MS-C1620 Statistical Inference

Exercise 10

Homework exercise

To be solved at home before the exercise session.

1. a. Consider the following linear model,

$$\mathbb{E}(y_i|\mathbf{x}_i) = \beta_0 + \beta_1 \mathrm{sex}_i + \beta_2 \mathrm{age}_i + \beta_3(\mathrm{sex}_i \times \mathrm{age}_i),$$

where sex_i is a binary variable (0 = male, 1 = female) and age_i is a continuous variable. Write down the model separately for males and females and using the two models give interpretations for the four parameters.

b. The data set `galaxy` from the package `ElemStatLearn` contains measurements on the position and radial velocity of the galaxy NGC7531. Fitting a model with the latter as a response, we get the following model summary and residual plot. Does the model fit well? If not, what could be tried next?

```
library(ElemStatLearn)
library(car)
```

```
## Loading required package: carData
```

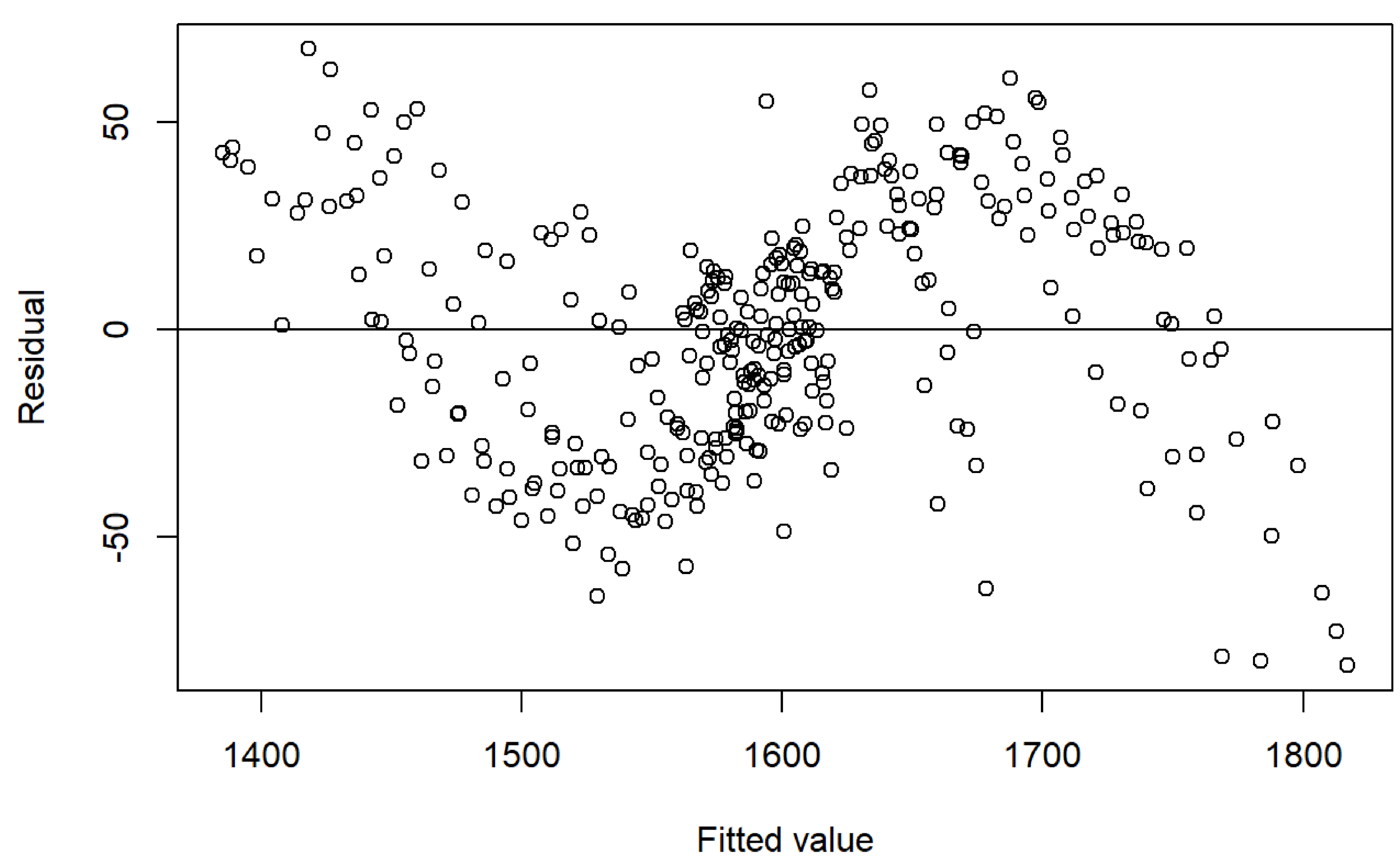
```
lm_galaxy <- lm(velocity ~ ., data = galaxy)
summary(lm_galaxy)
```

```
##
## Call:
## lm(formula = velocity ~ ., data = galaxy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.988 -23.673   0.442  22.770  67.527
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1589.42295    3.92939 404.496  < 2e-16 ***
## east.west         0.77410    0.31202   2.481  0.01362 *
## north.south      -3.19179    0.09697 -32.914  < 2e-16 ***
## angle             0.12454    0.04396   2.833  0.00491 **
## radial.position   0.90118    0.16042   5.618 4.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.13 on 318 degrees of freedom
## Multiple R-squared:  0.8991, Adjusted R-squared:  0.8979
## F-statistic: 708.6 on 4 and 318 DF,  p-value: < 2.2e-16
```

```
vif(lm_galaxy)
```

```
##       east.west     north.south           angle radial.position
##        4.996114        1.747546        1.002817        6.118775
```

```
plot(fitted(lm_galaxy), resid(lm_galaxy), xlab = "Fitted value", ylab = "Residual")
abline(h = 0)
```



Class exercise

To be solved at the exercise session.

1. The data set `Chirot` from the package `carData` contains statistics on the 1907 Romanian peasant rebellion. Each row of the data is a county for which the `intensity` of the rebellion has been measured, along with various socio-economic variables. Investigate using linear regression whether there is dependency between `intensity` and the explanatory variables.
   a. Visualize the data.
   b. Fit a linear regression model to the data.
   c. Assess the adequacy of the model and its assumptions through the model summary, VIFs and model diagnostics.
   d. Make changes to the model, if needed.
   e. Interpret the results.

2. The data set `longley` contains measurements of economic variables from the years 1947-1962. We are interested in predicting the number of people employed (`Employed` , in thousands) using the other variables.
   a. Visualize the data.
   b. Fit a linear regression model to the data.
   c. Assess the adequacy of the model through the model summary and VIFs.
   d. Make changes to the model, if needed.

3. **(Optional)** While general non-linear regression is beyond this course, fitting such models with R is quite straightforward. Try out the following code where a non-linear Generalized Additive Model (GAM) is fitted between temperature and ozone content in the `airquality` data.

```
# install.packages("mgcv")
library(mgcv)

x <- data.frame(ozone = airquality[, 1], temp = airquality[, 4])
gam_1 <- gam(temp ~ s(ozone), data = x)

plot(x, xlab = "Ozone", ylab = "Temperature")
ozone_grid <- data.frame(ozone = seq(min(x$ozone, na.rm = TRUE), max(x$ozone, na.rm = TRUE),length.out = 1000))
points(ozone_grid[, 1], predict(gam_1, ozone_grid), type = 'l')
```

Investigate especially what the final three lines do and what is the meaning of `ozone_grid` . Try also to fit a non-linear model to some other data set using the above as a template.