# TMA4245 Statistikk

Week 14 - Wednesday

## Week 14 (Mon. April 1 - Friday April 5)

**Monday**
- Free day!

**Wednesday**
- Intro videos (20-45 min)
- Lecture 2x45 min
- Obligatory exercises

**Thursday & Friday**
- Statistics lab 8:15 - 12:00 S6
- Statistics lab 12:15 - 16:00 S6
- STACK 12 Deadline: 05.04.24 23:59

## Week 15 (Mon. April 8 - Friday April 12)

**Monday**
- Intro videos (20-45 min)
- Lecture 2x45 min
- Obligatory exercises

**Wednesday**
- Intro videos (20-45 min)
- Lecture 2x45 min
- Obligatory exercises

**Thursday & Friday**
- Statistics lab 8:15 - 12:00 S6
- Statistics lab 12:15 - 16:00 S6
- STACK 13 Deadline 12.04.24 23:59

## Week 16 (Mon. April 15 - Friday April 19)

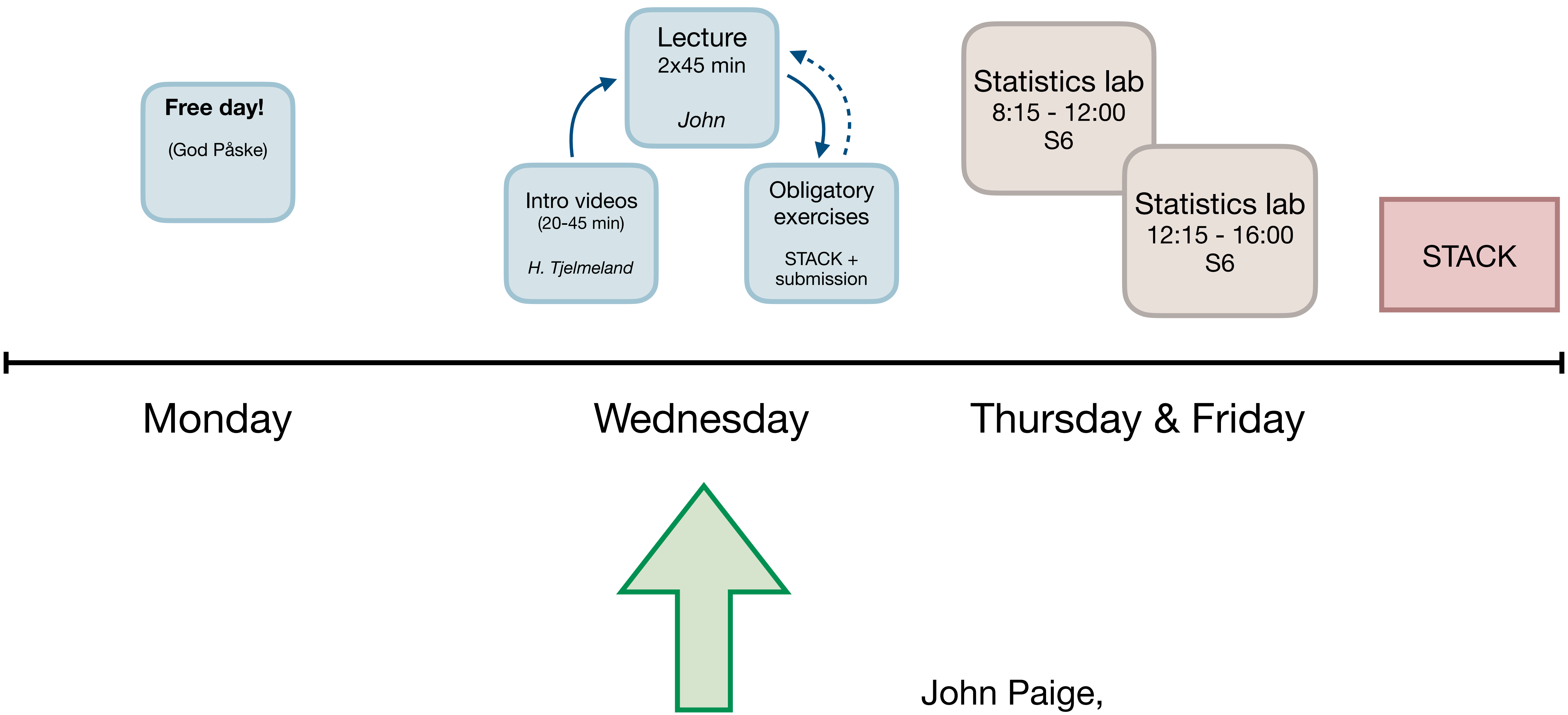
**Monday**
- Intro videos (20-45 min)
- Lecture 2x45 min
- Obligatory exercises

**Wednesday**
- Summary 2x45 min
- Obligatory exercises

**Thursday & Friday**
- Statistics lab 8:15 - 12:00 S6
- Statistics lab 12:15 - 16:00 S6
- **Submission 6** Deadline 19.04.24 23:59
- STACK 14 Deadline 19.04.24 23:59

# TMA4245 Statistikk

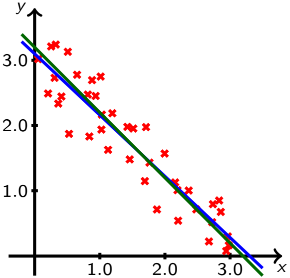
## Week 14 - Wednesday





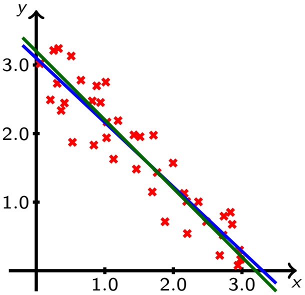
John Paige,
Department of mathematical sciences, NTNU

# Simple linear regression: motivation



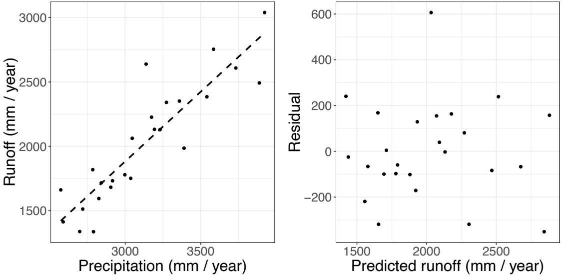

- Goal: we wish to predict a random variable, $Y$, based on a related quantity, $x$
  - $Y$ is the **'dependent' variable, or 'response'**
  - $x$ is the **'independent' or 'explanatory' variable**
- Assume $E[Y] = \beta_0 + \beta_1 x$
  - i.e. assume $Y = \beta_0 + \beta_1 x + \varepsilon$ with $E[\varepsilon] = 0$
- Estimate $\beta_0$ and $\beta_1$ as $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing sum of squared errors (maximum likelihood is an alternative with slightly different assumptions)

# Simple linear regression: motivation




- Simple linear regression is **central** to modern statistics
- Simple linear regression can be generalized to allow for:
  - Multiple explanatory variables (linear regression)
  - Nonlinear relationships (generalized additive models)
  - Non-Gaussian responses (generalized linear models)
  - Random explanatory variables (linear and generalized linear mixed models)
  - etc.

# Runoff (problem 3a, Spring 2019)

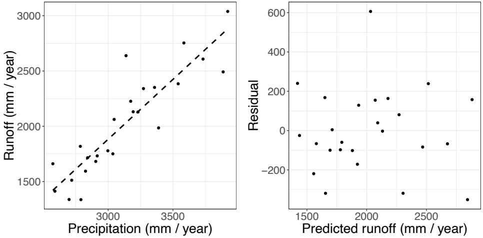



Assume the following linear relationship between annual runoff $Y$ and annual precipitation $x$ within a drainage basin,

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

where $\beta_0$ and $\beta_1$ are unknown constants and $\varepsilon$ is normally distributed with expected value (mean) 0 and unknown variance $\sigma^2$.

# Runoff (problem 3a, Spring 2019)

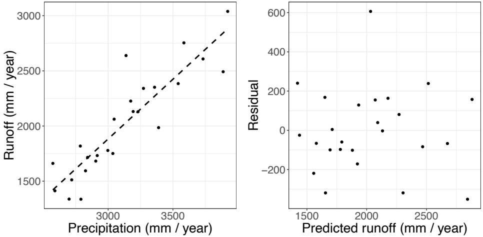



Questions:

1. How can 'least squares' be used to find estimators for $\beta_0$ and $\beta_1$ under the assumption that, for $i = 1, \ldots, 25$,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i?$$

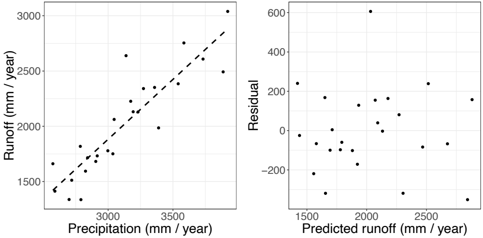
# Runoff (problem 3a, Spring 2019)




Questions:

1. How can 'least squares' be used to find estimators for $\beta_0$ and $\beta_1$ under the assumption that, for $i = 1, \ldots, 25$,
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i?$$
2. What assumptions are necessary for linear regression?

# Runoff (problem 3a, Spring 2019)




Questions:

1. How can 'least squares' be used to find estimators for $\beta_0$ and $\beta_1$ under the assumption that, for $i = 1, \ldots, 25$,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i?$$

2. What assumptions are necessary for linear regression?
3. Is linear regression reasonable for this problem? (Recall that $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, where 'iid' means independent and identically distributed)

# Hot chocolate sales (problem 4b, Fall 2015)

Last winter, on Sundays, Alexander sold cups of hot chocolate by the ski tracks near his house. This winter he plans to have a similar business. Alexander experienced that the sales changed dramatically with the weather and skiing conditions. He made a condition index, $x$, where $x = 1$ means “bad conditions”, $x = 2$ means “good conditions”, $x = 3$ means “very good conditions” and $x = 4$ means “excellent conditions”.

# Hot chocolate sales (problem 4b, Fall 2015)

For 20 Sundays, $i = 1, \ldots, 20$, he registered both the number of cups sold, denoted $y_i$, and the associated condition index, $x_i$. We will phrase the sales as a regression model taking condition as an explanatory variable:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, 20,$$

where $\epsilon_1, \ldots, \epsilon_{20}$ are independent normal distributed variables with expected value 0 and variance $\sigma^2$, and $\beta_0$ and $\beta_1$ are fixed but unknown regression parameters.

# Hot chocolate sales (problem 4b, Fall 2015)

Based on the data:

$\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 2.45$

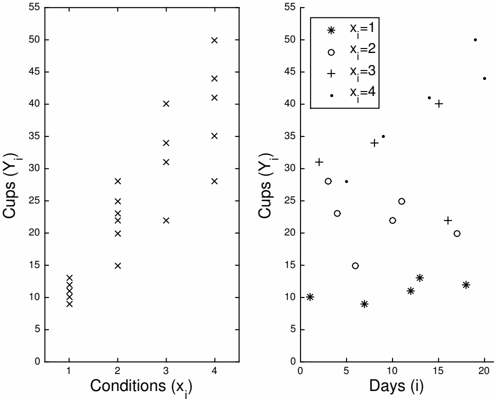
$\bar{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 25.65$

$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 24.95$

$\sum_{i=1}^{20}(x_i - \bar{x})y_i = 237.15$

$\frac{1}{18}\sum_{i=1}^{20}(y_i - \hat{\beta}_0 - \hat{\beta}_1)^2 = 5.65^2$

Questions:



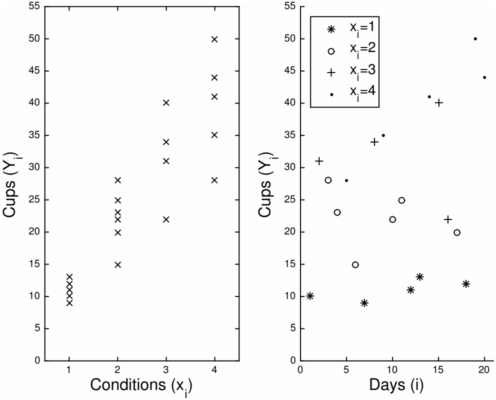
# Hot chocolate sales (problem 4b, Fall 2015)

Based on the data:

$$\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 2.45$$
$$\bar{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 25.65$$
$$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 24.95$$
$$\sum_{i=1}^{20}(x_i - \bar{x})y_i = 237.15$$
$$\frac{1}{18}\sum_{i=1}^{20}(y_i - \hat{\beta}_0 - \hat{\beta}_1)^2 = 5.65^2$$

Questions:

1. Do the modeling assumptions look reasonable from this plot? Why or why not?



# Hot chocolate sales (problem 4b, Fall 2015)

Based on the data:

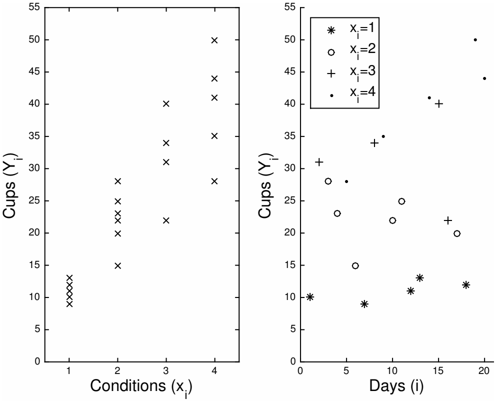
$\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 2.45$

$\bar{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 25.65$

$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 24.95$

$\sum_{i=1}^{20}(x_i - \bar{x})y_i = 237.15$

$\frac{1}{18}\sum_{i=1}^{20}(y_i - \hat{\beta}_0 - \hat{\beta}_1)^2 = 5.65^2$

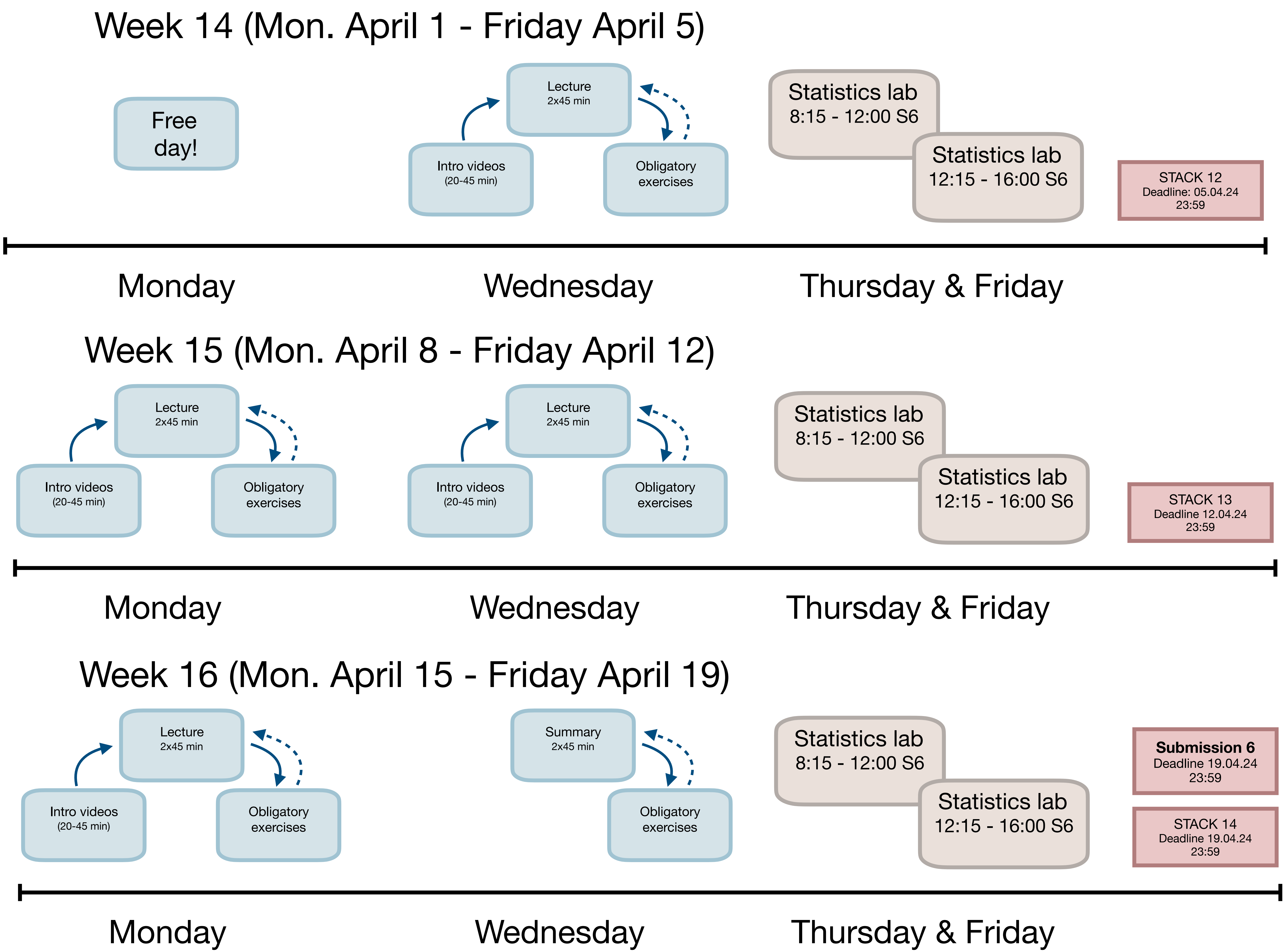



Questions:

1. Do the modeling assumptions look reasonable from this plot? Why or why not?
2. Calculate $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$.

# Logistics

- Karina Lilleborge will lecture Monday April 8
- Team based learning on Monday April 15
  - This session may be recorded to some extent (via video, photos). I will ask for your permission for this at the class session.


Week 14 (Mon. April 1 - Friday April 5)
Free day!
Lecture 2x45 min
Intro videos (20-45 min)
Obligatory exercises
Statistics lab 8:15 - 12:00 S6
Statistics lab 12:15 - 16:00 S6
STACK 12 Deadline: 05.04.24 23:59
Monday
Wednesday
Thursday & Friday
Week 15 (Mon. April 8 - Friday April 12)
Lecture 2x45 min
Intro videos (20-45 min)
Obligatory exercises
Lecture 2x45 min
Intro videos (20-45 min)
Obligatory exercises
Statistics lab 8:15 - 12:00 S6
Statistics lab 12:15 - 16:00 S6
STACK 13 Deadline 12.04.24 23:59
Monday
Wednesday
Thursday & Friday
Week 16 (Mon. April 15 - Friday April 19)
Lecture 2x45 min
Intro videos (20-45 min)
Obligatory exercises
Summary 2x45 min
Obligatory exercises
Statistics lab 8:15 - 12:00 S6
Statistics lab 12:15 - 16:00 S6
Submission 6 Deadline 19.04.24 23:59
STACK 14 Deadline 19.04.24 23:59
Monday
Wednesday
Thursday & Friday