Regression Analysis Student Project

Danish Iqbal - F 2007 RA Course
Email address: danish@alumni.uwaterloo.ca
Updated Email Address: danishmiqbal@gmail.com
Registration ID: 10093812

Objective

This project examines the correlation between several factors and the wage rate. The best fit model is determined based on highest adjusted R^2 using Ordinary Least Squares Regression with a 95% confidence interval.

Data

The data for this analysis is in appendix. This was obtained from:
http://lib.stat.cmu.edu/datasets/CPS_85_Wages

The data consist of a random sample of 534 persons from the Current Population Survey (CPS), with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership.

The Age, years of experience and numbers of years of education were grouped into buckets. The other variables were given indicators. The following is the list of indicators and groupings along with average wage rate for each split. The average wage rate is $9.02 per hour

Average Wage Rate per Hour by each Data Splits

	Indicator	Avg Wage Rate Per Hour
Management	1	$ 12.70
Professional	2	$ 11.95
Other	3	$ 8.43
Sales	4	$ 7.59
Clerical	5	$ 7.42
Service	6	$ 6.54
White	0	$ 9.28
Other	1	$ 8.06
Hispanic	2	$ 7.28
South	0	$ 9.49
North	1	$ 7.90
Female	0	$ 7.88
Male	1	$ 9.99
Manufacturing	1	$ 9.60
Construction	2	$ 9.22
Other	3	$ 8.87
UnMarried	0	$ 8.31
Marriage	1	$ 9.40
Non Union	0	$ 8.64
Union	1	$ 10.80

	Indicator	Avg Wage Rate Per Hour
Number of Yrs of Education	2	$ 5.38
	4	$ 10.00
	6	$ 5.28
	8	$ 6.58
	10	$ 6.87
	12	$ 7.79
	14	$ 10.96
	16	$ 11.53
	18	$ 13.53
Age	15	$ 6.22
	25	$ 8.77
	35	$ 10.39
	45	$ 9.53
	55	$ 9.48
Number of Yrs of Education	0	$ 6.73
	5	$ 8.21
	10	$ 10.17
	15	$ 9.63
	20	$ 9.38
	25	$ 9.50
	30	$ 10.21
	35	$ 9.43
	40	$ 9.64
	45	$ 5.91
	50	$ 6.00
	55	$ 7.00

Methodology

Step 1

The regression equation is: $Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6 + B_7X_7 + B_8X_8 + B_9X_9 + B_{10}X_{10}$

The Wage is the response variable (Y) and the other 10 columns are the explanatory variables (X), with

X_1 = Number of Yrs of Education
X_2 = Indicator variable for Location
X_3 = Indicator variable for Gender
X_4 = Number of years of work experience
X_5 = Indicator variable for union membership
X_6 = Age (years)
X_7 = Race
X_8 = Occupational category
X_9 = Sector
X_{10} = Marital Status

Additionally, A is the intercept of the regression equation and the B_i's are the regression coefficients (or slopes) for the X_i variables. The following is the Summary out for the Regression along with the Residual Plots.

~

SUMMARY OUTPUT

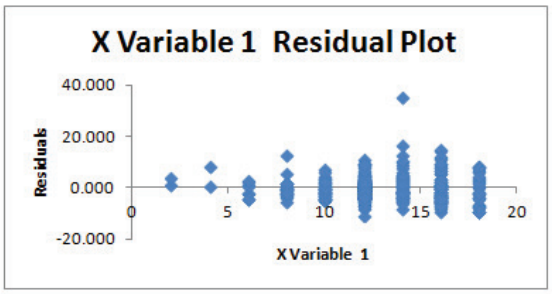
Regression Statistics	
Multiple R	0.555
R Square	0.308
Adjusted R Square	0.295
Standard Error	4.316
Observations	534

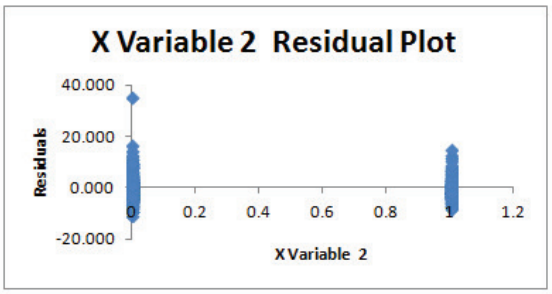
ANOVA

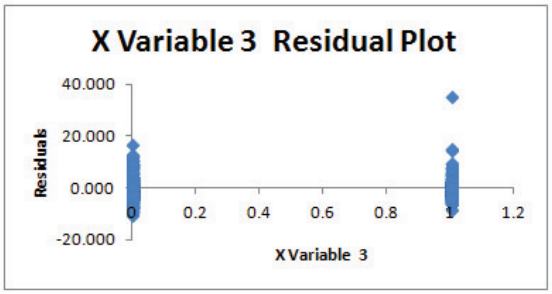
	df	*SS*	*MS*	*F*	*Significance F*
Regression	10	4336.106	433.611	23.282	0.000
Residual	523	9740.593	18.624		
Total	533	14076.699			

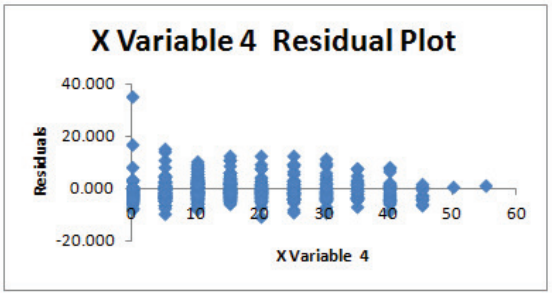
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.189	1.477	1.482	0.139	-0.713	5.090	-0.713	5.090
X Variable 1	0.742	0.099	7.499	0.000	0.548	0.936	0.548	0.936
X Variable 2	-0.601	0.422	-1.424	0.155	-1.431	0.228	-1.431	0.228
X Variable 3	-1.618	0.396	-4.082	0.000	-2.397	-0.840	-2.397	-0.840
X Variable 4	0.123	0.056	2.189	0.029	0.013	0.233	0.013	0.233
X Variable 5	1.489	0.502	2.966	0.003	0.503	2.474	0.503	2.474
X Variable 6	-0.033	0.056	-0.590	0.556	-0.142	0.076	-0.142	0.076
X Variable 7	-0.382	0.363	-1.054	0.292	-1.095	0.330	-1.095	0.330
X Variable 8	-0.702	0.134	-5.256	0.000	-0.965	-0.440	-0.965	-0.440
X Variable 9	-0.209	0.250	-0.837	0.403	-0.699	0.281	-0.699	0.281
X Variable 10	0.294	0.413	0.711	0.477	-0.518	1.105	-0.518	1.105

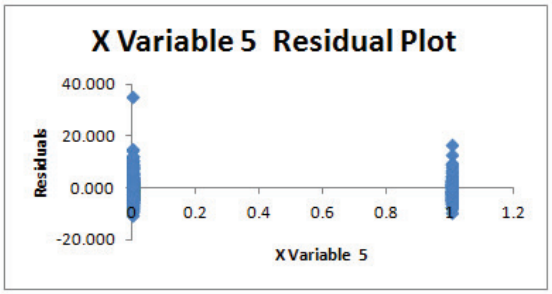
Residual Plots

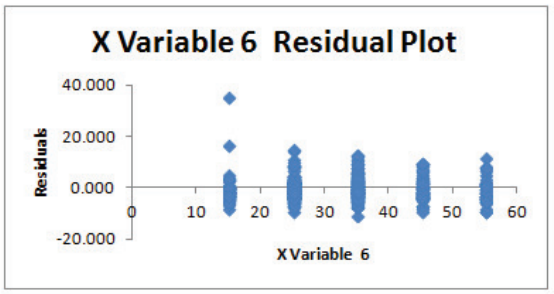


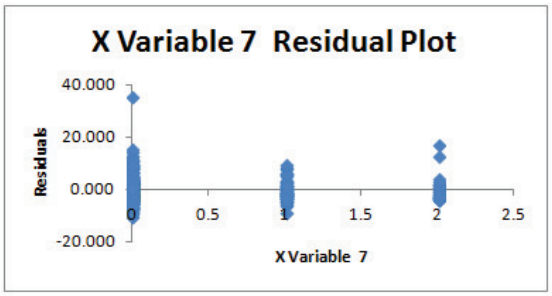


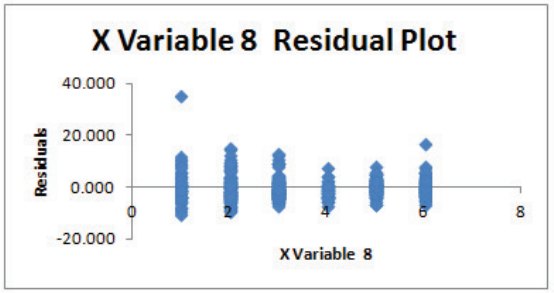


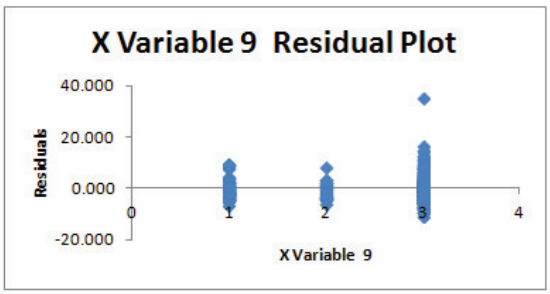


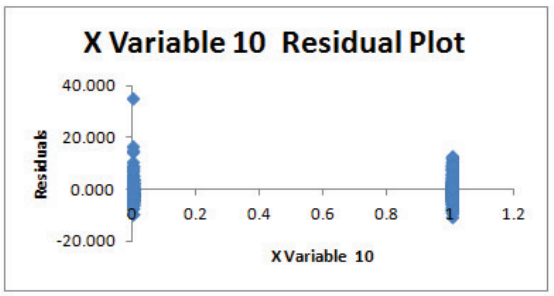












Step 2

The result of the regression suggests that the response variable Y is correlated to the explanatory variables. The p value for the F Statistics is ~0.000. However the p-value for the explanatory variables X_2, X_6, X_7, X_9 and X_{10} is very high and these need to be removed. Based on the Residual plots we can see we need to do a transformation to stabilize the variance.

Therefore the Wages were Log-Transformed and regression was re-run on the following formula:

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5$$

The Wage is the response variable (Y) and the other 10 columns are the explanatory variables (X), with

X_1 = Number of Yrs of Education
X_2 = Indicator variable for Gender
X_3 = Number of years of work experience
X_4 = Indicator variable for union membership
X_5 = Occupational category

Additionally, A is the intercept of the regression equation and the B_i's are the regression coefficients (or slopes) for the X_i variables. The following is the Summary out for the Regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.5720
R Square	0.3272
Adjusted R Square	0.3208
Standard Error	0.4349
Observations	534

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	48.572	9.714	51.356	0.000
Residual	528	99.875	0.189		
Total	533	148.447			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.203	0.137	8.778	0.000	0.934	1.472	0.934	1.472
X Variable 1	0.077	0.008	9.488	0.000	0.061	0.093	0.061	0.093
X Variable 2	(0.174)	0.040	(4.372)	0.000	(0.252)	(0.096)	(0.252)	(0.096)
X Variable 3	0.011	0.002	6.494	0.000	0.007	0.014	0.007	0.014
X Variable 4	0.219	0.050	4.369	0.000	0.120	0.317	0.120	0.317
X Variable 5	(0.073)	0.013	(5.604)	0.000	(0.099)	(0.048)	(0.099)	(0.048)

Conclusion

Analyzing the summary output for Step 2 we can see the p-value for all the explanatory variables is ~0.000 and the p-value for the F Statistic is ~0.000. This suggests that wages are dependent on Numbers of Yrs of Education, Gender, Work Experience, Union membership and Occupation.