Statistics MM6: Regression

Lecturer: Israel Leyva-Mayorga

email: ilm@es.aau.dk

AALBORG UNIVERSITY
DENMARK

Connectivity

Schedule

1. Introduction to statistics
2. Parameter estimation
3. Confidence intervals
4. Hypothesis testing 1
5. Hypothesis testing 2
6. **Regression**
7. Workshop: wrap-up and exam problems

Outline

Recap on hypothesis testing

Linear regression

Least squares estimators

Inference

Residual analysis

Summary

Recap on hypothesis testing

Types of tests based on the populations

Parameter testing with 1 population

There is some idea about the value of a parameter

Is that idea correct?

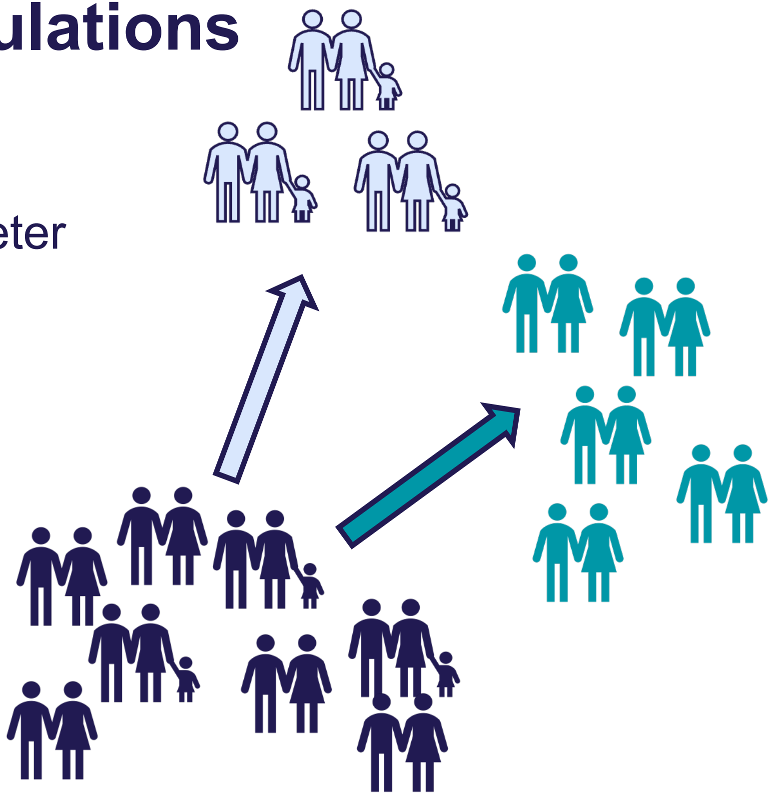
Example: Is it true that the average age is 20?

Compare 2 populations with each other

No parameter known a priori

- Begin with different populations
- One initial population divided into 2

Can we find differences between populations?



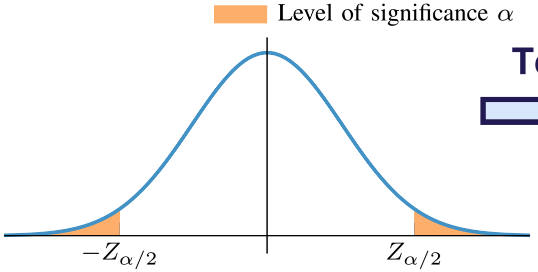
The limits of hypothesis testing

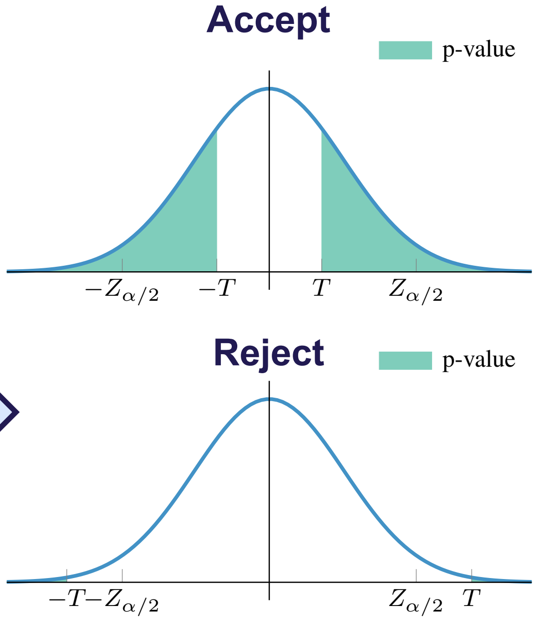
H_0: The **null hypothesis,** the one assumed to be true

H_1: The **alternative hypothesis,** which contradicts H_0

We try to find evidence to reject H_0

But we don't want to reject H_0 when it's true: α





What if we want to find relations between populations?

Linear regression

Where is life expectancy higher?

No binary distinction

Independent variable (x-axis):

- Life satisfaction index

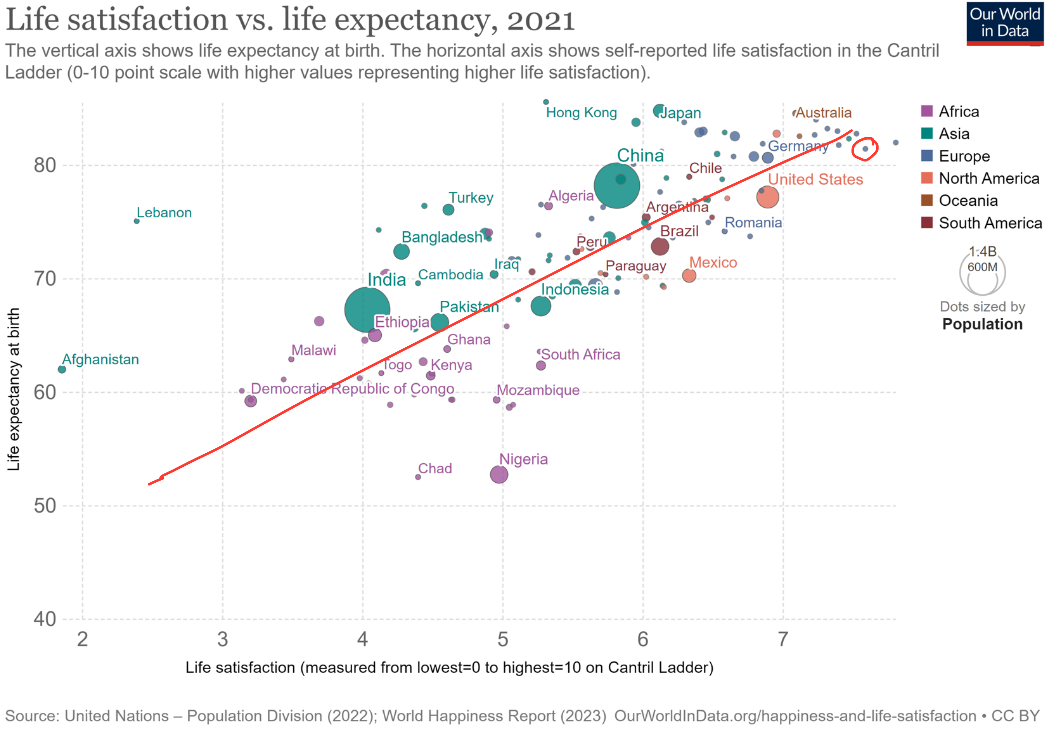
Dependent variable (y-axis):

- Life expectancy at birth

We start from scatter plot

Can we see a trend?

Can we use it for prediction?



Linear regression

We have data in the form $(X_1, Y_1), \dots, (X_n Y_n)$

And assume that the relationship between variables is **linear**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

Where

- β_0 is the y-intercept
- β_1 is the slope
- ϵ_i is the error (also called the noise)
- $\mathbb{E}(\epsilon_i | X_i) = 0$
- $\mathrm{var}(\epsilon_i | X_i) = \sigma^2$

Simple linear regression

The parameters β_0 and β_1 are **unknown**: **we have to estimate them**

We end up with $\hat{\beta}_0$ and $\hat{\beta}_1$ so the **fitted line** is

$$\hat{r}(x) = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x}$$

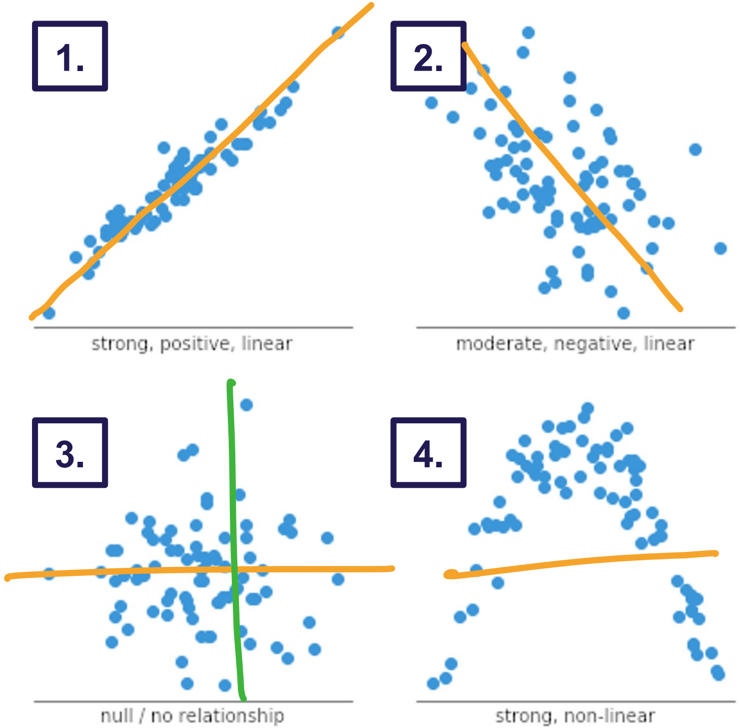
The **predicted values** are

$$\hat{Y}_i = \hat{r}(X_i)$$

And the **residuals** are

$$\hat{\epsilon}_i = \underbrace{Y_i}_{\text{real}} - \underbrace{\hat{Y}_i} = \underbrace{Y_i}_{\text{real}} - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$$

When to use linear regression



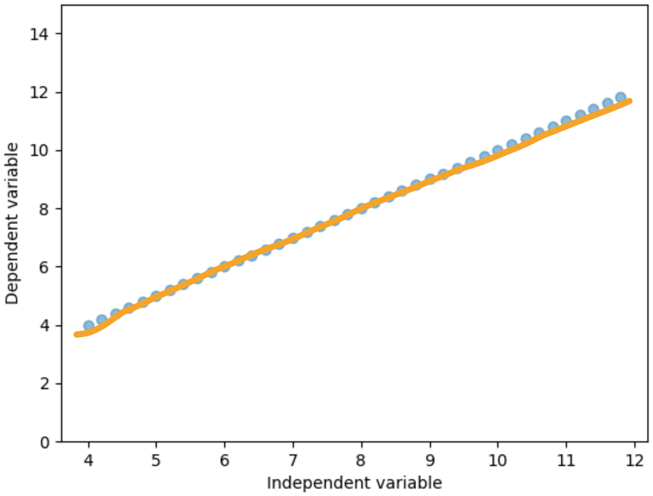
1. Good prediction with linear regression
2. Good average prediction but high variance
3. No observable trend
4. Not linear, leading to wrong predictions

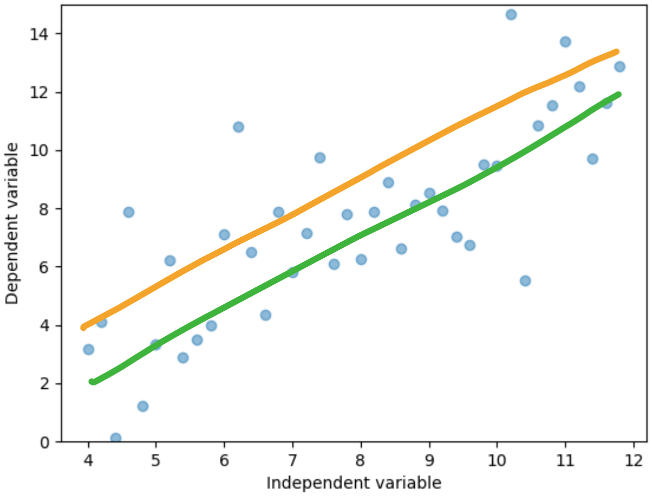
There is also multiple regression

$$Y_i = \beta_0 + \beta_1 X_i^{(1)} + \cdots + \beta_k X_i^{(k)} + \epsilon_i$$

Intuitive explanation

We are recovering the underlying function after removing the noise





Least squares estimators

Determining the best estimate

In general, it is impossible to find a line that passes over all the points

What is the best fitted line?

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

The one with the best estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

In general, it is impossible to find a line that passes over all the points

What is the best fitted line?

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

The one with the best estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

Least squares regression

Approach: Minimize the Mean Square Error (MSE) for the residuals

$$\hat{\epsilon}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$$

Specifically, we minimize the Residual Sum of Squares (RSS)

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2$$

Finding the least squares estimates

$[f(g(x))]' = f'(g(x))g'(x)$

Estimators

$$\hat{\beta}_1: \quad \frac{\partial \sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{\partial \underline{\beta_1}} = 0 \quad \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i)$$

$$\hat{\beta}_0: \quad \frac{\partial \sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{\partial \underline{\beta_0}} = 0 \Rightarrow \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

$$-\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \beta_0 + \sum_{i=1}^{n} \beta_1 X_i = 0$$

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{Y}_n - \beta_1 \bar{X}_n$$

The least squares estimates

Estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

Unbiased estimator for σ^2

$$\hat{\sigma}^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

Distribution of the estimators

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are RVs

Usually, we assume Gaussian noise:

$$\epsilon_i \sim N(0, \sigma^2)$$

Therefore, for a given X_i we have

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Least squares regression makes sense with Gaussian noise

Not so much if the noise is heavy-tailed

Distribution of $\hat{\beta}_1$

Recall that $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$

Sample variance $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_X^2 n}\right)$$

Distribution of $\hat{\beta}_0$

Recall that $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n = \bar{Y}_n - \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}\right)\bar{X}_n$

And $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

Sample variance $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$$\hat{\beta}_0 \sim N\left(\beta_0, \boxed{\frac{\sigma^2}{S_X^2 n}}\left(\frac{\sum_{i=1}^{n} X_i^2}{n}\right)\right)$$

Example

Underlying model

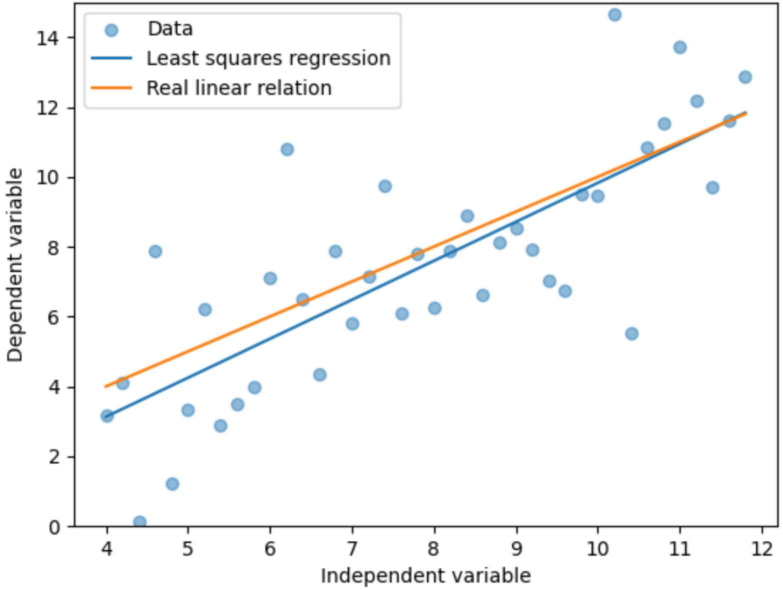
$$Y_i = X_i + \epsilon_i$$

$\beta_0 = 0$ $\beta_1 = 1$

Estimated model with least squares

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = -1.332 + 1.115 X_i$$

$\hat{\beta}_0$ $\hat{\beta}_1$



Example

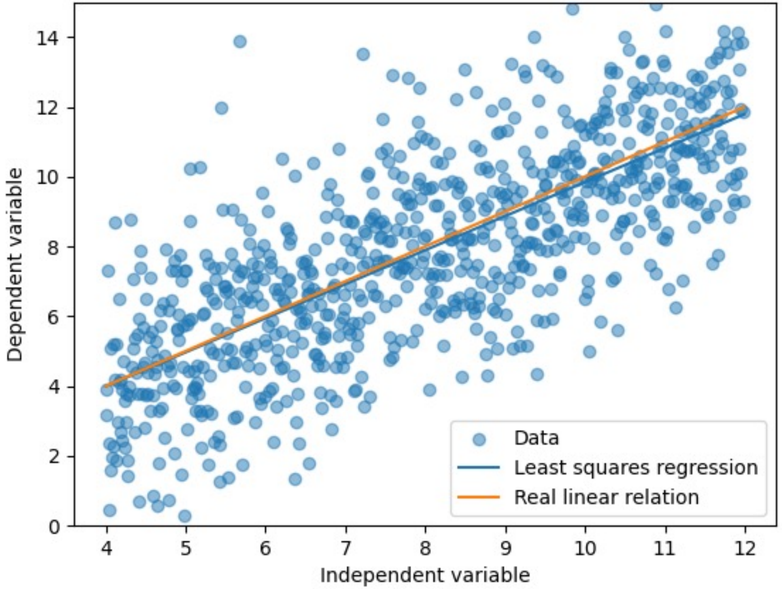
Underlying model

$$Y_i = X_i + \epsilon_i$$

Estimated model with least squares

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 0.094 + 0.974 X_i$$

$\beta_0 = 0 \quad \beta_1 = 1$



Inference using β_1

Example

A guy claims that how fast one drives has no impact on fuel consumption of a car
To test this, measurements were collected for a car driving at different speeds
Can we reject the claim with these data?

Speed (mph)	Miles per gallon
45	24.2
50	25.0
55	23.3
60	22.0
65	21.5
70	20.6
75	19.8

Estimated standard errors

We already had an unbiased estimator

$$\hat{\sigma}^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2$$

$$\operatorname{var}\left(\hat{\beta}_0\right) = \frac{\hat{\sigma}^2}{S_X^2 n}\left(\frac{\sum_{i=1}^{n} X_i^2}{n}\right) \qquad \widehat{\operatorname{se}}\left(\hat{\beta}_0\right) = \frac{\hat{\sigma}}{S_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}}$$

$$\operatorname{var}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}^2}{S_X^2 n} \qquad \widehat{\operatorname{se}}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}}{S_X\sqrt{n}}$$

Properties of the estimators

Under appropriate conditions we have

1. Consistency: $\hat{\beta}_0 \xrightarrow{P} \beta_0$ and $\hat{\beta}_1 \xrightarrow{P} \beta_1$
2. Asymptotic Normality:

   Both $\frac{\hat{\beta}_0 - \beta_0}{\widehat{\text{se}}(\hat{\beta}_0)}$ and $\frac{\hat{\beta}_1 - \beta_1}{\widehat{\text{se}}(\hat{\beta}_1)}$ are asymptotically standard Normal RVs: $N(0,1)$

3. Approximate $1-\alpha$ confidence intervals for β_0 and β_1

$$\hat{\beta}_0 \pm Z_{\alpha/2}\widehat{\text{se}}(\hat{\beta}_0) \text{ and } \hat{\beta}_1 \pm Z_{\alpha/2}\widehat{\text{se}}(\hat{\beta}_1)$$

Inference on the slope

Does X_i have an effect on Y_i?
If $\beta_1 = 0$, there is no effect

But we don't have β_1, we're stuck with its estimator $\hat{\beta}_1$

We can do hypothesis testing **H$_0$:** $\beta_1 = \beta_{\mathrm{H}} = 0$
Due to asymptotic normality

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_X^2 n}\right), \text{ then } \frac{\hat{\beta}_1 - \beta_1}{\widehat{\mathrm{se}}(\hat{\beta}_1)} \sim N(0,1)$$

Hypothesis testing

We use

$$\widehat{\mathrm{se}}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}}{S_X\sqrt{n}}$$

To get

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{\mathrm{se}}\left(\hat{\beta}_1\right)} = \frac{\left(\hat{\beta}_1 - \beta_1\right)S_X\sqrt{n}}{\hat{\sigma}} = \frac{\left(\hat{\beta}_1 - \beta_1\right)S_X\sqrt{n}}{\sqrt{\left(\frac{1}{n-2}\right)\sum_{i=1}^{n}\hat{\epsilon}_i^2}} \sim T_{n-2}$$

Degrees of freedom: no. of independent variables minus the no. of equations

We have n values and one equation for each β_0 and β_1, so $n-2$ dofs

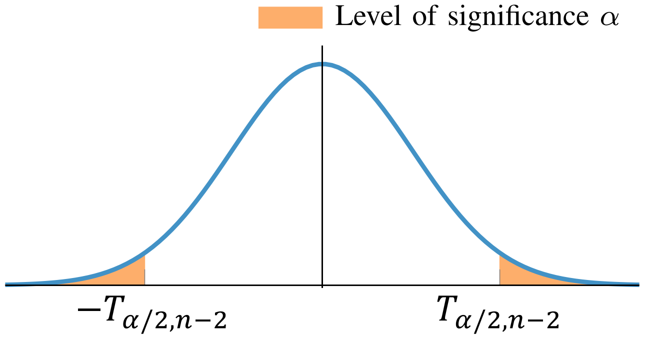
Test statistic and rejection region

Test statistic for two-sided test is

$$T = \frac{\left|\hat{\beta}_1 - \beta_{\mathrm{H}}\right|}{\widehat{\mathrm{se}}\left(\hat{\beta}_1\right)} = \frac{\left|\hat{\beta}_1 - \beta_{\mathrm{H}}\right| S_X \sqrt{n}}{\left(\frac{1}{n-2}\right) \sum_{i=1}^{n} \hat{\epsilon}_i^2}$$

Rejection region for level of significance α is $R = \left\{\hat{\beta}_1 : T > T_{n-2,\alpha/2}\right\}$

P-value: same as before



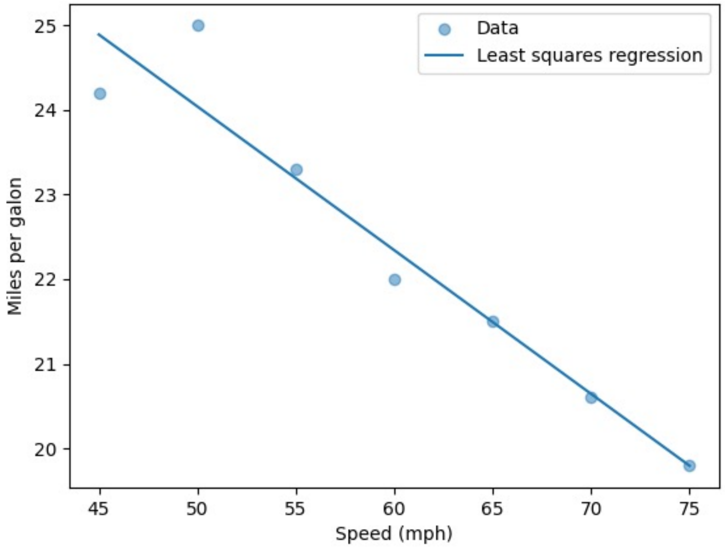
Solving the example

H_0: $\beta_1 = \beta_H = 0$

Estimators: $\hat{\beta}_0 = 32.542$ and $\hat{\beta}_1 = -0.169$ and $\widehat{\mathrm{se}}(\hat{\beta}_1) = 0.0208$

$T = \frac{|\hat{\beta}_1 - 0|}{\widehat{\mathrm{se}}(\hat{\beta}_1)} = 8.139 > 2.5705$: **reject H_0**

Speed (mph)	Miles per gallon
45	24.2
50	25.0
55	23.3
60	22.0
65	21.5
70	20.6
75	19.8



Mean response

We "train" the regression model with n data points and a new one x_* appears
The estimate of Y_* is also called the **mean response**

$$Y_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

What is the confidence interval for the mean response with a new value x_*?

$$C_{1-\alpha} = \hat{\beta}_0 + \hat{\beta}_1 x_* \pm T_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \left(\boxed{\frac{1}{n}} + \frac{(x_* - \bar{X}_n)^2}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \right)}$$

Example: What would be the **mean price** for a 100 m² house in Copenhagen?

Prediction interval for a new response

We "train" the regression model with n data points and a new one x_* appears

The estimate of Y_* is also called the **mean response**

$$Y_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

What is its **prediction interval** for a new response?

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm T_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \left(\underline{\boxed{1 + \frac{1}{n}}} + \frac{(x_* - \bar{X}_n)^2}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \right)}$$

Example: What would be the **price** for a 100 m^2 house in Copenhagen?

Residual analysis

Good scenario for linear regression

Standardized residuals ϵ_i

$$Z_i = \frac{(Y_i - \beta_0 - \beta_1 X_i)}{\hat{\sigma}}$$

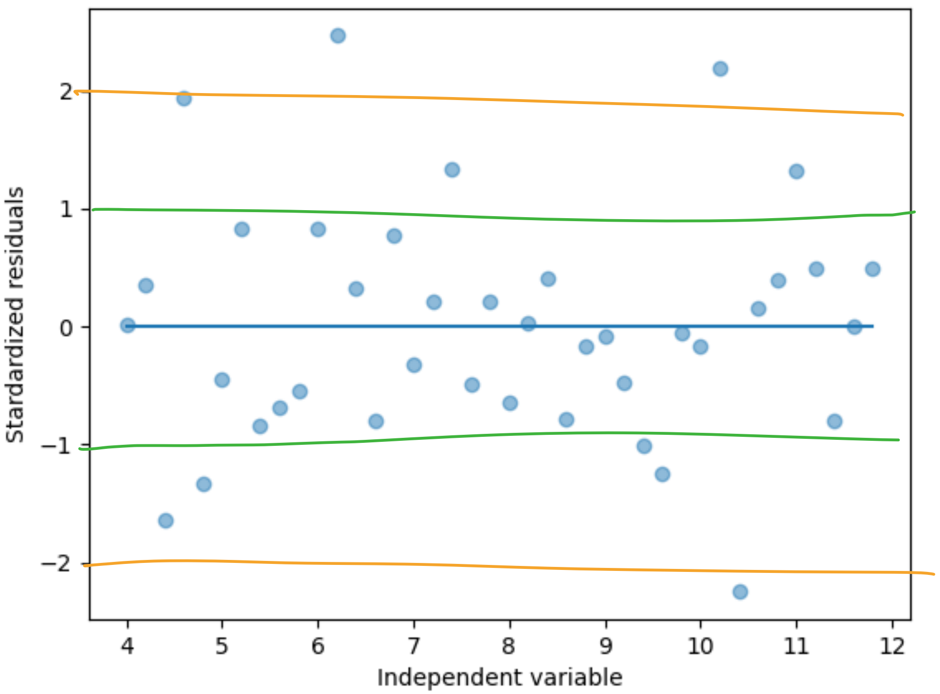
Measures in std. deviation units

$$\hat{\sigma}^2 = \left(\frac{1}{n-2}\right) \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

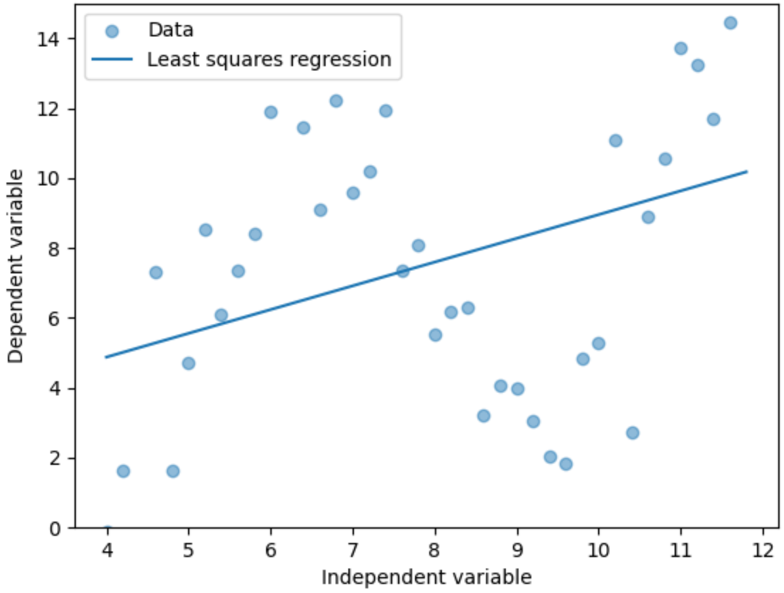
Should be no correlation with x-axis

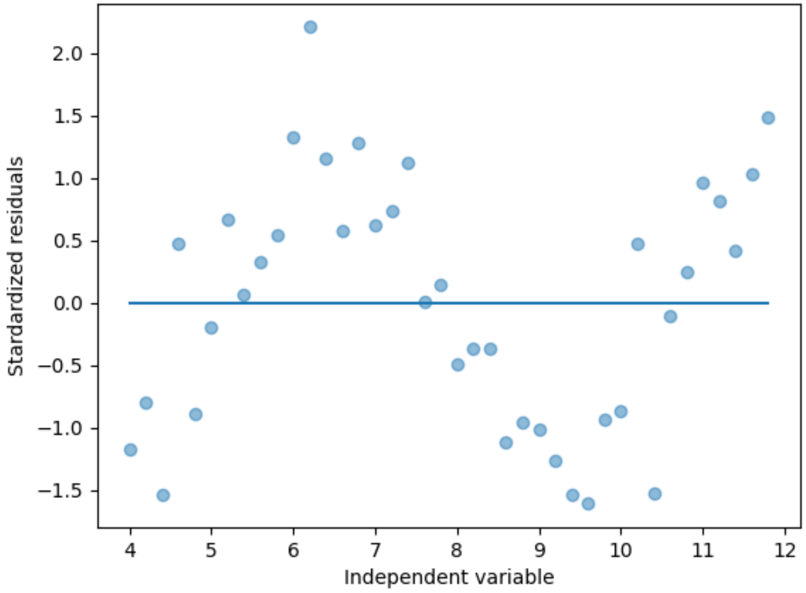
68% between -1 and 1

95% between -2 and 2



Bad scenario for linear regression: model is not linear

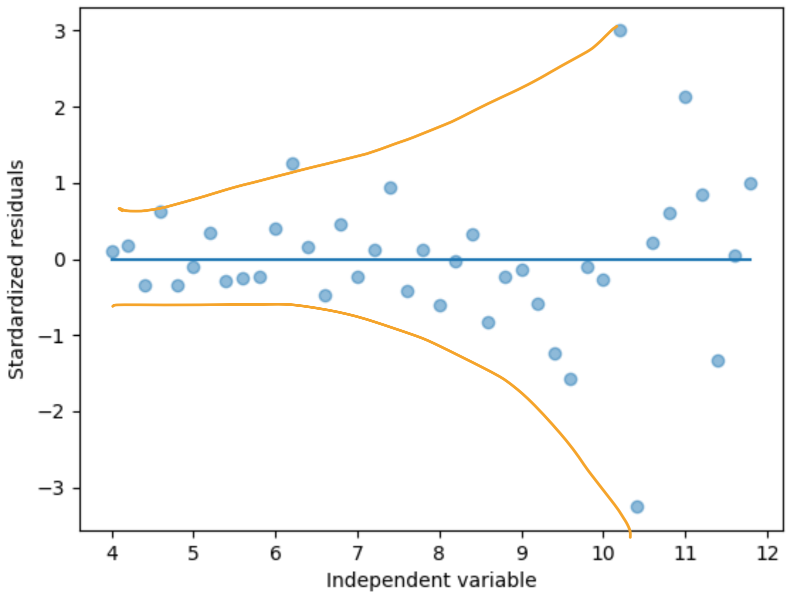




Another bad scenario: correlation of ϵ_i with X_i

The variance grows with X_i

Heteroscedastic samples



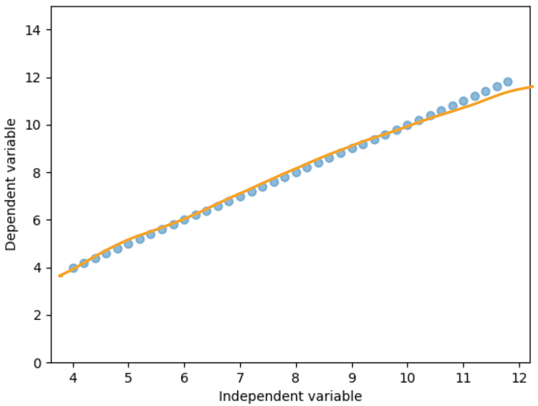
Coefficient of determination

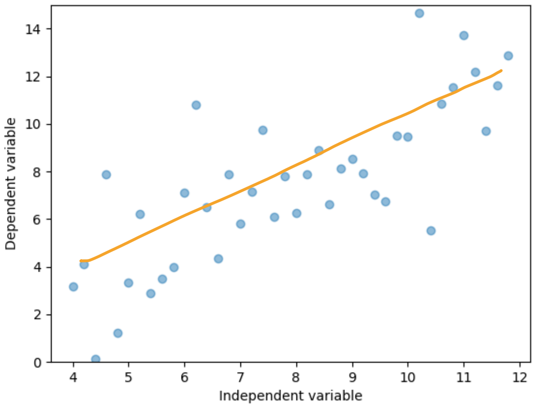
Measure of how good is the fit in a single value

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2} = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{nS_Y^2}$$

Maximum value is 1: the best possible fit

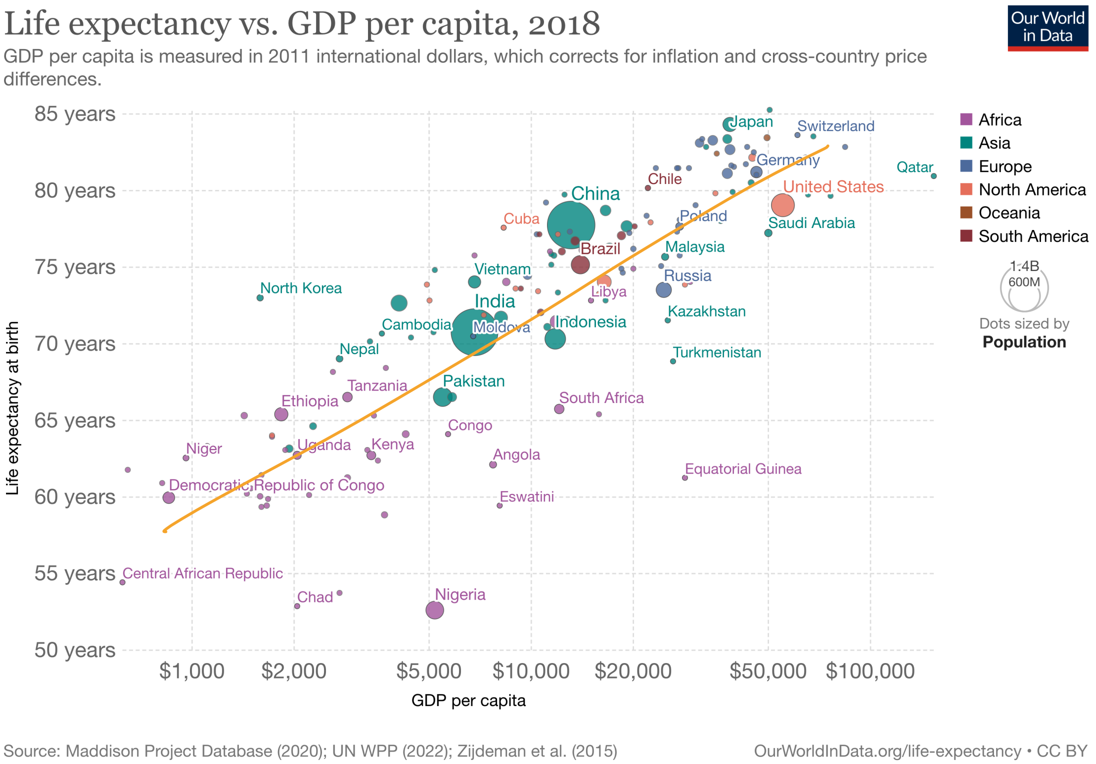
The lower the value, the worse the fit



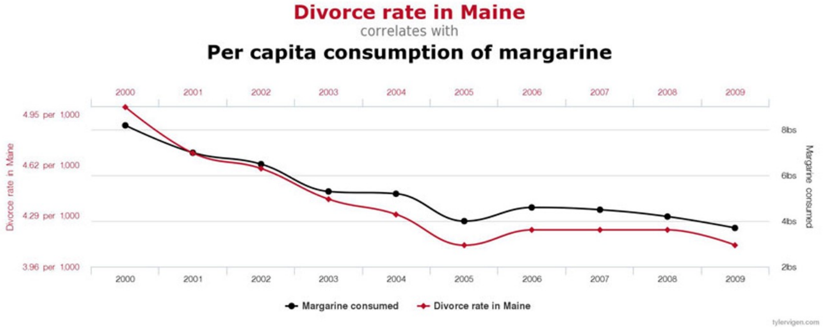


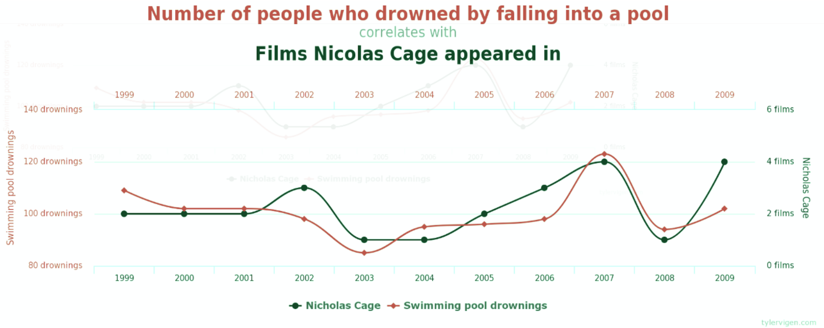
Transforming the data

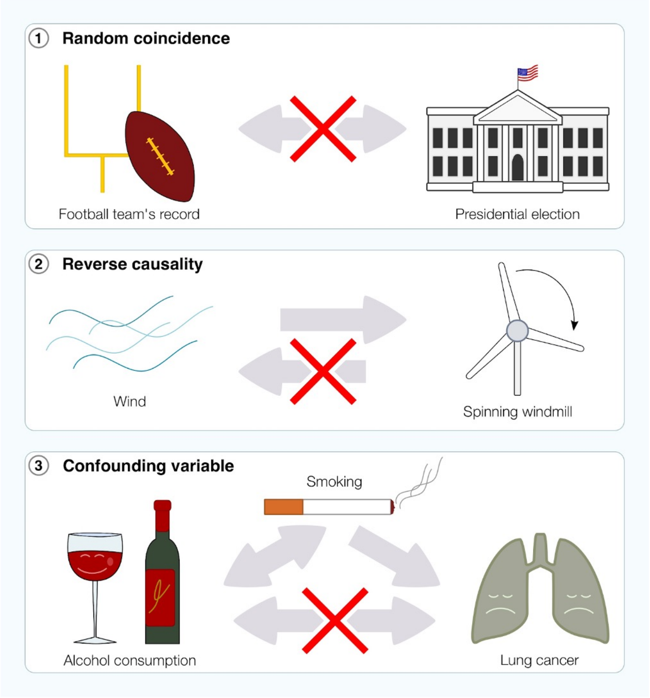
X-axis is not linear!



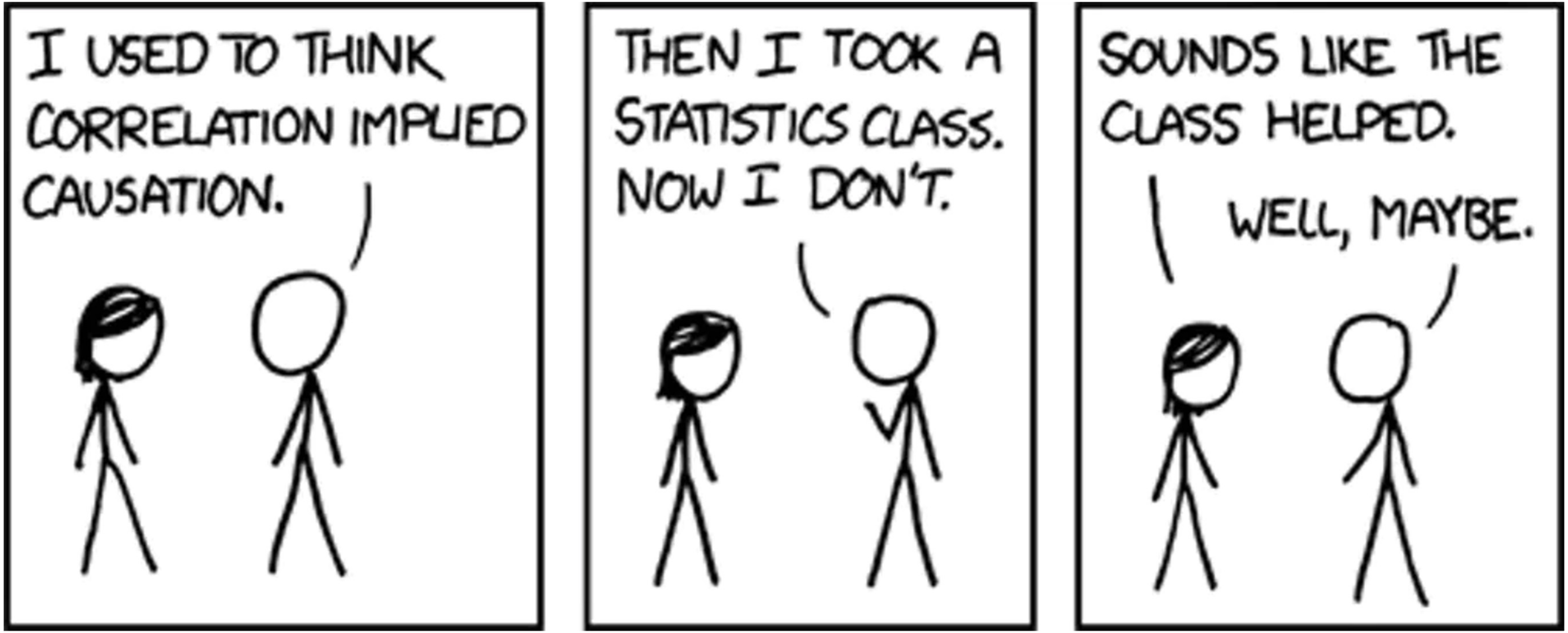
Correlation does not imply causation







Words of wisdom



Summary

Summary

Regression helps us identify correlation

Linear regression assumes the underlying model is linear

Besides fitting, we can perform:

- Hypothesis testing
- Prediction of the mean
- Prediction of a single new value

Residuals help us determine goodness of fit: is the model linear?

As initial test, try regression with a linear model with known parameters

That’s all for the lectures