15/20

Homework 3 - STAT 511

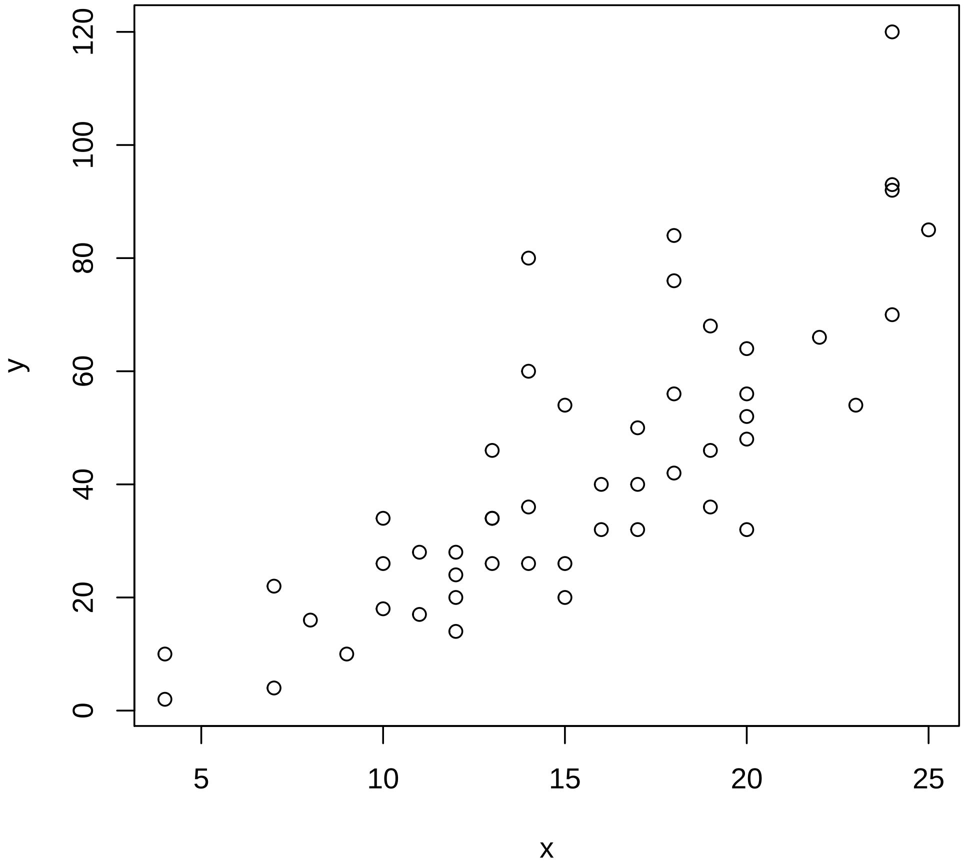
Amal Agarwal

Answer 1

(a) Let us call y as the stopping distance (response variable) and x as the speed (predictor variable). For Exploratory Data Analysis (EDA) following scatter plots were obtained:

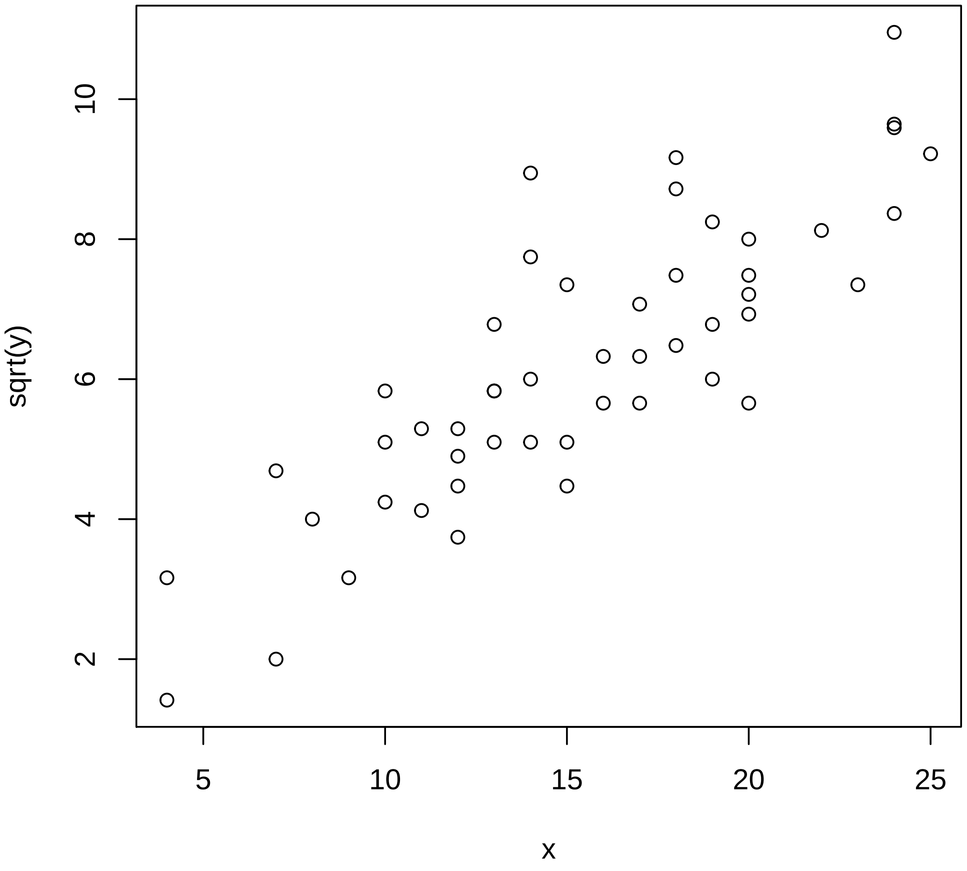
- Plot of y vs. x

4/5



✓

- Plot of $\sqrt{y}$ vs. x



It can be observed that the scatter plot of y vs. x indicates a non linear relationship between y and x. Further tha density of data points for $x < 7$ and $x > 21$ is relatively lower than for $8 < x < 20$. The transformations $log(y)$ vs. x can be ruled out since it does not look linear. The transformation y vs. x^2 looks promising in terms of fitting a simple linear model. But, the scatter plot that shows the most closest linear relationship after the transformation is $\sqrt{y}$ vs. x. Note: The transformations that didn't seem to work are not shown explicitly in the form of scatter plots.

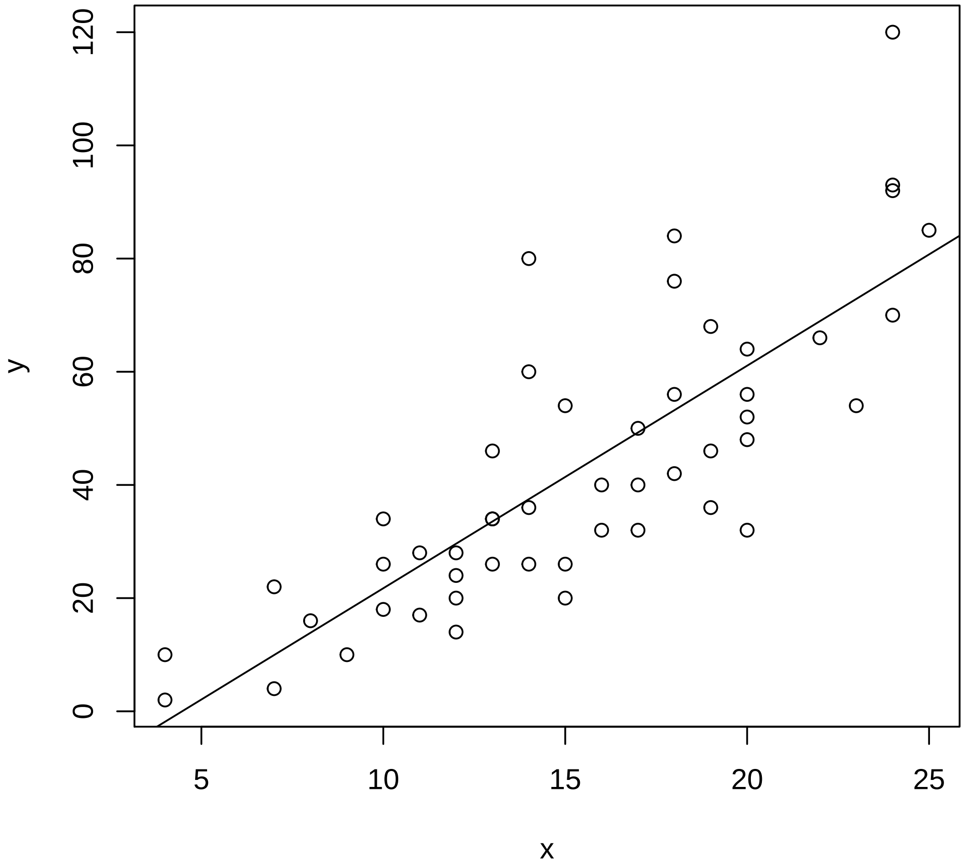
(b) The simple linear regression model in R that can be used to fit the given data is

$$y = \beta_0 + \beta_1 x + \epsilon$$

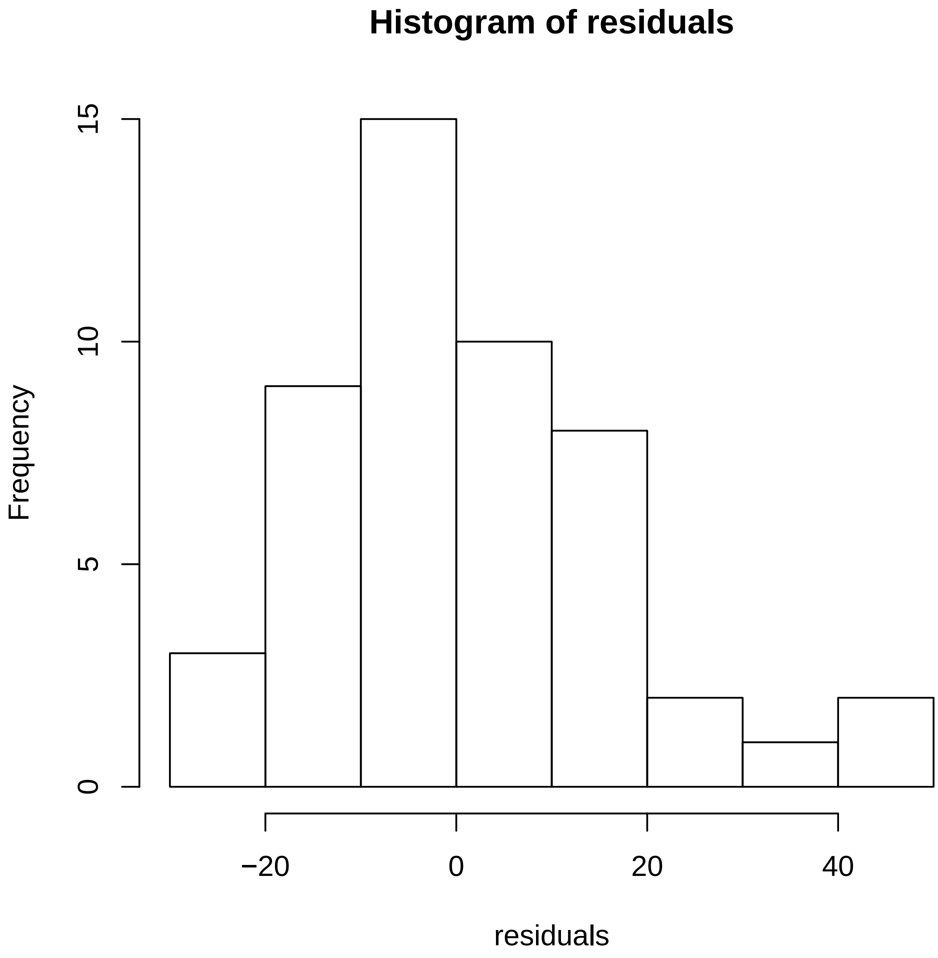
where $\epsilon \sim N(0, \sigma^2)$

(c) Using lm in R, to estimate model parameters and residuals.

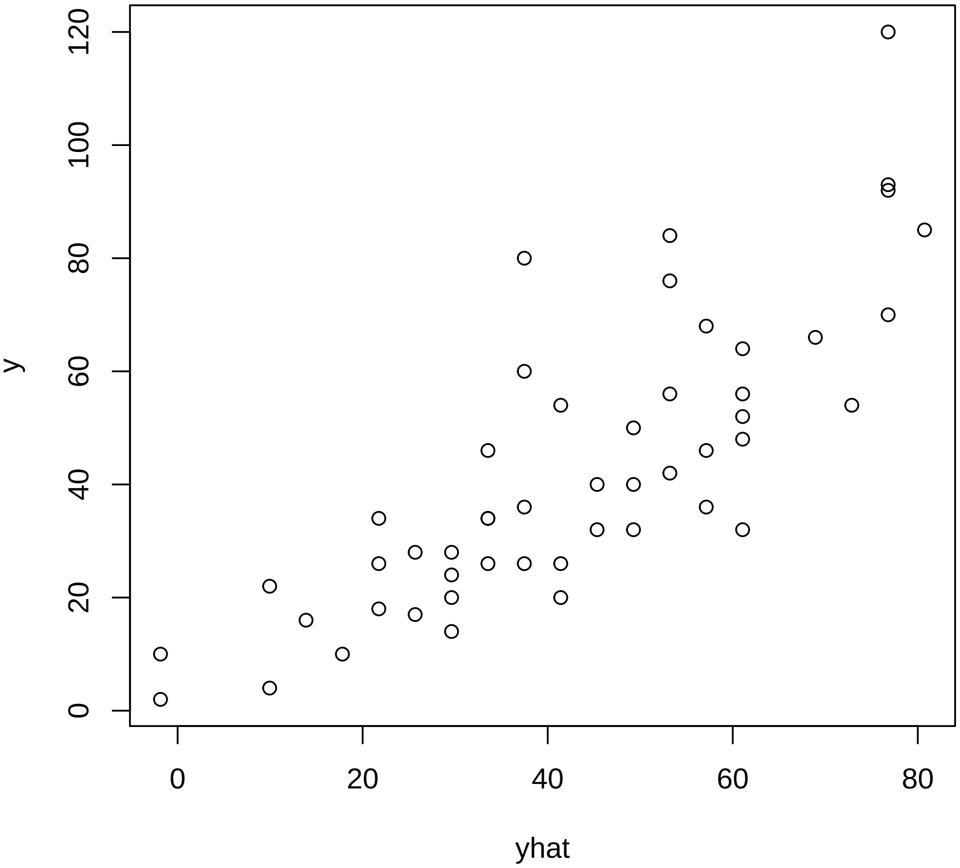
```
> fit1=lm(y~x)
> plot(x,y)
> abline(fit1)
```



```
> residuals=fit1$resid
> hist(residuals)
> ## the table given in summary(fit)
> summary(fit1)$coef

               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -17.579095  6.7584402 -2.601058 1.231882e-02
x             3.932409  0.4155128  9.463990 1.489836e-12

> summary(fit1)$sigma

[1] 15.37959
```



(d)
```
> yhat=fit1$fitted.values
> plot(yhat,y)
```



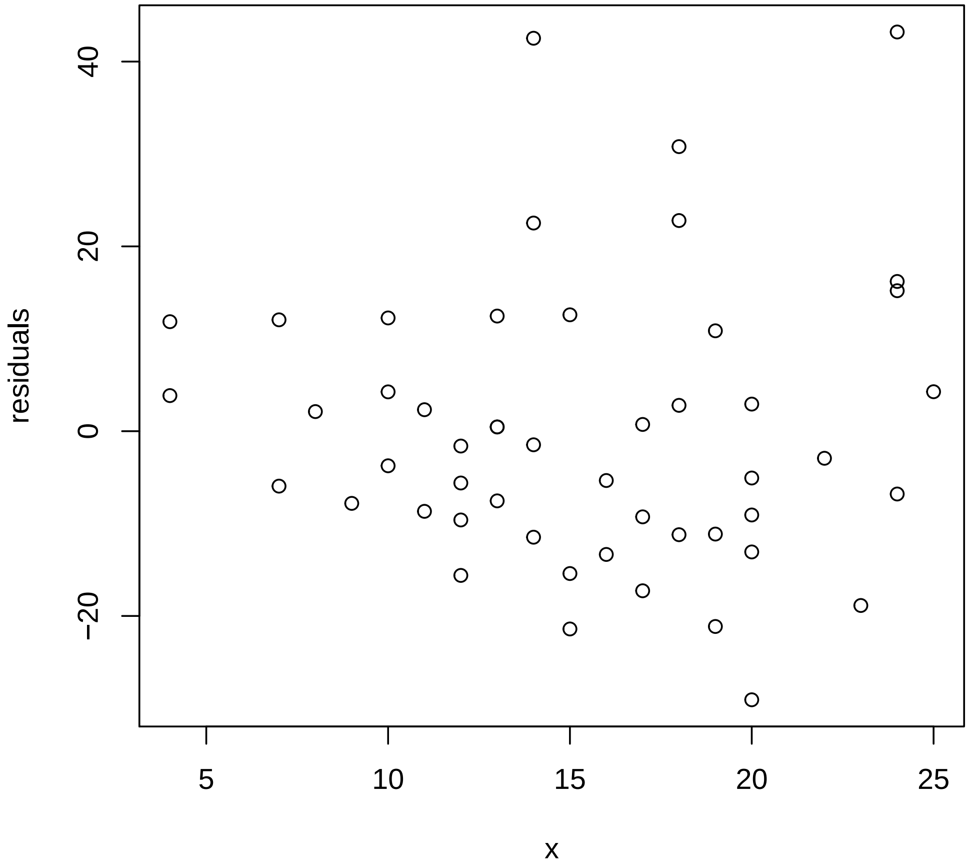
The mean of the residuals is given as

```
> mean(residuals)
```

```
[1] 8.65974e-17
```

This is very close to zero which indicates that our assumption of $E(\epsilon) = 0$ is satisfied. To check for homoscedasticity, residuals can be plotted against x as

```
> plot(x,residuals)
```



Clearly, as x increases, the fluctuations of residuals around zero increases which implies that the estimated variance of the residuals increase with the covariate ”speed”. This funnel shaped behaviour is called heteroscedascity which clearly violates our assumption of homoscedascity in the classical linear model.

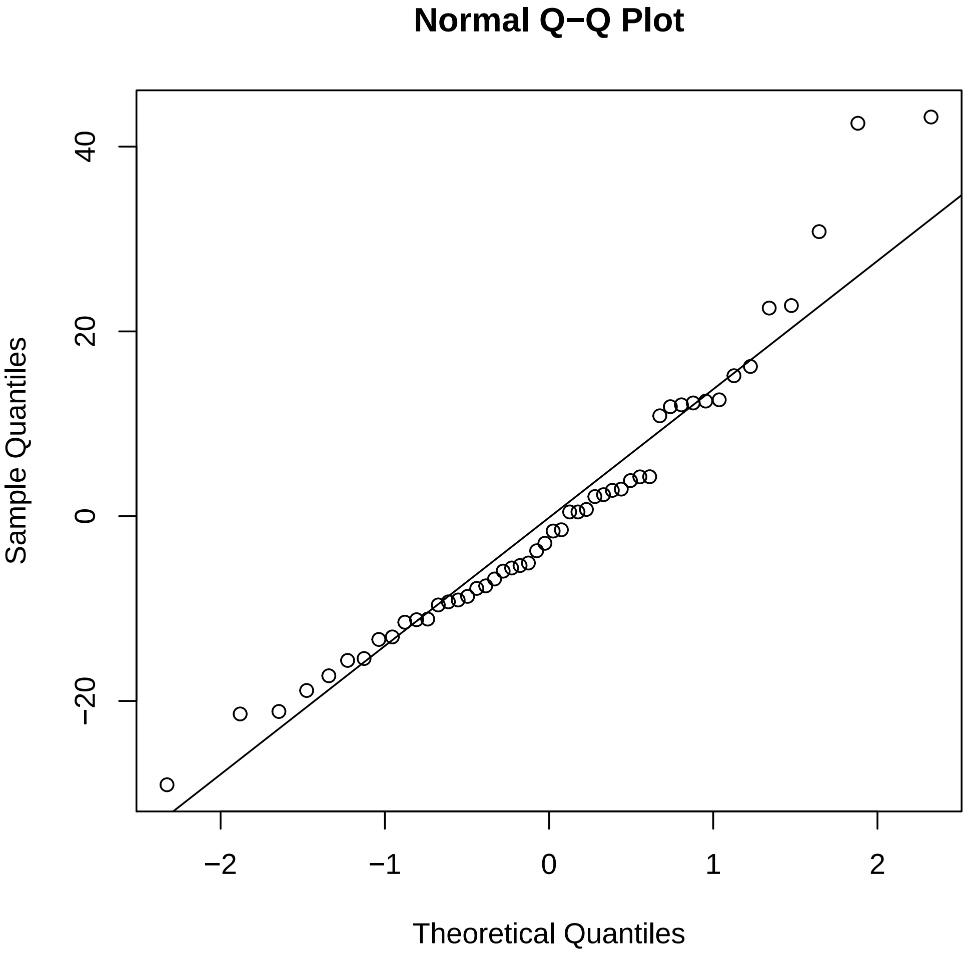
The column rank of the design matrix X can be calculated as:

```
[1] 2
```

which confirms our assumption of full column rank design matrix.

- QQ Plot for residuals

```
> qqnorm(residuals)
> qqline(residuals)
```



The above QQ plot shows significant deviation from our normality assumption.

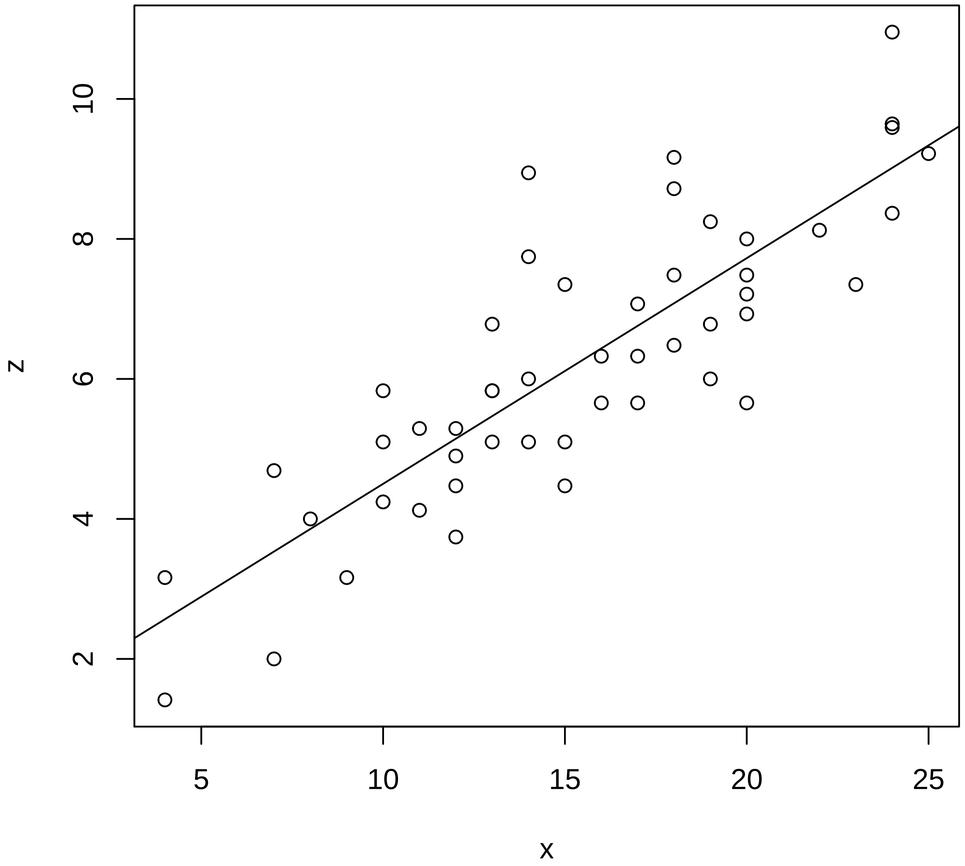
(e) The modified linear regression model in R that can be used to fit the given data is

$$z = \beta_0 + \beta_1 x + \epsilon$$

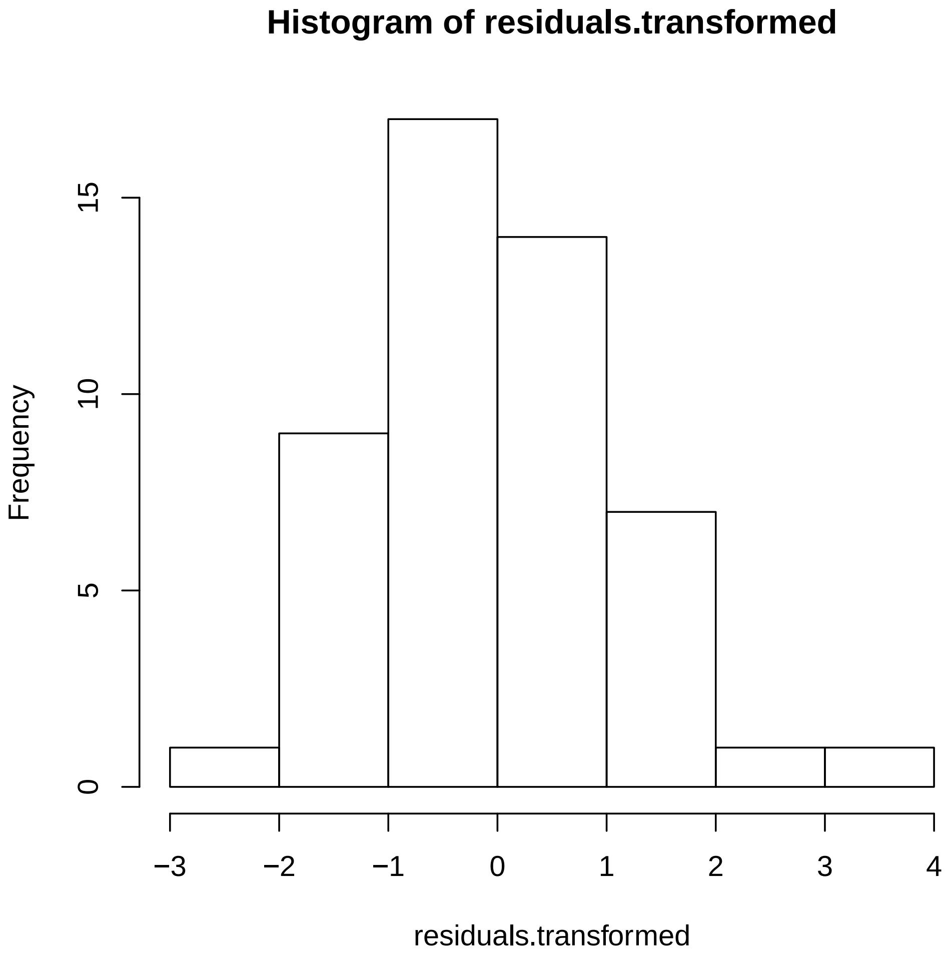
where $z = \sqrt{y}$ is the transformed response variable. Here we are using the power transform on the response variable as indicated by EDA.

(f) Using lm in R, to estimate modified model parameters and residuals.

```
> z<-sqrt(y)
> fit2=lm(z~x)
> plot(x,z)
> abline(fit2)
```

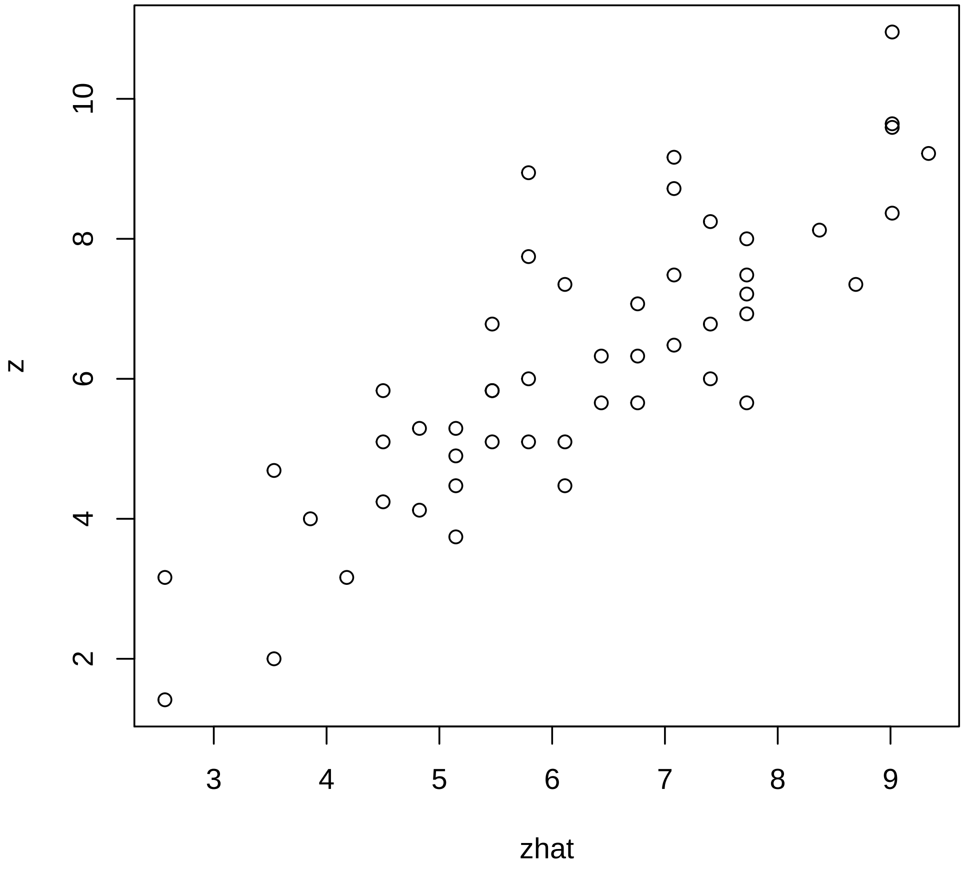


```
> residuals.transformed=fit2$resid
> hist(residuals.transformed)
```



```
> summary(fit2)$coef

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 1.2770502 0.48444202  2.636126 1.126220e-02
x           0.3224125 0.02978377 10.825106 1.773141e-14

> summary(fit2)$sigma

[1] 1.102402
```

```
> zhat=fit2$fitted.values
> plot(zhat,z)
```

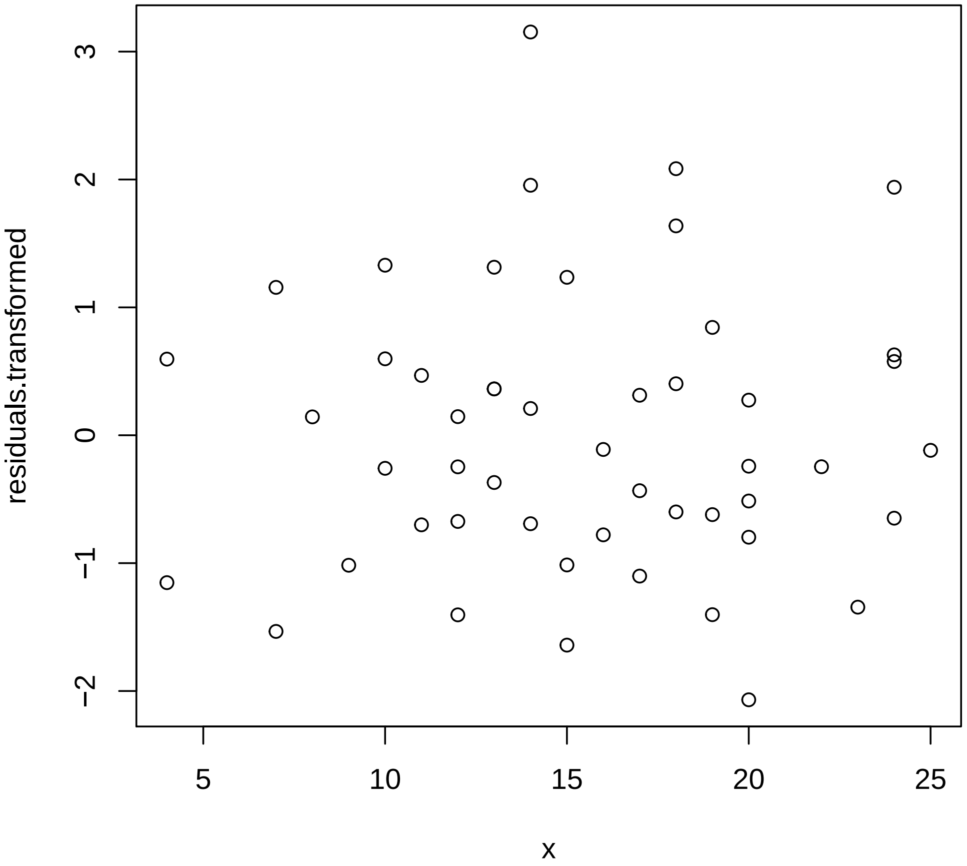


The mean of transformed residuals is given as

```
> mean(residuals.transformed)

[1] 3.109546e-17
```

This is again very close to zero which indicates that our assumption of $E(\epsilon) = 0$ is satisfied. To check for homoscedasticity, the transformed residuals can be plotted against x as

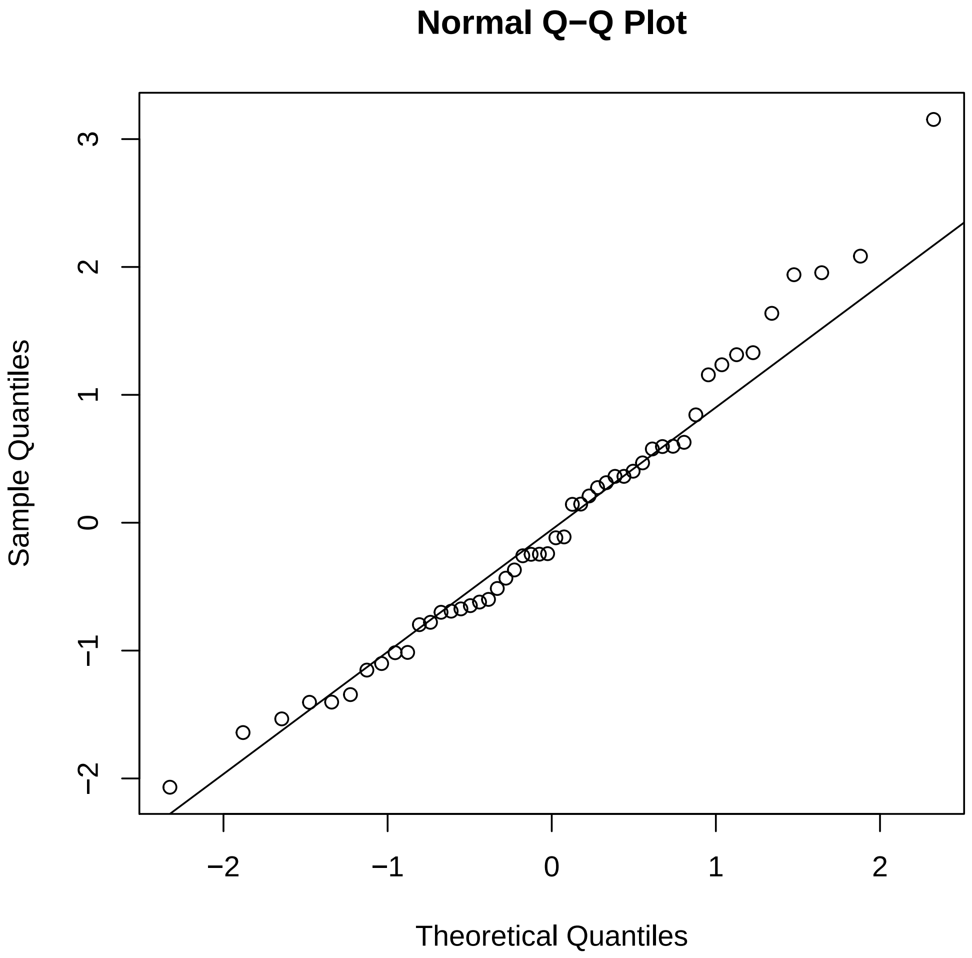
```
> plot(x,residuals.transformed)
```



This looks much better since the fluctuations of the transformed residuals around zero is more or less uniform and does not depend on x. Thus our assumption of homoscedascity in the classical linear model is satisfied after the transformation.

- QQ Plot for residuals

```
> qqnorm(residuals.transformed)
> qqline(residuals.transformed)
```



Except for a little fat tail on the right side, the above QQ plot is much better than the previous one. This shows that after transformation, the errors are approximately normally distributed.

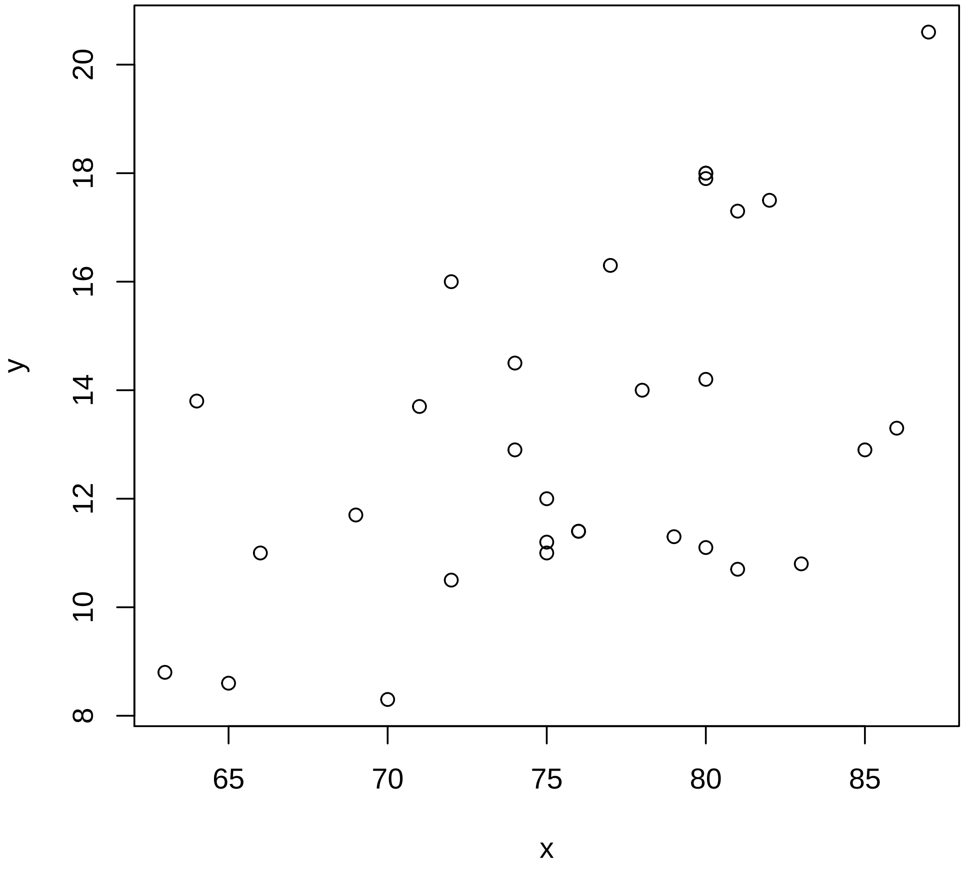
Interpretation of $\tilde{\beta}$, $\tilde{\sigma}^2$?

Answer 2

(a) Let us call y as the tree girth (response variable) and x as the tree height (predictor variable). For Exploratory Data Analysis (EDA) following scatter plots were obtained:

- Plot of y vs. x

4/5



✓

It can be observed that the scatter plot of y vs. x doesn't indicate a well defined non linear relationship between y and x. The various power transformations and log transformation also do not help much. Again the scatter plots corresponding to various transformations that don't work, are not shown explicitly.

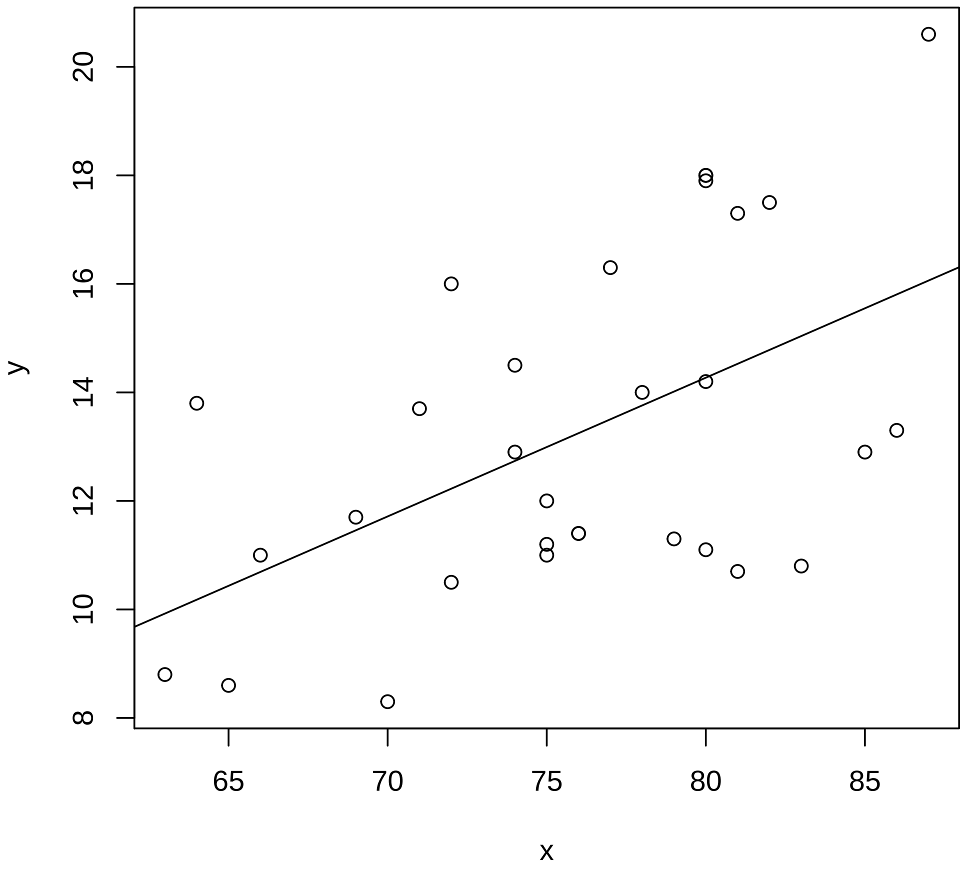
(b) The simple linear regression model in R that can be used to fit the given data is

$$y = \beta_0 + \beta_1 x + \epsilon$$

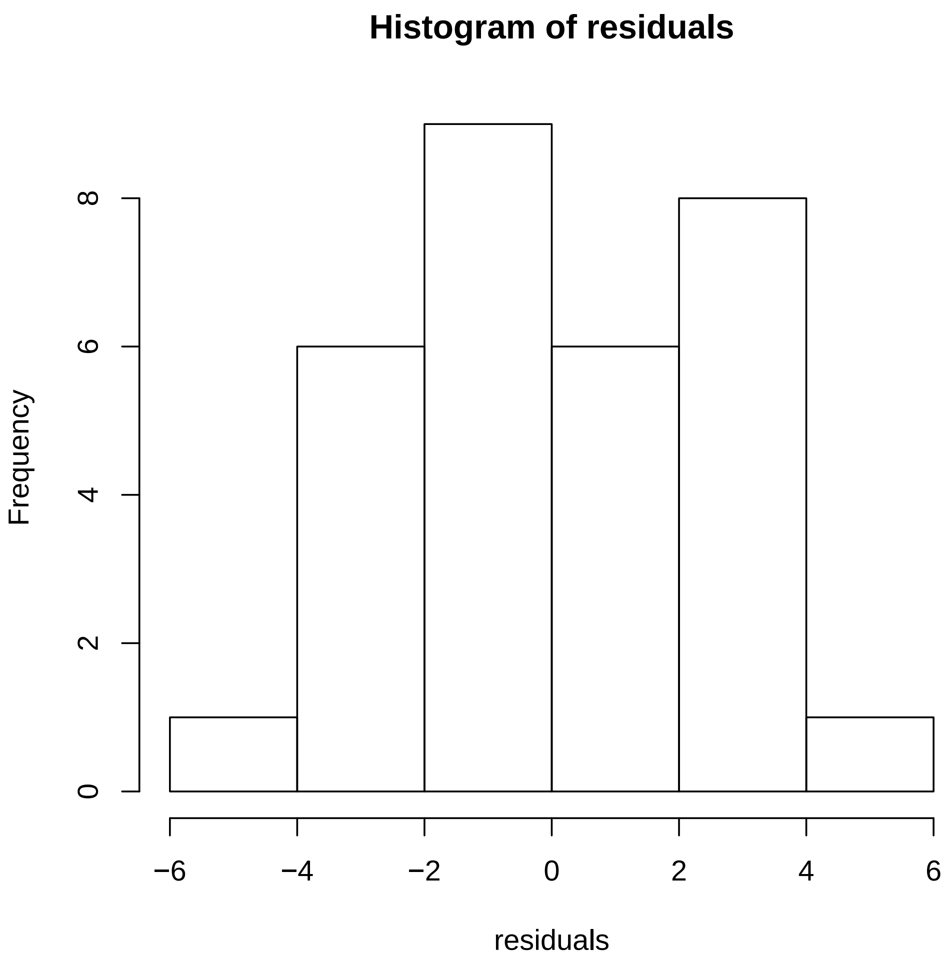
where $\epsilon \sim N(0, \sigma^2)$

(c) Using lm in R, to estimate model parameters and residuals.

```
> fit1=lm(y~x)
> plot(x,y)
> abline(fit1)
```

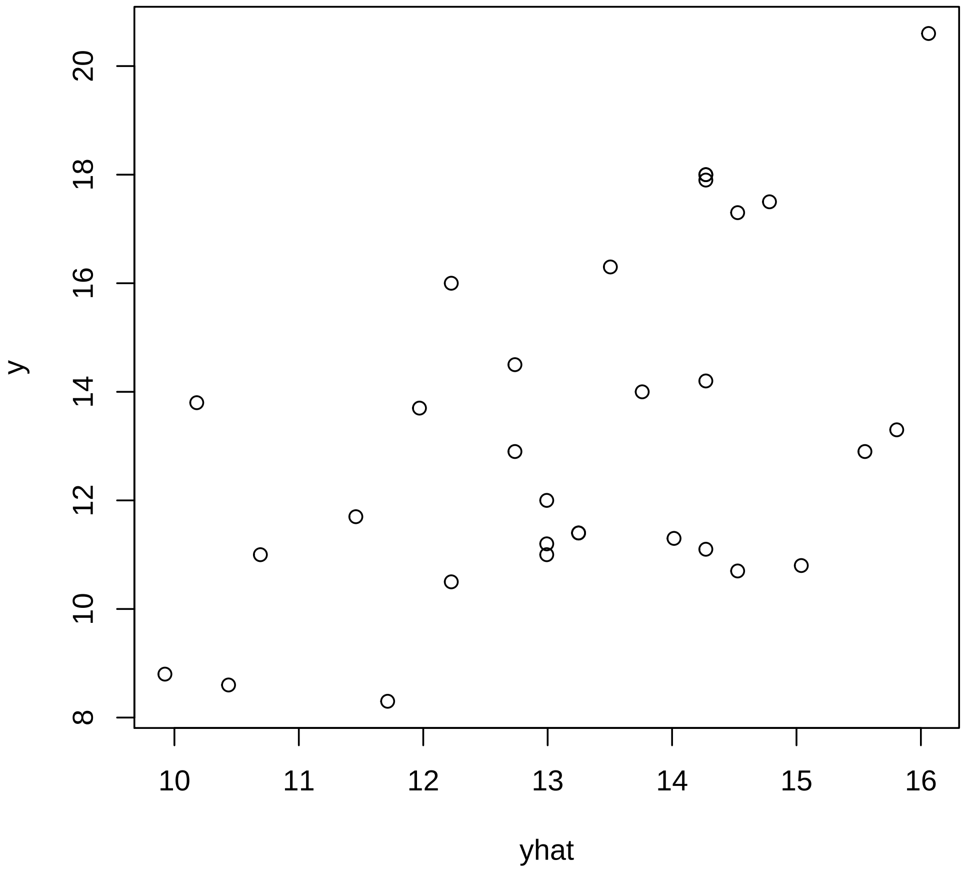


```
> residuals=fit1$resid
> hist(residuals)
```



```
> summary(fit1)$coef

              Estimate Std. Error   t value    Pr(>|t|)
(Intercept) -6.1883945  5.9601994 -1.038286 0.307716768
x            0.2557471  0.0781583  3.272169 0.002757815

> summary(fit1)$sigma

[1] 2.727713
```

(d)
```
> yhat=fit1$fitted.values
> plot(yhat,y)
```



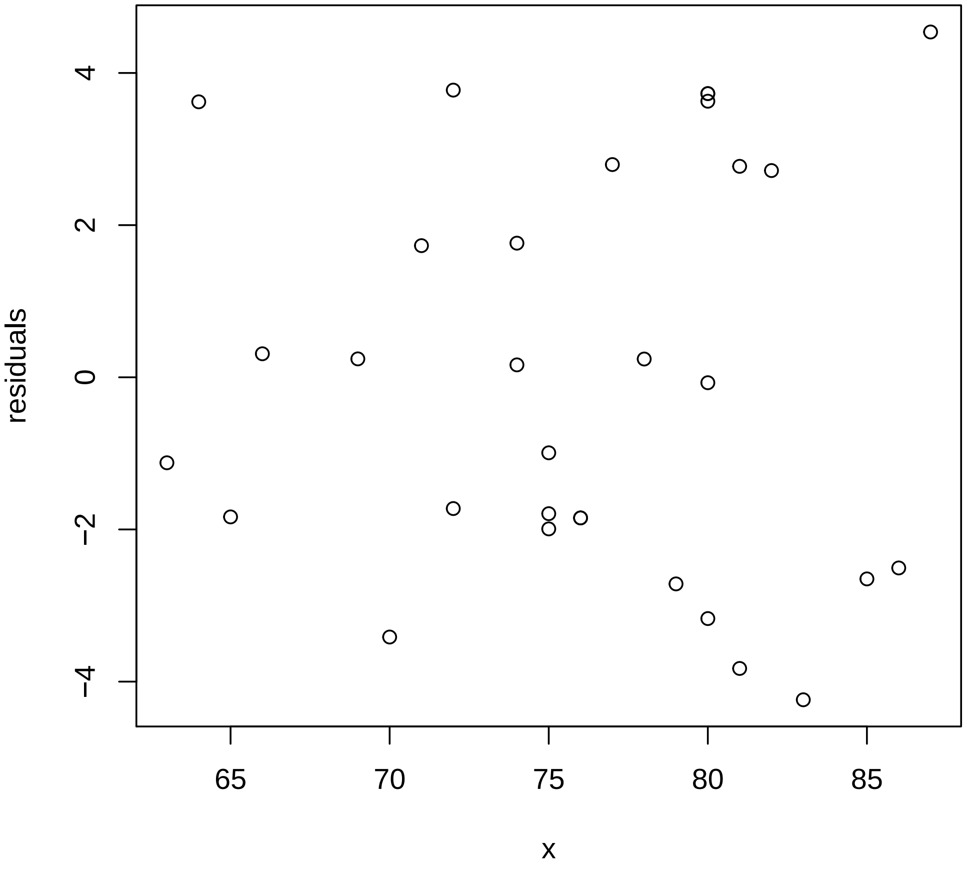
The mean of residuals is given as

```
> mean(residuals)
```

```
[1] -3.192451e-17
```

This is very close to zero which indicates that our assumption of $E(\epsilon) = 0$ is satisfied. To check for homoscedasticity, residuals can be plotted against x as

```
> plot(x,residuals)
```



The fluctuations of residuals around zero look uniform and do not seem to be dependent on the predictor. Thus our assumption of homoscedascity in the classical linear model is verified here.

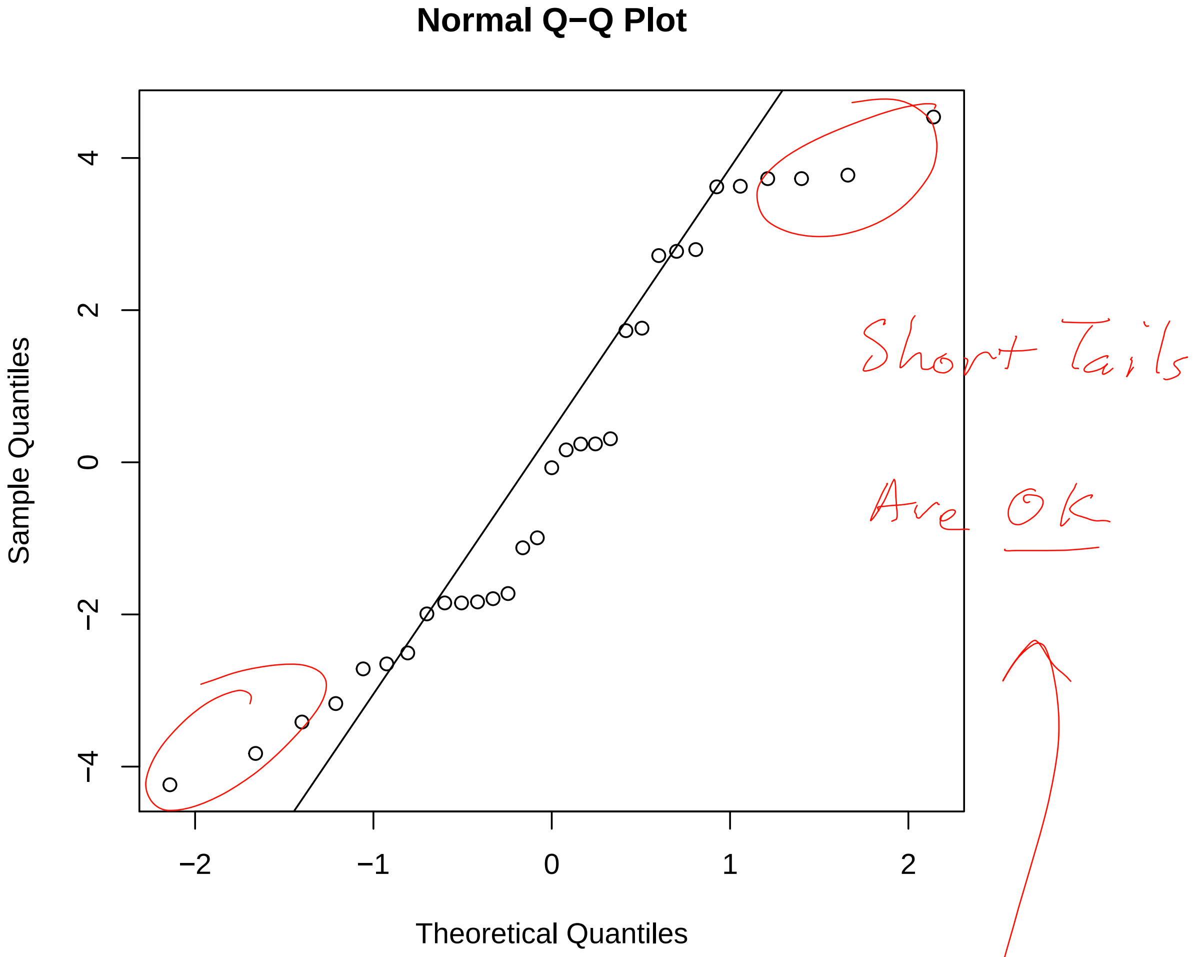
The column rank of the design matrix X can be calculated as:

```
[1] 2
```

which confirms our assumption of full column rank design matrix.

- QQ Plot for residuals

```
> qqnorm(residuals)
> qqline(residuals)
```



The above QQ plot shows significant deviation from our normality assumption.

(e) From the scatter plots of various transformations, it seems that there is no transformation that works better than the existing model which suggests that there is no need for modify the model.

Interpretation?

Answer 3

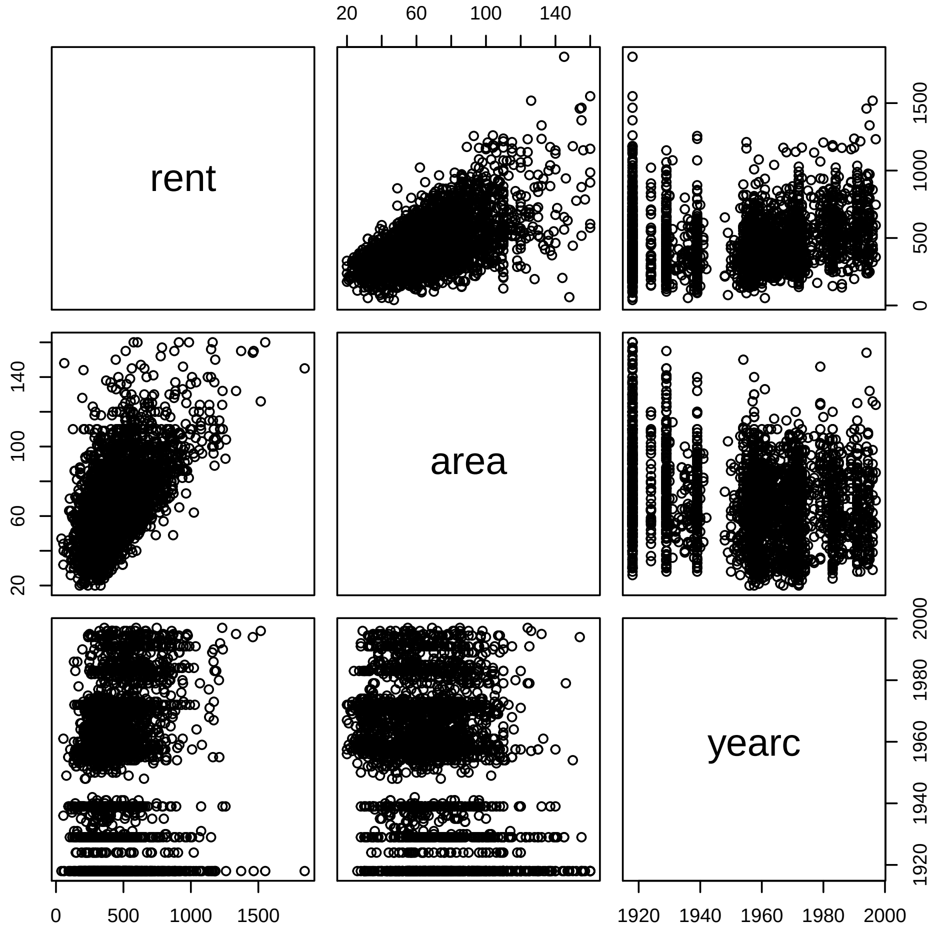
7/10

(a)

```
> Munich=read.csv("rent99.raw",sep=" ")
> Munich.rent<-Munich[,c(1,3,4)]
> str(Munich.rent)
```

```
'data.frame':        3082 obs. of  3 variables:
 $ rent : num  121 437 356 283 807 ...
 $ area : int  35 104 29 39 97 62 31 61 72 75 ...
 $ yearc: num  1939 1939 1971 1972 1985 ...
```

```
> pairs(Munich.rent)
> rent<-Munich.rent[,1]
> area<-Munich.rent[,2]
> yearc<-Munich.rent[,3]
> rentsqm<-Munich[,2]
```



✓

For exploratory data analysis, pairwise scatter plots between the rent, area and yearc variables have been drawn here. The scatter plot between rent and area suggests that as area is increased, the variability of rent increases i.e. for larger area apartments there are more options available in terms of rent, as compared to the smaller area apartments. This is the well known phenomenon of heteroscedascity and will result in the improper estimation of the variances of estimated regression coefficients $\hat{\beta}$. Further, the scatterplots involving yearc variable variable suggests that there are clear cut breaks during the time intervals of 1920-1930 and 1940-1950. The former points out that the data during the period might be extrapolated and therefore not that reliable to analyze while the latter might be accounted by noting that there might be a scarcity of apartments built during World War 2.

(b) Fitting the given linear regression model,

$$rent = \beta_0 + \beta_1 area + \beta_2 yearc + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

```
> fit=lm(rent~area+yearc)
> summary(fit)

Call:
lm(formula = rent ~ area + yearc)

Residuals:
    Min      1Q  Median      3Q     Max
-734.76  -94.75  -10.87   82.55 1063.17

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4775.5942   244.3549  -19.54   <2e-16 ***
area            5.3618     0.1165   46.01   <2e-16 ***
yearc           2.4913     0.1239   20.11   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 149.3 on 3079 degrees of freedom
Multiple R-squared: 0.4181,        Adjusted R-squared: 0.4177
F-statistic:  1106 on 2 and 3079 DF,  p-value: < 2.2e-16
```

The summary of the model fitted shows the estimated values of the parameters and the various quantiles for residuals. Further the mean of the residuals vector is calculated as

```
> residuals=fit$resid
> mean(residuals)

[1] -4.664461e-15
```

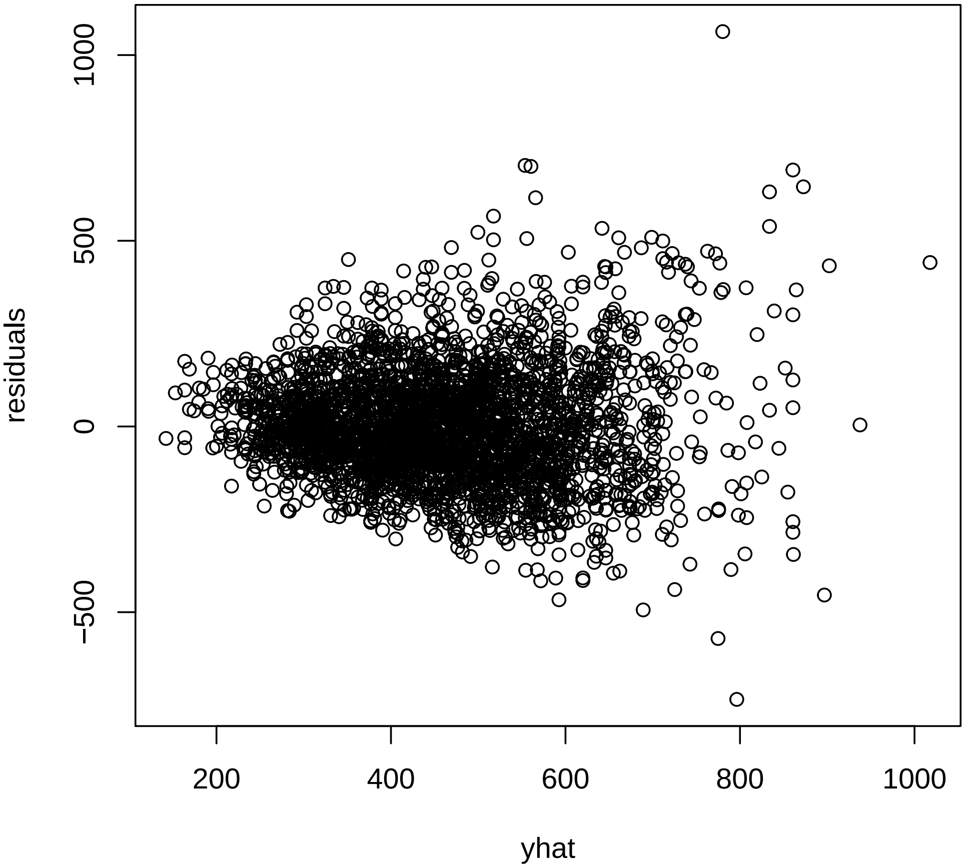
which is very close to zero, thus verifying our assumption of $E(\epsilon) = 0$. The column rank of the design matrix X can be calculated as:

```
[1] 3
```

which confirms our assumption of full column rank design matrix. The relevant plots after fitting the model and their interpretations are given as follows:

- Plot of model residuals vs. fitted values:

```
> yhat=fit$fitted.values
> plot(yhat,residuals)
```

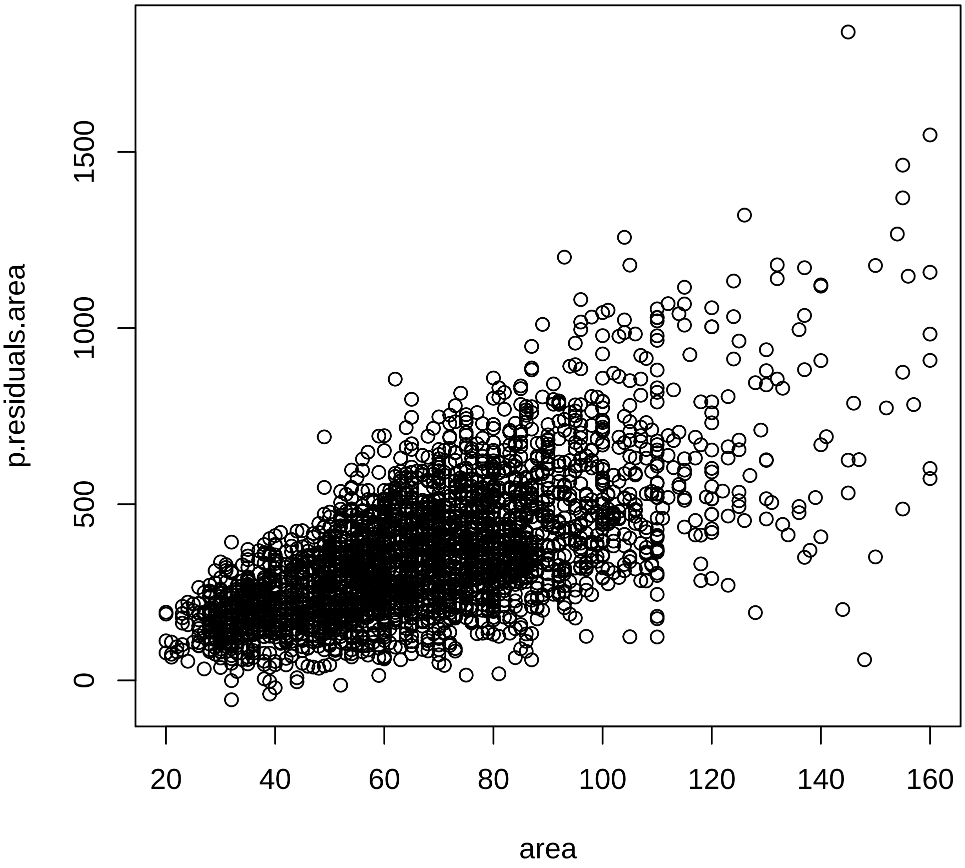


✓

The funnel shaped behaviour of the above residual plot indicates increase in variance of the residuals with the fitted values (or the covariates) which implies heteroscedascity. Thus here the assumption of homoscedascity in our classical linear model is violated.

- Partial residual plot for area

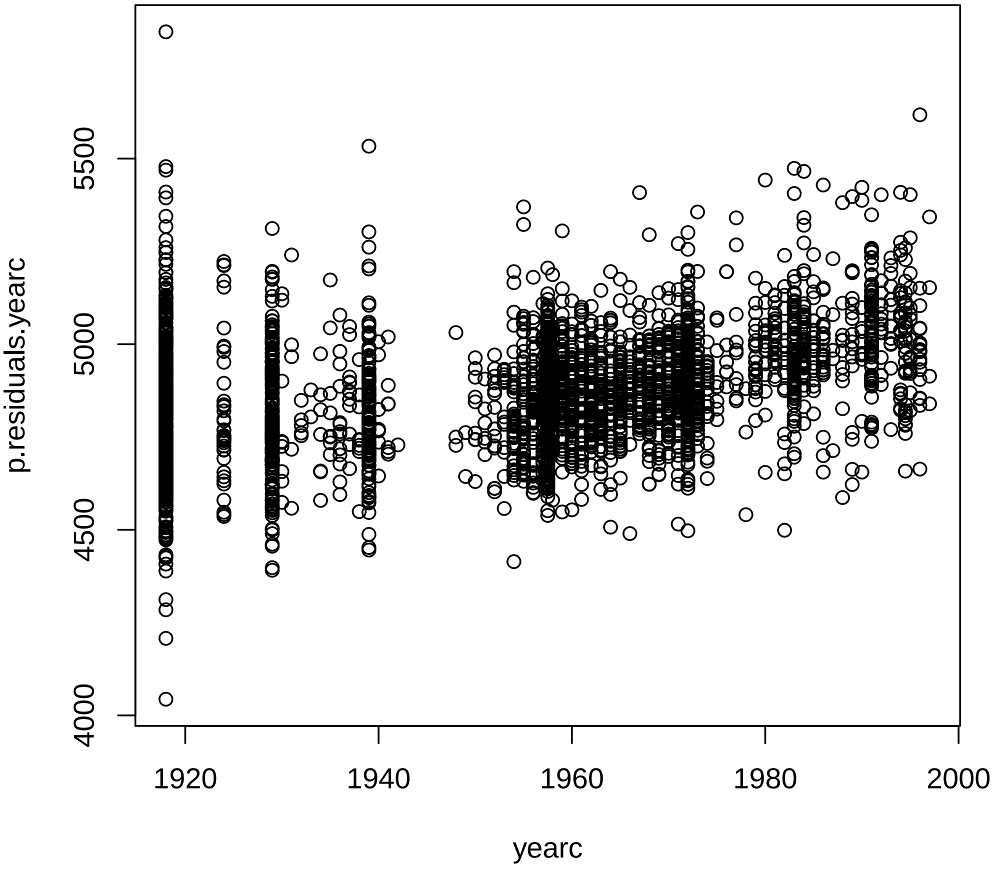
```
> beta=fit$coef
> p.residuals.area=residuals+beta[2]*area
> plot(area,p.residuals.area)
```



The above partial residual plot of area shows that the variance of rent increases as the area is increased. Further the points are much more separated towards larger areas which shows that there are lesser options in terms of choosing rent values from a particular interval for larger area apartments.

- Partial residual plot for yearc

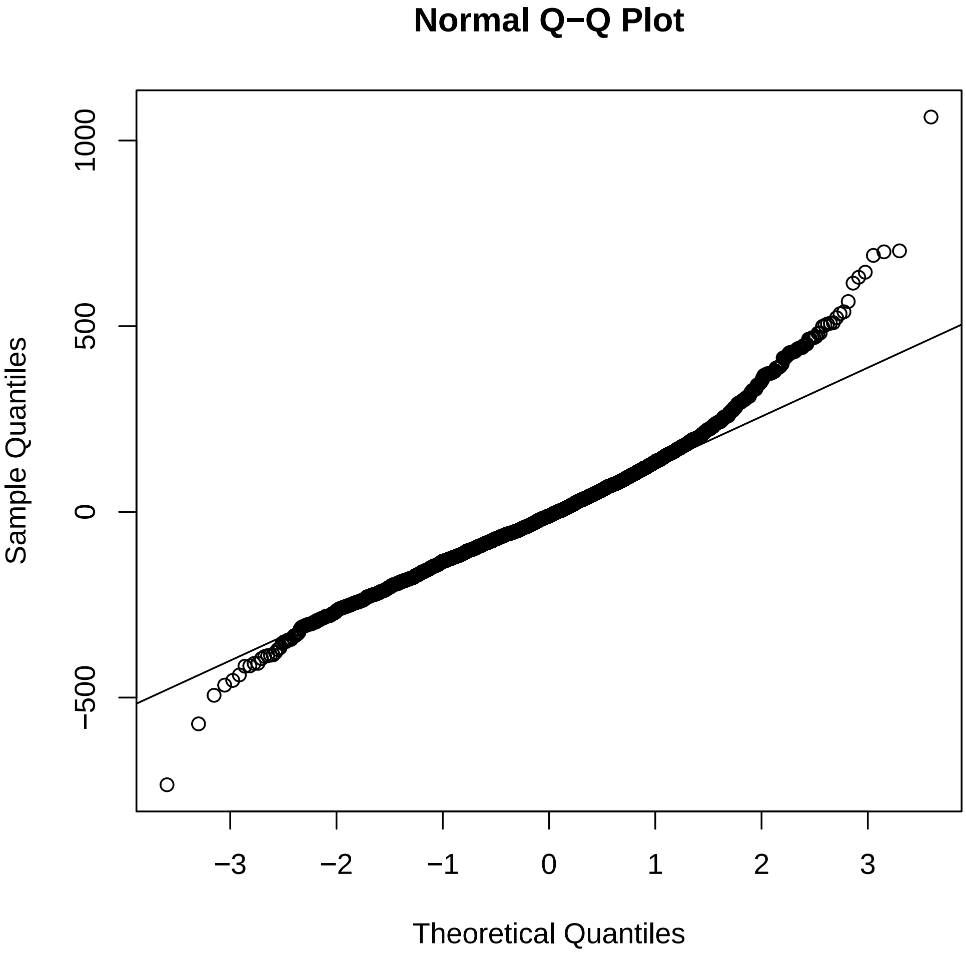
```
> p.residuals.yearc=residuals+beta[3]*yearc
> plot(yearc,p.residuals.yearc)
```



The above partial residual plot of yearc shows that there are blank spaces approximately in the intervals 1920-1930 and 1940-1950 which shows that the number of houses constructed during this period was much smaller.

- QQ Plot for residuals

```
> qqnorm(residuals)
> qqline(residuals)
```



The fat tails on both the ends clearly indicates deviation from our normality assumption.

(c) Now fitting the model

$$log(rent) = \beta_0 + \beta_1 area + \beta_2 yearc + \epsilon, \epsilon \sim N(0, \sigma^2)$$

```
> lrent<-log(rent)
> fit<-lm(lrent~area+yearc)
> summary(fit)
```

```
Call:
lm(formula = lrent ~ area + yearc)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5900 -0.1890  0.0236  0.2213  0.9213 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.6385556  0.5446836  -12.19   <2e-16 ***
area         0.0111738  0.0002598   43.01   <2e-16 ***
yearc        0.0060972  0.0002762   22.07   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.3328 on 3079 degrees of freedom
Multiple R-squared: 0.3945,	Adjusted R-squared: 0.3941 
F-statistic:  1003 on 2 and 3079 DF,  p-value: < 2.2e-16
```

The summary of the model fitted shows the estimated values of the parameters and the various quantiles for residuals.

(d) The mean of the residuals vector is calculated as

```
> residuals=fit$resid
> mean(residuals)
```

```
[1] 1.342919e-17
```

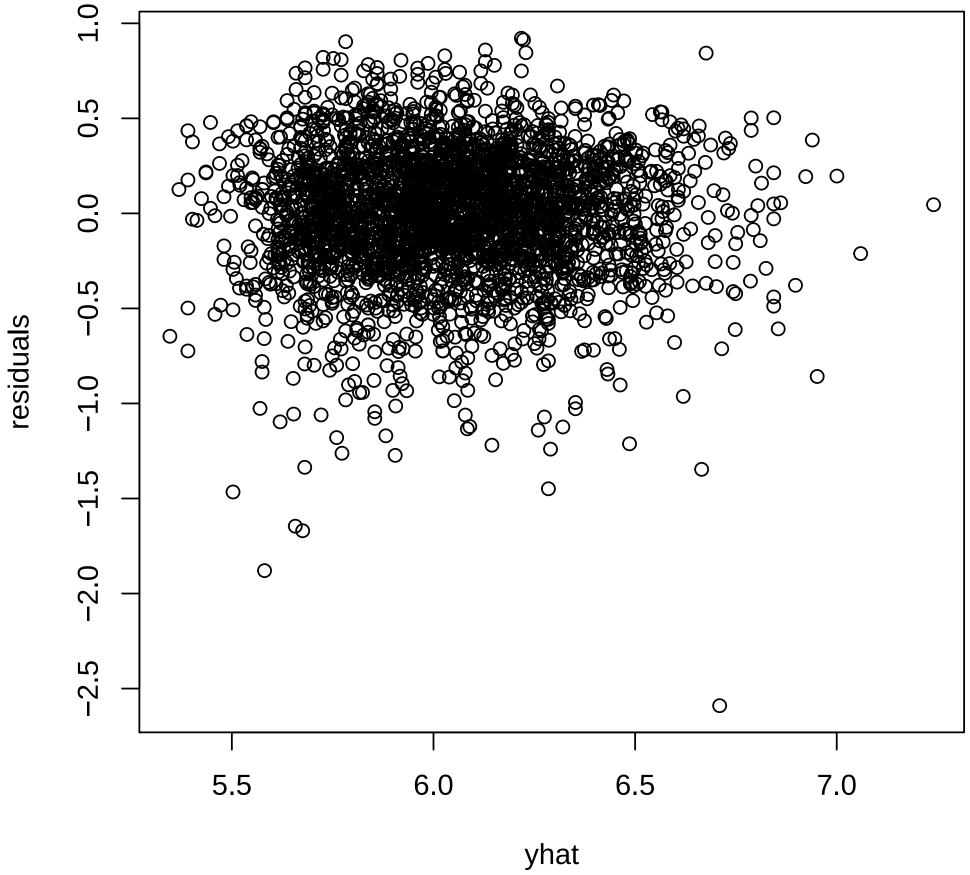
which is very close to zero, thus verifying our assumption of $E(\epsilon) = 0$. The column rank of the design matrix X can be calculated as:

```
[1] 3
```

which confirms our assumption of full column rank design matrix. The relevant plots after fitting the model and their interpretations are given as follows:

- Plot of model residuals vs. fitted values:

```
> yhat<-fit$fitted.values
> plot(yhat,residuals)
```

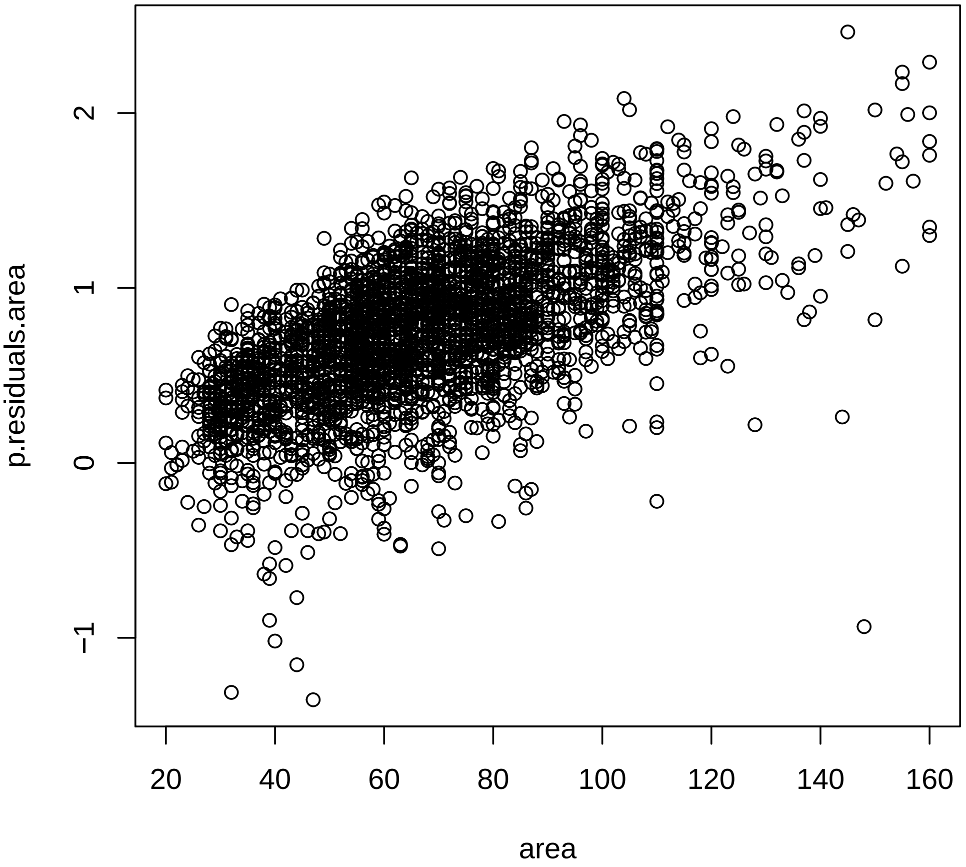


The above residual plot is much better than the previous one. However still there are some points that go as down as -2. Further the density of points towards the right is also smaller. But rest of the points are fairly uniformly distributed and thus we can say that our assumption of homoscedascity is approximately satisfied here.

✓

- Partial residual plot for area

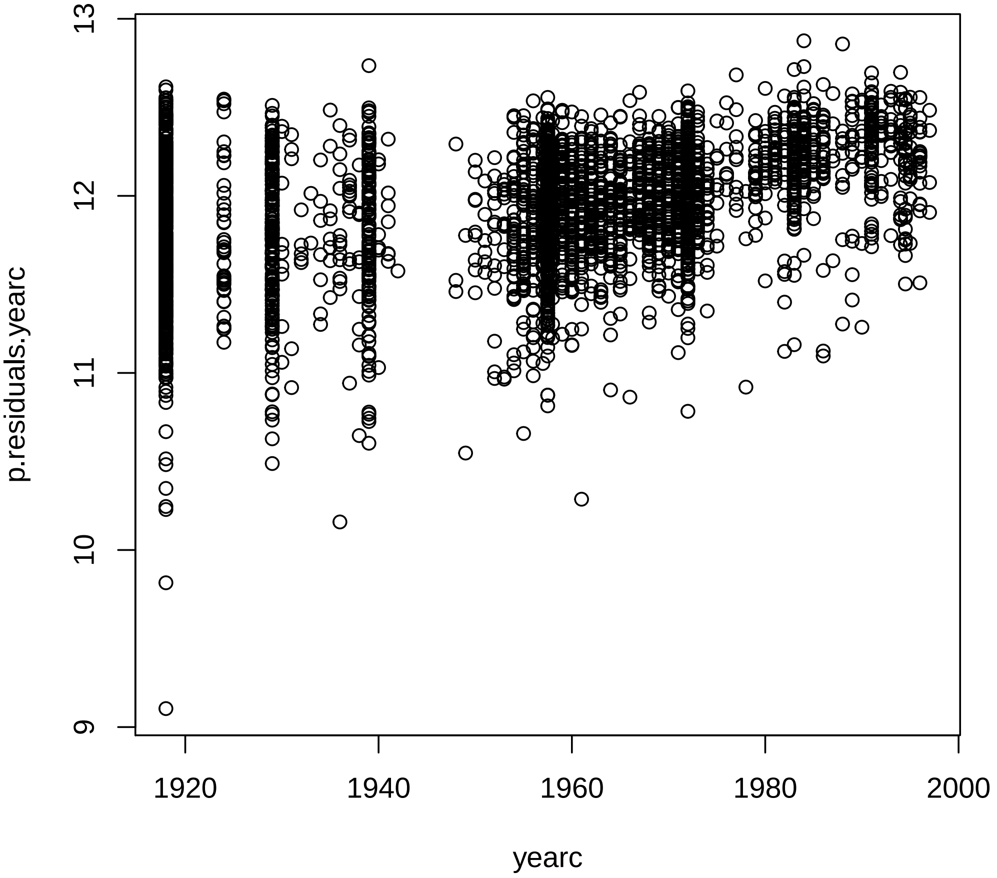
```
> beta<-fit$coef
> p.residuals.area<-residuals+beta[2]*area
> plot(area,p.residuals.area)
```



The above partial residual plot indicates that unlike the previous one, the variance of log of rent does not vary with area. and follows a linear relationship.

- Partial residual plot for yearc

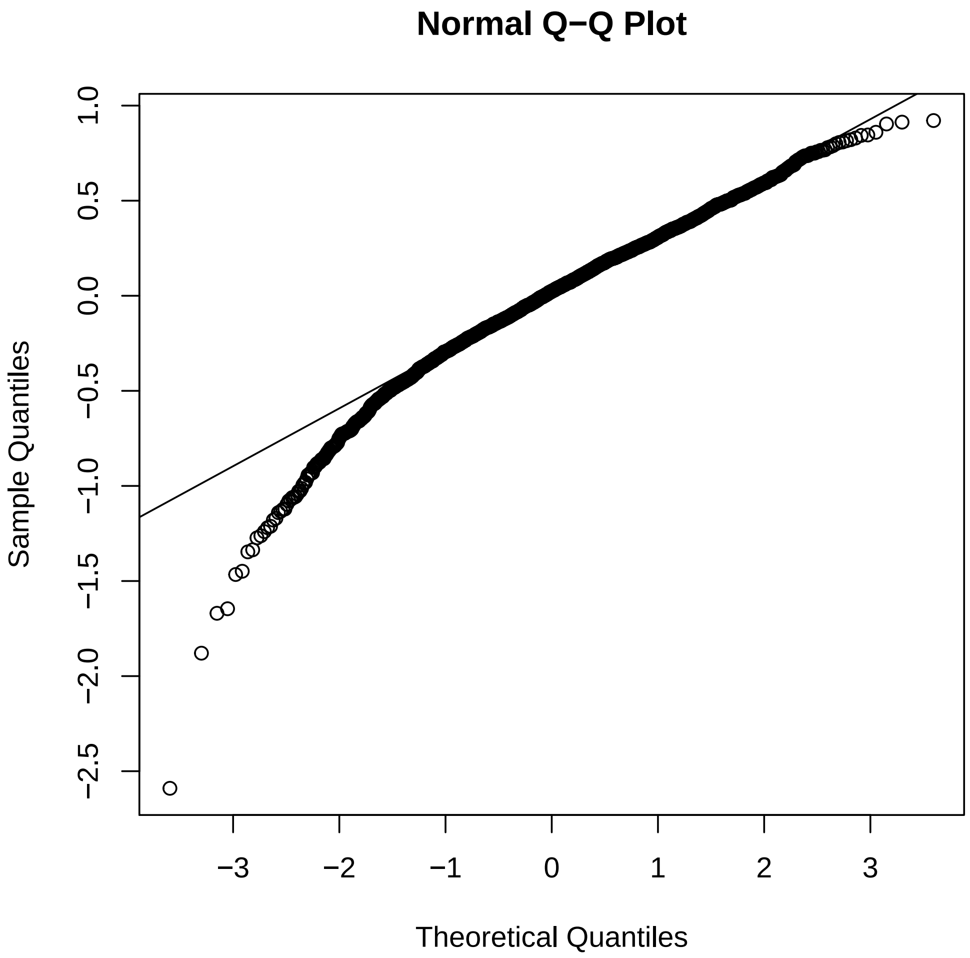
```
> p.residuals.yearc<-residuals+beta[3]*yearc
> plot(yearc,p.residuals.yearc)
```



The above residual plot again clearly shows missing data points in the intervals 1920-1925 and 1940-1950 which indicates that there were not many apartments that were constructed during these periods. What About Linearity?

- QQ Plot for residuals

```
> qqnorm(residuals)
> qqline(residuals)
```



The fat tail on the left end and less prominent short tail on the right end clearly indicates significant deviation from our normality assumption.

(e) Now fitting the model

$$rentsqm = \beta_0 + \beta_1 area + \beta_2 yearc + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

```
> fit<-lm(rentsqm~area+yearc)
> #par (mfrow=c(2,2))
> #plot(fit,which=1:4)
> summary(fit)

Call:
lm(formula = rentsqm ~ area + yearc)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5261 -1.5604 -0.1481  1.4134  9.0863 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -61.058133   3.527255  -17.31   <2e-16 ***
area         -0.027231   0.001682  -16.19   <2e-16 ***
yearc         0.035784   0.001789   20.01   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 2.155 on 3079 degrees of freedom
Multiple R-squared: 0.2178,	Adjusted R-squared: 0.2173 
F-statistic: 428.6 on 2 and 3079 DF,  p-value: < 2.2e-16
```

The summary of the model fitted shows the estimated values of the parameters and the various quantiles for residuals.

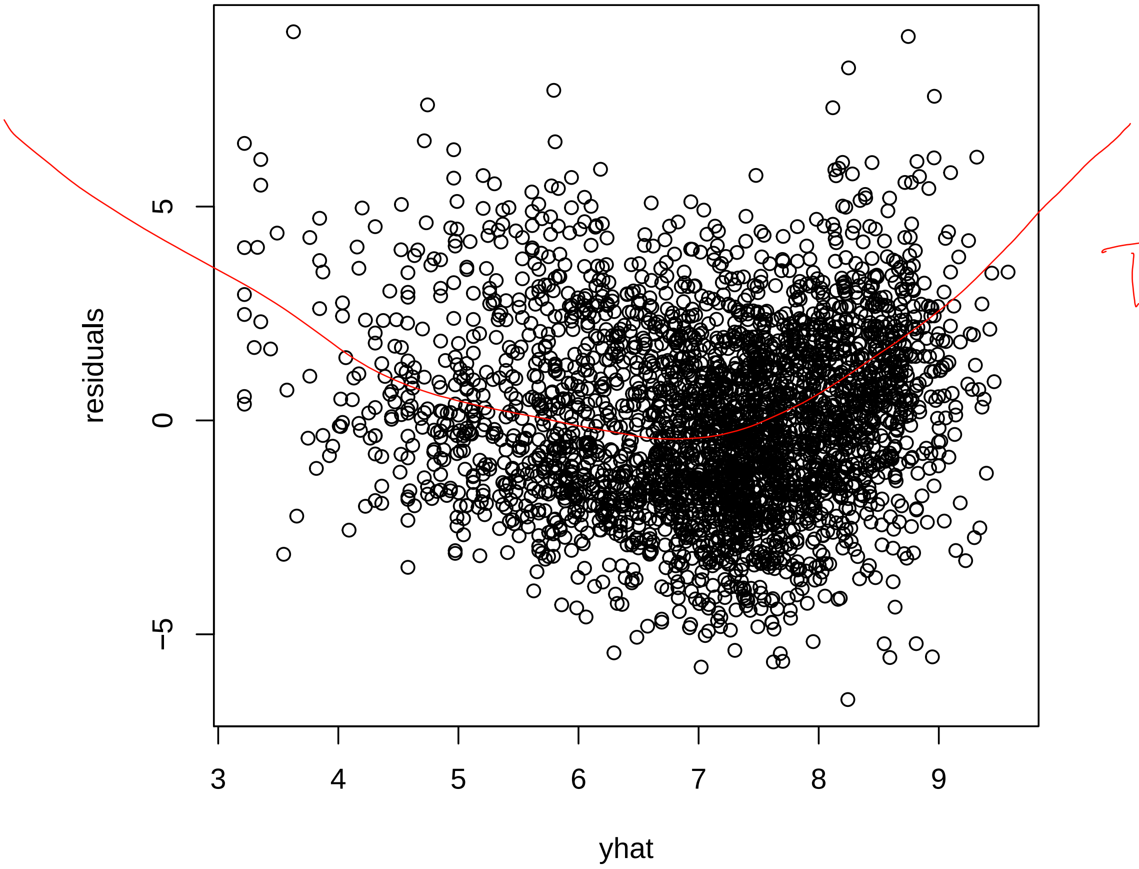
The mean of the residuals vector is calculated as

```
> residuals<-fit$resid
> mean(residuals)

[1] -1.58401e-16
```

which is very close to zero, thus verifying our assumption of $E(\epsilon) = 0$. The design matrix is same as last part and satisfies the full rank assumption. The relevant plots after fitting the model and their interpretations are given as follows:

- Plot of model residuals vs. fitted values:

```
> yhat<-fit$fitted.values
> plot(yhat,residuals)
```

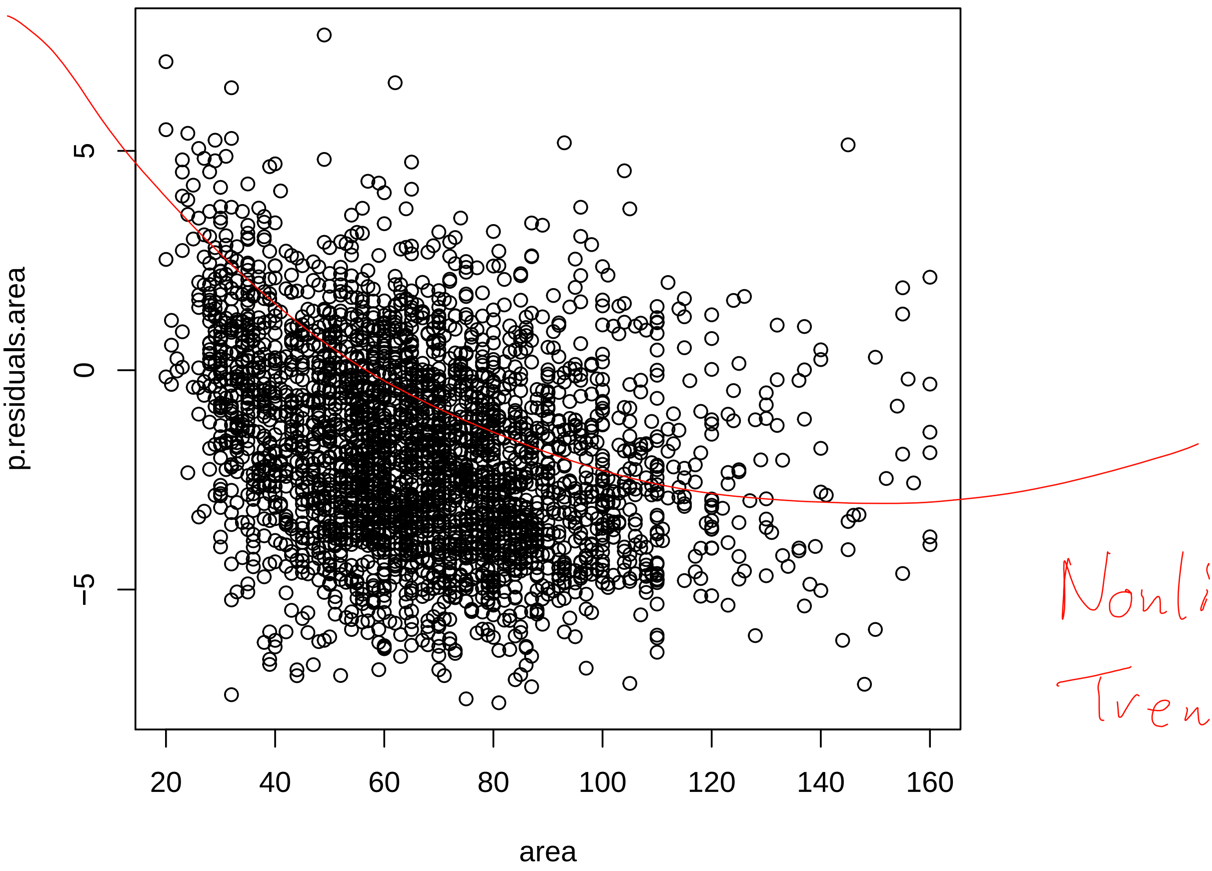


Trend !

The above residual plot shows residuals are much more uniformly distributed compared to the last two models. Thus our assumption of homoscdascity is approximately satisfied here.

- Partial residual plot for area

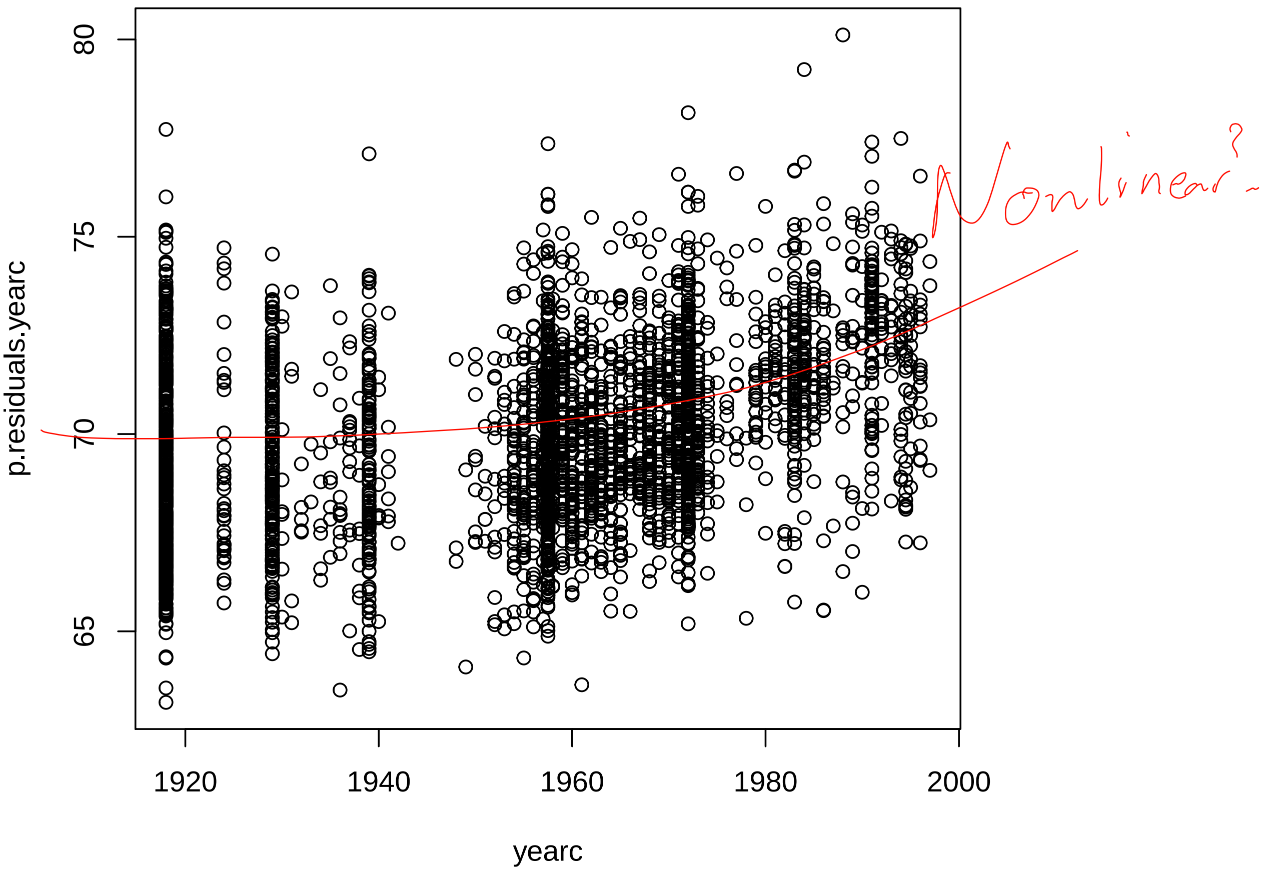
```
> beta<-fit$coef
> p.residuals.area<-residuals+beta[2]*area
> plot(area,p.residuals.area)
```



Nonlinear Trend!

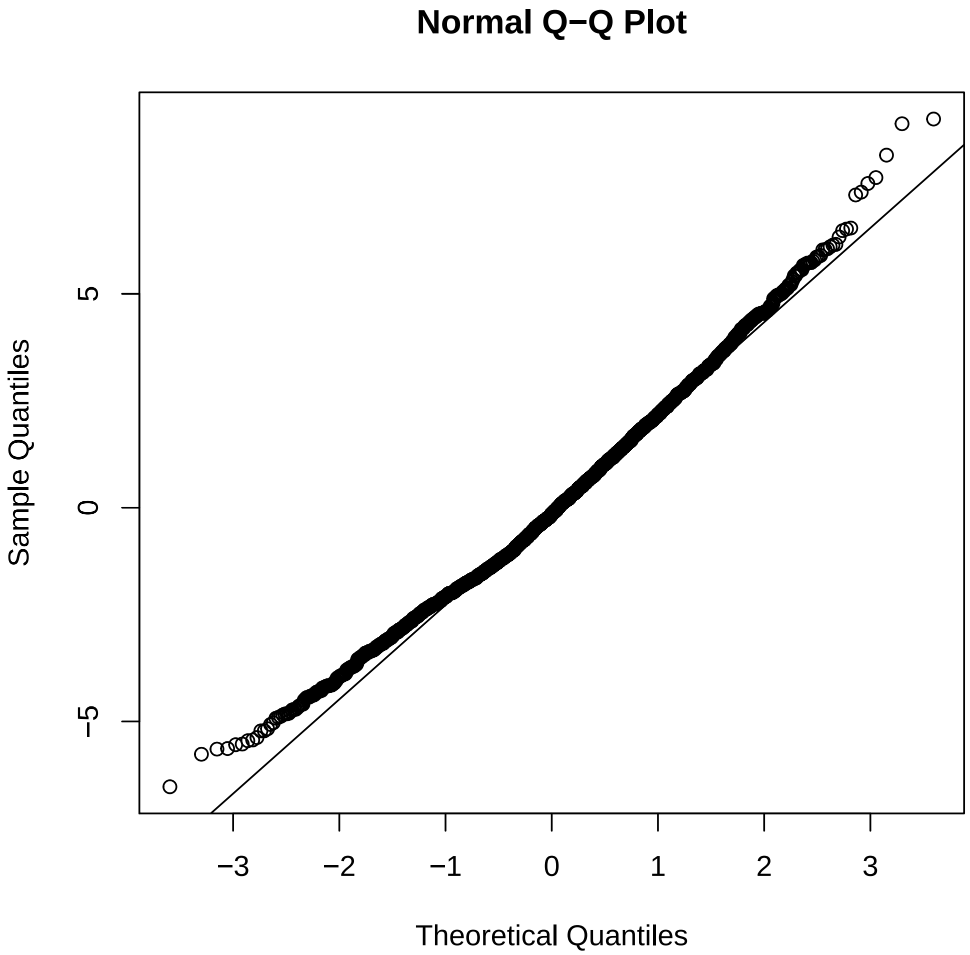
- Partial residual plot for yearc

```
> p.residuals.yearc<-residuals+beta[3]*yearc
> plot(yearc,p.residuals.yearc)
```



- QQ Plot for residuals

```
> qqnorm(residuals)
> qqline(residuals)
```



The short tail towards the left end and a fat tail towards the right end indicates a very slight deviation from our normality assumption. This is much better than our earlier models.

(f) Choosing $f(area) = \beta_1 \dfrac{1}{area}$ and $g(yearc) = \beta_2 yearc + \beta_2 yearc^2 + \beta_3 yearc^3$, we obtain the model as

$$rentsqm = \beta_0 + \beta_1 \frac{1}{area} + \beta_2 yearc + \beta_3 yearc^2 + \beta_4 yearc^3 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Fitting this model we get:

```
> fit3.4<-lm(rentsqm~I(1/area)+poly(yearc,3))
> #par (mfrow=c(2,2))
> #plot(fit,which=1:4)
> summary(fit3.4)

Call:
lm(formula = rentsqm ~ I(1/area) + poly(yearc, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-6.9920 -1.3705 -0.1354  1.3711  8.2834

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.9121     0.1009  48.692   <2e-16 ***
I(1/area)       129.5717     5.5358  23.406   <2e-16 ***
poly(yearc, 3)1  43.9384     2.0726  21.200   <2e-16 ***
poly(yearc, 3)2  27.5389     2.0596  13.371   <2e-16 ***
poly(yearc, 3)3  -1.7558     2.0400  -0.861    0.389
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 2.039 on 3077 degrees of freedom
Multiple R-squared:   0.3,        Adjusted R-squared: 0.2991
F-statistic: 329.7 on 4 and 3077 DF,  p-value: < 2.2e-16
```

The summary of the model fitted shows the estimated values of the parameters and the various quantiles for residuals.

The mean of the residuals vector is calculated as

```
> residuals3.4<-fit3.4$resid
> mean(residuals3.4)

[1] 1.90071e-17
```

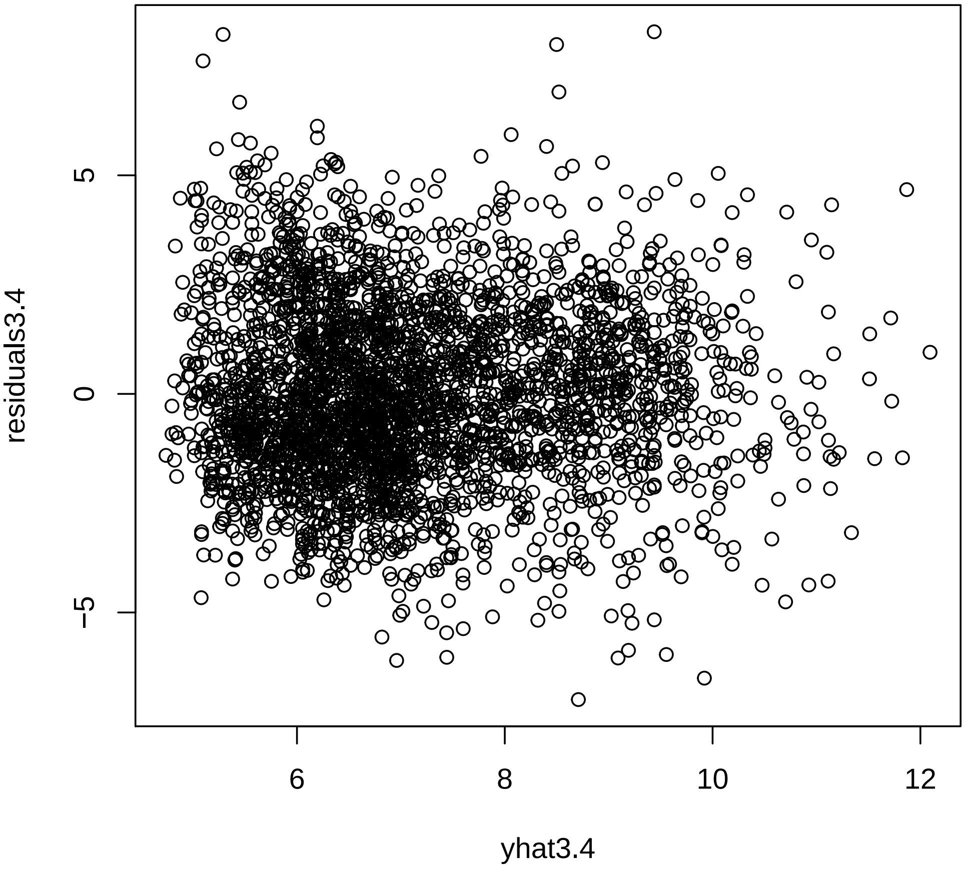
which is very close to zero, thus verifying our assumption of $E(\epsilon) = 0$. The column rank of the design matrix X can be calculated as:

```
[1] 5
```

which confirms our assumption of full column rank design matrix. The relevant plots after fitting the model and their interpretations are given as follows:

- Plot of model residuals vs. fitted values:

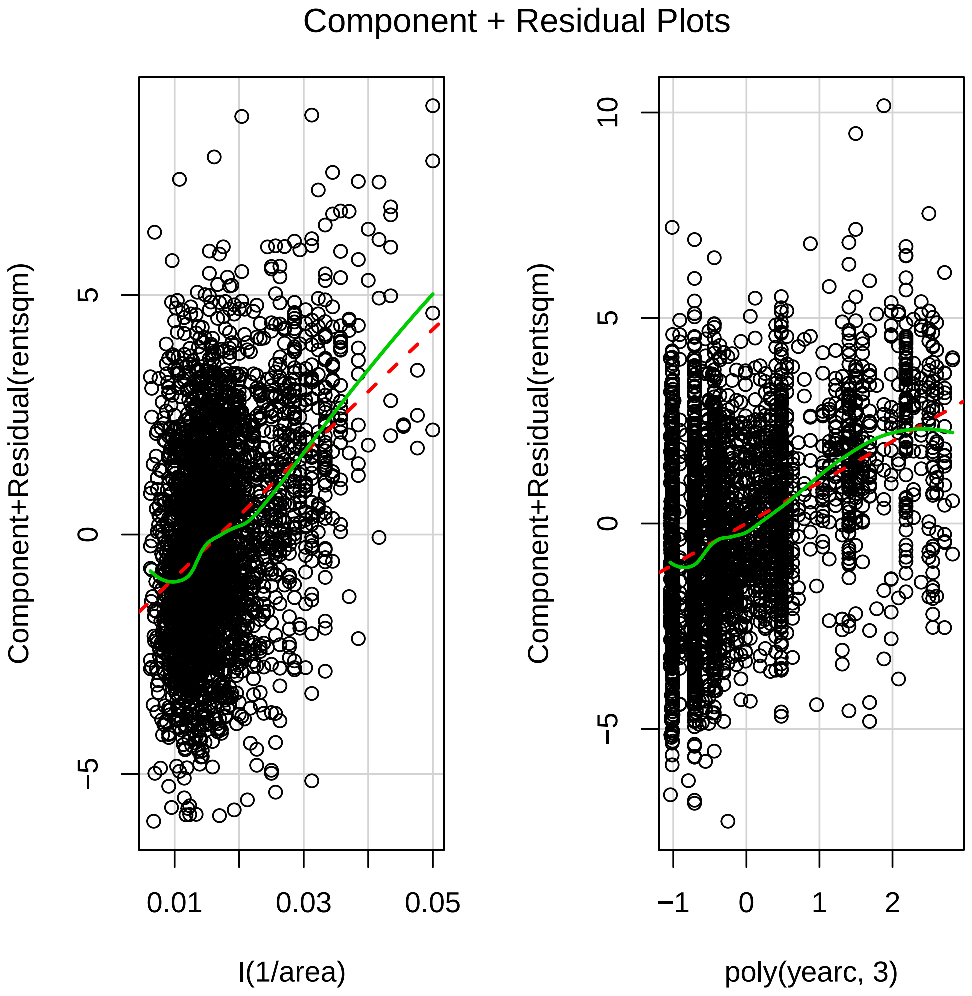
```
> yhat3.4<-fit3.4$fitted.values
> plot(yhat3.4,residuals3.4)
```



The above residual plot shows residuals are much more uniformly distributed compared to the last three models. Thus our assumption of homoscedascity is satisfied here to a great extent.

- Partial residual plots for area and yearc

```
> library(car)
> crPlots(fit3.4)
```



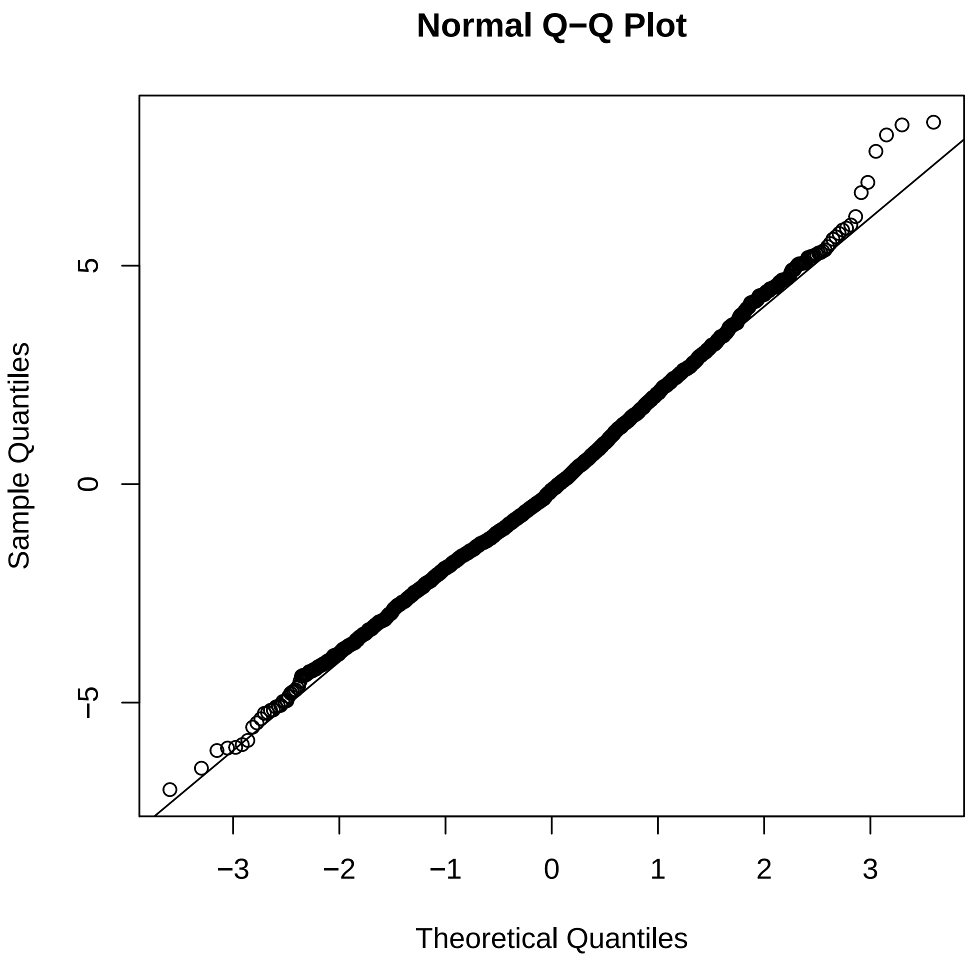
Plots of f(area), g(yearc)!

From the partial residual plot of rentsqm vs. area, it is clear that as area is increased the variability in rent per square meter increases. For a smaller area apartment, the number of options in terms of rent per square meter are densely concentrated in smaller intervals while for a larger area apartment they span a much larger interval but are sparsely distributed. Further, the partial residual plot of rentsqm vs. yearc again shows an increasing trend of rent per square meter as time of construction (in years) passes. Notably there are a few spots in the middle where the number of options in terms of rent per square meter are much lesser, indicating a decrease in the construction of number of houses during that particular period.

I Disagree

- QQ Plot for residuals

```
> qqnorm(residuals3.4)
> qqline(residuals3.4)
```



The small fat tail towards the right end indicates a very slight deviation from our normality assumption.

(g) Multiplying the finalized model in part (f) by area and taking the expectation, we get,

$rentsqm \times area = \beta_0 \times area + \beta_1 \frac{1}{area} \times area + \beta_2 yearc \times area + \beta_3 yearc^2 \times area + \beta_4 yearc^3 \times area + \epsilon \times area, \quad \epsilon \sim N(0, \sigma^2)$

which simplifies as

$rent = \beta_0 \times area + \beta_1 + \beta_2 yearc \times area + \beta_3 yearc^2 \times area + \beta_4 yearc^3 \times area + \epsilon \times area, \quad \epsilon \sim N(0, \sigma^2)$

The estimated expected rent thus becomes

$\mathrm{E}[\widehat{rent}] = \hat{\beta}_0 \times area + \hat{\beta}_1 + \hat{\beta}_2 yearc \times area + \hat{\beta}_3 yearc^2 \times area + \hat{\beta}_4 yearc^3 \times area + E[\hat{\epsilon}] \times area, \quad \epsilon \sim N(0, \sigma^2)$

Now $E[\hat{\epsilon}] = 0$. Thus we obtain

$\widehat{E[rent]} = \hat{\beta}_0 \times area + \hat{\beta}_1 + \hat{\beta}_2 yearc \times area + \hat{\beta}_3 yearc^2 \times area + \hat{\beta}_4 yearc^3 \times area$

The interaction terms of the area and yearc covariates clearly indicates that the rate of increase of rent with an unit increase in area is higher for a later time period as compared to an earlier time period. The estimated parameter values are already shown in the summary before.

The variance of rent can be calculated as

$$
\begin{aligned}
Var[\widehat{rent}] &= Var[\hat{\epsilon}] \times area^2, \quad \epsilon \sim N(0, \sigma^2) \\
&= \widehat{\sigma^2} \times area^2
\end{aligned}
$$

✓

It is clear from the above expression that there is no effect of yearc on the variance of rent. However the variance of rent is proprtional to the square of area.

(h) Following models were tried before finalizing this particular model.

What About Models From b, c, ...

- $$rentsqm = \beta_0 + \beta_1 \frac{1}{area} + \beta_2 yearc + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$
- $$rentsqm = \beta_0 + \beta_1 \frac{1}{area^2} + \beta_2 yearc^2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$
- $$rentsqm = \beta_0 + \beta_1 log(area) + \beta_2 yearc + \beta_3 yearc^2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

There were one or more of the following major problems with the above models:

- Our assumption of homoscedascity was not satisfied (visually inferred from the residual plots)
- Our assumption of normal errors was not satisfied (visually inferred from QQ plot)
- The partial residual plots were bizzare.

If I were to choose one of these I would choose the first one since it most closely satisfies our assumptions of homoscedascity and normality under the classical linear model.