Computational Statistics Exercise session 2

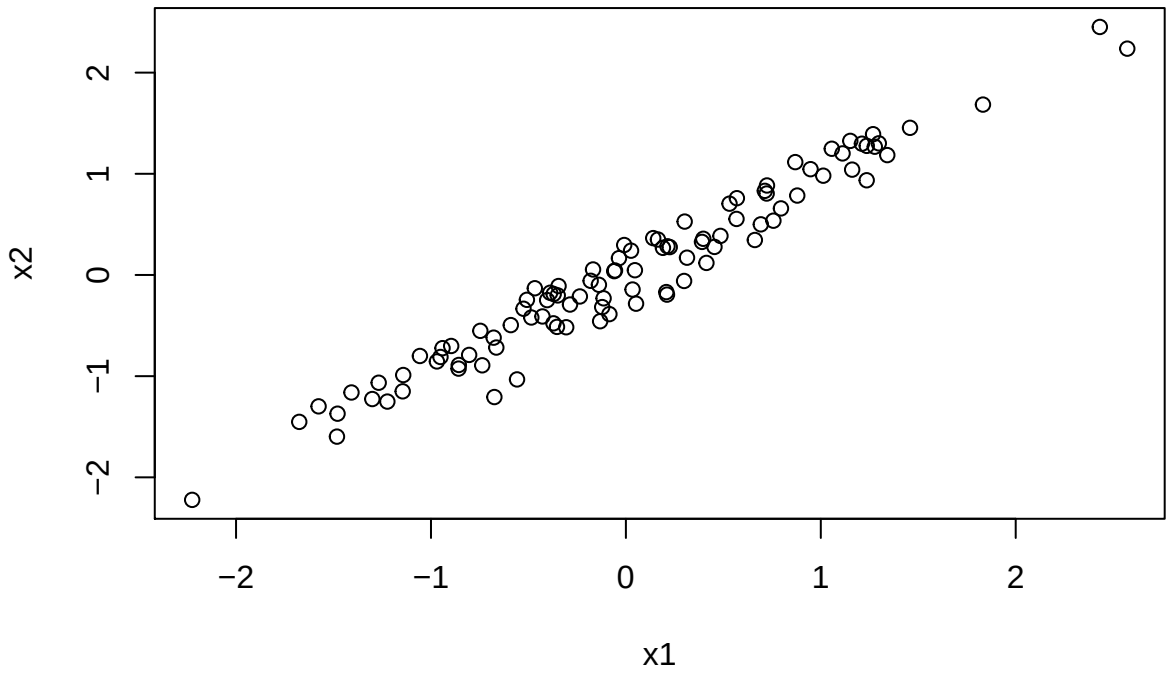
Heiko Kromer

2019-03-15

Problem 1

a) Create a plot of the observations of the two predictor variables x1 and x2.

```
plot(x1,x2)
```



b) Fit a linear model fit1<-lm(y~x1+x2) and print the summary using summary(fit1).

```
fit1 <- lm(y~x1+x2)
(s1 <- summary(fit1))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89540 -0.73467 -0.01828  0.58897  2.43687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0645     0.1062  47.678   <2e-16 ***
```

```
## x1            0.4440     0.5521   0.804    0.423
## x2           -0.8638     0.5674  -1.522    0.131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.061 on 97 degrees of freedom
## Multiple R-squared:  0.1137, Adjusted R-squared:  0.09542
## F-statistic: 6.222 on 2 and 97 DF,  p-value: 0.002869
```

c) Recompute the t-value corresponding to betahat1 by hand using the estimate betahat1 and its estimated standard error se(betahat1).

t value by hand

```
beta1.hat <- s1$coefficients["x1", "Estimate"]
se.beta1.hat <- s1$coefficients["x1", "Std. Error"]
(tval.beta1 <- beta1.hat / se.beta1.hat)
```

```
## [1] 0.8041907
```

t value from R

```
(tval.beta1.fromR <- s1$coefficients["x1", "t value"])
```

```
## [1] 0.8041907
```

Check if the values are different.

```
abs(tval.beta1 -  tval.beta1.fromR)
```

```
## [1] 0
```

d) Give the definition of a p-value. Then compute the p-value corresponding to betahat1 using the t-value from part c) and the quantile function of the t-distribution pt().

Note: You need to provide the correct number of degrees of freedom.

Definition of a p-value

The p-value is the probability of observing any value equal to |t| or larger, where $t = \frac{(\hat{\beta_1} - 0)}{SE(\hat{\beta_1})}$ under the null hypothesis which tests $\beta_1 = 0$ (there is no relationship between X and Y) versus $\beta_1 \neq 0$ (there is a relationship between X and Y).

```
# n are the number of observations, already defined
# p is the intercept and two variables x1 and x2
p <- 2
(pval.beta1 <- 2*pt(abs(tval.beta1), df=n-p, lower=FALSE))
```

```
## [1] 0.4232334
```

```
(pval.beta1.fromR <- s1$coefficients["x1", "Pr(>|t|)"])
```

```
## [1] 0.4232534
```

check difference

```
(abs(tval.beta1 -  tval.beta1.fromR))
```

```
## [1] 0
```

e) Report the p-value of the overall F-test and reproduce it using anova().

```
options(digits=10)
s1
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##         Min          1Q      Median          3Q         Max
## -2.89539537 -0.73467157 -0.01827616  0.58897411  2.43686760
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  5.0645272  0.1062239 47.67785  < 2e-16 ***
## x1           0.4439596  0.5520576  0.80419  0.42325
## x2          -0.8637536  0.5674145 -1.52226  0.13120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.060502 on 97 degrees of freedom
## Multiple R-squared:  0.1136987,  Adjusted R-squared:  0.09542449
## F-statistic:   6.2218 on 2 and 97 DF,  p-value: 0.002868773
```

```
pOverall.s1 <- 0.002868773
# overall p-value is 0.002869


# We can also reproduce the p-value by comparing two models:
fit1.small <- lm(y~1)
# and conducting a partial F-test:
s1.anova <- anova(fit1.small, fit1)
(pOverall.s1.anova <- s1.anova$`Pr(>F)`[2])
```

```
## [1] 0.002868772659
```

```
(abs(pOverall.s1.anova-pOverall.s1))
```

```
## [1] 3.405067079e-10
```

f) The overall F-test is significant. However, the p-values for x1 and x2 are not significant. Explain how this can be true.

The overall F-test compares two models: the constant model and a model with both predictors x1 and x2 present. The p-values in the table compare a model with one of the predictors versus a model without the predictor (i.e. model one is $y \sim x1$, model two is $y \sim x1 + x2$). Only one predictor is enough to make a significant prediction. This is not very surprising since we saw that the x1 and x2 are highly correlated, i.e. low x1 correspond to low values of x2 and high values of x1 correspond to high values of x2.

g) Report the residual standard error, interpret it, and recompute it based on residuals(fit1).

```
options(digits=5)
res.fromR <- residuals(fit1)
```

Re-compute summary statistics of residuals

summary(res.fromR, digits=3)

```
# Compare to summary(fit1)
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8954 -0.7347 -0.0183  0.5890  2.4369
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.065      0.106   47.68   <2e-16 ***
## x1             0.444      0.552    0.80     0.42
## x2            -0.864      0.567   -1.52     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 97 degrees of freedom
## Multiple R-squared:  0.114,  Adjusted R-squared:  0.0954
## F-statistic: 6.22 on 2 and 97 DF,  p-value: 0.00287
```

The residual standard error is 1.06.

It is

$$\hat{\sigma}^2 = \frac{RSS}{n-p}$$

where RSS is the residual sum of squares

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

n is the number of observations (rows) and p the number of parameters (3 in this exercise).

The residual standard error (RSE) is given as:

$$RSE = \hat{\sigma} = \sqrt{\frac{RSS}{n-p}}$$

$\hat{\sigma}^2$ is a measure of goodness of fit. It is an estimate of σ^2, the variance of the statistical errors. The smaller the number, the better the fit (points closer to the line). The smaller the better in relation to the scale of the dependent variable. The RSE is measured in the same units as the dependent variable.

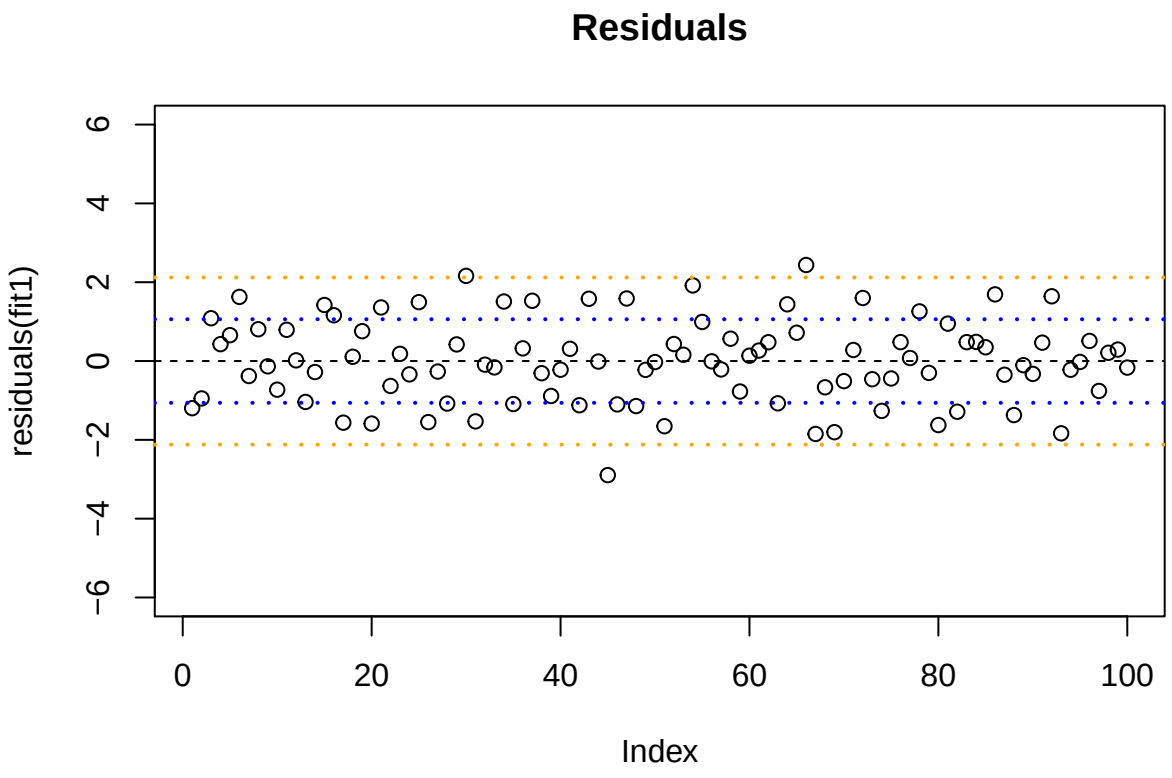
```
# Re-compute residual standard error (RSE):
p <- 3                                  # intercept and two variables
sum(res.fromR^2)/(n-p)                        # sigmahat^2 = RSS/(n-p)
```

```
## [1] 1.1247
```

```
(RSE <- sqrt( sum(res.fromR^2)/(n-p) ))      # RSE = sqrt(RSS/(n-p))
```

```
## [1] 1.0605
```

```
plot(residuals(fit1), ylim=c(-6,6), main="Residuals")
# In case of normally distributed errors, we expect:
#   About 66% of the points are within +/- hat.sigma
#     from the regression plane (blue dotted lines)
#   Aboute 95% of the points are within +/- 2*hat.sigma
#     from the regression plane (orange dotted lines)
abline(h=0, lty=2)
abline(h=RSE, lty=3, col="blue", lwd=2)
abline(h=-RSE, lty=3, col="blue", lwd=2)
abline(h=2*RSE, lty=3, col="orange", lwd=2)
abline(h=-2*RSE, lty=3, col="orange", lwd=2)
```



h) Report the R2 value, interpret it, and recompute it using residuals(fit1).

```
s1
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8954 -0.7347 -0.0183  0.5890  2.4369
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.065      0.106   47.68   <2e-16 ***
## x1             0.444      0.552    0.80     0.42
## x2            -0.864      0.567   -1.52     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 97 degrees of freedom
## Multiple R-squared:  0.114,  Adjusted R-squared:  0.0954
## F-statistic: 6.22 on 2 and 97 DF,  p-value: 0.00287
```

The R2 value is 0.114. The adjusted R2 value is 0.0954.

The R2 value represents the proportion of variance explained by regression model. R2 is the proportion of the variance in y that is explained by the model. If R2 = 0, then the model is useless; if R2 = 1, model explains everything (errors are the smallest).

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where TSS is the total sum of squares $TSS = \sum_{i=1}^{n}(y_i - y)^2$.

The adjusted R2 value penalizes larger models. A bigger model with more freedom has a better R2 than a smaller model. When adding variables (not more points, but more columns) to a model, the R2 can only go up. If p goes up and RSS stay the same —> RSS/(n-p) becomes larger —> adjusted R2 becomes smaller. So if the gain in a decrease of RSS when adding another variable outweighs the decrease in (n-p), the model will be better.

$$R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

```
# Re-compute R^2:
RSS <- sum( residuals(fit1)^2 )  # residuals = y - yhat
TSS <- sum( (y-mean(y))^2 )
(Rsquared <- 1 - RSS/TSS)
```

```
## [1] 0.1137
```

```
# Re-compute adjusted R^2:
(Rsquared.adj <- 1 - (RSS/(n-p))/(TSS/(n-1)))
```

```
## [1] 0.095424
```

```
# Adjusted for number of variables in the model
```

i) Assume now that we only observed the values for x1 and y whereas x2 is a hidden predictor that we do not observe. Fit the model fit3<-lm(y~x1) and print the summary summary(fit3). Compare the estimated coefficient corresponding to x1 to the one in part b). Interpret the coefficient of x1 in both models.

```
fit3<-lm(y~x1)
(s3 <- summary(fit3))
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0078 -0.6350 -0.0781  0.6853  2.5553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.056      0.107   47.35   <2e-16 ***
## x1            -0.377      0.119   -3.16   0.0021 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 98 degrees of freedom
## Multiple R-squared:  0.0925, Adjusted R-squared:  0.0833
## F-statistic: 9.99 on 1 and 98 DF,  p-value: 0.00209
```

In the first model it is $\beta_1 = 0.444$ and in this model (fit3) it is $\beta_1 = -0.377$. The sign of β_1 flipped between the 1st and 3rd model. This shows that the interpretation of each β_k depends on the other variables in the model!

Problem 2

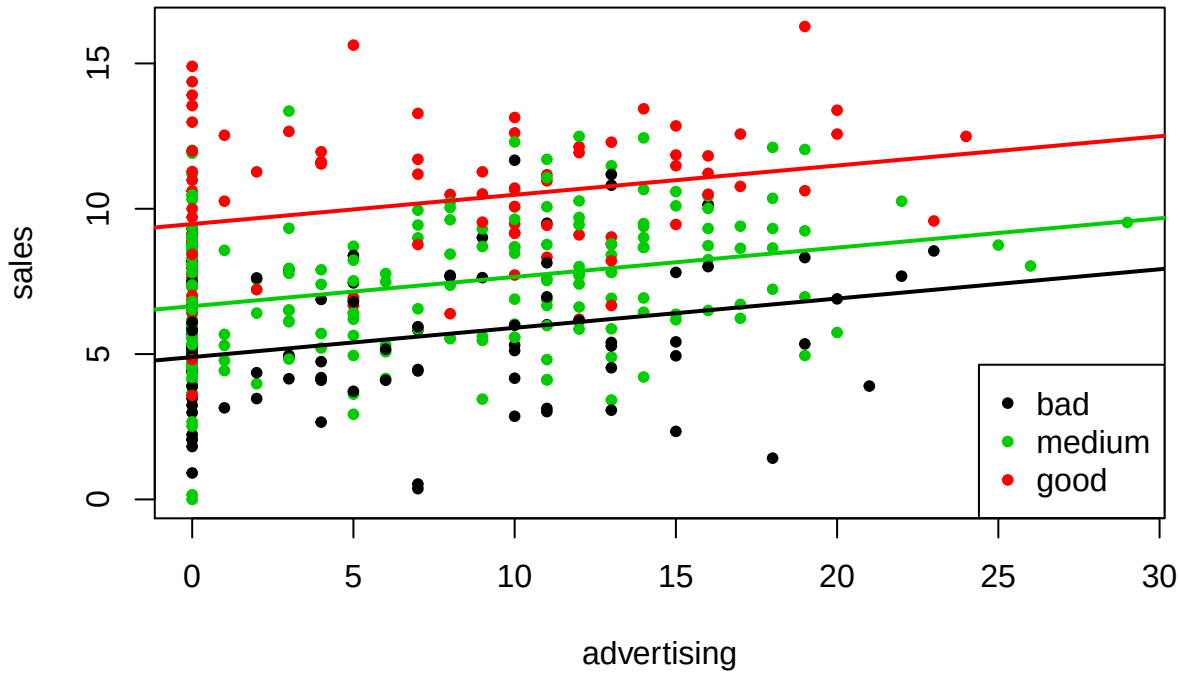
In this exercise, we will code a categorical variable by hand. The dataset Carseats contains the number of child car seat sales and several predictors in 400 locations. We will only use the quantitative predictor advertising (local advertising budget for company at each location in thousands of dollars) and the qualitative predictor shelveloc (a factor with levels 'Bad', 'Good' and 'Medium' indicating the quality of the shelving location for the car seats at each site).

```
# plot
fit<-lm(sales~shelveloc+advertising)
(s <- summary(fit))
```

```
##
## Call:
## lm(formula = sales ~ shelveloc + advertising)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.648 -1.620 -0.048  1.531  6.410
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.8966     0.2521   19.43  < 2e-16 ***
## shelvelocGood     4.5769     0.3348   13.67  < 2e-16 ***
## shelvelocMedium   1.7514     0.2748    6.37  5.1e-10 ***
## advertising       0.1007     0.0169    5.95  5.9e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.24 on 396 degrees of freedom
## Multiple R-squared:  0.373,  Adjusted R-squared:  0.369
## F-statistic: 78.6 on 3 and 396 DF,  p-value: <2e-16
```

```
coeff <- coefficients(fit)
```

```r
plot(advertising, sales, col=shelveloc, pch=20)
legend("bottomright", c("bad", "medium", "good"), col=c(1,3,2), pch=20)
abline(a=coeff[1], b=coeff[4], col=1, lwd=2)
abline(a=coeff[1]+coeff[3], b=coeff[4], col=3, lwd=2)
abline(a=coeff[1]+coeff[2], b=coeff[4], col=2, lwd=2)
```



a) Encode the factor variable shelveloc in the same way as done automatically by R by constructing appropriate predictors a1 and a2. a1 shall be 1 when the level of shelveloc is medium and a2 shall be 1 if its level is good. The so-called contrast coding in this case can be seen in Table 1. Fit the model $fit_a < -lm(sales\ a1 + a2 + advertising)$. Verify that fit and fit_a are indeed equal and give an interpretation of the coefficients corresponding to a1 and a2.

```r
# boolean vectors for easy construction of a1, a2
bad <- levels(shelveloc)[1]==shelveloc
medium <- levels(shelveloc)[3]==shelveloc
good <- levels(shelveloc)[2]==shelveloc
a1 <- medium*1
a2 <- good*1

fit_a<-lm(sales~a1+a2+advertising)
summary(fit_a)
```

```
##
## Call:
## lm(formula = sales ~ a1 + a2 + advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.648 -1.620 -0.048  1.531  6.410
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.8966     0.2521   19.43  < 2e-16 ***
## a1            1.7514     0.2748    6.37  5.1e-10 ***
## a2            4.5769     0.3348   13.67  < 2e-16 ***
## advertising   0.1007     0.0169    5.95  5.9e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.24 on 396 degrees of freedom
## Multiple R-squared:  0.373,  Adjusted R-squared:  0.369
## F-statistic: 78.6 on 3 and 396 DF,  p-value: <2e-16
```

There are three fitted parallel regression planes. The intercepts of these planes are given by $\hat{\beta}_1 = 4.89662$ for shelveloc bad data, $\hat{\beta}_1 + \hat{\gamma}_1 = 4.89662 + 1.75142$ for shelveloc medium data and $\hat{\beta}_1 + \hat{\gamma}_2 = 4.89662 + 4.57686$ for shelveloc good data.

The model in this case reads

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_1 * a1 + \gamma_2 * a2 + \epsilon$$

Which is hence for the three categories

$$y_i = \beta_1 + \beta_2 * x_i + \epsilon \text{ for the i's corresponding to shelveloc bad}$$

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_1 * a1 + \epsilon \text{ for the i's corresponding to shelveloc medium}$$

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_2 * a2 + \epsilon \text{ for the i's corresponding to shelveloc good}$$

b) Construct predictor variables b1 and b2 according to the contrast coding in Table 1 and fit the model $fit_b < -lm(sales\ b1 + b2 + advertising)$. Give an interpretation of the coefficients of b1 and b2. boolean vectors for easy construction of b1, b2

```
b1 <- bad*1
b2 <- good*1

fit_b<-lm(sales~b1+b2+advertising)
summary(fit_b)
```

```
##
## Call:
## lm(formula = sales ~ b1 + b2 + advertising)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.648 -1.620 -0.048  1.531  6.410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6480     0.1877   35.41  < 2e-16 ***
## b1           -1.7514     0.2748   -6.37  5.1e-10 ***
```

```
## b2            2.8254     0.2871    9.84  < 2e-16 ***
## advertising   0.1007     0.0169    5.95  5.9e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.24 on 396 degrees of freedom
## Multiple R-squared:  0.373,  Adjusted R-squared:  0.369
## F-statistic: 78.6 on 3 and 396 DF,  p-value: <2e-16
```

There are again three fitted parallel regression planes. The coefficient of the intercepts of these planes in the summary are different now (because the reference intersect, the baseline $\hat{\beta}_1$, is a different one) and are given by $\hat{\beta}_1 = 6.64805$ for shelveloc medium data, $\hat{\beta}_1 + \hat{\gamma}_1 = 6.64805 - 1.75142$ for shelveloc bad data and $\hat{\beta}_1 + \hat{\gamma}_2 = 6.64805 + 2.82543$ for shelveloc good data.

The model in this case reads

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_1 * b1 + \gamma_2 * b2 + \epsilon$$

Which is hence for the three categories

$$y_i = \beta_1 + \beta_2 * x_i + \epsilon \text{ for the i's corresponding to shelveloc medium}$$

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_1 * b1 + \epsilon \text{ for the i's corresponding to shelveloc bad}$$

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_2 * b2 + \epsilon \text{ for the i's corresponding to shelveloc good}$$

Note that the γ have different values compared to the ones in the previous exercise!

c) Construct predictor variables c1, c2 and c3 according to Table 1. Then fit the model $fit_c < -lm(sales\ c1 + c2 + c3 + advertising)$. This causes a problem. Why?

```
c1 <- bad*1
c2 <- medium*1
c3 <- good*1

fit_c<-lm(sales~c1+c2+c3+advertising)
summary(fit_c)
```

```
##
## Call:
## lm(formula = sales ~ c1 + c2 + c3 + advertising)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.648 -1.620 -0.048  1.531  6.410
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4735     0.2734   34.65  < 2e-16 ***
## c1           -4.5769     0.3348  -13.67  < 2e-16 ***
## c2           -2.8254     0.2871   -9.84  < 2e-16 ***
## c3                NA         NA      NA       NA
## advertising   0.1007     0.0169    5.95  5.9e-09 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.24 on 396 degrees of freedom
## Multiple R-squared:  0.373,  Adjusted R-squared:  0.369
## F-statistic: 78.6 on 3 and 396 DF,  p-value: <2e-16
```

The problem is evident looking at the equations given in the previous two exercises. The third c is already taken into account for by the equation for the intercept and there are only 3 categories!

$$y_i = \beta_1 + \beta_2 * x_i + \epsilon \text{ for the i's corresponding to shelveloc good}$$

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_1 * c1 + \epsilon \text{ for the i's corresponding to shelveloc bad}$$

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_2 * c2 + \epsilon \text{ for the i's corresponding to shelveloc medium}$$

Hence γ_3 must be equal to 0 in the following model equation (there are no categories left for not good, not bad and not medium):

$$y_i = \beta_1 + \beta_2 * x_i + \gamma_3 * c3 + \epsilon$$

d) Remove the intercept by using fit_c<-lm(-1+...). Interpret the coefficients corresponding to c1, c2 and c3.

```
fit_c<-lm(sales~-1+c1+c2+c3+advertising)
summary(fit_c)
```

```
##
## Call:
## lm(formula = sales ~ -1 + c1 + c2 + c3 + advertising)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.648 -1.620 -0.048  1.531  6.410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## c1            4.8966     0.2521   19.43  < 2e-16 ***
## c2            6.6480     0.1877   35.41  < 2e-16 ***
## c3            9.4735     0.2734   34.65  < 2e-16 ***
## advertising   0.1007     0.0169    5.95  5.9e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.24 on 396 degrees of freedom
## Multiple R-squared:  0.922,  Adjusted R-squared:  0.921
## F-statistic: 1.17e+03 on 4 and 396 DF,  p-value: <2e-16
```

In this case it is

$$y_i = \beta_2 * x_i + \gamma_1 * c1 + \gamma_2 * c2 + \gamma_3 * c3 + \epsilon$$

because the intercept was removed, i.e. $\beta_1 = 0$, compared to the previous model. Separated for the different categories:

$$y_i = \beta_2 * x_i + \gamma_1 * c1 + \epsilon \text{ for the i's corresponding to shelveloc bad}$$

$$y_i = \beta_2 * x_i + \gamma_2 * c2 + \epsilon \text{ for the i's corresponding to shelveloc medium}$$
$$y_i = \beta_2 * x_i + \gamma_3 * c3 + \epsilon \text{ for the i's corresponding to shelveloc good}$$

There are again three fitted parallel regression planes. The coefficient of the intercepts of these planes in the summary are also different now. In this case there is no reference baseline and the interesects of the regression planes are $\gamma_1 = 4.8966$ for shelveloc bad data, $\gamma_2 = 6.6480$ for shelveloc medium data, and $\gamma_3 = 9.4735$ for shelveloc good data.

e) Show that the fitted values are the same for fit_a, fit_b and fit_c. Note: Due to rounding errors the values are not exactly the same. Show that they are very close. R-hint: max(abs(fitted(fit_a)-fitted(fit_b)))

```
max(abs(fitted(fit_a)-fitted(fit_b)))
```

```
## [1] 2.0961e-13
```

```
max(abs(fitted(fit_b)-fitted(fit_c)))
```

```
## [1] 1.9007e-13
```

```
max(abs(fitted(fit_a)-fitted(fit_c)))
```

```
## [1] 1.7408e-13
```

The fitted values are the same.

f) We now want to know if distinguishing between all three categories is significantly better than distinguishing only between "bad" (level bad) and "not bad" (level medium or good) each time alsoaccounting for advertising. In which of the summaries of the fits fit_a, fit_b, fit_c can we see this directly? Explain.

We can see this directly in the summary of the p-value for the "shelveloc-bad-line" (note that this depends on the baseline in the fits). The p-value for the line of the respective coefficient is significant, meaning that if test (i.e. for fit_c)

$$H_0 : y_i = \beta_2 * x_i + \gamma_2 * c2 + \gamma_3 * c3 + \epsilon$$

(=Distinguishing only between bad and not bad)

$$H_a : y_i = \beta_2 * x_i + \gamma_1 * c1 + \gamma_2 * c2 + \gamma_3 * c3 + \epsilon$$

(=Distinguishing between bad, medium and good)

We find a significant p-value!

g) Suppose we used the coding from fit_a. Conduct a partial F-test to check if we need to distinguish between medium and good by fitting a model fit_d with a new dummy variable.

```
a1 <- medium*1
a2 <- good*1
# large model: distinguishes between medium and good
fit_a<-lm(sales~a1+a2+advertising)

# small model: does not distinguish between medium and good
```

```r
d1 <- bad*1
d2 <- bad*0   # not bad --> medium or good
fit_d <- lm(sales~d1+d2+advertising)

# and conducting a partial F-test:
(fit.anova <- anova(fit_d, fit_a) )
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ d1 + d2 + advertising
## Model 2: sales ~ a1 + a2 + advertising
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    397 2482
## 2    396 1994  1       488 96.8 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

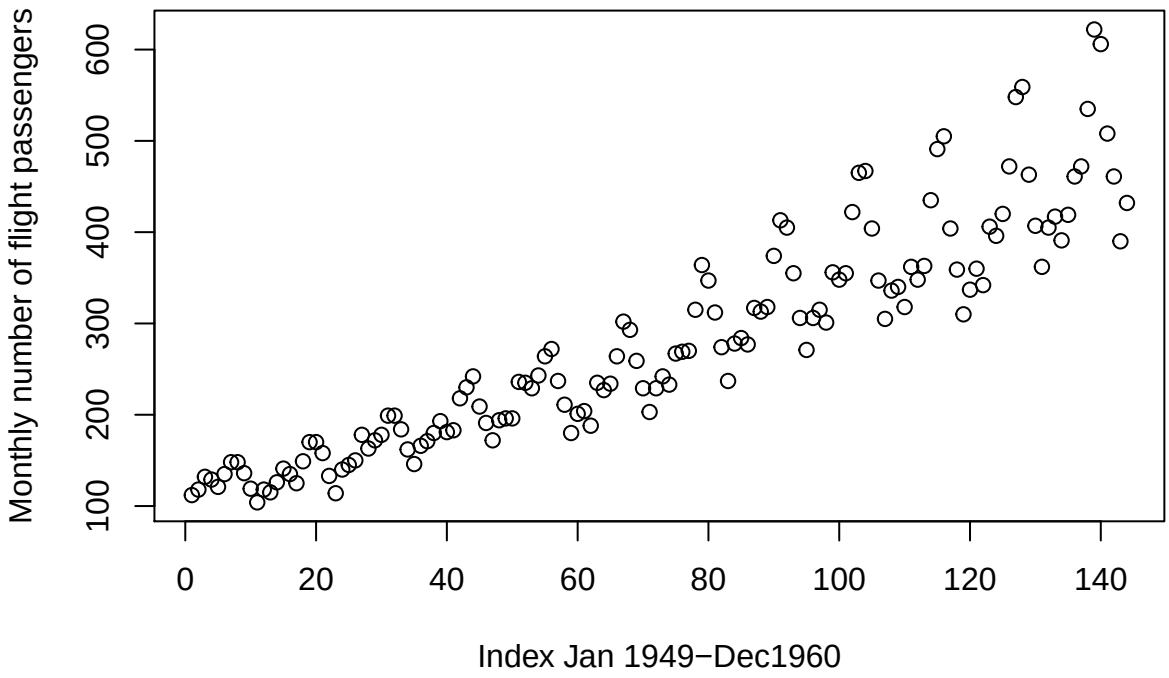
This significant very small p-value means that we can reject the hypothesis of the smaller model, so we need to distinguish between the medium and good as done in model fit_a.

Problem 3

The dataset airline contains the monthly number of flight passengers in the USA in the years 1949-1960 ranging from January 1949 to December 1960.

a) Plot the data against time and describe what you observe.

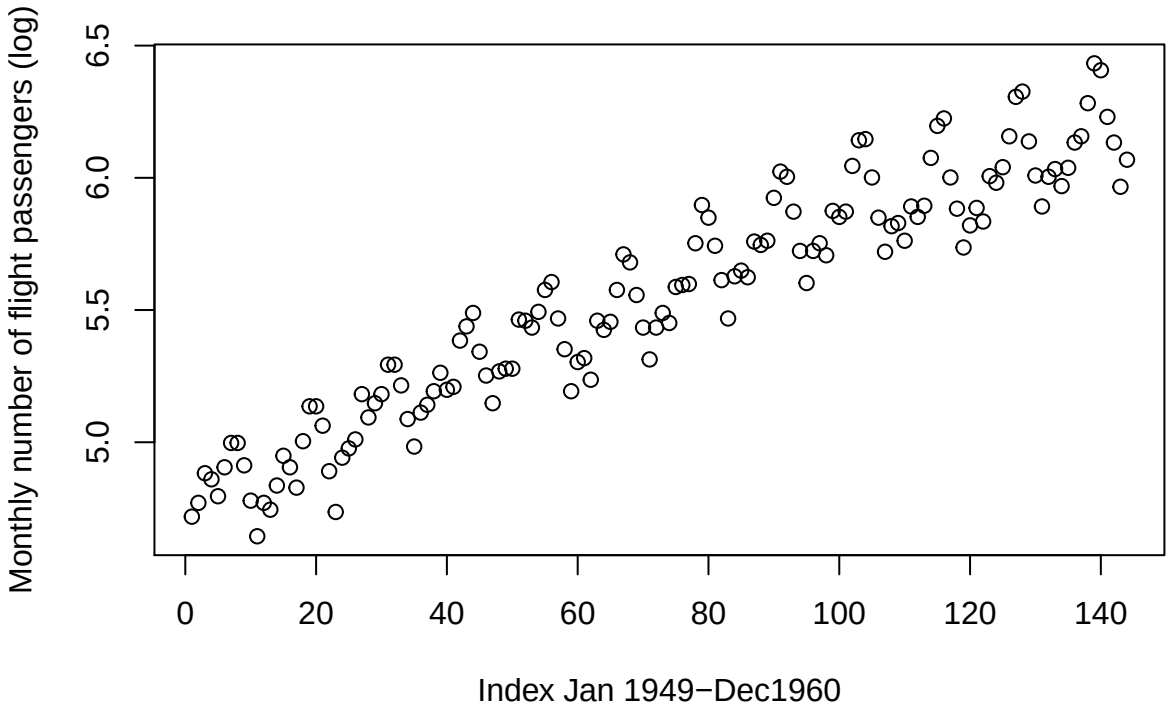
```
plot(airline, xlab="Index Jan 1949-Dec1960", ylab="Monthly number of flight passengers")
```



The monthly number of flight passengers in the USA increases from Jan 1949 to Dec 1960 (144 months). At the same time the spread between one timestamp and the next increases, meaning that for larger x values the spread in the y values increases.

b) Compute the logarithm of the data and plot against time. Comment on the difference.

```
log.airline <- log(airline)
plot(log.airline , xlab="Index Jan 1949-Dec1960", ylab="Monthly number of flight passengers (log)")
```



The monthly number of flight passengers in the USA increases from Jan 1949 to Dec 1960. However in this representation the spread in y is not increasing with increasing x.

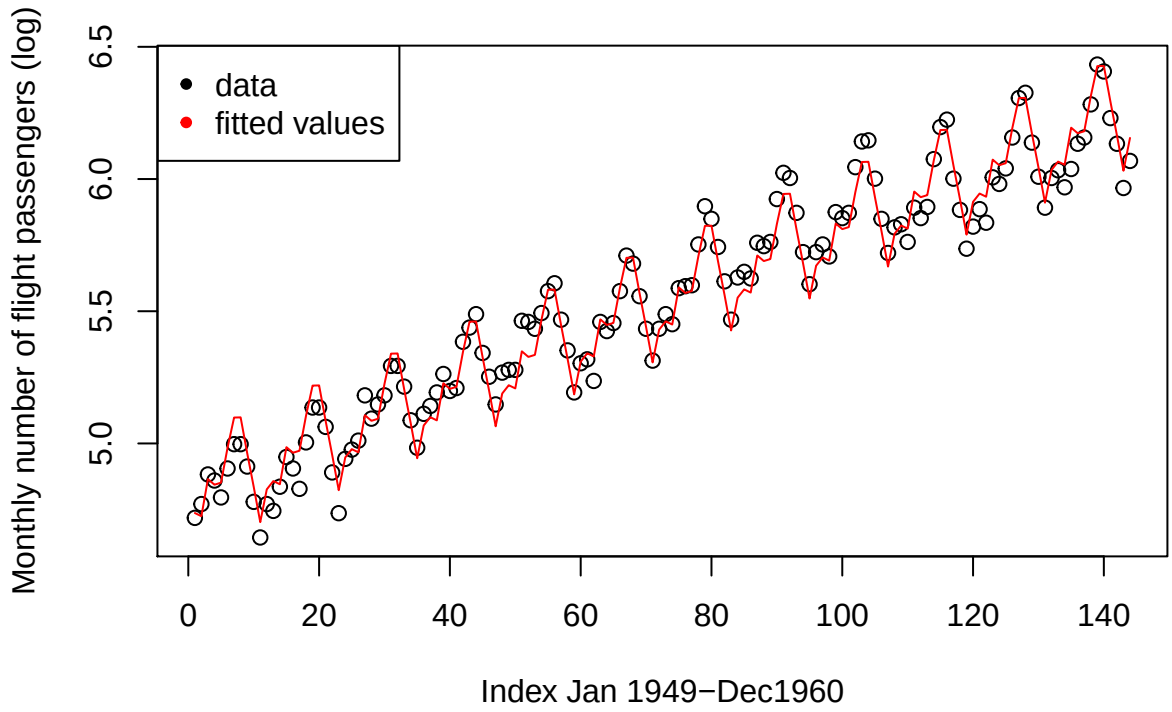
c) Define a linear model of the form ...

$$log(y_t) = \beta * t + \sum_{j=1}^{12} \gamma_j * x_{tj} + \epsilon_t$$

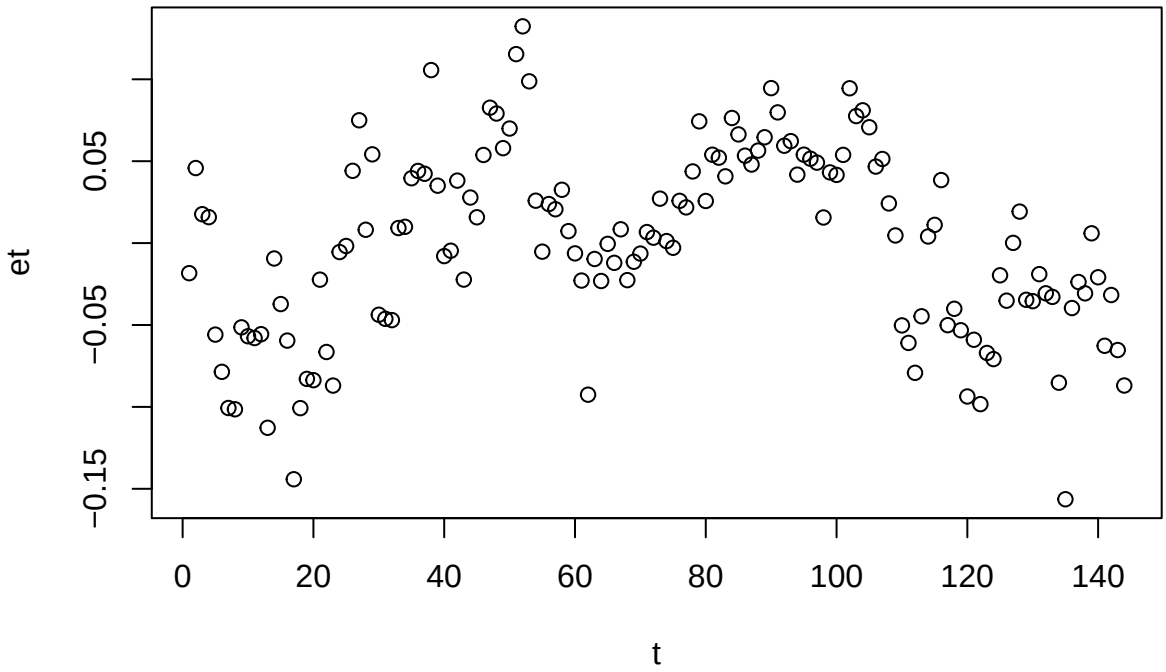
```
x1<-rep(c(1,rep(0,11)),12)
x2<-rep(c(0,1,rep(0,10)),12)
x3<-rep(c(0,0,1,rep(0,9)),12)
x4<-rep(c(0,0,0,1,rep(0,8)),12)
x5<-rep(c(0,0,0,0,1,rep(0,7)),12)
x6<-rep(c(0,0,0,0,0,1,rep(0,6)),12)
x7<-rep(c(rep(0,6),1,rep(0,5)),12)
x8<-rep(c(rep(0,7),1,rep(0,4)),12)
x9<-rep(c(rep(0,8),1,rep(0,3)),12)
x10<-rep(c(rep(0,9),1,rep(0,2)),12)
x11<-rep(c(rep(0,10),1,0),12)
x12<-rep(c(rep(0,11),1),12)
t<-1:144
fit <- lm(log.airline~-1+t+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12)
s <- summary(fit)
```

d) Plot the fitted values and residuals against time. Do you think that the model assumptions hold?

```r
# fitted values
yhat <- fitted.values(fit)
# residuals
et <-residuals(fit)
plot(log.airline , xlab="Index Jan 1949-Dec1960", ylab="Monthly number of flight passengers (log)")
lines(t, yhat, col=c(2))
legend("topleft", c("data", "fitted values"), col=c(1,2), pch=20)
```




```r
plot(t, et, main="Residuals vs time")
```



I do not think that the model assumptions hold. For low and high values of t the residuals are small (negative) and for medium values of t the residuals are large. This means that $E(\epsilon) \neq 0$.

e) Give an interpretation of the parameter beta in the above model if we consider the original scale.

It is

$$log(y_t) = \beta * t + \sum_{j=1}^{12} \gamma_j * x_{tj} + \epsilon_t$$

Hence for the expected value

$$\hat{y_t} = exp(\hat{\beta} * t + \sum_{j=1}^{12} \hat{\gamma_j} * x_{tj})$$

and for increasing t by 12, $t_1 = t + 12$:

$$\hat{y_{t+12}} = exp(\hat{\beta} * (t + 12) + \sum_{j=1}^{12} \hat{\gamma_j} * x_{tj})$$

$$= exp(\hat{\beta} * t + \hat{\beta} * 12 + \sum_{j=1}^{12} \hat{\gamma_j} * x_{tj})$$

$$= exp(\hat{\beta} * 12) * exp(\hat{\beta} * t + \sum_{j=1}^{12} \hat{\gamma}_j * x_{tj})$$

$$= exp(\hat{\beta} * 12) * \hat{y}_t$$

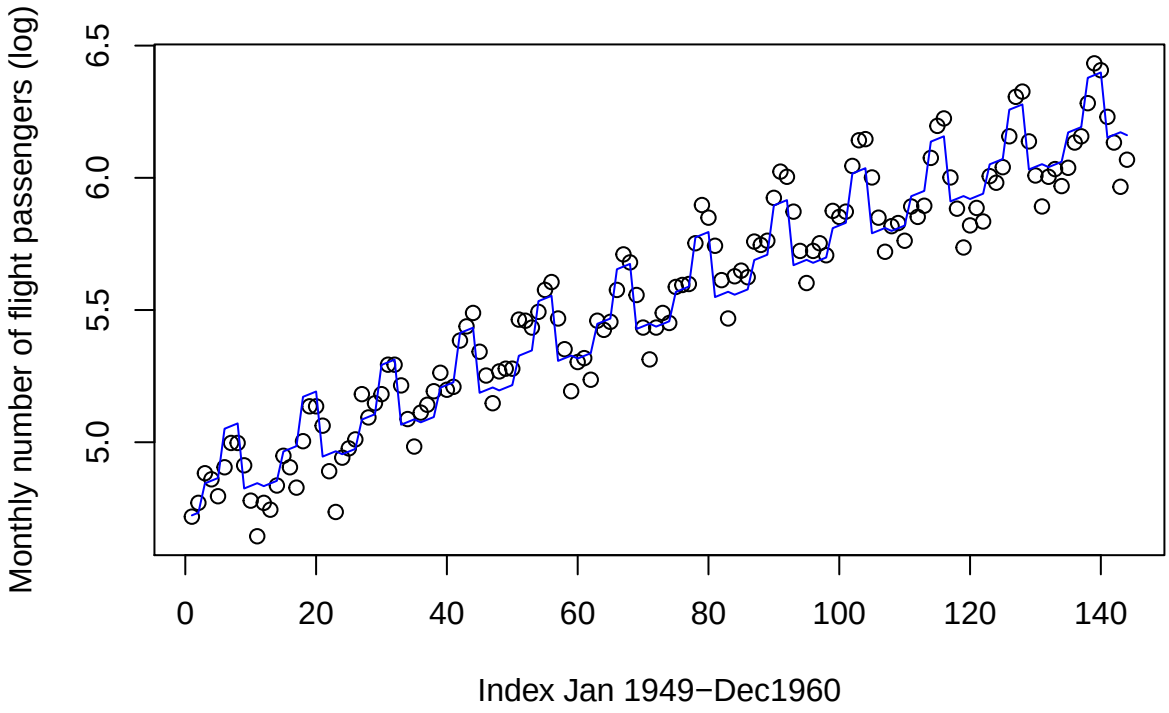
This means that the β parameter in the above model (considering the original scale) can be seen as some sort of slope of a "regression"" line, the "trend" of the number of passengers versus the time.

f) Conduct a partial F-test to check whether we can use four predictors indicating the seasons s1 , · · · , s4 (s1 for spring (month 3,4,5),. . . , s4 for winter (month 12,1,2)) instead of twelve indicators x1, · · · , x12 encoding the month.

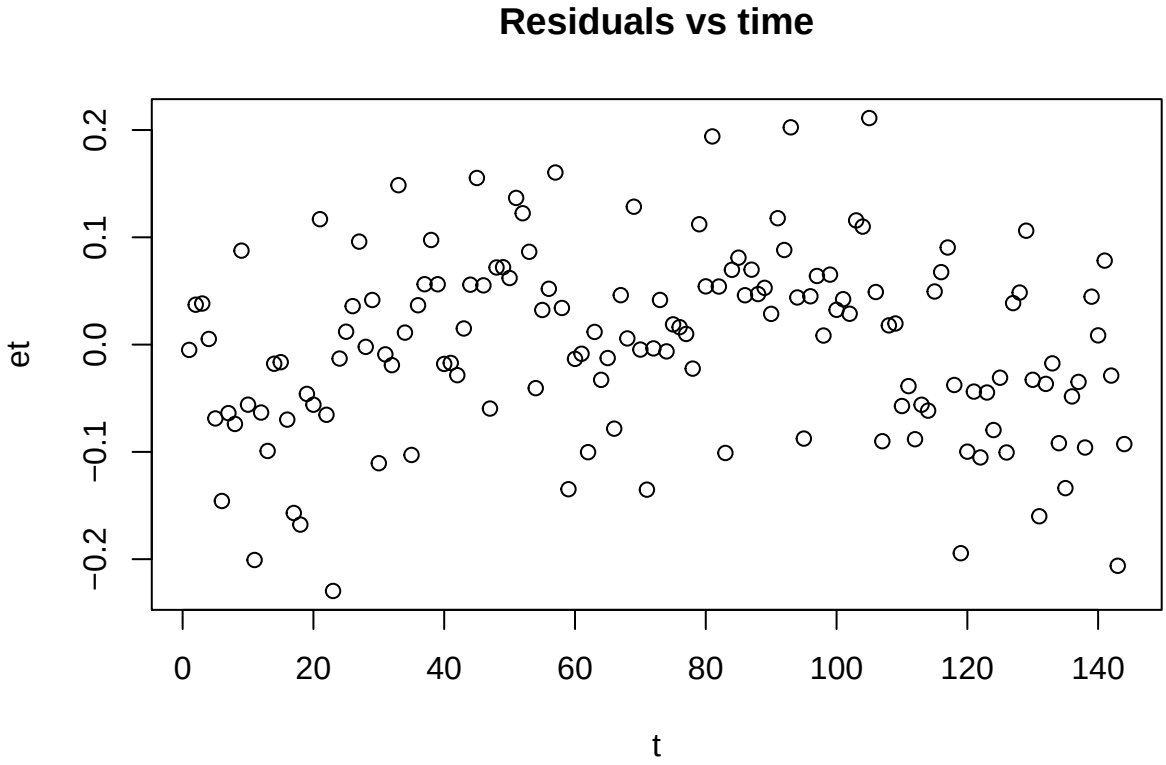
```
s1 <- rep(c(rep(0,2),rep(1,3),rep(0,7)),12) # spring
s2 <- rep(c(rep(0,5),rep(1,3),rep(0,4)),12) # summer
s3 <- rep(c(rep(0,8),rep(1,3),0),12) # herbst
s4 <- rep(c(1,1,rep(0,9),1),12) # summer
fit.seasons <- lm(log.airline~-1+t+s1+s2+s3+s4)
(s.seasons <- summary(fit.seasons))
```

```
##
## Call:
## lm(formula = log.airline ~ -1 + t + s1 + s2 + s3 + s4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2296 -0.0564  0.0071  0.0541  0.2112
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## t  0.010054   0.000173    58.2   <2e-16 ***
## s1 4.814345   0.018759   256.6   <2e-16 ***
## s2 4.990727   0.019097   261.3   <2e-16 ***
## s3 4.734576   0.019443   243.5   <2e-16 ***
## s4 4.713366   0.018871   249.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.086 on 139 degrees of freedom
## Multiple R-squared:     1,   Adjusted R-squared:     1
## F-statistic: 1.2e+05 on 5 and 139 DF,  p-value: <2e-16
```

```
# fitted values
yhat <- fitted.values(fit.seasons)
# residuals
et <-residuals(fit.seasons)
plot(log.airline , xlab="Index Jan 1949-Dec1960", ylab="Monthly number of flight passengers (log)")
lines(t, yhat, col=c(4))
```




```
plot(t, et, main="Residuals vs time")
```




```
# and conducting a partial F-test:
(fit.anova <- anova(fit.seasons, fit) )
```

```
## Analysis of Variance Table
##
## Model 1: log.airline ~ -1 + t + s1 + s2 + s3 + s4
```

```
## Model 2: log.airline ~ -1 + t + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 +
##     x9 + x10 + x11 + x12
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1    139 1.029
## 2    131 0.461  8     0.568 20.2 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The partial F test has a very low p value indicating that we cannot use the four predictors of the seasons but should use the model with the 12 indicators encoding each month.