UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK4900 — Statistical methods and applications.
Day of examination:	Tuesday June 4th 2019.
Examination hours:	9.00 – 13.00.
This problem set consists of 4 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 model used can be specified as responses $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where x_i is the height variable, β_0 is the theoretical intercept parameter, β_1 the theoretical slope parameter and the ε_i are independent error terms with equal variance σ^2 (and often assumed being drawn from a normal distribution).

Here we estimate β_0 by $\hat{\beta}_0 = -2.29$ which is the fitted value (extrapolated) with height $x_i = 0$, β_1 is estimated as $\hat{\beta}_1 = 0.031463$ which is the estimated mean increase in FVC when height increases one centimeter. The third parameter σ is estimated by the residual standard error $s = 0.2016$ and has the interpretation as the standard deviation of the observation given the covariate value.

We find that the t-value for testing $\mathrm{H}_0 : \beta_1 = 0$ using the test-statistic $t_1 = \hat{\beta}_1/\mathrm{se}(\hat{\beta}_1) = 0.031463/0.001082 = 29.07$ which is drawn for t-distribution with $n - 2 = 598$ degrees of freedom if the null is correct. A value of $t_1 = 29$ is indeed large if the null is true, and we can conclude that null hypothesis is very likely false, so it is safe to say that FVC tends to be larger for the taller children.

b) The first resiudal plot depicts the residuals $e_i = Y_i - \hat{Y}_i$ plotted against fitted (or predicted values) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. There is no clear curvature in the residuals as $\hat{Y}_i$ changes, thus there is probably little deviation from the linearity assumption of the model (R finds a slight curvature with the red line, though, so perhaps including a square term could improve the model somewhat).

The second plot is a qq-plot where percentiles from a standard normal distribution are given along the x-axis and the ordered (standardized)

(Continued on page 2.)

residuals along the y-axis. The points lie very close to a straight line, so there is not strong deviation from the normality assumption. However, there is one apparent outlier relative to the normality assumption.

In the third plot we see $\sqrt{|e_i^*|}$ where e_i^* are standardized residuals plotted against the fitted values. If these y-axis values tends to be bigger or smaller as the fitted values changes the assumption of constant variance σ^2 for all observations would not hold. This does not seem to be the case here, although perhaps the variance is slightly smaller for the very low expected values.

c) The coefficient of determination (or shorter just R2) is given 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 average of the Y_i's. This measures how large proportion of the variance that is explained by the regression model. Although a very useful measure it can only increase if more covariates are included in the model, thus it does not give a clearcut way of choosing a model.

The adjusted R2, however, penalizes for including more covariates p. It is defined as $R^2_{adj} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2/(n-p-1)}{\sum_{i=1}^n (Y_i - \bar{Y})^2)/(n-1)}$ and it can obtain a maximum when increasing p and gives a formal method for choosing a model by choosing this model with the maximum R^2_{adj}.

The third measure is the crossvalidated R2 defined as $R^2_{cross} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^{-i})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2)}$ where $\hat{Y}_i^{-i}$ is a predicted value for observation no. i where this observation has been omitted when fitting the model. Again the R^2_{cross} can have maximum over possible models and in the same way as adjusted R2 gives a formal method for choosing a model.

In this case we see that the model with the highest adjusted and crossvalidated R2's is the model where all four covariates are included, M4, so according to the model selection criteria above we would choose model M4 with all four covariates.

d) The interpretation of the regression coefficients in the model is the estimated change in the response if one covariate is changed with one unit *and* the other covariate is kept constant.

Here weight and height are strongly correlated, about 0.7, and both regression coefficients are found clearly significant and also the R2 is increased by including the new covariate. Then the estimated effect in the univariate model will consist of the bivariate effect of that covariate plus an effect of the other covariate mediated through the correlation. Specifically we have $\hat{b}_1 = \hat{\beta}_1 + r_{12}\hat{\beta}_2 \frac{s_2}{s_1}$ where r_{12} is the correlation between the two covariates and s_j is the standard deviation of covariate j, which can be verified $\hat{\beta}_1 + r_{12}\hat{\beta}_2 \frac{s_2}{s_1} = 0.261 + 0.701 * 0.0178 * 3.26/7.61 = 0.0314 = \hat{b}_1$.

(Continued on page 3.)

e) The predicted values of Y_i is given by $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 x_i$.

The confidence interval is for the expected values $\mu_x = \beta_0 + \beta_1 x_i$ which has the interpretation that if similar data were collected many times over again the interval would cover μ_x in 95% of the data sets.

The prediction interval is about uncertainty for a new observation $Y_{new} = \beta_0 + \beta_1 x_{new} + \varepsilon_{new}$ where ε_{new} is independent of previous ε_i. We then have a rule that will cover such new values Y_{new} with probability 95% .

Regarding the prediction with several covariates we see that prediction interval is more narrow than that from the simple linear regresssion, but the difference is not very large. This is related to a relatively modest increase in R2 from the simple regression to the full model (from 0.59 to 0.63).

Problem 2

a) It could be a reasonable assumption that the spam mails arrive according to a Poisson process: (i) The rate of spams λ is constant, (ii) The number of spams in non-overlapping intervals are independent and (iii) Only one spam will arrive at a single time-point. Then the number of spams in unit time intervals will be Poisson distributed and the probability of k spams will be given as $p(k;\lambda) = \frac{\lambda^k}{k!}\exp(-\lambda)$.

The E's are estimates of how many hours there are with a specific number k of spams, specifically $E = np(k;\hat{\lambda})$. When the values of E differs much from the observed number O we have indication that the assumed Poisson distribution does not hold after all. The deviations can be summarized by a test statistic $X^2 = \sum_{k=0}^{5} \frac{(O_k - E_k)^2}{E_k}$ which here equals 24.85. If the Poisson assumption is true then X^2 will approximately follow a χ^2 distribution with degrees of freedom given by no. of groups - 2 = 4 (here). From a table of the χ^2-distribution it turns out that the p-value for the test is less than 0.005 and so there is a significant deviation from a Poisson-distribution.

b) A Poisson regression model is defineb by Poisson distributed responses Y_i that are independent with rate (equal to expectations) λ_i and where the λ_i with covariates $x_{i1}, \ldots, x_{ip}$ can be written as $\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$. When we only have one categorical covariate this can be written as $\exp(\alpha_0 + \beta_k)$ where α_0 is an intercept term and β_k the log-linear difference in expectation between a reference level and level k of the categorical covariate.

The estimate for the rate of spams in August is then $\exp(\hat{\alpha}_0) = \exp(-0.03976) \approx 0.96$ spams per hour.

(Continued on page 4.)

The rate ratio is obtained as $\exp(\hat{\beta}_k) = \exp(0.35821) = 1.43$, so the rate of spams were 43% higher in October than in August. The confidence interval for $\exp(\beta_k)$ is given as $\exp(\hat{\beta}_k \pm 1.96 se_k)$ where se_k is the standard error of $\hat{\beta}_k$, thus we get the 95% CI by $\exp(0.35821 \pm 1.96 * 0.06733) = (1.25, 1.63)$.

c) One can test whether a new covariate gives a significant improvement to the model by calculating the difference in residual deviances between the models without and with the new (categorical) covariate. If the new covariate has no impact on the outcomes (β-s are equal to zero) then this difference in deviances is approximately χ^2-distributed with degrees of freedom equal to the number of parameters that are set equal to zero (no. of levels for the categorical covariate) and the p-value is the probability that such a χ^2-variate will exceed the observed difference.

By including first month the residual deviance is decreased by 35.6 on 4 degrees of freedom, corresponding to a p-value 0.005, so cleary significant.

Then including day of the week the residual deviance i decreased 7.76 on $6 = 7$ days -1 degrees of freedom, thus a p-value above 0.05, so the rate of spam on different days of the week are not found significantly different.

However, hour of the day gives a reduction in residual deviance of 44.2 with $24 - 1 = 23$ degrees of freedom with p-value of 0.005, thus there appear to daily variation.

d) The log-rate is a function $\beta_0 + \beta_1 h + \beta_2 h^2$, that is a quadratic, in h = hour of the day and since the estimated β_2 is greater than zero this is a function with a minimum. One may find this minimum by plotting the function or by noting that its derivative equals $\beta_1 + 2\beta_2 h$ which is zero when $h = -\beta_1/(2\beta_2)$. Plugging in the estimates we find that the time with the minimal spam rate is estimated as $-(-0.0389/(2 * 0.0015)) \approx 13$, thus around noon is the time with the minimum rate. The highest rates are around midnight.

e) We then estimate the overdispersion term as $\hat{\phi} = 3546.2/(n - p - 1) = 1.22$ (with $n = 2928$ and $p = 1 + 2 + 6 + 4 = 13$, 2 for hour and hour^2, 6 for weekday and 4 for month), this means that the variance Y_i is estimated as $1.22\lambda_i$. The parameter estimates given by Poisson-regression are consistent (valid) for the rates, but the standard errors need to be corrected by a formula $se_j/\sqrt{\hat{\phi}}$ where se_j are the standard errors given by Poisson regression. In turn this means that z-values are decreased by a factor $\sqrt{1.22} = 1.1$, thus 10% smaller than those reported by the Poisson regression and so p-values will be somewhat larger and borderline significant parameters reported by the Poisson-regression can not anymore be considered significant.

END