Compulsory exercise 1: Group X

TMA4268 Statistical Learning V2018

NN1, NN and NN3

Date when you hand in

Problem 2 - Linear regression

```
##
## Call:
## lm(formula = -1/sqrt(SYSBP) ~ ., data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0207366 -0.0039157 -0.0000304  0.0038293  0.0189747
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.103e-01  1.383e-03 -79.745  < 2e-16 ***
## SEX         -2.989e-04  2.390e-04  -1.251 0.211176
## AGE          2.378e-04  1.434e-05  16.586  < 2e-16 ***
## CURSMOKE    -2.504e-04  2.527e-04  -0.991 0.321723
## BMI          3.087e-04  2.955e-05  10.447  < 2e-16 ***
## TOTCHOL      9.288e-06  2.602e-06   3.569 0.000365 ***
## BPMEDS       5.469e-03  3.265e-04  16.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005819 on 2593 degrees of freedom
## Multiple R-squared:  0.2494, Adjusted R-squared:  0.2476
## F-statistic: 143.6 on 6 and 2593 DF,  p-value: < 2.2e-16
```

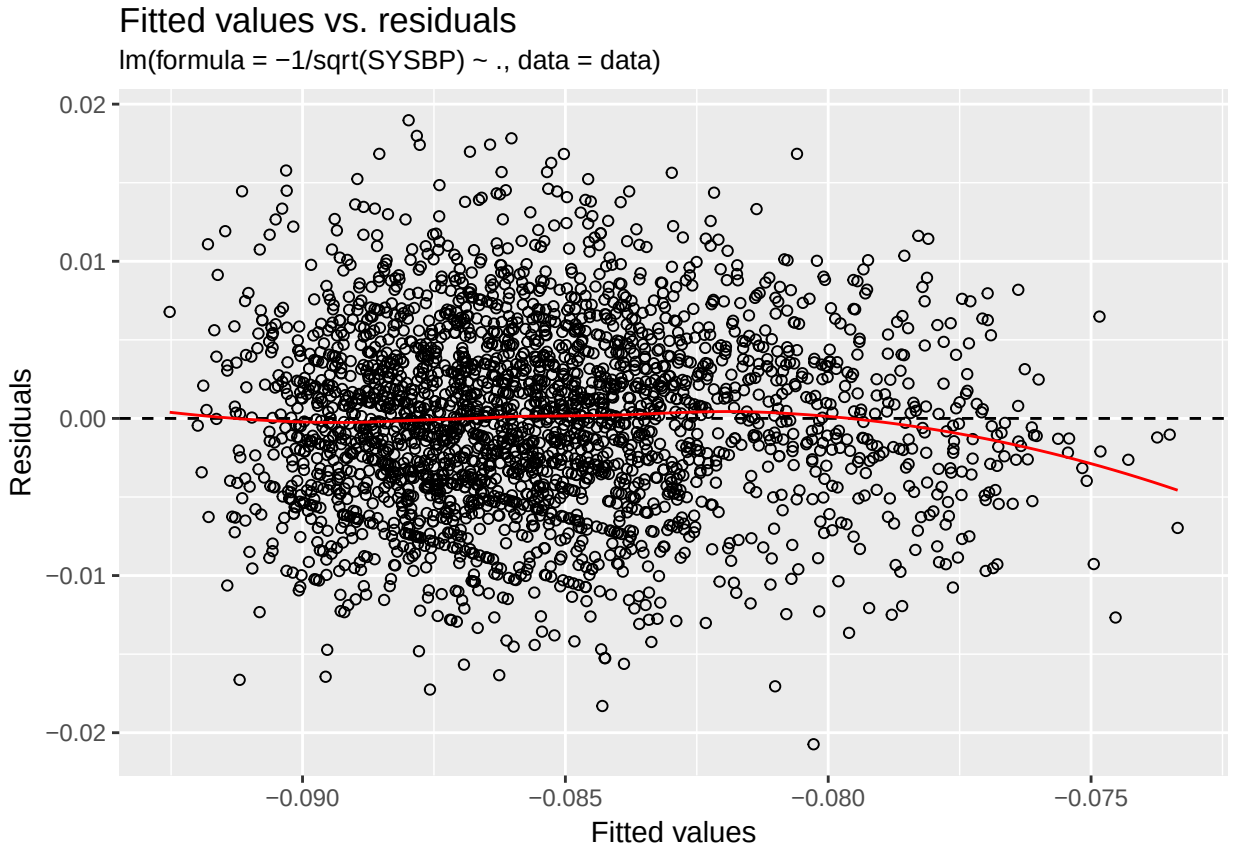
a)

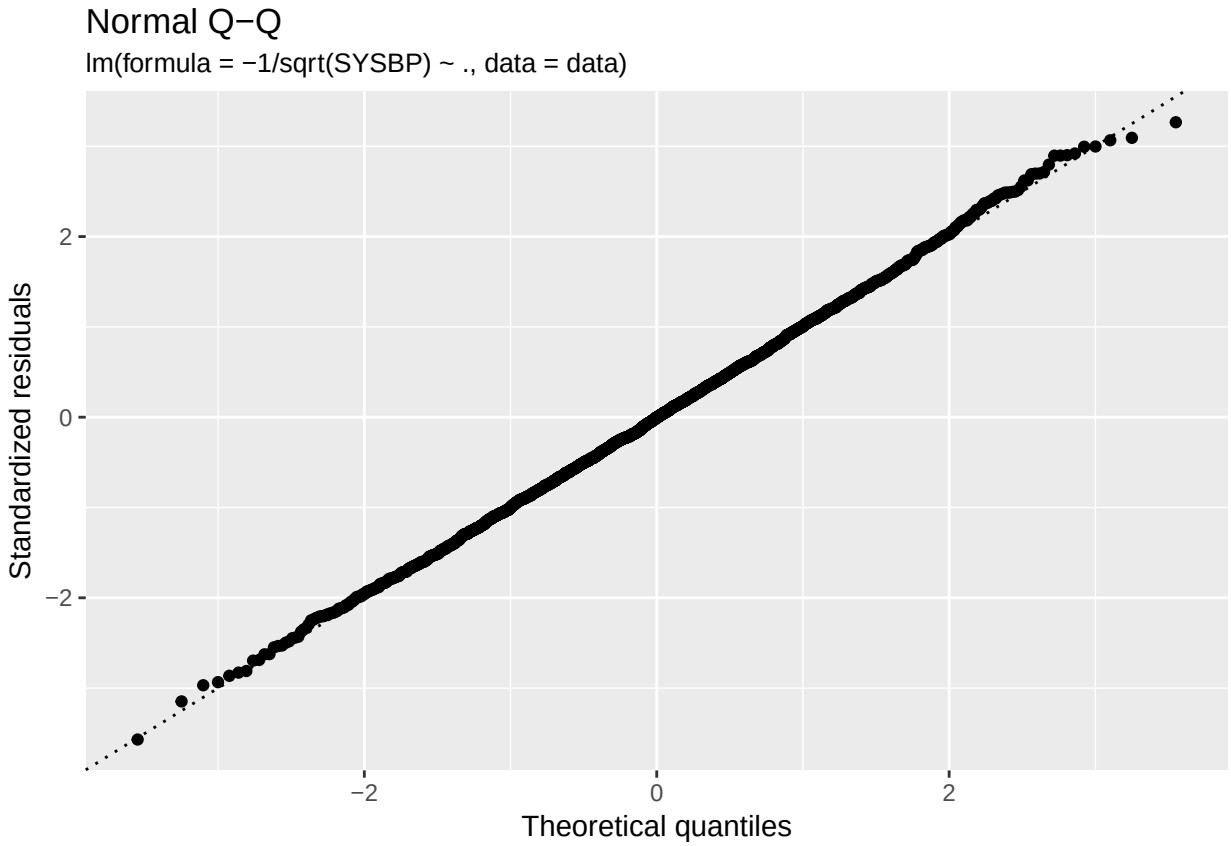
- The fittet model has the equation $\hat{Y} = X\hat{\beta}$, where $\hat{\beta} = (X^T X)^{-1} X^T Y$, X is the $n \times (p+1)$ design matrix, and Y is the corresponding response values.
- "Estimate" is the estimated coeffictients obtaining the minimum residual square error with the data set. The "intercept" is the constant term in the regression model.
- The "standard error" is the estimated standard deviation in the estimated coefficients. It is given as the square root of $\hat{Var}(\hat{\beta}_j) = c_{jj}\hat{\sigma}^2$, where $c_{ij} = ((X^T X)^{-1})_{ij}$ and $\hat{\sigma}^2 = (Y - \hat{Y})^T (Y - \hat{Y})/(n - p - 1)$.
- The "t value" is for every coefficent j, $\frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$ which is t distributed with $n - p - 1$ degrees of freedom under the assumtion that H_0 is true, that is, β_j truly is 0. "Pr(t > |t|)" is then the probability of obeserving such an extreme t value given that H_0 is true. Hence $\Pr(>|t|) := P(|T_{n-p-1}| \geq |\frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}|) = 2P(T_{n-p-1} \geq |\frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}|)$.
- The "Residual standard error" is our estimate for the variance of Y. The standard error squared is given as $\hat{\sigma}^2 = (Y - \hat{Y})^T (Y - \hat{Y})/(n - p - 1)$.

- The "F - statistic" is used to check the hypothesis of all betas being 0. In the table it is given as $\frac{(TSS-RSS)/p}{RSS/(n-p-1)}$, which is Fisher distributed with degrees of freedom p and $n-p-1$, where $TSS := \sum_{i=1}^{n}(y_i - \bar{y})^2$, and $RSS := \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

b)

- The proportion of variability explained by the model is given by the R^2- statistic $:= (TSS-RSS)/TSS$, here being equal to 0.2494. Hence our model explains approximately 25% of the variance in the response value.



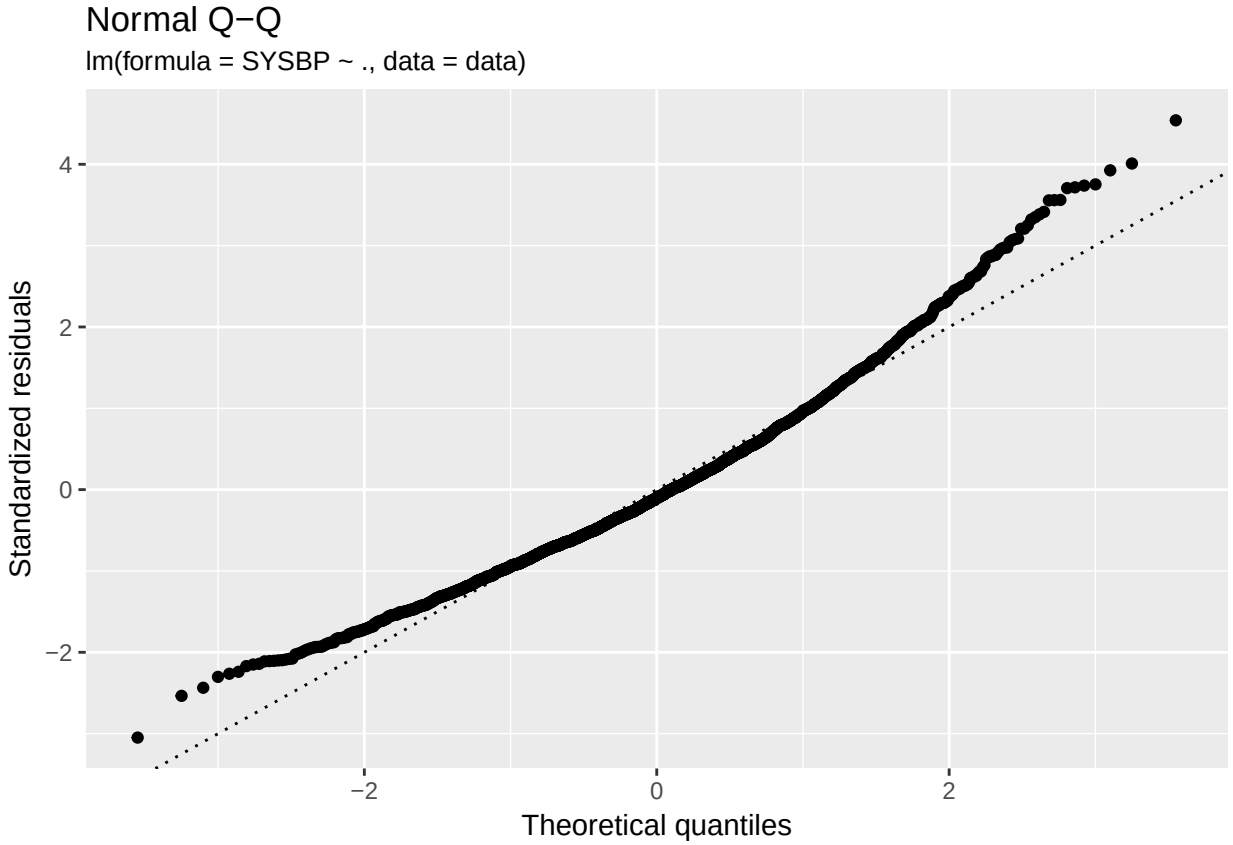


- Looking at the plot of residuals vs. fitted values we note that it does not appear to be a correlation between the value of the response and the variance of the response, and the mean appears to be 0. This fits well with the assumtion of the noise being normally distributed with mean 0 and constant variance.

  The QQ-plots strengthens our belief in this assumtion, as the points form a linear line.

```
##
## Call:
## lm(formula = SYSBP ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.800 -13.471  -1.982  11.063  88.959
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 56.505170   4.668798  12.103  < 2e-16 ***
## SEX         -0.429973   0.807048  -0.533  0.59424
## AGE          0.795810   0.048413  16.438  < 2e-16 ***
## CURSMOKE    -0.518742   0.853190  -0.608  0.54324
## BMI          1.010550   0.099770  10.129  < 2e-16 ***
## TOTCHOL      0.028786   0.008787   3.276  0.00107 **
## BPMEDS      19.203706   1.102547  17.418  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.65 on 2593 degrees of freedom
## Multiple R-squared:  0.2508, Adjusted R-squared:  0.249
```

```
## F-statistic: 144.6 on 6 and 2593 DF,  p-value: < 2.2e-16
```



- The RSE is considerably larger for model B, that is, the estimated variance in both the response and our estimated coefficients are larger in this model. Looking at the diagnostic plots of model B we also note that the QQ-plot suggests that these residuals are not normally distributed. Clearly we prefer model A to make inference about systolic blood pressure, for this model brings more likely coefficent estimates and follows the noise assumptions better.

c)

- The estimate for $\hat{\beta}_{BMI}$ is $3.087 \cdot 10^{-4}$.
- We interpret the estimated coefficient $\hat{\beta}_{BMI}$ as the coefficient of the variable containing the value of BMI in the linear expression for $-1/\sqrt{SYSBP}$, that is, the impact of change in BMI on the response

$$\hat{\beta}_{BMI} = \frac{\partial(-1/\sqrt{SYSBP})}{\partial BMI}$$

- Since $\hat{\beta}_{BMI} \sim N(\beta_{BMI}, \sigma^2 c_{BMI})$, where $c_{BMI} :=$ diagonal entry corresponding to BMI of $(X^T X)^{-1}$ we have

$$\frac{(\hat{\beta}_{BMI} - \beta_{BMI})/(\sigma\sqrt{c_{BMI}})}{\sqrt{\frac{1}{\sigma^2} RSS/(n-p-1)}} = \frac{\hat{\beta}_{BMI} - \beta_{BMI}}{\sqrt{\frac{RSS}{n-p-1} c_{BMI}}} \sim T_{n-p-1}$$

It follows that

$$Pr(\beta_{BMI} \in (\hat{\beta}_{BMI} - \hat{\sigma}\sqrt{c_{BMI}}t_{0.995,2593}, \hat{\beta}_{BMI} - \hat{\sigma}\sqrt{c_{BMI}}t_{0.005,2593})) = 0.99$$

Setting $t_{0.005,2593} = -2.577727$ and $t_{0.995,2593} = 2.577727$, we compute the interval to be $(2.325282 \cdot 10^{-4}, 3.848718 \cdot 10^{-4})$. This interval tells us that with probability 0.99, the true value of the coefficient is contained in this interval.

- We note that if H_0 is true, the center of the t distribution for prediction of $\hat{\beta}_{BMI}$ would be 0, but the degrees of freedom the same as for this prediction. Hence, a 99% prediction interval for the estimated coefficient would in this case be $(-|2.325282 \cdot 10^{-4} - 3.087 \cdot 10^{-4}|, |3.848718 \cdot 10^{-4} - 3.087 \cdot 10^{-4}|) = (-7.61718 \cdot 10^{-5}, 7.61718 \cdot 10^{-5})$. Clearly our observed value is outside the interval, meaning that the p value must be less than or equal to 0.01.

d)

- Model A predicts the response of these values to be -0.08667246, which corresponds to a SYSBP of 133.1183.
- Let $\tilde{Y_0}$ be a new observation of $-1/\sqrt{SYSBP}$ corresponding to the point x_0. Since we have $\tilde{Y_0} - {x_0}^T\beta \sim N(0, \sigma^2(1 + {x_0}^T(X^TX)^{-1}x_0))$ we get

$$\frac{(\tilde{Y_0} - {x_0}^T\beta)/(\sigma\sqrt{1 + {x_0}^T(X^TX)^{-1}x_0})}{\sqrt{\frac{1}{\sigma^2}RSS/(n-p-1)}} = \frac{\tilde{Y_0} - {x_0}^T\beta}{\hat{\sigma}\sqrt{1 + {x_0}^T(X^TX)^{-1}x_0}} \sim T_{n-p-1}$$

letting $\tilde{Y_0} = -\frac{1}{\sqrt{Y_0}}$ we obtain the following prediction interval for SYSBP at x_0

$$Pr(Y_0 \in \left(\frac{1}{({x_0}^T\beta + \hat{\sigma}kt_{0.05,2593})^2}, \frac{1}{({x_0}^T + \hat{\sigma}kt_{0.95,2593})^2}\right)) = 0.90, k = \sqrt{1 + {x_0}^T(X^TX)^{-1}x_0}$$

Setting $t_{0.05,2593} = -1.645441$ and $t_{0.95,2593} = 1.645441$ we compute the following prediction interval $(107.9250, 168.2845)$.

- ??????????????????????????

Problem 3 - Classification

a)

- We want to show that

$$logit(p_i) = log(\frac{p_i}{1-p_i})$$

is a linear function, where

$$p_i = \frac{e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}}{1+e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}}.$$

We see that

$$1 - p_i = 1 - \frac{e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}}{1+e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}} = \frac{1}{1+e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}}.$$

Thus

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = log\left(\frac{\frac{e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}}{1+e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}}}{\frac{1}{1+e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}}}\right) = log(e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}}) = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2}.$$

So $logit(p_i)$ is linear.

-
- $\hat{\beta}_1$ and $\hat{\beta}_2$ can be interpreted as how the odds vary. The odds is given as $\frac{p_i}{1-p_i}$. As shown in the class notes, if the covariate x_{1i} is increased by one unit, the odds is multiplied by $exp(\beta_1)$. The same is true for x_{1i} and $exp(\beta_1)$. $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimators of β_1 and β_2 and are estimated using the training data.
- We find the formula for the class boundary by solving $\hat{Pr}(Y = 1|\mathbf{x}) = 0.5$. This gives

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}}} = 0.5, e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}}$$

so

$$0.5e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}} = 0.5.$$

This means that we need $\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} = 0$. Thus

$$\mathbf{x}_2 = -\frac{\hat{\beta}_0}{\hat{\beta}_2} - \frac{\hat{\beta}_1}{\hat{\beta}_2}\mathbf{x}_1,$$

and the formula is linear.

- The training data is plotted with the class boundary.