Regression

Master Data Science
Summer 2023

Prof. Dr. Marlene Müller
marlene.mueller@bht-berlin.de

Exercises 3

Exercise 1

Use the Munich appartments dataset (`MunichRent2003.csv`) to estimate two different linear models with the (`netrent`) as dependent variable.

(a) Determine (with R) the predicted values $\widehat{y}_i$ and the residuals $\widehat{\varepsilon}_i$. Display them in a graph by plotting $\widehat{y}_i$ against $\widehat{\varepsilon}_i$.

(b) Check with R that the following properties hold:

$$\overline{y} = \overline{\widehat{y}} \ = \ \frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i \quad \text{and} \quad \overline{\widehat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i \ = \ 0\,.$$

i.e., the mean of the $\widehat{y}_i$'s equals $\overline{y}$ and the residuals have mean 0.

(c) Prove the equations in (b) for the simple linear regression case.

Exercise 2

Generate artificial regression data using the true model:

$$Y = 1 + 2X + X^2 + \varepsilon$$

where $X \sim N(0,1)$ and $\varepsilon \sim N(0, 0.25)$. Choose a sufficiently large sample size, e.g. $n = 500$ and estimate the following regression models:

- simple linear regressin, i.e. regress Y on X,
- quadratic regression, i.e. regress Y on X and X^2,
- cubic regression, i.e. regress Y on X, X^2, and X^3.

(a) Determine the coefficients of determination (R^2) of your three regressions. Do the same for the RSS values. (See tables on the last page of this exercise sheet.)

(b) Analyse and compare the residuals for all 3 model fits (e.g. by residual plots or by boxplots). What do you conclude with respect to which of the 3 models seems to be appropriate?

Exercise 3

Use the data generating process and the models of Exercise 3 again, but now with a small sample size (e.g. $n = 10$).

(a) Construct (in R) the matrix $\mathcal{X}$ for all 3 models (the matrix $\mathcal{X}$ is called “design matrix”).

(b) Check, if $\mathcal{X}$ and $\mathcal{X}^\top\mathcal{X}$ are of full rank. (Recall: What is the rank of a matrix? How could you determine this value using R?)

(c) Do also calculate the following matrix (the “hat matrix”) for each of the 3 models:

$$\mathbf{P} = \mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top .$$

Determine (for each model) the trace and the eigenvalues of $\mathbf{P}$. (Do you remember the relation between them?)

(d) By $\mathbf{I}$ we denote the identity matrix (a matrix with diagonal elements 1 and 0 otherwise). Show with the help of R and without R that it holds:

$$\mathbf{P}^2 = \mathbf{P} \quad \text{and} \quad (\mathbf{I} - \mathbf{P})^2 = \mathbf{I} - \mathbf{P}$$

Exercise 4

We use the CPS1985 dataset which is included in an R package:

```
require(AER)
data(CPS1985)
```

Check the contents of the dataset (`?CPS1985`) and consider the variables $Y =$ `wage`, $X_1 =$ `education` and $X_2 =$ `experience`. Estimate the following models:

- regression of $\log(Y)$ on X_1,
- regression of $\log(Y)$ on X_1 and X_2,
- regression of $\log(Y)$ on X_1, X_2 and X_2^2.

By $\log$ we denote the natural logarithm (denoted by $\ln$ in mathematics, the function is `log` in R as in statistics we merely use this logarithm).

(a) Interprete the estimated coefficients. Do they make sense? (Plot a graph for the quadratic part of the 3rd model.)

(b) Analyse and compare the residuals for all 3 model fits (e.g. by comparing R^2 and by residual plots or by boxplots). What do you conclude with respect to which of the 3 models seems to be appropriate?

Components of a Linear Model Estimated in R

The following R functions can be used to extract components from a linear model.
Example: `model <- lm(y ~ x); summary(model)`

Function	Description
`summary`	summary output (see also `?summary.lm`)
`coef`	estimated coefficients
`residuals`	residuals $\widehat{\varepsilon}_i$
`fitted`	predicted values $\widehat{y}_i$
`predict`	predicted values, useful for new data (see also `?predict.lm`)
`anova`	test for comparing two nested models
`plot`	some diagnostic plots
`confint`	confidence intervals for the coefficients
`deviance`	residual sum of squares RSS
`vcov`	estimated covariance matrix (of the coefficients)
`logLik`	log-likelihood (under normality assumption)
`AIC`	Akaike's information criterion (for model choice)

Further terms can be extracted from `summary`. Example: `summary(model)$call`

Function	Description
`call`	call of `lm`
`terms`	information on the explanatory variables
`residuals`	residuals $\widehat{\varepsilon}_i$
`coefficients`	table of coefficients, standard errors, t values and p values
`sigma`	estimated standard deviation $\widehat{\sigma}$
`df`	degrees of freedom
`r.squared`	coefficient of determination R^2
`adj.r.squared`	adjusted coefficient of determination
`fstatistic`	F statistic with acc. degress of freedom
`cov.unscaled`	unscaled covariance matrix (results in `vcov` when multiplied with $\widehat{\sigma}^2$)