CS&SS 321 - Data Science and Statistics for Social Sciences

Module IV - Hypothesis test and multivariate regression

Ramses Llobet

Module IV

- This module introduces and reviews the topic of causation in science.
  - *Statistical Inference*.
  - *Hypothesis test*.
  - *Multivariate regression*.

Statistical inference: estimation

- In statistical inference, we are concerned with making **predictions** (inferences) about a **DGP** or *population* based on information obtained from a *sample*.
- This involves the following key concepts:
  - **Estimand**: The **quantity of interest** from the data-generating process that we aim to estimate or infer.
  - **Estimator**: A statistical **method** or **formula** used to estimate the estimand based on sample data.
  - **Estimate**: it is the calculated value that serves as the **best guess** or approximation of the estimand based on the available information from the sample.

Estimand, estimator, and esitmate

- Statistical inference involves using **estimators** to obtain **estimates** of **estimands** from sample data to make predictions about the population.

- Analogy: have you ever heard about the *ecce homo*?

Estimand, estimator, and esitmate

Estimand, estimator, and esitmate

Estimand, estimator, and esitmate

Estimand, estimator, and esitmate

- **Estimates** are *best guesses*, but they never return you the "*true*".

Populations and samples

Population Parameter:

- A population **parameter** is a numerical value that describes a characteristic of a **population**.
- It is a **fixed and unknown** value that we aim to estimate or infer using statistical methods.

Sample Statistic:

- A sample **statistic** is a numerical value that describes a characteristic of a **sample**.
- It is calculated from the data of a sample and is used to estimate or make **inferences** about population parameters.

Sample statistics

- A **sample mean** that represents a social process:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

- The **sample variance** that we estimate:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2}$$

Sample statistics

	Grade_i	Grade_i - Grade_Mean	(Grade_i - Grade_Mean)^2
Student 1	2.4	-0.76	0.5776
Student 2	2	-1.16	1.3456
Student 3	3.8	0.64	0.4096
Student 4	3.6	0.44	0.1936
Student 5	3.4	0.24	0.0576
Student 6	2.9	-0.26	0.0676
Student 7	3.3	0.14	0.0196
Student 8	3.8	0.64	0.4096
Student 9	3.4	0.24	0.0576
Student 10	3	-0.16	0.0256
n	**Mean**		**Variance**
10	**3.16**		**0.3164**

Populations and samples

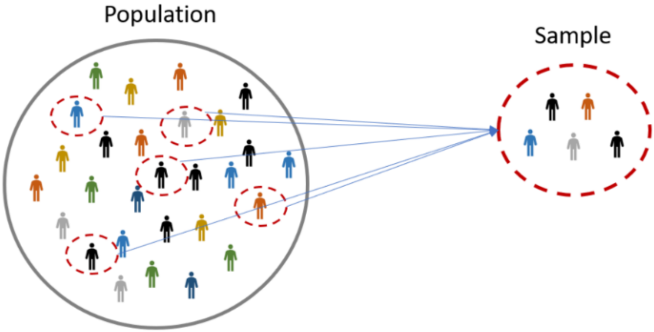
- Typically, we seek to learn features from **populations**, but studying the entire population is unfeasible.
- Thus, we rely on **samples** to make **inferences** under different **assumptions**.

Parameter/Statistic	Population	Sample
Mean	μ	$\bar{X}$
Variance	σ^2	$\hat{\sigma}^2$ or s^2
Standard deviation	σ	$\hat{\sigma}$ or s
Slope/coefficient	β	$\hat{\beta}$ or b

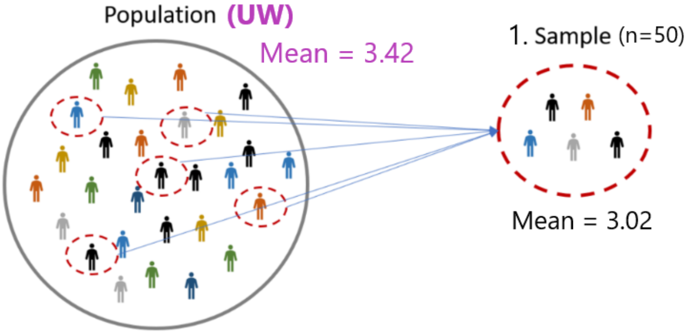
Table 1: Comparison of Population Parameters and Sample Statistics

Populations and samples

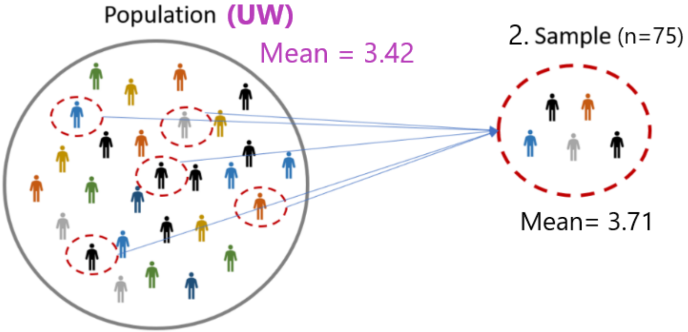
- *Example*: We want to learn the mean GPA of the University of Washington (population) through random sampling students.



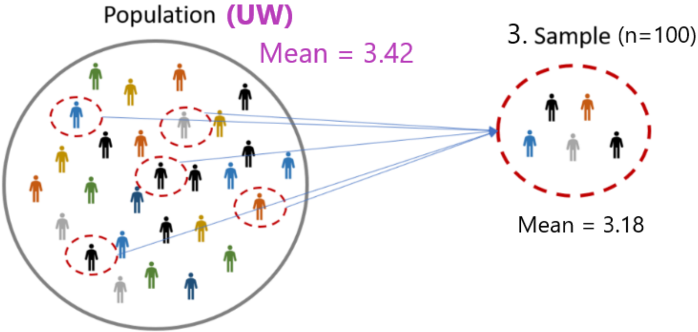
Populations and samples



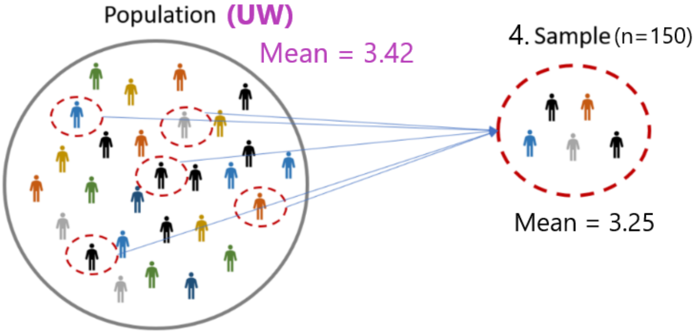
Populations and samples



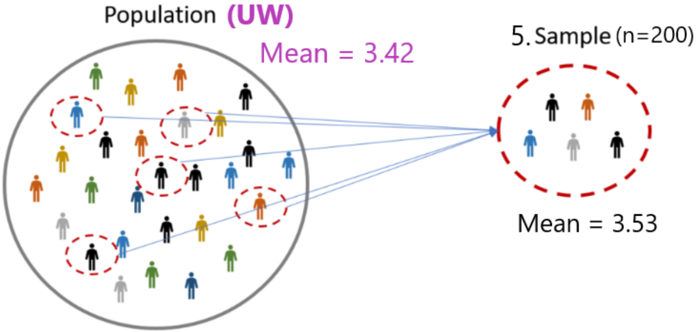
Populations and samples



Populations and samples



Populations and samples



Estimation: Bias

- However, how can we tell if these are good estimates?
  - Ideally, we would compute the estimation error or **bias**.

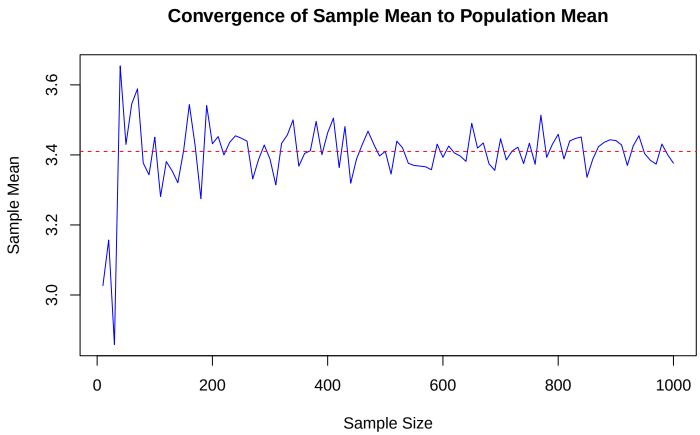
$$\text{bias} = \textit{estimate} - \textit{truth} = \bar{X} - \mu \tag{3}$$

n	**bias**	$\bar{X} - \mu$
50	-0.40	$3.02 - 3.42$
75	0.29	$3.71 - 3.42$
100	-0.24	$3.18 - 3.42$
150	-0.17	$3.25 - 3.42$
200	0.11	$3.53 - 3.42$

Table 2: What is the extent of bias in our estimates?

Estimation: Consistency

- What may happen if we repeat this "experiment" and we increase the sample in each iteration?



Estimation: Bias and Consistency

- **Unbiasedness**: an estimator $\bar{X}$ of a parameter μ is unbiased if and only if:

$$E(\bar{X}) = \mu \tag{4}$$

- **Consistency**: an estimator is consistent if for a sequence $\{X_n\}$ to converge to a limit μ as $n \to \infty$, we have:

$$\lim_{n \to \infty} X_n = \mu \tag{5}$$

However, an unbiased estimator with high variability is impractical because it will return **high prediction error** (MSE) as:

$$MSE = Var + bias^2 \tag{6}$$

Estimation

- Furthermore, they do not provide information about the **uncertainty** or precision of the estimate.
- **Confidence intervals** (CIs) address this issue by providing a range of plausible values for the **estimate**.
  - CIs are based on the principles of probability and sampling variability.
  - Different samples from the **same** population will yield different confidence intervals.

To construct **confidence intervals**, we need to estimate the standard deviation to determine the standard error.

Uncertainty: standard errors.

- The **sample standard deviation** is simply the square root of the variance (see second slide).

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{7}$$

- To characterize the variability of an estimator, we compute the **standard error**:

$$SE(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}} \tag{8}$$

Uncertainty: critical values.

To calculate the margin of error, we need to choose a **critical value**. Critical values influence the interpretation and outcome of the analysis because:

- constructing **confidence intervals**, and
- determining the **significance level** in hypothesis tests.

Significance Level	Critical Value	Confidence Interval
0.1	1.645	$1 - 0.1 = 0.9 \quad (90\%)$
0.05	1.96	$1 - 0.05 = 0.95 \quad (95\%)$
0.01	2.576	$1 - 0.01 = 0.99 \quad (99\%)$

Table 3: Common Critical Values and Confidence Intervals

Uncertainty: margin of error.

Once we have the standard error and select a critical value, the **margin error**, ME, and the **confidence intervals** are estimated as follows:

$$ME = \text{critical value} \times SE(\bar{X}) \tag{9}$$

$$\begin{aligned} \text{Confidence Interval} &= \left(\bar{X} - ME, \bar{X} + ME\right) \\ &= (CI_{lower}, CI_{upper}) \end{aligned} \tag{10}$$

Uncertainty: example

```r
dat <- read_csv("data/students.csv")
names(dat)
```

```
## [1] "GPA"    "gaming" "study"  "quiz"
```

```r
# Randomly sample 40 observations
sampled_data <- sample(dat$GPA, size = 40, replace = F)

(GPA_mean <- mean(sampled_data) ) # sample mean
```

```
## [1] 3.132462
```

```r
(GPA_sd <- sd(sampled_data)) # sample standard deviation
```

```
## [1] 1.348265
```

```r
(GPA_se <- GPA_sd / sqrt( length(sampled_data) ) ) # sample standard errors
```

```
## [1] 0.2131794
```

Uncertainty: example

Question: Is the sample mean biased estimator? Is the population mean within the confidence interval of our estimator?

```
statistics <- tibble(
  mean = GPA_mean,
  CI_lwr = GPA_mean - (1.96 * GPA_se),
  CI_upr = GPA_mean + (1.96 * GPA_se)
)

mean(dat$GPA) # population mean of GPA
```

```
## [1] 3.203627
```

```
statistics
```

```
## # A tibble: 1 x 3
##    mean CI_lwr CI_upr
##   <dbl>  <dbl>  <dbl>
## 1  3.13   2.71   3.55
```

Uncertainty: interpreting confidence intervals

- We rely on samples for making **inferences**. To determine if our estimations approach the true population parameter, we use **confidence intervals.**
- A **confidence interval** is a range of plausible estimates.
- The **confidence level**, denoted as $(1 - \alpha)$ or simply $1 -$ significance level, is **interpreted** as the probability that the confidence interval **will contain** the true population parameter **over hypothetical replications**.

Uncertainty: interpreting confidence intervals

- *Example*: a **95%** confidence interval implies that if we were to **hypothetically repeat** the estimation and construct confidence intervals for each sample/estimate, approximately 95% of those intervals **would** contain the *true* parameter.

Imai (2018, p. 328) - critical values

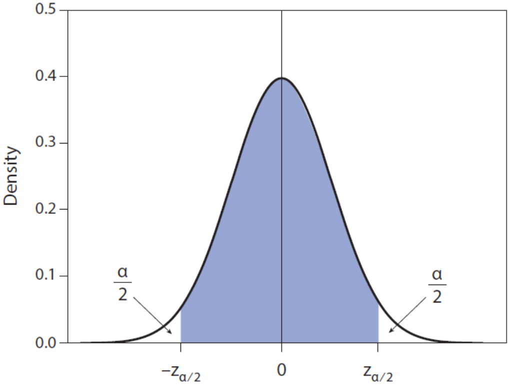
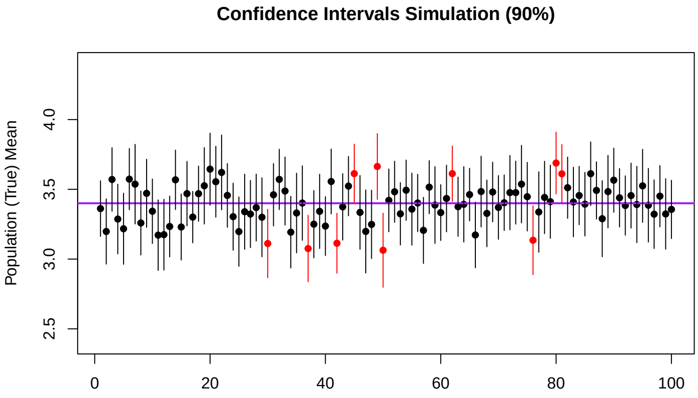


Figure 7.1. Critical Values Based on the Standard Normal Distribution. The lower and upper critical values, $-z_{\alpha/2}$ and $z_{\alpha/2}$, are shown on the horizontal axis. The area under the density curve between these critical values (highlighted in blue) equals $1 - \alpha$. These critical values are symmetric.

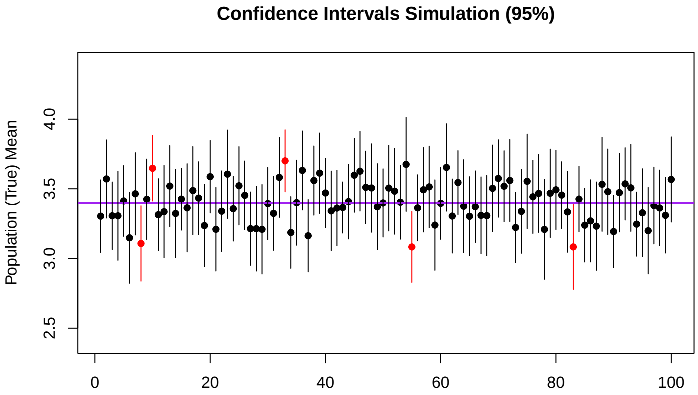
Uncertainty: interpreting confidence intervals

- Resampling and estimating the GPA of the same population (UW) over 100 iterations with a significance **level of 0.1** (90% confidence intervals).



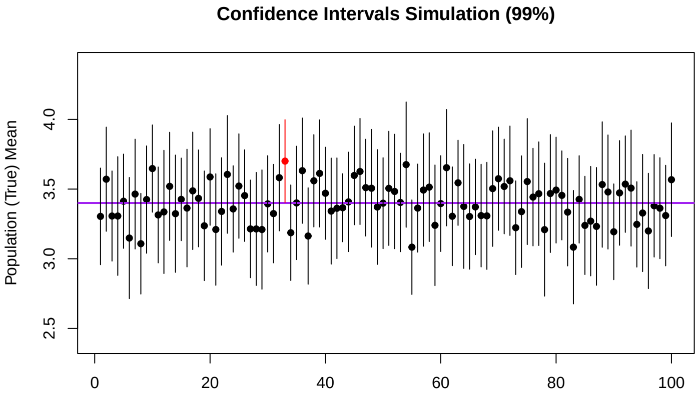
Uncertainty: interpreting confidence intervals

- Resampling and estimating the GPA of the same population (UW) over 100 iterations with a significance **level of 0.05** (95% confidence intervals).



Uncertainty: interpreting confidence intervals

- Resampling and estimating the GPA of the same population (UW) over 100 iterations with a significance **level of 0.01** (99% confidence intervals).



Takeaways

- Understand **bias** and **consistency**.
- Estimates must always inform of **uncertainty**.
- The impact of the **critical value** (α) on constructing confidence intervals.
- Wider confidence intervals increase the likelihood of the *"true value"* being within the intervals over **hypothetical replications**.
  - **Question**: Why might someone want to calculate narrower confidence intervals?

Time to code a little bit!

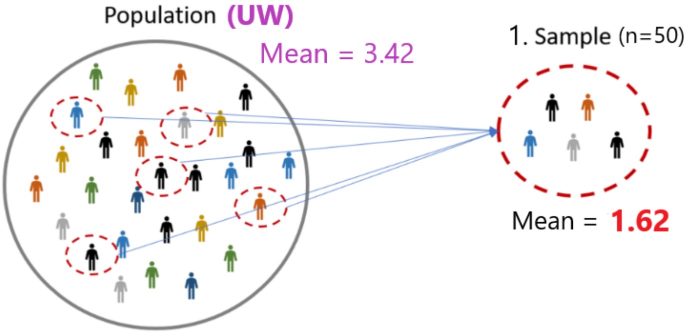
- Open the file `Confint.rmd`

Hypothesis Testing: motivation

- We have drawn a distinction between a population and a sample. However, how do we know that the sample reflects the population of interest?
- Due to inherent **variability** in the data, the sample may **not perfectly reflect** the entire population.
- Through a **t-test**, we assess whether the observed difference between the sample mean and the **hypothesized value** exceeds what is expected due to chance (aka random sampling variability alone).

Hypothesis Testing: motivation

- What if the sample mean is really off from the population mean?



Hypothesis Testing

- Hypothesis testing is used to make inferences about population **parameters** based on **sample** data.
- It involves formulating **null** and **alternative hypotheses** and evaluating the evidence against the null hypothesis.
  - **Null Hypothesis** (H_0): a statement of no effect or no difference between groups or variables (*proof by contradiction*).
  - **Alternative Hypothesis** (H_a): contradicts the null hypothesis and suggests the presence of an effect or a difference between groups or variables.
- **Goal**: Does the *evidence* from the sample supports the **null** hypothesis or provides evidence for the **alternative** hypothesis?

Hypothesis Testing

- **T-test**: quantifies the difference between the **sample** statistic and the hypothesized **population** parameter relative to the variability within the data.
  - It takes into account the **sample size** (N) and the **standard error** (SE) of the statistic to assess the likelihood of observing such a difference by chance.
- **Significance Level** (α): The predetermined **threshold** for rejecting the null hypothesis.

Hypothesis Testing: p-values

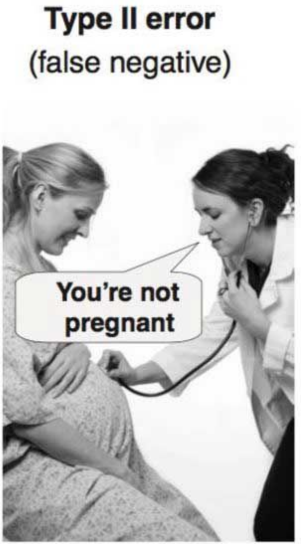
- **P-value**: it measures the **strength of evidence** against the null hypothesis, we compare it with the significance level to determine if we **reject or fail to reject** the null (H_0).
  - p-value is **large**: suggest insufficient evidence to reject the null hypothesis.
  - p-value is **low**: stronger evidence against the null, favoring the alternative(H_a).

Hypothesis Testing: error types

- There is a clear trade-off between **Type I** and **Type II** errors in that minimizing type I error usually increases the risk of type II error.

Decision	H_0 is True	H_0 is False
Retain H_0	Correct	**Type II Error**
Reject H_0	**Type I Error**	Correct

Hypothesis Testing: error types

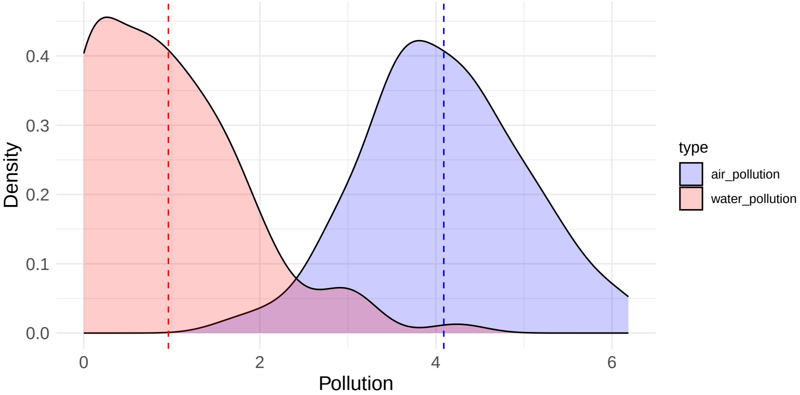


Hypothesis Testing Process

1. State the **null** and **alternative** hypotheses.
2. Choose a test statistic and the **significance level** (α).
3. Estimate the test statistic, in our case the **t-value**.
4. Compute the **p-value**, and compare it with with the significance level.
   - For example, is p-value $< \alpha$?
5. Reject the null hypothesis if the p-value is less than or equal to α.

Hypothesis Testing Process

- We will focus on a scenario where we want to assess the **association** of air and water **pollution** on **climate change**.
  - *Disclaimer*: this data was simulated.



Hypothesis Testing Process

- We define a theoretical model:

$$cc = \alpha + \beta_1 \text{air} + \beta_2 \text{water} + \epsilon \tag{11}$$

1. State the null and alternative hypotheses:
   - **Null Hypothesis** (H_0): air (β_1) or water (β_2) pollution are **not** associated with climate change. In other words, $\beta_1 = 0$ or $\beta_2 = 0$.
   - **Alternative Hypothesis** (H_a): air or water are associated with climate change. In other words, $\beta_1 \neq 0$ or $\beta_2 \neq 0$
2. Set the **significance level**, the default in social sciences is 0.05.

Hypothesis Testing Process

- The `lm()` function estimates the t-statistic and p-values (steps 3 and 4) using the fitted model and sample data argument.

```
model <- lm(climate_change ~ air_pollution + water_pollution)
round(coef(model), digits=2)
```

```
##     (Intercept)   air_pollution water_pollution
##            0.65            1.87            0.18
```

- Estimated model, are the coefficients statistically significant?

$$cc = 0.65 + 1.87\text{air} + 0.18\text{water} \tag{12}$$

Model summary

- Use the function `summary()` for the t-test and the p-value.
- Can we reject H_0?
  - Remember that the **significant level** that we choose was 0.05 *(critical value = 1.96)*.

```r
summary(model)
```

```
##
## Call:
## lm(formula = climate_change ~ air_pollution + water_pollution)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8735 -0.6615 -0.1320  0.6208  2.0701
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.6491     0.4555   1.425   0.1574
## air_pollution     1.8663     0.1048  17.802   <2e-16 ***
## water_pollution   0.1840     0.1093   1.683   0.0956 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9514 on 97 degrees of freedom
## Multiple R-squared:  0.7662, Adjusted R-squared:  0.7614
## F-statistic: 158.9 on 2 and 97 DF,  p-value: < 2.2e-16
```

Hypothesis Testing and Confidence Intervals

- Can we reject H_0?

```
(p_value <- summary(model)$coefficients[, "Pr(>|t|)"])
```

```
##     (Intercept)   air_pollution water_pollution
##    1.573807e-01    2.525692e-32    9.558931e-02
```

```
(t_value <- summary(model)$coefficients[, "t value"])
```

```
##     (Intercept)   air_pollution water_pollution
##        1.424952       17.802076        1.683010
```

```
p_value < 0.05 # is p-value < significant level?
```

```
##     (Intercept)   air_pollution water_pollution
##           FALSE            TRUE           FALSE
```

```
t_value > 1.96 # is t-value > critical value?
```

```
##     (Intercept)   air_pollution water_pollution
##           FALSE            TRUE           FALSE
```

Hypothesis Testing and Confidence Intervals

- Can we reject H_0?
  - H_0 air pollution: sufficient evidence to reject the null hypothesis.
  - H_0 water pollution: insufficient evidence to reject the null hypothesis.
- **Conclusion**: air pollution has a positive significant **association** with climate change. However, water pollution is **not statistically significant**.
  - When an estimated coefficient is not statistically significant, we mean that it is not **significantly different from 0**. In this case, $\beta_2 = 0 \neq 0.18$, because we fail to reject the null H_0 for water pollution.
- However...

Hypothesis Testing and Confidence Intervals

- Can we really reject H_0 if we instead use a significant level of **0.10**?

```
p_value < 0.1 # is p-value < significant level?
```

```
##     (Intercept)   air_pollution water_pollution
##           FALSE            TRUE            TRUE
```

```
t_value > 1.645 # is t-value > critical value?
```

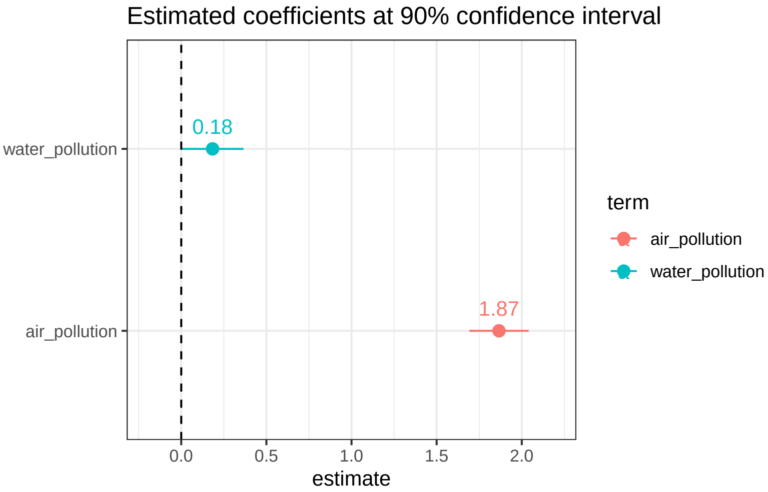
```
##     (Intercept)   air_pollution water_pollution
##           FALSE            TRUE            TRUE
```

- Type I and II error trade-off.

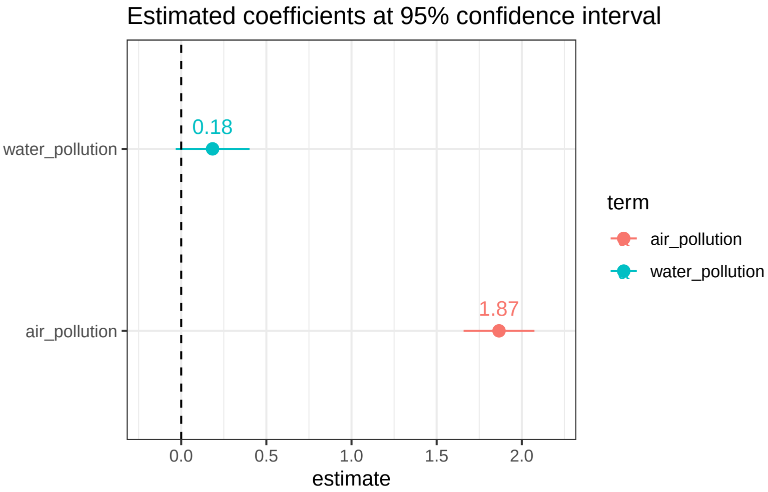
Hypothesis Testing and Confidence Intervals

- **Confidence intervals** and **hypothesis testing** are closely related.
- If the confidence interval **contains the null** value, $\beta_2 = 0$, the null hypothesis cannot be rejected.
- The p-value in hypothesis testing **quantifies** the strength of evidence against the null hypothesis, similar to how confidence intervals provide a range of **plausible** parameter values.
  - **Important**: the p-value is **NOT** the probability that the null is true.

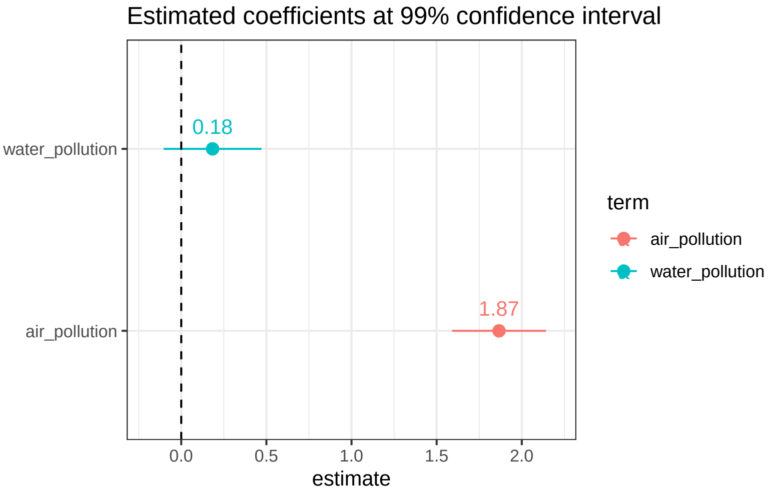
Hypothesis Testing and Confidence Intervals



Hypothesis Testing and Confidence Intervals



Hypothesis Testing and Confidence Intervals



Preview of Problem Set 4

- Problem of missingness
- Merging datasets
- Log transformations

Problem of missingness

- Default option: Listwise deletion
  - The whole observation (row) is deleted if **any** variable is missing
  - Even just 1 variable!
  - Can introduce bias

Investigate missingness

- Many functions/packages allow to check missingness

```
gapminder <- read_csv("data/gapminder2.csv")

# check missingness
questionr::freq.na(gapminder)
```

```
##             missing  %
## lifeExp         341 20
## gdpPercap       170 10
## cntry             0  0
## continent         0  0
## year              0  0
## pop               0  0
```

Investigate missingness

```
## For a single variable
mean(is.na(gapminder$gdpPercap))
```

```
## [1] 0.09976526
```

```
## For multiple variables
gapminder %>% summarize_all(~ mean(is.na(.)))
```

```
## # A tibble: 1 x 6
##   cntry continent  year lifeExp   pop gdpPercap
##   <dbl>     <dbl> <dbl>   <dbl> <dbl>     <dbl>
## 1     0         0     0   0.200     0    0.0998
```

```
## A dummy if any variable is NA
gapminder$missing_dummy <- ifelse(apply(gapminder, 1, anyNA), 1, 0)
mean(gapminder$missing_dummy)
```

```
## [1] 0.2764085
```

Investigate missingness

```
# with dplyr
gapminder %>%
  group_by(year) %>%
  summarize(missing=mean(missing_dummy,na.rm=T)) %>% print(n=8)
```

```
## # A tibble: 12 x 2
##    year missing
##   <dbl>   <dbl>
## 1  1952   0.303
## 2  1957   0.324
## 3  1962   0.261
## 4  1967   0.254
## 5  1972   0.239
## 6  1977   0.261
## 7  1982   0.275
## 8  1987   0.282
## # i 4 more rows
```

```
# with base R
tapply(gapminder$missing_dummy,gapminder$year,mean,na.rm=T)
```

```
##      1952      1957      1962      1967      1972      1977      1982      1
```

Handling missingness

```r
# Listwise deletion
gapminder_noNA <- na.omit(gapminder)
nrow(gapminder) - nrow(gapminder_noNA)
```

```
## [1] 471
```

```r
# Be more careful and selective when you drop NA
gapminder_noNA <- drop_na(gapminder, gdpPercap)
nrow(gapminder) - nrow(gapminder_noNA)
```

```
## [1] 170
```

Log transformations

There are several reasons why someone might choose to **transform** a variable using a **logarithm function** before fitting a model:

1. **Non-linear relationships**: taking the logarithm of a variable can help to linearize the relationship
2. **Interpretability**: when dealing with exponential growth or decay, taking the logarithm can convert it to a linear relationship.
3. **Multiplicative relationships**: by transforming the variables using logarithms, these relationships can be simplified to additive relationships

The interpretation of a log transformation varies **depending on the transformed variables**: dependent, independent, or both.

Log transformations: dependent/response variable

$$log(Y_i) = \alpha + \beta * X_i + \epsilon_i$$

- **Exponentiate** the coefficient (β) of X.
  - This gives the multiplicative factor for every one-unit increase in the independent variable.

Log transformations: dependent/response variable

$$log(Y_i) = \alpha + 0.198 * X_i + \epsilon_i$$

- *Example*: the coefficient (β) is 0.198. $exp(0.198) = 1.218962$.
- **Interpretation**: for every one-unit increase in the independent variable, our dependent variable increases by a factor of about 1.22, or 22%.
  - When ($\beta > 1$): multiplying a number by 1.22 is the same as **increasing** the number by 22%.
  - When ($\beta < 1$): multiplying a number by, say 0.84, is the same as **decreasing** the number by $1 - 0.84 = 0.16$, or 16%.

Log transformations: independent/predictor variable

$$Y_i = \alpha + \beta * log(X_i) + \epsilon_i$$

- **Divide the coefficient by 100**.
  - This tells us that a 1% increase in the independent variable **increases** (or decreases) the dependent variable by (coefficient/100) units.

Log transformations: independent/predictor variable

$$Y_i = \alpha + 0.198 * log(X_i) + \epsilon_i$$

- **Example**: the coefficient (β) is 0.198. $0.198/100 = 0.00198$.
  - **Interpretation**: For every 1% increase in the independent variable, our dependent variable increases by about 0.002.
- **Interpreting X**: For x percent increase, multiply the coefficient by log(1.x).
  - **Example**: For every 10% increase in the independent variable, our dependent variable increases by about $0.198 * log(1.10) = 0.02$.

Log transformations: dependent/response variable

$$log(Y_i) = \alpha + \beta * log(X_i) + \epsilon_i$$

- Interpret the coefficient (β) as the **percent increase** in the dependent variable for every **1% increase** in the independent variable.

Log transformations: dependent/response variable

$$log(Y_i) = \alpha + 0.198 * log(X_i) + \epsilon_i$$

- *Example*: the coefficient (β) is 0.198. For **every 1% increase** in the independent variable (X), our dependent variable (Y) **increases** by about 0.20%.
- **Interpreting X**: for x percent increase in X, calculate 1.x to the **power of the coefficient**, subtract 1, and multiply by 100.
  - **Example**: For every 20% increase in the independent variable, our dependent variable increases by about:
  - $(1.20^{0.198}$ - $1) * 100 = 3.7$ percent.

Fixed effects

Let's look at some data that has units (N) and for each unit several time periods (T).

```
data %>% head(12)
```

```
##     id time           y           x
## 1    1    1  0.65674831  0.04367577
## 2    1    2  0.33917190  0.56187415
## 3    1    3 -0.64075850  0.35627259
## 4    1    4 -1.44009620 -1.00115829
## 5    1    5  0.22082243  1.40279577
## 6    1    6 -0.40975236  0.41626347
## 7    1    7 -0.37248074 -1.14088682
## 8    1    8 -2.12337475  0.46023882
## 9    2    1  0.05988054  0.99210062
## 10   2    2 -1.93975204  0.50146125
## 11   2    3 -1.13816678 -0.23413534
## 12   2    4 -0.75639556  0.02076991
```

Fixed effects

```
data %>% group_by(id) %>% summarize(x=mean(x))
```

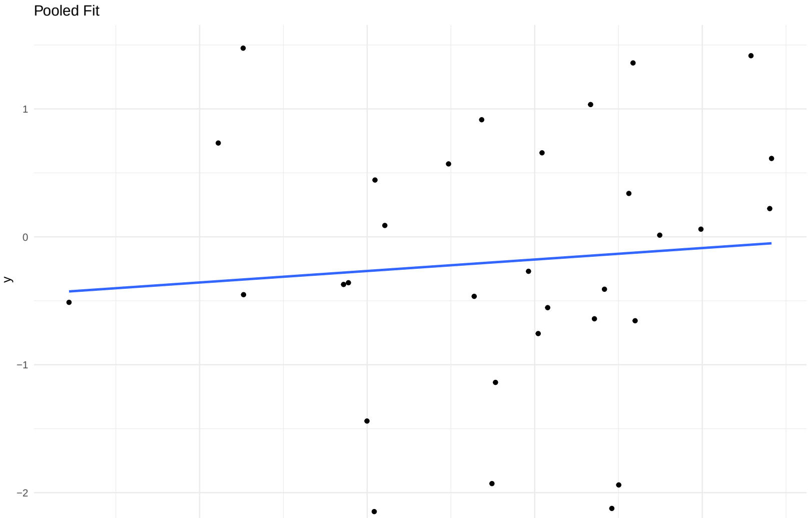
```
## # A tibble: 4 x 2
##   id          x
##   <fct>    <dbl>
## 1 1      0.137
## 2 2     -0.00282
## 3 3     -0.358
## 4 4     -0.541
```

```
data %>% group_by(time) %>% summarize(x=mean(x))
```

```
## # A tibble: 8 x 2
##   time        x
##   <fct>   <dbl>
## 1 1     -0.277
## 2 2     -0.101
## 3 3      0.0753
## 4 4     -0.305
## 5 5     -0.240
## 6 6      0.310
## 7 7      0.245
```

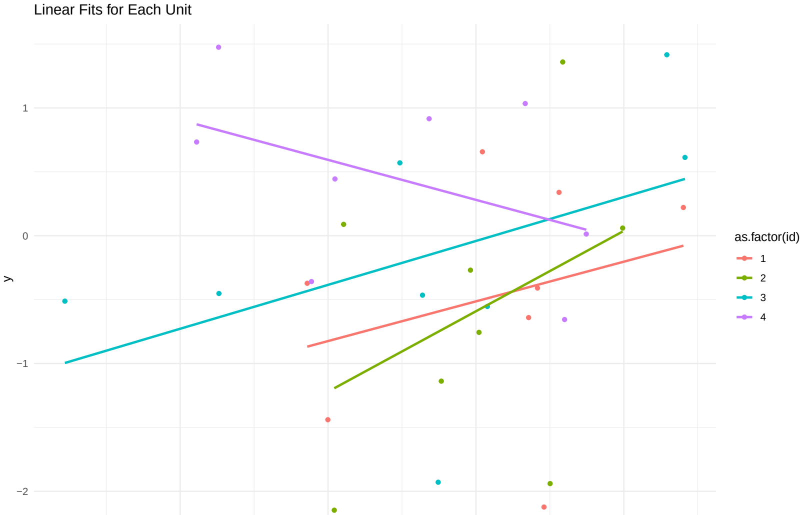
Fixed effects

Now, let's fit a slope between Y and X.



Fixed effects

What happens if instead we fit a slope for each **unit** (N)?



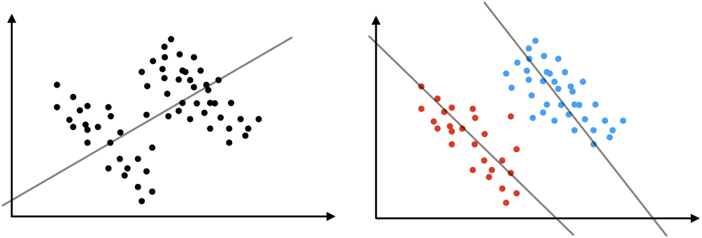
Fixed effects

Motivation: does a general relationship holds at the unit level?

- **Fixed effects** estimation is employed to investigate whether a general relationship holds at the **unit level**.
- By fitting a slope **within** each unit instead of pooling all the data, we can identify distinct patterns, which may sometimes be conflicting.
  - This phenomenon is referred to as the Simpson paradox.

Fixed effects

- Simpson's paradox



Fixed effects

- Account for unobserved time-invariant confounders
- Let's say, the relationship between democracy and economic development
  - Some country-specific but time-invariant characteristics can be confounders
  - For example, culture or legal institutions rarely change during short periods of time (T).
- Pooled regression model:

$$Y_{it} = \alpha + \beta X_{it} + \epsilon_{it}$$

- Fixed effects regression model:

$$Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}$$

for each $i = 1, 2, ..., N$, where α_i is a fixed but unknown intercept

Fixed effects in R

```
worldbank <- read.csv("data/world_bank.csv")
lm1_res <- lm(inf_mort ~ gdp_per_capita + factor(country_code), worldbank)
summary(lm1_res)
```

```
##
## Call:
## lm(formula = inf_mort ~ gdp_per_capita + factor(country_code),
##     data = worldbank)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.140  -11.482   -1.314    9.447  161.140
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              9.705e+01  8.790e+00  11.040  < 2e-16 ***
## gdp_per_capita          -6.795e-04  7.977e-05  -8.518  < 2e-16 ***
## factor(country_code)AGO  8.633e+01  1.059e+01   8.151 4.25e-16 ***
## factor(country_code)ALB -6.163e+01  1.036e+01  -5.948 2.84e-09 ***
## factor(country_code)AND -6.460e+01  1.103e+01  -5.855 4.99e-09 ***
## factor(country_code)ARE -3.309e+01  1.145e+01  -2.889 0.003875 **
## factor(country_code)ARG -5.984e+01  1.006e+01  -5.950 2.81e-09 ***
```

Fixed effects in R

```r
library(stargazer)
stargazer(lm1_res,
          omit = "country_code",
          add.lines = list(c("Year fixed effects", "Yes")),
          header = FALSE)
```

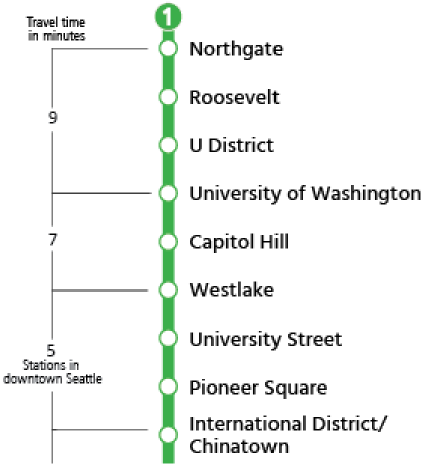
Table 4:

	Dependent variable:
	inf_mort
gdp_per_capita	-0.001^{***}
	(0.0001)
Constant	97.049^{***}
	(8.790)
Year fixed effects	Yes
Observations	7,176
R^2	0.779
Adjusted R^2	0.773

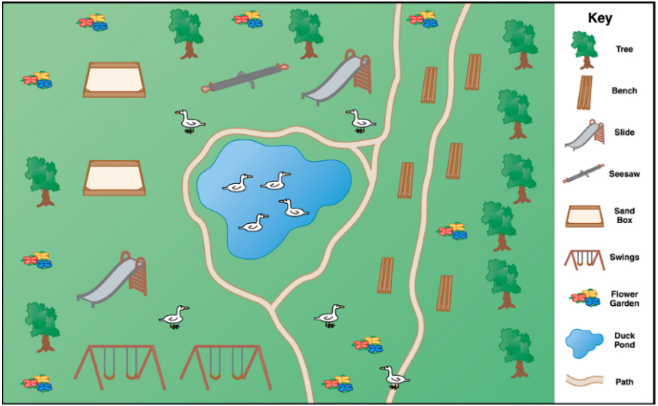
Lab Coding Demonstration:

- Open today's lab markdown file to view the coding demonstration.

Tools for learning

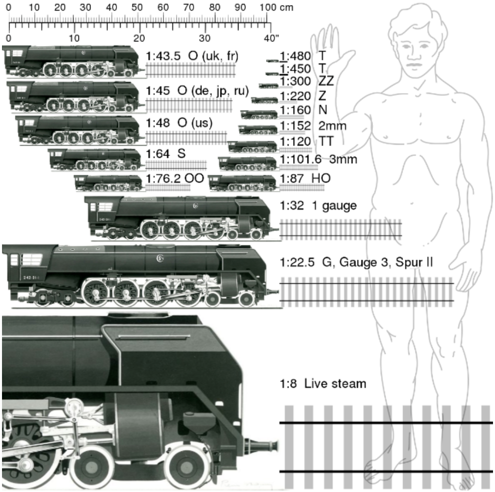


Tools for learning



Tools for learning

Tools for learning: trade-offs



Tools for learning

- **Question**: what do these pictures have in common?

Tools for learning: models

- These pictures depict various **models**; they are tools for **learning.**

"All models are wrong but some are useful"

George Box, 1976

- Models are
  - **simplified representations** of real-world systems.
  - facilitate **clear and concise** communication of complex ideas.
  - help to understand **complex systems** by highlighting significant variables.

Good luck on your journey!

- Please, be aware that you have only scratched the surface of a vast array of methodologies for scientific inference, including
  - Time series and panel data.
  - Multilevel models.
  - Bayesian inference.
  - machine learning for prediction and discovery.
  - Deep learning.
  - Applied causal inference, among others ...

Good luck on your journey!

- Remember the importance of consistency, bias, and efficiency in statistical inference.
  - **Consistency** is the most important property, before bias and efficiency!

Good luck on your journey!

- I hope you gained valuable insights from my labs.
- **Best of luck** with your future endeavors, and please...
  - take a moment to fill out the **course evaluations** if you haven't done so already!

Best wishes,

Ramses