Linear Regression Assumptions

Applied Regression in R

Aleš Vomáčka

23. 03. 2026

Faculty of Arts, Charles University

Why Make Assumptions?

Why Make Assumptions?

Our friend just messaged us they are leaving Celetná and going to meet us at Standard Cafe.

Can we tell how long it will take them just from the information they gave us?

Why Make Assumptions?

We can't!

We don't kow what path they'll take, whether they'll walk or take public transport, get lost along the way, etc.

To make an estimate for how long it will take before our friend meets us, we need to make assuptions.

Why Make Assumptions?

Assumptions are **information not included in our data**.

We make assumptions for two reasons:

- To **provide information** neccessary to answer our research question.
- To **get rid of uneccesary details**.

Why Make Assumptions?

Making assumptions isn't a bad thing. You can't do research without them.

However, you need to make assumptions responsibly, because incorrect assumptions will lead you to a wrong answer.

Why Make Assumptions?

Don’t think about assumptions in binary - they are not either correct or violated.

What matters is **how much our assumptions stray away from reality**.

Remember: Models are simplified approximations of reality.

Questions?

Linear Regression Assumptions

Linear Regression Assumptions

Quick vocabulary lesson:

Errors are differences between true and predicted values.

Residuals are differences between observed and predicted values.

Total error can be separeted into systemic (**bias**) and random (**variance**) parts.

Precision is reciprocal to random error (lower error → higher precision.)

Efficiency is how many observations we need to achieve desired model performance.

Linear Regression Assumptions

What are the assumptions of linear regression?

Linear Regression Assumptions

What are the assumptions of linear regression?

1. Validity & reliability of measurment
2. Representativity of data
3. Linearity & additivity
4. Independence of errors
5. Homoskedasticity of errors
6. Normality of errors

Linear Regression Assumptions

What are the assumptions of linear regression?

1. Validity & reliability of measurment
2. Representativity of data
3. Linearity & additivity
4. Independence of errors
5. Homoskedasticity of errors
6. Normality of errors

The assumptions are listed in **order of general importance**.

Validity & Rliability of Measurment

Validity & Rliability of Measurment

Boils down to two questions:

- Have we **measured all variables** needed?
- Are our data **measured with sufficient quality**?

Validity & Rliability of Measurment

We have already discussed why controlling (or not) for all relevant variables is important.

Remebmer the DAG (Directed Acyclic Graphs)!

Validity & Rliability of Measurment

Measurment quality is also of extreme importance.

There are many aspects to measurment quality - validity, reliability, invariance. Too many to cover here.

One example: Low reliability makes you underestimate relationship strength.

Validity & Rliability of Measurment

What is reliability?

Validity & Rliability of Measurment

What is reliability?

Roughly speaking, reliability is how much of the observed differences in a variable are real and not due to random noise.

Validity & Rliability of Measurment

According to Classical Test Theory:

$$\text{Observed score} = \text{True score} + \text{Error}$$

Similarly:

$$\text{Observed variance} = \text{True variance} + \text{Error variance}$$

Reliability then:

$$\text{Reliability} = \frac{\text{True variance}}{\text{Observed variance}}$$

Validity & Rliability of Measurment

The problem: To compute correlations, we use observed variance:

$$\text{Cor}_{x,y} = \frac{\text{Cov}_{x,y}}{\sqrt{\text{Var}_x \cdot \text{Var}_y}} = \frac{\text{Cov}_{x,y}}{\sqrt{(\text{True Var}_x + \text{Error Var}_x) \cdot (\text{True Var}_y + \text{Error Var}_y)}}$$

Lower reliability → Higher error variance → Higher observed variance → Lower observed correlation.

Validity & Rliability of Measurment

This problem is called **attenuation** or regression dilution.

Example: The true correlation between wellbeing and school performance is 0.8. Both variables are measured with $\text{rel} = 0.6$.

$$0.48 = 0.8 \cdot \sqrt{0.6 \cdot 0.6}$$

Reliabilities

Observed correlation

True correlation

Validity & Rliability of Measurment

Attenuation is as much a problem for regression models as is for correlation.

Low reliability of the outcome variable **increases random error**.

Low reliability of predictors **biases regression coefficients**.

Low validity of any variable **biases regression coefficients**.

Questions?

Representativity of Data

Representativity of Data

Representativity: how closely the distribution of variables in our sample matches the distribution in population.

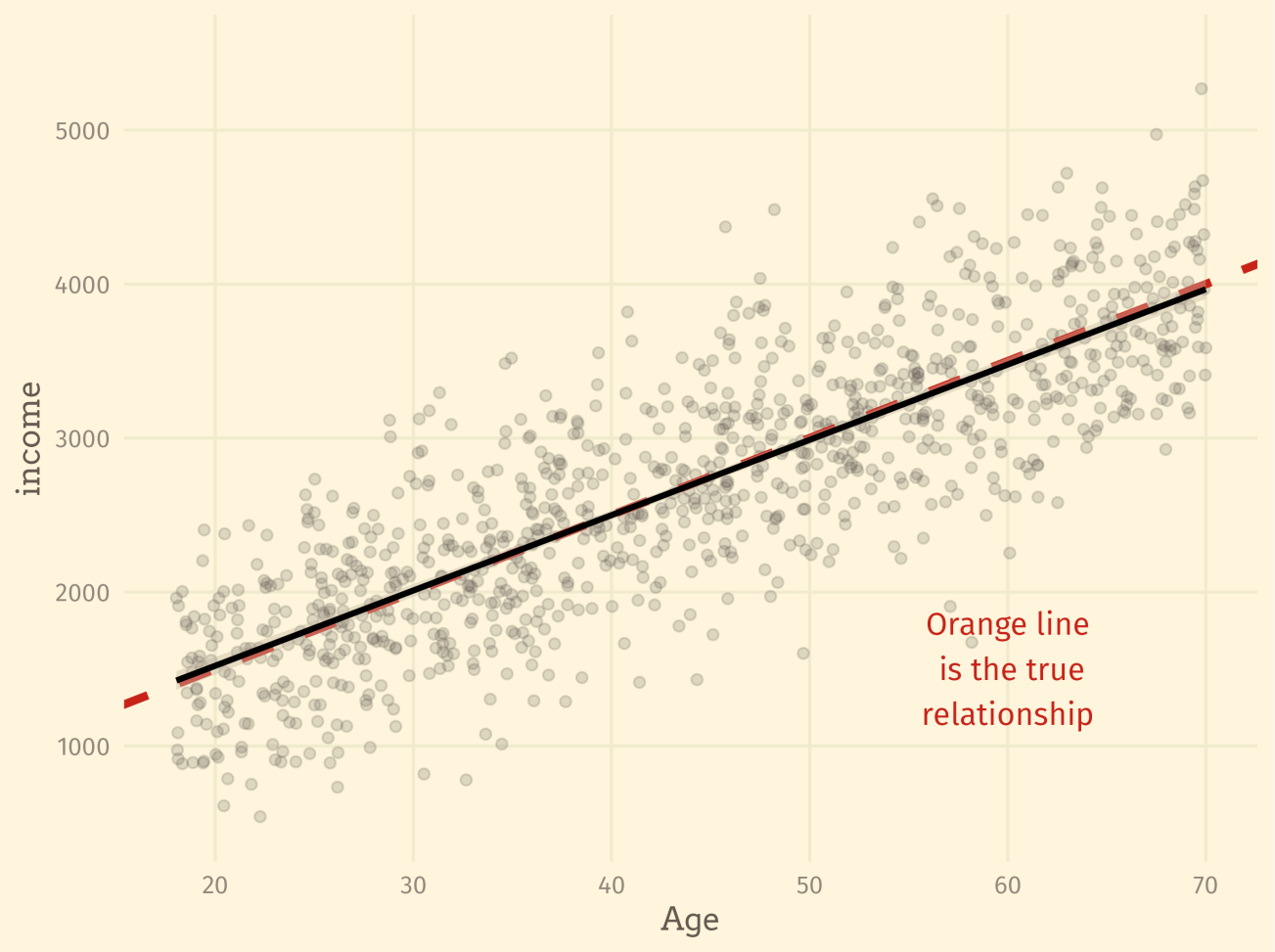
Nonrepresentative data **biases our regression coefficients**.

But not always...

Representativity of Data

We want to estimate relationship between age and income.

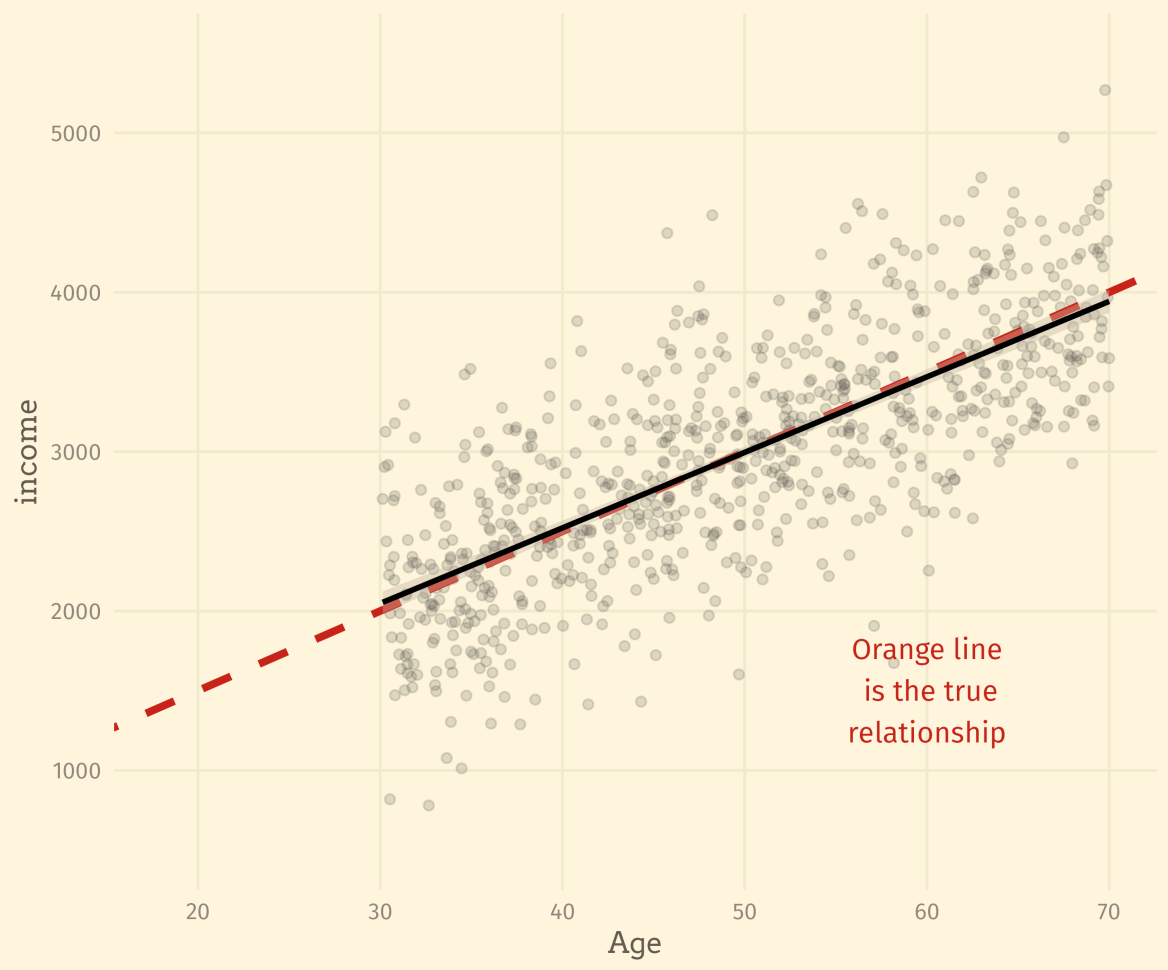
What happens if people below 30 refused to participate?



Representativity of Data

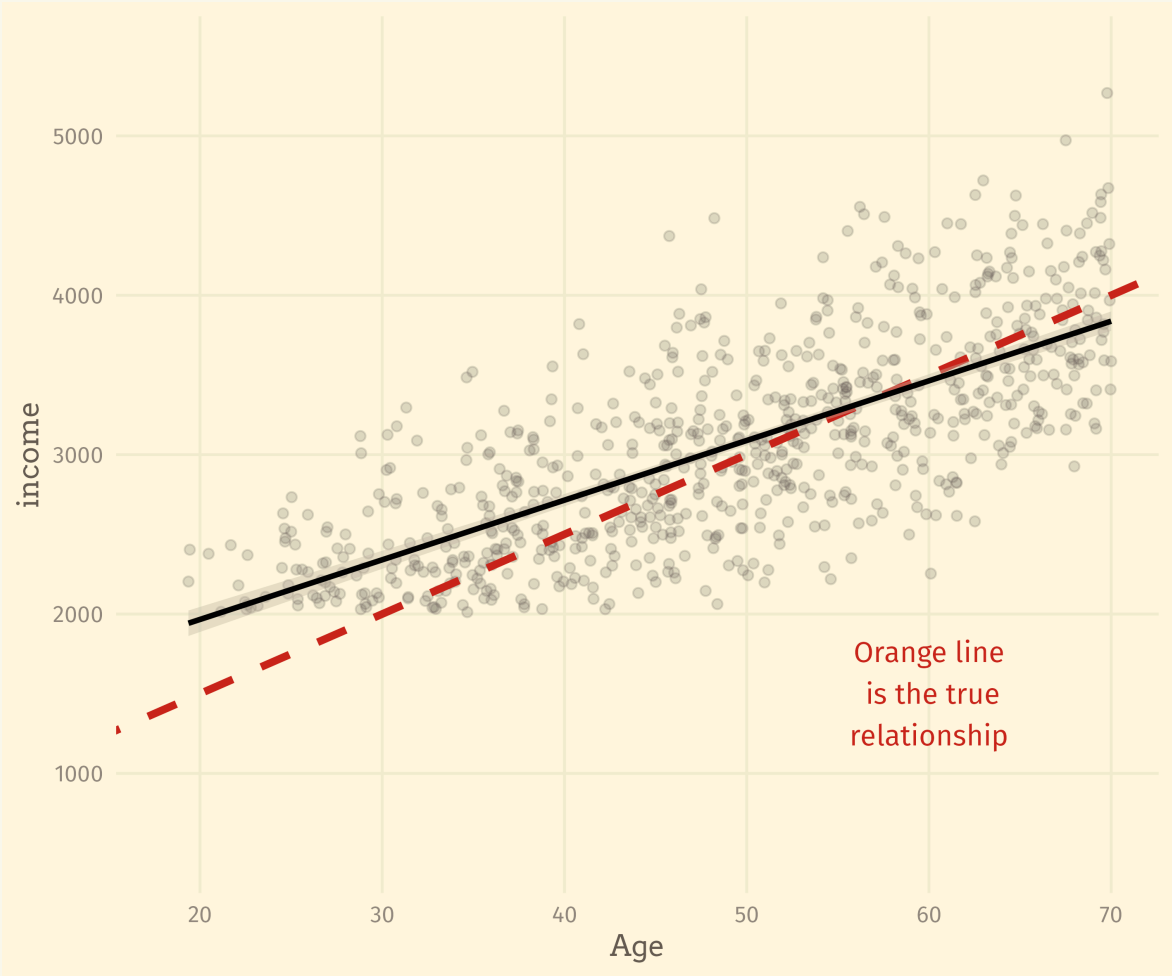
Nonrepresentativity of predictors doesn't matter for slope estimates!

What if people in the lowest income quartile refused to participate?



Representativity of Data

Nonrepresentativity in the outcome does introduce bias!



Representativity of Data

Nonrepresentativity in the outcome leads to **biased regression coefficients**.

Questions?

Linearity (& Additivity)

Linearity (& Additivity)

The linarity assumptions states that the outcome can be predicted using a **linear combination of predictors**.

In other words, the model has to be able to predict the outcome by **summing up effects of our predictors**.

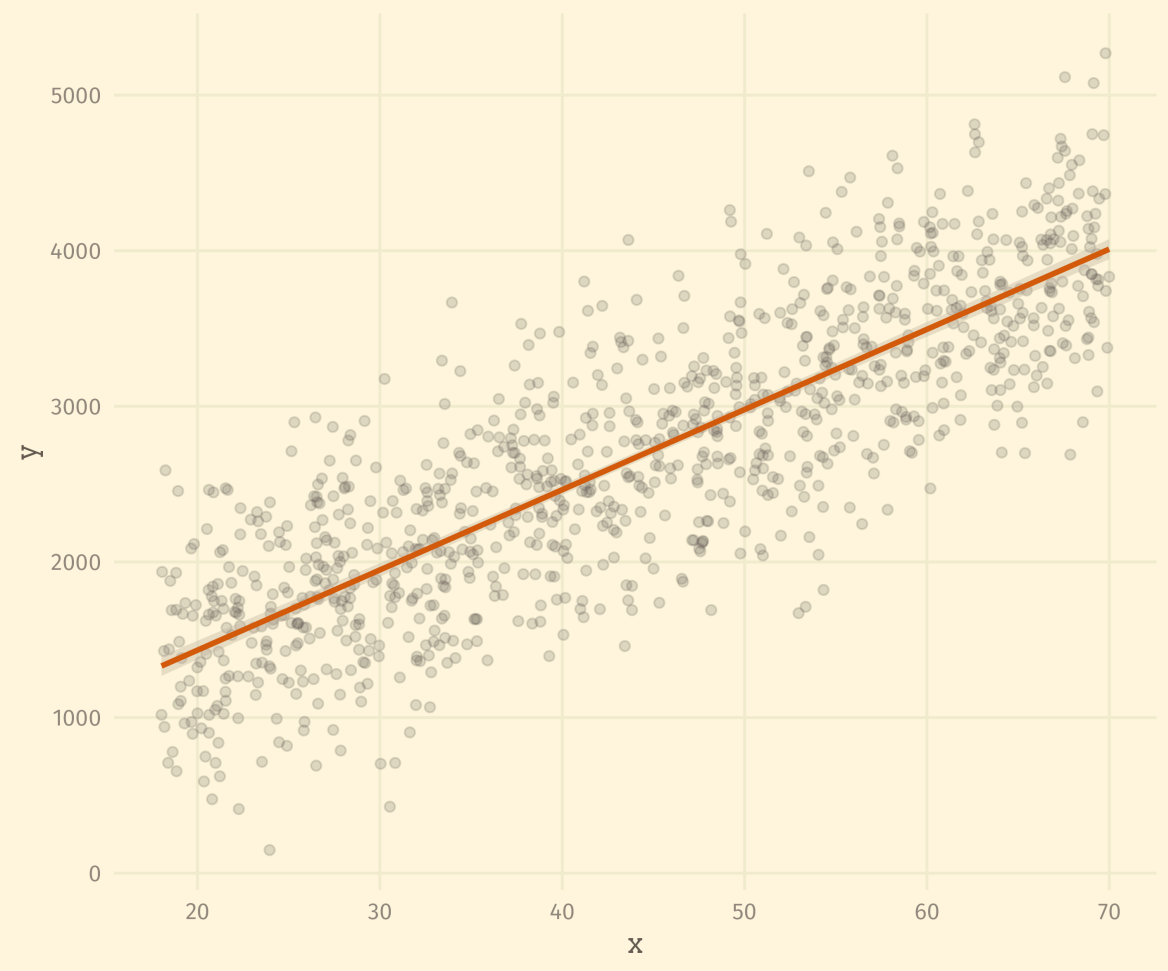
It *doesn't* mean the relationships have to be linear.

Linearity (& Additivity)

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

Linearity assumptions is met.

y is successfuly predicted by x.

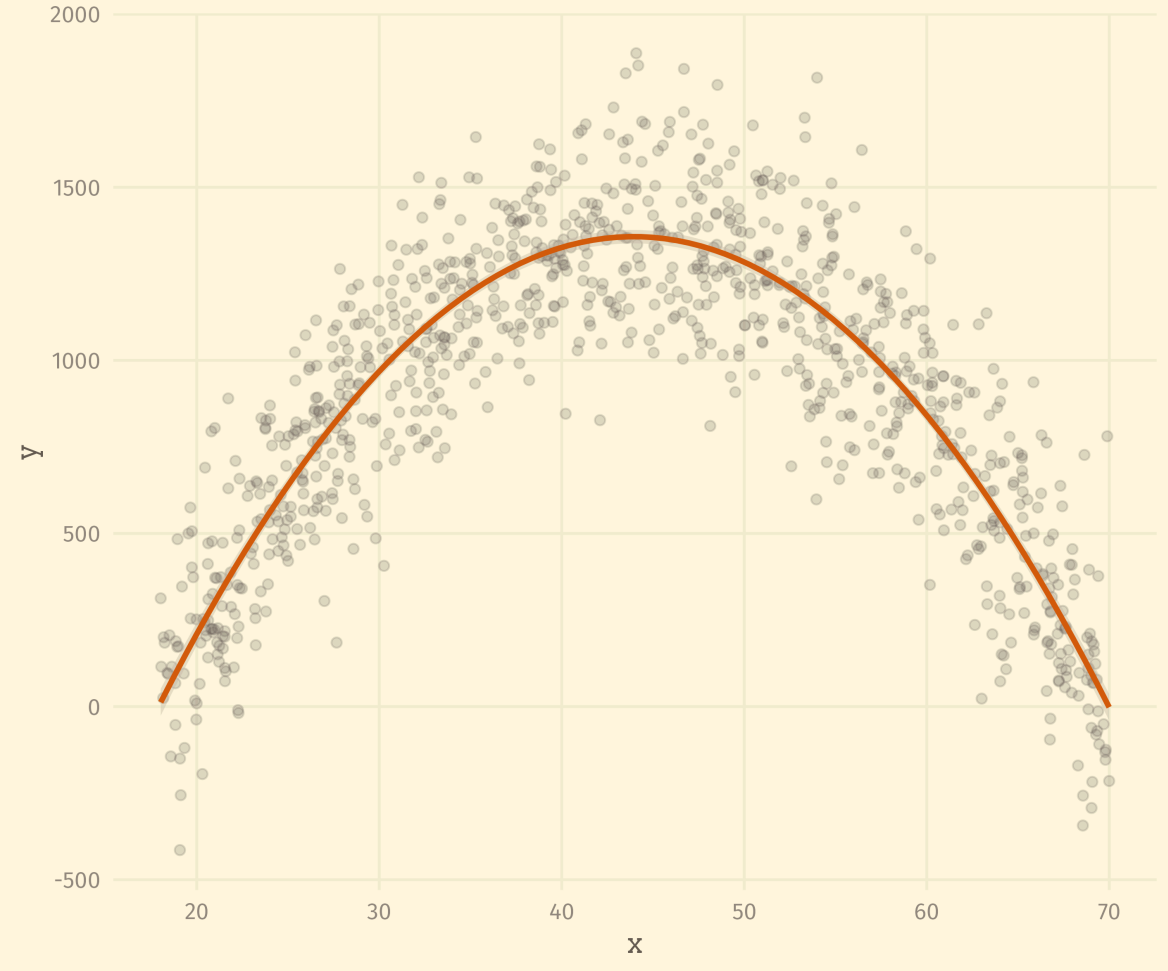


Linearity (& Additivity)

$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$$

Linearity assumptions is met.

Y is successfuly predicted by a sum of X and X^2.

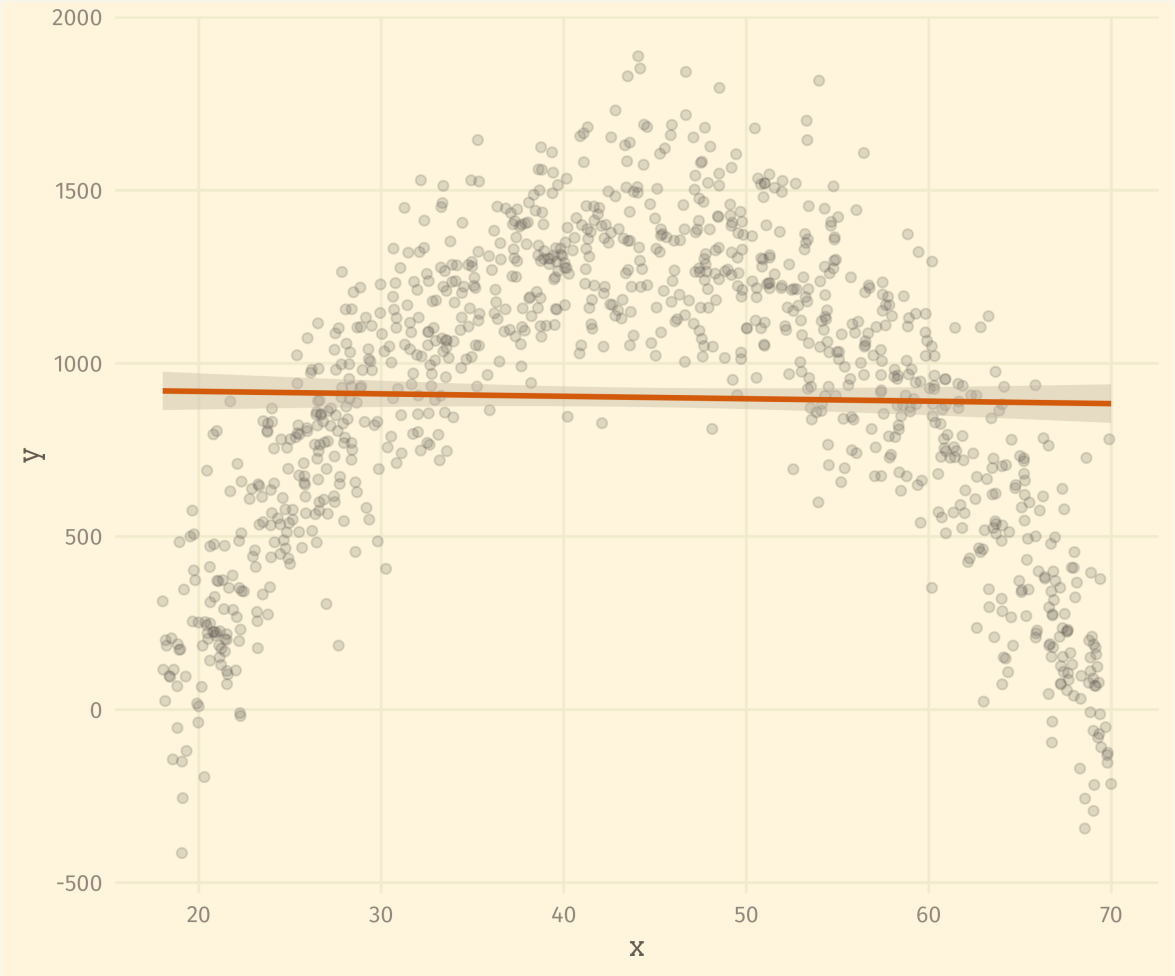


Linearity (& Additivity)

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

Linearity assumptions is not met.

x alone is not enough to predict y.



Linearity (& Additivity)

Roughly speaking, the linearity assumptions states we have included all important predictors in correct form.

Violating linearity assumptions leads to **biased regression coefficients**.

Questions?

Independence of Errors

Independence of Errors

Independence assumption states the errors shouldn't be correlated with each other.

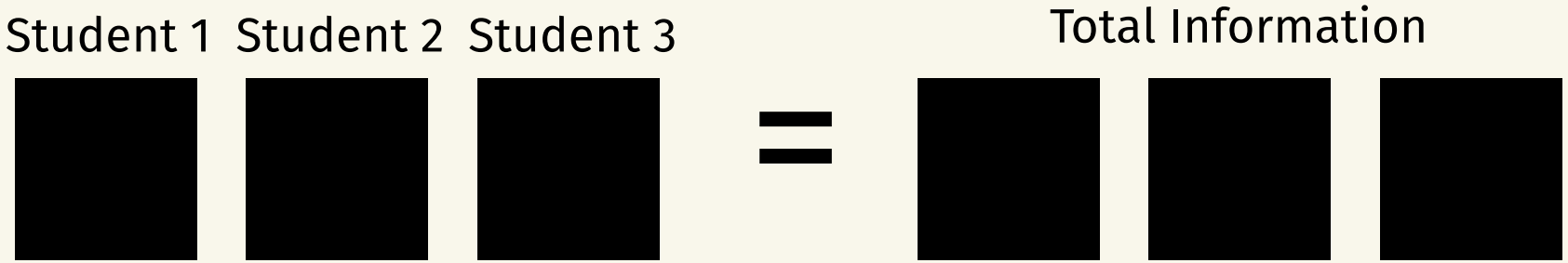
In other words, we assume there is no relationship between observations beyond our model.

Independence of Errors

Example: We are doing research on among high school students.

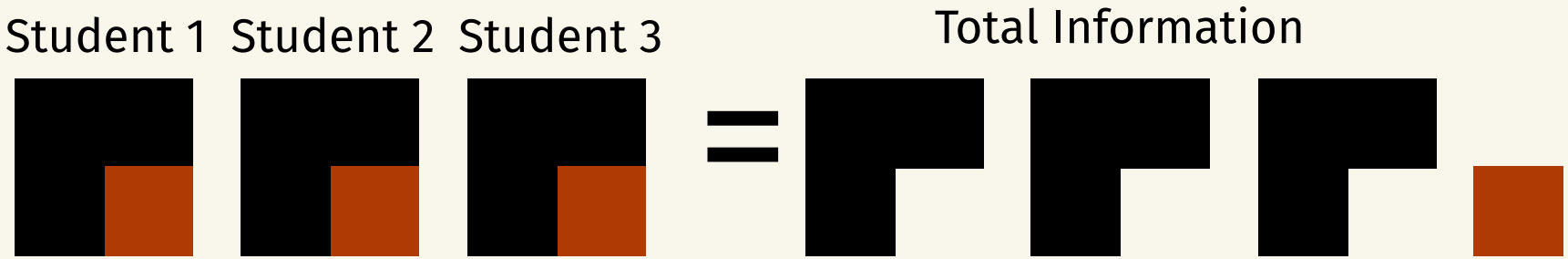
Every student in our study provides a „bit“ of information.

Students from **different schools all provide unique information**:



Independence of Errors

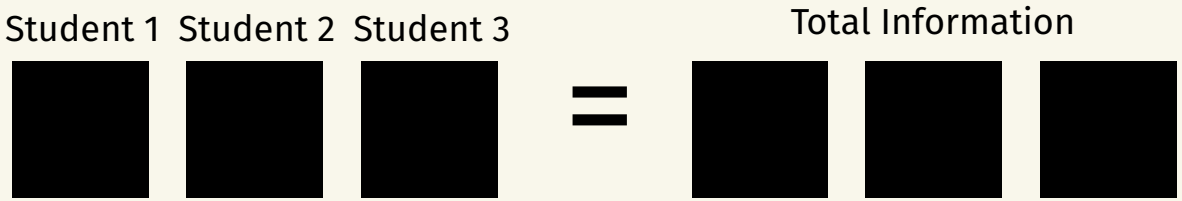
But students from the **same school** share the experience → the information they provide is **not completly unique**:



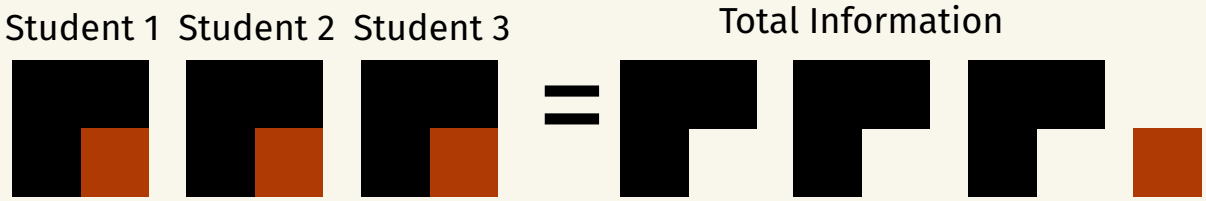
Independence of Errors

Violating independence assumption means **we think our data contain more information then they actually do**.

Different schools:



Same school:



Questions?

Homoskedasticity of Errors

Homoskedasticity of Errors

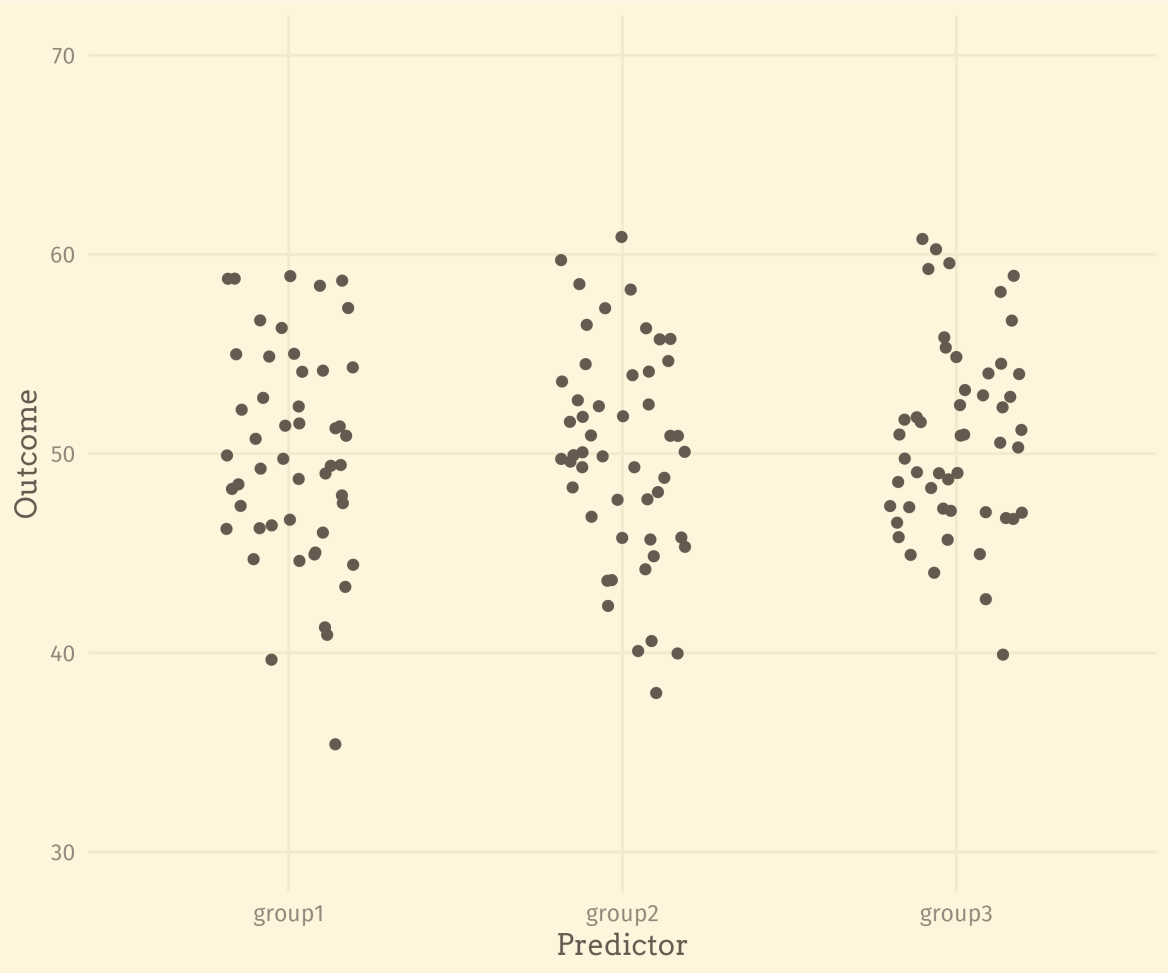
Homoskedasticity assumption states the **error variance is constant**.

In other words, we assume the variance of the dependent variable is the same for all predictors values.

Homoskedasticity of Errors

Homoskedasticity assumption is met.

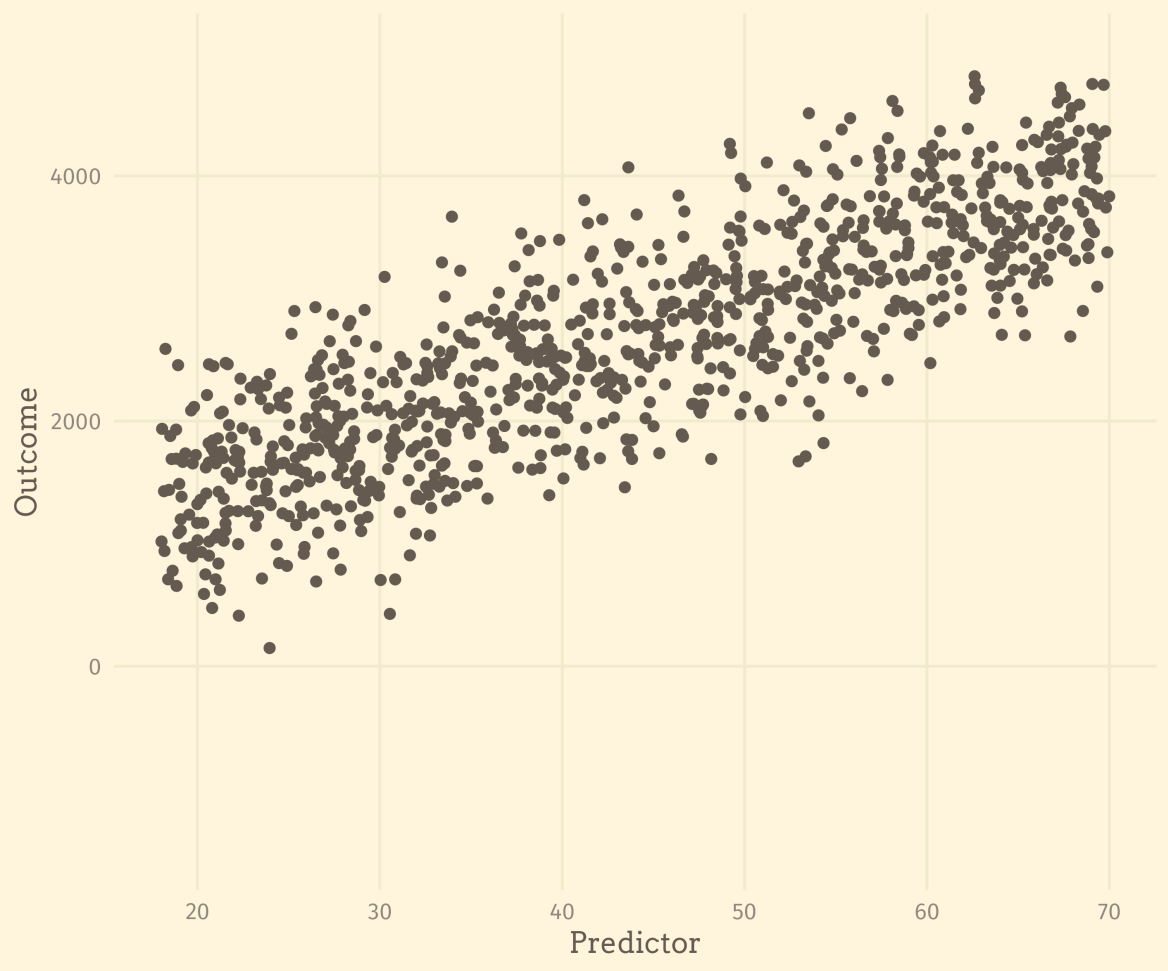
All groups have the same variance.



Homoskedasticity of Errors

Homoskedasticity assumption is met.

Variance is the same for all values of the predictor.



Homoskedasticity of Errors

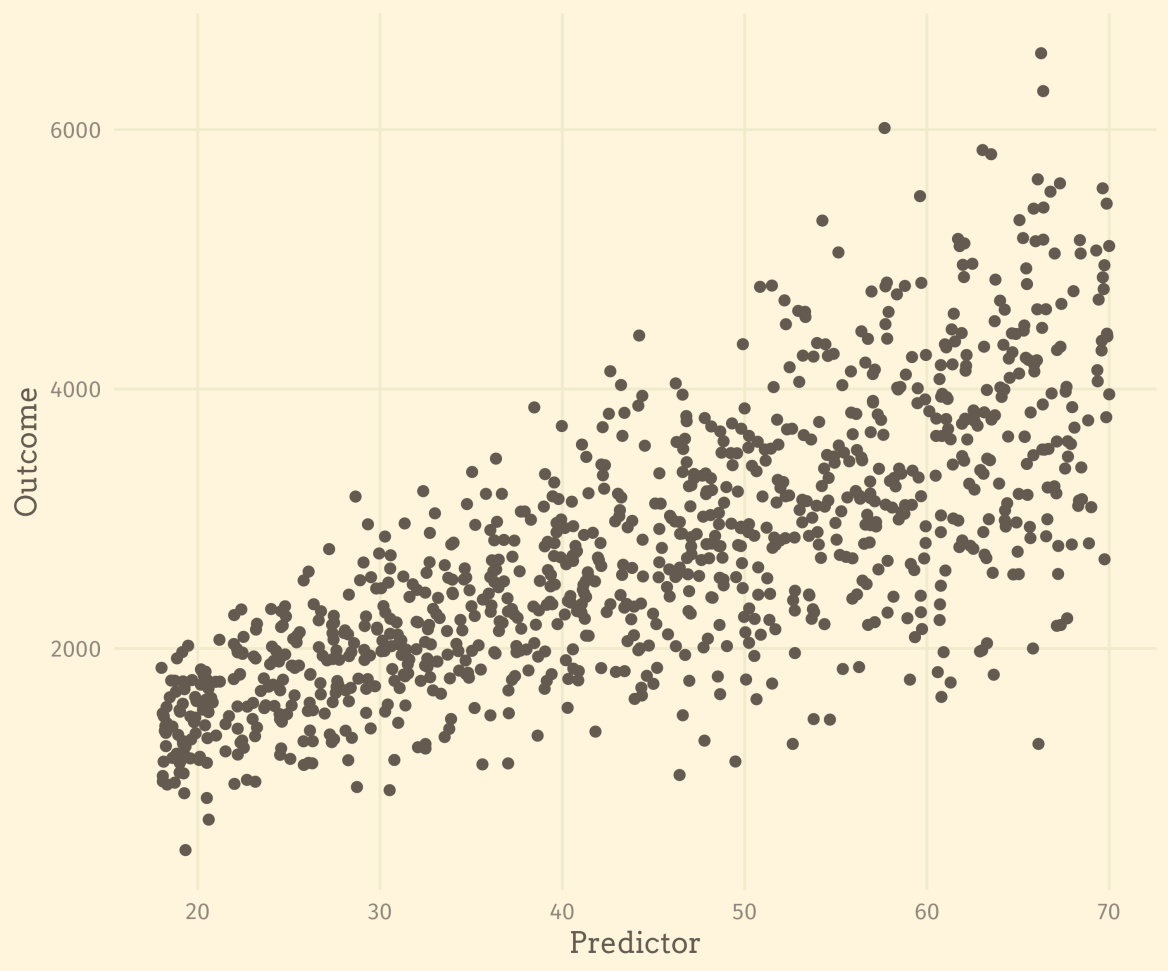
Homoskedasticity assumption is violated.

Variance depends based on group.

Homoskedasticity of Errors

Homoskedasticity assumption is violated.

Variance depends based on predictor value.



Homoskedasticity of Errors

Violating the homoskedasticity assumptions leads to **bias in standard errors**.

Questions?

Normality of Errors

Normality of Errors

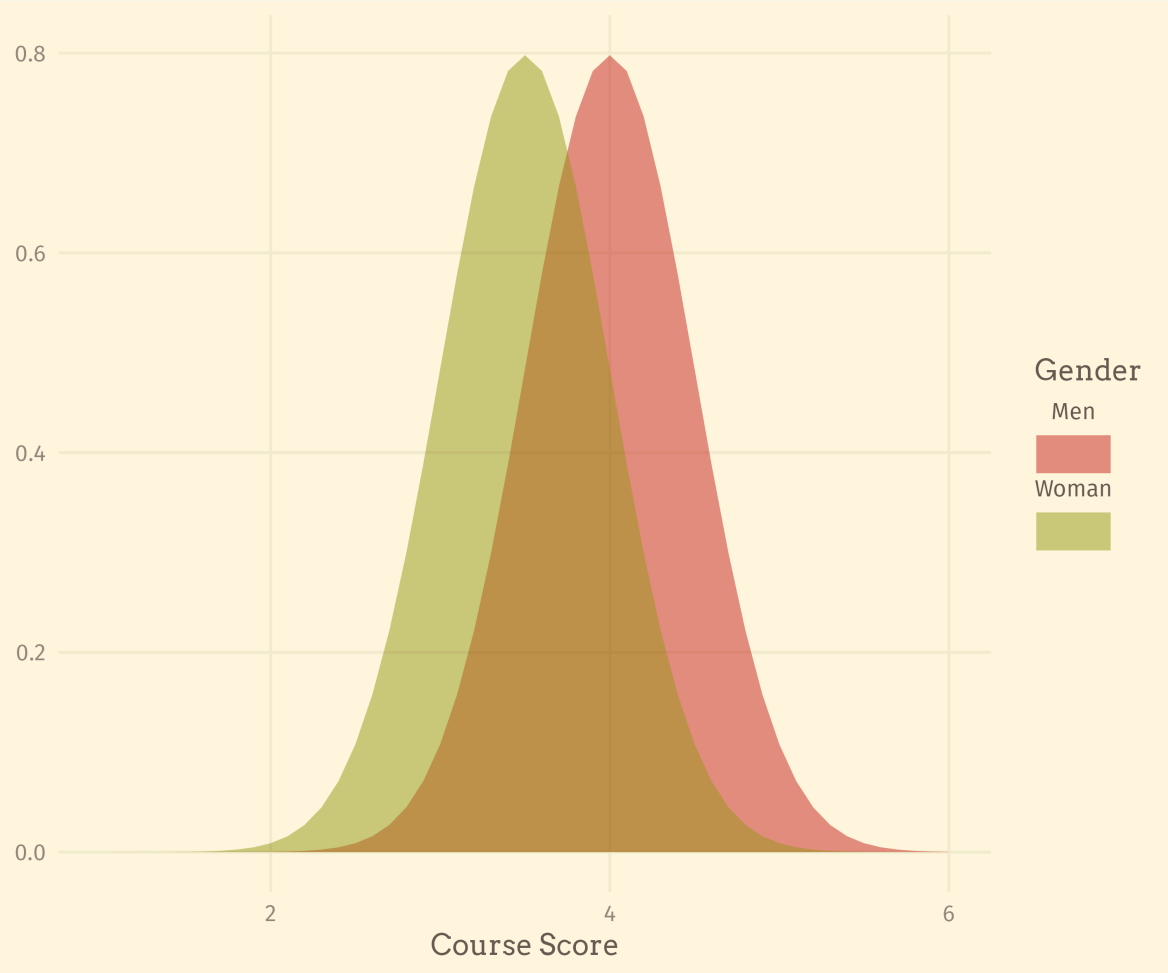
Normality assumption states that **errors are normally distributed**.

In other words, the individual observations should be normally distributed around the expected value.

Normality of Errors

Normality assumption is met.

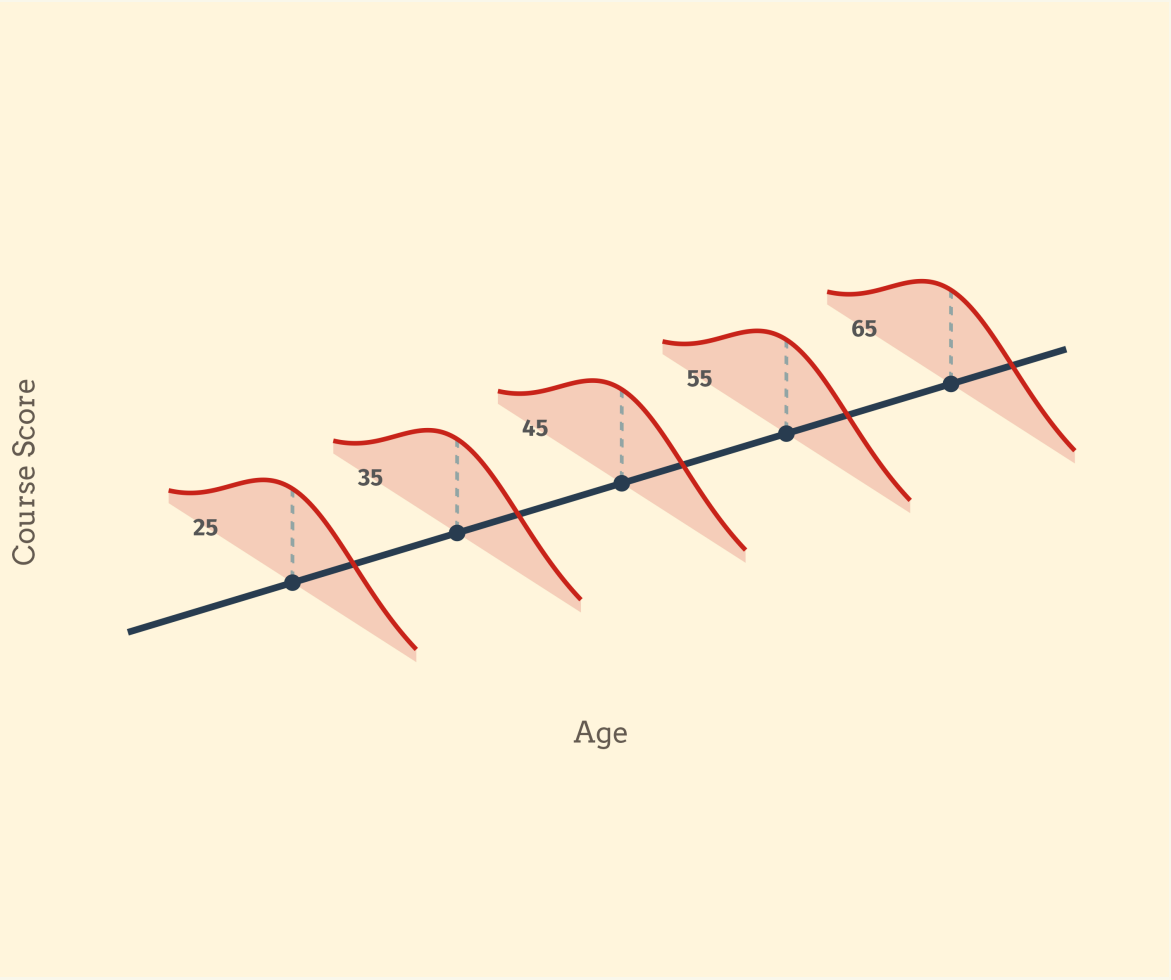
Both genders have normally distributed course scores.



Normality of Errors

Normality assumptions is met.

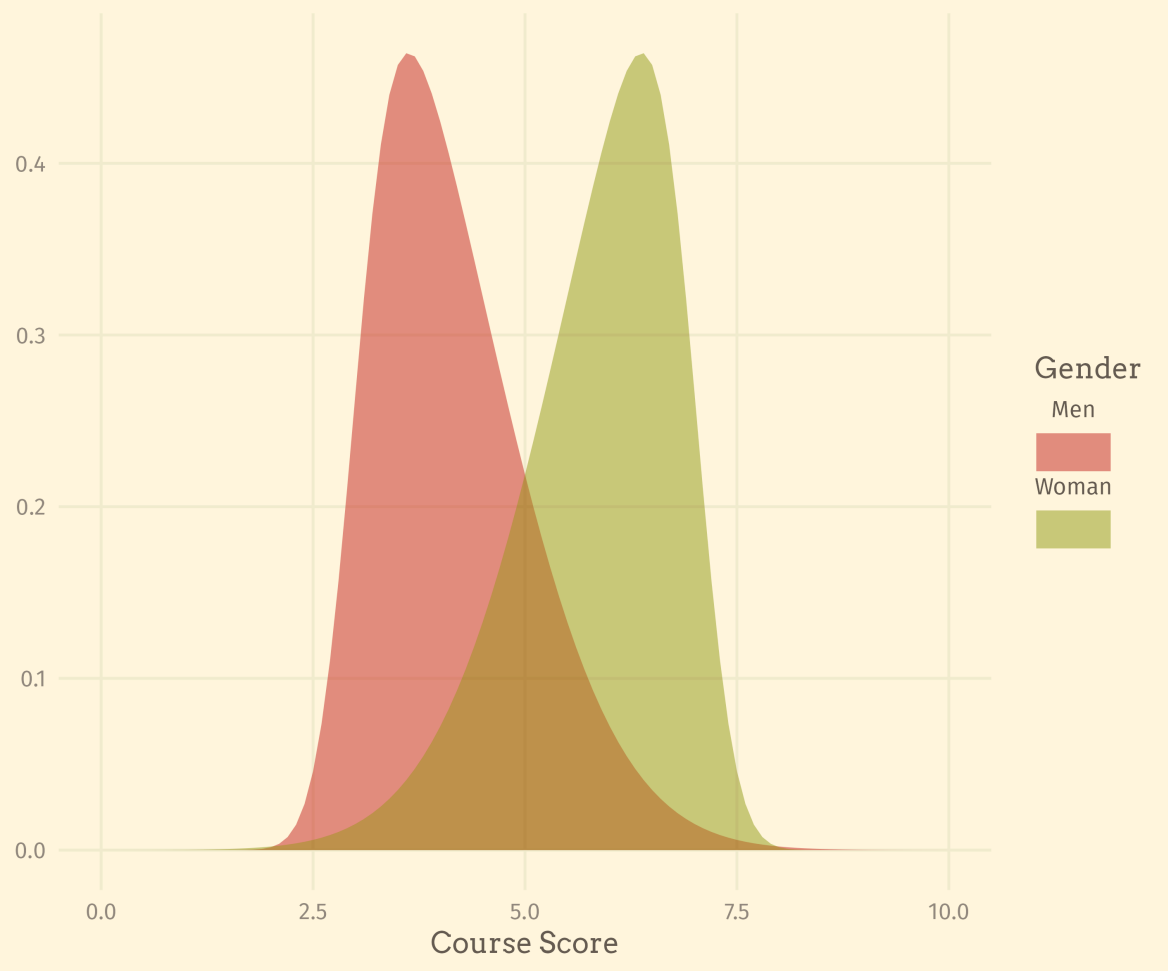
All age groups have normally distributed course scores.



Normality of Errors

Normality assumption is violated.

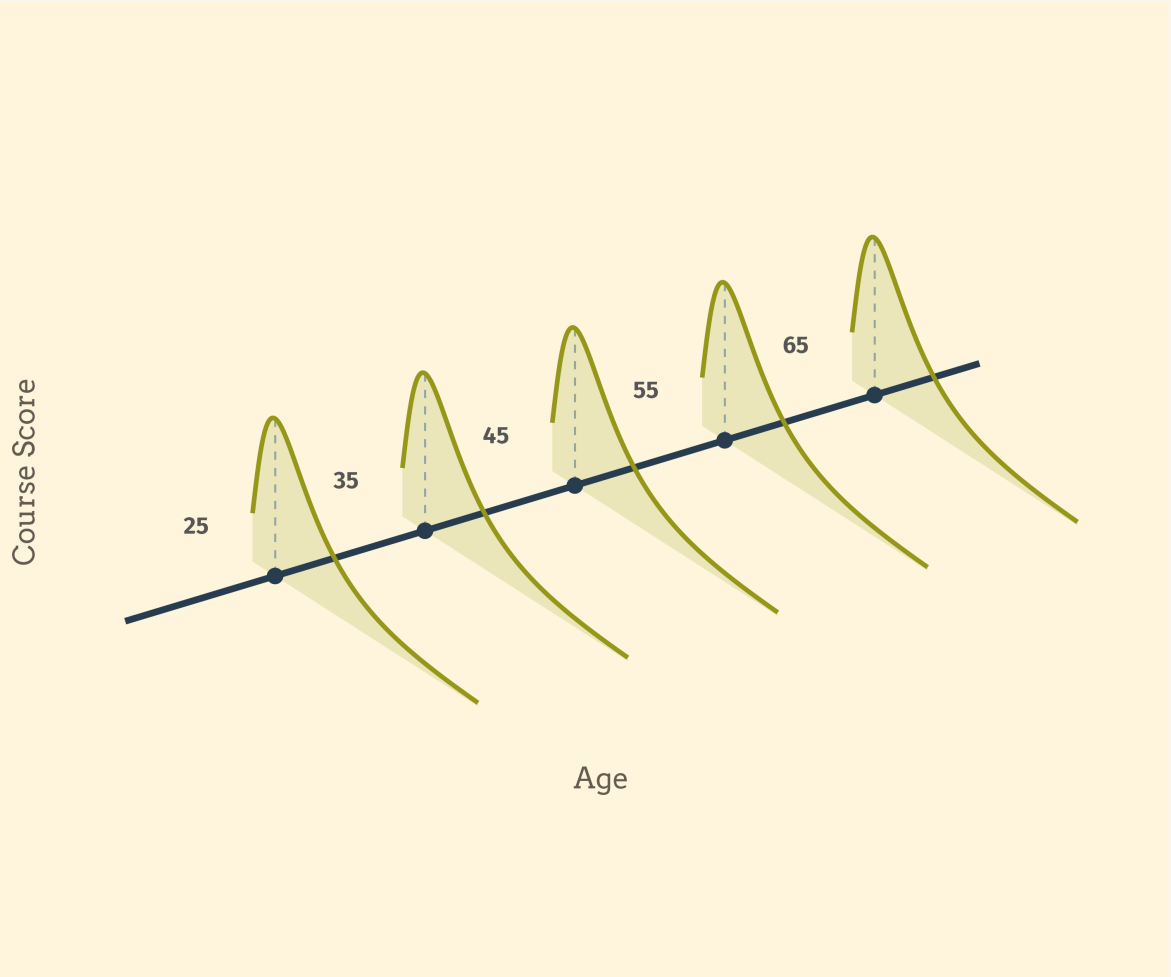
Both groups have skewed distributions.



Normality of Errors

Normality assumption is violated.

Dependent variable is skewed for all ages.



Normality of Errors

Violating normality assumption leads to incorrect predictions and **bias in standard errrors (in small samples)**.

Questions?

Assumptions Roundup

Assumption	When violated
Validity & Reliability	Biased regression coefficients
Representativity	Biased regression coefficients
Linearity	Biased regression coefficients
Independence	Biased standard errors
Homoskedasticity	Biased standard errors
Normality	Biased standard errors (in small samples)

Questions?

Checking Assumptions

Checking Assumptions

Not all assumptions can be checked.

Not all assumptions can be checked easily.

For the assumptions that can be checked, diagnostic plots are your best option.

Checking Assumptions - Validity & Reliability of Measurment

Can’t be easily checked.

It’s possible to check indirectly, but it’s laborous and non-trivial.

Predictive validity, factor analysis, Cronbach’s alpha, etc.

Checking Assumptions - Representativity

Can't be easily checked.

You need to know your population and think carefully about research design.

Partially can be checked against known population data (e.g. census).

Checking Assumptions - Linearity

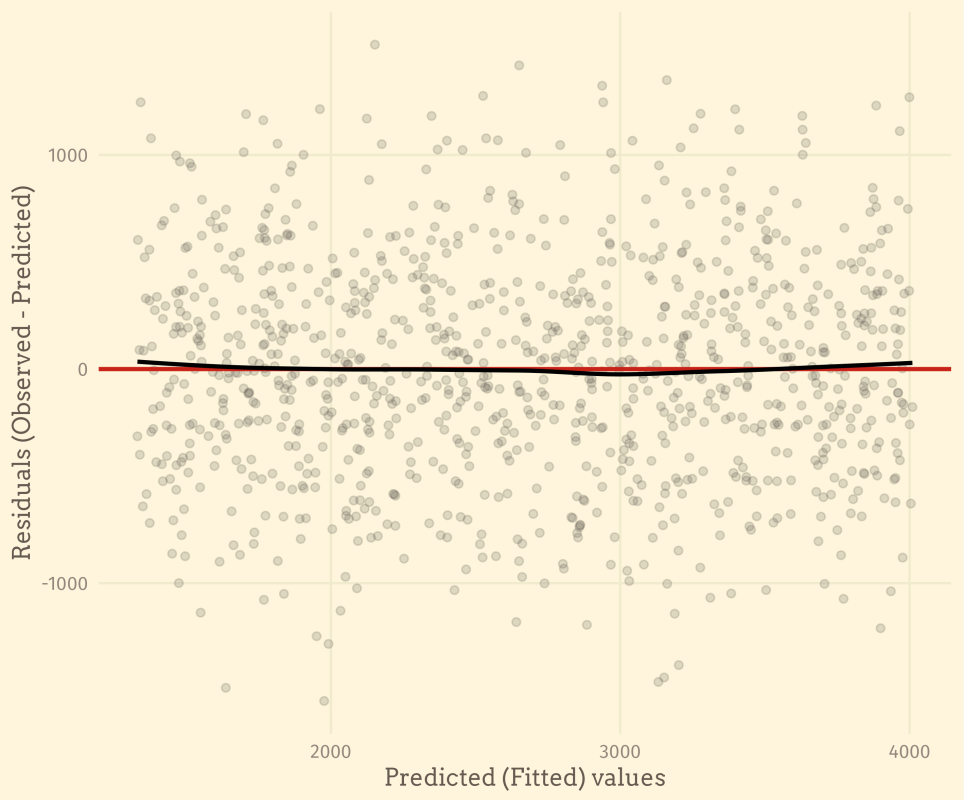
Can be checked!

We can look at the relationship between predicted values and residuals using **residual plot**.

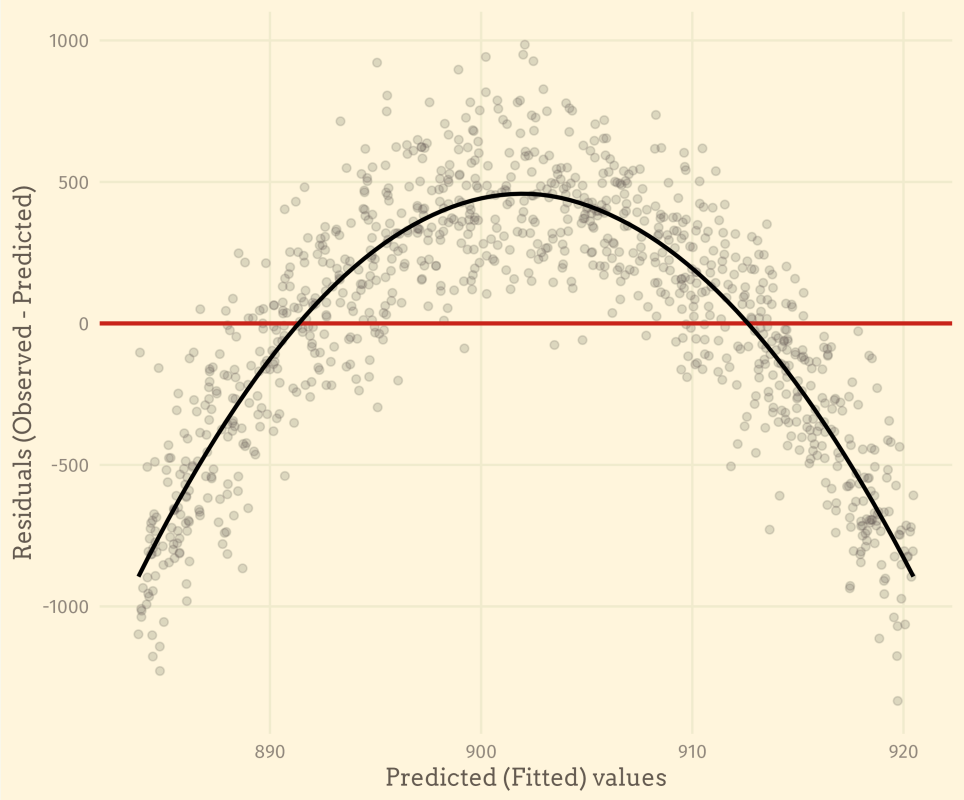
If the linearity assumption holds, the **residuals will be spread around zero without a pattern** (i.e. the model doesn't systematically over- or underestimate the data).

Checking Assumptions - Linearity

Linearity holds



Linearity violated



Checking Assumptions - Independence

Can't be easily checked.

Can happen for many reasons - clustering, time series.

You need to think about the nature of your data and structure your model accordingly.

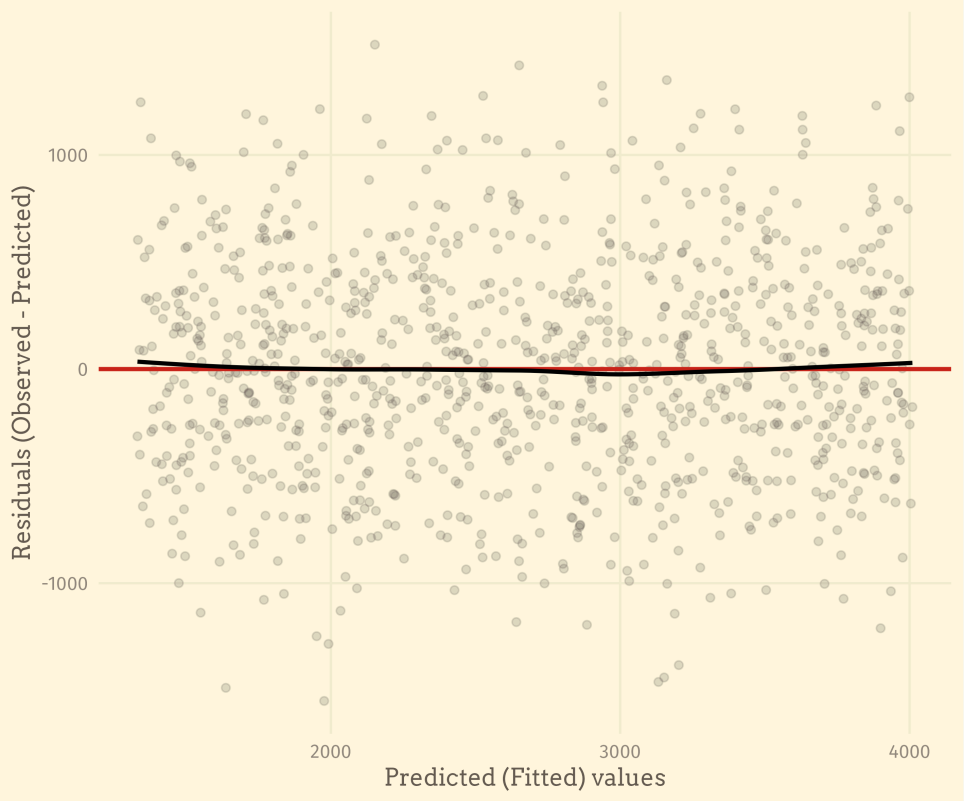
Checking Assumptions - Homoskedasticity

Can be check the same **residual plot** as linearity.

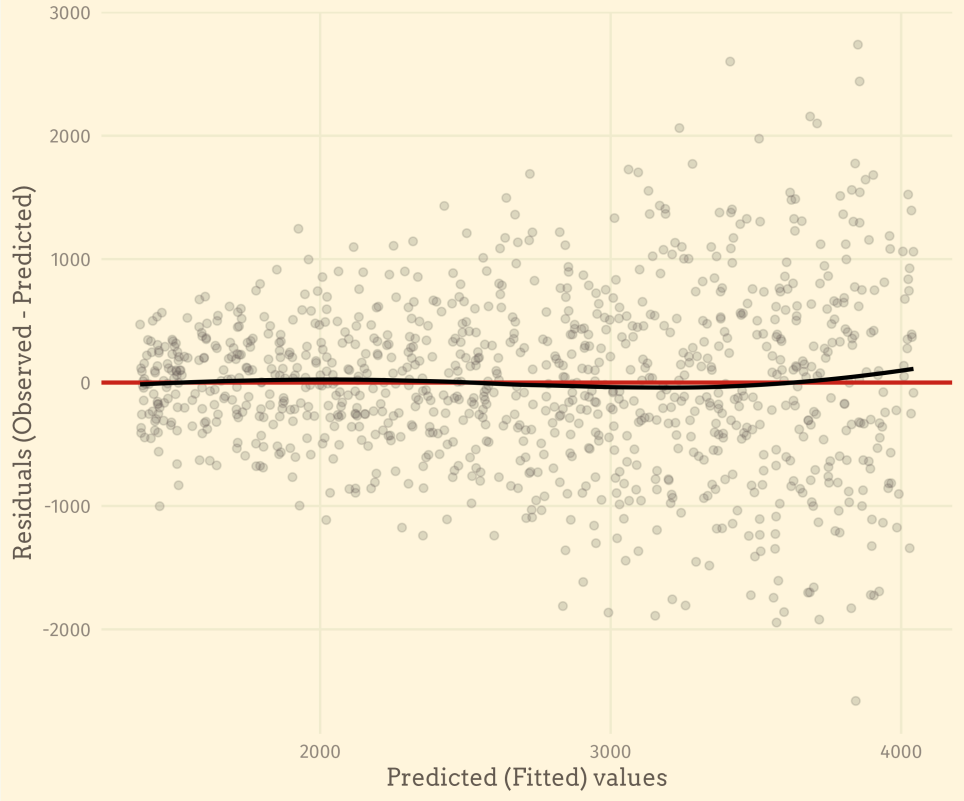
The spread/ **variance of residuals should be the same** for all predicted values.

Checking Assumptions - Homoskedasticity

Homoskedasticity holds



Homoskedasticity violated



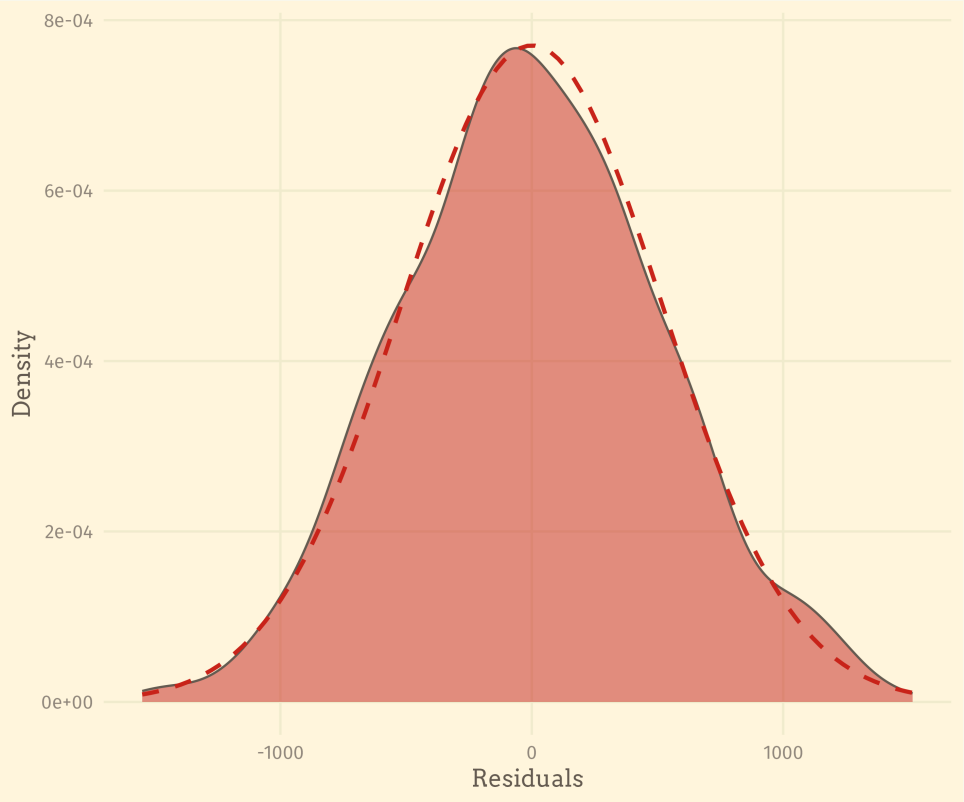
Checking Assumptions - Normality

Can be checked using either **density/histogram plots** or Q-Q plots.

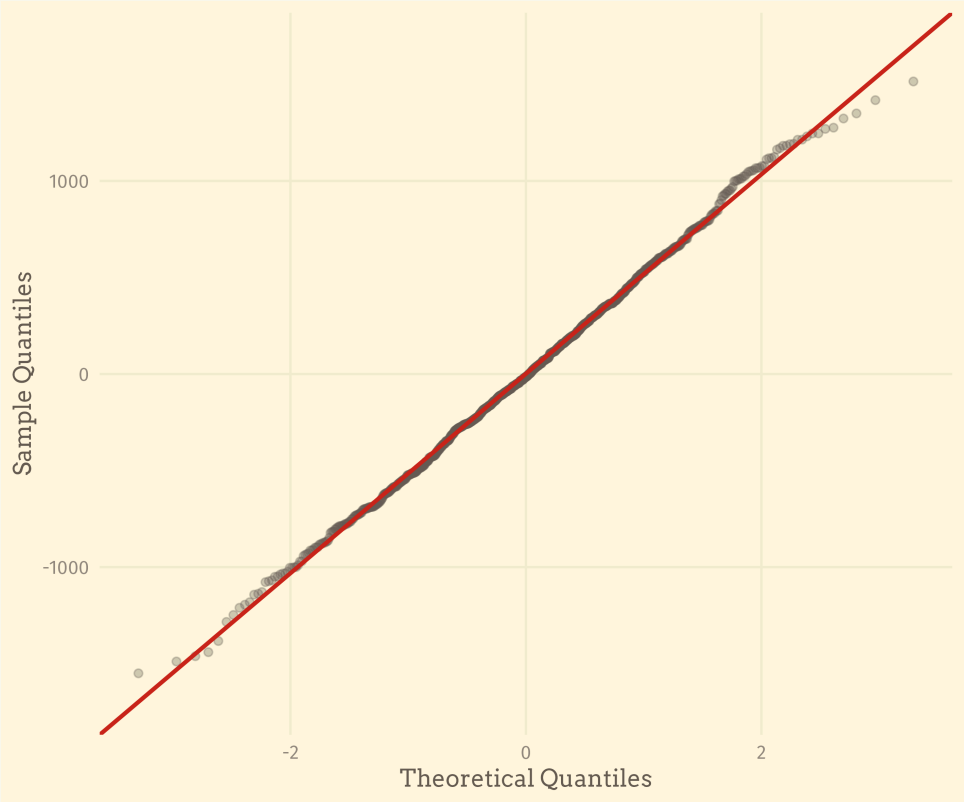
If the normality assumption is met, the density **plot of residuals should show normal distribution** and Q-Q plot should show linear relationship between theoretical and observed quantiles.

Checking Assumptions - Normality

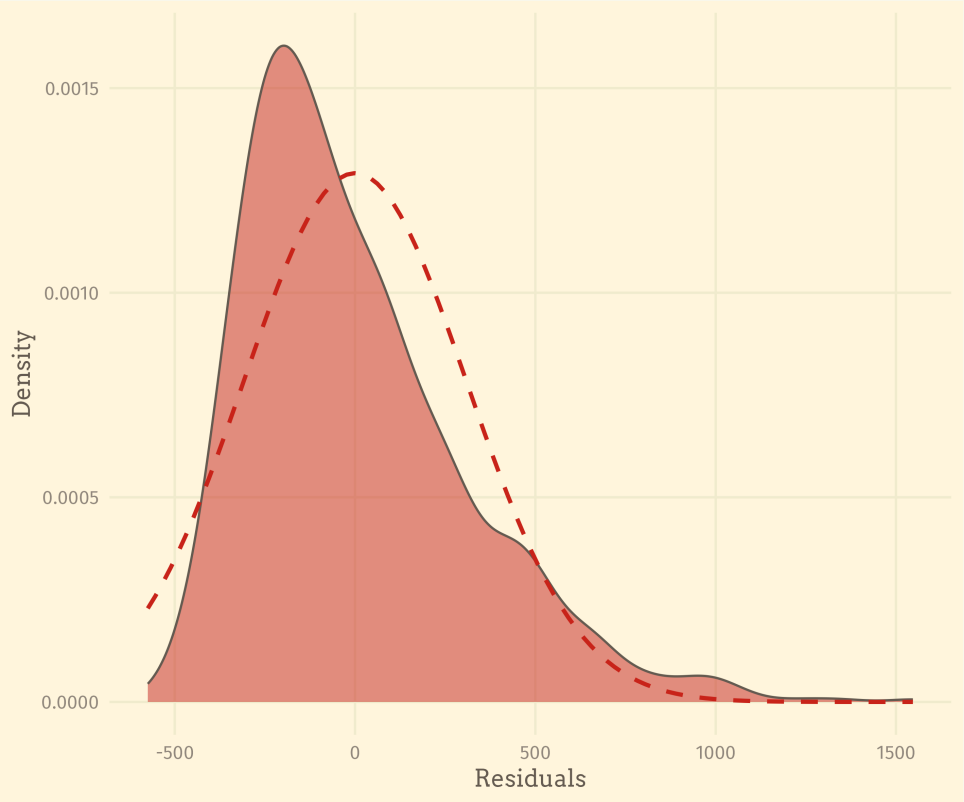
Normality met

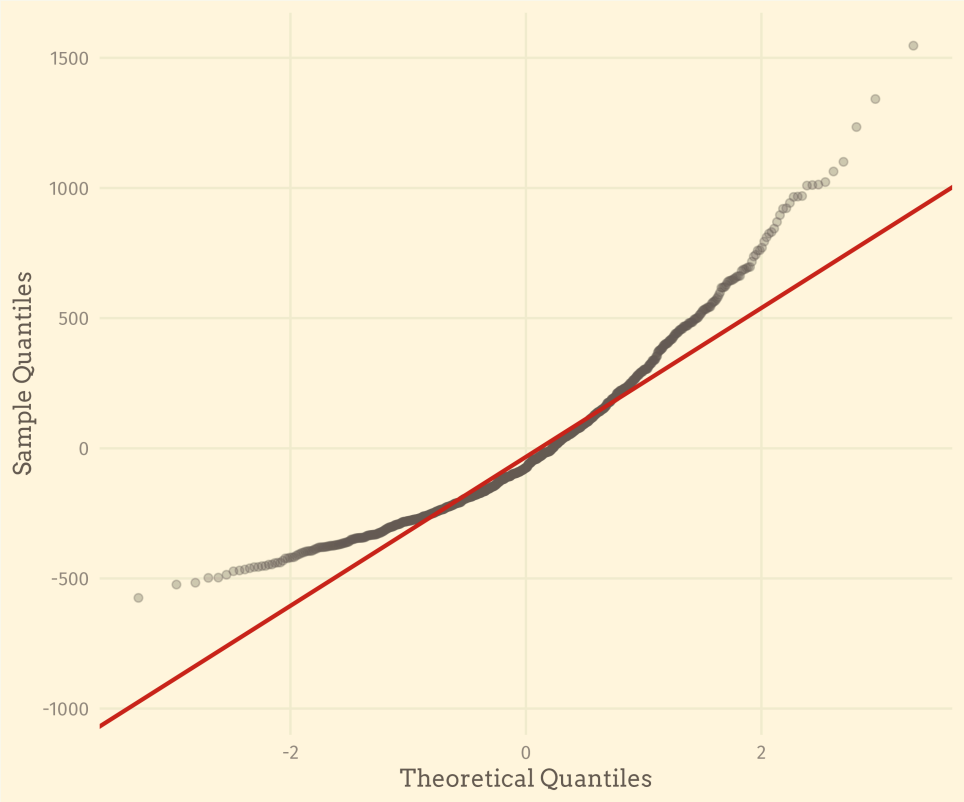


Normality met



Checking Assumptions





Assumptions Checking Roundup

Assumption	How to check
Validity & Reliability	No easy way (factor analysis, Cronbach's alpha)
Representativity	No easy way to check
Linearity	Residual plot
Independence	No easy way to check
Homoskedasticity	Residual plot
Normality	Density plot/Q-Q plot

Questions?

InteRmezzo!

Caveats (& More Rants)

Caveats (& More Rants)

Sometimes it's not clear whether an assumption is „close enough".

My advice: *If you catch yourself not being sure about an assumption, try to replace it and check how much the results.*

Caveats (& More Rants)

Some people will tell you to use hypothesis tests to check model assumptions.

Shapiro-Wilks normality test, Levene's homoskedasticity test, etc.

This is universaly a bad idea.

Don't use tests to check model assumptions!

Caveats (& More Rants)

Example: Normality tests check whether residuals are perfectly normal.

But normal distribution is a theoretical model, it doesn't actually exist in real life.

The null hypothesis is guaranteed to be wrong.

Caveats (& More Rants)

Normal distribution

Caveats (& More Rants)

Normal distribution

- Continuous
- Unbounded
- Symmetric

Likert Items

Caveats (& More Rants)

Normal distribution

- Continuous
- Unbounded
- Symmetric

Likert Items

- Discrete
- Bounded
- Almost always skewed.

Questions?

Caveats (& More Rants)

Collinearity is the correlation between predictors.

Some people care a lot about it.

It's mostly not neccesary.

Caveats (& More Rants)

Collinearity increases standard errors of regression coefficients.

This is correct - when predictors are correlated, it's harder to disentangle their effects.

(Near) perfect collinearity leads to instability during estimation, but cases like this are rare.

Questions?