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

The data in this problem comes from an experiment on degradation of dielectric strength of an insulator that has been exposed to different levels of temperature for varying lengths of time. The response variable is dielectric strength in kilo-volts, and the predictor variables are time in weeks and temperature in degrees Celcius. The total number of observations is $n = 128$ and there are four measurements of strength for each combination of temperature (with levels 180, 225, 250 and 275 degrees Celcius) and time (with 8 levels 1, 2, 4, 8, 16, 32 and 64 weeks), thus we have a balanced design. We will work with numerical covariates x_{1i} = degrees Celcius and $x_{2i} = \log(\text{Time})$ and the response variable Y_i = dielectric strength in KVolts for observation no. i.

a) In a first model consider a simple linear regression model $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$ for $i = 1, \ldots, n$ where ε_i are assumed independent and normally distributed with expectation zero and unknown variance σ^2.

   Determine the estimates for β_0, β_1 and σ from the `R`-output on the top of the next page.

   Explain the concept of multiple R-squared.

   Use the output to find the empirical correlation coefficient between the x_{1i} and Y_i.

(Continued on page 2.)

```
> summary(lm(Strength~Temperature))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.511906   1.767079   17.27   <2e-16
Temperature -0.082898   0.007515  -11.03   <2e-16
---
Residual standard error: 2.983 on 126 degrees of freedom
Multiple R-squared:  0.4913,Adjusted R-squared:  0.4872
F-statistic: 121.7 on 1 and 126 DF,  p-value: < 2.2e-16
```

b) Below you see output from a model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, thus with covariate x_{2i} added to the model and the same assumptions about ε_i as in question a).

Explain why the estimate for β_1 is the same as in question a).

Discuss what happened to the estimate of σ and to the multiple R-squared.

Also explain the effect on the standard error of $\hat{\beta}_1$.

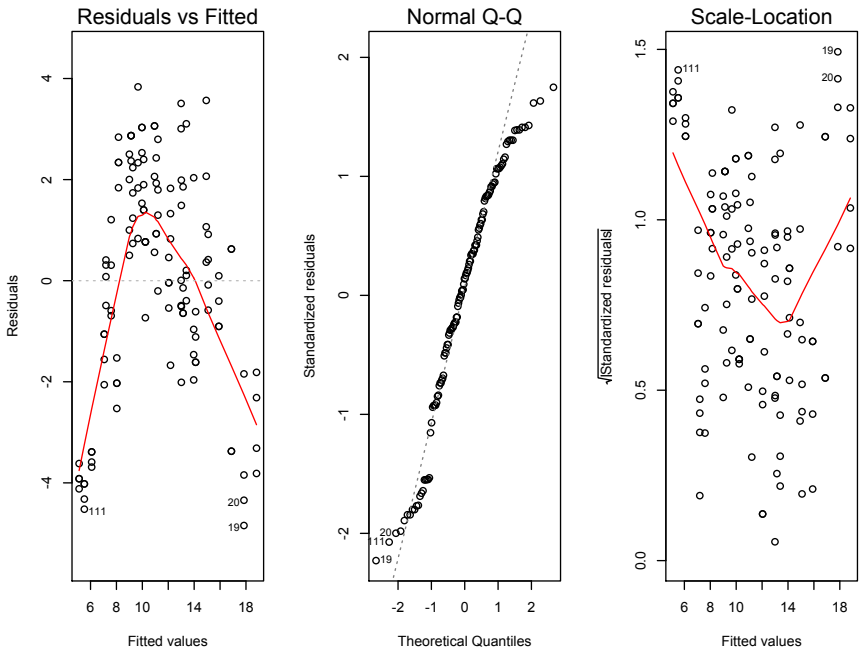
```
> summary(lm(Strength~Temperature+log(Time)))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.735643   1.347989   25.03   <2e-16
Temperature -0.082898   0.005573  -14.87   <2e-16
log(Time)   -1.399549   0.137163  -10.20   <2e-16
---
Residual standard error: 2.212 on 125 degrees of freedom
Multiple R-squared:  0.7224,Adjusted R-squared:  0.718
F-statistic: 162.7 on 2 and 125 DF,  p-value: < 2.2e-16
```

c) Use the residual plots below, that are from the model in question b), to examine the validity of the assumptions for this model.

Do you find room for improvement of the model?

(Continued on page 3.)



d) In order to improve the model one could include an interaction term $x_{i1}x_{i2}$ and second order terms x_{i1}^2 and x_{i2}^2. It will then be useful to apply model selection procedure.

Discuss the problem of using the standard multiple R-squared method as a model selection criterion to choose a model.

Below you see a table over predicted R-square values for various models. Give an explanation of this measure and discuss why it has desirable properties.

Choose a model based on the predicted R-square terms.

Model	Predicted R^2
M1 = Temperature (only)	0.473
M2 = M1 + log(Time)	0.707
M3 = M2 + Temperature $*$ log(Time)	0.820
M4 = M3 + Temperature^2	0.860
M5 = M4 + log(Time)^2	0.862

Problem 2

In this problem you will consider data about passengers aboard the Titanic and investigate factors important in whether they survived or died in the disaster. The outcome variable is whether the passenger died and explanatory variables are age, sex and passenger class. The analyses are restricted to the 756 passengers for whom age is recorded.

(Continued on page 4.)

a) Below you see a table of passengers that survived or died according to sex. Carry out a test for whether the risk of dying was the same for both sexes.

	Survived	Died	Total
Women	217	71	288
Men	96	372	468

b) Define the odds-ratio for mortality between the men and women.

Estimate the odds-ratio based on the table *and* on the R-output below.

Also find an estimate for the relative risk of dying between men and women. Compare with the odds-ratio. Comment on the differences.

```
> glm(Died~Sex,family=binomial)

Coefficients:
(Intercept)      Sexmale
     -1.117        2.472
```

c) The passerenger ages ranged from 0.17 years (infants 2 months) to 74 years. Below you find output from logistic regression model with the binary variable death as outcome and age as covariate. Interpret the regression parameter estimate for age.

Find a 95% confidence interval for the odds ratio of dying between an individual aged 20 and an individual aged 30 years.

Test whether the covariate age is significantly related to mortality.

```
> summary(glm(Died~Age,family=binomial))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.081428   0.173862   0.468   0.6395
Age         0.008795   0.005232   1.681   0.0928
---
    Null deviance: 1025.6  on 755  degrees of freedom
Residual deviance: 1022.7  on 754  degrees of freedom
AIC: 1026.7
```

d) Perhaps the covariate age should be entered into the model in a different way. On the next page you find output for a model with covariate log(Age) instead of Age directly as in question c).

Compare the relation between log(Age) and mortality from this model with the model from the previous question.

Discuss which model best captures the age dependency.

(Continued on page 5.)

```
> summary(glm(Died~log(Age),family=binomial))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8124     0.3462  -2.347 0.018927
log(Age)      0.3586     0.1044   3.435 0.000593
---
    Null deviance: 1025.6  on 755  degrees of freedom
Residual deviance: 1013.0  on 754  degrees of freedom
AIC: 1017
```

e) Below you find a deviance table from models with numerical covariate log(Age) and the categorical covariates Sex (levels Female and Male) and Passenger Class (`PClass`) (with levels 1st, 2nd and 3rd class). In addition a model with interaction between Sex and Passenger class is included. These terms are entered sequentially into the models, thus extending the previous with one covariate per line.

Give a brief description of the deviance test.

Consider the different submodels and conclude about which covariates and whether the interaction should be included in the model.

```
> anova(glm(Died~log(Age)+Sex*PClass,family=binomial))

           Df Deviance Resid. Df Resid. Dev
NULL                         755    1025.57
log(Age)    1   12.558       754    1013.01
Sex         1  225.990       753     787.02
PClass      2   99.369       751     687.66
Sex:PClass  2   30.371       749     657.28
```

f) Below you see the parameter estimates from the full model in question e). Describe what the results tell about the interaction between Sex and Passenger class.

```
> summary(glm(Died~log(Age)+Sex*PClass,family=binomial))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.9545     0.6956  -8.560  < 2e-16 ***
log(Age)            0.8309     0.1419   5.857 4.71e-09 ***
Sexmale             3.6051     0.4987   7.229 4.86e-13 ***
PClass2nd           1.2260     0.5735   2.138 0.032531 *
PClass3rd           3.7319     0.5149   7.247 4.25e-13 ***
Sexmale:PClass2nd   0.2547     0.6575   0.387 0.698521
Sexmale:PClass3rd  -2.1673     0.5760  -3.763 0.000168 ***
```

END