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 6 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

Lung function can be measured by FVC (Forced vital capacity), which is the volume of air (in litres) that a person is able to breathe out after full inhalation. In this problem we will discuss data on FVC for 600 children aged 3–5 years. In particular we will study how this measure depends on the height in cm, the weight in kg, age in years and sex coded as 1 for boys and 0 for girls.

a) Below you find output from a simple linear regression model with FVC as outcome variable and height as covariate (later referenced model M1).

Specify the model used in the analysis.

Identify and interpret all parameter estimates in the model.

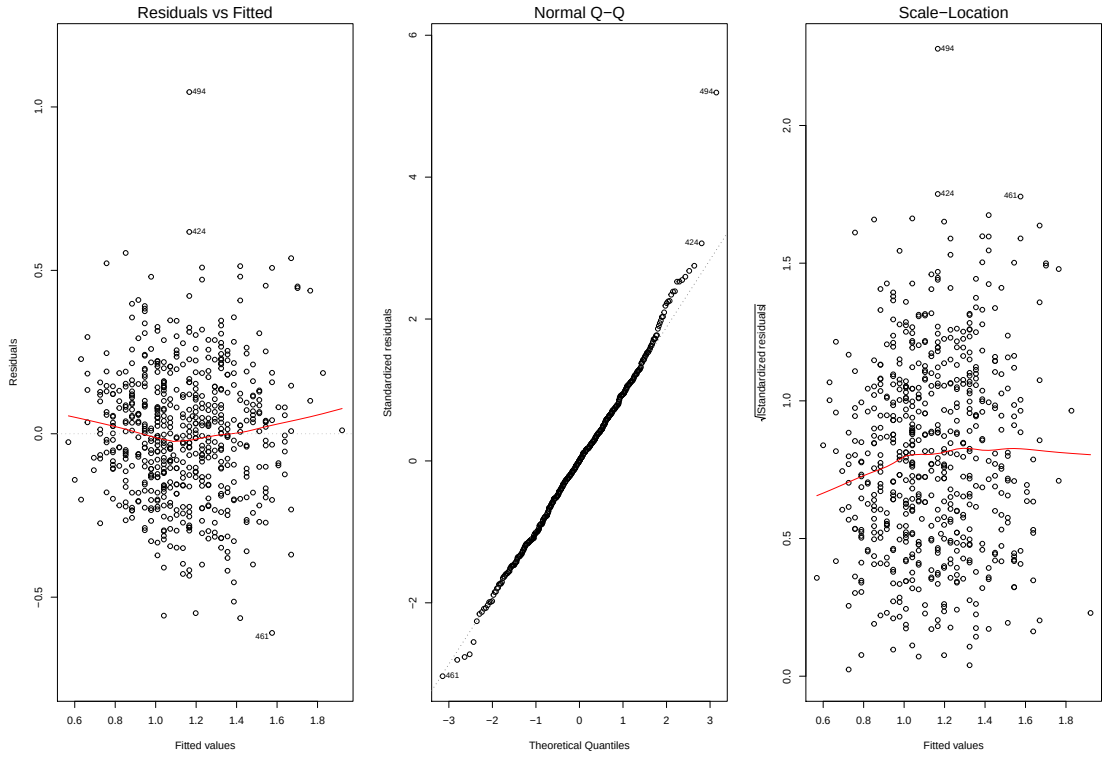
Conclude about the association between height and FVC.

```
Call:
lm(formula = fvc ~ height, data = lungfunction)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.294578   0.118523  -19.36   <2e-16 ***
height       0.031463   0.001082   29.07   <2e-16 ***
---
Residual standard error: 0.2016 on 598 degrees of freedom
Multiple R-squared:  0.5856,    Adjusted R-squared:  0.5849
F-statistic: 844.9 on 1 and 598 DF,  p-value: < 2.2e-16
```

(Continued on page 2.)

b) Residual plots from the analysis in question a) are given below. Explain the plots and discuss what they tell about the adequacy of the model.



c) Define and explain the measures R^2, adjusted R^2 and crossvalidated R^2 and discuss which can be used for model selection.

Below is a table of these three measures for nested linear models for how FVC depend on M1: height alone, M2: height and weight, M3: height, weight and age and M4: height, weight, age and sex. Use the table to choose a model. Give an argument for your choice.

Model	R^2	adjusted R^2	crossvalidated R^2
M1	0.586	0.585	0.583
M2	0.603	0.602	0.600
M3	0.625	0.623	0.620
M4	0.633	0.631	0.627

d) The output on the next page is from model M2. Interpret the estimates.

Explain the difference between the regression estimates for height in M1, $\hat{b}_1 = 0.03146$ (question a)) with the corresponding estimate

(Continued on page 3.)

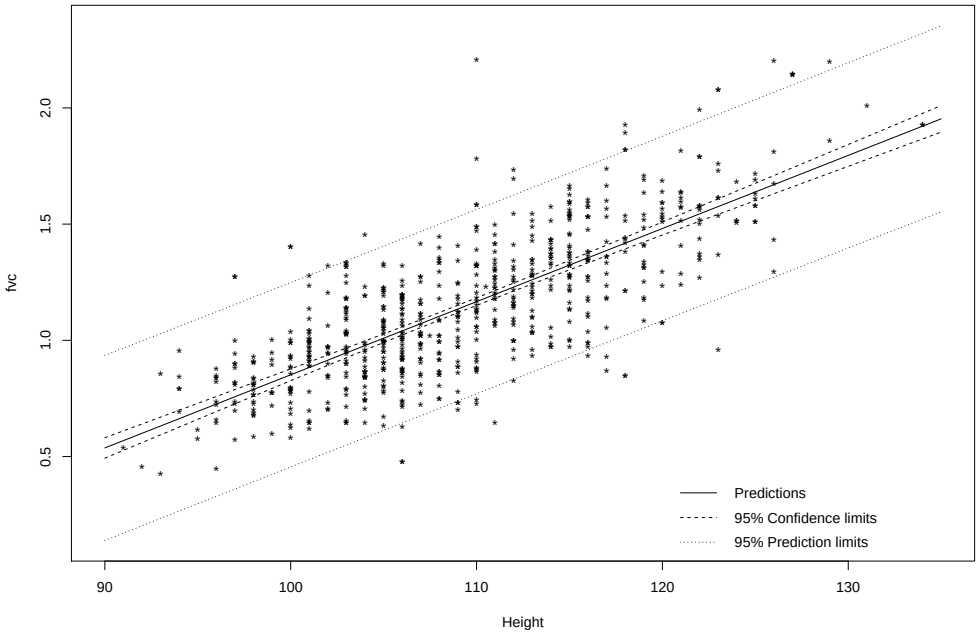
$\hat{\beta}_1 = 0.02610$ in M2. You can use that the empirical correlation between height and weight was equal to 0.701 and that the empirical standard deviations for height and weight were, respectively, 7.61 and 3.26.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.032393   0.126777 -16.031  < 2e-16 ***
height       0.026107   0.001486  17.569  < 2e-16 ***
weight       0.017838   0.003468   5.143 3.66e-07 ***
```

e) The plot below is from model M1 where we use only height to explain FVC. Give the formula for the predicted value of FVC.

Interpret and explain the difference between the 95% confidence intervals and the 95% prediction intervals.

Furthermore compare the prediction intervals from models M1 and M4 at height 110 cm and the other covariates set so that the predictions become approximately equal (see the R code and output below).




```
> newheight=data.frame(height=110)
> predict(lmh,newdata=newheight,interval="pred")
       fit      lwr      upr
1 1.166365 0.770198 1.562532
> newdat=data.frame(height=110,weight=20,age=4.575,sex=0)
> predict(lmhwas,newdata=newdat,interval="pred")
       fit       lwr      upr
1 1.166394 0.7922561 1.540533
```

(Continued on page 4.)

Problem 2

The number of spam mails at one email account were recorded for each hour over a period of 122 consecutive days with observations in each of the five months June to October. Over $n = 2928$ single hours there were recorded totally 3091 spams. The numbers of spam mails per hour are summarized in the following table

Count of spams	0	1	2	3	4	5	6	15	17
Number of hours	1099	1012	532	219	67	19	2	1	1

a) Explain why it may reasonable to consider a Poisson-distribution for the hourly counts of spam emails.

A test for whether the Poisson assumption actually holds can be based on the next table where the observed number of hours O from the table above are repeated in the first column, although with number of spams ≥ 5 lumped together, the second column E is the 'expected' number of hours with specified number of spams given that the Poisson assumption holds and the third column gives the values of $(\text{O-E})^2/\text{E}$.

Explain more carefully how the E's are calculated (using the mean number of spams per hour $\hat{\lambda} = 3091/2928 = 1.05$).

Carry out a test for whether there is evidence of departure from the Poisson assumption.

Counts	O	E	$(\text{O-E})^2/\text{E}$
0	1099	1018.1	6.43
1	1012	1075.2	3.71
2	532	567.7	2.24
3	219	199.8	1.84
4	67	52.8	3.84
5+	23	13.4	6.79

b) One reason for departure from a Poisson assumption could be that the rate depends on background variables or covariates such as hour of the day, day of the week or month. State a Poisson regression model where the rate depends on month only.

Output from fitting such a model is found on the next page. Determine the estimated rate of spams in August (which is the reference level).

Moreover find the estimated rate ratio between the spam rates in October and August.

Calculate also a 95% confidence interval for this rate ratio.

(Continued on page 5.)

```
Call:
glm(formula = spamhour ~ factor(mnth), family = poisson)

Coefficients:
                Estimate Std. Error z value
(Intercept)     -0.03976    0.03740  -1.063  0.28772
factor(mnth)Jul  0.14552    0.05107   2.850  0.00438 **
factor(mnth)Jun -0.02151    0.05981  -0.360  0.71916
factor(mnth)Oct  0.35821    0.06733   5.320 1.04e-07 ***
factor(mnth)Sep  0.10820    0.05192   2.084  0.03717 *
```

c) In an extended analysis we also include hour of the day as a 24 level categorical covariate and weekday as a 7 level categorical covariate. Output (slightly edited) from this model is provided below.

Which of the three categorical covariates hour of the day, week day and month show a significant association with the rate of spams? Explain your answer.

```
Analysis of Deviance Table

Model 1: spamhour ~ 1
Model 2: spamhour ~ factor(mnth)
Model 3: spamhour ~ factor(weekday) + factor(mnth)
Model 4: spamhour ~ factor(hour) + factor(weekday) + factor(mnth)
  Resid. Df Resid. Dev Df Deviance
1      2927     3756.2
2      2923     3720.6  4   35.630
3      2917     3712.8  6    7.757
4      2894     3668.6 23   44.248
```

d) It turned out that the hourly dependence on the spam rate could be well modeled by a second order polynomial. The relevant part of the results are given under (categorical covariates `weekday` and `mnth` are also part of the model). Here hour of the day is a numerical covariate with values 0 to 23.

Describe the variation in spam rate throughout the day. When is the spam rate at the highest and lowest values?

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.1904010  0.0707155   2.692 0.007092 **
hour              -0.0388726  0.0097657  -3.981 6.88e-05 ***
I(hour^2)          0.0015090  0.0004131   3.653 0.000259 ***
```

(Continued on page 6.)

e) With Y_i as the number spam emails for $i = 1, \ldots, n = 2928$ and $\hat{Y}_i$ the corresponding predicted number of spam emails from the model in question d) it turned out that $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/\hat{Y}_i = 3546.2$.

Estimate the overdispersion term relative to the Poisson model.

Explain how standard errors should be corrected for overdispersion and how inference should accordingly be performed.

Discuss if important changes will occur after this modification (detailed calculations are not required).

END