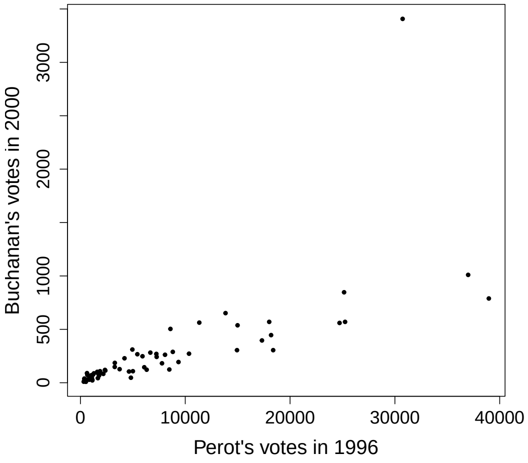
University of Edinburgh
School of Mathematics
Applied Statistics (MATH10096)

Assignment: sketch of the solutions

1. We can start by looking at the scatterplot of the data.

```
elections <- read.table("BushGore.dat", header = TRUE)
plot(elections$Perot, elections$Buchanan, pch = 20,
     xlab = "Perot's votes in 1996", ylab = "Buchanan's votes in 2000",
     cex.lab = 1.6, cex.axis = 1.4)
```



We can clearly notice an outlying observation. Let us verify that this observation indeed corresponds to Palm Beach county.

```
which(elections$Buchanan > 1500)

## [1] 50

elections[50,]

##       County Perot Buchanan
## 50 PalmBeach 30739     3407
```

To assess the potential impact of Palm Beach county, we will fit a simple linear regression model, using Perot's cotes in 1996 to predict Buchanan's votes in 2000, both with and without this observation. In both cases the model is of the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n,$$

where y_i denotes the ith county votes for Buchanan in 2000 and x_i denotes the ith county votes for Perot in 1996. The total number of counties n is 67 (66 when excluding Palm Beach). Let's begin by fitting the model to the entire dataset, which includes Palm Beach.

```r
fit_elections_1 <- lm(Buchanan ~ Perot, data = elections)
summary(fit_elections_1)
```

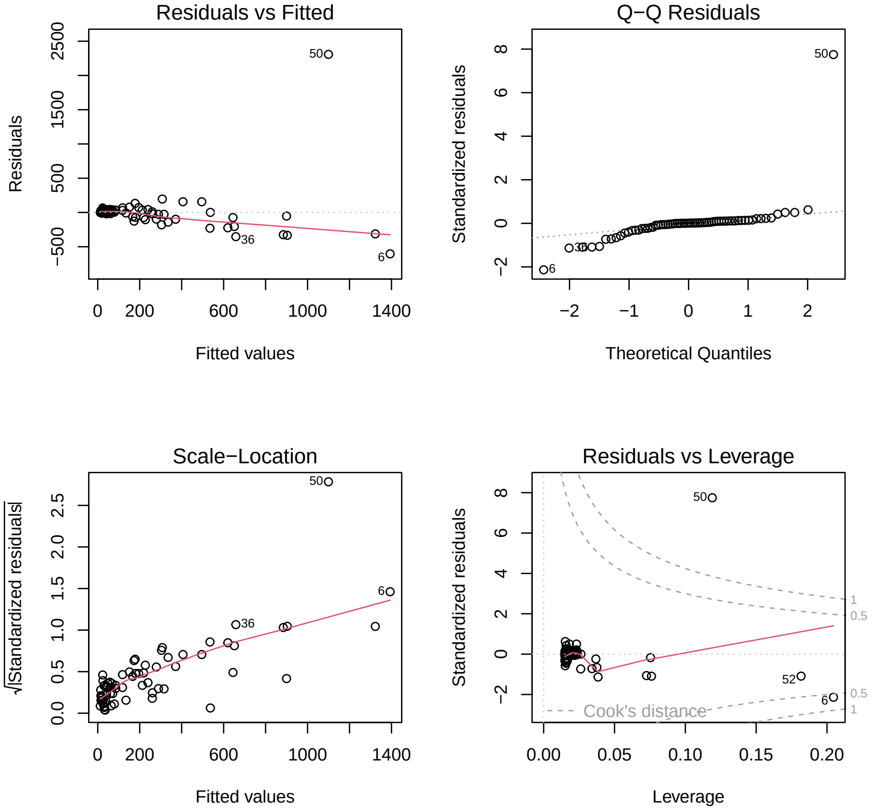
```
##
## Call:
## lm(formula = Buchanan ~ Perot, data = elections)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -604.44  -66.82    2.53   33.45 2307.47
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.081433  49.899462   0.022    0.983
## Perot        0.035735   0.004352   8.211 1.23e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 317.3 on 65 degrees of freedom
## Multiple R-squared:  0.5091,Adjusted R-squared:  0.5016
## F-statistic: 67.41 on 1 and 65 DF,  p-value: 1.234e-11
```

```r
confint(fit_elections_1)
```

```
##                    2.5 %       97.5 %
## (Intercept) -98.57467491 100.73754044
## Perot         0.02704248   0.04442673
```

We have that $\hat{\beta}_0 = 1.081$ (95% CI: (-98.575, 100.738)) and $\hat{\beta}_1 = 0.036$ $(0.027, 0.044)$ and $r^2 = 0.51$. The result shows that 51% of the variation of Buchanan's 2000 votes can be explained by Perot's 1996 votes. This value appears relatively low given that we are predicting votes for a candidate from the same party using the previous election result. Let us look now at the residual plots.

```r
par(mfrow = c(2, 2))
plot(fit_elections_1)
```



```
cooks.distance(fit_elections_1)[50]

##        50
## 4.055556
```

As expected, Palm Beach has an extremely high Cook's distance and so this observation has to have substantial influence on the model results. We should nonetheless check this and the extent of the impact. We also note from the residual plots that there is a strong mean variance relationship. Let us now fit the model without the observation corresponding to Palm Beach county.

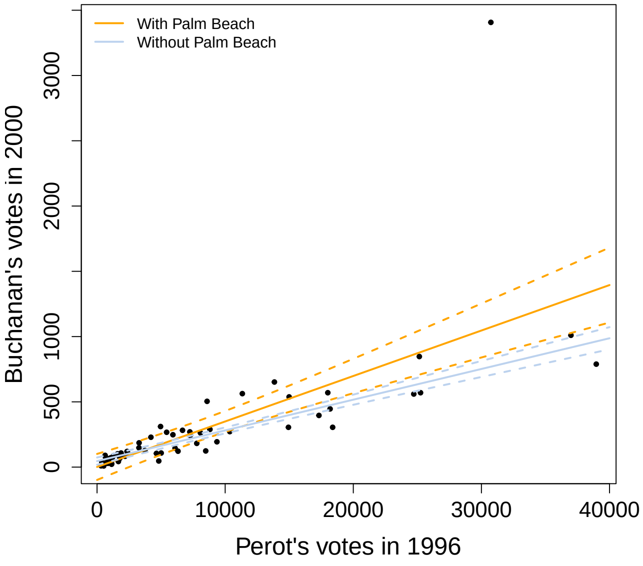
```
elections_pb <- elections[-50,]
fit_elections_2 <- lm(Buchanan ˜ Perot, data = elections_pb)
summary(fit_elections_2)

##
## Call:
## lm(formula = Buchanan ˜ Perot, data = elections_pb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197.38  -45.49  -18.64   22.84  272.08
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.689948  13.981299   3.268  0.00174 **
## Perot        0.024143   0.001281  18.847  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.31 on 64 degrees of freedom
## Multiple R-squared:  0.8473,Adjusted R-squared:  0.8449
## F-statistic: 355.2 on 1 and 64 DF,  p-value: < 2.2e-16
```

```r
confint(fit_elections_2)
```

```
##                  2.5 %      97.5 %
## (Intercept) 17.75909304 73.62080386
## Perot        0.02158345  0.02670163
```

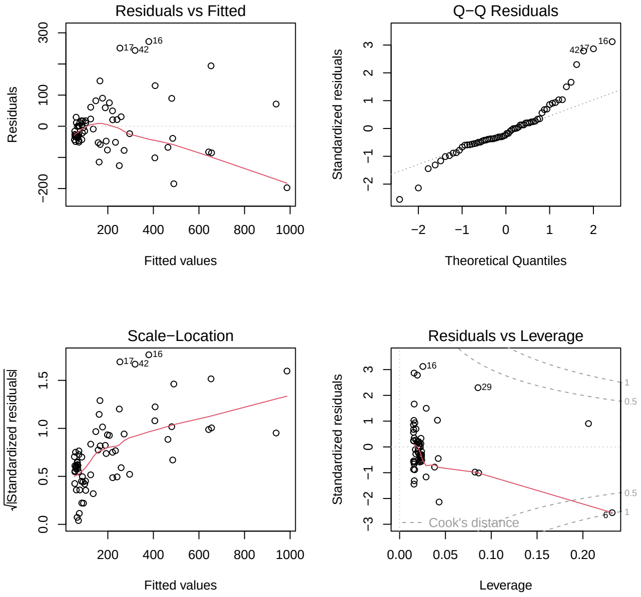
We immediately notice that the r^2 value has increased substantially, from 0.51 to 0.85. We further have that now $\hat{\beta}_0 = 45.690\ (17.759, 73.621)$ and $\hat{\beta}_1 = 0.024\ (0.022, 0.027)$. The impact of the changes in the intercept and slope is better understood through visualisation of the regression line.

```r
df_pred <- data.frame("Perot" = seq(10, 39000, len = 1000))
preds_1 <- predict(fit_elections_1, newdata = df_pred,
                   interval = "confidence")
preds_2 <- predict(fit_elections_2, newdata = df_pred,
                   interval = "confidence")

plot(elections$Perot, elections$Buchanan, pch = 20,
     xlab = "Perot's votes in 1996", ylab = "Buchanan's votes in 2000",
     cex.lab = 1.6, cex.axis = 1.4)

lines(seq(0, 40000, len = 1000), preds_1[, 1],
      col = "orange", lwd = 2)
lines(seq(0, 40000, len = 1000), preds_1[, 2], lty = 2,
      col = "orange", lwd = 2)
lines(seq(0, 40000, len = 1000), preds_1[, 3], lty = 2,
      col = "orange", lwd = 2)

lines(seq(0, 40000, len = 1000), preds_2[, 1],
      col = "lightsteelblue2", lwd = 2)
lines(seq(0, 40000, len = 1000), preds_2[, 2], lty = 2,
      col = "lightsteelblue2", lwd = 2)
lines(seq(0, 40000, len = 1000), preds_2[, 3], lty = 2,
      col = "lightsteelblue2", lwd = 2)

legend("topleft", legend = c("With Palm Beach", "Without Palm Beach"),
       col = c( "orange","lightsteelblue2"), lwd = c(2, 2), bty = "n")
```



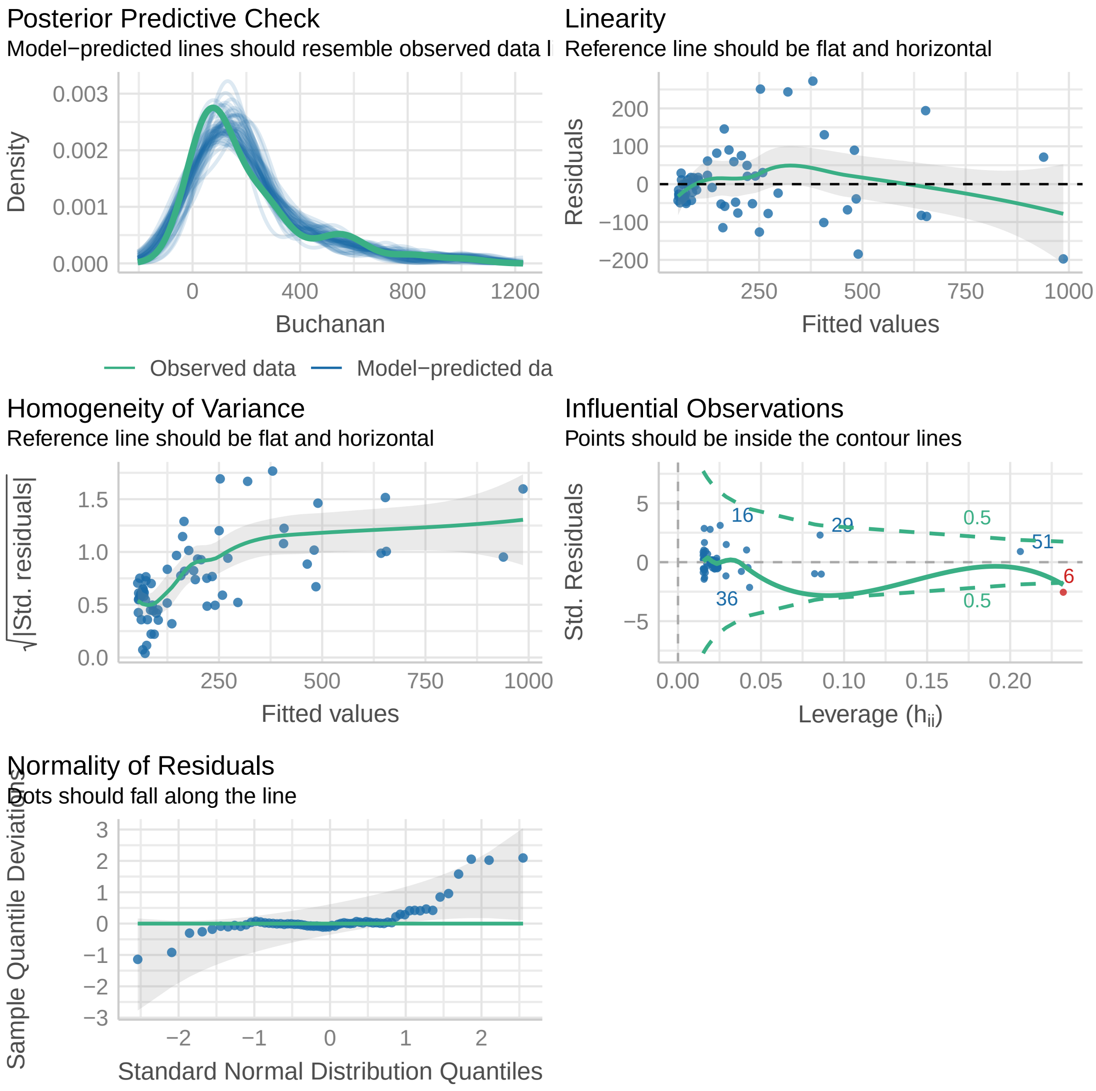
We find that the regression line is influenced by Palm Beach: removing it shifts the regression line considerably. The new regression line fits the remaining observations better. The exclusion of Palm Beach also leads to a great reduction in the width of the confidence intervals around the (conditional) mean. We also observe that there is very little overlap between the two confidence bands. Let us check the residual plots for the model without Palm Beach.

```r
par(mfrow = c(2, 2))
plot(fit_elections_2)
```



We still observe a mean-variance relationship, indicating that the prediction intervals included in the above scatterplot, as well as the confidence intervals for the slope and intercept, may not be entirely accurate. Additionally, there appears to be a trend in the mean of the residuals. However, we must note that the lowess estimate (red line), being a nonparametric estimate, is less reliable in areas with sparse data. I recently discovered the package `performance` which quantifies the uncertainty around the nonparametric estimates of the trend in both plots (fitted values against residuals and fitted values against $\sqrt{|\text{Std residuals}|}$). As we can see in the plot below, a horizontal trend would fall entirely within the uncertainty band surrounding the nonparametric estimate of the trend in the plot of fitted values versus residuals. The same is not true for the trend in the plot of fitted values against $\sqrt{|\text{Std residuals}|}$. A fix that could work (but I have not tried!) is to either transform the response (e.g., using the square root transformation) or to fit a GLM with a Poisson distribution, as our response is the number of votes Buchanan received. This latter approach would allow the variance to be proportional to the mean.

```
library(performance)
check_model(fit_elections_2)
```



Finally, using the model adjusted without Palma Beach, let us use Perot's votes in 1996 in that county to predict Buchanan's votes.

```
predict(fit_elections_2,
        newdata =  data.frame("Perot" = elections[50,]$Perot),
        interval = "prediction")
```

```
##        fit      lwr      upr
## 1 787.8075 599.8505 975.7644

elections[50, ]$Buchanan

## [1] 3407
```

The number of votes for Buchanan in Palm Beach, 3407, falls way outside the prediction interval $(600, 976)$. All of this is evidence that the observation from Palm Beach county has an impact on the results of the regression model.

2. (a) I start by inspecting the nature of the variables.

```
library(faraway)
str(teengamb)

## 'data.frame': 47 obs. of  5 variables:
##  $ sex   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ status: int  51 28 37 28 65 61 28 27 43 18 ...
##  $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
##  $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
##  $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

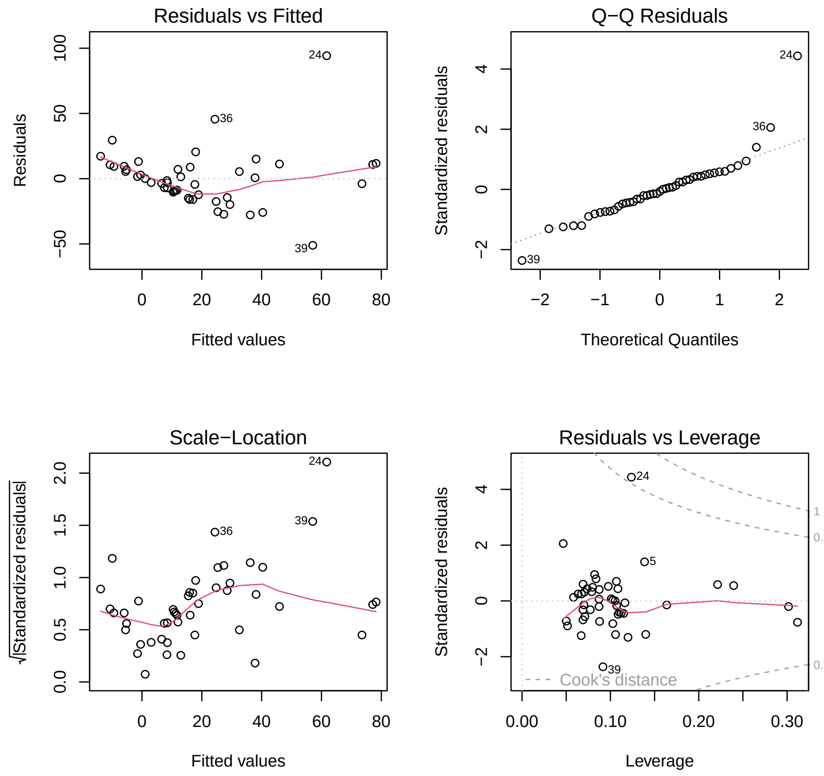
The variable sex is not coded as a factor, and although this does not make any difference here (as it only has two levels), it is good practice to code factor variables as such.

```
teengamb$sex <- as.factor(teengamb$sex)
```

Let us now fit the model with the expenditure on gambling as the response and sex, socioeconomic status, income, and verbal score as predictors. We will then inspect the residual plots.

```
fit_gamble <- lm(gamble ~ sex + status + income + verbal,
                 data = teengamb)
summary(fit_gamble)

##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex1        -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267,Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

```r
par(mfrow = c(2, 2))
plot(fit_gamble)
```

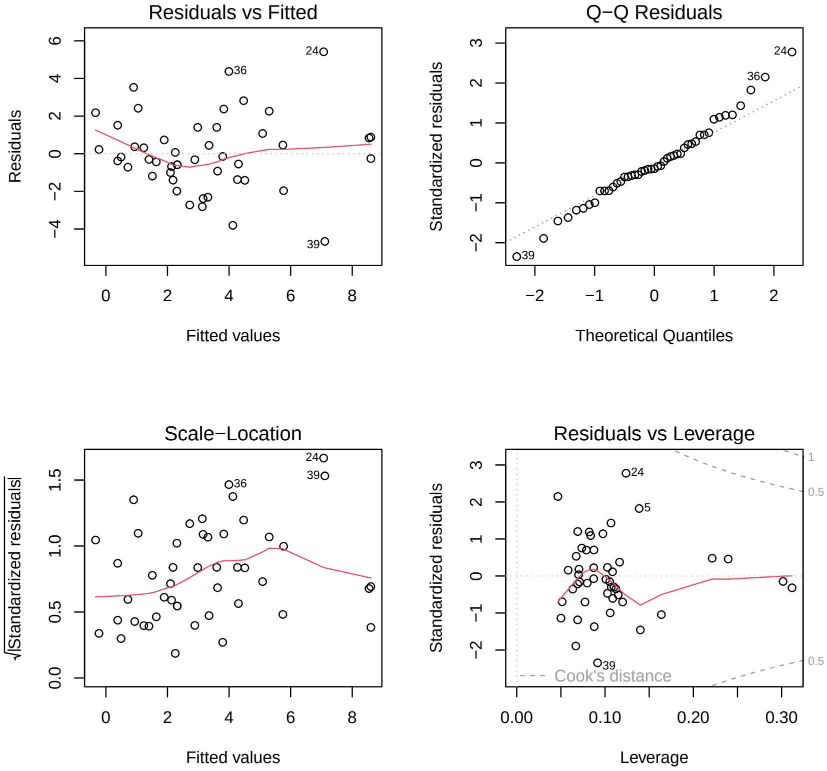


The variance of the residuals seems to increase with the fitted values showing a sign of non constant variance. Observation 24 has also a large standardized residual and a borderline Cook's distance value. Let us now fit the model with a transformed response variable, i.e., with the square root of the expenditure on gambling.

```r
fit_gamble_sqrt <- lm(sqrt(gamble) ~ sex + status + income + verbal,
                      data = teengamb)
summary(fit_gamble_sqrt)
```

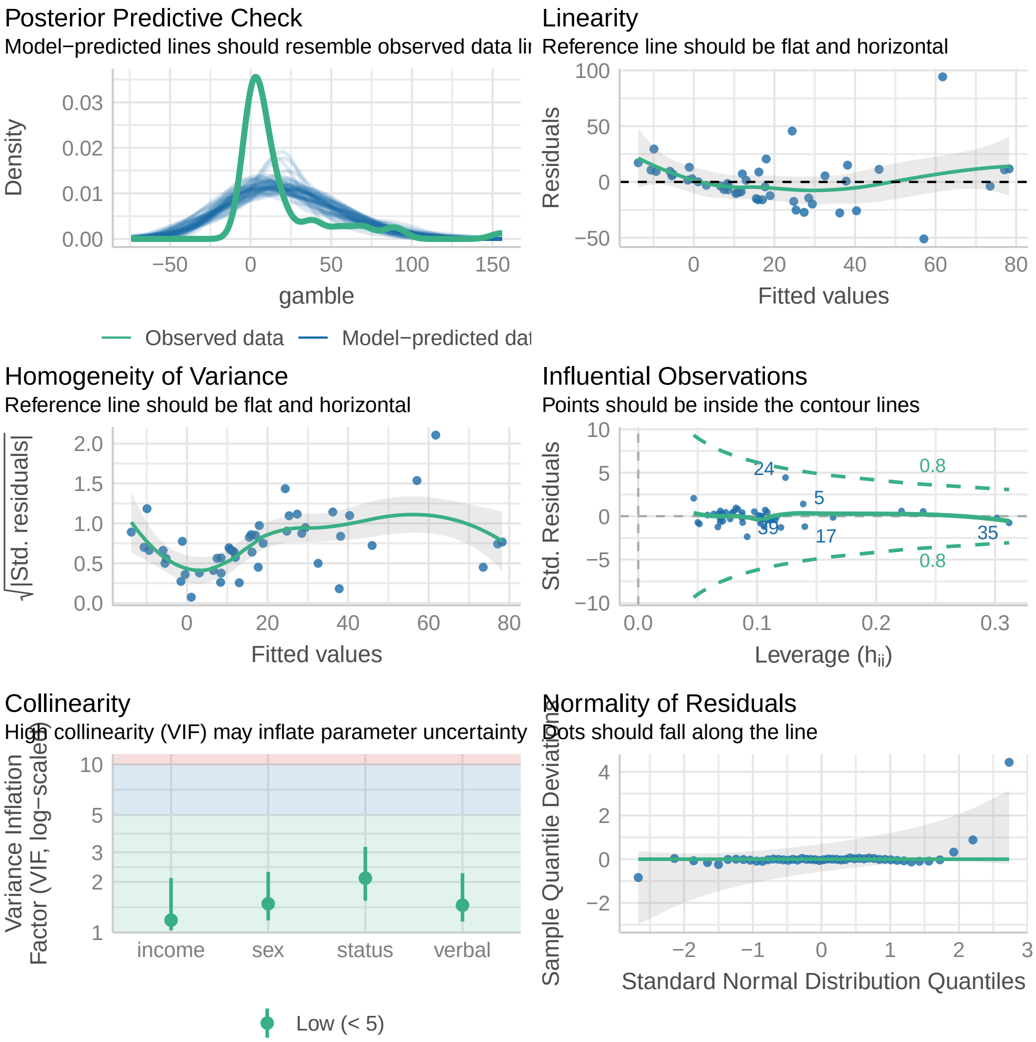
```
##
## Call:
## lm(formula = sqrt(gamble) ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6606 -1.0961 -0.2564  0.9786  5.4178
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.97707    1.57947   1.885  0.06638 .
## sex1        -2.04450    0.75416  -2.711  0.00968 **
## status       0.03688    0.02582   1.428  0.16057
## income       0.47938    0.09418   5.090 7.94e-06 ***
## verbal      -0.42360    0.19950  -2.123  0.03967 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 42 degrees of freedom
## Multiple R-squared:  0.5646,Adjusted R-squared:  0.5231
## F-statistic: 13.61 on 4 and 42 DF,  p-value: 3.362e-07
```

```
par(mfrow = c(2, 2))
plot(fit_gamble_sqrt)
```

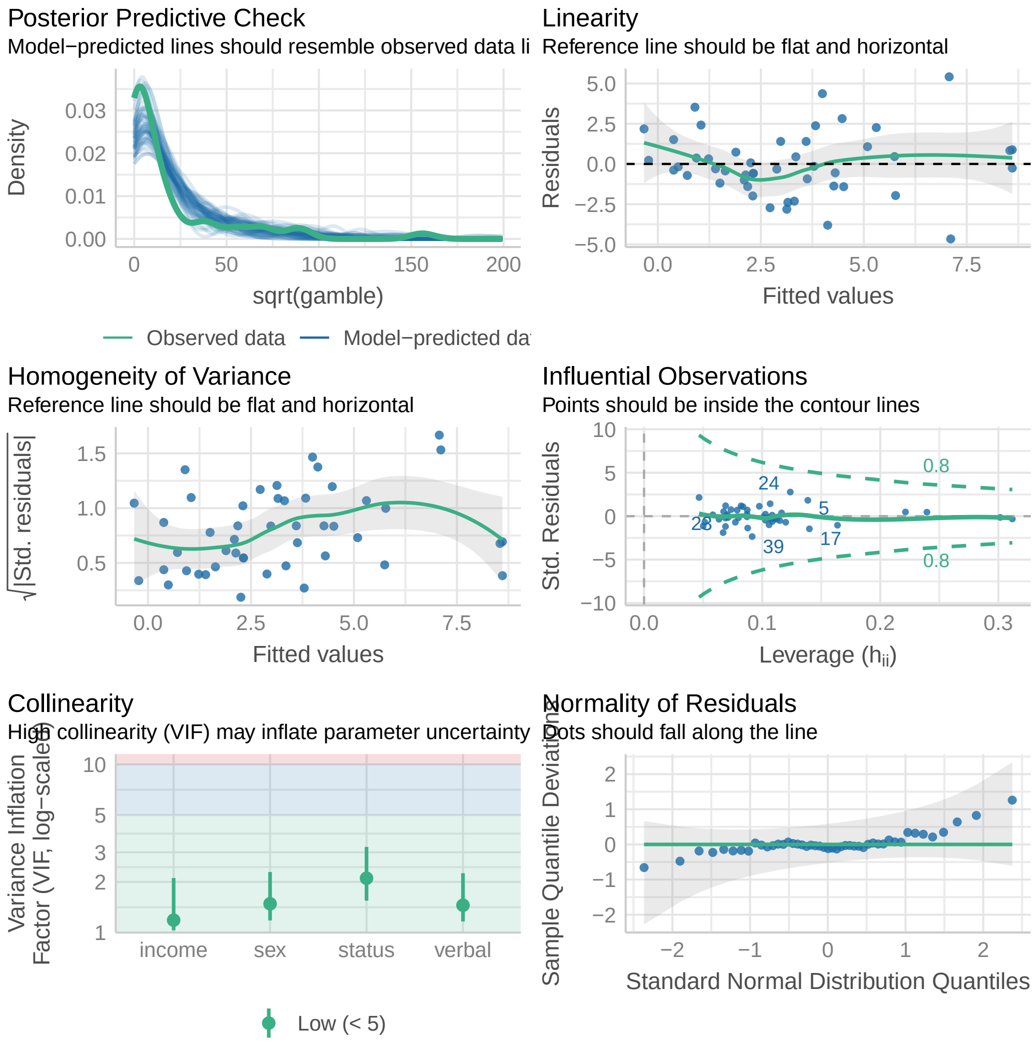


Although there are still signs of non-constant variance, the square root transformation of the expenditure on gambling seem to slightly improve this issue. Observation 24 is no longer 'problematic'. The main concern regarding this second model is that there are still signs of heteroscedasticity. Just out of curiosity, I have look at the plots produced by the package `performance` for both models.

```
check_model(fit_gamble)
```



```r
check_model(fit_gamble_sqrt)
```



Of course, by transforming the data, regression coefficients will need to be interpreted with respect to the transformed scale. Except for the log transformation, there is no straightforward way of back transforming them to values that can be interpreted in the original scale.

(b) Let us first write the model

$$\sqrt{\text{gamble}}_i = \beta_1 + \beta_2 \text{sex}_{M,i} + \beta_3 \text{status}_i + \beta_4 \text{income}_i + \beta_5 \text{verbal}_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

where $\text{sex}_{M,i}$ is a dummy variable taking the value 1 if teenager i is a female and 0 if a male. With all other predictor variables held constant, say at values status*, income*, and verbal*, the estimated difference in the mean (square root) expenditure on gambling for a male compared to a female is

$$\begin{aligned}
&(\hat\beta_1 + \hat\beta_2 \times 0 + \hat\beta_3 \times \text{status}^* + \hat\beta_4 \times \text{income}^* + \hat\beta_5 \times \text{verbal}^*) - \\
&(\hat\beta_1 + \hat\beta_2 \times 1 + \hat\beta_3 \times \text{status}^* + \hat\beta_4 \times \text{income}^* + \hat\beta_5 \times \text{verbal}^*) \\
&= -\hat\beta_2 \\
&= 2.04450
\end{aligned}$$

A 95% confidence interval for this difference is given by

$$(-\hat\beta_2 \pm t_{n-p,0.975}\hat\sigma_{\hat\beta_2}), \quad n - p = 47 - 5 = 42, \quad \hat\sigma_{\hat\beta_2} = 0.75416.$$

```
lb <- -coef(fit_gamble_sqrt)[2] - qt(0.975, df = 42)*0.75416
ub <- -coef(fit_gamble_sqrt)[2] + qt(0.975, df = 42)*0.75416
lb; ub

##      sex1
## 0.5225457
##     sex1
## 3.566459
```

The required 95% CI is $(0.523, 3.566)$.
Note: If you have obtained the symmetric estimate and interval, that is fine as well.

(c) We just need to find what the maximum values are for a male (which, in this case, coincide with the maximum values for the dataset).

```
ind_male <- which(teengamb$sex == "0")
df_male <- data.frame(sex = "0",
                      status = max(teengamb$status[ind_male]),
                      income = max(teengamb$income[ind_male]),
                      verbal = max(teengamb$verbal[ind_male])
                      )
predict(fit_gamble_sqrt, newdata = df_male,
        interval = "prediction", level = 0.95)^2

##        fit      lwr      upr
## 1 75.65038 13.80692 187.1336
```

For such hypothetical observation, the predicted expenditure on gamble is £75.65 with a 95% prediction interval of (£13.81, £187.13).

(d) Let us start by fitting the model where income is the only predictor, i.e., let us start by fitting the following model:

$$\sqrt{\text{gamble}_i} = \beta_1 + \beta_4 \text{income}_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

where we note that although we are using the same notation, the coefficients estimates of β_1 and β_4 in this model will be different from those in the full model (and their interpretation is also, of course, different!).

```
fit_gamble_sqrt_income <- lm(sqrt(gamble) ~ income, data = teengamb)
```

We are then interested in testing the following hypothesis:

$$H_0 = \begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad H_1 \begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_5 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This hypothesis can be tested through an F-ratio test:

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(n - p_1)} \sim F_{p_1 - p_0, n - p_1},$$

where RSS_0 is the residual sum of squares of the model including income as the only predictor and RSS_1 is the residual sum of squares of the full model. Here $p_1 - p_0$ is equal to three. We know that in practice, there is no need to obtain all these quantities 'manually', and that the `anova` command can be used for this purpose. This is what we will do next.

```
anova(fit_gamble_sqrt_income, fit_gamble_sqrt)

## Analysis of Variance Table
##
## Model 1: sqrt(gamble) ~ income
## Model 2: sqrt(gamble) ~ sex + status + income + verbal
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     45 272.76
## 2     42 182.41  3    90.346 6.9339 0.0006774 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.0006774 and therefore we reject the null hypothesis: that is, at least, one of the coefficients β_2, β_3, and β_5 is different than zero (does not necessarily mean that the three coefficients are all different than zero).

(e) Excluding the possibility of interactions between variables, there are eight possible models. The estimated effect, 95% CI, and p-value of ‘sex’ across these eight models are as follows.

```
fit_gamble_sqrt_1 <- lm(sqrt(gamble) ~ sex, data = teengamb)
fit_gamble_sqrt_2 <- lm(sqrt(gamble) ~ sex + income, data = teengamb)
fit_gamble_sqrt_3 <- lm(sqrt(gamble) ~ sex + status, data = teengamb)
fit_gamble_sqrt_4 <- lm(sqrt(gamble) ~ sex + verbal, data = teengamb)
fit_gamble_sqrt_5 <- lm(sqrt(gamble) ~ sex + status + income,
                        data = teengamb)
fit_gamble_sqrt_6 <- lm(sqrt(gamble) ~ sex + income + verbal,
                        data = teengamb)
fit_gamble_sqrt_7 <- lm(sqrt(gamble) ~ sex + status + verbal,
                        data = teengamb)
fit_gamble_sqrt_8 <- lm(sqrt(gamble) ~ sex + income + verbal + status,
                        data = teengamb)

summary_1 <- summary(fit_gamble_sqrt_1)
summary_2 <- summary(fit_gamble_sqrt_2)
summary_3 <- summary(fit_gamble_sqrt_3)
summary_4 <- summary(fit_gamble_sqrt_4)
summary_5 <- summary(fit_gamble_sqrt_5)
summary_6 <- summary(fit_gamble_sqrt_6)
summary_7 <- summary(fit_gamble_sqrt_7)
summary_8 <- summary(fit_gamble_sqrt_8)

ci_1 <- confint(fit_gamble_sqrt_1)
ci_2 <- confint(fit_gamble_sqrt_2)
ci_3 <- confint(fit_gamble_sqrt_3)
ci_4 <- confint(fit_gamble_sqrt_4)
ci_5 <- confint(fit_gamble_sqrt_5)
ci_6 <- confint(fit_gamble_sqrt_6)
ci_7 <- confint(fit_gamble_sqrt_7)
ci_8 <- confint(fit_gamble_sqrt_8)


df <- data.frame("Estimate" = c(summary_1$coefficients[2, 1],
                                summary_2$coefficients[2, 1],
                                summary_3$coefficients[2, 1],
                                summary_4$coefficients[2, 1],
                                summary_5$coefficients[2, 1],
                                summary_6$coefficients[2, 1],
                                summary_7$coefficients[2, 1],
                                summary_8$coefficients[2, 1]),
                 "LB" = c(ci_1[2, 1],
                          ci_2[2, 1],
                          ci_3[2, 1],
                          ci_4[2, 1],
                          ci_5[2, 1],
                          ci_6[2, 1],
                          ci_7[2, 1],
                          ci_8[2, 1]),
                  "UB" = c(ci_1[2, 2],
                          ci_2[2, 2],
                          ci_3[2, 2],
                          ci_4[2, 2],
                          ci_5[2, 2],
                          ci_6[2, 2],
                          ci_7[2, 2],
                          ci_8[2, 2]),
                 "Pval" = c(summary_1$coefficients[2, 4],
                            summary_2$coefficients[2, 4],
                            summary_3$coefficients[2, 4],
                            summary_4$coefficients[2, 4],
                            summary_5$coefficients[2, 4],
                            summary_6$coefficients[2, 4],
                            summary_7$coefficients[2, 4],
                            summary_8$coefficients[2, 4]))

colnames(df) <- c("Estimate", "Lower Bound 95 CI",
                  "Upper Bound 95 CI","P-value")
rownames(df) <- c("Sex", "Sex + Income", "Sex + Status",
                  "Sex + Verbal",
                  "Sex + Status + Income", "Sex + Income + Verbal",
                  "Sex + Status + Verbal",
                  "Sex + Income + Verbal + Status")
library(kableExtra)
knitr::kable(df, escape = FALSE, digits = 4,
             caption = "Question 2(e): estimate of sex coefficient,
             95 CI, and p-value.") %>%
  kable_styling(position = "center")
```

We can observe in the table below that although the estimated effect ranges from -2.0445 to -3.4572, the corresponding 95% CIs have good overlap, and the p-value is below 0.05 for all models. Therefore, the estimated effect of sex appears to be quite stable across the different models.

(f) There is no single answer here. Not all eight models are nested and so we resort to AIC for model selection (although all models are nested within the full model that includes all four predictors).

```
AIC(fit_gamble_sqrt_1, fit_gamble_sqrt_2, fit_gamble_sqrt_3,
    fit_gamble_sqrt_4, fit_gamble_sqrt_5, fit_gamble_sqrt_6,
    fit_gamble_sqrt_7, fit_gamble_sqrt_8)
```

```
##                   df      AIC
## fit_gamble_sqrt_1  3 230.1886
```

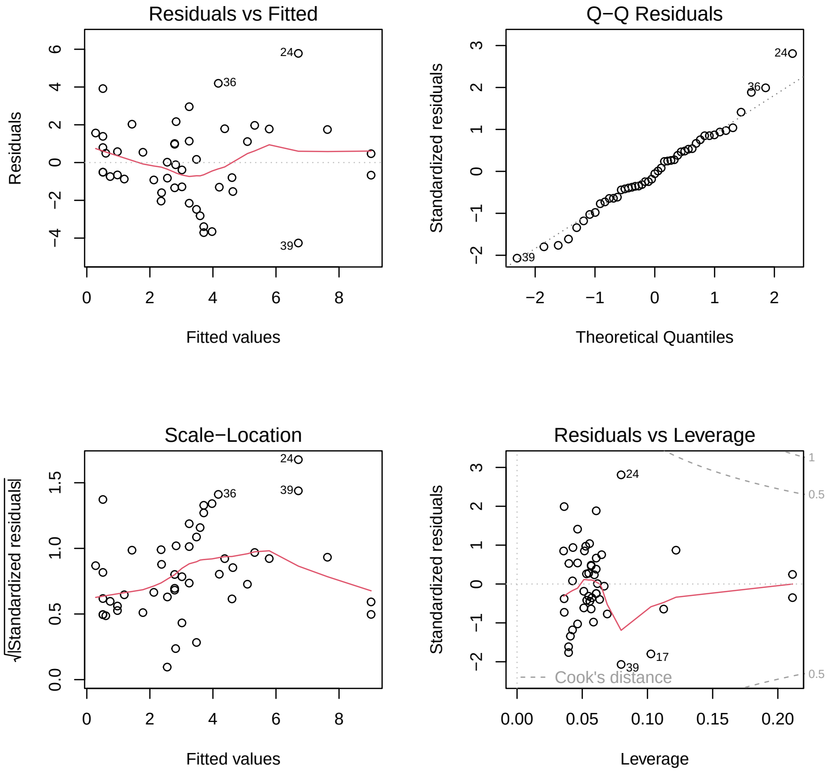
Table 1: Question 2(e): estimate of sex coefficient, 95 CI, and p-value.

	Estimate	Lower Bound 95 CI	Upper Bound 95 CI	P-value
Sex	-2.8884	-4.4960	-1.2809	0.0007
Sex + Income	-2.5069	-3.8009	-1.2130	0.0003
Sex + Status	-3.4572	-5.2775	-1.6369	0.0004
Sex + Verbal	-3.0611	-4.6220	-1.5002	0.0003
Sex + Status + Income	-2.3624	-3.9127	-0.8121	0.0037
Sex + Income + Verbal	-2.6389	-3.9226	-1.3553	0.0002
Sex + Status + Verbal	-3.1684	-4.9960	-1.3409	0.0011
Sex + Income + Verbal + Status	-2.0445	-3.5665	-0.5225	0.0097

```
## fit_gamble_sqrt_2  4 210.0445
## fit_gamble_sqrt_3  4 230.3917
## fit_gamble_sqrt_4  4 227.7623
## fit_gamble_sqrt_5  5 211.9106
## fit_gamble_sqrt_6  5 209.3479
## fit_gamble_sqrt_7  5 229.7019
## fit_gamble_sqrt_8  6 209.1185
```

We see that the AIC is lowest for the full model with the four predictors. However, model 6 (which includes sex, income, and verbal score), model 5 (which includes sex, socioeconomic status, and income), and model 2 (which includes only sex and income) have basically the same AIC. Given we only have 47 observations, I would personally choose model 2 (which has an adjusted r^2 of 0.49 compared to 0.52 of the full model). Let us look at the summary of this model and the residual plots.

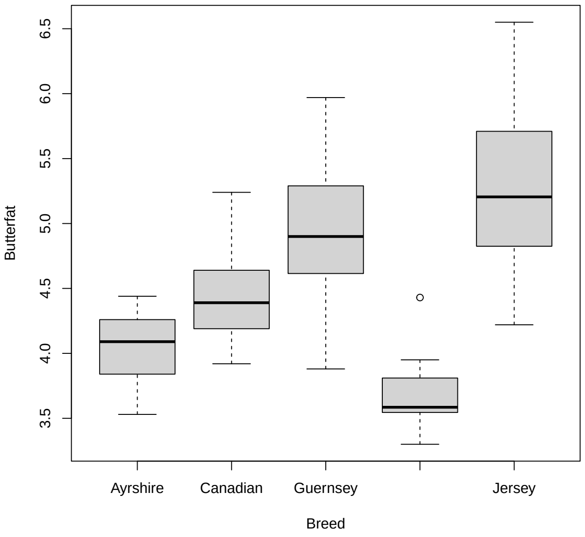
```
summary(fit_gamble_sqrt_2)

##
## Call:
## lm(formula = sqrt(gamble) ~ sex + income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2597 -1.2946 -0.1159  1.2604  5.7808
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.09425    0.60297   3.473 0.001166 **
## sex1        -2.50694    0.64205  -3.905 0.000321 ***
## income       0.46149    0.08968   5.146 5.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.146 on 44 degrees of freedom
## Multiple R-squared:  0.5164,Adjusted R-squared:  0.4945
## F-statistic:  23.5 on 2 and 44 DF,  p-value: 1.142e-07

par(mfrow = c(2, 2))
plot(fit_gamble_sqrt_2)
```



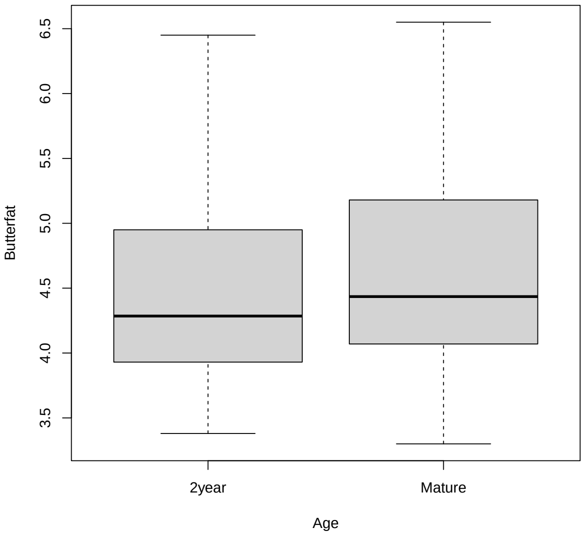
Residual plots do not seem worse than the ones from the full model. The estimated effect of sex is -2.51 indicating that, keeping income fixed, females spend less on gambling than males. On the other hand, the estimated effect of income is 0.46, meaning that, keeping gender fixed, each pound increase (per week) leads to an additional expenditure on gambling of 0.46 (in square root scale). In a more meaningful way, a £10 increase in income leads to a £4.6 increase of (square root) expenditure on gambling.

3. (a) We start by plotting the data to have an idea of differences in butterfat across the different breeds and age groups and to assess a possible interaction effect between breeds and age.

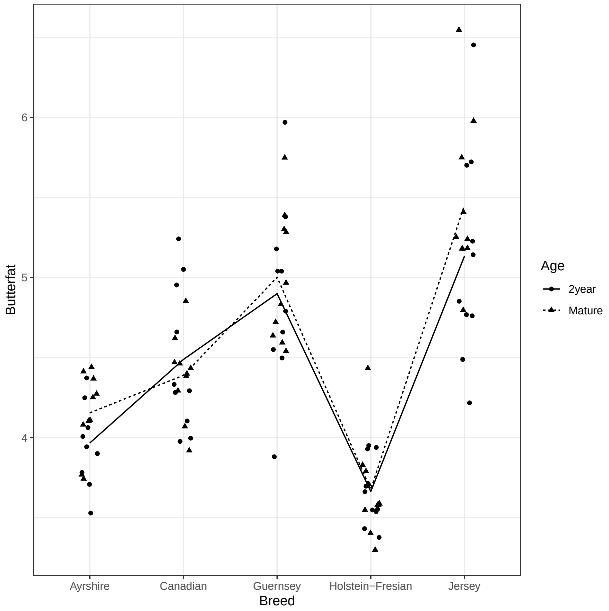
```
boxplot(Butterfat ~ Breed, data = butterfat, ylab = "Butterfat", xlab = "Breed")
```



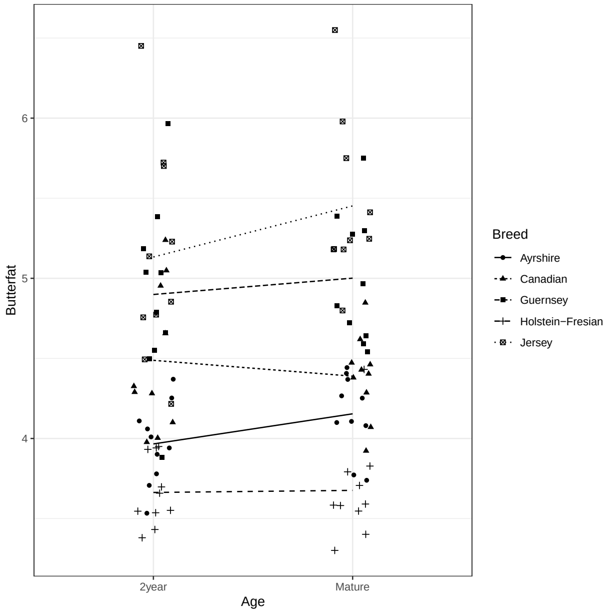
```r
boxplot(Butterfat ~ Age, data = butterfat, ylab = "Butterfat", xlab = "Age")
```



```r
library(ggplot2)
ggplot(butterfat, aes(x = Breed, y = Butterfat, shape = Age)) +
  geom_point( position = position_jitter(width = .1)) +
  stat_summary(fun = "mean", geom = "line", aes(group = Age, linetype = Age)) +
  theme(legend.position = "top", legend.direction = "horizontal") +
  theme_bw()
```



```
ggplot(butterfat, aes(x = Age, y = Butterfat, shape = Breed)) +
  geom_point( position = position_jitter(width = .1)) +
  stat_summary(fun = "mean", geom = "line", aes(group = Breed, linetype = Breed)) +
  theme(legend.position = "top", legend.direction = "horizontal") +
  theme_bw()
```



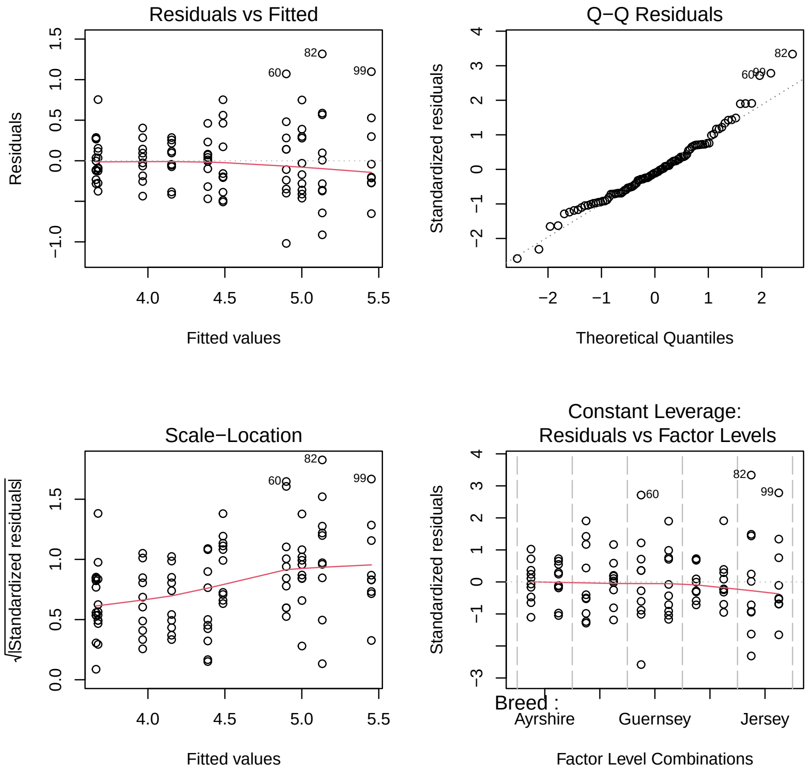
The breed of the cow appears to affect the butterfat content, whereas age does not. The interaction plots show lines that are nearly parallel, suggesting that there might not be an interaction between age and breed (but nonetheless, we should still test this).

(b) Let us start by fitting the most general model that contains an interaction effect between age and breed. Letting y_{ijk} denote the butterfat content $(\%)$ for the kth cow of breed type i and at age group j, the model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad k = 1, \ldots, 10,$$

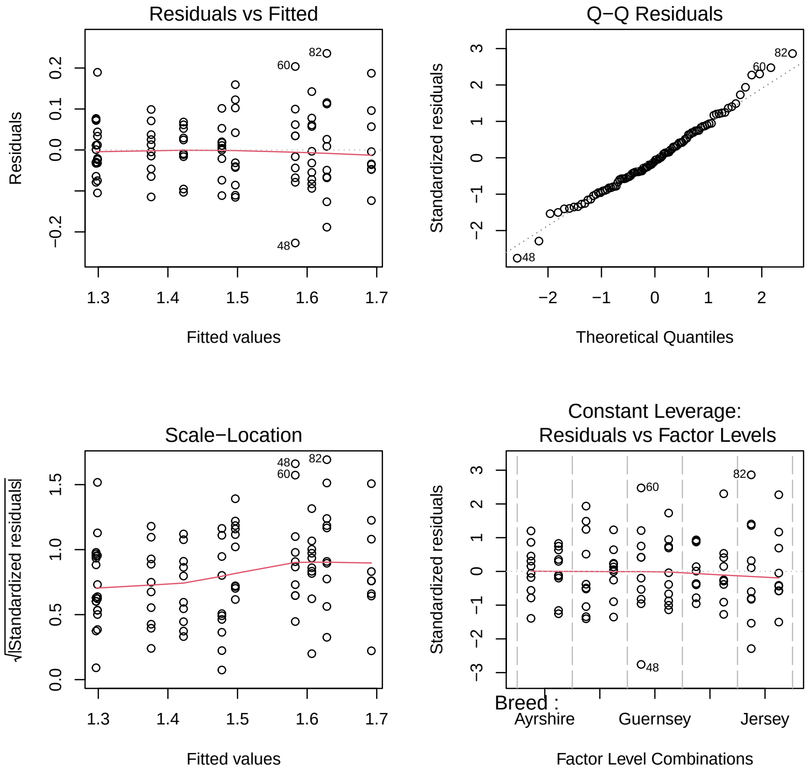
where α_i represents the main effect of breed type, $i \in \{A, C, G, HF, J\}$, β_j represents the main effect of age group, $j \in \{2Y, M\}$, and γ_{ij} represents the interaction effect of breed and age. We know that this model is not identifiable and using the identifiability constraints used by default in R we set $\alpha_A = 0$, $\beta_{2Y} = 0$, and $\gamma_{A,2Y} = \gamma_{A,M} = \gamma_{C,2Y} = \gamma_{G,2Y} = \gamma_{HF,2Y} = \gamma_{J,2Y} = 0$. This leaves us with ten identifiable parameters. Note that we could have alternatively written the model with indicator variables.

```r
fit_inter <- lm(Butterfat ~ Breed*Age, data = butterfat)
par(mfrow = c(2, 2))
plot(fit_inter)
```



There is some indication that the variance increases with the mean. Experimenting with the square root and log transformations, the log seems to work slightly better, but still does not resolve the heteroscedasticity issue completely.

```r
fit_inter_log <- lm(log(Butterfat) ~ Breed*Age, data = butterfat)
par(mfrow = c(2, 2))
plot(fit_inter_log)
```



We proceed in the log scale. Let us now fit the model without the interaction term and conduct an F-ratio test to test the null hypothesis that this simpler model is acceptable.

```
fit_no_inter_log <- lm(log(Butterfat) ~ Breed + Age, data = butterfat)
anova(fit_no_inter_log, fit_inter_log)
```

```
## Analysis of Variance Table
##
## Model 1: log(Butterfat) ~ Breed + Age
## Model 2: log(Butterfat) ~ Breed * Age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     94 0.70043
## 2     90 0.67811  4  0.022321 0.7406 0.5668
```

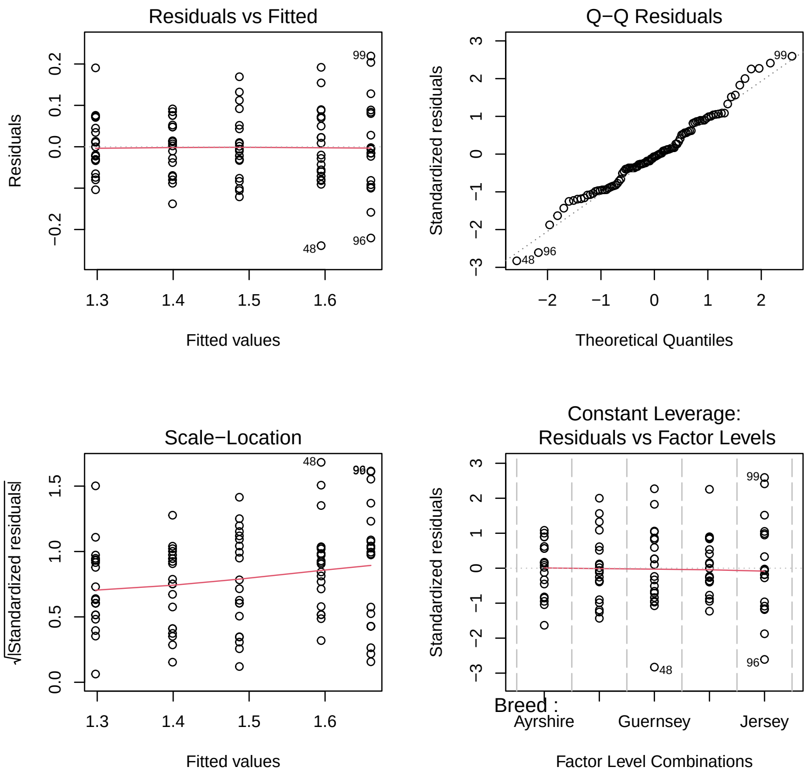
The p-value is 0.5668 and so we do not reject the null hypothesis, meaning that the model without the interaction term is not rejected. So the data does not offer evidence that any potential breed effect on the % of butterfat content changes with age. We can still check whether each term individually is significant.

```
drop1(fit_no_inter_log, test = "F")
```

```
## Single term deletions
##
## Model:
## log(Butterfat) ~ Breed + Age
##        Df Sum of Sq     RSS     AIC F value Pr(>F)
## <none>              0.70043 -484.12
## Breed   4   1.70334 2.40377 -368.81 57.1486 <2e-16 ***
## Age     1   0.01367 0.71410 -484.19  1.8343 0.1789
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we already strongly suspected from our exploratory analysis, the age effect is not significant and so our final model only includes the breed effect. There is thus evidence from this dataset that butterfat content (in %) changes with the breed type but not with age. Before writing the model down mathematically, let us look at the residual plots of the model containing only the breed type effect. I will exclude the intercept so that (because this is a balanced designed experiment), the (estimated) covariance matrix of the estimated breed effects is diagonal (simplifying calculations later in part (c)).

```
fit_no_inter_no_age_log <- lm(log(Butterfat) ~ Breed - 1, data = butterfat)
par(mfrow = c(2, 2))
plot(fit_no_inter_no_age_log)
```



It looks fine (of course, the slight non-constant variance pattern persists).

We then write the final model as follows

$$\log(y_{ik}) = \alpha_i + \epsilon_{ik}, \quad \epsilon_{ik} \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad i \in \{A, C, G, HF, J\}, \quad k = 1, \ldots, 20.$$

Alternatively, using indicator variables we could write the model, in an identifiable form, as

$$\log(y_l) = \alpha_{Ayrshire}\texttt{breed}_{Ayrshire,l} + \alpha_{Canadian}\texttt{breed}_{Canadian,l} + \alpha_{Guernsey}\texttt{breed}_{Guernsey,l} \\ + \alpha_{Holstein-Fresian}\texttt{breed}_{Holstein-Fresian,l} + \alpha_{Jersey}\texttt{breed}_{Jersey,l} + \epsilon_l,$$

with $\epsilon_l \overset{\text{iid}}{\sim} N(0, \sigma^2)$, for $l = 1, \ldots, 100$. Here $\texttt{breed}_{Ayrshire,l}$ is a dummy/indicator variable that takes the value 1 if cow l is of breed type Ayrshire and 0 otherwise. The other indicator variables are defined similarly.

(c) Let us first look at the summary of our selected model in part (b).

```
summary(fit_no_inter_no_age_log)

## 
## Call:
## lm(formula = log(Butterfat) ~ Breed - 1, data = butterfat)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.238995 -0.061397 -0.005733  0.051988  0.219157 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## BreedAyrshire          1.39919    0.01939   72.17   <2e-16 ***
## BreedCanadian          1.48717    0.01939   76.71   <2e-16 ***
## BreedGuernsey          1.59483    0.01939   82.27   <2e-16 ***
## BreedHolstein-Fresian  1.29780    0.01939   66.94   <2e-16 ***
## BreedJersey            1.66031    0.01939   85.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.0867 on 95 degrees of freedom
## Multiple R-squared:  0.9968,Adjusted R-squared:  0.9966 
## F-statistic:  5935 on 5 and 95 DF,  p-value: < 2.2e-16
```

We see that the best breed is Jersey followed by Guernsey. This is based on a point estimate only, which leads to the following question: But is the Jersey breed type clearly superior to Guernsey type? One possibility to answer this question is to find a confidence interval for the difference $\alpha_{Jersey} - \alpha_{Guernsey}$:

$$(\hat{\alpha}_{Jersey} - \hat{\alpha}_{Guernsey}) \pm t_{0.975,100-5}\sqrt{\widehat{\text{var}}(\hat{\alpha}_{Jersey} - \hat{\alpha}_{Guernsey})}.$$

We know that

$$\widehat{\text{var}}(\hat{\alpha}_{Jersey} - \hat{\alpha}_{Guernsey}) = \widehat{\text{var}}(\hat{\alpha}_{Jersey}) + \widehat{\text{var}}(\hat{\alpha}_{Guernsey}) - 2\widehat{\text{cov}}(\hat{\alpha}_{Jersey}, \hat{\alpha}_{Guernsey})$$

Let's confirm now that the estimated covariance matrix of the coefficients, under this parametrization, is in fact diagonal.

```
vcov(fit_no_inter_no_age_log)

##                       BreedAyrshire BreedCanadian BreedGuernsey
## BreedAyrshire          0.0003758409  0.0000000000  0.0000000000
## BreedCanadian          0.0000000000  0.0003758409  0.0000000000
## BreedGuernsey          0.0000000000  0.0000000000  0.0003758409
## BreedHolstein-Fresian  0.0000000000  0.0000000000  0.0000000000
## BreedJersey            0.0000000000  0.0000000000  0.0000000000
##                       BreedHolstein-Fresian  BreedJersey
## BreedAyrshire                  0.0000000000 0.0000000000
## BreedCanadian                  0.0000000000 0.0000000000
## BreedGuernsey                  0.0000000000 0.0000000000
## BreedHolstein-Fresian          0.0003758409 0.0000000000
## BreedJersey                    0.0000000000 0.0003758409
```

Therefore, $\widehat{\text{var}}(\hat{\alpha}_{Jersey} - \hat{\alpha}_{Guernsey}) = 2 \times 0.0003758409$, leading to the interval $(0.0111, 0.1199)$, which is located all above zero (or all above one in the original scale) and therefore we can conclude that Jersey is clearly superior to Guernsey.

```
(1.66031 -  1.59483) - qt(0.975, nrow(butterfat) - 5)*sqrt(2*0.0003758409)

## [1] 0.01105074

(1.66031 -  1.59483) + qt(0.975, nrow(butterfat) - 5)*sqrt(2*0.0003758409)

## [1] 0.1199093
```

Alternatively, we could conduct an hypothesis test:

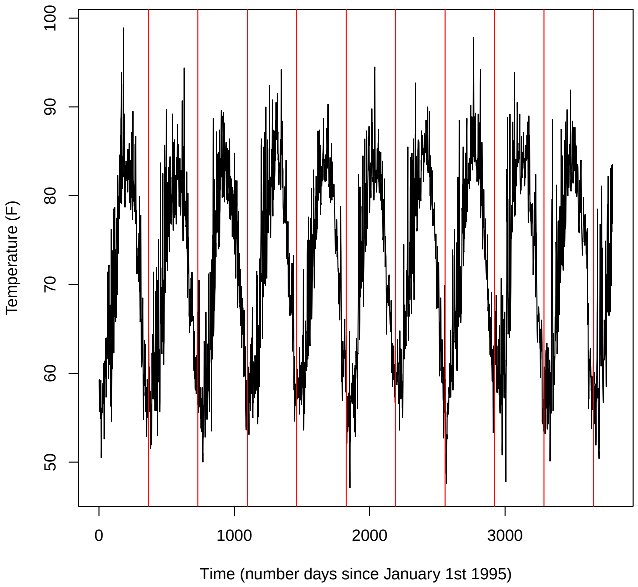
$$H_0 : \alpha_{Jersey} - \alpha_{Guernsey} = 0 \quad \text{vs} \quad H_1 : \alpha_{Jersey} - \alpha_{Guernsey} > 0.$$

```
obs_t <- (1.66031 -  1.59483)/sqrt(2*0.0003758409)
pt(obs_t, df = (nrow(butterfat) - 5), lower.tail = F)

## [1] 0.009451694
```

We reject the null hypothesis and so there is evidence that $\alpha_{Jersey} - \alpha_{Guernsey} > 0$, i.e., that Jersey breed type is superior to Guernsey type.

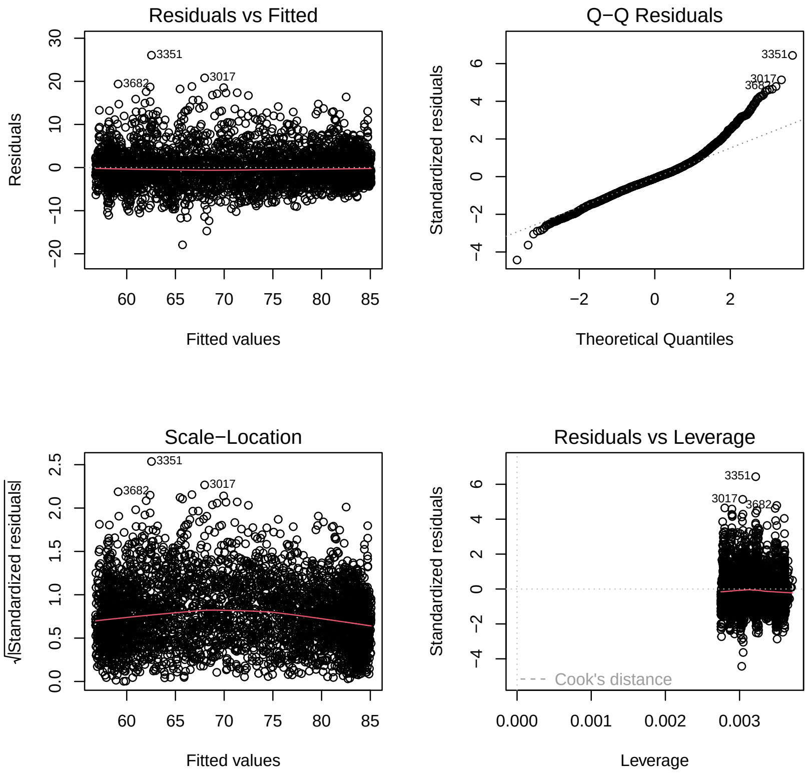
4. (a) Although it was not asked in the question, I started by plotting the data.

```
library(gamair)
data(cairo)

plot(cairo$time, cairo$temp, type = "l",
     xlab = "Time (number days since January 1st 1995)",
     ylab = "Temperature (F)")
abline(v = seq(365.25, max(cairo$time), by = 365.25), col = "red")
```



We can clearly observe the previously described annual cycle. We will use $T = 365.25$ (but $T = 365$, or even, $T = 366$, would work as well, obviously).

```r
fit_5 <- lm(temp ~ time + sin(2*time*pi/365.25) + cos(2*time*pi/365.25) +
            sin(4*time*pi/365.25) + cos(4*time*pi/365.25) +
            sin(6*time*pi/365.25) + cos(6*time*pi/365.25) +
            sin(8*time*pi/365.25) + cos(8*time*pi/365.25) +
            sin(10*time*pi/365.25) + cos(10*time*pi/365.25),
            data = cairo)

par(mfrow = c(2, 2))
plot(fit_5)
```



The residual plots look reasonable apart from the Normal QQ-plot, which shows evidence that the Normal distribution is not quite right. However, given the large sample size, we can probably get away with this.

(b) We will proceed as required in the question. In the first step, we start by testing

$$H_0 : \mathbb{E}(\text{temp}) = \beta_0 + \beta_1 t + \sum_{i=1}^{4} \beta_{2i} \sin(2i\pi t/T) + \beta_{2i+1} \cos(2i\pi t/T),$$

against

$$H_1 : \mathbb{E}(\text{temp}) = \beta_0 + \beta_1 t + \sum_{i=1}^{5} \beta_{2i} \sin(2i\pi t/T) + \beta_{2i+1} \cos(2i\pi t/T).$$

If we do not reject the null hypothesis we then proceed to try dropping the next pair of terms, and so on until no more terms can be dropped.

```r
fit_4 <- lm(temp ~ time + sin(2*time*pi/365.25) + cos(2*time*pi/365.25) +
            sin(4*time*pi/365.25) + cos(4*time*pi/365.25) +
            sin(6*time*pi/365.25) + cos(6*time*pi/365.25) +
            sin(8*time*pi/365.25) + cos(8*time*pi/365.25),
            data = cairo)

anova(fit_4, fit_5)

## Analysis of Variance Table
##
## Model 1: temp ~ time + sin(2 * time * pi/365.25) + cos(2 * time * pi/365.25) +
##     sin(4 * time * pi/365.25) + cos(4 * time * pi/365.25) + sin(6 *
##     time * pi/365.25) + cos(6 * time * pi/365.25) + sin(8 * time *
##     pi/365.25) + cos(8 * time * pi/365.25)
## Model 2: temp ~ time + sin(2 * time * pi/365.25) + cos(2 * time * pi/365.25) +
##     sin(4 * time * pi/365.25) + cos(4 * time * pi/365.25) + sin(6 *
##     time * pi/365.25) + cos(6 * time * pi/365.25) + sin(8 * time *
##     pi/365.25) + cos(8 * time * pi/365.25) + sin(10 * time *
##     pi/365.25) + cos(10 * time * pi/365.25)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   3770 61983
## 2   3768 61979  2    3.8937 0.1184 0.8884
```

The p-value is 0.8884 and therefore we do not reject the null hypothesis. We therefore now test

$$H_0 : \mathbb{E}(\text{temp}) = \beta_0 + \beta_1 t + \sum_{i=1}^{3} \beta_{2i} \sin(2i\pi t/T) + \beta_{2i+1} \cos(2i\pi t/T),$$

against

$$H_1 : \mathbb{E}(\text{temp}) = \beta_0 + \beta_1 t + \sum_{i=1}^{4} \beta_{2i} \sin(2i\pi t/T) + \beta_{2i+1} \cos(2i\pi t/T).$$

```
fit_3 <- lm(temp ~ time + sin(2*time*pi/365.25) + cos(2*time*pi/365.25) +
            sin(4*time*pi/365.25) + cos(4*time*pi/365.25) +
            sin(6*time*pi/365.25) + cos(6*time*pi/365.25),
            data = cairo)

anova(fit_3, fit_4)

## Analysis of Variance Table
##
## Model 1: temp ~ time + sin(2 * time * pi/365.25) + cos(2 * time * pi/365.25) +
##     sin(4 * time * pi/365.25) + cos(4 * time * pi/365.25) + sin(6 *
##     time * pi/365.25) + cos(6 * time * pi/365.25)
## Model 2: temp ~ time + sin(2 * time * pi/365.25) + cos(2 * time * pi/365.25) +
##     sin(4 * time * pi/365.25) + cos(4 * time * pi/365.25) + sin(6 *
##     time * pi/365.25) + cos(6 * time * pi/365.25) + sin(8 * time *
##     pi/365.25) + cos(8 * time * pi/365.25)
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   3772 62080
## 2   3770 61983  2    97.016 2.9504 0.05244 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.05244 and so we can either stop here and use the following model as our final model

$$\mathbb{E}(\text{temp}) = \beta_0 + \beta_1 t + \sum_{i=1}^{4} \beta_{2i} \sin(2i\pi t/T) + \beta_{2i+1} \cos(2i\pi t/T),$$

or proceed to the next step. Both options are sensible. I will proceed.

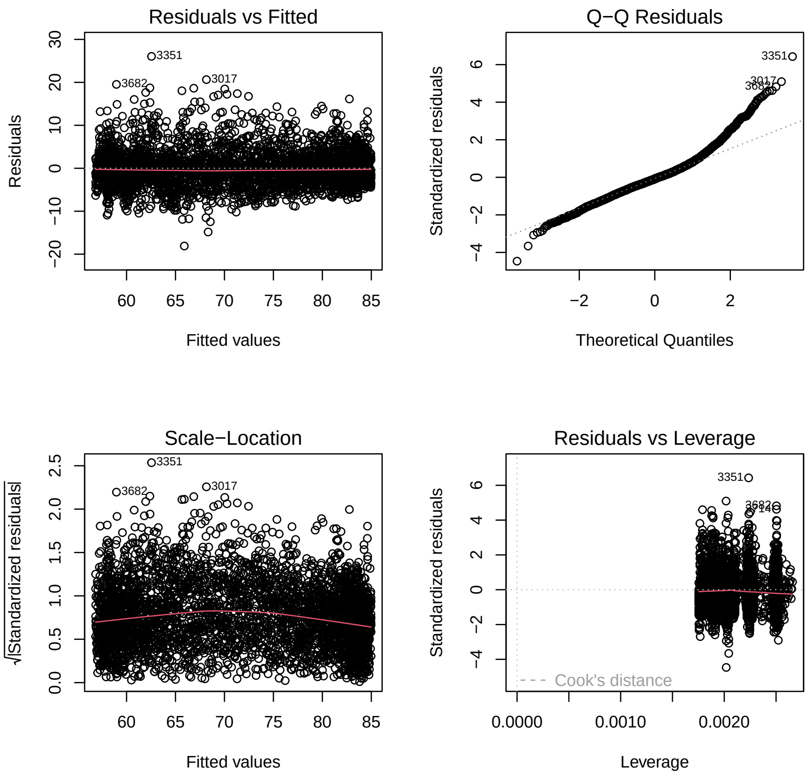
```
fit_2 <- lm(temp ~ time + sin(2*time*pi/365.25) + cos(2*time*pi/365.25) +
            sin(4*time*pi/365.25) + cos(4*time*pi/365.25),
            data = cairo)

anova(fit_2, fit_3)
```

```
## Analysis of Variance Table
##
## Model 1: temp ~ time + sin(2 * time * pi/365.25) + cos(2 * time * pi/365.25) +
##     sin(4 * time * pi/365.25) + cos(4 * time * pi/365.25)
## Model 2: temp ~ time + sin(2 * time * pi/365.25) + cos(2 * time * pi/365.25) +
##     sin(4 * time * pi/365.25) + cos(4 * time * pi/365.25) + sin(6 *
##     time * pi/365.25) + cos(6 * time * pi/365.25)
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1   3774 62626
## 2   3772 62080  2    545.74 16.58 6.775e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now firmly reject the null hypothesis and so our final model is

$$\mathbb{E}(\text{temp}) = \beta_0 + \beta_1 t + \sum_{i=1}^{3} \beta_{2i} \sin(2i\pi t/T) + \beta_{2i+1} \cos(2i\pi t/T).$$

```
par(mfrow = c(2, 2))
plot(fit_3)
```



As before the plots look fine, apart from the QQ plot.

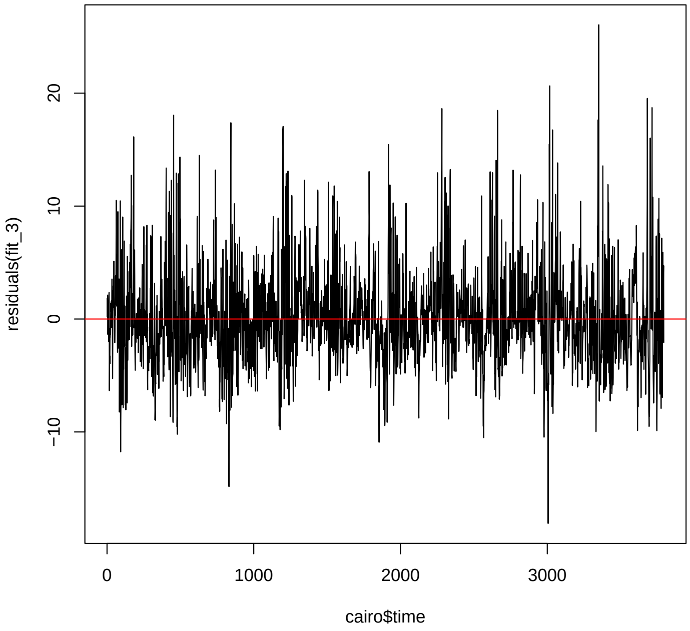
(c) Let us look at the summary of our selected model.

```
summary(fit_3)

##
## Call:
## lm(formula = temp ~ time + sin(2 * time * pi/365.25) + cos(2 *
##     time * pi/365.25) + sin(4 * time * pi/365.25) + cos(4 * time *
##     pi/365.25) + sin(6 * time * pi/365.25) + cos(6 * time * pi/365.25),
##     data = cairo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1013  -2.4155  -0.3446   1.9116  26.0514
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 7.102e+01  1.320e-01  537.973  < 2e-16 ***
## time                        4.675e-04  6.024e-05    7.761 1.08e-14 ***
## sin(2 * time * pi/365.25) -5.520e+00  9.300e-02  -59.359  < 2e-16 ***
## cos(2 * time * pi/365.25) -1.234e+01  9.384e-02 -131.497  < 2e-16 ***
## sin(4 * time * pi/365.25) -3.031e-01  9.325e-02   -3.251  0.00116 **
## cos(4 * time * pi/365.25) -9.277e-01  9.359e-02   -9.912  < 2e-16 ***
## sin(6 * time * pi/365.25)  4.568e-01  9.343e-02    4.889 1.06e-06 ***
## cos(6 * time * pi/365.25) -2.869e-01  9.332e-02   -3.075  0.00212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.057 on 3772 degrees of freedom
## Multiple R-squared:  0.8485,Adjusted R-squared:  0.8482
## F-statistic:  3018 on 7 and 3772 DF,  p-value: < 2.2e-16
```

The estimated coefficient of time is $\hat{\beta}_1 = 4.675 \times 10^{-4}$, which corresponds to an yearly increase of $4.675 \times 10^{-4} \times 365.25 \approx 0.17$ degrees F. The p-value corresponding to $H_0 : \beta_1 = 0$ (against $H_1 : \beta_1 \neq 0$) is 1.08×10^{-14} leading to a strong rejection of the null hypothesis. This suggests that the mean temperature over the years has been increasing beyond annual variation, which has already been accounted for in the model.

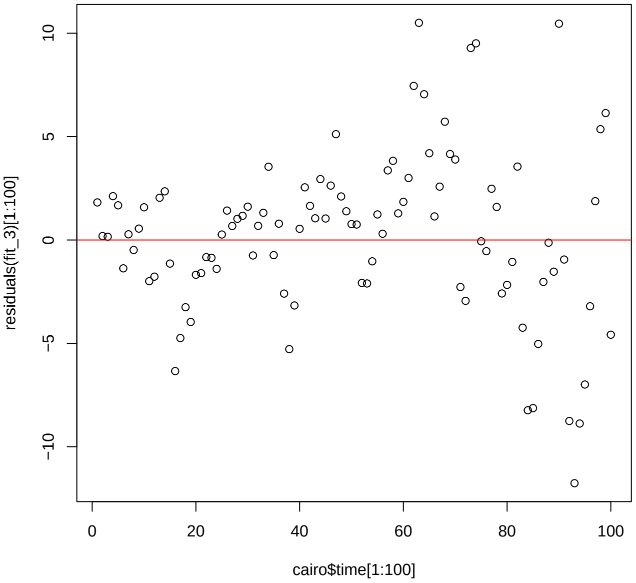
(d) For temporal data such as these, it is sensible to plot the residuals against time.

```
plot(cairo$time, residuals(fit_3), type = "l")
abline(h = 0, col = "red")
```



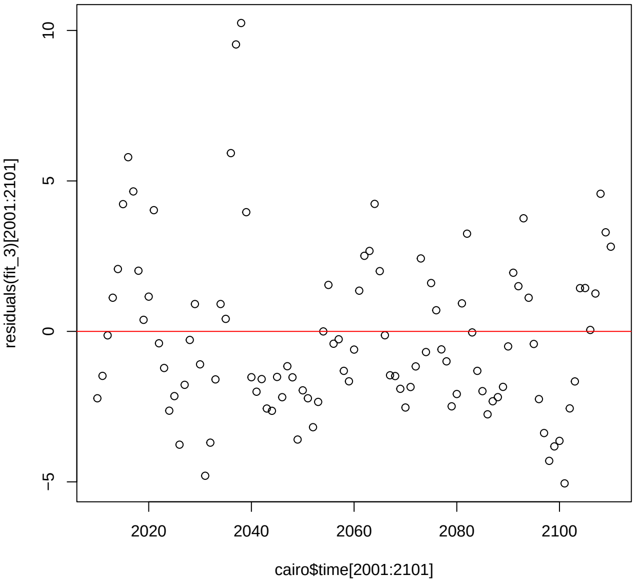
If the errors were uncorrelated, we would expect a random scatter of points above and below the $\epsilon = 0$ line. This is a dense plot, so let us ‘zoom in’ for the first say 100 observations.

```
plot(cairo$time[1:100], residuals(fit_3)[1:100])
abline(h = 0, col = "red")
```



This does not look like a random scatter about zero: we see long sequences of points above or below the $\epsilon = 0$ line. This is an indication of positive serial correlation. We can choose another 'period' of 100 observations.

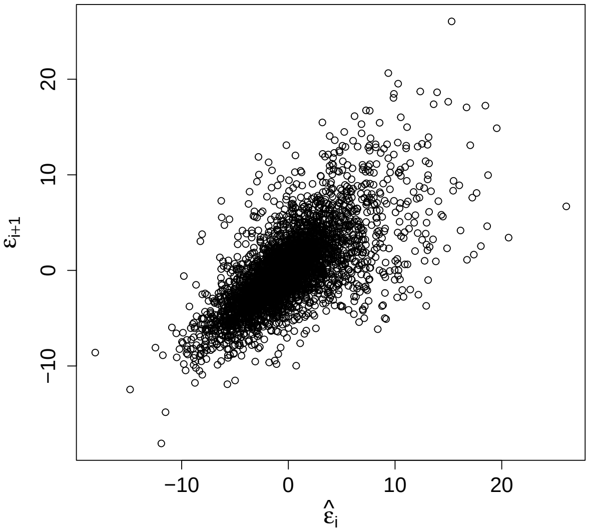
```
plot(cairo$time[2001:2101], residuals(fit_3)[2001:2101])
abline(h = 0, col = "red")
```



The same pattern of runs of high and low residuals is observed. As also asked in the question, an alternative approach to checking for serial correlation is to plot successive pairs of residuals, which we plot below.

```
rsd <- residuals(fit_3)
n <- length(rsd)
```

```
rsd.lag <- rsd[1 : (n-1)]
rsd <- rsd[2 : n]
plot(rsd.lag, rsd,
     xlab = expression(hat(epsilon)[i]),
     ylab = expression(hat(epsilon)[i+1]),
     cex.lab = 1.6, cex.axis = 1.4)
```



We can see a positive correlation again indicating positive serial correlation. Indeed the correlation between successive pairs of residuals is 0.68.

```
cor(rsd.lag, rsd)

## [1] 0.6824304
```

All of this is evidence that the assumption of independent errors of the linear model is violated.

(e) In addition to the two plots produced in the previous point, we can fit the required model to check for further numerical evidence of positive correlation (which is the case for a positive value of β that is statistically significant). We omitted the intercept term because the residuals have mean zero.

```
fit_lag <- lm(rsd ~ rsd.lag - 1)
summary(fit_lag)

##
## Call:
## lm(formula = rsd ~ rsd.lag - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5265  -1.5812  -0.1328   1.3906  15.6082
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## rsd.lag  0.68244    0.01189   57.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.963 on 3778 degrees of freedom
## Multiple R-squared:  0.4657,Adjusted R-squared:  0.4656
## F-statistic:  3293 on 1 and 3778 DF,  p-value: < 2.2e-16

confint(fit_lag)

##             2.5 %    97.5 %
## rsd.lag 0.6591202 0.7057515
```

From the output we obtain that $\hat{\beta} = 0.68\ (0.66, 0.71)$ and the p-value for testing $H_0 : \beta = 0$ is $< 2e - 16$ and hence there is clear evidence of correlation at lag 1. Of course, one could plot more than just successive pairs if we suspect of a more complex dependence.

(f) There is obvious auto-correlation in the residuals, breaking the independence assumption. Given the clear violation of the modelling assumptions, the p-value associated with the test for a long term trend in temperature can only be treated as very approximate. However, given the minute size of the calculated p-value, it is still fairly likely that there is some real increase in temperature.