STAT 210
Applied Statistics and Data Analysis:
Homework 8

Due on November 23/2025

You cannot use artificial intelligence tools to solve this homework.

Show complete solutions to get full credit. Writing code is not enough to answer a question. Your comments are more important than the code. Do not write comments in chunks. Label your graphs appropriately

For all tests in this homework use a significance level of $\alpha = 0.02$.

Question 1

A labor economist studies weekly wages (`Y`) and how they depend on several worker attributes. The researcher collected a sample of workers and recorded 7 candidate predictors. Your job is to build a regression model. The data is available in the file `HW825FQ1.csv`.

```
q1_data <- read.csv("HW825FQ1.csv")

str(q1_data)
```

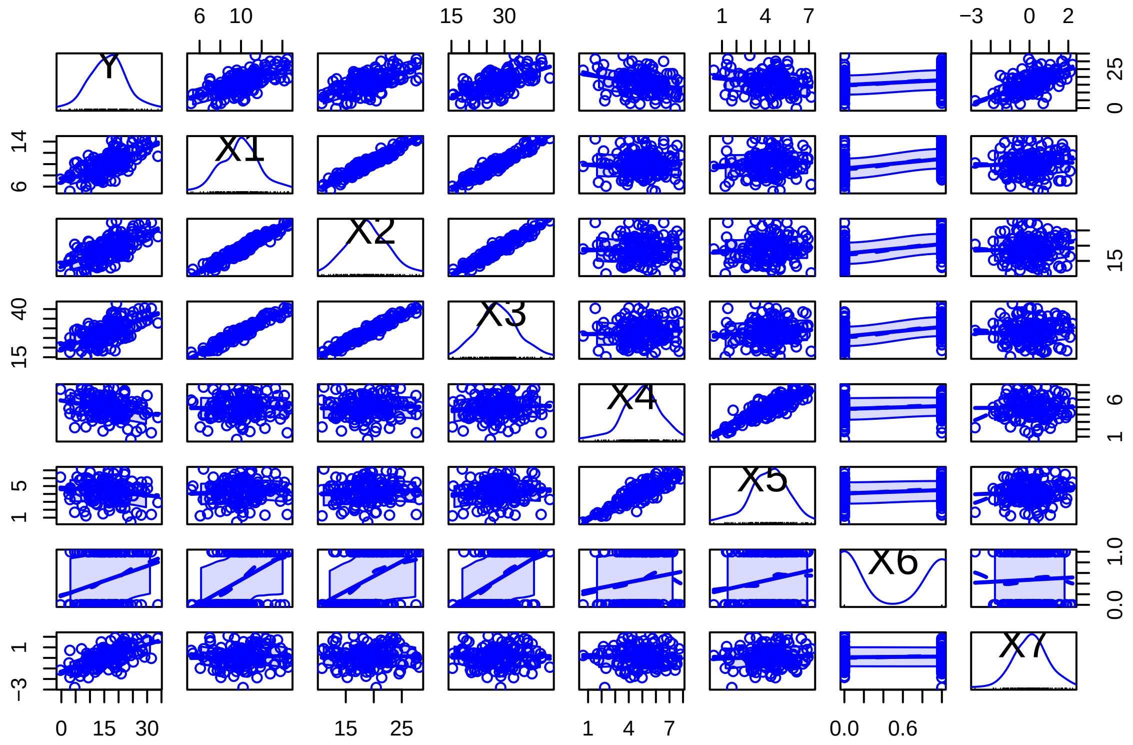
```
## 'data.frame':    150 obs. of  8 variables:
##  $ Y : num  28.3 17.1 18.9 16.8 18.3 ...
##  $ X1: num  13.28 12.05 10.44 9.74 13.32 ...
##  $ X2: num  24.2 21.8 18.5 19.5 24.7 ...
##  $ X3: num  36.3 34.5 27.9 29.3 38.5 ...
##  $ X4: num  5.3 5.58 3.26 4.92 7.55 5.76 5.13 5.68 5.1 4.77 ...
##  $ X5: num  4.74 4.13 2.98 2.87 6.59 4.89 4.65 5.06 3.3 3.81 ...
##  $ X6: int  1 1 0 0 0 1 1 1 1 1 ...
##  $ X7: num  0.61 0.51 -0.5 -0.14 -0.57 1.83 0.77 -0.05 -0.88 1.33 ...
```

(a) Do a exploratory analysis of this data, including a scatterplot matrix and a graphical representation of the correlation matrix. Comment on your results.

```
library(car)
```

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

```r
scatterplotMatrix(q1_data)
```



we can see from the scatterplot matrix that X1 X2 and X3 seem to have positive correlation indicating they have a very similar effect on y.

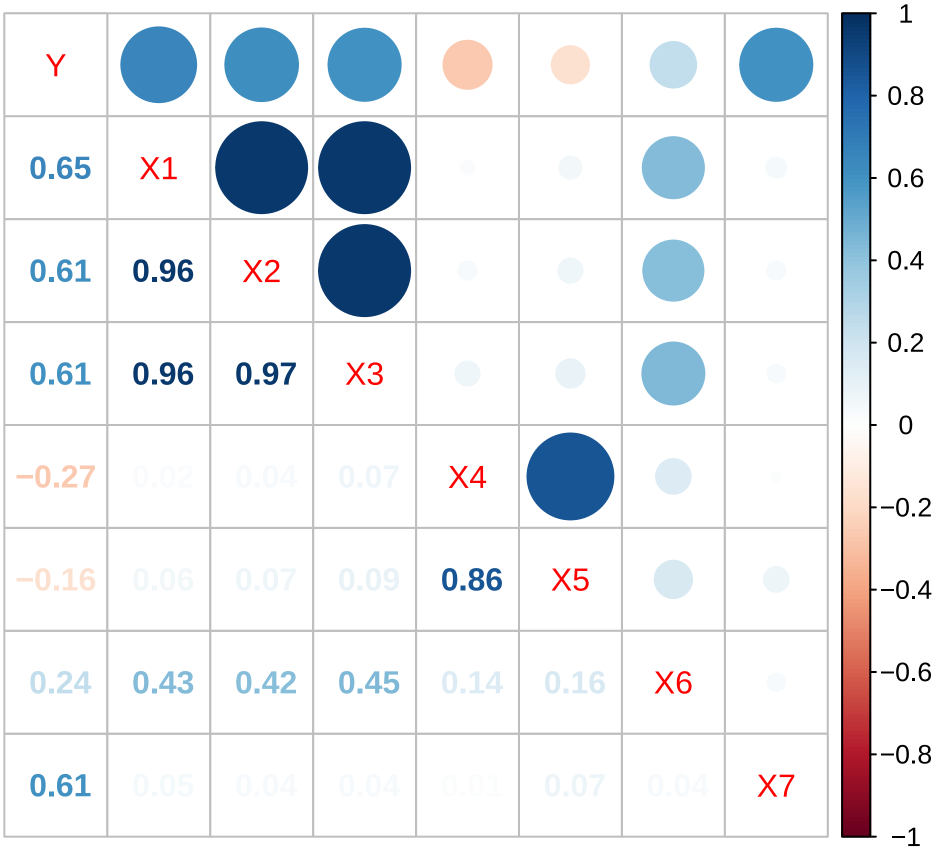
We also see that in general Y is linearly affected by all variables, with the exception of x6.

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.5.2
```

```
## corrplot 0.95 loaded
```

```r
cor_1 <- cor(q1_data)
corrplot.mixed(cor_1)
```



same conclusions we got from earlier with x1,x2,x3 having very similar effects on Y.

we also see strong positive relationship of Y with x1,x2,x3 and X7, and a negative relationship with X4 and x5.

(b) Fit a complete model for `Y` including all the other variables. Produce a summary table and interpret the t tests in the table. What is the p-value for the overall significance test for the regression?

```
full_model <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, data = q1_data)
summary(full_model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, data = q1_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3504 -1.7645 -0.0845  1.7590  6.3147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.71368    1.41503   1.211    0.228
## X1           2.26138    0.46965   4.815 3.73e-06 ***
## X2          -0.08099    0.25043  -0.323    0.747
## X3          -0.01075    0.17219  -0.062    0.950
## X4          -1.49758    0.30950  -4.839 3.37e-06 ***
## X5           0.18949    0.32162   0.589    0.557
## X6          -0.21165    0.48852  -0.433    0.665
## X7           4.09896    0.24099  17.009  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.64 on 142 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.832
## F-statistic: 106.4 on 7 and 142 DF,  p-value: < 2.2e-16
```

from the t tests we see that variables x2, x3, x5, x6 seem to not be significant to the model (p > 0.02) also the p-value of the entire model is <2.2e-16, thus we reject the null hypothesis and say that at least one variable is useful for predicting Y

(c) Check for multicollinearity and drop variables as needed until this problem is resolved. Use a threshold value of 2.

```r
vif(full_model)
```

```
##        X1        X2        X3        X4        X5        X6        X7
## 18.244189 18.874281 19.975250  3.908086  3.911549  1.280474  1.018029
```

```r
mod1 <- update(full_model, .~. - X3)
vif(mod1)
```

```
##        X1        X2        X4        X5        X6        X7
## 13.974219 13.737647  3.876618  3.910763  1.263552  1.017298
```

```r
mod2 <- update(mod1, .~. - X1)
vif(mod2)
```

```
##       X2       X4       X5       X6       X7
## 1.217712 3.866653 3.909562 1.242589 1.016281
```

```r
mod3 <- update(mod2, .~. - X5)
vif(mod3)
```

```
##       X2       X4       X6       X7
## 1.216255 1.020561 1.238332 1.002032
```

now all remaining variables are under the threshold so we keep them.

(d) Starting with the model obtained in section (c), get a minimal model using a backward selection procedure with a critical α of 0.1. Use the function `drop1` for this.

```r
drop1(mod3, test = "F")
```

```
## Single term deletions
##
## Model:
## Y ~ X2 + X4 + X6 + X7
##        Df Sum of Sq    RSS    AIC  F value    Pr(>F)
## <none>              1201.3 322.07
## X2      1   1820.52 3021.8 458.44 219.7494 < 2.2e-16 ***
## X4      1    548.85 1750.1 376.52  66.2499  1.65e-13 ***
## X6      1      0.64 1201.9 320.15   0.0776     0.781
## X7      1   2110.46 3311.7 472.19 254.7464 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
mod4 <- update(mod3, .~. - X6)
drop1(mod4, test = "F")
```

```
## Single term deletions
##
## Model:
## Y ~ X2 + X4 + X7
##        Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>              1201.9 320.15
## X2      1   2242.48 3444.4 476.08 272.403 < 2.2e-16 ***
## X4      1    554.03 1755.9 375.02  67.301 1.108e-13 ***
## X7      1   2112.79 3314.7 470.32 256.650 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(mod4)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X4 + X7, data = q1_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1147 -2.0446  0.0755  1.9858  6.7435
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.24369    1.44380   2.247   0.0262 *
## X2           1.03556    0.06274  16.505  < 2e-16 ***
## X4          -1.39703    0.17029  -8.204 1.11e-13 ***
## X7           4.16202    0.25980  16.020  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.869 on 146 degrees of freedom
## Multiple R-squared:  0.8055, Adjusted R-squared:  0.8015
## F-statistic: 201.6 on 3 and 146 DF,  p-value: < 2.2e-16
```

since all values are > alpha_critical, we found our minimal adequate model. using variables X2, X4, X7 to predict Y

(e) Starting with the full model obtained in section (b), fit a model using Akaike's Information Criterion (AIC). You can use the function `stepAIC` in the `MASS` library or the function `step` in the base package. Check the resulting model for multicollinearity. Compare with the model obtained in (d).

```r
library(MASS)
aic_model <- stepAIC(full_model)
```

```
## Start:  AIC=299.01
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7
##
##        Df Sum of Sq     RSS    AIC
## - X3    1      0.03  989.67 297.01
## - X2    1      0.73  990.37 297.12
## - X6    1      1.31  990.95 297.20
## - X5    1      2.42  992.06 297.37
## <none>               989.64 299.01
## - X1    1    161.58 1151.22 319.69
## - X4    1    163.17 1152.81 319.90
## - X7    1   2016.23 3005.88 463.65
##
## Step:  AIC=297.01
## Y ~ X1 + X2 + X4 + X5 + X6 + X7
##
##        Df Sum of Sq     RSS    AIC
## - X2    1      1.21  990.88 295.19
## - X6    1      1.37  991.04 295.22
## - X5    1      2.43  992.09 295.38
## <none>               989.67 297.01
## - X4    1    164.87 1154.54 318.12
## - X1    1    208.31 1197.98 323.66
## - X7    1   2018.08 3007.75 461.75
##
## Step:  AIC=295.19
## Y ~ X1 + X4 + X5 + X6 + X7
##
##        Df Sum of Sq     RSS    AIC
## - X6    1      1.37  992.25 293.40
## - X5    1      2.45  993.33 293.56
## <none>               990.88 295.19
## - X4    1    166.11 1157.00 316.44
## - X1    1   2020.10 3010.98 459.91
## - X7    1   2021.53 3012.41 459.98
##
## Step:  AIC=293.4
## Y ~ X1 + X4 + X5 + X7
##
##        Df Sum of Sq    RSS    AIC
## - X5    1      2.26  994.5 291.74
## <none>               992.3 293.40
## - X4    1    167.01 1159.3 314.74
## - X7    1   2020.76 3013.0 458.01
## - X1    1   2427.88 3420.1 477.02
##
## Step:  AIC=291.74
## Y ~ X1 + X4 + X7
##
##        Df Sum of Sq    RSS    AIC
## <none>               994.5 291.74
## - X4    1    522.97 1517.5 353.12
## - X7    1   2065.90 3060.4 458.35
## - X1    1   2449.87 3444.4 476.08
```

```
summary(aic_model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X4 + X7, data = q1_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4827 -1.7426 -0.0806  1.8846  6.3747
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8711     1.3350   1.401    0.163
## X1            2.0644     0.1089  18.965  < 2e-16 ***
## X4           -1.3567     0.1548  -8.762 4.45e-15 ***
## X7            4.1170     0.2364  17.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 146 degrees of freedom
## Multiple R-squared:  0.8391, Adjusted R-squared:  0.8358
## F-statistic: 253.8 on 3 and 146 DF,  p-value: < 2.2e-16
```

this model usees X1, X4, and X7 to predict Y with an R squared value of 0.8391 the model from part d uses X2 instead of X1 and has an R squared value of 0.8055, this means the AIC model is a better fit to the data

(f) Starting with the full model obtained in section (b), fit a model by maximizing the adjusted R^2. Check the resulting model for multicollinearity. Compare with the models obtained in (d) and (e). Are all the models the same? If not, which one would you choose and why?

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.5.2
```

```
subsets <- regsubsets(q1_data[ , -1], q1_data[,1])
summary(subsets)
```

```
## Subset selection object
## 7 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## X3     FALSE      FALSE
## X4      FALSE      FALSE
## X5      FALSE      FALSE
## X6      FALSE      FALSE
## X7      FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          X1  X2  X3  X4  X5  X6  X7 
## 1  ( 1 ) "*" " " " " " " " " " " " "
## 2  ( 1 ) "*" " " " " " " " " " " "*"
## 3  ( 1 ) "*" " " " " "*" " " " " "*"
## 4  ( 1 ) "*" " " " " "*" "*" " " "*"
## 5  ( 1 ) "*" " " " " "*" "*" "*" "*"
## 6  ( 1 ) "*" "*" " " "*" "*" "*" "*"
## 7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
```

```
summary(subsets)$adjr2
```

```
## [1] 0.4198368 0.7511401 0.8357875 0.8350301 0.8341137 0.8331582 0.8319879
```

```
best_adjR2_model <- lm(Y ~ X1 + X4 + X7, data = q1_data)
vif(best_adjR2_model)
```

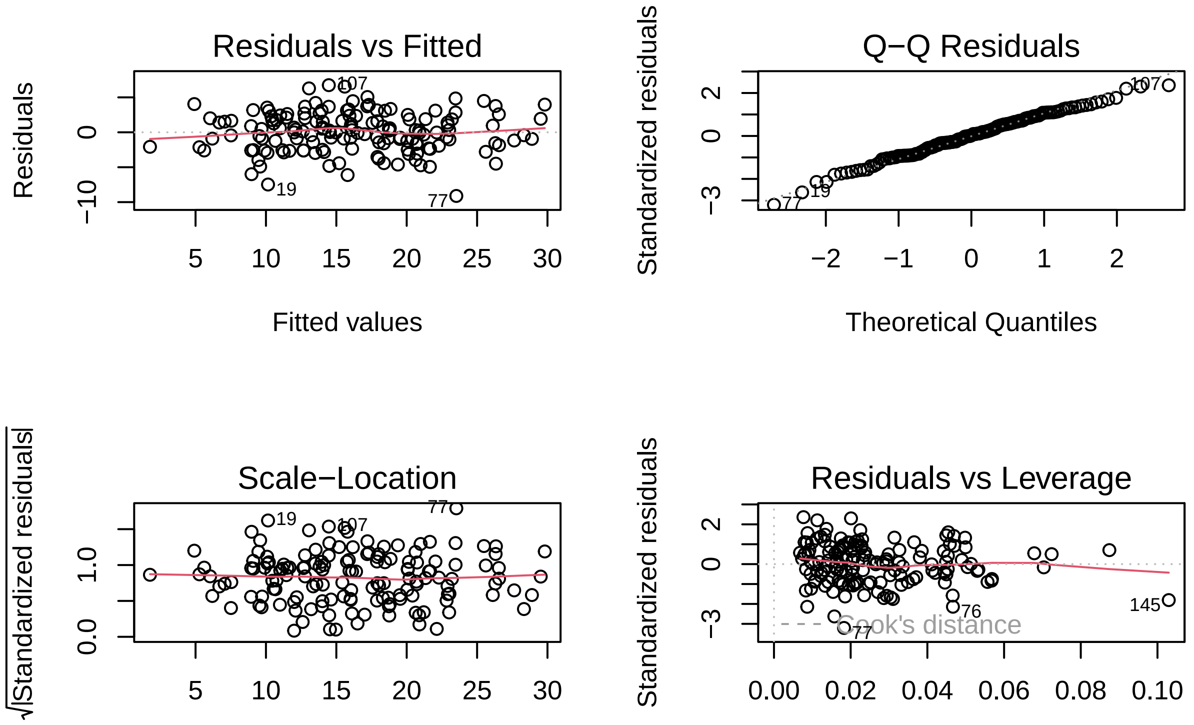
```
##       X1       X4       X7 
## 1.002760 1.000662 1.002344
```

we get that the maximum adjusted R value is 0.8357875 which is the 3 predictor model which includes predictors X1, X4, and X7. also checking for multicollinearity with threshold 2, we see that all predictors are below that so no issues here. This model is an exact replica of the AIC_model, which means we already got the best fitting model using the AIC method. and of course it follows the same comparison to part D which used X2 instead of X1 and has a lower R^2 and adj R^2 values.

Regarding choice i would choose the model we found with best R^2 and AIC model which is the same, this gives us the best fit with the best predictors.

(g) Plot the standard diagnostic graphs for the model that you fitted in (d) and comment on what you observe. Use also the Shapiro-Wilk and ncv tests and comment on the results.

```
par(mfrow = c(2, 2))
plot(mod4)
```



```r
par(mfrow = c(1, 1))
```

we see from residuals vs fitted that the line is nearly horizantal and is near 0, indicating linearity. the QQ residuals plot shows the points roughly lying on the line, indicating normality the scale-location plot is also horizantal with an even spread of points which indicates that we have equal variances. residuals vs leverage shows a few points with higher leverage but no signficant influential points.

```r
shapiro.test(residuals(mod4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod4)
## W = 0.99282, p-value = 0.6585
```

```r
ncvTest(mod4)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.04541626, Df = 1, p = 0.83124
```

with a p-value of 0.6585 in shapiro test we can say that the resdiuals are normally distributed with a p-value of 0.83124 in the ncv test we can say that the resdiausl have equal variances.

(h) Predict the `Y` value for a subject with covariates

(`X1, X2, X3, X4, X5, X6, X7`) = (15.2, 9.6, 10.7, 9.2, 5.33, 1, -0.8)

using the model you fitted in (d). Add a confidence interval at level 98%.

```r
predict_data <- data.frame(X2 = 9.6, X4 = 9.2, X7 = -0.8)
predict(mod4, newdata = predict_data, interval = "c", level = 0.98)
```

```
##         fit      lwr        upr
## 1 -2.997233 -5.34565 -0.6488172
```

Question 2

A city transportation department is studying bike rental duration. They believe that the effect of age on rental duration differs between:

- Casual riders (`member` = 0)
- Registered members (`member` = 1)

To study this, they collect data on 30 riders. The data is in the file `HW825FQ2.csv`. Remember to transform `member` into a factor.

```r
q2_data <- read.csv("HW825FQ2.csv")
str(q2_data)
```

```
## ’data.frame’:    30 obs. of  3 variables:
##  $ age     : int  51 24 41 24 39 37 29 31 34 41 ...
##  $ duration: num  101.1 63.8 87.9 61.8 83.1 ...
##  $ member  : int  0 0 0 0 0 0 0 0 0 0 ...
```

```r
q2_data$member <- as.factor(q2_data$member)
str(q2_data)
```

```
## ’data.frame’:    30 obs. of  3 variables:
##  $ age     : int  51 24 41 24 39 37 29 31 34 41 ...
##  $ duration: num  101.1 63.8 87.9 61.8 83.1 ...
##  $ member  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

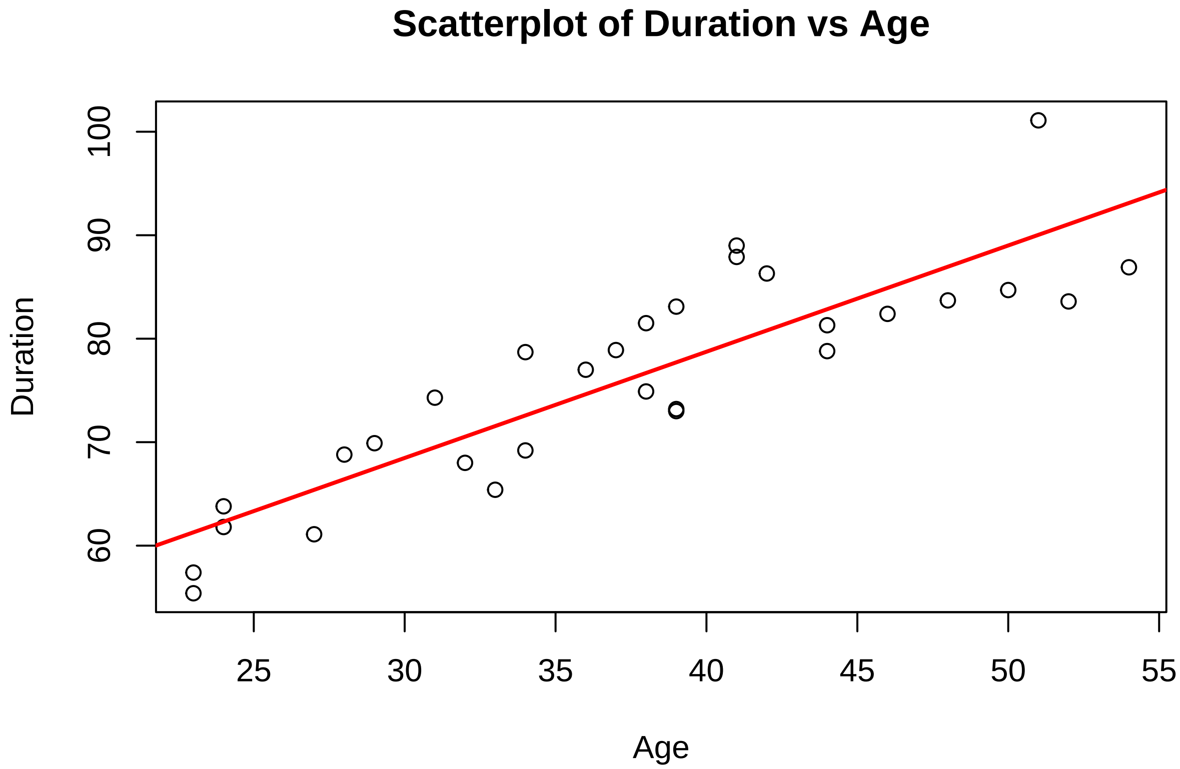
(a) Fit a simple regression model for `duration` in terms of `age`. Print the summary table and comment on the results. Draw a scatterplot and add the regression line. Comment. Using diagnostic plots and test, check whether the assumptions for the model are satisfied. Predict the value of `duration` for a value of `age = 45` with this model and include confidence intervals at the 98% level.

```r
mod_a <- lm(duration ~ age, data = q2_data)
summary(mod_a)
```

```
##
## Call:
## lm(formula = duration ~ age, data = q2_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.462 -4.231 -1.667  4.410 11.065
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6696     4.1689   9.036 8.59e-10 ***
## age           1.0268     0.1086   9.454 3.27e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.228 on 28 degrees of freedom
## Multiple R-squared:  0.7615, Adjusted R-squared:  0.7529
## F-statistic: 89.38 on 1 and 28 DF,  p-value: 3.273e-10
```

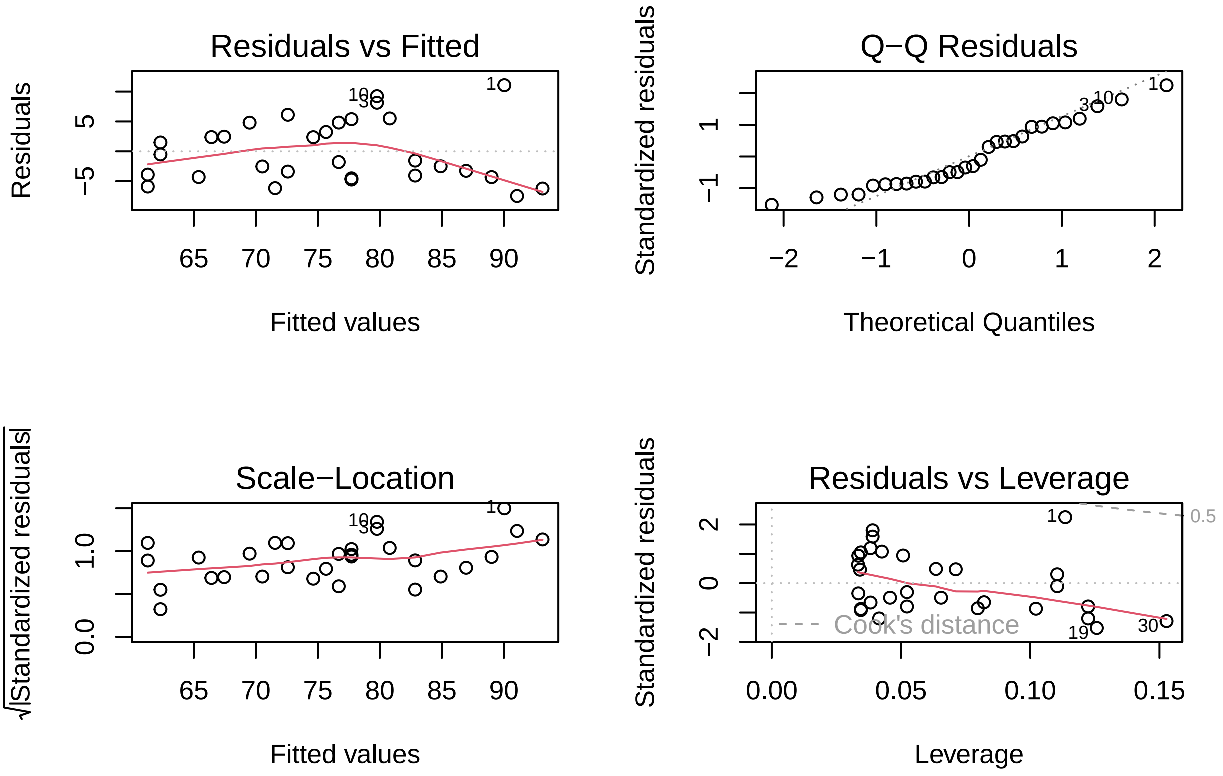
we can see the p-values being very low indicating the coefficients != 0, we also see the adj r^2 value of 0.7529

```
plot(q2_data$age, q2_data$duration,
xlab = "Age",
ylab = "Duration",
main = "Scatterplot of Duration vs Age")
abline(mod_a, col = "red", lwd = 2)
```



we see that the data points show a positive relationship between duration and age, we also see that the model fits that general positive trend.

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



```
par(mfrow = c(1, 1))
```

residuals vs fitted shows a very curved line with an inverted U shaped, this indicaites non-linearity. q-qresiduals shows the points lying close to the line but not perfectly fitting, indicating non-normality scale-location is mostly horizantal with a small positive slope, indicates constant variance but requires further testing residuals vs leverage shows a few points with high leverage (1, 19) but no influential points.

```
shapiro.test(residuals(mod_a))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod_a)
## W = 0.9343, p-value = 0.06394
```

```
ncvTest(mod_a)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.812606, Df = 1, p = 0.093526
```

Shapiro test p-value of 0.06394 indicates linearity using since our signficance level is 0.02 and 0.06394 > 0.02 ncv test p-value of 0.093 > 0.02 which means we have constant variance at that level

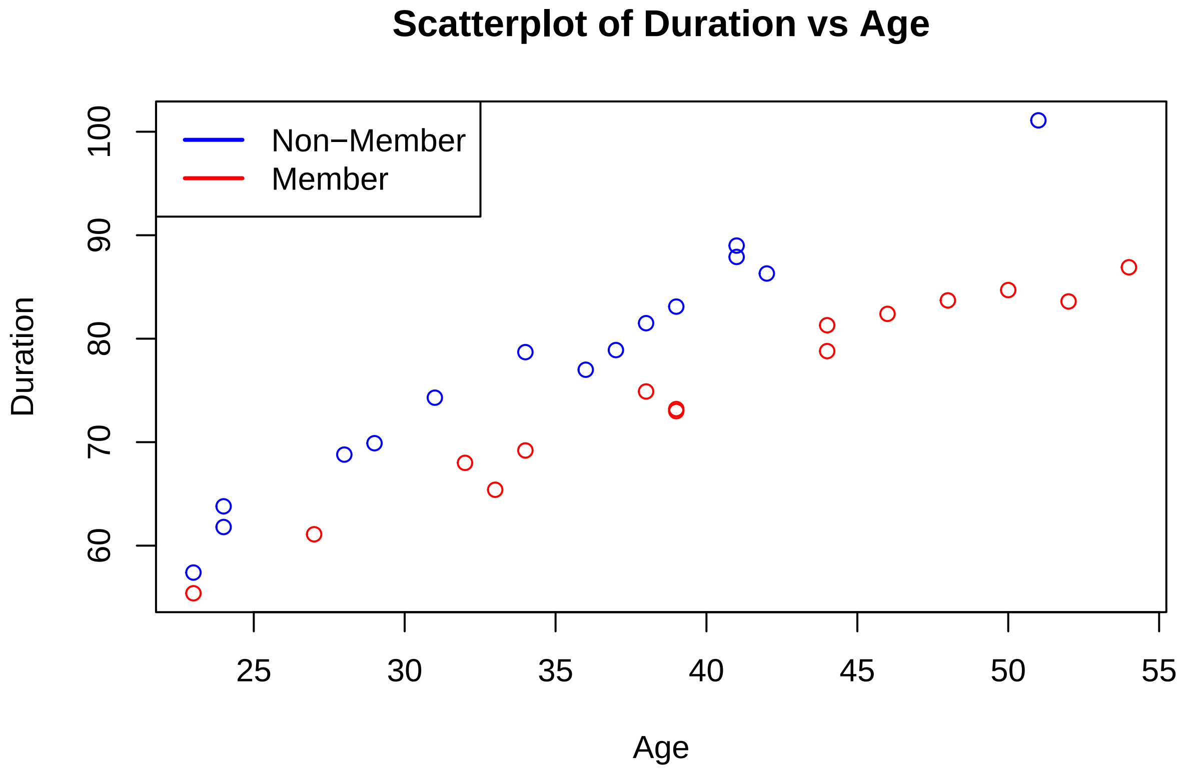
```
predict(mod_a, newdata = data.frame(age = 45), interval = "c", level = 0.98)
```

```
##        fit      lwr     upr
## 1 83.87436 80.75542 86.9933
```

we predict a 45 years old rider to have a duration of 83.8744, with interval [80.7554, 86.9933]

(b) Draw a new scatterplot and color the points according to the value of `member`. Comment on what you observe.

```
plot(q2_data$age, q2_data$duration,
col = c("blue", "red")[q2_data$member],
xlab = "Age",
ylab = "Duration",
main = "Scatterplot of Duration vs Age")

legend("topleft",
legend = c("Non-Member", "Member"),
col = c("blue", "red"),
lwd = 2)
```



we see that the data points seem to split into two different trendlines based on the member variable.

(c) Fit a new model for `duration` as a function of `age` and `member`, including an interaction term. Print an anova table for this model and interpret the p-values in the table. If necessary, fit a new model dropping the terms that have a non-significant p-value. Print a summary table for the final model and interpret the coefficients. What is the value for the estimated variance of the errors? What is the R^2, how do they compare with the previous model?

```r
mod_b <- lm(duration ~ age * member, data = q2_data)
anova(mod_b)
```

```
## Analysis of Variance Table
##
## Response: duration
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## age         1 2442.69 2442.69  733.49 < 2.2e-16 ***
## member      1  582.88  582.88  175.03 4.681e-13 ***
## age:member  1   95.71   95.71   28.74 1.299e-05 ***
## Residuals  26   86.59    3.33
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
summary(mod_b)
```

```
##
## Call:
## lm(formula = duration ~ age * member, data = q2_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2337 -1.5447  0.0104  1.4533  2.6837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.39758    2.16233  12.670 1.24e-12 ***
## age          1.44505    0.06111  23.646  < 2e-16 ***
## member1      6.72162    3.07271   2.188   0.0379 *
## age:member1 -0.43375    0.08091  -5.361 1.30e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.825 on 26 degrees of freedom
## Multiple R-squared:  0.973,  Adjusted R-squared:  0.9699
## F-statistic: 312.4 on 3 and 26 DF,  p-value: < 2.2e-16
```

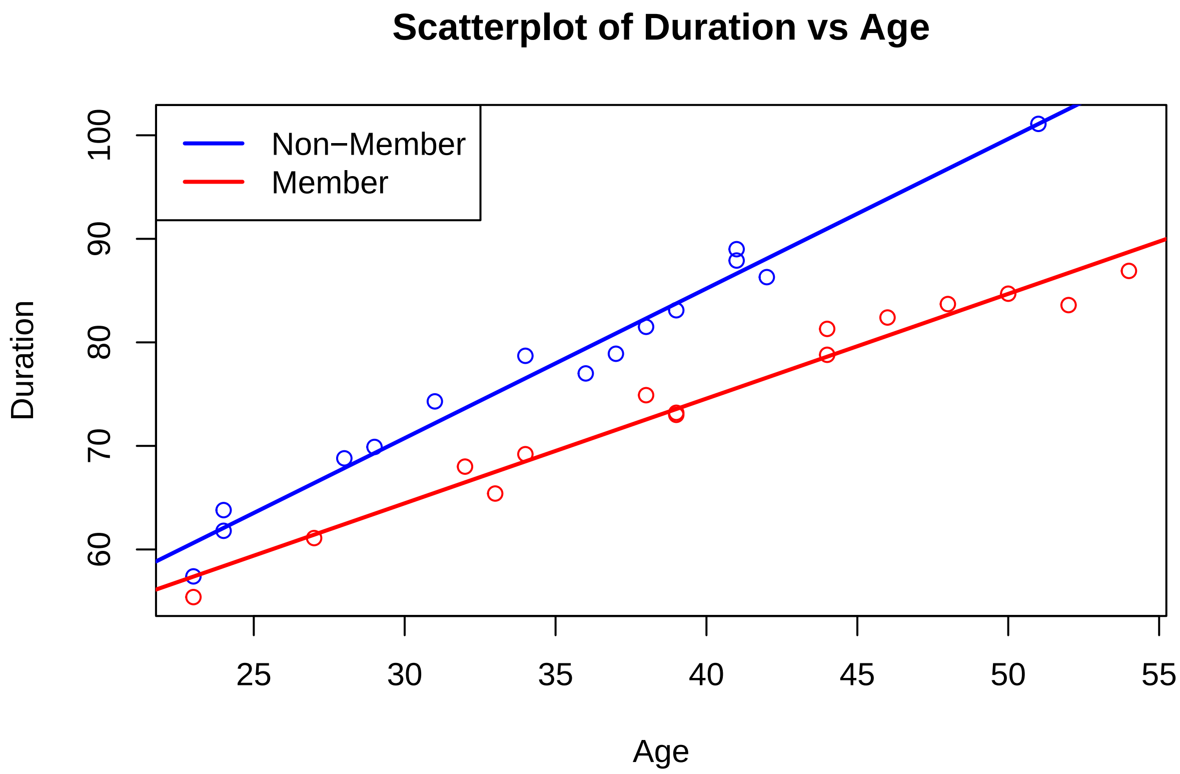
looking at the P values of the predictors, we see that age, member, and their interactions are signficant for predicting duration. therefore we don’t drop any variable and take this to be the minimal adequate fitting model.

the equation we get is: duration = beta1 + beta2 * age + beta3 * member1 + beta4 (age * member1), with member1 being a dummy variable indicating whether they are a member or not and the 4 beta values are the coefficients of the model, (b1 = 27.40, b2 = 1.45, b3 = 6.72, b4 = -0.43)

the estimated variance of errors is found to be 3.33 from the ANOVA table. R^2 is found to be 0.973 from the model summary which indicates a very good fit of the data.

(d) Draw a scatterplot and color the points according to the value of type. Add the regression lines corresponding to the model you fitted in (c). Write down an equation for this model.

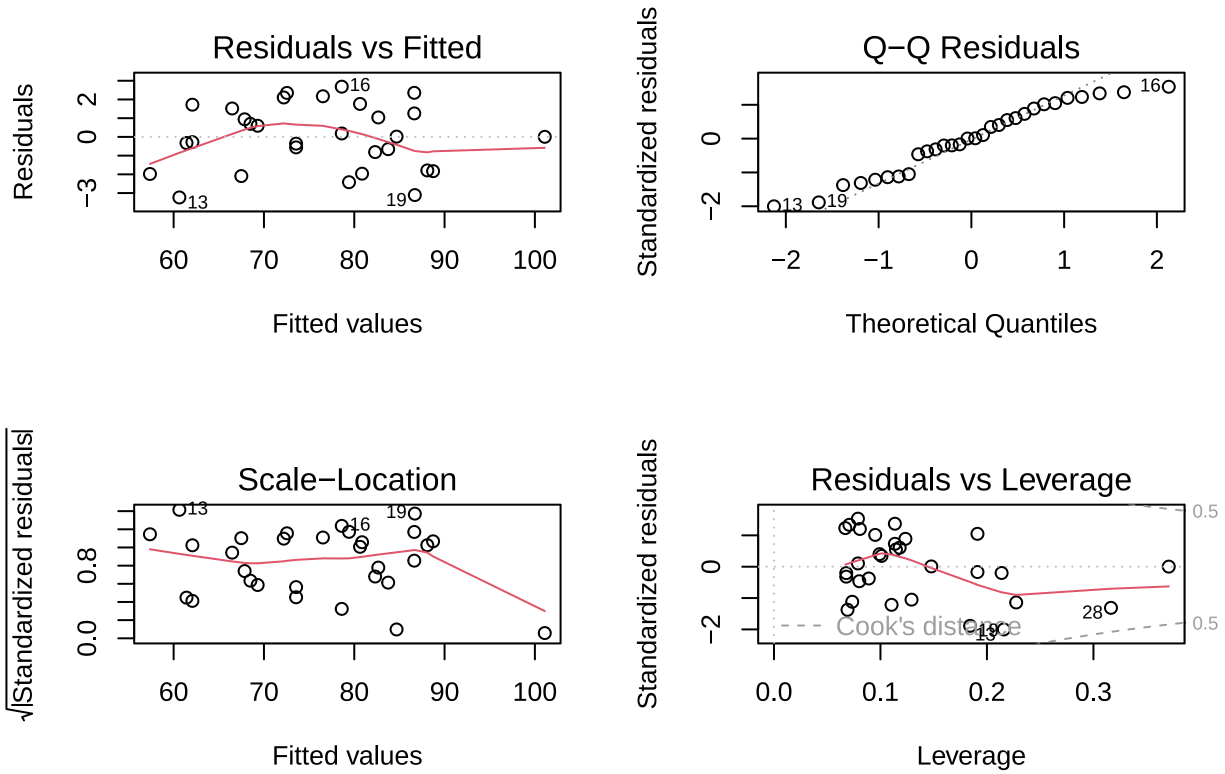
```r
plot(q2_data$age, q2_data$duration,
col = c("blue", "red")[q2_data$member],
xlab = "Age",
ylab = "Duration",
main = "Scatterplot of Duration vs Age")

legend("topleft",
legend = c("Non-Member", "Member"),
col = c("blue", "red"),
lwd = 2)

abline(a = coef(mod_b)[1], b = coef(mod_b)[2], col = "blue", lwd = 2)

abline(a = coef(mod_b)[1] + coef(mod_b)[3],
       b = coef(mod_b)[2] + coef(mod_b)[4], col = "red", lwd = 2)
```



The equation for this model is: Duration_Non_Member = 27.39758 + 1.44505 * age Duration Member = 27.39758 + 1.44505 * age + 6.72162 - 0.43375 * age = 34.1192 + 1.0113 * age

(e) Plot the standard diagnostic graphs for the model that you fitted in (c) and comment on what you observe. Use also the Shapiro-Wilk and ncv tests and comment on the results.

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



```
par(mfrow = c(1, 1))
```

from residuals vs fitted we see the line curves a lot and isn't horizantal, indicating non-linearity from qq residuals we see the points roughly fit the dashed line, indicating normality from scale-locaiton we see the line curving a lot and and has a strong slope at t he end, indicating no constant variance from residuals vs leverage we see no influential points but some points with high leveratge (13, 28)

```
shapiro.test(residuals(mod_b))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod_b)
## W = 0.95491, p-value = 0.2285
```

```
ncvTest(mod_b)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.03139018, Df = 1, p = 0.85937
```

with a high p-value for both shapiro and ncv test, we say that the residuals follow a normal distribution and have constant variance.

(f) Predict the value of `duration` for a value of `age = 45` and for the two levels of `member`. Include confidence intervals at the 98% level. Compare with the prediction in (a).

```
(predict_non <- predict(mod_b, newdata = data.frame(age = 45, member = "0"), interval = "c", level = 0.9
```

```
##        fit      lwr      upr
## 1 92.42487 90.45574 94.39399
```

```
(predict_member <- predict(mod_b, newdata = data.frame(age = 45, member = "1"), interval = "c", level =
```

```
##        fit      lwr      upr
## 1 79.62756 78.30017 80.95495
```

we predict that a non member aged 45 would have a rental duration of 92.425 with a 98% confidence interval [90.456, 94.394] and a member with the sage age to have a duration of 79.627 with a 98% confidence interval of [78.300, 80.955]