Edna Diaz
Computer Project II
80396667
Stats 4385

Baseball Data Salary

1.

(a) In this section, I have chosen a Stepwise Procedure:

Forward Addition

```
#SELECTING STEPWISE PROCEDURE: FORWARD ADDITION

fit.forward <-  step(fit.null, scope=list(lower=fit.null, upper=fit.whole),
                     direction="forward", k = 2 )
fit.forward$anova
summary(fit.forward)
fit <- lm(fit.forward)
```

```
> fit.forward$anova
   Step Df     Deviance Resid. Df Resid. Dev       AIC
1         NA           NA       336 0.52547467 -2176.212
2  + x2 -1 0.3413313278       335 0.18414334 -2527.586
3  + x9 -1 0.0309893461       334 0.15315400 -2587.685
4  + x4 -1 0.0842436539       333 0.06891034 -2854.826
5 + x10 -1 0.0009680921       332 0.06794225 -2857.594
6 + x13 -1 0.0006421774       331 0.06730007 -2858.794
7  + x8 -1 0.0006093884       330 0.06669068 -2859.859
8 + x15 -1 0.0004417272       329 0.06624896 -2860.099
> summary(fit.forward)

Call:
lm(formula = x1 ~ x2 + x9 + x4 + x10 + x13 + x8 + x15, data = dat)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.064046 -0.005495 -0.000792  0.005048  0.073422 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.061e-03  6.646e-03   0.762  0.44687    
x2           7.805e-01  2.183e-02  35.757  < 2e-16 ***
x9          -1.049e-03  5.841e-05 -17.964  < 2e-16 ***
x4           4.404e-04  3.071e-05  14.341  < 2e-16 ***
x10         -1.125e-04  3.812e-05  -2.951  0.00339 ** 
x13         -4.616e-03  1.931e-03  -2.391  0.01737 *  
x8           1.053e-04  5.818e-05   1.811  0.07110 .  
x15         -3.404e-03  2.298e-03  -1.481  0.13954    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01419 on 329 degrees of freedom
Multiple R-squared:  0.8739,	Adjusted R-squared:  0.8712 
F-statistic: 325.8 on 7 and 329 DF,  p-value: < 2.2e-16
```

- Based on low values of AIC, one can conclude that forward addition model will provide a parsimonious model of the data.

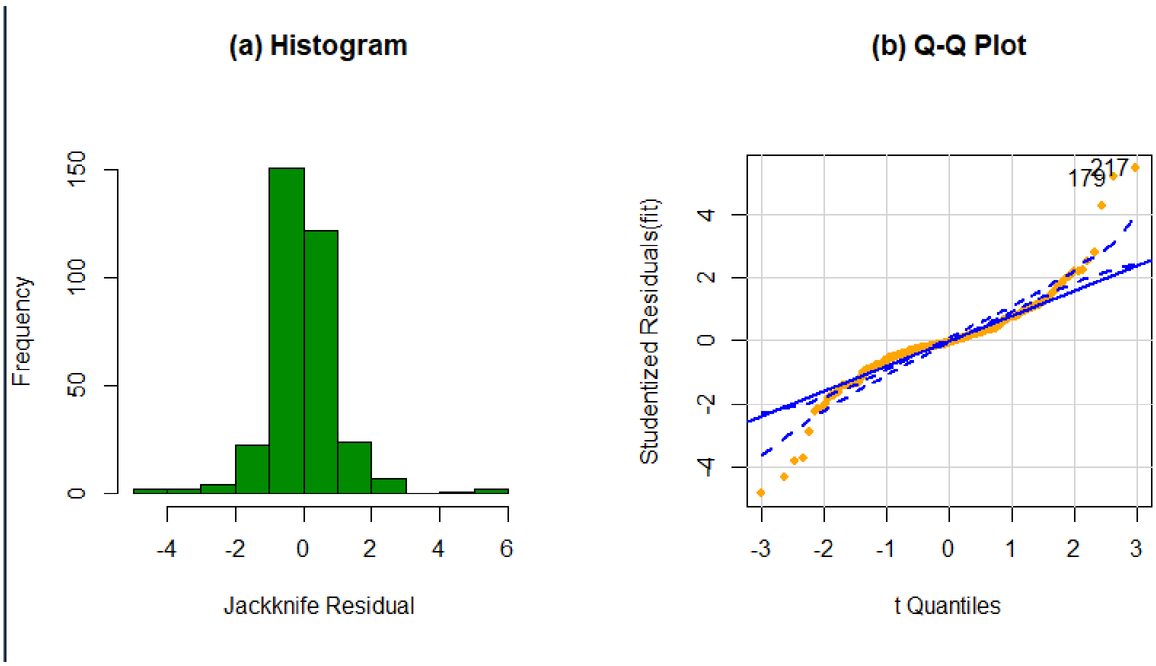
Edna Diaz
Computer Project II
80396667
Stats 4385

(b). R^2: .87 this means that 87% of the data is represented by the best model in this situation.

- Looking at the summary of parameter estimates, it shows that most of the predictors are meaningful addition to the model, except for x8 and x15 due to the fact that the p-values here are higher than .05.

1. (a)

    i.




```
        Shapiro-Wilk normality test

data:  r.jack
W = 0.8749, p-value = 6.591e-16
```

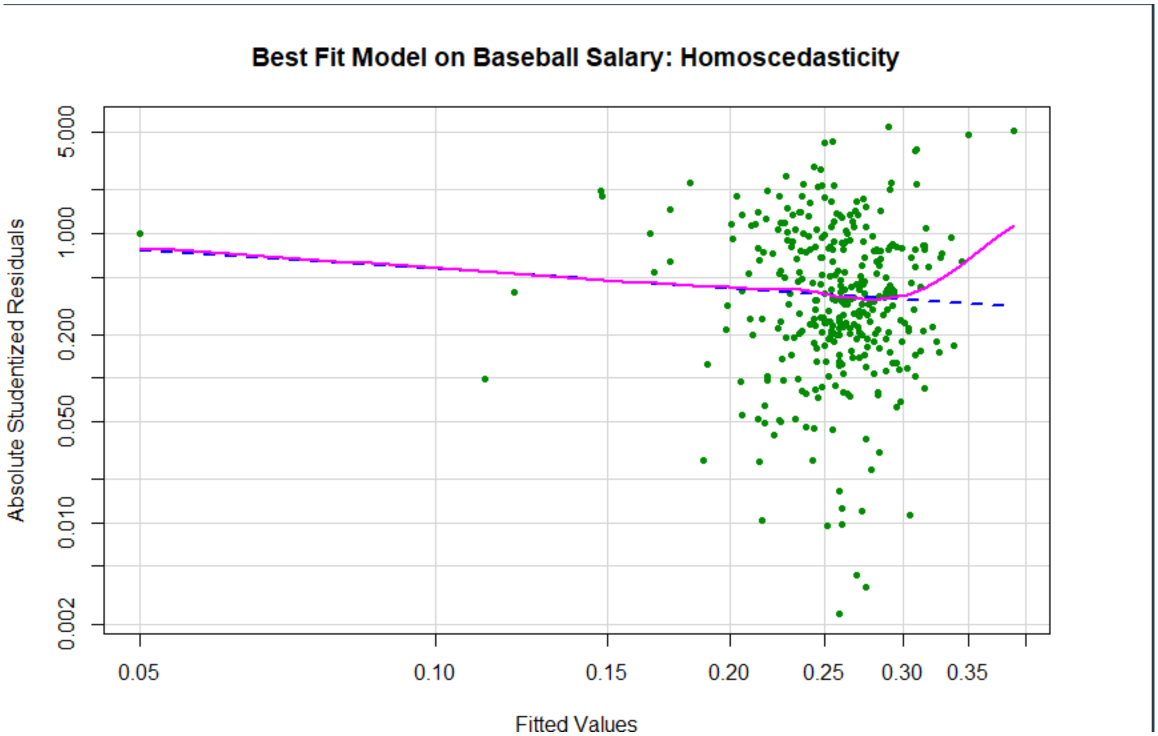
-When looking at the histogram, we can see that there is a bell shaped curve, which looks relatively normal, however, there are a few outliers which are influencing the data on both ends of the curve.

- The Q-Q plot shows us that there are quantities outside of the bounds, which can cause concern when it comes to normality. These points are outliers.

Edna Diaz
Computer Project II
80396667
Stats 4385

-Shapiro-Wilk test shows that the p-value = 6.59e-16, which indicates that this value is smaller than .05. In this case we reject Normality.

ii.



- Based on this graph, both lines should be straight, but there is actually a curve towards the end of this graph. This indicates that there is a violation on equal variance. After running the program, there is a suggested power transformation which will allow for a better fit.

```
Suggested power transformation:  1.444162
> par(mfrow=c(1,1),mar=c(4, 4, 4, 4))
> spreadLevelPlot(fit, pch=20, cex=0.5, col="green4",
+                 main="Best Fit Model on Baseball Salary: Homoscedasticity")
```

-

Edna Diaz
Computer Project II
80396667
Stats 4385

iii.

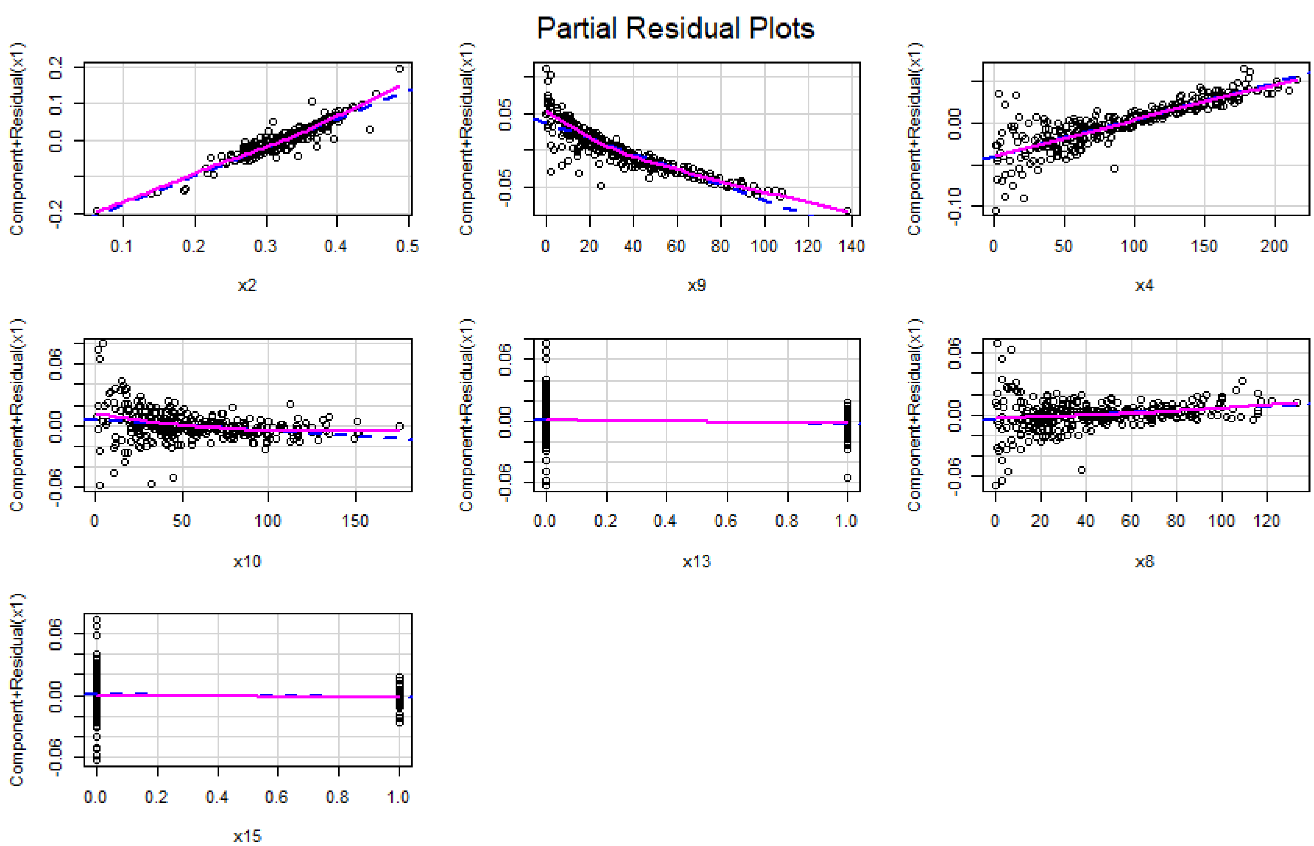
```
> #INDEPENDENCE
> durbinWatsonTest(fit)
 lag Autocorrelation D-W Statistic p-value
   1     -0.03081245      2.020839   0.888
 Alternative hypothesis: rho != 0
> # LARGE P-VALUE (>0.05) JUSTIFIES INDEPENDENCE
>
```

-
- When running the Durbin Watson test, one gets a value of:

  .888.

  .888 > .05, hence this value justifies independence.

  iv.



- When analyzing these partial residual plots, we can see that the smooth line is mostly in line with our model, therefore linearity is justified. The lines are mostly straight in all of the plots except for plot two, the line is a bit curved, but for the most part all of the lines are aligned with the blue line.

Edna Diaz
Computer Project II
80396667
Stats 4385

2b.

Potentially influential observations of

lm(formula = fit.forward) :

	dfb.1_	dfb.x2	dfb.x9	dfb.x4	dfb.x10	dfb.x13	dfb.x8	dfb.x15	dffit	cov.r	cook.d	hat
7	0.39	-0.43	0.45	-0.01	-0.16	-0.34	0.01	-0.06	-0.62_*	0.74_*	0.05	0.03
21	0.01	-0.03	0.13	0.04	0.16	-0.10	-0.16	-0.05	0.30	1.08_*	0.01	0.07_*
52	0.03	-0.02	0.17	-0.15	-0.15	-0.07	0.17	0.08	0.30	1.10_*	0.01	0.09_*
102	0.02	0.06	-0.07	-0.05	-0.02	-0.07	-0.01	-0.07	0.30	0.86_*	0.01	0.01
115	0.02	-0.05	0.27	0.22	-0.01	0.02	-0.33	-0.03	0.46	1.06	0.03	0.08_*
122	0.06	0.00	-0.02	-0.06	-0.05	-0.07	0.03	-0.06	0.24	0.92_*	0.01	0.01
124	0.11	-0.21	-0.10	0.15	-0.10	0.15	0.17	0.08	-0.56_*	0.67_*	0.04	0.02
134	0.95	-1.15_*	0.55	0.36	-0.19	0.02	-0.06	0.02	-1.29_*	0.63_*	0.20	0.07
151	-0.10	0.09	-0.05	0.01	0.04	0.02	-0.01	0.02	-0.10	1.09_*	0.00	0.06
176	0.14	-0.08	-0.02	0.00	-0.08	-0.09	0.02	-0.09	0.27	0.89_*	0.01	0.01
179	-1.34_*	1.57_*	-0.82	-0.37	0.24	-0.03	0.12	-0.04	1.69_*	0.61_*	0.33	0.10_*
192	0.00	0.00	0.00	-0.02	0.01	0.00	0.03	0.00	0.05	1.08_*	0.00	0.05
196	0.00	-0.01	0.10	-0.17	0.21	0.07	-0.04	0.05	0.33	1.11_*	0.01	0.10_*
201	0.00	0.00	0.01	0.00	-0.02	0.01	0.00	0.01	-0.02	1.07_*	0.00	0.05
217	-0.42	0.61	-0.32	-0.29	0.01	-0.09	0.10	-0.08	0.90_*	0.52_*	0.09	0.03
230	0.04	0.10	-0.05	-0.17	-0.15	-0.12	0.10	-0.11	0.52_*	0.68_*	0.03	0.01
234	-0.01	0.00	-0.01	0.01	0.02	0.01	-0.02	0.00	0.03	1.08_*	0.00	0.05
251	-0.02	-0.06	-0.01	0.10	0.03	0.09	0.01	0.07	-0.32	0.85_*	0.01	0.01
261	0.01	-0.02	0.45	-0.14	-0.16	-0.32	0.11	-0.17	0.64_*	1.03	0.05	0.09_*
303	0.38	-0.36	0.17	-0.01	-0.15	0.08	0.03	0.01	0.40	1.17_*	0.02	0.14_*
333	-0.03	0.02	-0.01	0.00	0.01	0.01	0.00	0.00	-0.03	1.10_*	0.00	0.07
337	0.43	-0.56	0.24	0.26	-0.06	0.07	-0.08	0.05	-0.71_*	0.76_*	0.06	0.04

Edna Diaz
Computer Project II
80396667
Stats 4385

Important outliers:

> infl.sum

	dfb.1_	dfb.x2	dfb.x9	dfb.x4	dfb.x10	dfb.x13	dfb.x8	dfb.x15	dffit	cov.r	cook.d
7	0.389995723	-0.4264355569	0.446562148	-0.0145343166	-0.161132777	-0.344951274	0.011235305	-0.059571883	-0.62085185	0.7438331	4.628049e-02
21	0.008493499	-0.0287339538	0.132495836	0.0350860699	0.157862100	-0.100487903	-0.163371563	-0.053863049	0.29689704	1.0764519	1.101477e-02
52	0.029525529	-0.0238868411	0.169603891	-0.1510352822	-0.145019986	-0.069866393	0.168729202	0.075835682	0.29868060	1.0992552	1.115401e-02
102	0.018253719	0.0569400212	-0.066145727	-0.0453093165	-0.016659885	-0.074248537	-0.005973737	-0.074666958	0.30101265	0.8594099	1.109751e-02
115	0.019459671	-0.0526711135	0.270221030	0.2159409337	-0.008284744	0.018113491	-0.327726008	-0.033615670	0.45692766	1.0559968	2.599403e-02
122	0.059753476	0.0015767282	-0.017842059	-0.0574602416	-0.052646434	-0.065329429	0.029428800	-0.061067425	0.23684279	0.9241216	6.932806e-03
124	0.110401130	-0.2132902403	-0.099938374	0.1495274280	-0.097191016	0.147079512	0.167171486	0.084667515	-0.55723352	0.6661960	3.681692e-02
134	0.946686175	-1.1460204594	0.548529842	0.3583796492	-0.190243106	0.024702228	-0.061960837	0.021796167	-1.29365147	0.6340432	1.959052e-01
151	-0.099095563	0.0896857071	-0.045835347	0.0059566716	0.043359851	0.019795675	-0.012616821	0.016860728	-0.10163219	1.0896829	1.294470e-03
176	0.140926851	-0.0812842051	-0.016807022	0.0031554311	-0.076310601	-0.091320189	0.018224961	-0.094079683	0.27354088	0.8905306	9.204903e-03
179	-1.344768055	1.5678508737	-0.815038266	-0.3654402449	0.242251704	-0.027198988	0.118258447	-0.040393616	1.69448223	0.6060888	3.328625e-01
192	-0.003295289	0.0007404214	-0.003848104	-0.0228838575	0.010024679	0.003858830	0.029293693	-0.002883989	0.04546873	1.0769015	2.591821e-04
196	-0.003315265	-0.0110305532	0.099203785	-0.1738280788	0.209030824	0.068156292	-0.035643762	0.053903723	0.33194235	1.1116741	1.377367e-02
201	0.003464739	-0.0022716973	0.010381717	-0.0002263683	-0.015546038	0.008195010	-0.003143824	0.007113748	-0.02403142	1.0749664	7.240618e-05
217	-0.418073625	0.6129039222	-0.319267055	-0.2906528217	0.006973057	-0.088956439	0.097504037	-0.084271468	0.89745260	0.5235272	9.254604e-02
230	0.037856778	0.0993891804	-0.053828533	-0.1659329843	-0.146952457	-0.116173046	0.102828695	-0.106095060	0.51637893	0.6755073	3.167814e-02
234	-0.007027221	0.0038237010	-0.007732400	0.0140896067	0.018157344	0.005959760	-0.016886290	-0.001342567	0.02521361	1.0801569	7.970521e-05
251	-0.022280598	-0.0575932353	-0.014743511	0.0975618844	0.028809650	0.087955751	0.009614324	0.068127398	-0.31660327	0.8469752	1.225408e-02
261	0.012747480	-0.0172698534	0.453283675	-0.1422688029	-0.159260702	-0.324825206	0.107500663	-0.174548763	0.64297190	1.0259622	5.120931e-02
303	0.377903583	-0.3645547606	0.167245354	-0.0149414794	-0.151119785	0.079882972	0.028056540	0.012849551	0.40345062	1.1652648	2.034735e-02
333	-0.026559207	0.0242699408	-0.012387069	0.0010577188	0.011096607	0.005112707	-0.002722924	0.004362686	-0.02707477	1.1008184	9.190700e-05
337	0.427624696	-0.5615144019	0.242786065	0.2572248766	-0.060610012	0.065215457	-0.078305132	0.047155675	-0.71059747	0.7624009	6.074204e-02

Edna Diaz
Computer Project II
80396667
Stats 4385

	hat
7	0.02586285
21	0.07352117
52	0.08849620
102	0.01151800
115	0.08275233
122	0.01167317
124	0.01621161
134	0.06697686
151	0.06318368
176	0.01174491
179	0.09695241
192	0.04944026
196	0.10022057
201	0.04707716
217	0.02622397
230	0.01446676
234	0.05165585
251	0.01179228
261	0.09362491
303	0.14155675
333	0.06941372
337	0.03512142

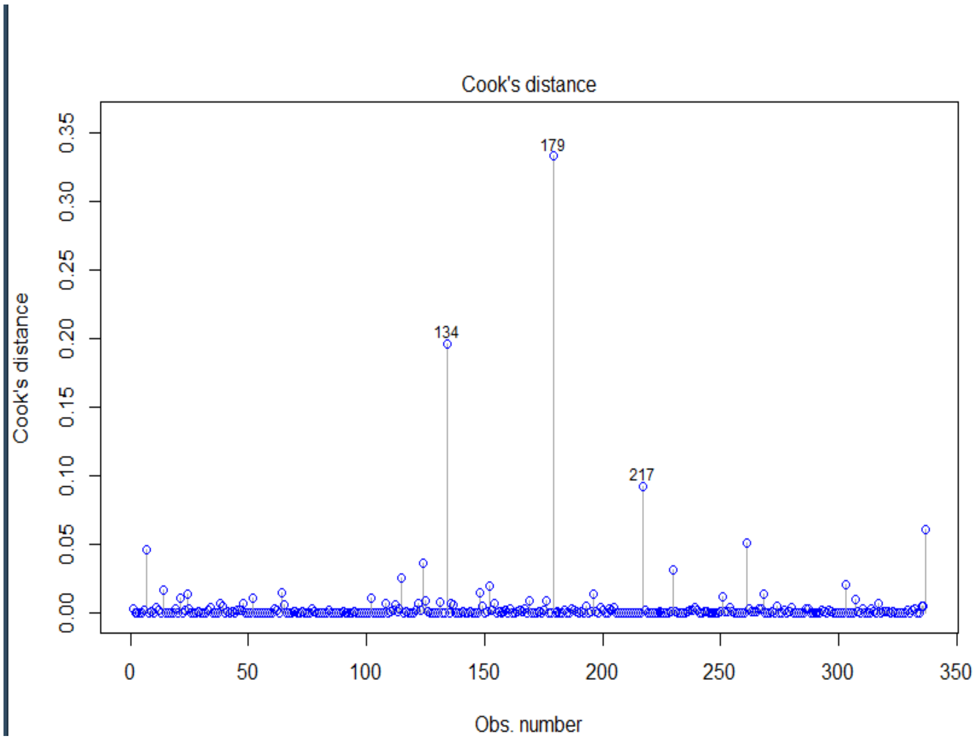
	rstudent	unadjusted p-value	Bonferroni p
217	5.468798	8.9974e-08	3.0321e-05
179	5.171458	4.0494e-07	1.3647e-04
134	-4.828376	2.1147e-06	7.1267e-04
124	-4.340852	1.8932e-05	6.3800e-03
230	4.262049	2.6509e-05	8.9334e-03

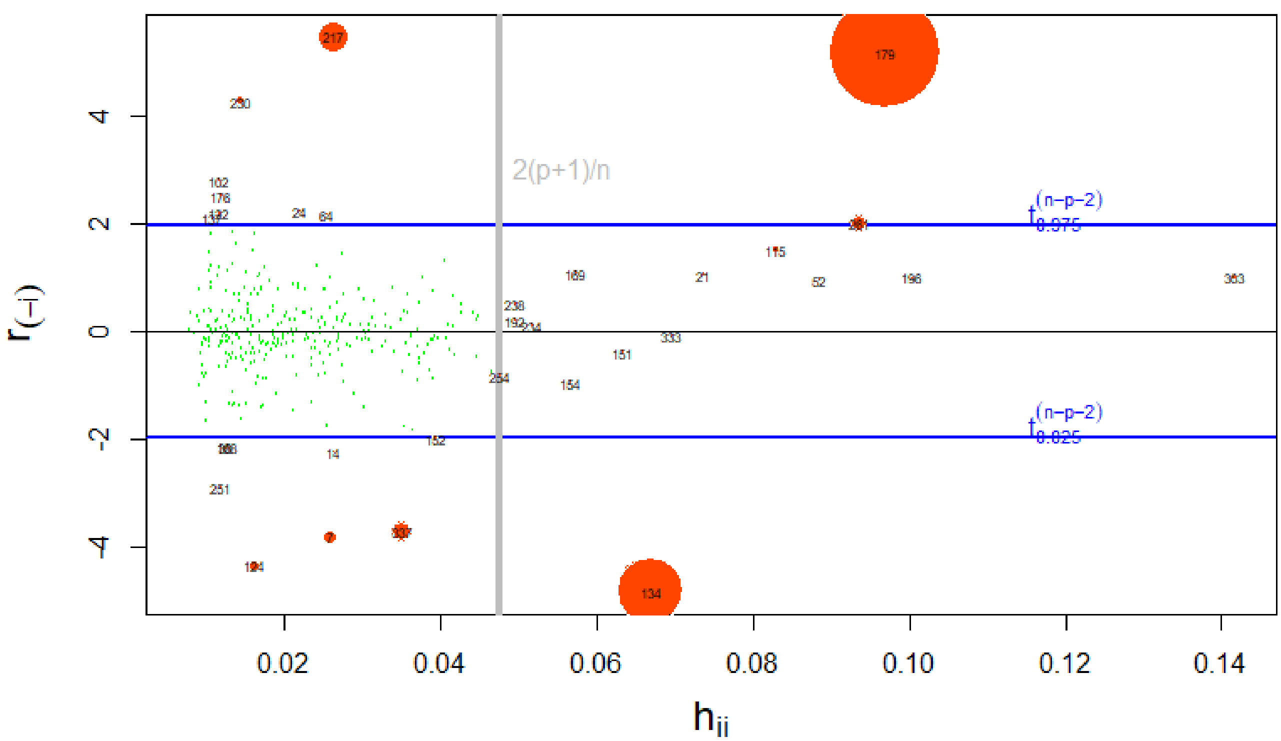
Edna Diaz
Computer Project II
80396667
Stats 4385

```
> # Cook's Distance
  [1] 3.428251e-03 2.481798e-08 4.789807e-04 2.541485e-07 7.128498e-07 1.660096e-03 4.628049e-02 1.182988e-05 7.328436e-04 3.623249e-04
 [11] 4.321672e-03 2.143798e-03 9.575290e-05 1.668970e-02 2.012447e-04 1.934399e-04 2.066900e-04 1.948460e-04 3.049215e-03 6.060846e-05
 [21] 1.101477e-02 4.051989e-04 2.200583e-03 1.390030e-02 2.819766e-03 1.860651e-05 1.115596e-04 5.682811e-04 1.151797e-03 8.110243e-05
 [31] 2.185384e-08 1.415936e-04 2.276641e-03 4.417753e-03 1.297592e-04 6.170737e-05 1.044773e-05 7.162437e-03 4.916026e-03 4.692957e-05
 [41] 7.610083e-04 1.610731e-05 9.979372e-04 7.422803e-06 2.030649e-03 1.714701e-03 2.130398e-03 6.554321e-03 1.188880e-04 1.394549e-04
 [51] 5.280179e-05 1.115401e-02 2.993721e-04 4.664407e-05 2.313736e-04 3.225186e-04 1.217999e-04 8.643652e-05 2.039750e-04 1.835116e-05
 [61] 3.110138e-03 2.083221e-03 3.044664e-08 1.514156e-02 5.562295e-03 7.547050e-05 1.979002e-04 4.178000e-04 6.362479e-04 1.385289e-03
 [71] 7.690182e-05 1.140687e-05 8.690381e-04 5.040992e-07 1.216183e-04 1.420685e-04 3.489237e-03 7.299398e-04 1.704087e-04 4.871876e-04
 [81] 5.315334e-05 1.393957e-04 4.533872e-04 1.945185e-03 3.209389e-04 3.716076e-04 3.513061e-04 6.017130e-05 4.746942e-04 1.199965e-03
 [91] 8.232436e-05 4.531868e-05 1.217393e-04 2.453838e-04 1.101937e-03 2.153299e-04 8.611619e-06 1.058709e-04 1.170284e-04 1.253610e-04
[101] 7.316698e-06 1.109751e-02 5.748774e-04 6.427017e-05 3.237819e-04 9.535880e-05 1.855902e-06 7.252132e-03 5.617621e-05 6.441428e-04
[111] 2.191370e-03 5.850574e-03 6.337661e-04 3.172601e-03 2.599403e-02 9.497698e-05 1.178377e-03 2.699755e-05 1.743121e-06 1.371558e-04
[121] 2.369439e-03 6.932806e-03 1.945869e-03 3.681692e-02 8.775536e-03 1.067404e-03 3.037528e-05 1.089081e-04 8.255154e-05 1.557733e-04
[131] 7.553663e-03 5.177491e-04 3.988195e-04 1.959052e-01 1.001008e-04 7.146617e-03 5.930847e-03 3.198069e-05 2.098361e-04 3.718852e-04
[141] 1.276205e-03 3.765677e-05 4.894603e-06 1.198385e-03 1.657074e-05 5.023486e-05 3.995610e-06 1.527406e-02 4.956714e-03 7.611639e-05
[151] 1.294470e-03 2.014951e-02 1.941662e-03 6.714554e-03 7.854733e-05 1.292575e-03 5.774596e-05 9.343386e-04 1.799285e-03 1.416536e-05
[161] 3.429379e-03 2.086357e-05 1.477952e-05 1.070518e-03 1.619679e-03 9.545779e-04 2.792318e-03 3.747001e-05 8.824365e-03 1.156553e-04
[171] 5.116703e-05 3.073223e-04 6.297294e-04 1.671036e-04 1.746086e-03 9.204903e-03 1.006275e-05 2.122129e-04 3.328625e-01 1.783623e-05
[181] 8.181923e-04 3.069742e-05 1.497456e-05 1.982388e-03 1.195032e-05 3.614812e-05 3.318697e-03 1.936450e-03 7.052122e-04 9.527009e-06
[191] 8.659887e-04 2.591821e-04 4.737324e-03 2.811022e-04 1.404238e-03 1.377367e-02 9.095602e-04 7.080448e-05 4.179762e-03 2.195191e-03
[201] 7.240618e-05 6.344668e-05 2.698050e-03 1.988798e-03 3.911801e-03 1.170721e-04 2.102019e-07 8.677168e-05 4.736738e-06 1.191446e-04
[211] 1.493537e-06 8.722601e-06 5.217655e-05 1.543391e-04 3.216406e-04 1.436451e-04 9.254604e-02 1.666984e-03 5.056771e-04 1.416832e-04
[221] 3.158635e-04 1.111328e-04 1.075427e-05 6.841249e-04 4.943655e-04 9.800860e-05 5.781315e-05 1.005284e-06 7.331811e-04 3.167814e-02
[231] 1.449000e-03 6.510105e-05 2.282379e-04 7.970521e-05 1.093156e-04 6.343210e-04 2.083607e-03 1.599809e-03 4.199374e-03 1.736035e-03
[241] 5.315661e-04 5.238665e-05 7.308174e-04 7.834138e-04 6.612895e-05 6.187331e-05 3.166388e-04 1.213943e-04 2.594375e-04 9.949821e-04
[251] 1.225408e-02 1.457806e-03 6.210506e-04 4.319977e-03 6.009819e-04 4.451961e-07 1.512839e-04 4.387588e-04 3.401960e-05 9.726536e-05
[261] 5.120931e-02 2.627321e-03 7.658791e-04 1.219903e-03 1.522231e-03 2.950603e-03 3.013276e-03 1.395624e-02 4.092455e-05 1.477634e-04
[271] 1.792710e-04 1.287597e-03 3.141972e-04 4.501672e-03 3.023288e-04 1.942800e-04 1.889718e-03 2.056897e-06 1.549983e-03 3.937179e-03
[281] 4.994443e-04 8.823175e-08 1.552935e-07 4.669797e-04 2.387253e-05 2.644244e-03 3.271656e-03 1.177105e-04 2.998526e-04 4.461541e-04
[291] 3.511973e-05 8.498584e-05 2.082669e-03 1.186530e-03 3.488432e-04 1.989187e-03 8.903605e-04 1.833569e-04 2.718209e-04 2.840674e-06
[301] 7.889183e-05 4.698860e-04 2.034735e-02 1.070434e-05 8.549526e-06 2.233192e-04 9.827976e-03 8.650494e-04 2.206026e-06 2.638270e-03
[311] 3.486841e-04 2.728600e-04 8.994041e-06 3.237919e-03 1.365135e-03 3.359904e-04 7.032374e-03 2.828365e-04 6.407341e-04 6.547888e-04
[321] 6.251743e-04 9.390103e-06 6.113018e-04 1.251517e-04 9.947106e-05 6.552539e-05 2.375062e-04 6.265204e-05 1.762888e-03 1.996794e-04
[331] 2.149555e-03 3.435186e-03 9.190700e-05 1.684836e-04 5.306631e-03 4.510433e-03 6.074204e-02
```

```
> # EXTRACT INFLUETIAL POINTS
> baseball[cook.d > 0.05, ]   # HIGH COOK'S DISTANCE
       x1    x2  x3  x4 x5 x6 x7  x8  x9 x10 x11 x12 x13 x14 x15 x16 logsalary
134 0.286 0.444   0   2  0  0  0   0   2   3   0   0   0   0   0   0  4.867534
179 0.457 0.486   6  16  4  2  0   7   2   2   0   2   0   0   0   0  4.976734
217 0.364 0.364   2   4  1  0  1   1   0   4   0   0   0   0   0   0  4.691348
261 0.318 0.453 104 178 31  2 32 109 138 112   1   2   0   0   0   0  6.429719
337 0.258 0.395   6   8  1  0  1   6   7  11   0   0   0   0   0   0  4.691348
>
```

Edna Diaz
Computer Project II
80396667
Stats 4385





Based on these values and the graph generated we see that our biggest outliers are 134, 179 and 217. These points are also highly influential on the data. The number 363 has the greatest leverage in this graph and the outliers under and over the bands are also considered outliers. Two of our greatest influencers also have high leverage as well.

2C.

```
> fitUpdate <- update(fit, ~. -1, data = dat);
>
> kappa(lm(fitUpdate, x=TRUE)$x);
[1] 632.4887
```

632.4887 > 100

-Based on this answer, we can see that the number we got is larger than the cut-off point of 100. It is actually a very serious concern in this case based on this number, as it is very big. There may be predictors that may be collinear with other predictors in this case.

```
> vif(fit)
      x2       x9       x4      x10      x13       x8      x15
1.766046 3.513289 4.237453 2.775130 1.493984 4.935015 1.376195
```

-When we look at VIF, we can see that all of our numbers are below ten, so there may not be an issue with multicollinearity.

3.

```
> ##95% interval
> player.new <- baseball[1, ]
> predict(fit, newdata=data.frame(player.new), se.fit=TRUE,interval="prediction", level=0.95);
$fit
        fit       lwr       upr
1 0.2824034 0.2538479 0.3109589

$se.fit
[1] 0.00305672

$df
[1] 329

$residual.scale
[1] 0.0141903
```

-Based on my best fit model, we can assume that there is a 95% chance the salary of this baseball player with the same stats as Andre Dawson, will fall between [.2538479, .3109589] in thousands of dollars.

Edna Diaz
Computer Project II
80396667
Stats 4385

Index

1. Set-up:

```
library(car)


baseball <- read.table(file=
                        "http://www.amstat.org/publications/jse/datasets/baseball.dat.txt", header
= F,
                       col.names=c("salary", "x1", "x2", "x3", "x4", "x5","x6", "x7","x8", "x9",
"x10", "x11", "x12", "x13", "x14", "x15", "x16", "ID"))
baseball$logsalary <- log(baseball$salary);
baseball <- baseball[, -c(1, 18)] # REMOVE salary AND ID dim(baseball);
head(baseball)
dat <- baseball
logsalary <- baseball$logsalary

fit.null <- lm(x1 ~ 1, data = dat)
fit.whole<- lm(baseball, data = dat, x = TRUE, y = TRUE)


n <- NROW(dat);
p <- length(coef(fit))-1 # NUMBER OF SLOPES
```

(a). #SELECTING STEPWISE PROCEDURE: BACKWARD DELETION

```
fit.forward <-  step(fit.null, scope=list(lower=fit.null, upper=fit.whole),
                  direction="forward", k = 2 )
fit.forward$anova
summary(fit.forward)
fit <- lm(fit.forward)
```

2.
(ai).
 #CHECKING NORMALITY: SHAPIRO, HISTOGRAM AND JACK RESIDUALS

```
r.jack <- rstudent(fit)

par(mfrow=c(1,2),mar=c(8,4,8,4))
```

Edna Diaz
Computer Project II
80396667
Stats 4385

```
# The fisrt plot: Histogram
hist(r.jack, xlab="Jackknife Residual", col="green4",
    main="(a) Histogram")
# FUNCION
qqPlot() in {car}: A fancier qq plot for studentized jackknife residuals
qqPlot(fit, pch=19, cex=.8, col="orange", main="(b) Q-Q Plot")
# ?qqPlot
shapiro.test(r.jack)
```

(aii).

```
#HOMOSCEDASTICITY


# Plot Absolute Jackknife Residuals vs. Fitted values
# Power Box-Cox Transformation on the response Y is suggested
# library(car)
par(mfrow=c(1,1),mar=c(4, 4, 4, 4))
spreadLevelPlot(fit, pch=20, cex=0.5, col="green4",
            main="Best Fit Model on Baseball Salary: Homoscedasticity")
# IF THE LINES ARE FLAT, THEN EQUAL VARIANCE IS JUSTIFIED
```

(aiii).

```
#INDEPENDENCE
durbinWatsonTest(fit)
# LARGE P-VALUE (>0.05) JUSTIFIES INDEPENDENCE
```

(aiv).

```
#LINEARITY
crPlots(fit, main="Partial Residual Plots")
```

2B.

```
infl <- influence.measures(fit);
infl.mat <- as.data.frame(infl$infmat)
infl.sum <- summary(influence.measures(fit));
infl.sum
write.csv(infl.sum, file="Infleunce-Mat.csv", row.names=TRUE)

# Cook's Distance
cook.d <- infl.mat$cook.d; cook.d

# library(car)
outlierTest(fit) # Bonferonni p-value for most extreme obs

# Plot of Cook's Distance
```

```
cutoff <- 4/(n-p-2)
plot(fit, which=4, cook.levels=cutoff, col="gray65", lwd=1.5)
points(1:n, cook.d, pch=1, cex=1, col="blue")

# EXTRACT INFLUETIAL POINTS
baseball[cook.d > 0.05, ]   # HIGH COOK'S DISTANCE


# Interactive Plot for Identifying Influential Points
# Press ESC to stop when you are done with identification
influencePlot(fit, id=list(method="identify"),
        col="blue",
        main="Influence Plot",
        sub="Circle size is proportial to Cook's d")
```

2C.

```
# CONDITION NUMBER (> 15, 30, 100?);

#WITH INTERCEPT
fitIntercept <- lm(fit, x=TRUE)
kappa(fitIntercept$x);


# WITHOUT INTERCEPT

fitUpdate <- update(fit, ~. -1, data = dat);

kappa(lm(fitUpdate, x=TRUE)$x);


# COMPUTE VIF USING FUNCTION vif DIRECTLY (> 10?)
vif(fit)
```

3.

```
##95% interval

player.new <- baseball[1, ]

predict(fit, newdata=data.frame(player.new), se.fit=TRUE,interval="prediction",
level=0.95);
```