# Predicting costs :: Part 2

*Planning and Regional Development*

*1/21/2020*

## 1. Summary

Large projects - defined as the biggest 20 percent of 239 projects evaluated in Part 1 — and ones subjected to closed bidding processes explain a significant degree of the inaccuracy observed in agency cost estimation. Part 2 searches for predictive factors from within those two project subpopulations. It also considers the relationship between accuracy and the number of bidders, which was generally omitted from Part 1 as internal estimators do not know how many bidders will respond as they develop estimates.

The agency's internal cost estimates predict 95 percent of variation in cost, using the second-lowest bid[1] as a predicting target. On an absolute basis[2], however, the gap between internal estimate and second-lowest bid averages $2.7 million, or 18 percent of the average project size. Reducing this gap would provide for stronger confidence in long-range capital capacity estimates and could reduce the need for project-level change orders.

## 2. Focus on largest projects

A glance at the list of 17 projects above $60 million immediately identifies two projects where the agency and bidders (second-lowest bidders) significantly disagreed over expected costs:

---

[1] The agency's internal regulations require, in all but a handful of cases, the acceptance of the lowest bid. The Engineering Department views the second-lowest bid as a better predictive target.

[2] Mean absolute error, MAE.
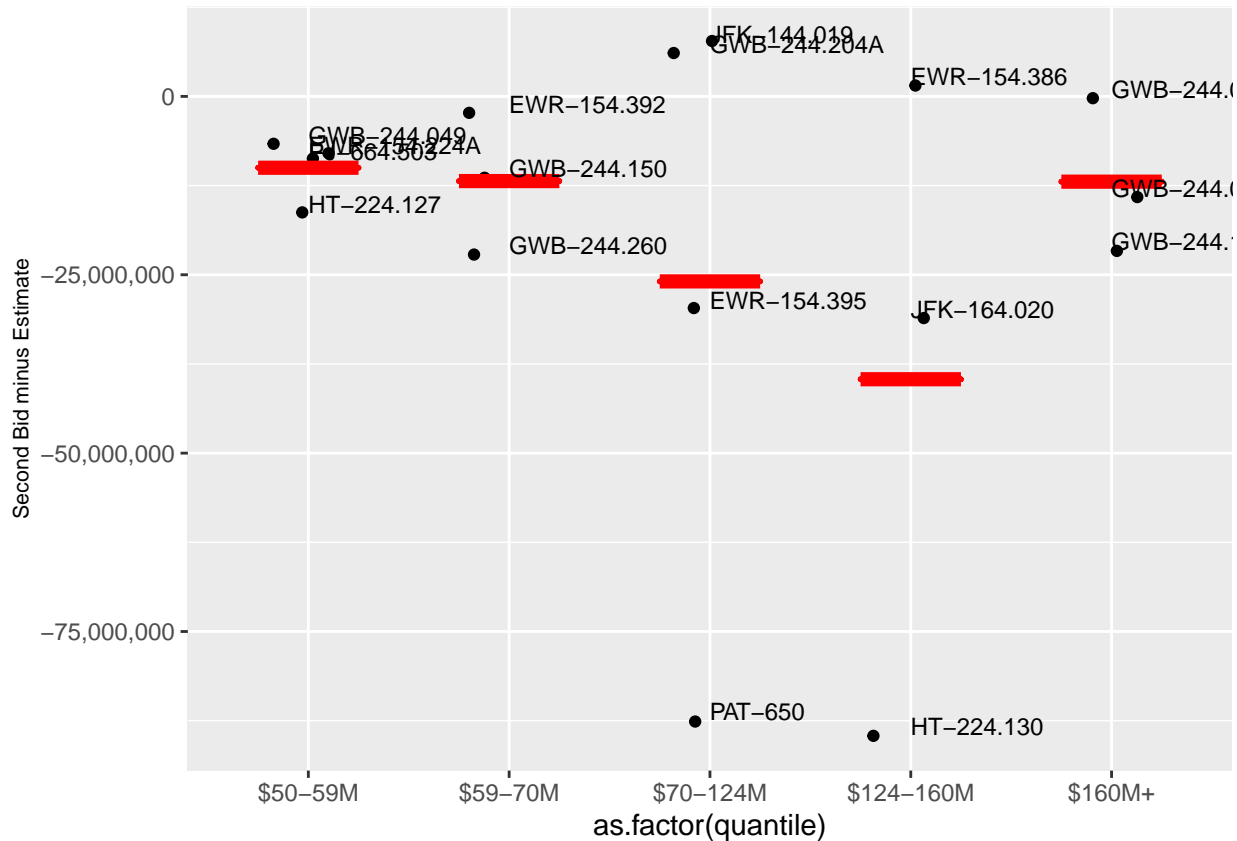
Consider how much of big projects' explanatory power may be due to the two large misses in the third and fourth quantiles above.

The Holland Tunnel (HT-224.130) project reflects major repairs to salt damage dated July 2019. It also collected five bids, of which the first ($135 million) and second ($159 million) were roughly two-fifths below the estimate ($249 million).

The PATH (PAT-650) project above represents repairs to PATH Tunnels E and F, dated January 2018, that was estimated at $200 million but attracted much lower bids — nearly half that figure for both the first ($108 million) and second ($112 million) bids.

Both are infrastructure projects.[3] While typeology did not emerge as one of the most significant predictors of estimation accuracy in Part 1, uncontrolled and non-penalized analysis had indicated infrastructure and building projects evaded accurate estimation more commonly than did paving projects; infrastructure projects also account for a larger share of the biggest work, making up one-half of the analyses' 239 projects but nearly all (14 of 17) of the largest projects:

```
##
## Building    Infra   Paving
## 25.52301 53.97490 20.50209

##
##  Building     Infra    Paving
##  5.882353 82.352941 11.764706
```

Bottom line: the two projects are large infrastructure projects, one of which is at a department (PATH) that has less typical projects (e.g. no relatively predictable paving work) and suffered significant damage from Hurricane Sandy.

```
##
```

---

[3]The data's "typeology" variable classifies projects as one of three categories: infrastructure, building, or paving.

```
##  Descriptive statistics by group
## group: Building
##    vars  n   mean        sd median trimmed     mad    min      max    range
## X1    1 61 8939624 14092759 2765000 5414215 2545847 415415 50950000 50534585
##    skew kurtosis      se
## X1 1.98     2.44 1804393
## --------------------------------------------------------------
## group: Infra
##    vars   n    mean         sd median trimmed     mad     min        max
## X1    1 129 18574212 52270017 3654000 6873718 3790341 142000 479645000
##        range skew kurtosis       se
## X1 479503000 6.06    45.93 4602118
## --------------------------------------------------------------
## group: Paving
##    vars  n   mean        sd median trimmed     mad    min       max      range
## X1    1 49 12771591 25376464 3769000 7614624 3543888 698371 155330252 154631881
##    skew kurtosis      se
## X1 4.04    18.48 3625209
```

Re-thinking the penalized algorithm after excluding the two major misses from the data set changes the regularized parameter set noticably. (See last page of output.) A number of variables that had vanished under penalization become significant, even when applying the same penalty. They're generally endogenous project characteristics (versus exogenous market characteristics): location, typology and an indicator for the very smallest projects, which may be difficult to estimate for their own reasons. But the original two variables of significance — bidding process and the indicator for the largest projects — remain significant, which suggests the results from Part 1's interation using all observations remains robust.

# 3. Number of bidders

All analyses has excluded the number of bidders. Including bidders would provide information not available to agency estimators prior to the bidding process and, as such, would confound results and detract from efforts to identify predictive metrics useful for operations. We anticipate, however, further interest in the relationship between bidders and accuracy and are prepared to explore the relationship further.

An uncontrolled look suggests the number of bids may be significantly related to project accuracy, prior to accounting for (holding constant) other variables when focusing on the second-lowest bids. (The more bids, the lower the second-lowest bid relative to agency estimate.) A glance at this relationship for projects above $60 million does not provide compelling evidence that a broad relationship between bid counts and accuracy would hold when trying to predict the largest projects:
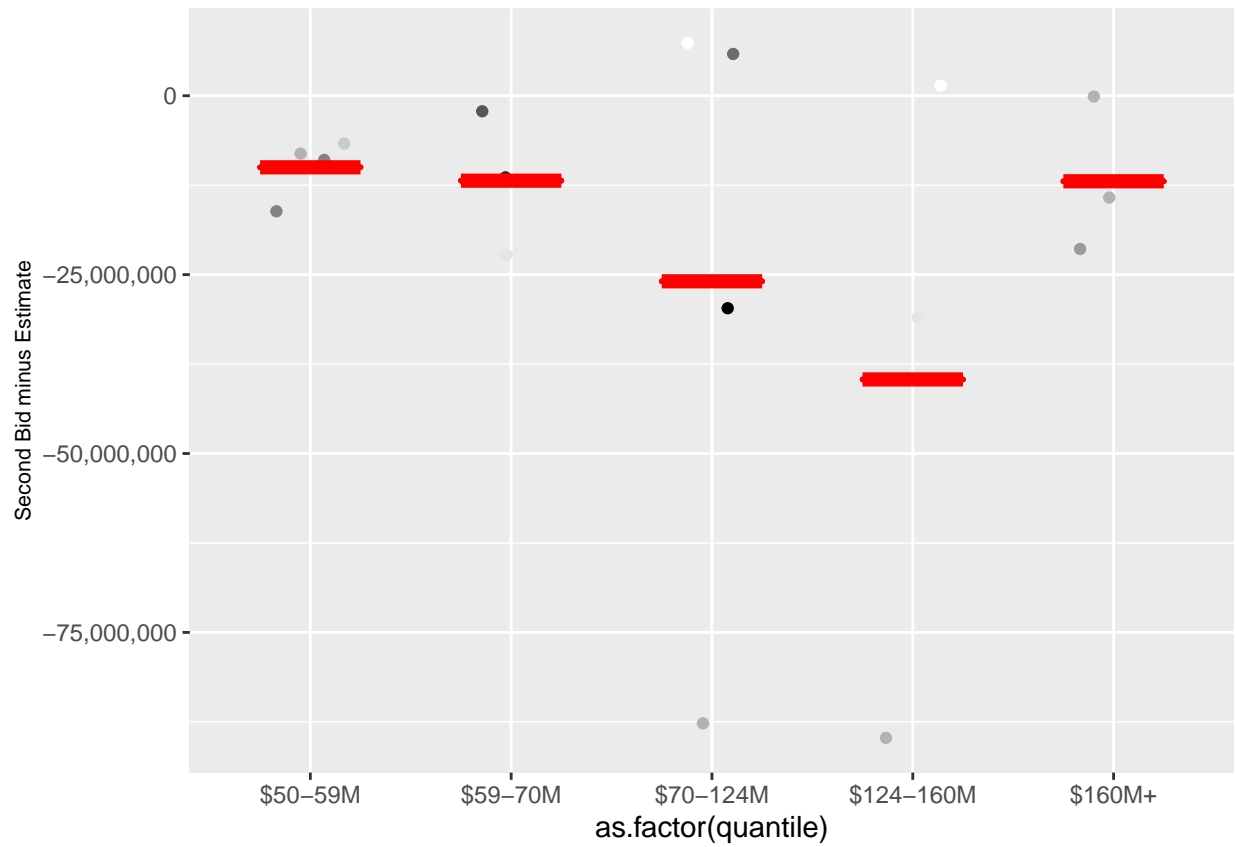
Future work can explicitly focus on bids' relationship to accuracy, with the working assumption that the presence of fewer bids may, after properly controlling for other factors, drive costs (bid size) higher. The potential relationship, however, is nuanced enough to defer more in-depth discussion for future work.

## Output: Variable Selection excluding PAT-650 and HT-224.130

```
## 48 x 1 sparse Matrix of class "dgCMatrix"
##                                                    s0
## (Intercept)                              3.556080e+06
## Engr.Est                                 9.004881e-01
## Loc                                     -2.678175e+05
## LD                                                  .
## Typeology                               -2.932840e+05
## 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             9.434122e+00
## 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                                 -1.370281e+06
## quantile_1                              -4.536997e+05
## quantile_2                                         .
## quantile_3                                         .
## quantile_4                                         .
## quantile_5                               3.476166e+06
```