Predicting costs

10/25/2019

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.

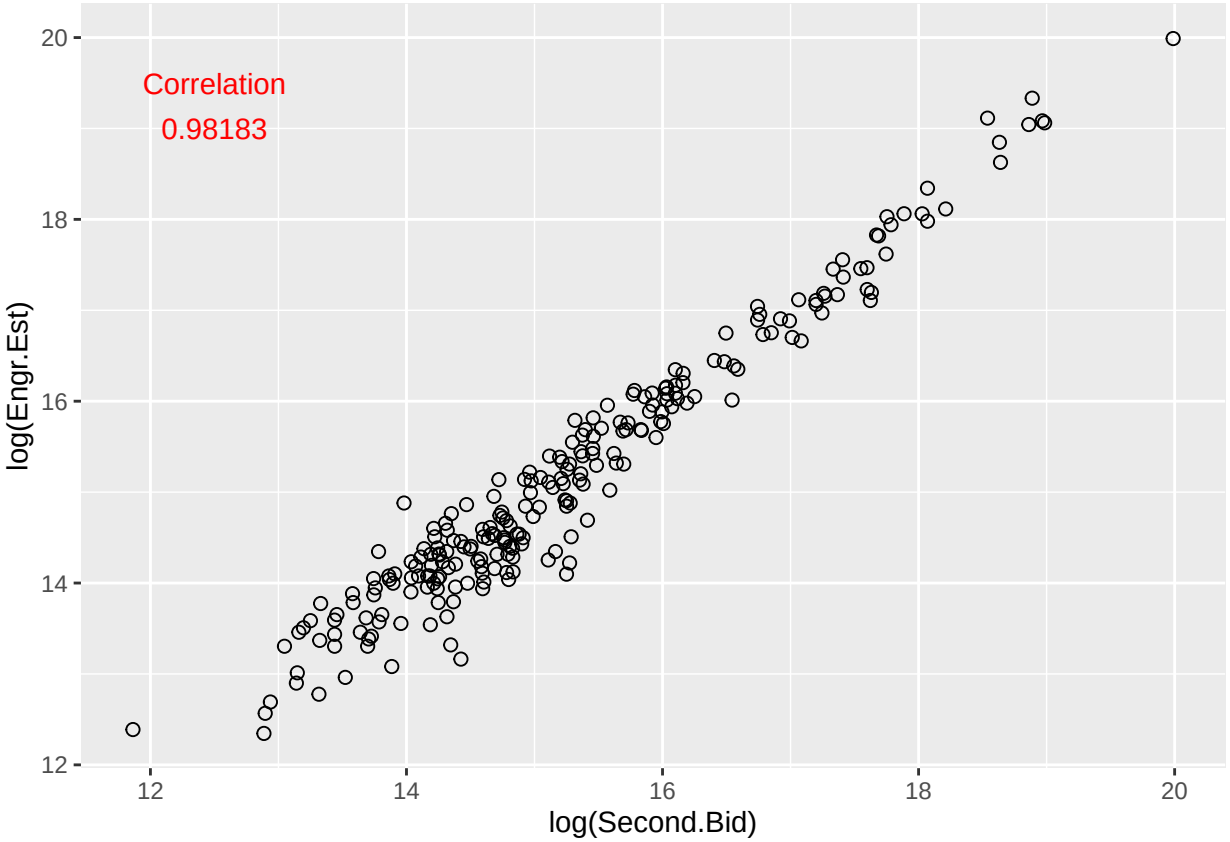
Agency projects last for months or years and actual costs do not exist for many of the observations, which at just over 260 projects already creates minor dimensionality concerns given the number of covariates. The second-lowest qualifying bid provides a reasonable target for evaluating internal estimates. 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. 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.

```
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 + c

  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 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:

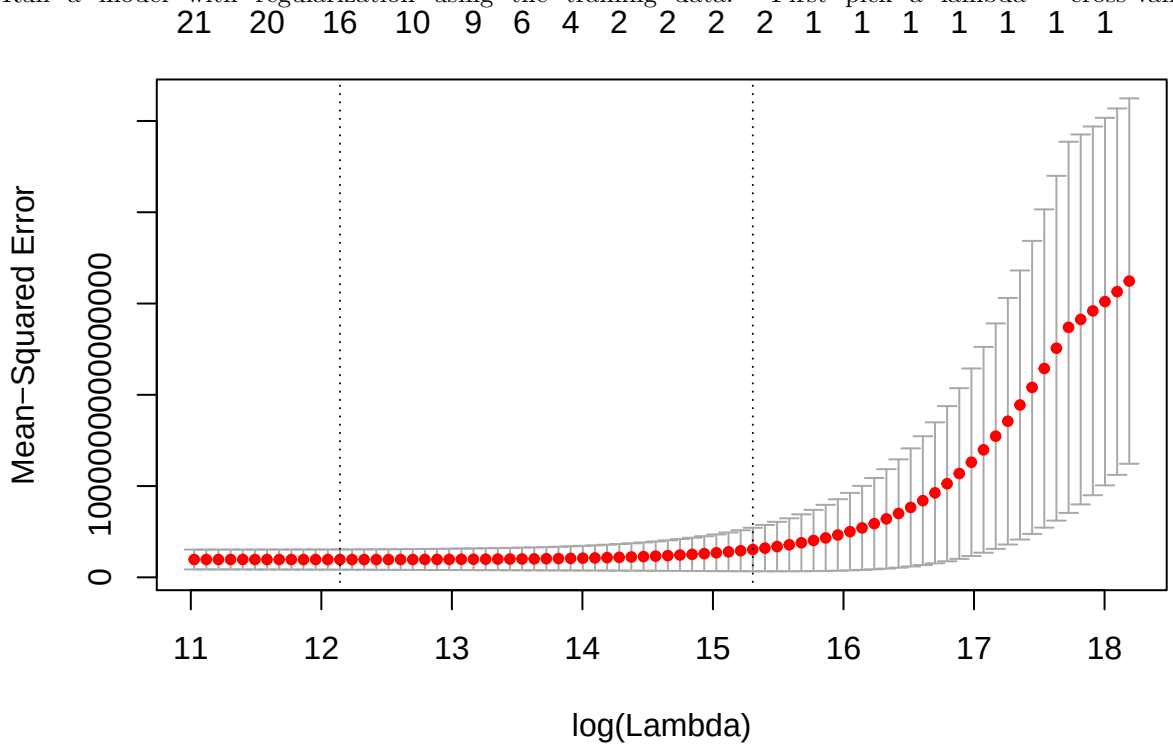
```
##
## Call:
## lm(formula = base$fitted.values ~ train$Second.Bid)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -23534799  -1210616   -386363    683955  74160211
##
## Coefficients:
##                      Estimate   Std. Error t value            Pr(>|t|)
```

```
## (Intercept)      377554.41688 474696.72204   0.795                0.427
## train$Second.Bid      0.97311      0.01114  87.375 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6541000 on 211 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.973
## F-statistic:  7634 on 1 and 211 DF,  p-value: < 0.00000000000000022
```

Maybe something more robust can come with a little creativity.

4b. Shrinkage and automated variable selection.

Run a model with regularization using the training data. First pick a lambda - cross-validation



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.

```
cv.model = glmnet(x, train2[, 5], alpha=0.5,lambda=cv$lambda.min)
coef(cv.model)
```

```
## 45 x 1 sparse Matrix of class "dgCMatrix"
##                                                      s0
## (Intercept)                                -278469010.7073543
## Engr.Est                                          0.9080085
## Loc                                          201909.8131642
## LD                                           203114.8000433
## Typeology                                   -650589.3004713
## Consumer.price.index                                      .
## Employment.in.communications                              .
## Employment.in.construction                                .
## Employment.in.education.and.health                        .
```

```
## Employment.in.financial.and.business.services          .
## Employment.in.financial.services                   88084.9848206
## Employment.in.government                          127086.7668630
## 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                            -2337.1634751
## Output.in.financial.services                        270.8294173
## Output.in.government                               2928.4752275
## Output.in.retail                                  -1365.6336785
## Output.in.education.and.health                         .
## Output.in.financial.and.business.services              .
## Output.in.other.services                           1001.8051982
## Output.in.production.industries                    -409.8175041
## Output.in.professional.services                        .
## Output.in.real.estate                                  .
## Output.in.transport.services                           .
## Output.in.wholesale                                 252.3127194
## 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                                        -1031445.3454324
## quantile                                         299333.6195180
```

```
cv.model
```

```
##
## Call:  glmnet(x = x, y = train2[, 5], alpha = 0.5, lambda = cv$lambda.min)
##
##      Df   %Dev Lambda
## [1,] 15 0.9734 187600
```

```
cv.predict = predict(cv, s=cv$lambda.min ,newx=data.matrix(test2[,-5]))
```

mean((cv.pred-test2$Second.Bid)^2)

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.

```
ols.model = glmnet(x, train2[, 5], alpha=0.5 ,lambda=0)
ols.predict = predict(ols.model, s=0, newx=data.matrix(test2[,-5]))
```

Some errors

```
mean((fitted.values(cv.model)-train2$Second.Bid)^2) # Regularized error, train
```

```
## [1] NaN
```

```
mean((cv.predict-test2$Second.Bid)^2) # Regularized error, test
```

```
## [1] 186757267233190
```

```
mean((fitted.values(ols.model) - train2$Second.Bid)^2) # OLS error, train
```

```
## [1] NaN
```

```
mean((ols.predict-test2$Second.Bid)^2) # OLS error, test
```

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

The model error is actually a bit higher. But consider the covariate survival process' implications for interpretability. Regress using the full data set.

```
full2 = rbind(train2,test2)

  x = train2[,-5]
  x = data.matrix(x)

full.model = glmnet(data.matrix(full2[,-5]), full2[, 5], alpha=0.5, lambda=cv$lambda.min)
full.predict = predict(full.model, type="coefficients", s = cv$lambda.min)
full.predict
```

```
## 45 x 1 sparse Matrix of class "dgCMatrix"
##                                                          1
## (Intercept)                                -179084210.8147489
## Engr.Est                                          0.8709706
## Loc                                          543456.2676754
## LD                                            93832.8115201
## Typeology                                   -695862.4772223
## 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                     112080.1143918
## Employment.in.other.services                              .
## Employment.in.production.industries                       .
## Employment.in.professional.services                       .
## Employment.in.real.estate                                 .
## Employment.in.retail                        -126435.6605323
## Employment.in.transport.services                          .
## Employment.in.wholesale                                   .
## Output.in.communications                                  .
## Output.in.construction                        -1808.1094795
## Output.in.financial.services                    214.3145145
## Output.in.government                           5900.1777600
## Output.in.retail                               -835.0423342
## Output.in.education.and.health                            .
## Output.in.financial.and.business.services                 .
## Output.in.other.services                        301.6995934
```

```
## Output.in.production.industries                    -662.5802556
## Output.in.professional.services                           .
## Output.in.real.estate                                     .
## Output.in.transport.services                              .
## Output.in.wholesale                                172.2911162
## 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                                          -61.2421838
## cci                                                       .
## Format2                                        -1842614.1586087
## quantile                                         437183.3131519
```

5. Discussion.

This suggests only targeted efforts can 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.

```
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)
```

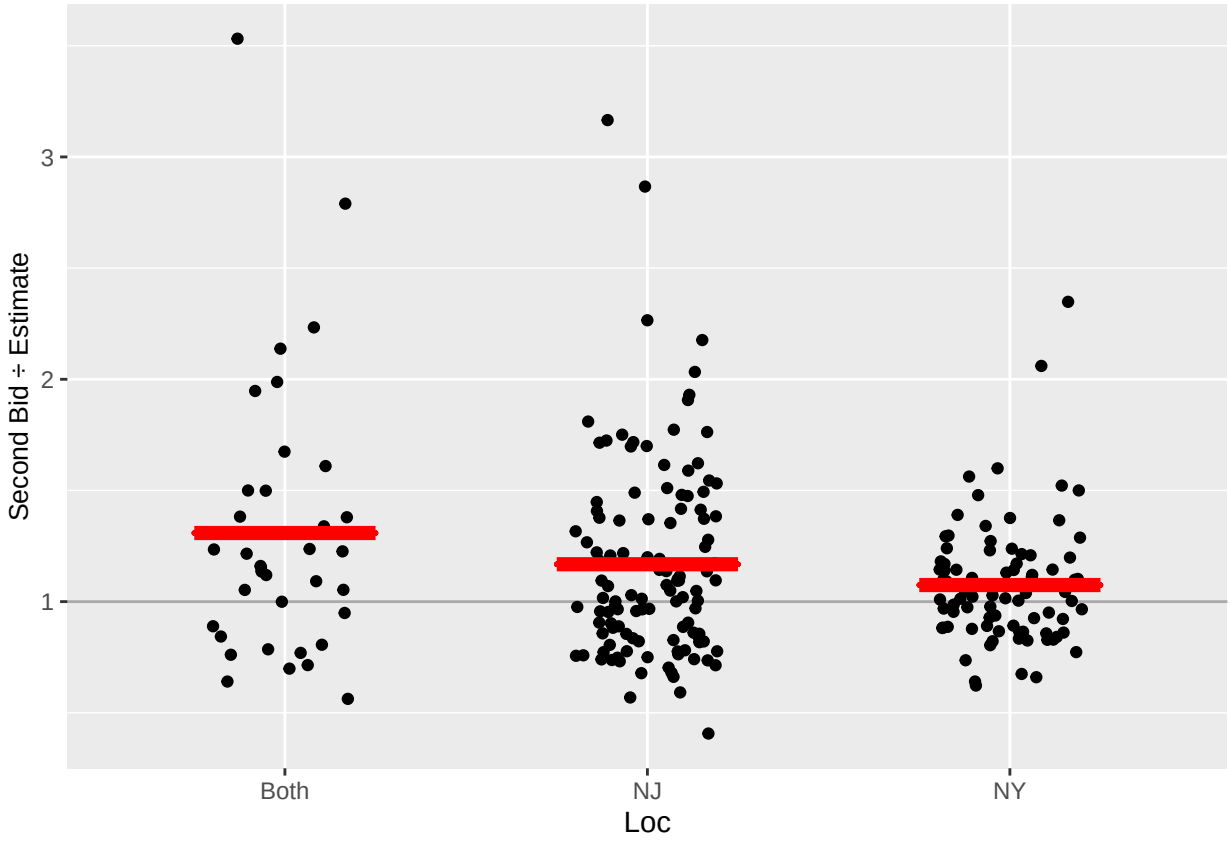
```
## bids$Engr.Est    ***
## bids$quantile_2
## bids$quantile_3
## bids$quantile_4
## bids$quantile_5
## bids$quantile_6
## bids$quantile_7
## bids$quantile_8
## bids$quantile_9  .
## bids$quantile_10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 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.

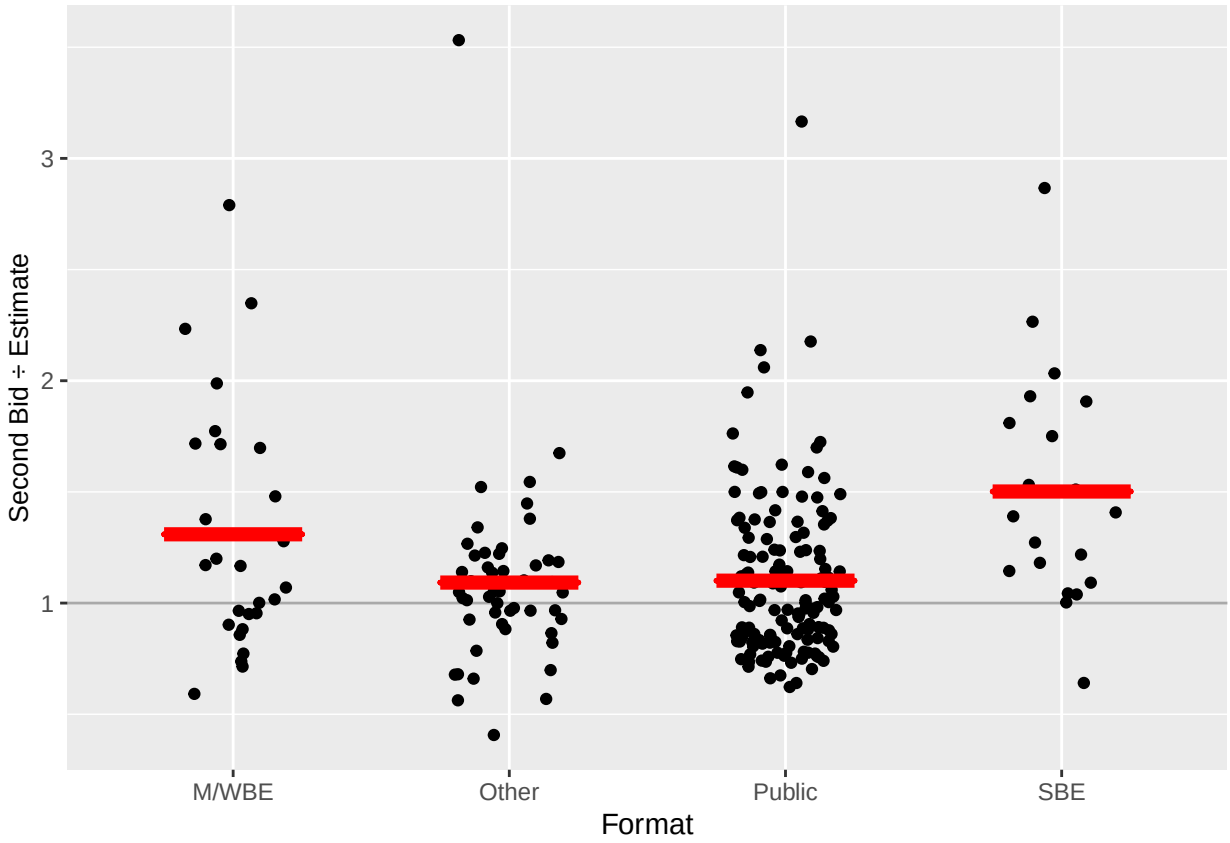
Location.




```
##
## 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
```

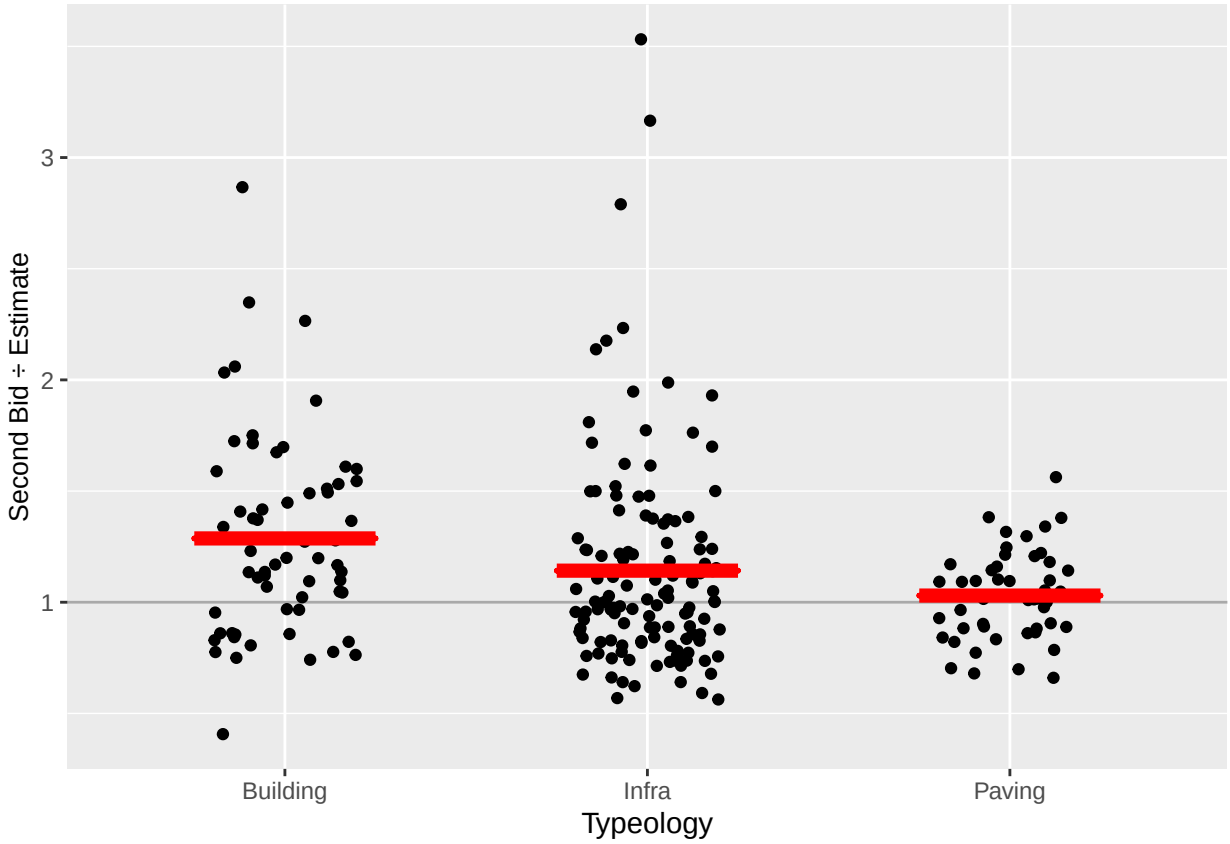
Bidding process (format).




```
##
## 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
```

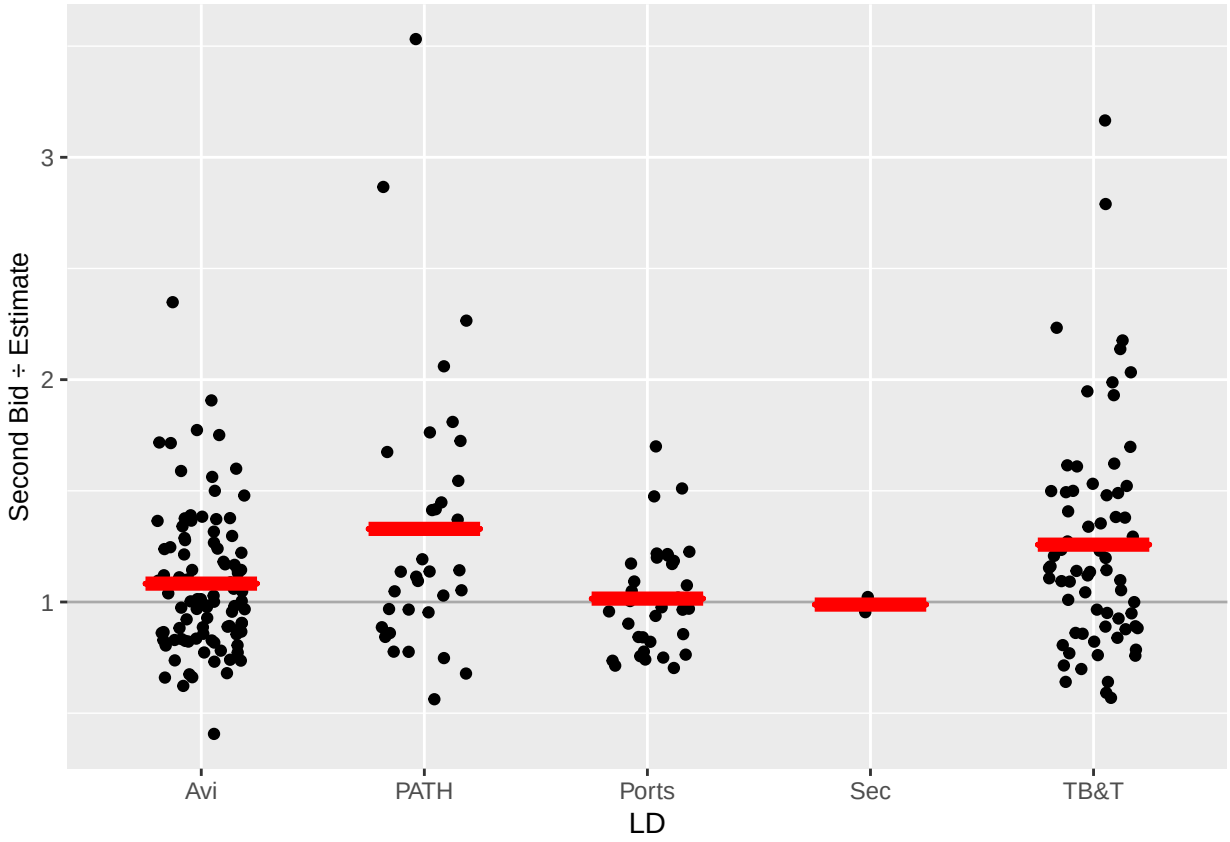
Project typeology.




```
##
## 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
```

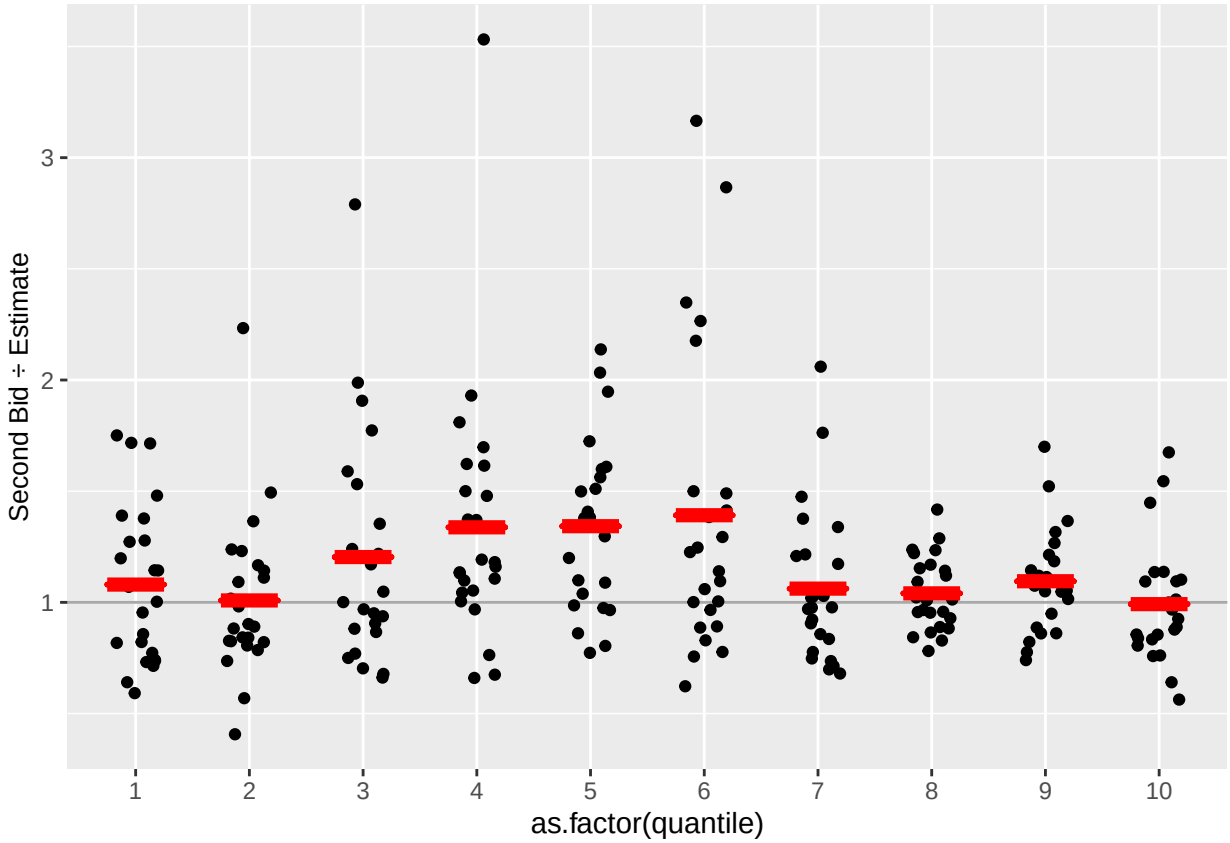
Department.




```
##
## 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 (dollars bid).




```
##
## 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
```