Hospitals' influence on health ...

⇒ Do hospitals improve people's health?

- National Health Interview Survey about how healthy people feel (1 excellent health; 5 poor health)
- Result suggests that hospitals make people sicker

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

⇒ Explanation: People who go to the hospital are sicker and people who never went think they are healthier in the first place

Describe problem more analytically

- Dummy variable:

$$D_i = \begin{cases} 0 & \text{if not treated in hospital} \\ 1 & \text{if treated in hospital} \end{cases}$$

- Health status: y_i
- Observed Outcome:

$$\begin{aligned} y_i &= \begin{cases} y_{1i} & \text{if } D_i = 1 \\ y_{0i} & \text{if } D_i = 0 \end{cases} \\ &= y_{0i} + (y_{1i} - y_{0i})D_i \end{aligned}$$

- Potential outcomes

$$y_i = \begin{cases} y_{1i} & \text{if individual has been to the hospital} \\ y_{0i} & \text{if individual has not been to the hospital} \end{cases}$$
$$= y_{0i} + (y_{1i} - y_{0i})D_i \quad (1)$$

- The causal effect of hospitalization for an individual is then:

$$y_{1i} - y_{0i}$$

- Problem: We never see both potential outcomes for any one person!

Comparing conditional means

- Difference in average health:

$$\underbrace{\mathbb{E}[y_i|D_i=1]-\mathbb{E}[y_i|D_i=0]}_{\text{Observed difference in average health}} = \underbrace{\mathbb{E}[y_{1i}|D_i=1]-\mathbb{E}[y_{0i}|D_i=1]}_{\text{average treatment effect on the treated}} + \underbrace{\mathbb{E}[y_{0i}|D_i=1]-\mathbb{E}[y_{0i}|D_i=0]}_{\text{selection bias}}$$

- Average causal effect of hospitalization on those who were hospitalized:

$$\mathbb{E}[y_{1i}|D_i=1]-\mathbb{E}[y_{i0}|D_i=1]=\mathbb{E}[y_{1i}-y_{0i}|D_i],$$

where $\mathbb{E}[y_{1i}|D_i=1]$ is the health of the hospitalized, and $\mathbb{E}[y_{0i}|D_i=1]$ is the average health of those had they not been hospitalized (unobservable)

- Sick people are more likely to get treated; causes *selection bias*
- Problem: Selection bias prevents us from studying *observable* quantitity $\mathbb{E}[y_i|D_i = 1] - \mathbb{E}[y_i|D_i = 0]$
- Goal of empirical economic research: Overcome selection bias and say something about the causal effect of a variable

Random Assignment

- Random assignment of D_i solves the selection problem because it makes D_i independent
- Note that

$$\begin{aligned}\mathbb{E}[y_i|D_i=1]-\mathbb{E}[y_i|D_i=0] &= \mathbb{E}[y_{1i}|D_i=1]-\mathbb{E}[y_{0i}|D_i=0] \\ &= \mathbb{E}[y_{1i}|D_i=1]-\mathbb{E}[y_{0i}|D_i=1] \\ &= \mathbb{E}[y_{1i}-y_{0i}|D_i=1]=\mathbb{E}[y_{1i}-y_{0i}]\end{aligned}$$

since under independence of y_{0i} and D_i, we have $\mathbb{E}[y_{0i}|D_i=0]=\mathbb{E}[y_{0i}|D_i=1]$.

$\Rightarrow$ Selection bias is eliminated!

Education Research: Are smaller classes better?

- Many non-experimental studies find no/little link between class size and children learning
- Typical error: just comparing test outcomes without accounting for selection bias
- Tennessee STAR experiment: Randomized trials; implemented in 1985/86; ran for 4 years; invovled $> 11,000$ children
- Key question: Did the randomization successfully balance subjects' characteristics across the different treatment groups?

Description

a cross-section from 1985-89

number of observations : 5748

observation : individuals

country : United States

Usage

```
data(Star)
```

Format

A dataframe containing :

tmathssk total math scaled score

treadssk total reading scaled score

classk type of class, a factor with levels (regular,small.class,regular.with.aide)

totexpk years of total teaching experience

sex a factor with levels (boy,girl)

freelunk qualified for free lunch ?

race a factor with levels (white,black,other)

schidkn school indicator variable

Does randomization work?

	Class size		
	Small	Regular	Regular/Aide
1. Free lunch	0.473	0.475	0.499
2. White/Asian	0.683	0.677	0.661
3. Score percentiles	0.744	0.733	0.733
4. Girls	0.485	0.486	0.487

- Differences are small $\Rightarrow$ Randomization seems to work!

Regression Analysis of Experiments

- Assume that treatment effect is the same for everyone:

$$y_{1i} - y_{0i} = \rho$$

- Then, (1) becomes

$$y_i = \underbrace{\alpha}_{\mathbb{E}[y_{0i}]} + \underbrace{\rho}_{(y_{1i}-y_{0i})} D_i + \underbrace{\eta_i}_{y_{0i}-\mathbb{E}[y_{0i}]},$$

where η_i is the random part of y_{0i}.

- Conditional expectations with treatment switched on and off are

$$\mathbb{E}[y_i|D_i = 1] = \alpha + \rho + \mathbb{E}[\eta_i|D_i = 1]$$
$$\mathbb{E}[y_i|D_i = 0] = \alpha + \mathbb{E}[\eta_i|D_i = 0]$$

so that

$$\mathbb{E}[y_i|D_1 = 1] - \mathbb{E}[y_i|D_i = 0] = \rho + \underbrace{\mathbb{E}[\eta_i|D_i = 1] - \mathbb{E}[\eta_i|D_i = 0]}_{\text{selection bias}},$$

where ρ is the treatment effect.

- Selection bias corresponds to correlation between error η_i and regressor D_i.
- Note that

$$\mathbb{E}[\eta_i|D_i = 1] - \mathbb{E}[\eta_i|D_i = 0] = \mathbb{E}[y_{0i}|D_i = 1] - \mathbb{E}[y_{0i}|D_i = 0].$$

- Hence, correlation reflects the difference in no-treatment potential outcomes between those who get treated and those who don't.
- Under random assignment, the selection bias is not present. Regressing y_i on D_i gives the causal effect of interest ρ.

	Dependent Variable: Score percentile		
	(1)	(2)	(3)
small	0.011***	0.011***	0.011***
	(0.002)	(0.002)	(0.002)
black		−0.024***	−0.009***
		(0.002)	(0.002)
girl		0.012***	0.011***
		(0.002)	(0.001)
freelunch			−0.027***
			(0.002)
totexpk			0.001***
			(0.0001)
Constant	0.733***	0.735***	0.735***
	(0.001)	(0.001)	(0.002)
Observations	5,748	5,748	5,748
R^2	0.007	0.051	0.102

Controls

- Covariates play two roles here:
  - Assignment to classes of different sizes was random within schools, but not across schools (different school types may have different average class sizes)

    $\Rightarrow$ Need to include school-specific fixed effects
  - Control for student characteristics
- Controls $\boldsymbol{x}_i$ uncorrelated with treatment $D_i \Rightarrow$ no affect on ρ, but may generate more precise estimates of causal effect
- Yields the model

$$y_i = \alpha + \rho D_i + \boldsymbol{x}_i'\gamma + \eta_i$$

Outline I

- We define the $(n \times 1)$ vector of dependent $\boldsymbol{y}$ and the $(n \times k)$ matrix of independent variables $\boldsymbol{X}$ as follows:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1},$$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}_{n \times k} = \begin{pmatrix} x_{11} & x_{12} & \vdots & x_{1k} \\ x_{21} & x_{22} & \vdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \vdots & x_{nk} \end{pmatrix}_{n \times k}.$$

Conditional Expectation Function

- Definition of the CEF for continuous y_i and conditional density $f_{y_i}(\cdot|\boldsymbol{x}_i = x)$

$$\mathbb{E}[y_i|\boldsymbol{x}_i = x] = \int t f_{y_i}(t|\boldsymbol{x}_i = x)dt$$

- Definition of the CEF for discrete y_i and conditional density $f_{y_i}(\cdot|\boldsymbol{x}_i = x)$

$$\mathbb{E}[y_i|\boldsymbol{x}_i = x] = \sum_t t f_{y_i}(t|\boldsymbol{x}_i = x)$$

- CEF is random because $\boldsymbol{x}_i$ is random and the CEF is a function of $\boldsymbol{x}_i$!

CEF-Decomposition Property

- An implication of the law of iterated expectations is the possibility to split up a random variable in two pieces:

$$y_i = \mathbb{E}[y_i|\boldsymbol{x}_i] + \varepsilon_i$$

  - (i) ε_i is "mean-independent" of $\boldsymbol{x}_i$, i.e. $\mathbb{E}[\varepsilon_i|\boldsymbol{x}_i] = 0$
  - (ii) ε_i is uncorrelated with any function of $\boldsymbol{x}_i$

- y_i can be decomposed into a piece that is perfectly correlated with $\boldsymbol{x}_i$ (the CEF) and a piece, which is orthogonal to any function of $\boldsymbol{x}_i$

CEF Prediction Property & ANOVA

- CEF-Prediction Property: Let $m(\boldsymbol{x}_i)$ be any function of $\boldsymbol{x}_i$. The CEF solves

$$\mathbb{E}[y_i|\boldsymbol{x}_i] = \underset{m(\boldsymbol{x}_i)}{\operatorname{argmin}}\, \mathbb{E}[(y_i - m(\boldsymbol{x}_i))^2],$$

  so it is the minimum mean squared error (MMSE) predictor of y_i given $\boldsymbol{x}_i$.

- ANOVA Theorem: The unconditional variance of y_i can be split up as follows:

$$\mathbb{V}[y_i] = \mathbb{V}(\mathbb{E}[y_i|\boldsymbol{x}_i]) + \mathbb{E}[\mathbb{V}(y_i|\boldsymbol{x}_i)].$$

CEF and linear regression

- *Population regression problem*: Least squares problem formulated in terms of non-random features of the joint distribution of dependent and independent variables.
- A population regression coefficient is the solution to a population least squares problem.
- Let the $(k \times 1)$ regression coefficient vector $\boldsymbol{\beta}$ be defined as

$$\boldsymbol{\beta} = \underset{\tilde{\boldsymbol{\beta}}}{\operatorname{argmin}}\, \mathbb{E}[(y_i - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})^2]$$

- First order condition:

$$\mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})] = 0$$

yielding

$$\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i']^{-1}\mathbb{E}[\boldsymbol{x}_i'y_i]. \tag{2}$$

- By construction:

$$\mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})] = \mathbf{0},$$

where

$$y_i - \boldsymbol{x}_i'\boldsymbol{\beta} = \varepsilon_i.$$

defines the population residual, which is uncorrelated with the regressors $\boldsymbol{x}_i$.

Anatomy of multivariate regression

- In a multivariate regression, i.e., with more than one (non-constant) regressor, the slope coefficient for the k-th regressor is given by coefficient

$$\beta_k = \frac{\mathbb{C}\text{ov}\left[y_i, \tilde{x}_{ki}\right]}{\mathbb{V}[\tilde{x}_{ki}]}$$

  where $\tilde{x}_{ki}$ is the residual from a population regression of x_{ki} on all other covariates.

- Hence, $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i']^{-1}\mathbb{E}[\boldsymbol{x}_i' y_i]$ is the $(k \times 1)$ vector with k-th element given by $\mathbb{C}\text{ov}\left[y_i, \tilde{x}_{ki}\right]/\mathbb{V}[\tilde{x}_{ki}]$.

Justification for regression

Theorem (Linear CEF). Suppose the CEF is linear. Then, the population regression function is it, i.e., $\mathbb{E}[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i'\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ given by (2).

But when is the CEF linear?

- If the vector $(\boldsymbol{y}, \boldsymbol{x}_i')$ is multivariate normally distributed (problem: variables are often discrete)
- If the regression models are saturated, i.e., there is a separate paramter for every possible combination of values the regressors can take

Theorem (Best Linear Predictor). The function $\boldsymbol{x}_i'\boldsymbol{\beta}$ is the best linear predictor of $\boldsymbol{y}$ given $\boldsymbol{x}_i$ in a MSE sense. Hence, $\mathbb{E}[y_i|\boldsymbol{x}_i]$ is the best MMSE predictor of y_i given $\boldsymbol{x}_i$ in the class of all functions of $\boldsymbol{x}_i$.

Theorem (Regression-CEF Theorem). The function $\boldsymbol{x}_i'\boldsymbol{\beta}$ provides the minimal MSE linear approximation to $\mathbb{E}[y_i|\boldsymbol{x}_i]$, that is

$$\boldsymbol{\beta} = \underset{\tilde{\boldsymbol{\beta}}}{\operatorname{argmin}}\, \mathbb{E}[(\mathbb{E}[y_i|\boldsymbol{x}_i] - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})^2]$$

- While the CEF is the best unrestricted predictor (in a MSE sense) of the dependent variable, the regression provides the best linear predictor for it.

- If we want to approximate $\mathbb{E}[y_i|\boldsymbol{x}_i]$ (as opposed to predicting y_i), regression provides the best (in a MSE sense) linear approximation to it.

How to estimate β?

- The population regression coefficient

$$\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i']^{-1} \mathbb{E}[\boldsymbol{x}_i' y_i].$$

  is estimated by replacing the population moments $\mathbb{E}[.]$ by their corresponding sample means:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \hat{\mathbb{E}}[\boldsymbol{x}_i \boldsymbol{x}_i']^{-1} \hat{\mathbb{E}}[\boldsymbol{x}_i' y_i] \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i' y_i \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{y}
\end{aligned}$$

- Ordinary Least Squares (OLS) estimator

Duncan Data

Duncan {carData} R Documentation

Duncan's Occupational Prestige Data

Description

The `Duncan` data frame has 45 rows and 4 columns. Data on the prestige and other characteristics of 45 U. S. occupations in 1950.

Usage

```
Duncan
```

Format

This data frame contains the following columns:

type

Type of occupation. A factor with the following levels: `prof`, professional and managerial; `wc`, white-collar; `bc`, blue-collar.

income

Percentage of occupational incumbents in the 1950 US Census who earned $3,500 or more per year (about $36,000 in 2017 US dollars).

education

Percentage of occupational incumbents in 1950 who were high school graduates (which, were we cynical, we would say is roughly equivalent to a PhD in 2017)

prestige

Percentage of respondents in a social survey who rated the occupation as "good" or better in prestige

Source

Duncan, O. D. (1961) A socioeconomic index for all occupations. In Reiss, A. J., Jr. (Ed.) *Occupations and Social Status.* Free Press [Table VI-1].

References

Fox, J. (2016) *Applied Regression Analysis and Generalized Linear Models*, Third Edition. Sage.

Fox, J. and Weisberg, S. (2019) *An R Companion to Applied Regression*, Third Edition, Sage.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
income	45	41.867	24.435	7	21	64	81
education	45	52.556	29.761	7	26	84	100
prestige	45	47.689	31.510	3	16	81	97
resmod2	45	−0.000	20.513	−59.088	−12.557	13.795	49.858

	Prestige
	prestige
education	0.546***
	(0.098)
income	0.599***
	(0.120)
Constant	−6.065
	(4.272)
Observations	45
R^2	0.828
Adjusted R^2	0.820
Residual Std. Error	13.369 (df = 42)
F Statistic	101.216*** (df = 2; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

How to estimate β_k?

- The population regression coefficient

$$\beta_k = \frac{\mathbb{C}\text{ov}\left[y_i, \tilde{x}_{ki}\right]}{\mathbb{V}[\tilde{x}_{ki}]}$$

is estimated by

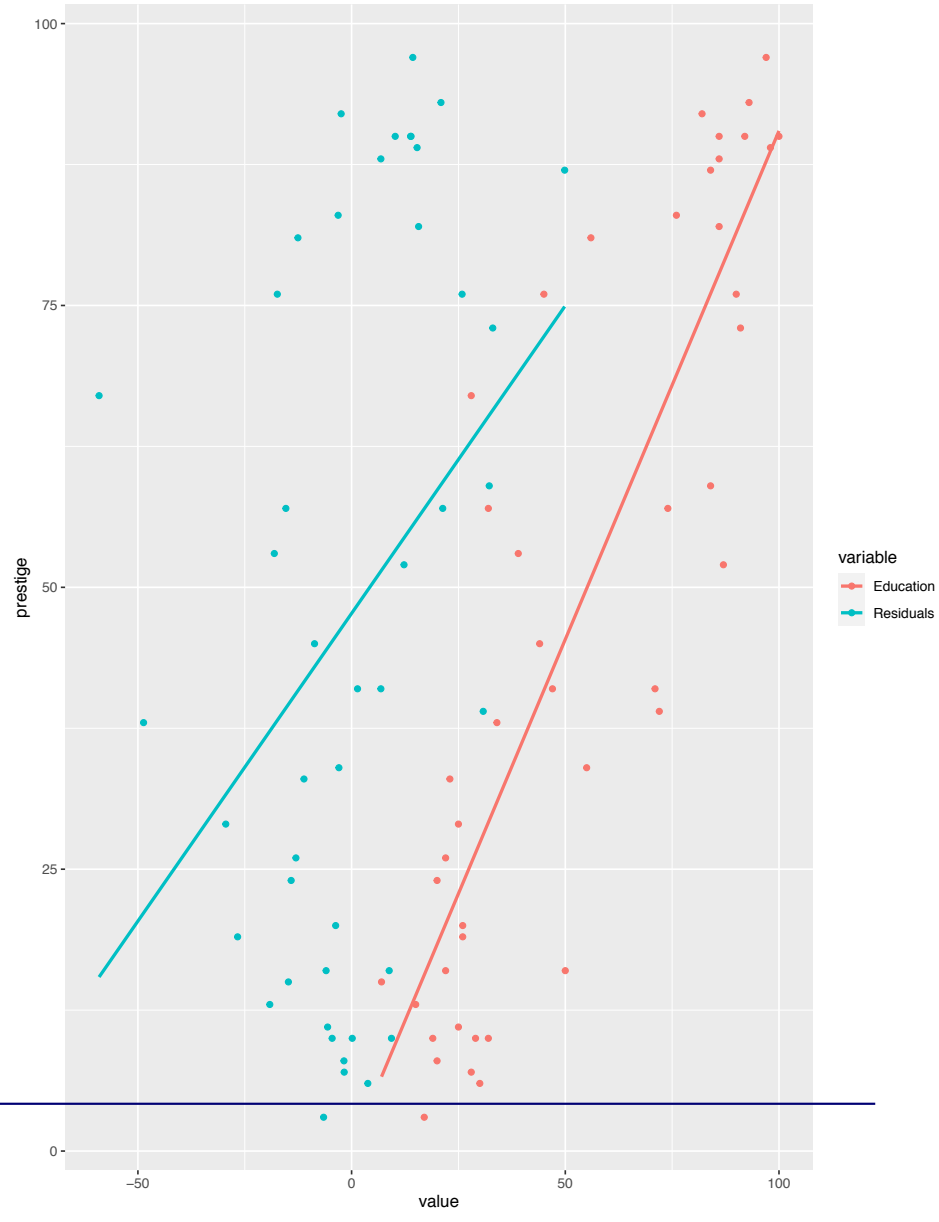
$$\hat{\beta}_k = \frac{\widehat{\mathbb{C}\text{ov}}[y_i, \tilde{x}_{ki}]}{\hat{\mathbb{V}}[\tilde{x}_{ki}]} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(\tilde{x}_{ki} - \bar{\tilde{x}}_k)}{\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_{ki} - \bar{\tilde{x}}_k)^2},$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.

Partialling out income

	prestige
	prestige
residuals from regressing education on income	0.546**
	(0.219)
Constant	47.689***
	(4.441)
Observations	45
R^2	0.126
Adjusted R^2	0.106
Residual Std. Error	29.794 (df = 43)
F Statistic	6.214** (df = 1; 43)
Note:	*p<0.1; **p<0.05; ***p<0.01

Illustration of regression coefficients



Wages

Wages {Ecdat} R Documentation

Panel Data of Individual Wages

Description

a panel of 595 observations from 1976 to 1982

number of observations : 4165

observation : individuals

country : United States

Usage

```
data(Wages)
```

Format

A dataframe containing :

exp
: years of full-time work experience

wks
: weeks worked

bluecol
: blue collar ?

ind
: works in a manufacturing industry ?

south
: resides in the south ?

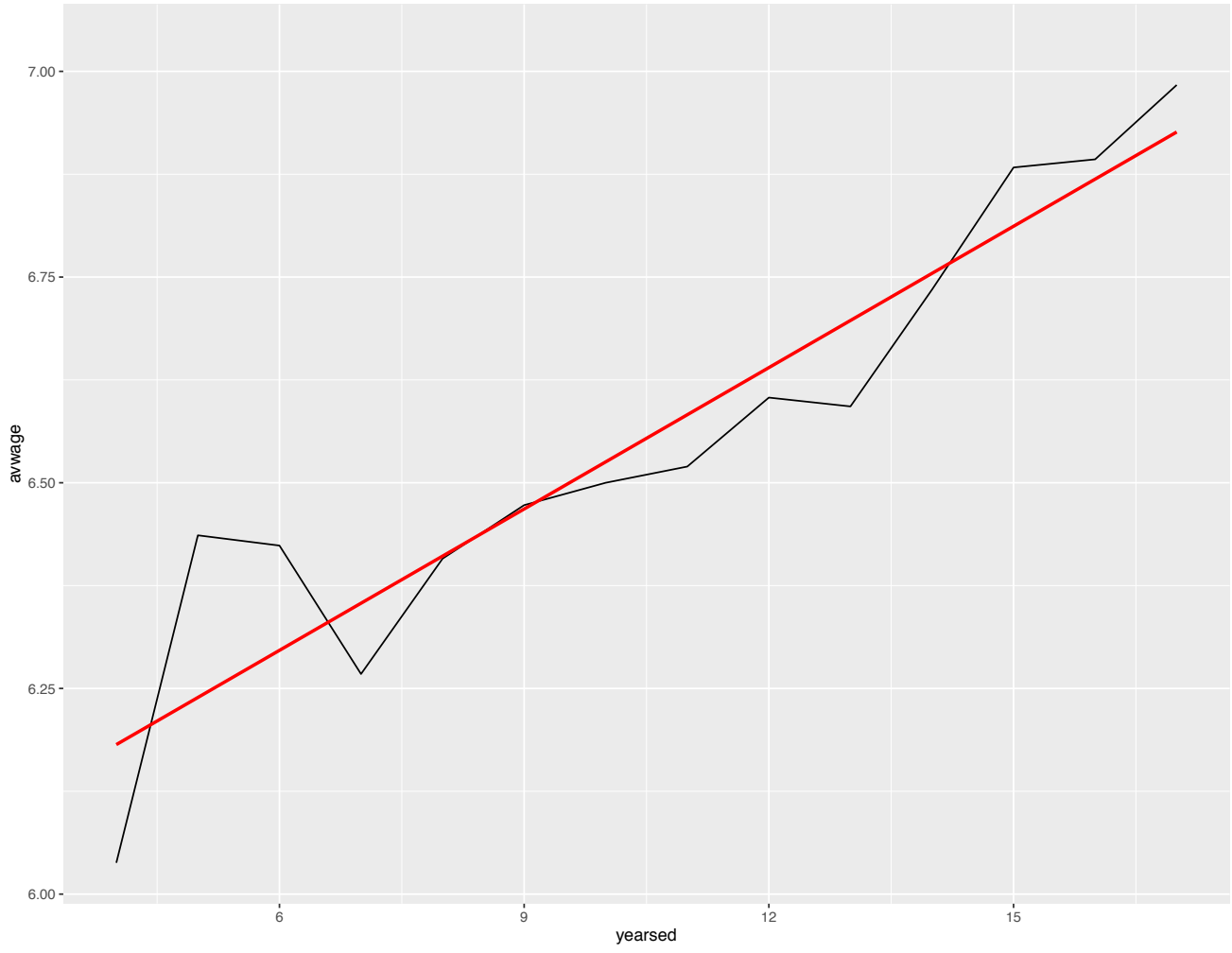
smsa
: resides in a standard metropolitan statistical are ?

married
: married ?

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
experience	4,165	19.854	10.966	1	11	29	51
weeks worked	4,165	46.812	5.129	5	46	50	52
industry	4,165	0.395	0.489	0	0	1	1
education	4,165	12.845	2.788	4	12	16	17
log wage	4,165	6.676	0.462	4.605	6.395	6.953	8.537

	Dependent variable:
	logarithmic Wage
years of education	0.065^{***}
	(0.002)
Constant	5.839^{***}
	(0.031)
Observations	4,165
R^2	0.155
Adjusted R^2	0.155
Residual Std. Error	0.424 (df = 4163)
F Statistic	764.526^{***} (df = 1; 4163)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression fit vs. conditional means



- Iterating expectations in the formula for $\boldsymbol{\beta}$ yields

$$\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i']^{-1} \mathbb{E}[\boldsymbol{x}_i y_i] = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i']^{-1} \mathbb{E}[\boldsymbol{x}_i \mathbb{E}[y_i | \boldsymbol{x}_i]].$$

- Hence, an implication of the regression-CEF theorem is that regression coefficients can be obtained by using $\mathbb{E}[y_i]\boldsymbol{x}_i]$ as dependent variable instead of y_i.
- Note that

$$\mathbb{E}[(\mathbb{E}[y_i|\boldsymbol{x}_i] - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})^2] = \sum_{u} (\mathbb{E}[y_i|\boldsymbol{x}_i = \boldsymbol{u}] - \boldsymbol{u}'\tilde{\boldsymbol{\beta}})^2 g_x(u),$$

  where $g_x(u)$ is the probability mass function.

$\Rightarrow$ $\boldsymbol{\beta}$ can be constructed from a weighted least squares regression of $\mathbb{E}[y_i|\boldsymbol{x}_i = \boldsymbol{u}]$ on $\boldsymbol{u}$, where $\boldsymbol{u}$ runs over the values of $\boldsymbol{x}_i$ and each observation is weighted based on the distribution of $\boldsymbol{x}_i$.

Non-weighted regression based on conditional means

	Dependent variable:
	Means by years of education
years of education	0.057***
	(0.006)
Constant	5.953***
	(0.071)
Observations	14
R^2	0.872
Adjusted R^2	0.862
Residual Std. Error	0.095 (df = 12)
F Statistic	81.912*** (df = 1; 12)
Note:	*p<0.1; **p<0.05; ***p<0.01

Weighted regression based on conditional means

	Dependent variable:
	Means by years of education
years of education	0.065^{***}
	(0.004)
Constant	5.839^{***}
	(0.056)
Observations	14
R^2	0.951
Adjusted R^2	0.947
Residual Std. Error	0.012 (df = 12)
F Statistic	235^{***} (df = 1; 12)
Note:	*p<0.1; **p<0.05; ***p<0.01

Outline I

Assumptions

(A1) The model is linear in the parameters and with the conditional mean specified by $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$.

(A2) The regressor matrix $\boldsymbol{X}$ is of full column rank k, i.e., $rk(\boldsymbol{X}) = k < n$ (with probability one).

(A3) The regressors $x_i^{(j)}$, $j = 1, \ldots, k$ are *strictly exogenous* with $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \boldsymbol{0}$ implying $\mathbb{E}[\varepsilon_i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = 0$ for all $i = 1, \ldots, n$.

Define the conditional covariance matrix of ε as

$$\mathbb{V}[\varepsilon|\boldsymbol{X}] = \mathbb{E}[\varepsilon\varepsilon'|\boldsymbol{X}] =: \boldsymbol{\Psi}.$$

(A4) The error terms are assumed to be uncorrelated, i.e., $\mathbb{E}[\varepsilon_i\varepsilon_j|\boldsymbol{X}] = 0$ for $i \neq j$. Then, $\boldsymbol{\Psi} = \boldsymbol{D}$ with

$$\boldsymbol{D} := diag[\sigma_1^2, \ldots, \sigma_n^2].$$

(A5) The error terms are assumed to be (conditionally) *homoscedastic*, i.e., they have equal variances across the sample, i.e., $\mathbb{V}[\varepsilon_i|\boldsymbol{X}] = \mathbb{E}[\varepsilon_i^2|\boldsymbol{X}] = \sigma^2$ for all $i = 1, \ldots, n$. Then, $\boldsymbol{D} = \sigma^2\boldsymbol{I}_n$, where $\boldsymbol{I}_n$ denoting an n-dimensional identity matrix.

Hat Matrix and Residual Maker

- The OLS residual vector is given by

$$\boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

- Vector of fitted dependent variables:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{P}\boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{P} := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is a *projection matrix*.

- Therefore:

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y},$$

  where

$$\boldsymbol{M} := \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

  is called "residual maker" or "annihilator".

- For the matrix $\boldsymbol{M}$ it can be shown that $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$, $\boldsymbol{M}' = \boldsymbol{M}$, and $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$
- Then, the relationship between $\boldsymbol{e}$ and ε is

$$\boldsymbol{e} = \boldsymbol{M}(\boldsymbol{X}\beta + \varepsilon) = \boldsymbol{M}\varepsilon,$$

Numerical Properties of OLS

- In terms of the residual vector, the F.O.C. is given by

$$\boldsymbol{X}'\boldsymbol{e} = \sum_{i=1}^{n} \boldsymbol{x}_i e_i = \mathbf{0}.$$

  - That is, each column of $\boldsymbol{X}$ is orthogonal to $\boldsymbol{e}$.

$\Rightarrow$ $\hat{\boldsymbol{y}}'\boldsymbol{e} = \mathbf{0}$

- If the regression equation has an intercept:

$$\sum_{i=1}^{n} x_{i1} \cdot e_i = \sum_{i=1}^{n} e_i = 0.$$

Frisch-Waugh Theorem

- Assume the regression model can be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \varepsilon$$

and the OLS estimator is given by
$\hat{\beta} = (\hat{\beta}_1', \hat{\beta}_2')' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

- **Theorem** (Frisch-Waugh) Define

$$\boldsymbol{M}_1 = \boldsymbol{I}_N - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1', \quad \boldsymbol{M}_2 = \boldsymbol{I}_N - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$$

as the residual makers based on regressions on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

Then, $\hat{\beta}_1$ ($\hat{\beta}_2$) is obtained as the OLS estimator in a regression of $\boldsymbol{M}_1\boldsymbol{y}$ ($\boldsymbol{M}_2\boldsymbol{y}$) on $\boldsymbol{M}_1\boldsymbol{X}_2$ ($\boldsymbol{M}_2\boldsymbol{X}_1$). Hence,

$$\hat{\beta}_1 = (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y})$$
$$\hat{\beta}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}).$$

- The Frisch-Waugh theorem proves the anatomy of multivariate regression discussed and illustrated in Section 2.1.

Mean and Covariance of OLS

- ***Proposition*** *(Unbiasedness)*: Under (A1)-(A3) , the OLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased, i.e.

$$\mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \boldsymbol{\beta} \qquad (\forall \boldsymbol{\beta}).$$

- ***Proposition*** *(Variance of the OLSE)*: Under assumptions (A1) - (A4), the conditional covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}}$ is given by

$$\boldsymbol{V}_{\hat{\beta}} := \mathbb{V}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{3}$$

If we additionally impose (A5), i.e., homoscedasticity, (3) collapses to

$$\boldsymbol{V}_{\hat{\beta}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{4}$$

Unbiased Estimation of σ^2 and $\mathbb{V}[\hat{\boldsymbol{\beta}}]$

- ***Proposition*** *(Mean of SSE=* $\sum_{i=1}^{n} e_i^2$*):*

$$\mathbb{E}[SSE|\boldsymbol{X}] = \mathbb{E}[\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{X}] = (n-k)\sigma^2.$$

- Then,

$$\hat{\sigma}^2 = \frac{SSE}{n-k} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k}$$

is an unbiased estimator of σ^2, i.e. $\mathbb{E}[\frac{SSE}{n-k}|\boldsymbol{X}] = \sigma^2$.

Estimating the OLS VC Matrix

- Under homoscedasticity, we have

$$\hat{\mathbb{V}}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

  which is an unbiased estimator of $\mathbb{V}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$,
  i.e. $\mathbb{E}[\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] = \mathbb{V}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$.
- Under heteroscedasticity, $\boldsymbol{V}_{\hat{\beta}}$ can be estimated by

$$\hat{\boldsymbol{V}}_{\hat{\beta}}^{EW} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'e_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

  Developed by Eicker (1963) and introduced in econometrics by White (1980) and is refered to as *Eicker-White robust* or *heteroscedasticity-consistent* covariance estimator

Computation of $\hat{\mathbb{V}}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$

```r
# Residuals standard error hat(sigma)
sqrt(1/(n-k) * as.numeric(t(e)%*%e)) # by hand
summary(lm(lwage ~ ed + wks + exp, data = Wages))$sigma # by package

# VC matrix under homoscedasticity
# ------------------------------------------
VCV = (1/(n-k) * as.numeric(t(e) %*% e) * solve(t(X) %*% X))
vc<-vcov(lm(lwage ~ ed + wks + exp, data = Wages))
xtable::xtable(vc, auto=TRUE)



# Standard errors of the estimated coefficients
(stder = sqrt(diag(VCV)))

# VC matrix under heteroskedasticity
# ------------------------------------------
model <- lm(lwage ~ wageedu + wks + exp, data = Wages)
vc<-vcovHC(model, type = 'HC0')
xtable::xtable(VCV, auto=TRUE)
```

Homoscedasticity:

	(Intercept)	ed	wks	exp
(Intercept)	0.0044591	-0.000074	-0.0000692	-0.0000115
ed	-0.0000740	0.000005	0.0000000	0.0000003
wks	-0.0000692	0.000000	0.0000015	0.0000000
exp	-0.0000115	0.000000	0.0000000	0.0000003

Heteroscedasticity:

	(Intercept)	ed	wks	exp
(Intercept)	0.0044591	-0.000074	-0.0000692	-0.0000115
ed	-0.0000740	0.000005	0.0000000	0.0000003
wks	-0.0000692	0.000000	0.0000015	0.0000000
exp	-0.0000115	0.000000	0.0000000	0.0000003

Breusch/Pagan/Godfrey Test

- Linear model: $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$
- Heteroscedastic error terms:

$$\mathbb{V}[\varepsilon_i] := \sigma_i^2 = h(\gamma_0 + \boldsymbol{z}_i'\boldsymbol{\gamma}),$$

  - $h(\cdot) > 0$ denotes an unknown function,
  - $\boldsymbol{z}_i$ denotes a $(p \times 1)$ vector of (known!) variance regressors.

- Null hypothesis: $H_0 : \boldsymbol{\gamma} = \boldsymbol{0}$.

LM-Testing Procedure

(i) Estimate the model under the null hypothesis.

(ii) Compute the centered R^2 of the auxiliary regression:

$$e_i^2/\hat{\sigma}^2 = \delta_0 + \delta_1 z_{i1} + \ldots + \delta_p z_{ip} + \xi_i,$$

where e_i is the OLS residual of step (i).

(iii) Compute $LM_{BPG} = nR^2$.

White Test

- Null hypothesis: $\sigma_i^2 = \sigma^2$.
- Idea: Comparing CV matrices $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

(i) Estimate the model under the null hypothesis.

(ii) Estimate the following auxiliary regression by OLS:

$$e_i^2 = \delta_0 + \boldsymbol{\delta}'\boldsymbol{z}_i,$$

where $\boldsymbol{z}_i$ consists of the regressors, their squares and all cross terms.

(iii) Then, the LM statistic is computed based on the R^2 of the auxiliary regression

$$n \cdot R^2 \underset{H_0}{\overset{a}{\sim}} \chi^2_{(q)},$$

where the degrees of freedom q are the number of auxiliary regressors in (ii) excluding the intercept.

```
> bptest(model, data = Wages)

studentized Breusch-Pagan test

data:  model
BP = 28.791, df = 3, p-value = 2.478e-06

> # White test
> bptest(model, ~wageedu*wks+wageedu*exp+wks*exp+I(wageedu)^2+I(exp)^2
+I(wks)^2, data = Wages)

studentized Breusch-Pagan test

data:  model
BP = 57.347, df = 6, p-value = 1.554e-10
```

Efficiency of OLS

- ***Theorem*** *(Gauss-Markov)*: Under the full ideal conditions $(i) - (v)$, the least squares estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, i.e.

$$\mathbb{V}[\tilde{\boldsymbol{\beta}}|\boldsymbol{X}] - \mathbb{V}[\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}] = \sigma^2 \boldsymbol{D}'\boldsymbol{D} \quad \text{pos. semi-def.},$$

  where $\tilde{\boldsymbol{\beta}}$ is any linear unbiased estimator of $\boldsymbol{\beta}$ and $\boldsymbol{D}$ is given by

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{D}'\boldsymbol{y}.$$

Statistical Properties under Normality

- ***Proposition*** *(The LSE under Normality)*: Under the full ideal conditions $(i) - (vi)$ the distribution of $\hat{\boldsymbol{\beta}}$ is given by:

$$\hat{\boldsymbol{\beta}} | \boldsymbol{X} \sim N(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

- ***Theorem***: For the normally distributed linear regression model, the distribution of the statistic $\boldsymbol{e}'\boldsymbol{e}/\sigma^2$ is

$$\frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} | \boldsymbol{X} \sim \chi^2_{(n-k)}.$$

- ***Proposition*** *(The variance of $\hat{\sigma}^2$)*: In a normally distributed regression model, the variance of $\hat{\sigma}^2$ is given by:

$$\mathbb{V}[\hat{\sigma}^2|\boldsymbol{X}] = \frac{2\sigma^4}{n-k}.$$

- ***Proposition*** *(Independence of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$)*: In the normally distributed linear regression model, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent random variables.

Goodness-of-Fit

- Coefficient of determination:

$$R^2 := \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\mathbb{V}}[\hat{y}_i]}{\hat{\mathbb{V}}[y_i]}.$$

- If the model contains an intercept term, then

$$\begin{aligned} \sum_i (y_i - \bar{y})^2 &= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i^2 \\ \text{SST} &= \text{SSR} + \text{SSE} \end{aligned}$$

- Then:

$$R^2 := 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\hat{\mathbb{V}}[e_i]}{\hat{\mathbb{V}}[y_i]}.$$

Properties of R^2

- $R^2 = \rho^2_{y,\hat{y}}$, where ρ denotes the empirical correlation.
- If the model has an intercept, $0 \leq R^2 \leq 1$.
- The R^2 follows an unknown distribution.
- Adding further variables leads to an increase in the R^2.
- An R^2 can have a reasonable size in spurious regressions if the regressors are non-stationary.
- Linear transformations of the regression model do not change the value of the R^2 coefficient.

Alternative R^2 Concepts

- Non-centered R^2:

$$\tilde{R}^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n y_i^2}.$$

- General definition of an R^2:

$$R^2 = \mathbb{C}\text{or}[y_i, \hat{y}_i]^2.$$

- Adjusted R^2:

$$\bar{R}^2 = 1 - \frac{(n-1)\sum_{i=1}^n e_i^2}{(n-k)\sum_{i=1}(y_i - \bar{y})^2},$$

where k denotes the number of regressors in the model.

Multicollinearity

Lemma(Variance of $\hat{\beta}_j$) Under assumptions (A1) to (A5), $\mathbb{V}[\hat{\beta}_j|\boldsymbol{X}]$ can be written as

$$\mathbb{V}[\hat{\beta}_j|\boldsymbol{X}] = \frac{\sigma^2}{S_{x^{(j)}x^{(j)}}} \frac{1}{1 - R_j^2},$$

where $S_{x^{(j)}x^{(j)}} := \sum_{i=1}^{n} (x_i^{(j)} - \bar{x}^{(j)})^2$ denotes the 'variation' of regressor $x^{(j)}$ and R_j^2 is the R^2 of an (auxiliary) regression of $x^{(j)}$ on all $x^{(i)}$, $(i \neq j)$.

Indicators for Multicollinearity

- Variance Inflation Factor:

$$VIF_j = \frac{1}{1 - R_j^2}$$

- Rule of thumb: $VIF_j > 10$ implies serious multicollinearity
- No problem if $\max_j(VIF_j) \leq 5$.
- Low values of $det(\boldsymbol{X}'\boldsymbol{X})$ indicate multicollinearity.
- If $\boldsymbol{X}$ if in correlation form: $0 \leq det(\boldsymbol{X}'\boldsymbol{X}) \leq 1$.