UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK4900 — Statistical methods and applications.

Day of examination: Friday 12. May 2017.

Examination hours: 9.00 – 13.00.

This problem set consists of 7 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.

Exercise 1. A study examined the effects of consuming three different varieties of onions on blood cholesterol levels. Five (Large White Landrace) pigs were randomly assigned to each of the treatments (five for each treatment). Change in cholesterol level was measured after a certain period on the diet. The researchers used an ANOVA in order to understand if type of onion has an effect on the change in cholesterol.

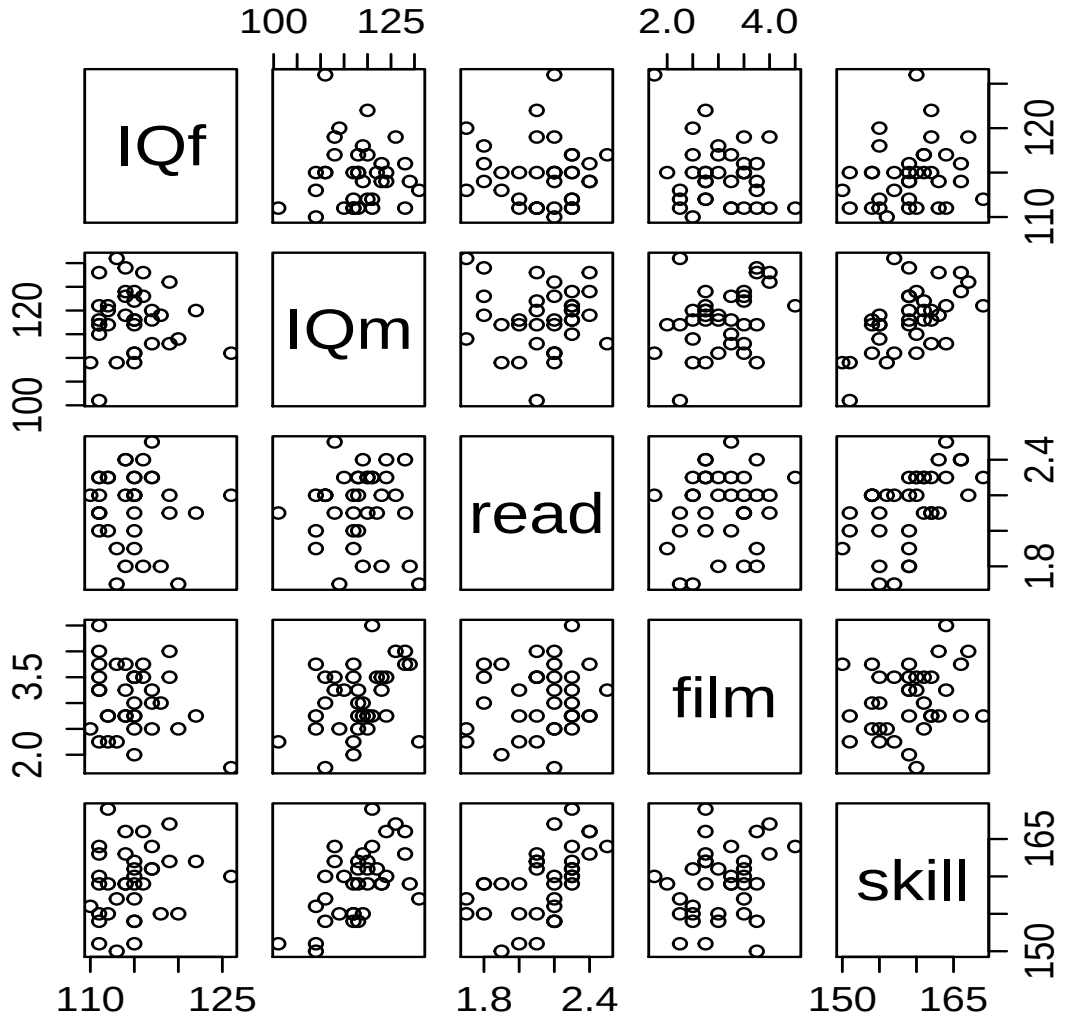
a) For the ANOVA, what is the null hypothesis and alternative hypothesis in this case? What are the degrees of freedom for the model sum of squares, residual sum of squares and total sum of squares?

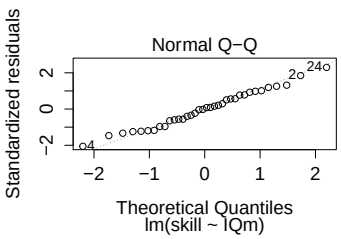
Exercise 2. The following variables were collected for 36 children in an observational study on child development:

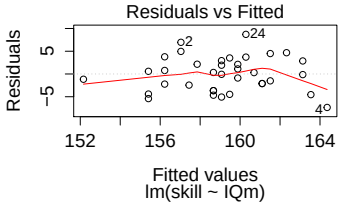
- skill: A score for analytical skills at age 4
- IQf: Father's IQ
- IQm: Mother's IQ
- read: Average number of hours per week the parents read books for the child
- film: Average number of hours per week the child has seen cartoons on TV the last 3 months

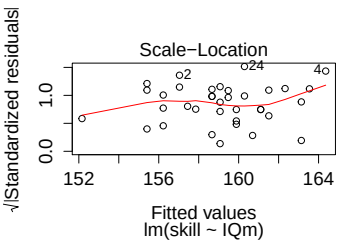
All families were recruited from a group with similar (high) educational level. Analytical skills were measured through a standard psychological test and the score will be used as response (dependent) variable y in the following. A scatterplot of the data can be found on the next page.

(Continued on page 2.)









(Continued on page 3.)

a) In Analysis 0 below, you find the results of fitting a simple linear regression model with mother's IQ as the only covariate. Write down a simple linear regression model, and specify all assumptions you have to make. Use the three diagnostic plots on the previous page to discuss shortly if these assumptions seem reasonable for this dataset. Find estimates for all the parameters in your model from R.

b) In a full analysis, all covariates are included in a multiple regression model (without interaction terms). The results of fitting the full model, are given in Analysis 1 below. Use the results from R below to argue why you would prefer the reduced model in Analysis 2 over the model in Analysis 1. Also, explain the difference between the definition of 'Multiple R-squared' for multiple regression and simple (univariate) regression.

c) In the preferred model (Analysis 2), what is the estimated effect of reading on the expected analytical skills of a 4 years old when adjusting for the parent's IQ scores? Construct a 99% confidence interval for this effect.

```
Analysis 0
Call:
lm(formula = skill ~ IQm, data = skilldata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3569 -2.7497  0.1157  2.8794  8.7091

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.0930    11.8567   9.370 6.02e-11 ***
IQm           0.4066     0.1002   4.058 0.000274 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.856 on 34 degrees of freedom
Multiple R-squared:  0.3263,Adjusted R-squared:  0.3065
F-statistic: 16.47 on 1 and 34 DF,  p-value: 0.000274

Analysis 1
Call:
lm(formula = skill ~ IQf + IQm + read + film, data = skilldata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7980 -1.6520  0.0472  1.5504  7.6081
```

(Continued on page 4.)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.49313   18.93091   2.297   0.0285 *
IQf          0.33290    0.13802   2.412   0.0220 *
IQm          0.41932    0.07631   5.495 5.19e-06 ***
read        12.66239    2.28405   5.544 4.52e-06 ***
film         0.27372    0.82029   0.334   0.7409
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.746 on 31 degrees of freedom
Multiple R-squared:  0.6884,Adjusted R-squared:  0.6482
F-statistic: 17.12 on 4 and 31 DF,  p-value: 1.654e-07

Analysis 2
Call:
lm(formula = skill ~ IQf + IQm + read, data = skilldata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9546 -1.6533  0.1292  1.5138  7.4485

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.36557   18.48732   2.400   0.0224 *
IQf          0.32148    0.13183   2.439   0.0205 *
IQm          0.42825    0.07046   6.078 8.66e-07 ***
read        12.76632    2.23107   5.722 2.43e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.708 on 32 degrees of freedom
Multiple R-squared:  0.6872,Adjusted R-squared:  0.6579
F-statistic: 23.44 on 3 and 32 DF,  p-value: 3.25e-08
```

Exercise 3. Spam emails represent a regular nuisance to most of us. In order to build your own simple spam filter, you collect 250 emails from your own inbox. You tag each of them as spam ($y_i = 1$) or not ($y_i = 0$), $i = 1, \cdots 250$, and register a list of p characteristics for each email. Call these characteristics $x_{1i}, x_{2i} \cdots x_{pi}$, $i = 1, \cdots 250$. Examples of such characteristics can be the presence or not of a certain word (like winner, loan, dollar, naked,...), if there are multiple recipients or not, the number of exclamation marks, etc.

a) Based on the data collected above, which regression model would be natural to use to model the outcome in terms of the possible characteristics? Write down the model. Explain shortly how you would fit such a model, and how you would use it for spam filtering of new emails. Please stick to a general formulation with $x_1, x_2 \cdots x_p$, you are not supposed to discuss what the possible covariates (characteristics) should be etc. here.

(Continued on page 5.)

b) Let the covariate x_1 represent if the email contains the word winner ($x_1 = 1$) or not ($x_1 = 0$). The estimated coefficient for this covariate has been found as $\hat{\beta}_1 = 1.4$. Give a precise interpretation of this value.

Exercise 4. The U.S. has a tradition for giving awards for special achievements to high school students. In a certain high school with 200 students in the same year, we have information on the number of awards each student has received, together with the student's math score and study program. Study programs can be either general (prog=1), academic (prog=2) or vocational (prog=3).
The number of awards students have received, can be summarised in the table below

no. of awards	0	1	2	3	4	5	6
no. observed	124	49	13	9	2	2	1

a) If these counts would follow a Poisson distribution with rate λ, and we estimate λ with the mean number of awards per student (=0.63), the expected numbers (rounded) in the table would be

no. of awards	0	1	2	3	4	5	6
no. observed	107	67	21	4	1	0	0

Perform a test in order to evaluate if the Poisson(λ) distribution fits these counts well.

b) Instead we could use a Poisson model where we let the rate depend on the math score. Below, you find the result of fitting this model to the data using R. What is the effect on the award rate of increasing the math score with one point? Is this effect significant? What is the effect of increasing the math score with 10 points?

c) We can also include study program in the Poisson regression. Explain why and how study program should be included as a factor in the model. Write down the full model and use the parameterization that corresponds to the analysis in R below.

d) Use the results from R to perform a deviance test in order to decide if study program should be a part of the model. Specify the hypotheses, test statistic, degrees of freedom and conclusion.

e) For students with the same math score, in which study program is the award rate largest? Find a 95% confidence interval for the rate ratio for students in the academic program compared to the general program, when math score is kept the same.

(Continued on page 6.)

```
> fitp0 = glm(num_awards~math,family=poisson,data=p)
> summary(fitp0)

Call:
glm(formula = num_awards ~ math, family = poisson, data = p)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1853  -0.9070  -0.6001   0.3246   2.9529

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.333532   0.591261  -9.021   <2e-16 ***
math         0.086166   0.009679   8.902   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 204.02  on 198  degrees of freedom
AIC: 384.08

Number of Fisher Scoring iterations: 6

> fitp = glm(num_awards~math+factor(prog),family=poisson,data=p)
> summary(fitp)

Call:
glm(formula = num_awards ~ math + factor(prog), family = poisson,
    data = p)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2043  -0.8436  -0.5106   0.2558   2.6796

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.24712    0.65845  -7.969 1.60e-15 ***
math           0.07015    0.01060   6.619 3.63e-11 ***
factor(prog)2  1.08386    0.35825   3.025  0.00248 **
factor(prog)3  0.36981    0.44107   0.838  0.40179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: 373.5
```

(Continued on page 7.)

```
Number of Fisher Scoring iterations: 6

> anova(fitp0,fitp,test="Chisq")
Analysis of Deviance Table

Model 1: num_awards ~ math
Model 2: num_awards ~ math + factor(prog)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       198     204.02
2       196     189.45  2   14.572 0.0006852 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The end