ASSIGNMENT #2 TMA4267 - LINEAR STATISTICAL MODELS

Problem 1.

a) 1. In this full model we use `prog` as response and the other measurements as covariates X_i for $i = 1, 2, \ldots, 10$. We assume a classical linear regression model. This implies that we can express the full model as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ corresponds to `prog` and is, together with the residual $\boldsymbol{\epsilon}$, an $n \times 1$-vector. In this expression, $\boldsymbol{\beta}$ is a vector of length 11 consisting of the parameters we want to estimate for the covariates. Finally $\boldsymbol{X}$ is an $n \times 11$-matrix consisting of the measured data of the covariates and also a column of ones at the start corresponding to the offset/intercept. In addition we also have the following assumptions for the classical linear regression model:

- We want:
$$E[\boldsymbol{\epsilon} \,|\, \boldsymbol{X}] = 0$$
which is to say that the errors and the covariates are uncorrelated and that the errors have zero mean.
- Further:
$$\mathrm{Cov}(\boldsymbol{\epsilon} \,|\, \boldsymbol{X}) = \sigma^2 I$$
which implies that the errors are homoscedastic and that the errors are pairwise uncorrelated (i.e. no autocorrelation).
- We want $\boldsymbol{X}$ to have full rank. This implies that there does not exist any linear dependence between the covariates.
- We also assume (for inferencing (see 5.)) that $\boldsymbol{\epsilon} \,|\, \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I)$. That is, the conditional error is multivariate normal. This also implies that ϵ_i and X_i are independent.

Say that SSE is given by

$$SSE(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$$

Then assumption 1 and 2 guarantees by the Gauss-Markov theorem that the estimators $\hat{\boldsymbol{\beta}}$ given by minimizing this quantity (existence is given by assumption 3) are indeed the "best" ones (i.e. $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and that $\hat{\boldsymbol{\beta}}$ has the smallest mean squared error). This vector is precisely the values in the "estimate" column, and it is given by:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

In fact the lm-algorithm in R uses QR-factorization, so it would be more correct to say that if $\boldsymbol{X} = QR$, then the values in the estimate column is the solution of:

$$R\hat{\boldsymbol{\beta}} = Q^T\boldsymbol{Y}$$

The next column is the column of "standard error", se_i, for each estimator $\hat{\beta}_i$. This is simply the square root of $\mathrm{Var}(\hat{\beta}_i)$. To compute these values, say that $A = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, then $\hat{\boldsymbol{\beta}} = A\boldsymbol{Y}$ and:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(A\boldsymbol{Y}) = A\,\mathrm{Var}(\boldsymbol{Y})A^T$$

We can estimate $\sigma^2 = \text{Var}(\boldsymbol{Y})$ unbiased by:

$$\sigma^2 \approx \hat{\sigma}^2 = \frac{SSE(\hat{\boldsymbol{\beta}})}{n-11} \approx 2933.3$$

where we have used $n - 11$ degrees of freedom. Note that this is the squared value of what is called "residual standard error" in R. Then:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}(A\boldsymbol{Y}) = \hat{\sigma}^2 AA^T = \hat{\sigma}^2 \boldsymbol{X}^T \boldsymbol{X}$$

This gives a covariance-matrix with the variances $\text{Var}(\beta_i)$ of each β_i down the diagonal.
The next column "t value" is just how many standard deviations (or "standard errors") se_i each β_i is away from zero, i.e. the t-value is simply β_i/se_i. The last column gives the probability that any "more obscure" event than the one that is measured could happen.

2. The estimate for the intercept gives the offset from the origin of the model plane. I.e. if all the covariates are zero, the predicted response would be this value.
3. The estimated coefficient for bmi may be interpreted as saying that an increase in bmi by 1 point leads to an increase in the disease progression of (at least) about 5.5 points.
4. See 1.
5. Intercept, `sex`, `bmi`, `map` and `ltg` are all significant at level 0.05. Additionally `tc` is not far from significant at this level either.
   Consider `bmi` and say that β is the estimate of this covariate. Then we have the following hypotheses:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

Under the fourth assumption that each error term ϵ_i is $\mathcal{N}(0, \sigma^2)$ it follows that:

$$t_{\beta_i} = \frac{\hat{\beta}_i - \beta}{\text{se}_i}$$

has a Student-t distribution with $n - 11$ degrees of freedom (since we are sampling from a normally distributed data set given that the errors are indeed normally distributed). Under the null hypothesis this reduces to the value given under the column "t value" in the output. Since the p-values are computed from this statistic it follows that the values are valid under these assumptions. The interpretation of this is that in the case of the covariates: Intercept, `sex`, `bmi`, `map` and `ltg` we reject the null hypothesis that the covariates show no linear relationship with `prog` at the significance level of 5%. For the other covariates we can not infer that there is such a linear relationship.

b) For the values in $\boldsymbol{Y}$ (i.e. the measured values for `prog`) we can form the average value $\mu_{\texttt{prog}}$ and then consider the variation in the data set from this value, that is the total sum of squares:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \mu_{\texttt{prog}})^2$$

Our model gives the predicted values $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, which again gives rise to a similar measure of variation from the mean in the predicted values given by:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \mu_{\texttt{prog}})^2$$

From 3.19 [1] it follows that:

$$\sum_{i=1}^{n}(y_i - \mu_{\texttt{prog}})^2 \geq \sum_{i=1}^{n}(\hat{y}_i - \mu_{\texttt{prog}})^2$$

and therefore that the measure:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\texttt{prog}})^2}{\sum_{i=1}^{n}(y_i - \mu_{\texttt{prog}})^2} = 1 - \frac{\text{SSE}}{\text{SST}} \in [0, 1]$$

where:

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

This measure can be interpreted as a way of measuring how much the variation is decreased by the regression; If the value is close to 1, then $\hat{y}_i$ is close to y_i for every i which is indicative of a good fit. If the value is close to 0 we must have that $\hat{y}_i$ is close to $\mu_{\texttt{prog}}$ for every i which indicates a bad fit. One could also think of this as the amount of total variation that is explained by the linear regression model. One problem with this value is that it will increase for each covariate added since you are adding more "explanation" to the model. This is where the adjusted value comes in and in our case, about 50% of the variation is explained by the linear regression model. This implies that 50 % of the variation is not explained by the linear regression model. The F-statistic (explained variance over unexplained variance) shows that these values for R^2 are significant. This in turn is saying that the underlying null hypothesis:

"The intercept-only model and the full fitted model have the same fit"

can be rejected, and we conclude with the alternative hypothesis that our model has a better fit than a model without any covariates. This is equivalent with saying that the regression model is significant and there exists some linear relation between the response and the predictors.

This in itself is not necessarily a good single measure to judge the goodness of fit. From the normal Q-Q plot based on the studentized residuals we see that there are some deviation from normality in the tails. This may affect some of the variation that is not explained by the model. However, the Anderson-Darling test shows that we can not reject the hypothesis of normality of the studentizes residuals. The scatter plot over studentizes residuals vs. fitted values does also show some heteroscedasticity, with the studentizes residuals being lower for lower values of the fitted value.

c) A model with unnecessarily many covariates may overfit. It may also add unnecessary complexity to a model by including covariates that does not effect the response. Both of these are adverse effects when the aim is to predict. Therefore a reduced model may be more suitable and easier to interpret.

In the best subset model selection every possible subset of the full model is considered and the ones with the same number of variables is compared with each other and selected on the basis of the lower SSE (sum of squares of the errors (residuals)). Since we in turn only are comparing models of the same size, this is not problematic. The output of the "regsubset" function is therefore the prediction model for `prog` with the given number of variables with the lowest SSE. To compare models of different sizes there are several criteria to consider. Amongst these we have adjusted R^2 and BIC. Adjusted R^2 is a using the regular R^2, but penalizes models with more covariates as the regular R^2 can only increase when using more and more covariates. The adjusted R^2 is given by:

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}}{\text{SST}} \cdot \frac{n-1}{n-p}$$

It can be argued that this measure does not penalise larger models enough, so another criterion to use is BIC (Bayesian Information Criterion):

$$\text{BIC} = \frac{\text{SSE}/n}{\hat{\sigma}^2} + \ln n \cdot \frac{p}{n}$$

Note: This is from the lecture notes. I can not see how this is equal to the quantity defined in the book. Further I can not find out which version of BIC the function "regsubset" uses ...

In figure 4 we see plots of R^2_{adj} and BIC relative to the model size. Larger R^2_{adj} is desired and smaller BIC. From these plots we see that the model with 5 covariates has the lowest BIC while the model with 8 covariates gives the highest R^2_{adj}. We also note that the models with 6-10 covariates have roughly the same R^2_{adj}. Similarly the models with 5-6 covariates have roughly the same BIC. A good trade-off between these measured may therefore be the model with the six covariates: `sex`, `bmi`, `map`, `tc`, `ldl`, `ltg` and an intercept. By fitting this model in R we get the following output:

```
> fitted6 <- lm(formula = prog ~ sex + bmi + map + tc + ldl + ltg, data = ds)
> summary(fitted6)

Call:
lm(formula = prog ~ sex + bmi + map + tc + ldl + ltg, data = ds)

Residuals:
     Min       1Q   Median       3Q      Max
-158.818  -39.184   -2.126   37.391  148.910

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -334.9019    25.3094 -13.232  < 2e-16 ***
sex          -21.5052     5.7045  -3.770 0.000186 ***
bmi            5.7040     0.7077   8.060 7.45e-15 ***
map            1.1260     0.2159   5.216 2.83e-07 ***
tc            -1.0391     0.2206  -4.710 3.33e-06 ***
ldl            0.8395     0.2296   3.656 0.000288 ***
ltg          168.5354    16.8138  10.024  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.07 on 435 degrees of freedom
Multiple R-squared:  0.5148,    Adjusted R-squared:  0.5081
F-statistic: 76.91 on 6 and 435 DF,  p-value: < 2.2e-16
```

Another possible model choice is the model with five covariates: `sex`, `bmi`, `map`, `hdl`, `ltg` and an intercept:

```
> fitted5 <- lm(formula = prog ~ sex + bmi + map + hdl + ltg, data = ds)
> summary(fitted5)

Call:
lm(formula = prog ~ sex + bmi + map + hdl + ltg, data = ds)

Residuals:
     Min       1Q   Median       3Q      Max
-150.361  -39.616   -0.412   37.119  148.513

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -240.0051    34.3139  -6.994 1.01e-11 ***
sex          -22.4291     5.7647  -3.891 0.000116 ***
bmi            5.6386     0.7040   8.010 1.06e-14 ***
map            1.1229     0.2172   5.170 3.58e-07 ***
hdl           -1.0629     0.2418  -4.396 1.39e-05 ***
ltg           99.4974    13.7887   7.216 2.39e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.35 on 436 degrees of freedom
Multiple R-squared:  0.5086,    Adjusted R-squared:  0.5029
F-statistic: 90.24 on 5 and 436 DF,  p-value: < 2.2e-16
```

Overall we see that this model have slightly higher standard deviations and a slightly higher residual standard error than the model with six covariates. The adjusted R^2 is also a bit lower.

By comparing the model with six covariates to the full fitted model we observe lowered standard deviations for all parameters. The estimates have not changed dramatically. All model parameters are now significant at the 5% level as we would expect. The residual standard error is also lowered, but not by much. We also have slightly higher adjusted R^2 as we saw from the model choice.

d) Consider now the model with five covariates: `sex`, `bmi`, `map`, `hdl`, `ltg` and an intercept. We have the hypotheses:

$$H_0 : \beta_{\text{age}} = \beta_{\text{tc}} = \beta_{\text{ldl}} = \beta_{\text{tch}} = \beta_{\text{glu}} = 0$$
$$H_1 : \text{at least one of the } \beta\text{'s are non-zero}$$

We proceed by the method described in the book (p. 128-129) [1]. We wish to test $H_0 : \boldsymbol{C\beta} = \mathbf{0}$ against $H_1 : \boldsymbol{C\beta} \neq \mathbf{0}$, where $\boldsymbol{C}$ is a 5×11 matrix with 1's and 0's so that:

$$\boldsymbol{C\beta} = \begin{bmatrix} \beta_{\text{age}} \\ \beta_{\text{tc}} \\ \beta_{\text{ldl}} \\ \beta_{\text{tch}} \\ \beta_{\text{glu}} \end{bmatrix}$$

The test statistic we want to use is given by:

$$F = \frac{n-p}{r} \cdot \frac{\Delta\text{SSE}}{\text{SSE}} \sim \text{F}_{r,n-p}$$

where $n = 442$, $p = 11$, $r = 5$, SSE is the residual sum of squares for the full model and ΔSSE is the difference between the residual sum of squares SSE for the full model and the residual sum of squares SSE_{H_0} for the model restricted under the null hypothesis (i.e. $\boldsymbol{C\beta} = \mathbf{0}$).

By computing "t(prog - full$fitted.values)%*%(prog - full$fitted.values)" in R we get SSE = 1264264. We can compute SSE_{H_0} in a similar manner by fitting the restricted model as "fitted5" and computing "t(prog - fitted5$fitted.values)%*%(prog - fitted5$fitted.values)". We then get $\text{SSE}_{H_0} = 1288082$. This gives:

$$F = \frac{n-p}{r} \cdot \frac{\Delta\text{SSE}}{\text{SSE}} = \frac{431}{5} \cdot \frac{1288082 - 1264264}{1264264} \approx 1.624$$

Since F is distributed with $\text{F}_{5,431}$ we have the p-value:

$$P(F > 1.624) \approx 0.1388$$

which is significantly larger than any sensible cut-off. The conclusion is that we do not reject H_0. In the case of the reduced model with six covariates from the previous exercise we get:

$$F \approx 0.7622$$

which gives the p-value:

$$P(F > 0.7622) \approx 0.578$$

which is even more significant than the model with five covariates. Either way there is strong evidence to use a reduced model.

Problem 2.

a) By storing the data as "pvalues" in R we can see how many entries that are below 0.05 by using the command "length(pvalues[pvalues ¡ 0.05])". This returns 155 which is to say that 155 of the p-values that we have read are below 0.05. This is equivalent to saying that 155 of the observations are very unlikely to happen if it is the case that the null hypothesis is true. By contradiction we then conclude that the null hypothesis must be false.

A false positive finding (type 1-error) is when we reject the null hypothesis when it is in fact true. Since the significance level $\alpha \in (0, 1)$ of a test is defined as the upper threshold of how many type 1-errors we accept on average, i.e.:

$$P(\text{reject } H_0 \,|\, H_0 \text{ true}) \leq \alpha$$

we will on average expect $\alpha \cdot 100\%$ type 1-errors. In this case, therefore, we would expect at most 50 type 1-errors. However, there is no way of knowing this number exact.

b) The Familywise Error Rate (FWER) is defined as the probability of obtaining at least one type 1-error. That is, if the number of type 1-errors is V, then:

$$\text{FWER} = P(V > 0)$$

To control FWER at level 0.05 means that we want $FWER \leq 0.05$. This is to say that the probability of at least one type 1-error is less than 5%.

Say now that we want to find a new cut-off α_{loc} for each of the tests. If R_j is the event that the j-th null hypothesis is rejected, then $P(R_j) = \alpha_{\text{loc}}$ and:

$$FWER = P(R_1 \cup R_2 \cup \cdots \cup R_{1000}) \leq \sum_{j=1}^{1000} P(R_j) = 1000 \cdot \alpha_{\text{loc}} \leq 0.05$$

which gives:

$$\alpha_{\text{loc}} \leq \frac{0.05}{1000} = 5 \cdot 10^{-5}$$

With this new level we will reject 50 null hypotheses and we can be fairly confident (95%) that none of these rejections produces type 1-errors.

c) We assume that the first 900 null hypotheses are true and the last 100 are false. Consider first $\alpha_{\text{loc}} = 0.05$. We then get the following table:

		H_0	
		True	False
Decision	Accept	845	0
	Reject	55	100

In the case where $\alpha_{\text{loc}} = 5 \cdot 10^{-5}$ we get this table:

		H_0	
		True	False
Decision	Accept	900	50
	Reject	0	50

From these tables we see that the number of total errors are 55 in the case where $\alpha_{\text{loc}} = 0.05$, whereas the total number of errors are 50 in the case where $\alpha_{\text{loc}} = 0.00005$. The distribution of the types of errors are, as expected, different. In the second case all the errors are of type 2 and in the first case all of the errors are of type 1.

References

[1] L. Fahrmeir, T. Kneib, S. Lang, B. Marx *Regression: Models, Methods and applications*, Springer-Verlag Berlin Heidelberg, 2013.