Regression basics

Lukáš Lafférs

Matej Bel University, Dept. of Mathematics

MUNI Brno

30.9.2021

Organization

- six meetings per 4hrs: 30.9., 1.10., 11.11., 12.11., 2.12., 3.12.

- three assignments/tests (per 20%) + written exam (40%)

- lukas.laffers@gmail.com

Some math tools

- random variable, expectation, variance, covariance

- probability density function, cumulative density function

- conditional expectation

- law of large numbers

- central limit theorem

New to R?

Here are some resources

- W. N. Venables, D. M. Smith and the R Core Team. "An Introduction to R"
  https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

- Wickham, Hadley, and Garrett Grolemund. R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.", 2016.
  https://r4ds.had.co.nz

- Wickham, Hadley. "Elegant graphics for data analysis." Springer.
  https://ggplot2-book.org

Today

- assumptions of the regression model
- geometry of linear squares
- confidence intervals
- connection to t-test
- stars and p-values - what do they mean
- purpose: prediction vs explanation
- correlated variables
- weighted regression
- heteroskedasticity
- model selection: bias-variance trade-off
- some practical considerations

Linear regression

Regression - What is it good for?

Prediction

(What will happen?)

easier, at least for repeated events

Explanation

(Why did something happen?)

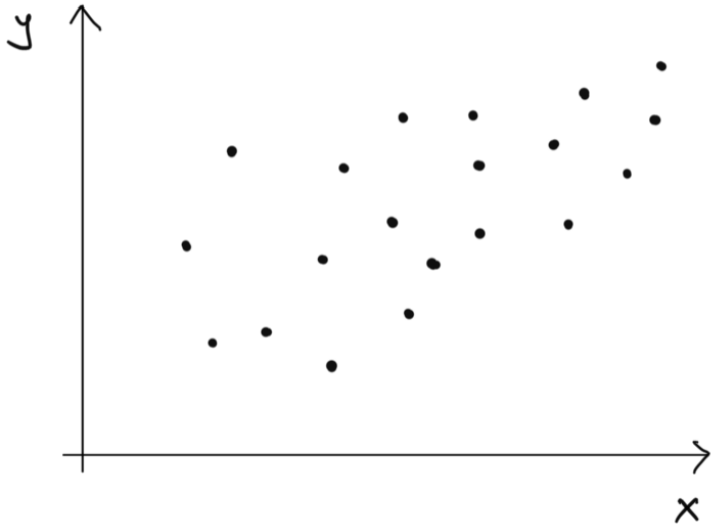difficult, requires deep institutional knowledge and subject matter expertise.

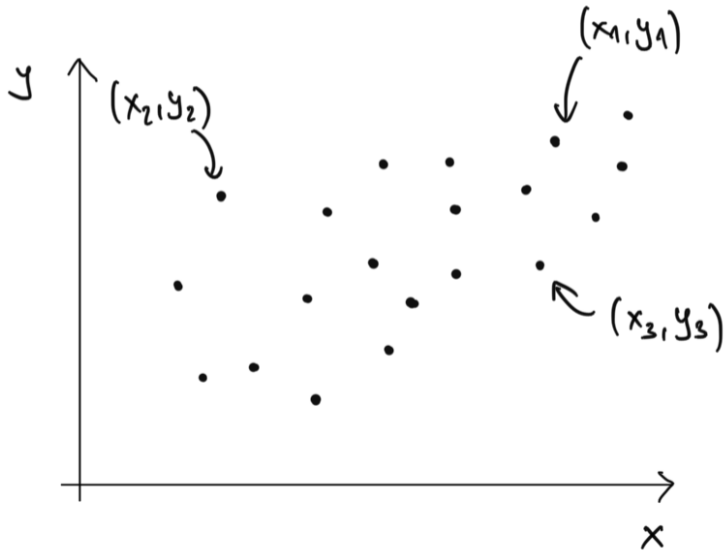These are fundamentally very different objectives.

Regression

- Prediction - easier, at least for repeated events.
- Explanation - difficult, requires deep institutional knowledge and subject matter expertise.

Data
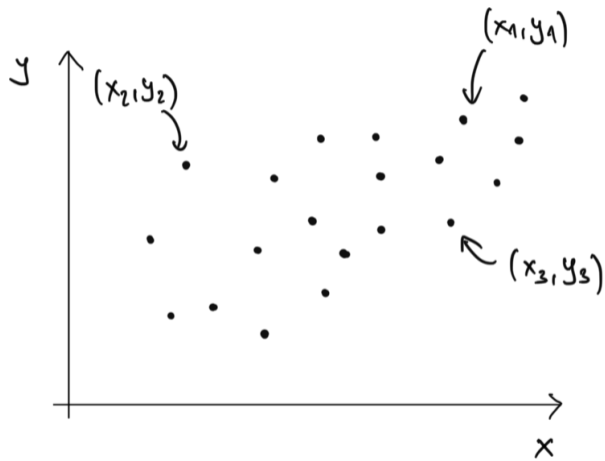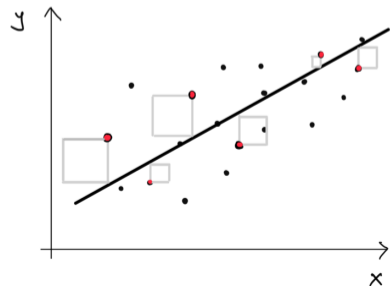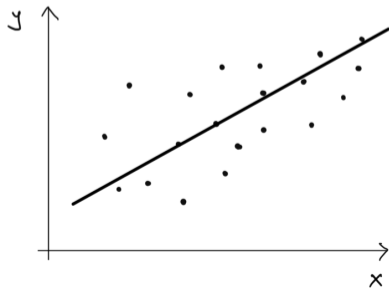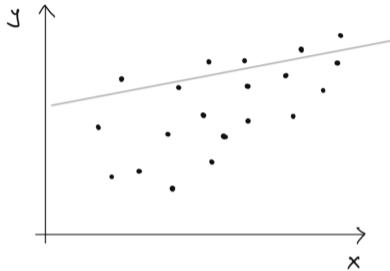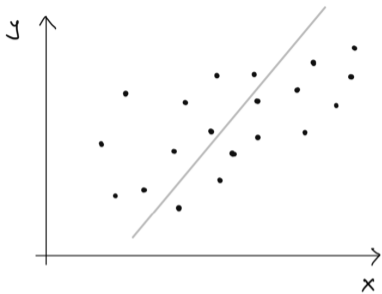
Data

- x - independent variable, regressor, factor, feature

- y - outcome, dependent variable

Least squares

Least squares
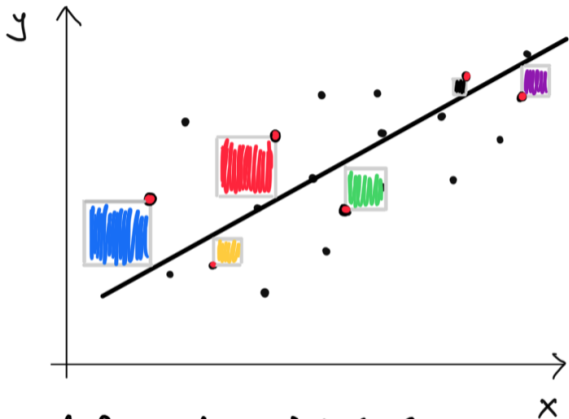
set line to minimize

 + + + + + + ...

Notation - Univariate regression

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

or

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

rewritten in matrix form

$$y = X\beta + \varepsilon$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, \ X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ ... \\ 1 & x_n \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{pmatrix}.$$

Ⓡ **Basic linear regression**

Notation - Multivariate regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$

or

$$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_{ip} + ... + \beta_p x_{ip} + \varepsilon_i$$

rewritten in matrix form

$$y = X\beta + \varepsilon$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, \ X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ ... & & ... & & \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{pmatrix}.$$

Model - philosophical interlude

The purpose of the model is to be **useful**.
It is **not** to be correct.
Model should be right about the relevant parts.
It simplifies the parts that it does not aim to model.

Ingredients:

- $y, x_1, x_2, ...,$ - observed random variables
- ε - unobserved random variables
- $\beta_0, \beta_1, ...$ - unknown variables we wish to estimate
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$ - assumption about how the different quantities are related to each other

Minimize sum of squares:

$$\sum_{i=1}^{n}(y_i - X_i\beta)^2 = \varepsilon^T\varepsilon = (y - X\beta)^T(y - X\beta)$$

differentiating via β, we get:

$$X^TX\hat{\beta} = X^Ty,$$

if X^TX is invertible, we get

$$\hat{\beta} = \underbrace{(X^TX)^{-1}X^T}_{H}y = Hy$$

Back to the model

Ingredients:

- $y, x_1, x_2, ...,$ - observed random variables
- ε - unobserved random variables
- $\beta_0, \beta_1, ...$ - unknown (fixed!) variables we wish to estimate
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$ - assumption about how the different quantities are related to each other
- $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^T$ - estimator = vector of random variables.
  - It is a function of our data sample, which is random.
  - Our best attempt to recover the true unknown β.
  - Different estimators possess different qualities (bias, variance, robustness). OLS estimator is just one of them (but pretty good).

Linear model $\neq$ simple

linear model **can** model non-linear relationships

linear = linear in parameters

Also a linear model
$$log(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

$$y = \beta_0 x^{\beta_1} \varepsilon \quad \rightarrow \quad \log(y) = \log(\beta_0) + \beta_1 \log(x) + \log(\varepsilon) = \beta_0^* + \beta_1^* x^* + \varepsilon^*$$

Goodness of fit measure

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} = \frac{\sum(\bar{y} - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2},$$

RSS *residual sum of squares*
TSS *total sum of squares*

$$R^2 = (cor(\hat{y}, y))^2 = \frac{\sum(\bar{y} - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}.$$

$R^2 = 0.65$ (for a simple linear model)
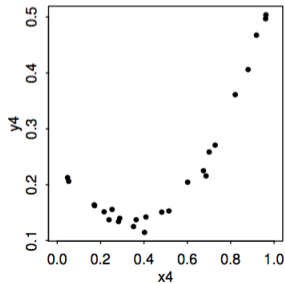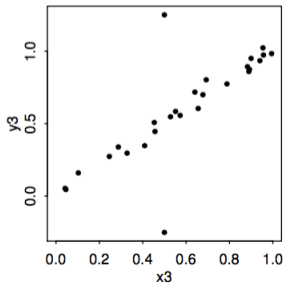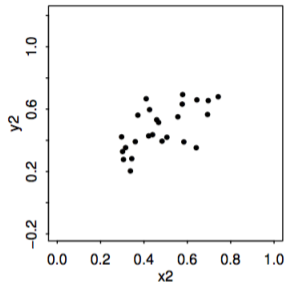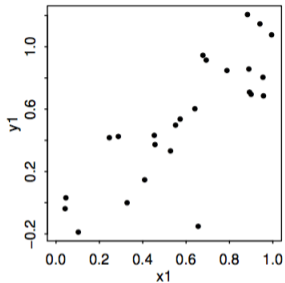
Geometry: It is a simple projection.

subspace generated by x-s.

$y = \hat{y} + \hat{\varepsilon}$

$y = \bar{y} + \bar{\varepsilon}$

TSS = ESS + RSS

$$\sum (y_i - \bar{y})^2 = \sum (\bar{y} - \hat{y}_i)^2 + \sum (y_i - \hat{y}_i)^2$$
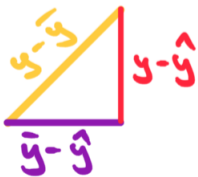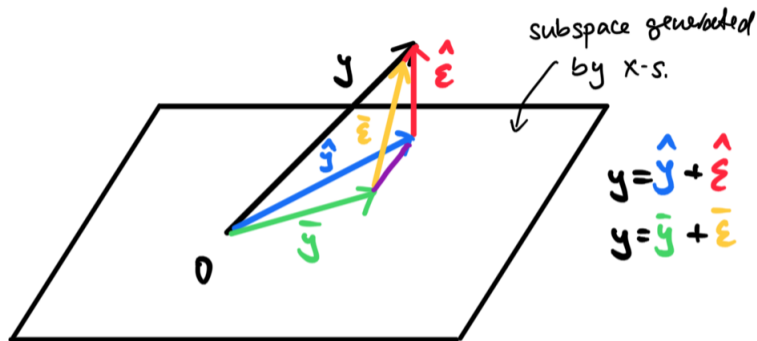
Goodness of fit?

- **Prediction** If you wish to predict well, you'd better explain the variation in y.
- **Explanation** Not necessary a problem if you have a small R^2.

Projection

$$\hat{y} = X\hat{\beta} = X\underbrace{\left(X^T X\right)^{-1} X^T y}_{\hat{\beta}} = \underbrace{X\left(X^T X\right)^{-1} X^T}_{P \,\equiv\, projection\ matrix} y = Py$$

$$\hat{\varepsilon} = y - X\hat{\beta} = \underbrace{(I - P)}_{M \,\equiv\, residual\ maker\ matrix} y = My$$

- symmetric $P^T = P$ and $M^T = M$
- idempotent $PP = P$ and $MM = M$
- $\hat{y} \perp \hat{\varepsilon}$

Different qualities Ordinary Least Squares estimator

- geometric interpretation
- easy analytic formula
  - $\hat{\beta} = (X^T X)^{-1} X^T y$
- among the linear, unbiased estimator it has the lowest variance (Gauss-Markov theorem)
  - $y = X\beta + \varepsilon$
  - $E[\varepsilon|X] = 0$
  - $Var[\varepsilon|X] = \sigma^2 I_n$
  $\implies$ For any unbiased estimator $\tilde{\beta}$ of β we have $var[\tilde{\beta}|X] \geq \sigma^2 (X^T X)^{-1}$
- Maximum Likelihood Estimator under normal errors (we will discuss latter)

Stastical inference

Assume that errors are normally distributed:

$$\varepsilon \sim N(0, \sigma^2 I) \quad + \quad y = X\beta + \varepsilon \quad \implies \quad y \sim N(X\beta, \sigma^2 I)$$

So we get that*

$$\hat{\beta} = (X^T X)^{-1} X^T y \sim N\left(\beta, (X^T X)^{-1} \sigma^2\right)$$

*$Var(Ay) = A Var(y) A^T \implies Var(\hat{\beta}) = (X^T X)^{-1} X^T (\sigma^2 I)((X^T X)^{-1} X^T)^T = (X^T X)^{-1} \sigma^2$

Hypothesis tests

$H_0 : \beta_i = 0$ can be tested:

$$t_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim t_{n-p}.$$

where

- $se(\hat{\beta}_i)$ is the standard error - sq. root of the diagonal of the matrix $(X^T X)^{-1} \hat{\sigma}^2$
- $\hat{\sigma}^2 = \frac{1}{n-p} \sum_i \hat{\varepsilon}_i^2$
- $RSS = \sum_i \hat{\varepsilon}_i^2$ is the residual sum of squares
- t_{n-p} is the Student's t-distribution with $n-p$ degrees of freedom

Hypothesis tests

$H_0 : \beta_i = \beta_j = 0$ can be tested:

$$F = \frac{(RSS_\omega - RSS_\Omega)/(2)}{RSS_\Omega/(n-p)} \sim F_{2,n-p}.$$

where

- Ω denotes a large model with p parameters
- ω denotes a small model with $p-2$ paramaters (a special case of Ω, the two models are **nested**)

Confidence intervals

$$CI^{\alpha} = [\hat{\beta}_i - t_{n-p}^{(\alpha/2)} se(\hat{\beta}_i), \hat{\beta}_i + t_{n-p}^{(\alpha/2)} se(\hat{\beta}_i)]$$

Parameter is a **fixed value** a confidence interval is **random interval**.

CIs for expected/future values

$$\hat{y}_0 = x_0^T \hat{\beta}$$

- CI for expected value

  $$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$$

- CI for future value

  $$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

  why "1+" ?

  $var(\hat{y}_0 + \varepsilon_0) = var(\hat{y}_0) + var(\varepsilon_0) = x_0^T (X^T X)^{-1} x_0 \sigma^2 + \sigma^2 = (1 + x_0^T (X^T X)^{-1} x_0) \sigma^2$

CIs for expected/future values
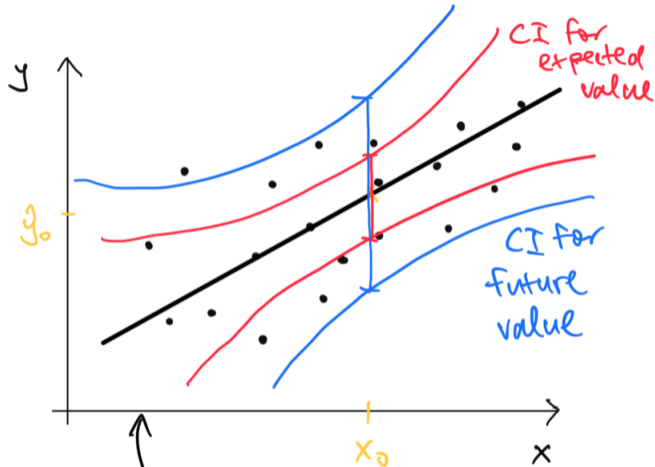
Ⓡ: **Returns to education.**

$$\log(\texttt{wage}) = \beta_0 + \beta_1 \texttt{education} + \beta_2 \texttt{age} + \varepsilon$$

~~an increase of education by one year increases wage by $\beta_1 \cdot 100\%$ percent.~~

our model predicts, that for individuals of the same age, an increase of education by 1 extra year is associated with an increase of wage by $\beta_1 \cdot 100\%$ percent.

Be careful with causal interpretations based on observational data.

Interpretation of parameters

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \varepsilon$$

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$[x_1 \rightarrow x_1 + \delta] \implies [y \rightarrow y + \beta_1 \delta]$$

our model predicts, that an increase of x_1 by one unit is associated with an increase in y by β_1 units, if the x_2 will not change.

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$[x_1 \to x_1 + \delta] \implies [y \to \exp(\beta_0 + \beta_1(x_1 + \delta) + \beta_2 x_2 + \varepsilon)) = y \cdot exp(\beta_1 \delta) \approx y(1 + \beta_1 \delta)]$$

our model predicts, that an increase in x_1 by one unit is associated with an increase in y by approximately $\beta_1 \cdot 100\%$, if x_2 will not change.

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \varepsilon$$

$$[x_1 \to x_1 \cdot (1+\delta)] \implies [y \to y + \beta_1 \log(1+\delta) \approx y + \beta_1 \delta]$$

our model predicts, that an increase in x_1 by 1% is associated with an increase in y by approximately $\beta_1/100$ units, if x_2 will not change.

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \varepsilon$$

$$[x_1 \to x_1 \cdot (1 + \delta)] \implies [y \to y \cdot \exp(\beta_1 \log(1 + \delta)) = y \cdot (1 + \delta)^{\beta_1} \approx y \cdot (1 + \beta_1 \delta)]$$

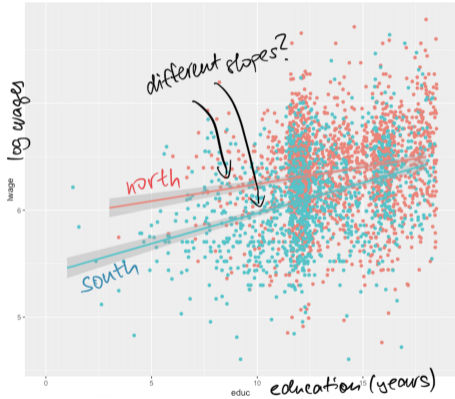*our model predicts, that an increase in x_1 by 1% is associated with an increase in y by approximately β_1%, if x_2 will not change.*

Interaction terms

$$log(wage) = \beta_0 + \beta_1\, educ + \beta_2\, south + \beta_3\, south \cdot educ + \varepsilon$$

$$log(wage) = \beta_0 + \beta_1\, educ + \varepsilon$$

$$log(wage) = \beta_0 + \beta_2 + (\beta_1 + \beta_3)\, educ + \varepsilon$$

Collinearity

Why is it a problem?
- Estimators have a high variance
- numerically unstable

How to detect it
(1) Look at the correlation matrix of regressors and look for numbers close to +1 or -1.
(2) Running a regression of x_i on other regressors, measure of linear fit R_i^2 is close to 1.
(3) Sort eigenvalues of $X^T X, \lambda_1 \geq \cdots \geq \lambda_p$. Condition number $\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}} \geq 30$ indicates problems.

How to quantify the effect of it

$$var(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2},$$

Collinearity

These two models

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

may lead to a **completely** different estimates of β_1.

®: **Pearl's Simpson Machine.**
http://dagitty.net/learn/simpson/index.html

Simpson's Simpson's paradox
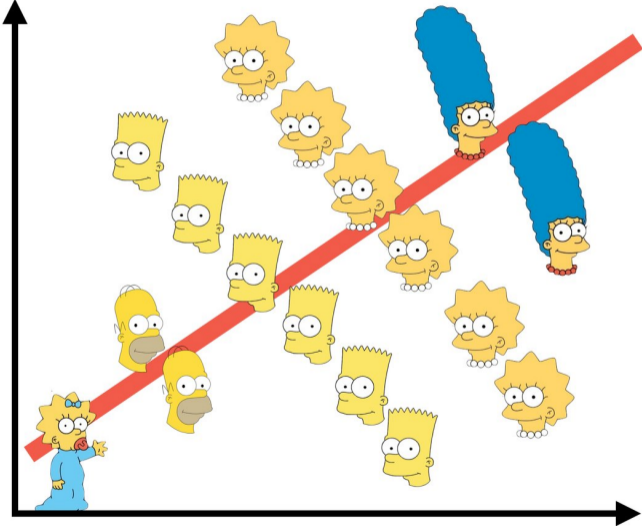
Ommitted variable bias

$$\log(\texttt{wage}) = \beta_0 + \beta_1 \texttt{education} + \beta_2 \texttt{age} + \varepsilon$$

$$\log(\texttt{wage}) = \beta_0^* + \beta_1^* \texttt{education} + \beta_2^* \texttt{age} + \beta_3^* \texttt{ability} + \varepsilon$$

$$\texttt{ability} = \gamma_0 + \gamma_1 \texttt{education} + \varepsilon'$$

$$\log(\texttt{wage}) = (\beta_0^* + \beta_3^* \gamma_0) + \underbrace{(\beta_1^* + \beta_3^* \gamma_1)}_{\beta_1} \texttt{education} + \beta_2^* \texttt{age} + (\beta_3^* \varepsilon' + \varepsilon)$$

Does not matter as long as $\texttt{ability}$ is either

- $\beta_3^* = 0$ - irrelevant
- $\gamma_1 = 0$ - uncorrelated with $\texttt{education}$

More on errors: Heteroskedasticity
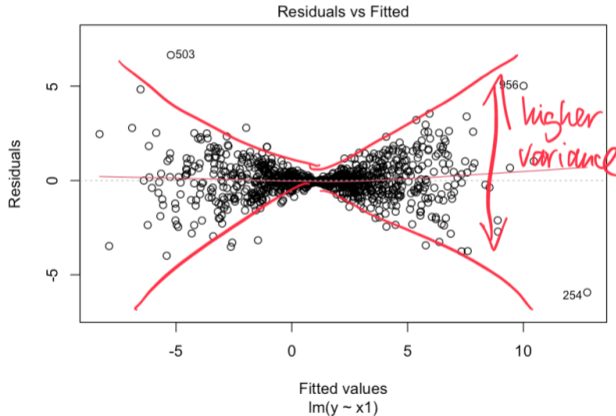
$$y_i = 1 + 3x_i + x_i \cdot \varepsilon$$

- The larger $|x_i|$ the larger the error

- But we typically assume $\sigma_i^2 = const$

So far we have assumed:

$$var(\varepsilon) = \sigma^2 I = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}.$$

But what if

$$var(\varepsilon) = \sigma^2 \Sigma?$$

$$\Sigma = SS^T.$$

Transform back

$$
\begin{aligned}
y &= X\beta + \varepsilon \\
S^{-1}y &= S^{-1}X\beta + S^{-1}\varepsilon \\
y' &= X'\beta + \varepsilon'
\end{aligned}
$$

Variance of the new transformed errors ε' is

$$var(\varepsilon') = var(S^{-1}\varepsilon) = S^{-1}var(\varepsilon)S^{-T} = \sigma^2 I.$$

We apply OLS on the transformed data $S^{-1}y$ a $S^{-1}X$.
We minimize

$$(y' - X'\beta)^T(y' - X'\beta) = (y - X\beta)^T\Sigma^{-1}(y - X\beta),$$

which is solved by

$$\hat{\beta}_W = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y.$$

The variance of this estimator is

$$var(\hat{\beta}_W) = \sigma^2(X^T\Sigma^{-1}X)^{-1}.$$

$$var(\varepsilon) = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{w_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{w_2} & \ldots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \ldots & \frac{1}{w_n} \end{pmatrix}.$$

$$S = \begin{pmatrix} \frac{1}{\sqrt{w_1}} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sqrt{w_2}} & \ldots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \ldots & \frac{1}{\sqrt{w_n}} \end{pmatrix} \qquad S^{-1} = \begin{pmatrix} \sqrt{w_1} & 0 & \ldots & 0 \\ 0 & \sqrt{w_2} & \ldots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \ldots & \sqrt{w_n} \end{pmatrix}.$$

Examples:

- If $var(\varepsilon_i) \propto x_i$, we make use of $w_i = x_i^{-1}$.
- If y_i are averages based on n_i obs, under LLN this is proportional to $1/n_i$. Hence $var(y_i) = var(\varepsilon_i) = \sigma^2/n_i$, so $w_i = n_i$. Example: average wage in different countries.
- In general we set $w_i = 1/var(y_i)$.

The covariance matrix is generally unknown, it may be estimated in different ways (R): **HR standard errors** :

- HC0: $\hat{V}_{\hat{\beta}} = (X^T X)^{-1} \left(\sum_i X_i X_i^T \hat{\varepsilon}_i^2 \right) (X^T X)^{-1}$
- HC1: $\hat{V}_{\hat{\beta}} = \left(\frac{n}{n-k} \right) (X^T X)^{-1} \left(\sum_i X_i X_i^T \hat{\varepsilon}_i^2 \right) (X^T X)^{-1}$
- HC2: $\hat{V}_{\hat{\beta}} = (X^T X)^{-1} \left(\sum_i X_i X_i^T \bar{\bar{\varepsilon}}_i^2 \right) (X^T X)^{-1}$
- HC3: $\hat{V}_{\hat{\beta}} = (X^T X)^{-1} \left(\sum_i X_i X_i^T \tilde{\varepsilon}_i^2 \right) (X^T X)^{-1}$

where (see Hansen 4.10 Residuals)

- $\hat{\varepsilon}_i = y_i - X_i^T \hat{\beta}$ is a vector of residuals
- $\bar{\bar{\varepsilon}} = (1 - h_{ii})^{-1/2} \hat{\varepsilon}$ is a vector of standardized residuals
  - h_{ii} is the leverage: the diagonal element of residual maker matrix M
- $\tilde{\varepsilon} = y_i - X_i^T \hat{\beta}_{(-i)} = (1 - h_{ii})^{-1} \hat{\varepsilon}$ is a vector prediction error
  - $\hat{\beta}_{(-i)}$ is estimated without i-th observation.

Weighting?

How do we deal with heteroskedasticity?

- Calculate $\hat{\beta}_W = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} y$ for some estimate $\hat{\Sigma}$
- Calculate $\hat{\beta} = (X^T X)^{-1} X^T y$ and then use $\hat{\Sigma}$ to adjust for standard errors

Angrist and Pischke (2008, section 3.4.1) argue for the second approach.

More efficient if $\hat{\Sigma}$ is estimated well.

But this requires good model for $E[\varepsilon^2 | X]$

Estimator for $E[\varepsilon^2 | X]$ may have bad finite sample properties.

Efficiency gains from using $\hat{\beta}_W$ are typically modest.

In case that $E[y_i | X_i]$ is not linear, the unweighted $\hat{\beta}$ estimates are at least the best linear minimum mean squared error.

Clustered standard errors

G groups. Within these groups, errors are allowed to be correlated.
But not between the groups

$$\Sigma = \begin{pmatrix} \Sigma_1 & \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 0 \end{pmatrix} & \cdots & \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 0 \end{pmatrix} & \Sigma_2 & \cdots & \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 0 \end{pmatrix} \\ \vdots & & \ddots & \\ \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 0 \end{pmatrix} & \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 0 \end{pmatrix} & \cdots & \Sigma_G \end{pmatrix}$$

When should you cluster your standard errors

- Assignment of a treatment is not on an individual level (but on a different level, say school level, town level etc)
- Sample is not random but first clusters are sampled and then observations within clusters are sampled.

Diagnostics ®
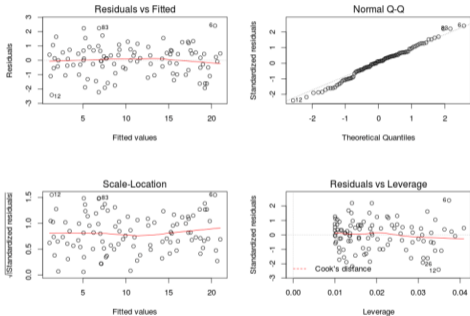
Correct model


```r
# correct model
x <- runif(100, 0, 10)
y <- 1 + 2 * x + rnorm(100, 0, 1)
m <- lm(y ~ x)
par(mfrow = c(2, 2))
plot(m)
```

Incorrect model
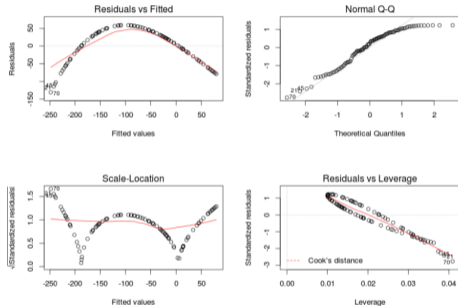

```r
# some wrong model
y <- 1 + 2 * x + 1 * x^2 - 0.5 * x^3
m <- lm(y ~ x)
par(mfrow = c(2, 2))
plot(m)
```

Model Selection - Occam's razor

- *'Among competing hypotheses, the one with the fewest assumptions should be selected.'*
- John Punch 1639: *'Entities must not be multiplied beyond necessity'*
- Aristoteles: *'We may assume the superiority ceteris paribus [other things being equal] of the demonstration which derives from fewer postulates or hypotheses.'*
- Ptolemaus: *'We consider it a good principle to explain the phenomena by the simplest hypothesis possible.'*
- Madhva: *'To make two suppositions when one is enough is to err by way of excessive supposition'*
- Isaac Newton: *'We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. Therefore, to the same natural effects we must, as far as possible, assign the same causes.'*

Small model vs Large model

- If we wish to **predict**, we may prefer **larger model** even if the smaller one is more parsimonious.
- When **explaining**, we prefer **smaller models**.

Automatic model selection based on p-values

- We add regressors with smallest p-values, until some threshold is met.
- We remove regressors with the largest p-values, until some threshold is met.
- The use of any automatic model selection tool is very risky.
- p-values are not valid as they are the result of multiple testing. Results look better than reality.
- removal $\neq$ no association
- model selection tools cannot replace deep subject matter expertise (R): **Model building**

If you have too many regressors relative to the sample size (eg. p=50, n=100).

You may find some patterns in the data just by chance!
ⓡ: **Model selection**

Freedman, David A., and David A. Freedman. "A note on screening regression equations." The American Statistician 37.2 (1983): 152-155.

$$IC = -([\text{FIT}] - [\text{COMPLEXITY}])$$

Akkaike Information criterion (AIC)

$$AIC = -2L(\hat{\theta}) + 2p,$$

we prefer models with a small AIC.

An alternative to AIC is **Bayes Information Criterion (BIC)**

$$BIC = -2L(\hat{\theta}) + 2p\log(n)$$

it penalizes large (more comples) models more.

Suppose you have a set of competing models with a similar fit

- Do they give qualitatively similar results?
- Do they predict similarly?
- How difficult/expensive is data collection?
- Are the model assumptions satisfied? (diagnostic graphs)

If we are in a situation, where models that fit data similarly well lead to very different results, it may be that we cannot answer the question of interest.

It is intelectually honest to appreciate this uncertainty, no matter how inconvenient/impractical it may be.

Model choice should not be data driven only (say looking at the statistical significance).

Subject matter expertise is always appreciated.

Regression and t-test connection

$$y_A \sim N(\mu_A, \sigma^2)$$
$$y_B \sim N(\mu_B, \sigma^2)$$

$$H_0 : \mu_A = \mu_B \quad vs \quad H_1 : \mu_A \neq \mu_B$$

$$T = \frac{(\bar{y}_A - \bar{y}_B) - (\mu_A - \mu_B)}{S_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_A + n_B - 2},$$

where $S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$ a S_A^2 a S_B^2 are sample variances.

$$y_i = \beta_A d_{iA} + \beta_B d_{iB} + \varepsilon$$

$$y_A \sim N(\beta_A, \sigma^2) \quad \text{and} \quad y_B \sim N(\beta_B, \sigma^2).$$

These are identical assumptions to the t-test.

Linear regression with dummy variables is the same thing as a t-test.

Statistical significance

Statistical vs. Practical significance (a.k.a. is the effect economically meaningful)

All point null hypotheses are wrong.

CHANGE MY MIND

More on Statistical significance
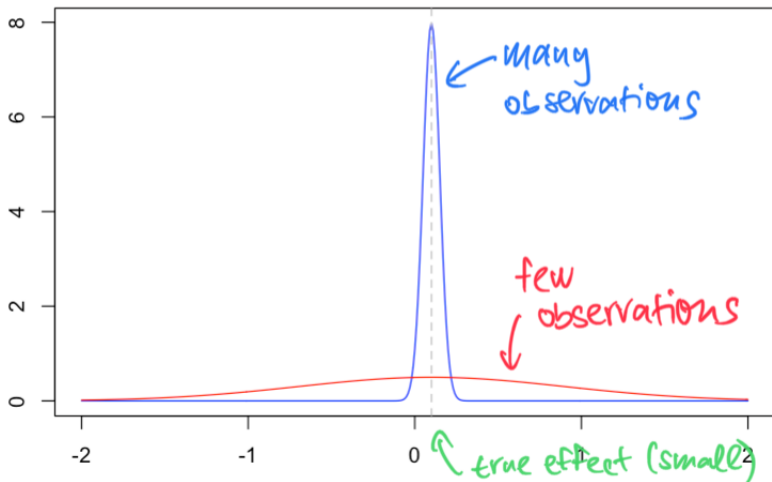
Statistical significance - p-values and confidence intervals

What they are not!
Greenland, Sander, et al. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations." European journal of epidemiology 31.4 (2016): 337-350.

Statement of American Statistical Association:
Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA statement on p-values: context, process, and purpose." (2016): 129-133.

P-values: Common misconceptions

The P value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P value for the null hypothesis is 0.08, there is an 8 % probability that chance alone produced the association.

No!

The P value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave P = 0.01, the null hypothesis has only a 1 % chance of being true; if instead it gave P = 0.40, the null hypothesis has a 40 % chance of being true.

Nope!

A significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected.

Also no!

There are 15 more variations in Greenland et al. (2016)

Confidence intervals: Common misconceptions

The specific 95 % confidence interval presented by a study has a 95 % chance of containing the true effect size.

No!

An effect size outside the 95 % confidence interval has been refuted (or excluded) by the data.

Not this one!

If two confidence intervals overlap, the difference between two estimates or studies is not significant.

Still no!

There are a few more in Greenland et al. (2016)

So what are they then??

These quantities are **model based**. They express a level of uncertainty about a particular statement (a hypothesis) assuming that a **particular model is correct**!

p-value: Assuming that the **model assumptions are correct** AND **assuming that a particular hypothesis is true**, then in a repeated sampling setup, you would observe more extreme values of a test-statistic in approximately p.100% cases.

p-value is a measure od compatibility of the calculated test statistic with the underlying model assumption and the null hypothesis.

So what are they then??

These quantities are **model based**. They express a level of uncertainty about a particular statement (a hypothesis) assuming that a **particular model is correct**!

95%CI: Assuming that the **model assumptions are correct** AND **assuming that a particular hypothesis is true**, in a repeated sampling, 95% CI is an interval estimate (that is, a random interval), that would cover the true effect in approximately 95% cases.
CI summarises the results of a hypothesis tests for multiple effect sizes.

Beautiful vis here:
https://seeing-theory.brown.edu/frequentist-inference/index.html

Misc: Principal components analysis

We often wish to reduce the dimension of X.

Source: https://www.quora.com/What-is-an-intuitive-explanation-for-PCA

Misc: Principal components analysis

Find the linear combination of regressors that maximize the variance:

$$var(Xu_1) \to \max \quad subject\ to \quad u_1^T u_1 = 1$$

then

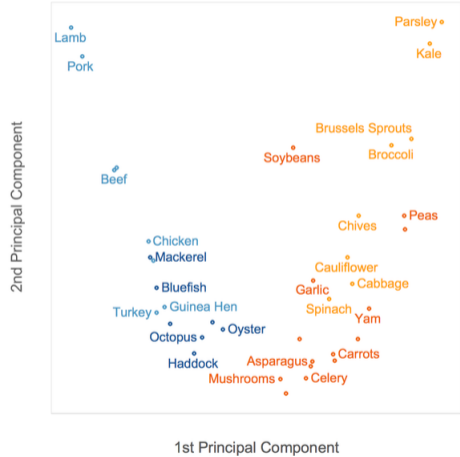$$var(Xu_2) \to \max \quad subject\ to \quad u_2^T u_2 = 1 \ \ and \ \ u_1^T u_2 = 1$$

...

Xu_1 is the first principal component
Xu_2 is the second principal component

- are orthogonal by construction
- compress most information (in terms of variance) into fewer variables
- may be interpretable
- may help us to identify "similar" points

Properties ® **PCA examples**

A few more comments...

- Ask about the data. How it was collected, handled, updated?
- Check for data discrepancies, summary statistics.
- Do plot the data. Yes, always.
- Talk to the experts.
- Comment your code, make it easily reproducible.
- Adhere to coding standards (http://adv-r.had.co.nz/Style.html).
- Do not alter your source dataset, od any preprocessing in a separate file.
- Consider interacting two most important regressors.

Thank you for your attention!

References

There are zillions of books when it comes to regression. The choices I made reflect my personal biases.

- An excellent book on linear regression with applications in R: Faraway, Julian J. Linear models with R. Chapman and Hall/CRC, 2004. Accompanying website: `https://julianfaraway.github.io/faraway/LMR/`
- Very accessible intro to regression: is ch1 in: Adams, Christopher P. Learning Microeconometrics with R. Chapman and Hall/CRC, 2020. Accompanying website: `https://sites.google.com/view/microeconometricswithr/home?authuser=0`
- A practical and very popular book with gear specifically towards economics is Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics. Princeton university press, 2008.
- Reference books (free to download) are two Bruce Hansen's books: Probability and Econometrics: `https://www.ssc.wisc.edu/ bhansen/econometrics/` The explanations are rather brief and succinct (which may not suit everyone).
- The paper that shows that you can find patterns in a noise: Freedman, David A. "A note on screening regression equations." The American Statistician 37.2 (1983): 152-155.
- When should one cluster standard errors: `https://blogs.worldbank.org/impactevaluations/when-should-you-cluster-standard-errors-new-wisdom-econometrics-oracle` is non-technical summary of: Abadie, Alberto, et al. When should you adjust standard errors for clustering?. No. w24003. National Bureau of Economic Research, 2017.
- The classic and book length account on Model selection by the pioneers of the field: Claeskens, Gerda, and Nils Lid Hjort. "Model selection and model averaging." Cambridge Books (2008).
- Discussions on statistical significance: Greenland, Sander, et al. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations." European journal of epidemiology 31.4 (2016): 337-350 and Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA statement on p-values: context, process, and purpose." (2016): 129-133.
- R resources: W. N. Venables, D. M. Smith and the R Core Team. "An Introduction to R `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`
- R resources: Wickham, Hadley, and Garrett Grolemund. R for data science: import, tidy, transform, visualize, and model data." O'Reilly Media, Inc.", 2016. `https://r4ds.had.co.nz`
- R resources: Wickham, Hadley. "Elegant graphics for data analysis." Springer. `https://ggplot2-book.org`
- For those interested in visualizations I would recommend Jonatan Schwabish's books: Schwabish, Jonathan. Better Data Visualizations: A Guide for Scholars, Researchers, and Wonks. Columbia University Press, 2021 and Schwabish, Jonathan. Better presentations. Columbia University Press, 2016. Also there is a short article Schwabish, Jonathan A. "An economist's guide to visualizing data." Journal of Economic Perspectives 28.1 (2014): 209-34.