Advanced Quantitative Methods

Course for the master programme in peace and conflict studies,
Uppsala University March–May 2019

Håvard Hegre and David Randahl

March 24, 2019

Introduction to the course

UPPSALA UNIVERSITET

Contents I

UPPSALA UNIVERSITET

Contents II

UPPSALA UNIVERSITET

The aim is that you at the end of the course will know how to

- understand and apply principles of statistical inference
- apply and interpret various linear and logistic regression models
- specify complex Monte Carlo simulation models and use to evaluate specification problems
- use basic programming and data-management techniques
- use the R statistical software package
- use techniques for simulating predictions, first differences, and other quantities of interests based on estimated models
- carry out simple cross-validation of predictive models
- independently write assignments within a given time frame

UPPSALA UNIVERSITET

Basic information

- All course communication will be sent out via the Student Portal, i.e. to the mail address that you have registered there
- Presence in all sessions is mandatory. Please contact me (Håvard) if you cannot attend a session
- For all lectures and seminars: bring your laptop with `R` installed

UPPSALA UNIVERSITET

Literature

- Main books: Gelman and Hill (2007) and Carsey and Harden (2014)
- Programming in R: Wickham and Grolemund (2017)
- Introductions to `R`: Torfs and Brauer (2014); Venables, Smith and the core R team (2016)
- Random forests: Hartshorn (2016)
- Classification: Witten, Frank and Hall (2011, pp. 159–177)
- Applications (see updated course guide)

UPPSALA UNIVERSITET

Examination

- Examination and final course grade (G/U) are based on four assignments and participation in class
- Assignments should be uploaded in the Student Portal:
  - Assignment 1: 5 April 17:00
  - Assignment 2: 12 April 17:00
  - Assignment 3: 23 April 08:00
  - Assignment 4: 3 May 17:00
- Dates also in student portal
- Deadlines for assignments are *hard*

UPPSALA UNIVERSITET

Assignments

Short presentations of experiments. Think of your non-AQM fellow students as readership. The assignments should contain the following:

1. A short motivation of the experiment (max 1 page)
2. Adaptations of `R` scripts discussed in the lectures and the textbook, attached as functioning and debugged scripts.
3. A short text (max 1 page) explaining the procedure employed in the assignment
4. A short text (max 1 page) presenting the results and your conclusions

We will comment on the assignments throughout the course

- For detailed information please consult the course guide

UPPSALA UNIVERSITET

Course evaluation

We would like to hear what you think about the course and how we could improve it

- Will be made available in the final weeks of the course
- Your input on the course will be much appreciated
- I will use the input to improve the course

Repetition of OLS

Gelman and Hill (2007, chs. 3, 4)
Carsey and Harden (2014, chs. 1, 5)

UPPSALA UNIVERSITET

The linear regression model (OLS)

- The **population** regression model:
  - $Y_i = \alpha + \beta X_i + u_i$
- A linear function
- Y is the dependent variable
- The linear function of independent variables X is called the systematic component
- Subscripts i represent what is unique to each observation

UPPSALA UNIVERSITET

The stochastic component u_i

- Remember that statistical inference is probabilistic
- u_i represents uncertainty: the 'stochastic'/'random' component
- Think of u_i as the sum of all unmodeled factors that might affect Y as well as any part of Y that is fundamentally stochastic
- The stochastic component is also called the residual, the disturbance term, or the error term (especially when estimated)
- In OLS we assume the stochastic component has a normal distribution with mean 0 and variance σ^2:

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

UPPSALA UNIVERSITET

Gelman and Hill (2007, p. 37) notation

Gelman and Hill (2007) use a different notation for the same:

- The **population** regression model:
  - $y_i = X_i\beta + \epsilon_i$
- where the errors/residuals ϵ_i have independent normal distributions with mean 0 and standard deviation σ
- $X_i\beta$ is here shorthand for several X_i with corresponding β
- The intercept is just another β (think of the corresponding X_i as a variable where all units have value 1 – a constant)
- Alternative formulation of the same:

$$y_i \sim N(X_i\beta, \sigma^2)$$

UPPSALA UNIVERSITET

Alternative formulation of the linear model I

Gelman and Hill's alternative formulation corresponds to the one used in Carsey and Harden (2014, section 6.2)

- The linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \epsilon \sim N(0, \sigma)$$

can also be written as Y following a normal distribution:

$$Y \sim N(\mu, \sigma), \mu = \beta_1 X_1 + \beta_2 X_2$$

- In the R script, this means that the formulation
  `Y <- b0 + b1*X1 + b2*X2 + rnorm(n, 0, 1)`
  is replaced with
  `Y <- rnorm(n, b0 + b1*X1 + b2*X2, 1)`

UPPSALA UNIVERSITET

The linear regression model (OLS)

The **population** regression model:

- $Y_i = \alpha + \beta X_i + u_i$
- The *Data-Generating Process*, an abstract model of what is out there in the real world

The **sample** regression model:

- $Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$
- The statistical model we estimate on the sample of data we have

Hats ($\hat{}$):

- Used to denote our *estimates* of the *true population parameters*
- No hats for Y_i and X_i – they are observations, not parameters, and a sample from the population

UPPSALA UNIVERSITET

Estimating the OLS model

- When we estimate OLS, we obtain

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + u_i$$

- $\hat{\alpha}$ is the intercept: the average value of Y when X is 0
- $\hat{\beta}$ is the slope: the average change in Y if X increases by 1 unit
- The residuals ($\hat{u_i}$) are the difference between the actual value for Y_i and the fitted value $\hat{Y_i}$:

$$\hat{u_i} = Y_i - \hat{Y_i}$$

UPPSALA UNIVERSITET

Estimating the OLS model

- Estimated by a method that minimizes the squared residuals, hence 'ordinary least squares'
- The residual sum of squares:

$$RSS = \sum_{i=1}^{n} \hat{u_i}^2$$

- OLS finds parameters that makes this sum as small as possible

UPPSALA UNIVERSITET

Estimating the OLS model

- OLS has a mathematical solution – it yields the following estimates for a two-variable model:

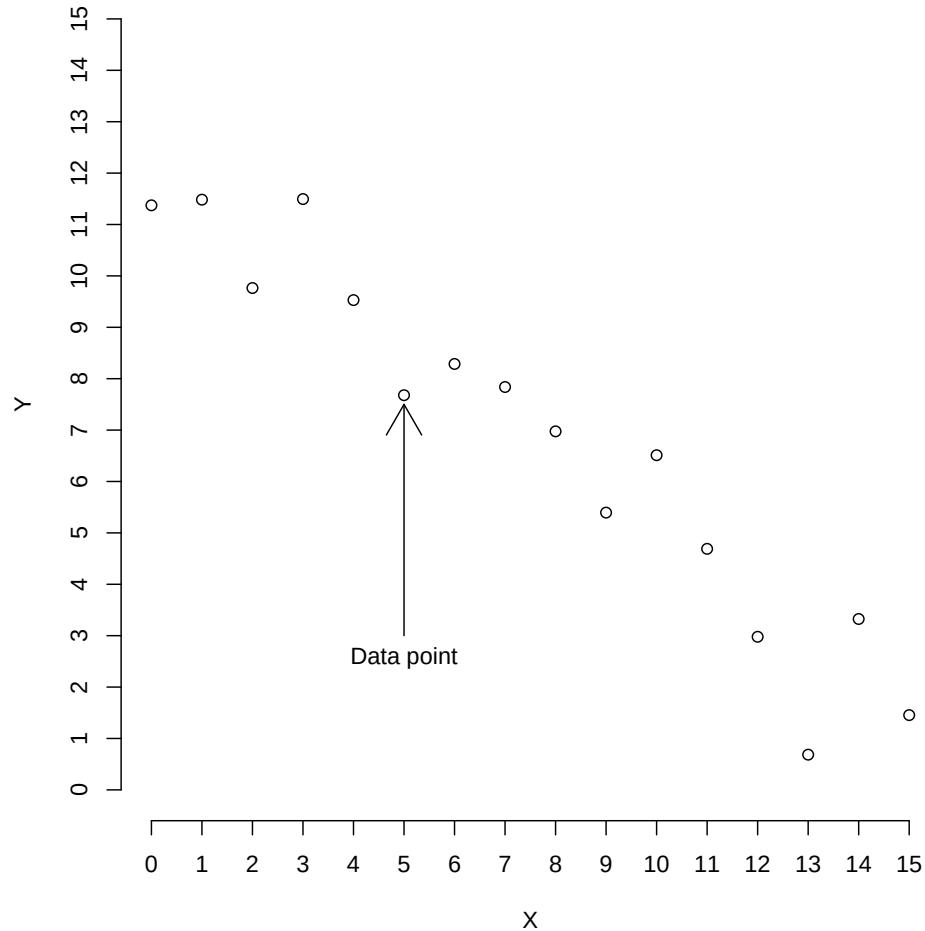
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \beta\bar{X}$$

UPPSALA UNIVERSITET

A Regression Model

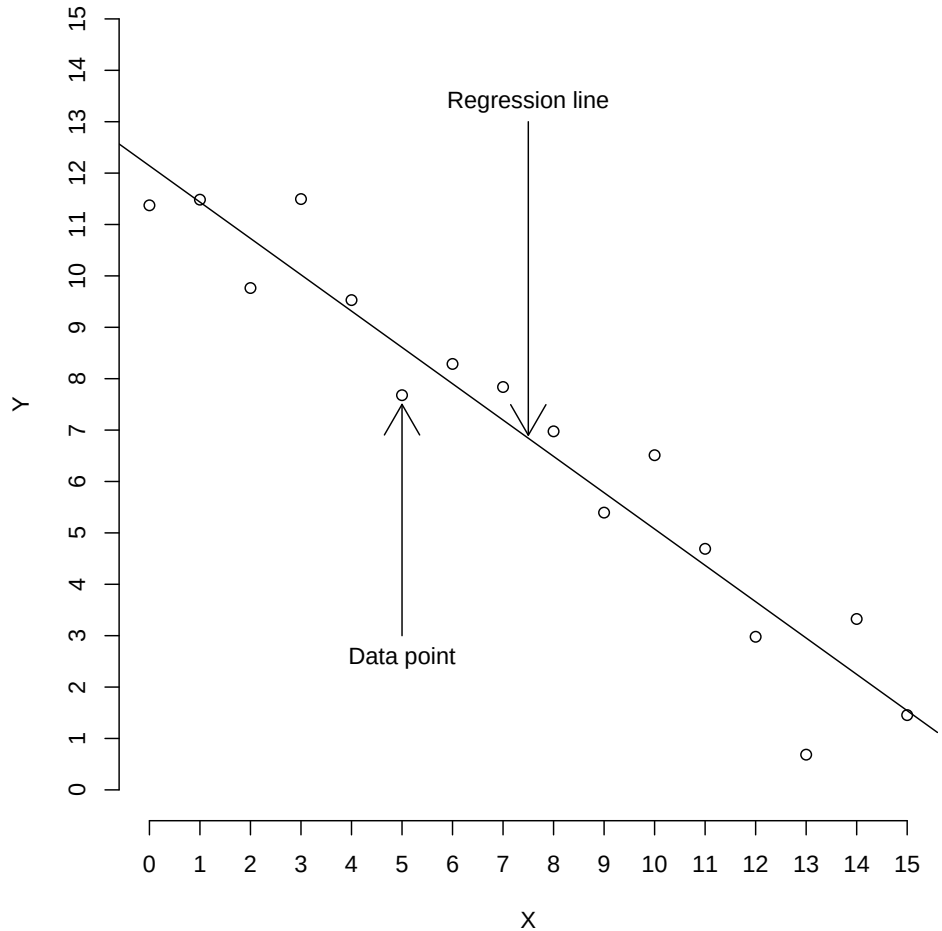
- Data



UPPSALA UNIVERSITET

A Regression Model

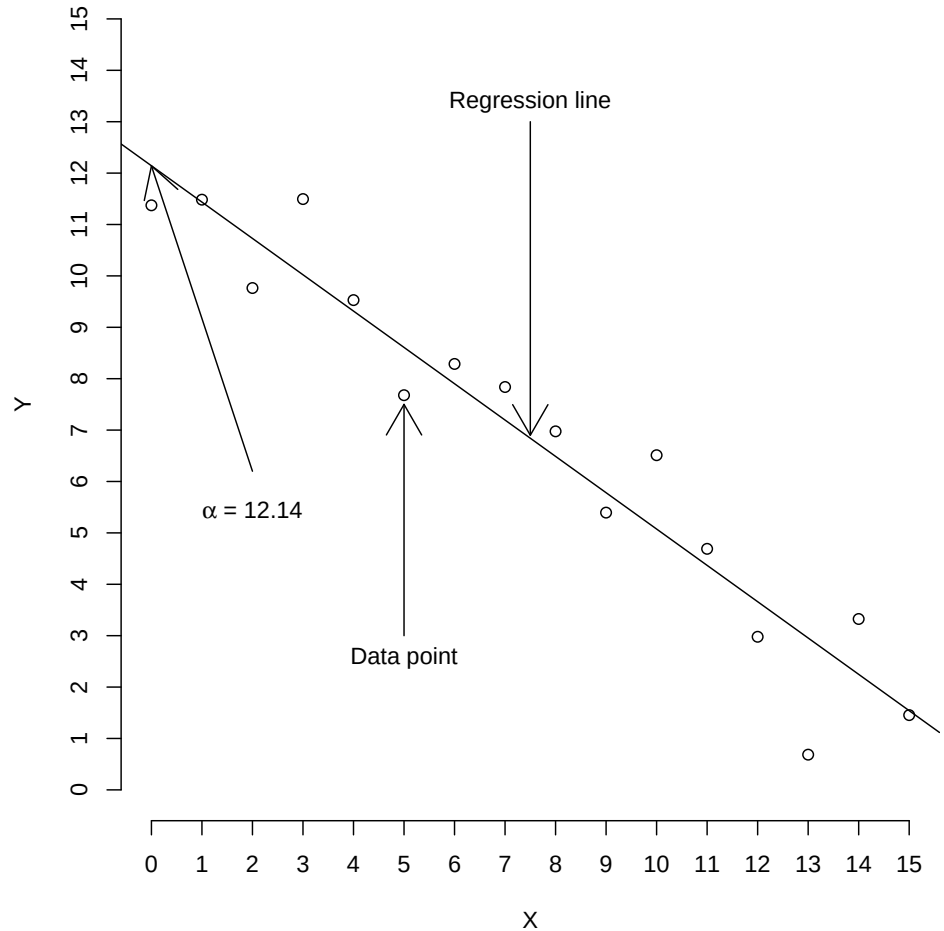
- Data
- Estimate OLS
- Obtain estimates:
  $Y_i = 12.14 + -0.70X_i$



UPPSALA UNIVERSITET

A Regression Model

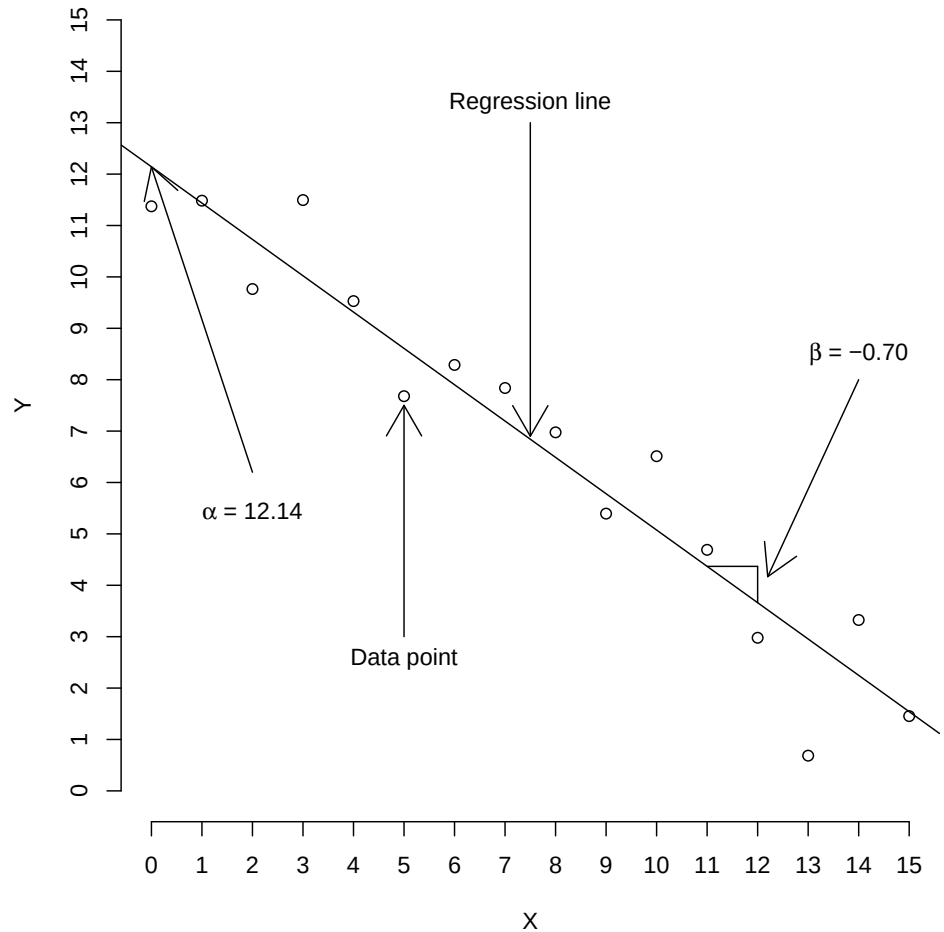
- Data
- Estimate OLS
- Obtain estimates:
  $Y_i = 12.14 + -0.70X_i$



UPPSALA UNIVERSITET

A Regression Model

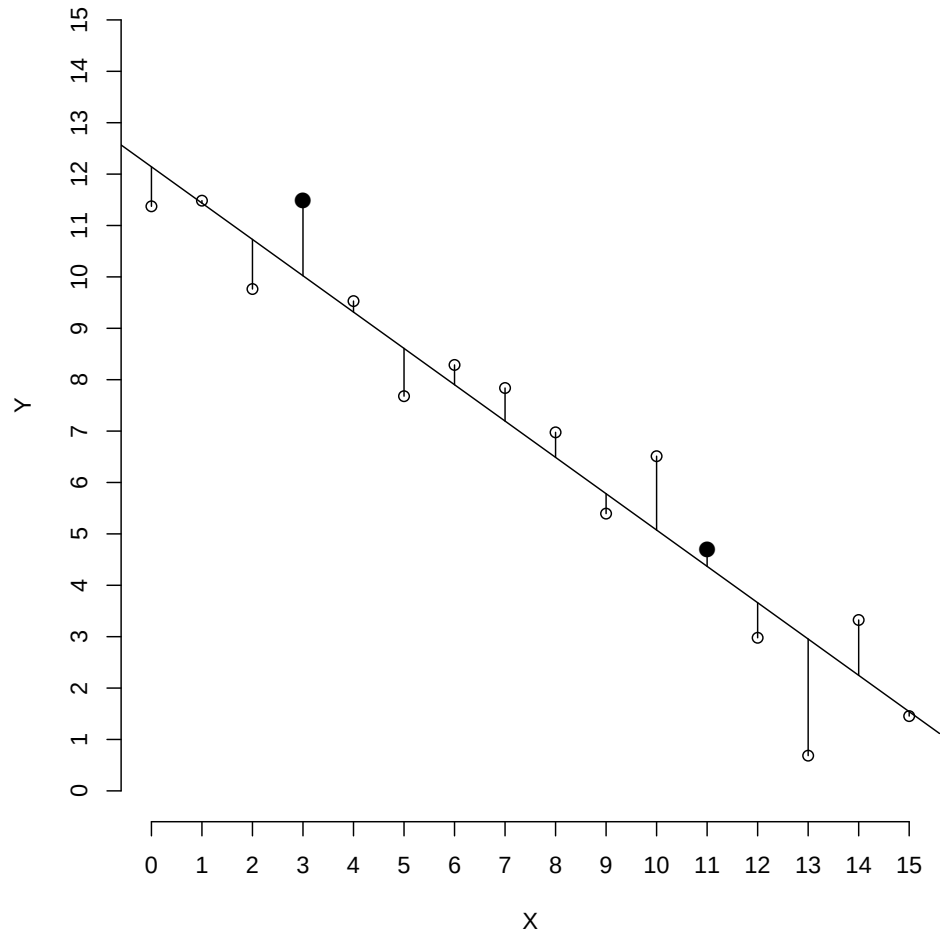
- Data
- Estimate OLS
- Obtain estimates:
  $Y_i = 12.14 + -0.70X_i$



UPPSALA UNIVERSITET

Residuals

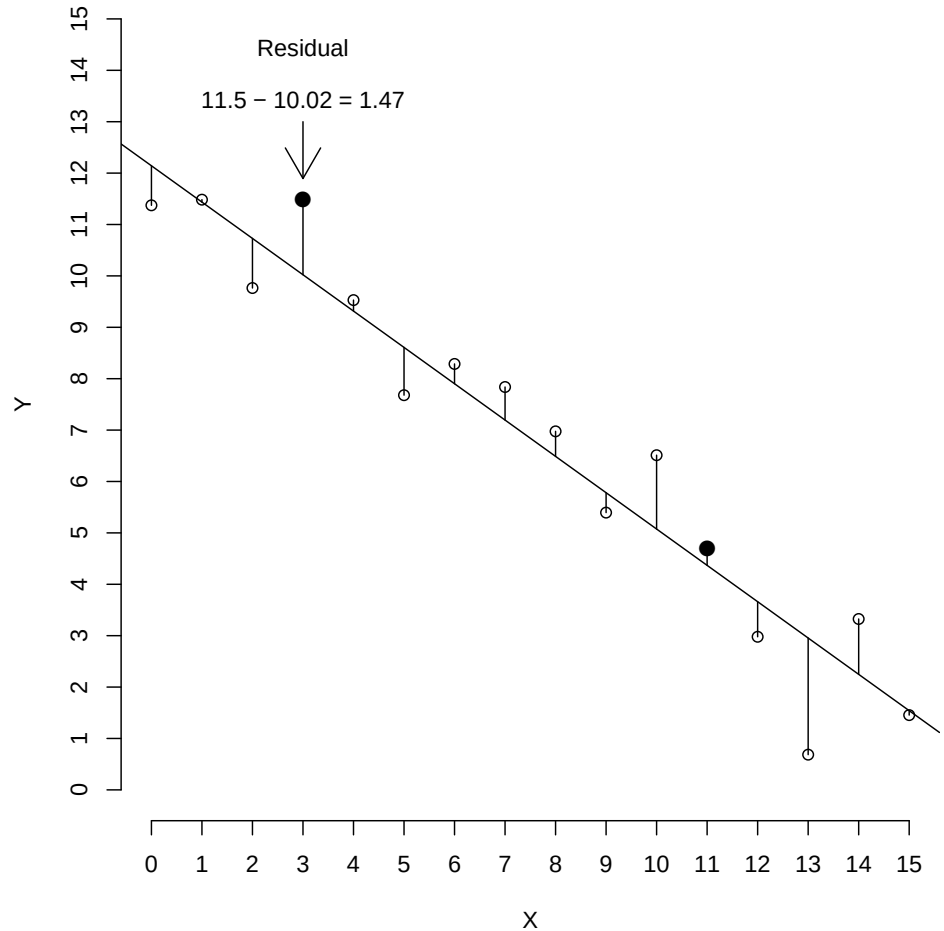
- $\hat{u}_i = Y_i - \hat{Y}_i$



UPPSALA UNIVERSITET

Residuals

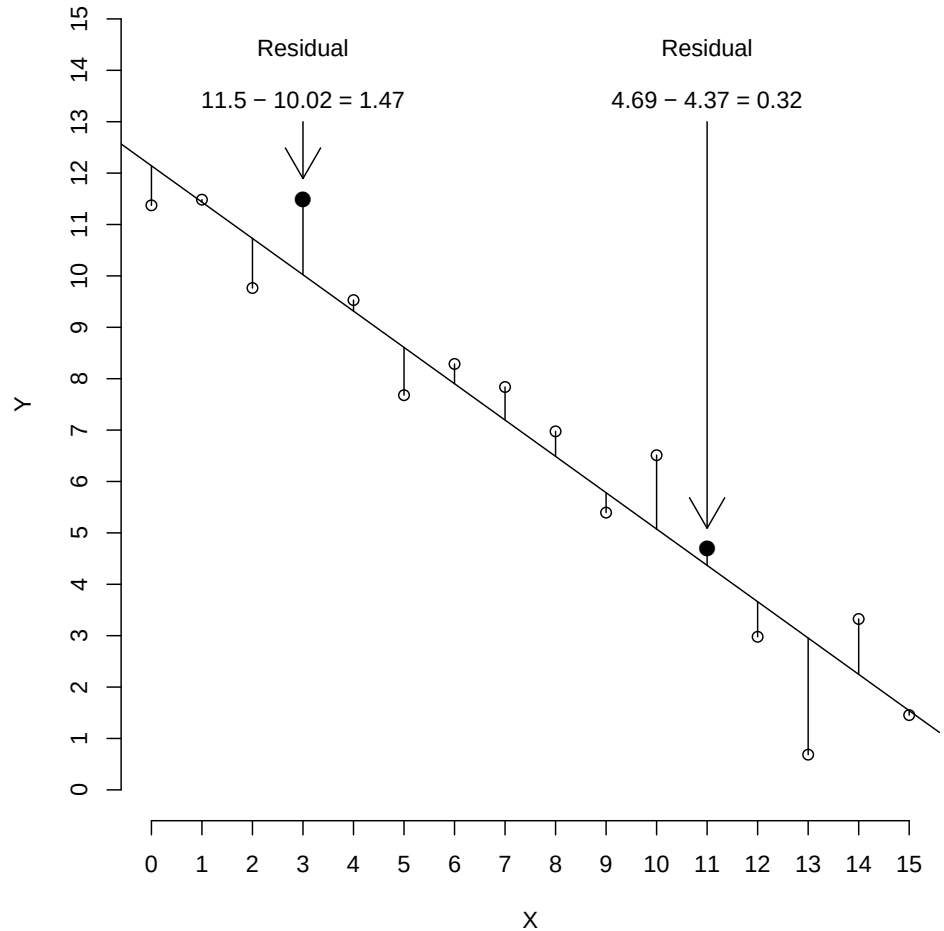
- $\hat{u_i} = Y_i - \hat{Y_i}$



UPPSALA UNIVERSITET

Residuals

- $\hat{u}_i = Y_i - \hat{Y}_i$



UPPSALA UNIVERSITET

Multiple Linear Regression Models

- What if we have multiple independent variables?
- New population regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- New interpretation of β_1:
  - β_1 represents the effect of X_1 on Y when controlling for X_2
- Alternatively:
  - β_1 represents how much Y changes when X_1 is increased by one unit while X_2 is held constant
- And vice versa for β_2
- New sample regression model: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{u}_i$

UPPSALA UNIVERSITET

Multiple Linear Regression Models

- New sample regression model: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{u}_i$
- The estimation of β-coefficients change
- The calculation of standard errors for the estimated β-coefficients change
- Remember: $se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$ in the bivariate regression model
- $se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{(\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2)(1 - R_j^2)}}$ in the multivariate regression model
  - R_j^2 is the R^2-value from a regression of X_j on X_{-j}

UPPSALA UNIVERSITET

Why do we want to add variables?

1. Some theoretical ideas require multiple variables to be operationalized properly, for example
   - a curvilinear relationship
   - an interactive relationship
2. We may be interested in maximizing explained variance (of Y), for example
   - for prediction
   - to reduce noise in order to highlight a signal
3. To avoid omitted variable bias (OVB)

UPPSALA UNIVERSITET

Omitted variable bias

See R file `Lecture2_4.R` for a little experiment...

- Arguably the biggest problem (that we know about) in quantitative social science research
- Is primarily a problem because we usually cannot randomly assign treatment in the social sciences: the majority of research is observational/non-experimental

UPPSALA UNIVERSITET

Omitted variable bias

- When we omit/exclude a relevant independent variable Z that is part of the true DGP from the statistical model
- Biases regression coefficients by producing correlation between X and the error term e_i, so that

$$E(u_i) \neq 0$$

- Oftentimes referred to as **endogeneity**
- A problem only if Z is correlated with both X and Y

UPPSALA UNIVERSITET

Omitted variable bias

An expression of the size and sign of the bias:

$$E(\hat{\beta}_1^*) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}_i\right)\left(Z_i - \bar{Z}_i\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}_i\right)^2}$$

- If β_2 or $\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}_i\right)\left(Z_i - \bar{Z}_i\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}_i\right)^2}$ is close to 0, bias is small
- If β_2 is positive, bias is positive if $corr(X_1, X_2)$ is positive, negative if $corr(X_1, X_2)$ is negative
- If β_2 is negative, bias is negative if $corr(X_1, X_2)$ is positive, positive if $corr(X_1, X_2)$ is negative

UPPSALA UNIVERSITET

Solutions to omitted variable bias

- Randomize treatment (usually not possible)
- Statistically control for confounders/cofounding variables
  - But data on Z can often not be collected..

Dummy variables and interactions

UPPSALA UNIVERSITET

Dummy variables

- Population regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- X_2 may be a dichotomous variable, e.g.:
  - man/woman
  - country at war/not at war
- We code the variable as e.g.
  - $X_2 = 0$ if male
  - $X_2 = 1$ if female
- $X_2 = 0$ (male) is the *reference category*
- Example in `R` script: The 'Hillary thermometer'

UPPSALA UNIVERSITET

Example: dummy variables in regression model

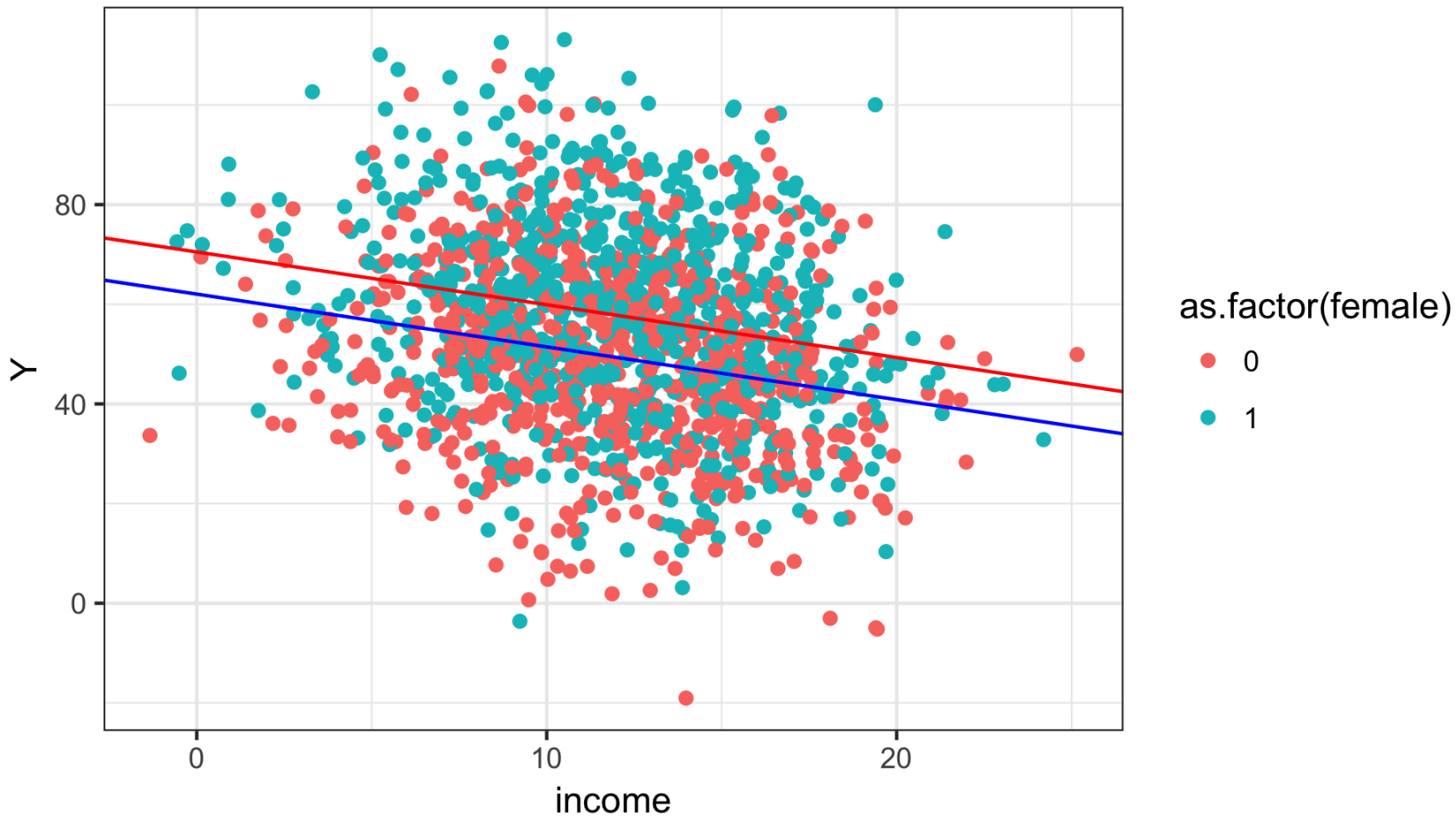


Figure 1: The 'Hillary thermometer', simulated data, as function of income and gender. Observed values and fitted regression lines for men and women

UPPSALA UNIVERSITET

Dummy variables

- A categorical variable with multiple categories may be represented as $k-1$ dummy variables (where k is the number of categories)
- E.g. four categories to represent peace and three types of conflict:
  - $sb = 0$, $ns = 0$, $os = 0$ if no conflict
  - $sb = 1$, $ns = 0$, $os = 0$ if state-based armed conflict
  - $sb = 0$, $ns = 1$, $os = 0$ if non-state armed conflict
  - $sb = 0$, $ns = 0$, $os = 1$ if one-sided violence
- (No conflict) is the reference category

UPPSALA UNIVERSITET

Interactions

- In some cases the effect of X_1 on Y is dependent on another independent variable X_2
- Example: If we study variation in Y: opinions regarding Hillary Clinton, the effect of 'feminist' attitudes on Y may be different for males and for females
- We can model this by a multiplicative interaction term:
- Population regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} \times X_{2_i} + u_i$$

- New interpretation of β_1:
  - β_1 represents how much Y changes when X_1 is increased by one unit **when** $X_2 = 0$
- And vice versa for β_2
- The interpretation for the interaction term β_3 is how much the presence of X_2 alters the effect of X_1

UPPSALA UNIVERSITET

Example: interaction continuous/dummy variable in regression model

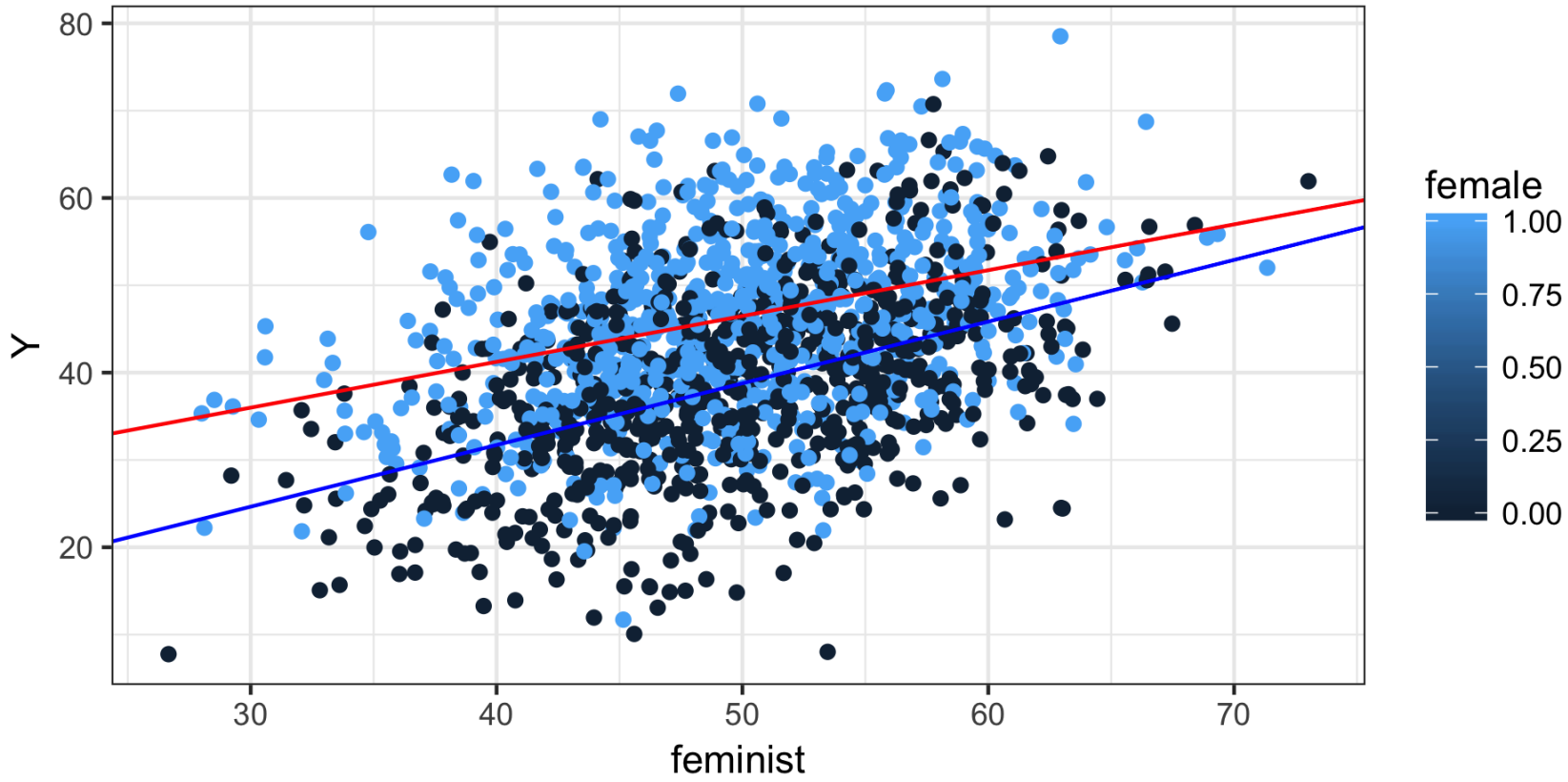


Figure 2: The 'Hillary thermometer', simulated data, as function of feminist index and gender. Observed values and fitted regression lines for men and women

UPPSALA UNIVERSITET

Multicollinearity

- Multicollinearity arises when independent variables are correlated
- *Perfect* multicollinearity is a violation of the OLS assumptions. The OLS model cannot be estimated in this case.
- Multicolinearity is still a problem in OLS even when it is not perfect
- More serious if there is a limited amount of data

UPPSALA UNIVERSITET

Multicollinearity

- Examples of situations where multicollinearity may arise:
  - Estimate the number of deaths in armed conflict as a function of GDP per capita and education levels in population
  - Estimate a country's influence in the global system as a function of total GDP, total population, and total military capabilities
  - Estimate a model where there is a large number of X variables relative to N
  - Estimate the growth rate of a country as a function of a four-category economic policy variable with mutually exclusive categories

UPPSALA UNIVERSITET

Multicollinearity: Consequences and diagnostics

- Inflated standard errors
- Unstable estimates for β_i for the correlated variables
- Remember: $se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{(\sum_{i=1}^{n}(X_{ji}-\bar{X}_j)^2)(1-R_j^2)}}$ in the multivariate regression model
- Implication: Only a problem with regards to statistical significance.

UPPSALA UNIVERSITET

Multicollinearity: Diagnostics and solution

- Variance inflation factor:
  - Run a set of regression models with X_i as the dependent variable and all the other X variables as independent variable
  - For each X_i, calculate $VIF_i = \frac{1}{1-R_i^2}$
  - Note: Same as the multiplier in the SE-estimation!
- Solving multicollinearity problems: redefine variables or collect more data
  - e.g., rather than use total GDP and total population as X variables use GDP per capita and total population

Assumptions in OLS

Gelman and Hill (2007, section 3.6)
Carsey and Harden (2014, section 1.3)

UPPSALA UNIVERSITET

Assumptions underlying the OLS model

- Assumptions about population stochastic component, with distribution

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

1. Normally distributed: $\sim N$
2. Mean equal to zero (no bias): $E(u_i) = 0$
   - u_i is random
3. Variance of $u_i = \sigma^2$ (homoskedasticity)
   - Heteroskedasticity: $u_i = \sigma_i^2$
4. No autocorrelation: $cov_{u_i,u_j} = 0 \forall i \neq j$
   - u_i is independently and identically distributed (iid)
   - Stochastic terms for two (or more) cases are unrelated
5. X variables do not contain stochastic components
   - uncertainty is due to u_i and not measurement error

UPPSALA UNIVERSITET

Assumptions underlying the OLS model

- Assumptions about model specification
  1. No causal variables left out, no non-causal variables included
     - Researchers therefore typically extend the two-variable regression:
       $Y_i = \alpha + \beta X_i + u_i$
       to multiple regression, e.g.
       $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_i$
  2. Relationship is linear (parametric linearity)
- Other requirements
  1. X must vary
  2. $n > k$

UPPSALA UNIVERSITET

Assumptions and what we know about uncertainty

- Statistical theory informs us about how certain or uncertain we can be about our parameter estimates
- Some detailed formulae in the next 3 slides
- This uncertainty assessment plays a crucial role in hypothesis testing
- However, statistical theory guarantees that the formulae are correct only if OLS assumptions are not violated
- Monte Carlo simulation will show what happens if the assumptions do not hold

UPPSALA UNIVERSITET

Uncertainty about u_i

- Estimate for the variance σ^2 of the population stochastic component u_i:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}$$

- Often, e.g. in `R` lm function, this is called 'Residual Standard Error'
- The variance for the estimate $\hat{\sigma}^2$ is increasing in the variance of the residual and decreasing in the size of the sample

UPPSALA UNIVERSITET

Uncertainty about β

- Estimate for the standard error of the slope parameter estimate β_1 :

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- In the numerator, we find the variance σ^2 of the population stochastic component u_i
- In the denominator, the sample variance for X:

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

- That is; we can be more certain about the slope parameter if the stochastic component is smaller and if we have more variation in X

UPPSALA UNIVERSITET

Uncertainty about α

- Estimate for the standard error of the intercept parameter estimate α :

$$se(\hat{\alpha}) = \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

Random variables

Gelman and Hill (2007, ch. 7)
Carsey and Harden (2014, ch. 2)

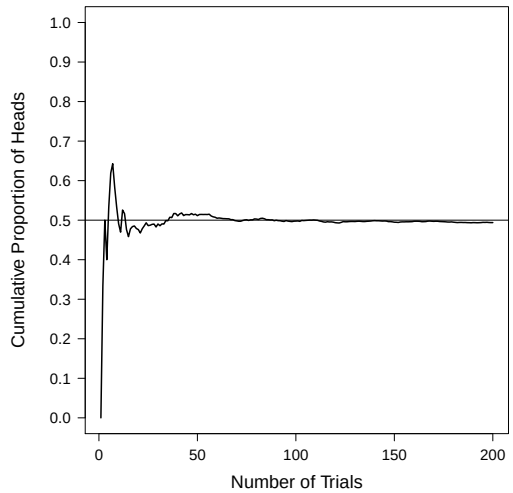
UPPSALA UNIVERSITET

Random variables and probability distributions

- **Random variable**: A variable which takes different real values mapped to some underlying distribution.
- The values the variable take are associated with probabilities which correspond to the underlying distribution of the values
- A random variable can be thought of as a random number drawn from a larger population that is described by a range of possible values
- An infinite number of draws from of the random variable will correspond exactly to the underlying distribution.

UPPSALA UNIVERSITET

Random variables and probability distributions



- Randomness occurs at the individual observation/trial level
- But the underlying distribution of probabilities associated with each possible outcome produces a predictable structure to the pattern of events in the long run
- E.g. the cumulative frequency of the proportion of coin flips that produce heads (`R` code)

UPPSALA UNIVERSITET

Discrete and continuous random variables

- Random variables may be **discrete** or **continuous**
- Each random variable has a set of possible values it can take (the range)
- A **discrete** random variable is a variable which can only take integers as values
- Each value in the range of possible integers has a probability greater than zero
- A **continuous** random variable can take any real number within the range of the variable
- Each value in the range has a probability is zero, but the probability of an interval in the range is larger than zero
- Random variables often come from a set of standard distributions. These distributions are defined by parameters.

UPPSALA UNIVERSITET

Discrete random variables

- If X is a discrete random variable the range of X is the set of all k where $P(X = k) > 0$, and k is an integer
- The probability distribution of X is the specification of these probabilities for all k values of X
- Example 1: With a fair dice $P(X = k) = 1/6 \; \forall \; k = 1, ..., 6$
- Example 2: With a loaded dice, you might have $P(X = 6) = 1/2$ and $P(X = k) = 1/10 \; \forall \; k = 1, ..., 5$
- The sum of all the probability of all possible outcomes is always 1, i.e., $\Sigma P_k(X = k) = 1$

UPPSALA UNIVERSITET

Continuous Random variables

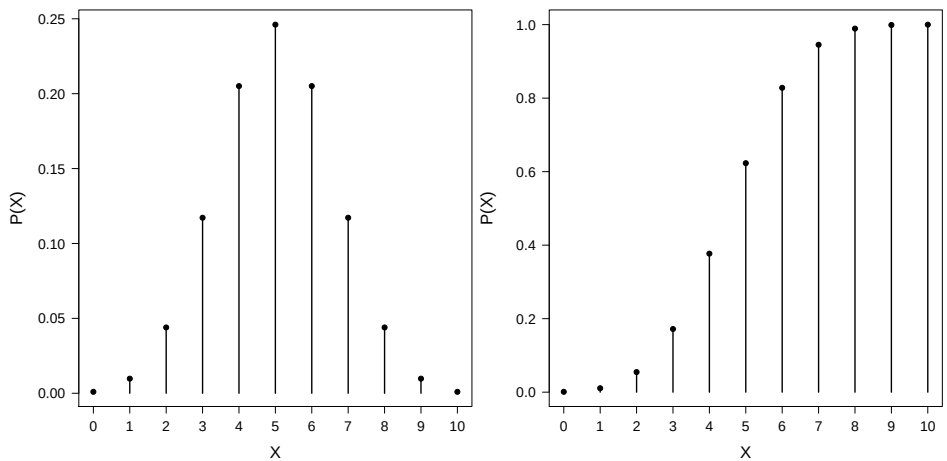
- If X is a continuous random variable, then the range of X are all real values of a where $P(a < X < b) > 0$
- However, $P(X = a) = 0 \; \forall \; a$
- Therefore, probabilities for continuous random variables are always conceptualized in ranges, for instance $P(X > 10)$
- Example: The probability that a randomly sampled female in Sweden is EXACTLY 1.80 meters tall is zero ($P(X = 1.8) = 0$), but the probability that a randomly sampled female in Sweden is taller than 1.80 meters is about 2.5% ($P(X > 1.8) \approx 0.025$

UPPSALA UNIVERSITET

PDF and CDF (Discrete variables)

- To get an overview a distribution, we can use the probability density function (PDF) and cumulative density function (CDF).
- The PDF shows, in the case of discrete variables, the probability the random variable taking a specific value.
- The CDF shows the probability that the random variable takes a value of less than or equal to the given value.

Figure 3: PDF and CDF, the binomial distribution, $n = 10, p = 0.50$



UPPSALA UNIVERSITET

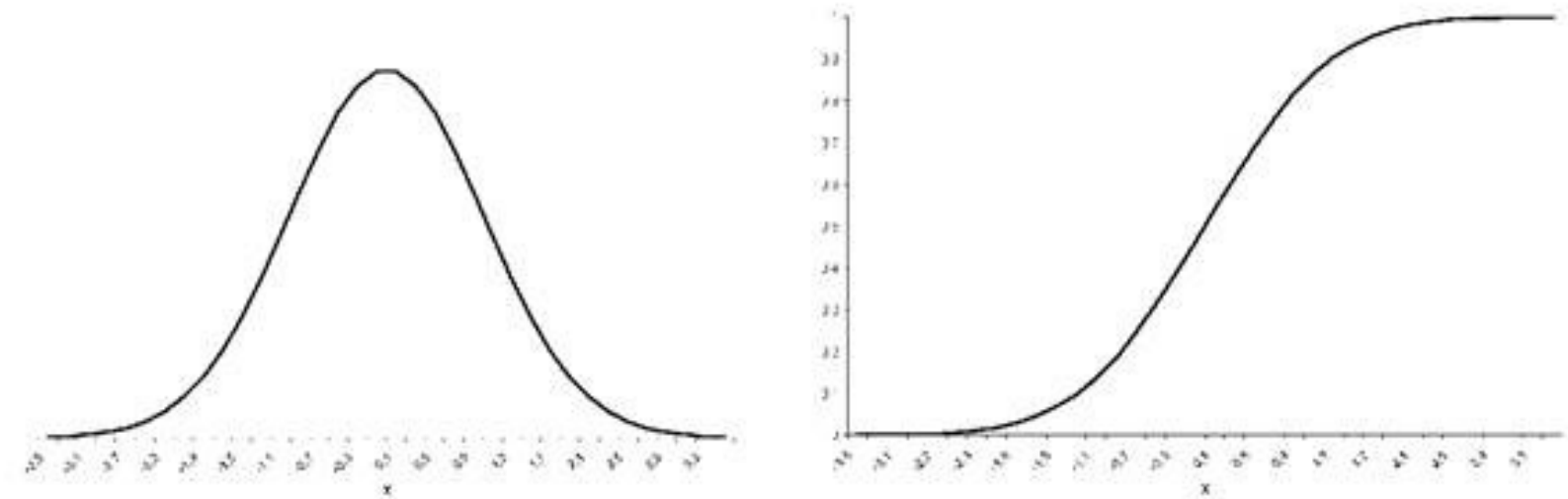
PDF and CDF (Continuous variables)

- In the case of a continuous variable, the PDF can be said to show the relative likelihood that the random variable will take a certain value.
- It is not a probability as in the discrete variable case, since the probability of a given value $P(X = a) = 0$
- We can obtain a probability by binning a range of values:
  - The area under the curve *between two points* is the probability that the random variable takes a value between those two points
- The CDF has the same interpretation as for the discrete variables, i.e. the probability that a random variable takes on a value less than or equal to a given value

UPPSALA UNIVERSITET

PDF and CDF (Continuous variables)

Figure 4: PDF and CDF, the Normal Distribution



a) Normal Probability Density Function

b) Normal Cumulative Distribution Function

Hypothesis testing: a recap

Gelman and Hill (2007, section 7.4)
Carsey and Harden (2014, section 2.1)
Also see
https://www.nature.com/articles/d41586-019-00857-9

UPPSALA UNIVERSITET

Statistical inference: from sample to population

- Population: data for every possible relevant case
- Sample: data for a subset of cases that is drawn from the population
- Sample should preferably be random
- Statistical inference is about using a sample (some information) to infer what is likely to be true about the population (all information)
- Statistical inference is probabilistic—conclusions are uncertain because we base them on limited information
- Examples?
- Why do we care whether the sample is random?

UPPSALA UNIVERSITET

Hypothesis testing

- X causes Y
  1. Internet access leads to protests in dictatorships
  2. Civil war is more likely in autocracies than in democracies
- Statements like these can be evaluated using various hypothesis tests
- Most common is to evaluate H_0: $\beta = 0$ against H_1: $\beta \neq 0$

UPPSALA UNIVERSITET

Hypothesis testing about β_1

- We know there is a stochastic component at play when we analyze data
- A seeming relationship may be due to chance
- Hypothesis testing apparatus:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- The hypotheses refer to the **population** (i.e., the DGP)
- Which of these hypotheses are supported by the data in a sample?

UPPSALA UNIVERSITET

Outcomes of hypothesis testing

Hypothesis accepted	True state of nature	
	H_0	H_1
H_0	Correct decision	False negative decision (Type II error)
H_1	False positive decision (Type I error)	Correct decision

- We want to reduce the danger of making Type I errors even more than that of Type II error
- We can estimate the probability of making a Type I error (p-value) and of a Type II error
- These probability estimates are not guaranteed to be correct if OLS assumptions are violated

UPPSALA UNIVERSITET

Hypothesis testing about β_1 : t tests

- To test a hypothesis, we formulate a t test, e.g.:

$$t_{n-k} = \frac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)} = \frac{-0.240 - 0}{0.181} = -1.326$$

- Test statistic compared to t distribution, $n - k = 98$ degrees of freedom: 1.326 is 'not statistically significant'
- Test says we should expect to observe estimates as far from 0 as this more in more than 95% of the samples we draw
- But what if our estimates $\hat{\beta}_1, se(\hat{\beta}_1)$ are systematically wrong?

UPPSALA UNIVERSITET

Hypothesis testing

- Typically compare the actual relationship between X and Y in sample data with what we would expect to find if X and Y were not related in the population
- The larger the discrepancy between the observed relationship and what we would observe if the was no relationship, the higher the confidence in the hypothesis

UPPSALA UNIVERSITET

p-values

- Most hypothesis tests rely on p-values, ranging from 0 to 1
- p-values give the probability that we would see the observed relationship between X and Y **in our sample data** if there is no relationship in the population
- Convention: p-values below 0.05 indicate a systematic relationship/statistical significance (0.1 and 0.01 are also common)
- p-values decrease as more data is collected (a precision metric)

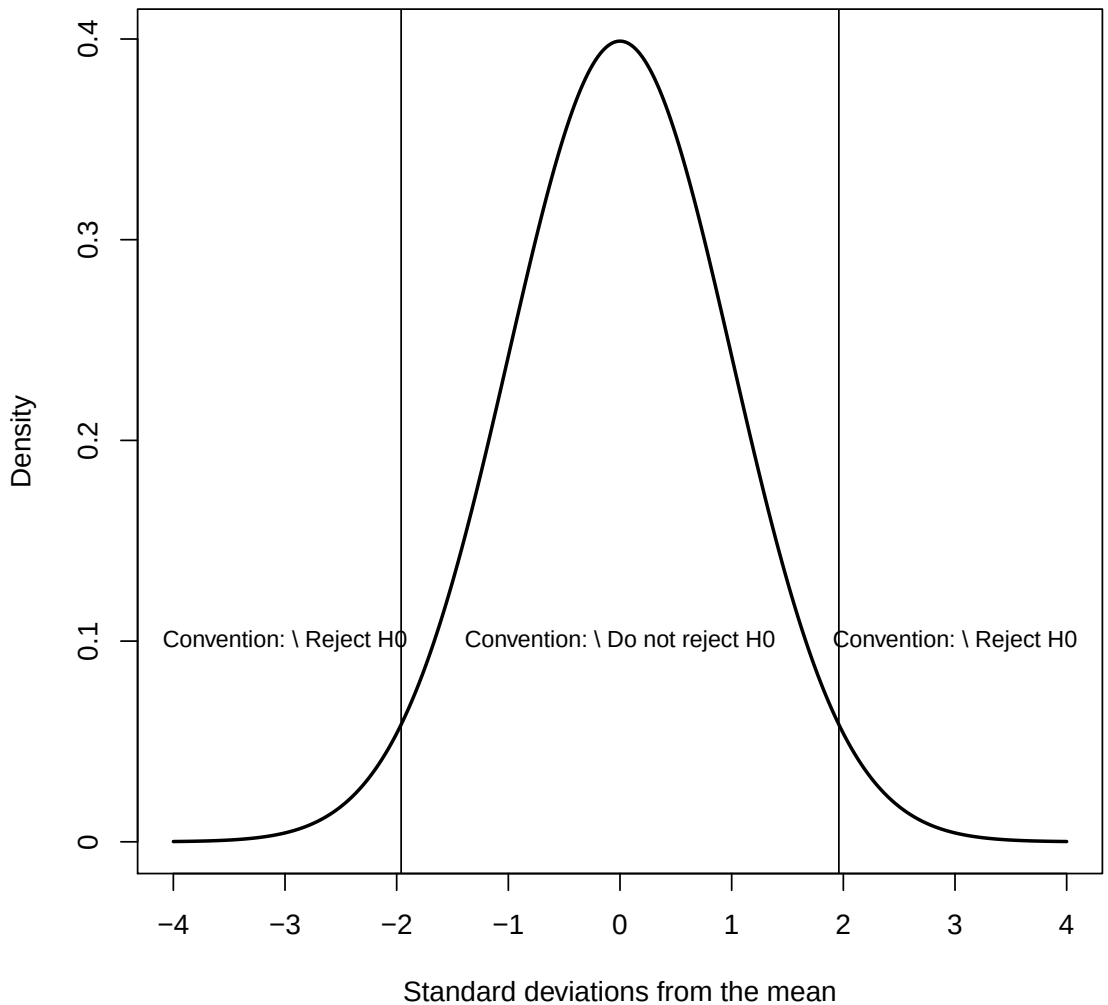
UPPSALA UNIVERSITET

p-values

- p-values do not **not** indicate:
  - Causality
  - Substantive importance of relationship

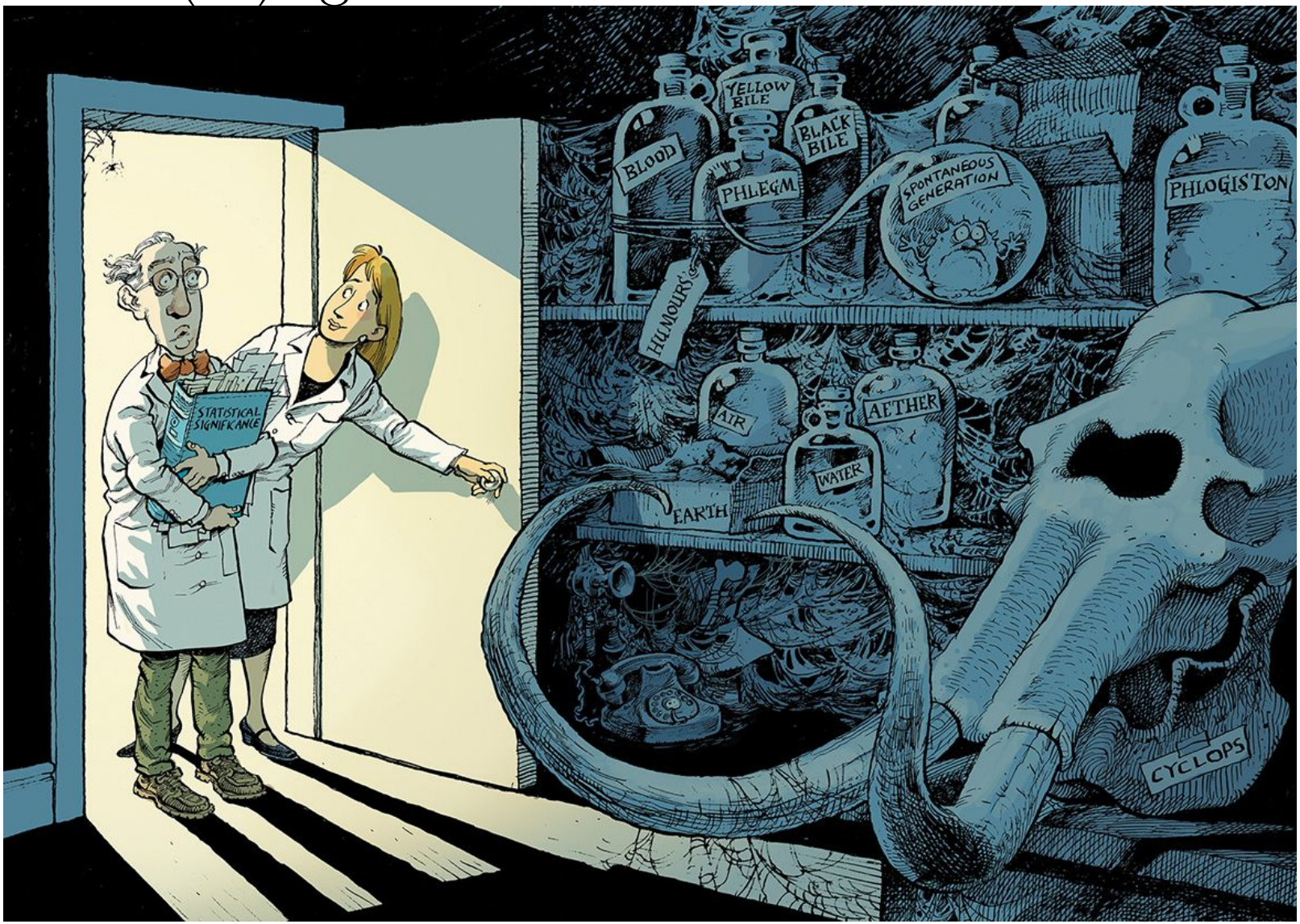
UPPSALA UNIVERSITET

Statistical significance



UPPSALA UNIVERSITET

Statistical (in)significance



Monte Carlo simulation

Gelman and Hill (2007, section 3.7)
Carsey and Harden (2014, chs. 4, 5)

UPPSALA UNIVERSITET

What is Monte Carlo simulation?

- Any computational algorithm that randomly generates multiple samples of data from a defined population based on an assumed data-generation process (DGP)

UPPSALA UNIVERSITET

What is Monte Carlo simulation?

- Any computational algorithm that randomly generates multiple samples of data from a defined population based on an assumed data-generation process (DGP)
- Emulates the process of drawing samples from a population – repeatedly
- Remember that statistical inference is about using a sample to infer what is likely to be true about the population
- But we rarely have access to repeated samples
- Monte Carlo simulation solves this problem
  - Through simulation, create many samples of data
  - Then assess patterns that appear across those repeated samples

UPPSALA UNIVERSITET

Why is Monte Carlo simulation useful (for us)?

- Allows us to assess the consequences of violating assumptions of regression models
  - For instance, what if the assumption of homoskedasticity is violated?
- A means of learning and understanding fundamental principles of statistical methods
  - **Without** requiring extensive mathematical background
- Develop intuition
- Evaluate performance of models
- Not only useful for learning the basics, but also very useful as models become complex
- Can also be used to evaluate substantive theories about social processes

UPPSALA UNIVERSITET

'In repeated samples'

- We estimate the size of a parameter in the population based on information in a sample
  - e.g., how much one degree Celsius heating increases the risk of civil war in a country
  - or, how state capacity influences log odds of government victory relative to no conflict
- Inference: we want to generalize the patterns we find in our sample of data to all of the observations that could have been in the sample
- 'In repeated samples'
  - e.g., when evaluating a sample of survey data, understand how data would change if we gave the same questions to a series of other samples
  - assuming the samples are drawn from the same population, with similar underlying characteristics
- Uncertainty: e.g. the standard error of the mean
- Also logic underlying hypothesis testing

UPPSALA UNIVERSITET

'In repeated samples'

- Use the information in a sample to draw inferences about a population using a set of assumptions and statistical theory
- That is, generalize from the sample to the observations that **could have been** in the sample
- But we rarely have access to repeated samples
- Monte Carlo simulation solves this problem:
  - Through simulation, create many samples of data
  - Then assess patterns that appear across those repeated samples
  - E.g. to assess the impact of violations of assumptions

UPPSALA UNIVERSITET

Examples

- (How) does the estimated relationship between X and Y change
  - when $E(u_i) = 0$ compared to $E(u_i) \neq 0$ (when assumption of no bias holds vs when it does not hold)?
  - when $u_i = \sigma^2$ compared to $u_i = \sigma_i^2$ (homoskedasticity vs heteroskedasticity)?
- What are the consequences of media reporting bias for the study of political violence? (Weidmann, 2016)

UPPSALA UNIVERSITET

Data-generating processes

- A data-generating process (DGP):
  - 'the mechanism that characterizes the population from which simulated samples are drawn'
  - 'the unobserved process in the larger population of how a phenomenon of interest is produced'
- Example:
  - A binomial distribution with $p = f(x), n$ trials – e.g. the number of heads when flipping a coin 10 times ($p = 0.5, 10$ trials)
  - A normal distribution, e.g. average student scores over 20 tests as a function of predictors

UPPSALA UNIVERSITET

Data-generating processes

- We rarely know what the true DGP is in the real world, we only see sample data
- Theories give an indication what the DGP is
- Most are a mix of a systematic/structural/deterministic component and a random/stochastic component
- MC simulation requires us to specify the DGP in the form of a mathematical expression
- By specifying the data-generating process, we have 'knowledge' of the true population DGP

UPPSALA UNIVERSITET

A data generating process (DGP, see Carsey and Harden, 2014)

- A DGP can be defined as
  - 'the mechanism that characterizes the population from which simulated samples are drawn'
  - 'the unobserved process in the larger population of how a phenomenon of interest is produced'
- For example: $Y_i = \alpha + \beta X_i + u_i$

UPPSALA UNIVERSITET

A data generating process (DGP)

- We rarely know what the true DGP is in the real world, we only see sample data
- Theories give an indication what the DGP is
- Most are a mix of a systematic/structural/deterministic component and a random/stochastic component
- Monte Carlo simulation requires us to specify the DGP in the form of a mathematical expression
- By specifying the data-generating process, we have 'knowledge' of the true population DGP

UPPSALA UNIVERSITET

Monte Carlo simulations as experiments

- In a laboratory, analysts can manipulate the environment to have leverage on causality
  - e.g., one set of patients gets a new drug vs. another set given a placebo
- Computer simulations follow the same logic, for example
  - Compare a set of simulations based on draws from a true DGP with homoskedasticity with a set based on draws from true DGP with heteroskedasticity
  - Compare simulations based on data with and without measurement error
- Note that 'causality' here refers to the causal effect of specification issues on the inferences we draw

UPPSALA UNIVERSITET

How to set up a script file; sequence I

For examples, see Carsey and Harden (2014, Chs. 4–5), and the `Lecture2_1.R` on github

1. Label the file
   - add a title that describes what it does, the name of the author, date last updated
2. Clear the workspace `rm(list = ls(all = TRUE))`
   - to remove objects that might be accidently referred to
3. Load packages `library()`
   - load those required for the script to run
4. Set the seed `set.seed()`
   - for replicability
5. Define objects
   - all objects that guide simulation, e.g. number of repetitions
6. Define the systematic part of the DGP

UPPSALA UNIVERSITET

How to set up a script file; sequence II

- coefficients, independent variables, sample size

7 The loop

- the stochastic component is within the loop

When does an estimator perform well?

Carsey and Harden (2014, chs. 5)

UPPSALA UNIVERSITET

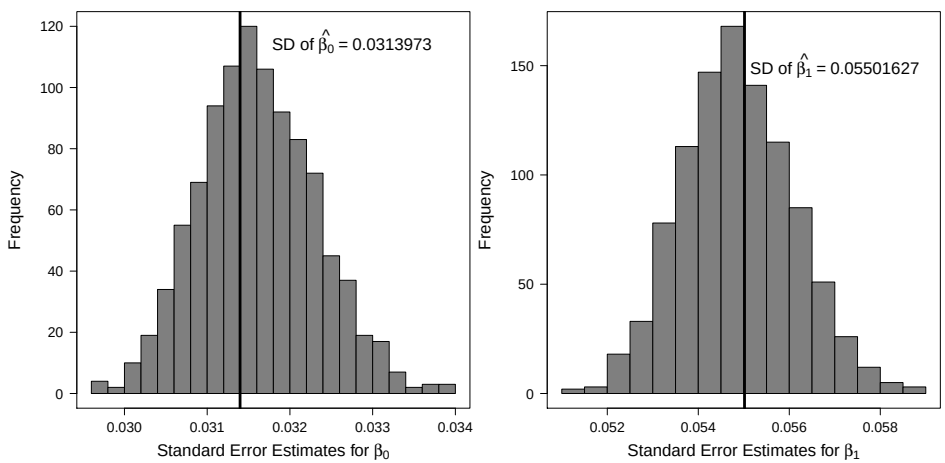
Quantities of interest: bias and efficiency

- We can quantify the consequences of violations of assumptions or other problems in terms of **bias** or **efficency**
- Bias:
  - Are the estimated parameters systematically off target compared with the true population parameters for the DGP?
- Efficiency:
  - Are the estimated parameters less precise than the true population parameters for the DGP?
- Often a trade-off between bias and inefficiency

UPPSALA UNIVERSITET

Bias

- An unbiased estimator gets the 'right answer' on average across repeated samples
- An estimator $\hat{\theta}$ for a true parameter θ is unbiased if $E(\hat{\theta}) = \theta$
- Example: Estimates of $SE(\theta)$ where the true values are 0.0314 and 0.0550, respectively



- What if the estimates of $se(\hat{\beta}_1)$ are systematically to small?

UPPSALA UNIVERSITET

Efficiency (AKA variance)

- An efficient estimator has low variance around the estimate
- The chance that it will be close to the 'right answer' in a single sample is higher
- Estimators may be both biased and inefficient:

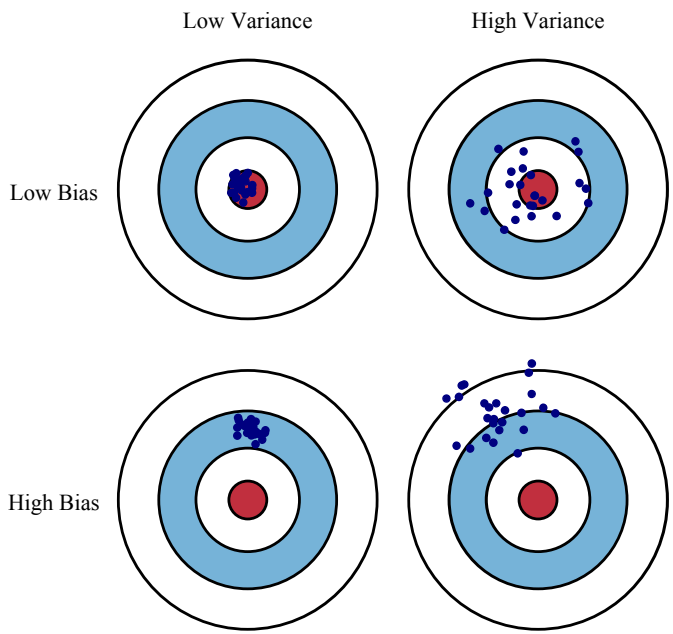


Fig. 1 Graphical illustration of bias and variance.

UPPSALA UNIVERSITET

Consistency

- A consistent estimator gets closer and closer to the 'right answer' as the sample size increases
- Taken to the extreme, a consistent estimator will be unbiased with a variance equal to zero

UPPSALA UNIVERSITET

Absolute bias vs. mean squared error

- How to quantify bias?
- Absolute bias (AKA 'error in estimation'):

$$AB = |\hat{\theta} - \theta|$$

- Mean squared error (MSE) is more commonly used:

$$MSE = E\left[(\hat{\theta} - \theta)^2\right]$$

- In a Monte Carlo analysis, calculate the mean AB or mean MSE across repeated samples
- Not informative in itself, but in comparison with another method to estimate the parameter

UPPSALA UNIVERSITET

Evaluating standard errors

- We may be interested in the estimates of coefficients $\hat{\beta}$ as well as the standard errors $\hat{SE}(\beta)$
  - SE: estimate of the variability in a parameter estimate
- We want the estimated standard error not to be too small (over-confidence) nor too large (under-confidence)
- In MC simulations, two methods to assess standard error performance:
  - the standard deviation method
    - the standard deviation of simulated coefficients should be similar to the mean of the estimated standard errors
  - the coverage probabilities method
    - the proportion of simulated samples for which the estimated confidence interval includes the true parameter should be consistent with the estimated confidence interval

Monte Carlo simulation specifics

Carsey and Harden (2014, chs. 4, 5)

UPPSALA UNIVERSITET

The coverage function

- Estimate confidence interval for the $\hat{\beta}$ estimate
- The (1-α)100% confidence interval for the slope parameter is

$$\hat{\beta} \pm t \times se(\hat{\beta}) \tag{1}$$

$$= \hat{\beta} \pm t_{\alpha/2,n-2} \times \sqrt{\frac{MSE}{\sum(x_i - \bar{x})^2}} \tag{2}$$

  where MSE is the estimate for the variance of the population stochastic component
- Coverage probability: The 95% CI will include the true β in 95% of samples if assumptions hold

UPPSALA UNIVERSITET

The coverage function

An `R` function

- Input parameters:
  - b: the vector of estimated $\hat{\beta}$
  - se: the vector of estimated $se(\hat{\beta}$
  - The true β
  - Significance level and degrees of freedom
- Output results:
  - A vector of confidence intervals
  - A vector of responses to 'is the true β within the confidence interval'?
  - Coverage probability: the proportion of estimated confidence intervals that include the true β
  - Monte Carlo error

UPPSALA UNIVERSITET

The coverage function plot

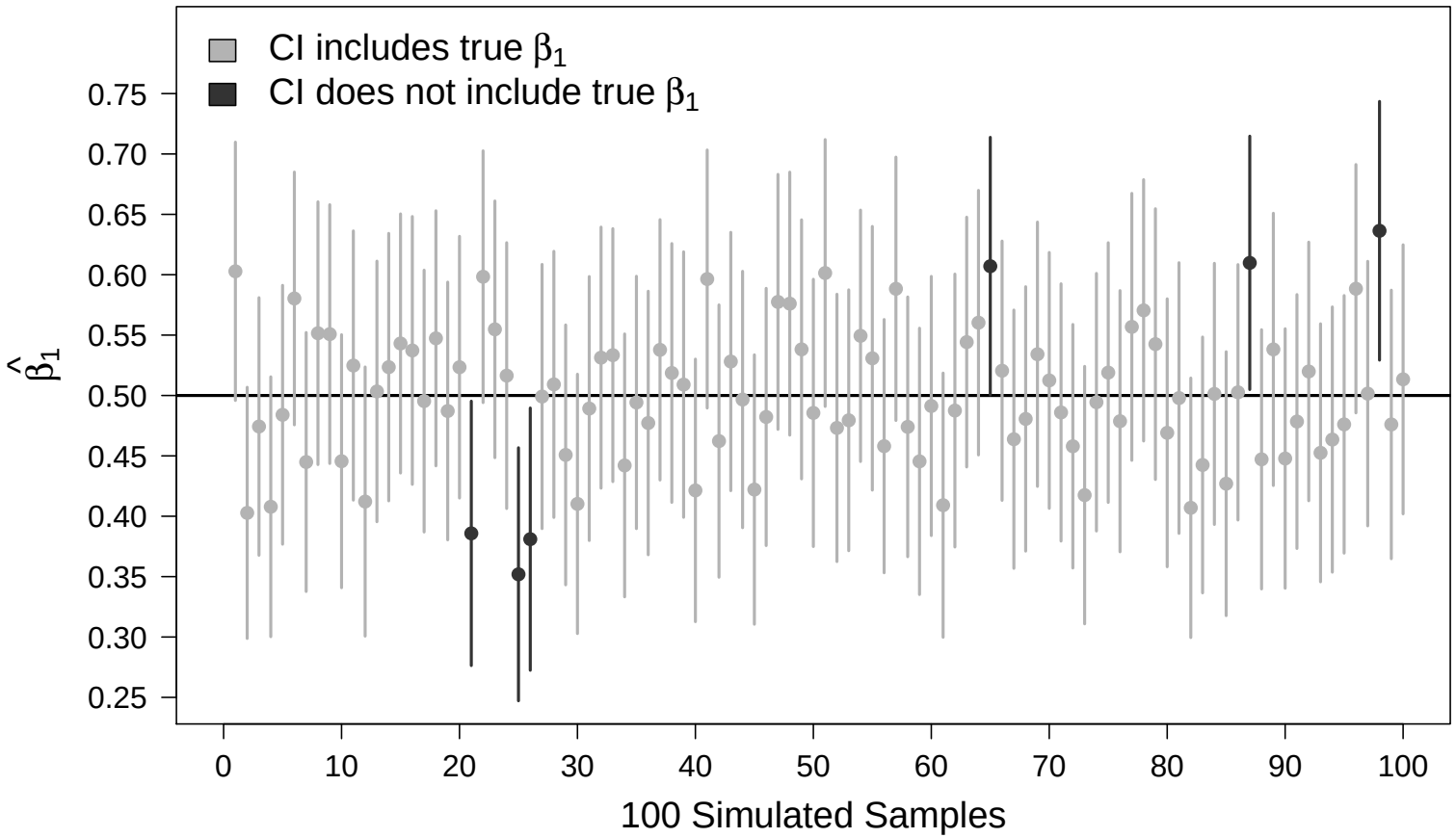


Figure 5: Estimated 95% confidence intervals for 100 simulated samples. Those that does not include the true β have black color

Some Monte Carlo simulation examples

See R files `Lecture2_2.R`, `Lecture2_3.R`, `Lecture2_4.R` in github

UPPSALA UNIVERSITET

Experiment: what are the consequences of violating the assumption of homoskedasticity?

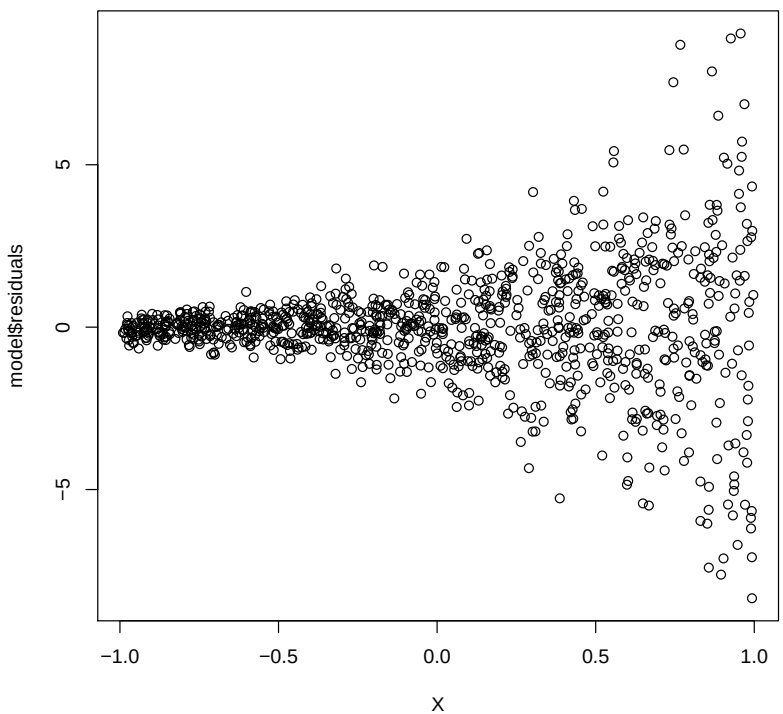
See R file `Lecture2_2.R`

- OLS assumption of homoskedasticity:
  - The variance of the dependent variable Y conditional on the model $\beta\mathbf{X}$ is constant
  - that is, that the variance of the residuals is constant, not a function of one or more X variables
- Experiment: What happens to the estimates in our model if we have a DGP where the errors have standard deviation $\sigma = exp(X \times \gamma)$?
- Run experiment, explore what happens to estimates

UPPSALA UNIVERSITET

More on homoskedasticity: What we normally see

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2579	0.0579	4.45	0.0000
X	0.6009	0.1004	5.99	0.0000



UPPSALA UNIVERSITET

How serious is this? The homoskedasticity MC analysis

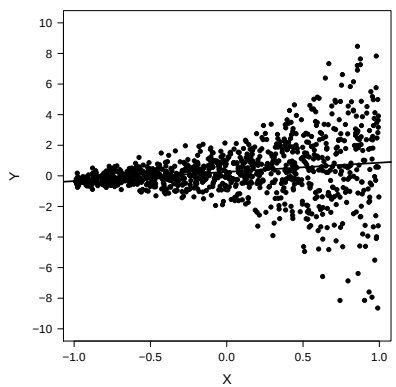
- We assess the magnitude of the violation and decide that `rnorm(n, 0, exp(X*gamma))` is an approximate model for the violation
- The formulation is convenient because it is always positive and has a parameter γ that we can adjust
- The DGP is then: `Y <- b0 + b1*X + rnorm(n, 0, exp(X*gamma))`
- We draw `reps=1000` samples from this DGP
- We want to know:
  - Are the estimated standard errors we just saw reliable?

UPPSALA UNIVERSITET

Reporting the homoskedasticity MC analysis I

1. State the DGP and other parameters of the MC analysis
   - DGP_1: $Y = .2 + .5 \times X + \epsilon, \epsilon \sim N(0, exp(X\gamma))$
   - X is drawn from a uniform distribution between –1 and 1
   - Sample size: 1000
   - Number of repetitions: 1000
2. 'Figure 11 reports the nature of the violation we are exploring consequences of:'

Figure 6: Residuals as a function of X



UPPSALA UNIVERSITET

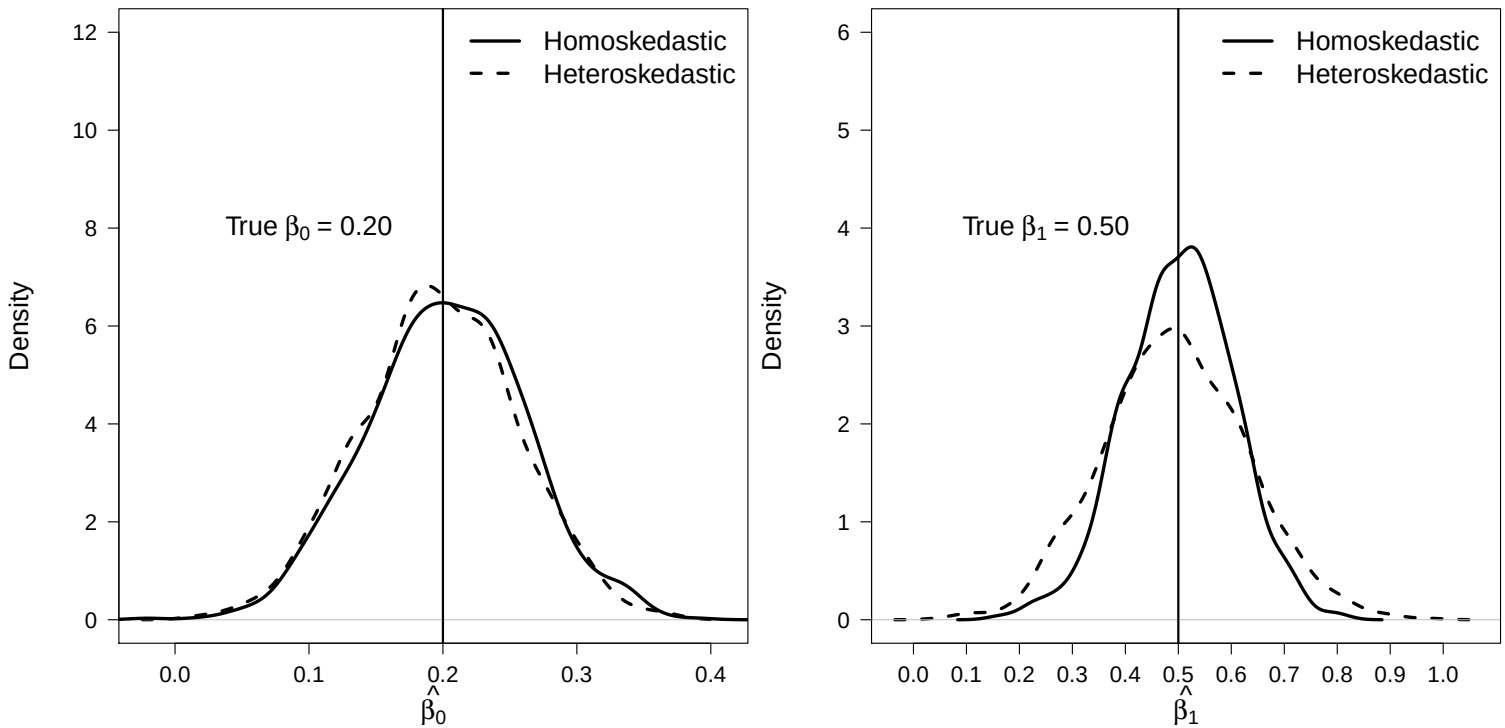
Reporting the homoskedasticity MC analysis II

3. Set up a comparison DGP that is similar except it does not violate that assumption
   - DGP_0: $Y = .2 + .5 \times X + \epsilon, \epsilon \sim N(0, \sigma)$
   - Where σ is the same as in DGP_1 but independent of X
4. Show the simulated coefficients under the two DGPs:

UPPSALA UNIVERSITET

Reporting the homoskedasticity MC analysis III

Figure 7: The effect of heteroskedasticity on the distribution of β_0 and β_1 estimates



UPPSALA UNIVERSITET

Reporting the homoskedasticity MC analysis IV

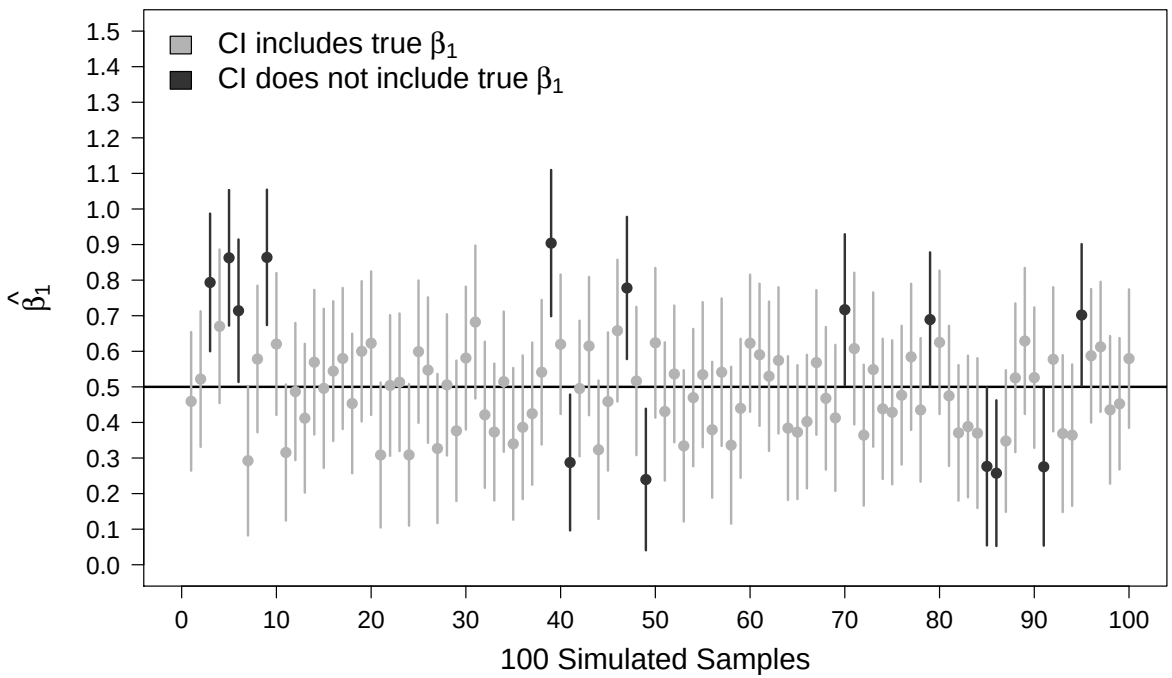
5. Since our main question was whether the estimates are over-confident, report:
   - Coverage probability and 95% error bounds for β_0 is $cp = 0.957(0.944, 0.970)$
   - Coverage probability and 95% error bounds for β_1 is $cp = 0.852(0.830, 0.874)$

UPPSALA UNIVERSITET

Reporting the homoskedasticity MC analysis V

6 Report: 'Figure 13 shows the coverage probability plot for the simulations'

Figure 8: Coefficient estimates and 95% confidence intervals of β_1 for 100 simulated samples



UPPSALA UNIVERSITET

Measurement error I

See R file `Lecture2_3.R`

- Measurement error in an independent variable
  - Analogous to having an imperfect proxy variable for X
  - E.g., Army size per 1,000 inhabitants as proxy for state capacity
- Textbooks tell us measurement error in X causes bias to regression coefficients because it produces correlation between X and the error term ϵ
  - – a situation that produces **endogeneity**
- But how to detect whether this is present in a sample of data?
  - Since non-correlation between X and ϵ is an assumption in OLS, estimated residuals and X_i are not correlated by construction $\rightarrow$ it cannot be detected

UPPSALA UNIVERSITET

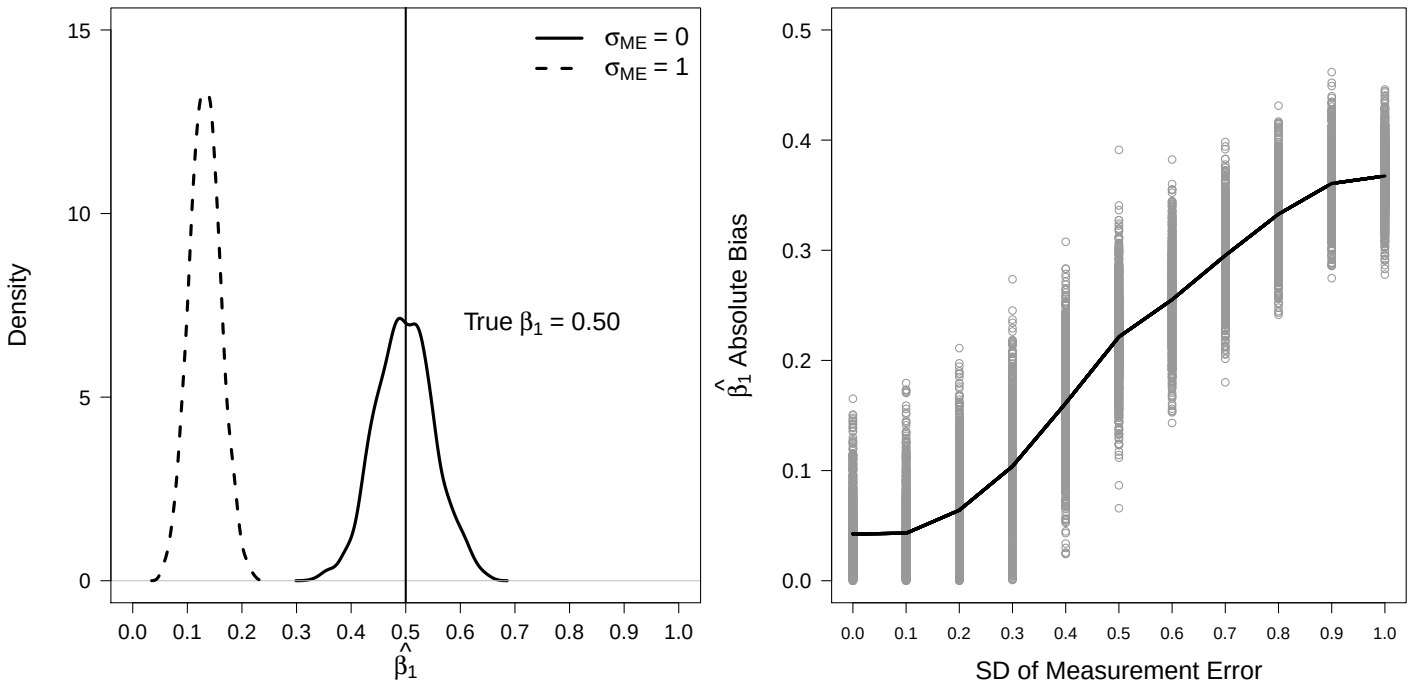
MC experiment: Measurement error I

- Strategy for MC experiment:
  1. DGP with an independent variable X **without** measurement error
  2. Generate another variable Xp which is X plus random noise/measurement error which goes into the `lm` component
  3. Repeat for different levels of measurement error `e.level`
  4. That is, change the DGP for X to $X_p = X + \epsilon_X, \epsilon_X \sim N(0, e.level)$
- True $\beta_1 = 0.5$
- `R Script`

UPPSALA UNIVERSITET

MC experiment: Measurement error II

Figure 9: The effect of measurement error in X on β_1 estimates for noise, for `e.level` $= 0$ and 1



UPPSALA UNIVERSITET

MC experiment: Measurement error III

- 'lowess line': LOcally WEighted Scatterplot Smoothing – fits simple linear regression models to segments of the X line and smooths the estimated segments
- Attenuating bias, increasing in the ratio of true variance in X relative to measurement error

UPPSALA UNIVERSITET

Measurement error: Effects I

- Pushes the absolute value of the affected betas to zero → makes it more difficult to prove effects
- p-values not correct, but significance levels are valid and possibly stronger
- Gives a harder test than correctly measured variables

Logarithms and other things you must know

The log transformation is extremely important to understand logistic regression, read up in Gelman and Hill (2007, section 4.4)

UPPSALA UNIVERSITET

Log transformation I

- Let a be a positive number $a \neq 1$. The logarithmic function $\log_a$ is defined as

$$\log_a x = y \Longleftrightarrow a^y = x$$

- If $a = 10$, $\log_{10} x$ is the number you have to raise 10 to in order to obtain x
- Hence, $\log_{10}(1000) = 3$ since $10^3 = 1000$.
- Logarithms transform what is multiplicative to something that is additive

UPPSALA UNIVERSITET

Log transformation II

- $\log_{10}(1000) = 3$
- $\log_{10}(100) = 2$
- $\log_{10}(10) = 1$
- $\log_{10}(1) =?$

UPPSALA UNIVERSITET

Log transformation III

- The 'natural logarithm' is log with base $e = 2.71828$.
- 'Natural' antilogarithm: $log(a) = x \Leftrightarrow e^x = a$
- $log(1) = 0 \Leftrightarrow e^0 = 1$

UPPSALA UNIVERSITET

Constructing regression tables

Table 1: The effects of gender and feelings toward the women's movement (WMT) on Hillary Clinton Thermometer (HCT) scores

	Dependent variable:	
	HCT scores	
	Additive model	Interactive model
WMT	0.614^{***}	0.707^{***}
	(0.035)	(0.050)
Female	7.764^{***}	16.800^{***}
	(0.491)	(3.531)
WMT x Female		-0.182^{***}
		(0.070)
Intercept	8.054^{***}	3.444
	(1.786)	(2.522)
Observations	1,500	1,500
R^2	0.271	0.274
Residual Std. Error	9.495 (df = 1497)	9.477 (df = 1496)
Note:		*p<0.1; **p<0.05; ***p<0.01

Standard errors in parentheses. *p*-values refer to two-sided tests. The dependent variable in both models is the respondent's thermometer scores for Hilary Clinton.

UPPSALA UNIVERSITET

Include in tables: I

- Title that communicates the purpost of the model and/or most important implications
- Names for independent variables that are as clear as possible
- Independent variables in order that fits your purposes (most important at the top)
- Estimated effect of each independent variable (usually $\hat{\beta}$
- Some indication of uncertainty for each estimate (standard errors or t-values)
- Some indication of which estimates are statistically significant
- What is the dependent variable

UPPSALA UNIVERSITET

Include in tables: II

- Overall diagnostics, including n and e.g. R^2
- Notes to explain symbols such as ***
- Any other information needed to convey the importance of finding

See R file Lecture2_5.R

Core question: what is the effect of multicollinearity on the stability of estimates for β_1?

Define 11 levels of multicollinearity: mc.level <- c(0, .1, .2, .3, .4, .5, .6, .7, .8, .9, .99)

Loop over the 11 levels, summarize the standard deviation of estimated β_1 for each

Repeat experiments for two sample sizes

UPPSALA UNIVERSITET

Assignment 1: Interactions I

In this assignment, we want you to look at interaction terms. Section 1 in this text specifies the initial DGP. In section 2, we want you to formulate the interpretation of a model with interaction term, and to look at the correlation that emerges when two variables that have the same sign are interacted. In section 3, we want you show that omitting the interaction term from the initial DGP leads to omitted variable bias. In section 4, you should show that centering reduces collinearity problems that interaction models have given the initial DGP. In section 5, you should demonstrate that these collinearity problem increase when the two interaction variables are highly correlated.
Remember to install and load the `car,mvtnorm` packages

1 Assume that the true data-generating process has the form

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 + u_i$$

where

- u_i has a normal distribution N(0,σ)
- X_1, X_2 are drawn from normal distributions $N(3, 1)$ (with mean = 3 and standard deviation = 1)
- $X_1 \times X_2$ is a multiplicative interaction term

UPPSALA UNIVERSITET

Assignment 1: Interactions II

- Set the true values for the β parameters to be an arbitrary combination of the values $(-1, -0.5, 0.5, 1)$, and the size of the sample to $N = 500$

2 Draw a realization from this DGP, fit two linear models – one including all the four terms and one omitting the interaction term

- Present the results in a joint regression table and interpret the results.
- Plot the predicted value for Y as a function of $X1$ for $X2 = 2$ for $X2 = 4$ and interpret the figures
- Present the joint distribution of X_1 and the multiplicative interaction term $X_1 \times X_2$ as three-dimensional density plots (one perspective plot and one contour plot). Describe and discuss.

3 Run a MC analysis to explore the consequences of omitting the interaction term from the model specification when fitting a regression model. Present results in terms of bias for the estimated parameter $\hat{\beta}_1$

UPPSALA UNIVERSITET

Assignment 1: Interactions III

4. Extend the MC analysis further by varying the correlation between X_1 and X_2 along the lines of the collinearity MC experiment from Chapter 5 in Carsey and Harden (2014). Again, focus on the fully specified model and concentrate on the collinearity issues in interaction models. For the centered and non-centered solution, report the following plots, discuss:
   1. The standard deviation of $\hat{\beta}_1$ as a function of `mc.level` – are estimates efficient?
   2. MSE for the β_1 estimates as a function of `mc.level` – are estimates biased?
   3. The mean variance inflation factor (VIF) for β_1 across simulations for each `mc.level`
   4. The estimated $\hat{\beta}_1$ against $\hat{\beta}_3$ for each of the repeated samples for `mc.level` = 0, 0.5 or .99. Choose either the centered or non-centered solution.
   5. Histograms of $\hat{\beta}_1$, $\hat{\beta}_3$, and the sum $\hat{\beta}_1 + \hat{\beta}_3$ for `mc.level` = 0, 0.5 or .99. Choose either the centered or non-centered solution. What happens here?

UPPSALA UNIVERSITET

Bibliography I

Carsey, Thomas M. and Jeffrey J. Harden. 2014. *Monte Carlo Simulation and Resampling. Methods for Social Science*. Los Angeles, London, and New Dehli: Sage.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge University Press.

Hartshorn, Scott. 2016. *Machine Learning With Random Forests And Decision Trees: A Mostly Intuitive Guide, But Also Some Python*. Goodreads.

Torfs, Paul and Claudia Brauer. 2014. "A (very) short introduction to R." Typescript, Hydrology and Quantitative Water Management Group, Wageningen University, The Netherlands.
URL: *https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf*

Venables, W.N., D.M. Smith and the core R team. 2016. "An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics. Version 3.3.1 (2016-06-21).".
URL: *https://cran.r-project.org/doc/manuals/R-intro.pdf*

Weidmann, Nils. 2016. "A closer look at reporting bias in conflict event data." *American Journal of Political Science* 60(1):206–218.

Wickham, Hadley and Garrett Grolemund. 2017. *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*. O'Reilly Media, Inc.
URL: *https://r4ds.had.co.nz*

Witten, Ian H., Eibe Frank and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann.