UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK9900 — Short solution

Day of examination: Friday 12. May 2017.

Examination hours: 9.00 – 13.00.

This problem set consists of 3 pages.

Appendices: Tabels for normal-, t-, χ^2- and F-distributions

Permitted aids: All printed and hand-written resources. Approved calculator.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

The exams in STK4900 and STK9900 have substantial overlap, but are not the same. This is a very brief solution to the STK9900 questions, which includes the questions for STK4900.

Exercise 1.

a) If the expected changes for each treatment are nominated μ_1, μ_2 and μ_3, we test $H_0 : \mu_1 = \mu_2 = \mu_3$ against H_a : not all are equal. K=3, n=15, so 2 df for model, 12 df for residual and 14 df for total sum of squares. It is important to randomize to avoid that other effects than the treatment systematically influence the response.

Exercise 2.

a) Let x_1 be the IQ of the mother. A simple linear regression model for the skills is then

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

where $i = 1, \ldots, 36$. The ϵ_i are assumed to be independent $N(0, \sigma^2)$. The first plot shows that normality seems reasonable, the second that the linear model is ok, and the third that we probably have constant variance (σ^2 does not depend on i). There seem to be some possible outliers, child 2, 4 and 24. We find $\hat{\beta}_0 = 111.09, \hat{\beta}_1 = 0.4066, \hat{\sigma} = 3.856$.

b) Testing $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against the alternative that at least one coefficient is not 0: We use an F-test. The test statistic $F = \frac{MSS/p}{RSS/(n-p-1)}$ is F-distributed with p and $n - p - 1$ degrees of freedom under H_0. Here $p = 4$ and $n = 36$. From R we find $F = 17.12$. The probability of being larger or equal to 17.12 in an F-distribution with 4 and 31 degrees of freedom is very small ($1.654 \cdot 10^{-7}$), hence we can reject the null hypothesis

(Continued on page 2.)

at any reasonable significance level, and conclude that at least one of the coefficients is significantly different from 0.

c) In Analysis 2, all coefficients are significant. In Analysis 1, we have a higher R-squared than that in Analysis 2, but we also have more covariates. When the number of covariates differs in the two models, we have to use R-squared-adjusted, which is adjusted for the number of parameters. R-squared-adjusted is slightly larger in Analysis 2 than in Analysis 1. In general, we prefer simpler models, and keep only significant covariates. 'Multiple R-squared' for multiple regression is the squared empirical correlation between the observed response and the modelled response, while in simple (univariate) regression, this is equal to the squared empirical correlation between the observed response and the covariate.

d) Keeping parents' IQ scores fixed, the estimated effect of an increase of one hour reading per week is an increase of 12.76632 points in expected skills. A 99% confidence interval for this effect is found as $12.76632 \pm 2.75 \cdot 2.23107$, that is $(6.63, 18.90)$ (using 30 degrees of freedom from the t-table rather than the correct 32).

Exercise 3.

a) We should use a logistic regression model. We have data $(x_1, x_2, ..., x_p, y)$ for $n = 250$ emails, where x_1 ... x_p are explanatory variables and the outcome y is binary (0 and 1). We are interested in modelling the probability of an email being spam mail, hence we model $p(x_1, x_2, ..., x_p) = P(y = 1|x_1, x_2, ..., x_p)$ with a logistic regression model

$$p(x_1, x_2, ..., x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}.$$

Such a model is fitted using the maximum likelihood principle; in R that is done with the glm function. For new emails arriving, we would collect the covariates x_1 ... x_p and plug them into the fitted model. That would give us an estimated probability of that email being spam. We would need to decide a cut off value in order to classify the email as either spam or ok, f.ex. we could classify an email as spam if the probability is larger than 90%, say.

b) If all the other covariates are the same, an email that contains the word winner ($x_1 = 1$) will have an estimated odds ratio compared to an email without the word winner ($x_1 = 0$) of $\exp(\hat{\beta}_1) = \exp(1.4) = 4.055$.

Exercise 4.

a) We use a χ^2-test where we merge the four last cells in order to have at least 5 expected counts in each cell. We find $\chi^2 = 26.79$ and with 2 degrees of freedom, we get a very small p-value and conclude that we do not have a good fit with a Poisson(λ) distribution. There

(Continued on page 3.)

is a misprint in the table in a), where 'no. observed' should be 'no. expected'. This should be clear from the text, but was announced to everybody at the exam.

b) The rate ratio corresponding to one unit's increase in math score is $\exp(\beta_1)$, where β_1 is the coefficient for math score x_1. From R we estimate this to $\exp(0.086) = 1.09$. We can check if this effect is significant by looking at the p-value for 'math' in R, which is very small, hence we can conclude that β_1 is significantly different from 0 (and the rate ratio different from 1). The Wald test is simply a z-test, since $z = \hat{\beta}_1 / SE_{\hat{\beta}_1}$ is approximately N(0,1) under $H_0 : \beta_1 = 0$. Here we have $z = 8.902$ which gives a (two-sided) p-value close to 0, and we reject the null hypothesis. The rate ratio corresponding to ten unit's increase in math score is $\exp(10 \cdot \beta_1)$, giving $\exp(0.86) = 2.36$, hence the rate is more than doubled.

c) Study program is a categorical variable with 3 levels. We can include it as a factor using 2 indicator variables. x_2 is 1 when the student is in program 2, 0 otherwise. x_3 is 1 when the student is in program 3, 0 otherwise. Then program 1 is the reference program, corresponding to the parametrization used in R. The model is hence $Y_i \sim Po(\lambda_i)$ where $\lambda_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$, $i = 1, \cdots, 200$.

d) $H_0 : \beta_2 = \beta_3 = 0$ against H_a: at least one of them is different from 0. The test statistic is $G = D_0 - D = 204.02 - 189.45 = 14.572$. The degrees of freedom is 2. The probability of observing 14.572 or larger in a χ^2-distribution with 2 degrees of freedom is 0.0006852 (from R) and we reject H_0 and conclude that study program should be a part of the model.

e) The award rate is largest in study program 2 (academic). The rate ratio here is $\exp(1.084) = 2.96$ compared to the reference program (general), for the same math score. A 95% confidence interval for β_2 is $1.084 \pm 1.96 \cdot 0.35825$, that is, $(0.38183, 1.78617)$. This gives the 95% confidence interval (1.4647, 5.9658) for the rate ratio for students in the academic program compared to the general program, when math score is kept the same.