Methods and Data Analysis 2

Sebastián Soriano Pérez [ss1072]

9/12/2019

Question 1: OLD FAITHFUL

- Fit a regression of interval on duration and day (treated as a categorical/factor variable). Is there a significant difference in mean intervals for any of the days (compared to the first day)? Interpret the effects of controlling for the days (do so only for the days with significant effects, if any).

```
OldFaithful <- read.csv('OldFaithful.csv')
lm_OldFaithful <- lm(Interval ~ Duration + as.factor(Date), OldFaithful)
summary(lm_OldFaithful)
```

```
##
## Call:
## lm(formula = Interval ~ Duration + as.factor(Date), data = OldFaithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3886  -4.7332  -0.5622   3.9759  15.9639
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       32.8770     3.0672  10.719   <2e-16 ***
## Duration          10.8813     0.6622  16.431   <2e-16 ***
## as.factor(Date)2   1.3275     2.7173   0.489    0.626
## as.factor(Date)3   0.7825     2.6994   0.290    0.773
## as.factor(Date)4   0.1625     2.6461   0.061    0.951
## as.factor(Date)5   0.2463     2.6459   0.093    0.926
## as.factor(Date)6   1.9918     2.6580   0.749    0.455
## as.factor(Date)7  -0.1700     2.7020  -0.063    0.950
## as.factor(Date)8  -0.6944     2.6957  -0.258    0.797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.866 on 98 degrees of freedom
## Multiple R-squared:  0.7408, Adjusted R-squared:  0.7196
## F-statistic:    35 on 8 and 98 DF,  p-value: < 2.2e-16
```

```
confint(lm_OldFaithful, level = 0.95)
```

```
##                      2.5 %    97.5 %
## (Intercept)      26.790153 38.963809
## Duration          9.567135 12.195545
## as.factor(Date)2 -4.064928  6.719930
## as.factor(Date)3 -4.574299  6.139349
## as.factor(Date)4 -5.088529  5.413622
## as.factor(Date)5 -5.004338  5.496977
## as.factor(Date)6 -3.282858  7.266529
## as.factor(Date)7 -5.532115  5.192078
## as.factor(Date)8 -6.043893  4.655112
```

From the model summary, we can observe that all of the categories of Date have a p-value greater than 0.455 and t-values that are close to 0. This means that for all of them, we cannot reject the null hypothesis that they are equal to zero and they do not seem to be useful predictors in the model. Compared to the first day, none of the days seem to have any impact on the response variable.

- Perform an F-test to compare this model to your model for this data from the last homework. In context of the question, what can you conclude from the results of the F-test?

```
lm_OldFaithful_old <- lm(Interval ~ Duration, OldFaithful)
anova(lm_OldFaithful, lm_OldFaithful_old)
```

```
## Analysis of Variance Table
##
## Model 1: Interval ~ Duration + as.factor(Date)
## Model 2: Interval ~ Duration
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     98 4620.2
## 2    105 4689.0 -7   -68.853 0.2086 0.9828
```

With a p-value of 0.9828 we cannot reject the null hypothesis that the RSS values for both models are equal to each other. Therefor we cannot determine that one of the models is more accurate than the other, although the new model that takes Date into account does have a slightly smaller RSS.

- Using k-fold cross validation (with k=10), compare the average RMSE for this model and the average RMSE for your model from the last homework. Which model appears to have higher predictive accuracy based on the average RMSE values?

```
set.seed(1) # Seed to ensure the results are reproducible
OldFaithful <- OldFaithful[sample(nrow(OldFaithful)),] # Randomly re-shuffle the data
K <- 10 # Define the number of folds
# Define a matrix to save the RSMEs
RSME <- matrix(0, nrow = K, ncol = 1) # Current model
RSME_old <- matrix(0, nrow = K, ncol = 1) # Old model
# Split the row indexes into k equal parts
kth_fold <- cut(seq(1, nrow(OldFaithful)), breaks = K, labels = FALSE)
for(k in 1:K){ #k-fold cross validation for-loop
  test_index <- which(kth_fold == k)
  train <- OldFaithful[-test_index,]
  test <- OldFaithful[test_index,]
  lm_train <- lm(Interval ~ Duration + as.factor(Date), train)
  lm_train_old <- lm(Interval ~ Duration, train)
  predicted_test_values <- predict(lm_train, test)
  predicted_test_values_old <- predict(lm_train_old, test)
  RSME[k,] <- sqrt(mean((test$Interval - predicted_test_values) ^ 2))
  RSME_old[k,] <- sqrt(mean((test$Interval - predicted_test_values_old) ^ 2))
}
mean(RSME) # Current model RSME
```

```
## [1] 6.977411
```

```
mean(RSME_old) # Old model RSME
```

```
## [1] 6.561417
```

The old model has a better predictive value since its RSME is lower.

Question 2: MATERNAL SMOKING AND BIRTH WEIGHTS

- **Summary**

By analyzing the data on 869 newborn male babies and their families, a model was created with forward selection using AIC to interpret and associate the variables that were found to be significant with the response variable of weight:

$$\hat{bwt.oz_i} = \hat{\beta}_0 + \hat{\beta}_1 smoke_i + \hat{\beta}_2 mht_i + \hat{\beta}_3 (mrace_{black})_i + \hat{\beta}_4 (mrace_{mexican})_i + \hat{\beta}_5 (mrace_{mix})_i$$
$$+\hat{\beta}_6 (mrace_{white})_i + \hat{\beta}_7 mpregwt_i + \hat{\beta}_8 parity_i$$

The model estimates that these variables are significant when trying to create a model for interpretation and association of the data. The specific coefficient values can be found in the "Model" section.

However, the final model's adjusted R-squared value of 0.1435 indicates that the model only explains a small proportion of the weight differences among newborn male babies with the variables that were considered in the final dataset.

- **Introduction**

This document presents a model to interpret the impact of several variables on a newborn's weight at birth. A dataset was analyzed considering the available data in order to find the best model to explain the association between the predictive variables and the response variable through an initial exploratory data analysis (EDA), and later with a forward selection in R. The main focus of this document is to find whether or not smoking during pregnancy had an impact in the child's weight.

- **Data**

The Child Health and Development Studies research was one of the first to collect data to understand and quantify the risk of smoking during pregnancy to the baby's health. The data was collected from 1960 to 1967, and a subset of that data is being analyzed in this document (the variables related to the father's information are neglected for this analysis). 869 cases of newborn male babies who lived at least 28 days are being analyzed (data set smoking.csv). The purpose of this document is to present a statistical model to interpret and understand the correlation between several variables and the baby's birth weight. The variables being considered for building the model, in association to the response variable of the baby's birth weight in ounces (bwt.oz), are the following:
- Total number of mother's previous pregnancies (parity)
- Mother's race or ethnicity (mrace)
- Mother's age in years at pregnancy termination (mage)
- Mother's education level (med)
- Mother's height in inches (mht)
- Mother's pre-pregnancy weight in pounds (mpregwt)
- Family yearly income in 2500-increment categories (inc)
- Indicator for the mother's smoking (smoke)

A summary of the data variables being analyzed can be found in Annex 2.1. A table of the correlation between all variables and plots for their interactions can be found in Annex 2.2.

- **Model**

Various methods for model selection were tested and interactions between predictors were considered as part of the full model. Ultimately, a forward selection with AIC model and a forward selection with BIC model were compared and the one with the higher adjusted R-squared value was selected as the final model. Using a forward selection approach with AIC as a selection criterion, the following model was obtained:

```
n <- nrow(smoking)
null_model <- lm(bwt.oz ~ 1, data = smoking)
full_model <- lm(bwt.oz ~ parity + mrace + mage + med + mht + mpregwt + inc + smoke
                 + parity:mage + parity:mpregwt + mage:mpregwt +mht:mpregwt,
                 data = smoking)
model_forward <- step(null_model, scope = formula(full_model),
                      direction = 'forward', trace = 0)
#model_forward$call
#model_forward2 <- step(null_model, scope = formula(full_model),
                       #direction = 'forward', trace = 0, k = log(n))
#model_forward2$call
summary(model_forward)
```

```
##
## Call:
## lm(formula = bwt.oz ~ smoke + mht + mrace + mpregwt + parity,
##     data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.969  -9.525  -0.336  10.131  50.206
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.70824   15.08992   2.764 0.005832 **
## smoke1       -9.35194    1.15218  -8.117 1.65e-15 ***
## mht           0.93387    0.26070   3.582 0.000360 ***
## mraceblack   -0.88802    3.28129  -0.271 0.786740
## mracemexican 11.23603    4.41828   2.543 0.011162 *
## mracemix      5.95467    5.21943   1.141 0.254244
## mracewhite    7.93888    3.03506   2.616 0.009060 **
## mpregwt       0.10808    0.03217   3.360 0.000814 ***
## parity        0.66507    0.31422   2.117 0.034584 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.71 on 860 degrees of freedom
## Multiple R-squared:  0.1514, Adjusted R-squared:  0.1435
## F-statistic: 19.17 on 8 and 860 DF,  p-value: < 2.2e-16
```

```
#summary(model_forward2)
confint(model_forward, level = 0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept)   12.09085636 71.3256207
## smoke1       -11.61335023 -7.0905277
## mht            0.42219278  1.4455490
## mraceblack    -7.32828642  5.5522492
## mracemexican   2.56416498 19.9079017
## mracemix      -4.28964952 16.1989888
```

```
## mracewhite     1.98189867 13.8958664
## mpregwt        0.04494574  0.1712097
## parity         0.04833843  1.2818088
```

$$\hat{bwt.oz_i} = \hat{\beta}_0 + \hat{\beta}_1 smoke_i + \hat{\beta}_2 mht_i + \hat{\beta}_3 (mrace_{black})_i + \hat{\beta}_4 (mrace_{mexican})_i + \hat{\beta}_5 (mrace_{mix})_i$$
$$+\hat{\beta}_6 (mrace_{white})_i + \hat{\beta}_7 mpregwt_i + \hat{\beta}_8 parity_i$$

Where $(mrace_{black})_i + (mrace_{mexican})_i + (mrace_{mix})_i + (mrace_{white})_i \in \{0, 1\}$, $(mrace_{black})_i \in \{0, 1\}; (mrace_{mexican})_i \in \{0, 1\}; (mrace_{mix})_i \in \{0, 1\}; (mrace_{white})_i \in \{0, 1\}$, $\hat{\beta}_0 = 41.70824$, $\hat{\beta}_1 = -9.35194$, $\hat{\beta}_2 = 0.93387$, $\hat{\beta}_3 = -0.88802$, $\hat{\beta}_4 = 11.23603$, $\hat{\beta}_5 = 5.95467$, $\hat{\beta}_6 = 7.93888$, $\hat{\beta}_7 = 0.10808$, and $\hat{\beta}_8 = 0.66507$.

The most significant variable is smoke with a p-value of -8.117. The model has a residual standard error of 16.71 on 860 degrees of freedom. It also has a low adjusted R-squared value of 0.1435, which indicates it does not explain most of the available data.

The model assumptions of linearity, independence and equal variance seem to be met. The normality assumptions may be a reason to worry about since the Q-Q shows tails on both sides. See Annex 2.4 to Annex 2.6.

- **Conclusions and Remarks**

It can be concluded that mothers who smoke do tend to give birth to babies with lower weights than mothers who do not smoke. The intercept value indicates that a non smoker mother, of Asian race, income category 1, with values of 0 for the rest of the variables, would give birth to male babies of 41.70824 ounces on average. Mothers who smoke would give birth to babies 9.35194 ounces lighter on average as indicated by $\hat{\beta}_1$. A likely range for the difference in birth weights for smokers and non-smokers would be the one indicated by the smoke1 confidence interval of (-11.61335023, -7.0905277) ounces with 95% confidence.

For the full model in the forward selection process the interaction between smoke and race were taken into account. However, this interaction was not statistically significant and the null hypotheses that their coefficients were equal to zero could not be rejected. However, when looking at the model itself, it can be noted that the interaction between smoke and mracemexican could be significant. There is no evidence in the data to support the claim that smoking and birth weight differ among the other races. Other than the variables of smoke, race, mpregwt and parity, no other variables or their interactions showed strong associations to the response variable.

```
test_model <- lm(formula = bwt.oz ~ smoke + mht + mrace + mpregwt + parity
                 + smoke:mrace,
    data = smoking)
summary(test_model)
```

```
##
## Call:
## lm(formula = bwt.oz ~ smoke + mht + mrace + mpregwt + parity +
##     smoke:mrace, data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.195  -9.584  -0.608  10.302  50.927
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           40.48977   15.21526   2.661 0.007934 **
## smoke1               -17.03758    6.50245  -2.620 0.008944 **
## mht                    0.98621    0.26198   3.764 0.000178 ***
## mraceblack            -3.78385    3.91069  -0.968 0.333534
## mracemexican           6.11520    5.09443   1.200 0.230328
## mracemix               6.39163    5.89566   1.084 0.278615
## mracewhite             5.90457    3.54251   1.667 0.095925 .
## mpregwt                0.10848    0.03216   3.373 0.000776 ***
## parity                 0.65441    0.31618   2.070 0.038776 *
## smoke1:mraceblack      9.34487    7.00948   1.333 0.182829
## smoke1:mracemexican   20.49765   10.19634   2.010 0.044713 *
## smoke1:mracemix       -5.15450   12.58528  -0.410 0.682227
## smoke1:mracewhite      7.39133    6.63312   1.114 0.265460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.69 on 856 degrees of freedom
## Multiple R-squared:  0.1572, Adjusted R-squared:  0.1454
## F-statistic:  13.3 on 12 and 856 DF,  p-value: < 2.2e-16
```

For the rest of the variables: each additional inch in mht adds an estimated 0.93387 ounces to bwt.oz; being black, Mexican, mix, or white insterad of Asian adds either -0.88802, 11.23603, 5.95467, or 7.93888 ounces to bwt.oz, respectively; each additional pound in mpregwt adds an extra 0.10808 ounces to bwt.oz; finally, each additional previous birth in parity adds an extra 0.66507 ounces to bwt.oz. Of all these associations, all are significant but mraceblack and mracemix, for which we cannot reject the null hypotheses that they are equal to zero and therefore they are most likely not statistically significant. This indicates that mothers of black or mixed race do not have any impact on their male babies birth weight because of their race.

Although this analysis suggests some of the predictors do have a very significant impact on the birth weight, the low R-squared value and the lack of a strong correlation between the predictor variables and the response variable suggest that there may be other factors that influence and predict birth weight much better. Maybe other biological or genetic factors have a stronger impact on the birth weight, or maybe a new model taking into account the father's information would provide better results.

- **Annex**

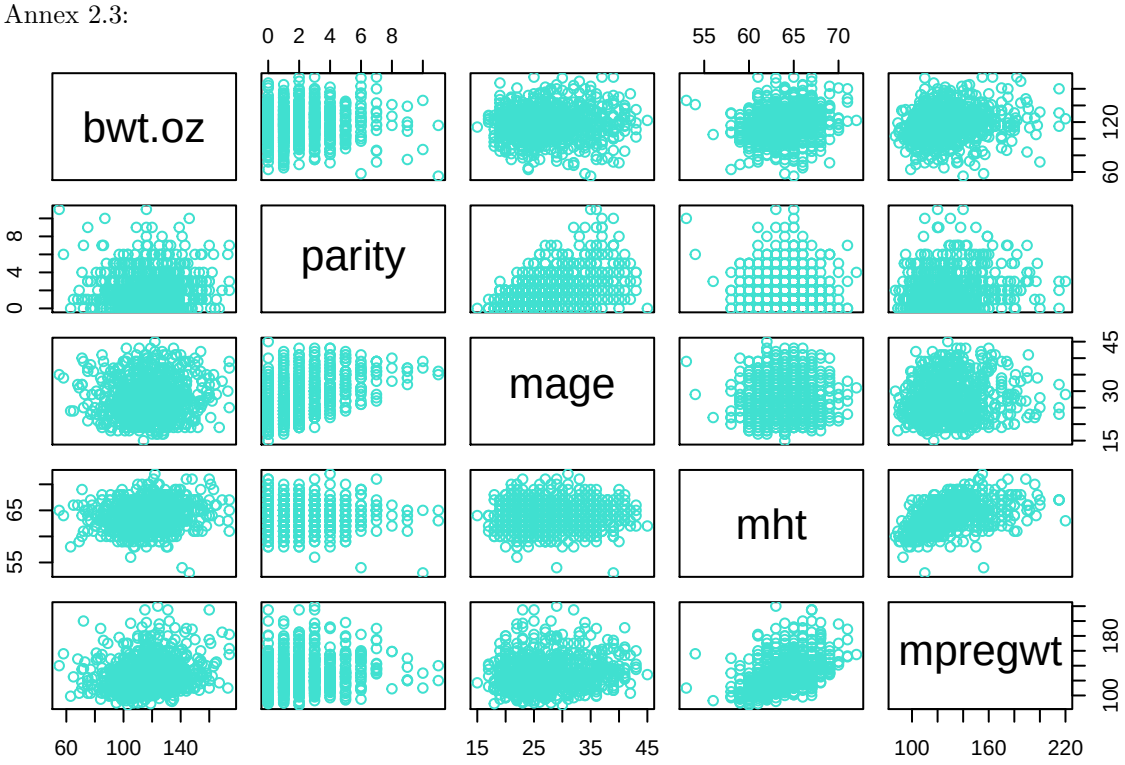
Annex 2.1:

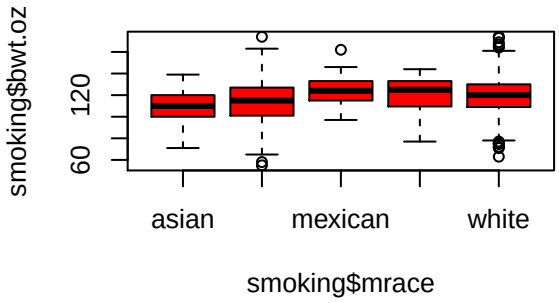
```
##      bwt.oz          parity           mrace          mage        med
##  Min.   : 55.0   Min.   : 0.000   asian  : 34   Min.   :15.00   0:  5
##  1st Qu.:108.0   1st Qu.: 1.000   black  :169   1st Qu.:23.00   1:130
##  Median :119.0   Median : 2.000   mexican: 25   Median :26.00   2:321
##  Mean   :118.4   Mean   : 1.953   mix    : 15   Mean   :27.29   3: 47
##  3rd Qu.:129.0   3rd Qu.: 3.000   white  :626   3rd Qu.:31.00   4:203
##  Max.   :174.0   Max.   :11.000                 Max.   :45.00   5:159
##                                                                 7:  4
##       mht           mpregwt          inc      smoke
##  Min.   :53.00   Min.   : 87.0   1      :153   0:466
##  1st Qu.:62.00   1st Qu.:113.0   2      :146   1:403
##  Median :64.00   Median :125.0   3      :136
##  Mean   :64.07   Mean   :128.5   7      :111
##  3rd Qu.:66.00   3rd Qu.:140.0   4      :105
##  Max.   :72.00   Max.   :220.0   5      : 98
##                                  (Other):120
```

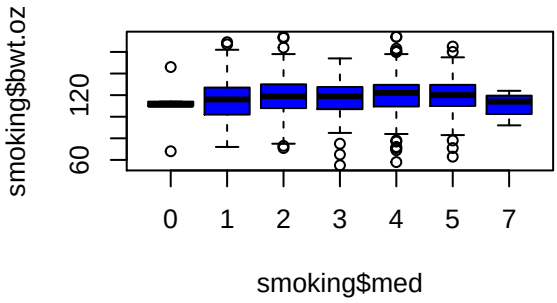
Annex 2.2

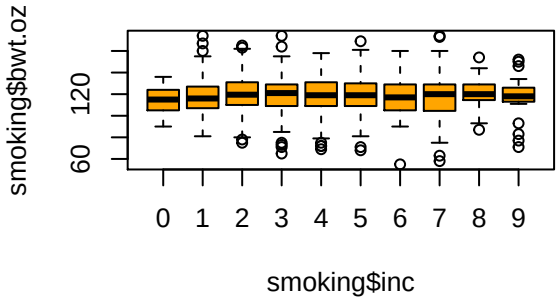
```
##              bwt.oz      parity         mage          mht   mpregwt
## bwt.oz   1.00000000  0.04106952  0.044343562  0.187758035 0.1821158
## parity   0.04106952  1.00000000  0.523690421 -0.042815618 0.1505379
## mage     0.04434356  0.52369042  1.000000000 -0.005470885 0.1461368
## mht      0.18775804 -0.04281562 -0.005470885  1.000000000 0.4604463
## mpregwt 0.18211578  0.15053789  0.146136818  0.460446304 1.0000000
```

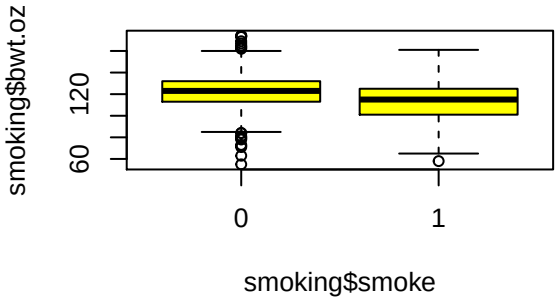
Annex 2.3:



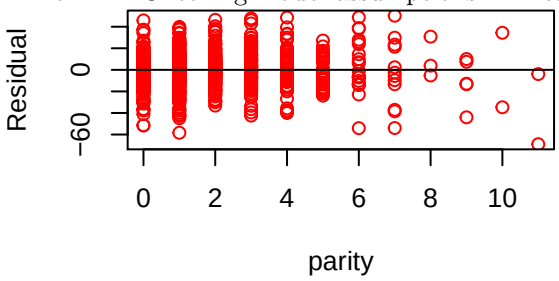


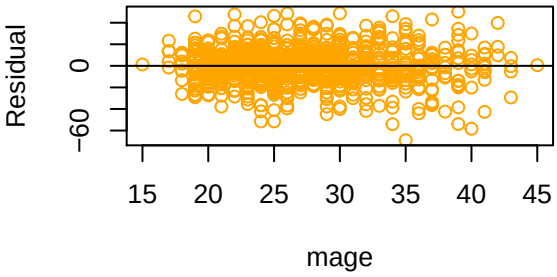


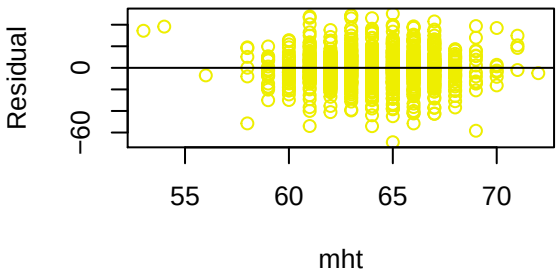


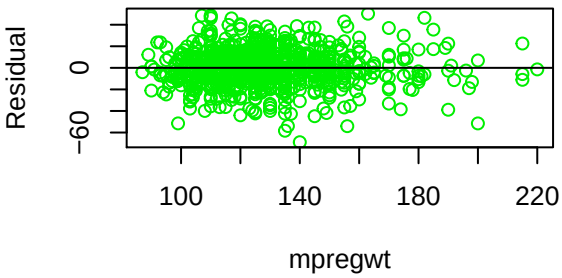


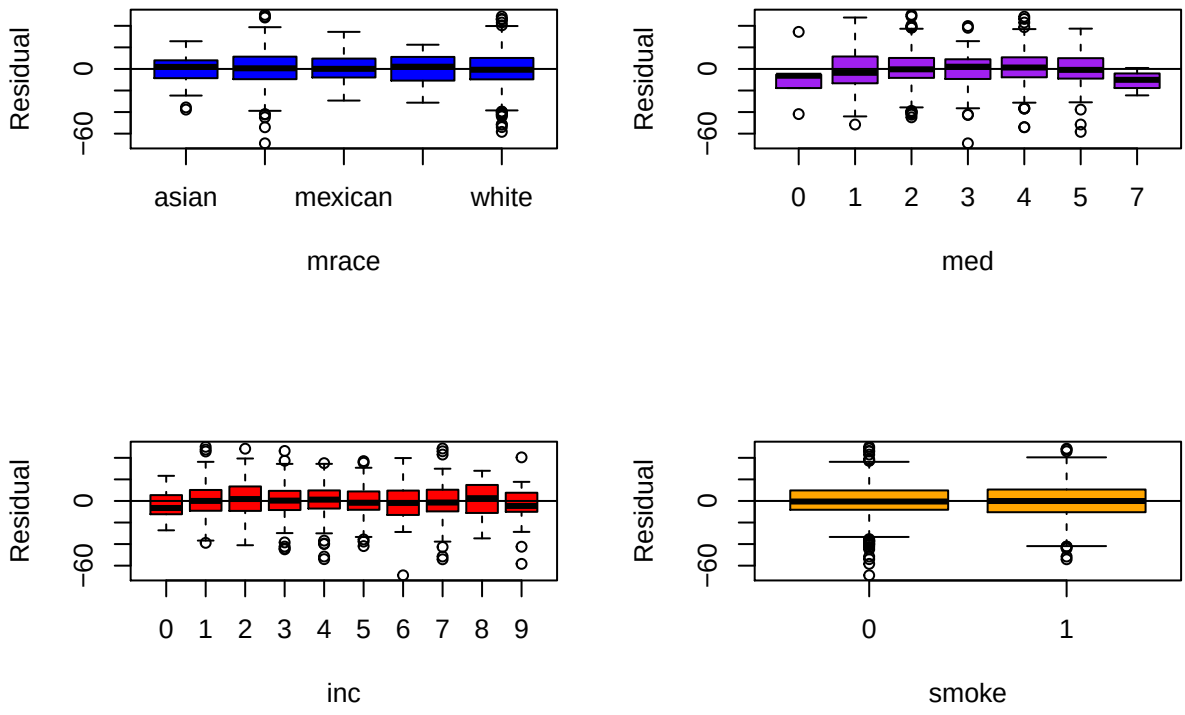
Annex 2.4: Checking model assumptions - Linearity



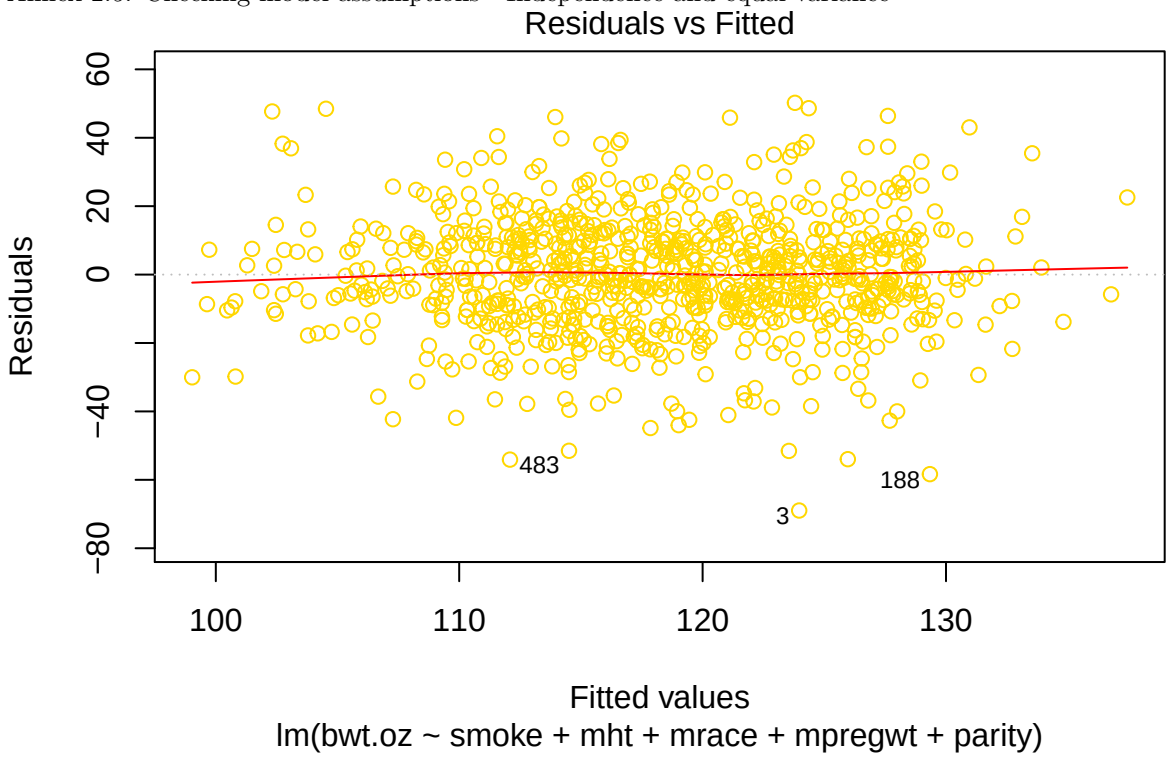




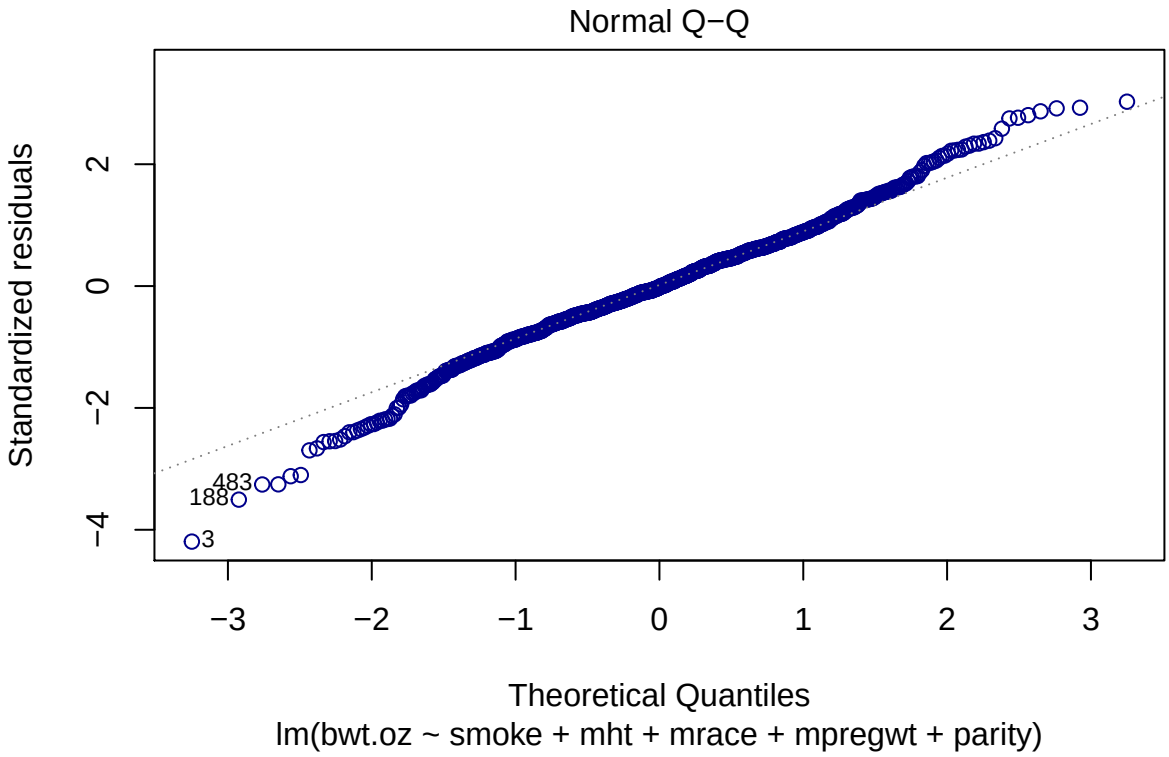




Annex 2.5: Checking model assumptions - Independence and equal variance



Annex 2.6: Checking model assumptions - Normality


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(bwt.oz ~ smoke + mht + mrace + mpregwt + parity)
483
188
3
−3
−2
−1
0
1
2
3
−4
−2
0
2