UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK4900 — Statistical methods and applications.

Day of examination: Friday June 15th 2018.

Examination hours: 14.30 – 18.30.

This problem set consists of 5 pages.

Appendices: Tables 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.

Problem 1

a) The estimates are $\hat{\beta}_0 = 30.511906, \hat{\beta}_1 = -0.082898$ for β_0 and β_1 and $s = 2.983$ for σ.

The (multiple) R-squared is defined as $R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$ where $\bar{Y}$ is the mean of the Y_i and $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ the predicted values of the Y_i, thus R^2 is proportion of the variation explained by the regression. Also R^2 is the squared correlation between the observed Y_i and the predicted values $\hat{Y}_i$.

In particular since this is a simple linear regression model we also have that R^2 equals the squared correlation between the Y_i and the covariate x_{i1}, thus the correlation becomes $-\sqrt{R^2} = -\sqrt{0.4913} = -0.701$ where the minus sign follows from $\hat{\beta}_1$ being negative.

b) The reason that $\hat{\beta}_1$ is the same in the simple linear regression in question a) and the multiple linear regression in this question is that the design is balanced, we have equally many observations of each value of x_{i1} for each value of x_{i2}. Then the value of x_{i2} gives no information of the value of x_{i2} and the two covariates must be uncorrelated. When we have two uncorrelated covariates omitting one of them will not change the estimated regression parameter of the other.

Although $\hat{\beta}_1$ did not change so did the estimate s of σ and of σ^2 given as $s^2 = \frac{1}{n-3} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ where now $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$ is the predicted values using both covariates. This estimate is smaller than

(Continued on page 2.)

s^2 from the simple linear regression since both covariates are important predictors of the Y_i.

This is also the reason that the new $R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$ is larger than that in simple linear regression.

The standard error se_1 of $\hat{\beta}_1$ is proportional to s and so this estimate has also become smaller in this situation where the covariates are uncorrelated, (can also be the case with moderately correlated covariates if the effect of the covariates is sufficiently small), something that has also led to t-value $\hat{\beta}_1/se_1$ deviating more from zero.

c) The first plot show the residuals $e_i = Y_i - \hat{Y}_i$ against the fitted values $\hat{Y}_i$. The smoothed line through the points $(\hat{Y}_i, e_i)$ shows a clear curvature. This indicates that transformations of the covariates or inclusion of square terms x_{i1}^2 and x_{i2}^2 and possibly also interaction terms $x_{i1}x_{i2}$ might improve the fit.

The second plot is a qqplot of the ordered (standardized) residuals e_i^* against percentiles in the standard normal distribution. When these points are not on a straight line we have an indication that the error terms are not normally distributed. This again could indicate that the use of the t-distribution when calculating p-values is not optimal (However, a close inspection reveals that the tails of the distribution of the residuals is "lighter" than that of the normal distribution, in this perspective this deviation is probably not very serious).

The third plot show $\sqrt{|e_i^*|}$ agains $\hat{Y}_i$ and is designed to show whether we have heteroscedasticity, i.e. the variance depends on the expected value of Y_i. Although the curve show some curvature it varies from values around 0.7 to 1.2 which is not considered very much, and so the heteroscedasticiy is likely not very serious.

This means that the part the model that needs to be worked on is the linear part, either including more terms or transforming those already included.

d) It is problematic to use the standard R^2 to choose a model because it will necessarily become larger when including a new term to the model (this improvement may be very modest, but R^2 does not tell anything about of what is an important change). It will then be better to use the adjusted R^2 or the predicted (cross-validated) R^2 which can have a maximum over the considered models and choose the model maximizing such modified R^2 measures (the predicted R^2 typically chooses a smaller model and also have other preferable properties).

The idea in the predicted R^2 is to calculate new predicted values $\hat{Y}_i^{(-i)}$ where the regression model is fitted without using value no. i

(Continued on page 3.)

and predicting Y_i from this model. The measure is then defined as $R^2_{pred} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^{(-i)})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$.

From the table we see that the model with the smallest predicted R^2 is the biggest model including both square terms and the interaction term (although the improvement for the last model including also the square term of log(Time) is quite modest and improves the model only marginally).

Problem 2

a) With p_W the probability that a woman died and p_M the probability that a man died we want to test $H_0 : p_W = p_M$. This can be done using a test statistic $Z = \frac{\hat{p}_M - \hat{p}_M}{se(\hat{p}_M - \hat{p}_M)}$ where $\hat{p}_M = 372/468 = 0.795$ and $\hat{p}_W = 71/288 = 0.247$ are the proportions dying among men and women and $se(\hat{p}_M - \hat{p}_M) = \sqrt{\hat{p}(1-\hat{p})/(1/468 + 1/288)}$ is the standard error of $\hat{p}_M - \hat{p}_M$ under the null that the mortality is the same $p_W = p_M = p$ and $\hat{p} = (71 + 372)/(288 + 468) = 0.585$.

Under the null hypothesis Z is drawn from a standard normal distribution, thus if $|Z| > 1.96$ we can reject the null at a 5% level. Here we find $Z = 14.86$ and so we reject the null with a very small p-value < 0.001.

b) The odds-ratio of dying between men and women is given by $OR = \frac{p_M/(1-p_M)}{p_W/(1-p_W)}$. We can estimate it by simply plugging in $\hat{p}_M$ and $\hat{p}_W$ from question a) giving us a value $(0.795/0.205)/(0.247/0.753) = 11.8$

However, this odds-ratio is also given as $\exp(\hat{\beta}_1)$ from the logistic regression with outcome Y = indicator of dying and x = indicator of man, thus again $OR = \exp(2.472) = 11.8$

The relative risk of dying between men and women becomes $RR = p_M/p_W = 0.795/0.247 = 3.2$ which is considerably smaller than the odds-ratio. The odds-ratio is a good approximation to the relative risk when the probabilities are both small, but here they are both large and there is a striking difference. Note, however, that if the $\hat{p}_M = \hat{p}_W$, then we have both OR and RR equal to 1 and if one of them is larger (smaller) than 1 then so is the other.

c) The model is given as, $Y = 0$ for a survivor and $Y = 1$ for passenger that died and x = age in years, $p(x) = P(Y = 1|x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$. Similarly to question a) we have the odds-ratio interpretation

$$OR(x+1, x) = \frac{p(x+1)/(1-p(x+1))}{p(x)/(1-p(x))} = \exp(\beta_1),$$

(Continued on page 4.)

thus the regression parameter for age $\hat{\beta}_1$ has the interpretation as the estimated odds-ratio of dying between two individuals with a one year difference in age.

For two individuals where one is 10 years older than the other we similarly have an odds-ratio given as $\exp(10\hat{\beta}_1) = \exp(0.08795) = 1.092$, thus an increase in odds of 9.2%.

The confidence interval for this odds-ratio $\exp(10\beta_1)$ is given by the formula $\exp(10\hat{\beta}_1 \pm 1.96 * 10se(\hat{\beta}_1) = (0.986, 1.210)$. Since this confidence interval overlaps OR=1 the risk of dying was *not* significantly associated with age.

This test could also be done considering the test statistic $Z = \hat{\beta}_1/se(\hat{\beta}_1) = 0.008795/0.005232 = 1.681$ which is less than the 97.5 percentile 1.96 in the standard normal distribution, thus the p-value is above 0.05.

d) The model with age entered directly did not show significantly association between age and mortality. In contrast the model with log(Age) has a p-value < 0.001 for the regression parameter being different from zero, thus there is a strongly significant correspondence between log-age and mortality. So clearly log-age is the better variable. This is also reflected in lower residual deviance and higher AIC-values for the log-age model than for the model with age entered without transformation.

e) The residual deviance is twice the difference between the log-likelihood of the actual model and the loglikelihood of the "saturated" model with "estimated" probabilities $\tilde{p}_i = Y_i$. Denoting D_1 and D_2 the deviances for two nested models M1 and M2, and where M1 is a special case of M2, we have that the log-likelihood ratio statistic $G = D_2 - D_1$ is approximately chi-square distributed assuming M1 as a null hypothesis. The degrees of freedom for G is equal to the difference of parameters in the two models.

From the deviance table we see that log(Age), Sex have G-statistics of 12.55 and 225.99 which is well above the 95 percentile of a chi-square distribution with one degree of freedom which equals 3.84 ($= 1.96^2$).

Furthermore the G-statistic for PClass equals 99.4 also well above the 95 percentile of a chi-square distribution with 2 degrees of freedom which equals 5.99, so again significant. We use 2 degrees of freedom here because PClass is a categorical covariate with three levels.

Finally we see that also the interaction between Sex and PClass is significant since G=30.4 > 5.99 with 2 degrees of freedom.

f) The log-oddsratio between a reference group woman at 1.class and the other groups can be described as adding the appropriate main effect parameters and the appropriate interaction parameter. We see that

(Continued on page 5.)

the interaction effect between Sex and 2. class is insignificant, so for 2. class a model with only the main effects seems to described the data well.

However, for the interaction between 3. class and sex the interaction effect is significant and negative. Thus the overall effect for men at 3. class is smaller than the sum of the two main effects, i.e. the log-oddsratio compared to a reference women at 1. class becomes $3.61 + 3.73 - 2.17 = 5.17 < 3.61 + 3.73 = 7.34$.

END