Compulsory exercise 2 TMA4267

TMA4267 V2022

August Arnstad

21 mars, 2022

1)

a)

Define the design matrix X as

$$X = \begin{bmatrix} 1 & X_{1,1} & \dots & X_{1,k} \\ 1 & X_{2,1} & \dots & X_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,k} \end{bmatrix}$$

and the response vector Y as

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

In our case, Y will be tha values from the "prog" column, β are the regression coefficients and X is the matrix with the other data columns as its columns.

i)

Estimate - This column contains the estimators $\hat{\beta} = (X^T X)^{-1} X^T Y$ where X is the design matrix containing the data, and Y is the response from the prog data.

Now define

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = Y - X\beta, \quad \hat{\epsilon} = \begin{bmatrix} \hat{\epsilon_1} \\ \hat{\epsilon_2} \\ \vdots \\ \hat{\epsilon_n} \end{bmatrix} = Y - X\hat{\beta}$$

Standard error - The standard error is defined as $SE(\hat{\beta}) = \sqrt{\hat{\sigma}^2 diag(X^T X)^{-1}}$ where
$\hat{\sigma}^2 = \frac{1}{n-p}(Y - X\hat{\beta})^T (Y - X\hat{\beta})$, $p = k + 1$ and k is the number of predictors.

t-value - The t-value is calculated by $T = \frac{\hat{\beta}}{SE(\hat{\beta})}$

$Pr\{> |t|\}$ - This is calculated as the probability that a new measurement $T = \frac{\hat{\beta}}{SE(\hat{\beta})}$ is greater than the absolute value of the t-value, given that T has a student-t distribution and that H_0 is true. This can be calculated as $Pr\{|T| > |t|\} = 1 - Pr\{|T| < |t|\}$ where the last probability can be found from the CDF of the student-t distribution.

ii) The estimate for the intercept can be thought of as the response, if all covariates are 0.

iii) We see that the coefficient is positive, this implies that an increase in BMI causes an increase in the response, i.e. the progression of diabetes in the last year. We interpret this to mean that a person with a high BMI will have a progressed more towards diabetes in a year than a person with a low BMI.

In other words, the coefficient is the slope of the response if all other values are fixed.

iv) We can look at the "Residual standard error" and this is $54.16 = \hat{\sigma}$. We can square this and see that the numerical value of the variance of the estimated error is $(54.16)^2 = \hat{\sigma}^2 = 2933.3$.

v)To be significant at level 0.05 we must have that the p-value of the coefficient is smaller than 0.05. Hence the variables "sex", "bmi", "map" and "ltg" are found to be significant at level 0.05.

Our hypothesis test is as follows:

$H_0 : \beta_j = 0$
$H_1 : \beta_j \neq 0 \ j = age, sex, ..., glu$

The assumptions that must be made are that all ϵ_i are independent and identically distributed, s.t. $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and homoschedastic.

b)

To check the fit, we see if the assumptions of a linear regression is fulfilled. The following assumptions are made:

1) $E[\epsilon_j] = 0 \quad \forall j$

2) $Var(\epsilon_j) = \sigma^2 \quad \forall j$

3) ϵ are iid normally distributed

From the top of figure 2, we see that there seems to be some covariates that have a linear correlation with the response. This is an indicator that linear regression might be used, but some of the covariates may be redundant.

The residual vs fitted plot shows no signs of correlation, which indicates that assumptions 1) and 2) holds.

The QQ plot tells us that the errors are normally distributed due to the straight line, and shows no correlation. This backs up assumption 3.

Further, four of the p-values are significantly low and we can conclude that these values affect the response. Hence it seems to be a correlation between covariates and response and regression is a good model for our data.

Conclusively the full model fits relatively good and the regression is significant, but we should look to reduce the model by removing some of the covariates that seem to be redundant.

$H_0 : \beta_j = 0 \quad \forall j$
$H_1 : \beta_j \neq 0$ for some $j = 1, ..., n$

Multiple R^2 is the proportion of the variance in the response variable that can be explained by the predictor value and is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

where $\hat{Y} = (\hat{y}_1, ..., \hat{y}_n)^T = X\hat{\beta}$

c)

A reduced model can perform better when the aim is prediction because a full model might overfit the data. That is, an overfit model fits the training data well but takes into account to much of the random noise. This leads to a greater variance which makes predictions worse.

The best subset model selection is used to target the predictors which give the best model. It fits a least square regression to every possible combination of predictors p and look at the resulting models. To choose the best model of these subset models, we use the adjusted R^2 and the BIC criteria. The adjusted R^2 and BIC criteria add a penalty to the model error that increases with the number of predictions used.

In figure 3 we fit models with one, two, three etc. parameters and find the best model with one, two three etc parameters. We then evaluate their performance based on the BIC and adjusted R^2 values. This is done by plotting the BIC value versus the number of predictors used in the model. By definition of the BIC criteria the goal is for this to be as small as possible. We see that the minimum BIC value corresponds to a model that uses 5 predictors. We can also see that the smallest value of the BIC uses the predictors "sex", "bmi", "map", "hdl" and "ltg"

After checking the BIC criteria we move on to the adjusted R^2. This value should be as high as possible, with values ranging from 0 to 1. The same procedure but with adjusted R^2 as the criteria finds that the optimal number(highest adjusted R^2) of predictors is 8. From the final plot we can see that these are "sex", "bmi", "map", "tc", "ldl", "tch", "ltg", "glu".

For the reduced model we should pick a model with a high adjusted R^2 and a low BIC. We see that the "hdl" predictor reduces our adjusted R^2 so we do not include this. We also see that the BIC model works nicely with 6 of the 8 paramters that give the highest adjusted R^2. From this we define a reduced model by

```
reduced_fit<-lm(prog~sex+bmi+map+tc+ldl+ltg, data=ds)

summary(reduced_fit)
```

```
##
## 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
```

i.e.

$$Y = -335.3586 - 21.5914x_{sex} + 5.7110x_{bmi} + 1.1266x_{map} - 1.0429x_{tc} + 0.8433x_{ldl} + 168.7953x_{ltg}$$

We observe that the residual standard error of the reduced model decreases, which is a good sign. This indicates that the variance of our reduced model is lower than in the full model. Further we notice that the adjusted R^2 is higher, meaning that we have successfully removed parameters that do not improve the model fit.

The regression parameters also differ slightly. Some, e.g. bmi and map, increase, while sex increases. This might be due to some correlation between the covariates and that the reduced model weigh these differently.

Generally we see that the p-values for the parameters in the reduced model is lower, which also indicates improvement. In fact, they all satisfy a significance level of $\alpha = 0.05$, which is the common standard for a significant correlation. This means that our model now consists of parameters that are highly correlated with the response.

Conclusively, the reduced model is a better fit.

d)

To perform the hypothesis test we can use the built in anova function. If the p-value tells us to reject H_0, this means that we should not reduce our model, as at least one of the paramters of the rejected covariates is not zero.

```
full_model<-lm(prog~., data=ds)
reduced_model<-lm(prog~sex+bmi+map+hdl+ltg, data=ds)

anova(full_model, reduced_model)
```

```
## Analysis of Variance Table
##
## Model 1: prog ~ X + age + sex + bmi + map + tc + ldl + hdl + tch + ltg +
##     glu
## Model 2: prog ~ sex + bmi + map + hdl + ltg
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    430 1263937
## 2    436 1287879 -6    -23942 1.3575 0.2304
```

From the print out, we see that we should not reject H_0 as the p-value, 0.2304, is greater than the standard 0.05. Hence the reduced model is preferred.

2)

a)

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

```
## [1] 155
```

From this code we find that there are 155 rejections of H_0.

A false positive (type I error) is when the null hypothesis is rejected, when it is in fact true.

We do not know the number of false positive findings in our data. However, had we know α, the significance level, we could say that in each test there would be an α chance of getting a type I error.

b)

The FWER is the probability of producing at least one false positive from the multiple hypothesis test.

If we reject H_0 for all adjusted p-values below $\alpha = 0.05$ the overall type I errors will be controlled at level $\alpha = 0.05$. In other words, the $FWER$ is controlled at level 0.05 if $FWER \leq 0.05$

The Bonferroni method gives us the cutoff p-value by the formula

$$\alpha = m\alpha_{loc}$$

where m is the number of tests and α_{loc} is the cutoff. In our case $m = 1000$, so $\alpha_{loc} = 5 * 10^{-5}$

```
pvalues <-
scan("https://www.math.ntnu.no/emner/TMA4267/2018v/pvalues.txt")
index<-which(pvalues<5*10**{-5})
length(index)
```

```
## [1] 50
```

With the Bonferroni cutoff, we get 50 rejections of H_0.

c)

Splitting the data as the test described yields

```
h0_true=pvalues[1:900]
type_1_error=length(which(h0_true<0.05))


h0_false=pvalues[901:1000]
type_2_error=length(which(h0_false>0.05))


h0_true=pvalues[1:900]
type_1_errorB=length(which(h0_true<5*10**{-5}))


h0_false=pvalues[901:1000]
type_2_errorB=length(which(h0_false>=5*10**{-5}))

notBonferroni=c(type_1_error, type_2_error)
Bonferroni=c(type_1_errorB, type_2_errorB)

notBonferroni
```

```
## [1] 55  0
```

```
Bonferroni
```

```
## [1]  0 50
```

From this we can see that in a) and b) we commit 55 type 1 error with $\alpha = 0.05$ and 0 type II errors. With the Bonferroni, $\alpha = 5 * 10^{-5}$, we commit 0 type 1 errors but 50 type II errors.