# UNIVERSITY OF OSLO

## Faculty of mathematics and natural sciences

| | |
|---|---|
| Examination in | STK4900/9900 — Statistical methods and applications. |
| Day of examination: | 8 June 2009. |
| Examination hours: | 09.00 – 12.00. |
| This problem set consists of 5 pages. | |
| Appendices: | Data for Problems 1 and 2 |
| 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.

*For completeness the data sets for the problems are given in the appendices. Note, however, that you do not need to inspect the data to solve the problems.*

## Problem 1

Foresters need to be able to asses the amount of timber in a part of a forest. Therefore they need to have a simple and quick method to estimate the volume a tree. It is difficult to estimate the volume of a living tree. But it is fairly easy to measure its height, and even easier to measure its diameter at ground level. Foresters therefore want to have a formula that relates the volume of a tree to its diameter and height.

In Appendix A are given measurements of the diameter, height and volume of a sample of 31 trees from a forest in the US. (These measurements were taken after the trees were cut down.) For each tree the measurements are:

| | |
|---|---|
| DIAMETER | Diameter in inches 4.5 feet above the ground |
| HEIGHT | Height in feet |
| VOLUME | Volume in cubic feet |

We have fitted the following linear regression models to the data:

Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_1$

Model 2: $\lg(y) = \tilde{\beta}_0 + \tilde{\beta}_1 \lg(x_1) + \tilde{\beta}_2 \lg(x_2) + \epsilon_2$

Here $y$ is the VOLUME, $x_1$ is the DIAMETER, and $x_2$ is the HEIGHT of a tree, while lg is the logarithm with base 10.

*(Continued on page 2.)*

The model fitting gave the following results:

**Model 1:**

```
Call:
lm(formula=VOLUME~DIAMETER+HEIGHT)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07
DIAMETER      4.7082     0.2643  17.816  < 2e-16
HEIGHT        0.3393     0.1302   2.607   0.0145

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

**Model 2:**

```
Call:
lm(formula=log10(VOLUME)~log10(DIAMETER)+log10(HEIGHT))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.88007    0.34734  -8.292 5.06e-09
log10(DIAMETER)  1.98265    0.07501  26.432  < 2e-16
log10(HEIGHT)    1.11712    0.20444   5.464 7.81e-06

Residual standard error: 0.03535 on 28 degrees of freedom
Multiple R-squared: 0.9777,     Adjusted R-squared: 0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

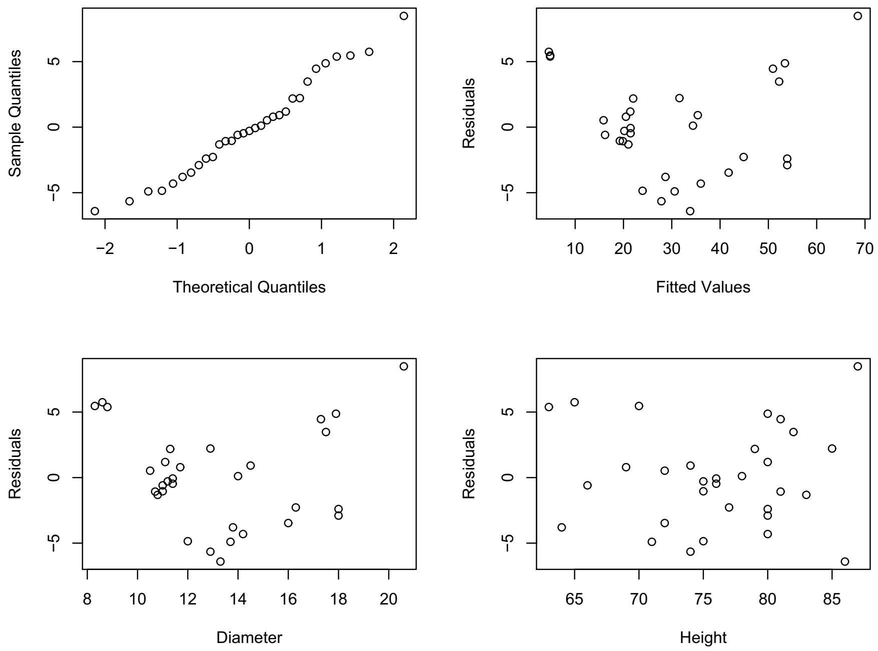
a) Which of the two models would you prefer to use for predicting the volume of a tree from its diameter and height? Give an argument for your answer.

b) Use the model of your choice to predict the volume of a tree with diameter 15 inches and height 70 feet.

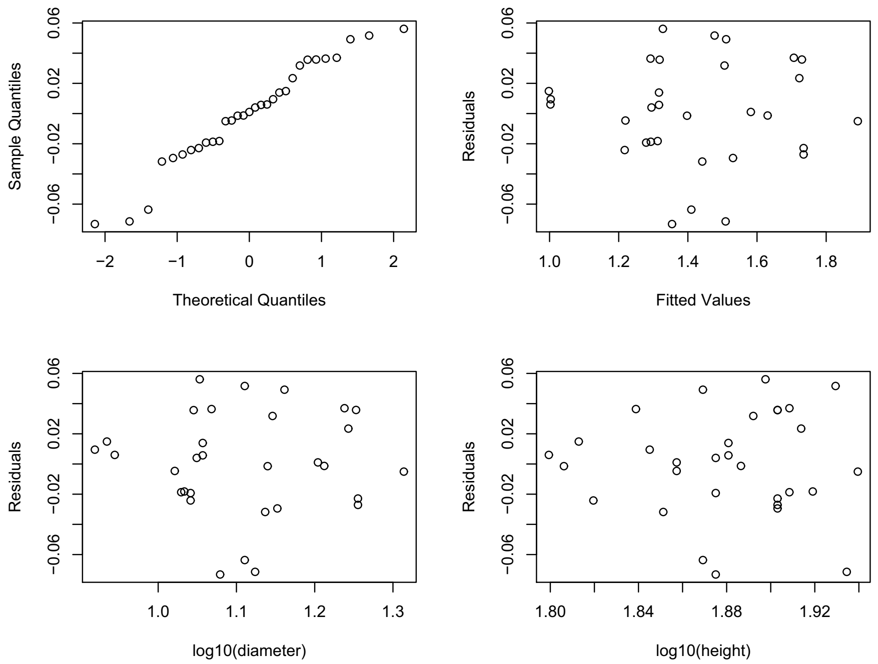
For each of the two models we have made four plots of the residuals. These plots are given on the following page.

*(Continued on page 3.)*

## Residual plots for model 1:




## Residual plots for model 2:




c) Describe the assumptions for the linear regression model, and discuss how the residual plots may be used to check these assumptions for model 1 and model 2.

*(Continued on page 4.)*

# Problem 2

In this problem, we will study data on accidents in a portfolio of private cars in a medium sized English insurance company during a three months period. The data are given in Appendix B and contain the number of insurance claims according to the age of the driver, the motor volume of the car, and the area where the driver lived. The data set also contains information on the number of insured persons in each group (defined by age, motor volume and area).

The variables in the data set are as follows:

| | |
|---|---|
| AGE | Age of the driver (1 = less than 25 year, 2 = 25–29 years, 3 = 30–35 years, 4 = more than 35 years) |
| VOLUME | Motor volume of the car (1 = less than 1 litre, 2 = 1–1.5 litres, 3 = 1.5–2 litres, 4 = more than 2 litres) |
| AREA | Area where the driver lived (4 = London and other big cities, 1–3 = other districts) |
| NUMBER | Number of insured persons in the group (defined by age, motor volume and area) |
| ACCIDENTS | Number of accidents in the group |

a) Explain why it is reasonable to assume that the number of accidents in a given group is Poisson distributed.

The data have been analysed using Poisson regression.
*In this analysis, all covariates have been defined as categorical (i.e. factors).*

First we fitted a model with VOLUME as the only covariate. This gave the results below. (Note that the output has been edited.)

```
Call:
glm(formula=ACCIDENTS~offset(log(NUMBER))+VOLUME,family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.21682    0.04307 -51.467  < 2e-16
VOLUME2      0.14926    0.05045
VOLUME3      0.38865    0.05490
VOLUME4      0.55272    0.07211

    Null deviance: 236.26  on 63  degrees of freedom
Residual deviance: 147.91  on 60  degrees of freedom
```

b) Compute estimates of the rate ratio (RR) for levels 2, 3, and 4 for VOLUME compared to level 1, and describe what these estimates tell you about the effect of motor volume on the risk of accidents.

*(Continued on page 5.)*

c) Derive a 95% confidence interval for the rate ratio of cars with motor volume 1–1.5 litres compared to cars with motor volume less than 1 litre.

We then fitted a model with VOLUME, AGE and AREA as covariates. This gave the following result (edited output):

```
Call:
glm(formula=ACCIDENTS~offset(log(NUMBER))+VOLUME+AGE+AREA,
            family=poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.82174    0.07679 -23.724  < 2e-16
VOLUME2      0.16134    0.05053
VOLUME3      0.39281    0.05500
VOLUME4      0.56341    0.07232
AGE2        -0.19101    0.08286
AGE3        -0.34495    0.08137
AGE4        -0.53667    0.06996
AREA2        0.02587    0.04302   0.601 0.547597
AREA3        0.03852    0.05051   0.763 0.445657
AREA4        0.23421    0.06167   3.798 0.000146

    Null deviance: 236.26  on 63  degrees of freedom
Residual deviance:  51.42  on 54  degrees of freedom
```

d) Describe how the risk of accidents depends on the age of the driver. Is there a significant difference in accidence risk between the two youngest age groups?

e) The estimate of the intercept for the model with VOLUME as the only covariate is smaller than the estimate of the intercept in the model with all three covariates. Explain why this is the case.

Finally we fitted a model with main effects for VOLUME, AGE and AREA and interaction between AGE and AREA. This model has a residual deviance of 40.91.

f) Explain what we mean by interaction between AGE and AREA, and use the results above to test the null hypothesis that there is no interaction between AGE and AREA.

**END**

## Appendix A: Diameter, height and volume of 31 trees

| DIAMETER | HEIGHT | VOLUME |
|---|---|---|
| 8.3 | 70 | 10.3 |
| 8.6 | 65 | 10.3 |
| 8.8 | 63 | 10.2 |
| 10.5 | 72 | 16.4 |
| 10.7 | 81 | 18.8 |
| 10.8 | 83 | 19.7 |
| 11.0 | 66 | 15.6 |
| 11.0 | 75 | 18.2 |
| 11.1 | 80 | 22.6 |
| 11.2 | 75 | 19.9 |
| 11.3 | 79 | 24.2 |
| 11.4 | 76 | 21.0 |
| 11.4 | 76 | 21.4 |
| 11.7 | 69 | 21.3 |
| 12.0 | 75 | 19.1 |
| 12.9 | 74 | 22.2 |
| 12.9 | 85 | 33.8 |
| 13.3 | 86 | 27.4 |
| 13.7 | 71 | 25.7 |
| 13.8 | 64 | 24.9 |
| 14.0 | 78 | 34.5 |
| 14.2 | 80 | 31.7 |
| 14.5 | 74 | 36.3 |
| 16.0 | 72 | 38.3 |
| 16.3 | 77 | 42.6 |
| 17.3 | 81 | 55.4 |
| 17.5 | 82 | 55.7 |
| 17.9 | 80 | 58.3 |
| 18.0 | 80 | 51.5 |
| 18.0 | 80 | 51.0 |
| 20.6 | 87 | 77.0 |

## Appendix B: Data on accidents of private cars

| AGE | VOLUME | AREA | NUMBER | ACCIDENTS |
|---|---|---|---|---|
| 1 | 1 | 1 | 197 | 38 |
| 2 | 1 | 1 | 264 | 35 |
| 3 | 1 | 1 | 246 | 20 |
| 4 | 1 | 1 | 1680 | 156 |
| 1 | 2 | 1 | 284 | 63 |
| 2 | 2 | 1 | 536 | 84 |
| 3 | 2 | 1 | 696 | 89 |
| 4 | 2 | 1 | 3582 | 400 |
| 1 | 3 | 1 | 133 | 19 |
| 2 | 3 | 1 | 286 | 52 |
| 3 | 3 | 1 | 355 | 74 |
| 4 | 3 | 1 | 1640 | 233 |
| 1 | 4 | 1 | 24 | 4 |
| 2 | 4 | 1 | 71 | 18 |
| 3 | 4 | 1 | 99 | 19 |
| 4 | 4 | 1 | 452 | 77 |
| 1 | 1 | 2 | 85 | 22 |
| 2 | 1 | 2 | 139 | 19 |
| 3 | 1 | 2 | 151 | 22 |
| 4 | 1 | 2 | 931 | 87 |
| 1 | 2 | 2 | 149 | 25 |
| 2 | 2 | 2 | 313 | 51 |
| 3 | 2 | 2 | 419 | 49 |
| 4 | 2 | 2 | 2443 | 290 |
| 1 | 3 | 2 | 66 | 14 |
| 2 | 3 | 2 | 175 | 46 |
| 3 | 3 | 2 | 221 | 39 |
| 4 | 3 | 2 | 1110 | 143 |
| 1 | 4 | 2 | 9 | 4 |
| 2 | 4 | 2 | 48 | 15 |
| 3 | 4 | 2 | 72 | 12 |
| 4 | 4 | 2 | 322 | 53 |
| 1 | 1 | 3 | 35 | 5 |
| 2 | 1 | 3 | 73 | 11 |
| 3 | 1 | 3 | 89 | 10 |
| 4 | 1 | 3 | 648 | 67 |
| 1 | 2 | 3 | 53 | 10 |
| 2 | 2 | 3 | 155 | 24 |
| 3 | 2 | 3 | 240 | 37 |
| 4 | 2 | 3 | 1635 | 187 |

## Appendix B, continued

| AGE | VOLUME | AREA | NUMBER | ACCIDENTS |
|---|---|---|---|---|
| 1 | 3 | 3 | 24 | 8 |
| 2 | 3 | 3 | 78 | 19 |
| 3 | 3 | 3 | 121 | 24 |
| 4 | 3 | 3 | 692 | 101 |
| 1 | 4 | 3 | 7 | 3 |
| 2 | 4 | 3 | 29 | 2 |
| 3 | 4 | 3 | 43 | 8 |
| 4 | 4 | 3 | 245 | 37 |
| 1 | 1 | 4 | 20 | 2 |
| 2 | 1 | 4 | 33 | 5 |
| 3 | 1 | 4 | 40 | 4 |
| 4 | 1 | 4 | 316 | 36 |
| 1 | 2 | 4 | 31 | 7 |
| 2 | 2 | 4 | 81 | 10 |
| 3 | 2 | 4 | 122 | 22 |
| 4 | 2 | 4 | 724 | 102 |
| 1 | 3 | 4 | 18 | 5 |
| 2 | 3 | 4 | 39 | 7 |
| 3 | 3 | 4 | 68 | 16 |
| 4 | 3 | 4 | 344 | 63 |
| 1 | 4 | 4 | 3 | 0 |
| 2 | 4 | 4 | 16 | 6 |
| 3 | 4 | 4 | 25 | 8 |
| 4 | 4 | 4 | 114 | 33 |