Predicting costs

10/25/2019

1. Overview

The agency estimates project costs internally prior to public bidding openings.

Model variation in those bids a couple of ways and try to find ways to improve upon the original estimates before future bidding openings. (Compare basic statistical methods with what are called "regularization" alternatives.)

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 engineers' estimation methods.

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 somewhat representative of 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 despite uncertainty regarding whether its implicit presence in agency estimates helps or hurts its role in any multivariable treatments. It is likely of second-order importance for now.

One variable of interest is the actual bidding process. Institutional discussions and earlier modeling suggests the bidding process may influence actual project costs. Limits placed on the range of bidders could, for example, 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 closed to firms that don't qualify, such as large enterprises. First I'll clean that variable a bit to consolidate categories very close in spirit.

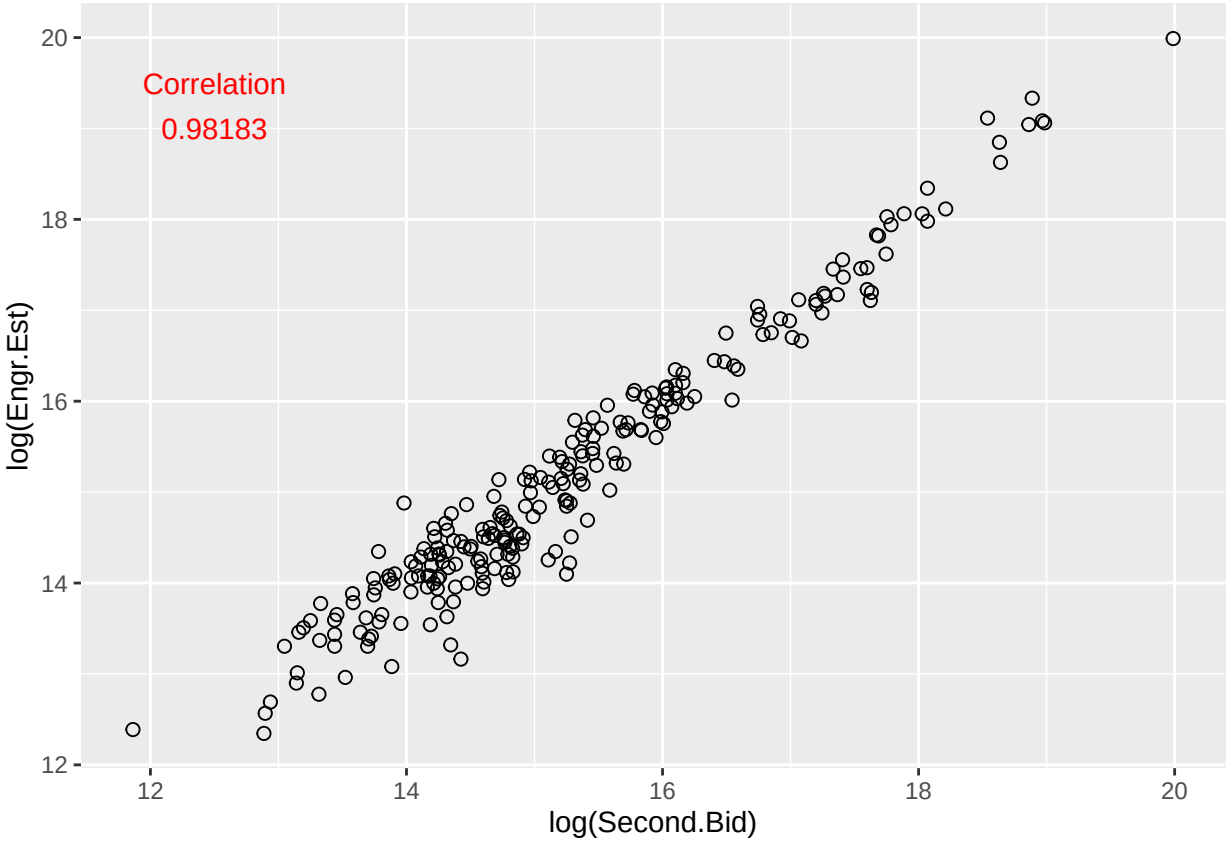
Earlier work suggests there isn't major variation across the individual developing the in-house estimate and unique employee (estimator) identifiers are omitted from this review.

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 potential predictive variables. 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.

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?

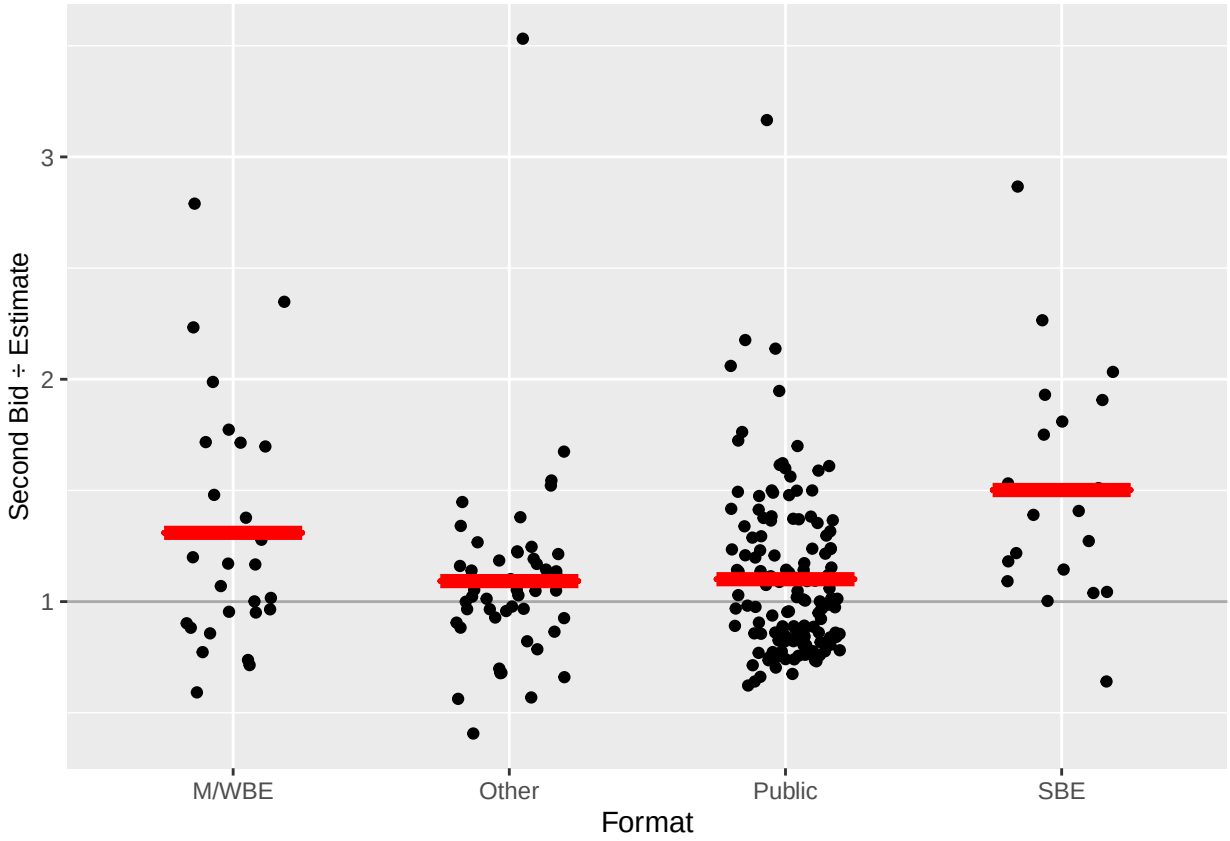


The in-house engineers' guesses predict more than 98% of the variation in second-lowest bids. Some of the remaining variation may be explained by institutional guesswork.

Uncontrolled bivariate relationships provide easy clues as to predictors' potential role in more controlled multivariate relationships. Plots for this exploratory work are in the appendix at the end of the document; all use the definition of "accuracy" defined a few paragraphs above.

- 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. (Visualization for that relationship is just below these bullet points. Plots for the other bullet points can be found in the appendix, along with corresponding test statistics.)
- 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.

Bidding process (format).



The implication of project size may provide the most valuable results. One might have expected the largest or smallest projects to be the toughest to predict - particularly the very smallest 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 second lowest bid information, that might be closer to the low bid. (Note: when projects only attracted one bid, that bid is used instead of a second-lowest bid. There are few of these projects in the data set.)

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 introduce some dual (reverse) causality, which could confuse models. Note: 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. And, just for modeling purposes, remove the accuracy metric. One would split randomly with cross-sectional data and since this effort aims to help predict temporally, split by date — use projects (n=213) from 2015 through the first quarter of 2019 to train models and projects (n=26) since the end of March to test them. That's not many projects for testing but the real test will come as another quarter or two of data rolls in.

Choose a handful of potential predictors and build a linear model. Go with the original estimate (always included) and (1) regional construction-specific employment, (2) bidding process (public or other), (3) typeology (infrastructure, building, or paving), (4) construction permits (lagged by one quarter), (5) construction costs (indexed), and (6) project size, with project dollars (from the second bid) categorized in deciles.

```
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
## -71511141   -965914    101910    928757  36057047
##
## Coefficients:
##                                 Estimate     Std. Error t value
## (Intercept)                 6480575.50070 18102711.09948   0.358
## Engr.Est                          0.92614        0.01558  59.460
## Employment.in.construction  -145346.86493   162525.05467  -0.894
## Format2Public               -1087152.85330  1048463.94815  -1.037
## TypeologyInfra              -1884496.88779  1196782.47632  -1.575
## TypeologyPaving             -1750022.03616  1469411.68848  -1.191
## permits_1                         34.58965       99.17871   0.349
## cci                             4503.02581     5979.35994   0.753
## quantile.L                   1615261.82377  1859239.37672   0.869
## quantile.Q                   -925988.08752  1801313.89445  -0.514
## quantile.C                   -787744.58376  1652165.94976  -0.477
## quantile^4                  -1459907.57205  1578328.58201  -0.925
## quantile^5                  -1612438.69112  1526686.15560  -1.056
## quantile^6                  -1520076.57709  1490945.78988  -1.020
## quantile^7                   -769296.64776  1507402.60790  -0.510
## quantile^8                   -318579.79071  1497814.39145  -0.213
## quantile^9                    125894.52031  1512632.20799   0.083
##                                    Pr(>|t|)
## (Intercept)                           0.721
## Engr.Est                  <0.0000000000000002 ***
## Employment.in.construction            0.372
## Format2Public                         0.301
## TypeologyInfra                        0.117
## TypeologyPaving                       0.235
## permits_1                             0.728
## cci                                   0.452
## quantile.L                            0.386
## quantile.Q                            0.608
```

```
## quantile.C                                0.634
## quantile^4                                0.356
## quantile^5                                0.292
## quantile^6                                0.309
## quantile^7                                0.610
## quantile^8                                0.832
## quantile^9                                0.934
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6880000 on 196 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9709
## F-statistic: 443.2 on 16 and 196 DF,  p-value: < 0.00000000000000022
```

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.

What is the summary of the predicted values? How does it compare to the summary of second-lowest bids? Looking at the two summaries, the original engineering estimate and the enhanced prediction do appear to be significantly (statistically) different.

.

Maybe something more robust can come with a little creativity.

4b. Shrinkage and automated variable selection.

Traditional methodology — tossing variables that made intuitive sense into a model and evaluating outcomes — might have identified a few reliable predictors. But it fell short and, even if it had, specification problems and covariate bias would have challenged the results. A survival method may help — automated variable selection can weed out weaker variables and identify one or more key items to help prediction. Regularization processes penalize variables that threaten to introduce more uncertainty than predictive power to a model. Regularization modeling builds atop ordinary least squares regression, which fits a curve to data that minimizes the distance between the curve and any given point in the data set. Traditional least squares modeling can be prone to overfitting — reading too much signal from what is essentially meaningless noise and providing a tool that creates a poor fit for data that, while tied to the underlying process being modeled, was absent during the fitting process; this presents obvious challenges for prediction.

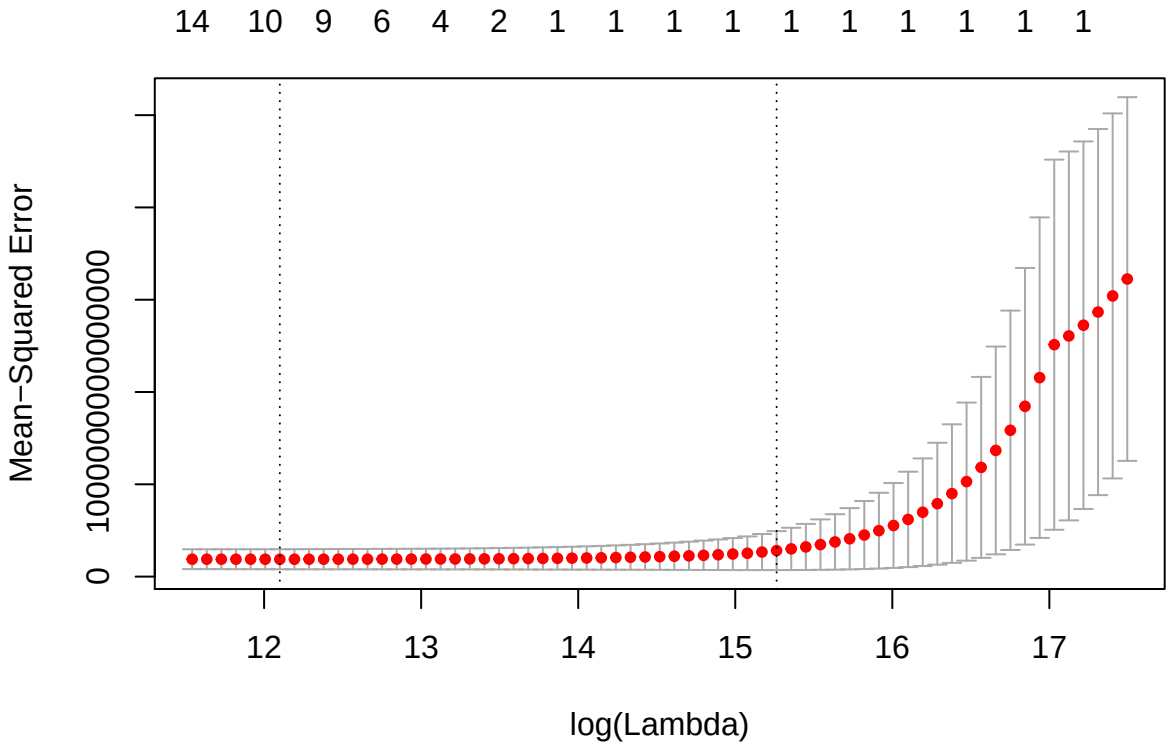
Regularization begins with the least squares method and adds a penalty. That penalty term is guided by a tuning parameter that essentially works as a dimmer switch — it can be cranked up to increase the penalty or down all the way to zero, which effectively turns off the penalty and produces the same results one would get from ordinary least squares. The tuning parameter is represented by the lambda in the second equation for error estimation (both from Hastie and Tibshirani):

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

Run cross-validation to pick the tuning parameter. Cross-validation runs modeling processes iteratively by resampling the data — splitting it into a series of tranches used systematically as training and test sets. The fit with the lowest test error identifies the tuning parameter:

Note: to help here with manageability and reproducibility, adopt some of the modeling software's basic assumptions (10 folds, standardized coefficients, use of mean squared error as an evaluation metric, et cetera).



The penalty term associated with an optimal tuning parameter (associated with the lowest model error) coincides with a model suppressing all but a handful of predictors.

This regularization process actually carries the prospect of increasing bias and, in the process, dulling coefficients' reflection of the real world relationship between the processes being modeled. (It can introduce or exacerbate bias.) Regularization is often run with a penalty that, prior to tuning, sums squared coefficients or their absolute values, the second of which is responsible for forcing weak predictors completely out of the model. Performing analysis with the latter shrunk all covariates (aside from the engineer's estimate) to zero, suppressing them completely. This is a harsh variable selection process but it might've been expected given the original estimates' proven strength. It doesn't leave enough clues for anyone to offer potential improvements. A little fine-tuning to the penalty combines the regularization treatments and gives the model a little leash, identifying predictors that survive as the model gets further from ordinary least squares without disappearing:

```
##
## Call:  glmnet(x = x, y = train2[, 5], alpha = blend, lambda = cv$lambda.min)
##
##      Df  %Dev Lambda
## [1,] 10 0.973 179900
```

Predict using the model and the test data.

4c. Evaluate regularization.

Compare this error with a basic OLS regression using all potential covariates; regularized interations can be compared with it. Turning the tuning parameter (the dimmer switch) off makes it easy.

Errors (test) for regularization and ordinary models, respectively:

```
## [1] 187517775220584
```

```
## [1] 221388743444856
```

Regularization not only reduced error for the model built using pre-Q2 2019 projects, but that model then provides a stronger fit using the withheld projects. And consider the implications for interpretability. Regress using the full data set:

```
## 45 x 1 sparse Matrix of class "dgCMatrix"
##                                                          1
## (Intercept)                                -74760508.0874747
## Engr.Est                                          0.8723463
## Loc                                          317669.4999643
## LD                                                        .
## Typeology                                   -490781.8904814
## 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                     6244.5720341
## Employment.in.other.services                              .
## Employment.in.production.industries                       .
## Employment.in.professional.services                       .
## Employment.in.real.estate                                 .
## Employment.in.retail                        -39231.1181364
## Employment.in.transport.services                          .
## Employment.in.wholesale                                   .
## Output.in.communications                                  .
## Output.in.construction                       -635.3690785
## Output.in.financial.services                  160.9802708
## Output.in.government                         3367.3157553
## Output.in.retail                                          .
## Output.in.education.and.health                            .
## Output.in.financial.and.business.services                 .
## Output.in.other.services                                  .
## Output.in.production.industries               -92.7091837
## 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                                  -1626498.6787224
## quantile                                   375622.5665574
```

5. Discussion.

This suggests some room for targeted efforts to add value to the average engineer's estimate, since it already correlated very highly with targets. The most obvious predictor is project size, which isn't too surprising given the exploratory work done earlier. Next steps include looking for best practices regarding post-estimation adjustments for project cost by size. Here are bivariate, uncontrolled relationship parameters between bids and project size:

```
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))
```

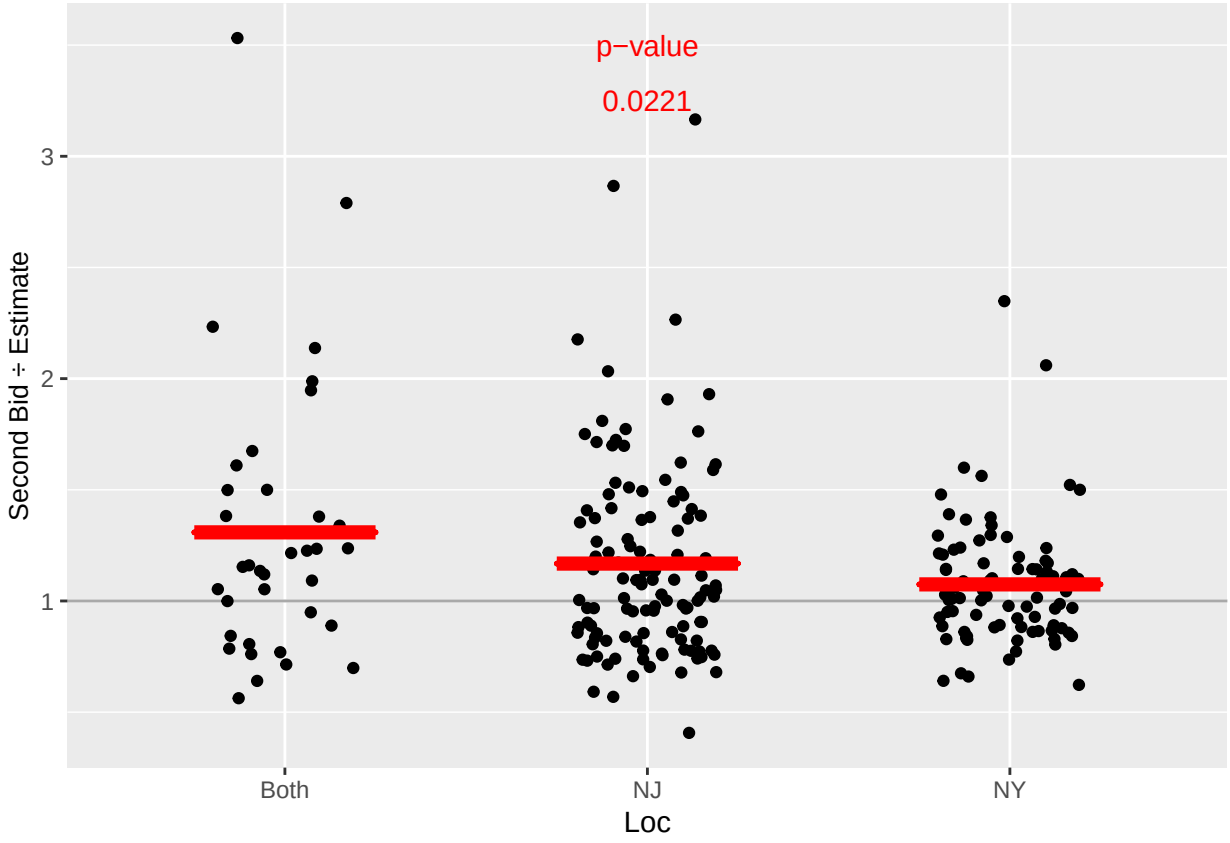
```
##
## Call:
## lm(formula = bids$Second.Bid ~ bids$Engr.Est + bids$quantile_2 +
##     bids$quantile_3 + bids$quantile_4 + bids$quantile_5 + bids$quantile_6 +
##     bids$quantile_7 + bids$quantile_8 + bids$quantile_9 + bids$quantile_10)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -64927474   -490838     31601    573534  56729942
##
## Coefficients:
##                       Estimate    Std. Error t value            Pr(>|t|)
## (Intercept)        76048.34661 1598483.07814   0.048               0.9621
## bids$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
```

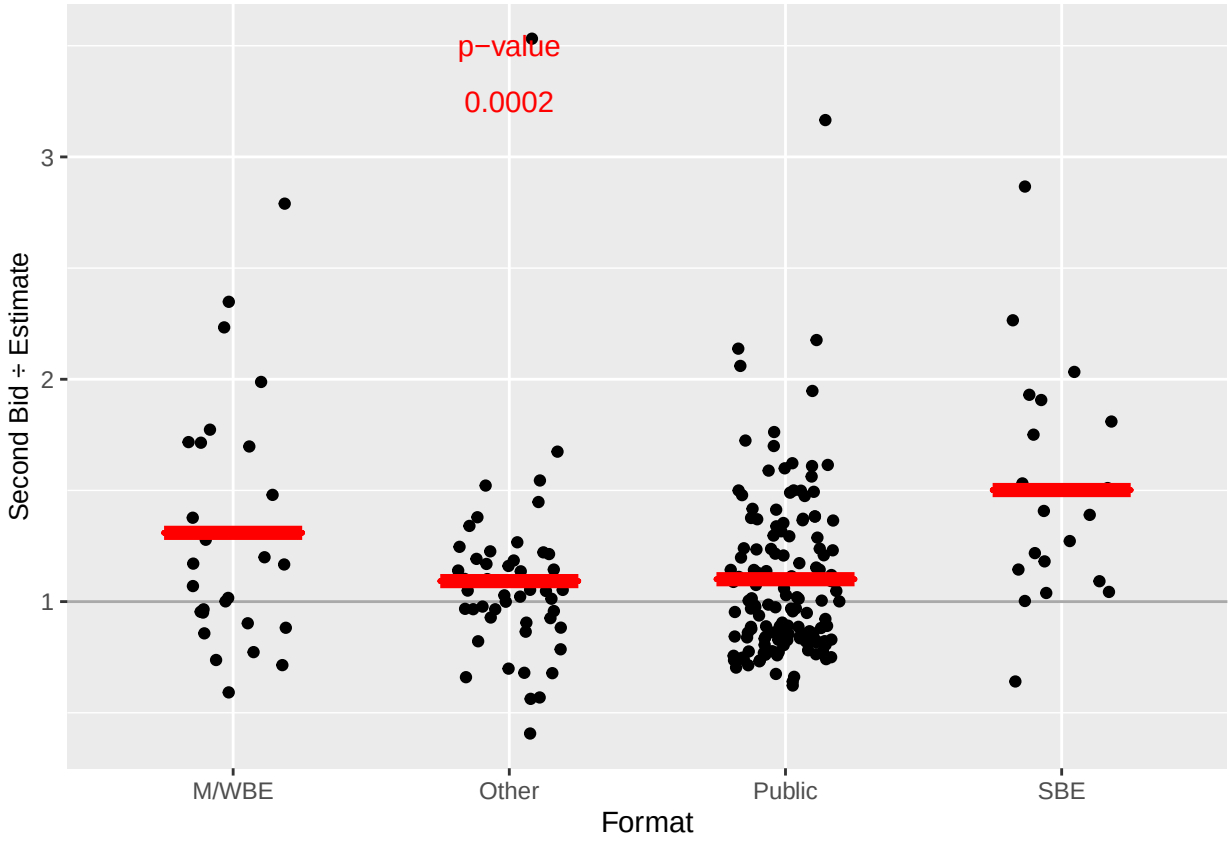
Appendix: exploratory work.

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.

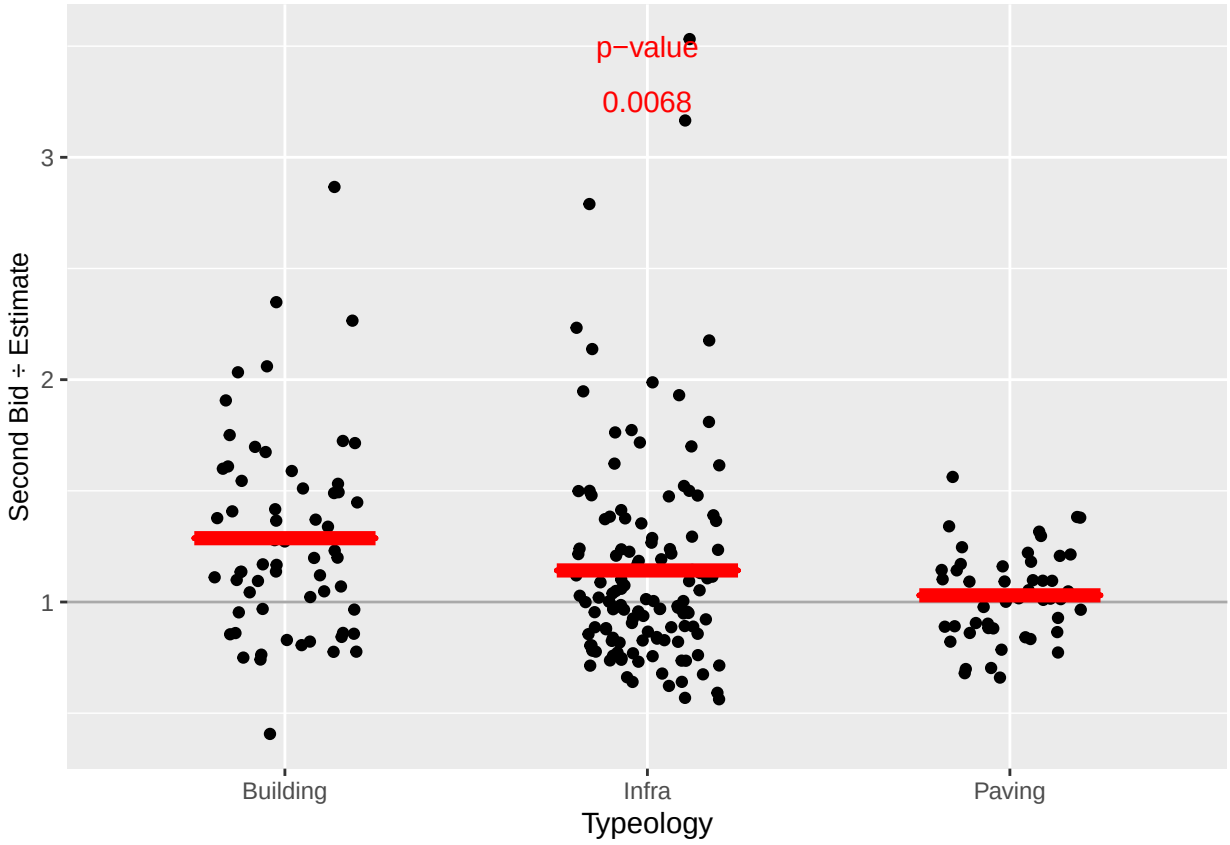
Location.



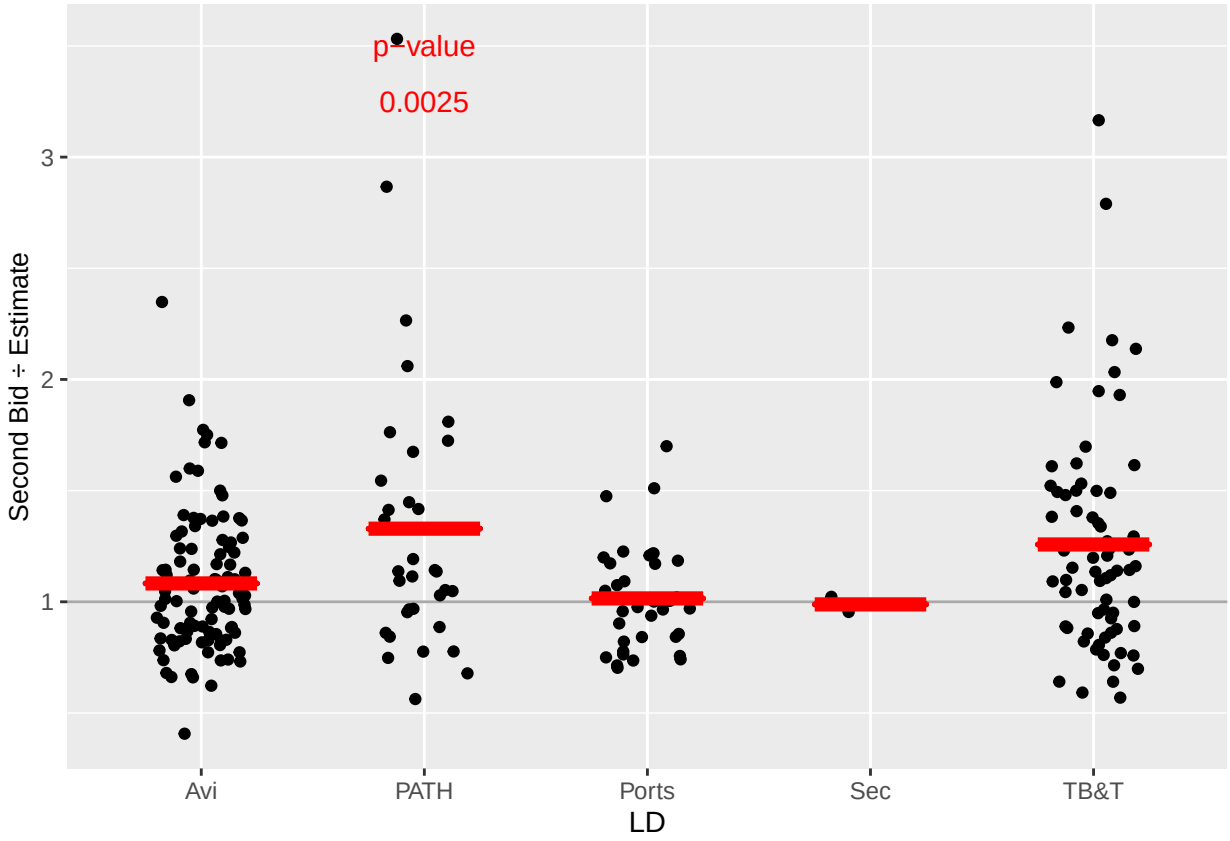
Bidding process ("format").



Project typeology.



Department.


p-value
0.0025
Second Bid ÷ Estimate
3
2
1
Avi
PATH
Ports
Sec
TB&T
LD

Project size (dollars bid).