TMA4267 - Compulsory exercise 2

Anders Fagerli

March 12, 2019

1 Problem 1

a) Refer to the print-out from `summary(full)` in Figure 1 and *briefly* answer the following questions:

1. For each column `Estimate`, `Std.Error`, `t value`, `Pr(>|t|)`, write down the formula that the numerical values are based on, and explain all quantities used (e.g., what is $\boldsymbol{Y}$?)

   - `Estimate` - $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$
     $\boldsymbol{X}$ is the $(n \times p)$ design matrix, containing observed values for the $p-1$ covariates. $\boldsymbol{Y}$ is the observed $(n \times 1)$ response vector from which we wish to derive a linear relationship with the covariates, in this case `prog`. $\hat{\boldsymbol{\beta}}$ are the estimates of the coefficients $\boldsymbol{\beta}$, found by minimizing the squared error terms $(y_i - \hat{y}_i)^2$ with respect to $\boldsymbol{\beta}$, resulting in the above equation.

   - `Std.Error` - $\sqrt{\text{Var}(\hat{\beta}_j)}$
     The standard deviation of each estimated coefficient, found by taking the root of the diagonal entries in $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, where σ^2 is the variance of each random component in the random $(n \times 1)$ vector $\boldsymbol{\epsilon}$. The variance is usually unknown and is estimated by $\hat{\sigma}^2 = SSE/(n-p)$, where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

   - `t value` - $T_0 = \hat{\boldsymbol{\beta}}_j / \sqrt{\text{Var}(\hat{\beta}_j)}$
     The t-statistic under the assumption of $H_0 : \beta_j = 0$, used for testing if the coefficient β_j is significant in the regression. Since β_j is normally distributed and σ^2 is estimated by $\hat{\sigma}^2$, it can be shown that the statistic is t-distributed.

   - `Pr(>|t|)` - $P(T > |t| \,|\, H_0 \text{ true})$
     The p-value of the test $H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$, denoting the probability of observing T_0 or something more extreme in the direction of the alternative hypotheses H_1, given that the null hypotheses H_0 is true. Small p-values indicate that the null hypotheses may be false, in this case meaning the regression coefficient β_j is not likely to be zero.

2. How do you interpret the estimate for the intercept? (That is, which values of the covariates would give this as the predicted response?)

   - The intercept is the expected response when all covariates are set to zero. In this case, the intercept alone has no meaningful information, as setting some covariates to zero (e.g `bmi`) is unrealistic for a patient.

3. How would you explain to someone unfamiliar with linear regression how the estimated regression coefficient for `bmi` can be interpreted?

   - The estimated coefficient $\hat{\beta}_{\texttt{bmi}}$ is an estimated numerical value for how much the response `prog` increases when the BMI of a patient increases by one unit, given that all other covariates are held constant. E.g an increase in `bmi` from 24 to 25 will result in an estimated increase in `prog` by 5.6 when all other covariates are unchanged.

4. Where (in the print-out) can you find the estimated error variance? What is the formula for the estimated error variance?

   - The estimated error variance can be found from the print-out in Figure 1 at,

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

     This is the root of the estimated error variance $\hat{\sigma}^2 = SSE/(n-p) = 54.16^2$

5. Which of the covariates are found to be significant at level 0.05? Write down the null- and alternative hypotheses associated with one such test. What are the assumptions need for the p-value to be valid?

- By reading the `Signif.codes` from the `summary`, we can see that `sex`, `bmi`, `map` and `ltg` are significant at level 0.05. This gives a quick overview of the covariates that may be significant in the regression. The null- and alternative hypotheses in these tests are on the form:

$$H_0 : \beta_j = 0 \qquad \text{versus} \qquad H_1 : \beta_j \neq 0$$

The assumptions needed for the p-value to be valid are the model assumptions under multiple linear regression. When conducting the hypotheses test itself we test for a linear relationship between the response and the covariates, so the assumption for the p-value to be valid is strictly only that the random errors are normally distributed with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

b) How would you, based on Figures 1 and 2 evaluate the fit of the full model? Is the regression significant? Write down the null- and alternative hypotheses for this test. Explain what the number called Multiple R-squared in Figure 1 means.

There is a lot of information from Figures 1 and 2 depicting the fit of the model:

- `Multiple R-squared -` This number gives the proportion of variability explained by the regression,

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 0.5176$$

This value ranges from 0 to 1, where $R^2 = 1$ means a perfect fit where all residuals are zero. We can see that almost half of the variance is unexplained, indicating a poor fit in terms of variance explained.

- `F-statistic -` This statistic is used to test if the regression is significant, given by the test

$$H_0 : \beta_{age} = \beta_{sex} = \cdots = \beta_{glu} = 0 \qquad \text{versus} \qquad H_1 : \text{at least one} \neq 0$$

The statistic,

$$F = \frac{(SSE_0 - SSE)/r}{SSE/(n-p)} \sim F_{r,n-p}$$

is used for general hypotheses testing, where SSE_0 is the error sum of squares for the restricted model, in this case when all coefficients are set to zero (except intercept). A p-value can be calculated from the statistic, which from the `summary` can be seen to be `p-value:2.2e-16`. This indicates a false null hypotheses, meaning the regression is likely to be significant and our model fits the data well.

- `Residuals vs. fitted values -` A plot of studentized residuals against the fitted values for `prog`. An assumption we make for a linear regression model is $\mathrm{Var}(\epsilon_i) = \sigma^2$, meaning the variance in each residual is constant for all fitted values. This should result in a fairly constant spread of residuals in the plot, with no recognizable patterns. We see from the plot that the studentized residuals tend to lower values and larger spread for larger values of `prog`, meaning the assumption of constant variance may not be valid for our model.

- `Q-Q plot -` A plot depicting whether our model comes from another known distribution, in this case the normal distribution. An assumption we make for a linear regression model is that the residuals are normally distributed, which should result in a straight line in the Q-Q plot. We see they form a relatively straight line alongside the red line depicting the normal distribution, but they deviate at the tails. It is difficult to conclude anything from the Q-Q plot alone, but the p-value for the `Anderson-Darling normality test` from the `summary` suggests that the assumption of normality is valid for our model. An Anderson-Darling normality test will test the null hypothesis that the model is normally distributed, where our p-value of 0.4176 means we cannot reject the null hypotheses.

- `Scatter plot -` A plot depicting the relationship between the different covariates and the response. The bottom row is of interest as it displays the relationship between each covariate and the response, where we are looking for a linear relationship for our model. From the plots we can see that `prog` may not be linear with `age`, `tc`, `ldl` and `tch`. This may indicate that a reduced model will give a better fit, with less variability.

In total, the model fits the data fairly well. The regression is significant, but a large proportion the variability is unexplained. We also see that some of the assumptions on the residuals may not be valid, and that some of the covariates may not have a linear relationship with the response.

c) Why might a reduced model have better performance than a full model when the aim is prediction? Explain briefly what is done in the best subset model selection, and give the reasoning behind the R^2_{adj} and BIC criteria. In particular, explain how the 10 models presented in Figure 3 was found. Results from using the R^2_{adj} and the BIC criteria are presented in Figures 3 and 4. Based on these results, choose a reduced regression model, fit this reduced model in R, and write down the fitted regression model for the model you choose. Compare the estimated regression parameters and the estimated standard deviations for the full model (Figure 1) and the reduced model that you choose. Explain what you observe.

A reduced model may perform better than the full model in prediction, if the full model contains covariates that have little to no relationship with the response. These covariates will act as noise, adding variability to the model, and thus adding uncertainty to a prediction. In practice, this means e.g a prediction interval will be larger for a given significance level.

The best subset model selection aims to produce the combination of covariates that minimizes some model choice criteria. The method goes through every combination of covariates $\binom{p-1}{j}$ for $j = 1, 2, \ldots, p-1$ and picks the combination for each j giving the smallest SSE or largest R^2. It then selects one of the $p-1$ models according to the model choice criteria. The SSE or R^2 alone will not function as a model choice criteria, as the bigger model will always produce a lower SSE or larger R^2. The criteria must therefore penalize larger models. The R^2_{adj} and BIC are examples of possible model choice criteria.

- `Adjusted` R^2`:`

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

  The larger coefficient of determination R^2_{adj}, the better fit. We see from the numerator of the fraction that an increase in covariates p will lead to a decrease in R^2_{adj}.

- `Bayesian information criterion:`

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

  The smaller Bayesian information criterion, the better fit. We can see from the last part in the equation how it penalizes larger models.

The print-out in Figure 3 presents the result of the best subset model selection for each model size $j = 1, 2, \ldots, 10$, using the R function `regsubsets`. From the plots in Figure 4 (or the print-out of `allsummary$bic` and `allsummary$adjr2`) we can see which models fit the data best. The lowest BIC is at 5 covariates, while the largest R^2_{adj} is at 8 covariates. Since the difference in R^2_{adj} from 8 covariates to 5 covariates is so small, we choose the model with the 5 covariates `sex`, `bmi`, `map`, `hdl` and `ltg`, as this combination produces the lowest BIC. It also has fewer covariates, which may result in a model with lower variance. The following R code may be used to fit the reduced model:

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

A print-out of the estimated parameters may be found from `summary`, attached below.

```
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
```

We can see a small change in estimated coefficients, which comes from the change in model. Most noticeable is the change in `hdl` and `ltg`, where `hdl` has changed sign and `ltg` has a smaller effect on `prog` in the reduced model. As the coefficient for `hdl` in both models takes a low value, with a relatively high `Std.Error` in the full model compared to its estimate, this may not be significant. The change in sign may also indicate that `hdl` is correlated to some of the covariates outside the reduced model. We can see that the `Std.Error` for all coefficients in the reduced model are reduced, giving slightly more accurate estimates. Most noteworthy may be the change in the `F-statistic`, which has an increase from 46.25 to 90.24. This may indicate that the reduced model will fit the data just as well as the full model.

d) Perform the test

$$H_0 : \beta_{age} = \beta_{tc} = \beta_{ldl} = \beta_{tch} = \beta_{glu} = 0 \qquad \text{versus} \qquad H_1 : \text{at least one} \neq 0$$

in the full model. Report a p-value of the test. Comment on the result. Would you prefer the full or the reduced model?

The test can be performed by calculating the `F-statistic` and corresponding p-value for a chosen significance level, or using the R function `anova` with the reduced and full model as inputs.

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

```
Analysis of Variance Table

Model 1: prog ~ sex + bmi + map + hdl + ltg
Model 2: prog ~ age + sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    436 1288082
2    431 1264264  5     23817 1.6239 0.1523
```

We can see from the print-out that the p-value of the test is 0.1523. Using a standard significance level of $\alpha = 0.05$, we cannot reject the null hypotheses, and therefore conclude that some of the covariates in H_0 may be significant in the regression. We therefore prefer the full model.

2 Problem 2

a) Assume that we reject all null-hypotheses with corresponding p-values below 0.05. How many null-hypotheses do we reject? What is a false positive finding (type I error)? Do we know the number of false positive findings in our data?

We use R to find the rejected hypotheses:

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

This results in 155 rejected null-hypotheses when the significance level is $\alpha = 0.05$.

A false positive finding is rejecting H_0 when it is true. We do not know the number of false positive findings, as we do not have the true values of the coefficients.

b) What is the definition of the familywise error rate, FWER? What does it mean to control the FWER at level 0.05? What cut-off on p-values (significance level) should we use if we want to control the FWER at level 0.05 for our data with the Bonferroni method? How many null-hypotheses will we reject with this new cut-off?

The family wise error rate, FWER, is the probability of one or more false positive findings.

$$\text{FWER} = P(V > 0), \qquad V \text{ number of false positive findings}$$

The number of false positive findings V is unknown, but we may set the cut-off on the individual p-values so that we control the FWER to a specific significance level α, where the cut-off is defined as α_{loc}. The Bonferroni method is a conservative way to control the FWER, where α_{loc} is given by

$$\alpha_{loc} \leq \frac{\alpha}{m}, \qquad m \text{ number of tests}$$

Setting $\alpha = 0.05$ and $m = 1000$, we get

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

Using this cut-off, we reject 50 null-hypotheses.

c) To see the effect of choosing different cut-offs on p-value on the number of false positive findings we need to know which null hypotheses are true and which are false. Let us assume that the first 900 null hypotheses are true and the last 100 are false. What does this imply about the number of type I and type II errors in (a) and (b)?

We calculate the number of type I errors by summing the number of p-values below the cut-off for the first 900 p-values, and the number of type II errors by summing the number of p-values above the cut-off for the last 100 p-values.

(a) Type I: 55
Type II: 0

(b) Type I: 0
Type II: 50

We now see the effect of choosing the cut-off, which here results in a trade-off between type I and type II errors. Normally, we choose the one with lowest type I error.