Prediction of Average Salary of Baseball players using SAS and MS Excel

- By Sandeep Ramesh

1.To find outliers in X, outliers in Y and influential points.

h_Salary	Hii value	rstudent_Salary	residuals	cookd_Salary	Cooks distance
0.133025	0.069264	3.463409579	1.970707	0.046875461	0.920787
0.107226		3.425504058		0.044624085	
0.09449		3.211653585		0.041940934	
0.091589		2.998087545		0.038989508	
		2.682853926		0.033389169	
		2.61806208		0.031998027	
		2.524863158		0.029211606	
		2.319496204		0.02531948	
		2.302184223			
		2.244728484			
		2.228883327			

Outliers in X:

Outliers in x can be found by finding out the hat matrix or hii value. Any value larger than 2*(k+1)/n are considered outliers in X . here the values above 0.069264 are considered outliers in X . some of the values are listed above.

Outliers in Y:

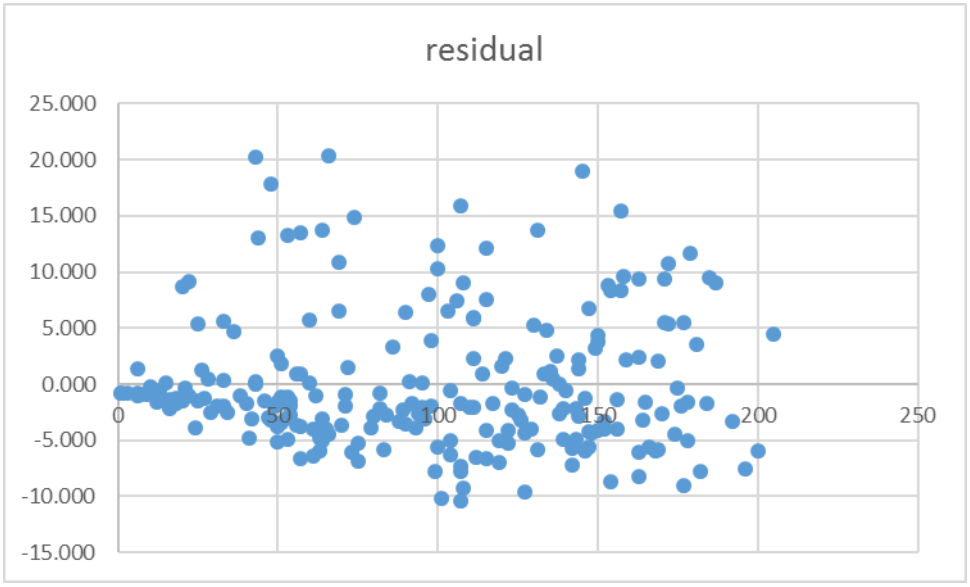
Outliers in Y can be found by finding out the studentized deleted residual values. Any value larger than t-table value with n-k-2 df are considered outliers in Y . here the values above 1.970707 are considered outliers in Y . some of the values are listed above.

Influential points:

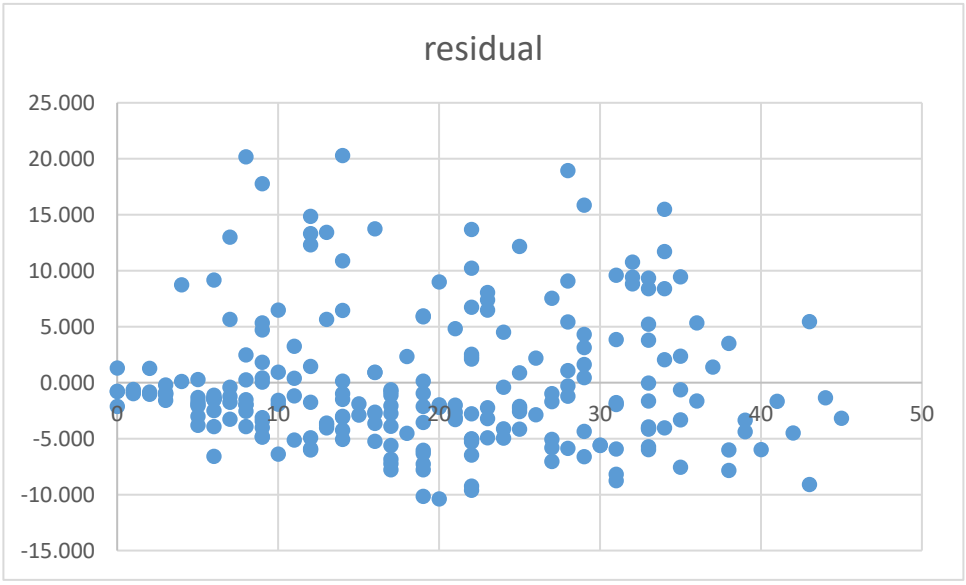
Values above the fiftieth percentile of an F distribution with k+1 and n-k-1 degrees of freedom are considered influential. Here our cook's distance value is 0.920787 and we don't have any points greater than this value. Hence we don't have influential points. Some of the values listed above are all less than 0.920787.

2. To check for assumption violations:

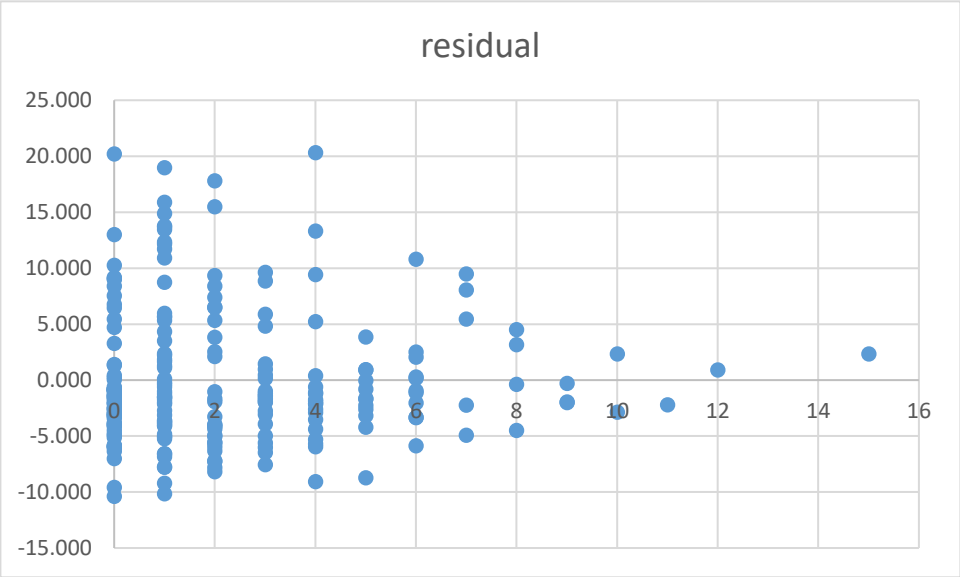
a. H vs salary



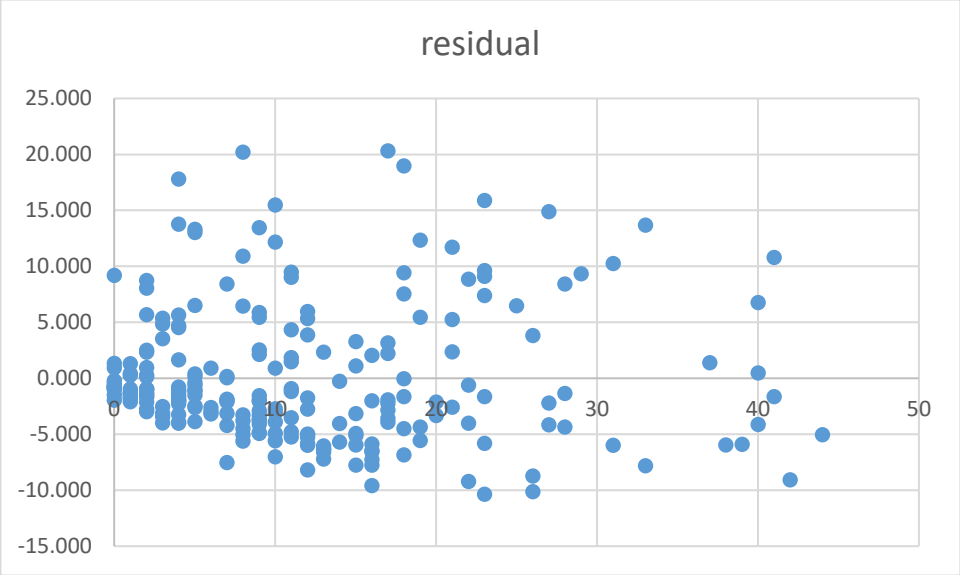
2B vs salary



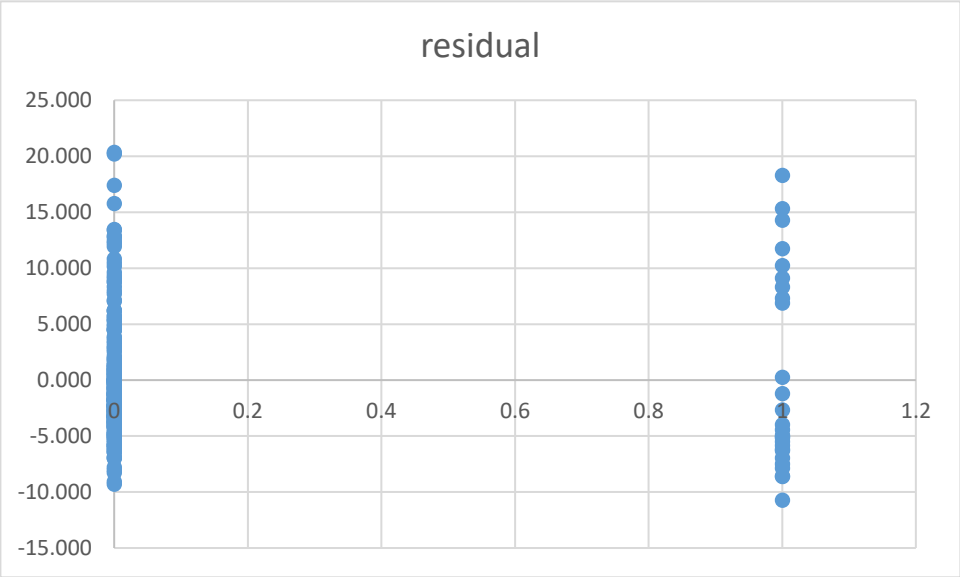
3B vs salary



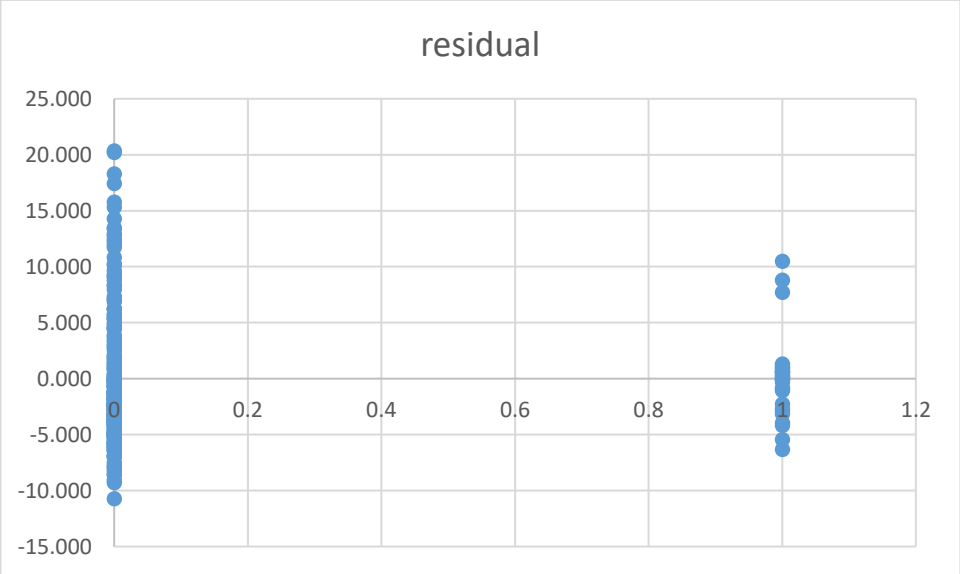
HR vs salary



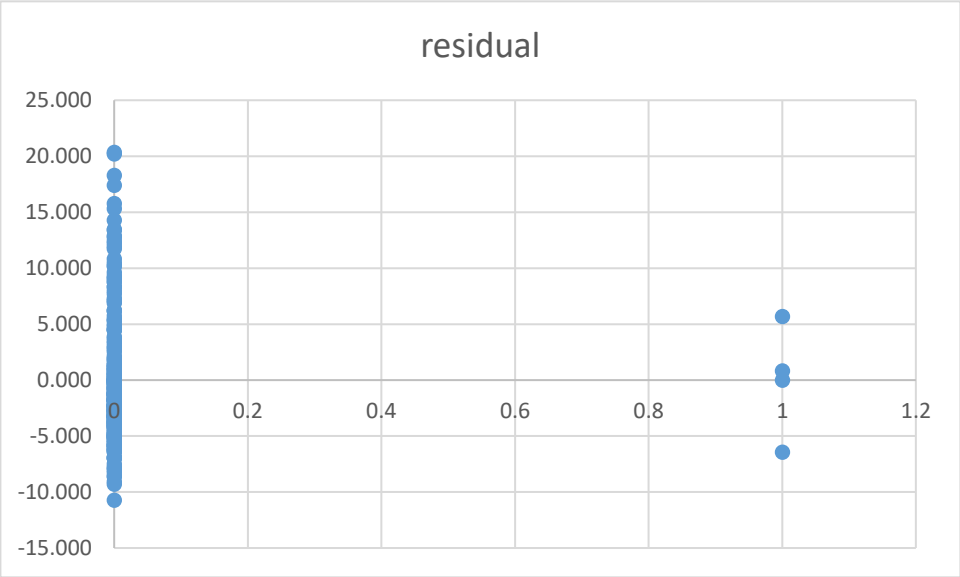
DV1B vs Salary



DVSS Vs Salary



DVDH Vs Salary



The independent variables H,2B,3B,HR,DV1B,DVSS and DVDH have been used to determine the assumption violations. The 4 assumptions are Linearity, Independence, Normality and Equal Variance. The intercepts and slopes of independent variables were used to get the value of y^ which in turn was used to get the value of y-y^.

These residual values were plotted against each independent variables to test the assumption violations of linearity and equal variance.

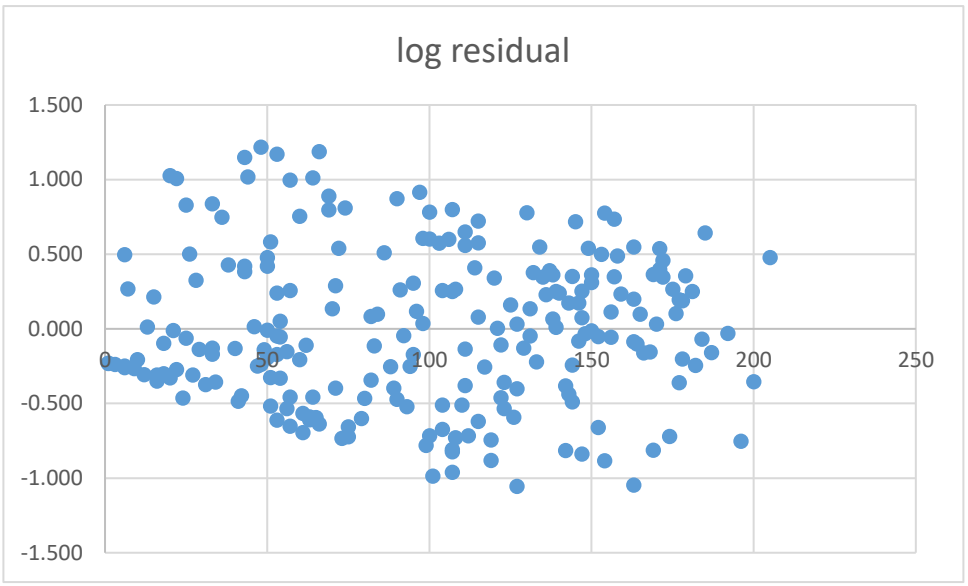
Independence cannot be tested as we don't have a time period data. Also normality is fine because the sample size is greater than 30 although the graph below shows a normality violation. We will now look at the 4 graphs above and check for linearity and equal variance assumption violation alone.

H,2B,HR graphs does not show any assumption violations. They show a random scattering of points and does not follow a definite pattern . hence there are no linearity and variance violation .

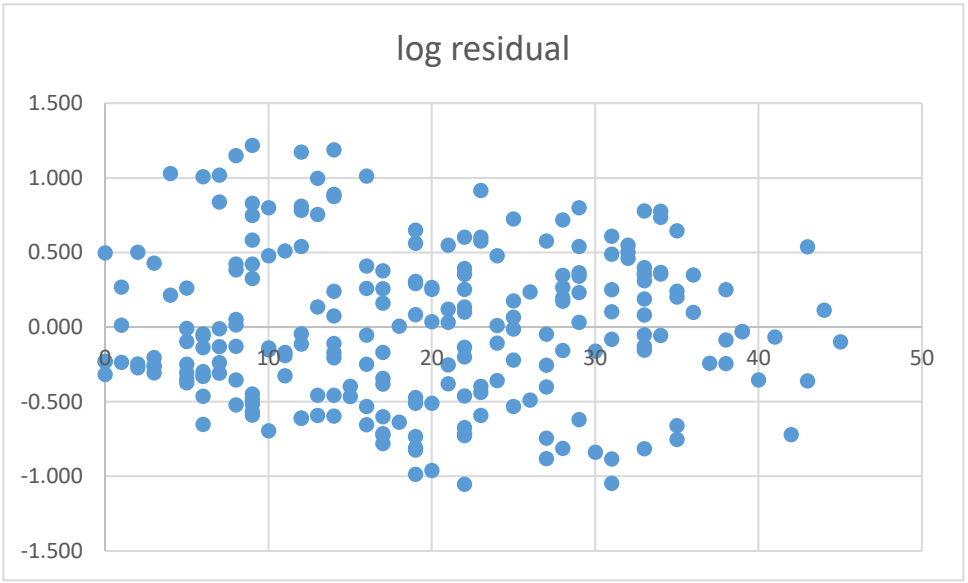
Dummy variable plots shows unequal variability. Also the plot of 3B vs residual shows a violation of equal variance. The spread of the residuals fall from larger values to smaller values. Hence we need to transform the dependent variable Salary. Our transformation includes transforming the dependent variable salary using logarithmic function.

The dependent variable salary is transformed to give Log(salary) and its residual values are plotted against all 7 independent variables. The graphs are listed below.

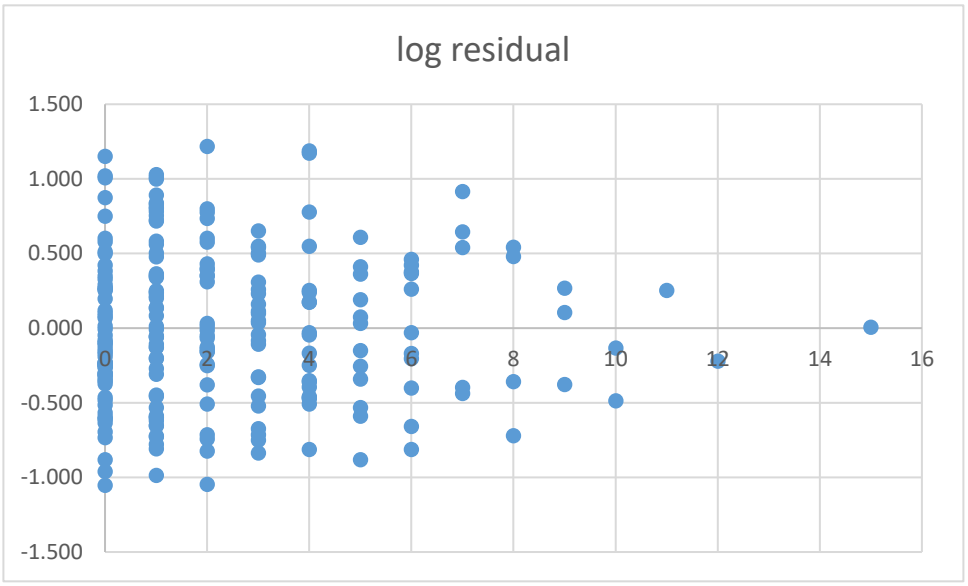
H vs Residual



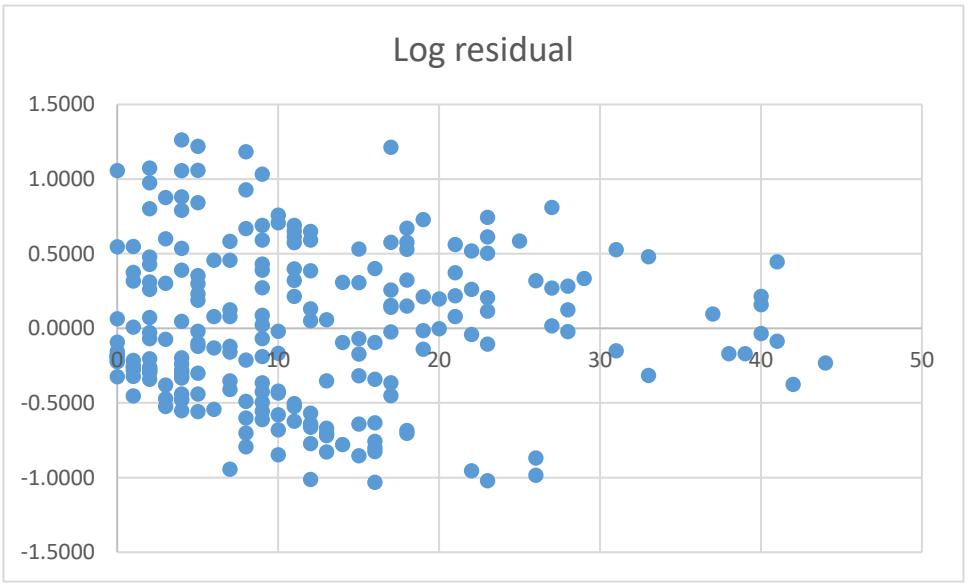
2B vs Residual:



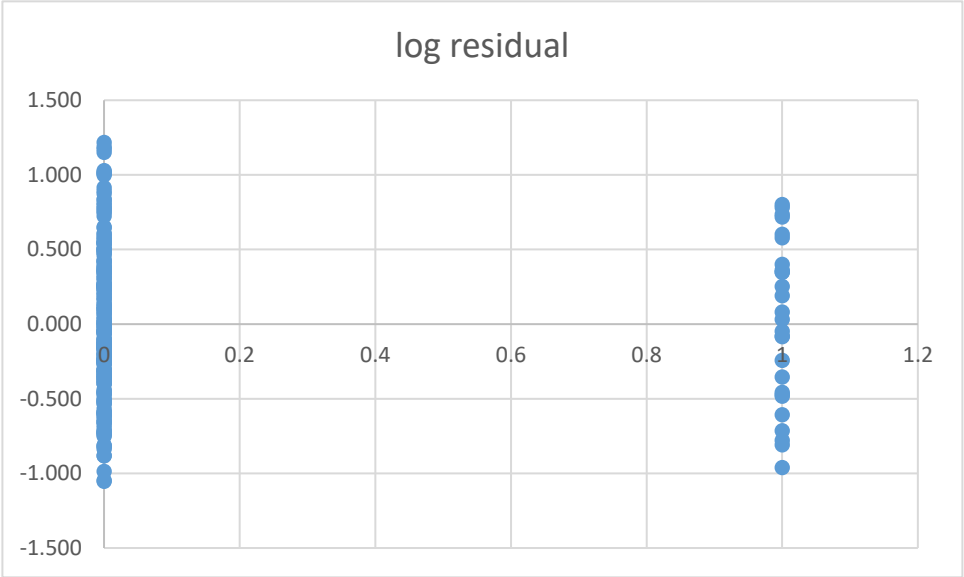
3B vs Residual:



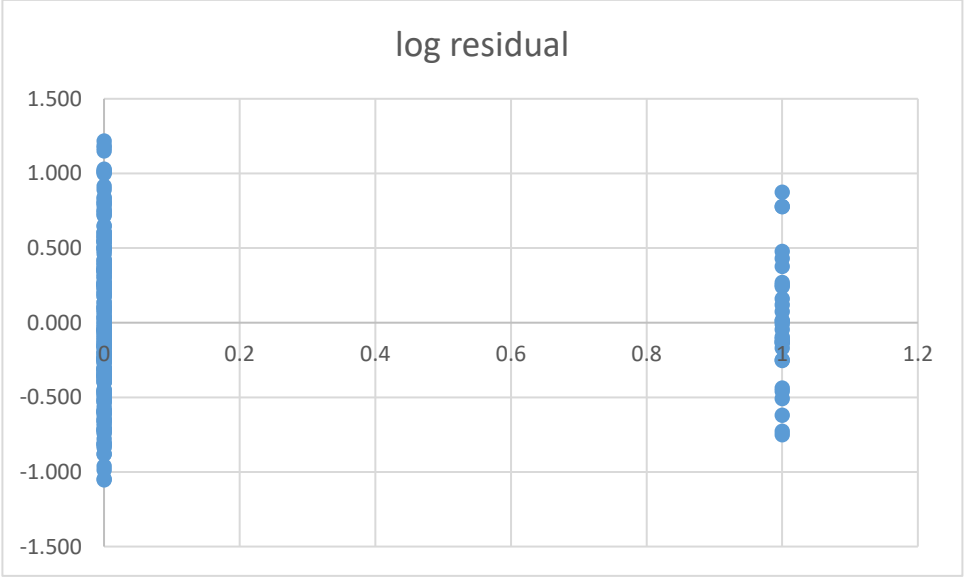
HR vs Residual:



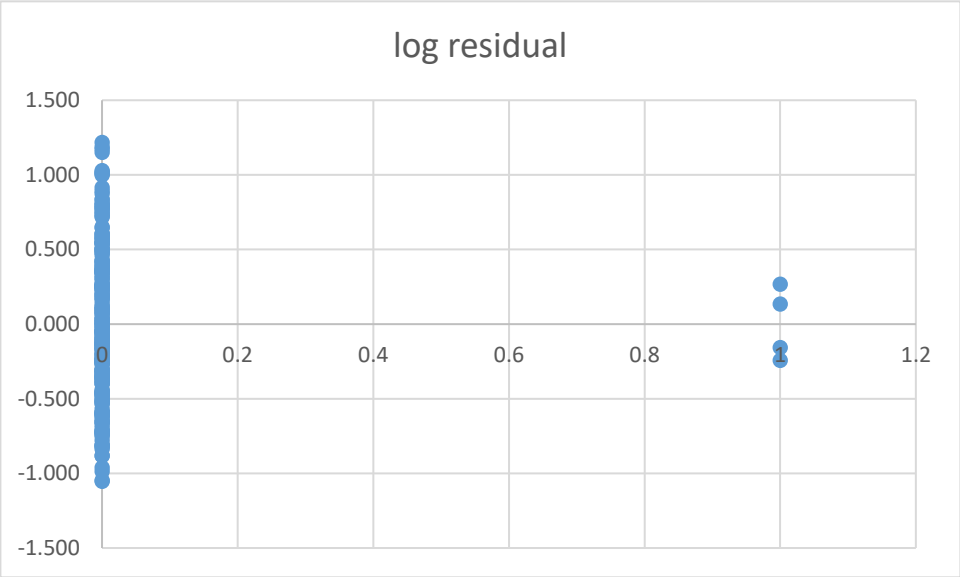
DV1b vs Residual



DVSS Vs Residual



DVDH Vs Residual



The assumptions are still the same even after transformation. All the independent variables except 3B and dummy variables holds good. 3B and dummy variables still has violation of equal variance . hence we revert back to the old dependent variable values. The normality seems to have been corrected a bit but it still has some values that does not follow a straight line .

The assumption transformations did not help as it still had variance violation .hence we revert back to the old data.

3.Model Selection using R- Square:

Using all-possible regression, reduce the list of possible models down to three, explaining your reasoning for choosing those three.

Model	Number in Model	R-Square	Variables in Model
1	**1**	0.233	HR
9	**2**	0.2582	3B HR
29	**3**	0.2934	H 3B HR

Listed above are the 3 models chosen based on their R - square values. Detailed list is given below in Appendix. We will now choose one among these 3.

Examining the three models in detail:

Model	Variables	R Square	Adjusted R Square	F test value	Root MSE
HR	1	0.233	0.2297	69.57	6.2963
3B,HR	2	0.2582	0.2517	39.68	6.2056
H,3B,HR	3	0.2934	0.2841	31.42	6.0699

The above values are the summary of the 3 models we have chosen. Some of the criterion used to predict the right model are also given above. Lets discuss each one now.

R – Square: The R – Square values tends to increase as each variable increase . but it necessarily doesn't predict the right model . we can use the adjusted R square value to determine the best model.

Adjusted R square: This value explains the model that contain different number of predictors. The value has increased with each increase in variable. Hence we can choose H,3B,HR model as of now.

F Test value: A significant F test value determines whether a relationship between the dependent variable salary and independent variables are statistically reliable. Although it is greater for 1st model we would choose the 3rd model because of adjusted R square values.

Root MSE: RMSE is the square root of the variance of the residuals. It is a good measure of how accurately it predicts the response variable 'salary'. Lower values of RMSE indicates better accuracy. The 3rd model which we had chosen earlier suffices this theory because of lower RMSE values than the other 2 models.

Hence we have chosen H,3B,HR model based on the reasons cited above.

Least Squares Line equation of H,3B,HR model:

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k$

$\hat{Y}$ = 1.15005 + 0.03800(x1) - 0.73781(x2) + 0.23554(x3)

Using the least squares equation found above from the training set, we can predict the values of the dependent variable in the validation data set. Average the absolute differences between the actual value and the predicted value.

Sample Average	4.105053
Sample Standard Dev(s)	3.771069

n = 58

standard error , $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

= 0.495166

Degree of freedom = n-k-1

= 58 – 3 – 1

Df = 54

Margin of error = t value * std error

= 2.0049*0.495166

= 0.992758

Average salary for all players = Sample mean ± M.O.E

= 4.105053 ± 0.992758

With 95% confidence, the average salary of all players in the future using H,3B,HR model will be between 3.112295 and 5.097811.

4.Using the validation data set, let us see if the model found in 4 is useful.

H0: H=3B=HR(no variable has an effect)

H1: At least one has an effect on salary

The F Test statistic value of this model is 7.20(from SAS)

F critical value is the value of F(3,54) = 2.77

Reject H0 If the F test value > F critical Value

Conclusion : With this F test, we can say that changes in values of at least one independent variable H,3B or HR is associated with changes in the average value of the dependent variable ‘Salary’.

`

Appendix:

2 . Before Transformation:

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	3991.38674	570.19811	16.21	<.0001
Error	223	7844.98444	35.17930		
Corrected Total	230	11836			

Root MSE	5.93121	R-Square	0.3372
Dependent Mean	5.88551	Adj R-Sq	0.3164
Coeff Var	100.77662		

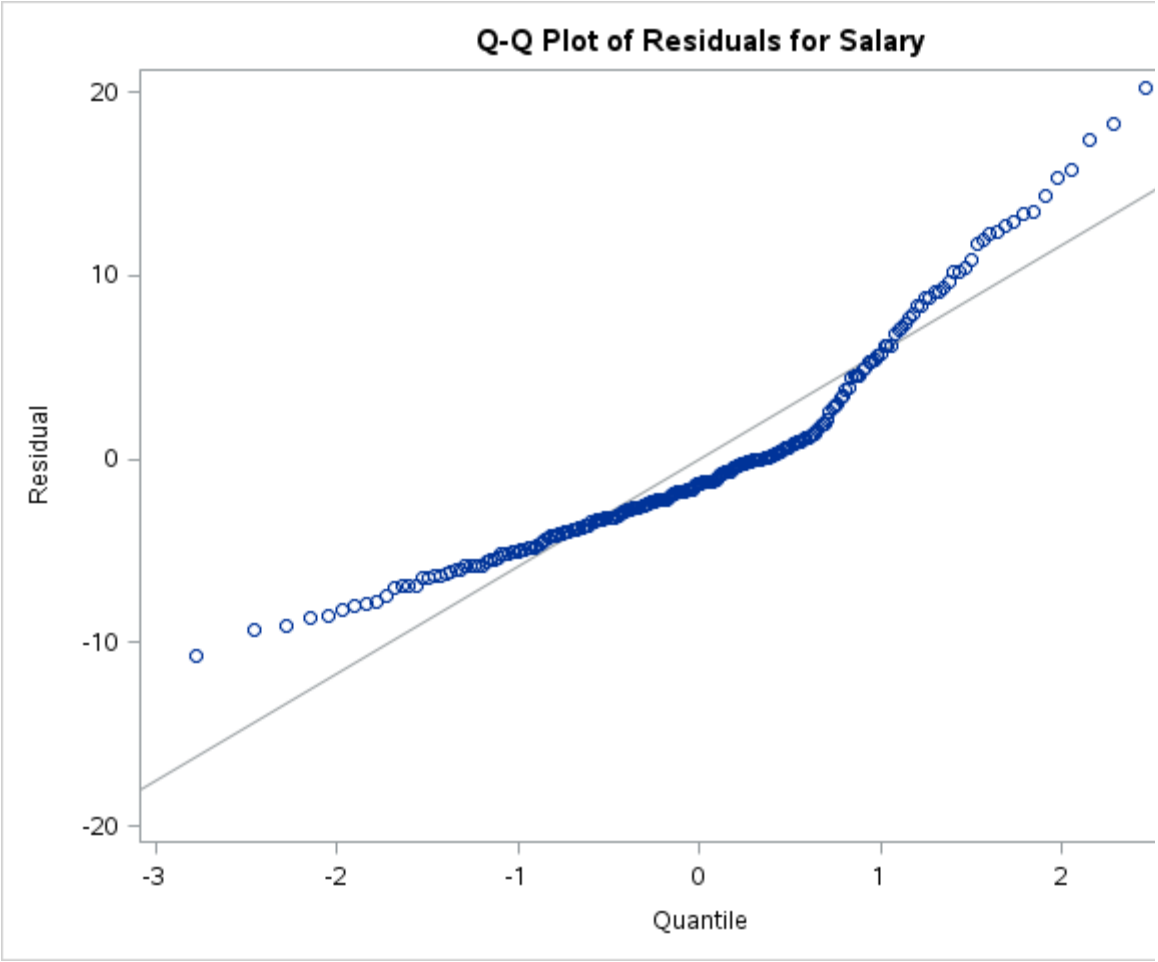
Parameter Estimates

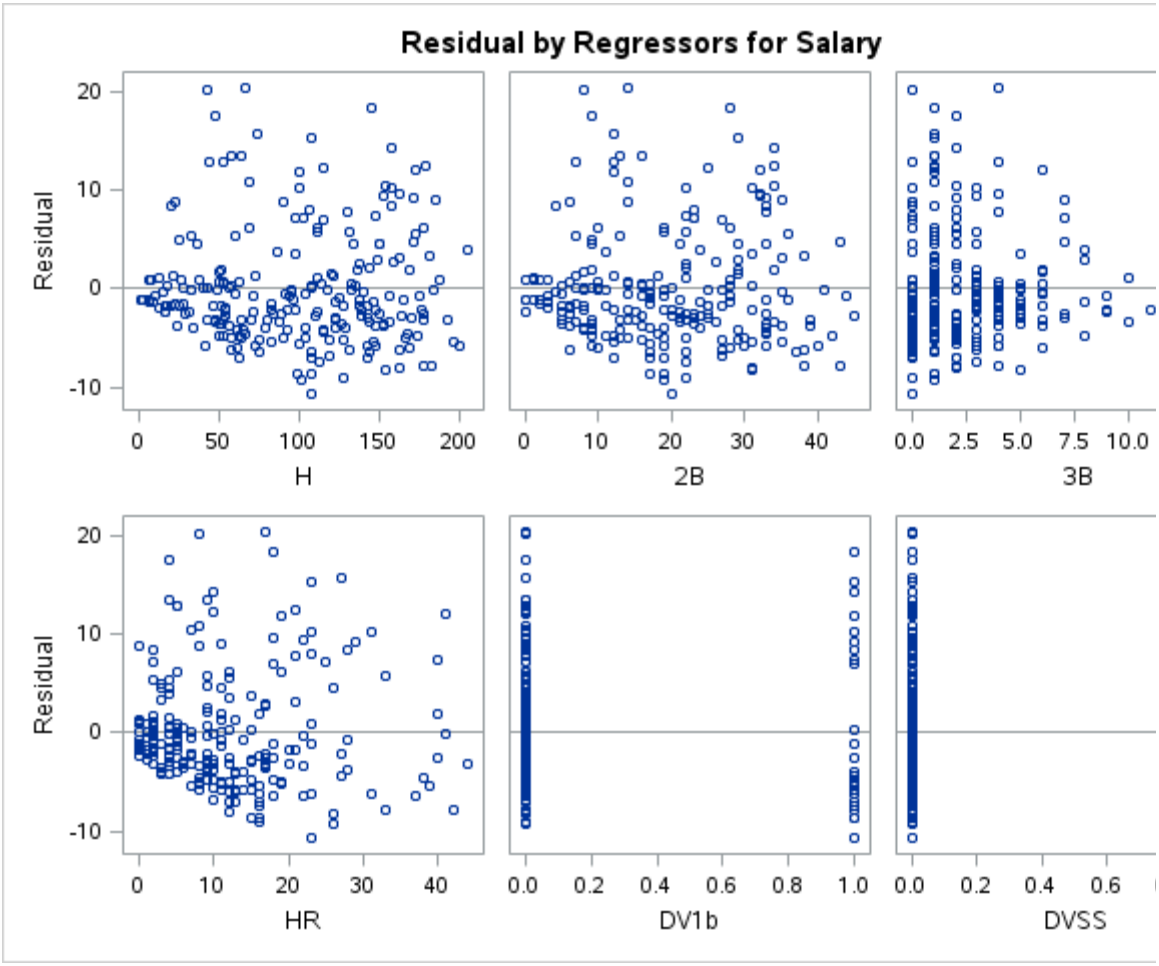
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.65079	0.85334	1.93	0.0543	0
H	1	0.05537	0.01970	2.81	0.0054	7.02137
2B	1	-0.10344	0.08981	-1.15	0.2506	6.36406
3B	1	-0.64777	0.17900	-3.62	0.0004	1.42474
HR	1	0.19958	0.05639	3.54	0.0005	2.05027
DV1b	1	1.18632	1.32395	0.90	0.3712	1.18807
DVSS	1	-2.23150	1.18868	-1.88	0.0618	1.04846
DVDH	1	9.26848	3.11914	2.97	0.0033	1.08707

Generated by the SAS System ('SASApp', Linux) on November 26, 2016 at 6:25:57 PM

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary


Q-Q Plot of Residuals for Salary
Residual
20
10
0
-10
-20
-3
-2
-1
0
1
2
Quantile


Residual by Regressors for Salary
Residual
20
10
0
-10
0
50
100
150
200
H
0
10
20
30
40
2B
0.0
2.5
5.0
7.5
10.0
3B
Residual
20
10
0
-10
0
10
20
30
40
HR
0.0
0.2
0.4
0.6
0.8
1.0
DV1b
0.0
0.2
0.4
0.6
DVSS


Residual by Regressors for Salary
Residual
20
10
0
-10
0.0
0.2
0.4
0.6
0.8
1.0
DVDH

2. After transformation:

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Log(salary)

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	30.40674	4.34382	16.39	<.0001
Error	223	59.09055	0.26498		
Corrected Total	230	89.49728			

Root MSE	0.51476	R-Square	0.3398
Dependent Mean	0.37717	Adj R-Sq	0.3190
Coeff Var	136.47920		

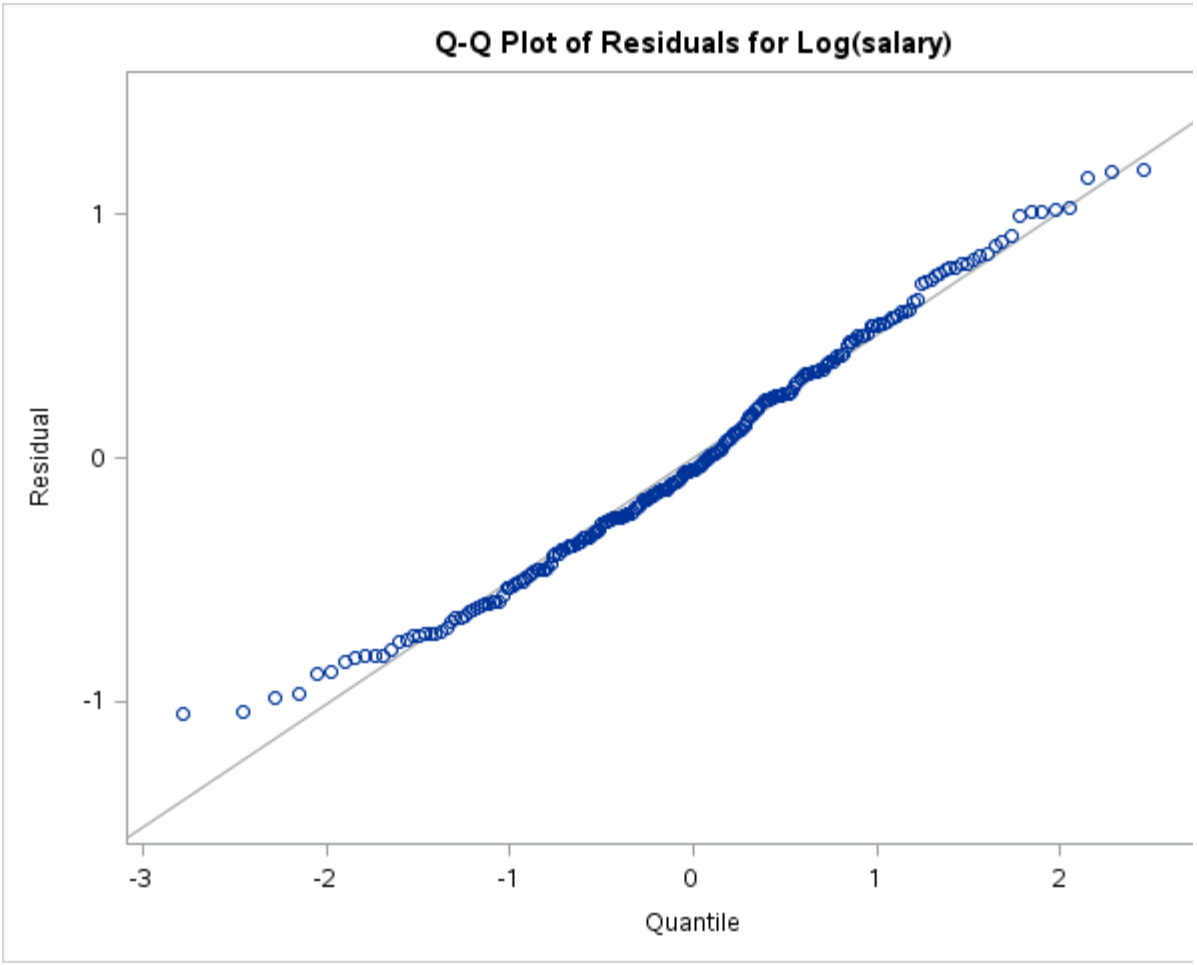
Parameter Estimates

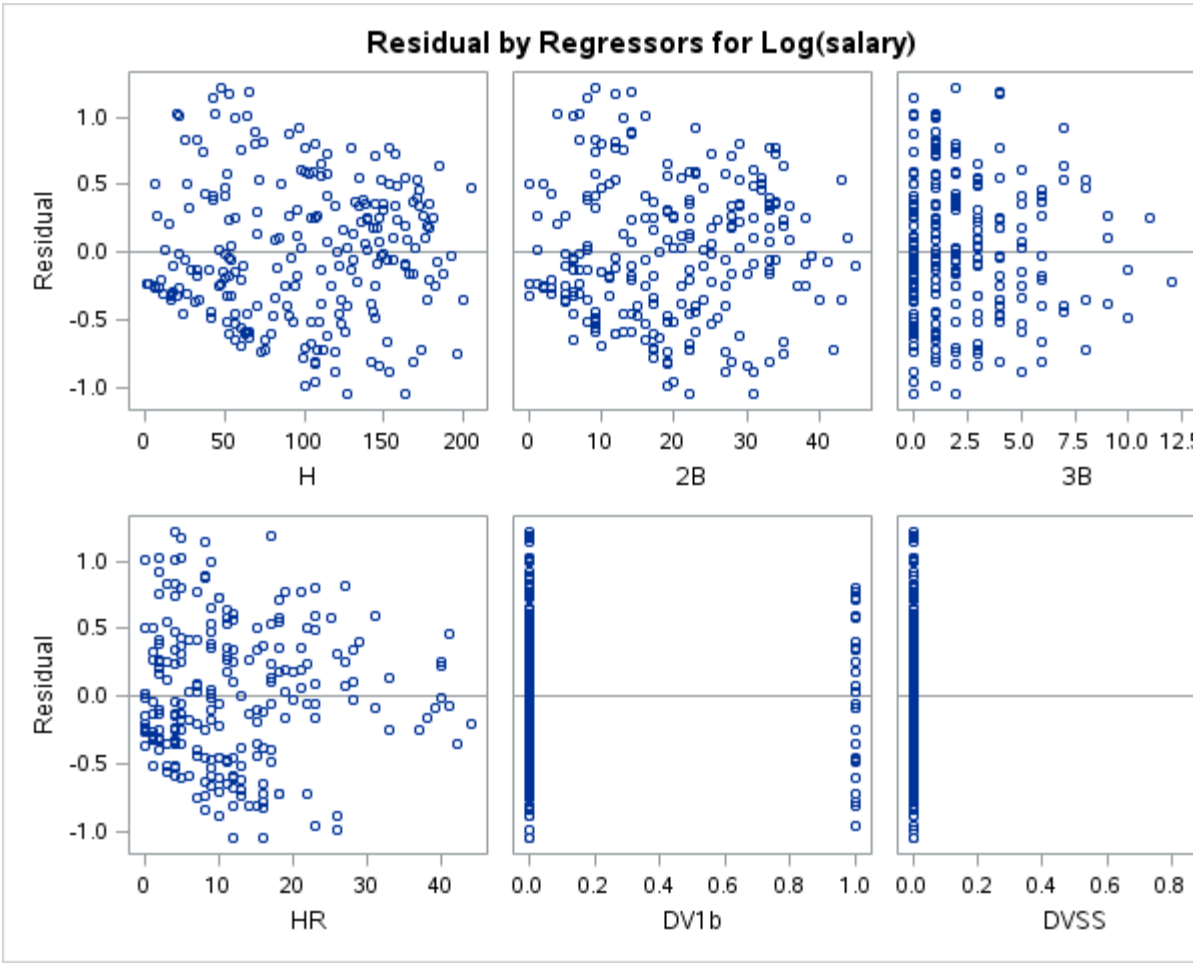
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.06926	0.07406	-0.94	0.3507	0
H	1	0.00556	0.00171	3.25	0.0013	7.02137
2B	1	-0.00313	0.00779	-0.40	0.6884	6.36406
3B	1	-0.06572	0.01554	-4.23	<.0001	1.42474
HR	1	0.01291	0.00489	2.64	0.0089	2.05027
DV1b	1	-0.06389	0.11490	-0.56	0.5787	1.18807
DVSS	1	-0.23941	0.10316	-2.32	0.0212	1.04846
DVDH	1	0.35362	0.27071	1.31	0.1928	1.08707

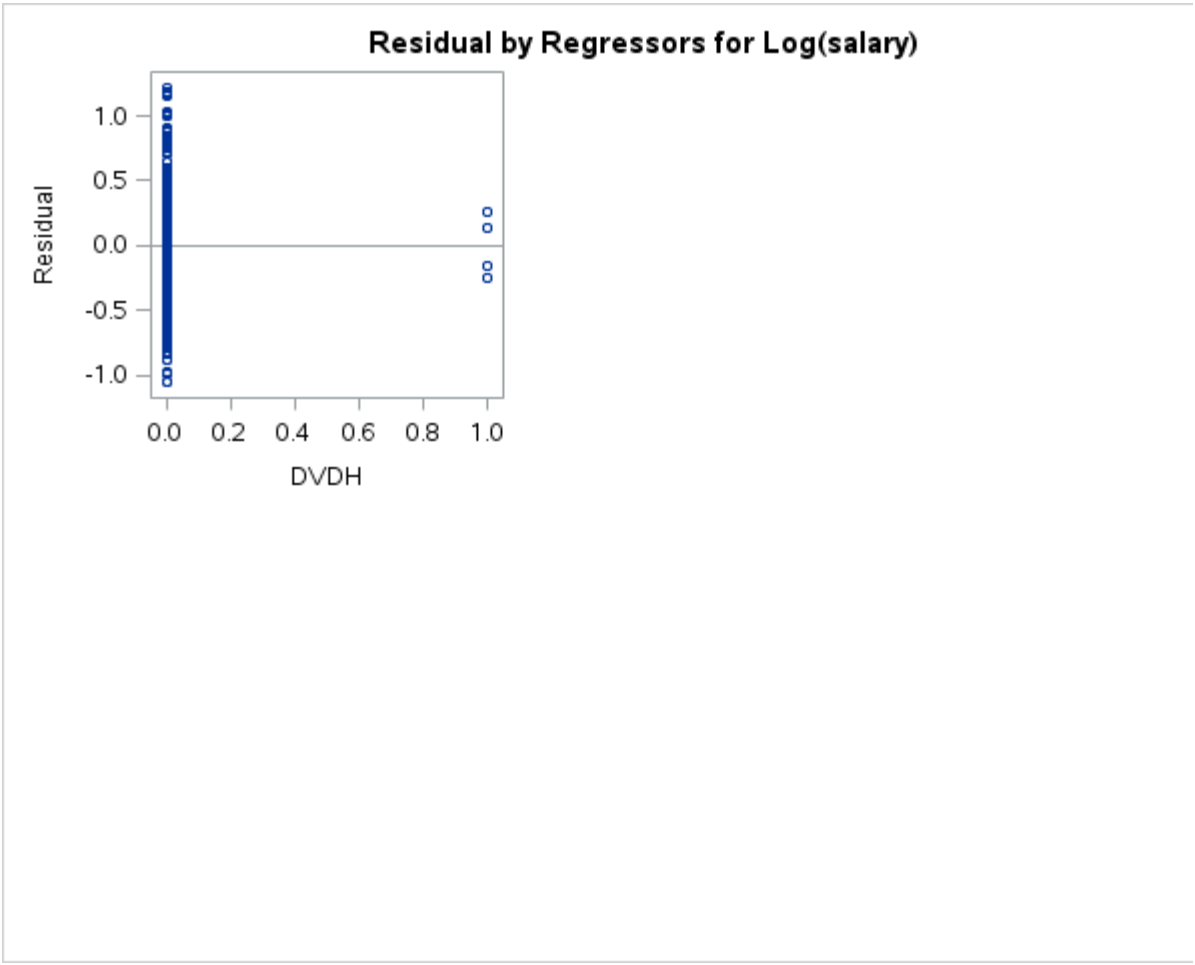
Generated by the SAS System ('SASApp', Linux) on November 26, 2016 at 7:08:38 PM

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Log(salary)


Q-Q Plot of Residuals for Log(salary)
Residual
1
0
-1
-3
-2
-1
0
1
2
Quantile


Residual by Regressors for Log(salary)
Residual
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
200
H
0
10
20
30
40
2B
0.0
2.5
5.0
7.5
10.0
3B
Residual
HR
0.0
0.2
0.4
0.6
0.8
1.0
DV1b
DVSS


Residual by Regressors for Log(salary)
Residual
1.0
0.5
0.0
-0.5
-1.0
0.0
0.2
0.4
0.6
0.8
1.0
DVDH

3. Model Selection

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary

R-Square Selection Method

Number of Observations Read	231
Number of Observations Used	231

Model Index	Number in Model	R-Square	Variables in Model
1	1	0.2330	HR
2	1	0.1349	H
3	1	0.1159	2B
4	1	0.0793	DVDH
5	1	0.0468	DV1b
6	1	0.0365	DVSS
7	1	0.0158	3B
8	2	0.2702	HR DVDH
9	2	0.2582	3B HR
10	2	0.2438	HR DVSS
11	2	0.2398	H HR
12	2	0.2371	HR DV1b
13	2	0.2348	2B HR
14	2	0.2344	H 3B
15	2	0.1931	H DVDH
16	2	0.1845	2B 3B
17	2	0.1792	2B DVDH
18	2	0.1627	H DVSS
19	2	0.1612	H DV1b
20	2	0.1418	2B DVSS
21	2	0.1404	2B DV1b
22	2	0.1350	H 2B
23	2	0.1323	DV1b DVDH
24	2	0.1105	DVSS DVDH
25	2	0.0885	3B DVDH
26	2	0.0731	DV1b DVSS
27	2	0.0558	3B DV1b
28	2	0.0509	3B DVSS
29	3	0.2934	H 3B HR
30	3	0.2888	3B HR DVDH

31	3	0.2804HR DVSS DVDH
32	3	0.2779HR DV1b DVDH
33	3	0.2773H HR DVDH
34	3	0.27272B HR DVDH
35	3	0.27162B 3B HR
36	3	0.2701H 3B DVDH
37	3	0.26773B HR DVSS
38	3	0.25953B HR DV1b
39	3	0.2561H 3B DVSS
40	3	0.2518H HR DVSS
41	3	0.2469HR DV1b DVSS
42	3	0.24592B HR DVSS
43	3	0.2447H HR DV1b
44	3	0.2429H 2B HR
45	3	0.2424H 3B DV1b
46	3	0.23912B HR DV1b
47	3	0.2350H 2B 3B
48	3	0.23072B 3B DVDH
49	3	0.2250H DV1b DVDH
50	3	0.2176H DVSS DVDH
51	3	0.20932B DV1b DVDH
52	3	0.20542B 3B DVSS
53	3	0.20182B DVSS DVDH
54	3	0.19432B 3B DV1b
55	3	0.1935H 2B DVDH
56	3	0.1827H DV1b DVSS
57	3	0.1627H 2B DVSS
58	3	0.1612H 2B DV1b
59	3	0.16062B DV1b DVSS
60	3	0.1534DV1b DVSS DVDH
61	3	0.13603B DV1b DVDH
62	3	0.11903B DVSS DVDH
63	3	0.08193B DV1b DVSS
64	4	0.3201H 3B HR DVDH
65	4	0.3049H 3B HR DVSS
66	4	0.30182B 3B HR DVDH
67	4	0.2983H 2B 3B HR
68	4	0.29793B HR DVSS DVDH
69	4	0.2944H 3B HR DV1b
70	4	0.29263B HR DV1b DVDH
71	4	0.2901H 3B DVSS DVDH
72	4	0.2886H HR DVSS DVDH
73	4	0.2866HR DV1b DVSS DVDH
74	4	0.2861H HR DV1b DVDH
75	4	0.28332B HR DVSS DVDH
76	4	0.2826H 3B DV1b DVDH

77	4	0.28162B 3B HR DVSS
78	4	0.28072B HR DV1b DVDH
79	4	0.2791H 2B HR DVDH
80	4	0.27252B 3B HR DV1b
81	4	0.2704H 2B 3B DVDH
82	4	0.26853B HR DV1b DVSS
83	4	0.2613H 3B DV1b DVSS
84	4	0.2573H 2B 3B DVSS
85	4	0.2556H HR DV1b DVSS
86	4	0.2553H 2B HR DVSS
87	4	0.24962B 3B DVSS DVDH
88	4	0.24922B HR DV1b DVSS
89	4	0.2482H 2B HR DV1b
90	4	0.24562B 3B DV1b DVDH
91	4	0.2434H 2B 3B DV1b
92	4	0.2428H DV1b DVSS DVDH
93	4	0.22592B DV1b DVSS DVDH
94	4	0.2250H 2B DV1b DVDH
95	4	0.2176H 2B DVSS DVDH
96	4	0.21232B 3B DV1b DVSS
97	4	0.1829H 2B DV1b DVSS
98	4	0.15713B DV1b DVSS DVDH
99	5	0.3311H 3B HR DVSS DVDH
100	5	0.3234H 2B 3B HR DVDH
101	5	0.3232H 3B HR DV1b DVDH
102	5	0.31142B 3B HR DVSS DVDH
103	5	0.3103H 2B 3B HR DVSS
104	5	0.3054H 3B HR DV1b DVSS
105	5	0.30502B 3B HR DV1b DVDH
106	5	0.30083B HR DV1b DVSS DVDH
107	5	0.2995H 2B 3B HR DV1b
108	5	0.2990H 3B DV1b DVSS DVDH
109	5	0.2958H HR DV1b DVSS DVDH
110	5	0.2908H 2B HR DVSS DVDH
111	5	0.2907H 2B 3B DVSS DVDH
112	5	0.28982B HR DV1b DVSS DVDH
113	5	0.2882H 2B HR DV1b DVDH
114	5	0.2832H 2B 3B DV1b DVDH
115	5	0.28212B 3B HR DV1b DVSS
116	5	0.2630H 2B 3B DV1b DVSS
117	5	0.26092B 3B DV1b DVSS DVDH
118	5	0.2594H 2B HR DV1b DVSS
119	5	0.2429H 2B DV1b DVSS DVDH
120	6	0.3348H 2B 3B HR DVSS DVDH
121	6	0.3333H 3B HR DV1b DVSS DVDH
122	6	0.3267H 2B 3B HR DV1b DVDH

123	6	0.31372B 3B HR DV1b DVSS DVDH
124	6	0.3110H 2B 3B HR DV1b DVSS
125	6	0.3000H 2B 3B DV1b DVSS DVDH
126	6	0.2983H 2B HR DV1b DVSS DVDH
127	7	0.3372H 2B 3B HR DV1b DVSS DVDH

HR Model Results

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2757.91800	2757.91800	69.57	<.0001
Error	229	9078.45318	39.64390		
Corrected Total	230	11836			

Root MSE	6.29634	R-Square	0.2330
Dependent Mean	5.88551	Adj R-Sq	0.2297
Coeff Var	106.98043		

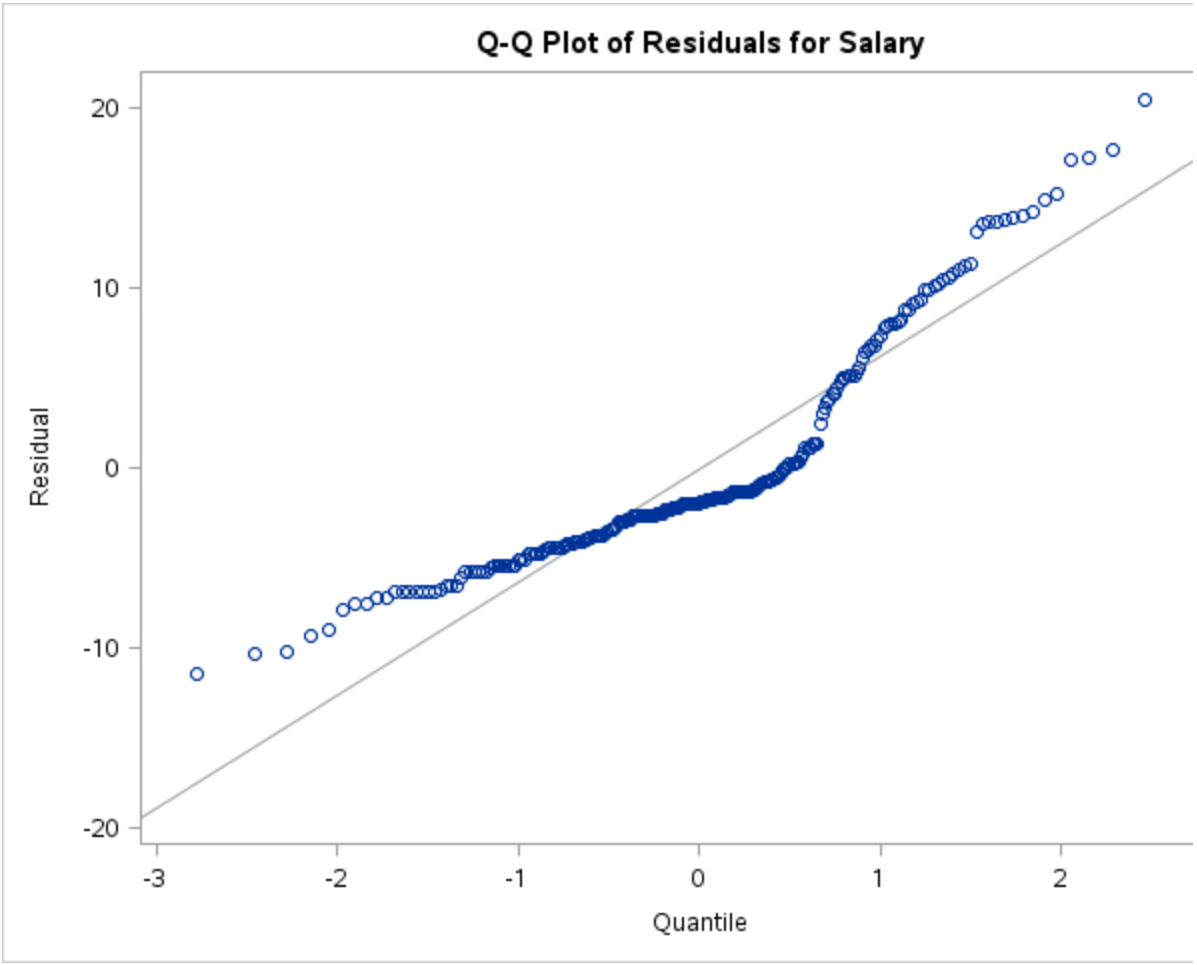
Parameter Estimates

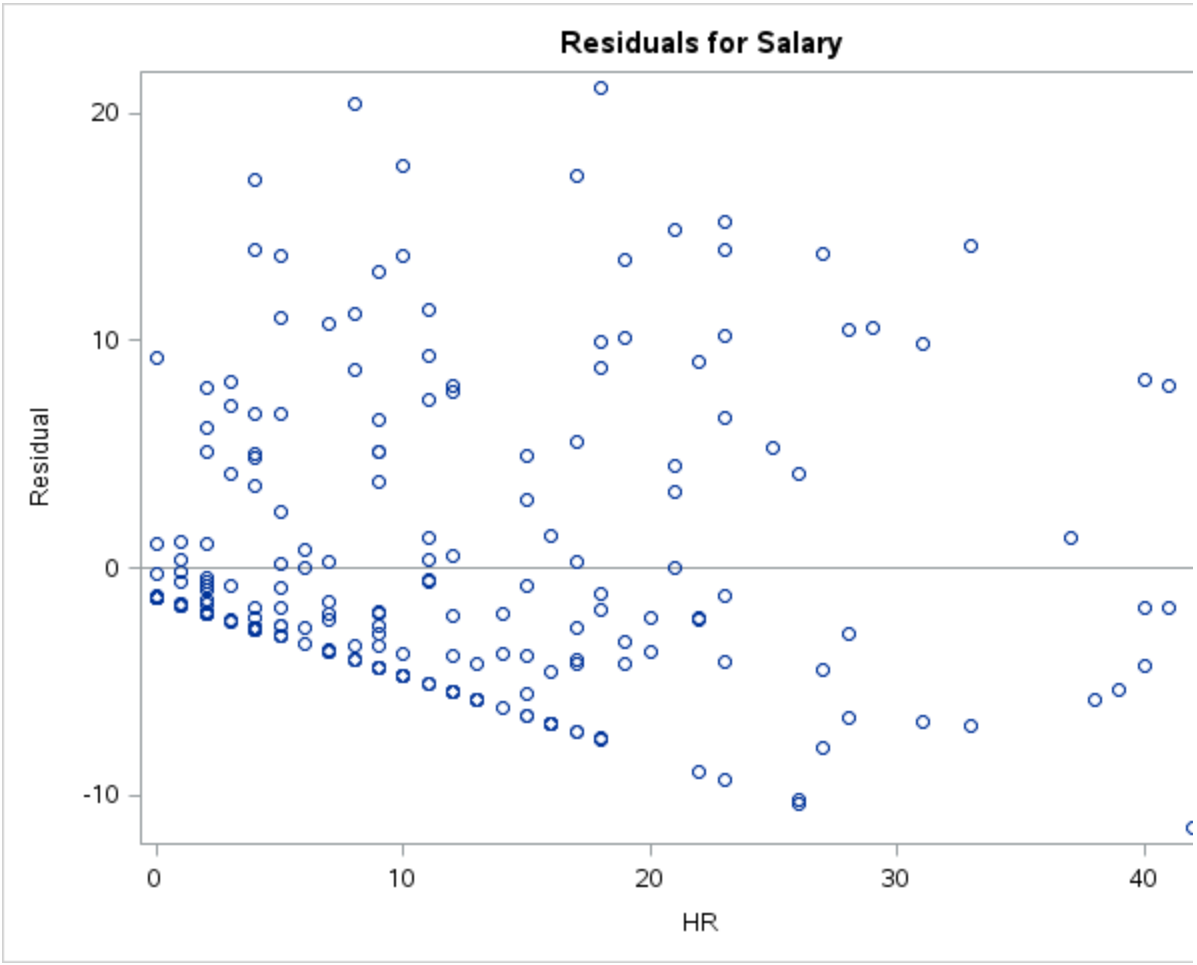
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.79018	0.64242	2.79	0.0058	0
HR	1	0.34870	0.04181	8.34	<.0001	1.00000

Generated by the SAS System ('SASApp', Linux) on November 25, 2016 at 6:25:08 PM

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary


Q-Q Plot of Residuals for Salary
Residual
20
10
0
-10
-20
-3
-2
-1
0
1
2
Quantile



Generated by the SAS System ('SASApp', Linux) on November 25, 2016 at 6:25:08 PM

- **3B,HR model Result**

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3056.15472	1528.07736	39.68	<.0001
Error	228	8780.21645	38.50972		
Corrected Total	230	11836			

Root MSE	6.20562	R-Square	0.2582
Dependent Mean	5.88551	Adj R-Sq	0.2517
Coeff Var	105.43902		

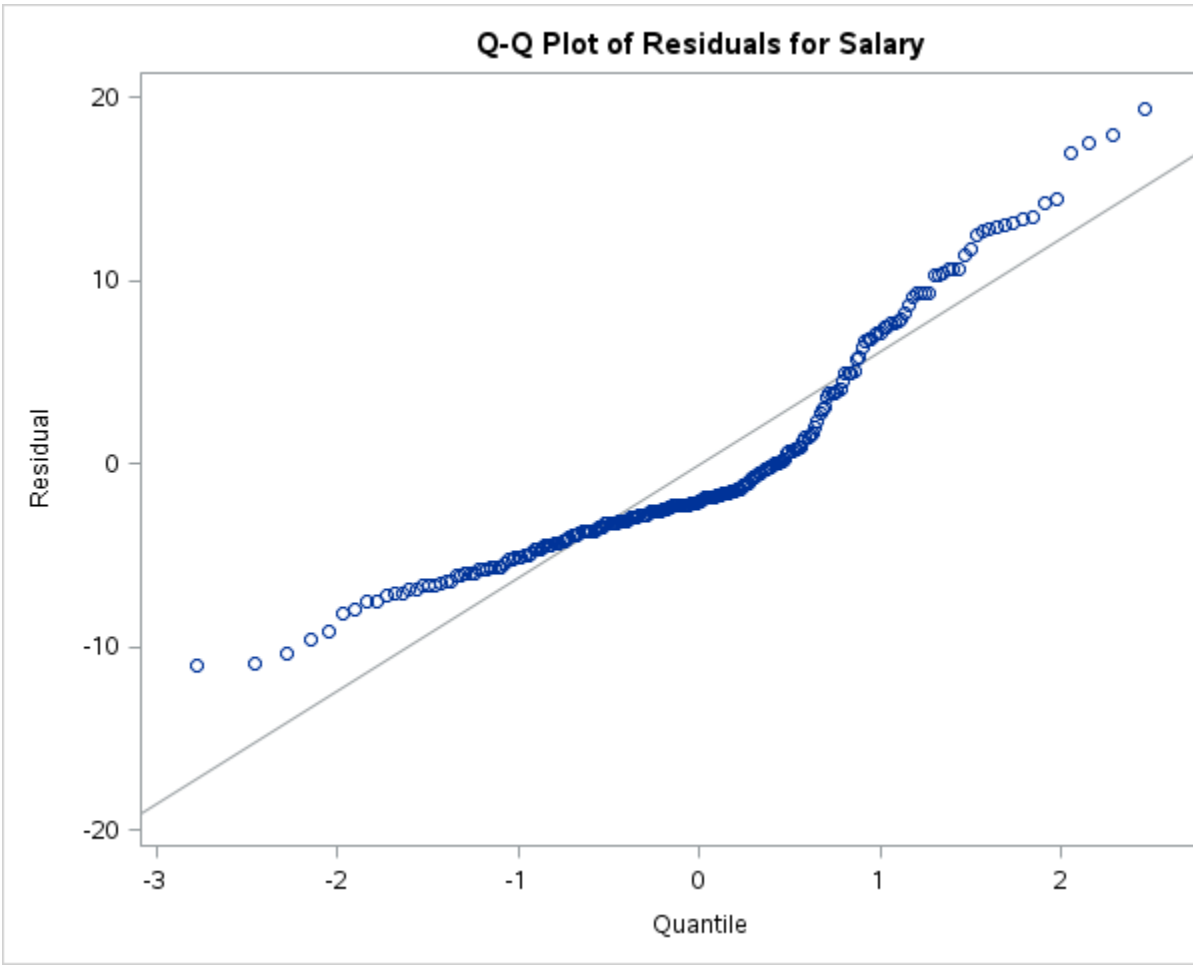
Parameter Estimates

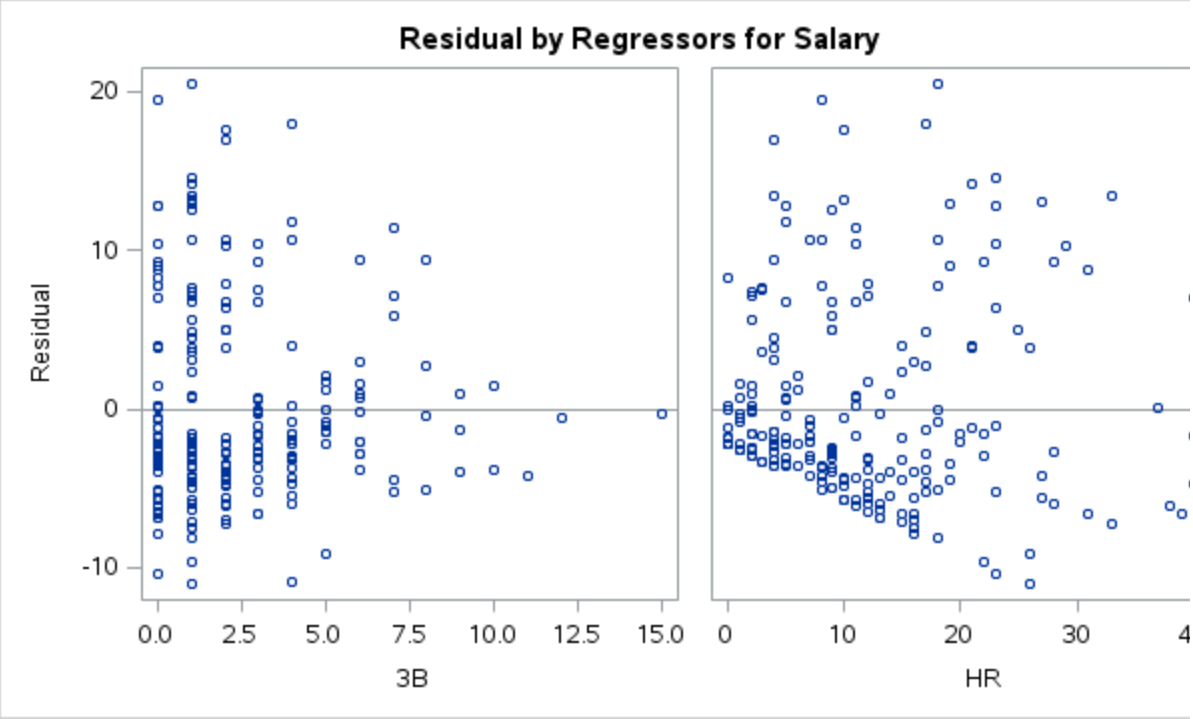
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2.72539	0.71682	3.80	0.0002	0
3B	1	-0.43765	0.15726	-2.78	0.0058	1.00463
HR	1	0.35650	0.04130	8.63	<.0001	1.00463

Generated by the SAS System ('SASApp', Linux) on November 25, 2016 at 6:30:12 PM

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary





Generated by the SAS System ('SASApp', Linux) on November 25, 2016 at 6:30:12 PM

- **H,3B,HR Model Result**

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary

Number of Observations Read	231
Number of Observations Used	231

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3472.84793	1157.61598	31.42	<.0001
Error	227	8363.52325	36.84371		
Corrected Total	230	11836			

Root MSE	6.06990	R-Square	0.2934
Dependent Mean	5.88551	Adj R-Sq	0.2841
Coeff Var	103.13305		

Parameter Estimates

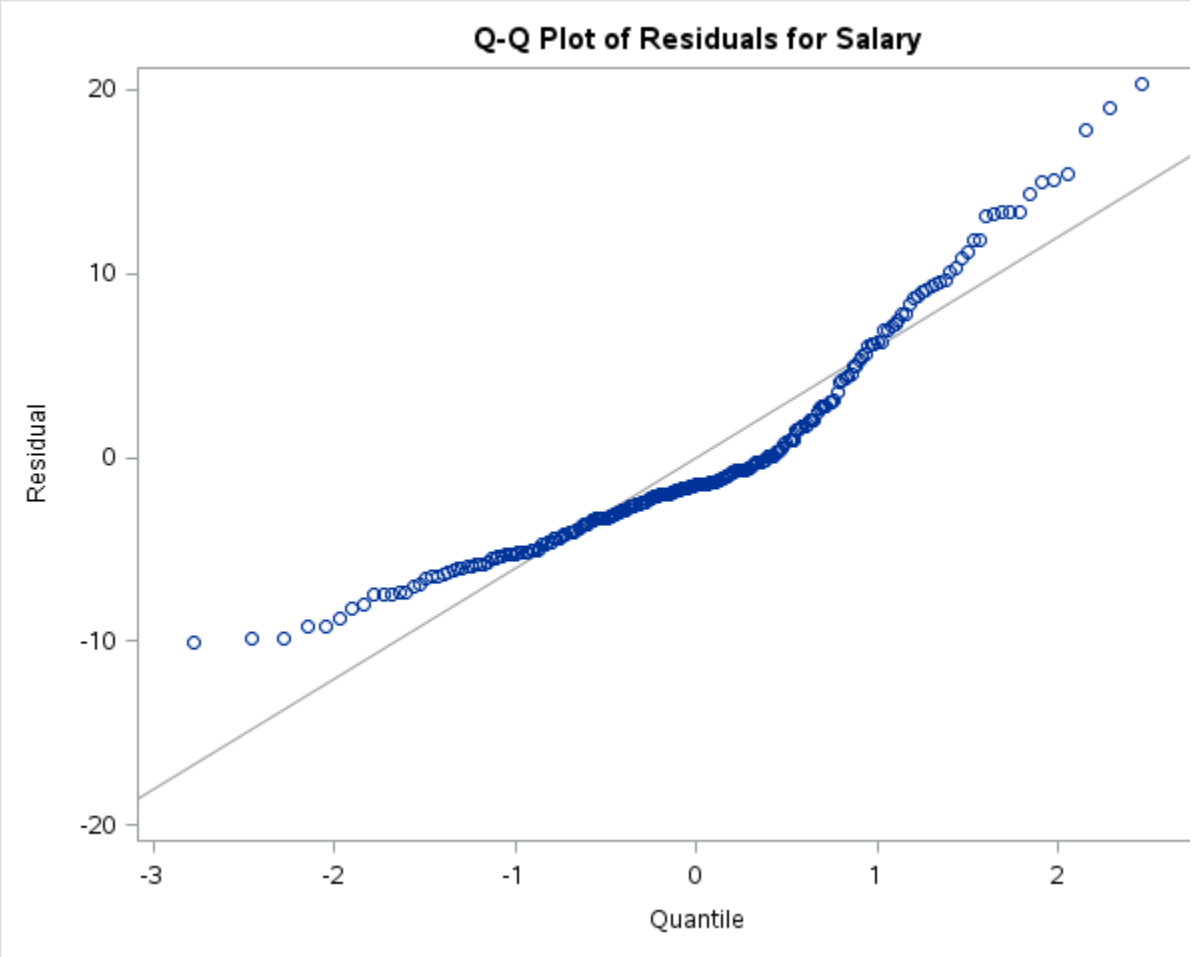
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.15005	0.84323	1.36	0.1740	0

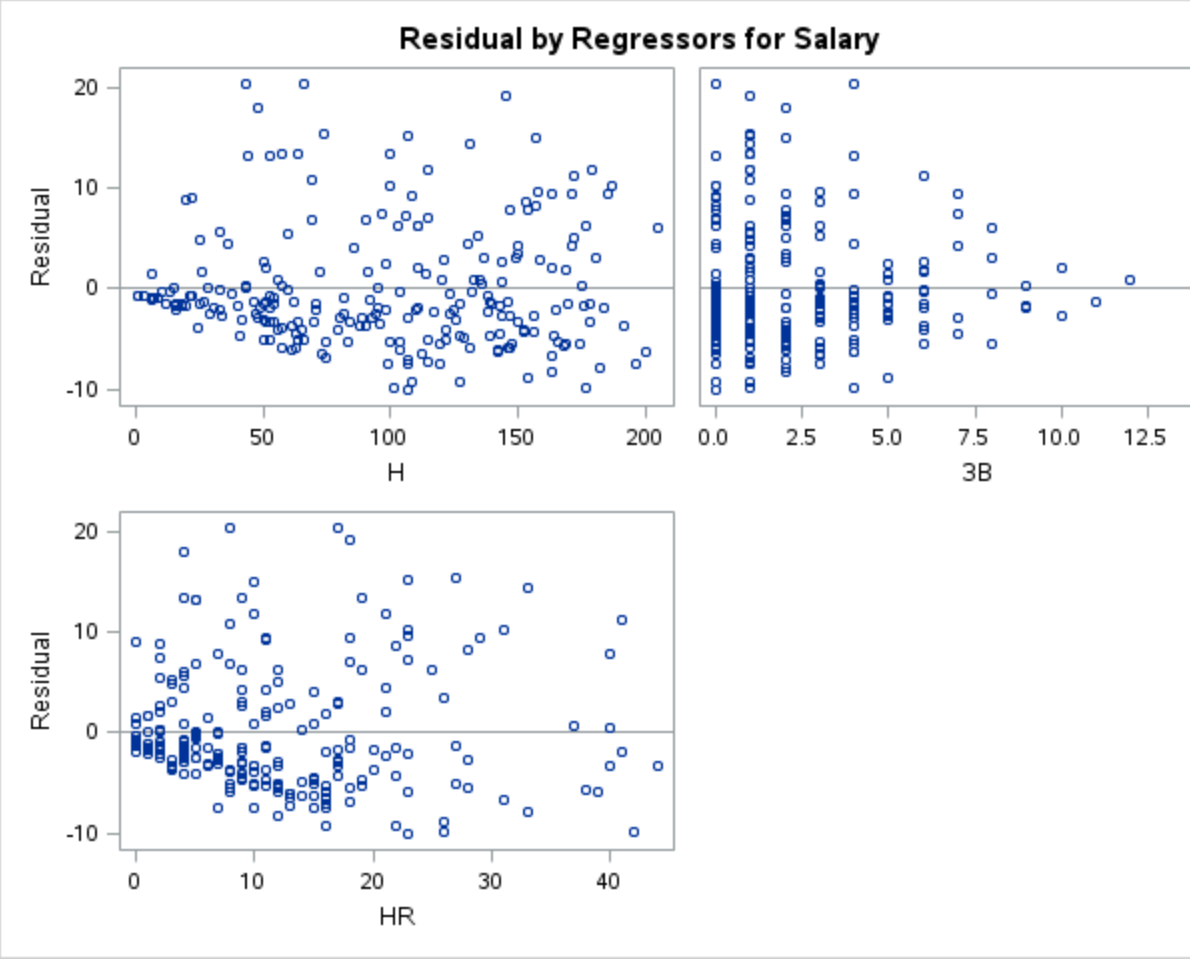
H	1	0.03800	0.01130	3.36	0.0009 2.20514
3B	1	-0.73781	0.17784	-4.15	<.0001 1.34286
HR	1	0.23554	0.05409	4.35	<.0001 1.80106

Generated by the SAS System ('SASApp', Linux) on November 25, 2016 at 6:35:52 PM

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary





Generated by the SAS System ('SASApp', Linux) on November 25, 2016 at 6:35:52 PM

Validation Data Set Results

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary

Number of Observations Read	58
Number of Observations Used	58

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	700.47065	233.49022	7.20	0.0004
Error	54	1750.18085	32.41076		
Corrected Total	57	2450.65150			

Root MSE	5.69304	R-Square	0.2858
Dependent Mean	5.71122	Adj R-Sq	0.2462
Coeff Var	99.68169		

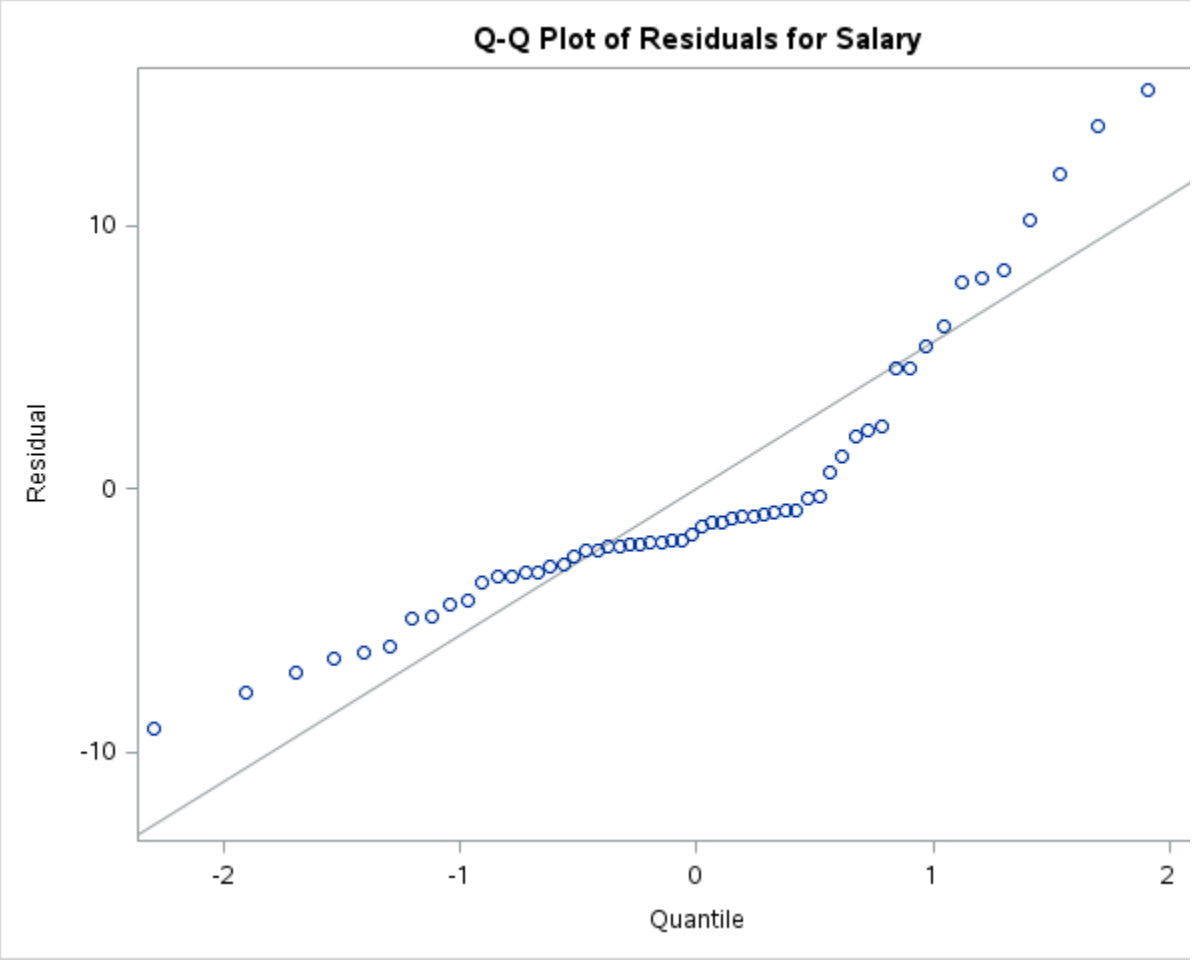
Parameter Estimates

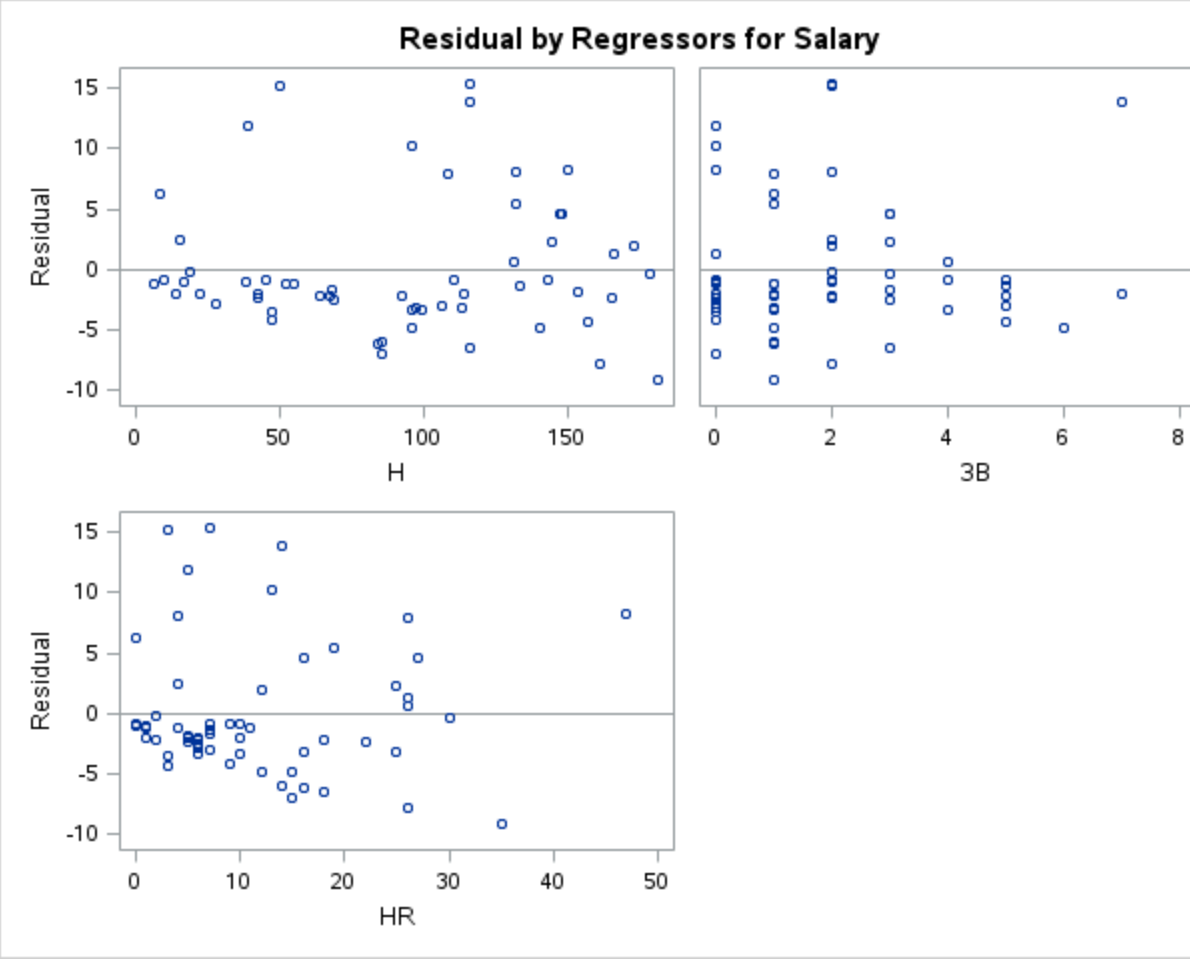
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.20109	1.55909	0.77	0.4444	0
H	1	0.05272	0.02530	2.08	0.0419	2.96193
3B	1	-0.82471	0.45653	-1.81	0.0764	1.61652
HR	1	0.11879	0.11797	1.01	0.3185	2.43935

Generated by the SAS System ('SASApp', Linux) on November 25, 2016 at 8:12:47 PM

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Salary




Residual by Regressors for Salary
Residual
15
10
5
0
-5
-10
0
50
100
150
H
0
2
4
6
8
3B
Residual
15
10
5
0
-5
-10
0
10
20
30
40
50
HR