R_Activity_Assignment_6

Dahee Ahn

2024-10-09

1. Download the loblolly_pines data set and load it into R.

```
loblolly_trees <- read.table(file= "C:/Users/chemk/OneDrive/Desktop/Classes/ENT6707_DataAnalysi
s/week8/loblolly_pines.txt", header=TRUE, sep="\t")
sum(is.na(loblolly_trees))
```

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

```
str(loblolly_trees)
```

```
## 'data.frame':    84 obs. of  3 variables:
##  $ height: num  4.51 10.89 28.72 41.74 52.7 ...
##  $ age   : int  3 5 10 15 20 25 3 5 10 15 ...
##  $ Seed  : int  301 301 301 301 301 301 303 303 303 303 ...
```

```
head(loblolly_trees)
```

```
##   height age Seed
## 1   4.51   3  301
## 2  10.89   5  301
## 3  28.72  10  301
## 4  41.74  15  301
## 5  52.70  20  301
## 6  60.92  25  301
```

```
tail(loblolly_trees)
```

```
##    height age Seed
## 79   3.46   3  331
## 80   9.05   5  331
## 81  25.85  10  331
## 82  39.15  15  331
## 83  49.12  20  331
## 84  59.49  25  331
```

```
summary(loblolly_trees)
```

```
##      height           age            Seed      
##  Min.   : 3.46   Min.   : 3.0   Min.   :301.0  
##  1st Qu.:10.47   1st Qu.: 5.0   1st Qu.:307.0  
##  Median :34.00   Median :12.5   Median :317.0  
##  Mean   :32.36   Mean   :13.0   Mean   :316.1  
##  3rd Qu.:51.36   3rd Qu.:20.0   3rd Qu.:325.0  
##  Max.   :64.10   Max.   :25.0   Max.   :331.0
```

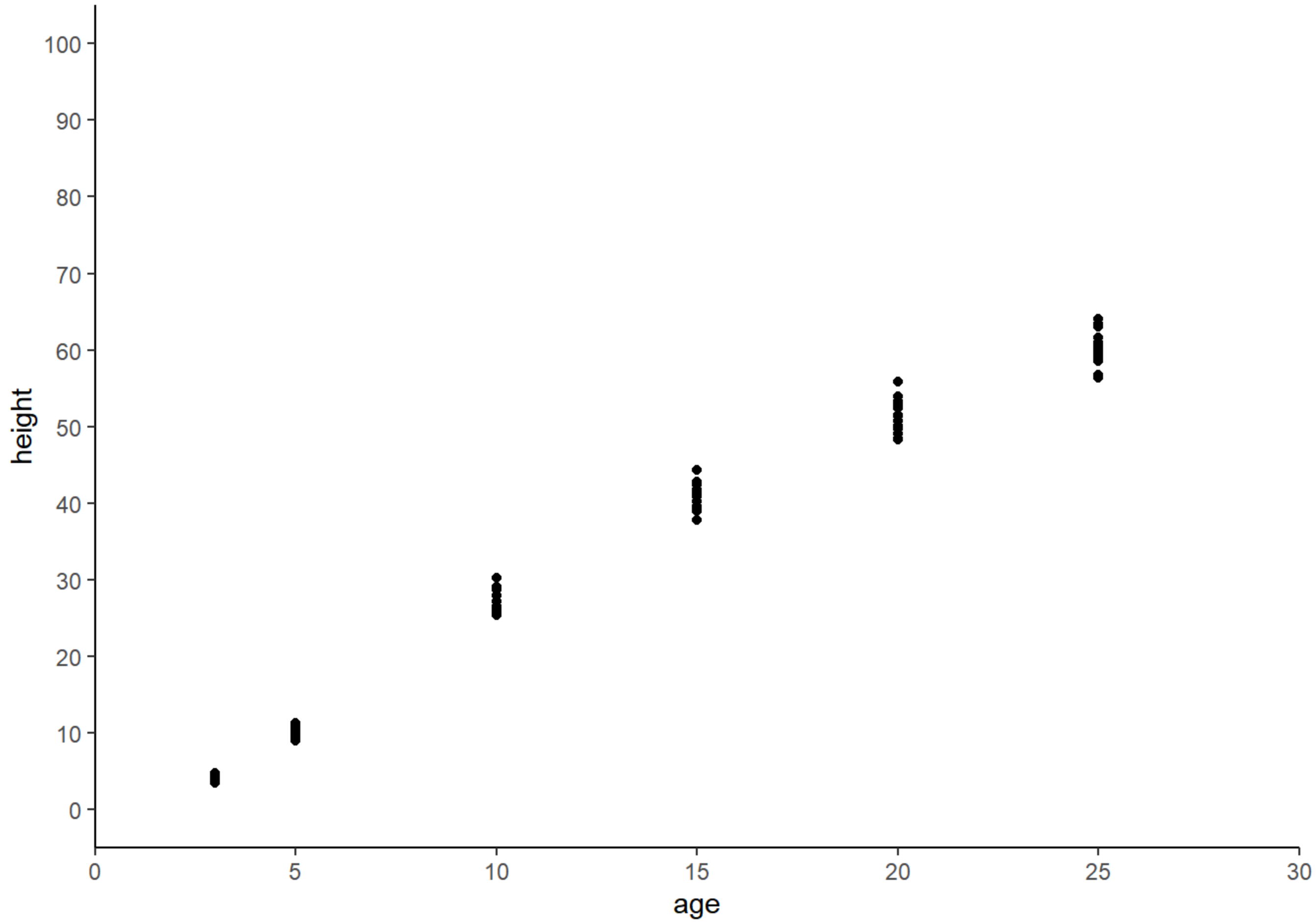
```
View(loblolly_trees)
```

2. Create a scatter plot of height (feet) as a function of age (years) using ggplot2.

Q. IF you fit a regression of height as a function of age, what would be your guesses for the Estimates of the Intercept and slope for Age? (i.e., please complete this step without doing any formal analyses).

A. The intercept would be close to 0 or exactly 0, indicating the height of seedling may be lower than 10cm (not sure for unit) or the seed may take time to germinate. From the graph below, it can be seen that height increases with age, particularly rapidly from ages 5 to 20, but the growth slows down from ages 20 to 25. Therefore, based on my estimation, the growth rate(slope) may continue to decelerate as age increases.

```
library(ggplot2)
ggplot(data=loblolly_trees, mapping=aes(x=age, y=height))+geom_point()+theme_classic()+scale_y_c
ontinuous(limits=c(0,100), breaks = seq(0, 100, by = 10))+scale_x_continuous(limits=c(0,30), bre
aks=seq(0,40, by=5), expand=c(0,0))
```



3. Fit a simple linear regression of height as a function of age. Name this model fit_linear.

```
fit_linear <- lm(height~age, data=loblolly_trees)
summary(fit_linear)
```

```
##
## Call:
## lm(formula = height ~ age, data = loblolly_trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0207 -2.1672 -0.4391  2.0539  6.8545
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.31240    0.62183  -2.111   0.0379 *
## age          2.59052    0.04094  63.272   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.947 on 82 degrees of freedom
## Multiple R-squared:  0.9799, Adjusted R-squared:  0.9797
## F-statistic:  4003 on 1 and 82 DF,  p-value: < 2.2e-16
```

4. According to fit_linear, how tall is the average loblolly pine at 0 years old? What about 15 years old? What does the model tell you about the average height gained per year by loblolly pines?

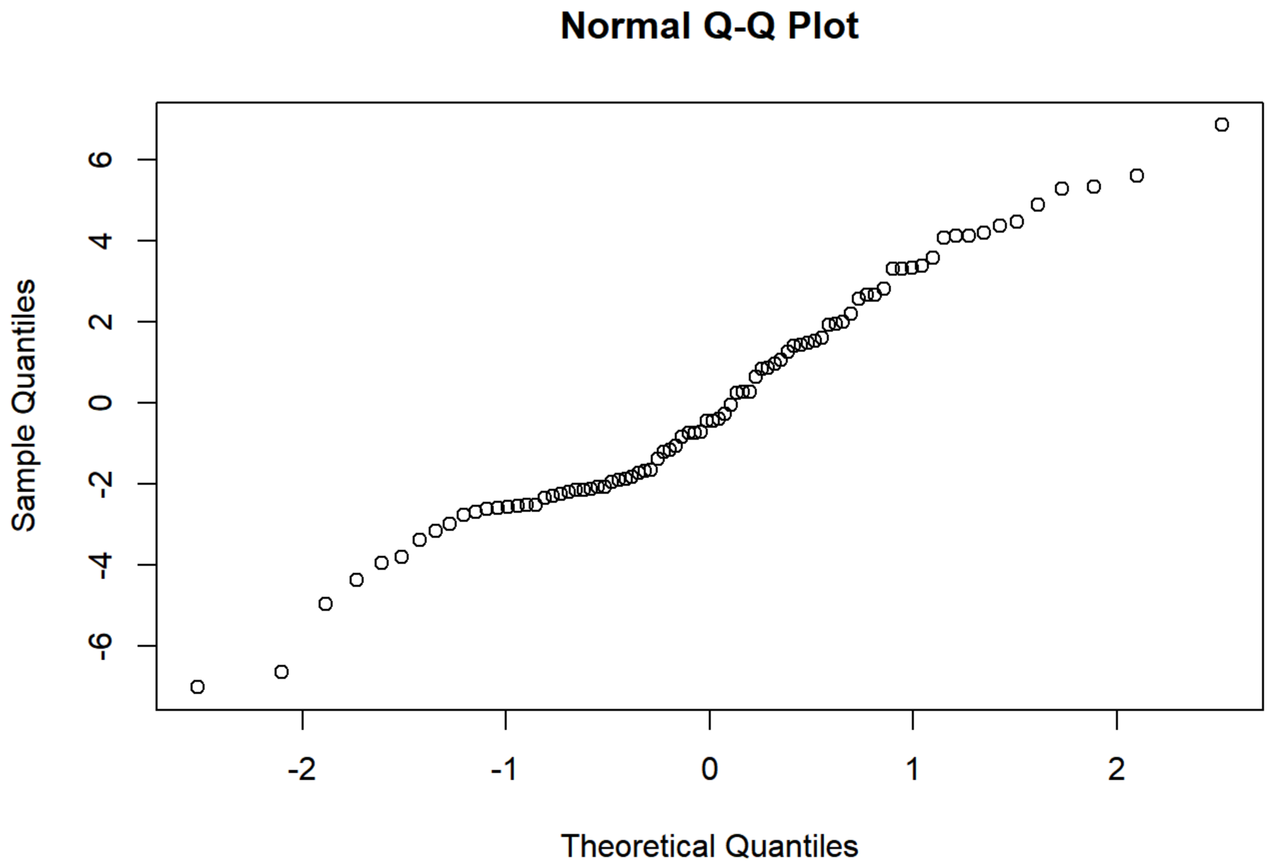
A1. Average height of loblolly pine at 0 years old : -1.312 (cm)

A2. Average height of loblolly pine at 15 years old : y = b + mx, y = -1.3 + (2.6 × 15), y = 37.7 (cm)

A3. It tells me the average height gained per year by loblolly pines is approximately 2.6, according to the Estimate of age.

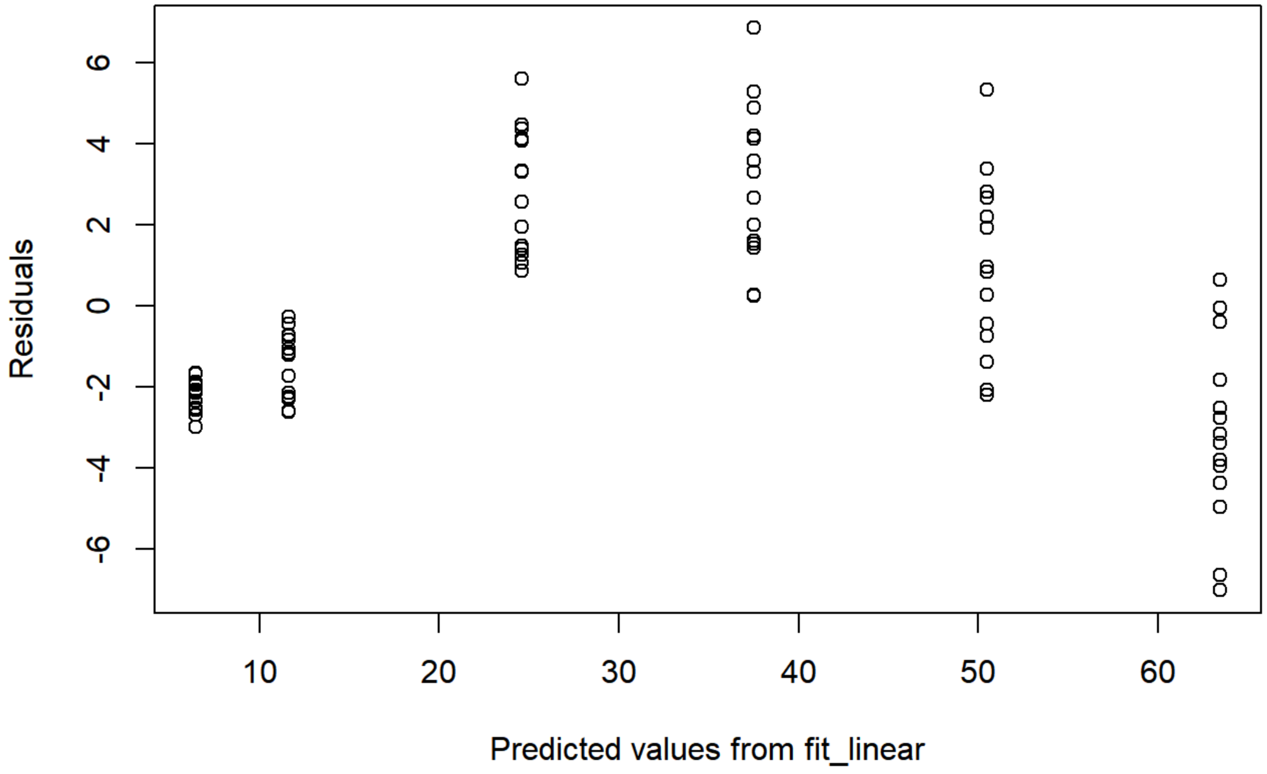
5. Provide a residual and a Q-Q plot for fit_linear. Do the residuals look normally distributed and homoscedastic? Are you happy with how well the model fits the data? Explain your reasoning.

```
qqnorm((residuals(fit_linear)))
```



A. In Q-Q plot, the points are closely aligned along the diagonal line and do not deviate significantly, suggesting that they follow a normal distribution.

```
plot(residuals(fit_linear)~fitted.values(fit_linear), xlab="Predicted values from fit_linear", ylab="Residuals")
```



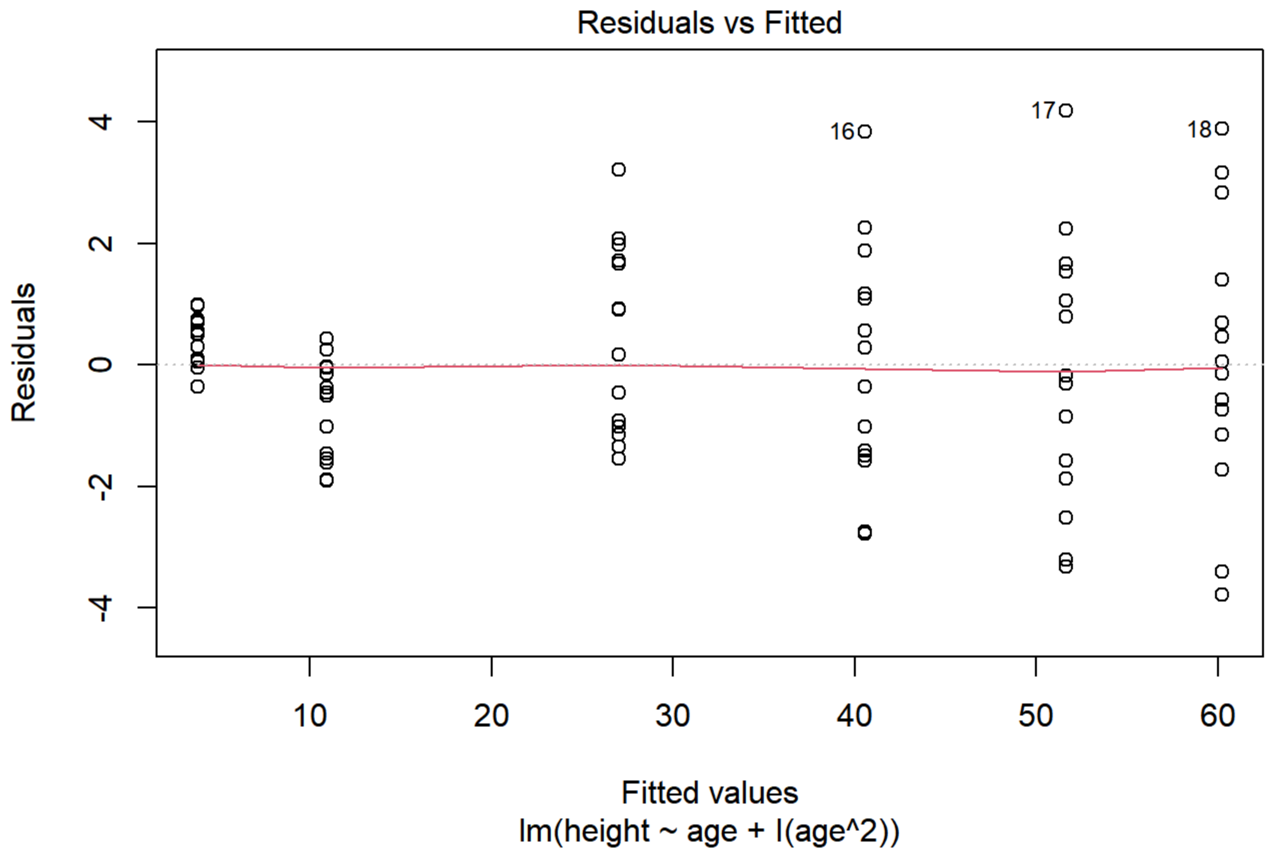
A. According to the result of the residual plot, the residuals appear to be normally distributed around the X-axis and Y-axis. However, the uneven clustering and the hump-shaped pattern suggest a violation of the assumption.

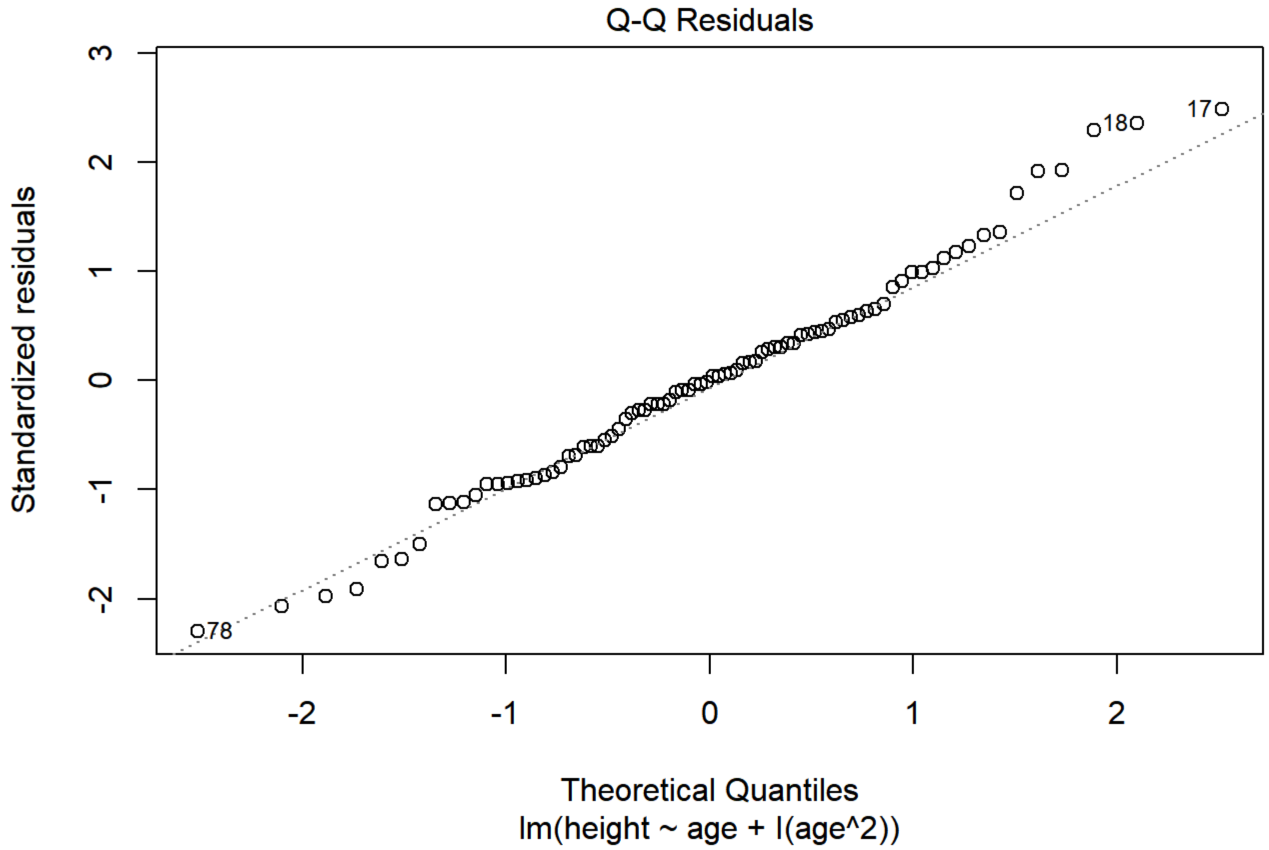
6. Fit the same model as fit_linear but add a polynomial term for age (i.e., a quadractic or squared version of age). Name this new model fit_poly (coding hint: see the supplemental document “Transformations and Curvilinear Models”).

```
fit_poly <- lm(height~age + I(age^2), data=loblolly_trees)
summary(fit_poly)
```

```
## 
## Call:
## lm(formula = height ~ age + I(age^2), data = loblolly_trees)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.7902 -1.1496  0.0183  0.9401  4.1815 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -7.607232   0.608017  -12.51   <2e-16 ***
## age          3.959044   0.109236   36.24   <2e-16 ***
## I(age^2)    -0.049838   0.003884  -12.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.702 on 81 degrees of freedom
## Multiple R-squared:  0.9934, Adjusted R-squared:  0.9932 
## F-statistic:  6079 on 2 and 81 DF,  p-value: < 2.2e-16
```

```
plot(fit_poly, which=c(1,2))
```


Residuals vs Fitted
Residuals
4
2
0
-2
-4
16
17
18
10
20
30
40
50
60
Fitted values
lm(height ~ age + I(age^2))


Q-Q Residuals
Standardized residuals
3
2
1
0
-1
-2
18
17
78
-2
-1
0
1
2
Theoretical Quantiles
lm(height ~ age + I(age^2))

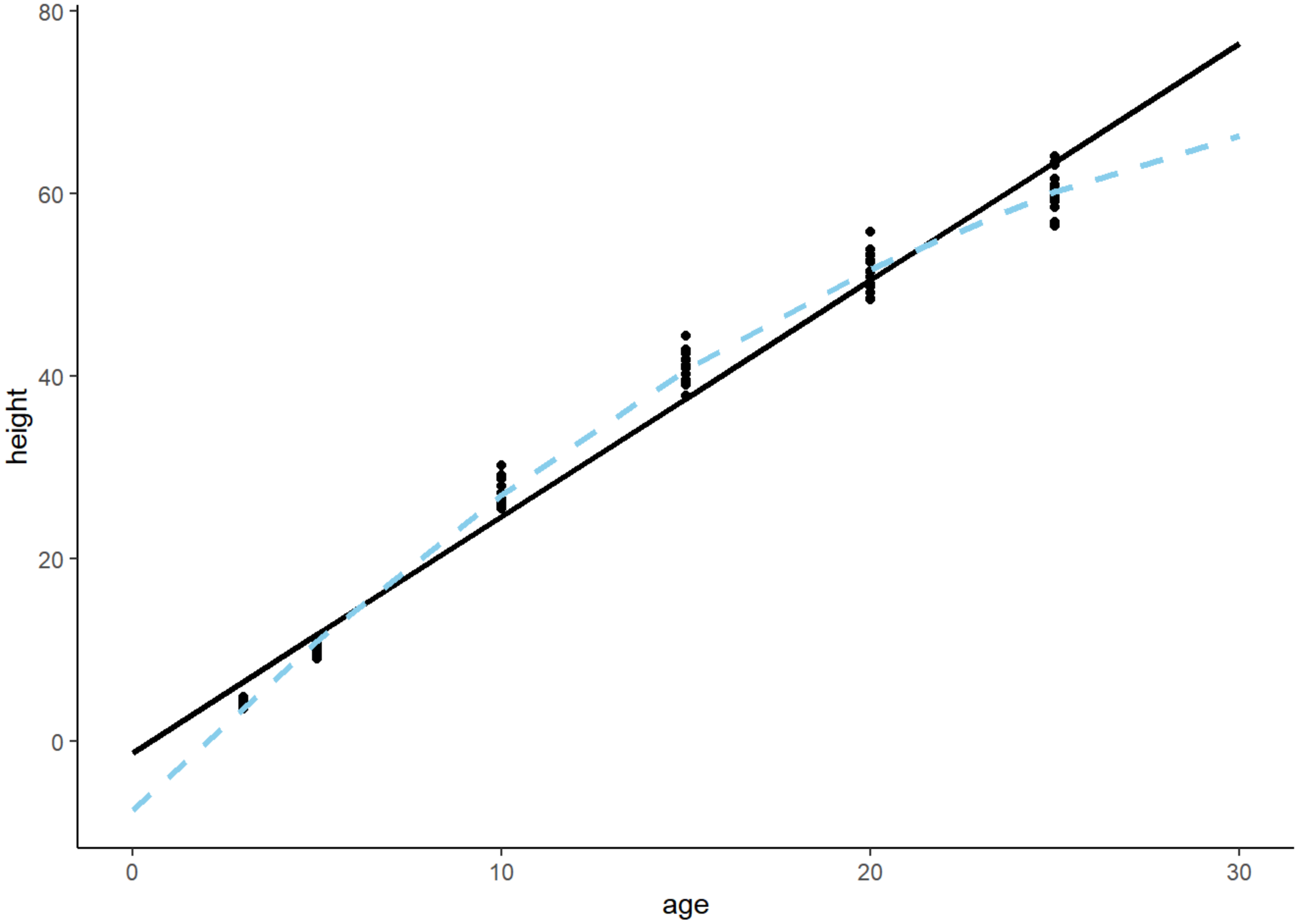
7. Look at the residuals of fit_poly. Do they look normally distributed and homoscedastic? You don’t need to provide diagnostic graphs.

A. Q-Q plot looks distributed better than the previous one. In this polynomial graph, the hump shaped pattern disappears while there are still some clustering points of the residuals. Overall, the residual plot looks unbiased and heteroscedastic.

8. Reproduce the plot you created above of height as a function of age and add fit lines from fit_linear and fit_poly to the graph (i.e., your graph should have two lines overlaid on the cloud of raw data points). Color the line from fit_poly as your favorite color.

```r
new_data <- data.frame(age= seq(0, 30, 5))
new_data$Pred_lm <- predict(fit_linear, newdata=new_data)
new_data$Pred_poly <- predict(fit_poly, newdata=new_data)
```

```r
library(ggplot2)
ggplot(data=loblolly_trees, mapping=aes(x=age, y=height))+geom_point()+theme_classic()+geom_line(data=new_data, aes(x=age, y=Pred_lm), linewidth=1.2) + geom_line(data=new_data, aes(x=age, y=Pred_poly), linetype="dashed", color="skyblue", linewidth=1.2)
```


80
60
40
20
0
height
0
10
20
30
age