STAT511 HW3

Ben Straub

September 26, 2015

(3) Patricle Displacement

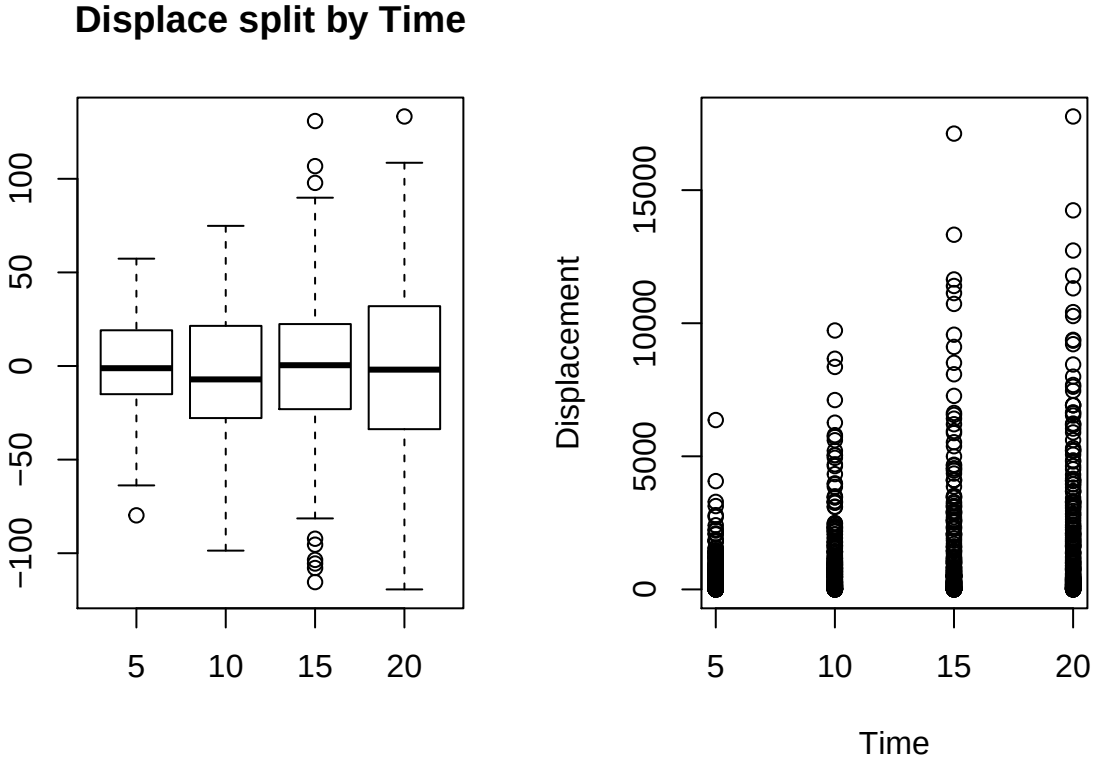
Exploratory Data Analysis of Fluid Data

Summary of Fluid Data

The Fluid Data has 800 observations and 2 variables.

t	x
Min. : 5.00	Min. :-119.332
1st Qu.: 8.75	1st Qu.: -24.615
Median :12.50	Median : -1.999
Mean :12.50	Mean : -1.906
3rd Qu.:16.25	3rd Qu.: 22.219
Max. :20.00	Max. : 133.292

Boxplots and XY-Plot of Fluid Data



Observation:

- As the fluid data increases in time it becomes more spread out.
- The Time Intervals are Discrete: 5, 10, 15, 20.

SUMMARY OF MODEL

ORIGINAL MODEL EQUATION:

$$x^2 = \beta(t)^{\beta} \cdot \epsilon$$

$$\epsilon - lognormal(0, \sigma^2)$$

MODEL TRANSFORMATION:

$$log(x^2) = log(\beta_0) + \beta_1 \cdot (log(t)) + log(\epsilon)$$

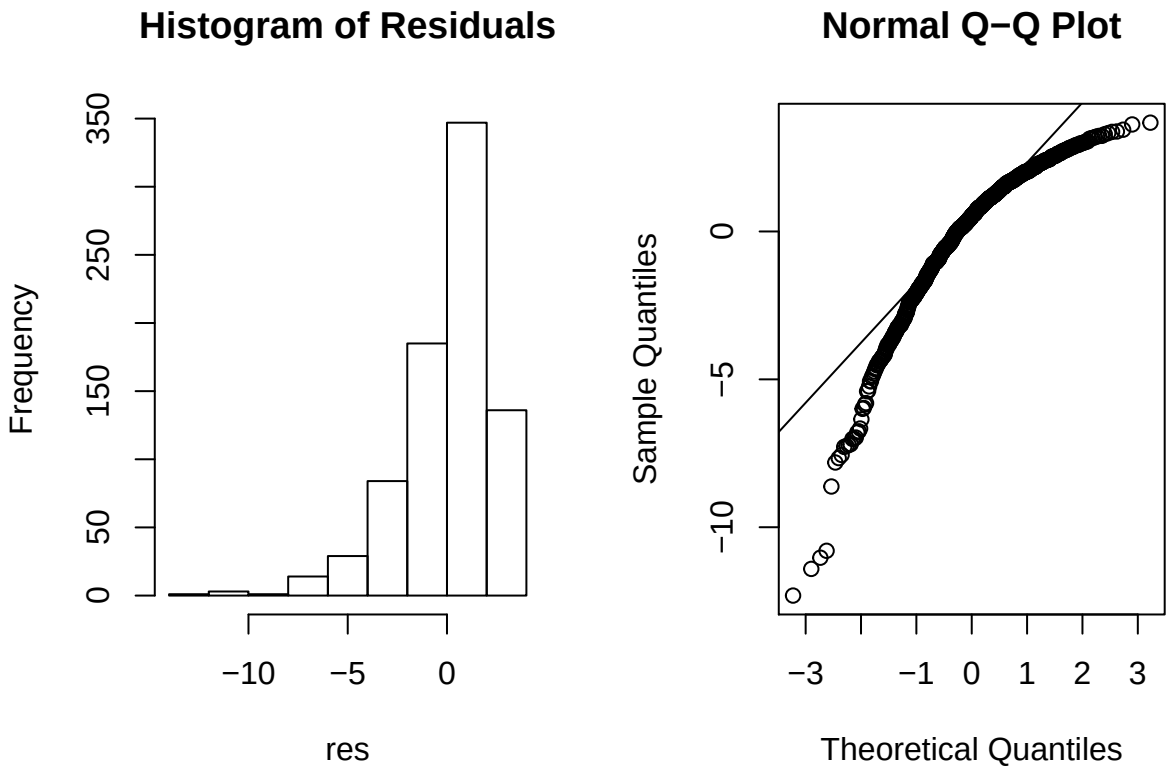
Estimated Coefficients of Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.54	0.39	9.16	0
log(t)	0.96	0.16	6.08	0

INTERPRETATION OF THE PARAMETERS:

- Taking the log of the Researcher's Model turns our model into a linear model with log(lognormal) which become standard normal errors. We can now investigate our assumptions of the Linear Model by checking the Model's Residuals using a histogram, QQ-plot and checking the residuals plotted against the yhats.
- Gamma is our Beta Hat 0 in our Log-Linear Model. The Gamma's value is 3.542
- I have since learned that there is way more to Gamma's calculation... E[Error]*Gamma'value, but I'm not there so no dice.
- Alpha is our Beta Hat One in our Log-Linear Model. The Alpha's value is 0.9555

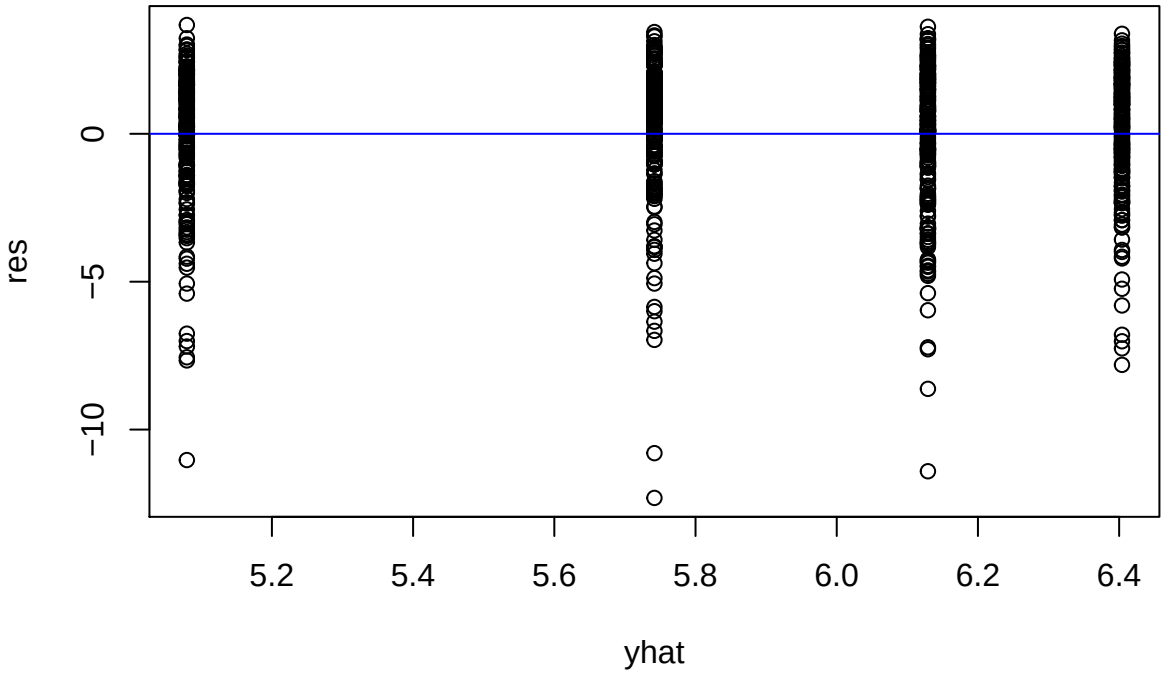
RESIDUAL CHECK OF OUR MODEL



OBSERVATIONS:

- The models residuals in the histogram are not in our sought-after Normal Distribution.
- The model's QQ-plot has thick tailss, which indicates that we can find a better model to fit the data.

Residuals of the Data



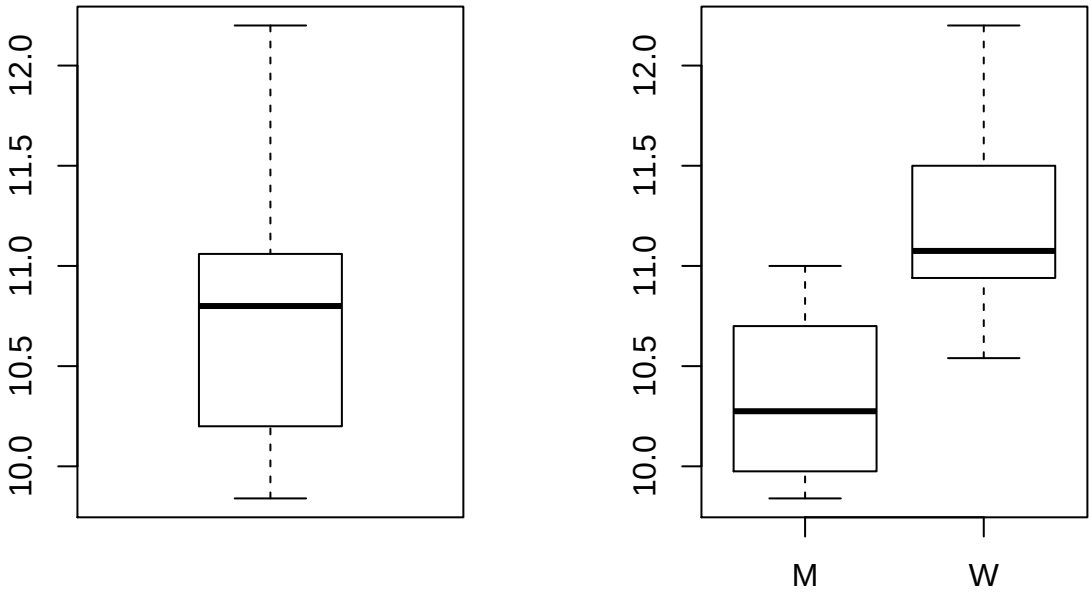
CONCLUSION:

- The models residuals plotted against its yhats shows non-constant error variance!
- We got it all here of a Linear Model that violates our asssumptions of constant error variance . We got heteroscadacity in our residuals , violations in the QQ plot and the Histogram of the residuals has some serious skewness in it. Time to go back to the drawing board.

(4) Olympic 100m Gold Medal Times

Exploratory Data Analysis of Olympics Data

The Olympic Data set has 42 observations with 3 variables: year, goldtime and gender.



OBSERVATIONS:

- Something fishy is going on here!!
- The Women's Data and Men's Data do not match up and the Women's data is skewed towards the first Quartile in the boxplot.

Six Number Summary seperated by gender

Table 3: MEN

year	goldtime	gender
Min. :1900	Min. : 9.840	M:24
1st Qu.:1927	1st Qu.: 9.982	W: 0
Median :1958	Median :10.275	NA
Mean :1954	Mean :10.318	NA
3rd Qu.:1981	3rd Qu.:10.650	NA
Max. :2004	Max. :11.000	NA

Table 4: WOMEN

year	goldtime	gender
Min. :1928	Min. :10.54	M: 0
1st Qu.:1953	1st Qu.:10.95	W:18

year	goldtime	gender
Median :1970	Median :11.07	NA
Mean :1969	Mean :11.23	NA
3rd Qu.:1987	3rd Qu.:11.50	NA
Max. :2004	Max. :12.20	NA

Summary of Model

Model Equation

$$goldtime = \beta + \beta_i \cdot year + \epsilon \; N(0, \sigma^2)$$

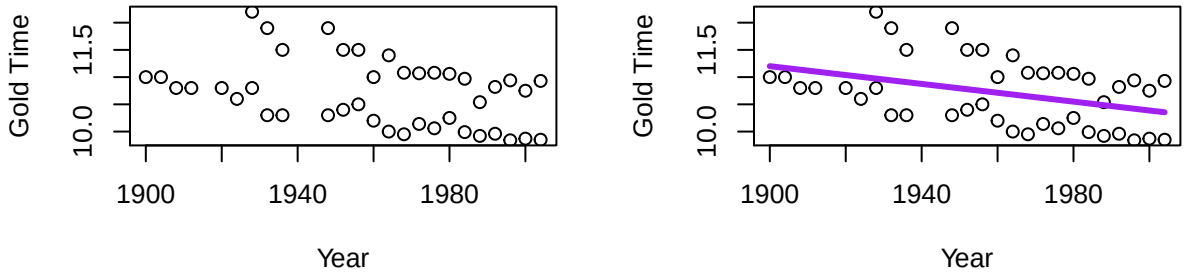
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.66	5.87	4.55	0.00
year	-0.01	0.00	-2.72	0.01

Model Equation with Estimated Coefficients

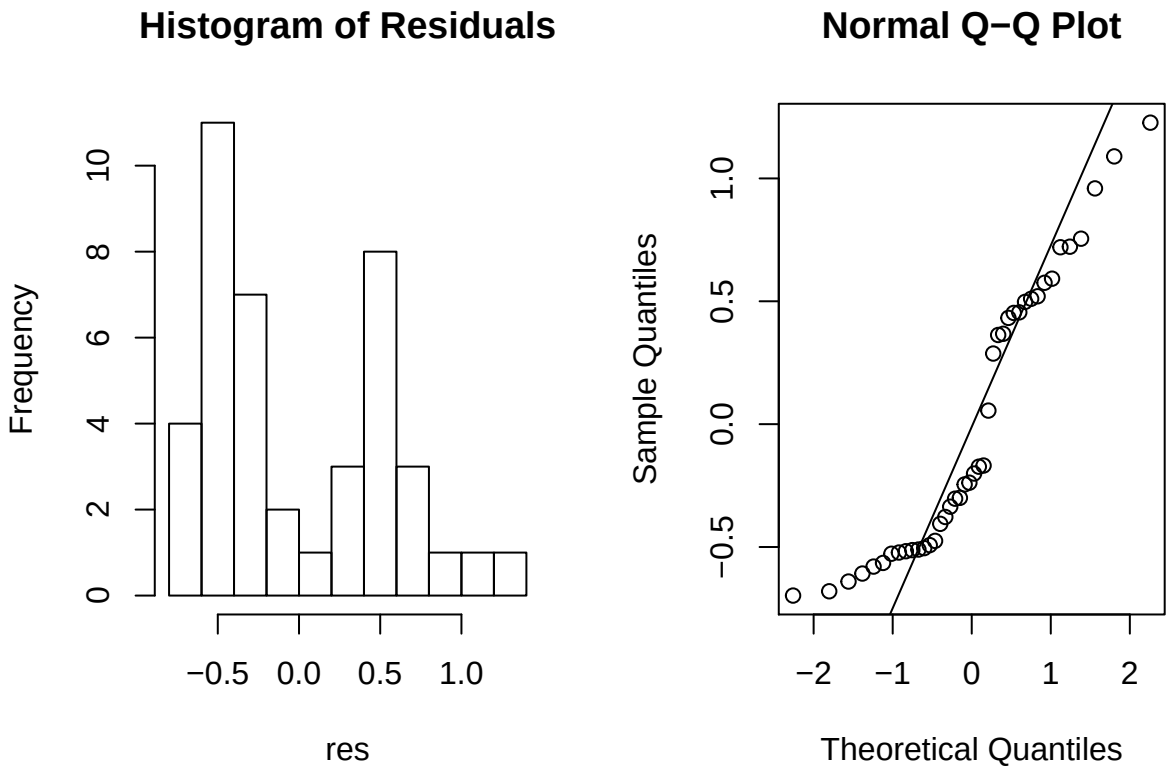
$$goldtime = 26.664 + -0.008 year$$

Interpretation: **Every fourth year we see a decrease of 0.008 of time it takes to win the gold.**

Olympic Data with an Estimated Regression Line



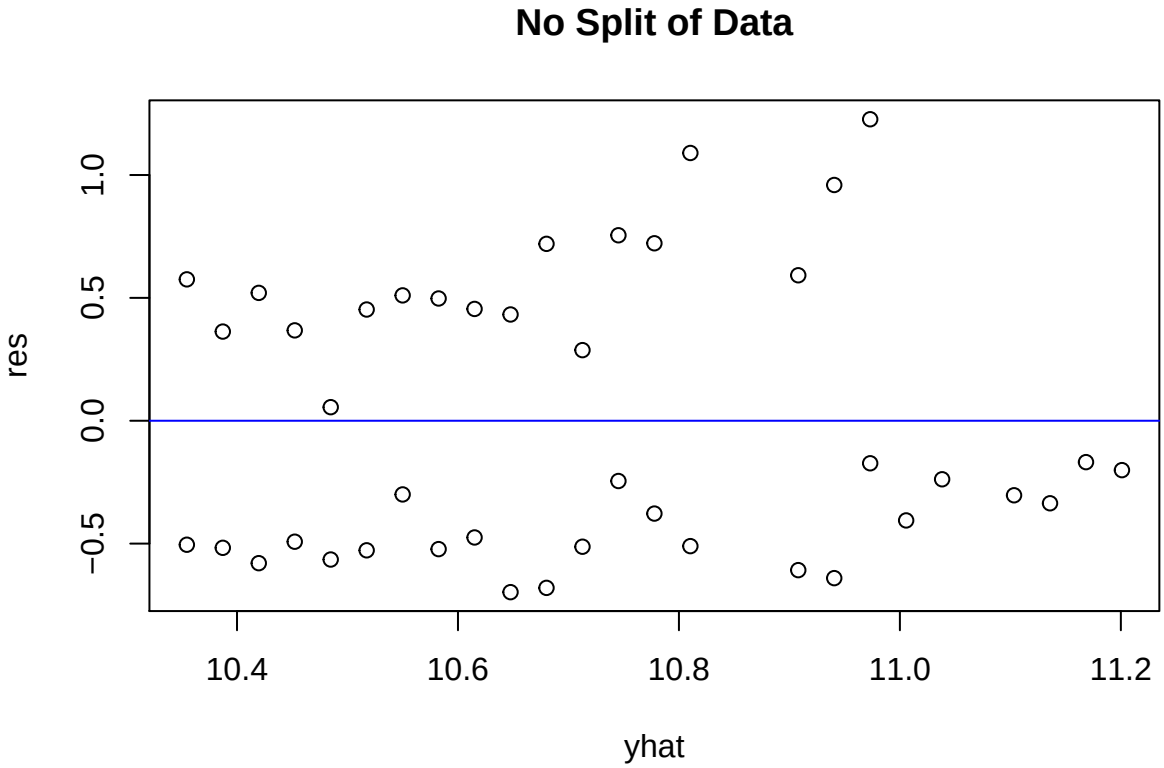
Check of the Residuals of Model

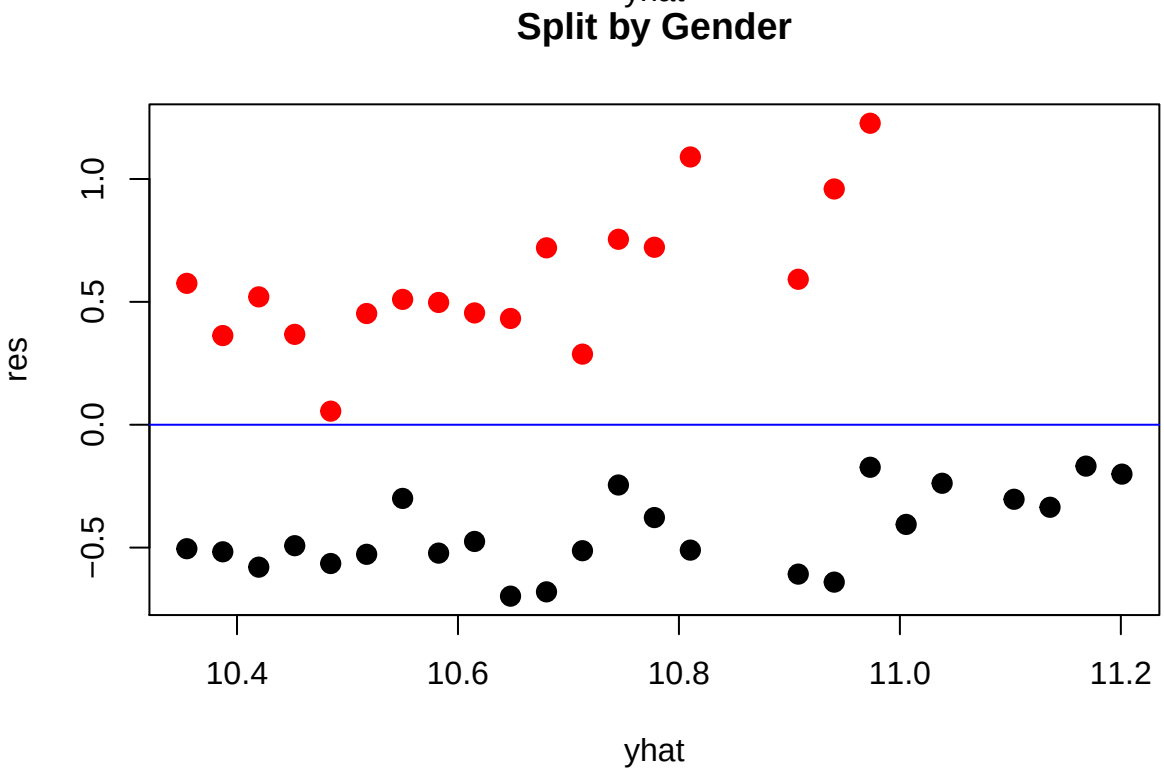


OBSERVATION:

- We have a bi-modal histogram of residuals!!!!!!
- QQ-Plot has some tails to it, which indicates that the errors again do not follow a normal distribution.

Residuals plotted against yhats





CONCLUSION

The current goldtime model is not a good fit for the data. We can see that the Histogram of the Residuals has a bi-modal distributions, which indicates two distinct groups in the data. The Residuals plotted against yhats indicates heteroscadacity and if we color code based on gender we can see an obvious split of the data into two distinct groups.

NEW MODEL with Year and Gender As an Interaction Effect

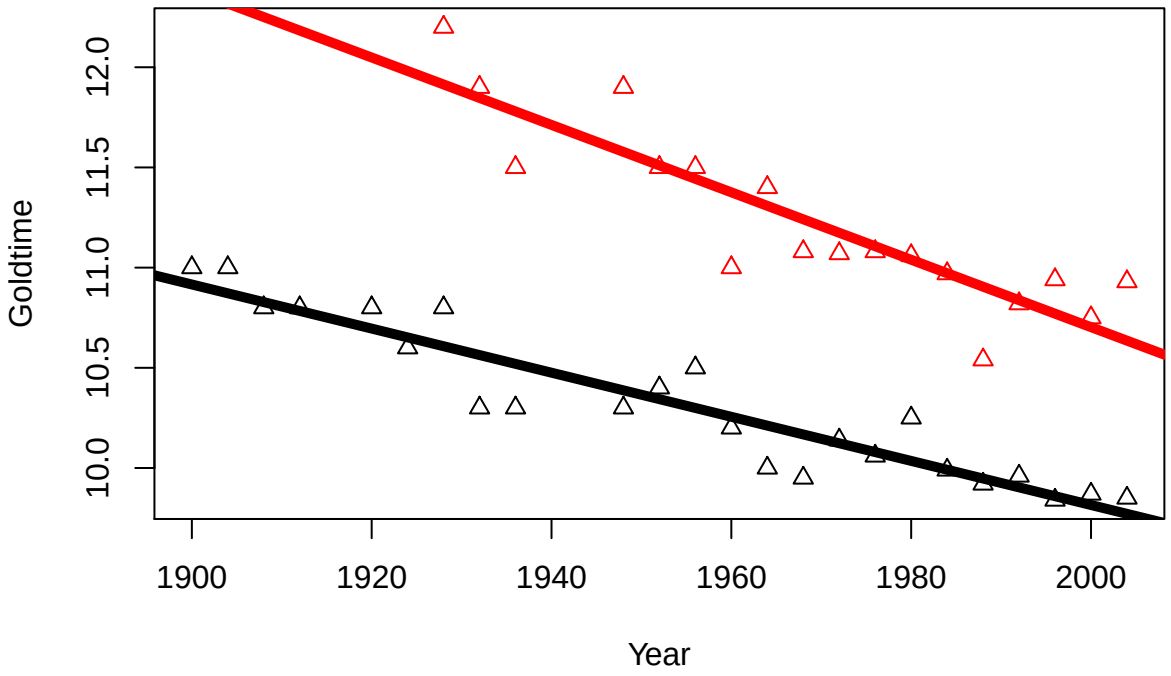
$$goldtime = \beta_0 + \beta_1 year + \beta_2 z + \beta \cdot year \cdot z$$

z is a either 0 for men or 1 for women

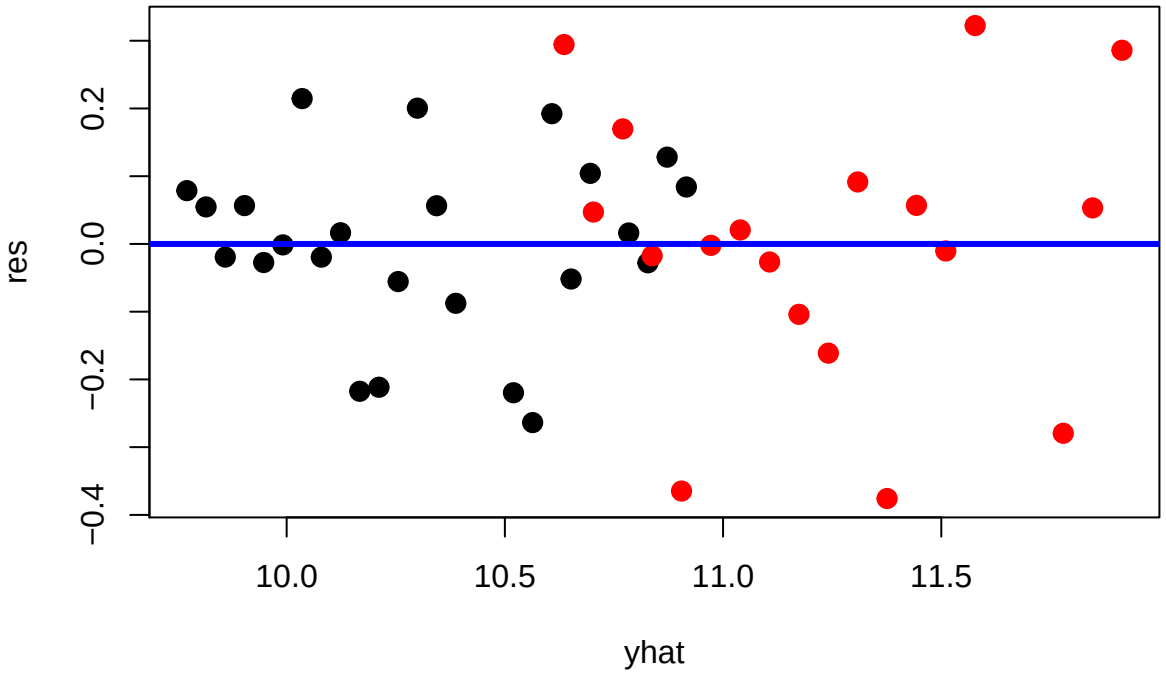
- Not sure how to latex this yet!! Sorry

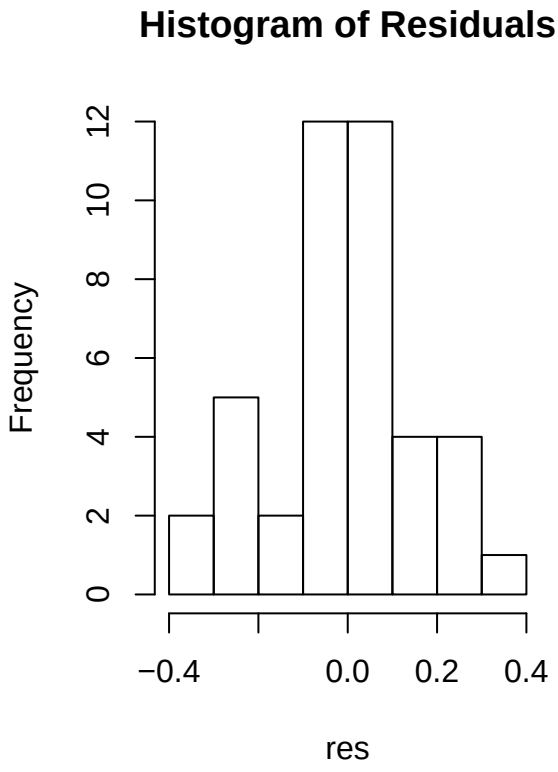
	Est	imate Std	. Error t v	alue Pr(	>|t|)
(Intercept)	31.826	2.129	14.950	0.000	
year	-0.011	0.001	-10.104	0.000	
factor(gender)W	12.521	4.076	3.072	0.004	
year:factor(gender)W	-0.006	0.002	-2.804	0.008	

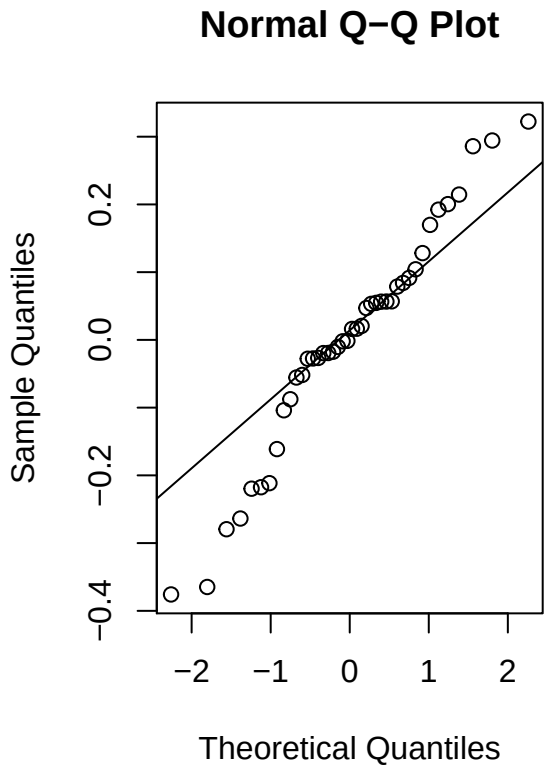
Our new model equation with coefficients is goldtime = 31.83 -0.011year + 12.52gender + -0.006gender*year



We have broken up the data by the variable gender. The lines for each corresponding variable are a much better fit, then the previous regression line where gender was not incorporated.







CONCLUSION

- The new model shows the residual errors having "better" constant error variance. However, it still looks like two groups of residual errors...
- The histogram of the residuals has a better looking normal distribtuion than our previous model.
- The QQ-plot still has tails in it, but the data set is so small.
- The New Model is an improvement from our last model, but we can still do better!

Our new model equation is:

$$goldtime = 31.83 - 0.011year + 12.52gender + -0.006gender * year$$

(4c)

Our new model equation with estimated coefficients:

$$goldtime = 31.83 - 0.011year + 12.52gender + -0.006gender * year$$

- For Men: 31.826452523 + -0.011005562(1944) = 10.43164
- For Women: 31.826452523 + -0.011005562(1944) + 12.520596237 - 0.006(1944) = 11.28824
- In 1944 the *men's* goldtime would of been *10.432* and the *women's* goldtime would of been *11.28824*