

# UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4900/9900 — Statistical Methods  
and Applications

Day of examination: Thursday 11. May 2023

Examination hours: 9.00–13.00

This problem set consists of 12 pages.

Appendices: Tables for normal,  $t$ -,  $\chi^2$ - and  $F$ -distributions

Permitted aids: All printed, 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

**a** (weight 10p)

Based on field experiments, a new variety of greengram is expected to give an yield of 13 quintals per hectare. The variety was tested on 12 randomly selected farmer fields. The yields (quintal/hectare) were recorded as 14.3, 12.6, 13.7, 10.9, 13.7, 12.0, 11.4, 12.0, 13.1, 12.6, 13.4, 13.1. Do the results conform the expectation? (Use 5 percent level of significance)

*Sketch of the solution Problem 1a.*

Null hypothesis  $H_0$ : mean is equal to a specific value  $\mu_0$ , the result conform the expectation, i.e.

$$H_0 : \mu = \mu_0 = 13$$

The test statistic become

$$t = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}}$$

where  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $s = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right)}$  is unbiased estimate of  $\sigma$ . Then

$$\bar{x} = \frac{152,8}{12} = 12,73;$$

$$s = \sqrt{\frac{1}{11} \left( 1956,94 - \frac{(152,8)^2}{12} \right)} = 1,01;$$

(Continued on page 2.)

$$t = \frac{|12,73 - 13|}{\frac{1,01}{\sqrt{12}}} = \frac{0,27}{0,29} = 0,93$$

$t$ -table value at  $(n - 1) = 11$  d.f. at 5 percent level of significance is 2,20. Calculated value of  $t < \text{table value of } t$ ,  $H_0$  is not rejected and we may conclude that the results conform to the expectation.

**b** (weight 20p)

A study to assess the capability of subsurface flow wetland systems to remove biochemical oxygen demand (**BOD**) and various other chemical constituents resulted in the accompanying data on  $x = \text{BOD mass loading}$  (kg/ha/d) and  $y = \text{BOD mass removal}$  (kg/ha/d).

|     |   |   |    |    |    |    |    |    |    |    |    |    |     |     |
|-----|---|---|----|----|----|----|----|----|----|----|----|----|-----|-----|
| $x$ | 3 | 8 | 10 | 11 | 13 | 16 | 27 | 30 | 35 | 37 | 38 | 44 | 103 | 142 |
| $y$ | 4 | 7 | 8  | 8  | 10 | 11 | 16 | 26 | 21 | 9  | 31 | 30 | 75  | 90  |

Values of relevant summary quantities are

$$\sum_{i=1}^{14} x_i = 517; \quad \sum_{j=1}^{14} y_j = 346; \quad \sum_{i=1}^{14} x_i^2 = 39095; \quad \sum_{j=1}^{14} y_j^2 = 17454; \quad \sum_{i=1}^{14} x_i y_i = 25825.$$

1. Construct boxplots of both mass loading and mass removal and comment on any interesting features.
2. Obtain the equation of the least squares line.
3. Predict the value of **BOD mass removal** for a single observation made when **BOD mass loading** is 35, and calculate the value of the corresponding residual. The residual should be computed with respect to the observation given in the table. Calculate RSS (*residual sum of squares*) and then a point estimate of  $\sigma$ .
4. What proportion of observed variation in removal can be explained by the approximate linear relationship between the two variables? The last two  $x$  values, 103 and 142, are much larger than the others. How are the equation of the least squares line and the value of the coefficient of determination affected by deletion of the two corresponding observations from the sample?

*Sketch of the solution Problem 1b.*

1. To create boxplot we need median ( $M = Q_2$ ), quartile 1 ( $Q_1$ ), quartile 3 ( $Q_3$ ).

Quartiles divide the data set into four equal segments after it has been arranged in ascending order. Approximately 25 percent of the data points will fall below the first quartile,  $Q_1$ . Approximately 50 percent of the data points will fall below the second

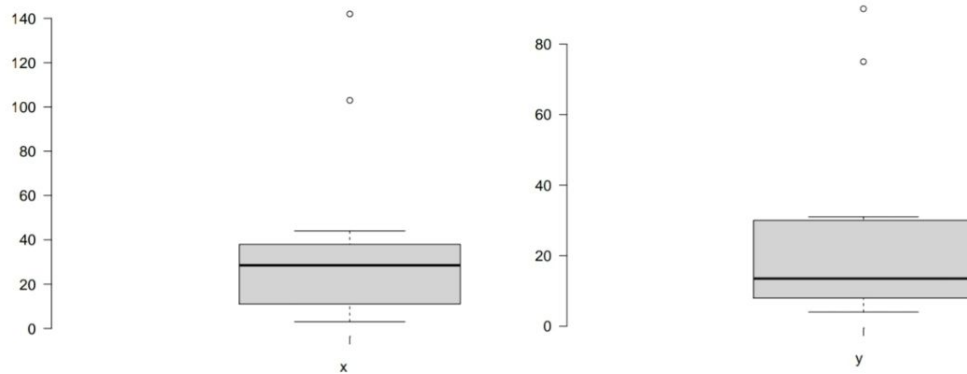
(Continued on page 3.)

quartile (median),  $M = Q_2$ . And 75 percent should fall below the third quartile,  $Q_3$ . Both samples have outliers:  $x_{13} = 103, x_{14} = 142$  and  $y_{13} = 75, y_{14} = 90$ .

Then,

$$M_x = Q_{2-x} = \frac{27 + 30}{2} = 28,5 \text{ and } M_y = Q_{2-y} = \frac{11 + 16}{2} = 13,5$$

$$Q_{1-x} = 11, Q_{3-x} = 38, Q_{1-y} = 8, Q_{3-y} = 30.$$



2. We will work with the simple linear regression model:

$$y = \beta_0 + \beta_1 \cdot x + \epsilon,$$

where  $y$  is **BOD mass removal**,  $x$  is **BOD mass loading**.

The least squares estimate for the slope is given by:

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x}$$

where

$$r = \frac{\sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{s_x \cdot s_y}$$

is the Pearson correlation coefficient, and  $s_x$  and  $s_y$  are the empirical standard deviations of the  $x$  and the  $y$ .

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x} = \frac{\sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{s_x \cdot s_y} \cdot \frac{s_y}{s_x} = \frac{\frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

where

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2 \cdot \bar{x} \cdot \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 =$$

(Continued on page 4.)

$$= \sum_{i=1}^n x_i^2 - \frac{2}{n} \cdot \left( \sum_{i=1}^n x_i \right)^2 + \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2.$$

Similarly,

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i$$

Thus,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2}$$

By substituting value we have

$$\hat{\beta}_1 = \frac{25825 - \frac{1}{14} \cdot 517 \cdot 346}{39095 - \frac{1}{14} (517)^2} = \frac{13047,714}{20002,93} \approx 0,65$$

It is possible to find  $\hat{\beta}_1$  an alternative way

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right)}$$

Then

$$s_x = \sqrt{\frac{1}{14-1} \cdot \left( 39095 - \frac{1}{14} \cdot (517)^2 \right)} \approx 39,22.$$

In a similar way, we get

$$s_y = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^n y_i^2 - \frac{1}{n(n-1)} \cdot \left( \sum_{i=1}^n y_i \right)^2 \right)},$$

$$s_y = \sqrt{\frac{1}{13} \cdot \left( 17454 - \frac{1}{14} \cdot (346)^2 \right)} \approx 26,17.$$

Let's represent  $r$

$$r = \frac{\sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{s_x \cdot s_y} = \frac{1}{s_x \cdot s_y} \cdot \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right) =$$

Thus

$$r = \frac{1}{39,22 \cdot 26,17 \cdot 13} \left( 25825 - \frac{1}{14} \cdot 517 \cdot 346 \right) \approx 0,98.$$

Finally, we obtain the same result

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x} = 0,98 \cdot \frac{26,17}{39,22} \approx 0,65$$

(Continued on page 5.)

Now we can find

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

then  $\hat{\beta}_0 = 0,63$ . Linear regression model:

$$y = 0,63 + 0,65x + \epsilon.$$

**3.** To find predicted value, we will use formula  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ , where  $x = 35$ . Substituting  $x$  we obtain

$$\hat{y} = 0,63 + 0,65 \cdot 35 = 23,46.$$

By definition  $r_i = y_i - \hat{y}_i$  are residuals. The residuals are estimates of the unobserved  $\epsilon$ . As  $x = 35$  can be found in the sample (it is  $x_9$ ), and corresponding value of  $y = 21$ , we can calculate residual as

$$\begin{aligned} r_9 &= y_9 - \hat{y}, \\ r_9 &= 21 - 23,46 = -2,46. \end{aligned}$$

Residual sum of squares:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

and the least squares estimate of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{RSS}{n-2}.$$

**4.** For the simple linear regression model  $R^2$  is just the square of the Pearson correlation coefficient:

$$R^2 = r^2 = (0,98)^2 = 0,96.$$

Let's find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  without last two  $x$  values, 103 and 142.

$$\sum_{i=1}^{12} x_i = 517 - 103 - 142 = 272; \quad \sum_{j=1}^{12} y_j = 346 - 75 - 90 = 181;$$

$$\sum_{i=1}^{12} x_i^2 = 39095 - 103^2 - 142^2 = 8322;$$

$$\sum_{j=1}^{12} y_j^2 = 17454 - 5625 - 8100 = 3729; \quad \sum_{i=1}^{12} x_i y_i = 25825 - 103 \cdot 75 - 148 \cdot 90 = 5320.$$

Then

$$\hat{\beta}_1 = \frac{5320 - \frac{1}{12} \cdot 272 \cdot 181}{8322 - \frac{1}{12} (272)^2} \approx 0,564;$$

$$\hat{\beta}_0 = 2,29.$$

$$r^2 = 0,69$$

Linear regression model:

$$y = 2,29 + 0,564x + \epsilon.$$

(Continued on page 6.)

| Parameters      | n=14 | n=12  |
|-----------------|------|-------|
| $\hat{\beta}_0$ | 0,63 | 2,29  |
| $\hat{\beta}_1$ | 0,65 | 0,564 |
| $r^2$           | 0,96 | 0,69  |

The intercept grows significantly. Slope is slightly decreased. Changes in the coefficient of determination shows that the new regression model explains only 69% of  $y$  variation, comparing the 96% of the initial model. So it is less efficient in explaining  $y$ .

## Problem 2

In this problem you will consider data about passengers aboard the Cruise Liner 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.

**a** (weight 10p)

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   |

*Sketch of the solution Problem 2a.*

With  $p_w$  the probability that a woman died and  $p_m$  the probability that a man died we want to test  $H_0 : p_w = p_m$ . This can be done using a test statistic

$$Z = \frac{\hat{p}_m - \hat{p}_w}{se(\hat{p}_m - \hat{p}_w)},$$

where  $\hat{p}_m = \frac{372}{468} = 0,795$  and  $\hat{p}_w = \frac{71}{288} = 0,247$  are the proportions dying among men and women and

$$se(\hat{p}_m - \hat{p}_w) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{\frac{1}{468} + \frac{1}{288}}}$$

is the standard error of  $\hat{p}_m - \hat{p}_w$  under the null that the mortality is the same  $p_w = p_m = p$  and

$$\hat{p} = \frac{(71 + 372)}{(288 + 468)} = 0,585.$$

Under the null hypothesis  $Z$  is drawn from a standard normal distribution, thus if  $|Z| > 1,96$  we can reject the null at a 5% level. Here we find  $Z = 14,86$  and so we reject the null with a very small p-value  $< 0,001$ .

(Continued on page 7.)

**b** (weight 10p)

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   |

*Sketch of the solution Problem 2b.*

The odds-ratio of dying between men and women is given by

$$OR = \frac{p_m/(1-p_m)}{p_w/(1-p_w)}.$$

We can estimate it

$$OR = \frac{p_m/(1-p_m)}{p_w/(1-p_w)} = 11,8.$$

However, this odds-ratio is also given as  $\exp(\hat{\beta}_1)$  from the logistic regression with outcome  $Y = \text{indicator of dying}$  and  $x = \text{indicator of man}$ , thus again  $OR = \exp(2,472) = 11,8$ . The relative risk of dying between men and women becomes  $RR = p_m / p_w = \frac{0,795}{0,247} = 3,2$  which is considerably smaller than the odds-ratio. The odds-ratio is a good approximation to the relative risk when the probabilities are both small, but here they are both large and there is a striking difference. Note, however, that if the  $\hat{p}_m = \hat{p}_w$ , then we have both OR and RR equal to 1 and if one of them is larger (smaller) than 1 then so is the other.

**c** (weight 10p)

The passenger 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.

*Sketch of the solution Problem 2 c.*

The model is given as,  $Y = 0$  for a survivor and  $Y = 1$  for passenger that died and  $x = \text{age in years}$ ,

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

(Continued on page 8.)

```
> 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
```

We have the odds-ratio interpretation

$$OR = \frac{p(x+1) \setminus (1-p(x+1))}{p(x) \setminus (1-p(x))} = \exp(\beta_1);$$

thus the regression parameter for age  $\hat{\beta}_1$  has the interpretation as the estimated odds-ratio of dying between two individuals with a one year difference in age. For two individuals where one is 10 years older than the other we similarly have an odds-ratio given as

$$\exp(10\hat{\beta}_1) = \exp(0,08795) = 1,092,$$

thus an increase in odds of 9.2%. The confidence interval for this odds-ratio  $\exp(10\hat{\beta}_1)$  is given by the formula

$$\exp(10\hat{\beta}_1 \pm 1,96 \cdot 10se(\hat{\beta}_1)) = (0,986; 1,210).$$

Since this confidence interval overlaps  $OR = 1$  the risk of dying was not significantly associated with age. This test could also be done considering the test statistic

$$Z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0,008795}{0,005232} = 1,681$$

which is less than the 97,5 percentile 1,96 in the standard normal distribution, thus the p-value is above 0,05.

**d** (weight 10p)

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  $\log(\text{Age})$  instead of Age directly as in question c). Compare the relation between  $\log(\text{Age})$  and mortality from this model with the model from the previous question. Discuss which model best captures the age dependency.

(Continued on page 9.)

```
> 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
```

*Sketch of the solution Problem 2 d.*

The model with age entered directly did not show significant association between age and mortality. In contrast the model with  $\log(\text{Age})$  has a p-value  $< 0,001$  for the regression parameter being different from zero, thus there is a strongly significant correspondence between log-age and mortality. So clearly log-age is the better variable. This is also reflected in lower residual deviance and higher AIC-values for the log-age model than for the model with age entered without transformation.

**e** (weight 10p)

Below you find a deviance table from models with numerical covariate  $\log(\text{Age})$  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     |

*(Continued on page 10.)*

*Sketch of the solution Problem 2 e.*

The residual deviance is twice the difference between the log-likelihood of the actual model and the loglikelihood of the "saturated" model with "estimated" probabilities  $\tilde{p}_i = Y_i$ . Denoting  $D_1$  and  $D_2$  the deviances for two nested models  $M_1$  and  $M_2$ , and where  $M_1$  is a special case of  $M_2$ , we have that the log-likelihood ratio statistic  $G = D_2 - D_1$  is approximately chi-square distributed assuming  $M_1$  as a null hypothesis. The degrees of freedom for  $G$  is equal to the difference of parameters in the two models. From the deviance table we see that log(Age), Sex have  $G$ -statistics of 12,55 and 225,99 which is well above the 95 percentile of a chi-square distribution with one degree of freedom which equals 3,84 ( $= 1,96^2$ ). Furthermore the  $G$ -statistic for PClass equals 99,4 also well above the 95 percentile of a chi-square distribution with 2 degrees of freedom which equals 5,99, so again significant. We use 2 degrees of freedom here because PClass is a categorical covariate with three levels. Finally we see that also the interaction between Sex and PClass is significant since  $G = 30,4 > 5,99$  with 2 degrees of freedom.

### Problem 3

Here is data on launch temperature and the incidence of failure for drill in 23 experiments for NASA's Volatiles Investigating Polar Exploration Rover (VIPER). VIPER uses drill to detect and analyze various lunar soil environments at a range of depths and temperatures.

| Temperature | Failure | Temperature | Failure | Temperature | Failure |
|-------------|---------|-------------|---------|-------------|---------|
| 53          | Y       | 68          | N       | 75          | N       |
| 57          | Y       | 69          | N       | 75          | Y       |
| 58          | Y       | 70          | N       | 76          | N       |
| 63          | Y       | 70          | N       | 76          | N       |
| 66          | N       | 70          | Y       | 78          | N       |
| 67          | N       | 70          | Y       | 79          | N       |
| 67          | N       | 72          | N       | 81          | N       |
| 67          | N       | 73          | N       |             |         |

We have chosen to let  $p$  denote the probability of failure.

**a** (weight 10p)

Choose a suitable regression model for studying how the probability of failure for drill depends on the temperature. Give the reasons for your choice of regression model.

*Sketch of the solution Problem 3 a.*

A binomial regression model is suitable for studying this data because of the binary outcome, that is either 0 or 1. The model should give us the probability of failure.

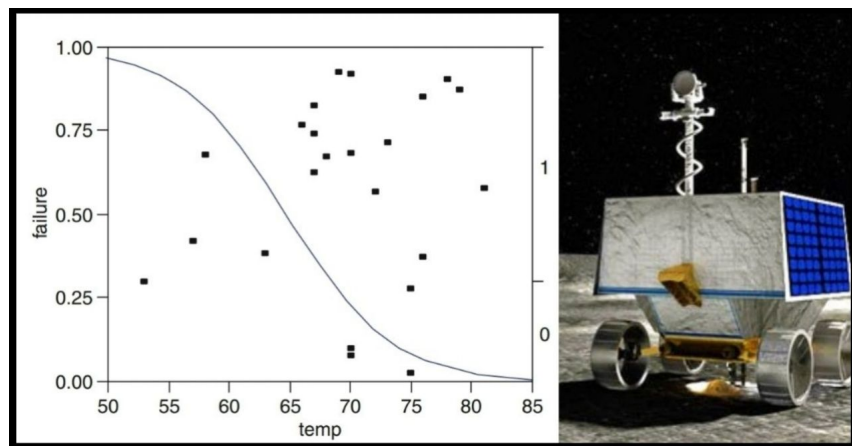
$$p(x) = \frac{e^{\beta_0 + \beta_1 \cdot x}}{1 + e^{\beta_0 + \beta_1 \cdot x}},$$

(Continued on page 11.)

where  $x$  corresponds to the temperature (outcome),  $\beta_0$  is the intercept,  $\beta_1$  is the estimated parameter of temperature.

**b** (weight 10p)

The launch temperature for the VIPER mission was  $31^\circ \text{F}$ . Using the model output (see Figure 1), estimate the probability of failure at  $x = 31$ . How will the situation change when the temperature rises by  $1^\circ \text{F}$ ? Can the odds ratio be considered as an approximation to a relative risk in this situation?



#### Parameter Estimates

| Term      | Estimate   | Std Error | ChiSquare | Prob>ChiSq |
|-----------|------------|-----------|-----------|------------|
| Intercept | 15.0422911 | 7.378391  | 4.16      | 0.0415     |
| temp      | -0.2321537 | 0.1082329 | 4.60      | 0.0320     |

Figure 1: Model output

*Sketch of the solution Problem 3 b.*

We have output for a logistic regression model. Let  $p$  denote the probability of failure. Failures tended to occur at lower temperatures and successes at higher temperatures, so the graph of  $\hat{p}$  decreases as temperature increases. The estimate of  $\beta_1$  is  $\hat{\beta}_1 = -0,2322$ , and the estimated standard deviation of  $\hat{\beta}_1$  is  $s_{\hat{\beta}_1} = 0,1082$ .

For each 1-degree increase in temperature, we estimate that the odds of failure decrease by a factor of  $e^{\hat{\beta}_1} = e^{-0,2322} \approx 0,79$ . The launch temperature for the mission was only  $31^\circ \text{F}$ . Because this value is much smaller than any temperature in our sample, it is dangerous to extrapolate the estimated relationship. Nevertheless, it appears that for a temperature this small, the drill failure is almost a sure thing. The logistic regression gives the estimated probability at  $x = 31$  as

$$p(31) = \frac{e^{\beta_0 + \beta_1 \cdot 31}}{1 + e^{\beta_0 + \beta_1 \cdot 31}} = \frac{e^{15,0423 - 0,23215 \cdot 31}}{1 + e^{15,0423 - 0,23215 \cdot 31}} = 0,99961,$$

(Continued on page 12.)

and the odds associated with this probability are  $\frac{0,99961}{(1-0,99961)} \approx 2563$ .

Note, the odds ratio can be considered an approximation to a relative risk (RR) if  $p(1)$  and  $p(0)$  are both small

$$OR = \frac{\frac{p(1)}{(1-p(1))}}{\frac{p(0)}{(1-p(0))}} \approx RR = \frac{p(1)}{p(0)}$$

In this case they are both large, so OR can not be considered as an approximation to RR.

Thus, if the logistic regression can be extrapolated down to 31, the probability of failure is 0,99961, the probability of success is 0,00039, and the predicted odds are 2563 to 1 against success.

### References

- [1] Jay L. Devore, Kenneth N. Berk (2012). Modern Mathematical Statistics with Applications, Springer New York Dordrecht Heidelberg London.
- [2] Cortez P., Morais A. d. J. R. (2007). A data mining approach to predict forest fires using meteorological data. *In Proceedings of the 13th Portuguese Conference on Artificial Intelligence.*
- [3] Vittinghoff E., Glidden David V. , Shiboski Stephen C., McCulloch, Charles E. (2012). Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, 2nd ed. Springer US.