Solution of assignment 4, ST2304

Problem 1 1. Analysis of Variance Table

```
Response: log.flighttime
          Df  Sum Sq Mean Sq F value    Pr(>F)
size       1 0.15360 0.15360  4.4202  0.049077 *
wing       2 1.78191 0.89096 25.6400 4.011e-06 ***
clip       1 0.50256 0.50256 14.4627  0.001202 **
Residuals 19 0.66022 0.03475
```

We first log transform the response variable, and then reanalyse the data, using all three explanatory variables. All explanatory variables have a significant effect, so we do not need to remove any.

```
Call:
lm(formula = log.flighttime ~ size + wing + clip)

Residuals:
      Min        1Q    Median        3Q       Max
-0.387636 -0.077667 -0.009399  0.092523  0.355112

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.43039    0.08508  28.565  < 2e-16 ***
sizesmall   -0.16000    0.07610  -2.102  0.04908 *
wingdown    -0.53599    0.09320  -5.751 1.53e-05 ***
wingup      -0.61244    0.09320  -6.571 2.73e-06 ***
clipyes     -0.28941    0.07610  -3.803  0.00120 **
---

Residual standard error: 0.1864 on 19 degrees of freedom
Multiple R-squared: 0.7869,     Adjusted R-squared: 0.742
F-statistic: 17.54 on 4 and 19 DF,  p-value: 3.545e-06
```

The adjusted R^2 of this model is 0.742, while the complete model without the log transformation had an adjusted R^2 of 0.7674. This 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. $R^2 = 1 - SS_{err}/SS_{tot}$ Adjusted R2 is a modification of R2 that adjusts for the number of explanatory terms in a model: R^2 adj $= 1 - SS_{err}/SS_{tot} * df_t/df_e$

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:

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

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 $e^{\alpha*0} = 1$.

Thus, the estimated effect of attaching a clip is $e^{-0.289} = 0.749$, or 75% of the flighttime without a clip.

The estimated effect of a small helicopter is $e^{-0.16} = 0.852$, thus a small helicopter falls to the ground 15% faster relative to a large helicopter.

3.
```
> confint(model.1)
                 2.5 %       97.5 %
                 2.5 %        97.5 %
(Intercept)  2.2523049  2.6084709934
sizesmall   -0.3192808 -0.0007161569
wingdown    -0.7310732 -0.3409128069
wingup      -0.8075212 -0.4173607984
clipyes     -0.4486948 -0.1301301724
```

as with the estimates, we take the exponential of those confidence intervals and multiply by 100,

```
                2.5 %     97.5 %
(Intercept) 950.96297 1357.82737
sizesmall    72.66715   99.92841
wingdown     48.13921   71.11209
wingup       44.59622   65.87832
clipyes      63.84609   87.79811
```

(note that this does not make any sense for the intercept)

Problem 2 1. The variables `stai` and `bdi` and variables `avoidant` and `depend` seem to be somewhat positively correlated.

2.
```
model.1 <- lm(subjpain ~ bdi + stai + age + edu + status + relax +
avoidant + depend + compuls + sleepqual + duration)

> anova(model.1)
Analysis of Variance Table

Response: subjpain
          Df  Sum Sq Mean Sq F value  Pr(>F)
bdi        1 2035.13 2035.13 11.7708 0.01097 *
stai       1  247.75  247.75  1.4330 0.27024
age        1 1157.90 1157.90  6.6970 0.03606 *
edu        1  414.55  414.55  2.3976 0.16545
status     1  212.04  212.04  1.2264 0.30470
relax      2  884.10  442.05  2.5567 0.14669
avoidant   1 1370.81 1370.81  7.9285 0.02593 *
```

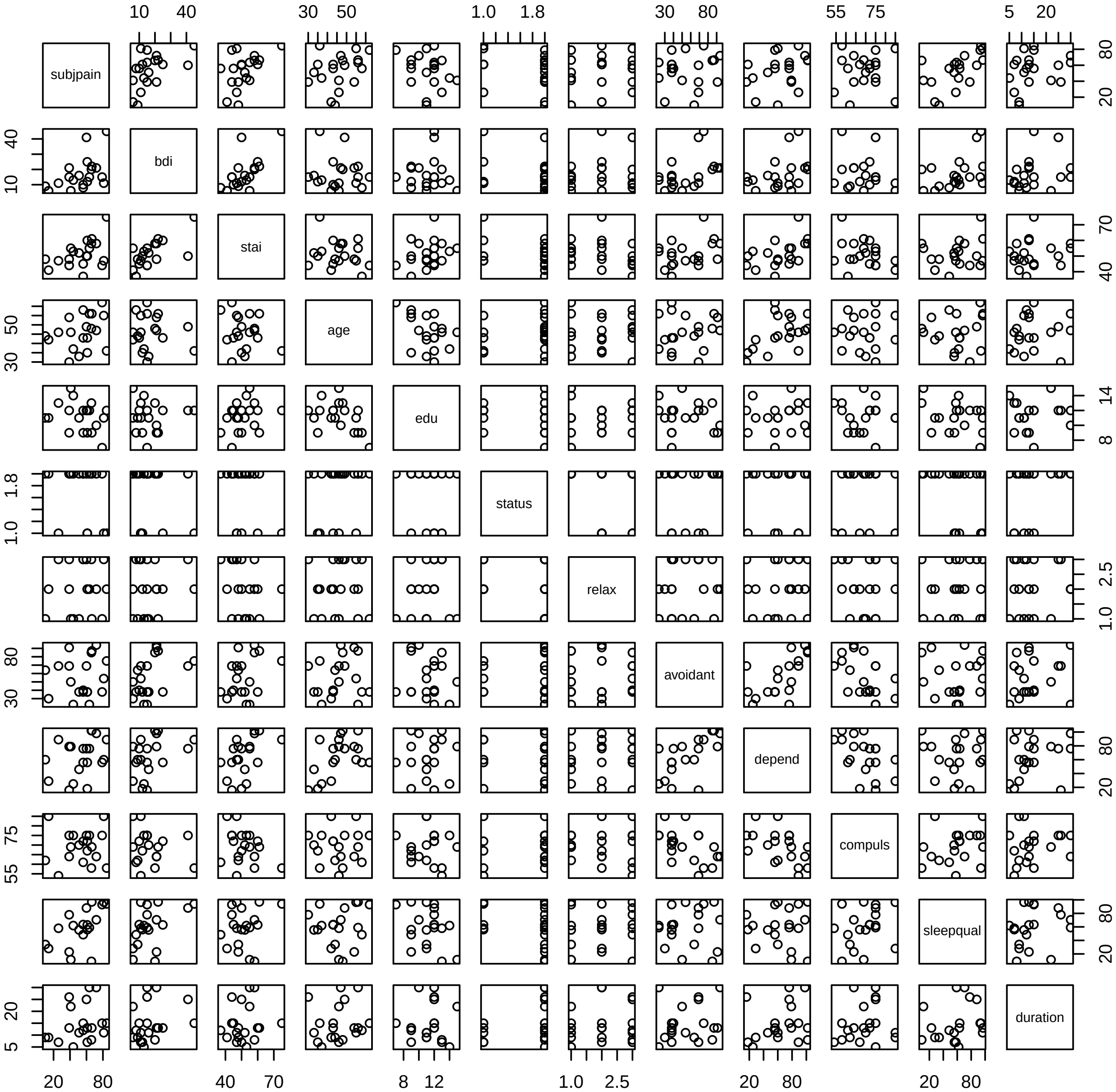

Figure 1: Ouput of function pairs()

```
depend     1  173.75  173.75  1.0050 0.34950
compuls    1  188.30  188.30  1.0891 0.33137
sleepqual  1    9.89    9.89  0.0572 0.81780
duration   1  279.73  279.73  1.6179 0.24401
Residuals  7 1210.28  172.90
```

>From the anova table of the full model it seems only `bdi`, `age` and `avoidant` have a significant effect on the subjective pain level. This full model has an adjusted R^2 of 0.5986.

```
> drop1(model.1,test="F")
Single term deletions

Model:
subjpain ~ bdi + stai + age + edu + status + relax + avoidant +
    depend + compuls + sleepqual + duration
          Df Sum of Sq    RSS   AIC F value   Pr(F)
<none>                 1210.3 108.1
bdi        1      37.1 1247.4 106.7  0.2148 0.65709
stai       1    2089.3 3299.6 126.1 12.0841 0.01032 *
age        1      89.8 1300.0 107.5  0.5192 0.49455
edu        1    1226.4 2436.7 120.1  7.0935 0.03231 *
status     1      22.8 1233.1 106.4  0.1320 0.72709
relax      2    1766.4 2976.7 122.1  5.1083 0.04286 *
avoidant   1     507.5 1717.7 113.1  2.9351 0.13040
depend     1      30.3 1240.6 106.6  0.1755 0.68785
compuls    1     128.7 1339.0 108.1  0.7445 0.41681
sleepqual  1      47.7 1258.0 106.8  0.2761 0.61549
duration   1     279.7 1490.0 110.2  1.6179 0.24401
```

The `F values` are from tests if the fit of your model changes if you would remove that explanatory variable, and the `Sum of Sq` how much the sum of squares would change; the smaller the change in sum in squares, the smaller the F value.

Each step, we

- remove the explanatory with the lowest F value in the `drop1(model.1)` table, using `model.2 <- update(model.1, .~.-status}`
- check the adjusted R^2 of the resulting model, using `summary(model.2)`
- find the explanatory to remove next, using `drop1(model.2)`

resulting in the following models:

```
model.2 <- update(model.1, .~.-status)        # adj. r2=0.642
model.3 <- update(model.2, .~.-sleepqual) # adj. r2=0.6754
model.4 <- update(model.3, .~.-depend) # adj. r2=0.6978
model.5 <- update(model.4, .~.-bdi) # adj. r2=0.7092
model.6 <- update(model.5, .~.-compuls) # adj. r2=0.704 (bit worse than model5)
model.6B <- update(model.5, .~.-age) # adj. r2=0.7044
model.7 <- update(model.6, .~.-age) # adj. r2=0.7003
anova(model.7)        #all except 'duration' sign. at 0.05 (duration sign. at 0.1)
model.8 <- update(model.7, .~.-duration) # adj. r2=0.6354 (worse than model7)
```

Since models 5 and 7 do not differ much in adjusted R^2, the preference goes to the one with a smaller number of explanatory variables - model 7. An other way to compare models is based on their AIC values (`AIC(model.5)`), those hardly differ.

```
Call:
lm(formula = subjpain ~ stai + edu + relax + avoidant + duration)
Residuals:
    Min      1Q  Median      3Q     Max
-18.734  -6.215  -3.464   8.538  18.924

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.5601    20.4028   0.665 0.517910
stai          2.8698     0.4537   6.325 2.64e-05 ***
edu          -8.8717     1.6657  -5.326 0.000138 ***
relaxtype1   -5.3931     6.6853  -0.807 0.434346
relaxtype2   29.8754     7.5163   3.975 0.001586 **
avoidant     -0.4908     0.1416  -3.467 0.004170 **
duration      0.7139     0.3555   2.008 0.065901 .
---

Residual standard error: 11.36 on 13 degrees of freedom
Multiple R-squared: 0.7949,     Adjusted R-squared: 0.7003
F-statistic: 8.399 on 6 and 13 DF,  p-value: 0.0007262
```

The model seems to make sense. The anxiety index `stai`, years of education and the number of years the patient had fibromyalgia `duration` have a significant effect on the subjective pain level. The estimated effect of the 2nd type of relaxation technique `relaxtype2` is larger than those of any of the other explanatory variables, while the 1st type has no significant effect.

3.
```
mymod <- model.7
predict(mymod,validationset)
        3         8         9        10        14        21
 38.08337  57.31477  -4.54104  44.71322 -18.74591  96.95964
```

Those predictions do not seem to make much sense. Two of them are negative, while the scale is from 0 to 100. Another one is very close to the maximum of the scale, while you would expect this to be very rare.

Alternatively, we can start from the minimal model with intercept only,

```
model.A <- lm(subjpain ~ 1)
```

and use `add1()` to test if additional terms should be added to the model.

```
add1(model.A, .~.+bdi + stai + age + edu + status + relax
+ avoidant + depend + compuls + sleepqual + duration, test="F")
```

We see that `bdi`, `stai` and `sleepqual` would change the model significantly if added. We start by adding `sleepqual`, as it has the largest `Sum of Sq`

```
model.B <- update(model.A, .~.+sleepqual)
```

This model has an adjusted R^2 of 0.3125 (`summary(model.B)`). Using `add1()` shows us that none of the explanatory variables will improve the model significantly at the 0.05 threshold when added, but the p-value of `stai` is close, so we try adding that one

```
model.C <- update(model.B, .~.+stai)
```

and see that model.C has a higher adjusted R^2 than model.B, namely R^2=0.4136. Following the same logic, we then add `age` to get model.D

```
model.D <- update(model.C, .~.+age)
```

which has adjusted R^2 of 0.5084. `add1()` shows us that none of the other explanatory variables would make an improvement to the model, so we end up with this model

```
> summary(model.D)

Call:
lm(formula = subjpain ~ sleepqual + stai + age)

Residuals:
     Min       1Q   Median       3Q      Max
-29.1560  -4.7590   0.1322   9.0804  21.8722

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -59.6851    29.8255  -2.001   0.0626 .
sleepqual     0.3590     0.1248   2.878   0.0109 *
stai          1.0772     0.4112   2.620   0.0186 *
age           0.7939     0.3838   2.068   0.0552 .
---

Residual standard error: 14.55 on 16 degrees of freedom
Multiple R-squared: 0.586,      Adjusted R-squared: 0.5084
F-statistic:  7.55 on 3 and 16 DF,  p-value: 0.002288
```

It is very different from the model obtained when starting from the full model and using `drop1()`, having only `stai` in common.

```
> predict(model.D,validationset)
       3        8        9       10       14       21
46.13594 40.49688 17.15055 43.04877 41.87598 68.70829
```

These results seem to make sense, since at least they are all between 0 and 100. However, since the adjusted R^2 of the best model using the first method was 0.70, compared to adj. R^2=0.5084 for the second method's best model, we would expect the first one to be better at predicting.