Exam Statistical Models

18 February 2021

You may use a simple calculator provided it is not part of a communicating device, and the following quantiles: $F_{2,75;0.95} = 3.12$, $t_{47;0.95} = 1.6779$, $t_{47;0.975} = 2.0117$, $\chi^2_{1;0.95} = 3.84$, $\chi^2_{2;0.95} = 5.99$, $\chi^2_{3;0.95} = 7.81$, $\chi^2_{44;0.95} = 60.48$, $\chi^2_{45;0.95} = 61.66$, $\chi^2_{46;0.95} = 62.83$, $\chi^2_{47;0.95} = 64.0$, $\chi^2_{48;0.95} = 65.17$, $\chi^2_{50;0.95} = 67.5$. *The significance level is always* $\alpha = 0.05$ *unless specified otherwise.*

1. To investigate the effect of 3 types of diet, 78 persons were divided randomly in 3 groups, the first group following diet 1, second group diet 2 and the third group diet 3. After 6 weeks of diet, both the weight and the lost weight were measured (in kg) for each person in the study. The collected data is summarized by the numerical columns `weight.lost` (the lost weight after 6 weeks of diet, `weight6weeks` (the weight after 6 weeks of diet), and the factor column `Diet` (the type of diet followed).

   (i) (5) In R, we create the model `mod1=lm(weight.lost~Diet)`. The R-output of `anova(mod1)[1,4]` delivers the `F value` 6.2. What is the studied model here? Specify model assumptions and constraints. Conclude on whether the factor `Diet` influences `weight.lost`.

   (ii) (6) Now including `weight6weeks` as an additional variable into the R-analysis, we create the model `mod2=lm(weight.lost~weight6weeks+Diet)`. The partial R-output of `anova(mod2)` is

```
             Df Sum Sq Mean Sq F value   Pr(>F)
weight6weeks  1  _____  24.460  4.4551 ______
Diet          2  70.52   _____  6.4216 0.002681 **
Residuals    74  _____   5.490
```

   What is the studied model here? Draw relevant conclusions using the above R-output.

   (iii) (8) Use the below R-output of `summary(mod2)` to draw relevant conclusions and to estimate lost weight for the three types of diet, for a person with `weight6weeks` 80 kg?

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.65232    2.14095   3.574 0.000623 ***
weight6weeks -0.06256    0.02999  -2.086 0.040463 *
Diet2        -0.36727    0.65888  -0.557 0.578924
Diet3         1.77974    0.65818   2.704 0.008494 **
```

   (iv) (8) Which of the two models, `mod1` or `mod2` (without or with variable `weight6weeks`), do you prefer? Why? Describe how you could investigate whether the dependence of `weight.lost` on `weight6weeks` is similar under all three types of diet.

2. Suppose we have a dataset $\{(x_1, Y_1), \dots, (x_{50}, Y_{50})\}$ which is modeled as follows:

$$Y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \dots, 50, \qquad (*)$$

   where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ is to be estimated, $f(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 \frac{\exp(\theta_3 x)}{1+\exp(\theta_3 x)}$, $\varepsilon_1, \dots \varepsilon_{50}$ are independent random errors such that $\mathrm{E}\varepsilon_i = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, \dots, 50$.

   (i) (5) For the model $(*)$, give the normal equations used for calculating the least squares estimator (LSE) of $\boldsymbol{\theta}$. Suppose we obtained the LSE $\hat{\boldsymbol{\theta}} = (-1, 2, 2)$ for the parameter $\boldsymbol{\theta}$. Explain how this estimate can be used to construct a consistent estimator of σ^2.

(ii) (10) Suppose we obtained an estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}$:

$$\widehat{\mathrm{Cov}(\hat{\boldsymbol{\theta}})} = \hat{\sigma}^2(\hat{V}^T\hat{V})^{-1} \approx \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{pmatrix}.$$

Use this matrix, the LSE $\hat{\boldsymbol{\theta}} = (-1, 2, 2)$ and relevant quantiles to construct a 95% (approximate) confidence interval for $f(0, \boldsymbol{\theta})$.

(iii) (8) Consider the reduced model where $\theta_1 = \theta_3 = 1$ and describe how you would test whether this reduced model is adequate for describing the data.

3. Suppose we observe $Y_1, \ldots, Y_n$, independent and normally distributed. Let $Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i \in \mathbb{R}$, $i = 1, \ldots, n$, $\sigma^2 > 0$, i.e., the density of Y_i is $f_i(y) = (2\pi\sigma^2)^{-1/2} \exp\{-(y - \mu_i)^2/(2\sigma^2)\}$. Besides, the values of three covariates are available, called, say, `X1`, `X2` and `X3`.

(i) (4) For observed data, propose a generalized linear model with three covariates.

(ii) (12) The general form of the exponential family is

$$f(y, \theta_i) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i)\right\}.$$

Show that the distribution of Y_i can be written in this form with an appropriate function $h(\mu_i) = \theta_i$. Identify the functions $b(\theta_i)$, $c(y, \phi/A)$, the parameters ϕ, A_i, and demonstrate how to compute $\mathrm{E}Y_1$ and $\mathrm{Var}(Y_1)$ by using these quantities. Derive the canonical link function $g(\mu)$.

(iii) (9) Suppose we obtained the following analysis of deviance table.

	Df	Deviance	Resid. Df	Resid. Dev
NULL			47	69.92
X1	1	3.08	46	66.84
X2	1	9.81	45	57.03
X3	1	6.43	44	50.60

What is the number of observations? What can you tell about the relevance of the covariates `X1`, `X2` and `X3` in the model? Let $\sigma^2 = 1$ be known and ω be the reduced sub-model with one covariate `X1`. Test H_ω: `the reduced model fits well`.

4. Let $\{Z_t\}$ be independent normal random variables such that $Z_t \sim N(0, \sigma^2)$, $\sigma > 0$.

(i) (7) Let $\sigma^2 = 1$. Is the time series $Y_t = Z_{t-1}^2 + 2Z_{t+1}^2$ weakly stationary?

(ii) (9) Consider a stationary AR(1) time series

$$X_t = \alpha X_{t-1} + Z_t.$$

Derive the Yule-Walker equations and argue how these can be used to estimate α and σ^2.

(iii) (9) A general ARMA(p, q) process is of the form $X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \ldots + \beta_q Z_{t-q}$. Consider the time series $\{Y_t\}$ given by $Y_t = \nabla_2 V_t$, where

$$V_t = 0.5V_{t-1} - 2Z_{t+1} - Z_{t-1}.$$

Does $\{Y_t'\} = \{aY_t\}$ follow an ARMA(p, q) model (for some $a \neq 0$)? If so, identify $s \in \mathbb{R}$ and the values of parameters p, q, α_i, $i = 1, \ldots, p$ and β_j, $j = 1, \ldots, q$.