Compulsory Exercise 2 in TMA4267 Statistical Linear Models, Spring 2021

Sander Ruud, Alexander J Ohrt

14 mars, 2021

Problem 1 Diabetes Progression

a)

1. Each column and the formula which it is based upon is given in the dotted list below.

- `Estimate`: $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{Y}$, where $\boldsymbol{Y}$ contains all the response variables, i.e. `prog` in this case. Moreover, the components β_j in $\boldsymbol{\beta}$ are estimates of the coefficients in the assumed linear relationship between the covariates and the response (`prog`): $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$.
- `Std. Error`: This column is calculated from the formula $\sqrt{\frac{\text{SSE}}{n-p}(X^TX)^{-1}_{ii}}$, where $(X^TX)^{-1}_{ii}$ are the diagonal elements of the matrix $(X^TX)^{-1}$. Moreover, SSE $= \sum \hat{\varepsilon}_i^2 = \boldsymbol{Y}^T(I-H)\boldsymbol{Y}$, where $H = X(X^TX)^{-1}X^T$ is called the Hat matrix and $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^T$. Also, n is the amount of observations in the data and p is the amount of covariates in the linear model. Since $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$, it is apparent that the column `Std. Error` is an estimate of the standard deviation of each component β_i, $i = 1, \ldots, p-1$, of $\boldsymbol{\beta}$. This is true because $\hat{\sigma}^2 = \frac{\text{SSE}}{n-p}$ is an unbiased estimator of σ^2 and, hence, $\hat{\sigma}^2(X^TX)^{-1}_{ii}$ is an unbiased estimator of $\text{Cov}(\hat{\boldsymbol{\beta}})$. Moreover, the covariance matrix has the variances on its diagonal, which is the square of what is estimated by `Std. Error`.
- `t value`: t value calculated in a t-test for each component of $\boldsymbol{\beta}$, given by $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$. This is used to test whether the null hypothesis seems reasonable or if there is enough evidence to discard it. The test statistic used in each test is $T = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{ii}}} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{ii}}} \sim t_{n-p}$, given H_0. The value of each of the test statistics are the values in the column `t value` from `summary(full)`.
- `Pr(>|t|)`: p value calculated from the t value in the previous column. It is calculated from the formula $2P\left(T \geq \frac{|\hat{\beta}_i|}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{ii}}}\right)$. This can be used to make a conclusion about the significance of each of the components β_j in $\boldsymbol{\beta}$, i.e. to reason about whether or not each of the estimators should be used in the model for predictive or inferential power. If the p-value is less than or equal to the chosen significance level, H_0 will be discarded.

2. We interpret the estimate for the intercept in the following way: This value is estimated to be attained when all the covariates are zero, i.e. when each patient is a female and the rest of the continuous predictors are zero-valued.

3. The estimated regression coefficient for `bmi` can be interpreted as being the estimated increase in the response, `prog`, when the value of `bmi` is increased by one unit.

4. The estimated error variance can be found in the field that says `Residual standard error` $= 54.16$. The formula for calculating this value is $\hat{\sigma}^2 = \frac{\text{SSE}}{n-p}$.

5. `tc` is found to be significant at level 0.05, and `sex`, `bmi`, `map` and `ltg` are significant at lower levels than 0.05. E.g. the null- and alternative hypothesis associated with a hypothesis test on `tc` are $H_0 : \beta_{tc} = 0$ vs. $H_1 : \beta_{tc} \neq 0$. The assumptions needed for the p-value to be valid are that the estimators for β and σ^2 are independent. Also, a valid p-value W fulfills $P(\text{discard } H_0) = P(W \leq \alpha) \leq \alpha$, given that the

null hypothesis is true. Here, α is the chosen significance level, i.e. the greatest probability of type I-errors we are willing to accept.

b)

Based on Figures 1 and 2 we would evaluate the fit of the full model as not very well aligned with the assumptions of the linear model. Immediately, when looking at the `pairs`-plot, there is no clear linear relationship between any of the predictors and `prog`. Also, it looks like the residuals increase slightly towards the middle of the x-axis of the residual versus fitted values-plot. This means that the homoscedasticity of the residuals is questionable. Moreover, the normal Q-Q plot shows some deviation between the theoretical quantiles and the sample quantiles at the endpoints, but it is hard to say whether or not they clearly break the assumption of normally distributed errors. Finally, the output of the Anderson-Darling normality test shows that the p-value is quite large, which means that the normality test does not fail. Hence, we cannot state that the data does not fit the normal distribution, but we cannot state that it fits either. No significant departure from normality was found in the data.

which signals that the errors are not normally distributed.Enig? Spesielt usikker på denne siste setningen. Ja, tolker det som at de ikke er normally distributed? Er ikke testen at han antar at det ikke følger en fordeling, og en lav p-verdi viser at det er usannsynlig å se det vi ser gitt at dataet ikke følger en normalfordeling. Tror det er motsatt: Dersom H0 forkastes så kan man være relativt sikker på at de ikke er normalfordelt, men dersom H0 ikke forkastes vet man ikke om de er det, men man kan ikke utelukke det heller. (leste på lenken nedenfor) https://variation.com/wp-content/distribution_analyzer_help/hs140.htm

As a second note, the p-value of the F-statistic is low, which means that there is enough evidence to discard the null hypothesis, i.e. the regression is significant. The null- and alternative hypotheses for this test are $H_0 : \beta_j = 0 \ \forall j \in \{1, ..., k\}$ vs. H_1 : At least one $\beta_j \neq 0$. This means that the model seems to have some merit - we cannot confidently say that the regression is insignificant.

Tenker du at modellen er god eller ikke da? Virker som at det finnes argumenter både for og imot her :(Også var det dette med gyldigheten til en p-verdi lenger oppe som jeg lurte på om vi kunne diskutere.

The `Multiple R-squared` is the ratio of total variance explained by the model. It is calculated as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\boldsymbol{Y}^T(I - H)\boldsymbol{Y}}{\boldsymbol{Y}^T(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Y}}.$$

The value of the multiple R-squared is 0.5176, which means that the model only explains approximately half of the total variance in the data.

c)

A reduced model might have better performance than a full model when the aim is prediction because the variance of the estimated covariates is lower for a reduced model. This means that we can make more accurate predictions with a smaller model compared to the larger model.

In best subset selection, all subsets of the k covariates are used to fit a linear model. In the end, the model that is considered the best among all these models is the one we choose. For each number j of covariates from 1 to k, we fit all models with j covariates and store the best model among all these, based on SSE or R^2. In the end, we choose the best among the zero-model, which is the model containing only the intercept, and all the stored models with j covariates. When choosing the best among these models, we need to use other model choice criteria than SSE or R^2, since these always decrease with the amount of covariates. Two options of criteria are R^2_{adj} or BIC, which penalize larger models in order to choose the best model more objectively.

With this in mind, the 10 models presented in Figure 3 were found by:

1) Fit j models for each j from 1 to k.
2) Among all these j models, choose the best one, based on SSE or R^2.

Based on the results presented in Figure 3 and 4, the best reduced regression model is the model with 5 covariates. This is because BIC has a minimum for this model. This model consists of the intercept and the covariates `sex`, `bmi`, `map`, `hdl` and `ltg`. This means that the linear model is given by

$$Y_i = \beta_0 + \beta_{\mathrm{s}} x_{i\mathrm{s}} + \beta_{\mathrm{b}} x_{i\mathrm{b}} + \beta_{\mathrm{m}} x_{i\mathrm{m}} + \beta_{\mathrm{h}} x_{i\mathrm{h}} + \beta_{\mathrm{l}} x_{i\mathrm{l}}$$

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

```
#>
#> 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
```

The estimated regression parameters have changed compared to the larger model. The intercept has been heavily increased, while the estimates of the covariates of `ldl` and `hdl` have been significantly reduced. The other estimates are very similar.

The estimated standard deviations are reduced in the reduced model compared to the larger model. They are quite similar for the parameters `sex`, `bmi` and `map`, and are heavily reduced for the intercept and the parameters `hdl` and `ltg`. This is a result of the fact that, as noted earlier, the variance of each estimated covariate is smaller in the reduced model compared to a larger model.

d)

```
full.fit <- lm(prog~., data = ds)
```

```
anova(full.fit, reduced.fit)
```

```
#> Analysis of Variance Table
#>
#> Model 1: prog ~ 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    431 1264264
#> 2    436 1288082 -5    -23817 1.6239 0.1523
```

From the test on the full model, it becomes apparent that it does not confirm that the reduced model is preferable, since the p-value is given by 0.1523. which is not significant in most cases.

Problem 2 Multiple testing

a)

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

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

When assuming that we reject all null-hypotheses with corresponding p-values below 0.05 we want to know how many null-hypotheses we end up rejecting. From the output above, we see that we end up rejecting 155 p-values in this case. A false positive finding, i.e. a type I error, is a case where the null-hypothesis is rejected despite it being true. The number of false positive findings is not known in our data, but it is possible to find a level that gives an upper limit to the probability of at least one false positive finding (FWER) (as we will discuss a bit later).

b)

The definition of familywise error rate (FWER) is the probability of one or more false positive findings $P(V > 0)$, where V is the number of false positive findings among the m tests ($m = 1000$ in this case).

To control FWER at level 0.05 means that the maximum probability of at least one false positive finding in the data is 0.05.

When using Bonferroni's method we should use $\alpha_{\text{loc}} = 0.05/1000 = 5 \cdot 10^{-5}$, if we want to control FWER at 0.05. With this new level we will reject `length(pvalues[pvalues<5e-5])` $= 50$ null-hypotheses in our data.

c)

```
pvaluesa <- ifelse(pvalues<0.05, "Reject", "Keep")
pvaluesb <- ifelse(pvalues<5e-5, "Reject", "Keep")
type1a <- sum(pvaluesa[1:900] == "Reject")
type2a <- sum(pvaluesa[900:1000] == "Keep")
type1a # Type 1 error in a)
```

```
#> [1] 55
```

```
type2a # Type 2 error in a)
```

```
#> [1] 1
```

```
type1b <- sum(pvaluesb[1:900] == "Reject")
type2b <- sum(pvaluesb[900:1000] == "Keep")
type1b # Type 1 error in b)
```

```
#> [1] 0
```

```
type2b # Type 2 error in b)
```

```
#> [1] 51
```