Statistical Modeling Homework 1 - Computational Statistics

```
######################################################################
# Umut Bayrak
# Statistical Modeling Homework
# Last changed: 26.11.2017 17:20
#
# R version 3.4.1 (2017-06-30) -- "Single Candle"
# Copyright (C) 2017 The R Foundation for Statistical Computing
# Platform: x86_64-w64-mingw32/x64 (64-bit)
######################################################################

# Question 1 #
  # (a) - i.

# Blood pressure is modelled as follows:
# Yi = B0 + B1*x1i + B2*x2i + B3*x1i*x2i + epsilon_i

# P(X2i = 1| X1i = 0) = q0
# P(X2i = 1| X1i = 1) = q1 with 70% of the patients taking the treatment


# Data generating true model coefficients are defined:
# I assume that this treatment is for people who have low blood pressure
# beta0 is below average systolic blood pressure
beta0 = 80
beta1 = 4
beta2 = 8   # It's known that the effect is extreme
beta3 = 2

# Standard deviation of error
sigma = 4


# The probability of not getting a treatment but gene exists is q0
# The probability of getting a treatment and gene exists is q1
q0 = 0.65
q1 = 0.50

# n is going to be used to create the dataset
n=100

# Since the x1 and x2 values are binomial(either treatment or gene exists or not),
# I can use rbinom to create dataset.
# However, since the distribution of x2 dependent on x1 with q0 and q1
# I will fill x2 in a for loop to make the distribution work

# Probability of having 1 in x1 is 0.70
p_x1 =  0.70
x1 <- rbinom(n, size = 1, prob = p_x1)
```

```r
x2 <- sample(NA, size = n, replace=TRUE)
epsilon = rnorm(n, mean = 0, sd = sigma)

for(i in 1:n)
{  if (x1[i] == 1){
    x2[i] = rbinom(1, size = 1, prob = q1)
  }
  else {
    x2[i] = rbinom(1, size = 1, prob = q0)
  }
}

# asim: number of simulations
asim <- 1000

# Parameter estimates vectors are created:
b0_model1 <- vector("numeric",asim)
b1_model1 <- vector("numeric",asim)
b2_model1 <- vector("numeric",asim)
b3_model1 <- vector("numeric",asim)

b0_model2 <- vector("numeric", asim)
b1_model2 <- vector("numeric", asim)

for(i in 1:asim)
{ set.seed(i)

  # True data generator equation is below:
  y <- beta0 + (beta1 * x1) + (beta2 * x2) + (beta3 * x1*x2) + epsilon

  # With gene all the parameters and coefficients are present
  model1 <- lm(formula = y ~ x1 + x2 + x1*x2)
  b0_model1[i] <- summary(model1)$coef[1,1]
  b1_model1[i] <- summary(model1)$coef[2,1]
  b2_model1[i] <- summary(model1)$coef[3,1]
  b3_model1[i] <- summary(model1)$coef[4,1]

  # Withouth gene (x2 = 0) only x1, b0 and b1 are present.
  model2 <- lm(formula = y ~ x1)
  b0_model2[i] <- summary(model2)$coef[1,1]
  b1_model2[i] <- summary(model2)$coef[2,1]
}

# When the gene exists:
# The average effect of the treatment is 4.504474
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x1 * x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0587  -3.1212   0.1333   2.9365  11.3517
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.659      1.667  48.985  < 2e-16 ***
## x1             2.402      1.876   1.280  0.20359
## x2             6.065      1.972   3.075  0.00274 **
## x1:x2          4.643      2.272   2.044  0.04371 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.715 on 96 degrees of freedom
## Multiple R-squared:  0.5485, Adjusted R-squared:  0.5344
## F-statistic: 38.87 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
# When the gene does not exist:
# The average effect of the treatment is 4.791741
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3054  -4.3877  -0.8947   4.5922  15.8136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   85.992      1.259  68.288   <2e-16 ***
## x1             4.316      1.484   2.908   0.0045 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.663 on 98 degrees of freedom
## Multiple R-squared:  0.07945,    Adjusted R-squared:  0.07005
## F-statistic: 8.458 on 1 and 98 DF,  p-value: 0.004497
```

```
  # (a) - ii.

b0_model3 <- vector("numeric",asim)
b1_model3 <- vector("numeric",asim)
b2_model3 <- vector("numeric",asim)

for(i in 1:asim)
{ set.seed(i)

  # True data generator equation is below:
  y <- beta0 + (beta1 * x1) + (beta2 * x2) + epsilon

  # With gene all the parameters and coefficients are present
  model3 <- lm(formula = y ~ x1 + x2)
  b0_model3[i] <- summary(model3)$coef[1,1]
  b1_model3[i] <- summary(model3)$coef[2,1]
  b2_model3[i] <- summary(model3)$coef[3,1]
```

```
}

summary(model3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4383  -3.3278   0.1477   3.2179  11.6228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.2362     1.1346  70.719  < 2e-16 ***
## x1            4.2045     1.0598   3.967 0.000139 ***
## x2            8.0579     0.9804   8.219 9.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.724 on 97 degrees of freedom
## Multiple R-squared:  0.441,  Adjusted R-squared:  0.4294
## F-statistic: 38.26 on 2 and 97 DF,  p-value: 5.631e-13
```

```
# True coefficient of treatment is defined as 4, but the mean of b1 is 5.425385
# Since they are not equal, the estimator is biased.

  # (a) - iii.
# To compare variances I used anova() function:
# According to this, p-values are very close to 0
# Null hypothesis claiming that coefficients are zero is rejected.

anova(model1, model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x1 * x2
## Model 2: y ~ x1
## Model 3: y ~ x1 + x2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     96 2134.2
## 2     98 4351.2 -2   -2217.0 49.861 1.413e-15 ***
## 3     97 2164.3  1    2186.9 98.369 2.230e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
  # (b) - i.
# This case is examined in the part (a) - i. withouth gene effect.
# The resulting coefficient of b1 is 4.791741 when true mean is defined as 4
# Thus, this estimation is biased as well.

  # (b) - ii.
anova(model2, model1)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x2 + x1 * x2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     98 4351.2
## 2     96 2134.2  2      2217 49.861 1.413e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Since p-value is really small, the null hypothesis is rejected.

  # (c)

b0_model4 <- vector("numeric",asim)
b1_model4 <- vector("numeric",asim)
b2_model4 <- vector("numeric",asim)

b0_model5 <- vector("numeric",asim)
b1_model5 <- vector("numeric",asim)


for(i in 1:asim)
{ set.seed(i)

  # True data generator equation is below:
  y <- beta0 + (beta1 * x1) + (beta2 * x2) + epsilon

  # With gene all the parameters and coefficients are present
  model4 <- lm(formula = y ~ x1 + x2)
  b0_model4[i] <- summary(model4)$coef[1,1]
  b1_model4[i] <- summary(model4)$coef[2,1]
  b2_model4[i] <- summary(model4)$coef[3,1]

  # Withouth gene (x2 = 0) only x1, b0 and b1 are present.
  model5 <- lm(formula = y ~ x1)
  b0_model5[i] <- summary(model5)$coef[1,1]
  b1_model5[i] <- summary(model5)$coef[2,1]
}

summary(model4)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4383  -3.3278   0.1477   3.2179  11.6228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.2362     1.1346  70.719  < 2e-16 ***
## x1            4.2045     1.0598   3.967 0.000139 ***
## x2            8.0579     0.9804   8.219 9.17e-13 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.724 on 97 degrees of freedom
## Multiple R-squared:  0.441,  Adjusted R-squared:  0.4294
## F-statistic: 38.26 on 2 and 97 DF,  p-value: 5.631e-13
```

```
summary(model5)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1387  -3.6721  -0.8869   3.9222  14.9802
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   85.992      1.157   74.34   <2e-16 ***
## x1             3.149      1.363    2.31    0.023 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.121 on 98 degrees of freedom
## Multiple R-squared:  0.05164,    Adjusted R-squared:  0.04197
## F-statistic: 5.337 on 1 and 98 DF,  p-value: 0.02298
```

```
# When b3 is 0 in the first model, first model becomes the second model
# and the estimates did not change, still biased.

# When the relation is simplified with a simple linear regression model,
# the coefficient of the treatment decreased to 3.740 and biased.

# Are the following statements right or wrong?

  # (a):

# It's FALSE because, rejecting the null hypothesis does not guarantee the quality of the model
# To decide on how good is your prediction, one needs to check the distribution
# of the residuals, if they are randomly distributed then we can say that
# model is predicting well.

  # (b):

# If the confounding covariate "Age" is stated in the hypothesis then removing it
# basically is not an option and the answer is False;
# However, if it is not there and there are other covariates with low p-values
# then age can be removed for simplicity and avoiding overfitting.
```