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.

```
# For males, sex = 0 and we have
# E(y|x) = b0 + 0 + b2*age + b3(0 * age) = b0 + b2*age

# For females, sex = 1 and we have
# E(y|x) = b0 + b1*1 + b2*age + b3(1 * age) = (b0 + b1) + (b2 + b3)*age

# That is,
# b0 = the intercept of males
# b1 = the difference between the intercepts of females and males
# b2 = the slope of age for males
# b3 = the difference between the age-slopes of females and males (e.g., if b3 = 0, the effect of age on the response is the same for both sexes).
```

```
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?
```

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

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

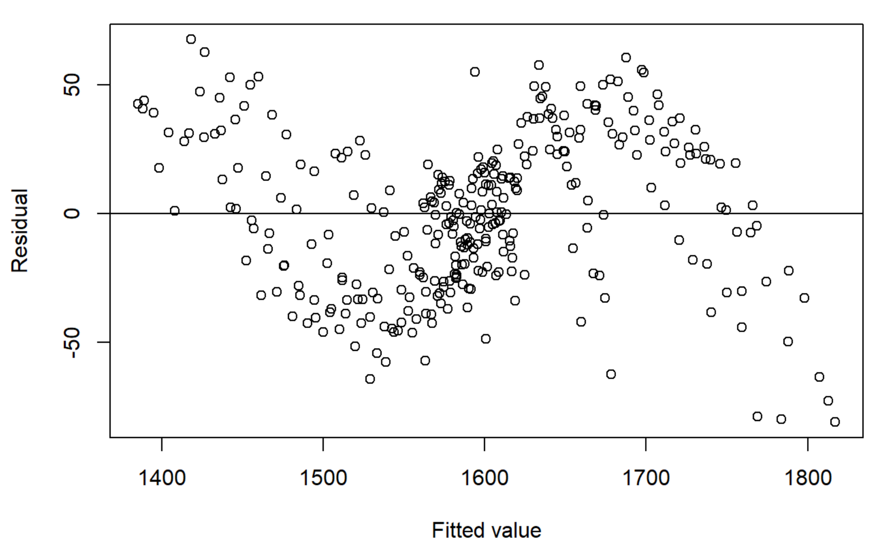
```r
lm_galaxy <- lm(velocity ~ ., data = galaxy)
summary(lm_galaxy)
```

```
##
## 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
```

```r
vif(lm_galaxy)
```

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

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



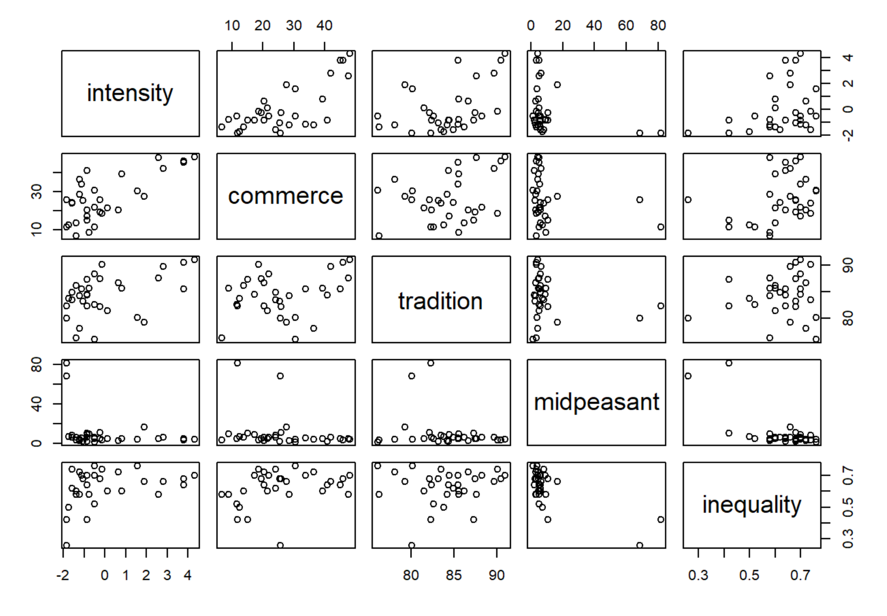

```
# Based on the model summary alone, the model fits nicely: all predictors are significant on the 0.05 level, the coefficient of determination is extremely good and the variance inflation factors stay quite well below 10 (although it would not harm to try dropping `radial.position` from the model to see if anything changes).

# However, the residual plot shows that there is clearly unaccounted non-linear structure in the response (one consequence of this is for example that the current model could behave extremely badly outside of the current range of x-values, although that is always risky). The next step in modelling could be to refit the model using interactions and transformations of the explanatory variables.
```

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.

```r
library(carData)

# a.
pairs(Chirot)
```



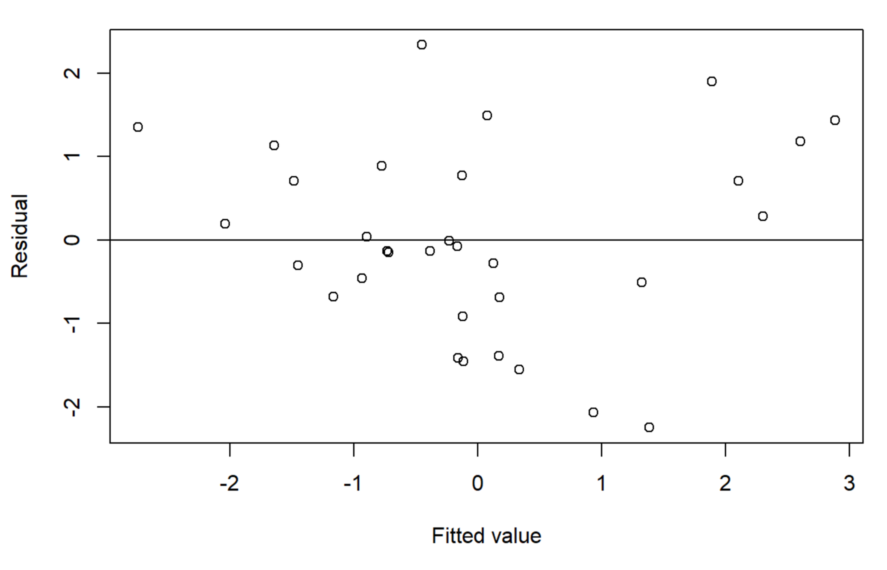

```
# Looks like there is positive correlation between intensity and all covariates except "midpeasant".
# The variable "midpeasant" has two outliers, and their influence on the model fit could be interesting to explore later (not done here).


# b.
lm_chirot <- lm(intensity ~ ., data = Chirot)


# c.
summary(lm_chirot)
```

```
##
## Call:
## lm(formula = intensity ~ ., data = Chirot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2460 -0.6781 -0.1013  0.8025  2.3378
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.919018   5.507499  -2.346 0.026587 *
## commerce      0.091140   0.020268   4.497 0.000118 ***
## tradition     0.116787   0.060688   1.924 0.064906 .
## midpeasant   -0.003342   0.017695  -0.189 0.851625
## inequality    1.137970   2.850304   0.399 0.692853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.227 on 27 degrees of freedom
## Multiple R-squared:  0.5836, Adjusted R-squared:  0.5219
## F-statistic: 9.462 on 4 and 27 DF,  p-value: 6.476e-05
```

```r
vif(lm_chirot)
```

```
##   commerce  tradition midpeasant inequality
##   1.201808   1.131518   1.948342   2.011226
```

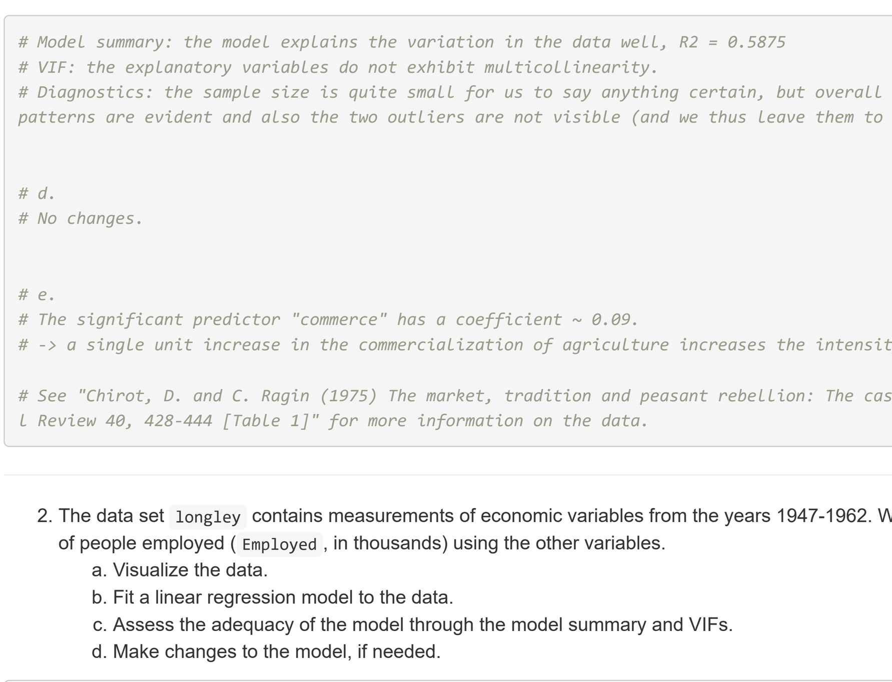
```r
plot(fitted(lm_chirot), resid(lm_chirot), xlab = "Fitted value", ylab = "Residual")
abline(h = 0)
```




```
# Model summary: the model explains the variation in the data well, R2 = 0.5875
# VIF: the explanatory variables do not exhibit multicollinearity.
# Diagnostics: the sample size is quite small for us to say anything certain, but overall the residuals look fine. No weird patterns are evident and also the two outliers are not visible (and we thus leave them to the model).


# d.
# No changes.


# e.
# The significant predictor "commerce" has a coefficient ~ 0.09.
# -> a single unit increase in the commercialization of agriculture increases the intensity of the rebellion by 0.09 units

# See "Chirot, D. and C. Ragin (1975) The market, tradition and peasant rebellion: The case of Romania. American Sociological Review 40, 428-444 [Table 1]" for more information on the data.
```

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.

```r
# a.
# The data is a time series so a possibly useful visualization could be:
plot.ts(longley)
```




```r
# b.
lm_longley <- lm(Employed ~ ., data = longley)


# c.
summary(lm_longley)
```

```
##
## Call:
## lm(formula = Employed ~ ., data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
## GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
## Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+00  4.555e-01   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3049 on 9 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

```r
vif(lm_longley)
```

```
## GNP.deflator          GNP   Unemployed Armed.Forces   Population
##    135.53244   1788.51348     33.61889      3.58893    399.15102
##         Year
##    758.98060
```

```r
# The model predicts the response extremely well (R2 ~ 1) but there is some significant multicollinearity between the predictors.


# d.
# We drop one-by-one the variables with the highest VIF, until no VIF is > 10, resulting into the model
lm_longley_2 <- lm(Employed ~ Unemployed + Armed.Forces + Population, data = longley)
summary(lm_longley_2)
```

```
##
## Call:
## lm(formula = Employed ~ Unemployed + Armed.Forces + Population,
##     data = longley)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3835 -0.2868 -0.1353  0.3596  1.3382
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.323091   4.211566  -0.314  0.75880
## Unemployed   -0.012292   0.003354  -3.665  0.00324 **
## Armed.Forces -0.001893   0.003516  -0.538  0.60019
## Population    0.605146   0.047617  12.709 2.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6843 on 12 degrees of freedom
## Multiple R-squared:  0.9696, Adjusted R-squared:  0.962
## F-statistic: 127.7 on 3 and 12 DF,  p-value: 2.272e-09
```

```
# The resulting model has still very high R2 = 0.9696.

# Diagnostics could now be checked and the model coefficients could be interpreted, although the latter is difficult due to the help file not properly defining the variables and their units...
```

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.

```r
# 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.