Exercise 7 - solution

Paul Blanche & Jolien Cremers

Exercise A

Question 1

We first load the data and look at the "summary", as always.

```
load(url("http://paulblanche.com/files/BW.rda"))
summary(BW)
```

```
##       age             lwt             race           smoke       
##  Min.   :14.00   Min.   : 80.0   Min.   :1.000   Min.   :0.0000  
##  1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :23.00   Median :121.0   Median :1.000   Median :0.0000  
##  Mean   :23.24   Mean   :129.8   Mean   :1.847   Mean   :0.3915  
##  3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :45.00   Max.   :250.0   Max.   :3.000   Max.   :1.0000  
##       ptl               ht                ui              fvt        
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000  
##  Mean   :0.1958   Mean   :0.06349   Mean   :0.1481   Mean   :0.7937  
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000  
##  Max.   :3.0000   Max.   :1.00000   Max.   :1.0000   Max.   :6.0000  
##       bwt        
##  Min.   : 709  
##  1st Qu.:2414  
##  Median :2977  
##  Mean   :2945  
##  3rd Qu.:3475  
##  Max.   :4990
```

Question 2

We look at the frequencies that were observed for the variables indicating the number of:

- previous premature births of the mother (ptl)
- medical consultations during the first trimester of pregnancy (fvt)

```
table(BW$fvt)
```

```
##
##   0   1   2   3   4   6
## 100  47  30   7   4   1
```

We see only few observations of medical consultations beyond 2. We therefore decide to define a group as "at least 2". Otherwise we would need to either base our analyses on very few observations or make additional assumptions to compensate.

```
table(BW$ptl)
```

```
##
##   0   1   2   3
## 159  24   5   1
```

Similar comment. Hence we create new variables as suggested.

```
BW$fvt2 <- factor(as.numeric(BW$fvt>0) + as.numeric(BW$fvt>1),
                  levels=c(0,1,2),
                  labels=c("0","1","2+"))
BW$ptl2 <- factor(as.numeric(BW$ptl>0),
                  levels=c(0,1),
                  labels=c("0","1+"))
```

We now make the variables **race** and **smoke** become factor variables and give appropriate labels.

```
BW$race <- factor(BW$race,
                  levels=c(1,2,3),
                  labels=c("white","black","other"))
BW$smoke <- factor(BW$smoke,
                   levels=c(0,1),
                   labels=c("no","yes"))
```

Finally we update the variables indicating the birth weight (**bwt**) and mother's weight (**lwt**) to have them in the same unit, in kilograms.

```
BW$lwt <- BW$lwt/2.205
BW$bwt <- BW$bwt/1000
```

We can now print the summary of our updated, well prepared, data.

```
summary(BW)
```

```
##       age             lwt             race     smoke          ptl
##  Min.   :14.00   Min.   : 36.28   white:96   no :115   Min.   :0.0000
```

```
##  1st Qu.:19.00   1st Qu.: 49.89   black:26   yes: 74   1st Qu.:0.0000
##  Median :23.00   Median : 54.88   other:67             Median :0.0000
##  Mean   :23.24   Mean   : 58.87                        Mean   :0.1958
##  3rd Qu.:26.00   3rd Qu.: 63.49                        3rd Qu.:0.0000
##  Max.   :45.00   Max.   :113.38                        Max.   :3.0000
##        ht                ui               fvt              bwt        fvt2
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.709   0 :100
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2.414   1 : 47
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :2.977   2+: 42
##  Mean   :0.06349   Mean   :0.1481   Mean   :0.7937   Mean   :2.945
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:3.475
##  Max.   :1.00000   Max.   :1.0000   Max.   :6.0000   Max.   :4.990
##  ptl2
##  0 :159
##  1+: 30
##
##
##
##
```

Question 3

We now produce a descriptive "Table 1" to compare the distribution of relevant variables between the two groups that we want to compare: mothers who smoke and those who do not. To do so, we use the convenient **Publish** package from R. Note that we use **Q()** to indicate that we want median and interquartile range instead of mean and standard deviation. This is a personal choice and presenting means and standard deviations instead would also be fine (although we find the interpretation easier with our choice).

```
library(Publish)
Tab <- univariateTable(smoke~Q(age) + Q(lwt) + race + ui + ht + ptl2 + fvt2,
                data=BW,
                compare.groups = FALSE,
                show.totals = FALSE)
Tab
```

```
##    Variable        Level        no (n=115)        yes (n=74)
## 1       age median [iqr]       23 [20, 26]       22 [19, 26]
## 2       lwt median [iqr] 56.2 [50.8, 64.2] 54.4 [48.6, 62.2]
## 3      race        white         44 (38.3)         52 (70.3)
## 4                  black         16 (13.9)         10 (13.5)
## 5                  other         55 (47.8)         12 (16.2)
## 6        ui            1         15 (13.0)         13 (17.6)
## 7                      0        100 (87.0)         61 (82.4)
## 8        ht            0        108 (93.9)         69 (93.2)
```

```
## 9                      1           7 (6.1)           5 (6.8)
## 10     ptl2            0        103 (89.6)         56 (75.7)
## 11                    1+         12 (10.4)         18 (24.3)
## 12     fvt2            0         55 (47.8)         45 (60.8)
## 13                     1         35 (30.4)         12 (16.2)
## 14                    2+         25 (21.7)         17 (23.0)
```

We observe a few differences in the distribution of the variables between the two groups. Some seem minor and clinically insignificant, such as the difference in age or mother's weight. Some are more pronounced and should probably not be neglected, such as the differences in race groups (38% of white mothers among non smokers versus 70% among smokers).

Question 4

We estimate a multiple linear model to compare the average birth weight of children born from smoking and nonsmoking mothers, whom mothers are otherwise similar with respect to the listed variables.

```
lm2 <- lm(bwt~smoke*age + lwt + race + ui + ht + ptl2 + fvt2,data=BW)
summary(lm2)
```

```
##
## Call:
## lm(formula = bwt ~ smoke * age + lwt + race + ui + ht + ptl2 +
##     fvt2, data = BW)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82005 -0.37453  0.02257  0.47931  1.54259
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.480249   0.355904   6.969 6.07e-11 ***
## smokeyes      0.709562   0.465811   1.523 0.129473
## age           0.012510   0.011791   1.061 0.290132
## lwt           0.009447   0.003760   2.512 0.012890 *
## raceblack    -0.395832   0.151269  -2.617 0.009645 **
## raceother    -0.275009   0.117711  -2.336 0.020595 *
## ui           -0.527161   0.137316  -3.839 0.000172 ***
## ht           -0.598505   0.198955  -3.008 0.003011 **
## ptl21+       -0.215840   0.136620  -1.580 0.115926
## fvt21         0.138229   0.122193   1.131 0.259487
## fvt22+       -0.031790   0.122280  -0.260 0.795186
## smokeyes:age -0.042886   0.019262  -2.227 0.027241 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6388 on 177 degrees of freedom
## Multiple R-squared:  0.277,  Adjusted R-squared:  0.2321
## F-statistic: 6.166 on 11 and 177 DF,  p-value: 1.529e-08
```

We note that the interaction term between smoking status and age is significant (p-value=0.027241). Hence, the results confirm that modeling an interaction between smoking status and age seems relevant.

To facilitate the interpretation (especially that of the intercept and parameters related to smoking), we refit the model after having removed 27 to all ages and 55 kg to all mother's weights.

```
BW$age27 <- BW$age-27
BW$lwt55 <- BW$lwt-55
lm3 <- lm(bwt~smoke*age27 + lwt55 + race + ui + ht + ptl2 + fvt2,data=BW)
summary(lm3)
```

```
##
## Call:
## lm(formula = bwt ~ smoke * age27 + lwt55 + race + ui + ht + ptl2 +
##     fvt2, data = BW)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82005 -0.37453  0.02257  0.47931  1.54259
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.337601   0.118751  28.106  < 2e-16 ***
## smokeyes       -0.448372   0.127460  -3.518 0.000553 ***
## age27           0.012510   0.011791   1.061 0.290132
## lwt55           0.009447   0.003760   2.512 0.012890 *
## raceblack      -0.395832   0.151269  -2.617 0.009645 **
## raceother      -0.275009   0.117711  -2.336 0.020595 *
## ui             -0.527161   0.137316  -3.839 0.000172 ***
## ht             -0.598505   0.198955  -3.008 0.003011 **
## ptl21+         -0.215840   0.136620  -1.580 0.115926
## fvt21           0.138229   0.122193   1.131 0.259487
## fvt22+         -0.031790   0.122280  -0.260 0.795186
## smokeyes:age27 -0.042886   0.019262  -2.227 0.027241 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6388 on 177 degrees of freedom
```

```
## Multiple R-squared:  0.277,  Adjusted R-squared:  0.2321
## F-statistic: 6.166 on 11 and 177 DF,  p-value: 1.529e-08
```

The estimated value for the intercept, 3.338 kg, is the estimated weight at birth for a baby born from a mother:

- that does not smoke
- is 27 years old
- weighs 55 kg
- is white
- has no history of uterine irritability
- has no history of hypertension
- has had 0 medical consultations during the first trimester of pregnancy
- has never given birth to premature baby

that is, a mother having the profile corresponding to all predictor variables being equal to 0.

Interestingly, the estimated value -0.448372 kg (second line), is the estimated mean difference in birth weight between two babies that are both born from a mother 27 years, one smokes, the other does not, and both mothers are otherwise similar with respect to race, weight, history of uterine irritability, previous premature birth and hypertension and number of medical consultations. In other words, -0.448 kg is the estimated mean difference in birth weight between babies born from mothers of the reference age 27, who are similar with respect to all variables adjusted on except smoking.

To get confidence intervals for all the parameters in the model we use:

```
confint(lm3)
```

```
##                       2.5 %       97.5 %
## (Intercept)     3.103250497  3.571950926
## smokeyes       -0.699908924 -0.196834234
## age27          -0.010758348  0.035778219
## lwt55           0.002026091  0.016867849
## raceblack      -0.694354797 -0.097309337
## raceother      -0.507306593 -0.042711892
## ui             -0.798147916 -0.256174196
## ht             -0.991135395 -0.205875511
## ptl21+         -0.485453393  0.053774199
## fvt21          -0.102913739  0.379371309
## fvt22+         -0.273103983  0.209524613
## smokeyes:age27 -0.080898473 -0.004874354
```

The confidence interval for the estimated birth weight mean difference between a baby born from a smoker and a non-smoker from mothers, both being 27 years old and similar with respect to all other variables is equal to $(-0.70; -0.20)$. It means means that the effect of smoking on birth weight is negative and significant for 27 year old mothers.

The estimated effect of smoking on birth weight for 20 or 30 year old mothers can be deduced

directly from the previous model estimates as follows.

- $-0.448372 + (-0.042886 * 3)$ = -0.57703 for 30 year old mothers
- $-0.448372 + (-0.042886 * -7)$ = -0.14817 for 20 year old mothers.

By "effect", here again we mean "mean difference in birth weight" when we compare two mothers, one smokes, the other does not, both are otherwise similar with respect to all other variables adjusted for.

To obtain the above values directly from the output of the summary function, we can use the same "trick" as before: we remove either 20 or 30 to the initial age variable. This trick is also particularly convenient to obtain confidence intervals.

```
BW$age30 <- BW$age-30
lm3b <- lm(bwt~smoke*age30 + lwt55 + race + ui + ht + ptl2 + fvt2,data=BW)
coef(lm3b)["smokeyes"]     # we print the relevant coeffecicient estimate
```

```
##   smokeyes
## -0.5770308
```

```
confint(lm3b)["smokeyes",] # we print the relevant confidence interval
```

```
##      2.5 %     97.5 %
## -0.9032985 -0.2507631
```

```
BW$age20 <- BW$age-20
lm3c <- lm(bwt~smoke*age20 + lwt55 + race + ui + ht + ptl2 + fvt2,data=BW)
coef(lm3c)["smokeyes"]
```

```
##   smokeyes
## -0.1481667
```

```
confint(lm3c)["smokeyes",]
```

```
##      2.5 %     97.5 %
## -0.4004816  0.1041482
```

The negative effect of smoking on birth weight is not significant for 20 year old mothers, but is for 30 year old mothers. From the estimated weight differences for smoking and non-smoking mothers at age 20, 27 and 30, we conclude that negative effect of smoking increases with age. We could have concluded this from the significant negative interaction between smoking and age (-0.042886, p-value=0.027). However, this estimated value is not easy to grasp and presenting differences for different ages is easier to understand and communicate. Strictly speaking, the estimate of the interaction term -0.042886 means that the mean difference in birth weight increases with the age of the mothers by -0.042886 kg each time the two mothers become older by one year (See graphical representation of question 6 for a visual interpretation).

Question 5

Yes, in the output of the summary function above we can read the significant negative coefficients for `black` (-0.395832) and `other` (-0.275009). This suggests that babies born from mothers from different racial groups who are otherwise similar with respect to the other covariates in the model differ in birth weight. Precisely, the mean birth weight difference between babies born from a black mother and those born from a white mother (here the arbitrary reference group), when the mothers are otherwise similar with respect to all variables adjusted for, is -0.3958 kg. The negative sign indicates that the babies from black mothers are in average smaller, again, when comparing babies from mothers who are otherwise similar with respect to the other variables. A 95% confidence interval was already computed above (with the **confint()** function). It is [-0.694;-0.0973].

Note, however, note that the above summary does show only two of the three possible comparisons that we can look at, when investigating a difference among the three race groups. In addition, when we look at the results unadjusted for multiple testing (because there are three possible comparisons), then the risk of finding at least one false significant results is not controlled at 5%. Hence the motivation to perform all-pairwise comparisons and to adjust for multiple comparisons, as exemplified in the lecture and R-demo. We now do that.

```
library(multcomp)
Res <- glht(lm3, mcp(race="Tukey")) # make all-pairwise comparisons
summary(Res) # print adjusted p-values (min-P method)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = bwt ~ smoke * age27 + lwt55 + race + ui + ht + ptl2 +
##     fvt2, data = BW)
##
## Linear Hypotheses:
##                     Estimate Std. Error t value Pr(>|t|)
## black - white == 0  -0.3958     0.1513  -2.617   0.0254 *
## other - white == 0  -0.2750     0.1177  -2.336   0.0524 .
## other - black == 0   0.1208     0.1582   0.764   0.7230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
confint(Res) # print adjusted 95% confidence intervals (min-P method)
```

```
##
##   Simultaneous Confidence Intervals
```

```
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = bwt ~ smoke * age27 + lwt55 + race + ui + ht + ptl2 +
##     fvt2, data = BW)
##
## Quantile = 2.3547
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                     Estimate  lwr       upr
## black - white == 0 -0.395832 -0.752030 -0.039634
## other - white == 0 -0.275009 -0.552187  0.002168
## other - black == 0  0.120823 -0.251627  0.493273
```

After adjusting for multiple comparisons because we make all the pairwise comparisons, we observe only one significant difference: between babies born from mothers from black and white groups, who are otherwise similar for the other variables. The p-value that we report to test the association between birth weight and race group is 0.0254 (i.e. the minimum of the adjusted p-values for each comparison). The other two differences, "other vs. white" and "other vs. black" are not significant, now that we have adjusted for multiple comparisons.

Testing the same hypothesis using an F-test instead could be done as follows, by comparing two models: the full model and the same model without the variable race.

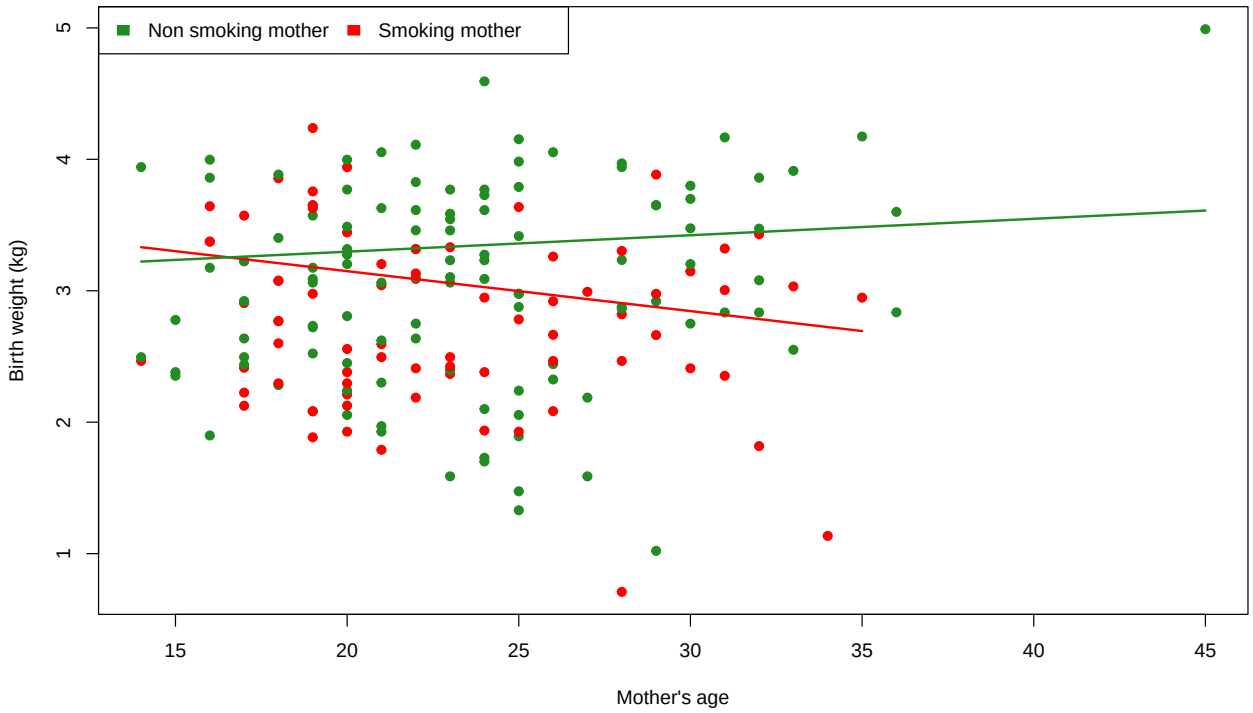
```
anova(lm(bwt~smoke*age27 + lwt + ui + ht + ptl2 + fvt2,data=BW),
      lm(bwt~smoke*age27 + lwt + race + ui + ht + ptl2 + fvt2,data=BW))
```

```
## Analysis of Variance Table
##
## Model 1: bwt ~ smoke * age27 + lwt + ui + ht + ptl2 + fvt2
## Model 2: bwt ~ smoke * age27 + lwt + race + ui + ht + ptl2 + fvt2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    179 76.027
## 2    177 72.237  2    3.7905 4.6438 0.01082 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test is significant too (p-value=0.01082). This means that, according to the F-test, we can also reject the hypothesis of no difference in mean birth weight of babies born from mothers from different racial groups that are otherwise similar for the adjusted variables. The result is similar to the all-pairwise comparisons except that based on the F-test we however cannot conclude anything on which race groups do or do not show a difference. The conclusion is only that some differences exist between some of the groups, tantalizing but not informative.

Question 6

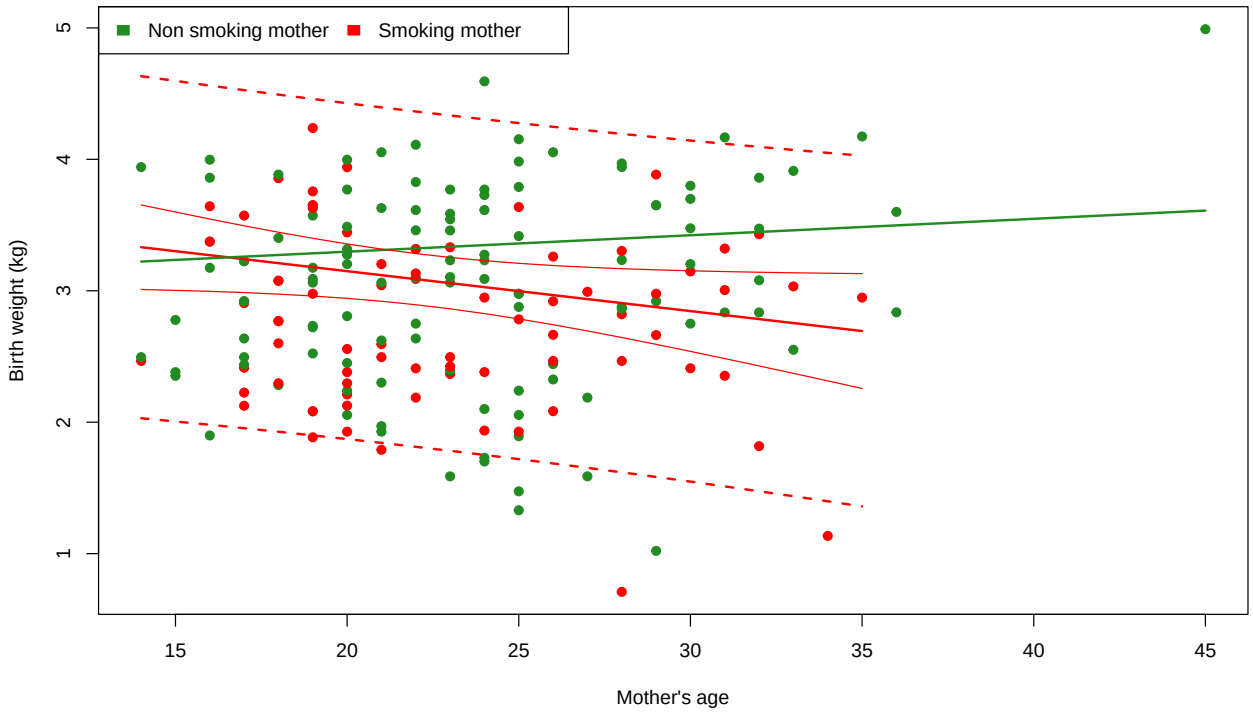
```
# Plot the observations
plot(BW$age,BW$bwt,xlab="Mother's age",ylab="Birth weight (kg)",
     col=ifelse(BW$smoke == "yes","red","forestgreen"),
     pch=19)
# Create "new" data to call the predict funtion
# and compute the estimated mean birth weight
# for specific mother profiles.
# First, profiles for nonsmokers.
dsmoker <- expand.grid(age=BW$age[BW$smoke=="yes"],
                       smoke="yes",
                       lwt=60,
                       race="white",
                       ht=0,
                       ui=0,
                       ptl2="0",
                       fvt2="0")
dsmoker <- dsmoker[order(dsmoker$age),] # order rows by age
# Second, profiles for smokers
dnonsmoker <- expand.grid(age=BW$age[BW$smoke=="no"],
                       smoke="no",
                       lwt=60,
                       race="white",
                       ht=0,
                       ui=0,
                       ptl2="0",
                       fvt2="0")
dnonsmoker <- dnonsmoker[order(dnonsmoker$age),] # order rows by age
# Add the estimated mean birth weight for smokers
lines(dsmoker$age, predict(lm2, dsmoker),col="red",lwd=2)
# Add the estimated mean birth weight for smokers
lines(dnonsmoker$age, predict(lm2, dnonsmoker),col="forestgreen",lwd=2)
# Add the legend
legend("topleft",
       ## pch=19,col=c("forestgreen","red"),
       fill=c("forestgreen","red"),
       border="white",
       legend=c("Non smoking mother","Smoking mother"),
       ncol=2)
```



If any of the assumptions about mother's weight, race, hypertension, uterine irritability, medical consultation or premature birth was changed, the lines would shift up or down but their slopes would not change as there are no interaction effects included for these variables (just age and smoking status).

We now add a graphical representation of confidence intervals and prediction intervals for the birth weights of babies born from smokers.

```
# First compute prediction intervals
predSpi <- predict(lm2, dsmoker,interval='pred')
# Then confidence intervals
predSci <- predict(lm2, dsmoker,interval='conf')
lines(dsmoker$age,predSci[,"upr"],col="red",lwd=1) # upper ci
lines(dsmoker$age,predSci[,"lwr"],col="red",lwd=1) # lower ci
lines(dsmoker$age,predSpi[,"upr"],col="red",lwd=2,lty=2) # upper pi
lines(dsmoker$age,predSpi[,"lwr"],col="red",lwd=2,lty=2) # lower pi
```



Exercise B

Question 1

We first load the data and look at the "summary", as always.

```
load(url("http://paulblanche.com/files/brain.rda"))
summary(brain)
```

```
##      litter           body            brain
##  Min.   : 3.0   Min.   :5.450   Min.   :0.3680
##  1st Qu.: 5.0   1st Qu.:6.641   1st Qu.:0.4065
##  Median : 7.5   Median :7.330   Median :0.4155
##  Mean   : 7.5   Mean   :7.747   Mean   :0.4168
##  3rd Qu.:10.0   3rd Qu.:8.926   3rd Qu.:0.4333
##  Max.   :12.0   Max.   :9.780   Max.   :0.4440
```

We now count the number of observations per litter size.

```
table(brain$litter)
```

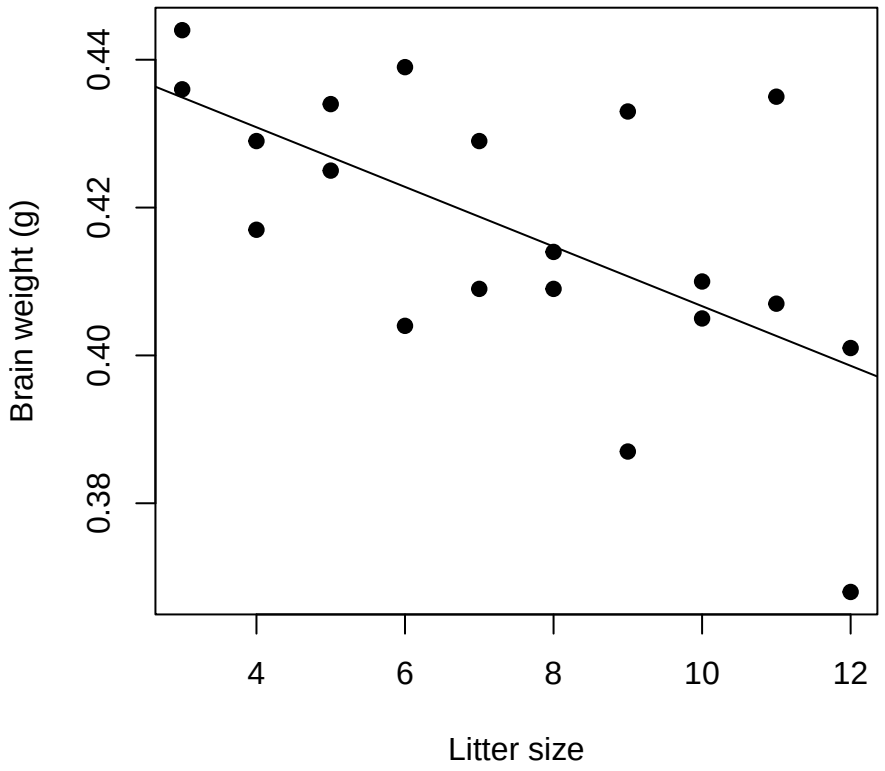
```
##
##  3  4  5  6  7  8  9 10 11 12
##  2  2  2  2  2  2  2  2  2  2
```

We have 2 observations per litter size.

Question 2

We plot the observations of brain weights against those of litter sizes, then add the regression line.

```
plot(brain ~ litter,
     data = brain,
     pch=19,
     xlab = "Litter size",
     ylab = "Brain weight (g)")
lm4 <- lm(brain ~ litter, data = brain)
abline(lm4)
```



We now look at the estimated slope and corresponding confidence interval.

```
summary(lm4)
```

```
##
## Call:
## lm(formula = brain ~ litter, data = brain)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030600 -0.006742  0.000183  0.007650  0.032367
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.447000   0.009625  46.443  < 2e-16 ***
## litter      -0.004033   0.001198  -3.366  0.00344 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01539 on 18 degrees of freedom
## Multiple R-squared:  0.3862, Adjusted R-squared:  0.3521
## F-statistic: 11.33 on 1 and 18 DF,  p-value: 0.003445
```

```
confint(lm4)
```

```
##                    2.5 %      97.5 %
## (Intercept)  0.426779125  0.46722087
## litter      -0.006551127 -0.00151554
```

We conclude that the litter size has a significant negative effect on the average brain weight (p-value=0.00344). With a one mouse increase in litter size, the mean brain weight decreases by -0.0040 grams (95% CI=[-0.0066; -0.0015]).

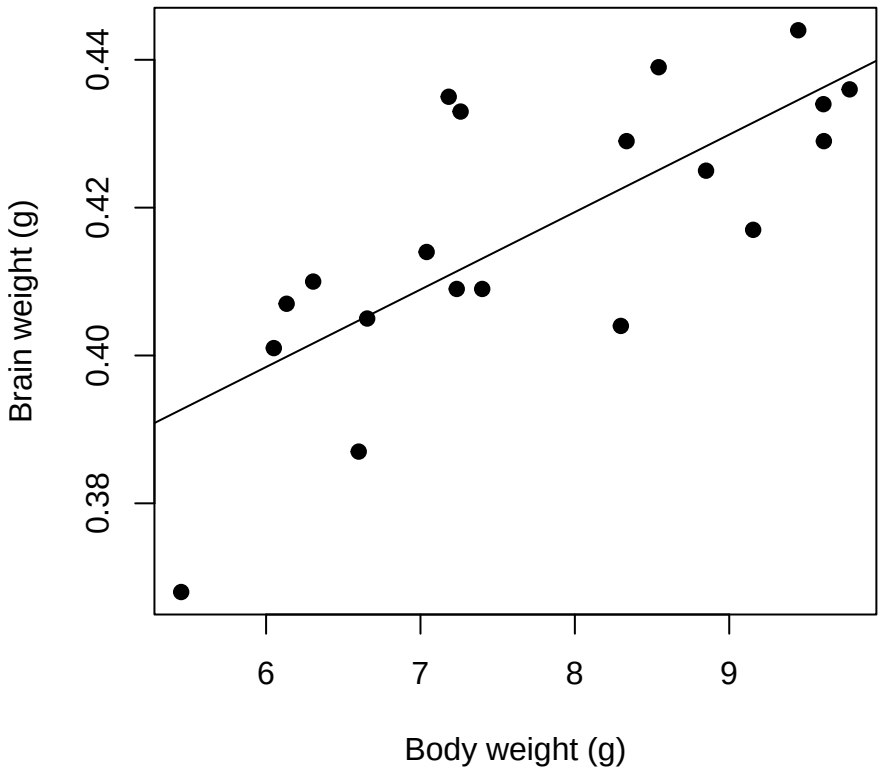
Question 3

Yes, it would be quite unreasonable. As discussed and illustrated in Lecture 3, when one of the two variables is not a random variable but a set of values chosen by the investigator in the design of the study (e.g. dose), it does not make much sense to report a correlation coeficient. Here the number of observations per litter size was chosen to be 2. It was not randomly observed. Hence it does not make much sense to report a correlation coeficient.

Question 4

We plot the observations of brain weights against those of body weights, then add the regression line.

```
plot(brain ~ body,
     data = brain,
     pch=19,
     xlab = "Body weight (g)",
     ylab = "Brain weight (g)")
lm5 <- lm(brain ~ body, data = brain)
abline(lm5)
```



We now look at the estimated slope and corresponding confidence interval.

```
summary(lm5)
```

```
##
## Call:
## lm(formula = brain ~ body, data = brain)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.024679 -0.004910 -0.001179  0.007464  0.024161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.335568   0.017318  19.377 1.66e-13 ***
## body        0.010479   0.002203   4.756 0.000158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01308 on 18 degrees of freedom
## Multiple R-squared:  0.5569, Adjusted R-squared:  0.5323
## F-statistic: 22.62 on 1 and 18 DF,  p-value: 0.0001578
```

```
confint(lm5)
```

```
##                 2.5 %     97.5 %
```

```
## (Intercept) 0.29918471 0.37195158
## body        0.00585007 0.01510793
```

We conclude that the body weight has a significant positive effect on the average brain weight (p-value= 0.000158). With a one gram increase in body weight, the mean brain weight increases by 0.010479 grams (95% CI= [0.0059; 0.0151]). The intercept is not useful in this model since it is an estimate of the average brain weight for body weight equal to 0.

Question 5

We now estimate a multiple linear model for the average brain weights using the two variables litter size and brain weight. We do not model an interaction between these two variables.

```
lm6 <- lm(brain ~ body + litter, data = brain)
summary(lm6)
```

```
##
## Call:
## lm(formula = brain ~ body + litter, data = brain)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0230428 -0.0098897  0.0006377  0.0092065  0.0180602
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.178771   0.075213   2.377  0.02947 *
## body        0.024260   0.006769   3.584  0.00229 **
## litter      0.006671   0.003128   2.132  0.04785 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01196 on 17 degrees of freedom
## Multiple R-squared:  0.6504, Adjusted R-squared:  0.6093
## F-statistic: 15.81 on 2 and 17 DF,  p-value: 0.000132
```

```
confint(lm6)
```

```
##                    2.5 %     97.5 %
## (Intercept) 2.008551e-02 0.33745581
## body        9.978796e-03 0.03854132
## litter      7.088313e-05 0.01327158
```

We conclude the following.

- When we compare the mean (average) brain weight of two litters, one with (average) body weight 1 gram heavier than the other, both having the same litter size, then the

difference is 0.024 grams (95% CI=[0.001;0.039], p-value=0.0229). That is, for the same litter size, the heavier the body weight the heavier the brain weight.

- When we compare the mean (average) brain weight of two litters, one larger by one mouse than the other, both having the same (average) brain weight, then the difference is 0.007 grams (95% CI=[0.001;0.013], p-value=0.04785). That is, for the same (average) body weight, the larger the litter size the heavier the brain weight.

At first sight these results can be surprising. Indeed, the coefficient for litter size switched sign from negative to positive, from the univariate (question 2) to the multivariate analysis results (second item above). The apparent contradiction between the results of the two models is actually not a contradiction. The results of each model have a different interpretation and they both make sense.

Indeed, we can believe/expect that the larger the litter the smaller the mice and hence the smaller the brains. Hence the result from questions 2. But, for any given litter size, we can believe that the larger the body the larger the brain, no matter the litter size. Hence the results of the first item above. Finally, we can expect that in litters having the same average body weight, mice from larger litters are expected to have a relatively larger brain, because the brain is one of the most important part of the body. In fact, this is a well known phenomenom: if offspring are breeded under growth constraints, the development of the vital organs is prioritized. Hence the results of the second item above.

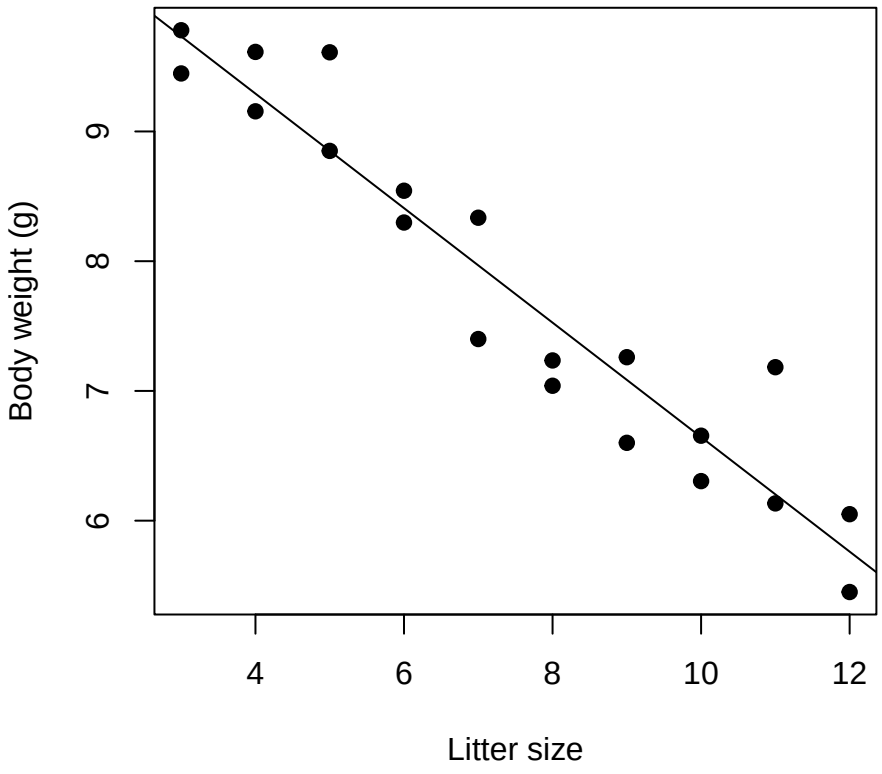
In the above explanation, we stated that "*we can believe/expect that the larger the litter the smaller the mice*". The topic of the next question is to check whether we indeed see that in the data.

Note: the intercept does not have a meaningful interpretation in this model, as it is an estimate of the average brain weight for an empty litter with body weight of 0 gram.

Question 6

We plot the observations of body weights against those of litter sizes, then add the regression line.

```
plot(body ~ litter,
     data = brain,
     pch=19,
     xlab = "Litter size",
     ylab = "Body weight (g)")
lm7 <- lm(body ~ litter, data = brain)
abline(lm7)
```



We now look at the estimated slope and corresponding confidence interval.

```
summary(lm7)
```

```
##
## Call:
## lm(formula = body ~ litter, data = brain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56772 -0.29649 -0.03498  0.20320  0.98025
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.05642    0.26027   42.48  < 2e-16 ***
## litter      -0.44124    0.03241  -13.62 6.44e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4163 on 18 degrees of freedom
## Multiple R-squared:  0.9115, Adjusted R-squared:  0.9066
## F-statistic: 185.4 on 1 and 18 DF,  p-value: 6.443e-11
```

```
confint(lm7)
```

```
##                  2.5 %     97.5 %
```

```
## (Intercept) 10.5096026 11.6032338
## litter      -0.5093289 -0.3731559
```

We conclude that, as hypothesized at the last question, the litter size is significantly negatively associated with the average body weight (p-value<0.0001). With a one unit increase in litter size the mean body weight decreases by -0.44124 grams (95% CI=[-0.509;-0.373]). Again, the intercept is not useful in this model, same reasons as above.

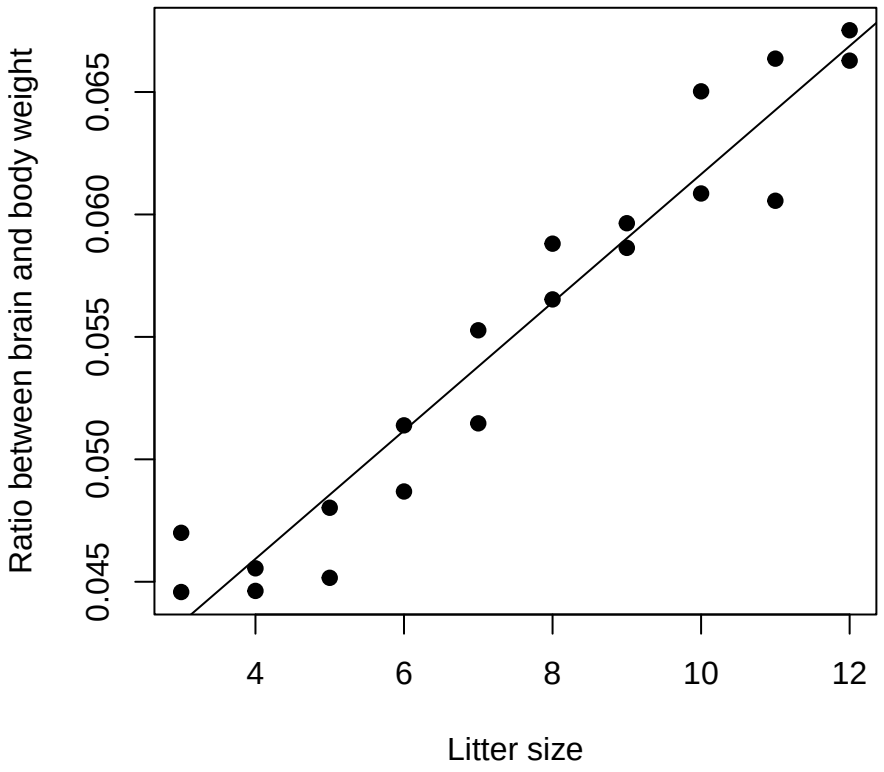
Question 7

Yes! In the explanation provided in our answer to question 5, we stated that "*we can believe/expect that the larger the litter the smaller the mice*". The results to the previous questions indeed support this statement.

Question 8

We plot the observations of the ratio of brain versus body weights against those of litter sizes, then add the regression line.

```
brain$ratio <- brain$brain/brain$body
plot(ratio ~ litter,
     data = brain,
     pch=19,
     xlab = "Litter size",
     ylab = "Ratio between brain and body weight")
lm8 <- lm(ratio ~ litter, data = brain)
abline(lm8)
```



We now look at the estimated slope and corresponding confidence interval.

```
summary(lm8)
```

```
##
## Call:
## lm(formula = ratio ~ litter, data = brain)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0037024 -0.0009179 -0.0001321  0.0013173  0.0036813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.035464   0.001317   26.93 5.38e-16 ***
## litter      0.002618   0.000164   15.97 4.52e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002106 on 18 degrees of freedom
## Multiple R-squared:  0.934,  Adjusted R-squared:  0.9304
## F-statistic: 254.9 on 1 and 18 DF,  p-value: 4.518e-12
```

```
confint(lm8)
```

```
##                   2.5 %      97.5 %
```

```
## (Intercept) 0.032696858 0.038230484
## litter      0.002273529 0.002962546
```

We conclude that the litter size has a significantly positive effect on the average brain/body weight ratio (p-value<0.001). With a one mouse increase in litter size the brain-body weight ratio increases by 0.0026 (95% CI=[0.0023;0.0030]), meaning that the average brain weight increases by 0.26% relative to the body weight relative to the body weight (95% CI=[0.23%;0.30%]).