Exercises for the lecture on

Statistical Methods

Prof. Dr. Vladimir Spokoiny
Vladimir.Spokoiny@wias-berlin.de
Dr. Thorsten Dickhaus
Thorsten.Dickhaus@wias-berlin.de
www.wias-berlin.de/people/dickhaus/

Humboldt-University Berlin
Department of Mathematics
Winter term 2013 / 2014

Sheet 7

Solutions are due on Monday, December 2nd, 2013, 3:15pm.
Every completely and correctly solved exercise gives 4 points.

Exercises

25. **Residuals and noise variance in multiple linear regression.**

    (a) Prove Theorem 3.24.(a).

    (b) Show that, under the assumptions of Theorem 3.24.(a), the following two assertions hold true.

        (i) For the distribution of the sum of squares of residuals, we obtain

        $$(n-p)\,\frac{\widehat{\sigma^2}}{\sigma^2} = \frac{\hat{\varepsilon}^t\hat{\varepsilon}}{\sigma^2} \sim \chi^2_{(n-p)}.$$

        (ii) The sum of squares of residuals $\hat{\varepsilon}^t\hat{\varepsilon}$ and the least squares estimator $\hat{\beta}$ are stochastically independent.

    (c) Show that, under the assumptions of Theorem 3.24.(b), the matrix $(I_n - H)$ is symmetric and idempotent.

26. **Dummy-coding of categorical covariates.** A scientific study comprising $n = 12$ participants was carried out to investigate how smoking affects food habits. To this end, the smoking status and the intake of anti-oxidative vitamins (response variable Y, measured on a standardized, continuous scale) was assessed for all study participants. The smoking status was represented by four indicator variables, corresponding to "never smoked" (covariate X_1), "ex-smoker" (covariate X_2), "mild smoker" (covariate X_3) and "strong smoker" (covariate X_4). Two study participants never smoked, four study participants were ex-smokers, and three study participants were mild and strong smokers, respectively. Three different statistical software systems and a method taken from a textbook were employed to fit a multiple linear regression model of the form

    $$\mathbb{E}[Y|\vec{X} = \vec{x}] = \beta_0 + \sum_{j=1}^{4} \beta_j x_j,\ \vec{X} = (X_1, X_2, X_3, X_4)^\top,$$

    yielding the following four sets of parameter estimates.

	GENSTAT	SAS	SPSS	Textbook
$\hat{\beta}_0$	1.8	0.3	1.025	1.0
$\hat{\beta}_1$	0	1.5	0.775	0.8
$\hat{\beta}_2$	-0.3	1.2	0.475	0.5
$\hat{\beta}_3$	-1.3	0.2	-0.525	-0.5
$\hat{\beta}_4$	-1.5	0	-0.725	-0.7

(a) Check that all four given sets of parameter estimates result in exactly the same fitted response values $(\hat{y}_i)_{1\le i\le n}$. Compute the average vitamin level for each of the four smoking classes separately.

(b) Assume that a further software package would estimate the regression coefficients under the constraint $\hat{\beta}_0 = 0$ (no intercept). What would be the resulting values for $(\hat{\beta}_j)_{1\le j\le 4}$?

(c) Compute the grand average of all $n = 12$ vitamin levels.

(d) Which of the given sets of parameter estimates corresponds to the coding that was recommended in the lecture?

27. **Programming exercise: Third Glasgow MONICA survey.**
Make yourself familiar with the dataset in the file `MONICA-Alkohol.txt` which you can download from the lecturer's homepage. If you should encounter problems in downloading the file or if you do not have access to the internet, you may alternatively get the file via USB stick during the lecturer's consulting hour.

The file comprises the data of $n = 40$ participants of the third Glasgow Monitoring Trends and Determinants on Cardiovascular Diseases (MONICA) survey and is structured as follows (there is no headline in the file!).

variable name	column number(s)	description
age	$1-2$	age in years
alcohol	5	drinking habits, coded into five categories: $1 \mathrel{\hat{=}}$ never drinks alcohol $2 \mathrel{\hat{=}}$ occasionally drinks alcohol $3 \mathrel{\hat{=}}$ mild alcoholic $4 \mathrel{\hat{=}}$ medium alcoholic $5 \mathrel{\hat{=}}$ heavy alcoholic
protc	$8-10$	protein C level (iu/dl)
prots	$13-18$	protein S level (%)

(a) Describe the univariate distributions of all variables in the dataset.

(b) The primary scientific interest of the study was to estimate parameters for a linear relationship between protein S level (the response) and the other (co-)variates. Fit a multiple linear regression model to the given data by making use of statistics software.

(c) Assess the quality of your model fit. Hint: Use Definition 3.21.

(d) Perform a (e. g., graphical) residual analysis.

28. **Multiple Select.** Which of the following statements are true and which are false?
Please give reasons for your respective decisions (one short sentence each is sufficient).

1. Under the assumptions of Theorem 3.18, the hat matrix H has rank n.
2. If, under the assumptions of Theorem 3.18, the sample size n is an even integer and $p = 2$ (only one covariate is considered), then $n/2$ of the data points $(x_i, y_i)_{1\le i\le n}$ lie above (in y-direction) the least squares regression line and $n/2$ data points lie below this line.
3. The centering operator C defined in Lemma 3.20 is invertible.
4. If, under the assumptions of Theorem 3.21, the coefficient of determination R^2 is equal to one, then the noise variance σ^2 will be estimated as zero by both methods discussed in Theorem 3.24.