Erasmus
School of
Economics

Statistics for Data Science
Lecture 4

Dennis Fok (Econometric Institute)

September – October, 2025

Erasmus University Rotterdam

Plan for Lecture 4

1. Last week's assignment
2. OLS regression
   - Univariate regression
   - Multivariate regression
   - Regression with interactions
3. Regression diagnostics: beginner

Before next time

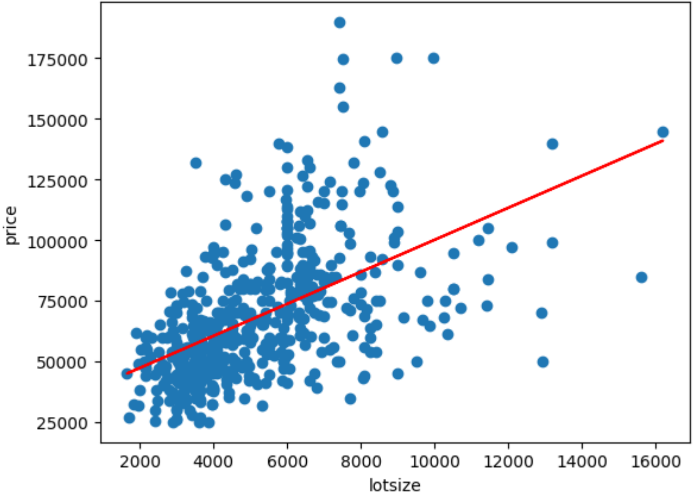
Assignment for next week

- Reread Chapter 2 & 3 (if needed)
- Read Chapter 4 (main material for next week)
- Reconsider/finish the in-class assignments
- Examples in book
- Small new programming assignment
  - Visualize the correlation between some (continuous) variables in the houseprice data using a scatter plot
  - Calculate the correlation
  - Perform a hypothesis test on this correlation (clearly formulate the hypotheses and the conclusion)
- Work on final assignment

Univariate regression

What can regression do?

- Data science is about exploring dependence across (multiple) variables
- The simplest model for dependence: linear relation (strong link with correlation)



The setup of a regression

Can see "regression line" as

- Predicted value of price (y) at certain value of lotsize (x)
- A fitted model that links y to x

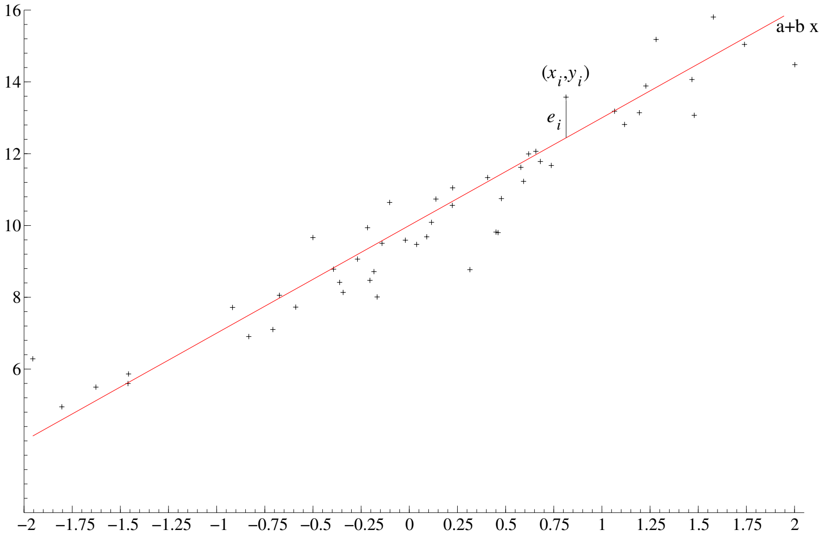
Mathematically,

$$y_i = a + bx_i + e_i$$

where

- y_i: dependent variable (for observation i)
- x_i: explanatory variable (for observation i)
- a and b: estimated coefficients (apply to all observations)
- e_i: residual, or prediction error (for observation i)

Graphical interpretation



Ordinary Least Squares [OLS]

How to find (estimate) a and b given data?

$$y_i = a + bx_i + e_i$$

Idea: Small values of e_i (close to zero) are preferred

$\rightarrow$ Minimize sum of squared e_i (=OLS)

$$\min_{a,b} S(a,b) = \sum_i e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Calculating the first derivatives and setting these to zero yields:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and}$$

$$a = \bar{y} - b\bar{x}$$

(Statistical) properties

Properties OLS:

+ Easy calculation
+ Well-known statistical properties
+ Optimal under some assumptions
− Sensitive to outliers
− Not optimal if assumptions are *not* true

Question

How to judge whether OLS is a good method?

Difficult! $\rightarrow$ Answer depends on the "true" relationship between y and x

To analyze properties of OLS we need to

- define the true (unknown) relationship
  (also known as the data generating process [DGP])

Is OLS a good method? 6 assumptions to answer this

The "true" relationship between y and x (data generating process [DGP])

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where α and β are unknown & ε_i is "pure" random variation (note the Greek letters)

Formal assumptions:

- A1 *Non-degeneration:* x_i are fixed (non-random) with $\sum(x_i - \bar{x})^2 > 0$
- A2 *Mean zero:* ε_i are random with $\mathrm{E}[\varepsilon_i] = 0$
- A3 *Linearity:* $y_i = \alpha + \beta x_i + \varepsilon_i$ holds exactly
- A4 *Homoskedasticity:* $\mathrm{E}[\varepsilon_i^2] = \mathrm{Var}(\varepsilon_i) = \sigma^2$
- A5 *No autocorrelation:* $\mathrm{E}[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$
- A6 *Normality:* $\varepsilon_i \sim N(0, \sigma^2)$

Gauss-Markov theorem:

Under these assumptions one can show that OLS is "best" (= smallest uncertainty)

$\rightarrow$ not all assumptions are really necessary

Summary

Linear regression (OLS):

- strong method
- often used
- building block for further analysis

Interpretation of coefficients:

Given the model $y = \alpha + \beta x + \varepsilon$

- α: Expected value of y if $x = 0$ (not always useful)
- β: Increase in expected value if x increases by 1

Linear model using Python

Packages

- `import statsmodels.api as sm`
- `import statsmodels.formula.api as smf`

Main function: `smf.ols()`

- `smf.ols(formula="y ~ x", data = yourframe)`: linear model with y explained by x (and a constant)
- Give the model a name, eg.: `m = smf.ols(formula="y ~ x", data = yourframe)`
- Estimate the parameters `res = m.fit()` and store the result

Useful functions using `res`, the result of `.fit()`

- `res.summary()`: give a summary of the results
- `sm.graphics.abline_plot(model_results=res, color='red', ax=plt.gca())`: add a fitted (straight) line to an existing plot

Other properties and methods (will be useful later)

- `res.params`: give estimated coefficients
- `res.conf_int(alpha=..)`: provide confidence intervals
- `res.fittedvalues`: given in-sample fitted values
- `res.predict(exog={'x': [1,2,3]})`: give predicted values for new data

Assignment

In-class assignment 4.1

- Use housing data
- Explain price using lotsize using a linear model
- Reproduce scatter with fitted linear line
- Interpret the results of your final model

You will also need

- `import pandas as pd`
- `import matplotlib.pyplot as plt`

Testing and model evaluation

Evaluate goodness of fit

For a *good* model:

- All scatter points are close to the line
- All residuals $e_i = y_i - a - bx_i$ are close to zero
- Fit is related to sum squared errors = SSE = $\sum_i e_i^2$

Goodness of fit:

- Relate SSE to "scale of data" = Total sum of squares = SST or SSY

$$\text{SSY} = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n\bar{y}^2$$

- Goodness of fit: R^2

$$R^2 = 1 - \frac{SSE}{SSY}$$

R^2

Alternative definition of R^2

$$R^2 = \frac{SSR}{SSY}$$

where SSR= "regression sum of squares" $= \sum_i (\hat{y}_i - \bar{\hat{y}})^2$

Interpretation

- R^2 is squared correlation between x and y
- R^2: proportion of variation explained
- $R^2 = 0$: nothing explained
- $R^2 = 1$: everything explained

`res.summary()` gives R^2 as standard output
(res is the result from `smf.ols(...).fit()`)

Estimating the variance of ε_i

Question: Which part of y can never be explained?

$\rightarrow$ The error term:

$$\varepsilon_i = y_i - \alpha - \beta x_i$$

Denote the variance of ε_i by σ^2

Estimation

- Estimate ε_i by $e_i = y_i - a - bx_i$
- Estimate σ^2 by s^2

$$s^2 = \frac{\sum_i e_i^2}{n-k} = \frac{SSE}{n-k} \quad \text{(here: } k = 2)$$

- In general: k=number of parameters

Estimation uncertainty

Note:

- We estimate α and β (with a and b)
- There is estimation uncertainty!
- How large is this?
- Does x have a *significant* impact?
  $\rightarrow$ Can we reject $H_0 : \beta = 0$?

Quantifying the uncertainty

- Recall: a and b are a function of y (and x) $\rightarrow$ a random variable
- In fact a linear function of y $\rightarrow$ can easily work out Var[a] and Var[b]

$$\text{Var}[a] = \frac{\sigma^2 \bar{x}^2}{SSX}$$

$$\text{Var}[b] = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{SSX}$$

Variance/standard error of b

Estimate σ^2 by s^2:

- Estimated variance of b: s^2/SSX
- Estimated standard deviation = standard error of b = $s/\sqrt{SSX}$

Small standard errors if

- small σ^2 (find a good fitting model)
- large SSX:
  - many observations
  - large spread in x

Hypothesis testing

Interesting hypothesis

$$H_0 : \beta = 0$$

(note: formulated in terms of β, not b)

Using standard error, we can formulate a t-test (as before)

$$\text{t-stat}_b = \frac{b}{SE_b} \sim t_{n-k}$$

```python
res = smf.ols(formula="y ~ x", data=df)
res.summary()
```

Distribution is really t_{n-k} if:
$\rightarrow$ all 6 assumptions are satisfied!

Non-linear models

Non-linearity

The basic model specifies

$$y_i = \alpha + \beta x_i + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2)$$

$\rightarrow$ Linear relation between x and y

Alternatives:

- Also use transformations of x as explanatory variable
  - x^2
  - $\log(x)$
  - $\sqrt{x}$
  - $\frac{1}{x}$
  - etc.
  - (can also use multiple transformations at the same time)
- Transformations of y
  - $\log(y)$ (most often used)
  - Simply write eg. `np.log(y)` inside the formula (`using numpy as np`)

Note: resulting models are still linear in the "econometric sense"

Interpretation in most commonly used models

If

$$y_i = \alpha + \beta \log(x_i) + \varepsilon_i$$

$\rightarrow$ increase x by 1% $\implies$ y increases by $\beta \log(1.01) \approx \beta/100$ units

If

$$\log(y_i) = \alpha + \beta \log(x_i) + \varepsilon_i$$
$$y_i = \exp(\alpha + \beta \log(x_i) + \varepsilon_i)$$

$\rightarrow$ increase x by 1% $\implies$ y increases by β% (elasticity)

If

$$\log(y_i) = \alpha + \beta x_i + \varepsilon_i$$
$$y_i = \exp(\alpha + \beta x_i + \varepsilon_i)$$

$\rightarrow$ increase x by 1 unit $\implies$ y increases by $100(\exp(\beta) - 1)\% \approx 100\beta\%$

Assignment

In-class assignment 4.2

- Consider the earlier regression model
- What is the R^2? Does this model fit well?
- Use the output to perform a hypothesis test for no impact of lotsize
- Compare this result to the result of

```python
from scipy import stats
stats.pearsonr(x,y)
```

- Also try a model for log(price) explained by log(lotsize).
- How should the parameters in this model be interpreted?

Other advanced methods

Many more techniques are available

- Estimate a truly non-linear model $y_i = \alpha + x_i^{\beta} + \varepsilon_i$
- Estimate with unknown/flexible functional form
  - Non-parametric estimation
  - General Additive Models
  - ...
- Estimate with multiple explanatory variables (next topic)
- Estimate with other types of dependent variables (later)

Multiple regression

Multiple explanatory variables

Why only 1 explanatory variable?

- Multiple factors influence y
- These factors are often related!
- What is the true influence?

Important questions

- What do parameters mean in a model with multiple x?
- What about interactions?
- Which variables to include? (later topic)

Econometrics of multiple regression

Consider the model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

(note there is no x_{i1})

New *additional* assumptions:

- All variables show variation
- No *perfect* linear relations between variables

Short-hand notation

If we introduce $x_{i1} = 1$, we can write

$$y_i = \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

or with matrix/vector notation

$$y_i = (x_{i1}, x_{i2}, \ldots, x_{ik}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \varepsilon_i$$

which we summarize as

$$y_i = x_i' \beta + \varepsilon_i$$

(x_i and β are both column vectors)

Grouping all observations

Next we collect all observations $i = 1, \dots, n$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

which we summarize as

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

$\rightarrow$ extremely general notation!

Estimating β

OLS can still be used to estimate β

Define $e_i = y_i - x_i'b$ and minimize

$$\text{SSE} = \sum_i e_i^2 = \begin{pmatrix} e_1 & e_2 & \dots & e_n \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = e'e = (y - Xb)'(y - Xb)$$

with $e = y - Xb$.

Can show that the solution is

$$b = (X'X)^{-1}X'y$$

$\rightarrow$ Most important formula in econometrics

Estimation uncertainty and Goodness of fit

For multiple regression $y = X\beta + \varepsilon$

- $b = (X'X)^{-1}X'y$ is an estimator of β
- Possible to estimate the variance of b: SE_{b_j} $j = 1, 2, \cdots, k$

Test hypothesis $H_0 : \beta_j = 0$ for a given j

- Test statistic

$$\text{t-stat}_{b_j} = \frac{b_j}{SE_{b_j}} \sim t_{n-k}$$

Test hypothesis $H_0 : \beta_2 = \cdots = \beta_k = 0$

- Apart from the "constant", no variable in x explains the variation of y
- Test statistic

$$F = \frac{RegressionSS/(k-1)}{ErrorSS/(n-k)} = \frac{SSR/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k)$$

- If F is large, then reject the null

Multiple regression in R

Executing multiple regression is easy

Examples

- `formula=y ~ x2 + x3` inside the smf.ols() method
- a constant is always added automatically

→ Next use same functions as before

Goodness of fit

For multiple regression

- R^2 same as before
- However: adding variables $\rightarrow$ guaranteed increase in R^2 (Q: why?)
- Adjusted R^2

$$\text{Adj}R^2 = 1 - \frac{SSE/(n-k)}{SSY/(n-1)}$$

  includes penalty on additional variables
- Information criteria, eg. AIC
  - Balances fit vs. no. parameters
  - Lower numbers are better
  - Also counts variance as parameter
  - `res = model.fit()` and `res.aic`

Parameter interpretation – ceteris paribus

Suppose an estimated model is

$$\log(\text{income}) = 7 + 0.01\text{age} + 0.025\text{educ} \; + e$$

with educ: number of years of education

How to interpret the coefficients?

- if age=educ=0 $\rightarrow$ log(income)=7 $\rightarrow$ income $\approx$ 1069 (does this mean anything?)
- if age=30, educ=12 $\rightarrow$ log(income)=7.6 $\rightarrow$ income $\approx$ 2000
- if age +1 $\rightarrow$ income +1% (holding educ constant!)
- if educ +1 $\rightarrow$ income +2.5% (holding age constant!)

Important: all results are ceteris paribus! (keeping other things fixed)

Interactions

Regression with interaction

Consider the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

What's the point to add the interaction term $(x_1 x_2)$?

- Interaction effect: there is a "synergy" (or "anti–synergy") regarding the impact of x_1 and x_2 on y
- Moderation effect: the impact of x_1 on y depends on x_2 (or the other way around)

Typical mistakes in interpreting interaction regressions

- β_1 (or its estimate b_1) is not the impact of x_1 on y!
  - An insignificant b_1 does not necessarily mean y_1 and x are not related!
  - A significant b_1 does not necessarily mean x_1 and y are related either!

$\rightarrow$ A significant b_3 does mean that there is an interaction effect!

How to interpret regression with interaction?

Rewrite the model!

$$y = \alpha + \beta_2 x_2 + (\beta_1 + \beta_3 x_2) x_1 + \varepsilon$$

- The intercept: $\alpha + \beta_2 x_2$
- The slope for x_1: $\beta_1 + \beta_3 x_2$
- Interpret in the context
  - choose a value for x_2
  - calculate impact of x_1 at that value of x_2

  (or plot impact as a function of x_2)

$\rightarrow$ Can of course also swap roles of x_1 and x_2

Interaction regression in Python

The model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Estimate the model

- Easy, as if run a three variables regression
- Even easier, you do not have to construct $x_1 x_2$, Python does it for you
  `.ols(formula = y ~ x1 + x2 + x1:x2, data=..)`
- A more convenient way: `.ols(formula = y ~ x1*x2, data=..)`: It means "all individual and interaction effects based on x1 and x2"

Visualizing the interaction effects (useful for interpretation, but requires some work):

1. Create predictions varying one of the variables, keeping the other(s) fixed
2. Repeat for various values of "the other(s)"
3. Create plot

Assignment

In-class Assignment 4.3

- Use the Murder rate data (state.x77)
  This standard R data file is available on Canvas as csv file.

```python
import pandas as pd
statex77 = pd.read_csv("statex77.csv")
```

- Explain Murder rate by Income, Population
- Interpret the coefficients
- **Optional**
  - Add the interaction effect between Income and Population
  - Plot the interaction effect
  - Interpret the interaction
  - Question after the exercise *Do you have a story behind the result?*

Preparing for diagnostics: Normality testing

Normality tests

Many models/tests rely on normality of error terms (not the y or x variable!)

Can we test whether a variable is normally distributed?

- Yes, if variables are identically normal distributed under H_0
- Not directly, if mean of variable depends on stuff that is not normally distributed (not iid)

Many tests exist, for example

- Shapiro-Wilk test
  `scipy.stats.shapiro(x)`
  based on so-called order-statistics (*smallest*, *next-to-smallest*,..., *largest* observation)
- Jarque-Bera test
  `scipy.stats.jarque_bera(x)`
  based on skewness and kurtosis
- ...

Graphical procedures: QQ-plots

QQ plots - more formal description

The idea of empirical distributions can be used to test for particular distributions.
$\rightarrow$ Main idea: Compare estimated cdf versus theoretical cdf

Given n observations:

- If data is really normal: what would you expect the smallest observation to be?
- and the next-to-smallest?
- ...

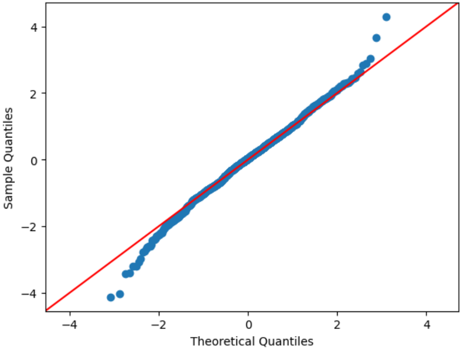
QQ plot

- Plot the observed quantiles vs. theoretical quantiles
- Should be nice straight line (intercept and slope depend on mean and variance, or are a least squares fit)

Example QQ plot

Options

- Using scipy
  `fit = scipy.stats.fit(stats.norm, data, bounds)`
  `fit.plot(plottype='qq')`
- Using `statsmodels.api.qqplot(x, line="45")`



→ Example: Data clearly not normal!

Model diagnostics

The assumptions

The Model (data generating process [DGP])

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

Formal assumptions (omitting A1 and A2):

A3 *Linearity:* $y_i = x_i\beta + \varepsilon_i$ holds exactly

A4 *Homoskedasticity:* $\mathrm{E}[\varepsilon_i^2] = \mathrm{Var}(\varepsilon_i) = \sigma^2$

A5 *No autocorrelation:* $\mathrm{E}[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$

A6 *Normality:* $\varepsilon_i \sim N(0, \sigma^2)$

Additional assumption in multivariate regression

A7 No perfect linear relationship in X

Model diagnosis: two key questions

- Are these assumptions valid?
- If an assumption fails, what to do?

Simple diagnosis in Python

- Fit the model:
  `model = smf.ols(formula=..., data=...).fit()`
- Download `olsdiagnostics.py` from Canvas into working folder and
  `from olsdiagnostics import *`
- Create OLSInfluence object (`from statsmodels.stats.outliers_influence`)
  `influence = OLSInfluence(model)`
- Run `diagnosticplots(influence)`
  → creates four plots

Checking assumptions using `olsdiagnostics`

4 diagnostic plots with: `olsdiagnostics`

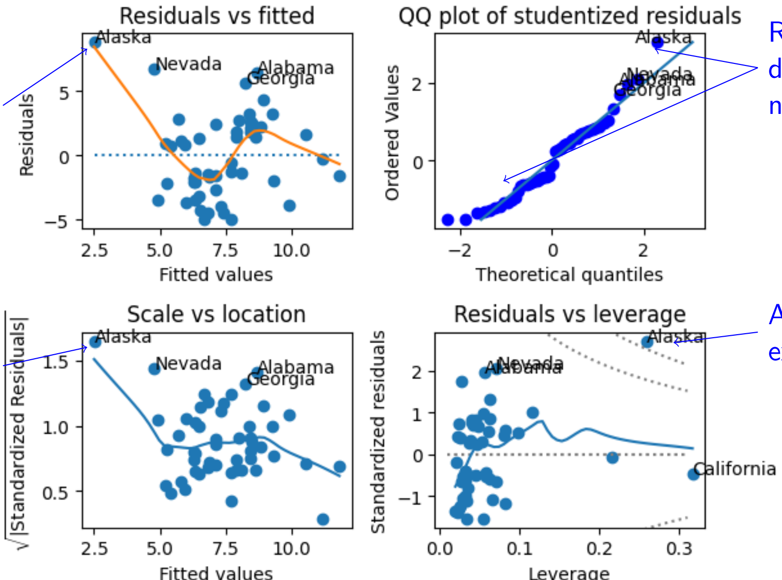
1. Plot of residuals vs. fitted values `residfitted(influence)`
   → Check for structure in mean of residuals (Linearity [A3])
   → Check for structure in absolute value of residuals (Heteroskedasticity [A4])
2. QQ plot of studentized residuals: Normality [A6] `qqresid(influence)`
   → Check for deviations from normality
3. Plot of sqrt(abs(stand. residual)) vs fitted: Heteroskedasticity [A4]
   `scalelocation(influence)`
   → Check whether magnitude of residuals depends on fitted value
4. Leverage (high if "extreme in terms of x") vs. standardized residual: Outliers
   `residleverage(influence)`
   → Does not correspond to one assumption, and is not very useful for outlier detection

(One missing assumption [A5])

Illustration on formula ="Murder ∼ Population + Income"

Some structure in the residuals (due to Alaska)

No sign of heteroskedasticity (only Alaska deviates)



Residuals do not look normal

Alaska is extreme

Take home assignment

- Use the Murder rate data (Murder as dependent variable)
- Start with four independent variables: Income, Population, Illiteracy, Frost
- Do some experimentation
  - If a variable is not significant, try to remove it
    - Does the R^2 go up or go down? What about Adjusted R^2?
    - What about AIC?
- Ultimate goal: find the best model (the lowest AIC)
- Finally: check the model assumptions using the diagnostics plot
  $\rightarrow$ What do you conclude?

Before next time

- Reread Chapter 4
- No new material for next week
- Reconsider/finish the in-class assignments
- Work on the take home assignment
- Final assignment (part 1 is due on Sunday)