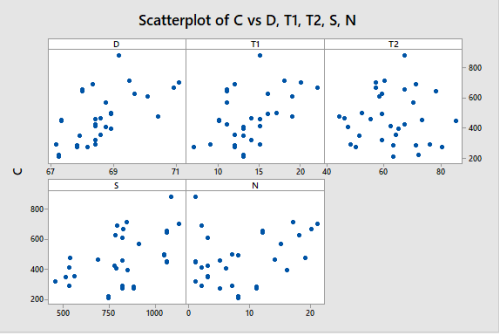
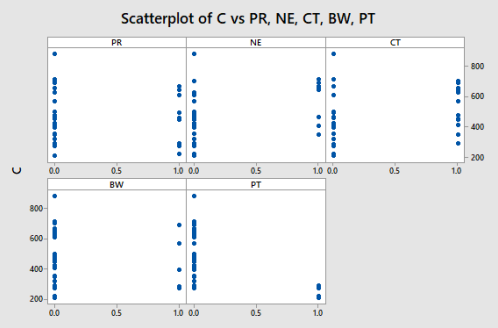
Problem Set 2 - Problem 1

Nick Handelman
4/25/2018

1 Visual Analysis of Cost vs Individual Variables





As an initial step, I wanted to get a feel for how cost varied depending on each of the explanatory variables alone. So, I made scatterplots to get an idea. In the scatterplots of cost vs each of the continuous variables, there are fairly clear patterns. T1 and S show linear relationships with a few outliers that possibly have significant influence. D and N also show linear relationships, though there are several points that appear to have a strong influence. T2 doesn't show a linear pattern, but the data do appear to have a triangle shape, indicating that higher cost is associated with middling levels of T2.

In the scatterplots of cost vs each of the categorical variables, there aren't any clear patterns in PR, NE, CT and BW. The variances are all roughly similar and none appear to be heavily skewed. If the plots are considered without the plant with the greatest cost, which appears to be a possible outlier in all of the 0 levels, then the existence of any pattern is even less clear. PT is a clear exception, where the plot clearly indicates that PT plants are all among the cheapest constructed. Also, the group variances are clearly different.

The scatterplot of PT suggest a transform, but a log transform did little to restore equal variance. I will keep this variable in mind in the analysis. None of the other variables were screaming for a transform, though some did have outliers that might have to be considered. Many of the explanatory variables appear related to the cost. Multiple regression will shed light on which of these (or some combination thereof) best explains the cost.

2 Multiple Regression

2.1 Best Subsets Regression on Given Explanatory Variables

Response is C

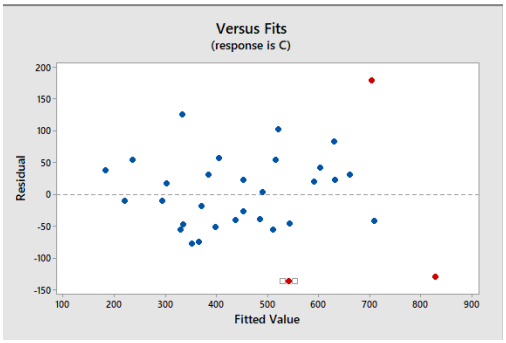
Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	D	T 1	T 2	S	P R	N E	C T	B W	N	P T
1	37.3	35.2	30.3	54.0	136.97	X									
1	34.0	31.8	28.4	58.3	140.46										X
2	58.4	55.5	49.7	28.4	113.43	X			X						
2	56.7	53.7	50.2	30.6	115.76				X						X
3	69.1	65.8	62.3	16.4	99.497	X			X						X
3	69.0	65.6	59.2	16.6	99.744	X			X		X				
4	75.2	71.5	67.2	10.5	90.802	X			X		X				X
4	74.8	71.1	62.9	10.9	91.434	X			X		X			X	
5	79.3	75.3	64.8	7.1	84.527	X			X		X	X		X	
5	77.3	73.0	66.1	9.6	88.445	X		X	X		X				X
6	81.8	77.4	68.6	5.8	80.801	X		X	X	X	X			X	
6	80.5	75.8	64.5	7.5	83.721	X		X	X		X	X		X	
7	83.5	78.7	67.4	5.6	78.534	X		X	X	X	X	X		X	
7	82.4	77.2	66.7	7.1	81.224	X		X	X	X	X		X	X	
8	83.8	78.2	65.8	7.2	79.446	X	X	X	X	X	X	X		X	
8	83.7	78.0	65.8	7.3	79.771	X		X	X	X	X	X	X	X	
9	83.9	77.3	61.6	9.0	80.971	X	X	X	X	X	X	X		X	X
9	83.9	77.3	64.3	9.1	81.098	X	X	X	X	X	X	X	X	X	
10	83.9	76.3	58.8	11.0	82.827	X	X	X	X	X	X	X	X	X	X

I performed a Best Subsets Regression on the 10 explanatory variables provided. Only first order terms were included in the regression. No interaction terms were included in the regression. The table shows the models with the highest R^2 for models using up through all 10 explanatory variables. Among these, it is important to also consider their $R^2(adj)$ and C_p statistics since they show which models have the best fits without having too many explanatory variables.
The model with 7 explanatory variables (D, T2, S, PR, NE, CT, N) has the largest $R^2(adj)$ and the smallest C_p. Both of these are desired in a model, so among the models considered, this model has the best fit to the given data without overfitting. Overall, the models consisting of 5-8 variables offer the smallest C_p statistics, while the models consisting of 6-10 variables offer the largest $R^2(adj)$ statistics. So, I would consider any model in the 6-8 variable range, depending on which presents the most reasonable interpretation. For further analysis, I will consider the aforementioned 7 variable model.

2.2 Multiple Regression on 7 Variable Model

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	7	749151	83.50%	749151	107022	17.35	0.000
D	1	334335	37.27%	359019	369019	59.83	0.000
T2	1	68161	7.60%	28539	28539	4.63	0.042
S	1	134926	15.04%	152846	162846	26.40	0.000
N	1	24818	2.77%	53288	53288	8.64	0.007
PR	1	44223	4.93%	27209	27209	4.41	0.046
NE	1	127488	14.21%	114114	114114	18.50	0.000
CT	1	15200	1.69%	15200	15200	2.46	0.130
Error	24	148021	16.50%	148021	6168		
Total	31	897172	100.00%				



Most of the variables have a significant contribution. The p-value of 0.13 for CT indicates there isn't strong evidence that it's associated coefficient isn't 0. However, considering the results of the best subsets regression, I'm leaving it in the equation. There are 3 outliers (marked in red) that are flagged as having a large residuals. The positive residual outlier is associated with the highest cost plant, so I'm not too surprised to see it flagged. while it does have the largest Cook's Distance and studentized residual (2.78), it doesn't have high leverage. In the scatterplots of C vs D and c vs T2, it's explanatory variable is middling and in C vs S and C vs N, several other observations share the same or similar explanatory values. Considering these

points, I can't exclude it from the data.
For the negative residual outliers, the one on the left doesn't particularly standout in any way with respect to the explanatory variables individually. It's leverage isn't high, so I don't see a strong a reason to exclude it from the data. The other negative residual stands out in three ways: it is the most recent constructed (D, but only by a few months), it has the largest net capacity (S, but comparable to several other observations) and it was constructed by the most experienced engineer (N). Also, it has the highest leverage and 2nd highest Cook's Distance. It doesn't standout in cost though, as might be expected consider its features. This is the likely reason it is flagged as an influential point. I reran the regression without this observation, but it didn't change much and introduced several new residuals. It did have slightly higher R_2 and $R_2(adj)$ statistics, but there wasn't enough difference to consider the data without this observation.

3 Conclusions

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-9606	1259	(-12206, -7007)	-7.63	0.000	
D	138.6	17.9	(101.6, 175.6)	7.74	0.000	1.66
T2	4.01	1.86	(0.16, 7.86)	2.15	0.042	1.89
S	0.4215	0.0820	(0.2522, 0.5908)	5.14	0.000	1.21
N	-8.22	2.80	(-13.99, -2.45)	-2.94	0.007	1.58
PR						
1	-73.7	35.1	(-146.2, -1.3)	-2.10	0.046	1.37
NE						
1	143.4	33.3	(74.6, 212.2)	4.30	0.000	1.08
CT						
1	47.6	30.3	(-15.0, 110.1)	1.57	0.130	1.15

Refer to the coefficients table. The following statements refer to average cost change with all other variables held constant. An increase of 1 year (D), increased cost by 138.6. Each extra month (T2) increased cost by 4.01. Each extra MWe (S) increased cost by 0.4215. Experienced engineers (N) decreased cost by 8.22. If an LWR existed on the site before, cost decreased by 73.7. If the plant was built in the Northeast (NE), cost increased by 143.4. Addition of a cooling tower (CT) increased cost by 47.6.
Before performing an analysis, from my limited knowledge on the subject of nuclear power plant construction, I expected D, S, PR, NE, CT, N and PT to have certain effects on the cost. Interestingly, T2 had an effect that I wasn't expecting. Most surprisingly, PT did not have a strong effect on the regression. In fact, in the Best Subsets Regression, none of the "best" (6-8 variable) models included PT as an explanatory variable.