# Goodness of fit of the model

NHH
TECH3

How well does the model describe the data?

How much of the variation in the data is explained by the model?

# How much of the variation in the data is explained by the model?

- What is the variation in the data?

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

- What variation is left after the model?

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

- What variation is explained by the regression?

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$SST = SSR + SSE$$

# How much of the variation in the data is explained by the model?

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Special case: $y = \beta_0 + \beta_1 x + \epsilon$

$$R^2 = r^2$$

# How much of the variation in the data is explained by the model?

$$0 \leq R^2 \leq 1$$

# Do not use for model selection

- If you have many independent variables, **model selection** means finding the optimal combination of explanatory variables for your regression model
- $R^2$ will always improve by adding more independent variables to model
- One should use metrics that penalize complicated models for model selection
  - Akaike's information criteria (AIC)
  - Bayesian information criteria (BIC)
- Or cross validation

NHH

TECH3

Sondre Hølleland
Geir Drage Berentsen