

# TMA4315: Project 2

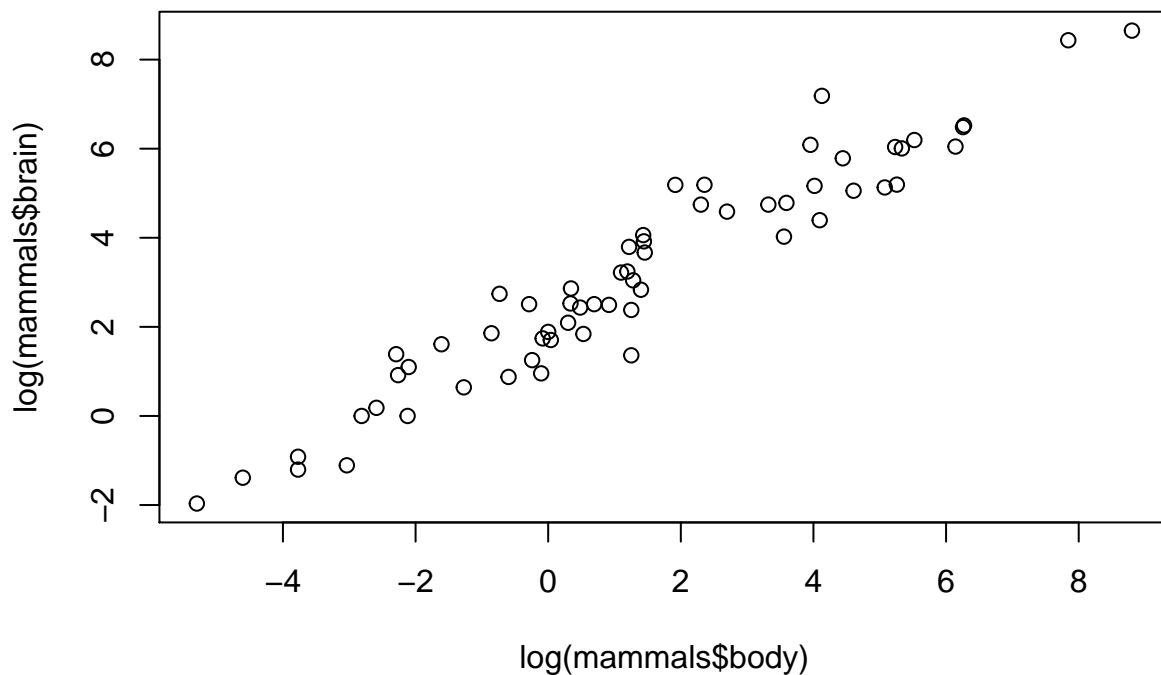
jototlan@stud.ntnu.no (10018), martigtu@stud.ntnu.no (10037)

## Problem 1

```
mammals <- read.table(  
  "https://www.math.ntnu.no/~jarlet/statmod/mammals.dat",  
  header=T)
```

a)

```
plot(log(mammals$body), log(mammals$brain)) # Seems pretty linear.
```



A log-log plot of the brain mass against body mass seems to reveal a linear trend. We thus fit the following model:

```
mod0 <- lm(log(brain) ~ log(body), data = mammals)  
summary(mod0)
```

```
##
## Call:
## lm(formula = log(brain) ~ log(body), data = mammals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71550 -0.49228 -0.06162  0.43597  1.94829
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.13479    0.09604   22.23  <2e-16 ***
## log(body)    0.75169    0.02846   26.41  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6943 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

b)

```
is.human = ifelse(mammals$species == "Human", 1, 0)
mammals$is.human = as.factor(is.human)

mod1 <- lm(log(brain) ~ log(body) + is.human, data = mammals)
summary(mod1)

##
## Call:
## lm(formula = log(brain) ~ log(body) + is.human, data = mammals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68392 -0.46764 -0.02398  0.47237  1.64949
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11500    0.09030   23.421  < 2e-16 ***
## log(body)    0.74228    0.02687   27.622  < 2e-16 ***
## is.human1    2.00691    0.66083    3.037  0.00356 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6511 on 59 degrees of freedom
## Multiple R-squared:  0.9315, Adjusted R-squared:  0.9292
## F-statistic: 401.1 on 2 and 59 DF,  p-value: < 2.2e-16
```

Let  $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2]^T$  be the coefficient estimates given in the summary above. Then the estimated effect on brain mass from being a human is  $\hat{\beta}_3 \approx$ .

We use the notation  $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  to represent the linear model. Here,  $X$  is the  $n \times p$  design matrix, where  $n$  is the number of observations and  $p$  is the number of parameters used in the model. As usual,  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ . This (along with the other usual assumptions [how much detail is required here??](#)) gives the well known result:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}).$$

Now we want to perform the hypothesis test

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_2 > 0.$$

Under  $H_0$ , we obtain that (we also index from 0 in the design matrix)

$$\frac{\hat{\beta}_2}{\sigma \sqrt{(X^T X)^{-1}_{2,2}}} \sim \mathcal{N}(0, 1).$$

Combining this with the fact that

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p},$$

where  $s^2 = RSS/(n-p)$ , we obtain the test statistic

$$\frac{\hat{\beta}_2}{s \sqrt{(X^T X)^{-1}_{2,2}}} \sim t_{n-p},$$

under  $H_0$ . We perform the calculations in R:

```
n <- nrow(mammals)
p <- 3
beta.2 <- mod1$coefficients[3]
s <- sqrt(deviance(mod1)/(n-p))
X <- model.matrix( ~ log(body) + is.human, data = mammals)
XtX.inv <- solve(t(X) %*% X)

T.stat <- beta.2/(s*sqrt(XtX.inv[3,3]))
p.val <- pt(T.stat, n - p, lower.tail = F)
p.val

## is.human1
## 0.001777696
```

The calculated p-value is 0.0017777, which suggest that...

## Problem 2

a)

In this problem we apply ordinal multinomial regression to data from Norway Chess 2021.

```
df <- read.csv('data/Norway\ Chess\ 2021.csv')
```

The response variable  $y_i$  is the outcome of the  $i$ 'th match. This can be considered an ordered categorical variable

$$y_i = \begin{cases} 1 & , \quad \text{white win} \\ 2 & , \quad \text{draw} \\ 3 & , \quad \text{black win} \end{cases}$$

which may depend on relative strength of different players, which player plays white and black and the type of game played. The response can be determined by an underlying latent variable  $u_i$ , where

$$u_i = x_i^T \beta + \epsilon_i.$$

The event  $y_i = r$  occurs if  $\theta_{r-1} < u_i \leq \theta_r$ , where

$$-\infty = \theta_0 < \theta_1 < \theta_2 < \theta_3 = \infty.$$

This means that our model returns that white wins whenever  $u_i \leq \theta_1$ , draw if  $\theta_1 < u_i \leq \theta_2$  and black win for  $u_i > \theta_2$ .