Exercise 4

Tasks 19 to 24

21.10.2023

Contents

Task 19: **2**
- 19a: Show that X^T r = 0: . . . 2
- 19b: sum is 0, where ri is the i-th residual. . . . 3
- 19c: standardized_resids . . . 4
- 19d: compute with built in function . . . 5
- 19e: studentized residuals: . . . 5
- 19f: built in studentized residuals: . . . 5

Task 20: **6**
- 20: residual vs. fitted plot . . . 6

Task 21: **9**
- 22a. In Model 1, the coefficient of TV . . . 9
- 22b. In Model 1, the intercept. . . . 9
- 22c. In Model effect of youtube advertising on sales . . . 9
- 22d. relationship between sales and youtube . . . 9
- 22e. In Model 4, the intercept. . . . 9
- 22f. In Model 4, the coefficient of (TV - $\overline{TV}$). . . . 9

Task 22: **10**
- 22a: Descriptives . . . 10
- 22b: Linear Model . . . 13
- 22c: Sum contrasts . . . 14
- 22d: Interactions . . . 15

Task 23: **16**
- 23a: Minimum age of mother 1 . . . 16
- 23b: Minimum age of mother 2 . . . 19
- 23c: Minimum age of mother 3 . . . 20

Feedback **21**

Task 19:

19a: Show that X^T r = 0:

(Thank god for R auto-formatting <3)

Consider the linear model:

$$y = X\beta + \epsilon$$

with $E(\epsilon) = 0$ and $\text{cov}(\epsilon) = \sigma^2 I_n$ for a sample of n observations. Assuming the first column of X is a vector of ones (i.e., the model contains an intercept), and we estimate the coefficients using OLS, we have:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

The residuals r from this OLS model are given by:

$$r = y - X\hat{\beta}$$

Substituting $\hat{\beta}$ we get:

$$r = y - X(X^\top X)^{-1} X^\top y$$

The residuals can also be expressed as:

$$r = (I - H)y$$

where H is the "hat" matrix, given by:

$$H = X(X^\top X)^{-1} X^\top$$

Now let's prove that $X^\top r = 0$:

$$X^\top r = X^\top (I - H)y$$

$$X^\top r = X^\top y - X^\top H y$$

Since $H = X(X^\top X)^{-1} X^\top$, we substitute H in the equation:

$$X^\top r = X^\top y - X^\top X(X^\top X)^{-1} X^\top y$$

$$X^\top r = X^\top y - X^\top y$$

$$X^\top r = 0$$

This shows that the residuals are orthogonal to the column space of X. In terms of the correlation between the covariates and the residuals, this implies there is no linear association between them, satisfying one of the Gauss-Markov assumptions for the OLS estimator to be the Best Linear Unbiased Estimator (BLUE).

19b: sum is 0, where ri is the i-th residual.

Given the residuals from an OLS regression $r_i = y_i - \hat{y}_i$, and the predicted values $\hat{y} = X\hat{\beta}$, we want to show that:

$$\sum_{i=1}^{n} r_i = 0$$

From the normal equations, we have:

$$X^\top y = X^\top X\hat{\beta}$$

Considering the first column of X is a vector of ones, denoted as x_0, then $x_0^\top y$ is the sum of the y values and $x_0^\top X\hat{\beta}$ is the sum of the predicted values $\hat{y}$. Therefore:

$$x_0^\top y = x_0^\top X\hat{\beta}$$

The sum of the residuals is given by:

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

And using the normal equations:

$$\sum_{i=1}^{n} r_i = x_0^\top y - x_0^\top X\hat{\beta}$$

$$\sum_{i=1}^{n} r_i = x_0^\top y - x_0^\top y$$

$$\sum_{i=1}^{n} r_i = 0$$

This shows that the sum of the residuals in an OLS regression model with an intercept term is always equal to zero.

19c: standardized_resids

```
data("mtcars", package = "datasets")
y <- mtcars$mpg
X1 <- mtcars$disp
X2 <- mtcars$hp
m <- lm(y ~ X1 + X2)
r <- residuals(m)
head(r)
```

```
##         1         2         3         4         5         6
## -2.148091 -2.148091 -2.348379  1.225844  3.235770 -3.199783
```

```
# Build the design matrix including an intercept
# Make sure to convert all columns to numeric explicitly if they are not already
design_matrix <- cbind(rep(1,length(X2)), X1, X2)


standardized_resids <- function(residuals, design_matrix) {
  df <- nrow(design_matrix) - ncol(design_matrix) # Calculate degrees of freedom

  sigma_hat_squared <- sum(residuals^2) / df # Estimate variance of the residuals
  sigma_hat <- sqrt(sigma_hat_squared) # use sqrt to get standard deviation

  H <- design_matrix %*% solve(t(design_matrix) %*% design_matrix) %*% t(design_matrix)
  leverage <- diag(H) # Extract the diagonal
  standardized_residuals <- residuals / (sigma_hat * sqrt(1 - leverage)) # Calculate standardized resid

  return(standardized_residuals)
}

# Calculate standardized residuals
standardized_res <- standardized_resids(r, design_matrix)

# Check the first few standardized residuals
standardized_res
```

```
##          1          2          3          4          5          6          7
## -0.7020670 -0.7020670 -0.7758853  0.4079066  1.0795485 -1.0552598  0.1935267
##          8          9         10         11         12         13         14
## -0.1152271 -0.4279150 -1.1080264 -1.5650067 -0.4877968 -0.1941765 -0.8792905
##         15         16         17         18         19         20         21
## -0.3279304 -0.3616238  1.0355939  1.9015231  1.0936326  2.3221987 -1.0479571
##         22         23         24         25         26         27         28
## -0.6166510 -0.8515789 -0.2457508  1.6970296  0.2007351  0.3872396  1.7881741
##         29         30         31         32
##  0.7833102 -0.7830296  0.7752712 -0.9745009
```

19d: compute with built in function

```
built_in_standardized_res <- rstandard(m)
standardized_res <- standardized_resids(residuals(m), design_matrix)

all.equal(standardized_res, built_in_standardized_res)
```

```
## [1] TRUE
```

19e: studentized residuals:

```
studentized_residuals <- function(r, x){
  ri <- standardized_resids(r, x)
  n <- nrow(x)
  p <- ncol(x) - 1
  return (ri *sqrt((n - p -2) / (n - p- 1 - ri^2) ))
}

studentized_residuals(r, design_matrix)
```

```
##          1          2          3          4          5          6          7
## -0.6957946 -0.6957946 -0.7704291  0.4019668  1.0827517 -1.0574065  0.1902837
##          8          9         10         11         12         13         14
## -0.1132489 -0.4218062 -1.1125598 -1.6071513 -0.4812913 -0.1909234 -0.8757501
##         15         16         17         18         19         20         21
## -0.3228260 -0.3561381  1.0369364  1.9970953  1.0974811  2.5290323 -1.0497999
##         22         23         24         25         26         27         28
## -0.6099379 -0.8474304 -0.2417284  1.7570378  0.1973810  0.3814920  1.8627667
##         29         30         31         32
##  0.7779603 -0.7776756  0.7698064 -0.9736260
```

19f: built in studentized residuals:

```
built_in_studentized_res <- rstudent(m)
my_studentized_residuals <- studentized_residuals(r, design_matrix)
all.equal(built_in_studentized_res, my_studentized_residuals)
```

```
## [1] TRUE
```

Task 20:

20: residual vs. fitted plot

Figure 1: Residuals vs. Fitted Plots

DataSet1: Residuals are centered around 0 and have approximately equal variance. They do not indicate any pattern and are independent of the fitted values.

DataSet2: Residuals are not centered around 0, but a quadratic distribution can be assumed.

DataSet3: The residuals are centered against 0, but the further they are from the fitted value of 2, the greater their variance. This behavior indicates that DataSet3 is heteroskedastic.

DataSet4: Residuals are again centered against 0 and have a constant variance, except for very large residuals.

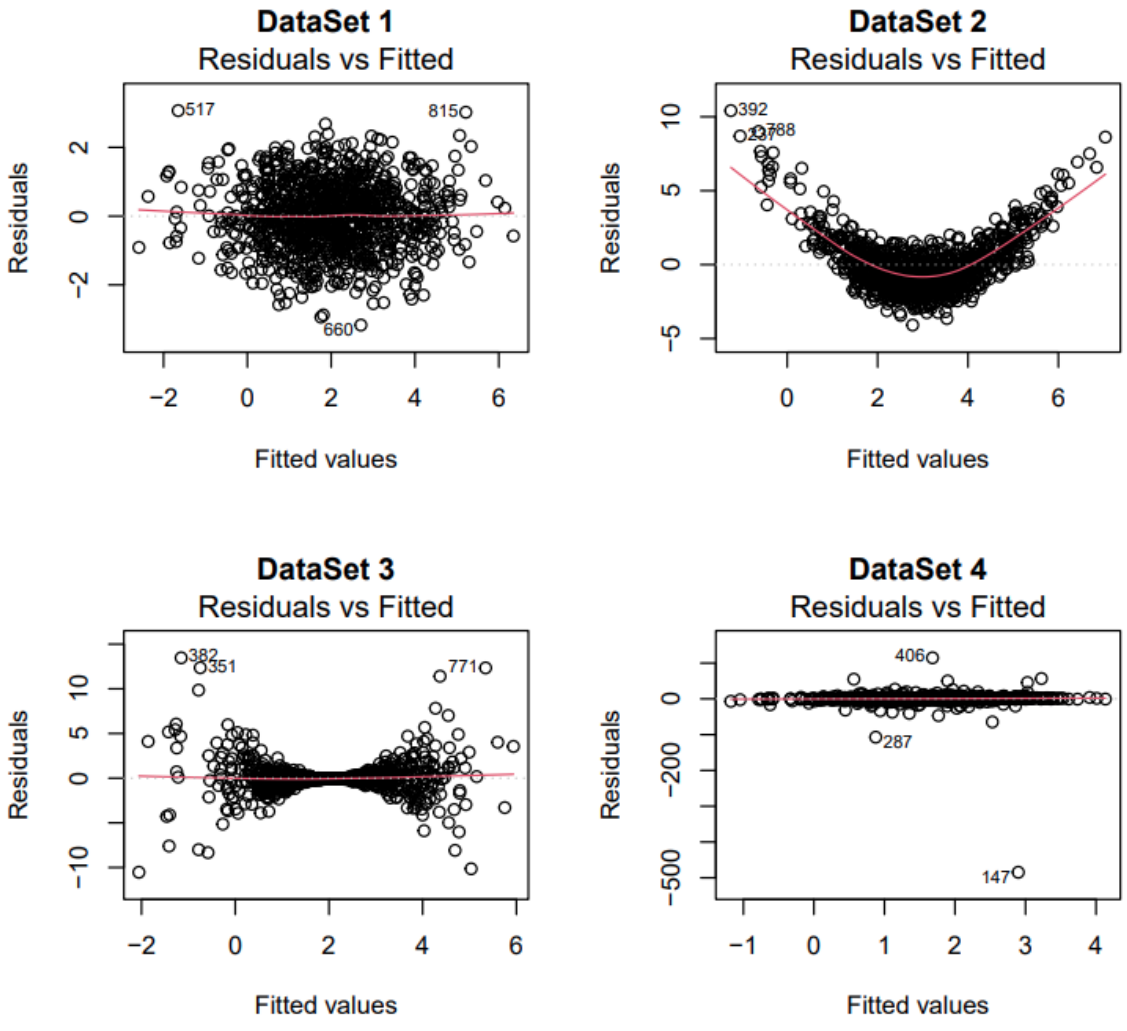


Figure 1: Residuals vs. fitted plot

Figure 2: QQ Plots for the Residuals

DataSet 1: The QQ plot shows that residuals deviate from the line at both ends, suggesting that the distribution of residuals is normal but not perfectly normal. But assumptions of linearity, homoscedasticity and noraml-distribution is not violated.

DataSet 2: The curvature in this plot indicates that the residuals are not normally distributed, suggesting a violation of the linearity and normality assumption.

DataSet 3: The residuals also deviate from the line at both tails in this plot, similar to DataSet 1, which indicates issues with normality and homoscedasticity, especially with outlier effects.

DataSet 4: The plot indicates that while the central part of the distribution appears to follow a normal distribution, the tails deviate significantly, indicating the presence of outliers which must not be ignored. Therefore we must declare that normality is being violated.

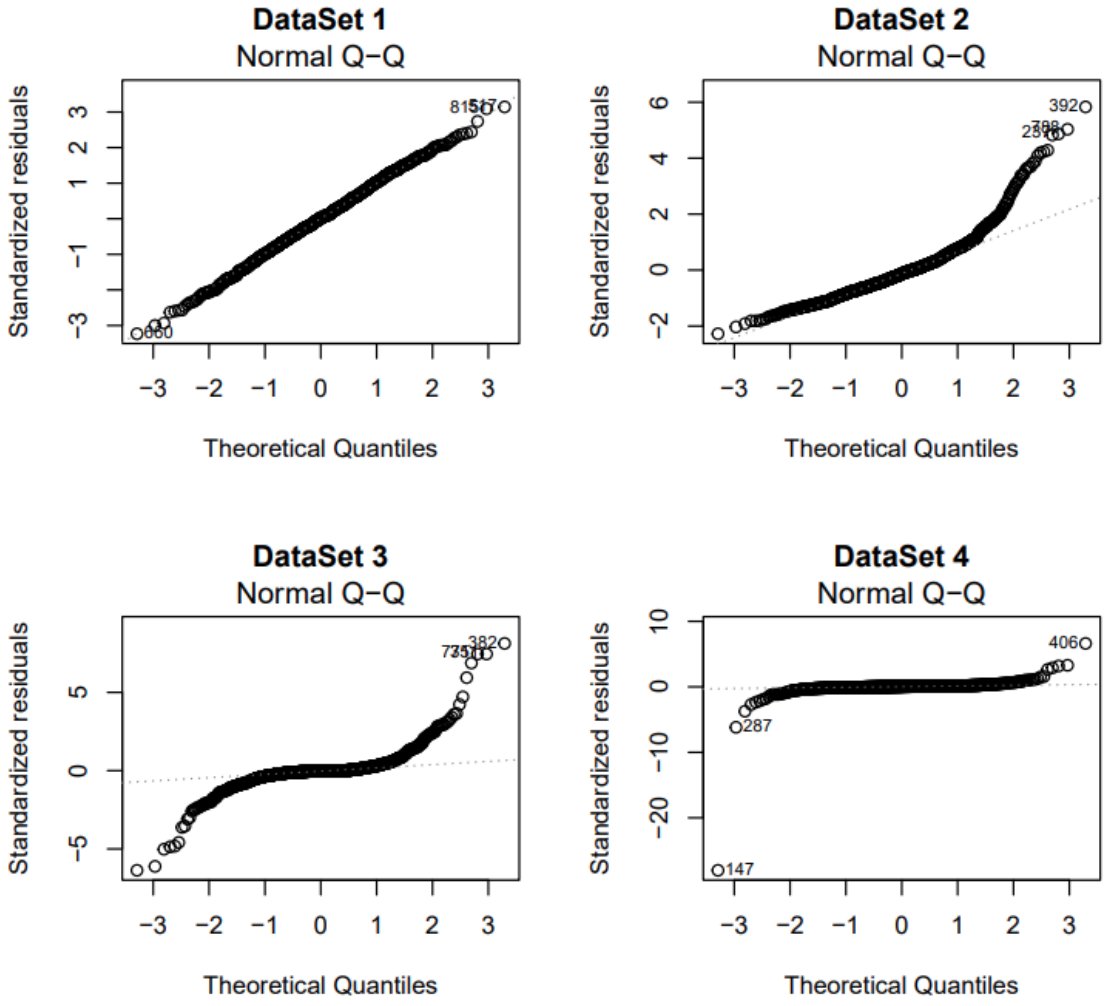
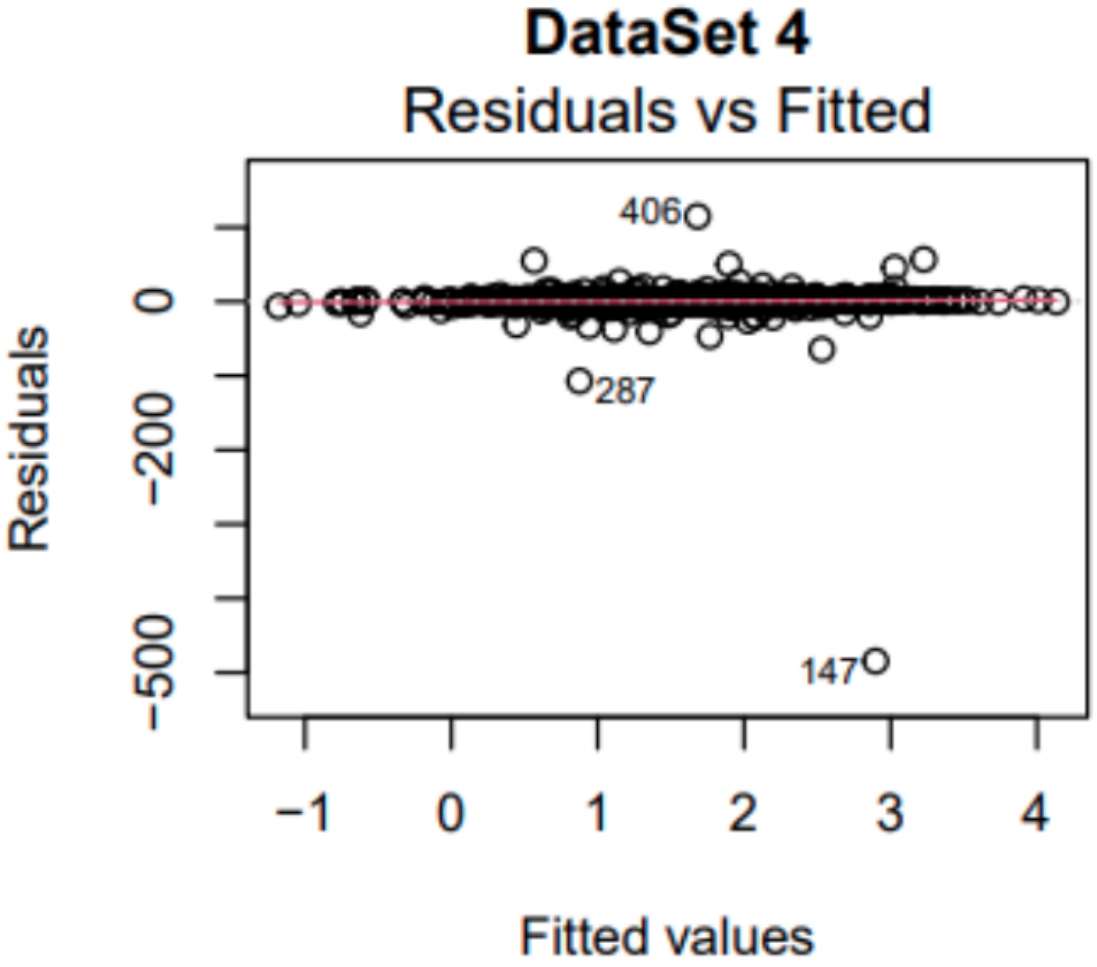


Figure 2: QQplots for the residuals

Best Fit and Assumption Violations

Best Fit: Based on the Residuals vs. Fitted plots (Figure 1), DataSet 4 seems to indicate the best fit among the four because the residuals appear to have the least pattern, despite the potential heteroscedasticity and outliers. This is also somewhat supported by the QQ plot, as the majority of data points follow the line.



Task 21:

Model 1: sales = 7 + 0.03 · T V + 0.2 · youtube + 0.5 · social Model 2: sales = 6 + 0.2 · youtube + 0.5 · social + 0.25 · social · youtube Model 3: sales = 5.5 + 0.2 · youtube + 0.3 · youtubeˆ2 + 0.1 · social Model 4: sales = 200 + 0.05 · (TV - $\overline{\text{TV}}$) + 0.2 · (youtube - $\overline{\text{youtube}}$) + 0.5 · social

22a. In Model 1, the coefficient of TV .

In Model 1, the coefficient of TV is 0.03. This suggests that, holding the effects of youtube and social media advertising constant, a one-dollar increase in TV advertising is associated with an increase of 0.03 hundred units in sales.

22b. In Model 1, the intercept.

In Model 1, the intercept is 700. This represents the expected sales in hundred units when the budgets for TV, youtube, and social media advertising are all zero == no marketing.

22c. In Model effect of youtube advertising on sales

In Model 2, the effect of youtube advertising on sales for a given value of social media advertising is initially 0.2. However, there is an interaction term between youtube and social which is 0.25. This means that the effect of youtube on sales increases by 0.25 for each additional dollar spent on social media advertising. As the social media ad budget increases, the incremental effectiveness of youtube advertising on sales also increases.

22d. relationship between sales and youtube

In Model 3, the relationship between sales and youtube is quadratic due to the presence of youtube and youtubeˆ2 terms. This suggests there is a non-linear relationship, and the preferable values for youtube advertising depend on the specific values of the coefficients.

With the given positive coefficient for youtubeˆ2, the sales would initially increase at an increasing rate with the youtube budget but after reaching a certain point, the effect would start to increase at a decreasing rate. If the coefficient of youtubeˆ2 were negative, the relationship would be a parabola opening downwards, indicating that there is a peak point for youtube advertising after which additional spending would lead to a decrease in sales.

22e. In Model 4, the intercept.

The intercept is the expected number of sales (in hundred units) when the TV budget is equal to the average TV budget ($\overline{\text{TV}}$), the youtube budget is equal to the average youtube budget ($\overline{\text{youtube}}$), and the social media budget is zero. Thus, one would expect to sell 20,000 units.

22f. In Model 4, the coefficient of (TV - $\overline{\text{TV}}$).

If the TV advertising budget is increased by x, the expected sales would increase by x * 0.05 * 100 = 5 units.

Task 22:

22a: Descriptives

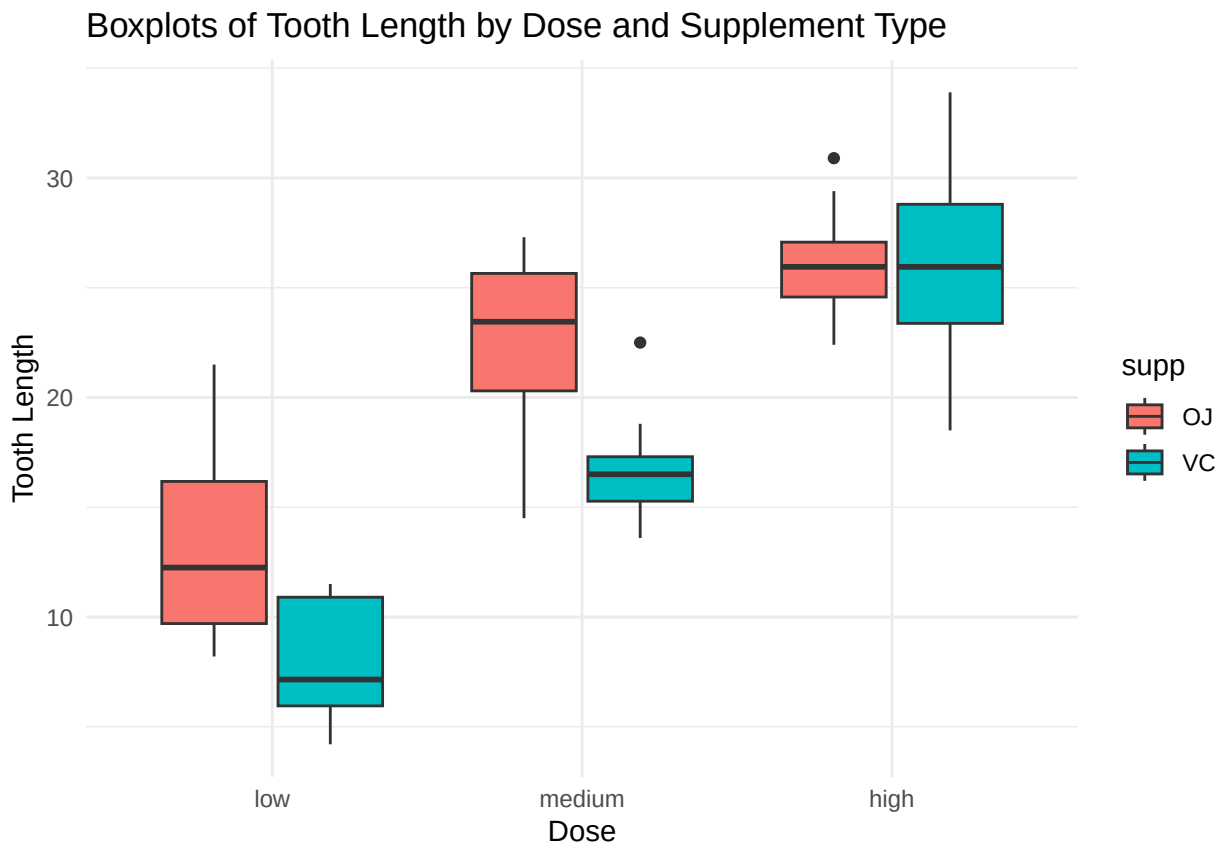
```
data("ToothGrowth", package = "datasets")
```

```
ToothGrowth$dose <- factor(ToothGrowth$dose,
                           levels = c(0.5, 1, 2),
                           labels = c("low", "medium", "high"))
```

```
summary(ToothGrowth)
```

```
##       len        supp        dose
##  Min.   : 4.20   OJ:30   low   :20
##  1st Qu.:13.07   VC:30   medium:20
##  Median :19.25           high  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

```
# Boxplots for length by dose and supp
library(ggplot2)
ggplot(ToothGrowth, aes(x = dose, y = len, fill = supp)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplots of Tooth Length by Dose and Supplement Type",
       x = "Dose",
       y = "Tooth Length")
```




```
#install.packages("dplyr")
library(dplyr)
```

```
## Warning: Paket 'dplyr' wurde unter R Version 4.3.2 erstellt
```

```
##
## Attache Paket: 'dplyr'
```

```
## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     filter, lag
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ToothGrowth |>
  group_by(supp) |>
  summarise(
    Mean = mean(len),
    SD = sd(len),
    N = n()
  )
```

```
## # A tibble: 2 x 4
##   supp   Mean    SD     N
##   <fct> <dbl> <dbl> <int>
## 1 OJ     20.7  6.61    30
## 2 VC     17.0  8.27    30
```

```
ToothGrowth |>
  group_by(dose) |>
  summarise(
    Mean = mean(len),
    SD = sd(len),
    N = n()
  )
```

```
## # A tibble: 3 x 4
##   dose    Mean    SD     N
##   <fct>  <dbl> <dbl> <int>
## 1 low     10.6  4.50    20
## 2 medium  19.7  4.42    20
## 3 high    26.1  3.77    20
```

```
ToothGrowth |>
  group_by(supp, dose) |>
  summarise(
    Mean = mean(len),
    SD = sd(len),
    N = n()
  )
```

```
## `summarise()` has grouped output by 'supp'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 6 x 5
## # Groups:   supp [2]
##   supp  dose    Mean    SD     N
##   <fct> <fct>  <dbl> <dbl> <int>
## 1 OJ    low    13.2   4.46    10
## 2 OJ    medium 22.7   3.91    10
## 3 OJ    high   26.1   2.66    10
## 4 VC    low     7.98  2.75    10
## 5 VC    medium 16.8   2.52    10
## 6 VC    high   26.1   4.80    10
```

22b: Linear Model

```
model1 <- lm(len ~ supp + dose, data = ToothGrowth)
summary(model1)
```

```
##
## Call:
## lm(formula = len ~ supp + dose, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.085 -2.751 -0.800  2.446  9.650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4550     0.9883  12.603  < 2e-16 ***
## suppVC       -3.7000     0.9883  -3.744 0.000429 ***
## dosemedium    9.1300     1.2104   7.543 4.38e-10 ***
## dosehigh     15.4950     1.2104  12.802  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 56 degrees of freedom
## Multiple R-squared:  0.7623, Adjusted R-squared:  0.7496
## F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

Interpretation: The intercept is the expected length of odontoblasts when a guinea pig has received a low dose of vitamin C through orange juice.

suppVC is a negative number, which is why it can actually be interpreted as the expected decrease in the length of odontoblasts if the guinea pig received vitamin C through ascorbic acid instead of orange juice.

Dosemedium represents the expected increase when a guinea pig has received a medium dose of vitamin C.

22c: Sum contrasts

```r
# Load the ToothGrowth data
data("ToothGrowth")

# Convert dose to a factor if it's not already
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ToothGrowth$supp <- as.factor(ToothGrowth$supp)

# Set sum contrasts for the factors
contrasts(ToothGrowth$supp) <- contr.sum(levels(ToothGrowth$supp)) # for the supp variable
contrasts(ToothGrowth$dose) <- contr.sum(levels(ToothGrowth$dose)) # for the dose variable

# Fit the linear model with sum contrasts
model <- lm(len ~ supp + dose, data = ToothGrowth)

# Summarize the model
summary(model)
```

```
##
## Call:
## lm(formula = len ~ supp + dose, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.085 -2.751 -0.800  2.446  9.650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.8133     0.4941  38.073  < 2e-16 ***
## supp1         1.8500     0.4941   3.744 0.000429 ***
## dose1        -8.2083     0.6988 -11.746  < 2e-16 ***
## dose2         0.9217     0.6988   1.319 0.192573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 56 degrees of freedom
## Multiple R-squared:  0.7623, Adjusted R-squared:  0.7496
## F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

Interpretation: “dose1” is the difference in the expected effect for a low dose relative to the overall mean.

“dose2” is the difference in the expected effect for a medium dose, incorporating the overall mean.

“supp1” is the difference in the expected effect when Vitamin C is administered via orange juice. The overall mean must be taken into account and should not change.

22d: Interactions

I am sorry. I don't know what to do here :(

```
model3 <- lm(len ~ dose * supp, data = ToothGrowth)
summary(model3);
```

```
##
## Call:
## lm(formula = len ~ dose * supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.8133     0.4688  40.130  < 2e-16 ***
## dose1        -8.2083     0.6630 -12.381  < 2e-16 ***
## dose2         0.9217     0.6630   1.390 0.170190
## supp1         1.8500     0.4688   3.946 0.000231 ***
## dose1:supp1   0.7750     0.6630   1.169 0.247568
## dose2:supp1   1.1150     0.6630   1.682 0.098394 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.7937, Adjusted R-squared:  0.7746
## F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

Task 23:

```
library(haven)
```

```
## Warning: Paket 'haven' wurde unter R Version 4.3.2 erstellt
```

```
child_iq <- read_dta("./Exercise_4/child.iq.dta")
(child_iq)
```

```
## # A tibble: 400 x 3
##     ppvt educ_cat momage
##    <dbl>    <dbl>  <dbl>
##  1   120        2     21
##  2    89        1     17
##  3    78        2     19
##  4    42        1     20
##  5   115        4     26
##  6    97        1     20
##  7    94        1     20
##  8    68        2     24
##  9   103        3     19
## 10    94        3     24
## # i 390 more rows
```

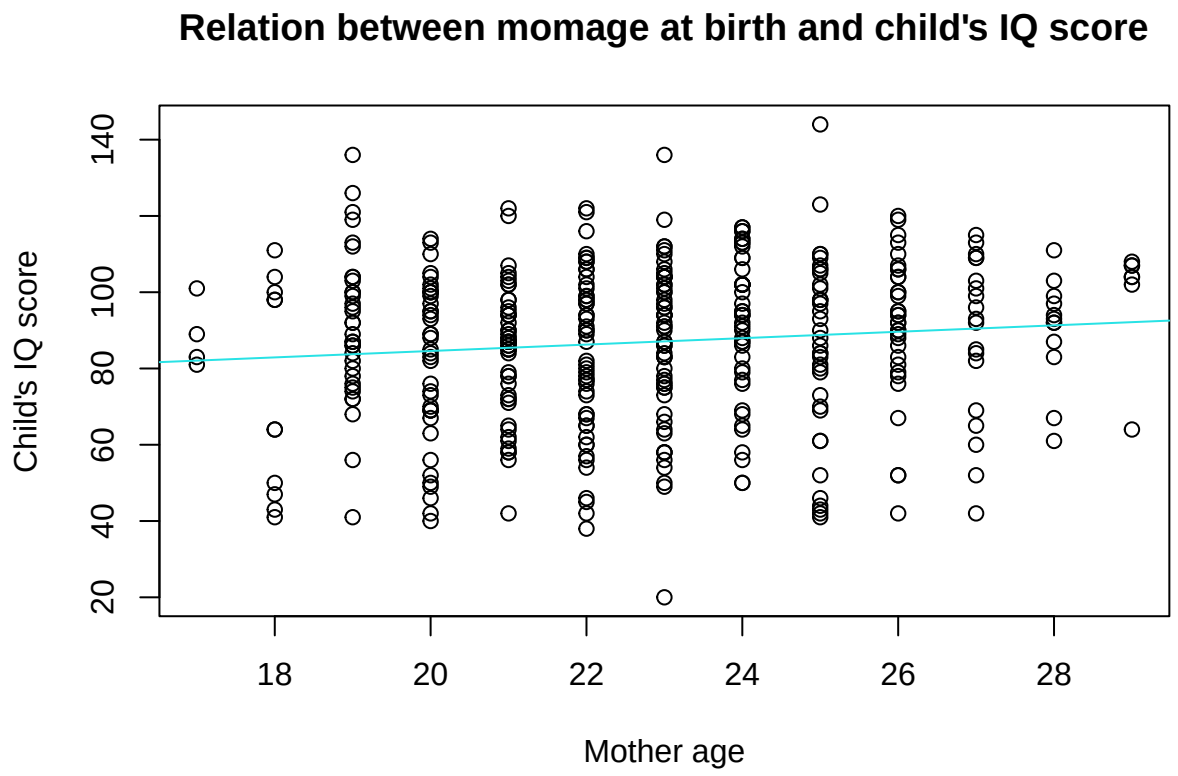
23a: Minimum age of mother 1

```
#library(ggplot2)

model <- lm(ppvt ~ momage, data = child_iq)

# Summary of the model to interpret coefficients
summary(model)
```

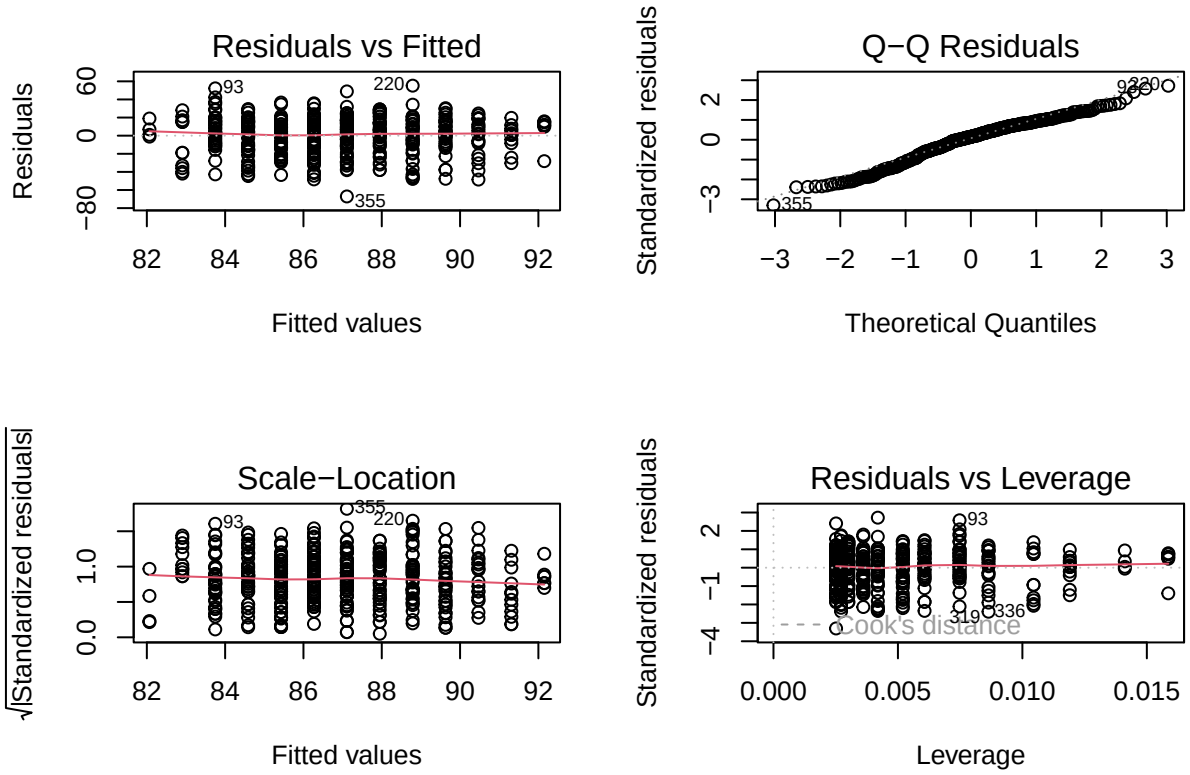
```
##
## Call:
## lm(formula = ppvt ~ momage, data = child_iq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.109 -11.798   2.971  14.860  55.210
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.7827     8.6880   7.802 5.42e-14 ***
## momage        0.8403     0.3786   2.219    0.027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.34 on 398 degrees of freedom
## Multiple R-squared:  0.01223,    Adjusted R-squared:  0.009743
## F-statistic: 4.926 on 1 and 398 DF,  p-value: 0.02702
```

```
plot(child_iq$momage, child_iq$ppvt,
     main = "Relation between momage at birth and child's IQ score",
     ylab = "Child's IQ score",
```

```
     xlab = "Mother age")

abline(model, col = 5)
```




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



Interpret slope coefficient: If the mother is older, then the child will have more IQ.

```
(90 - coefficients(model)[1]) / coefficients(model)[2]
```

```
## (Intercept)
##     26.4406
```

Based on this data, I would reccommend that the mother is older than 26.4406 years. But since the oldest mother in our dataset is 29, we can't say about anything older than that. To get the maximum IQ, the mother should be 29.

23b: Minimum age of mother 2

```
model2 <- lm(ppvt ~ I(momage) + as.numeric(educ_cat),
             data = child_iq)
summary(model2)
```

```
##
## Call:
## lm(formula = ppvt ~ I(momage) + as.numeric(educ_cat), data = child_iq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.763 -13.130   2.495  14.620  55.610
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            69.1554     8.5706   8.069 8.51e-15 ***
## I(momage)               0.3433     0.3981   0.862 0.389003
## as.numeric(educ_cat)    4.7114     1.3165   3.579 0.000388 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.05 on 397 degrees of freedom
## Multiple R-squared:  0.04309,    Adjusted R-squared:  0.03827
## F-statistic: 8.939 on 2 and 397 DF,  p-value: 0.0001594
```

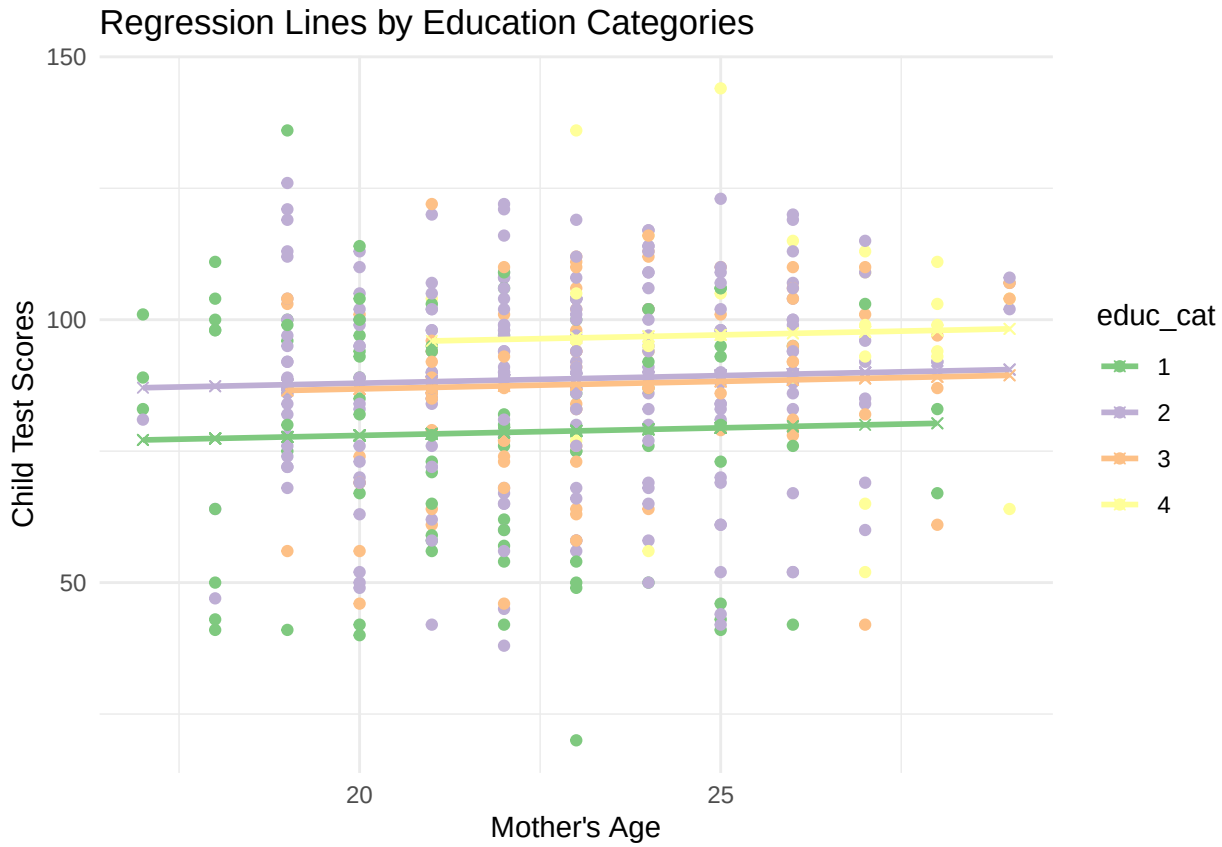
If the mother is a year higher, then the IQ would rise 0.3433.

If the education is higher, then the IQ would rise 4.7114.

Based on this, I would reccommend not using the age but using the education level. The higher the education, the better. If I absolutley have to give an answer, then I would say 29. We have no data on Anyone older.

23c: Minimum age of mother 3

```r
# Assuming 'child_iq' is your data frame and it contains the variables 'momage', 'test_scores', and 'ed

# Convert 'educ_cat' to a factor if it's not already
child_iq$educ_cat <- as.factor(child_iq$educ_cat)

# Fit the regression model
model <- lm(ppvt ~ momage + educ_cat, data = child_iq)

# Create predictions for the fitted model
child_iq$predicted_scores <- predict(model)

# Load the ggplot2 library for plotting
library(ggplot2)

# Plot the data with regression lines for different education levels
ggplot(child_iq, aes(x = momage, y = ppvt, color = educ_cat)) +
  geom_point() + # Plot the actual points
  geom_point(aes(y = predicted_scores), shape = 4) + # Add 'X' marks for the fitted values
  geom_line(aes(y = predicted_scores), linewidth = 1) + # Add regression lines
  labs(x = "Mother's Age", y = "Child Test Scores", title = "Regression Lines by Education Categories") +
  theme_minimal() +
  scale_color_brewer(type = 'qual') # Use a qualitative color scale for clarity
```



Feedback

Task 22d incomplete: -2

Task 23b: Important here is that the variable momage is not significant anymore. -2 Task 23b: You cannot say some variable has a larger impact than another variable with the argument, that it has the larger estimated coefficient. The magnitude of the coefficient depends on the scale of the variable itself, not on the impact the variable has. Only with scaled data, you could make such argument.

No Task 24: -16