

Where are we?

Introduction

Interpretation

Two sources of variation in the data

Study technique

Interpretation

Silje Synnøve Lyder Hermansen

16 februar 2026

Where are we?

Where are we?

- ▶ week 1:
 - ▶ purpose of the course
 - ▶ R as an object oriented language
- ▶ week 2:
 - ▶ dialects in R
 - ▶ descriptive statistics:
 - ▶ measurement level and choice of descriptives
 - ▶ data exploration
- ▶ week 3: (this week)
 - ▶ linear regression (OLS)
 - ▶ interpretation
 - ▶ non-linear effects

Plan for the day

- ▶ lecture: uncertainty and interpretation of linear models
 - ▶ substantive interest: the size of the effect
 - ▶ statistical significance: sources of variation/uncertainty
 - ▶ R notebook 1: interpretation (Gelman and Hill, King et al)
- ▶ study technique: how to use AI/LLMs in this class
- ▶ Thursday:
 - ▶ implementation in R
 - ▶ R notebook 2: non-linear effects (Berry et al)

Introduction

Today's example

What is the effect of electoral systems on parliamentarians resource allocation?

- ▶ Members of the European Parliament (MEPs) sit together in one institution, but run for election under different rules
- ▶ expectation: more local investment among MEPs in candidate-centered systems (compared to party-centered systems), because of their need for a personal brand
- ▶ variables:
 - ▶ y: number of constituency-level assistants employed (metric)
 - ▶ x : candidate vs. party-centered systems (binary)

Two views on linear regression

Two views on linear regression

Linear regression summarizes how the average values of a numerical outcome variable vary over subpopulations defined by linear functions of predictors. (Gelman and Hill, 2007, ch 3)

- ▶ **comparison of means:** descriptive approach to regression; makes sense for categorical predictors
- ▶ **relationship between variables:** their correlation; more causal, makes sense for numerical predictors

A comparison of means: group means

Most obvious when my predictor is categorical

```
df %>%
  group_by(OpenList) %>%
  reframe("mean_y" = mean(LocalAssistants)) %>%
  ungroup %>%
  mutate(diff = mean_y - lag(mean_y))
```

```
## # A tibble: 2 x 3
##   OpenList mean_y   diff
##       <int>   <dbl>   <dbl>
## 1       0     2.47    NA
## 2       1     3.42    0.949
```

- ▶ MEPs from *party-centered* systems employ on average 2.47 local assistants
- ▶ MEPs from *candidate-centered* systems employ on average 3.42 local assistants.
- ▶ The difference is 0.95

A comparison of means: regression

Most obvious when my predictor is categorical

```

#Estimate the equation
mod <- lm(LocalAssistants ~ OpenList,
           df)
#Summarize the results
summary(mod)

##
## Call:
## lm(formula = LocalAssistants ~ OpenList, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.42  -2.42  -0.47   1.53  36.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.468     0.161   15.35 < 2e-16 ***
## OpenList      0.949     0.234    4.05  5.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.2 on 737 degrees of freedom
## Multiple R-squared:  0.0218, Adjusted R-squared:  0.0204 
## F-statistic: 16.4 on 1 and 737 DF,  p-value: 5.68e-05

```

- ▶ MEPs from *party-centered* systems employ on average 2.47 local assistants
- ▶ The *difference* is 0.95.
- ▶ MEPs from *candidate-centered* systems employ on average $2.47 + 0.95 = 3.42$ local assistants.

Relationship between variables: regression

More descriptive statistics

```
mod2 <- lm(LocalAssistants ~ OpenList + LaborCost,
            df)

summary(mod2)

##
## Call:
## lm(formula = LocalAssistants ~ OpenList + LaborCost, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.49  -1.94  -0.41   1.08  35.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.1266    0.2861   14.42 < 2e-16 ***
## OpenList     0.8288    0.2278    3.64  0.00029 ***
## LaborCost   -0.0702    0.0102   -6.91   1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.1 on 736 degrees of freedom
## Multiple R-squared:  0.0814, Adjusted R-squared:  0.0789
## F-statistic: 32.6 on 2 and 736 DF,  p-value: 2.69e-14
```

- ▶ the relationship (correlation)
- ▶ net of other variable's influence (controlling for...)
- ▶ the precision (uncertainty)
- ▶ the shared variation (R^2)
- ▶ the remaining variation (residuals, σ^2)

Interpretation

Stages of interpretation

- ▶ **hypothesis testing:** direction and significance
- ▶ **marginal effect:** the relative increase in your predictor w/o accounting for the value of other predictors.
- ▶ **prediction:** fill in the equation for all predictors and calculate the predicted effect
- ▶ **first difference:** fill in the equation for two *scenarios* and calculate the difference in y
- ▶ **effect plot:** fill in the equation for all scenarios relevant to your predictor

⇒ as we move to GLMs, the importance of stages 3-6 becomes important

Hypothesis testing

Hypothesis testing

Hypotheses are mostly about direction and significance

```
summary(mod2)

##
## Call:
## lm(formula = LocalAssistants ~ OpenList + LaborCost, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.49  -1.94  -0.41   1.08  35.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.1266    0.2861  14.42 < 2e-16 ***
## OpenList     0.8288    0.2278   3.64  0.00029 ***
## LaborCost   -0.0702    0.0102  -6.91   1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.1 on 736 degrees of freedom
## Multiple R-squared:  0.0814, Adjusted R-squared:  0.0789
## F-statistic: 32.6 on 2 and 736 DF,  p-value: 2.69e-14
```

- ▶ **direction:** MEPs from candidate-centered systems have on average more local assistants on their payroll
- ▶ **significance:** this is unlikely to be random
 - ⇒ ... *but what is the substantive effect?*

Marginal effect: change in x

Marginal effect: change in x

The relative (marginal) increase in your predictor (difference in means)

- ▶ without accounting for the value of other predictors
 - ▶ important once we move to GLMs

- ▶ regression is the estimation of an equation

$$y = \alpha + \beta x$$

- ▶ marginal effects focus on βx
 - ▶ β : from the model (you estimated it)
 - ▶ x : from the data (you pick it)

Marginal effect: example of change in x

The relative (marginal) increase in your predictor (difference in means) without accounting for the value of other predictors.

```
summary(mod2)
```

```
##  
## Call:  
## lm(formula = LocalAssistants ~ OpenList + LaborCost, data = df)  
##  
## Residuals:  
##     Min      1Q Median      3Q     Max  
## -4.49  -1.94  -0.41   1.08  35.00  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  4.1266    0.2861   14.42 < 2e-16 ***  
## OpenList      0.8288    0.2278    3.64  0.00029 ***  
## LaborCost   -0.0702    0.0102   -6.91  1e-11 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ','  
##  
## Residual standard error: 3.1 on 736 degrees of freedom  
## Multiple R-squared:  0.0814, Adjusted R-squared:  0.0789  
## F-statistic: 32.6 on 2 and 736 DF,  p-value: 2.69e-14
```

► interpretation:

- β : “when x increases with one unit, y increases with β units”
- x : when labor cost increases with *one* unit (x , here 1000 euros), the average number of assistants decreases by 0.07

⇒ *but is this what we want to know?*

Partial scenario: set values for x

Find an increment (change) in x that makes sense for your story

```
##Summary of x
summary(df$LaborCost)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##        4       10     26       23     31       41
```

```
##Find two typical values
summary(df$LaborCost)[c(4,5)]
```

```
##      Mean 3rd Qu.
##        23     31
```

```
## E.g. change from mean to 3rd quartile
summary(df$LaborCost)[c(4,5)] %>% diff
```

```
## 3rd Qu.
##     8.6
```

- ▶ univariate statistics / data exploration helps you find interesting changes in x
- ▶ calculate βx by filling in a realistic *change* in x.
- ▶ 8580 euro increase (increase by 8.58) corresponds to a 0.6 decrease in assistants ($\beta x = -0.07 \times 8.58$).

⇒ use the univariate statistics to find an interesting increments

Prediction: fill in all x's

Prediction: fill in all x's

We estimated an equation with the help of our data

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i$$

data (observed)

- ▶ variables: X and Y
- ▶ observations: i is a counter for the observations, refers to the i^{th} observation. $i \dots N$

parameters (estimated)

- ▶ α intercept, the value of Y when X == 0
- ▶ β slope, the increase in Y when X increases by one unit

We make predictions by filling in data points for that equation

$$Y_i = 4.13 + 0.83 \times OpenList + -0.07 \times LaborCost$$

If all x's were 1:

$$4.89 = 4.13 + 0.83 \times 1 + -0.07 \times 1$$

Why prediction?

- ▶ data description: “out-of-sample”
 - ▶ forecasting: e. g. election
 - ▶ machine learning: e.g create a new variable
- ▶ model statistics: “in-sample”
 - ▶ compare observed and predicted y
- ▶ interpretation:
 - ▶ set scenarios (fill in x)
 - ▶ predict outcomes (using β)

Creating one full scenario

You create a predicted scenario when you fill in values for *all* the predictors (x).

In R:

```
##Create variables
x1 = 1; x2 = 22

# or a data frame
scenario <- data.frame(
  OpenList = 1,
  LaborCost = 22)

# extract coefficients and apply to new data
predict(mod2, newdata = scenario)
```

```
##    1
## 3.4
```

⇒ MEPs from candidate-centered electoral systems with average labor cost, are predicted to have – on average – a local staff of 3.41 people.

By hand:

$$Y_i = \alpha + \beta_1 OpenList + \beta_2 LaborCost$$

$$Y_i = \alpha + \beta_1 \times 1 + \beta_2 \times 22$$

$$3.41 = 4.13 + 0.83 \times 1 + -0.07 \times 22$$

When would you be interested in full scenarios

When we use prediction for interpretation, we are interested in three metrics:

- ▶ two assymmetric scenarios: describe two typical value constellations (Ward and Ahlquist, ch 3)
- ▶ first difference: the difference in y between two predicted scenarios
- ▶ effect plots: the predicted y , as x increases, holding all other x constant.

First difference

First difference

First difference compares the predicted outcomes of two scenarios where one x changes, holding all other predictors constant

- ▶ first difference: difference between the two
- ▶ marginal effect vs first difference:
 - ▶ linear effects: marginal effect with partial scenario is the same as first difference
 - ▶ non-linear effects: the two are different

How to calculate a first difference

You create **two scenarios** and calculate the difference in y between the two

In R:

```
x1 = c(0, 1); x2 = 22

# or data frame
scenario <- data.frame(OpenList = c(0, 1),
                        LaborCost = 22)
#Predict both
predict(mod2, scenario)
```

```
##    1    2
## 2.6 3.4
```

```
#Take the difference
predict(mod2, scenario) %>% diff
```

```
##    2
## 0.83
```

By hand:

$$Y_i = \alpha + \beta_1 OpenList_{1:2} + \beta_2 LaborCost$$

$$\text{scenario 1: } 2.58 = 4.13 + 0.83 \times 0 + -0.07 \times 22$$

$$\text{scenario 1: } 3.41 = 4.13 + 0.83 \times 1 + -0.07 \times 22$$

$$\text{First difference: } 0.83 = 2.58 - 3.41$$

⇒ The first difference can be calculated for any two scenarios of your choice!

Effect plot

Effect plot

Effect plots allow us to visualize our effects

- ▶ choice depends on the measurement level of x
 - ▶ categorical predictor: a tilted coefficient plot
 - ▶ numeric predictor: effectplot

Prediction using many scenarios

You create a bunch of scenarios covering the entire range of x

In R:

```
#Scenario
scenario <- data.frame(OpenList = c(0),
                      LaborCost = min(df$LaborCost): max(df$LaborCost))
```

#Inspect the first three scenarios

```
scenario[1:3,]
```

```
##   OpenList LaborCost
## 1         0      3.8
## 2         0      4.8
## 3         0      5.8
```

#Predict

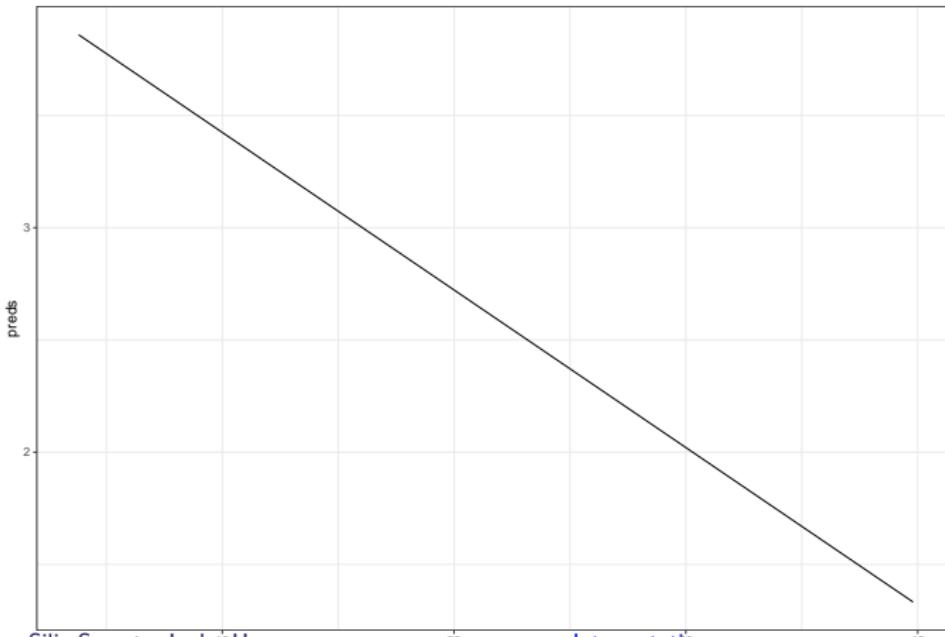
```
scenario <- scenario %>% mutate(preds = predict(mod2, newdata = scenario))
scenario[1:3, ]
```

```
##   OpenList LaborCost preds
## 1         0      3.8  3.9
## 2         0      4.8  3.8
## 3         0      5.8  3.7
```

Visualize

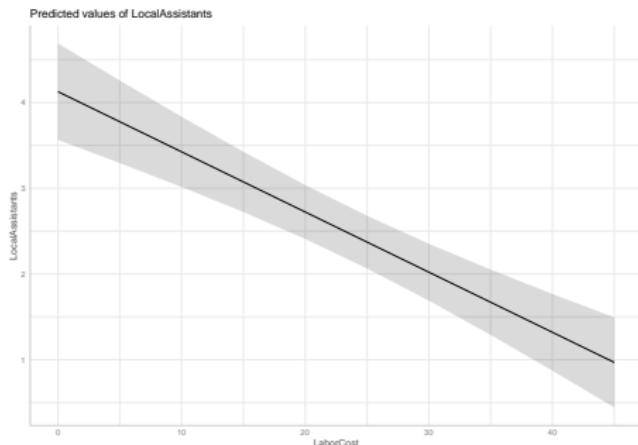
Visualize by plotting your x against your predicted y

```
scenario %>%
  ggplot +
  geom_line(aes(x = LaborCost,
                 y = preds))
```



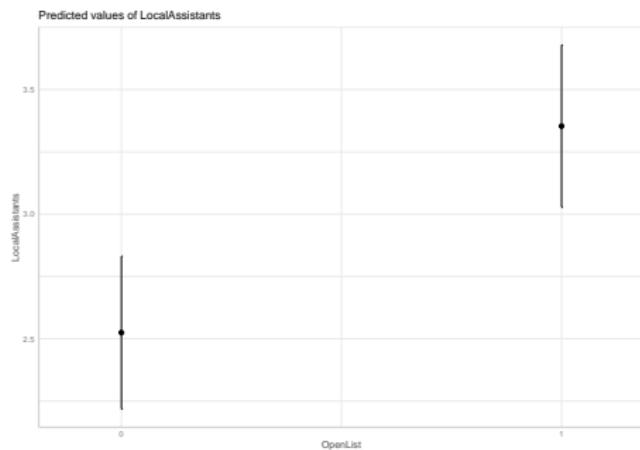
Visualization for metric predictors

Metric predictors are best visualized with an effect plot (a line).



Visualization: categorical predictor

Categorical predictors are best visualized with a titled coefficient plot



Two sources of variation in the data

Two sources of variation in the data

But are these effects statistically significant?

- ▶ **Fundamental uncertainty:** The natural randomness in outcomes, even if the true parameters were known (Captured by residual variance).
- ▶ **Estimation uncertainty:** How precisely are the coefficients estimated? (Captured by the variance-covariance matrix)

⇒ *the uncertainty of your predictions depend on both*

Fundamental uncertainty

Fundamental uncertainty

$$Y_i = \alpha + \beta X1_i + \beta X2_i + \sigma^2$$

data (observed)

- ▶ variables: X and Y
- ▶ observations: i is a counter for the observations, refers to the i^{th} observation. $i \dots N$

parameters (estimated)

- ▶ α intercept, the value of Y when X == 0
- ▶ β slope, the increase in Y when X increases by one unit
- ▶ σ^2 variance in the error term; $\sqrt{\sigma^2}$ = standard deviation

Let's rewrite

Any model in this class

$$Y \sim g(\theta, \sigma^2)$$

$$\theta = \alpha + \beta X_i$$

- ▶ $g()$: probability distribution (the link between parameters and outcomes)
- ▶ θ : central tendency (systematic component)
- ▶ σ^2 : spread around that systematic component

The normal (linear) model

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta X_i$$

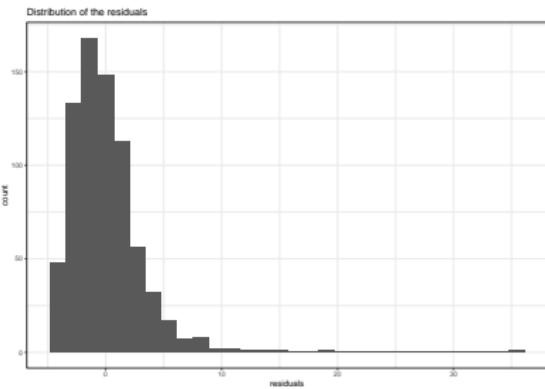
- ▶ $N()$: the normal distribution (identity link; linear; OLS)
- ▶ μ : mean predicted value
- ▶ σ^2 : residual variance (variation of Y around μ)

What are the residuals?

We are always wrong in our predictions, but how wrong are we (in-sample)?

- ▶ residuals: difference between expected and observed outcomes ($Y_i - \mu_i$)
- ▶ a variable: one error / distance per observation in the data

```
df <- df %>% mutate(  
  #Predict in sample  
  pred = predict(mod2, newdata = .),  
  #Difference between expected and observed  
  residuals = LocalAssistants - pred  
)
```



How to describe the residuals?

We describe the residuals by their spread (standard deviation/residual standard error)

```
mean(df$residuals)
```

```
## [1] -9.8e-15
```

- ▶ mean: with an unbiased estimator, their average is 0

```
sd(df$residuals)
```

```
## [1] 3.1
```

- ▶ standard deviation: but their spread can be more or less high
- ▶ here, the average distance from their mean is a staff size of 3.08 local assistants.

⇒ *residual standard error*

Where is it reported?

```
summary(mod2)
```

```
##  
## Call:  
## lm(formula = LocalAssistants ~ OpenList + LaborCost, data = df)  
##  
## Residuals:  
##    Min      1Q  Median      3Q     Max  
## -4.49 -1.94 -0.41  1.08 35.00  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 4.1266    0.2861   14.42 < 2e-16 ***  
## OpenList     0.8288    0.2278    3.64 0.00029 ***  
## LaborCost   -0.0702    0.0102   -6.91 1e-11 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '  
##  
## Residual standard error: 3.1 on 736 degrees of freedom  
## Multiple R-squared:  0.0814, Adjusted R-squared:  0.0789  
## F-statistic: 32.6 on 2 and 736 DF,  p-value: 2.69e-14
```

```
summary(mod2)$sigma
```

```
## [1] 3.1
```

- ▶ we may have predicted an average staff size of 3.86 per MEP in candidate-centered systems

- ▶ ... but individual MEPs typically deviate by 3.08 staffers

- ▶ from the variance, we can calculate the standard deviation, and thus prediction intervals around the mean prediction.

- ▶ Residual standard error: 3.08

Conclusion: fundamental error

- ▶ important for predictions and model statistics
- ▶ not really for the uncertainty of the estimation of our effect

Estimation uncertainty

Estimation uncertainty

- ▶ most research is about the *effect of x on y*
- ▶ so, we're interested in the uncertainty of β

The central limit theorem and sampling

A fiction: the assumptions underpinning the uncertainty of the parameters

- ▶ assumption that data is a sample from a population
- ▶ we *could* sample many times
- ▶ we calculate the same parameter (e.g. mean, differences in means...) in each sample
- ▶ they will vary, but will follow a *normal distribution*

⇒ *each parameter is a distribution with a mean and a standard deviation*

Standard errors

```
summary(mod2)

##
## Call:
## lm(formula = LocalAssistants ~ OpenList + LaborCost, data = df)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
##  -4.49  -1.94  -0.41   1.08  35.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.1266    0.2861   14.42 < 2e-16 ***
## OpenList     0.8288    0.2278    3.64  0.00029 ***
## LaborCost   -0.0702    0.0102   -6.91   1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.1 on 736 degrees of freedom
## Multiple R-squared:  0.0814, Adjusted R-squared:  0.0789
## F-statistic: 32.6 on 2 and 736 DF,  p-value: 2.69e-14
```

- ▶ mean: average of all the differences in means between the two groups of MEPs: 0.95
- ▶ spread: the standard deviation of this distribution is 0.23

\Rightarrow a standard error is the standard deviation of a hypothetical distribution (parameters)

Colinearities

Colinearities

Regression parameters may be correlated

```
mat <-vcov(mod2)
mat

##           (Intercept) OpenList LaborCost
## (Intercept)    0.0819 -0.02846 -0.00244
## OpenList      -0.0285  0.05191  0.00018
## LaborCost     -0.0024  0.00018  0.00010
```

- ▶ reported in the *variance-covariance matrix*
- ▶ diagonal: the variance of the parameter.
 - ▶ variance in effect of electoral system: $\sigma^2 = 0.05$
 - ▶ standard error in effect of electoral system: $\sqrt{\sigma^2} = 0.23$
- ▶ off-diagonal: the covariance of the parameters
 - ▶ low correlation between labor cost and electoral system

Estimate

King et al. (2000) make two points

- ▶ find interesting scenarios when you interpret
- ▶ estimate the uncertainty for the scenarios including
 - ▶ standard error (diagonal)
 - ▶ covariance (off-diagonal)

⇒ *the correlation between variables may mean higher or lower uncertainty than only using the standard error*

Simulation

They do this using simulation

- ▶ set scenario for all predictors
- ▶ draw from the distribution of parameters
- ▶ make prediction
- ▶ repeat many times
- ▶ extract the information and report
 - ▶ mean
 - ▶ median
 - ▶ mode
 - ▶ standard deviation
 - ▶ plot the distribution!

How we do it

We will see two ways of doing this in R

- ▶ `ggeffects` package: simulates scenarios for us and can be plotted seamlessly → *effect plots, coefplots and point predictions*
- ▶ `MASS` package: the “manual” simulation from a multivariate normal distribution using the variance-covariance matrix. → *entire vector of simulations; for other plots/purposes*

Recap I: interpretation

Once your model is estimated, ask what it can tell you.

- ▶ point estimates:
 - ▶ interpret in numbers, text and figures.
 - ▶ use scenarios
- ▶ uncertainty:
 - ▶ fundamental : in residuals
 - ▶ estimation: standard errors of coefficients

Recap II: interpretation

Stages of interpretation	Uses	Coefficient
Hypothesis testing	direction and significance	sign of β ; binary uncertainty
Marginal effects	size of the slope coefficient; partial scenario / increment in x	βx
Prediction	predicted outcomes; full scenario	$\mu = \alpha + \beta x$
First difference	two or more predicted outcomes; full scenario	$\mu = \alpha + \beta x$
Graphical display	predicted outcomes along entire range of x; full scenario	$\mu = \alpha + \beta x$

For Thursday

At home

- ▶ work through the first R-notebook: it implement what I talked about here
- ▶ finish reading: esp Berry et al.

In class

- ▶ we focus on second R-notebook and non-linear effects

Study technique

For this class

- ▶ learn by doing!
 - ▶ all readings include R examples; code along!
 - ▶ my R notebooks
 - ▶ then play around with the concepts; also with your own data/former exams
- ▶ dialogue with AI (ChatGPT, Claude)

What to ask and not to ask chat for?

R codes

- ▶ dont ask for complex codes
 - ▶ requires quirey competence on your end
 - ▶ you don't learn
- ▶ ask it to annotate your scripts
 - ▶ explain what each line means
 - ▶ dissect all code chunks you find and ask

What to ask and not to ask chat for?

Statistics

- ▶ don't ask for a summary of the reading
 - ▶ it's not necessarily what we will focus on
 - ▶ you don't learn
- ▶ ask for definitions
 - ▶ ask it to define key concepts you don't understand while you read
 - ▶ rephrase definitions and ask it this is a good understanding
- ▶ match with your readings
 - ▶ upload the PDF and ask specific questions
 - ▶ ask for examples, possibly with R codes
- ▶ interpretation
 - ▶ copy-paste your model output and ask for an explainer
 - ▶ use descriptive statistics to find interesting scenarios, ask it to help you find a plain English intuitive sentence