Homework 4 - STAT 511

Amal Agarwal

Answer 3

(a) It is given that $MSD(t) = E[X^2(t)]$. From the given data, it can be observed that the position of different particles have been given for 4 time points viz. 5, 10, 15, 20. Thus for each time point, we can estimate the MSD (mean of the squared distances) by approximating it with the sample mean of the squared distances at each of the time points.

```
[1]  574.3304 1272.7540 1755.0098 2327.1401
```

(b) Now MSD scales according to the following power law:

$$MSD(t) = \gamma t^{\alpha}$$

Taking log of the above equation, we get

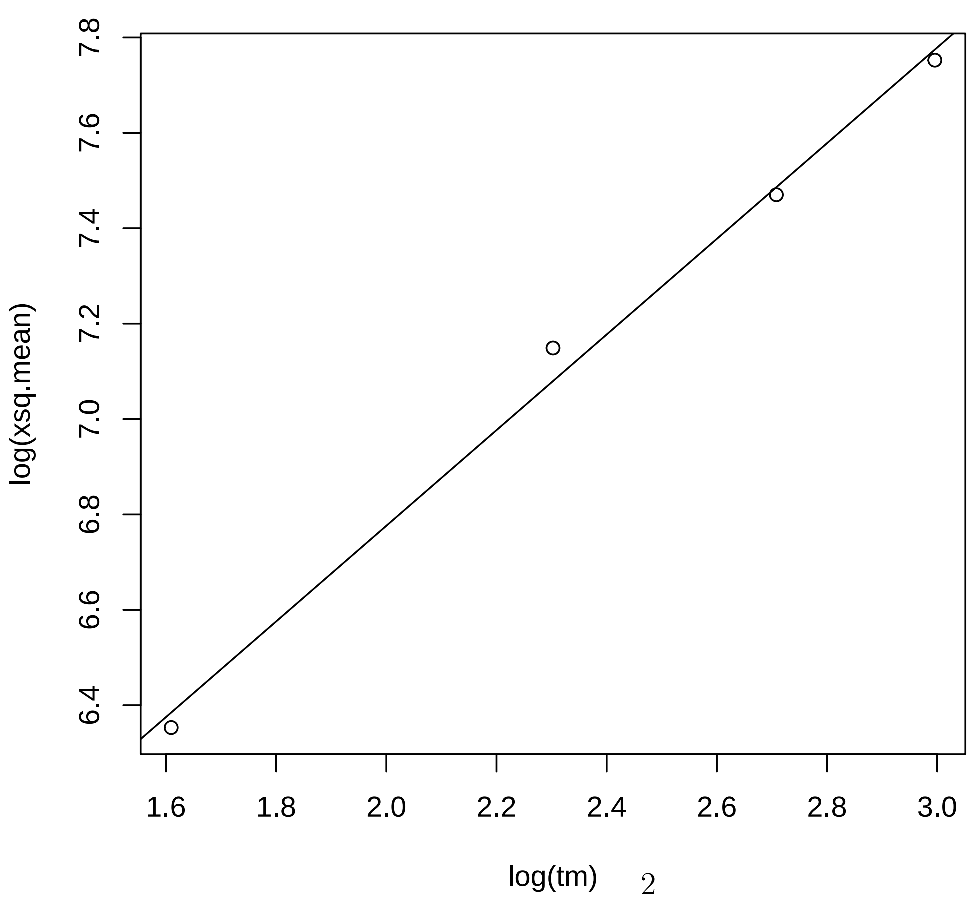
$$log(MSD(t)) = log(\gamma) + \alpha log(t)$$

Now since we know the estimated MSD at each of the time points from the above data, we can fit the following model

$$log(\widehat{MSD}_i(t_i)) = \beta_0 + \alpha log(t_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

The summary of the model fit and the plot of $log(\widehat{MSD})$ vs. $log(t)$ with the fitted line are given as follows

```
Call:
lm(formula = I(log(xsq.mean)) ~ I(log(tm)))

Residuals:
       1        2        3        4 
-0.03140  0.06938 -0.01586 -0.02212 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.77096    0.13508   35.32 0.000801 ***
I(log(tm))   1.00261    0.05492   18.26 0.002987 ** 
---
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: 0.05718 on 2 degrees of freedom
Multiple R-squared: 0.994,	Adjusted R-squared: 0.9911 
F-statistic: 333.3 on 1 and 2 DF,  p-value: 0.002987

(Intercept)  I(log(tm)) 
   4.770961    1.002612 
```



(c) Clearly the estimated value of α is

```
I(log(tm))
  1.002612
```

It is not needed to estimate γ here. However, even if it was needed we could not have done it since $\widehat{log(\gamma)} \neq log(\hat{\gamma})$. This can be verified by Jensen's inequality which gives $E(\widehat{log(\gamma)}) \geq log(E(\hat{\gamma}))$.

(d) The assumptions of the model can verified as follows:

- The mean of the residuals can be calculated as

  ```
  [1] -1.734723e-18
  ```

  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:

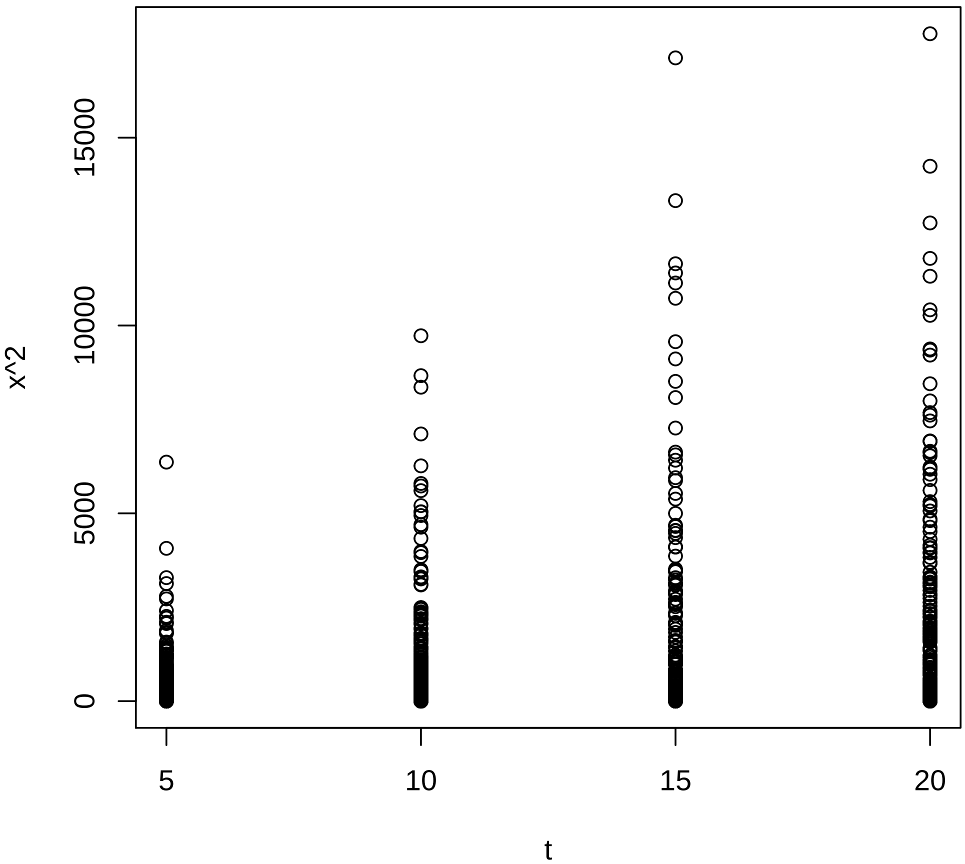
  ```
  [1] 2
  ```

  which confirms our assumption of full column rank design matrix.

- To check for homoscedasticity, the plot of residuals against the predictor doesn't help much here since we only have 4 points. Similarily the QQ plot of residuals which is the diagnostic for normality also fails.

- Plot of x^2 vs. t

```
> plot(t,x^2)
```



The above plot shows a heteroscedastic behaviour of x^2 with t. Further tha variance increases as t is increased and thus the errors are correlated.

- We can't comment much on normality assumption in our model but still we can say that individually the mean of x^2 at different time points must follow a normal distribution (since $n = 200 >> 30$ for each t_i) by central limit theorem.

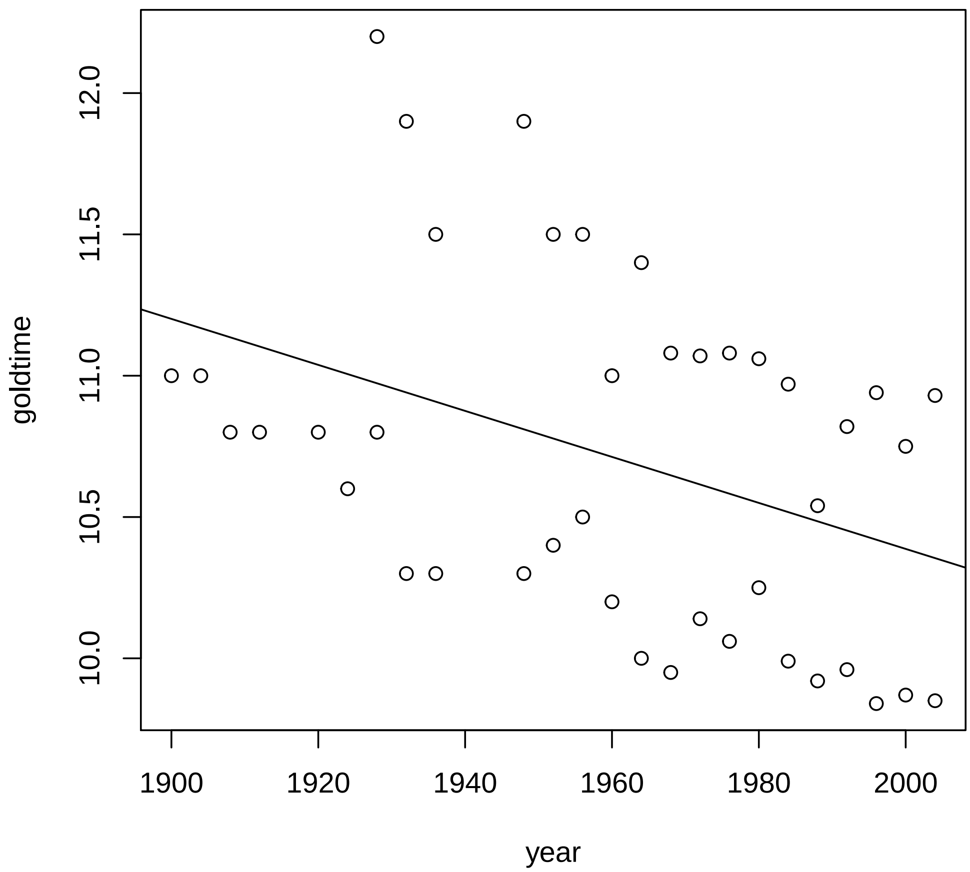
Answer 4

(a) The following model was fitted to the given data:

$$goldtime_i = \beta_0 + \beta_1 year_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

```
'data.frame':        42 obs. of  3 variables:
 $ year    : int  1900 1904 1908 1912 1920 1924 1928 1932 1936 1948 ...
 $ goldtime: num  11 11 10.8 10.8 10.8 10.6 10.8 10.3 10.3 10.3 ...
 $ gender  : Factor w/ 2 levels "M","W": 1 1 1 1 1 1 1 1 1 1 ...

Call:
lm(formula = oly$goldtime ~ oly$year)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6976 -0.5089 -0.2196  0.4869  1.2269

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.663985   5.865382   4.546 4.97e-05 ***
oly$year    -0.008138   0.002991  -2.721   0.0096 **
---
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: 0.5679 on 40 degrees of freedom
Multiple R-squared: 0.1561,       Adjusted R-squared: 0.135
F-statistic: 7.401 on 1 and 40 DF,  p-value: 0.0096

[1] 0.5678969
```

```
> plot(year,goldtime)
> abline(fit)
```



Clearly there is a set of points above fitted line and there is another set below it which shows that we are missing the effect of a categorical covariate.
Diagnostics:
The mean of the residuals can be calculated as

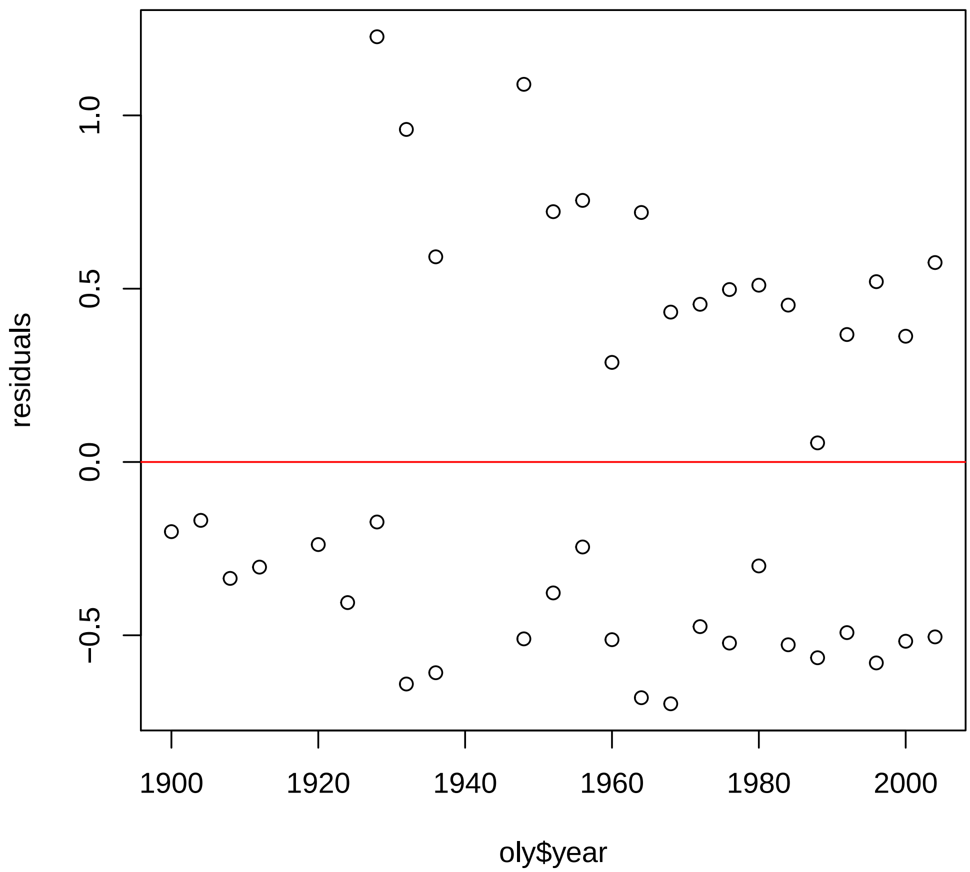
```
[1] 3.003498e-18
```

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] 2
```

which confirms our assumption of full column rank design matrix.

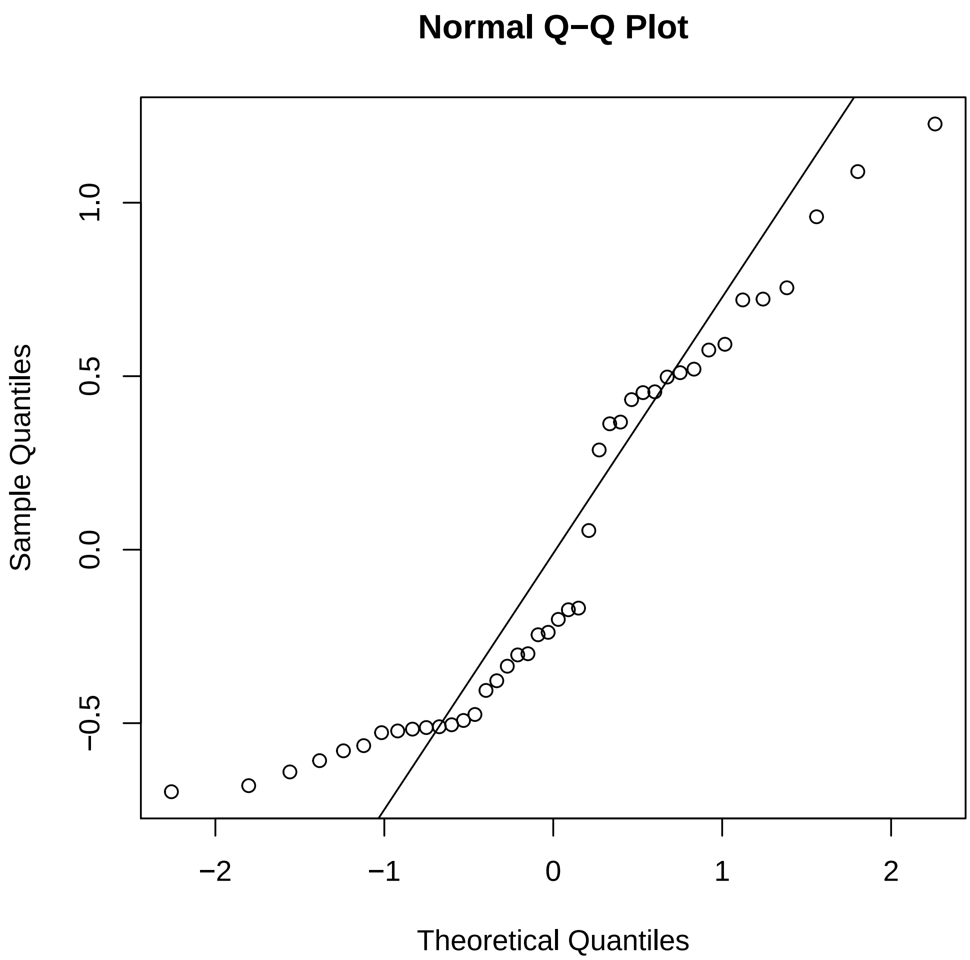
The residual plot of the corresponding fit is given as follows:



This residuals are not uniformly distributed around zero which shows that our assumption of homoscedascity is not satisfied here.

To check for our normality assumption, QQ plot of the residuals is given as follows:

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



Although the short tails on both the left and right sides don't matter much, still this is not very close to normality. Interpretation of the effect of year on gold medal time in 100 m race: Estimated value of the β_1, i.e. $\hat{\beta}_1$ shows the mean decrease of $-0.008s$ in goldtime per unit increase in year. Note that the effect of gender has not been taken into account in this model.

(b) Considering the transformations and interactions, the final model that I think is most appropriate for the given data is:

$$goldtime_i = \beta_0 + \beta_1\left(\frac{1}{year_i}\right) + \beta_2(gender_i) + \beta_3(year_i * gender_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

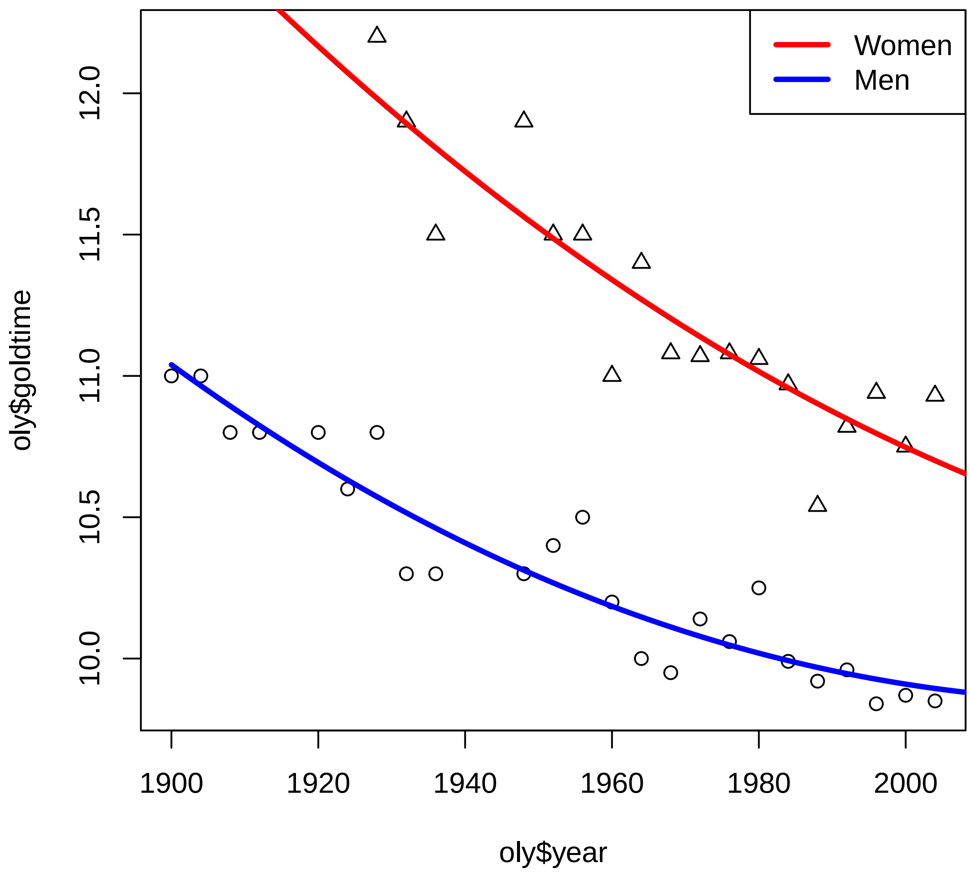
The summary of the fit including the estimated parameters are given as follows:

```
Call:
lm(formula = goldtime ~ I(1/year) + gender + (year:gender), data = oly)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.36163 -0.06361  0.00333  0.08119  0.33698 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -5.295e+02  2.600e+02  -2.036 0.048931 *  
I(1/year)     5.476e+05  2.537e+05   2.159 0.037430 *  
genderW       1.671e+01  4.350e+00   3.841 0.000464 ***
genderM:year  1.328e-01  6.663e-02   1.993 0.053646 .  
genderW:year  1.249e-01  6.566e-02   1.902 0.065007 .  
---
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: 0.1631 on 37 degrees of freedom
Multiple R-squared: 0.9356,	Adjusted R-squared: 0.9287 
F-statistic: 134.5 on 4 and 37 DF,  p-value: < 2.2e-16
```

The estimated variance is

```
[1] 0.1630664
```

The fitted model is given as:

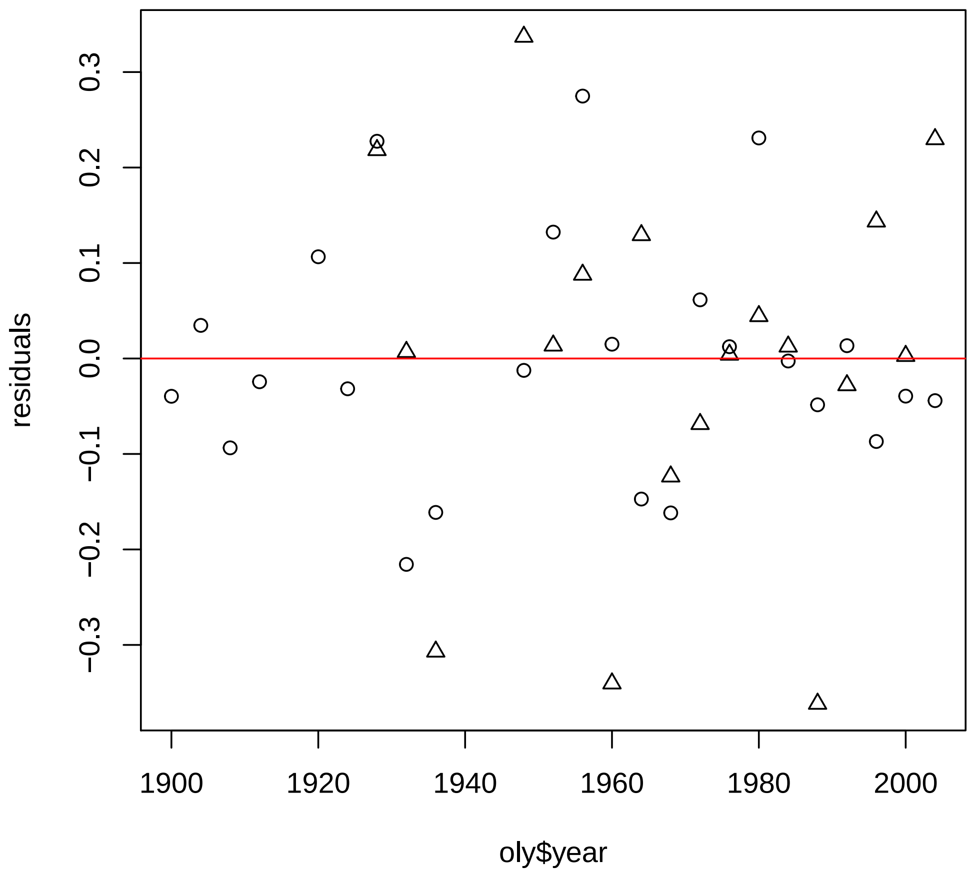


Diagnostics:
The mean of the residuals can be calculated as

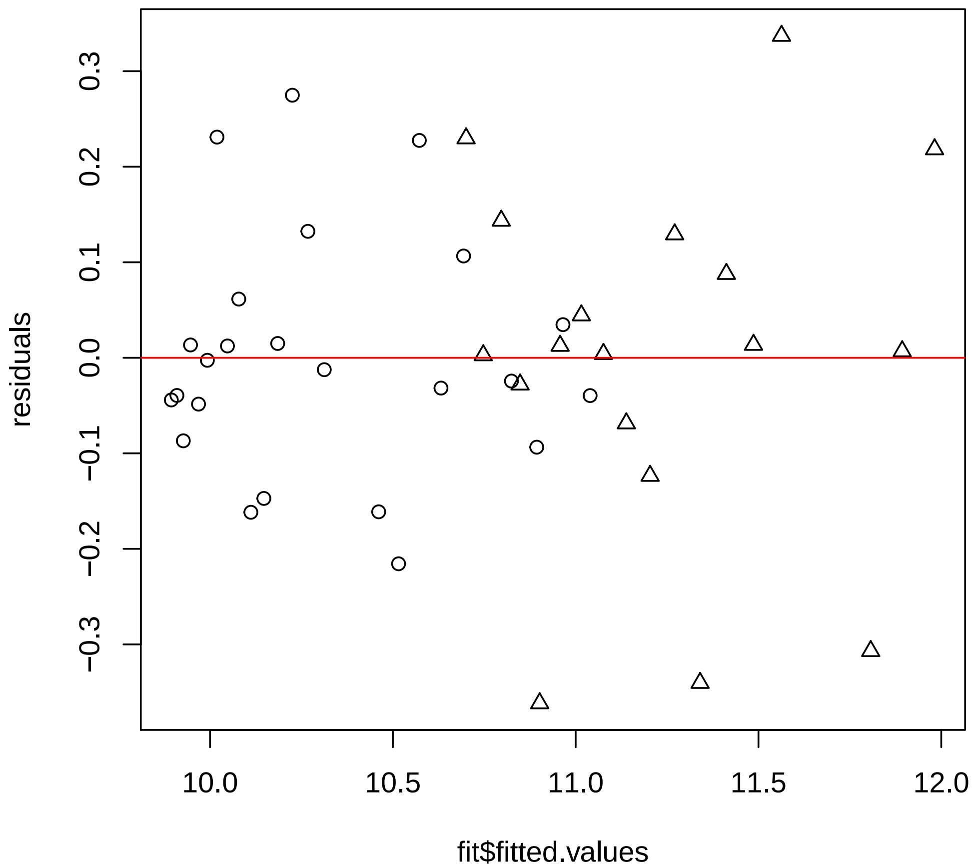
```
[1] -4.252912e-18
```

which is very close to zero thus verifying our assumption of $E(\epsilon) = 0$.

The residual plots of the corresponding fit is given as follows:



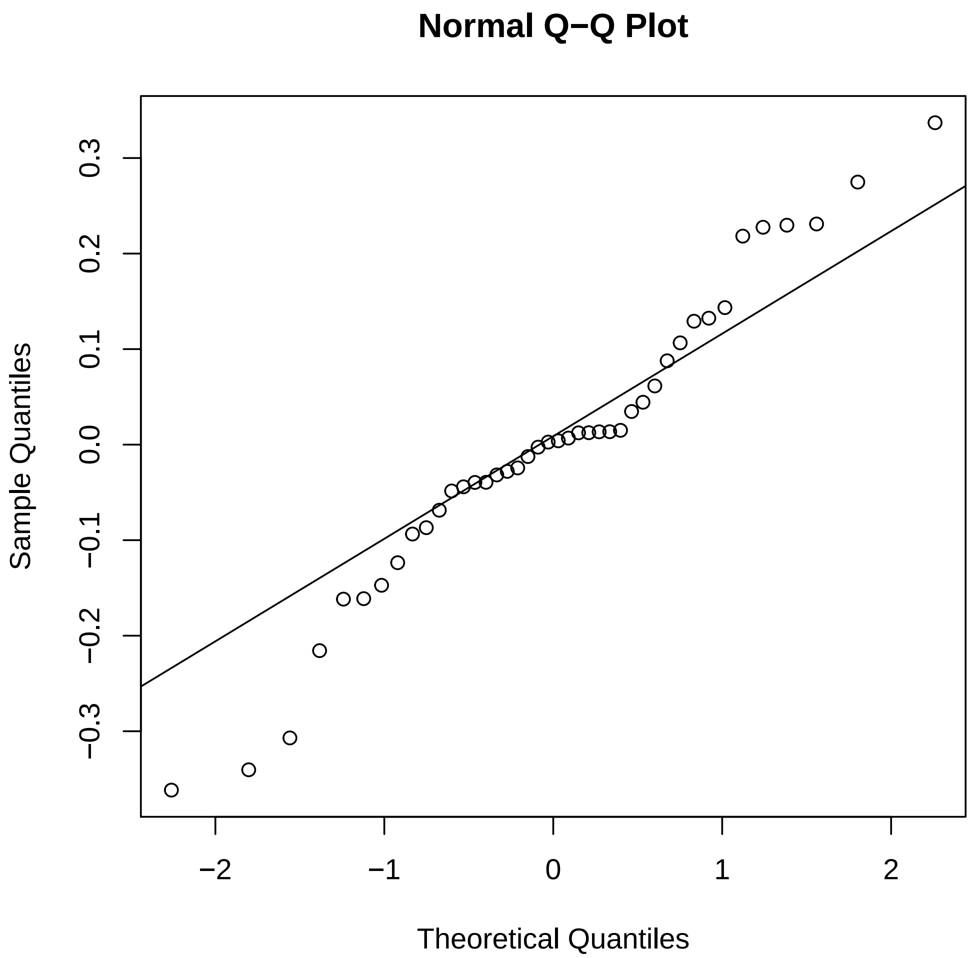
This shows that the residuals are uniformly distributed with respect to year and hence uncorrelated. Thus our assmption of independent errors is satisfied.



This shows that the residuals are uniformly distributed with respect to fitted values and hence our assmption of homoscedascity is satisfied.

To check for our normality assumption, QQ plot of the residuals is given as follows:

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



The fat tails on both the ends clearly shows significant deviation from normality.

Note that this model seemed to be best since both our assumptions of homoscedascity and independent errors are best verified in this case. However our assumption of normality does not hold good, which is fine since it is not that important.

It is not possible to construct partial residual plots and check the assumption of linearity since we have an interaction term.

Interpretation of the effect of predictor variables on gold medal time in 100 m race: Estimated value of the β_1, i.e. $\hat{\beta}_1 = 5.476 \times 10^5$ shows that the mean gold medal time decreases with increase in year for both Men's and Women's races. $\hat{\beta}_2 = 16.71$ shows that the mean gold medal time is higher for Women than Men. Further $\hat{\beta}_3$ is 0.1328 for Men and 0.1249 for Women which shows that the mean gold time decreases at a faster rate with respect to year for women than men.

(c) The predicted gold medal time values for Mens and Womens 100 m races in 1944 are respectively given as 10.32404 s and 11.58253 s. The corresponding R code is given as follows

```
> for (i in 1:length(x.vals)) {
+   if (x.vals[i]==1944) {
+     Women.1944.predicted<-f.vals.0[i]
+     Men.1944.predicted<-f.vals.1[i]
+   }
+ }
```