Homework 2

Hyunjoon Rhee

2/5/2021

Problem 1

```r
indicators <- read.table('indicators.txt', header = TRUE)
#head(indicators)
```

```r
summary(indicators)
```

```
##   MetroArea          PriceChange     LoanPaymentsOverdue
##  Length:18          Min.   :-9.700   Min.   :1.650
##  Class :character   1st Qu.:-7.000   1st Qu.:3.020
##  Mode  :character   Median :-3.950   Median :3.300
##                     Mean   :-3.428   Mean   :3.532
##                     3rd Qu.:-0.750   3rd Qu.:4.478
##                     Max.   : 6.900   Max.   :5.630
```

```r
price <- lm(PriceChange ~ LoanPaymentsOverdue, indicators)
summary(price)
```

```
##
## Call:
## lm(formula = PriceChange ~ LoanPaymentsOverdue, data = indicators)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6541 -3.3419 -0.6944  2.5288  6.9163
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.5145     3.3240   1.358   0.1933
## LoanPaymentsOverdue  -2.2485     0.9033  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.954 on 16 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2341
## F-statistic: 6.196 on 1 and 16 DF,  p-value: 0.02419
```

###Problem 1 a)

```r
summary(price)$r.squared
```

```
## [1] 0.2791527
```

```r
summary(price)$adj.r.squared
```

```
## [1] 0.2340997
```

R squared value shows that there is 27.9% of the data that fits the regression model. The adjusted R squared shows a smaller value because it takes consideration of the predictors into the value, having 23.4% of the data fitting to the regression model.

###Problem 1 b)

```r
slope <- summary(price)$coef
slope[2,1]
```

```
## [1] -2.24852
```

```r
confint(price, 'LoanPaymentsOverdue', level=0.95)
```

```
##                         2.5 %     97.5 %
## LoanPaymentsOverdue -4.163454 -0.3335853
```

There is an evidence of a significant negative linear association, with 95% confidence interval showing negative numbers.

###Problem 1 c)

```r
(predict(price, data.frame(LoanPaymentsOverdue = 4), interval="confidence"))
```

```
##         fit       lwr       upr
## 1 -4.479585 -6.648849 -2.310322
```

0% is not feasible because 0 is not in the 95% confidence interval.

##Problem 2

```r
library(faraway)
data(sat)
head(sat)
```

```
##            expend ratio salary takers verbal math total
## Alabama     4.405  17.2 31.144      8    491  538  1029
## Alaska      8.963  17.6 47.951     47    445  489   934
## Arizona     4.778  19.3 32.175     27    448  496   944
## Arkansas    4.459  17.1 28.934      6    482  523  1005
## California  4.992  24.0 41.078     45    417  485   902
## Colorado    5.443  18.4 34.571     29    462  518   980
```

###Problem 2 a)

```r
score <- lm(total ~ expend+ratio+salary, data = sat)
summary(score)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

###Problem 2 b)

Suppose that $\alpha = 0.05$. For the hypothesis of $H_0 : \beta_{salary} = 0$, because the p value is 0.0667, $0.0667 > 0.05$. This shows that we fail to reject the null hypothesis, meaning we can consider $\beta_{salary} = 0$.

###Problem 2 c) $H_0 : \beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. Suppose that $\alpha = 0.05$. The following hypothesis could be tested with F statistic, which shows a p value of 0.01209. Because $0.01209 < 0.05$, we can reject the null hypothesis and say that there are predictors that have effect on the response.

###Problem 2 d)

```r
score_d <- lm(total ~ expend+ratio+salary+takers, data = sat)
summary(score_d)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

Suppose that $\alpha = 0.05$. For the hypothesis of $H_0 : \beta_{takers} = 0$, because the p value is $2.61e^{-16}$, $2.61e^{-16} < 0.05$. This shows that we reject the null hypothesis. Moreover, considering the F test, because the p value for the F test is $2.2e^{-16}} meaning we can consider that the model with takers have predictors that affects the response. Also the p value is smaller than the previous model, which means that this fits better.

##Problem 3

###Problem 3 a)

```r
data(prostate)
prob3 <- lm(lpsa ~ ., data = prostate)
summary(prob3)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```r
confint(prob3, c("age"), .95)
```

```
##           2.5 %      97.5 %
## age -0.04184062 0.002566267
```

```r
confint(prob3, c("age"), .90)
```

```
##             5 %         95 %
## age -0.0382102 -0.001064151
```

The confidence interval for age shows that when the interval is 90%, it does not include 0, which means that age is significant, whereas when the interval is 95%, it does include 0 meaning that it is may not be significant. For the hypothesis testing for $H_0 : \beta_{age} = 0$, when $\alpha = 0.05$, the p value, which is 0.08229 is greater than α, meaning that it fails to reject the null hypothesis.

###Problem 2 b)

```
prob3_2 <- update(prob3, . ~ lcavol + lweight + svi)
anova(prob3, prob3_2)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
## Model 2: lpsa ~ lcavol + lweight + svi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     88 44.163
## 2     93 47.785 -5   -3.6218 1.4434 0.2167
```

THe null hypothesis $H_0 : \beta_{age} = 0$ Because the p value of the anova analysis is 0.2167, which is greater than 0.05, this tells that reduced model is not significantly better than the original model. Therefore, the original model with all the predictors is preferred.

###Problem 2 c)

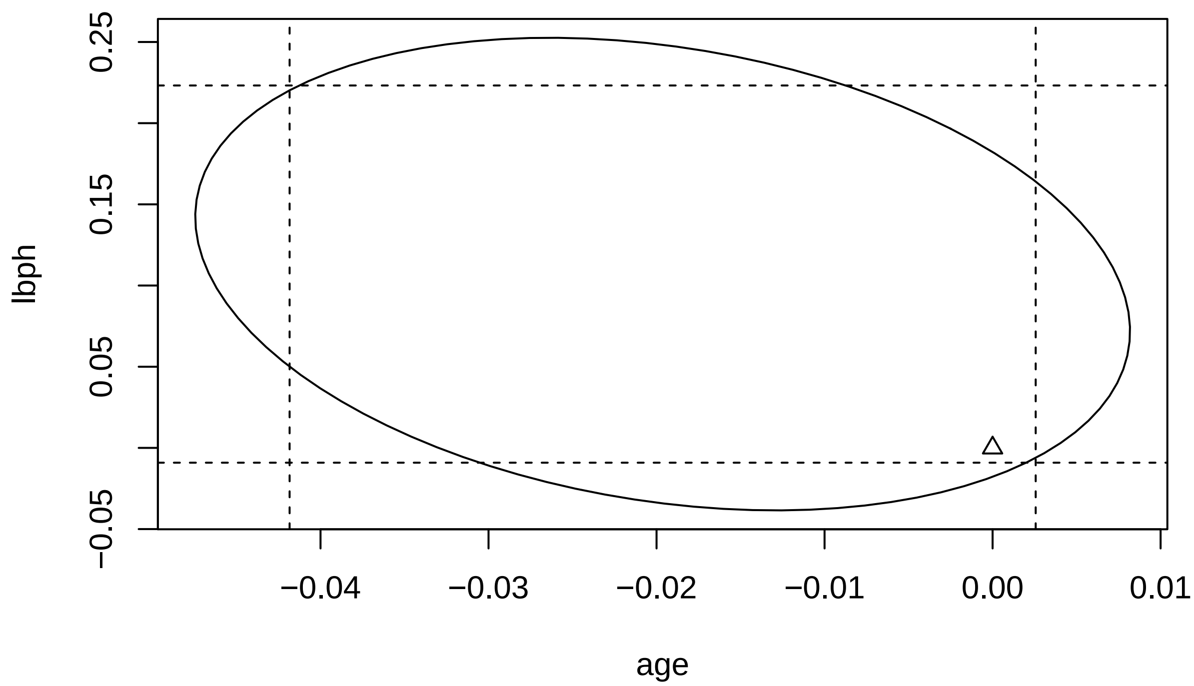
```
library(ellipse)
```

```
## Warning: package 'ellipse' was built under R version 4.0.2
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

```
plot(ellipse(prob3, c('age', 'lbph')), type = "l")
points(0, 0, pch = 2)
abline(v= confint(prob3)['age',], lty = 2)
abline(h= confint(prob3)['lbph',], lty = 2)
```



The joint null hypothesis is $H_0 : \beta_{age} = \beta_{lbph} = 0$. It fails to reject the null hypothesis because the origin is inside the confidence region. With the same reason, when the null hypothesis are $H_0 : \beta_{age} = 0$ or $H_0 : \beta_{lbph} = 0$, they both fail to reject because 0 is inside the 95% confidence region.

###Problem 3 d)

```
n.iter = 5000;
tt = numeric(n.iter);
for(i in 1:n.iter){
    newprostate=prostate;
    newprostate[,c(3)]=prostate[sample(97),c(3)];
    ge = lm(lpsa ~., data=newprostate);
    tt[i] = summary(ge)$coef[4,3]
}
#Estimated p-value
length(abs(tt[abs(tt) > abs(summary(prob3)$coef[4,3])]))/n.iter
```

```
## [1] 0.0892
```

The p value for age is 0.08229. The permutation of the t value shows that the estimated p value is getting closer to the value of 0.08229 as the number of iteration increases.

Problem 4

###Problem 4 a)

```r
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------ tid
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
data(fat)
View(fat)
fat <-  select(fat, age, weight, height, neck, chest,
                abdom, hip, thigh, knee, ankle,
                biceps, forearm, wrist, brozek)
modelf <-  lm(brozek ~ ., data = fat)
```

```r
modelf1 <- lm(brozek ~ chest + abdom, data = fat)
summary(modelf1)
```

```
##
## Call:
## lm(formula = brozek ~ chest + abdom, data = fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6863  -3.3891   0.2494   3.0501  11.6890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.85796    3.75494  -7.153 9.46e-12 ***
## chest        -0.24133    0.08294  -2.910  0.00394 **
## abdom         0.75769    0.06484  11.685  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.448 on 249 degrees of freedom
## Multiple R-squared:  0.6732, Adjusted R-squared:  0.6706
## F-statistic: 256.5 on 2 and 249 DF,  p-value: < 2.2e-16
```

For the null hypothesis, F test should be looked. The p value for the f test is almost equal to zero. This means that we reject the null hypothesis that $H_0 : \beta_{chest} = \beta_{abdom}$.

###Problem 4 b)

```r
modelf2 <- lm(brozek ~ age + weight + height + abdom, data = fat)
anova(modelf, modelf2)
```

```
## Analysis of Variance Table
##
## Model 1: brozek ~ age + weight + height + neck + chest + abdom + hip +
##     thigh + knee + ankle + biceps + forearm + wrist
## Model 2: brozek ~ age + weight + height + abdom
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    238 3785.1
## 2    247 4205.0 -9    -419.9 2.9336 0.002558 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test shows that the p value is 0.002558. This shows that the reduced model is not significantly better than the original model.

###Problem 4 c)

```r
medianvalue=apply(fat,2,median)
x=data.frame(t(medianvalue))

predict.lm(modelf, newdata = x, interval = "prediction")
```

```
##        fit     lwr      upr
## 1 17.49322 9.61783 25.36861
```

```r
predict.lm(modelf2, newdata = x, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 17.84028 9.696631 25.98392
```

The intervals are almost the same, so they do not differ an important amount.

###Problem 4 d)

```r
fatd <- fat[25:50,c(1,2,3,6)]
predict(modelf2, new=data.frame(fatd), interval="prediction")
```

```
##          fit         lwr      upr
## 25  8.298418  0.08103281 16.51580
## 26  9.903086  1.71322935 18.09294
## 27  9.216292  1.01552526 17.41706
## 28 19.740864 11.48411519 27.99761
## 29  8.747253  0.49903716 16.99547
## 30 13.376012  5.19875861 21.55326
## 31 14.797935  6.62619278 22.96968
## 32 14.065015  5.88455369 22.24548
## 33  8.315872  0.10061987 16.53112
## 34 21.038046 12.84083747 29.23525
## 35 30.623842 22.38906643 38.85862
## 36 36.201628 27.83841989 44.56484
## 37 23.528151 15.36925601 31.68705
## 38 22.473944 14.31151961 30.63637
## 39 45.310482 36.47199166 54.14897
```

```
## 40 30.120799 21.92013711 38.32146
## 41 38.663109 30.34468421 46.98153
## 42 30.910811 20.40248479 41.41914
## 43 30.810797 22.61170204 39.00989
## 44 25.164766 16.99699489 33.33254
## 45 11.141535  2.93651578 19.34655
## 46 10.568910  2.38586739 18.75195
## 47  8.125807 -0.07361926 16.32523
## 48 10.999713  2.82955826 19.16987
## 49 16.325612  8.12181572 24.52941
## 50  5.970276 -2.26886915 14.20942
```

47 and 50 includes 0 in their intervals with even a possibility of having negative fat. These two data points could be considered as anomalous.

###Problem 4 e)

```
fate <- fat[c(25:46, 48, 49),]
predict(modelf2, new=data.frame(fate), interval="prediction")
```

```
##          fit         lwr      upr
## 25  8.298418  0.08103281 16.51580
## 26  9.903086  1.71322935 18.09294
## 27  9.216292  1.01552526 17.41706
## 28 19.740864 11.48411519 27.99761
## 29  8.747253  0.49903716 16.99547
## 30 13.376012  5.19875861 21.55326
## 31 14.797935  6.62619278 22.96968
## 32 14.065015  5.88455369 22.24548
## 33  8.315872  0.10061987 16.53112
## 34 21.038046 12.84083747 29.23525
## 35 30.623842 22.38906643 38.85862
## 36 36.201628 27.83841989 44.56484
## 37 23.528151 15.36925601 31.68705
## 38 22.473944 14.31151961 30.63637
## 39 45.310482 36.47199166 54.14897
## 40 30.120799 21.92013711 38.32146
## 41 38.663109 30.34468421 46.98153
## 42 30.910811 20.40248479 41.41914
## 43 30.810797 22.61170204 39.00989
## 44 25.164766 16.99699489 33.33254
## 45 11.141535  2.93651578 19.34655
## 46 10.568910  2.38586739 18.75195
## 48 10.999713  2.82955826 19.16987
## 49 16.325612  8.12181572 24.52941
```

```
medianvalue2=apply(fate,2,median)
y=data.frame(t(medianvalue2))

predict.lm(modelf, newdata = y, interval = "prediction")
```

```
##        fit     lwr      upr
## 1 15.54726 7.65764 23.43688
```

```r
predict.lm(modelf2, newdata = y, interval = "prediction")
```

```
##    fit      lwr      upr
## 1 16.1 7.949797 24.25021
```

It changed the lower and upper bound of the interval. But the difference is not huge.

problem 5

$$\mathrm{var}(\hat{Y} \mid X) = \mathrm{var}(HY \mid X)$$

$$\mathrm{var}(HY \mid X) = H\,\mathrm{var}(Y \mid X)\,H'$$

$$= \sigma^2 I \cdot H \cdot H'$$

$$= \sigma^2 X(X'X)^{-1} X'X(X'X)^{-1} X'$$

$$= \sigma^2 X(X'X)^{-1} X'$$

$$= \sigma^2 H$$

Figure 1: Problem5