Predicting costs

11/06/2019

1. Summary

The Port Authority estimates project costs internally prior to bidding. Its accuracy varies from project to project. The statistical analysis described here indicates two factors — project size and bidding process — explain a noteworthy degree of accuracy. The agency may be able to more accurately predict costs by adjusting initial estimates to account for those two variables.

We considered a number of data sets including project-level characteristics, construction costs, and regional and national indicators specific to the economy, labor market, and construction activity.

Discussion is broken down into: - Overview of data sets, - Exploratory analysis, - Methodology, and - Discussion and next steps.

The appendix includes more detailed findings from the exploration and full model results.

2. Data

Project-level characteristics, referred to here as the "bids" data set, 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 permitting data comes from the City of New York's Department of Buildings. The Engineering News-Record's construction cost index (CCI) serves as proxy for broader construction costs. Data not specific to project, such as economic variables and permitting, is quarterly.

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

One variable of interest prior to this analysis is the actual bidding process itself. Institutional discussions and earlier modeling work suggested the bidding process may increase real (actualized) project costs. This is intuitive: limits placed on the range of bidders could, on average and holding other things constant, increase the average (and lowest qualifying) bid - this is basic supply and demand. The Engineering Department's data set included a range of categorical designations for bidding processes, and this analysis simplifies the 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.

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

Note: costs and estimation accuracy

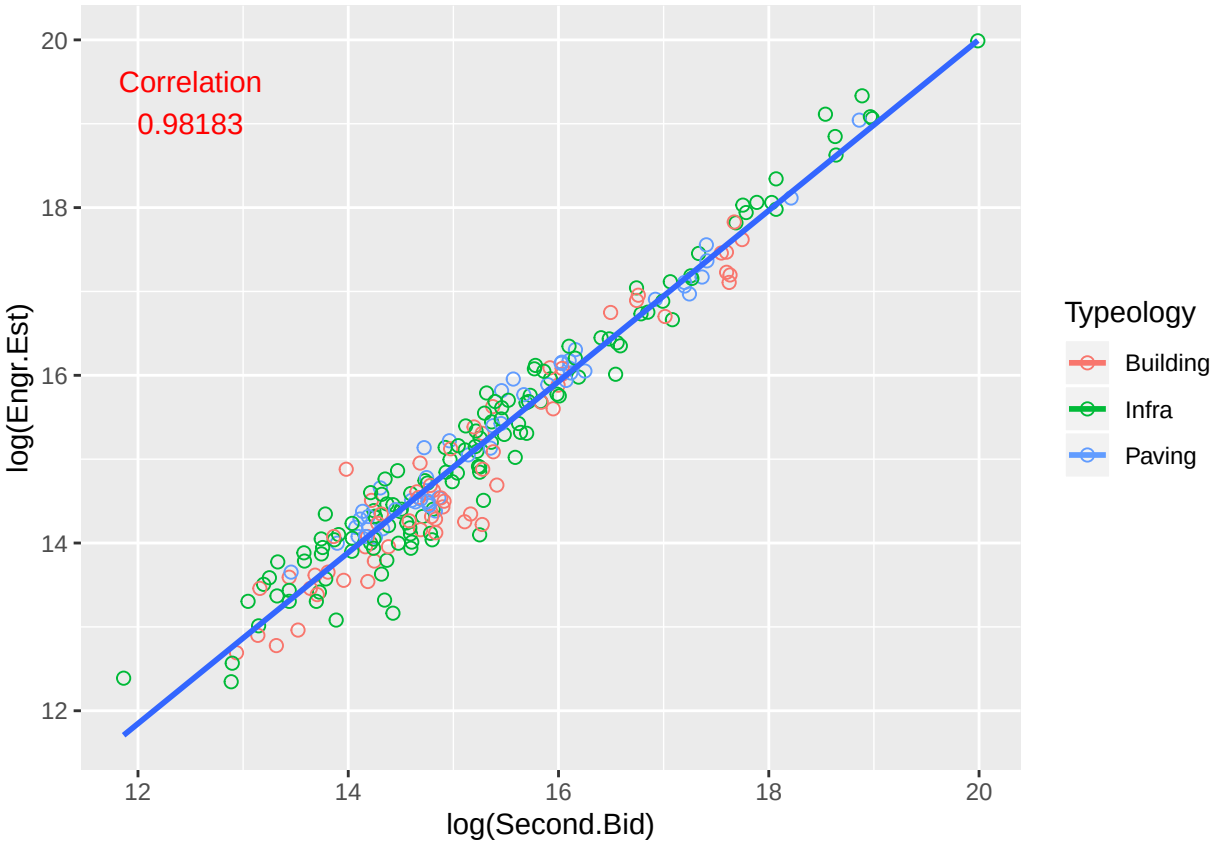
The agency's construction projects span months or years of time. Actual cost data does not exist for many of the observations, which at 239 projects already creates minor dimensionality concerns given the number of potential predictive variables. This analysis thus evaluates agency estimates as they compare to the second-lowest bid. This provides a metric — a reasonable target for evaluating internal estimates. The "accuracy" metric occasionally referenced through the exploratory section and appendix refers to a ratio of the second-lowest bid divided by the initial agency estimate, both in dollars. For example, an accuracy score of 1.0 would represent a case where the second-lowest bid (numerator) provided a near-precise match to the internal estimate (denominator); an accuracy of 0.94 would mean the second-lowest bid included 94 cents for every dollar estimated internally; et cetera.

Five projects attracted only one bid. They averaged only $3.2 million (second-lowest bid). In those cases, that bid is used in place of a second-lowest bid.

3. Exploratory analysis

Why might developing a conrolled multivariate model help improve predictions? The average gap between second-lowest bids and initial agency estimates is less than $900,000, or roughly 5 percent (the average project bid was around $15 million).

How much inconsistency does that represent?



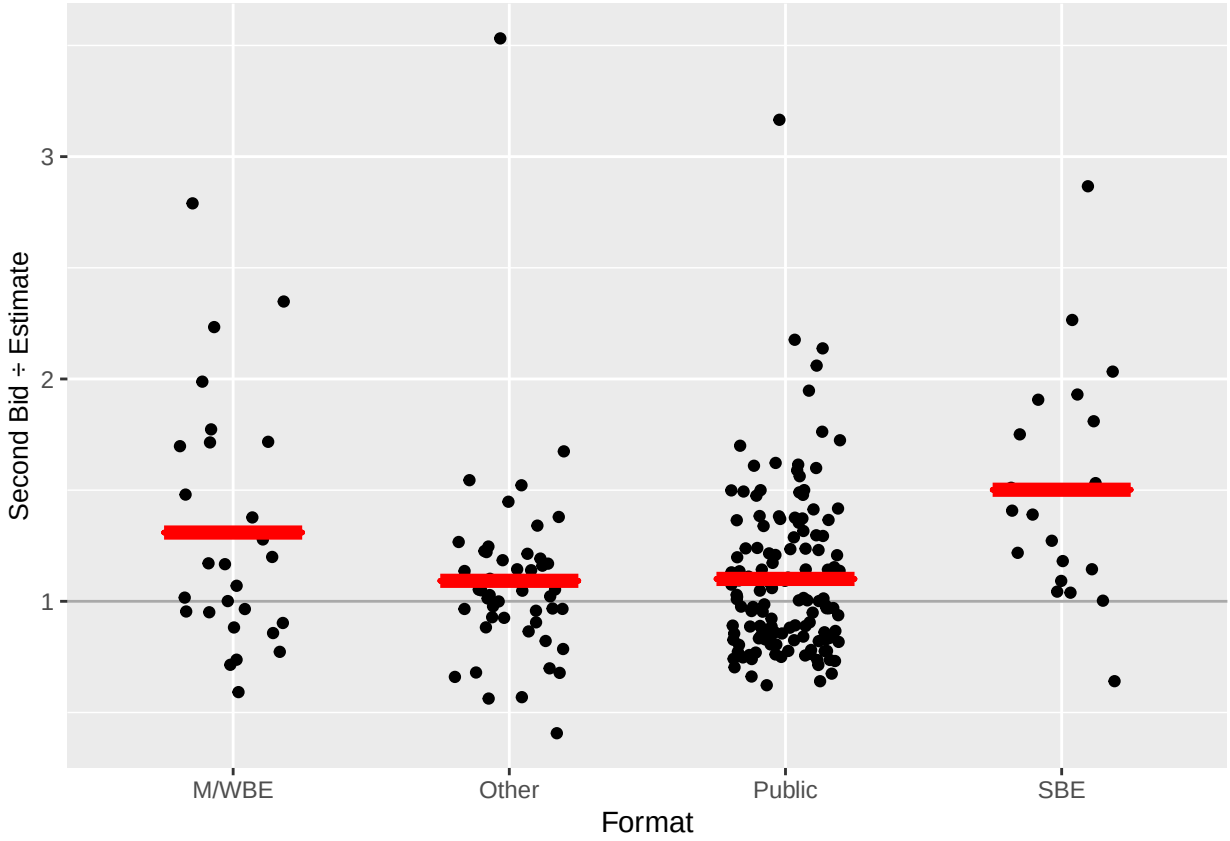
The agency's in-house engineering estimates predict more than 98% of the variation in second-lowest bids. Some of the remaining variation may be explained by institutional guesswork or targeted considerations, as discussed below.

Uncontrolled variable-to-variable relationships provide easy clues as to different data points' potential role in any prediction exercises. Plots for this exploratory work are in the appendix; all use the definition of "accuracy" defined a few paragraphs above. A few things stand out as promising predictors before getting to modeling:

- Location: 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 and appear to be slightly tougher to predict;
- Bidding (see plot below): the bidding process can be constrained or open, with potential ramifications on the ability to estimate costs.
- Typeology: project type (infrastructure, building, pavement) has an identifiable, if yet uncontrolled, relationship with estimating accuracy. Of those three categories, the agency appears to predict paving

projects the best. (Note of interest: the story changes when considering the lowest bid, where bidding paving projects varies significantly.)

- Variance: outliers for PATH and TB&T push the average accuracy metric up for each, otherwise there's little apparent difference across departments.
- Size: strong signals emerge that projects of medium size (in dollar terms) may be more evasive than much larger or smaller projects.



Project size and bidding format provide strong one-to-one signals of predictive power and 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 seems to be the projects ranging from the 40th to 70th percentiles that are the most challenging, although this insight may change subsequent to review by and feedback from the Engineering Department.

4. Modeling and prediction

This section provides a technical discussion of statistical modeling and is included as documentation for reproducibility and transparency. Supporting code and data sets have accompanied delivery of this report and are otherwise available.

The goal of this project is to enlist the help of various data points to try and enhance internal agency estimates of project cost. The data set carries dimensionality challenges given the number of variables (absolutely and relative to the number of observations). It includes a mix of continuous variables and qualitative factors, both ordinal and nominal and all treated categorically without conversion to binary constituencies — the two modeling processes (OLS and regularization) employed here do that automatically.

4a. Base model (manual selection and linear regression)

Model specification can be manual and intuitive. Given the fact that estimators' already try and take many of the variables considered here into account when they're doing their work, a model with even a handful of extra covariates could overfit the test data.

Note: we ensure the accuracy variable calculated earlier is dropped before modeling or it would introduce some dual (reverse) causality, which could confuse models.

Split data into training / test sets. One would split randomly with cross-sectional data but 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) from March through today to test. That's not many projects for testing but the real test will come as another quarter or two of data arrive.

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 + quantile, data = train)

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
## -0.000000136689 -0.000000001608  0.000000000461  0.000000002493
##             Max
##  0.000000025253
##
## Coefficients:
##                                        Estimate              Std. Error
## (Intercept)                -0.000000030630322240 0.000000027534193668 66
## Engr.Est                    1.000000000000000220 0.000000000000000020 17
## Employment.in.construction  0.000000000858768746 0.000000000246019676 12
## Format2Public               0.000000001595600791 0.000000001554100752 92
## TypeologyInfra             -0.000000000251773208 0.000000001802514403 94
## TypeologyPaving            -0.000000002177508062 0.000000002202503445 02
## permits_1                  -0.000000000000066888 0.000000000000150639 43
## cci                        -0.000000000026898590 0.000000000009039077 23
## quantile.L                 -0.000000000958206091 0.000000001860068125 21
## quantile.Q                  0.000000001588807492 0.000000001754402271 40
## quantile.C                  0.000000002049779860 0.000000001633778523 957
## quantile^4                  0.000000001563302283 0.000000001619117309 63
##                                       t value            Pr(>|t|)
## (Intercept)                            -1.112            0.267275
## Engr.Est                  4958766119263184 8.000 < 0.0000000000000002 ***
## Employment.in.construction             3.491            0.000592 ***
## Format2Public                          1.027            0.305794
## TypeologyInfra                        -0.140            0.889054
## TypeologyPaving                       -0.989            0.324022
## permits_1                             -0.444            0.657497
## cci                                   -2.976            0.003281 **
## quantile.L                            -0.515            0.607017
## quantile.Q                             0.906            0.366226
## quantile.C                             1.252            0.212186
## quantile^4                             0.966            0.335441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00000001051 on 201 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.254e+32 on 11 and 201 DF,  p-value: < 0.00000000000000022
```

The equation above throws a decent amount of information at the engineering estimate to try and enhance its predictive accuracy; the result might be called an "enhanced" prediction.

Some regularization may be able to do better by selecting variables for us.

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 omitted variable 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. It builds atop ordinary least squares regression, which fits a curve 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 (from James, Witten, Hastie and Tibshirani, 2013):

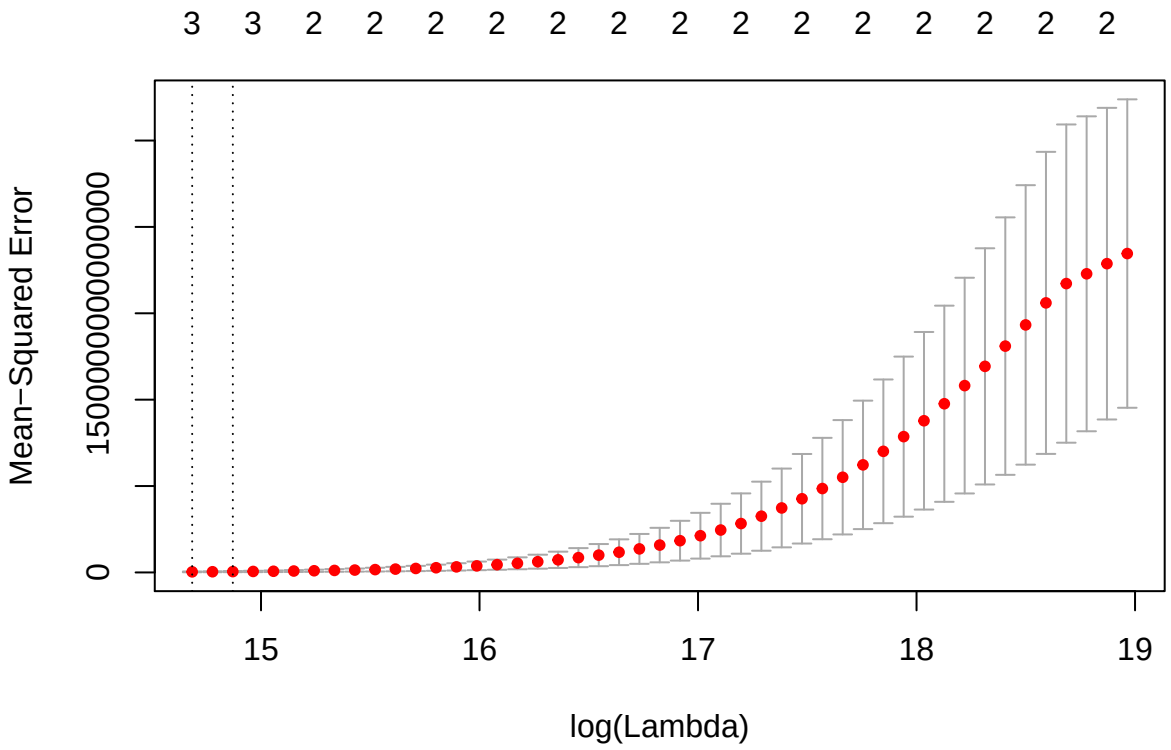
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

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

Run cross-validation to pick the tuning parameter. Cross-validation uses iterative resampling — splitting the data into a series of tranches used systematically as training and test sets. The fit with the lowest test error identifies a cut-off point regarding reliable predictive variables, which in tune guides selection of 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).

This analysis uses a blended regularization penalty that falls between ridge and lasso (the "alpha" is set at 0.25):



The tuning parameter that gets chosen isn't actually the one associated with the model exhibiting the lowest test error, but with the one coinciding with a sister model carrying the lowest test error plus a standard error. If a few models' test errors are with a standard error of one another, I default to the simplest of the group.

The result suppresses all but a pair of predictors: project size and bidding process.

(Very technical note: 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.) This harsh variable selection process is appropriate given the original estimates' obvious raw predictive strength. A little fine-tuning to the penalty combines the regularization treatments and gives the model just enough leash to identify predictors that survive as the model gets further from ordinary least squares without disappearing.

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 ordinary model and regularized model, respectively:

```
## [1] 0.00000000000000001297269
```

```
## [1] 5884507017053
```

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. The second model is also much easier to interpret. Now regress using the second model and the full data set:

5. Discussion.

This suggests some room for targeted efforts to add value to the average engineer's estimate, which already correlates very highly with targets. The most obvious predictor is project size, which isn't too surprising given the exploratory work done earlier. The bidding process, or "format" using the Engineering Department's terminology, was the other predictor to survive the modeling process.

Both project size and format represent avenues for post-estimation adjustments that could improve accuracy.

Next steps include looking for best practices regarding post-estimation adjustments for project cost by size and/or bidding process. Relationship parameters between bids and project size:

Project size and bidding process

```
summary(lm(bids$Second.Bid ~ bids$Engr.Est + bids$quantile_2 + bids$quantile_3 +
    bids$quantile_4 + bids$quantile_5 + bid$Format2))
```

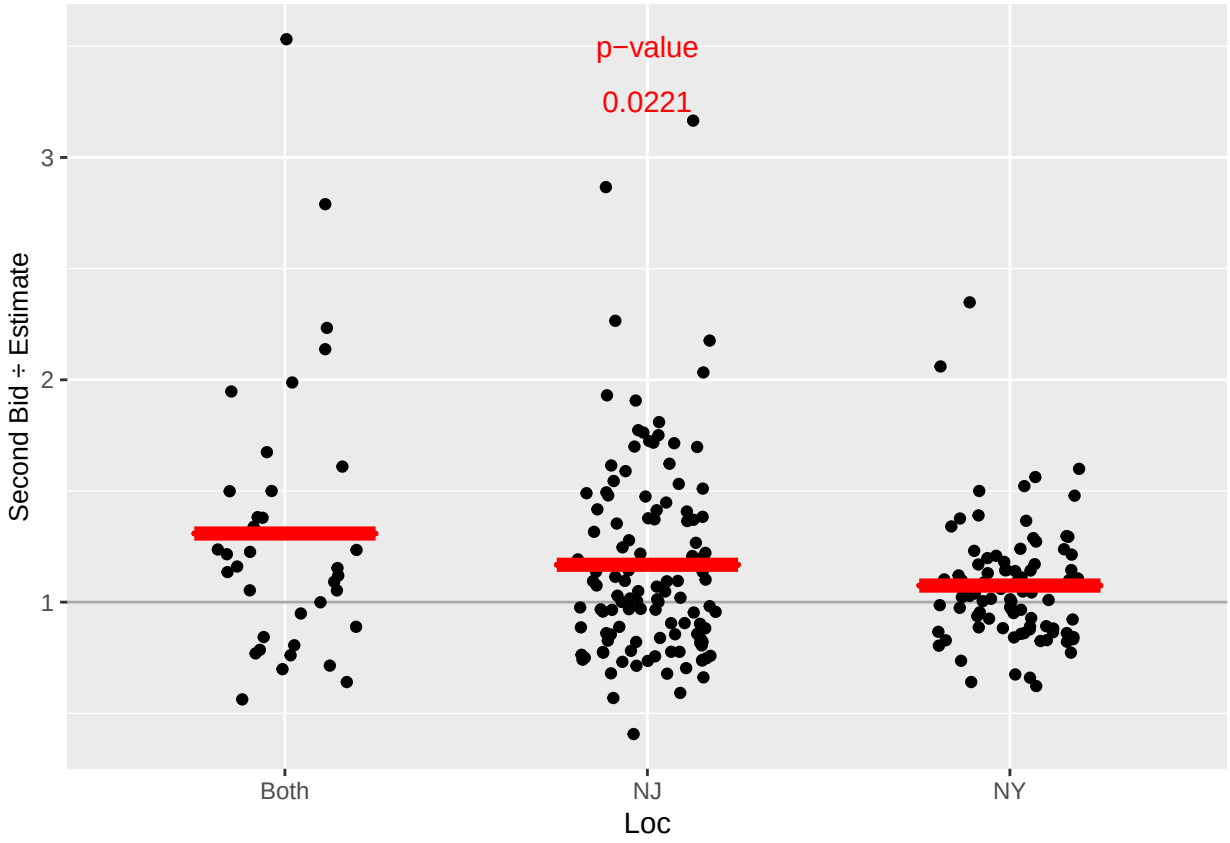
```
##
## Call:
## lm(formula = bids$Second.Bid ~ bids$Engr.Est + bids$quantile_2 +
##     bids$quantile_3 + bids$quantile_4 + bids$quantile_5 + bid$Format2)
##
## Residuals:
##            Min             1Q         Median             3Q            Max
## -0.00000002795 -0.00000000416 -0.00000000149  0.00000000248  0.00000044173
##
## Coefficients:
##                                Estimate              Std. Error
## (Intercept)        0.00000001156651188171  0.00000000457319979977
## bids$Engr.Est      0.99999999999999933387  0.00000000000000005119
## bids$quantile_2    0.00000000064270036179  0.00000000605565534973
## bids$quantile_3    0.00000000136420304818  0.00000000612408561939
## bids$quantile_4    0.00000001099185172539  0.00000000614823692750
## bids$quantile_5    0.00000000040097118746  0.00000000697290103951
## bid$Format2Public -0.00000000579095111839  0.00000000397499034427
##                                  t value            Pr(>|t|)
## (Intercept)                        2.529              0.0121 *
## bids$Engr.Est     19533217454346460.000 <0.0000000000000002 ***
## bids$quantile_2                    0.106              0.9156
## bids$quantile_3                    0.223              0.8239
## bids$quantile_4                    1.788              0.0751 .
## bids$quantile_5                    0.058              0.9542
## bid$Format2Public                 -1.457              0.1465
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00000002953 on 232 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 9.321e+31 on 6 and 232 DF,  p-value: < 0.00000000000000022
```

Appendix: data summary, exploratory work, model outputs.

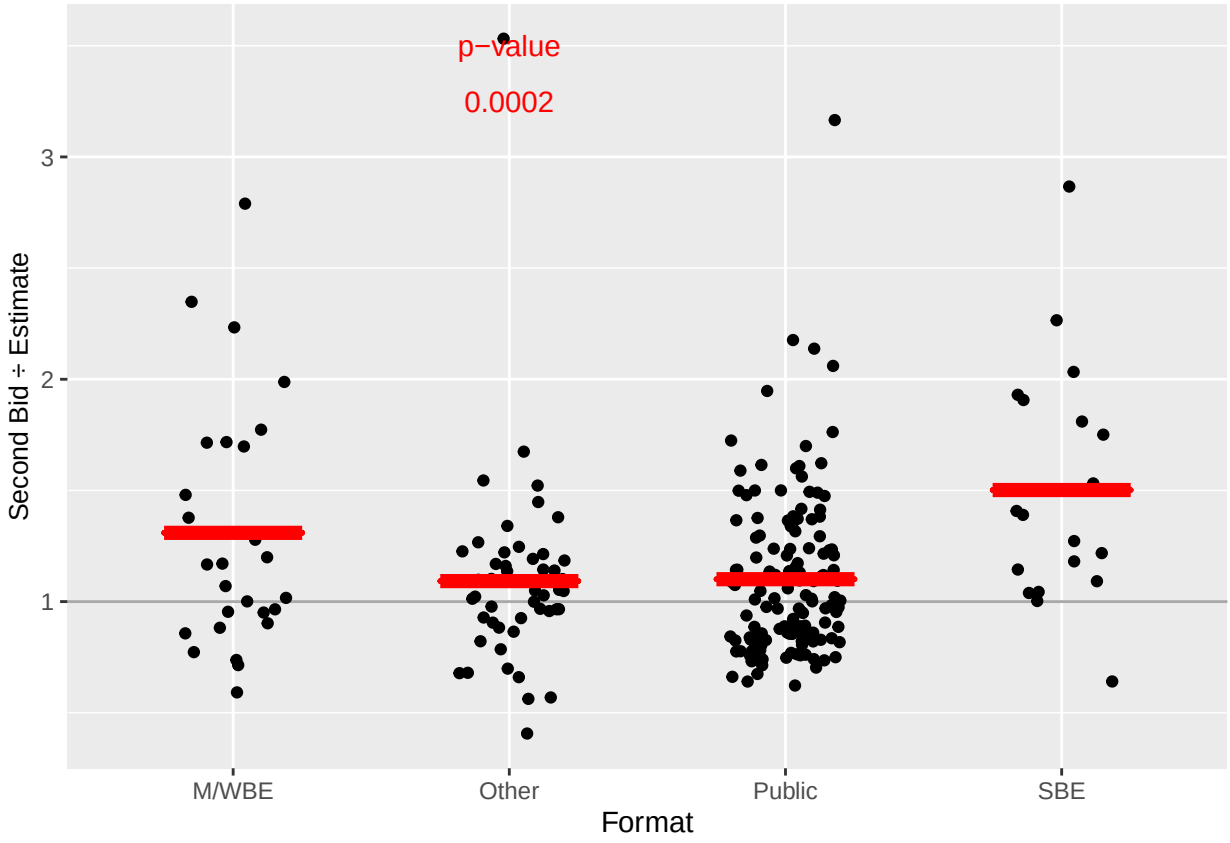
Data summary to come.

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.

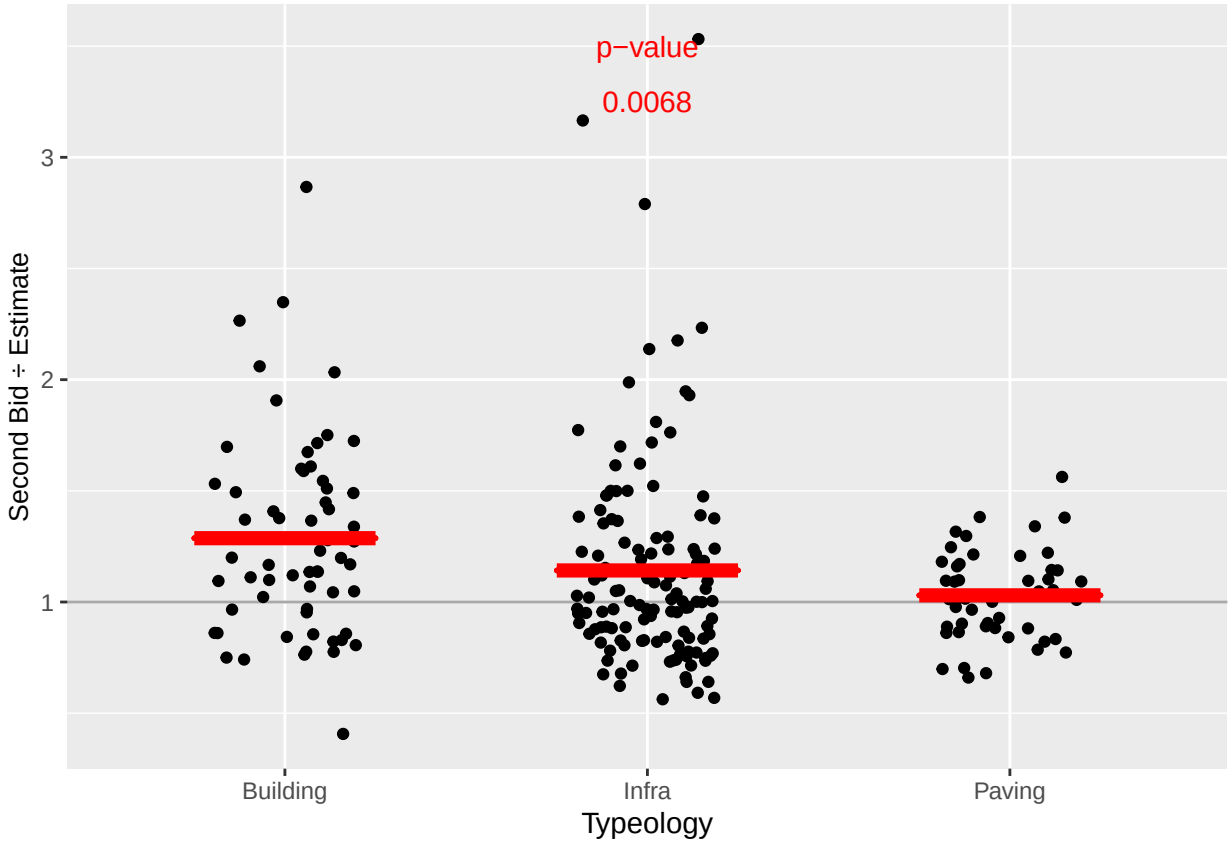
Location



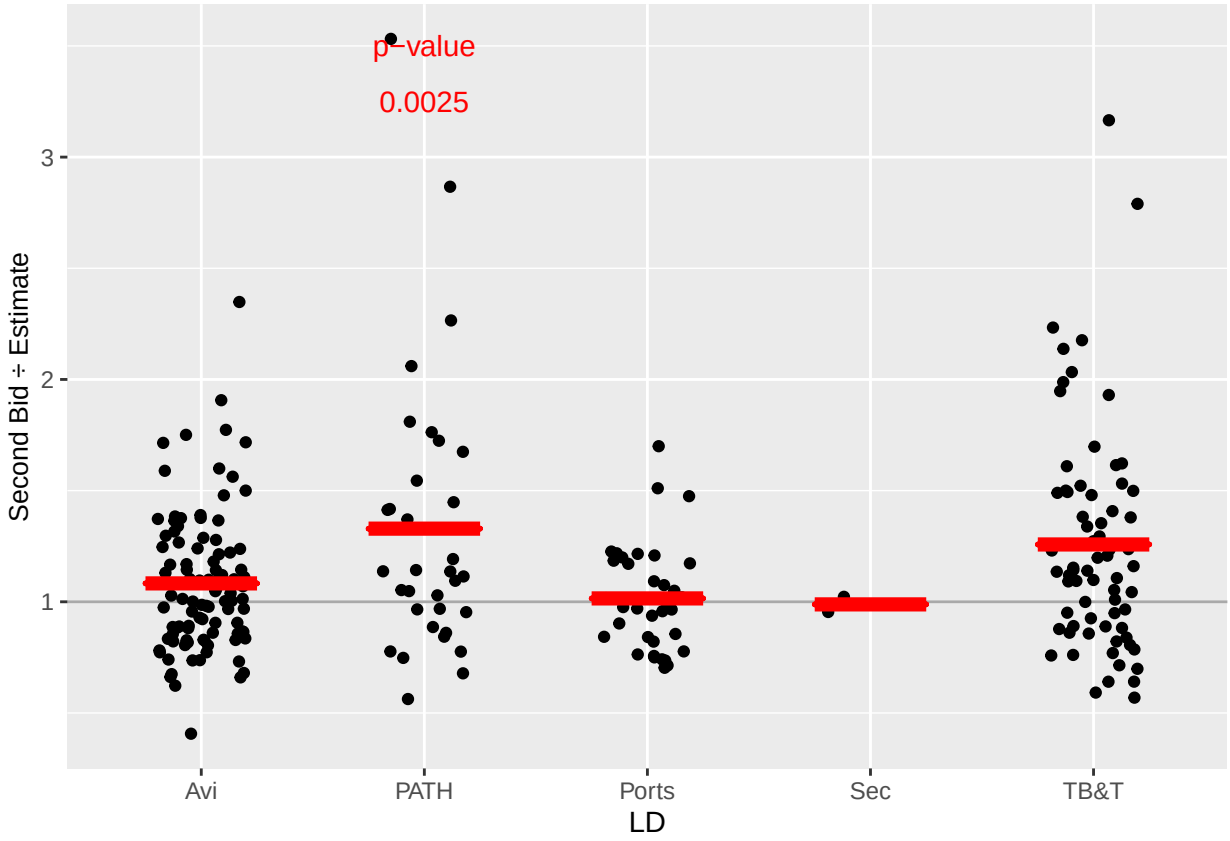
Bidding process (format)



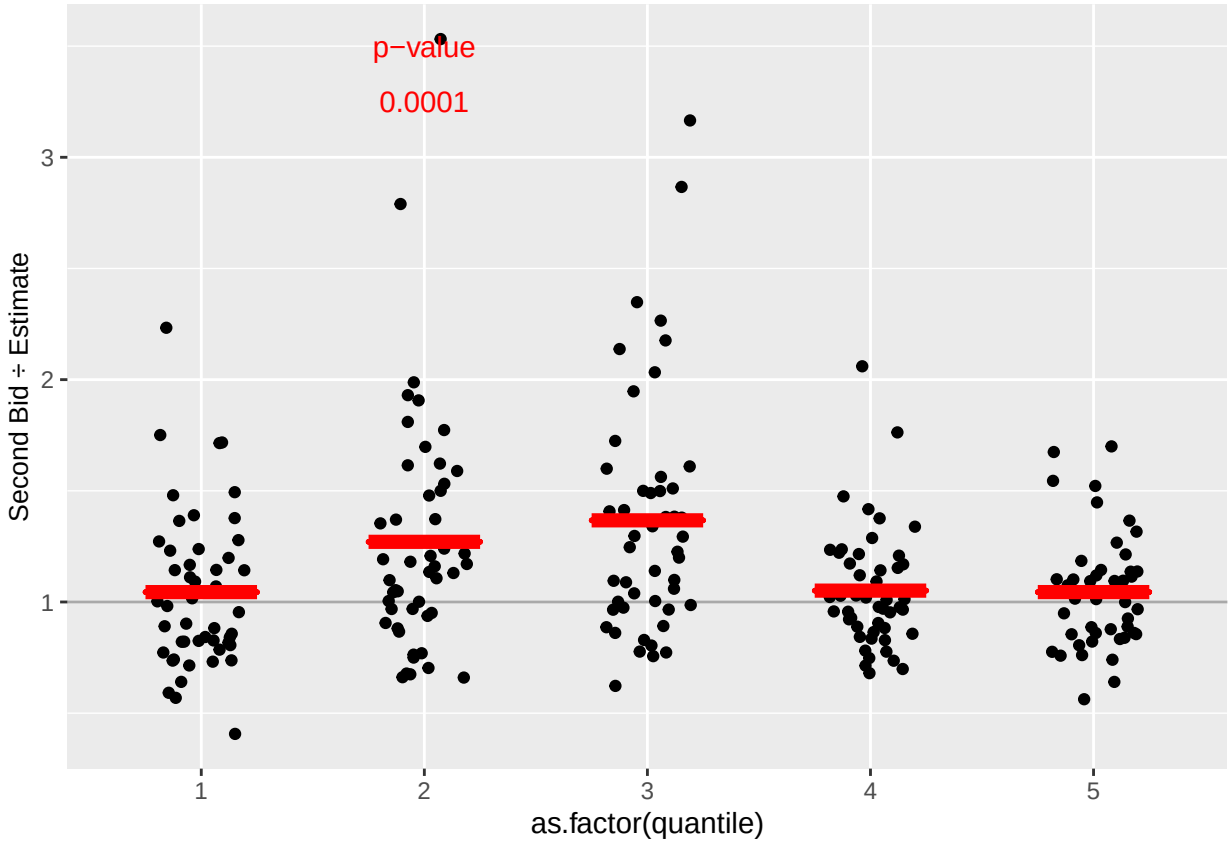
Project typeology



Department

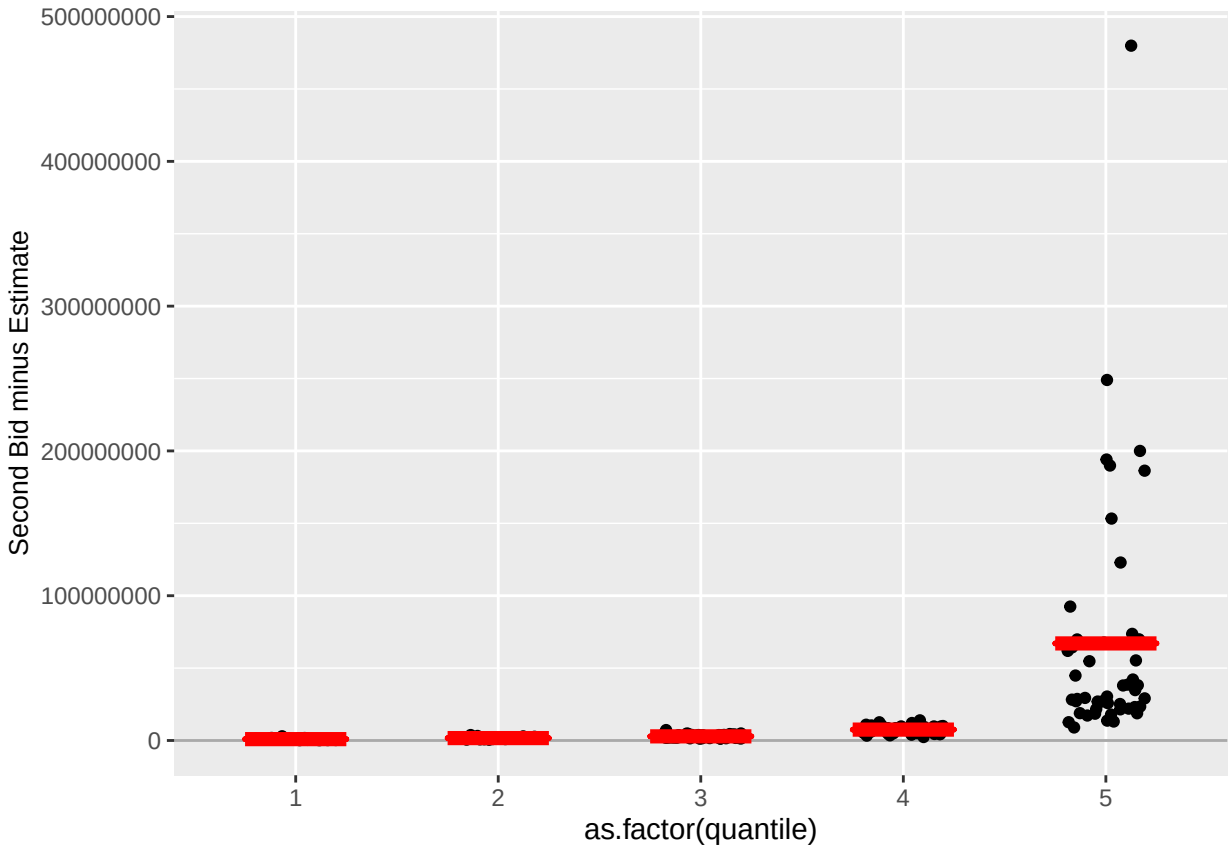

p-value
0.0025
Second Bid ÷ Estimate
3
2
1
Avi
PATH
Ports
Sec
TB&T
LD

Project size

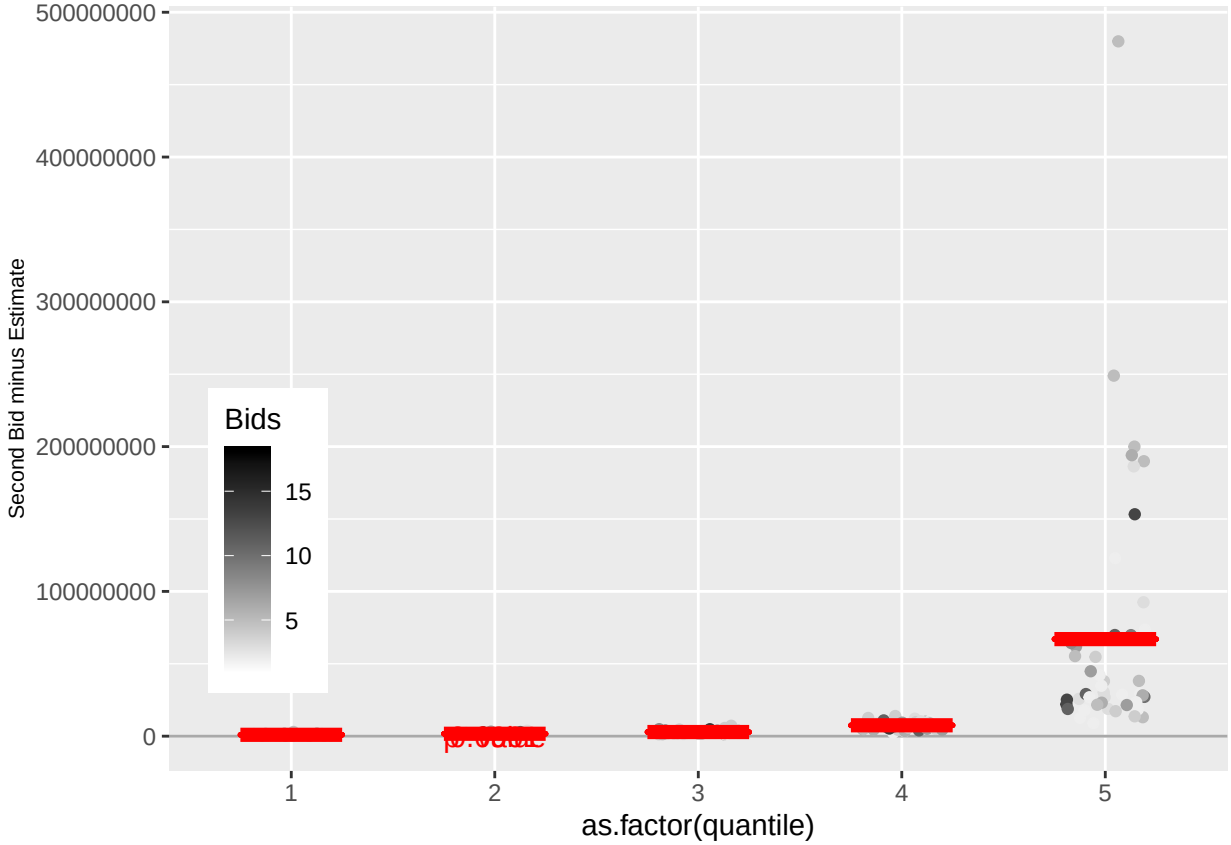


Project size (dollars bid)

```
ggplot(bids, aes(x = as.factor(quantile), y = bal)) + geom_hline(colour = "dark gray",
    yintercept = 1) + geom_jitter(width = 0.2) + geom_crossbar(data = decile.summary2,
    aes(ymin = bal, ymax = bal), size = 1, col = "red", width = 0.5) + ylab("Second Bid minus Estimate")
    theme(axis.title.y = element_text(size = 10))
```


500000000
400000000
300000000
200000000
100000000
0
Second Bid minus Estimate
1
2
3
4
5
as.factor(quantile)

Project size (dollars bid) with number of bidders



Output: OLS

```
##
## Call:
## lm(formula = Second.Bid ~ Engr.Est + Employment.in.construction +
##     Format2 + Typeology + permits_1 + cci + quantile, data = train)
##
## Residuals:
##             Min              1Q          Median              3Q
## -0.000000136689 -0.000000001608  0.000000000461  0.000000002493
##             Max
##  0.000000025253
##
## Coefficients:
##                                        Estimate              Std. Error
## (Intercept)                -0.00000003063032224003  0.00000002753419366866
## Engr.Est                    1.00000000000000022204  0.00000000000000002017
## Employment.in.construction  0.00000000085876874633  0.00000000024601967612
## Format2Public               0.00000000159560079121  0.00000000155410075292
## TypeologyInfra             -0.00000000025177320866  0.00000000180251440394
## TypeologyPaving            -0.00000000217750806212  0.00000000220250344502
## permits_1                  -0.00000000000006688873  0.00000000000015063943
## cci                        -0.00000000002689859019  0.00000000000903907723
## quantile.L                 -0.00000000095820609137  0.00000000186006812521
## quantile.Q                  0.00000000158880749204  0.00000000175440227140
## quantile.C                  0.00000000204977986081  0.00000000163778523957
## quantile^4                  0.00000000156330228354  0.00000000161911730963
##                                       t value             Pr(>|t|)
## (Intercept)                            -1.112             0.267275
## Engr.Est                   49587661192631848.000 < 0.0000000000000002 ***
## Employment.in.construction                3.491             0.000592 ***
## Format2Public                             1.027             0.305794
## TypeologyInfra                           -0.140             0.889054
## TypeologyPaving                          -0.989             0.324022
## permits_1                                -0.444             0.657497
## cci                                      -2.976             0.003281 **
## quantile.L                               -0.515             0.607017
## quantile.Q                                0.906             0.366226
## quantile.C                                1.252             0.212186
## quantile^4                                0.966             0.335441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00000001051 on 201 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.254e+32 on 11 and 201 DF,  p-value: < 0.00000000000000022
```

Output: Variable Selection

```
## 46 x 1 sparse Matrix of class "dgCMatrix"
##                                                        1
## (Intercept)                                   530639.1880413
## Engr.Est                                           0.4833214
## 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                                       28887.1111906
## bal                                                0.4776021
```