Assignment 2

Ola Rasmussen

Problem 1

a.

1.
    - Estimate: $\hat{\beta} = (X^T X)^{-1} X^T Y$
    - Std. Error: $\frac{\sigma}{\sqrt{n}}$
    - T-value: $T = \frac{Estimate}{Std.Error}$
    - Pr(>|t|): $P - value(t) = P(T \geq t)$
    - $Y = X\beta + \epsilon$ is a quantitative measurement of disease progression one year after baseline.
    - n is the number of observations and σ is standard deviation.
2. Estimate of the intercept we interpret is at $\hat{\beta}_0$. We get this when all the other covariates are zero.

3. When the bmi covariate increases by 1, the Y-value increases by 5.59548.

4. Residual standard error: 54.16 on 431 degrees of freedom.

    The formula is: $\frac{1}{n}\sum_{i=1}^{\infty}\left(Y_i - \hat{Y}_i\right)^2$

5. With a significant level at 0.05 we would consider sex, bmi, map and ltg so be significant.

    $H_0 : p - value_{covariate} \leq \alpha$

    versus

    $H_1 : p - value_{covariate} > \alpha$

    For the p-value to be valid, we need to assume that the null-hypothesis is correct.

b. I would say that the fit of the full model is OK. The adjusted R-squared value is about 0.5, and since humans are hard to predict, that is a good adjusted R-squared value.

Less than half of the null-hypotheses are rejected, so i would not say that the model is significant at level $\alpha = 0.05$.

$H_0 : \beta_{age} = \beta_{sex} = \beta_{bmi} = \beta_{map} = \beta_{tc} = \beta_{ldl} = \beta_{hdl} = \beta_{tch} = \beta_{ltg} = \beta_{glu} = 0$

versus

H_1: at least one $\neq 0$

Multiple R-squared is a measurement for the fit of the model. The value 0.5176 means that the full model explains roughly 50% of the data.

c. A reduced model can have a better performance than a full model because it might reduce overfitting when we want to predict the future.

In the best subset model selection, all the possible combinations of the independent variables are considered. They are tested by some criterion.

Some of these criterion are adjusted R-squared and Bayesian Information Criterion (BIC).

Adjusted R-squared is used because it includes a correction term for the number of parameters. The bigger the better.

BIC introduces a penalty term for the number of parameters. The lower the better.

Based on the results of the adjusted R-squared and the BIC criteria, I choose model 6. It has the second lowest BIC value while also having the second highest adjusted R-squared value. The other contenders are model 5 and 7. Model 5 has a much lower adjusted R-squared value and model 7 has a much higher BIC value.

```
ds <-
read.csv("https://web.stanford.edu/~hastie/CASI_files/DATA/diabetes.csv
", sep = ",")
model6 <- lm(prog ~ sex + bmi + map + tc + ldl + ltg, data = ds)
summary(model6)

##
## Call:
## lm(formula = prog ~ sex + bmi + map + tc + ldl + ltg, data = ds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -158.277  -39.476   -2.068   37.221  148.693
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -335.3586    25.3234 -13.243  < 2e-16 ***
## sex          -21.5914     5.7056  -3.784 0.000176 ***
## bmi            5.7110     0.7073   8.075 6.69e-15 ***
## map            1.1266     0.2158   5.219 2.79e-07 ***
## tc            -1.0429     0.2208  -4.724 3.12e-06 ***
## ldl            0.8433     0.2298   3.670 0.000272 ***
## ltg          168.7953    16.8279  10.031  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 54.06 on 435 degrees of freedom
## Multiple R-squared:  0.5149,  Adjusted R-squared:  0.5082
## F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16
```

Comparing Model 6 with the full model we observe that the adjusted R-squared value has increased, implying that the model fit more to the data. We also observe that all of the null-hypotheses are rejected, which means that the model is significant.

d. Test:

$H_0: \beta_{age} = \beta_{tc} = \beta_{ldl} = \beta_{tch} = \beta_{glu} = 0$

versus

H_1: at least one $\neq 0$

```
ds <-
read.csv("https://web.stanford.edu/~hastie/CASI_files/DATA/diabetes.csv
", sep = ",")
reduced <- lm(prog ~ age + tc + ldl + tch + glu, data = ds)
summary(reduced)

##
## Call:
## lm(formula = prog ~ age + tc + ldl + tch + glu, data = ds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156.22  -51.39   -7.70   47.15  191.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -94.1711    28.0725  -3.355 0.000864 ***
## age           0.3166     0.2547   1.243 0.214455
## tc            0.7294     0.2128   3.428 0.000666 ***
## ldl          -1.3202     0.2669  -4.946 1.08e-06 ***
## tch          29.9553     3.4764   8.617  < 2e-16 ***
## glu           1.3528     0.3140   4.308 2.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.7 on 436 degrees of freedom
## Multiple R-squared:  0.2819,  Adjusted R-squared:  0.2737
## F-statistic: 34.23 on 5 and 436 DF,  p-value: < 2.2e-16
```

We can see that the adjusted R-squared value has drastically gone down, so i would prefer the full model of this reduces model.

Problem 2

a.

```
pvalues <-
scan("https://www.math.ntnu.no/emner/TMA4267/2018v/pvalues.txt")
sum(pvalues < 0.05)

## [1] 155
```

We reject 155 null-hypotheses.

A type 1 error occurs when we reject a null-hypothesis when said null-hypothesis is true.

We do not know the number of false positive findings in out data, but we can assume it given $\alpha = 0.05$

b. The familywise error rate (FWER) is defined as the probability of one or more false positive findings.

Controlling the FWER at level 0.05 means that we set an upper limit to the FWER. That is we set a cut-off on the p-value.

Given $\alpha = 0.05$ and $m = 1000$, we want the cut-off on p-values to be $\alpha_{loc} = \frac{\alpha}{m} = \frac{0.05}{1000} = 0.00005$ for out data using the Bonferroni method.

```
pvalues <-
scan("https://www.math.ntnu.no/emner/TMA4267/2018v/pvalues.txt")
sum(pvalues < 0.00005)

## [1] 50
```

Using the Bonferroni method with $\alpha_{loc} = 0.00005$, we reject 50 null-hypotheses.

c. Assuming that the first 900 null-hypotheses are true ant the last 100 are false, it implies that 10% of the number of type 1 and type 2 errors are false.