HW7_1700

Yue Zhang

2024-11-13

Hypertension #11.13

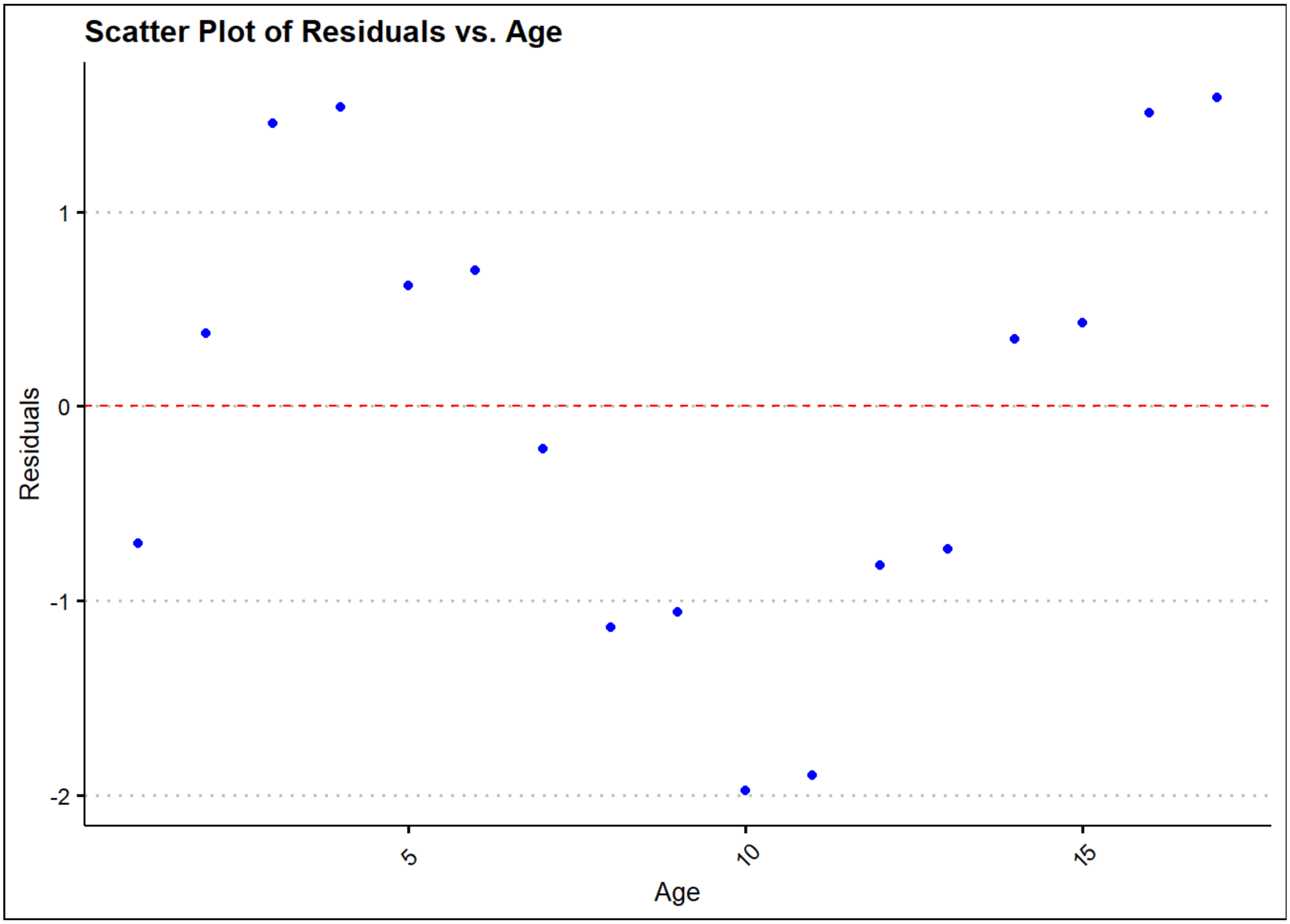
```
sbp = read.csv("E:/Biostat/Biostatistics/PHL_1700/Data/Raw/sbp-dat-1.csv")
sbp_model = lm(sbp ~ age, data = sbp)
summary(sbp_model)
```

```
##
## Call:
## lm(formula = sbp ~ age, data = sbp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9779 -0.8162  0.3456  0.6985  1.5882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 97.78676    0.61816  158.19  < 2e-16 ***
## age          1.91912    0.06033   31.81 3.49e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.219 on 15 degrees of freedom
## Multiple R-squared:  0.9854, Adjusted R-squared:  0.9844
## F-statistic:  1012 on 1 and 15 DF,  p-value: 3.494e-15
```

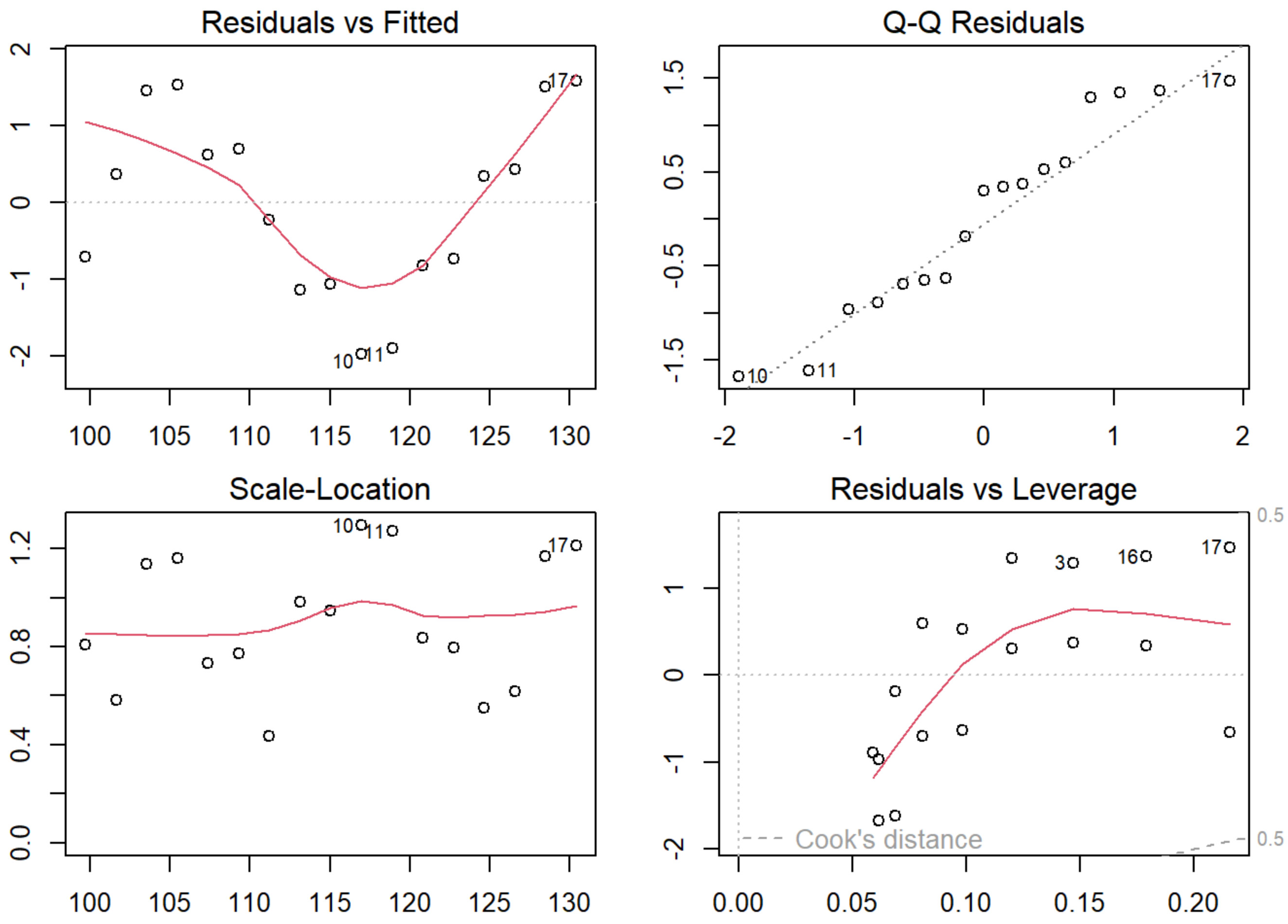
The equation is $sbp = 97.78676 + 1.91912 * age$.

#11.18

```
# Scatter plot of residuals vs. age
ggplot(sbp, aes(x = age, y = resid(sbp_model))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Age", x = "Age",
        y = "Residuals")
```



```
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(sbp_model, cex.axis = 1, cex.lab = 1)
```



Scatter plot of residuals vs. age: There's no indication of heteroscedasticity. Also, it shows that the residuals are independent. However, there's indication of non-linearity. Residuals vs. Fitted Plot: The curved pattern indicates that the relationship between age and SBP might not be linear as the residuals are not evenly distributed around zero. This suggests that a non-linear model might be a good fit for the data, Q-Q Plot: The residuals deviate slightly from the line, especially at the tails, suggesting mild departures from normality. However, since the dataset is small, we can treat it as normally distributed. Scale-Location Plot: The spread of residuals remains consistent, although there is a slight increase in variability at higher fitted values. This suggests that the assumption of homoscedasticity is still reasonable. Residuals vs. Leverage Plot: Observations 10, 17, and 3 show higher leverage and Cook's distance, indicating they may be influential points. These observations could disproportionately impact the model's fit. Overall, the linear model satisfies the assumption of normality and homoscedasticity. It does not satisfy the assumption of linearity. Therefore, the linear regression does not provide a good fit to the data.

#11.14

```
confint(sbp_model, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 96.469193 99.104336
## age          1.790536  2.047699
```

The 95% CI for intercept is (96.4692, 99.1043), and the 95% CI for coefficient of age is (1.7905, 2.0477). If the residuals display a significant deviation from normality, the CI might not accurately reflect the uncertainty of the estimated blood pressure as it might be too narrow or too wide.The CIs could be misleading if the linearity assumption is violated as they rely on the model being correctly specified.

#11.15

```
predict(sbp_model, newdata = data.frame(age = 13), se.fit = TRUE)
```

```
## $fit
##        1
## 122.7353
##
## $se.fit
## [1] 0.3815351
##
## $df
## [1] 15
##
## $residual.scale
## [1] 1.218525
```

The predicted blood pressure is 122.7353. The validity of the prediction depends on the assumption of linearity and whether the model is a good fit across the ages. If the model shows a non-linearity pattern in the residuals, predictions made by the model could be biased.

#11.16 The standard error of the estimated blood pressure is 0.3815. The standard error assumes that the residuals are homoscedastic. If heteroscedasticity is shown in the model, the standard error might be underestimate or overestimate and thus affecting the CI.

#11.17

```
predict(sbp_model, newdata = data.frame(age = 17), se.fit = TRUE)
```

```
## $fit
##        1
## 130.4118
##
## $se.fit
## [1] 0.565908
##
## $df
## [1] 15
##
## $residual.scale
## [1] 1.218525
```

The predicted blood pressure is 130.4118 and the standard error is 0.5659. The standard error is based on the homoscedastic assumption and the predicted value is based on the linearity assumption.

#11,25 The $R^2 = 0.27$ means that 27% of the variance in the 24-hour urinary Na (y) can be explained by the estimated 24-hour urinary Na(x).

Prediatrics, Endocrinology #11.49

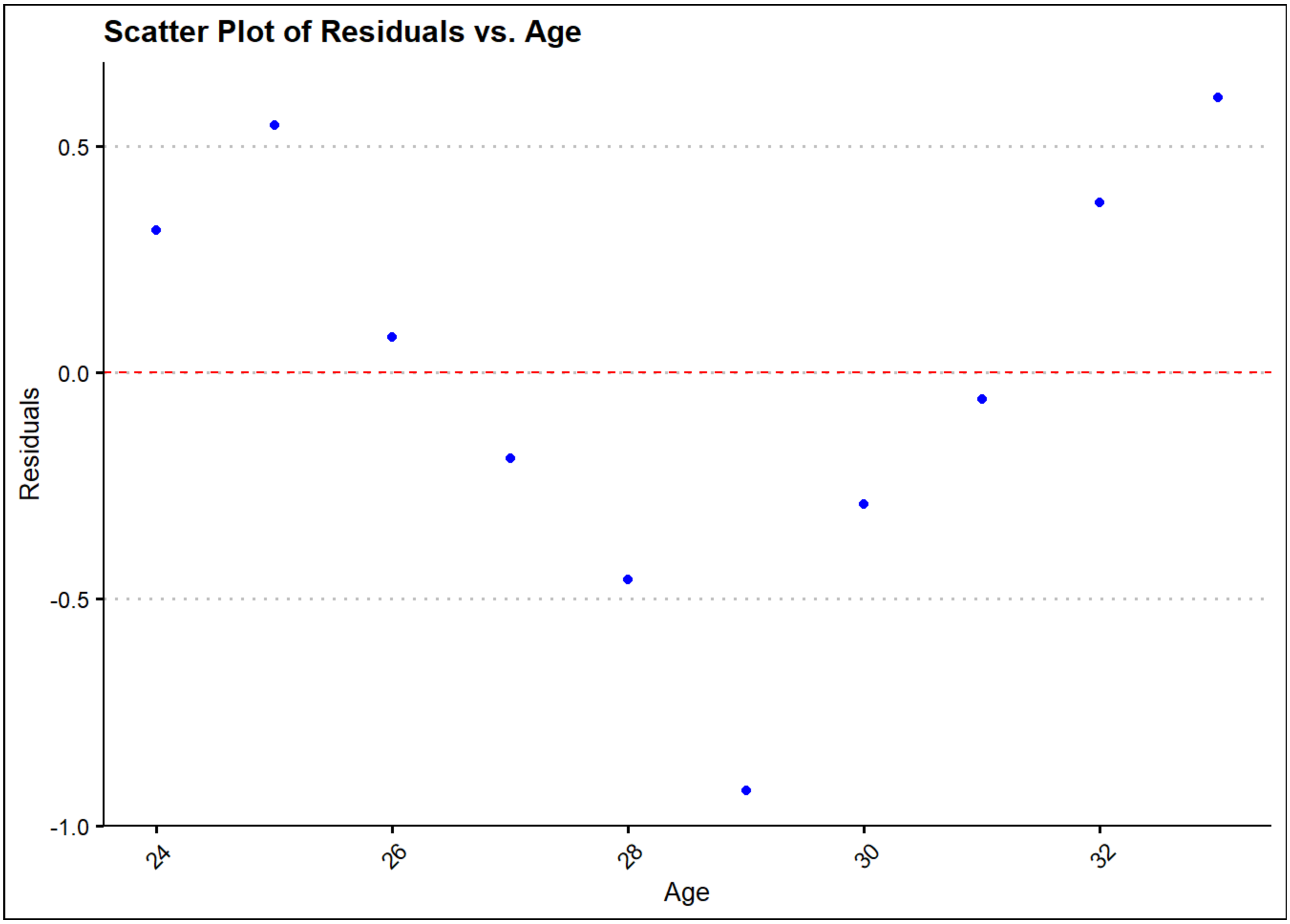
```
thy = read.csv("E:/Biostat/Biostatistics/PHL_1700/Data/Raw/thyroxine-dat-1.csv")
n = 10
sum_x = 285
sum_x2 = 8205
sum_y = 78.4
sum_y2 = 627.88
sum_xy = 2264.7
Lxx = sum_x2 - (sum_x)^2/n
Lyy = sum_y2 - (sum_y)^2/n
Lxy = sum_xy - sum_x * sum_y/n
b = Lxy/Lxx
a = (sum_y - b * sum_x)/n
```

The equation is $y = -2.6273 + 0.3673 * x$.

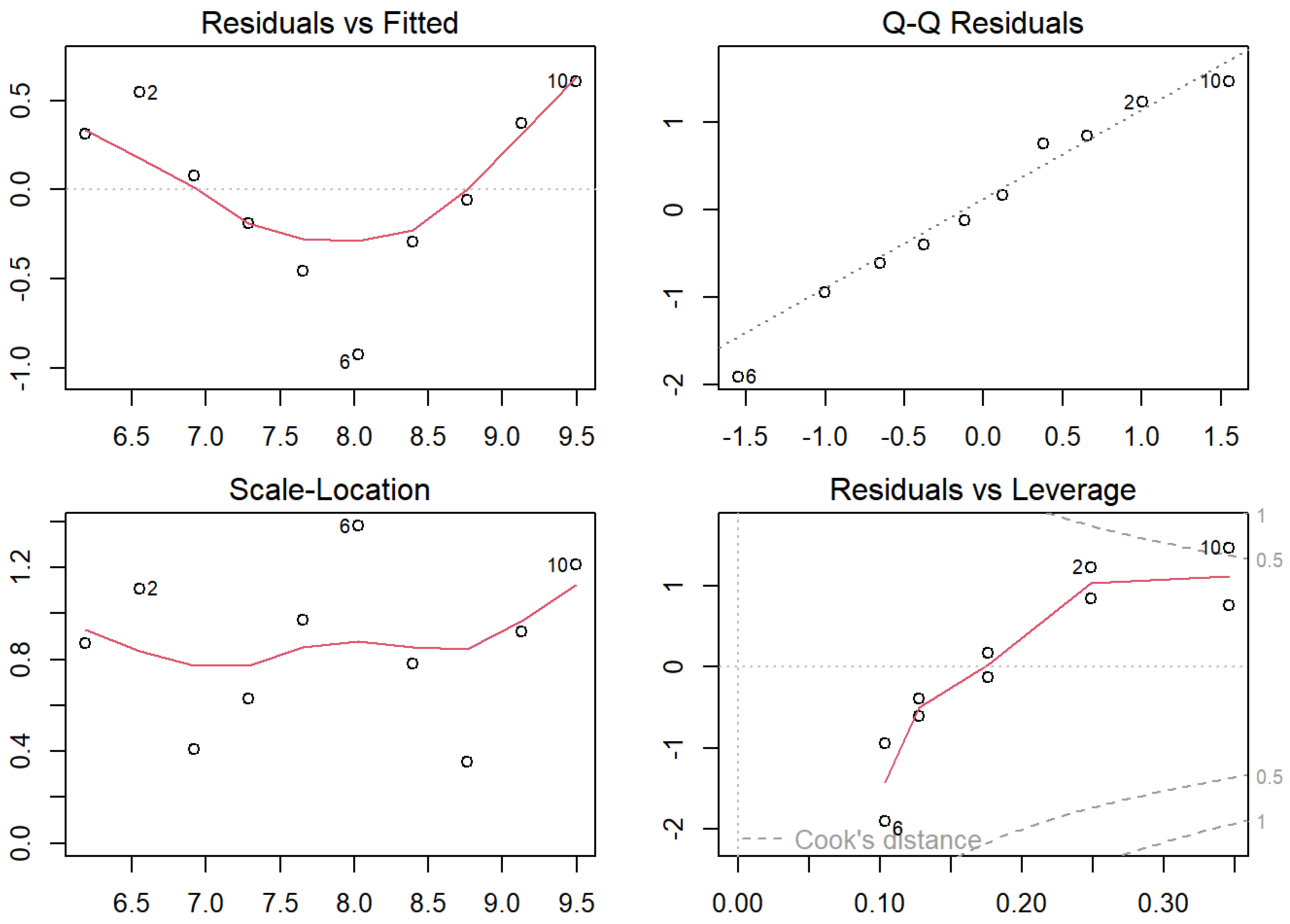
#11.51

```
thy_model = lm(thyroxine ~ age, thy)

# Scatter plot of residuals vs. age
ggplot(thy, aes(x = age, y = resid(thy_model))) + geom_point(color = "blue") +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(title = "Scatter Plot of Residuals vs. Age", x = "Age",
        y = "Residuals")
```



```
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(thy_model, cex.axis = 1, cex.lab = 1)
```



Scatter plot of residuals vs. age: There's no indication of heteroscedasticity. Also, it shows that the residuals are independent.However, there's indication of non-linearity Residuals vs. Fitted Plot: The curved pattern indicates that the relationship between age and thyroxine might not be linear as the residuals are not evenly distributed around zero. This suggests that a non-linear model might be a good fit for the data, Q-Q Plot: The residuals deviate slightly from the line, especially at the tails, suggesting mild departures from normality. However, since the dataset is small, we can treat it as normally distributed. Scale-Location Plot: The spread of residuals almost remains consistent. This suggests that the assumption of homoscedasticity is still reasonable. Residuals vs. Leverage Plot: Observations 6, 2, and 10 show higher leverage and Cook's distance, indicating they may be influential points. These observations could disproportionately impact the model's fit. Overall, the linear model satisfies the assumption of normality and homoscedasticity. It does not satisfy the assumption of linearity. Therefore, the linear regression does not provide a good fit to the data.

#11.50

```
anova(thy_model)
```

```
## Analysis of Variance Table
##
## Response: thyroxine
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## age        1 11.1284  11.128  42.482 0.0001846 ***
## Residuals  8  2.0956   0.262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value is less than 0.05, we can reject the null hypothesis and state that there’s a significant relationship between the mean thyroxine level and gestational age. The ANOVA test indicates that there’s a statistically significant association but it does not specify whether this relationship is linear or non-linear. It only states that age has an effect on thyroxine level. As we’ve concluded that there’s a no-linear relationship between the two variables, the p-value might be biased.