Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Final Exam
SA.01

Saturday, 19. 1. 2019

Duration: 90 Minutes

Surname, Forename: _____

IDS Nr.: (On the back of your HSLU-Card) L _____

Stick number: _____

Signature: _____

Problem	1	2	3	4	Total
max. Points	5	10	10	10	35
achieved Points					

Important Information

- Allowed aids:
  a) Printed R *reference card*.
  b) Ten (10) pages DIN-A4, with arbitrary content. This corresponds to five sheets DIN-A4 paper written on both sides.
  c) Calculator
  d) Formulary
- Other aids are not allowed! Switch off all phones.
- Place your HSLU-Card in front of you on your table.
- All of your answers should come with explanations! Solutions without understandable justifications obtain no credit.

Good luck!
Peter Büchel

Saturday, 19. 1. 2019

Problem 1: ..... (5 Points)

A test with a lie detector is done routinely to employees working in sensitive positions. Let $+$ denote the event of a positive test, meaning the lie detector claims the employee is lying. Let W denote the event, that the employee is actually telling the truth, and L the event, that the employee is actually lying.

Test with the lie detector show that

$$P(+|L) = 0.88 \quad \text{and} \quad P(-|W) = 0.86$$

Further it is assumed that

$$P(W) = 0.99$$

(a) Interpret the meaning of $P(W)$ and $P(+|L)$.

(b) The lie detector claims a person is lying. What is the probability that the employee actually lied?

(c) Interpret the result of (b) in 2-3 sentences. How convincing is the use of the lie detector?

Solution 1:

a) $P(W)$: Probability that an employee is telling the truth.
$P(+|L)$: Probability that the lie detector claims the person is lying, under the assumption that the employee is actually lying.

b) Sought: $P(L|+)$

We are using the Bayes' theorem:

$$P(L|+) = \frac{P(+|L) \cdot P(L)}{P(+)}$$

On the right hand side of the equation, there are three expressions ($P(+|L)$, $P(L)$, $P(+)$) which are known or which we can calculate:

- The probability $P(+|L) = 0.88$ is given.
- Then
$$P(L) = 1 - P(W) = 1 - 0.99 = 0.01$$
- We can calculate $P(+)$ with the law of total probability:

$$\begin{aligned} P(+) &= P(+|W)P(W) + P(+|L)P(L) \\ &= (1 - P(-|W))P(W) + P(+|L)P(L) \\ &= 0.14 \cdot 0.99 + 0.88 \cdot 0.01 \\ &= 0.1474 \end{aligned}$$

Above, we used

$$P(+|W) = 1 - P(-|W)$$

and $P(-|W)$ is given.

Finally

$$P(L|+) = \frac{P(+|L) \cdot P(L)}{P(+)} = \frac{0.88 \cdot 0.01}{0.1474} = 0.0597$$

c) If the test shows a person is lying, the probability that the person is actually lying is only 6 %. In other words 94 % of the employees that the test claims to be lying are actually telling the truth.

The test hence accuses someone wrongly of lying with high probability.

Problem 2: ..... (10 Points)

On the 14th of April 1912, the *Titanic* hit an iceberg and sank. Of about 2245 passengers approximately 1500 died. The data sets **survived.dat** and **died.dat** contain the age of the (known) passengers who survided or died respectively.

We want to investigate whether there is an age difference between those passengers who survived and those who died.

Hint: The commands for loading the data sets are included in the **R**-file on the stick.

(a) We want to investigate the age difference with an hypothesis test. *[2]*

i) Do you choose a test for paired or unpaired samples? Justify your answer!

ii) Do you choose a one-sided or two-sided test? Justify your answer!

(b) Calculate the average and standard deviation and give an interpretation of these values. *[2]*

(c) Initially we perform a t-test. *[4]*

i) What are the requirements to justify a t-test?

ii) Describe the null hypothesis and the alternative hypothesis.

iii) Perform the test and use the p-value for the test decision on a significance level of 5 %.

iv) Use the confidence interval for the test decision.

(d) We perform a Wilcoxon-test. *[2]*

a) Give an argument to use the Wilcoxon-test instead of a t-test.

b) What is your test decision in this case?

c) Explain the difference of the p-values from the t- and Wilcoxon-test?

Solution 2:

(a) i) We make an unpaired test because there is no connection between those who survided and those who died. Furthermore, the data sets have different sizes so they cannot be paired.

ii) We don't know in which direction there is a difference (if there is one). So we make a two-sided test.

(b) Averages and standard deviations:

```
survived <- read.table("survived.dat")$x
died <- read.table("died.dat")$x

mean(survived)
## [1] 28.40839
sd(survived)
## [1] 14.42786
mean(died)
## [1] 30.13853
sd(died)
## [1] 13.89832
```

It *seems* that the survivors have a smaller average age than the dead.

The standard deviation is very similar in both case, 14 years. That means on "avergage", the deviation from the average is 14 years. This is quite a lot compared to the average. We can justified this by the fact, that there were children and older people on the *Titanic*.

(c) i) A t-test is appropriate if the data is normally distributed. The variance is unknown and is estimated from the data.

ii) Notation:

- μ_s: Average age of the survivors
- μ_d: Average age of the dead

Null hypothesis:

$$H_0: \quad \mu_s = \mu_d$$

Alternativ hypothesis:

$$H_0: \quad \mu_s \neq \mu_d$$

iii) t-test:

```r
t.test(survived, died, paired = F, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  survived and died
## t = -1.763, df = 704.1, p-value = 0.07834
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.6569075  0.1966269
## sample estimates:
## mean of x mean of y
##  28.40839  30.13853
```

The p-Wert is 0.078 and therefore slightly bigger than the significance level of 5 %. Thus the null hypothesis is not rejected. There is no statistically significant difference in the average age of the survivors and the dead.

iv) We can formulate the null hypothesis as follows:

$$H_0: \quad \mu_s - \mu_d = 0$$

Now, the confidence interval is

$$[-3.66,\ 0.2]$$

0 is not in this interval, thus the null hypothese is not rejected.

(d) i) The assumption that the data are normally distributed is often too strong and not justified. The Wilcoxon-test assumes less, particulary that the distribution is symmetrical. This test is often more appropriate than the t-test.

ii) Wilcoxon-Test:

```r
wilcox.test(survived, died, paired = F, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity
##  correction
##
## data:  survived and died
## W = 89850, p-value = 0.3677
## alternative hypothesis: true location shift is not equal to 0
```

In this case, the p-value 0.37 is even further above the significance level than with the t-test.

iii) The p-value is bigger than the one from the t-test. This is (almost) always the case. The Wilcoxon-test assumes less than the t-test, therefore there is a built-in greater uncertainty. That means, the difference between the averages must be bigger for the Wilcoxon-test to reject the null hypothesis.

Problem 3: ..... (10 Points)

The data set `potato.csv` relates to a study which investigated the quality of potatoes with respect to texture, flavor and moisture depending on several factors (A. Mackey and J. Stockman (1958). „Cooking Quality of Oregon-Grown Russet Potatoes“, American Potato Journal, Vol.35, pp395-407).

In this problem, we just consider the factors `Size` (size of the potato: Large, Medium) and `CookingMethod` (Boil, Steam, Mash, Bake.350 (Fahrenheit), Bake.450).

Hint: The command for loading the data set is included in the `R`-file on the stick.

(a) Generate two boxplots with response variable `MoistnessScore` (the smaller this value, the drier the potato) depending on the predictors `Size` and `CookingMethod` respectively. *[2]*

Interpret these plots using the notions of medium value.

(b) Generate an interaction plot of `Size`, `CookingMethod` and `MoistnessScore`. Interpret this plot. *[3]*

(c) Perform an analysis of variance for the response variable `MoistnessScore` depending on the predictors `Size` and `CookingMethod`. *[3]*

Take interaction into account. Describe null hypotheses, alternative hypotheses and give an interpretion of your test decision on a significance level of 5 %.

(d) Repeat (c) for `FlavorScore` (the higher this value, the better the taste of the potato). Interpret this result. *[2]*

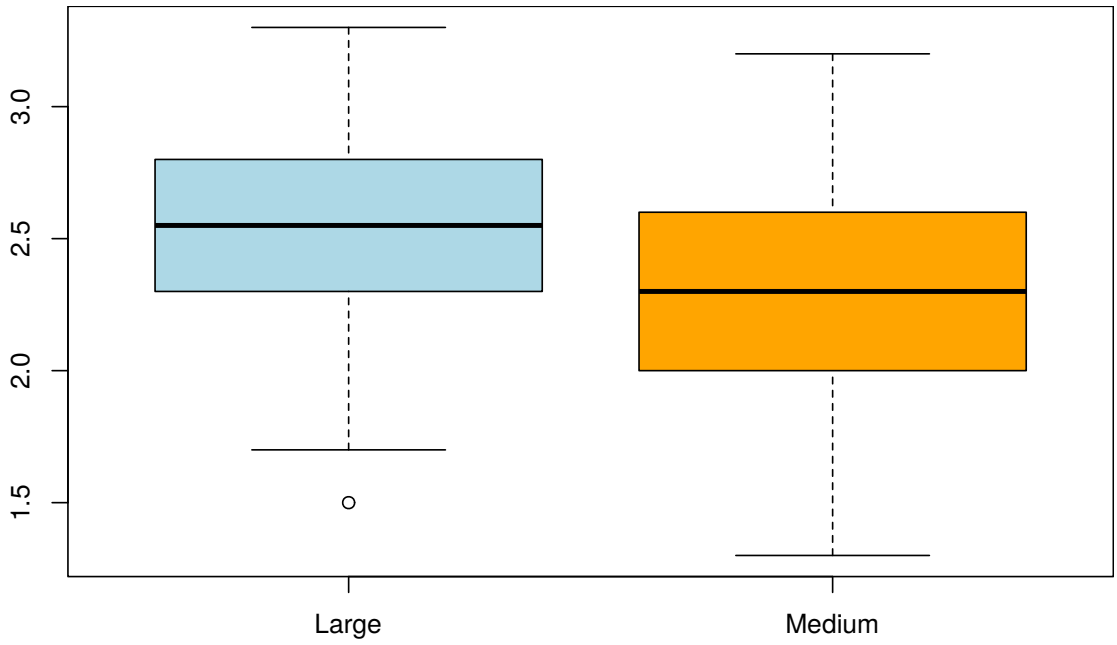
Solution 3:
Datei einlesen

```r
df <- read.csv("potato.csv")[, -1]
head(df)
```

```
##   GrowingArea HoldingTemp  Size StoragePeriod
## 1           1           1 Large             1
## 2           1           1 Large             1
## 3           1           1 Large             1
## 4           1           1 Large             1
## 5           1           1 Large             1
## 6           1           1 Large             2
##   CookingMethod TextureScore FlavorScore
## 1          Boil          2.9         3.2
## 2         Steam          2.3         2.5
## 3          Mash          2.5         2.8
## 4      Bake.350          2.1         2.9
## 5      Bake.450          1.9         2.8
## 6          Boil          1.8         3.0
##   MoistnessScore
## 1            3.0
## 2            2.6
## 3            2.8
## 4            2.4
## 5            2.2
## 6            1.7
```

(a) Boxplot

```r
boxplot(MoistnessScore ~ Size,
        data = df,
        col = c("lightblue", "orange", "yellow3", "violetred3",
                "seagreen2")
)
```

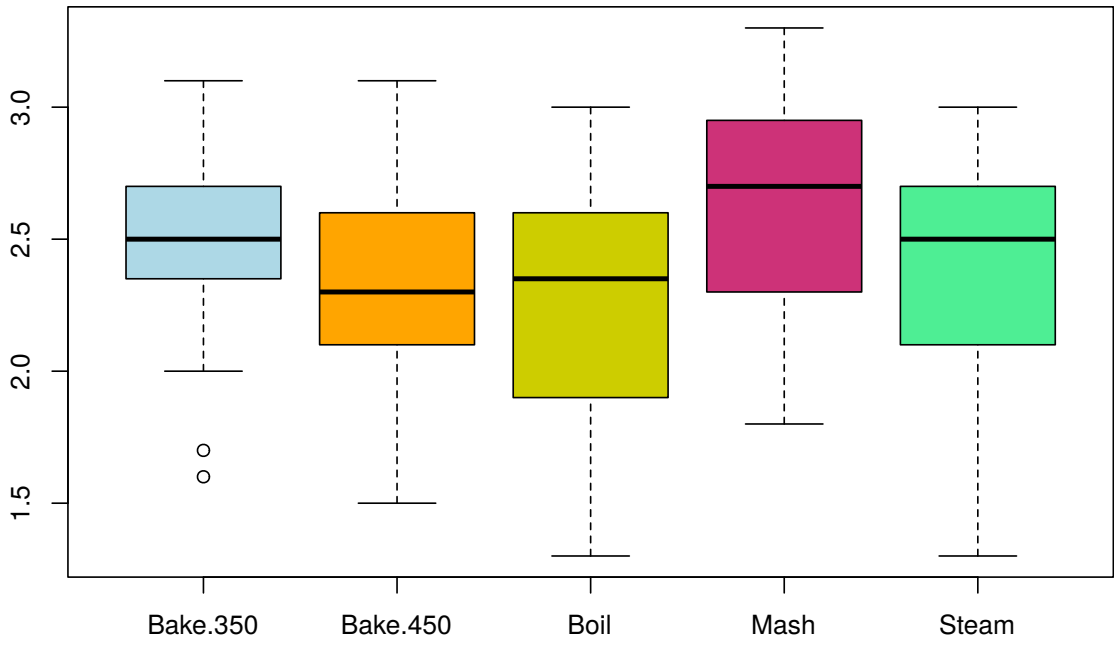


The median corresponding to the moisture of the potatoes is for large potato-

es slightly bigger ($\approx$ 2.6) than for medium potatoes ($\approx$ 2.4). The variation is smaller for large potatoes than for medium potatoes. Large potatoes seems to be more uniform in terms of moisture than medium potatoes.

Boxplot

```
boxplot(MoistnessScore ~ CookingMethod,
        data = df,
        col = c("lightblue", "orange", "yellow3", "violetred3",
                "seagreen2")
)
```

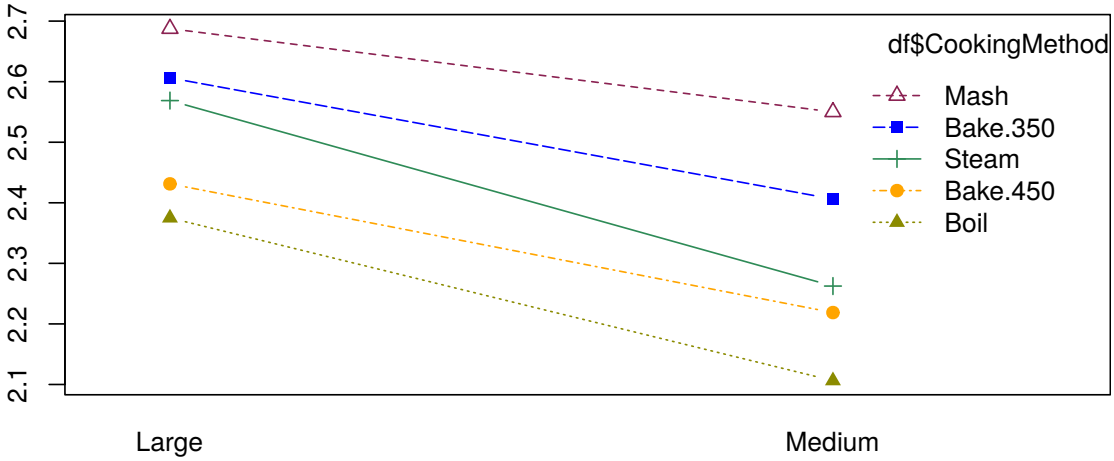


The median is for **Mash** the biggest, which is not surprising because moisture is added with milk/water. The smallest median has **Bake.450**, which is also not surprising, because the potatoes are loosing moisture.

However, die differences are not particularly large (which, of course, we have to verify with an analysis of variance).

(b) Interaktionplot:

```
interaction.plot(df$Size,
                 df$CookingMethod,
                 df$MoistnessScore,
                 type = "b",
                 col = c("blue", "orange", "yellow4", "violetred4",
                         "seagreen4"),
                 pch = c(15,19,17,2,3)
)
```



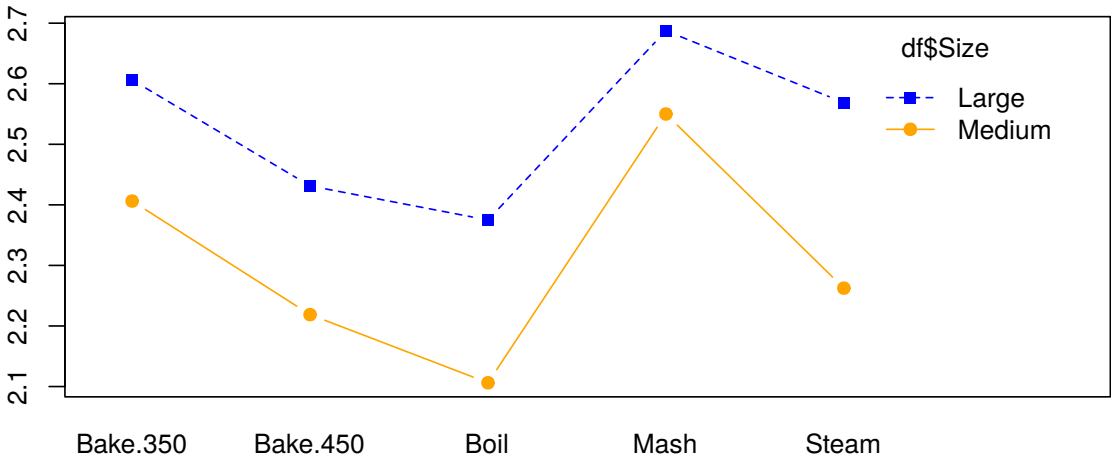
The interaction plot kind of summarizes the boxplots above but it uses averages instead of medians.

The moisture is lower for medium potatoes than for large potatoes. This is indicated by the descending lines for all cooking methods.

Again, there are differences between each cooking methods. The lines lie apart (instead on each other).

Additionally, we see that lines are almost parallel. This indicates that there is no interaction.

```
interaction.plot(df$CookingMethod,
                 df$Size,
                 df$MoistnessScore,
                 type = "b",
                 col = c("blue", "orange", "yellow4", "violetred4",
                         "seagreen4"),
                 pch = c(15,19,17,2,3)
)
```



The interpretation in this case is accordingly.

(c) We have 3 null hypothesis with their corresponding alternative hypothesis:

- Null hypothesis H_{01}: There is *no* difference in the average cooking method on the moisture of the potatoes.
  Alternative hypothesis H_{A1}: There is a difference in the average cooking method of the moisture of the potatoes.
- Null hypothesis H_{02}: There is *no* difference in the average potato size of the moisture of the potatoes.
  Alternative hypothesis H_{A1}: There is a difference in the average potato size on the moisture of the potatoes.
- Null hypothesis H_{03}: There is *no* interaction between cooking method and potato size on the moisture of the potatoes.
  Alternativ hypothesis H_{A3}: There is interaction between cooking method and potato size on the moisture of the potatoes.

R-output

```
fit <- aov(MoistnessScore ~ CookingMethod * Size, data = df)

summary(fit)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## CookingMethod         4  2.821  0.7052   4.474 0.001919
## Size                  1  2.025  2.0250  12.846 0.000457
## CookingMethod:Size    4  0.136  0.0339   0.215 0.929727
## Residuals           150 23.646  0.1576
##
## CookingMethod      **
## Size               ***
## CookingMethod:Size
## Residuals
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *p*-value for the cooking methods is 0.002 and therefore far below the significance level of 5 %. The null hypothesis is rejected. There is a statistically significant difference of the cooking methods on the moisture of the potatoes although the boxplots above didn't indicate this. However, there are 160 measurements which is quite large and a rather small difference is significant.

```
dim(df)
```

```
## [1] 160   8
```

The *p*-value for the potato size is 0.0005 and therefore far below the significance level of 5 %. The null hypothesis is rejected. There is a statistically significant difference of the potato size on the moisture of the potatoes although the boxplots above didn't indicate this.

The *p*-value for the interaction is 0.93 and far above the significance level of 5 %. The null hypothesis is not rejected. There is no statistically difference between the potato size and the cooking methods on the moisture of the potatoes.

(d) We have 3 null hypothesis with their corresponding alternative hypothesis:

- Null hypothesis H_{01}: There is *no* difference in the average cooking method on the flavor of the potatoes.
  Alternative hypothesis H_{A1}: There is a difference in the average cooking method of the flavor of the potatoes.
- Null hypothesis H_{02}: There is *no* difference in the average potato size of the flavor of the potatoes.
  Alternative hypothesis H_{A1}: There is a difference in the average potato size on the flavor of the potatoes.
- Null hypothesis H_{03}: There is *no* interaction between cooking method and potato size on the flavor of the potatoes.
  Alternativ hypothesis H_{A3}: There is interaction between cooking method and potato size on the flavor of the potatoes.

R-output

```
fit <- aov(FlavorScore ~ CookingMethod * Size, data = df)

summary(fit)
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## CookingMethod        4  1.344  0.3359   4.459 0.00197
## Size                 1  0.000  0.0002   0.003 0.95414
## CookingMethod:Size   4  0.076  0.0190   0.252 0.90795
## Residuals          150 11.300  0.0753
##
## CookingMethod      **
## Size
## CookingMethod:Size
## Residuals
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the cooking methods is 0.002 and therefore far below the significance level of 5 %. The null hypothesis is rejected. There is a statistically significant difference of the cooking methods on the flavor.

The p-value for the potato size is 0.95 and therefore far above the significance level of 5 %. The null hypothesis is not rejected. There is a no statistically significant difference of the potato size on the flavor of the potatoes.

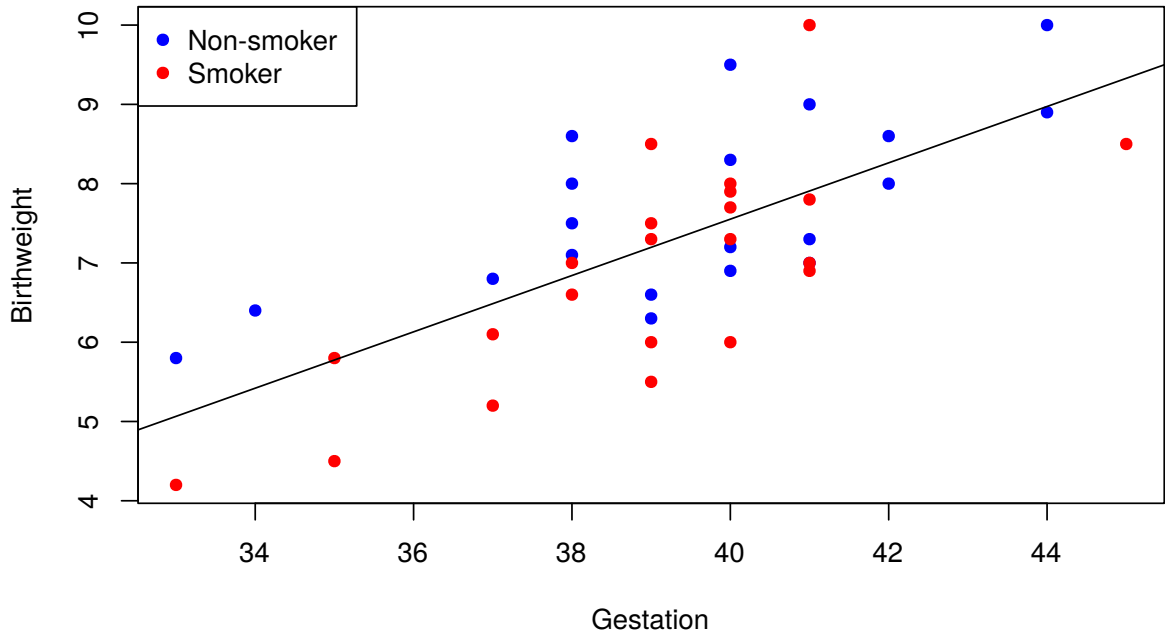
The p-value for the interaction is 0.91 and far above the significance level of 5 %. The null hypothesis is not rejected. There is no statistically difference between the potato size and the cooking methods on the flavor of the potatoes.

Problem 4: ..... (10 Points)

The file **birthweight.csv** contains information of newborn babies and their parents. We want to investigate which variables affect the birthweight of the babies. We consider just a few variables.

Hint: The command for loading the data set is included in the **R**-file on the stick.

(a) Consider the following scatter plot of **Gestation** (length of pregnancy in weeks) and **Birthweight** (in lbs). The black line is the regression line. *[3]*



i) Interpret the scatterplot *without* taking colours into account.

ii) Interpret the scatterplot taking colours into account.

(b) We use a regression model to investigate whether the birthweight (**Birthweight** in lbs) of babies depends on the age of the mother (**motherage** in years), the length of pregnancy (**Gestation** in weeks), the smoking behaviour of the mother (**smoker**) and the weight of the mother before birth (**mppwt** in lbs). *[7]*

Note that **smoker** is a factorial variable: 0=Non smoker, 1=smoker

i) Write out an equation describing the multiple linear regression model for the variables mentioned above.

ii) Determine the parameters of this model and give an interpretation of these values?

iii) What part of the variance is explained by the regression model?

iv) Interpret the p-value for the corresponding F-value.

v) We consider the individual regression coefficents. Is there any indication that we can remove any variables from the model? Justify your answer with p-values on a significance level of 5 %.

Solution 4:

Load the data set

```
df <- read.csv("birthweight.csv")
head(df)
```

```
##     id headcirumference length Birthweight Gestation
## 1 1313                12     17         5.8        33
## 2  431                12     19         4.2        33
## 3  808                13     19         6.4        34
## 4  300                12     18         4.5        35
## 5  516                13     18         5.8        35
## 6  321                13     19         6.8        37
##   smoker motherage mnocig mheight mppwt fage fedyrs
## 1      0        24      0      58    99   26     16
## 2      1        20      7      63   109   20     10
## 3      0        26      0      65   140   25     12
## 4      1        41      7      65   125   37     14
## 5      1        20     35      67   125   23     12
## 6      0        28      0      62   118   39     10
##   fnocig fheight lowbwt mage35 LowBirthWeight
## 1      0      66      1      0            Low
## 2     35      71      1      0            Low
## 3     25      69      0      0         Normal
## 4     25      68      1      1            Low
## 5     50      73      1      0            Low
## 6      0      67      0      0         Normal
```

(a) i) The plot indicates that the longer the pregnancy the heavier the baby, which makes sense.

ii) The red dots (smokers) are more or less below the regression line, the blue dots (non-smokers) more or less above. That means that the babies of smoker have a sligthly smaller birthweight than the babies of non-smokers.

(b) i) Model:

$$\texttt{Birthweight} = \beta_0 + \beta_1 \cdot \texttt{motherage} + \beta_2 \cdot \texttt{Gestastion} + \beta_3 \cdot \texttt{smoker} + \beta_4 \cdot \texttt{mppwt}$$

ii) Output:

```r
fit <- lm(Birthweight ~ motherage + Gestation + smoker + mppwt,
          data = df
)

summary(fit)

##
## Call:
## lm(formula = Birthweight ~ motherage + Gestation + smoker + mppwt,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33758 -0.64529 -0.07245  0.69843  1.94012
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.093956   2.169154  -3.270  0.00233 **
## motherage   -0.004720   0.025693  -0.184  0.85525
## Gestation    0.313003   0.053620   5.837 1.04e-06 ***
## smoker      -0.654434   0.277582  -2.358  0.02379 *
## mppwt        0.020300   0.009256   2.193  0.03465 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8733 on 37 degrees of freedom
## Multiple R-squared:  0.6107,Adjusted R-squared:  0.5687
## F-statistic: 14.51 on 4 and 37 DF,  p-value: 3.23e-07
```

- The parameter for **`Intercept`** is -7.1. That means, the birthweight of a baby whose mother is 0 years old, for whom the pregnancy lastet 0 weeks, whose mother is a non-smoker and has a weight of 0 lbs, is -7.1 lbs. Of course, this doesn't make any sense at all.
- The parameter for **`motherage`** is -0.005. That means, for each year which the mother is older the birthweight of the babies is 0.005 lbs less. However, the corresponding p-value is not signigicant.
- The parameter for **`Gestation`** is 0.31. That means, for each additional week of pregnancy, the babies weigh 0.31 lbs more.
- The parameter for **`smoker`** is -0.65. That means, is the mother a smoker the birthweight is 0.65 lbs less than for non-smokers.
- The parameter for **`mppwt`** is 0.02. That means, for each additional lbs of the motherweight, the birthweight is 0.02 lbs bigger.

iii) The R^2-value is 0.61. That means, 61 % of the variation is explained by the model.

iv) The p-value of the F-Wertes is $3 \cdot 10^{-7}$, which is highly significant. That means, that at most one predictor contributes to the birthweight of the babies.

v) The p-value for **`motherage`** is 0.86, which is way above the significance level. It seems that the age of the mother has no influence on the birthweight

of the babies.