Predicting costs

10/20/2019

1. Overview

The agency estimates project costs internally. Model variation in bids using a training data set (2015 through Q1 2019) and test it on a holdout set (Q2 and Q3). Use OLS and a penalized alternative.

Project-level data includes a number of characteristics. Economic and demographic variables are specific to national and regional economic and labor market conditions. City of New York building permits can help proxy for activity in the broader regional construction market. ENR's cost index serves as proxy for broader hard construction costs.

The goal is to bridge exploration of the agency's data with potential exogenous predictors, some of which the internal estimation process may underestimate or inadvertently miss. If those predictors can add measurable value to the agency's cost estimation methods it may inform potential changes in methodology.

2. Data.

Project data, called "bids" here, come from the Engineering Department. Economic and demographic indicators are specific to Greater New York (18 counties on both sides of the Hudson River) and come from the Planning and Regional Development Department; underlying data is from Oxford Economics.

Construction data is better from some parts of the region than others, but Jersey City's construction data is not yet as dependable as the City of New York's. NYC dominates regional construction anyway and it's justifiable for now to use its permitting data as a representative for the broader regional construction market.

Prices of construction materials and labor already figure directly into the agency's internal cost estimation, and this analysis borrows the same index for predictive powers now despite uncertainty regarding whether its implicit presence in agency estimates helps or hurts its role in any multivariable considerations. It is likely of second-order importance for now.

One variable of interest is the bidding process. Institutional discussions and earlier modeling suggests the bidding process may influence the bids. Limits placed on the range of bidders, for example, could, on average and holding other things constant, increase the average (and lowest qualifying) bid - this is basic microeconomics. I'll simplify the bidding format variable by making it binary: "public" for projects without significant constraints and "other" for ones, such as projects closed to firms not deemed "small business enterprises," that aren't. First I'll clean it a bit to consolidate near-dupliate categories.

Earlier work suggests there isn't major causal variation across the individual developing the in-house estimate and unique identifiers are omitted from this review.

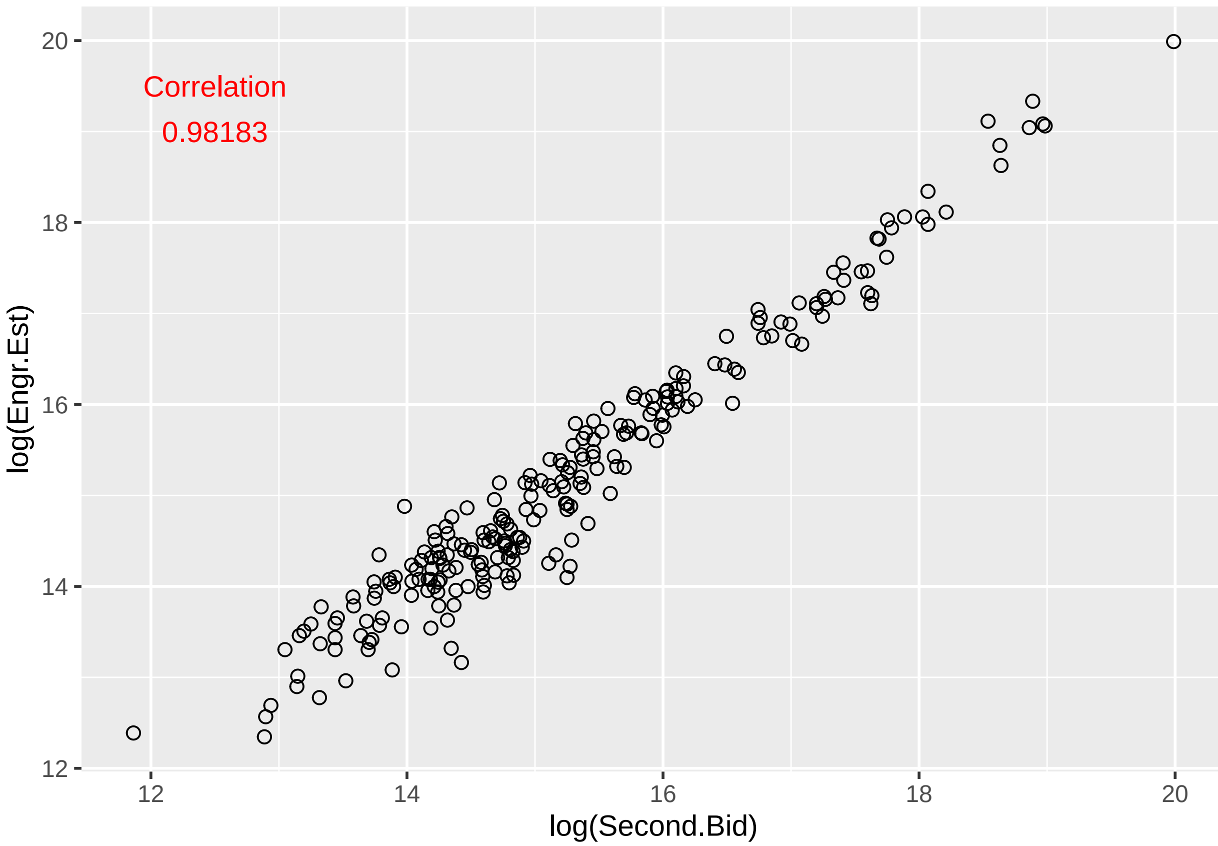
Agency projects last for months or years and actual costs do not exist for many of the observations, which at just over 260 projects already creates minor dimensionality concerns given the number of covariates. The second-lowest qualifying bid provides a reasonable target for evaluating internal estimates. The accuracy metric referenced through the exploratory discussion below and appendixed plots represents a ratio of that second-lowest bid over the estimate, both in dollars. A 1 would represent a case where the internal estimate (denominator) precisely matched the second-lowest bid; a 0.94 would mean the bidder bid 94 cents for every dollar estimated internally, et cetera.

```
bid = bids # backup my data frame
```

3. Exploratory analysis.

Why might developing a conrolled multivariate model will be worth it? The average gap between bids and estimates is less than $900,000, or around 5% - the average project bid was $15 million.

How much inconsistency does that represent?



The in-house engineers' guesses predict more than 98% of the variation in second-lowest bids. Some of the remaining variation may be explained by institutional guesswork.

Uncontrolled bivariate relationships provide easy clues as to predictors' potential role in more controlled multivariate relationships. Plots for this and subsequent exploratory work is at the end of the document; all consider accuracy, defined here as a ratio of the second-highest bid (a rule-of-thumb target) over the internal agency estimate.

- Location may matter: projects that span the Hudson River seem to come in, on average, higher than expected. Projects that span the Hudson River wind up costing (with respect to estimates) more, on average, than ones plunked squarely in either New York or New Jersey.
- The bidding process can be constrained or open, with potential ramifications on the ability to estimate costs.
- The signal is stronger regarding the type of project, which has an identifiable (if yet uncontrolled) relationship with estimating accuracy. Of three categories — infrastructure, paving, building — the agency appears to predict paving projects the best. Note: the story changes when considering the lowest bid, where paving projects' relationship to estimates varies significantly.
- Outliers for PATH and TB&T push the average accuracy metric up for each, otherwise there's little apparent difference across departments.
- Strong signals emerge that projects of medium size (in dollar terms) may be less evasive than much larger or smaller projects.

The implication of project size may provide the most valuable results. One might have expected the largest or smallest projects to be the tough to predict - particularly smaller projects. But it's the projects ranging from the 40th to 70th percentiles that seem to be the most challenging.

4. Modeling and prediction

Try and use exogenous covariates to predict an alternative engineering estimate, without using the low bid information, that might be closer to the low bid. Call it "expected low bid" or something so we can remember what we're trying to get.

The data set carries dimensionality challenges, with a number of variables (absolutely and relative to the number of observations). It includes mostly continuous variables but also a number of qualitative factors, both ordinal and nominal and all treated categorically without conversion to binary subvariables - the modeling processes used here do that automatically.

4a. Base model (manual selection).

Interpretation: specification was manual and intuitive. Given the fact that estimators' already try and take much of this information into account, however, a model with even a handful of extra covariates could represent overfitting - trying to hard.

Note

Ensure the accuracy variable calculated earlier is dropped before modeling or introducesome dual (reverse) causality, which could confuse models. ###Note 2 When a number appears in the output without context, it is likely an information criterion (and AIC), which may or may not provide value post-modeling.

Split data into training / test sets.

```
train = subset(bids,bids$Date<"2019-03-31")
test = subset(bids,bids$Date>="2019-04-01")
```

Choose a handful of potential predictors and build a linear model.

```
base = lm(Second.Bid ~ Engr.Est + Employment.in.construction + Format2 + Typeology + permits_1 + cci + q

  options(scipen=999)
  summary(base)
```

```
##
## Call:
## lm(formula = Second.Bid ~ Engr.Est + Employment.in.construction +
##     Format2 + Typeology + permits_1 + cci + quantile, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -65506055  -1141743     81828   1117149  54487762
##
## Coefficients:
##                                Estimate     Std. Error t value
## (Intercept)                  9010053.56125 20406029.67896   0.442
## Engr.Est                           0.88129        0.01627  54.155
## Employment.in.construction  -101671.69710   182067.69616  -0.558
## Format2Public              -1851502.29841  1125776.90936  -1.645
## TypeologyInfra             -1819770.72508  1274485.23111  -1.428
```

```
## TypeologyPaving           -1637036.88837  1574779.31649  -1.040
## permits_1                       55.19035      101.37426   0.544
## cci                           2767.59329     6713.67033   0.412
## quantile.L                 3489085.27332  1969863.89429   1.771
## quantile.Q                  -70865.32473  1919233.10004  -0.037
## quantile.C                  -76741.29146  1760463.28646  -0.044
## quantile^4               -1230490.29392  1695879.99702  -0.726
## quantile^5               -1224478.90714  1635257.21895  -0.749
## quantile^6               -1435368.89893  1604093.22396  -0.895
## quantile^7                -588789.04394  1634083.18058  -0.360
## quantile^8                -299412.85795  1607742.43624  -0.186
## quantile^9                 235847.01840  1613415.93362   0.146
##                                      Pr(>|t|)
## (Intercept)                            0.6593
## Engr.Est                 <0.0000000000000002 ***
## Employment.in.construction             0.5771
## Format2Public                          0.1015
## TypeologyInfra                         0.1547
## TypeologyPaving                        0.2997
## permits_1                              0.5867
## cci                                    0.6806
## quantile.L                             0.0779 .
## quantile.Q                             0.9706
## quantile.C                             0.9653
## quantile^4                             0.4689
## quantile^5                             0.4548
## quantile^6                             0.3719
## quantile^7                             0.7190
## quantile^8                             0.8524
## quantile^9                             0.8839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7817000 on 222 degrees of freedom
## Multiple R-squared:  0.9658, Adjusted R-squared:  0.9633
## F-statistic: 391.9 on 16 and 222 DF,  p-value: < 0.00000000000000022
```

```
mse_b = round(mse(train$Second.Bid,base$fitted.values),0)
adjr_b = round(summary(base)$adj.r.squared,3)
```

The equation above throws a decent amount of information at the engineering estimate and tries to predict the second bid. If the result is noticably closer to the low qualifying bid than the original estimate, you can use the delta as a post-estimation fudge factor to adjust the final estimate.

The numbers calculated above include a few metrics to use in comparing the three estimates against the objective data point at the low bid, which is what we're trying to predict. Ghose three estimates I'm talking about are: 1. The original, raw engineering cost estimate, 2. The first alternative, where we built a model by hand to try and use a few more data points to enhance the original estimate, and; 3. The second alternative, a kitchen sink model that throws even more data points at the question. This followed an effort to use a penalized regression to identify the best covariates, but that penalization algorithm actually suggested there isn't much we can do to enhance the original estimate. (Note: this will prove prescient.)

I'll build a table near the very end of this script that summarize the metrics I'm using to understand how well these modeling efforts work. The metrics will be: A. A basic t-test to understand whether there's even a statistically significant difference between the estimate I'm getting and the enhanced estimate I'm modeling with it, B. A correlation between the two numbers, to try and understand the magnitude of that difference (if we can trust it really exists), C. Two measures of the predictive modeling power of the models, an adjusted R-squared and the mean squared error (MSE). Both are common metrics of power. The first can be viewed discretely for each model but the second only provides a relative measure between models.

What is the summary of the predicted values? How does it compare to the summary of second-lowest bids?

Looking at the two summaries above, what's more accurate, the original engineering estimate or the first enhanced prediction? Neither, really. In fact the estimate and prediction aren't even (statistically) significantly different:

```
##
## Call:
## lm(formula = base$fitted.values ~ train$Second.Bid)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -38596289  -1454493   -455877    873116  68843962
##
## Coefficients:
##                     Estimate  Std. Error t value            Pr(>|t|)
## (Intercept)      510390.5537 512180.8632   0.997                0.32
## train$Second.Bid      0.9658      0.0118  81.815 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7435000 on 237 degrees of freedom
## Multiple R-squared:  0.9658, Adjusted R-squared:  0.9657
## F-statistic:  6694 on 1 and 237 DF,  p-value: < 0.00000000000000022
```

Maybe something more robust can come with a little creativity.

(Note: the model should control to prevent negative values. To be done next time.)

It might have been worth trying with the log of prices, only because the statistical fit becomes multiplicative instead of linear, but that produced similar results.

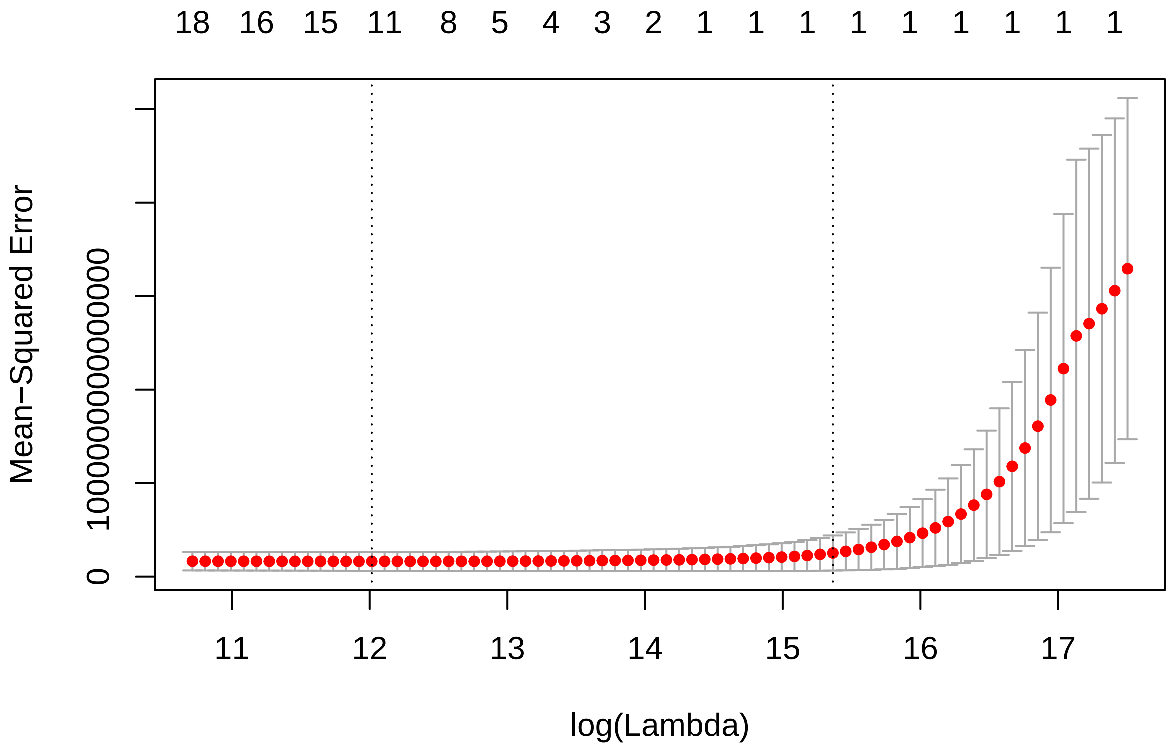
4b. Regularization.

Strip the bids data of unusable stuff, then set controls and run. The package I'm using, glmnet, requires a little extra preparation.

Could running a regularization process weed out weaker variables and identify one or two items that help prediction? The process would shrink (toward zero) the estimates for covariates that threaten to introduce more uncertainty than predictive power. This chooses a model automatically.

Running a regularized regression below adopts base assumptions in the modeling software package (10 folds, standardized coefficients, gaussian distribution, MSE evaluation metric, et cetera).

How many variables survive the penalty process as the penalty grows?



At a lower penalty (higher tuning parameter) there are a dozen non-zero predictors remaining, once the parameter cranks up a little all but one or two fall away.

Initial attempt supresses all covariates. A little fine-tuning identifies the two parameters that survive as the model gets further from ordinary least squares without disappearing:

```
lasso2 = cv.glmnet(x=XP, y=YP, alpha=.25)
coef(lasso2, s=lasso2$lambda.1se)
```

```
## 45 x 1 sparse Matrix of class "dgCMatrix"
##                                                           1
## (Intercept)                                  -552568.9426178
## Engr.Est                                           0.7759872
## Loc                                                .
## LD                                                 .
## Typeology                                          .
## Consumer.price.index                               .
## Employment.in.communications                       .
## Employment.in.construction                         .
## Employment.in.education.and.health                 .
## Employment.in.financial.and.business.services      .
## Employment.in.financial.services                   .
## Employment.in.government                           .
## Employment.in.other.services                       .
## Employment.in.production.industries                .
## Employment.in.professional.services                .
## Employment.in.real.estate                          .
## Employment.in.retail                               .
## Employment.in.transport.services                   .
## Employment.in.wholesale                            .
## Output.in.communications                           .
## Output.in.construction                             .
```

```
## Output.in.financial.services                          .
## Output.in.government                                  .
## Output.in.retail                                      .
## Output.in.education.and.health                        .
## Output.in.financial.and.business.services             .
## Output.in.other.services                              .
## Output.in.production.industries                       .
## Output.in.professional.services                       .
## Output.in.real.estate                                 .
## Output.in.transport.services                          .
## Output.in.wholesale                                   .
## Personal.disposable.income..nominal                   .
## Personal.disposable.income..real                      .
## Personal.income..nominal                              .
## Retail.sales..nominal                                 .
## Retail.sales..real                                    .
## Total.employment                                      .
## Total.office.based.employment                         .
## Total.output                                          .
## Total.population                                      .
## permits_1                                             .
## cci                                                   .
## Format2                                               .
## quantile                                         588275.0021597
```

5. Discussion.

This suggests using anything beyond the engineer's estimate itself to better predict the lowest qualifying good adds more uncertainty (in the form of noise that's tough to explain) than it adds value. (The "penalty" associated with adding variables is greater than the extra predictive power they bring.) The most obvious exception is project size, which isn't too surprising given the exploratory work done earlier.

```
summary(lm(bids$Second.Bid ~ bids$Engr.Est + bids$quantile_2 + bids$quantile_3 + bids$quantile_4 + bids$
             bids$quantile_8 + bids$quantile_9 + bids$quantile_10))
```

```
##
## Call:
## lm(formula = bids$Second.Bid ~ bids$Engr.Est + bids$quantile_2 +
##     bids$quantile_3 + bids$quantile_4 + bids$quantile_5 + bids$quantile_6 +
##     bids$quantile_7 + bids$quantile_8 + bids$quantile_9 + bids$quantile_10)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -64927474   -490838     31601    573534  56729942
##
## Coefficients:
##                      Estimate    Std. Error t value            Pr(>|t|)
## (Intercept)        76048.34661 1598483.07814   0.048              0.9621
## bids$Engr.Est          0.87692       0.01604  54.685 <0.0000000000000002
## bids$quantile_2   -34059.25935 2260576.56828  -0.015              0.9880
## bids$quantile_3   191658.26895 2260597.38380   0.085              0.9325
## bids$quantile_4   467164.92654 2260635.79910   0.207              0.8365
## bids$quantile_5   724044.57116 2260706.51997   0.320              0.7491
```

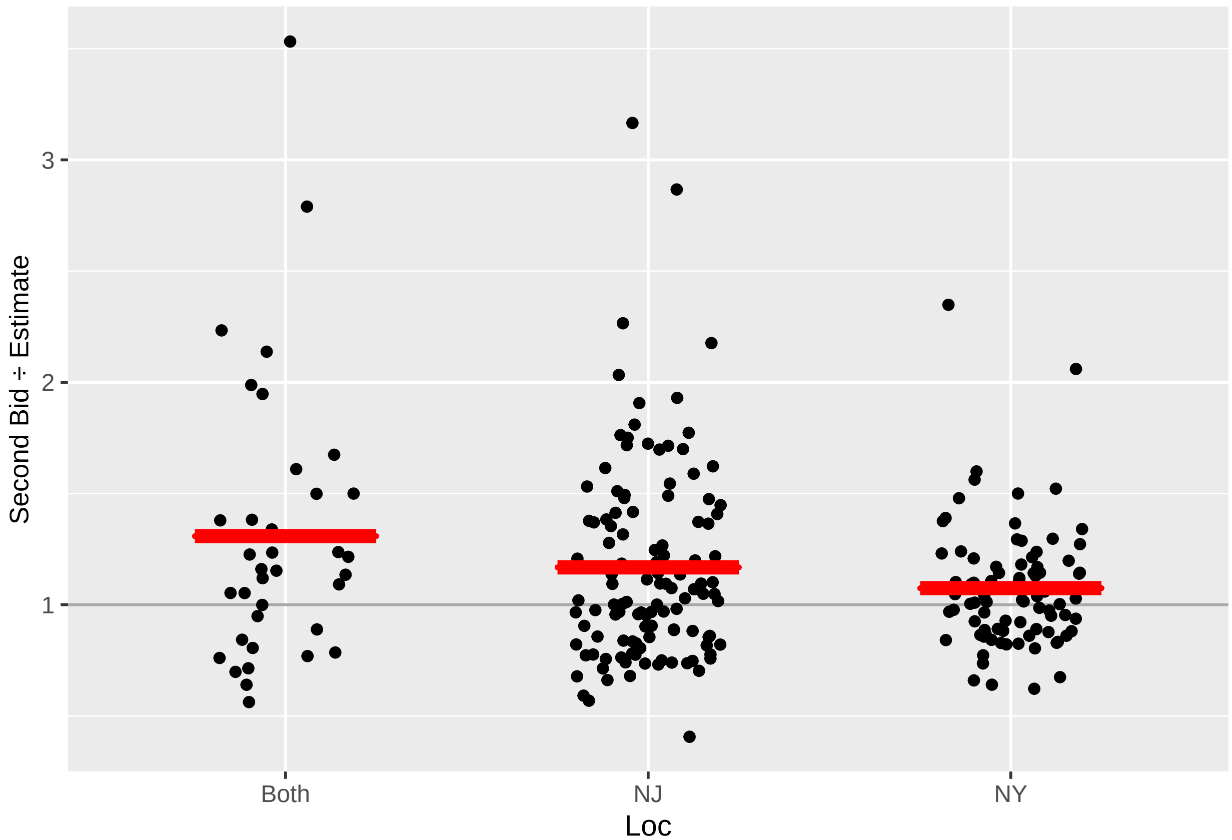
```
## bids$quantile_6   921683.48531 2261008.27650   0.408              0.6839
## bids$quantile_7   491236.92917 2262088.51151   0.217              0.8283
## bids$quantile_8  1195745.54423 2264742.28422   0.528              0.5980
## bids$quantile_9  4099789.39909 2289440.33695   1.791              0.0747
## bids$quantile_10 2003942.85274 2907737.97455   0.689              0.4914
##
## (Intercept)
## bids$Engr.Est    ***
## bids$quantile_2
## bids$quantile_3
## bids$quantile_4
## bids$quantile_5
## bids$quantile_6
## bids$quantile_7
## bids$quantile_8
## bids$quantile_9  .
## bids$quantile_10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7831000 on 228 degrees of freedom
## Multiple R-squared:  0.9648, Adjusted R-squared:  0.9632
## F-statistic: 624.1 on 10 and 228 DF,  p-value: < 0.00000000000000022
```

Exploratory work.

Plots and output from earlier exploratory bivariate work are below. Each considers a potential predictor's relationship to the agency's cost estimation accuracy, defined here as the second-lowest bid over the internal agency estimate.

Location.

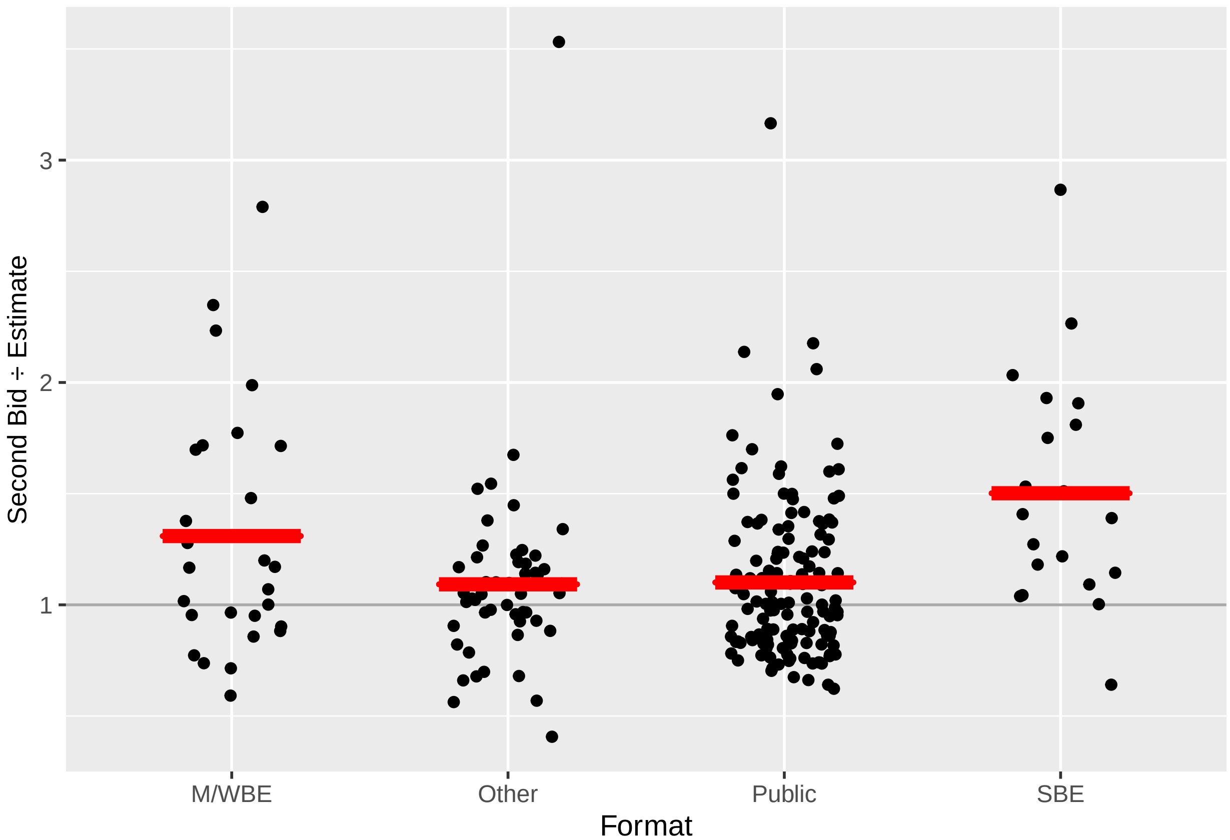
```
##
## Call:
## lm(formula = bids$accuracy ~ bids$Loc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76135 -0.25572 -0.07396  0.18784  2.22307
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  1.30860    0.07132  18.349 < 0.0000000000000002 ***
## bids$LocNJ  -0.14044    0.08147  -1.724              0.08606 .
## bids$LocNY  -0.23401    0.08509  -2.750              0.00642 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4279 on 236 degrees of freedom
## Multiple R-squared:  0.0318, Adjusted R-squared:  0.02359
## F-statistic: 3.875 on 2 and 236 DF,  p-value: 0.02208
```



Bidding process.

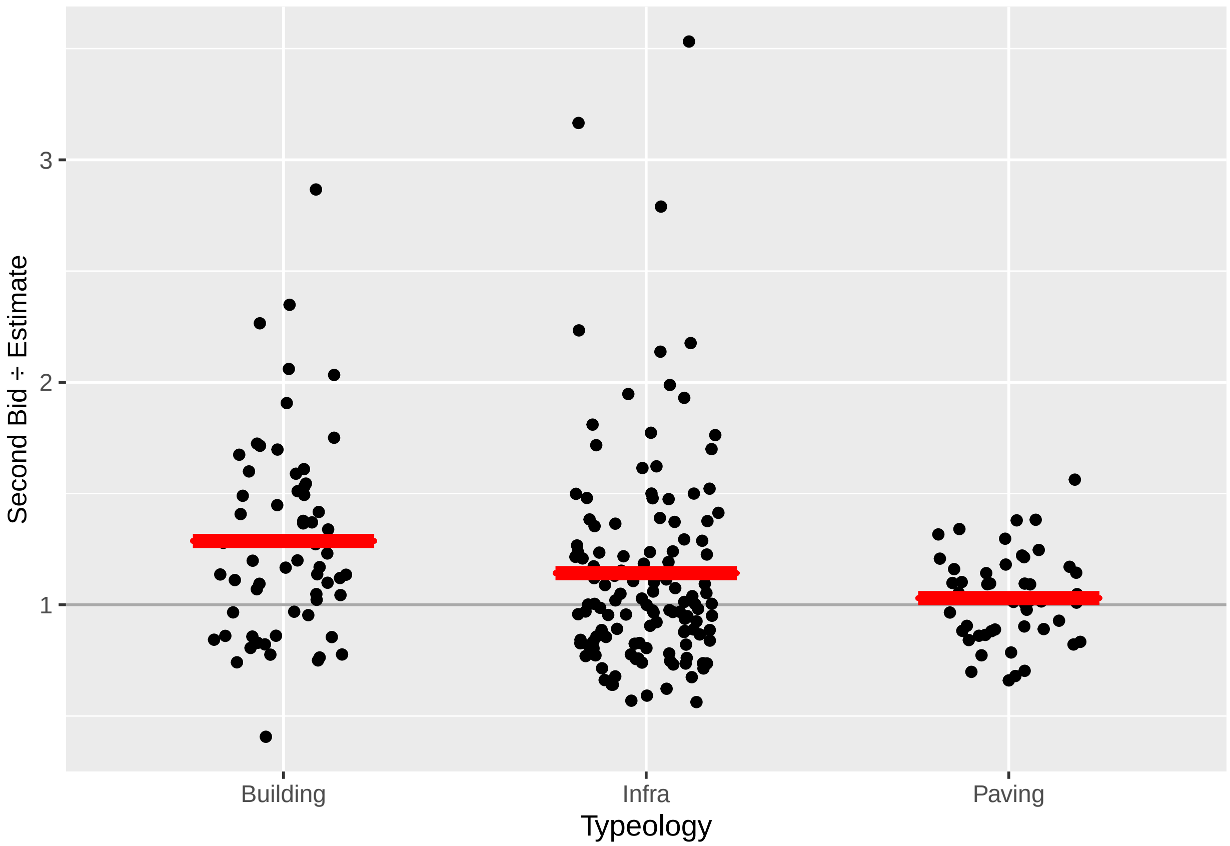
```
##
## Call:
## lm(formula = bids$accuracy ~ as.factor(bids$Format))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86097 -0.27198 -0.08562  0.16241  2.43943
##
## Coefficients:
##                              Estimate Std. Error t value
## (Intercept)                   1.30933    0.08035  16.295
## as.factor(bids$Format)Other  -0.21709    0.09937  -2.185
## as.factor(bids$Format)Public -0.20862    0.08771  -2.379
## as.factor(bids$Format)SBE     0.19246    0.12318   1.563
##                                         Pr(>|t|)
## (Intercept)                  <0.0000000000000002 ***
## as.factor(bids$Format)Other               0.0299 *
## as.factor(bids$Format)Public              0.0182 *
## as.factor(bids$Format)SBE                 0.1195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4175 on 235 degrees of freedom
## Multiple R-squared:  0.08211,    Adjusted R-squared:  0.07039
```

```
## F-statistic: 7.007 on 3 and 235 DF,  p-value: 0.000156
```



Project typeology.

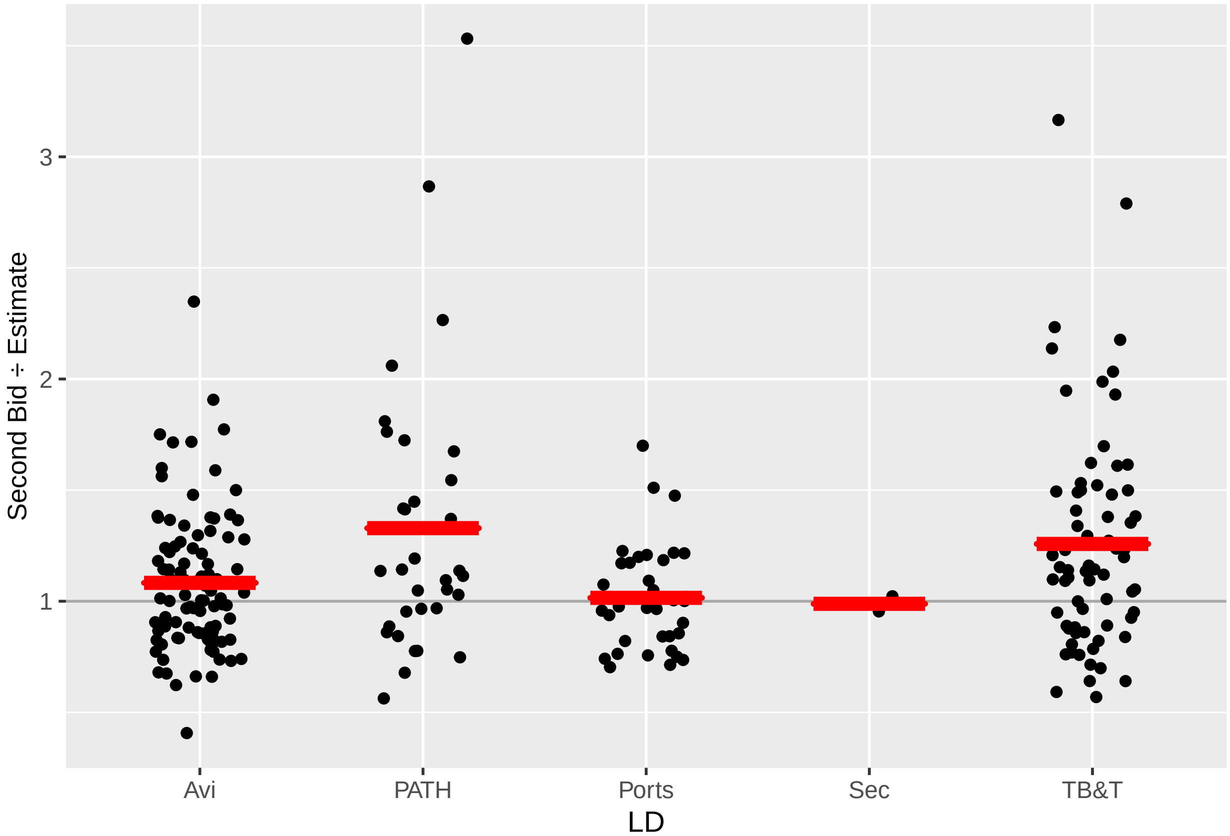
```
##
## Call:
## lm(formula = bids$accuracy ~ as.factor(bids$Typeology))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88015 -0.26202 -0.08883  0.15642  2.38969
##
## Coefficients:
##                                 Estimate Std. Error t value
## (Intercept)                      1.28697    0.05451  23.609
## as.factor(bids$Typeology)Infra  -0.14499    0.06616  -2.192
## as.factor(bids$Typeology)Paving -0.25693    0.08168  -3.146
##                                             Pr(>|t|)
## (Intercept)                     < 0.0000000000000002 ***
## as.factor(bids$Typeology)Infra               0.02938 *
## as.factor(bids$Typeology)Paving              0.00187 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4258 on 236 degrees of freedom
## Multiple R-squared:  0.04143,    Adjusted R-squared:  0.03331
## F-statistic:   5.1 on 2 and 236 DF,  p-value: 0.006784
```



Department.

```
##
## Call:
## lm(formula = bids$accuracy ~ as.factor(bids$LD))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76624 -0.25864 -0.06976  0.18420  2.20275
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)               1.08270    0.04260  25.417 < 0.0000000000000002
## as.factor(bids$LD)PATH    0.24622    0.08487   2.901              0.00407
## as.factor(bids$LD)Ports  -0.06742    0.08137  -0.829              0.40819
## as.factor(bids$LD)Sec    -0.09435    0.30121  -0.313              0.75439
## as.factor(bids$LD)TB&T    0.17514    0.06627   2.643              0.00878
##
## (Intercept)              ***
## as.factor(bids$LD)PATH   **
## as.factor(bids$LD)Ports
## as.factor(bids$LD)Sec
## as.factor(bids$LD)TB&T   **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4217 on 234 degrees of freedom
## Multiple R-squared:  0.06762,    Adjusted R-squared:  0.05168
## F-statistic: 4.243 on 4 and 234 DF,  p-value: 0.002468
```



Project size.

```
##
## Call:
## lm(formula = bids$accuracy ~ as.factor(bids$quantile))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76921 -0.25251 -0.07487  0.15180  2.19418
##
## Coefficients:
##                              Estimate Std. Error t value
## (Intercept)                   1.15535    0.02694  42.888
## as.factor(bids$quantile).L -0.09813    0.08546  -1.148
## as.factor(bids$quantile).Q -0.32559    0.08551  -3.808
## as.factor(bids$quantile).C  0.09462    0.08538   1.108
## as.factor(bids$quantile)^4  0.17352    0.08521   2.036
## as.factor(bids$quantile)^5 -0.16602    0.08509  -1.951
## as.factor(bids$quantile)^6 -0.07635    0.08503  -0.898
## as.factor(bids$quantile)^7 -0.05900    0.08501  -0.694
## as.factor(bids$quantile)^8  0.06234    0.08500   0.733
## as.factor(bids$quantile)^9  0.10264    0.08500   1.208
##                            Pr(>|t|)
```

```
## (Intercept)                 < 0.0000000000000002 ***
## as.factor(bids$quantile).L              0.25203
## as.factor(bids$quantile).Q              0.00018 ***
## as.factor(bids$quantile).C              0.26892
## as.factor(bids$quantile)^4              0.04287 *
## as.factor(bids$quantile)^5              0.05227 .
## as.factor(bids$quantile)^6              0.37018
## as.factor(bids$quantile)^7              0.48834
## as.factor(bids$quantile)^8              0.46411
## as.factor(bids$quantile)^9              0.22847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4164 on 229 degrees of freedom
## Multiple R-squared:  0.1102, Adjusted R-squared:  0.0752
## F-statistic:  3.15 on 9 and 229 DF,  p-value: 0.001329
```