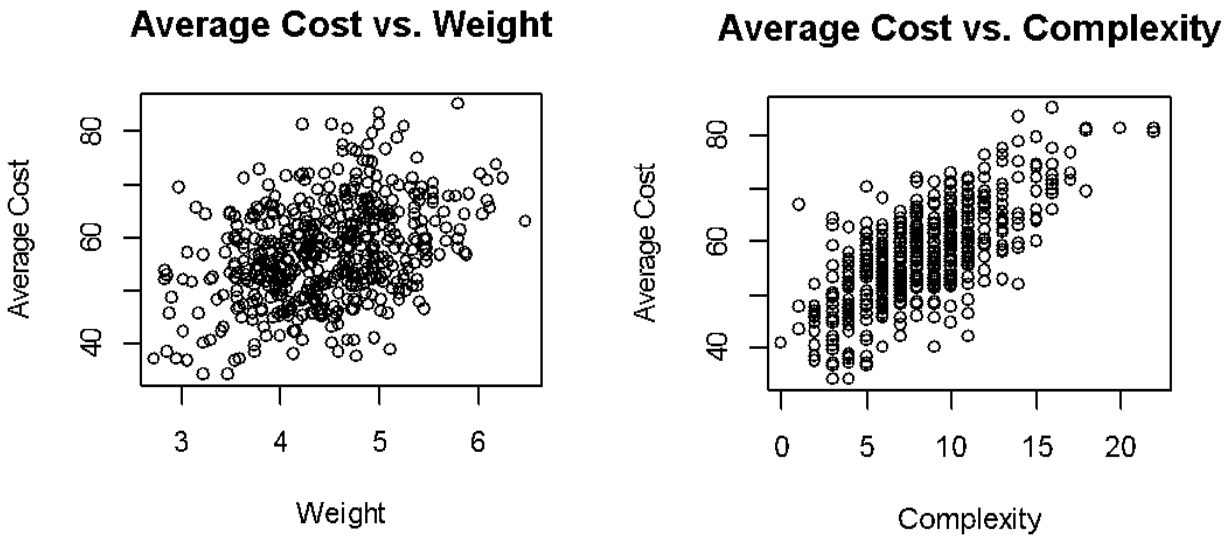
Appendix

A. Selection of Variables

A.1 Selection of Variables for Cost Estimation Model

The quantitative variables *Units*, *Weight*, *Chisel* and *Stamp* appear to be good predictors of average cost; however, the variables *Chisel* and *Stamp* are highly correlated ($r = 0.87$). Adding both of them as two separate variables to the model causes multicollinearity. Since the number of chiseling and stamping operations represents the complexity of the block, adding them together to become one variable called *Complexity* will help minimize multicollinearity and improve the model. Figure 1 shows that *Weight* and *Complexity* both have positive linear relationships with average cost. Therefore, together with *Units*, *Weight* and *Complexity* are good predictors of average cost. Given the quantitative variables above are already in the model, the three categorical variables *Goal.sd*, *Rush*, and *Detail* are not statistically significant at any reasonable significance level.

Figure 1. Linear Relationship Between Response and Predictors



A.2 Selection of Variables for Cost Analysis Model

Beside the three variables in the estimation model, *Labor*, *Cost*, *Lost* and categorical variable for the managers also yield significance as the last predictor added. Although treating *Weight*, *Lost* and *Cost* as three separate variables provides better predictions, it does not make economic sense. Combining these variables to create *Material Cost* is more appropriate because *Material Cost* is expressed in the same units as average cost. Figure 2 demonstrates the positive linear relationship of *Material Cost* and *Labor* with average cost.

The variables *Manager*, *Music*, *Shift*, *Plant*, *Rework* and *Breakdown* are closely related to one another. As a result, including more than one of these variables in the model introduced strong multicollinearity and in a few cases perfect linear dependencies. After thoroughly investigating these variables, it was determined that *Manager* was the most appropriate indicator of average cost. However, when *Manager* was included as a categorical variable with one level for each individual, only the dummy variable for Devon was significant (Table 1.). The plot of *Manager* against average cost (Figure 3.) clearly shows that jobs with Devon as production manager have lower average costs than jobs with any other manager. Thus, we created a new variable called *Devon* which has the value one when Devon is on duty and zero otherwise.

Finally, *Room Temperature* and average cost did not appear to have any relationship whatsoever, thus it was not included in the model.

Table 1. Coefficients Summary

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     24.81177    3.35591   7.393 6.25e-13 ***
units.log       -2.90405    0.46921  -6.189 1.28e-09 ***
material.cost    1.05616    0.04401  23.999  < 2e-16 ***
labor           11.18862    1.54163   7.258 1.56e-12 ***
managerBeatrice -0.09149    0.59058  -0.155   0.8769
managerCarl     -0.92771    0.56736  -1.635   0.1027
managerDevon    -4.28213    0.59566  -7.189 2.46e-12 ***
managerEbrahim  -1.03069    0.56942  -1.810   0.0709 .
complexity       1.08210    0.08425  12.843  < 2e-16 ***
```

Figure 2. Plots of Average Cost against Material Cost and Labor

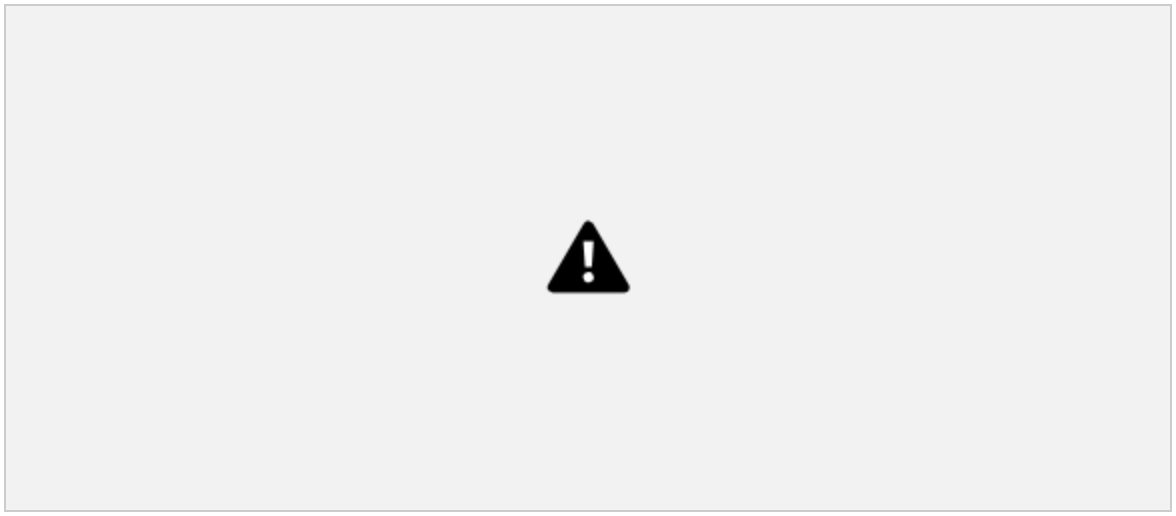
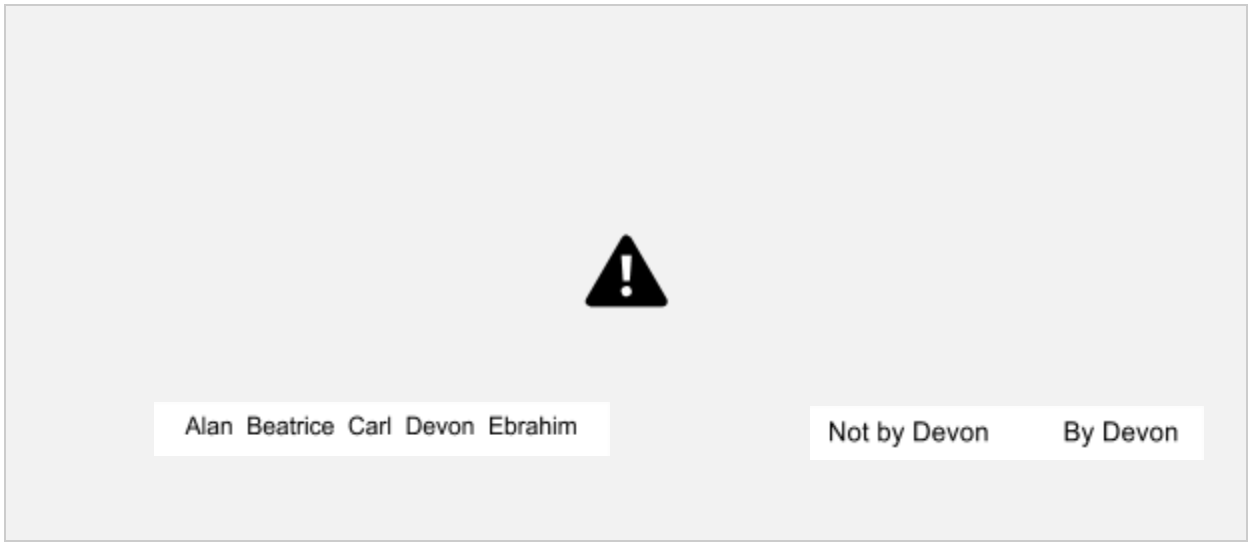


Figure 3. Plot of Average Cost against Managers



B. Transformations

Because the sales representatives believe *Units* is not a straight-line predictor, several transformations on *Units* were considered. Using Tukey's Bulge Rule as a reference, the transformations considered were square root, natural logarithm, inverse square root, and inverse, all of which account for diminishing marginal costs. Ultimately, the natural logarithm was chosen for reasons discussed in the following section.

C. Candidate models

The only difference between the three candidate models within each table is transformation of the unit term. Although the three candidate models are all significant at the $\alpha = 0.001$ significant level, their statistics are slightly different. Since the cost estimation model will be used to estimate the cost for future orders, the PRESS statistic is used as criteria to evaluate the models. Among these models, the second model has the smallest PRESS statistic and AIC as well as the largest adjusted R^2. Because the differences in evaluation criteria between model (1) and model (2) are small, we chose model (1) for better interpretation of coefficient. The coefficient of the logarithmic term can be interpreted in the following way: holding all other variables constant, a small percentage change in the number of units results in a change in the average cost by that percentage of the coefficient. The negative sign of the coefficient for the logarithmic term implies that the more units that customer orders the lower the average cost per finished block will be.

Table 2. Candidate Models for Cost Estimation

Dependent Variable: Average Cost per Finished Block			
Predictors	(1)	(2)	(3)
ln(Units)	-2.69482 (0.58697)		
$\frac{1}{Units}$		833.62851 (153.33791)	
Units			-0.00580 (0.00156)
Weight	5.55205 (0.33837)	5.55591 (0.33562)	5.54377 (0.34076)
Complexity (stamp + chisel operations)	1.84245 (0.06199)	1.84217 (0.06149)	1.84316 (0.06243)
Intercept	33.49543 (3.87388)	15.02801 (1.66593)	19.84501 (1.76735)
Summary Statistics and Joint Tests			
PRESS statistic	12840.28	12604.65	13027.12
$R^2_{Prediction}$	0.804	0.807	0.800
Adjusted R^2	0.6972	0.7021	0.6929
AIC	3033.62	3025.49	3040.65
F-statistic	382.5	391.5	374.8
n	498	498	498

Notes:
(1) Standard errors are in parentheses.
(2) All of coefficients are statistically significant at the $\alpha = 0.001$ significant level.

Table 3. Candidate Models for Cost Analysis

Dependent Variable: Average Cost per Finished Block Predictors	(1)	(2)	(3)
ln(Units)	-2.85736 (0.46945)		
$\frac{1}{Units}$		819.24260 (122.57492)	
Units			-0.00637 (0.00125)
Material Cost	1.05708 (0.04402)	1.05474 (0.04371)	1.05761 (0.04450)
Labor	11.14798 (1.54329)	10.90252 (1.53003)	11.05802 (1.55975)
Complexity (stamp + chisel operations)	1.07926 (0.08402)	1.09035 (0.08333)	1.08393 (0.08491)
Devon (if Devon is Manager on Duty)	-3.73533 (0.46971)	-3.72144 (1.09035)	-3.72733 (0.47472)
Intercept	24.03600 (3.34115)	4.94279 (1.95961)	9.70070 (2.00592)
Summary Statistics and Joint Tests			
PRESS statistic	8217.3	8086.0	8396.8
$R^2_{Prediction}$	0.69	0.70	0.69
Adjusted R^2	0.8071	0.8098	0.803
AIC	2811.11	2803.98	2821.64
F-statistic	416.9	424.3	406.1
n	498	498	498

Notes:
(1) Standard errors are in parentheses.
(2) All of coefficients are statistically significant at the $\alpha = 0.001$ significant level.

D. Predictions

Table 4. Predictions for Typical Values

Units		**Weight**		**Complexity**		**Average Cost ($/block)**	**95% PI Lower Bound**	**95% PI Upper Bound**
347	L	3.99	L	6	L	$50.46	$39.66	$64.26

347	L	3.99	L	8	M	$54.34	$43.55	$65.13
347	L	3.99	L	11	H	$60.15	$49.55	$70.95
347	L	4.47	M	6	L	$53.11	$42.31	$63.90
347	L	4.47	M	8	M	$56.98	$46.20	$67.76
347	L	4.47	M	11	H	$62.79	$52.00	$73.58
347	L	4.92	H	6	L	$55.58	$44.78	$66.38
347	L	4.92	H	8	M	$59.46	$48.67	$70.25
347	L	4.92	H	11	H	$65.27	$54.47	$76.07
442	M	3.99	L	6	L	$49.96	$39.17	$60.76
442	M	3.99	L	8	M	$53.48	$43.05	$64.62
442	M	3.99	L	11	H	$59.65	$48.85	$70.44
442	M	4.47	M	6	L	$52.60	$41.82	$63.39
442	M	4.47	M	8	M	$56.48	$45.70	$67.26
442	M	4.47	M	11	H	$62.59	$51.50	$73.08
442	M	4.92	H	6	L	$55.08	$44.29	$65.87
442	M	4.92	H	8	M	$58.95	$58.17	$69.74
442	M	4.92	H	11	H	$64.77	$53.97	$75.56
548	H	3.99	L	6	L	$49.51	$38.71	$60.32
548	H	3.99	L	8	M	$53.39	$42.59	$64.19
548	H	3.99	L	11	H	$59.20	$48.40	$70.01
548	H	4.47	M	6	L	$52.16	$41.36	$62.95
548	H	4.47	M	8	M	$56.03	$45.24	$66.82
548	H	4.47	M	11	H	$61.84	$51.05	$72.64
548	H	4.92	H	6	L	$54.63	$43.83	$65.43
548	H	4.92	H	8	M	$58.51	$47.71	$69.30
548	H	4.92	H	11	H	$64.32	$53.52	$75.12

Notes:
L = Low (first quartile of variable)
M = Medium (median of variable)
H = High (third quartile of variable)

E. Outliers

Evaluating the models on the training data revealed two peculiar observations: 19 and 311. These observations exerted significant influence on the fitted values, the coefficient estimates and the estimate of precision as measured by DFFITS, DFBETAS and COVRATIO. Figure 4 shows the plot of residuals against fitted values, and it is clear that the residuals for observations 19 and 311 are extremely unusual. Taking a closer look into the data set, average cost of these two observations is abnormally low. While the mean of average cost for all the data is $56.98, the average costs for observations 19 and 311 are $6.74 and $5.50, respectively. This is due to the fact that they have unusually low total costs. Observation 19 has 499 units and total cost of

$3,362.97 and observation 311 has 674 units and total cost of $3,706.08. When the number of units for a job is in the range of 450 to 700, the total cost varies from approximately $30,000 to $40,000. It looks like there were errors while recording the total costs of these two observations. Thus, these observations were eliminated, and this helped to improve the models. Additionally, Tables 5 and 6 contain a comprehensive list of potentially influential points that need more attention. These points were identified after the final model was run on the data excluding observations 19 and 311.

Figure 4. Residuals Plotted Against Fitted Values

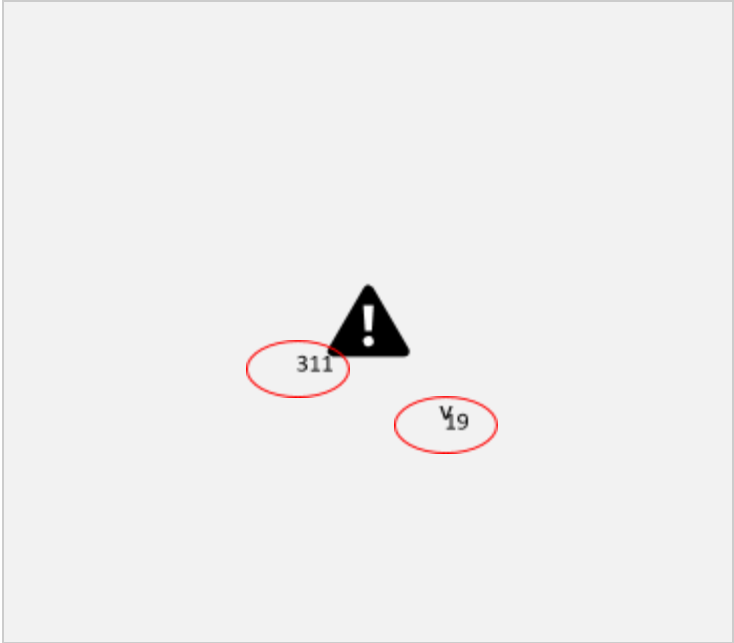


Table 5. Additional Potential Leverage / Influential Points in Cost Estimation Model

	Observations
h_{ii} (0.016)	400, 204, 149, 21, 215, 123, 227, 453, 487, 347, 151, 489, 216, 94, 350, 326, 465, 247, 301, 408, 473, 179, 342, 90, 187, 304, 93, 134, 162, 196.
DFFITS (0.18)	70, 216, 400, 204, 149, 266, 84, 290, 159, 433, 37, 226, 42, 248, 325, 342, 421, 423, 430, 480, 169, 247, 473, 307.
DFBETAS $_{intercept}$ (0.09)	400, 70, 204, 149, 421, 480, 266, 307, 248, 453, 42, 162, 169, 306, 430, 152, 326, 94, 140, 215, 450, 417, 275, 404, 72, 68, 486.
DFBETAS $_{ln(Units)}$ (0.09)	400, 204, 149, 70, 266, 480, 421, 453, 430, 307, 248, 423, 162, 94, 404, 306, 72, 152, 417, 37, 209, 140, 350, 84, 169, 342, 178, 122, 215, 148.
DFBETAS $_{Weight}$ (0.09)	216, 70, 159, 84, 37, 342, 42, 459, 102, 290, 326, 424, 473, 44, 433, 247, 99, 97, 172, 275, 381, 15, 200, 59, 309, 399, 308, 62, 346, 47, 169, 110, 226.
DFBETAS $_{Complexity}$ (0.09)	216, 226, 325, 290, 433, 70, 247, 58, 176, 30, 491, 123, 248, 266, 199, 9, 381, 244, 496, 212, 168, 159, 117, 369, 473, 67, 187.

COVRATIO (<0.98 and >1.02)	21, 70, 80, 84, 90, 93, 123, 149, 151, 159, 167, 169, 204, 215, 216, 226, 227, 248, 261, 266, 290, 301, 304, 347, 372, 400, 408, 433, 453, 465, 487, 489, 498.
Cook's Distance	Although all of observations have Cook's distance less than 1, which means there is no potential leverage points by using this criteria.

Note: The suggested thresholds are in parentheses.

Table 6. Additional Potential Leverage / Influential Points in Cost Analysis Model

	Observations
h_{ii} (0.024)	400, 204, 21, 149, 247, 215, 297, 123, 453, 162, 301, 476, 473, 227, 382, 151, 347, 480, 91, 245, 487, 109, 358, 408, 489, 27, 159, 61.
DFFITS (0.22)	216, 297, 266, 215, 70, 279, 59, 124, 204, 312, 496, 317, 80, 400, 487, 350, 364, 247, 423, 84, 218, 372, 187, 399, 44, 342, 169, 274, 38, 258.
$DFBETAS_{Intercept}$ (0.09)	215, 400, 204, 297, 134, 248, 149, 279, 350, 486, 84, 317, 307, 266, 70, 68, 430, 195, 363, 247, 487, 118, 393, 8, 325, 223, 80, 216, 417, 496, 21, 399.
$DFBETAS_{ln(Units)}$ (0.09)	204, 400, 215, 266, 350, 149, 134, 21, 70, 317, 404, 423, 364, 307, 399, 342, 68, 151, 430, 223, 312, 53, 421, 189, 258, 417, 178, 152, 486.
$DFBETAS_{Material\ Cost}$ (0.09)	297, 216, 266, 124, 59, 342, 91, 44, 290, 487, 97, 381, 364, 99, 215, 247, 179, 159, 195, 42, 347, 312, 110, 393, 178, 118, 80, 226, 244, 45, 258, 37, 38, 11, 350, 109, 22, 423.
$DFBETAS_{Labor}$ (0.09)	216, 80, 84, 59, 297, 422, 218, 274, 279, 404, 261, 248, 487, 472, 372, 419, 169, 247, 215, 364, 253, 451, 244, 498, 8, 357, 496, 231.
$DFBETAS_{Complexity}$ (0.09)	216, 297, 274, 187, 279, 218, 472, 419, 244, 80, 84, 169, 422, 261, 32, 347, 325, 349, 256, 253, 266, 215, 204, 128, 241, 168, 159, 214, 4, 59, 430.
$DFBETAS_{Devon}$ (0.09)	70, 279, 496, 124, 372, 218, 317, 423, 312, 152, 475, 38, 451, 307, 175, 169, 258, 189, 254, 419, 291, 498, 247, 53, 448, 185, 28, 73, 37, 133, 433, 301, 245, 259, 335, 184, 30, 361.
COVRATIO (<0.96 and >1.04)	21, 27, 44, 45, 68, 84, 123, 149, 162, 204, 216, 227, 247, 256, 261, 266, 358, 364, 382, 384, 400, 408, 453, 473, 476, 480, 486, 489.
Cook's Distance	Although all of observations have Cook's distance less than 1, which means there is no potential leverage points by using this criteria.

Note: The suggested thresholds are in parentheses.

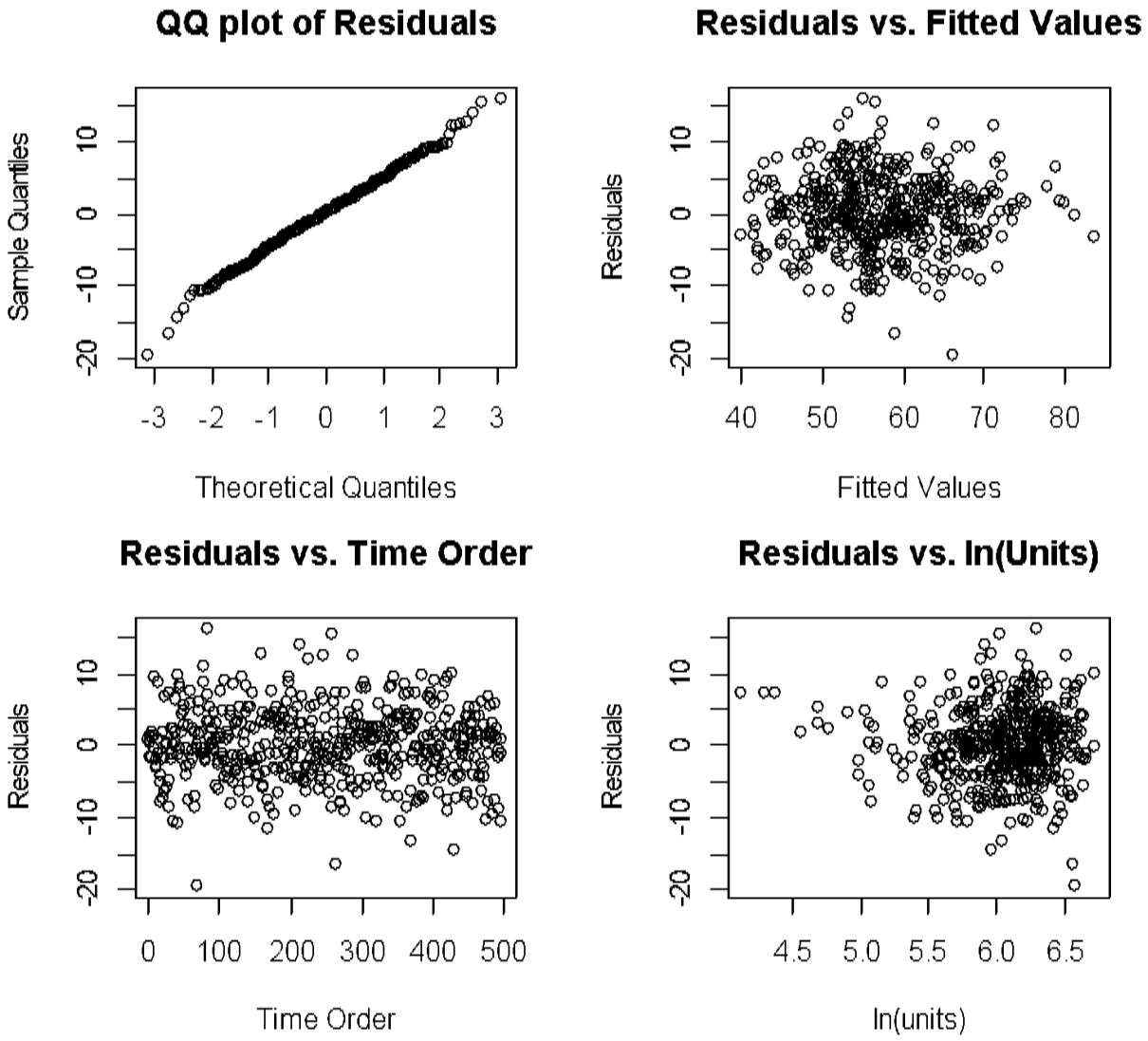
E. Assumptions Checking

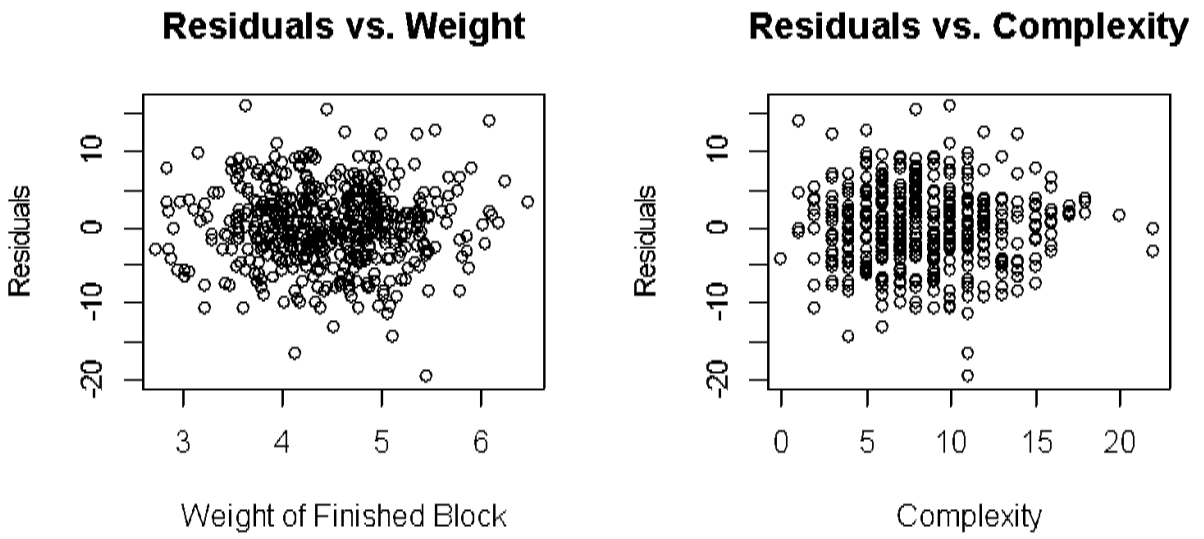
E.1 Assumptions Checking for Cost Estimation Model

The QQ plot of the residuals shows an approximately straight line, and resembles the QQ plot of randomly generate normal data. Thus, there is no problem with the normality assumption. Also, plots of the residuals against fitted values, time order and predictors do not have any patterns. Therefore none of the assumptions necessary for linear regression are violated. Additionally,

plots of the residuals against predictors not included in the model do not exhibit any patterns, so the decision to omit them is appropriate.

Figure 5. Plots of Residuals

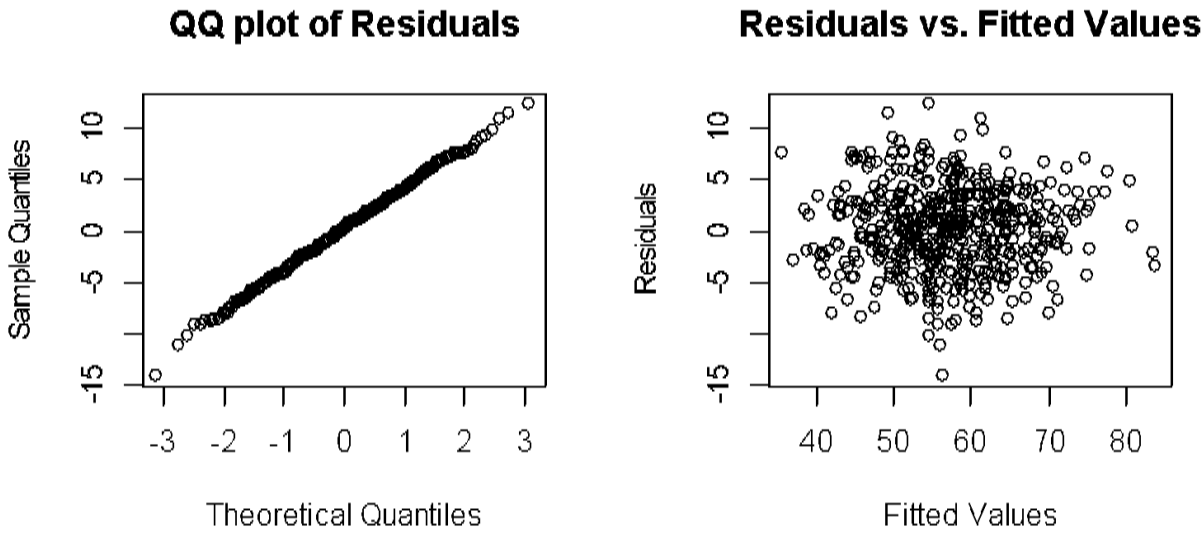


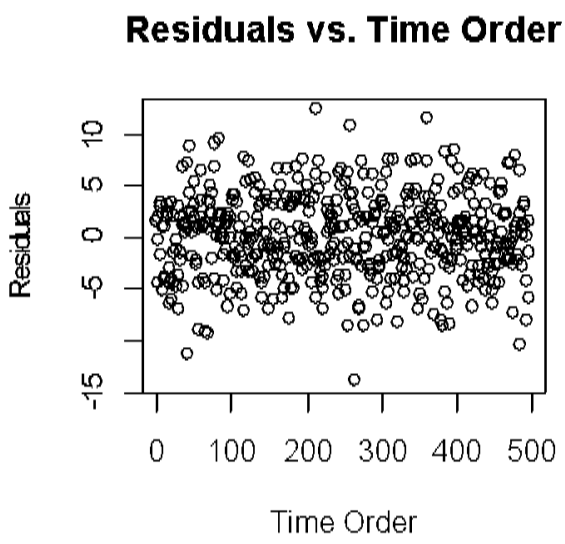


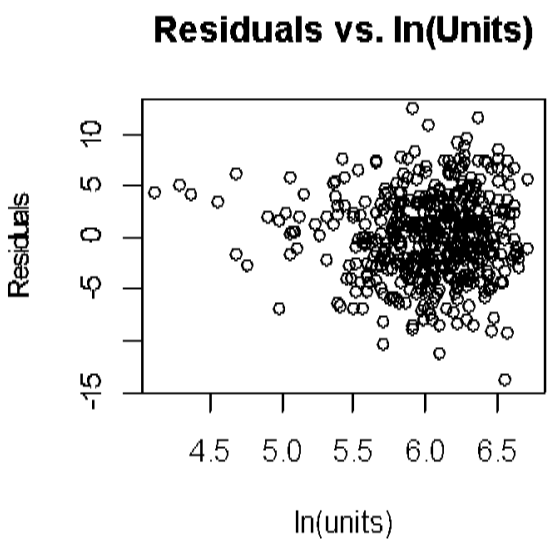
E.2 Assumptions Checking for Cost Analysis Model

Similar to the estimation model, the QQ plot of the residuals and the plots of residuals against fitted values, time order and predictors show that all assumptions hold.

Figure 6. Plots of Residuals




Residuals vs. Time Order
Residuals
10
5
0
-5
-15
0
100
200
300
400
500
Time Order


Residuals vs. ln(Units)
Residuals
10
5
0
-5
-15
4.5
5.0
5.5
6.0
6.5
ln(units)


Not by Devon
By Devon