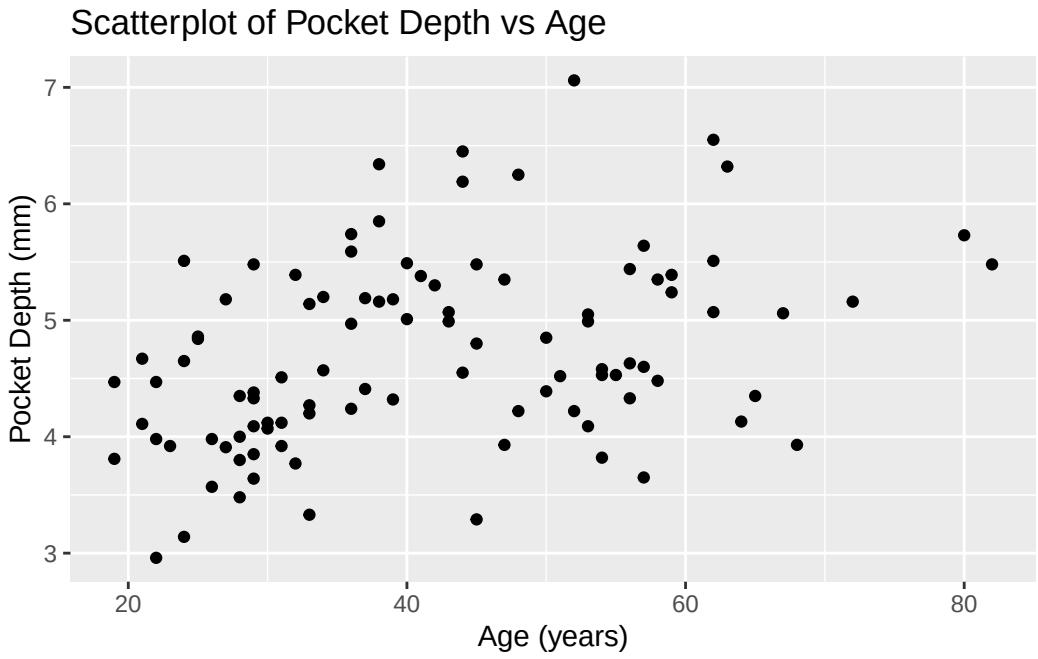
Medical Statistics – Answers lab 6

Part 1: Pearson's correlation coefficient and simple linear regression



Question 1

Based on the scatterplot, is there an indication of a linear association between `age` and `pocketdepth`? If so, is this association positive or negative?

Answer question 1

The scatterplot suggests that there is a positive linear association between `age` and `pocketdepth`. As age increases, pocket depth tends to increase.

Pearson's Correlation Coefficient

```
    Pearson's product-moment correlation

data:  pockets$age and pockets$pocketdepth
t = 3.9647, df = 98, p-value = 0.0001398
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1891831 0.5295346
sample estimates:
     cor
0.371786
```

> Question 2
>
> What does the correlation coefficient tell us about the relationship between `age` and `pocketdepth`? Does this align with your interpretation of the scatterplot?

> Answer question 2
>
> The correlation coefficient is approximately 0.37, indicating a moderate positive linear relationship between `age` and `pocketdepth`. This aligns with the interpretation of the scatterplot.

> Question 3
>
> What is the p-value for the correlation coefficient test? Based on this p-value, do we have sufficient evidence to reject the null hypothesis?

> Answer question 3
>
> The p-value for the correlation coefficient test is approximately 0.00014, indicating that we have sufficient evidence to reject the null hypothesis of no correlation between `age` and `pocketdepth`.

Fitting a Simple Linear Regression Model

```
Call:
lm(formula = pocketdepth ~ age, data = pockets)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.49289 -0.44376 -0.07903  0.49597  2.13335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.85872    0.22862  16.878  < 2e-16 ***
age          0.02054    0.00518   3.965  0.00014 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7535 on 98 degrees of freedom
Multiple R-squared:  0.1382,    Adjusted R-squared:  0.1294
F-statistic: 15.72 on 1 and 98 DF,  p-value: 0.0001398
```

Question 4

Is the relationship between `age` and `pocketdepth` statistically significant (at $= 0.05$)?

Answer question 4

The estimated coefficient for `age` (0.021) is statistically significant with a p-value of 0.00014, indicating that the relationship between `age` and `pocketdepth` is statistically significant at $= 0.05$.

Question 5

How does the p-value for `age` in the regression output compare to the p-value for the correlation coefficient test? Are they consistent with each other?

Answer question 5

The p-value for `age` in the regression output (0.00014) and the p-value for the correlation coefficient test (0.00014) are consistent with each other. They both indicate a statistically significant relationship between `age` and `pocketdepth`.

Question 6

What is the interpretation of the intercept and the coefficient for `age` in the regression output?

Answer question 6

The intercept (3.859) represents the estimated pocket depth when age is 0, which may not have a meaningful interpretation in this context. The coefficient for `age` (0.021) represents the estimated change in pocket depth for each additional year of age.

Question 7

Based on the fitted model, what is the expected pocket depth for a person who is 40 years old?

Answer question 7

The expected pocket depth for a person who is 40 years old can be calculated using the intercept and coefficient from the regression output: 3.859 + 0.021 * 40 = 4.699 mm

Question 8

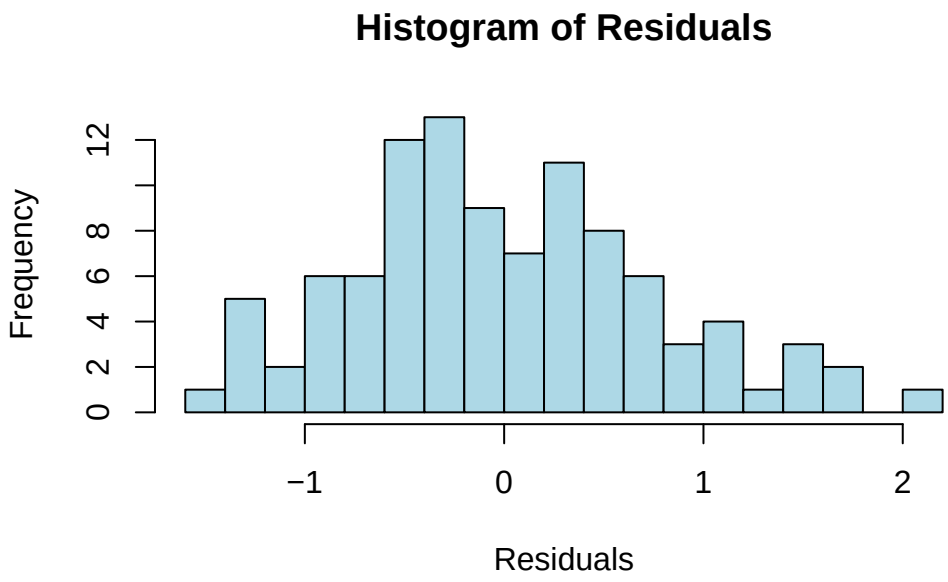
How much of the variation in pocket depth is explained by age in this model?

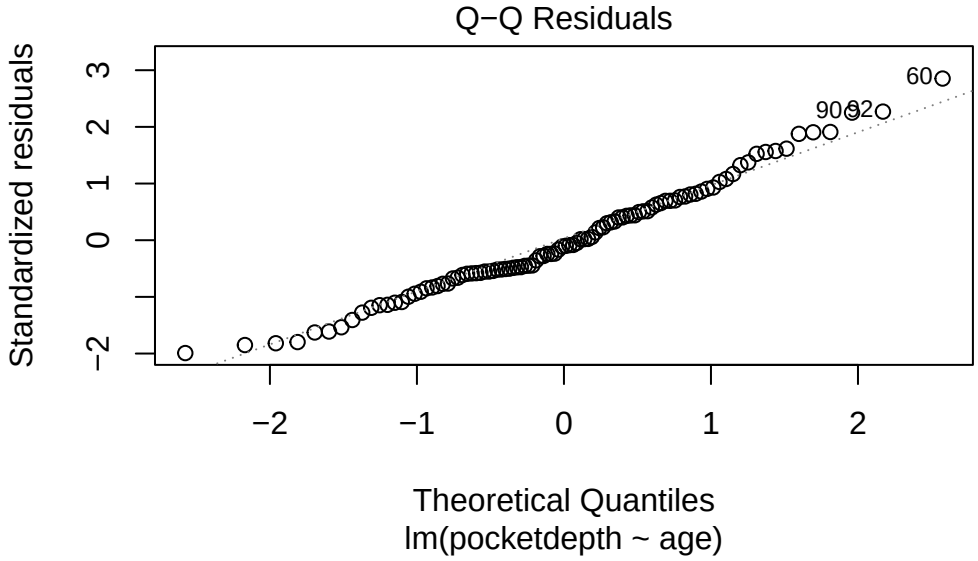
Answer question 8

The R-squared value of 0.138 indicates that approximately 14% of the variation in pocket depth is explained by age in this model.

Assumption Checking

Normality of Residuals


Histogram of Residuals
Frequency
0
2
4
6
8
12
−1
0
1
2
Residuals


Q−Q Residuals
Standardized residuals
−2
0
1
2
3
60
90
82
−2
−1
0
1
2
Theoretical Quantiles
lm(pocketdepth ~ age)

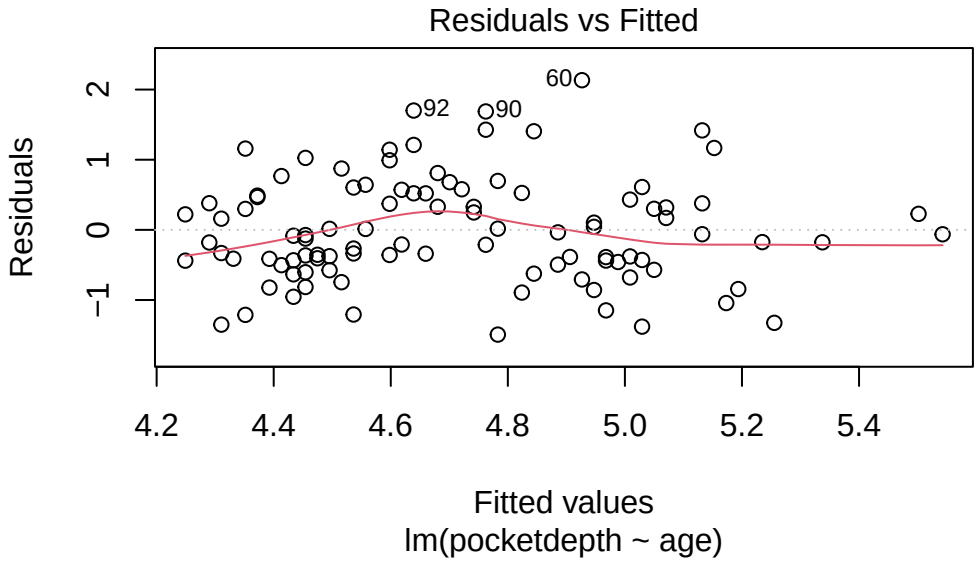
Question 9

Do the histogram and Q-Q plot suggest that the residuals are reasonably normally distributed?

Answer question 9

The two plots suggest that the residuals are approximately normally distributed, with the histogram showing a roughly symmetric shape and the Q-Q plot showing the residuals closely following the diagonal line.

Homoscedasticity and linearity



Question 10

Does the residual-versus-fitted plot suggest constant variance?

Answer question 10

The residual-versus-fitted plot suggests that the variance of the residuals is relatively constant across the range of fitted values, indicating homoscedasticity.

Question 11

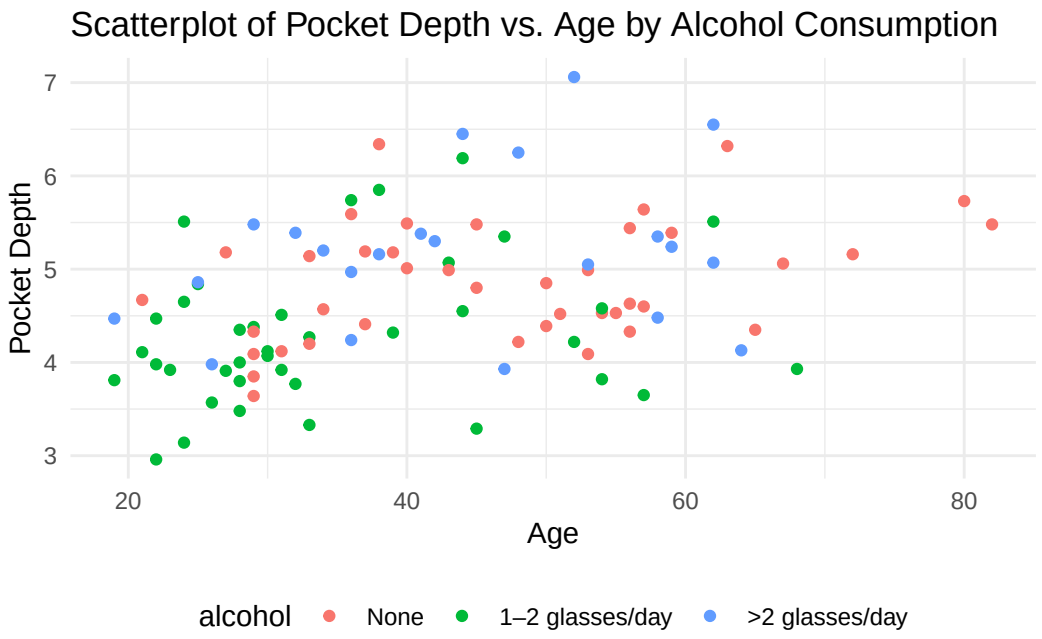
Does the residual-versus-fitted plot suggest any violation of the linearity assumption?

Answer question 11

There is no obvious pattern or curvature in the residual-versus-fitted plot, indicating that the linearity assumption is not violated.

Part 2: ANCOVA (Analysis of Covariance)

Exploratory Data Analysis



Question 12

What can you infer from the scatterplot about the relationship between `age`, `pocketdepth`, and `alcohol` consumption?

Answer question 12

As discussed previously, the scatterplot suggests a positive relationship between `age` and `pocketdepth`, with higher pocket depths observed for older individuals. It also suggests that there is a U-shaped relationship between `alcohol` consumption and `pocketdepth`, with individuals consuming "1-2 glasses/day" having lower pocket depths compared to those consuming "None" or ">2 glasses/day".

Fitting the ANCOVA Model

R output

```
Call:
lm(formula = pocketdepth ~ age + alcohol, data = pockets)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35037 -0.45299 -0.06626  0.39746  1.76573

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.165461   0.270688  15.388  < 2e-16 ***
age                      0.014849   0.005248   2.829  0.00568 **
alcohol1-2 glasses/day -0.394551   0.172383  -2.289  0.02428 *
alcohol>2 glasses/day    0.364565   0.188347   1.936  0.05586 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7069 on 96 degrees of freedom
Multiple R-squared:  0.257, Adjusted R-squared:  0.2338
F-statistic: 11.07 on 3 and 96 DF,  p-value: 2.638e-06
```

SPSS output

Between-Subjects Factors

		Value Label	N
Alcohol consumption categories	0	None	40
	1	1–2 glasses/day	38
	2	>2 glasses/day	22

Tests of Between-Subjects Effects

Dependent Variable: Average pocket depth (mm)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16,593[a]	3	5,531	11,068	<,001
Intercept	160,713	1	160,713	321,601	<,001
age	4,001	1	4,001	8,005	,006
alcohol	7,669	2	3,834	7,673	<,001
Error	47,974	96	,500		
Total	2287,218	100			
Corrected Total	64,567	99			

a. R Squared = ,257 (Adjusted R Squared = ,234)

Parameter Estimates

Dependent Variable: Average pocket depth (mm)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4,530	,275	16,464	<,001	3,984	5,076
age	,015	,005	2,829	,006	,004	,025
[alcohol=0]	-,365	,188	-1,936	,056	-,738	,009
[alcohol=1]	-,759	,195	-3,888	<,001	-1,147	-,372
[alcohol=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Figure 1: Screenshot of the SPSS output tables

Question 13

Based on the ANCOVA model output, what is the expected difference in pocket depth between individuals who consume `"None"` and those who consume `">2 glasses/day"`,

while controlling for age?

Answer question 13

By default, the reference level for the `alcohol` variable is set to `"None"` in R and to `">2 glasses/day"` in SPSS.
Looking at the R output, the coefficient for `">2 glasses/day"` is 0.365, which means that individuals who consume `">2 glasses/day"` are expected to have a pocket depth that is 0.365 mm higher than those who consume `"None"`, while controlling for age.
Looking at the SPSS output, the coefficient for `"None"` is -0.365, which means that individuals who consume `"None"` are expected to have a pocket depth that is 0.365 mm lower than those who consume `">2 glasses/day"`, while controlling for age.
Therefore, the expected difference in pocket depth between individuals who consume `"None"` and those who consume `">2 glasses/day"` is 0.365 mm, independent of the reference level chosen.

Question 14

Based on the ANCOVA model output, what is the expected difference in pocket depth between individuals who consume `"1-2 glasses/day"` and those who consume `">2 glasses/day"`, while controlling for age?

Answer question 14

By default, the reference level for the `alcohol` variable is set to `"None"` in R and to `">2 glasses/day"` in SPSS.
Looking at the R output, the coefficient for `">2 glasses/day"` is 0.365 and the coefficient for `"1-2 glasses/day"` is -0.395. This means that individuals who consume `">2 glasses/day"` are expected to have a pocket depth that is 0.365 - -0.395 = 0.76 mm higher than those who consume `"1-2 glasses/day"`, while controlling for age.
Looking at the SPSS output, the coefficient for `"1-2 glasses/day"` is -0.759, which means that individuals who consume `"1-2 glasses/day"` are expected to have a pocket depth that is 0.759 mm lower than those who consume `">2 glasses/day"`, while controlling for age.
Therefore, the expected difference in pocket depth between individuals who consume `"1-2 glasses/day"` and those who consume `">2 glasses/day"` is 0.76 mm, independent of the reference level chosen.

```
Anova Table (Type III tests)

Response: pocketdepth
             Sum Sq Df  F value    Pr(>F)
```

```
(Intercept) 118.337  1 236.8042 < 2.2e-16 ***
age           4.001  1   8.0055 0.0056789 **
alcohol       7.669  2   7.6729 0.0008104 ***
Residuals    47.974 96
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 15

Based on the ANOVA table, is the `alcohol` variable significantly associated with `pocketdepth` after accounting for `age`?

Answer question 15

The p-value for the `alcohol` variable in the ANOVA table is less than 0.05, indicating that the `alcohol` variable is significantly associated with `pocketdepth` after accounting for `age`.

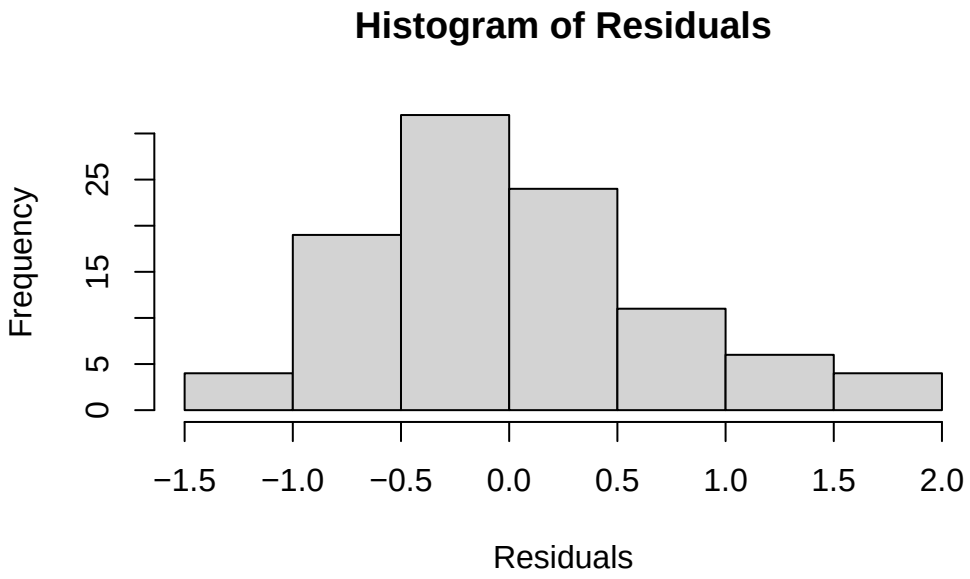
Model Diagnostics

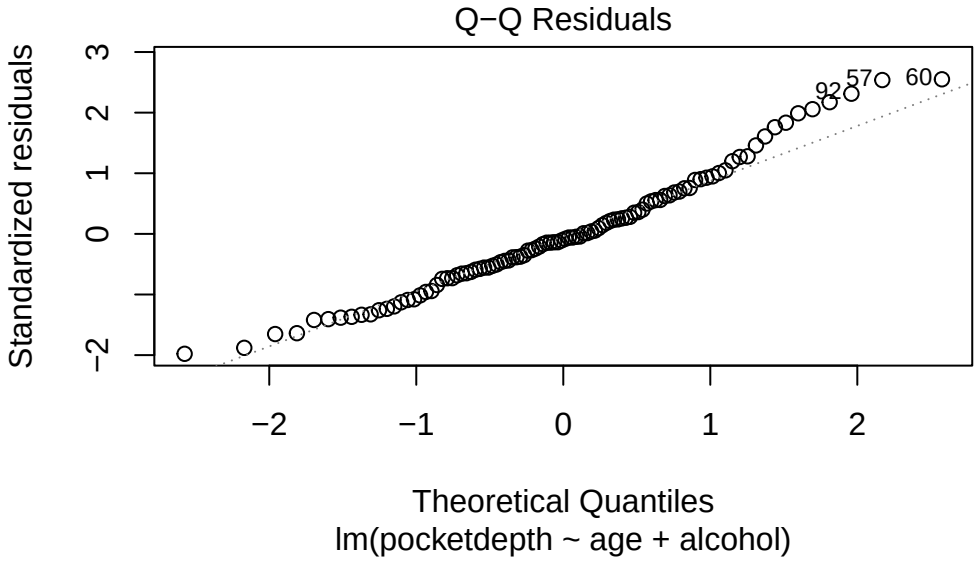
Exercise

Check the normality of residuals and homoscedasticity assumptions for the ANCOVA model.

Answer question 15

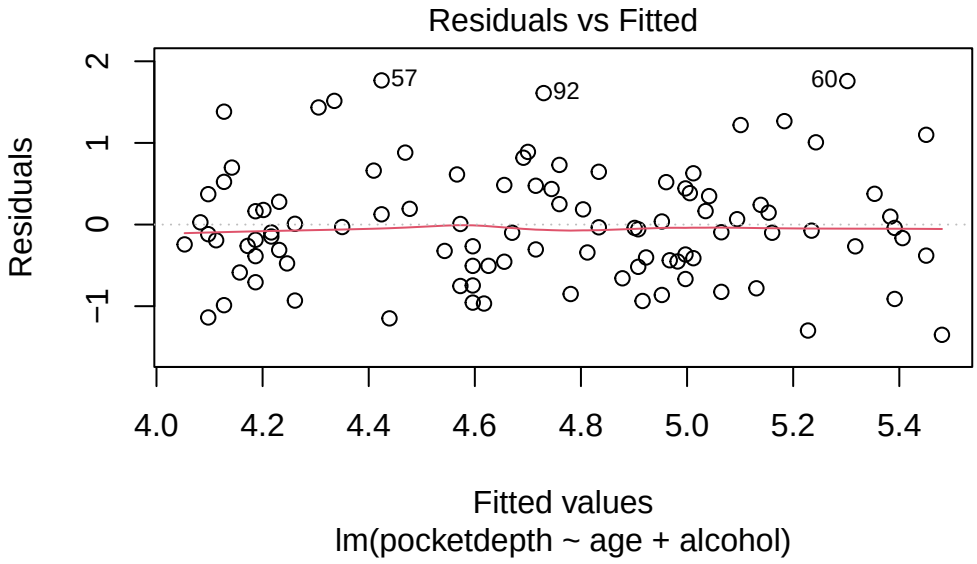
Normality of Residuals





The histogram and normal Q-Q plot of residuals shows that the residuals are approximately normally distributed, indicating that the normality assumption is not violated.

Homoscedasticity



The residuals vs. fitted plot shows that the residuals are randomly scattered around zero, indicating that the homoscedasticity assumption is not violated.

Part 3: Interactions in ANCOVA

Fitting the Interaction Model

```
Call:
lm(formula = pocketdepth ~ age * alcohol, data = pockets)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35020 -0.45321 -0.06748  0.39755  1.76684

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.161e+00  3.811e-01  10.921   <2e-16 ***
age                          1.494e-02  7.747e-03   1.928   0.0569 .
alcohol1-2 glasses/day      -3.862e-01  5.142e-01  -0.751   0.4544
alcohol>2 glasses/day        3.690e-01  6.538e-01   0.564   0.5738
age:alcohol1-2 glasses/day -2.082e-04  1.214e-02  -0.017   0.9863
```

```
age:alcohol>2 glasses/day  -9.585e-05  1.395e-02  -0.007   0.9945
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7144 on 94 degrees of freedom
Multiple R-squared:  0.257, Adjusted R-squared:  0.2175
F-statistic: 6.503 on 5 and 94 DF,  p-value: 3.102e-05


Anova Table (Type III tests)

Response: pocketdepth
             Sum Sq Df  F value  Pr(>F)
(Intercept) 60.865  1 119.2592 < 2e-16 ***
age          1.897  1   3.7169 0.05688 .
alcohol      0.784  2   0.7678 0.46691
age:alcohol  0.000  2   0.0001 0.99985
Residuals   47.974 94
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 16

Based on the output in the ANOVA table, is there a significant interaction between `age` and `alcohol` in predicting `pocketdepth`?

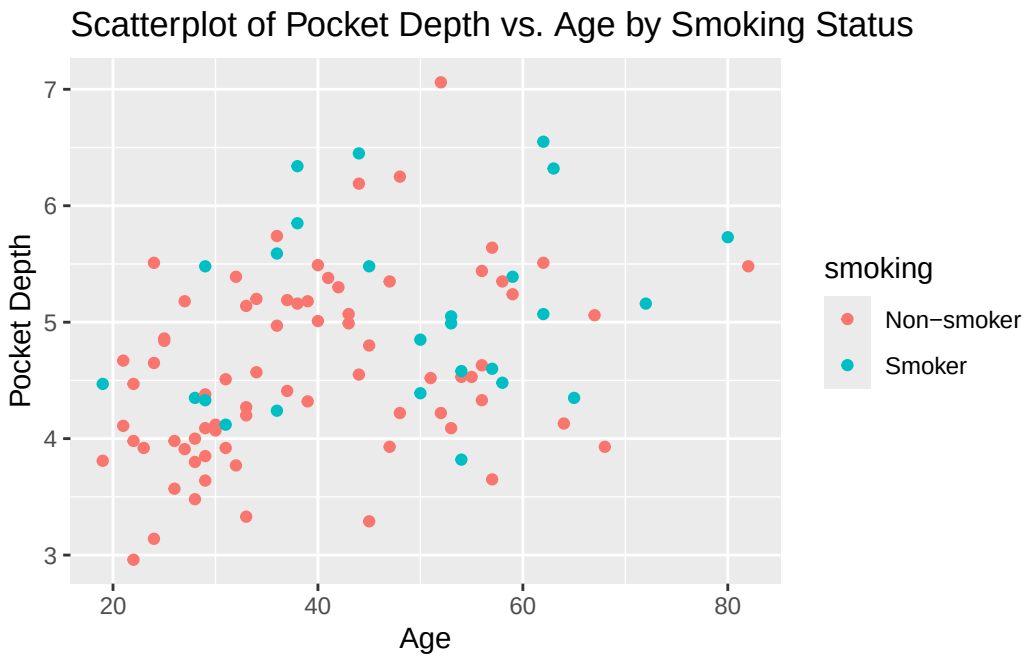
Answer question 16

The p-value for the interaction term `age:alcohol` in the ANOVA table is approximately equal to 1, indicating that there is no interaction between `age` and `alcohol` in predicting `pocketdepth`.

Part 4: Relationship Between Smoking and Pocket Depth

In addition to information about alcohol consumption, the dataset also contains information about smoking habits. Explore the relationship between smoking and pocket depth, and how it interacts with age. You can use the same approach as in the previous sections to fit models, test for significance, and check assumptions.

Exploratory Data Analysis



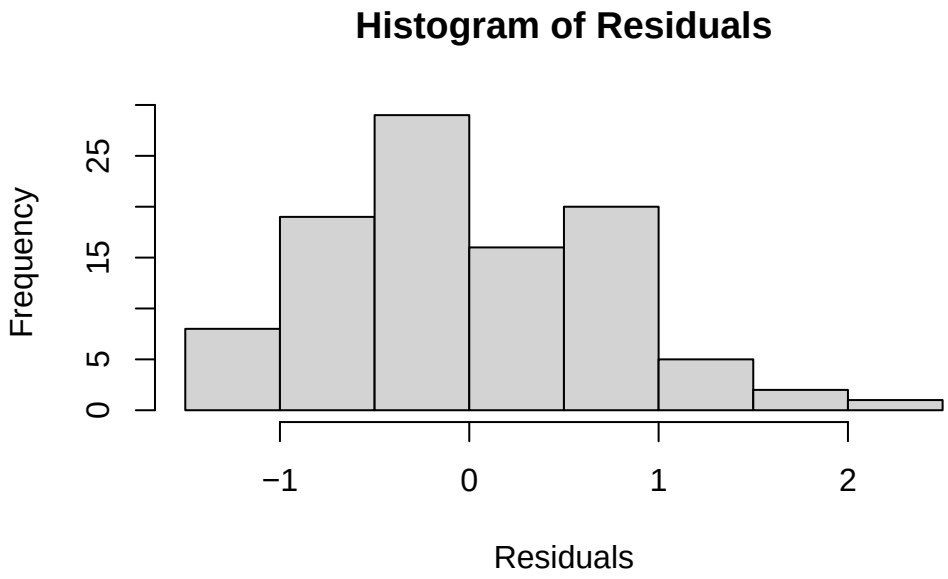
Fitting the ANCOVA Model

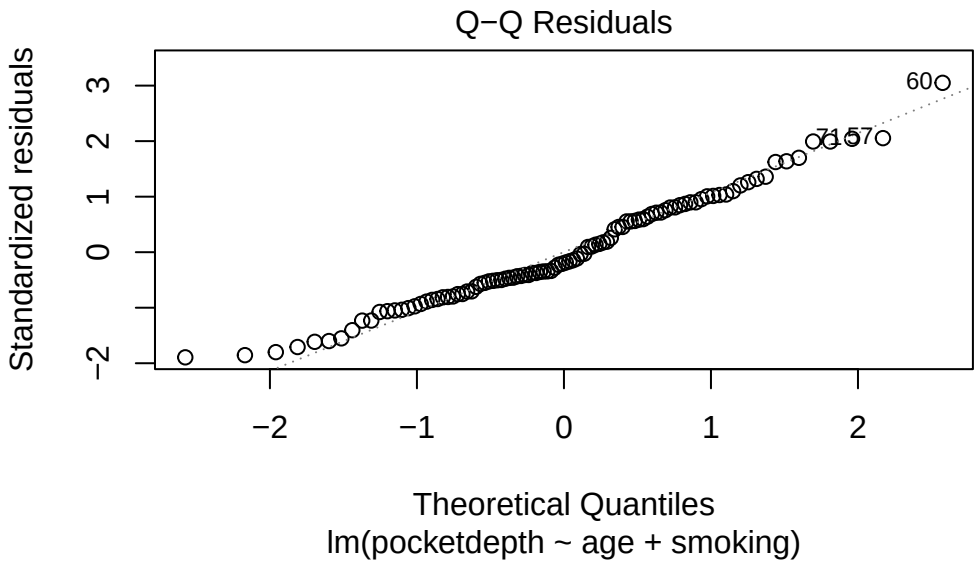
```
Call:
lm(formula = pocketdepth ~ age + smoking, data = pockets)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3995 -0.5308 -0.1453  0.5331  2.2462

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.89031    0.22660  17.168  < 2e-16 ***
age            0.01776    0.00534   3.326  0.00125 **
smokingSmoker  0.32371    0.17709   1.828  0.07063 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7447 on 97 degrees of freedom
Multiple R-squared:  0.1669,    Adjusted R-squared:  0.1497
```

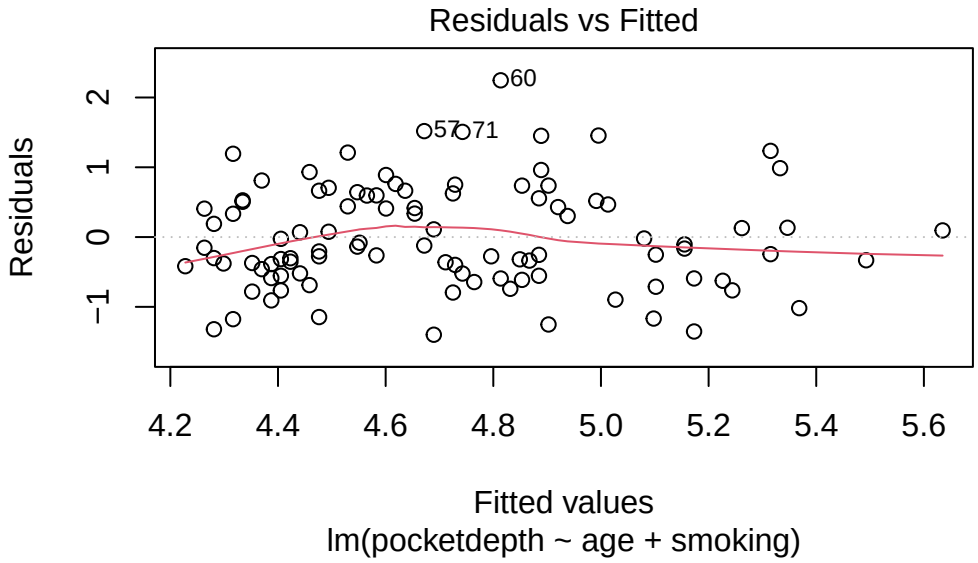
```
F-statistic: 9.718 on 2 and 97 DF,  p-value: 0.0001423
```

The p-value for the `smoking` variable in the ANOVA table is 0.071, indicating that the `smoking` variable is not significantly associated with `pocketdepth` after accounting for `age`.

Model Diagnostics







The plots do not show any violations of the normality or homoscedasticity assumptions.

Fitting the Interaction Model

```
Call:
lm(formula = pocketdepth ~ age * smoking, data = pockets)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4160 -0.5382 -0.1353  0.5514  2.2096

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.778052   0.264041  14.309  < 2e-16 ***
age                0.020622   0.006360   3.243  0.00163 **
smokingSmoker      0.772509   0.567866   1.360  0.17690
age:smokingSmoker -0.009779   0.011755  -0.832  0.40751
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7459 on 96 degrees of freedom
Multiple R-squared:  0.1729,    Adjusted R-squared:  0.147
F-statistic: 6.689 on 3 and 96 DF,  p-value: 0.0003787


Anova Table (Type III tests)

Response: pocketdepth
             Sum Sq Df  F value    Pr(>F)
(Intercept) 113.894  1 204.7350 < 2.2e-16 ***
age           5.849  1  10.5145  0.001629 **
smoking       1.029  1   1.8506  0.176898
age:smoking   0.385  1   0.6921  0.407507
Residuals    53.405 96
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the interaction term `age:smoking` is 0.407, indicating that there is no significant interaction between `age` and `smoking` in predicting `pocketdepth`.