UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK4900 — Statistical methods and applications.
Day of examination:	Wednesday June 4th 2014.
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 correlation between ozone and temperature equals $+\sqrt{R^2} = +\sqrt{0.488} = 0.699$. The correlation is positive since $\hat{\beta}_1 = 2.4391 > 0$.

The Wald-statistic for testing $H_0 : \beta_1 = 0$ is given by $t = \hat{\beta}_1 / se(\hat{\beta}_1) = 2.4391/0.2393 = 10.2$. Under the null hypothesis this statistic has t-distribution with $n - 2 = 109$ degrees of freedom, thus approximately standard normal. Since $|t| > 3$ this association between temperature and ozone is strongly significant.

b) R^2 is proportion of the variation in the responses that is explained by the regression. Specifically $R^2 = 1 - \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2}$. A high R^2 indicates that the covariates predict the outcome well.

Here the R^2 is increase from 0.488 to 0.581, and prediction power is clearly increased by including the covariate `wind`. Also we see that `wind` is clearly significant, p¡0.001, so including `wind` in the model is obviously an improvement.

The estimate for `temp` has changed from 2.44 to 1.83. Then there must be a (non-zero) correlation between `wind` and `temperature` and in the simple linear regression the estimated regression coefficient is confounded with the effect of wind through this correlation.

c) The assumption for the linear regression is that $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ where the x_{ji} are the two covariates and the error term ε_i is normal with mean zero and a variance σ^2 and independent. This can be written out in four issues:

(Continued on page 2.)

- Linearity $\mathrm{E}[Y_i] = beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$
- Constant variance: $\mathrm{Var}[Y_i] = \sigma^2$ for all i
- Independence of the Y_i
- ε_i are $N(0, \sigma^2)$ normally distributed

The first plot of residuals versus predicted variables, i.e. $(\hat{Y}_i, \hat{\epsilon}_i) = (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}, Y_i - \hat{Y}_i)$ shows some curvature. This indicates that the second order terms or other transformations of the covariates can improve on the model. (We can also see some negative predicted values, which does not make sense, this also indicates potential for improving the model).

The Normal Q-Q plot are the ordered residuals versus theoretical percentiles in the normal distribution should lie close to a straight line if the error terms were normally distributed. Here we see that there is a somewhat heavy upper, relative to the normal, tail in distribution. A transformation, for instance log, of the outcome ozone could improve this assumption. There might also be an outlier, observation 77, as can be seen also from other plots. A separate analysis omitting this observation could be a good idea.

The third plot is used to check if the variance is constant. If the variation in the (square root of the) absolute value of the residuals on seems to be roughly the same over all predicted values along the x-axis then this assumption is OK. Here there is not a clear indication of non-constant variance.

The two last plot are "component plus residual" plots. In these the covariates are along the x-axes and the points are $\beta_j x_{ij} + \hat{\varepsilon}_i$. The dashed lines are simply the linear component $\beta_j x_{ij}$. If there is a clear non-linearity in the plots this indicates a need for quadratic terms or (log-)transformations of the covariates. In accordence with the first plot there are indications of deviation from linearity for both covariates.

d) Since the data are taken over time it could be reason to think that observations taken close in time correspond better than observation farther apart, thus there are dependencies in the data, which is the third assumption on the list in question c). This could have the consequence that the standard errors are biased and hence confidence intervals and hypothesis test may be in error.

Problem 2

a) The lizards can choose to stay either in the shade or in the sun, thus the outcome is binary and a logistic regression is the standard model for binary data.

(Continued on page 3.)

The model used in the problem has two factor covariates, time of day with 3 levels and species with two levels. Thus the model can be stated

P_{ij} = P(A lizard prefers the sun | levels i and j) = $\frac{\exp(\alpha+\beta_i+\gamma_j)}{1+\exp(\alpha+\beta_i+\gamma_j)}$

where β_i is the effect of time of day, level i ($\beta_1 = 0$) and γ_j the effect of species ($\gamma_1 = 0$)

Then $\exp(\beta_i)$ and γ_j has interpretations as odds-ratios

$\exp(\beta_i) = \frac{P_{ij}/(1-P_{ij})}{P_{1j}/(1-P_{1j})}$ and $\exp(\gamma_j) = \frac{P_{ij}/(1-P_{ij})}{P_{i1}/(1-P_{i1})}$

relative to the reference category $i = 1$ or $j = 1$ which in the output is the early in the day and species Grahami. ($\exp(\alpha)$ is the odds in the reference group, $i = j = 0$.)

When all P_{ij} are small we have that odds-ratios approximate relative risks (such as P_{ij}/P_{1j}). Here we have that P_{ij} range from about 5% to about 25%, so odds-ratios should deviate somewhat more from 1 (=no difference) than relative risks.

b) Odds-ratios for time Midday versus time Early: $\exp(-1.484) = 0.22$, so lizards tend to stay in the sun much less during midday than early in the day (and with a relative risk interpretation only one fourth as often).

Similarly Odds-ratios for time late versus time early: $\exp(0.2429) = 1.28$, thus somewhat more, maybe about 25% more.

And for Opalinus vs Grahami: $\exp(-0.748) = 0.47$, thus Opalinus is more rarely in the sun, maybe about only half as often.

The 95% confidence interval for the log-odds of Opalinus vs Grahami is given as $\hat{\gamma}_2 \pm 1.96 se(\hat{\gamma}_2) = -0.748 + 1.96 * 0.3037 = (-1.34, -0.15)$, thus the 95% confidence interval for the odds-ratio becomes $\exp(-1.34), \exp(-0.15)) = (0.26, 0.86)$. Since this confidence interval does not contain the value one, the difference is significant with a p-value less than 5 percent (from the output we find a p-value of 1.4%).

c) Deviances are generalizations of sums of squares for linear regression models. With l being the log-likelihood of a given model and $\tilde{l}$ the log-likelihood of a saturated model, that is a model with as many parameters as observations, the deviance is defined as $D = 2(\tilde{l} - l)$.

When comparing two nested models M0 and M1 (with M0 a special case of M1) with deviances D_0 and D_1 respectively we have that $G = D1 - D0$ is approximately χ^2-distributed with degrees of freedom given as the difference in the number of parameters under the null hypothesis that M0 is true.

Here we get a change in deviance $G = 43.624$ comparing a model with differences over the day to a model where lizards are equally likely

(Continued on page 4.)

to be in the sun any time of the day. The upper 0.5% cutoff for the χ^2 distribution with 2 degrees of freedom (three groups, with one reference) is found in the chi-square table to be 10.6, thus much smaller than G and hence there is a strongly significant difference between times of the day.

d) A model with interaction can be written as $P_{ij} = \frac{\exp(\alpha+\beta_i+\gamma_j+(\alpha\beta)_{ij})}{1+\exp(\alpha+\beta_i+\gamma_j+(\alpha\beta)_{ij})}$.

Here we must require that several of the interaction parameters $(\alpha\beta)_{ij} = 0$. This corresponds to model with different probabilities for each combination of time and species.

However with grouped data the saturated model is actually the one with different probabilities P_{ij} not satisfying the equation in question a). Thus the residual deviance for the interaction model equals 0 and hence the change in deviance G from the model with the main effect to the interaction model is the residual deviance in the table, thus $G = 0.733$. Furthermore there are two residual degrees of freedom left after including the main effects, thus the degrees of freedom for G equals 2. From the table this gives a p-value between 5% and 95%, thus the interaction is not significant (from R: p=0.69).

END

Problem 3

23 patients with acute myeloid leukemia were included in a study. The patient were randomized into 11 given chemotherapy and 12 who were not given this treatment. They were followed until time of remission, that is until they were free of cancer, or to a censoring time.

a) The Kaplan-Meier estimates, in the presence of censored data, the probabilities that survival time exceed different values. (Censored data occur when some lifetimes are only known to exceed given censoring times). Thus when the Kaplan-Meier in one groups lies above another, as for the chemotherapy group compared to the non-chemo group here, then the survival in the chemo group is higher than in the non-chemo. Thus survival is higher in the chemo group.

We estimate medians as the times when the Kaplan-Meiers cross the value 0.5. Here we then get a median about 31 for the chemo group and about 23 for the non-chemo, thus a difference of 9 (months, days, years? I will check) in favor of the chemotherapy group.

b) The Cox proportional hazards model with one covariate is given by $h_x(t) = h_0(t)\exp(betax)$. Here $h_x(t)$ is a hazard function, with the interpretation that the probability of an event in a short interval

(Continued on page 5.)

$(t, t+dt]$ *given* no event before t is approximately $h_x(t)dt$. The $h_0(t)$ is thus the hazard with $x = 0$. We thus get that $\exp(beta) = h_1(t)/h_0(t) = HR$ is the hazard ratio between the exposed ($x = 1$, non-chemo) group and the reference ($x = 0$, chemotherapy). (typo in Problem, have fixed).

Cox-regression gives $\hat{\beta} = 0.916$ and $HR = 2.5$. Thus we estimate that the non-chemotherapy group has 2.5 times higer hazard than the chemotherapy group.

However, when considering the confidence interval for the hazard ratio we find it to be

$exp(\hat{\beta} \pm 1.96se(\hat{\beta})) = \exp(0.916 \pm 1.96 * 0.512) = (0.91, 6.81)$, thus overlapping 1 and we can not conclude that the survival is significantly higher without chemotherapy.