.)

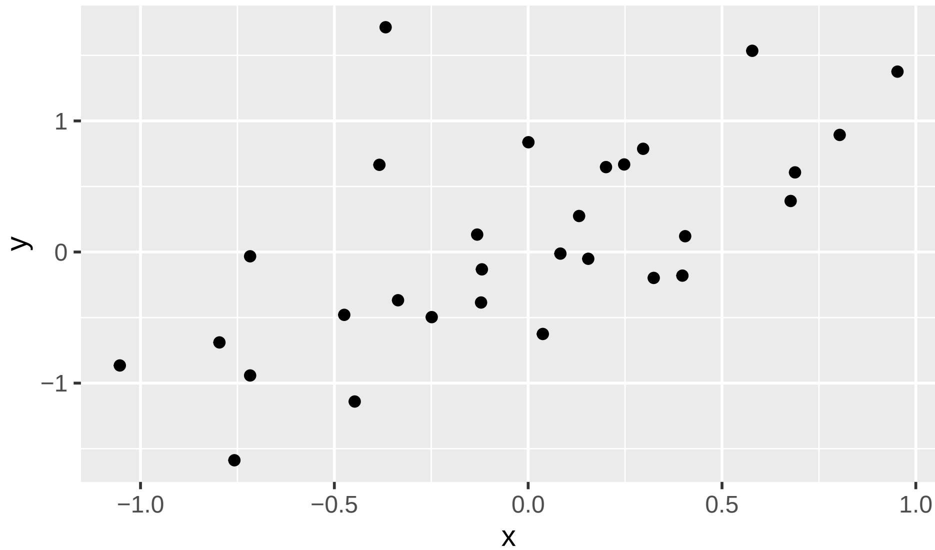


- None - No reasonable pattern can be detected

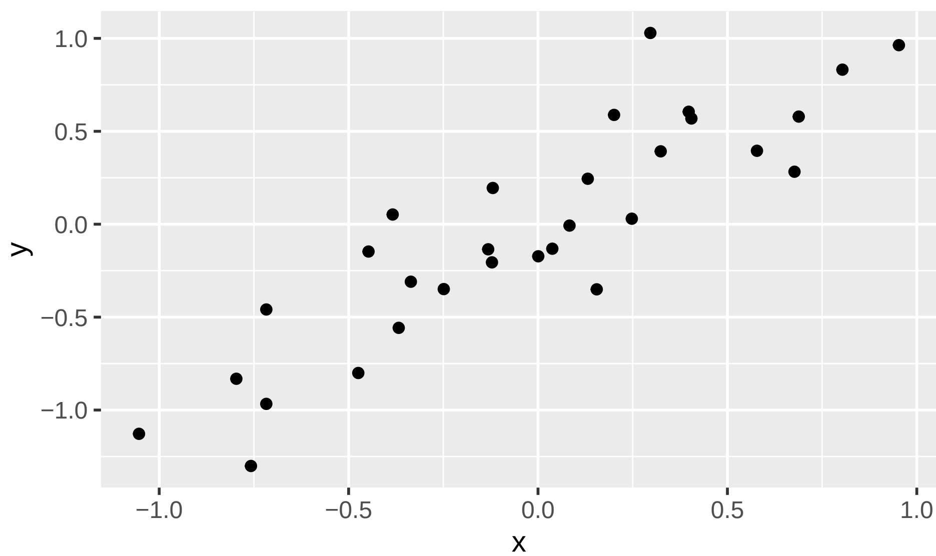


Strength:

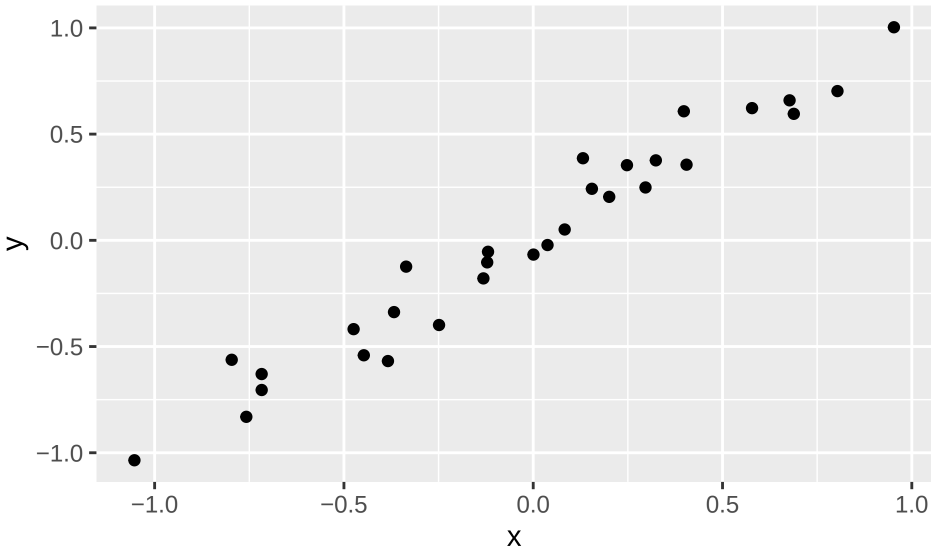
- Weak - Pattern is not very pronounced



- Moderate - Pattern is slightly more pronounced



- Strong - Pattern is highly pronounced



Potential Outliers - points in the scatterplot that deviate highly from the rest of the data

Example

Consider the blood pressure dataset, where we are interested quantifying the variation of systemic blood pressure based on the subjects' age.

Draw a scatterplot of systolic blood pressure against age.

```
bloodpressure <- read.csv("../Data/bloodpressure.csv")
```

Describe the scatterplot.

Now that we have analyzed the scatterplot, we need to answer the following questions:

- What is the appropriate mathematical model to use – straight line, logarithmic function, exponential function, etc.?
- Given a specific model form, what criteria do we use and how do we obtain the best fitting line to the data?

We will start by answering these questions for a **straight line dataset with one explanatory (independent) variable**. We will then use this to expand into non-linear models with more than one potential explanatory variable.

Straight Line Model

Mathematically, a straight line is defined as

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

where

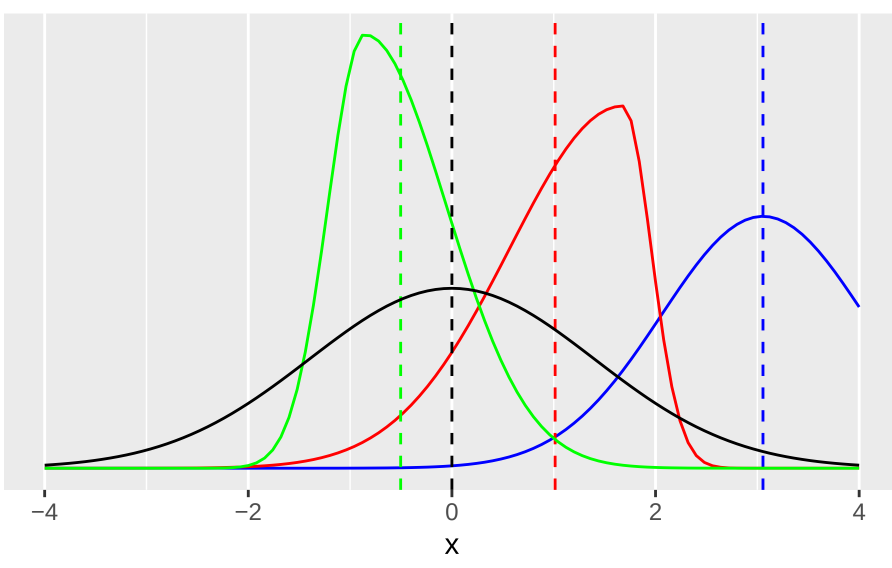
- β_0 is the intercept – the value of y when $x = 0$
- β_1 is the slope – the change in y for a one unit change in x

Let's say that we are tentatively assuming a straight line model for our given dataset.

Assumptions Needed for Linear Models

For all the plots in this section, note that green: $x = -1$, black: $x = 0$, red: $x = 1$, and blue: $x = 2$. In addition, the dashed vertical lines represent the mean of that distribution.

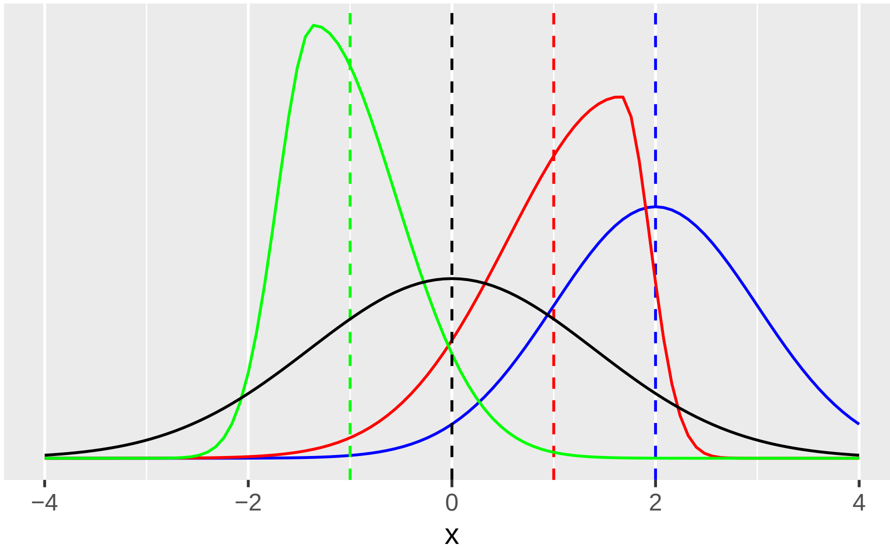
- Existence: For any given value of X, Y is a random variable with a certain probability distribution with a finite mean and variance. Define:
  - $\mu_{Y|X}$ - the population mean of Y for a fixed X
  - $\sigma^2_{Y|X}$ - the population variance of Y for a fixed X



- Independence: The observed values of Y are statistically independent of one another given X Counterexample:
  - X = Amount of
- Linearity: $\mu_{Y|X}$ is a straight line function of X. In other words we say that

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

  where β_0 and β_1 are defined here as the population intercept and slope, respectively.



However, there is still some difference between the random variable Y and its mean $\mu_{Y|X}$ that we have yet to account for. Therefore, the complete linear model is now typically expressed as complete statistical linear model

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

where ϵ is a random variable with zero mean $\mu_{\epsilon|X} = 0$.

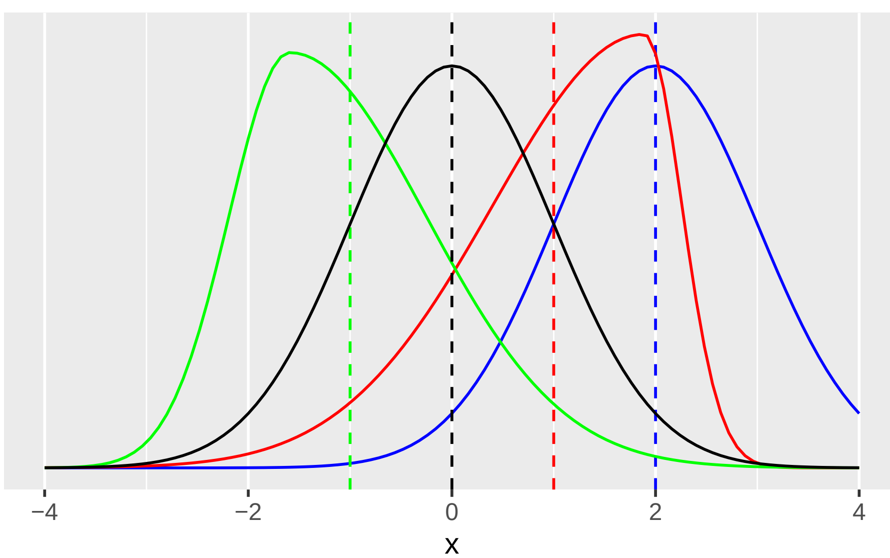
ϵ is commonly referred to as the **errors/residuals** of the linear model.

The next two assumptions discuss the distribution of ϵ.

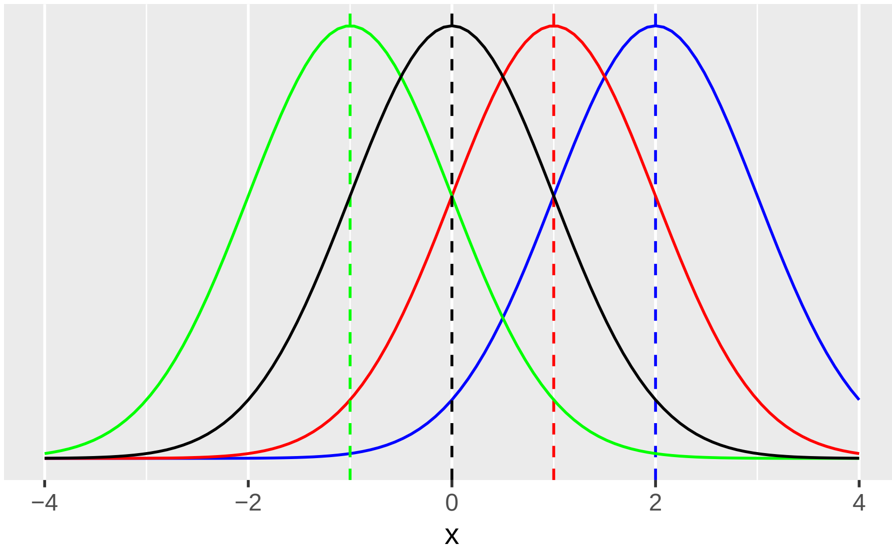
- Homoscedasticity: The variance of Y is the same for different given values of X. Mathematically, this is equivalent to saying

  $$\sigma^2_{Y|X} = \sigma^2.$$

  , or in other words $\sigma^2_{Y|X_i} = \sigma^2_{Y|X_j}$ for different i and j.



- Normality: For any fixed value of X, Y is normally distributed. This fact makes analysis of the data easier.



All of these assumptions put together lead us to our full mathematical model we will assume:

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

.

We will discuss checking the validity of these assumptions later in the semester.

"Least Squares" Regression Model

Define $\hat{\beta}_0$ and $\hat{\beta}_1$ as the linear regression estimates of β_0 and β_1, respectively. How can we choose the "best" estimates of these population parameters?

The most obvious is to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the average error is zero. That is

$$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)).$$

What is the main issue with this method?

The "least squares" method provides estimates that minimizes the sum of the squared differences between observed y and its estimates from the regression line. In other words, the least squares methods finds $\hat{\beta}_0$ and $\hat{\beta}_1$ that satifisies

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

The least squares method produces the estimates

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

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Example:

Suppose I obtain data with the following information:

$$n = 20, \sum x_i = 40, \sum y_i = 230, \sum x_i^2 = 100, \sum y_i^2 = 2500, \sum x_i y_i = 500.$$

Find $\hat{\beta}_0$ and $\hat{\beta}_1$.

These estimates can be combined to provide an estimate for Y for a given value $X = x$,

$$\hat{Y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

.

Performing Least Squares regression in R

Example

According to the least squares regression line, what do we expect a randomly selected 50 year old's systolic blood pressure to be?

What do we expect a randomly selected 80 year old's systolic blood pressure to be?

How do outliers impact the Least Squares regression line?

Estimating the variance $\sigma^2_{Y|X}$

Recall that, without knowledge of X, our estimate of the variance of Y is

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

The summation can be thought of as the sum of squared distance between an observed y_i and its prediction $\bar{y}$, or in other words, the **sum of squared errors**. The estimate for the variance of the linear regression model, $\sigma^2_{Y|X} = \sigma^2$ is calculated in a similar manner

$$s^2_{Y|X} = \frac{1}{n-2} SSE = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

Note that, for $s^2_{Y|X}$ we divide the sum of squared errors by $n-2$, whereas for s_Y^2 we divide the sum of squared errors by $n-1$.

Inference on linear regression model

Recall from Chapter 3, we said that

$$\frac{\bar{Y} - \mu}{s_Y^2} \sim t_{df=n-1}$$

where $\bar{Y}$ is the random variable associated with our estimate of μ ($\bar{y}$). We can obtain a similar conclusion regarding the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} \sim t_{df=n-2}$$

and

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{df=n-2}$$

where $s^2_{\hat{\beta}_0} = s^2_{Y|X}\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s^2_X}\right)$ and $s^2_{\hat{\beta}_1} = \frac{s^2_{Y|X}}{(n-1)s^2_X}$ (Note: Often times, inference on β_0 is not meaningful to us.)

Confidence intervals for β_1

A $C = 100 \times (1-\alpha)\%$ confidence interval for the population slope, β_1, is

$$\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2},n-2}s_{\hat{\beta}_1}$$

where $t_{1-\frac{\alpha}{2},n-2}$ is the $1-\frac{\alpha}{2}$ quantile of the t-distribution with $n-2$ degrees of freedom.

Example Calculate and interpret a 90% confidence interval for β_1 for the blood pressure dataset.

Hypothesis testing for β_1

In terms of hypothesis testing, we are often concerned with testing three different types of alternative hypotheses with $H_0 : \beta_1 = 0$.

- Testing for a positive linear relationship between x and y
- Testing for a negative linear relationship between x and y
- Testing for a linear relationship between x and y

Test Statistic:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

p-values: Note that T represents a t-distributed random variable with $n-2$ degrees of freedom.

- $Pr(T > t)$
- $Pr(T < t)$
- $Pr(T > |t|)$

Decision:

- If $p - value \leq \alpha$, reject H_0
- If $p - value > \alpha$, do not reject H_0

Example Conduct a hypothesis test for a significant positive linear relationship between age and systolic blood pressure.

Interpretations of hypothesis test If $H_0 : \beta_1 = 0$ is rejected, this does **NOT** necessarily mean that the underlying relationship between X and Y is linear. Similarly, if $H_0 : \beta_1 = 0$ is not rejected, this does **NOT** necessarily mean that their is no relationship between X and Y.

Consider the drug concentration dataset where we are interested in modeling the amount of concentration of the drug left in the body after a certain number of hours. Let's look at a scatterplot of the dataset.

Now, let's look at the results of the linear model regressing drug concentration on number of hours.

Confidence Intervals for $\mu_{Y|X}$ for $X = x_0$

Suppose we are interested in inference for $\mu_{Y|X=x_0}$, the mean of Y for a given value of X (x_0). We have already shown an estimate of $\mu_{Y|X=x_0}$,

$$\hat{Y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

We can also say that

$$\frac{\hat{Y}_{x_0} - \mu_{Y|x_0}}{s_{\hat{Y}_{x_0}}} \sim t_{df=n-2}$$

where $s^2_{\hat{Y}_{x_0}} = s^2_{Y|X}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s^2_X}\right)$. Therefore, we can calculate a $C = 100 \times (1-\alpha)\%$ confidence interval for $\mu_{Y|x_0}$ as

$$\hat{Y}_{x_0} \pm t_{1-\frac{\alpha}{2},n-2} \times s_{\hat{Y}_{x_0}}.$$

Example Calculate and interpret a 90% confidence interval for the mean systolic blood pressure for all 55 year olds.

Visualization of confidence intervals (bands) for $\mu_{Y|X}$

Prediction Intervals for Y for $X = x_0$

Perhaps instead of calculating an interval for the mean of Y for all individuals where $X = x_0$, we are interesting in an interval for a single individual where $X = x_0$. Note that the variance of an estimate for a single individual naturally is **larger** than the variance of an estimate for a group of individuals. More precisely,

$$\underbrace{\text{Var}(Y - \hat{Y}_{x_0})}_{\text{error by predicting an individual } Y \text{ by } \hat{Y}_{x_0}} = \underbrace{\text{Var}(Y - \mu_{Y|x_0})}_{\text{deviation of an individual } Y \text{ from its true mean}} + \underbrace{\text{Var}(\mu_{Y|x_0} - \hat{Y}_{x_0})}_{\text{deviation of a prediction } \hat{Y}_{x_0} \text{ from its true mean}}.$$

Recall earlier, we stated that

$$Y \sim \mathcal{N}(\mu_{Y|X} = \beta_0 + \beta_1 X, \sigma^2_{Y|X} = \sigma^2),$$

which means that for any $X = x_0$, an estimate of the variance of Y is $s^2_{Y|X}$. We also showed earlier from the section on confidence intervals for $\mu_{Y|x_0}$ that an estimate for $\text{Var}(\hat{Y}_{x_0})$ is $s^2_{\hat{Y}_{x_0}}$.

Based on all of this information, we can say that

$$\frac{\hat{Y}_{x_0} - Y}{\sqrt{s^2_{Y|X} + s^2_{\hat{Y_{x_0}}}}} \sim t_{df=n-2}.$$

Therefore, we can calculate a $C = 100 \times (1-\alpha)\%$ **prediction** interval for and individual Y as

$$\hat{Y}_{x_0} \pm t_{1-\frac{\alpha}{2},n-2} \times \sqrt{s^2_{Y|X} + s^2_{\hat{Y_{x_0}}}}.$$

Example Calculate and interpret a 90% prediction interval for the systolic blood pressure of a randomly selected 55 year old.

Visualization of prediction intervals (bands) for Y