STAT 308 – Chapter 6

Background Information

We have previously discussed how we can model observed values of Y by our knowledge of the **independent variable** X. We discussed methods to assess if $\beta_1 > 0$ and create confidence intervals for β_1 and $\mu_{Y|X}$ and prediction intervals for Y. We also discussed **graphical methods** to asssess the goodness of fit of our linear model. We will now discuss **numerical methods** to make inference on our linear model.

Important Definitions

Correlation Coefficient: A number describing the *strength* and *direction* of the *linear association* between X and Y.

The sample correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2}} = \frac{s_{xy}}{s_x s_y},$$

where s_{xy} is the sample **covariance** between the observed x and y.

It can be noted that r is directly related to the estimated regression slope $\hat{\beta}_1$,

$$r = \frac{s_x}{s_y}\hat{\beta}_1.$$

Example:

Recall the `bloodpressure` dataset we have previously used. Calculate the correlation coefficient between Age and Systolic Blood Pressure.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Properties of the sample correlation coefficient

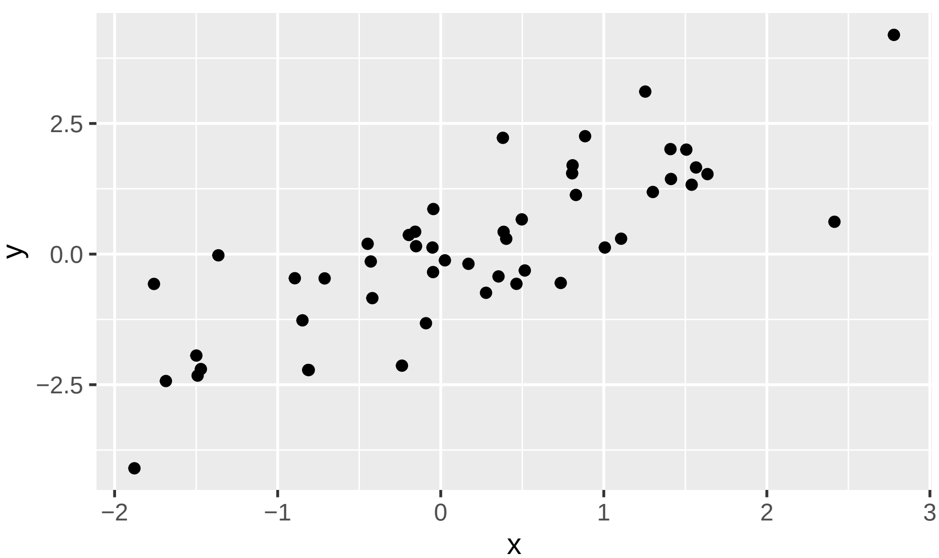
- r is a value between -1 and 1.

– $r = 1$ means there is a direct positive linear relationship between X and Y.

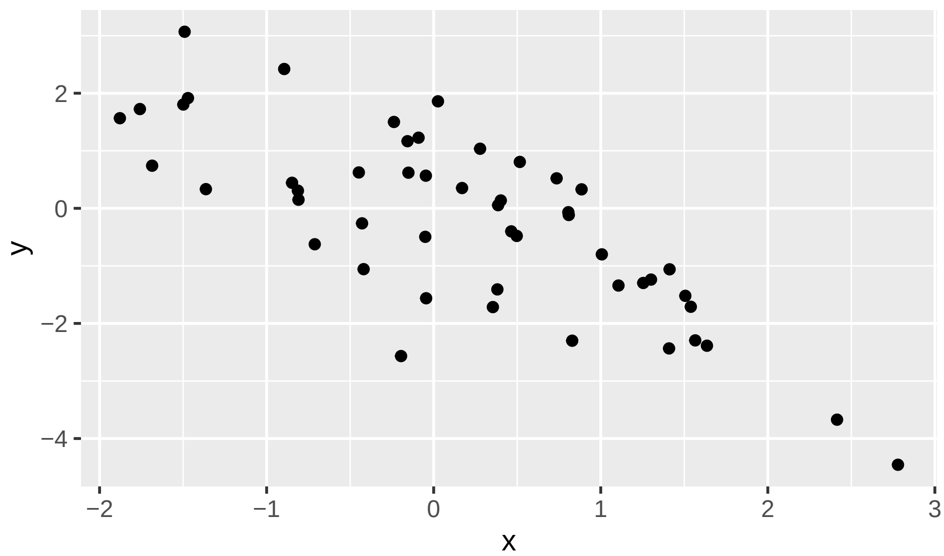
– $r = -1$ means there is a direct negative linear relationship between X and Y.

- r is a *unitless* measure
- r has the same sign as $\hat{\beta}_1$. That is
$$r > 0 \longleftrightarrow \hat{\beta}_1 > 0$$
and
$$r < 0 \longleftrightarrow \hat{\beta}_1 < 0$$
- $r = 0$ means there is no **linear** association between X and Y. That does not mean there is no pattern at all that can be made out by the graph of X and Y.

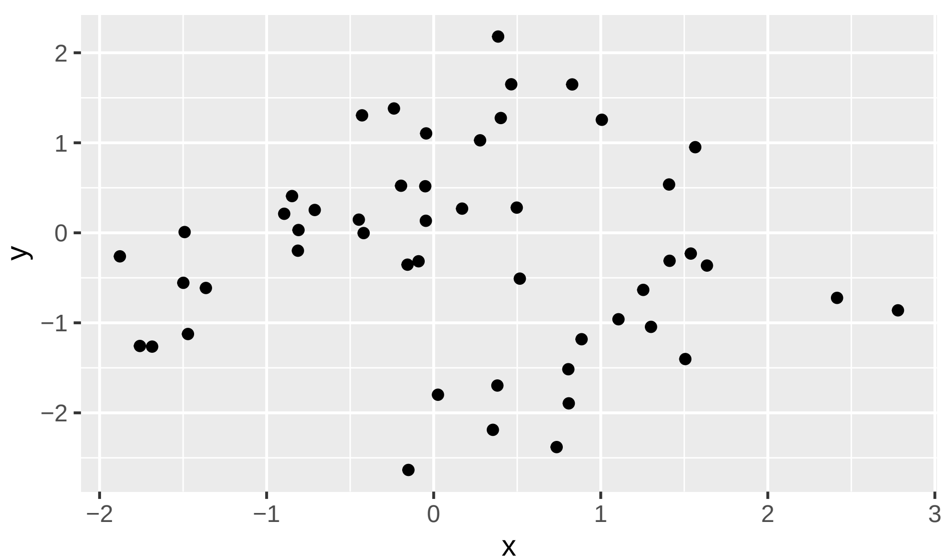
Examples of scatterplots and the correlation coefficient



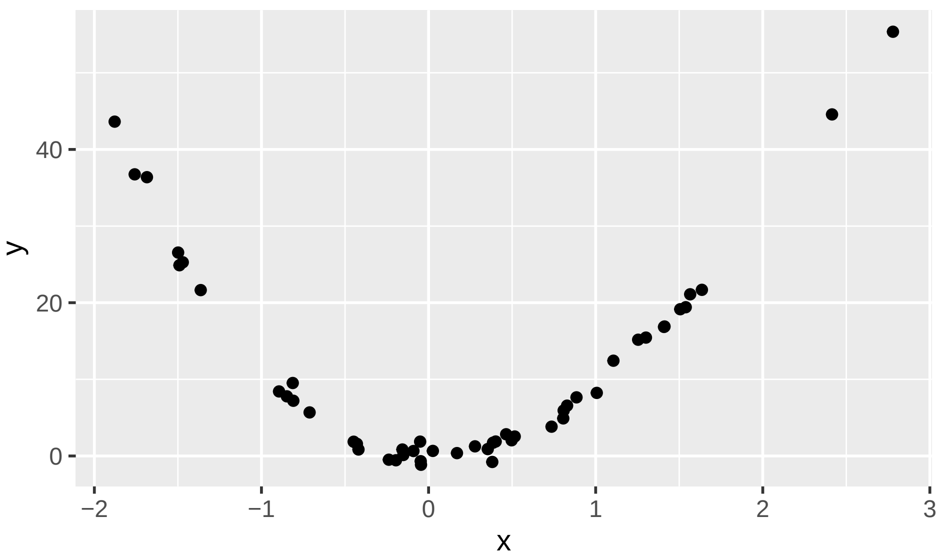

```
## [1] "Correlation between X and Y is 0.81."
```




```
## [1] "Correlation between X and Y is -0.77."
```




```
## [1] "Correlation between X and Y is -0.03."
```




```
## [1] "Correlation between X and Y is 0.05."
```

Example

What is the correlation coefficient of the blood pressure dataset?

Bivariate Normal Distribution: A distribution that describes the joint relationship between two different normally distributed random variables X and Y

Parameters of Bivariate Normal Distribution:

- μ_X: univariate mean of X
- μ_Y: univariate mean of Y
- σ^2_X: univariate variance of X
- σ^2_Y: univariate variance of Y
- ρ_{XY}: correlation between X and Y

A nice property of the bivariate normal distribution is that we can slice the distribution at a fixed value of X to obtain the *conditional distribution* of Y at a given value of X. This distribution is also normally distributed with

- $\mu_{Y|X} = \mu_Y + \frac{\rho_{XY}\sigma_Y}{\sigma_X}(X - \mu_X)$ and
- $\sigma^2_{Y|X} = \sigma^2_Y(1 - \rho^2_{XY})$.

Recall from the Chapter 5 and 6 notes, we can say that

- $\hat{\beta}_1 = \frac{rs_y}{s_x}$ and

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

If we substitute in the parameters for their respective estimates, we have

- $\beta_1 = \frac{\rho_{XY}\sigma_Y}{\sigma_X}$ and
- $\beta_0 = \mu_y - \beta_1 \mu_x$.

Then, using some substitution, we have

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

showing the relationship between the least squares regression line and the bivariate normal distribution!

Now, if we were to take the formula for $\sigma^2_{Y|X}$ and solve for ρ^2_{XY}, we would get

$$\rho^2_{XY} = \frac{\sigma^2_Y - \sigma^2_{Y|X}}{\sigma^2_Y}.$$

In other words, ρ^2_{XY} is the

R-squared

R-Squared(r^2): the percent of variation in the response variable Y that can be explained through its linear relationship with the explanatory variable X

Formally,

$$r^2 = \frac{SSY - SSE}{SSY},$$

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

Example

Find the r^2 of the systolic blood pressure dataset. Interpret this value in the context of the given problem.

About R-squared

- Naturally, r^2 is the square of the sample correlation coefficient, so $0 \leq r^2 \leq 1$
- The larger the value of r^2, the more variance in Y we can explain through its linear relationship with X, and thus, the stronger the linear relationship between the two variables

– If $r^2 = 1$, all of the variation in Y can be explained linearly by X (in other words, $SSE = 0$)

– If $r^2 = 0$, no variation in Y can be explained linearly by X

- r^2 does NOT measure the magnitude of $\hat{\beta}_1$ (i.e. r^2 can be close to one, but $\hat{\beta}_1$ may still be close to zero, or r^2 can be close to zero, but $\hat{\beta}_1$ may be large.)
- r^2 is NOT a measure of the appropriateness of the linear model.