Solution of assignment 4, ST2304

August 1, 2016

Problem 1

Reload the data from assignment 3.

```
> heli <- read.csv("z:/folder/helicopterdata.csv")
> attach(heli)
```

1. We first log-transform the response variable, and then reanalyse the data using a three-way analysis of variance (ANOVA).

```
> logflighttime <- log(flighttime)
> loghelimod <- lm(logflighttime ~ size + wing + clip)
> anova(loghelimod)
Analysis of Variance Table

Response: logflighttime
           Df Sum Sq Mean Sq F value    Pr(>F)    
size        1 0.0783  0.0783  1.8078    0.1814    
wing        2 8.1317  4.0659 93.8251 < 2.2e-16 ***
clip        1 2.0896  2.0896 48.2198 2.414e-10 ***
Residuals 115 4.9835  0.0433                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Just like in assignment 3, *size* is non-significant. Thus, we remove this variable and reanalyse the data.

```
> loghelimod2 <- lm(logflighttime ~ wing + clip)
> anova(loghelimod2)
Analysis of Variance Table

Response: logflighttime
           Df Sum Sq Mean Sq F value    Pr(>F)    
wing        2 8.1317  4.0659  93.176 < 2.2e-16 ***
clip        1 2.0896  2.0896  47.886 2.646e-10 ***
Residuals 116 5.0618  0.0436                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then we compare the adjusted R^2 of the logtransformed and original model (`helimod` from assignment 3). We compare the complete models.

```
> summary(loghelimod)$adj.r.squared
[1] 0.6625815
> helimod <- lm(flighttime ~ size + wing + clip)
> summary(helimod)$adj.r.squared
[1] 0.7028989
```

The adjusted R^2 of this model is 0.663, while the complete model without the log transformation had an adjusted R^2 of 0.703. The alternative model does thus have a worse fit.

A short reminder (from Wikipedia): R^2 is the proportion of variability in a data set that is accounted for by the statistical model, and it provides a measure of how well future outcomes are likely to be predicted by the model. $R^2 = 1 - SS_{err}/SS_{tot}$, where SS_{tot} and SS_{err} = the total and residual sums of squares. Adjusted R^2 is a modification of R^2 that adjusts for the number of explanatory terms in a model: R^2 adj $= 1 - SS_{err}/SS_{tot} * df_t/df_e$, where df_t and df_e = the total and residual degrees of freedom.

2. The regression can again be written in the form of a multiple regression model

$$\begin{aligned}\log(\texttt{flighttime}) = \mu + &\alpha_{small}x_{small} \\ &+\beta_{up}x_{up} + \beta_{down}x_{down} \\ &+\gamma_{yes}x_{yes} \\ &+\epsilon\end{aligned}$$

We can look at the untransformed response by taking the exponential of both sides:

$$\begin{aligned}\texttt{flighttime} &= e^{\mu+\alpha_{small}x_{small}+\beta_{up}x_{up}+\beta_{down}x_{down}+\gamma_{yes}x_{yes}} \\ &= e^{\mu+\alpha_{small}x_{small}}e^{\beta_{up}x_{up}}e^{\beta_{down}x_{down}}e^{\gamma_{yes}x_{yes}}\end{aligned}$$

Because each x is either 0 or 1, each component of the formula will multiply the flighttime by for example either $e^{\alpha*1} = e^{\alpha}$ or $e^{\alpha*0} = 1$.

The summary table provides all estimates.

```
> summary(loghelimod)

Call:
lm(formula = logflighttime ~ size + wing + clip)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47272 -0.14327  0.03741  0.12800  0.56625

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2.35093    0.04249  55.326  < 2e-16 ***
sizesmall   -0.05110    0.03801  -1.345    0.181
wingdown    -0.52747    0.04655 -11.332  < 2e-16 ***
wingup      -0.57401    0.04655 -12.332  < 2e-16 ***
clipyes     -0.26392    0.03801  -6.944 2.41e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2082 on 115 degrees of freedom
Multiple R-squared:  0.6739,Adjusted R-squared:  0.6626
F-statistic: 59.42 on 4 and 115 DF,  p-value: < 2.2e-16
```

We see that the estimated effect of attaching a clip is $e^{-0.264} = 0.768$, or 76.8% of the flighttime without a clip.

The estimated effect of a small helicopter is $e^{-0.051} = 0.95$, thus a small helicopter falls to the ground 5% faster relative to a large helicopter.

3. Confidence intervals are computed using `confint()`. The optional argument (`level`) allows one to specify the confidence interval required, but we will accept the default which give us 95% confidence intervals.

```
> confint(loghelimod)
                 2.5 %      97.5 %
(Intercept)  2.2667584  2.43509676
sizesmall   -0.1263839  0.02418249
wingdown    -0.6196716 -0.43526624
wingup      -0.6662135 -0.48180810
clipyes     -0.3392007 -0.18863437
```

To transform these to confidence intervals in percent, we take the exponential and multiply by 100.

```
> exp(confint(loghelimod))*100
                2.5 %     97.5 %
(Intercept) 964.80753 1141.69233
sizesmall    88.12765  102.44773
wingdown     53.81211   64.70924
wingup       51.36498   61.76656
clipyes      71.23394   82.80892
```

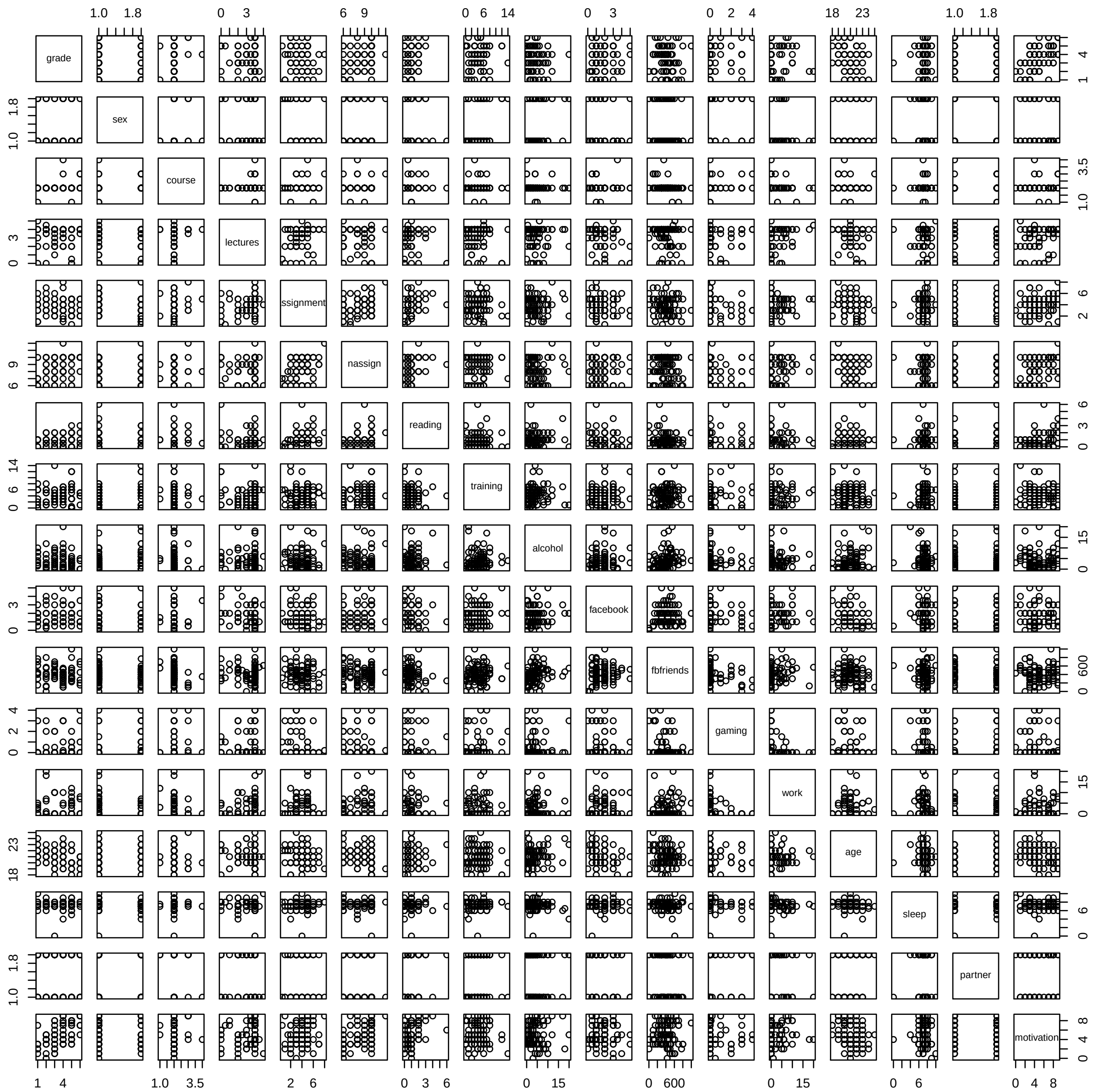
Note that this does not make any sense for the intercept.

Problem 2
We download the data set and split it into two parts. `skip = 1` allows us to skip one row (the one with variable explanations) in the data set betfore we start reading data into R.

```
> grades <- read.csv("Z:/folder/Grade_prediction_data_2015.csv", skip = 1)
> trainingset <- grades[complete.cases(grades),]
> validationset <- grades[complete.cases(grades[,-2])&is.na(grades$grade),]
> attach(trainingset)
```

1. We first make a scatterplot of all variables.

```
> pairs(grades[, -1])
```



Then we estimate all pairwise correlations.

```
> round(cor(grades[,sapply(grades,is.numeric)],use="complete.obs"), 2)
            grade lectures assignments nassign reading training alcohol
grade        1.00     0.00       -0.08    0.46    0.13     0.07   -0.13
lectures     0.00     1.00        0.22    0.29    0.13     0.06   -0.23
assignments -0.08     0.22        1.00    0.31    0.39     0.08   -0.07
nassign      0.46     0.29        0.31    1.00    0.26     0.15   -0.13
reading      0.13     0.13        0.39    0.26    1.00    -0.12    0.11
training     0.07     0.06        0.08    0.15   -0.12     1.00    0.02
alcohol     -0.13    -0.23       -0.07   -0.13    0.11     0.02    1.00
facebook     0.10    -0.25       -0.09    0.02   -0.05     0.13    0.04
fbfriends   -0.06    -0.14       -0.08   -0.10   -0.05     0.03    0.29
gaming       0.14    -0.12       -0.30   -0.26   -0.28    -0.04    0.02
work        -0.01     0.05        0.07    0.08    0.08     0.00    0.07
age         -0.33    -0.13       -0.09   -0.28    0.04    -0.08    0.20
sleep        0.02     0.05        0.04   -0.12    0.09     0.17   -0.15
motivation   0.58     0.26        0.18    0.54    0.42     0.00   -0.03
            facebook fbfriends gaming  work   age sleep motivation
grade           0.10     -0.06   0.14 -0.01 -0.33  0.02       0.58
lectures       -0.25     -0.14  -0.12  0.05 -0.13  0.05       0.26
assignments    -0.09     -0.08  -0.30  0.07 -0.09  0.04       0.18
nassign         0.02     -0.10  -0.26  0.08 -0.28 -0.12       0.54
reading        -0.05     -0.05  -0.28  0.08  0.04  0.09       0.42
training        0.13      0.03  -0.04  0.00 -0.08  0.17       0.00
alcohol         0.04      0.29   0.02  0.07  0.20 -0.15      -0.03
facebook        1.00      0.04  -0.01  0.05 -0.27  0.13      -0.06
fbfriends       0.04      1.00  -0.25  0.14 -0.12 -0.09      -0.07
gaming         -0.01     -0.25   1.00 -0.22 -0.05 -0.02      -0.05
work            0.05      0.14  -0.22  1.00 -0.01 -0.06       0.04
age            -0.27     -0.12  -0.05 -0.01  1.00 -0.03      -0.33
sleep           0.13     -0.09  -0.02 -0.06 -0.03  1.00       0.02
motivation     -0.06     -0.07  -0.05  0.04 -0.33  0.02       1.00
```

In multiple regression we assume no or little multicollinearity (correlaion among explanatory variables). The variables *motivation* and *nassign*, *motivation* and *reading*, and *assignments* and *reading* seem to be somewhat positively correlated. All other correlations are lower and should not cause any problems in our multiple regression. However, even correlations as low as 0.28 has been found to bias analyses (Graham 2003 Ecology). Collinearity may cause (1) inaccurate model parameterization, (2) decreased statistical power, and (3) exclusion of significant predictor variables during model selection. With highly correlated explanatory variables care should be taken when contructing models. Here we will just keep all variables in the analyses. However, with very highly correlated variables (e.g. $\geqslant 0.7$) one should use biological knowledge to construct the model that makes best sense while avoiding highly correlated variables in the same model.

2. We start by fitting the full model with all relevant additive terms. Then we examine

the parameter estimates (using `summary()`) and the significance of each term using F-tests (`drop1()` command). *A priori* we exclude the variables *facebook*, *fbfriends* and *work* as we do not see how these variables may affect the grade of students when we have direct measures of time spent studying for the course. *Partner* was considered irrelevant for the grades of student.

```
> # Fit
> mod0 <- lm(grade ~ sex + course + lectures + assignments +
+              nassign + reading + training + alcohol +
+              gaming + age + sleep + motivation)
> # Paramter estimates
> summary(mod0)

Call:
lm(formula = grade ~ sex + course + lectures + assignments +
    nassign + reading + training + alcohol + gaming + age + sleep +
    motivation)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5858 -0.6735 -0.0567  0.5901  2.5633

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.32257    2.72378   0.853 0.397160
sexmale       -0.63877    0.39095  -1.634 0.107434
courseMA0001   0.40448    0.80786   0.501 0.618396
courseMA1101   1.25320    1.11032   1.129 0.263449
courseTMA4100  1.83546    1.36699   1.343 0.184342
lectures      -0.28484    0.12221  -2.331 0.023090 *
assignments   -0.29257    0.10832  -2.701 0.008939 **
nassign        0.27118    0.11402   2.378 0.020538 *
reading        0.12302    0.22753   0.541 0.590698
training       0.05379    0.05096   1.056 0.295348
alcohol       -0.03470    0.03975  -0.873 0.386039
gaming         0.31647    0.16513   1.917 0.059989 .
age           -0.06778    0.10227  -0.663 0.510027
sleep          0.03474    0.11098   0.313 0.755309
motivation     0.29104    0.07730   3.765 0.000377 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.091 on 61 degrees of freedom
Multiple R-squared:  0.5426,Adjusted R-squared:  0.4376
F-statistic: 5.169 on 14 and 61 DF,  p-value: 2.8e-06
> # F-tests
> drop1(mod0, test="F")
Single term deletions

Model:
```

```
grade ~ sex + course + lectures + assignments + nassign + reading +
    training + alcohol + gaming + age + sleep + motivation
            Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                   72.671 26.596
sex          1    3.1804 75.852 27.852  2.6696 0.1074337
course       3    3.6305 76.302 24.301  1.0158 0.3919556
lectures     1    6.4717 79.143 31.080  5.4323 0.0230898 *
assignments  1    8.6906 81.362 33.181  7.2949 0.0089393 **
nassign      1    6.7388 79.410 31.336  5.6565 0.0205378 *
reading      1    0.3483 73.020 24.960  0.2923 0.5906980
training     1    1.3273 73.999 25.972  1.1141 0.2953475
alcohol      1    0.9081 73.579 25.540  0.7623 0.3860387
gaming       1    4.3758 77.047 29.040  3.6730 0.0599891 .
age          1    0.5232 73.195 25.141  0.4392 0.5100269
sleep        1    0.1168 72.788 24.718  0.0980 0.7553094
motivation   1   16.8897 89.561 40.478 14.1771 0.0003765 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-values tests if the fit of your model changes if you would remove that explanatory variable, and the sums of squares tells us how much the sum of squares would change; the smaller the change in sum in squares, the smaller the F-value.

We decided to use a step-wise approach where we remove all non-significant terms in one go, then try to add each of the removed variables to the new model using `add1()`.

Another popular approach is to remove the explanatory variable with the lowest F-value in the `drop1()`-table in each step (e.g. using `mod1 <- update(mod0, .~.-sleep)` until all variables are significant. Stepwise approaches are generally problematic due to the problem of multiple testing (i.e. 1 out of 20 tests will be significant by chance).

```
> mod1 <- lm(grade ~ lectures + assignments +
+               nassign + motivation)
> add1(mod1, .~. + sex + course + reading + training + alcohol +
+               gaming + age + sleep, test="F")
Single term additions

Model:
grade ~ lectures + assignments + nassign + motivation
         Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>                 87.021 20.292
sex       1    0.5118 86.509 21.843  0.4141 0.52199
course    3    4.2055 82.816 22.527  1.1510 0.33493
reading   1    0.4452 86.576 21.902  0.3599 0.55048
training  1    0.4315 86.590 21.914  0.3488 0.55667
alcohol   1    2.7686 84.253 19.834  2.3003 0.13386
gaming    1    4.1135 82.908 18.612  3.4731 0.06657 .
```

```
age       1    2.4535 84.568 20.118  2.0309 0.15857
sleep     1    0.7693 86.252 21.617  0.6244 0.43210
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

None of the excluded variables seems to affect the students grades. We then check whether the model may be further simplified.

```
> drop1(mod1, test="F")
Single term deletions

Model:
grade ~ lectures + assignments + nassign + motivation
            Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>                    87.021 20.292
lectures     1    3.7109  90.732 21.466  3.0277  0.086188 .
assignments  1    7.7689  94.790 24.791  6.3386  0.014071 *
nassign      1   10.1606  97.182 26.685  8.2900  0.005265 **
motivation   1   27.2131 114.234 38.971 22.2030 1.185e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lectures no longer significantly affect students grades and we remove this variable. We then examine the parameter estimates of the selected model.

```
> mod2 <- lm(grade ~ assignments + nassign + motivation)
> summary(mod2)

Call:
lm(formula = grade ~ assignments + nassign + motivation)

Residuals:
     Min       1Q   Median       3Q      Max
-2.12601 -0.81429  0.06387  0.65871  3.08808

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.66362    0.74646   0.889  0.37695
assignments -0.25233    0.09194  -2.745  0.00764 **
nassign      0.26859    0.10270   2.615  0.01085 *
motivation   0.29946    0.06698   4.471 2.84e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.123 on 72 degrees of freedom
```

```
Multiple R-squared:  0.4289,Adjusted R-squared:  0.4051
F-statistic: 18.03 on 3 and 72 DF,  p-value: 7.937e-09
```

We see that highly motivated students which complete many assignments perform better on the exam. On the other hand, students which spend a lot of time with assignments each week perform less well on the exam. This effect might at first glance seem suprising. However, it is likely that the students which spend a lot of time to complete assignments are those that struggle a lot with the course. These students most likely perform much better on the exam than if they had not spent a lot of time with assignments. Still, on average they perform less well than students which conquer the course stright away.

3. Finally we predict the grades of students with missing values in the data set.

```
> round(predict(mod2,validationset))
 3  5  6 10 15 16 17 18 19 21 23 24 25 27 29 31 33 40 42 43
 2  4  4  3  2  3  2  2  3  3  3  2  3  3  4  4  4  5  4  4
```

The predictions are all between 1 and 6 and seems to make sense. However, generally this model could predict grades above 6. The adjusted R^2 of the final model is 0.405, not much worse than the full model in task 1 (0.438). With only 40.5% of the variation in student grades explained by our model we should not expect the predictions to be to accurate.