

UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK4900 — Statistical methods and applications.

Day of examination: 8 June 2010.

Examination hours: 09.00–12.00.

This problem set consists of 5 pages.

Appendices: Tables for the standard normal distribution, the chi-square distributions, the t distributions, and the 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

In this problem we will study the influence that some variables have on the birth weight of a child. Our analysis is based on a sample of 500 birth weights from pregnancies where the pregnancy lasted at least 38 weeks.

The response variable, `WEIGHT`, is the weight of a child (in kg), and the covariates we will consider are the following:

`SEX` Sex of child (0: boy; 1: girl)
`WEEKS` Length of pregnancy (in weeks)
`AGE` Age of mother at start of pregnancy (in years)
`FIRST` Firstborn child or not (0: firstborn child; 1: not firstborn child)

First we fit a linear regression model with the covariates `SEX`, `WEEKS`, and `AGE` (model 1). This gives the following result:

Model 1:

Call:

```
lm(formula=WEIGHT~SEX+WEEKS+AGE)
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | -3.064382 | 0.909318 | -3.370 | 0.00081 |
| SEX | -0.118337 | 0.038491 | -3.074 | 0.00223 |
| WEEKS | 0.161909 | 0.022345 | 7.246 | 1.65e-12 |
| AGE | 0.008538 | 0.003985 | | |

(edited output)

(Continued on page 2.)

- a) Use an appropriate hypothesis test to decide if there is a significant effect of the age of the mother.

We then fit a model where also the covariate `FIRST` is taken into account (model 2):

Model 2:

Call:

```
lm(formula=WEIGHT~SEX+WEEKS+AGE+FIRST)
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | -2.941529 | 0.896679 | -3.280 | 0.00111 |
| SEX | -0.115891 | 0.037938 | -3.055 | 0.00237 |
| WEEKS | 0.160624 | 0.022023 | 7.293 | 1.21e-12 |
| AGE | 0.002061 | 0.004254 | | |
| FIRST | 0.164852 | 0.041616 | 3.961 | 8.55e-05 |

(edited output)

- b) Is there a significant effect of the age of the mother in this model? Discuss why model 1 and model 2 give different estimates for the effect of mother's age.

Finally we fit a model with the three covariates `SEX`, `WEEKS`, and `FIRST` (model 3):

Model 3:

Call:

```
lm(formula=WEIGHT~SEX+WEEKS+FIRST)
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -2.85704 | 0.87888 | -3.251 | 0.00123 |
| SEX | -0.11428 | 0.03776 | -3.026 | 0.00261 |
| WEEKS | 0.15971 | 0.02193 | 7.284 | 1.28e-12 |
| FIRST | 0.17260 | 0.03839 | 4.496 | 8.63e-06 |

(edited output)

- c) Give an interpretation of the effects of `SEX`, `WEEKS`, and `FIRST`.
- d) Compute a 95% confidence interval for the effect of `SEX` and give an interpretation of the confidence interval.
- e) Predict the weight of a newborn girl who is the second child of her mother, and where the length of the pregnancy is 40 weeks.

(Continued on page 3.)

Problem 2

Byssinosis, also called “brown lung disease”, is a chronic asthma-like disease of the lungs caused by breathing in cotton dust or dusts from other vegetable fibers. In this problem we will look at data from a study of workers in the cotton industry. The variable of interest is presence of byssinosis, and we will relate this to the following categorical covariates (factors):

| | |
|--------|--|
| DUST | Dustiness of the workplace (1: high; 2: medium; 3: low) |
| EMPLOY | Length of employment in the cotton industry (1: less than 10 years; 2: between 10 and 20 years; 3: more than 20 years) |
| SMOKE | Smoking status (1: smoker, 2: nonsmoker) |

For all combinations of the levels of the three factors, we know the number of workers who suffer from byssinosis (NOBYS) as well as the total number of workers (NOTOT).

- a) Explain why logistic regression is an appropriate model for analysing the data. Give an explicit formulation of the logistic regression model when we only consider the factor DUST.

When we fit the logistic regression model with DUST as the only covariate (model 1), we get the result:

Model 1:

Call:

```
glm(formula=cbind(NOBYS,NOTOT-NOBYS)~factor(DUST),family=binomial)
```

| | Estimate | Std. Error | z value | Pr(> z) |
|---------------|----------|------------|---------|----------|
| (Intercept) | -1.6811 | 0.1063 | -15.817 | <2e-16 |
| factor(DUST)2 | -2.5847 | 0.2601 | -9.939 | <2e-16 |
| factor(DUST)3 | -2.7151 | 0.1881 | -14.431 | <2e-16 |

| | | | |
|--------------------|---------|-------|--------------------|
| Null deviance: | 290.739 | on 17 | degrees of freedom |
| Residual deviance: | 38.630 | on 15 | degrees of freedom |

(edited output)

- b) Describe how the dustiness of the workplace influences the probability that a worker will suffer from byssinosis.

(Continued on page 4.)

Next we fit a model with the covariates DUST and EMPLOY (model 2):

Model 2:

Call:

```
glm(formula=cbind(NOBY5,NOTOT-NOBY5)~factor(DUST)+factor(EMPLOY),
     family=binomial)
```

| | Estimate | Std. Error | z value | Pr(> z) |
|-----------------|----------|------------|---------|----------|
| (Intercept) | -2.0146 | 0.1446 | -13.934 | < 2e-16 |
| factor(DUST)2 | -2.6083 | 0.2608 | -10.002 | < 2e-16 |
| factor(DUST)3 | -2.7613 | 0.1893 | -14.589 | < 2e-16 |
| factor(EMPLOY)2 | 0.5643 | 0.2479 | 2.277 | 0.022810 |
| factor(EMPLOY)3 | 0.6732 | 0.1808 | 3.724 | 0.000196 |

Null deviance: 290.739 on 17 degrees of freedom
Residual deviance: 23.527 on 13 degrees of freedom

(edited output)

- c) Define the odds ratio between workers who have been employed between 10 and 20 years and those who have been employed less than 10 years. Estimate the odds ratio and derive a 95% confidence interval for it. Describe what the estimated odds ratio and the confidence interval tell you.
- d) Estimate the odds ratio between workers who have been employed more than 20 years and those who have been employed between 10 and 20 years, and interpret the estimate. (No confidence interval is required.)

Finally we fit a model with the three covariates DUST, EMPLOY, and SMOKE (model 3):

(Continued on page 5.)

Model 3:

Call:

```
glm(formula=cbind(NOBY,NOTOT-NOBY)~factor(DUST)+factor(EMPLOY)
     +factor(SMOKE),family = binomial)
```

| | Estimate | Std. Error | z value | Pr(> z) |
|-----------------|----------|------------|---------|----------|
| (Intercept) | -1.8336 | 0.1525 | -12.026 | < 2e-16 |
| factor(DUST)2 | -2.5493 | 0.2614 | -9.753 | < 2e-16 |
| factor(DUST)3 | -2.7175 | 0.1898 | -14.314 | < 2e-16 |
| factor(EMPLOY)2 | 0.5060 | 0.2490 | 2.032 | 0.042119 |
| factor(EMPLOY)3 | 0.6728 | 0.1813 | 3.710 | 0.000207 |
| factor(SMOKE)2 | -0.6210 | | | |

Null deviance: 290.739 on 17 degrees of freedom
Residual deviance: 12.094 on 12 degrees of freedom

(edited output)

- e) Use an appropriate hypothesis test to decide if smoking has a significant effect on the risk of suffering from byssinosis.