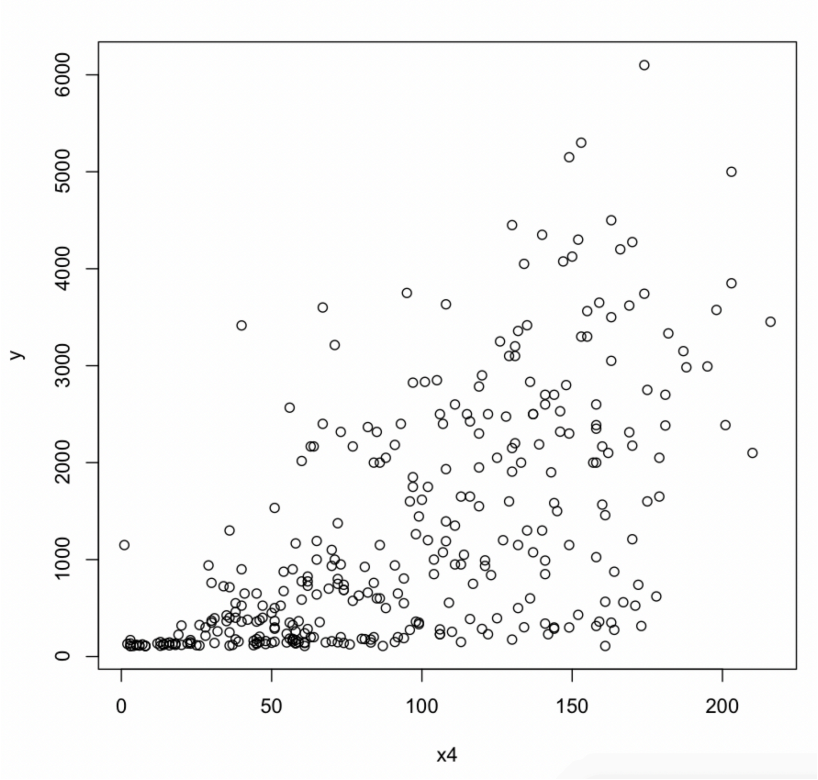
HW 5 baseball data

1.) done in r

2.) What percentage of the variation in salaries is explained by the linear model?

the amount of variation explained by the model is equal to the coefficient of determination , which in the model with all predictors included would be a R^2 = .70, and an R^2 adj of = .68

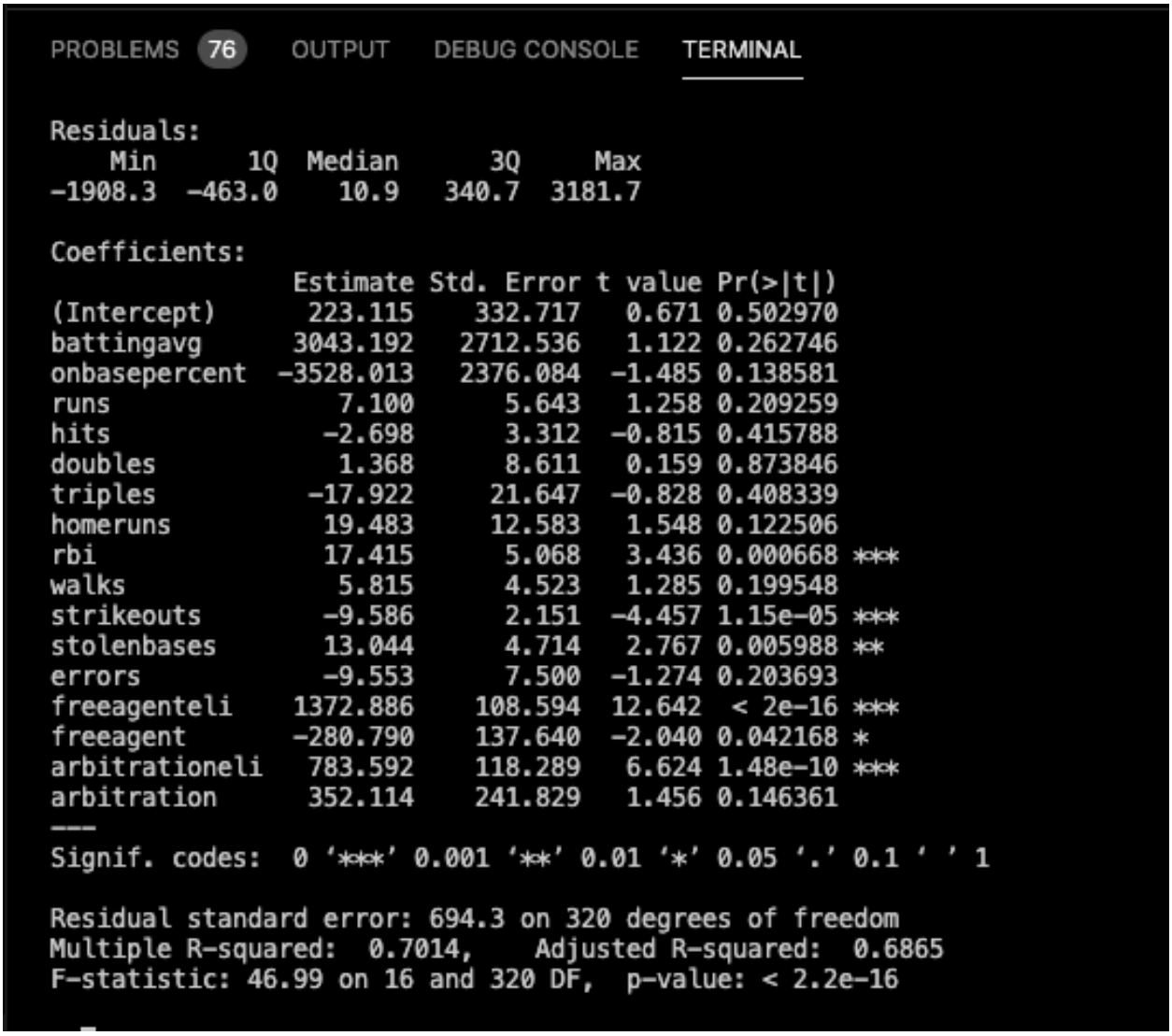
3.) after making a model with only 1 independent variable ‘hits’ and plotting the data , the plot seemed to against my intuition that a player with more hits would make a bigger salary when in fact that while there does seem to be a trend that a higher hit count does suggest a higher salary , the R^2 for this model is only .38.



The non constant variance suggest that there are other variable explaining salary.

However the coefficient of hits with in the multivariate model is a bit more confusing . A negative coefficient of -2.698 is not congruent with my beliefs/hypothesis that the more hits a player has the more their salary is but it is also no consistent with the plot above .

4.) After entering the summary command the p value for the whole model is quite small :



This means we can reject the null hypothesis that all the coefficients in the model are 0.

5.) to do this we need to do the reduction method and we need a model that has every variable except batting average , on base percentage , hits, doubles, and triples. I called this model fitreduced . We need to then get the see for both the model with very variable and the reduced model.

basebal.r ● baseball.csv

HW5 > basebal.r

```
44  fithits = lm(y ~ x4)
45  plot(x4,y)
46
47
48  #testing weather or not we need certain variables.
49  fitreduced = lm(y~x3+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16)
50
51  ssefitfull = anova(fit)
52  ssefitreduced = anova(fitreduced)
```

PROBLEMS 85 OUTPUT DEBUG CONSOLE TERMINAL

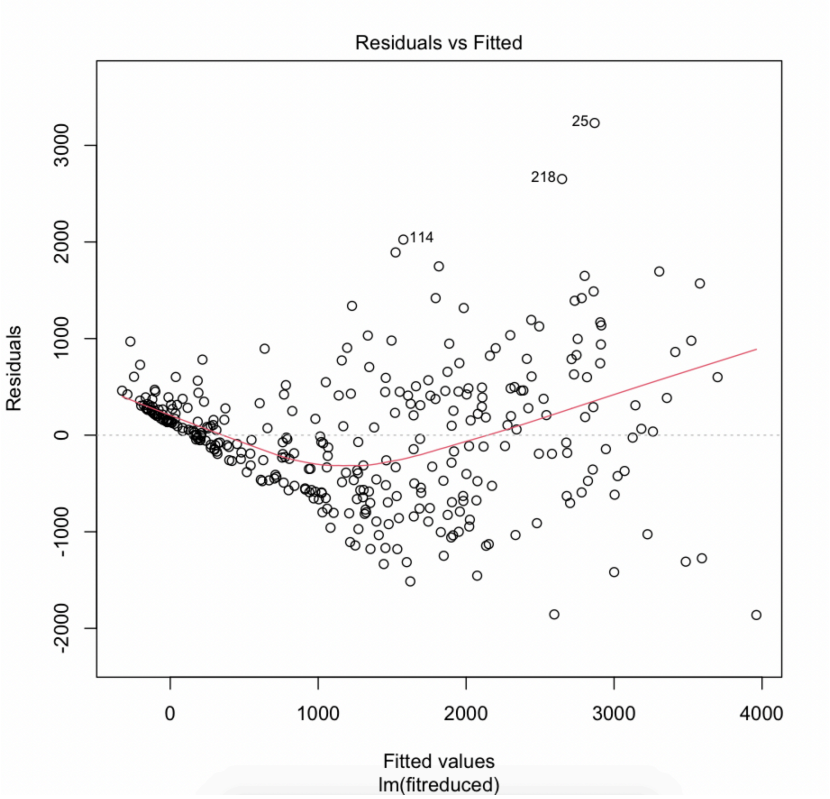
```
> ssefitfull
Analysis of Variance Table

Response: y
           Df    Sum Sq   Mean Sq  F value    Pr(>F)
x1          1  39460222  39460222  81.8623 < 2.2e-16 ***
x2          1  15926699  15926699  33.0408 2.105e-08 ***
x3          1 158167818 158167818 328.1274 < 2.2e-16 ***
x4          1   3832874   3832874   7.9515  0.005104 **
x5          1   1054160   1054160   2.1869  0.140172
x6          1  10093140  10093140  20.9387 6.793e-06 ***
x7          1  13819560  13819560  28.6694 1.642e-07 ***
x8          1   4657607   4657607   9.6624  0.002050 **
x9          1     33021     33021   0.0685  0.793696
x10         1  23036568  23036568  47.7906 2.579e-11 ***
x11         1   2878639   2878639   5.9719  0.015076 *
x12         1   2949330   2949330   6.1185  0.013896 *
x13         1  57916996  57916996 120.1518 < 2.2e-16 ***
x14         1   1666893   1666893   3.4581  0.063862 .
x15         1  25879052  25879052  53.6874 1.929e-12 ***
x16         1   1021939   1021939   2.1201  0.146361
Residuals 320 154250172    482032
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ssefitreduced
Analysis of Variance Table

Response: y
           Df    Sum Sq   Mean Sq  F value    Pr(>F)
x3          1 213542106 213542106 445.0161 < 2.2e-16 ***
x7          1  22438900  22438900  46.7621 3.981e-11 ***
x8          1   8066010   8066010  16.8094 5.227e-05 ***
x9          1    468616    468616   0.9766   0.32378
x10         1  21272565  21272565  44.3315 1.182e-10 ***
x11         1   2511982   2511982   5.2349   0.02278 *
x12         1   1713200   1713200   3.5703   0.05971 .
x13         1  62038861  62038861 129.2874 < 2.2e-16 ***
x14         1   1647942   1647942   3.4343   0.06476 .
x15         1  26014466  26014466  54.2135 1.489e-12 ***
x16         1    977976    977976   2.0381   0.15436
Residuals 325 155952067    479853
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The next step is to plug the sse for both models into the equation. $F = \frac{SSE_r - SSE_f/L}{SSE_f/n-k-1}$ this

value turns out to be f = 0.7061339 . We we look at the f-table the critical value for our numerator DOF (5) and out denominator DOF (320) is 2.31 since F is < $F_{5,320,.05}$ we fail to reject the null hypothesis meaning that we don't really need the variables in the model and in fact when we do a summary of the reduced model and compare the r^2 to the full model we find that they are not very different at all where $R^2_{reduc} = 0.6981$ and $R^2_{full} = 0.7014$. This could make sense because while it would make sense that the higher these stats are the more a player would be paid , salary may be determined more by other factors like seniority and home run count.

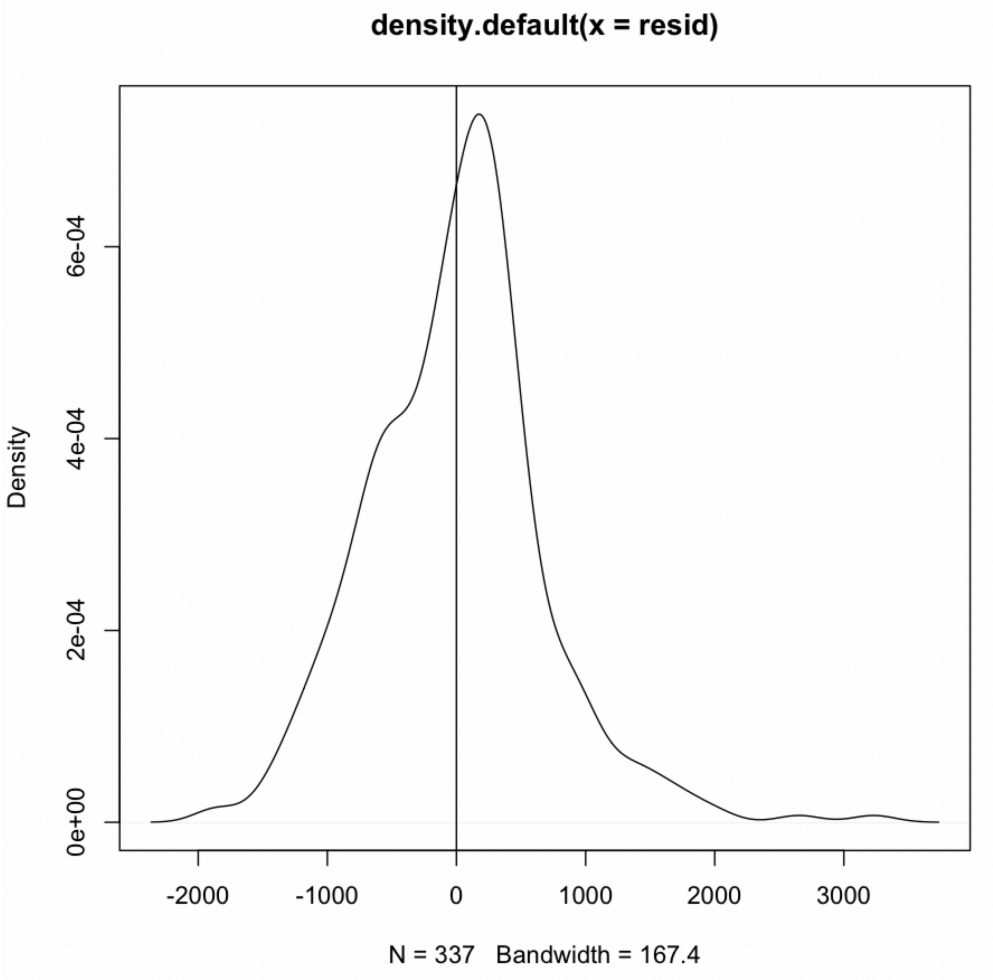
6.) $R^2_{reduc} = 0.6981$

7.) **residuals vs predicted values**



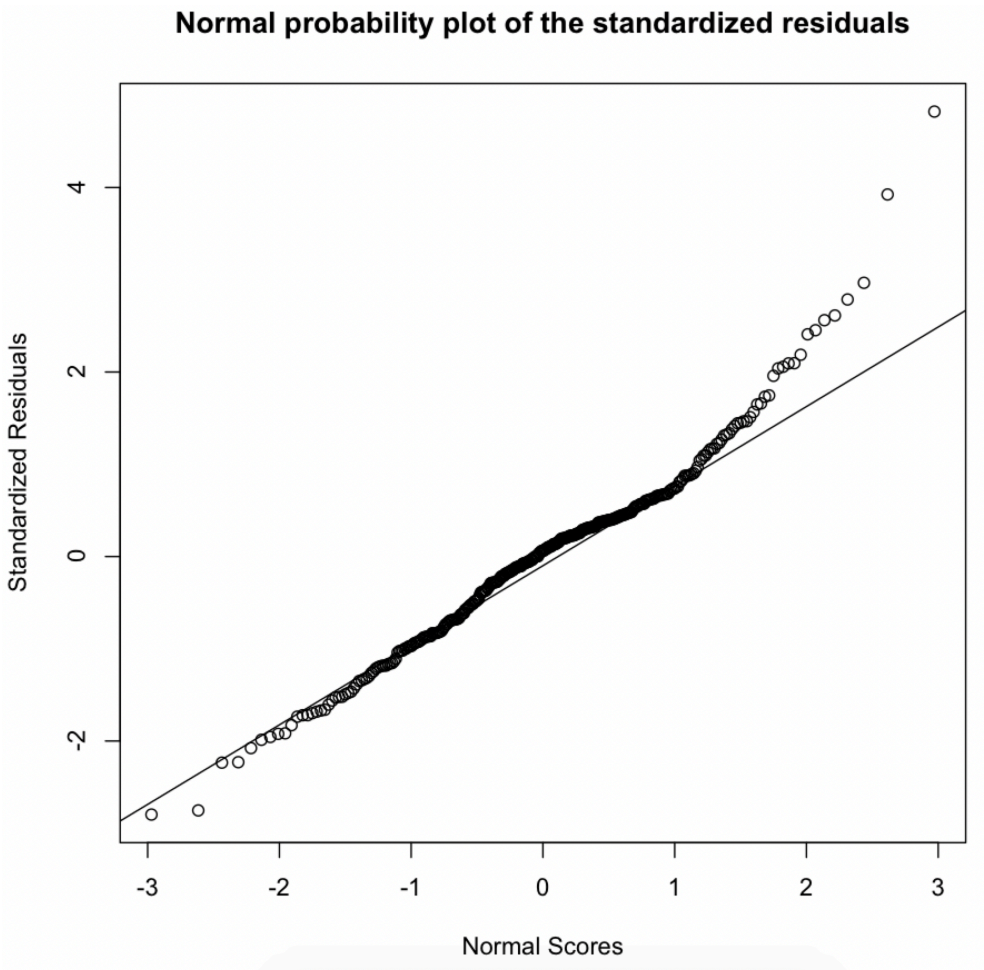
For the most part the plot above is fairly random except for the smaller predicted values that follow a somewhat negative trend . This might suggest that salary is not linearly related to the model with a reduced number of variables.

Kernel density estimate of the residuals



This plot checks the normality assumption the the error terms are normally distributed with an expected value of 0 and this plot does support this assumption.

Normal probability plot of the standardized residuals



Again this plot is trying to show that if a random variable (residuals) follows a normal distribution with mean μ and variance σ^2 . In our assumptions are that error terms should have a normal distribution with mean =0 . This plot would support these assumptions if the plot is linear and for the most part this is true of this plot.