Module 3: Recommended Exercises

TMA4268 Statistical Learning V2020

Stefanie Muff, Department of Mathematical Sciences, NTNU

January xx, 2020

Contents

Problem 1: Compulsory exercise in 2018

There will be a very similar regression problem in the compulsory exercise 1 in 2019!

The Framingham Heart Study is a study of the etiology (i.e. underlying causes) of cardiovascular disease, with participants from the community of Framingham in Massachusetts, USA. For more more information about the Framingham Heart Study visit https://www.framinghamheartstudy.org/. The dataset used in here is subset of a teaching version of the Framingham data, used with permission from the Framingham Heart Study.

We will focus on modelling systolic blood pressure using data from n = 2600 persons. For each person in the data set we have measurements of the seven variables

- `SYSBP` systolic blood pressure,
- `SEX` 1=male, 2=female,
- `AGE` age in years at examination,
- `CURSMOKE` current cigarette smoking at examination: 0=not current smoker, 1= current smoker,
- `BMI` body mass index,
- `TOTCHOL` serum total cholesterol, and
- `BPMEDS` use of anti-hypertensive medication at examination: 0=not currently using, 1=currently using.

A multiple normal linear regression model was fitted to the data set with `-1/sqrt(SYSBP)` as response and all the other variables as covariates.

```r
library(ggplot2)
data = read.table("https://www.math.ntnu.no/emner/TMA4268/2018v/data/SYSBPreg3uid.txt")
dim(data)
colnames(data)
modelA = lm(-1/sqrt(SYSBP) ~ ., data = data)
summary(modelA)
```

a) Understanding model output

We name the model fitted above `modelA`.

- Write down the equation for the fitted `modelA`.
- Explain (with words and formula) what the following in the `summary`-output means.
- `Estimate` - in particular interpretation of `Intercept`
- `Std.Error`
- `t value`
- `Pr(>|t|)`
- `Residual standard error`
- `F-statistic`

b) Model fit

- What is the proportion of variability explained by the fitted `modelA`? Comment.
- Use diagnostic plots of "fitted values vs. standardized residuals"" and "QQ-plot of standardized residuals" (see code below) to assess the model fit.
- Now fit a model, call this `modelB`, with `SYSBP` as response, and the same covariates as for `modelA`. Would you prefer to use `modelA` or `modelB` when the aim is to make inference about the systolic blood pressure?

```r
# residuls vs fitted
ggplot(modelA, aes(.fitted, .resid)) + geom_point(pch = 21) + geom_hline(yintercept = 0,
    linetype = "dashed") + geom_smooth(se = FALSE, col = "red", size = 0.5,
    method = "loess") + labs(x = "Fitted values", y = "Residuals", title = "Fitted values vs. residuals
    subtitle = deparse(modelA$call))

# qq-plot of residuals
ggplot(modelA, aes(sample = .stdresid)) + stat_qq(pch = 19) + geom_abline(intercept = 0,
    slope = 1, linetype = "dotted") + labs(x = "Theoretical quantiles",
    y = "Standardized residuals", title = "Normal Q-Q", subtitle = deparse(modelA$call))
```

c) Confidence interval and hypothesis test

We use `modelA` and focus on addressing the association between BMI and the response.

- What is the estimate $\hat{\beta}_{\text{BMI}}$ (numerically)?
- Explain how to interpret the estimated coefficient $\hat{\beta}_{\text{BMI}}$.
- Construct a 99% confidence interval for β_{BMI} (write out the formula and calculate the interval numerically). Explain what this interval tells you.
- From this confidence interval, is it possible for you know anything about the value of the p-value for the test $H_0 : \beta_{\text{BMI}} = 0$ vs. $H_1 : \beta_{\text{BMI}} \neq 0$? Explain.

d) Prediction

Consider a 56 year old man who is smoking. He is 1.75 meters tall and his weight is 89 kilograms. His serum total cholesterol is 200 mg/dl and he is not using anti-hypertensive medication.

```r
names(data)
new = data.frame(SEX = 1, AGE = 56, CURSMOKE = 1, BMI = 89/1.75^2, TOTCHOL = 200,
    BPMEDS = 0)
```

- What is your best guess for his `-1/sqrt(SYSBP)`? To get a best guess for his `SYSBP` you may take the inverse function of `-1/sqrt`.

(Comment: Is that allowed - to only do the inverse? Yes, that could be the result of a first order Taylor expansion approximation.)

- Construct a 90% prediction interval for his systolic blood pressure `SYSBP`. Comment. Hint: first contruct values on the scale of the response `-1/sqrt(SYSBP)` and then transform the upper and lower limits of the prediction interval.
- Do you find this prediction interval useful? Comment.

Problem 2: Theoretical questions

a)

A core finding is $\hat{\boldsymbol\beta}$.

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

with $\hat{\boldsymbol\beta} \sim N_p(\boldsymbol\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

- Show that $\hat{\boldsymbol\beta}$ has this distribution with the given mean and covariance matrix.
- What do you need to assume to get to this result?
- What does this imply for the distribution of the jth element of $\hat\beta$?
- In particular, how can we calculate the variance of $\hat\beta_j$?

b)

What is the interpretation of a 95% confidence interval? Hint: repeat experiment (on Y), on average how many CIs cover the true β_j?

c)

What is the interpretation of a 95% prediction interval? Hint: repeat experiment (on Y) for a given $\mathbf{x}_0$.

d)

Construct a 95% CI for $\mathbf{x}_0^T\beta$. Explain what is the connections between a CI for β_j, a CI for $\mathbf{x}_0^T\beta$ and a PI for Y at $\mathbf{x}_0$.

e)

Explain the difference between *error* and *residual*. What are the properties of the raw residuals? Why don't we want to use the raw residuals for model check? What is our solution to this?

f)

Consider a multiple linear regression model A and a submodel B (all parameters in B are in A also). We say that B is nested within A. Assume that regression parameters are estimated using least squares. Why is then the following true: RSS for model A will always be smaller or equal to RSS for model B. And thus, R^2 for model A can never be worse than R^2 for model B. (See also Problem 3d below.)

Problem 3: Munich Rent index

a)

Fit the regression model with first `rent` and then `rentsqm` as response and following covariates: `area`, `location` (dummy variable coding using location2 and location3, just write `as.factor(location)`), `bath`, `kitchen` and `cheating` (central heating).

Look at diagnostic plots for the two fits. Which response do you prefer?

Consentrate on the response-model you choose for the rest of the tasks.

b)

Explain what the parameter estimates mean in practice. In particular, what is the interpretation of the intercept?

c)

Go through the summary printout and explain all parts.

d)

Now we add random noise as a covariance, but simulating the IQ of the landlord of each appartment. Observe that R^2 increases (or stays unchanged) and RSS decreases (or stays the same) if we add IQ as covariate, but R^2_{adj} decreases. What does this tell you about model selection and overfitting?

For the code - what is the connection between `sigma` and RSS?

```r
library(gamlss.data)
orgfit = lm(rent ~ area + as.factor(location) + bath + kitchen + cheating,
    data = rent99)
summary(orgfit)
set.seed(1)  #to be able to reproduce results
n = dim(rent99)[1]
IQ = rnorm(n, 100, 16)
fitIQ = lm(rent ~ area + as.factor(location) + bath + kitchen + cheating +
    IQ, data = rent99)
summary(fitIQ)

summary(orgfit)$sigma
summary(fitIQ)$sigma

summary(orgfit)$r.squared
summary(fitIQ)$r.squared
summary(orgfit)$adj.r.squared
summary(fitIQ)$adj.r.squared
```

```
##
## Call:
## lm(formula = rent ~ area + as.factor(location) + bath + kitchen +
##     cheating, data = rent99)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -633.41  -89.17   -6.26   82.96 1000.76
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -21.9733    11.6549  -1.885   0.0595 .
## area                   4.5788     0.1143  40.055  < 2e-16 ***
## as.factor(location)2  39.2602     5.4471   7.208 7.14e-13 ***
## as.factor(location)3 126.0575    16.8747   7.470 1.04e-13 ***
## bath1                 74.0538    11.2087   6.607 4.61e-11 ***
## kitchen1             120.4349    13.0192   9.251  < 2e-16 ***
## cheating1            161.4138     8.6632  18.632  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 145.2 on 3075 degrees of freedom
## Multiple R-squared:  0.4504, Adjusted R-squared:  0.4494
## F-statistic:   420 on 6 and 3075 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = rent ~ area + as.factor(location) + bath + kitchen +
##     cheating + IQ, data = rent99)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -630.95  -89.50   -6.12   82.62  995.76
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -41.3879    19.5957  -2.112   0.0348 *
## area                   4.5785     0.1143  40.056  < 2e-16 ***
## as.factor(location)2  39.2830     5.4467   7.212 6.90e-13 ***
## as.factor(location)3 126.3356    16.8748   7.487 9.18e-14 ***
## bath1                 74.1979    11.2084   6.620 4.23e-11 ***
## kitchen1             120.0756    13.0214   9.221  < 2e-16 ***
## cheating1            161.4450     8.6625  18.637  < 2e-16 ***
## IQ                     0.1940     0.1574   1.232   0.2179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 145.2 on 3074 degrees of freedom
## Multiple R-squared:  0.4507, Adjusted R-squared:  0.4494
## F-statistic: 360.3 on 7 and 3074 DF,  p-value: < 2.2e-16
##
## [1] 145.1879
## [1] 145.1757
## [1] 0.4504273
## [1] 0.4506987
## [1] 0.449355
## [1] 0.4494479
```

Problem 4: Simulations in R

a

Make R code that shows the interpretation of a 95% CI for β_j. Hint: Theoretical question a.

b

Make R code that shows the interpretation of a 95% PI for a new response at $\mathbf{x}_0$. Hint: Theoretical question b.

c.

For simple linear regression, simulate at data set with homoscedastic errors and with heteroscedastic errors. Here is a suggestion of one solution - not using `ggplot`. You use `ggplot`. Why this? To see how things looks when the model is correct and wrong.

```r
# Homoscedastic errore
n = 1000
x = seq(-3, 3, length = n)
beta0 = -1
beta1 = 2
xbeta = beta0 + beta1 * x
sigma = 1
e1 = rnorm(n, mean = 0, sd = sigma)
y1 = xbeta + e1
ehat1 = residuals(lm(y1 ~ x))
plot(x, y1, pch = 20)
abline(beta0, beta1, col = 1)
plot(x, e1, pch = 20)
abline(h = 0, col = 2)
# Heteroscedastic errors
sigma = (0.1 + 0.3 * (x + 3))^2
e2 = rnorm(n, 0, sd = sigma)
y2 = xbeta + e2
ehat2 = residuals(lm(y2 ~ x))
plot(x, y2, pch = 20)
abline(beta0, beta1, col = 2)
plot(x, e2, pch = 20)
abline(h = 0, col = 2)
```

Problem 5

A problem with an interaction term between two continuous variables, and between a continuous and a factor covariable with more than two levels.

Important: In the latter case, we need again the F-test to check if the interaction is relevant.