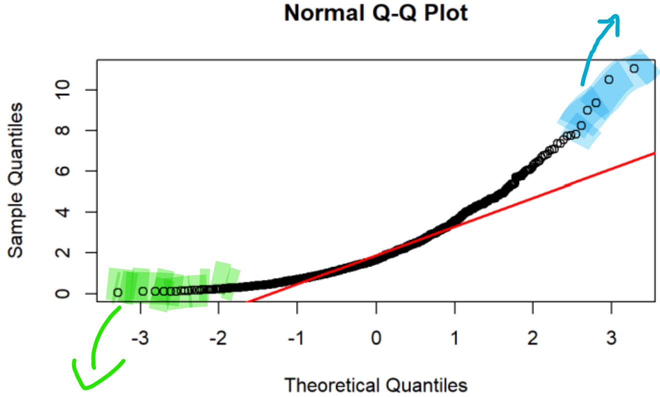
2nd April, 2024

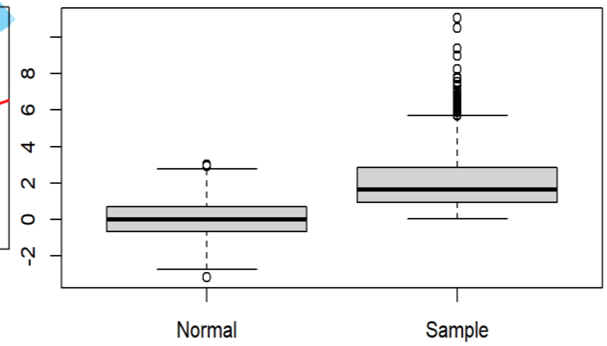
This file is made by me, to highlight different Q-Q plot shapes.

**Part 1: Q-Q plots**

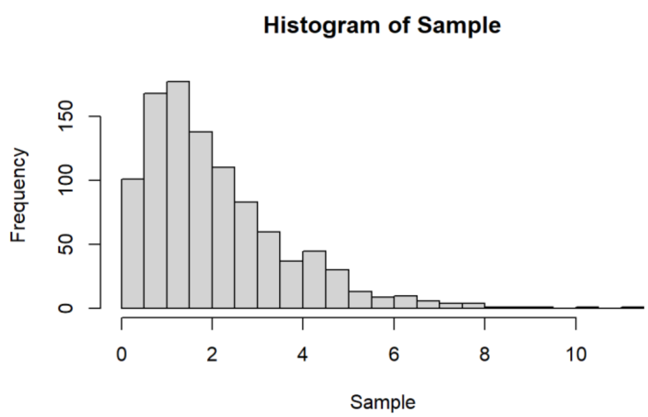
**Right/positive skewed distribution:**

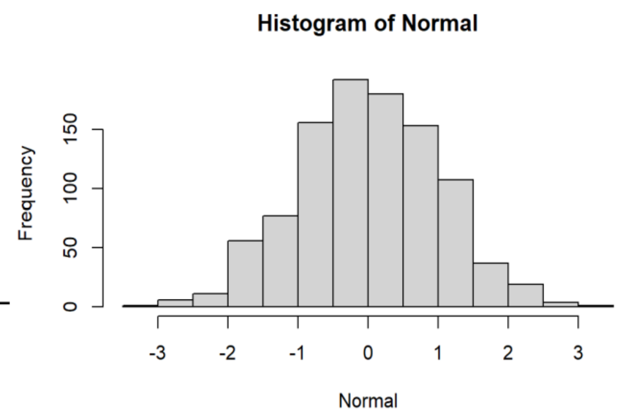
Right tail ABOVE line is called
a Heavy tail.




Left tail ABOVE line is called
a Light tail.



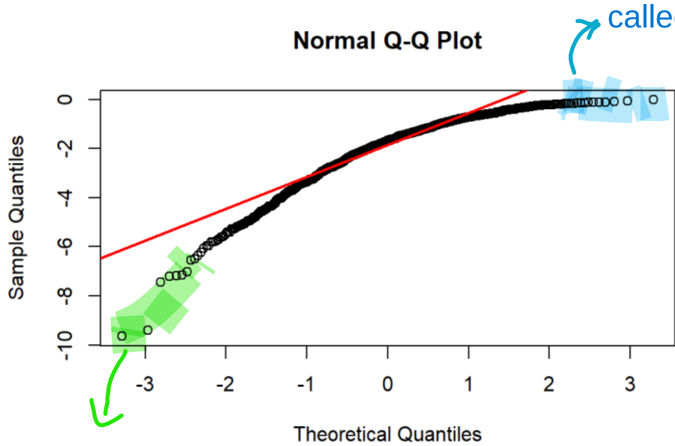

```
     Normal                Sample
 Min.   :-2.85402     Min.   :0.02769
 1st Qu.:-0.68906     1st Qu.:0.92246
 Median : 0.01147     Median :1.65324
 Mean   : 0.01399     Mean   :1.93777
 3rd Qu.: 0.73186     3rd Qu.:2.58775
 Max.   : 2.75883     Max.   :9.76855
```

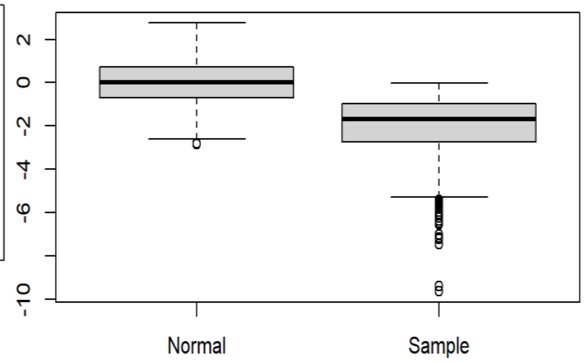
**In this case, the Q-Q plot (backed up the histograms and boxplots) shows us that we observed more positive extreme values than would be expected under the Normal distribution.**

**<u>Hence, distribution is right-skewed.</u>**

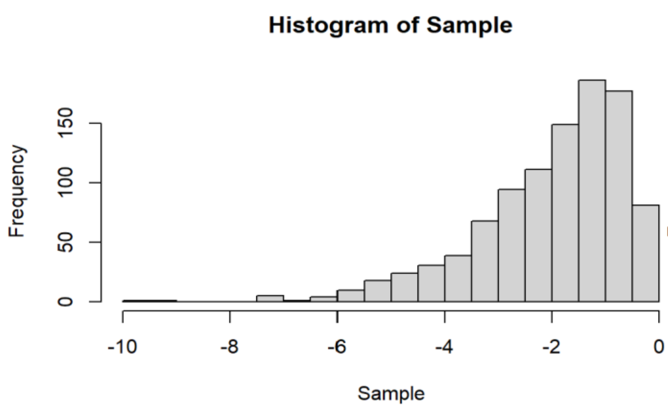
**Left/Negative skewed distribution:**

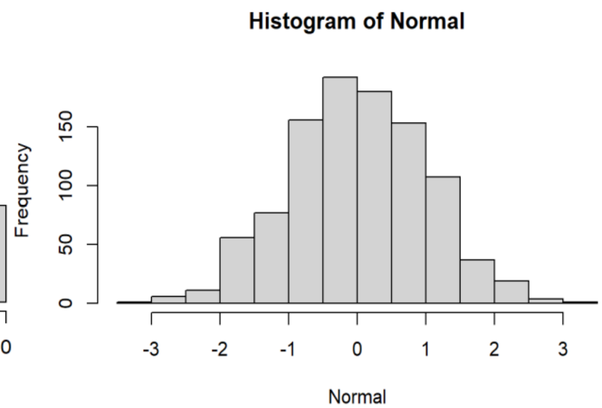
Right tail BELOW line is called a Light tail.




Left tail BELOW line is called a Heavy tail.



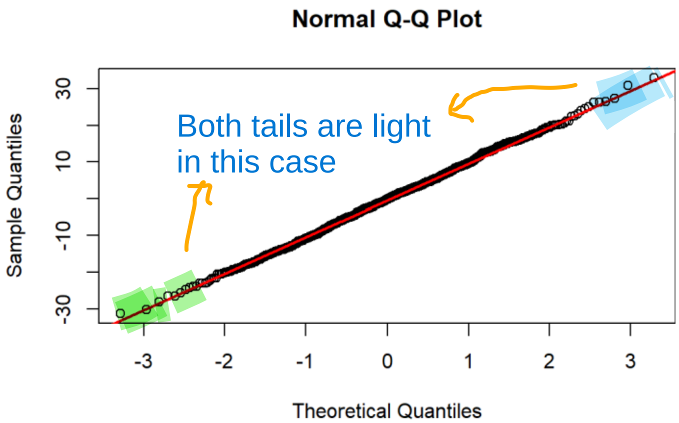



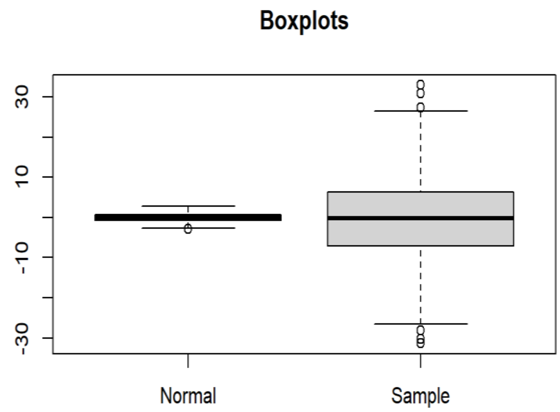

```
     Normal                  Sample
 Min.   :-2.85402      Min.   :-9.64428
 1st Qu.:-0.68906      1st Qu.:-2.73323
 Median : 0.01147      Median :-1.68258
 Mean   : 0.01399      Mean   :-2.01163
 3rd Qu.: 0.73186      3rd Qu.:-0.97950
 Max.   : 2.75883      Max.   :-0.02393
```

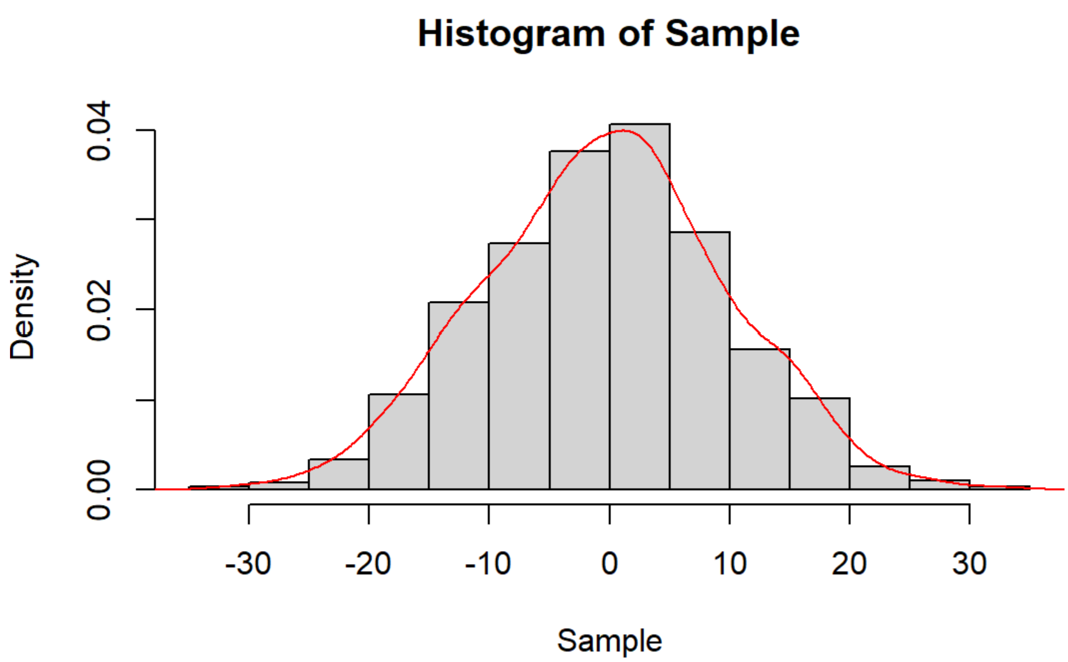
**In this case, the Q-Q plot (backed up the histograms and boxplots) shows us that we observed more negative extreme values than would be expected under the Normal distribution.**

**<u>Hence, distribution is left-skewed.</u>**

**Light-tailed distribution:**



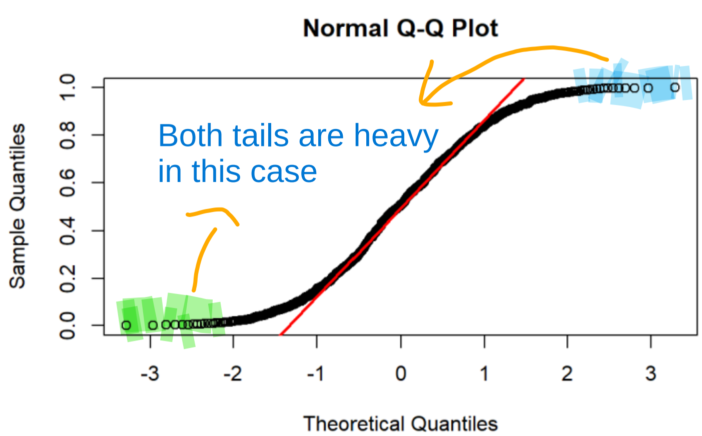



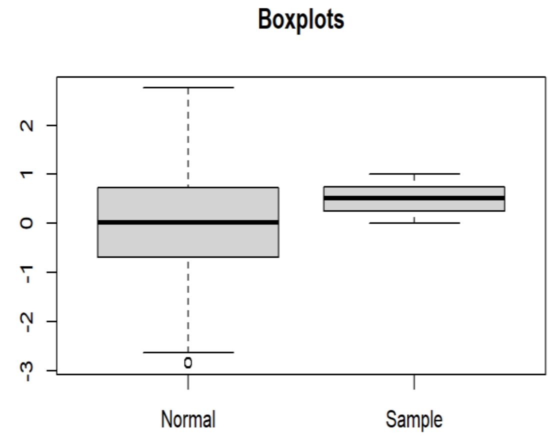



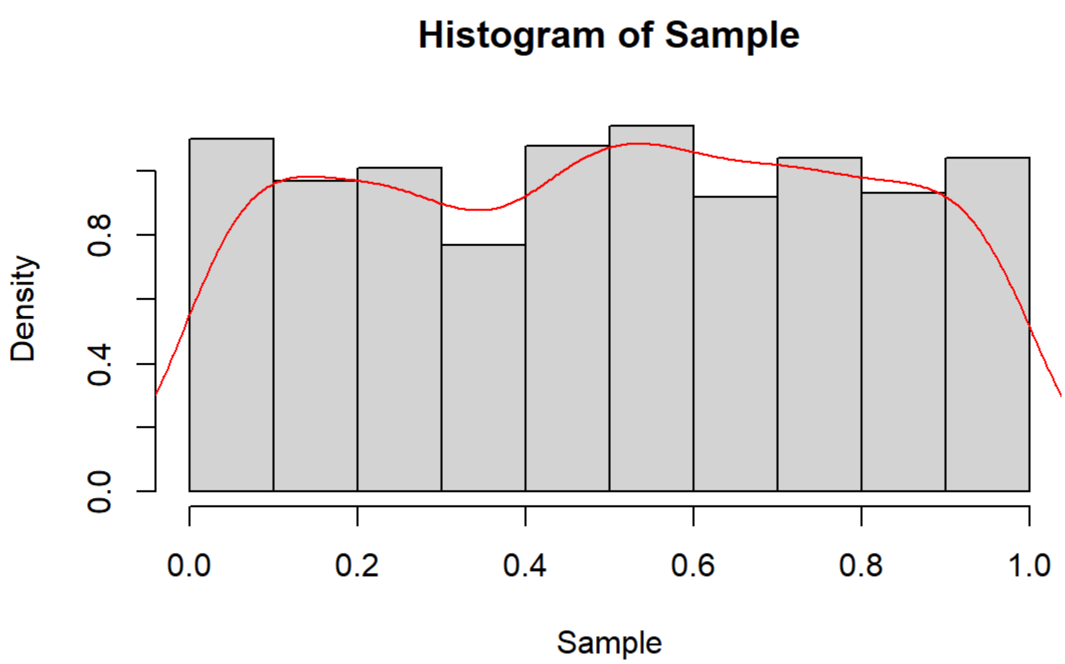

```
    Normal                Sample
Min.   :-2.85402    Min.   :-31.2951
1st Qu.:-0.68906    1st Qu.: -7.1509
Median : 0.01147    Median : -0.1400
Mean   : 0.01399    Mean   : -0.3535
3rd Qu.: 0.73186    3rd Qu.:  6.2861
Max.   : 2.75883    Max.   : 32.8952
```

**In this case, we don't observe as many extreme values as expected under it's Normal distribution.**

**Heavy-tailed distribution:**



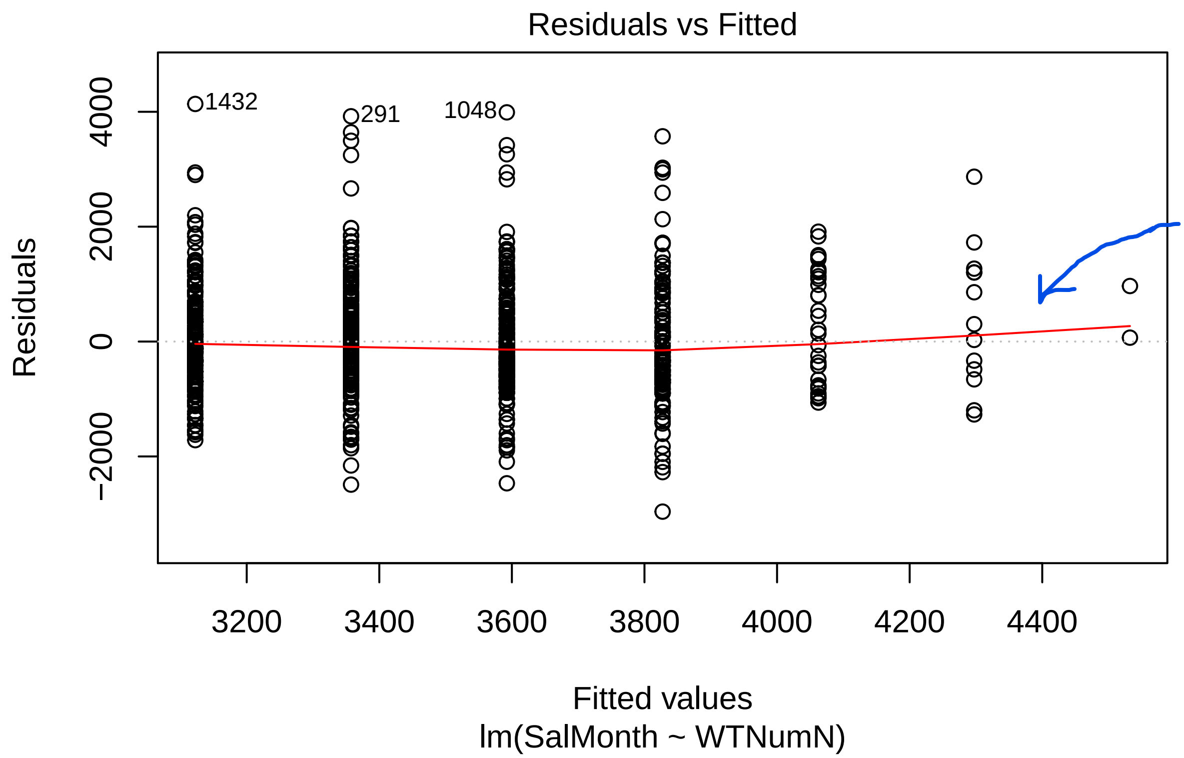

```
     Normal                Sample
 Min.    :-2.85402    Min.    :0.0000412
 1st Qu.:-0.68906     1st Qu.:0.2425101
 Median : 0.01147     Median :0.4919487
 Mean   : 0.01399     Mean   :0.4965139
 3rd Qu.: 0.73186     3rd Qu.:0.7463338
 Max.   : 2.75883     Max.   :0.9999098
```

**In this case, we observe too many extreme values than we'd expect under the Normal distribution.**

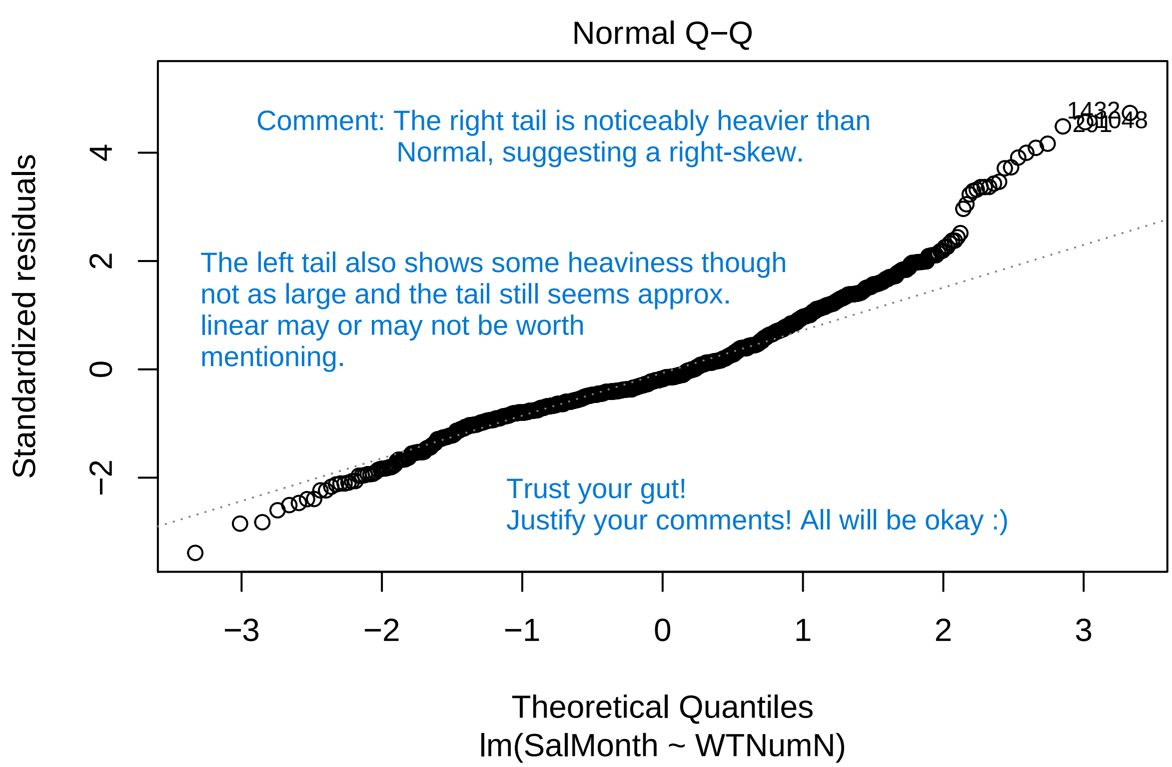
This file is the "COOP SALARIES" PDF in the supplementary material (CH. 6)

Returning to the Co-op Salary ~ Work term model:

- No trends/patterns, so linearity assumptions seems valid
- Constant variance assumption looks good
- Residuals are scattered uniformly around 0, so normality seems ok.




This red line is called Cook's distance. You don't need to worry about it, but it can be helpful in demonstrating that there is no non-linearity present in the residuals.

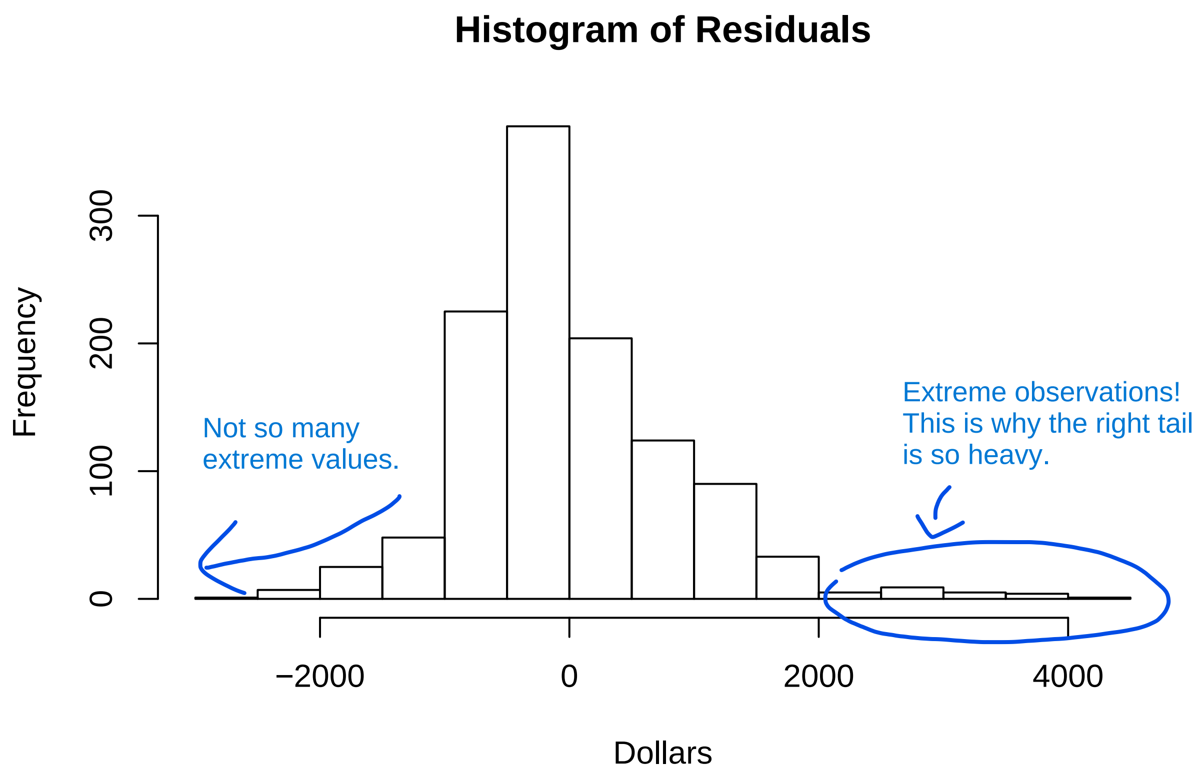



Comment: The right tail is noticeably heavier than Normal, suggesting a right-skew.

The left tail also shows some heaviness though not as large and the tail still seems approx. linear may or may not be worth mentioning.

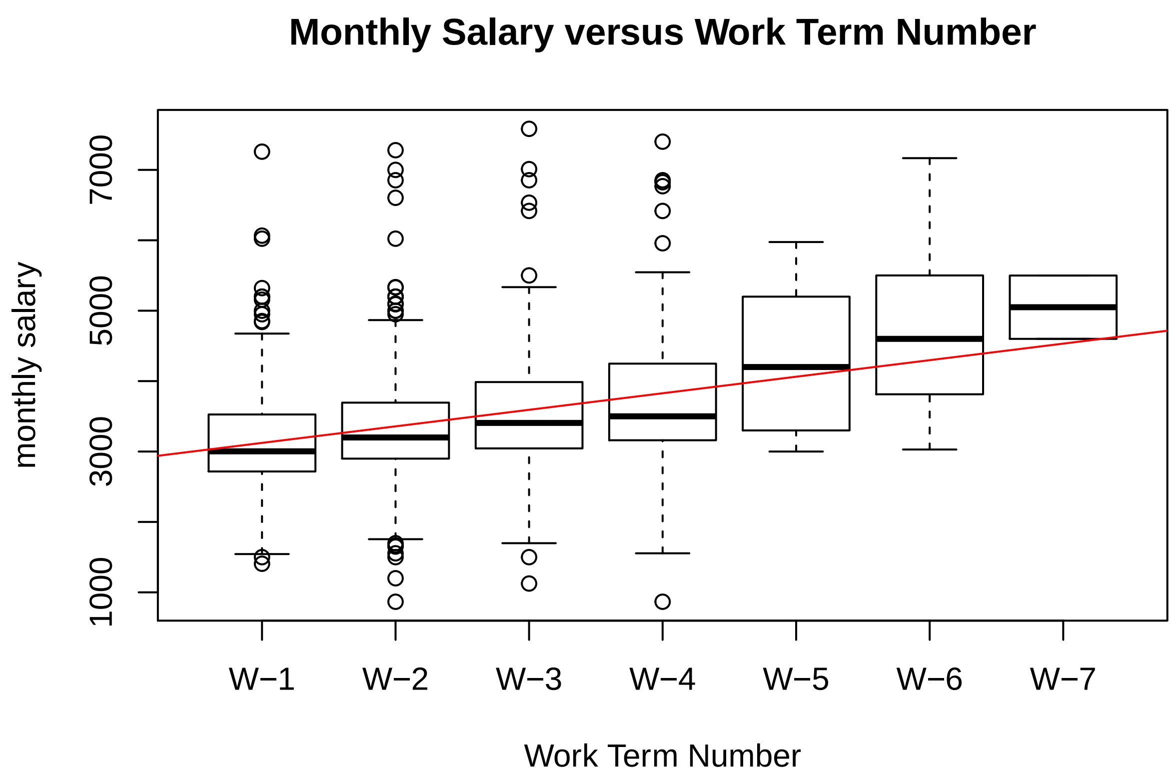
Trust your gut!
Justify your comments! All will be okay :)

```r
hist(resid(Sal.lm),main='Histogram of Residuals',xlab='Dollars')
```

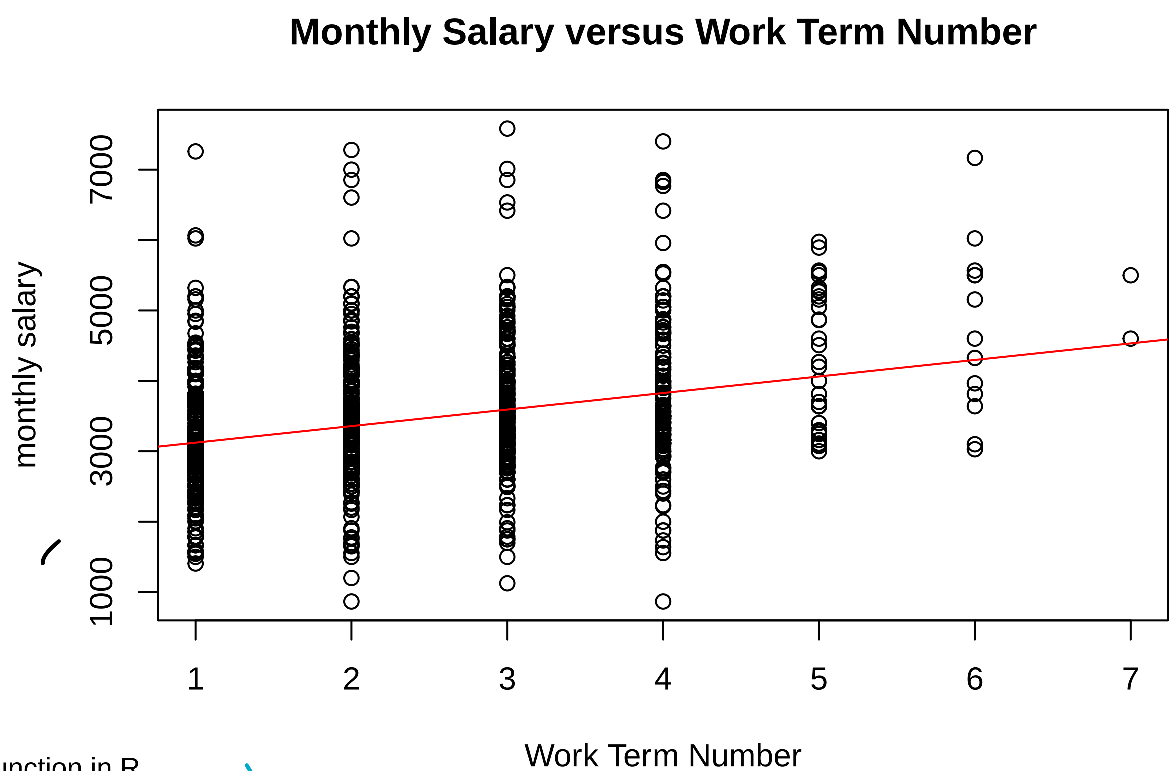



```r
boxplot(salarynz$SalMonth~salarynz$WTNum, xlab='Work Term Number',ylab='monthly salary',
     main='Monthly Salary versus Work Term Number')
abline(Sal.lm,col=2)
```

The very first plot we looked at in this section!




```
plot(salarynz$WTNumN, salarynz$SalMonth, xlab='Work Term Number',ylab='monthly salary',
     main='Monthly Salary versus Work Term Number')
abline(Sal.lm,col=2)
```




CI function in R

this object is where we stored our lm() results.

```
confint(Sal.lm)
```

```
##                    2.5 %     97.5 %
## (Intercept) 2777.026 2997.7795
## WTNumN          193.677   276.3077
```

α β

Present 95% CI for both parameters by default

```
confint(Sal.lm, level=.99)
```

```
##                     0.5 %    99.5 %
## (Intercept) 2742.2548 3032.551
## WTNumN          180.6617   289.323
```

If you want a different % CI, you can specify it.

```
qt(.005,1149)
```

99% critical value for t-distribution

```
## [1] -2.580115
```

```
qnorm(.005)
```

99% Critical value for N(0,1)

```
## [1] -2.575829
```

Given a critical value and lm() output, you should be able to manually compute CI'S for alpha and beta.

```
anova(Sal.lm)
```

```
## Analysis of Variance Table
##
## Response: SalMonth
##                Df      Sum Sq  Mean Sq F value     Pr(>F)
## WTNumN          1    95290565 95290565  124.54 < 2.2e-16 ***
## Residuals 1149 879171908    765163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```