Predicting costs

Chris Eshleman

10/20/2019

The agency estimates project costs internally. It may be able to use statistics to improve those predictions.

Model variation in bids, withholding the two most recent quarters of observations. Then test those models on the withheld observations. How well do they help with predictions?

Methods and Data.

1. Basic (manually constructed linear model),
2. Lasso (penalized regression - machine learning).

The project-level data comes from internal agency cost estimation.
The economic data is just quarterly stuff from the usual suspects and is specific to national and regional economic and labor market conditions.

Load some programming tools that are commonly used for this analysis. Not all of these packages will be used, and at some point it'll be worth backing up and cleaning the list.

Overview.

This effort bridges exploration of the agency's data with potential predictors from exogenous sources. The key idea is that if economic indicators (numbers) can add measurable value to the agency's cost estimation methods, it can help set the agency up for better informed next steps. Those next steps are yet to be defined.

```
#![Here's what we have in mind.](Estimator_Data.png)
```

The measures of "accuracy" are, for now, bivariate correlation. We discuss below why it's not yet time for fancier metrics, but the process above represents a first cut at trying to add a little more statistical value to the process.

Data comes from a few sources and needs to be merged. Two usual suspects are the agency's internal project information and a set of economic indicators specific to Greater New York (18 counties on both sides of the Hudson River).

Add economic data.

Add permits and steel prices.

The City of New York's database on permitting covers comercial and residential activity.

I'd like to ask have robust data on the rest of the region, including (and namely) Jersey City, but it's weaker than the City's, which by itself is a decent barometer of construction activity in greater New York.

Prices of construction materials and labor also figure into the agency's internal cost estimation, and I'll use steel prices for now. Future models might try and rope in other pricing data points, but the estimators are generally already taking prices into account when setting their numbers so this is likely of second- and third-order importance but the modeling selection algorithms may suggest using them.

The Engineering News-Record tracks and aggregates construction cost data through an index. Use that to cover prices.

```r
cci = read.csv("./cci.csv")
names(cci) = tolower(names(cci))
cci$avg. = NULL
cci = gather(cci, month, cci, jan:dec, factor_key=TRUE)
cci$month = gsub("(^[[:alpha:]])", "\\U\\1", cci$month, perl=TRUE)
cci$month = as.Date(paste(cci$month,"01",cci$year, sep="-"), format="%b-%d-%Y")
cci = cci[month(cci$month) == 2 | month(cci$month) == 5 | month(cci$month) == 8 | month(cci$month) == 1

cci$Quarter = paste("Q",quarter(cci$month), sep="")
cci$Q = paste(year(cci$month),cci$Quarter,sep="-")
cci$Quarter = NULL
cci$month = NULL
cci$year=NULL

bids = merge(bids, cci, by = "Q", all.x=TRUE)

rm(cci)
```

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.

There's room to also eventually include the names (anonymized is fine) of each project estimator to help modeling. Past work has suggested there isn't major causal variation between estimators — they generally do a pretty equivalent job in estimating bids. But having their names included nonetheless may prove to offer some control value. We can leave that to future modeling.

Restating objective.

We want to understand whether and how we might help the agency estimate the actual cost of a project. That's invariably going to be represented by the low qualifying bid, and our starting point is the estimate coming from the Engineering Department. From here on we'll define "accuracy" as the ratio of dollars estimated over dollars bid. So a "1" would mean the engineering team nailed it, a "0.94" would mean they estimated 94 cents for every 1 dollar in the low bid, et cetera.

If there are outliers in there - projects that, for an unexplainable reason, was way off, consider removing them.

We may want to think of project size categorically.

```r
bids = bids %>%
  mutate(quantile = ntile(as.numeric(Second.Bid), 10))
```

```
## Warning: `as_dictionary()` is soft-deprecated as of rlang 0.3.0.
## Please use `as_data_pronoun()` instead
## This warning is displayed once per session.

## Warning: `new_overscope()` is soft-deprecated as of rlang 0.2.0.
## Please use `new_data_mask()` instead
## This warning is displayed once per session.

## Warning: The `parent` argument of `new_data_mask()` is deprecated.
## The parent of the data mask is determined from either:
```

```
##
##   * The `env` argument of `eval_tidy()`
##   * Quosure environments when applicable
## This warning is displayed once per session.
```

```
## Warning: `overscope_clean()` is soft-deprecated as of rlang 0.2.0.
## This warning is displayed once per session.
```

```r
bids$quantile = as.factor(bids$quantile)
bids$quantile = ordered(bids$quantile, levels = 1:10)
```

Munge and back up data

Stats software treats different variables in different ways depending on individual formatting. It's worth taking a look at the data structure.

I'll need to tell the software to reformat some of the variables, namely the economic indicators.

Back up data.

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

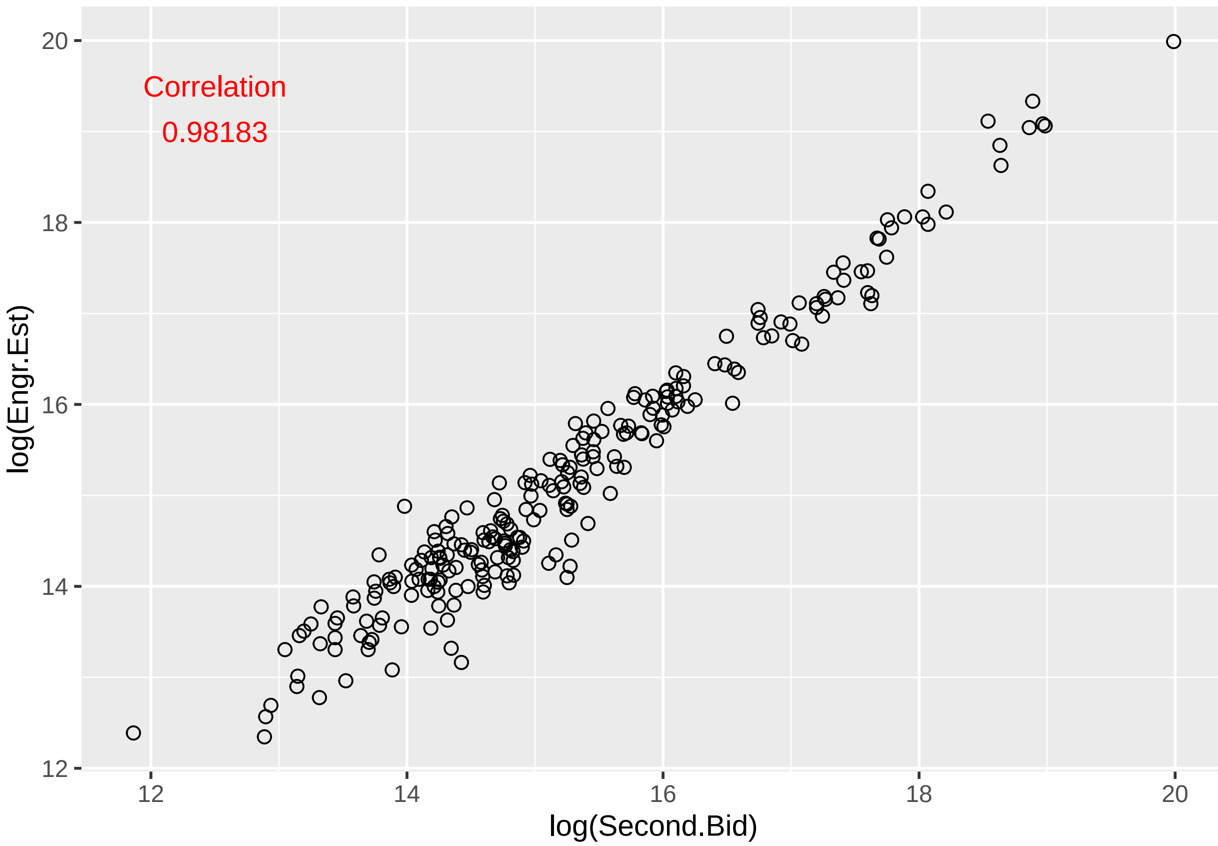
Analysis

Now the data is prepped. So split it into the training / test sets we talked about at the start. The bids start in 2015.

Motivation.

Why do we think developing a conrolled multivariate (complicated) model will be worth it? Well, the average gap is $2.5 million, or 20%, off of our estimates. What's the raw (uncontrolled) bivariate relationship between engineering estimates and low bids?

```
##
## Call:
## lm(formula = bids$Second.Bid ~ bids$Engr.Est)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -65502890   -996757   -514372    419628  53757694
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.428e+05  5.319e+05   1.772   0.0776 .
## bids$Engr.Est 8.855e-01  1.112e-02  79.663   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7762000 on 237 degrees of freedom
## Multiple R-squared:  0.964,  Adjusted R-squared:  0.9638
## F-statistic:  6346 on 1 and 237 DF,  p-value: < 2.2e-16
```



(The logarithm treatment is just to distribute it across the plot (one of the observations is an outlier).)

The in-house engineers' guesses predict more than 98% of the variation in low bids.

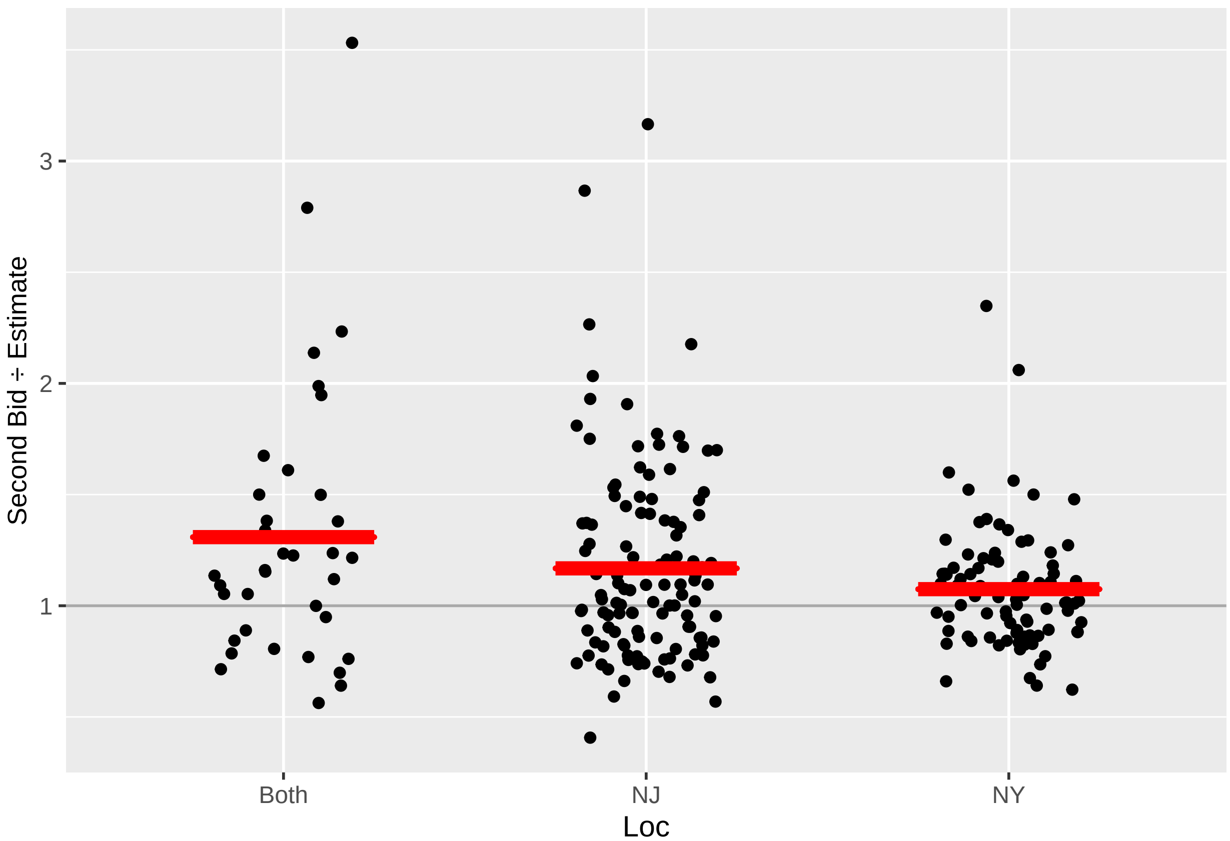
Some of the remaining variation can be explained with some guesswork. Calculate summary stats for key variables - this will help later:

```
loc.summary = aggregate(accuracy ~ Loc, mean, data=na.omit(bids))
type.summary = aggregate(accuracy ~ Typeology, mean, data=na.omit(bids))
format.summary = aggregate(accuracy ~ Format, mean, data=na.omit(bids))
decile.summary = aggregate(accuracy ~ quantile, mean, data=na.omit(bids))
ld.summary = aggregate(accuracy ~ LD, mean, data=na.omit(bids))
```

Location likely has some predictive power that engineers may not be able to capture or fully predict. Basically, projects that span the Hudson River wind up costing more, on average, than ones plunked squarely in either New York or New Jersey. What's the raw (uncontrolled) relationship between bid accuracy and 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  < 2e-16 ***
## 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
```



Projects that span the Hudson River seem to come in, on average, higher than expected.

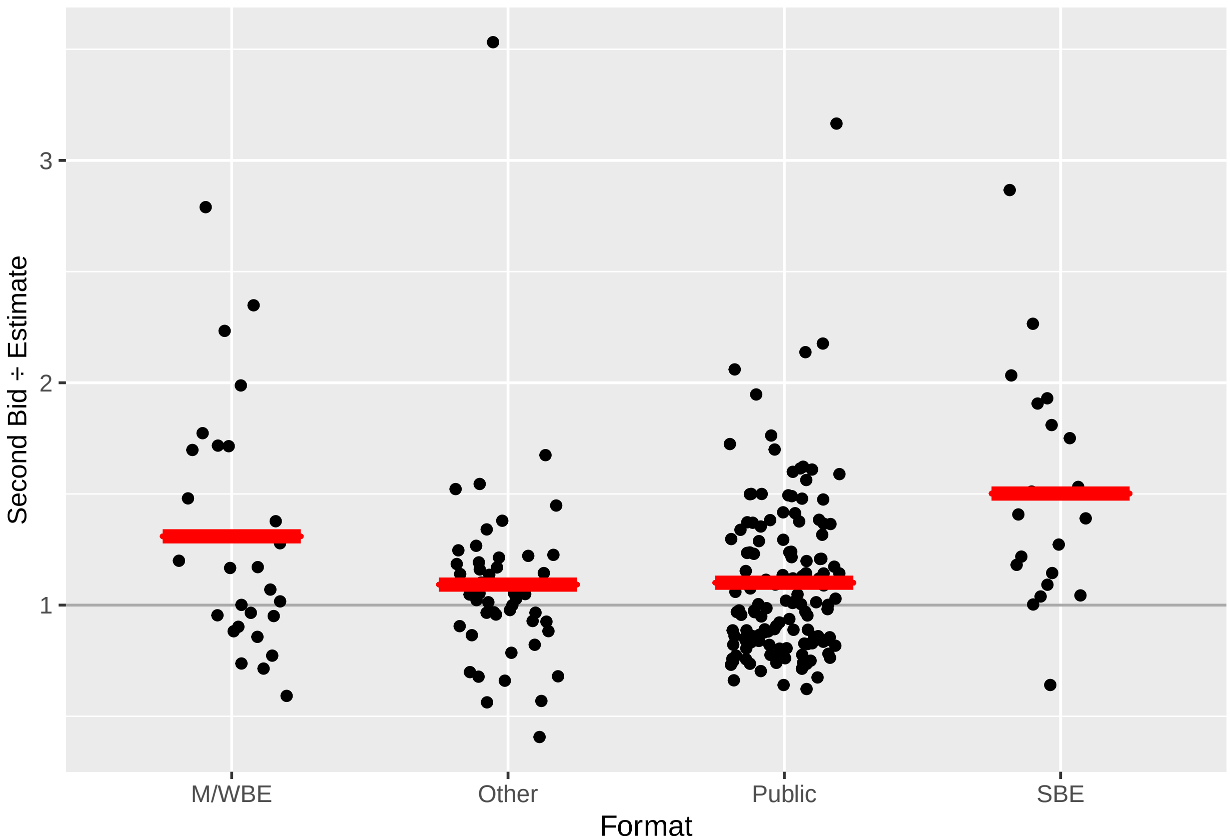
A project's bidding process can be constrained or open, with potential ramifications on the ability to estimate costs:

```
summary(lm(bids$accuracy ~ as.factor(bids$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 Pr(>|t|)
## (Intercept)                   1.30933    0.08035  16.295   <2e-16 ***
## as.factor(bids$Format)Other  -0.21709    0.09937  -2.185   0.0299 *
## as.factor(bids$Format)Public -0.20862    0.08771  -2.379   0.0182 *
## as.factor(bids$Format)SBE     0.19246    0.12318   1.563   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
```

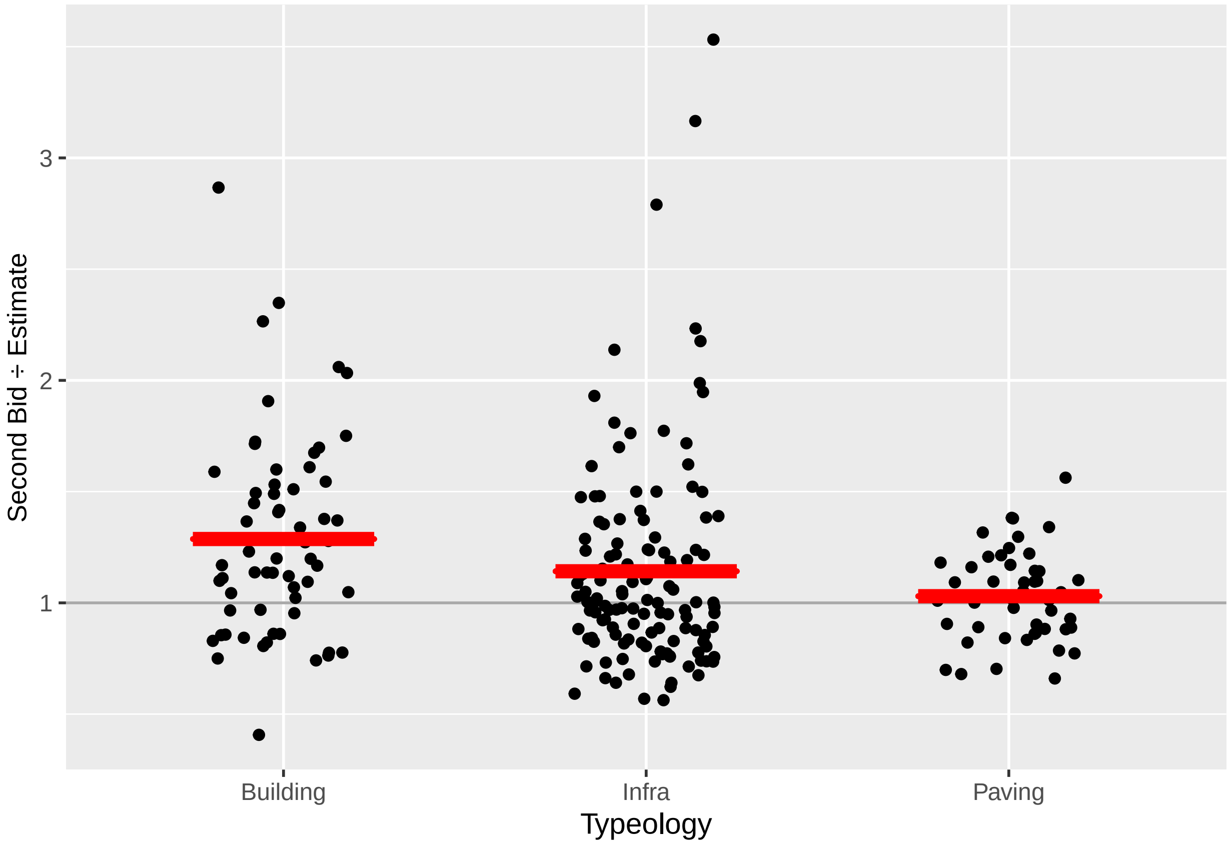
```
ggplot(bids, aes(x = Format, y = accuracy)) + #, shape = Loc)) +
    geom_hline(colour="dark gray", yintercept=1) +
    geom_jitter(width=0.2) +
    geom_crossbar(data=format.summary, aes(ymin = accuracy, ymax = accuracy),
                  size=1,col="red", width = .5) +
      ylab("Second Bid ÷ Estimate") +
    theme(axis.title.y=element_text(size=10))
```



The signal is stronger regarding the type of project, which has an identifiable (if yet uncontrolled) relationship with estimating accuracy. What's the raw (uncontrolled) relationship between estimation accuracy and the type of project?

```
##
## 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 Pr(>|t|)
## (Intercept)                        1.28697    0.05451  23.609  < 2e-16 ***
```

```
## as.factor(bids$Typeology)Infra  -0.14499    0.06616  -2.192  0.02938 *
## as.factor(bids$Typeology)Paving -0.25693    0.08168  -3.146  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
```

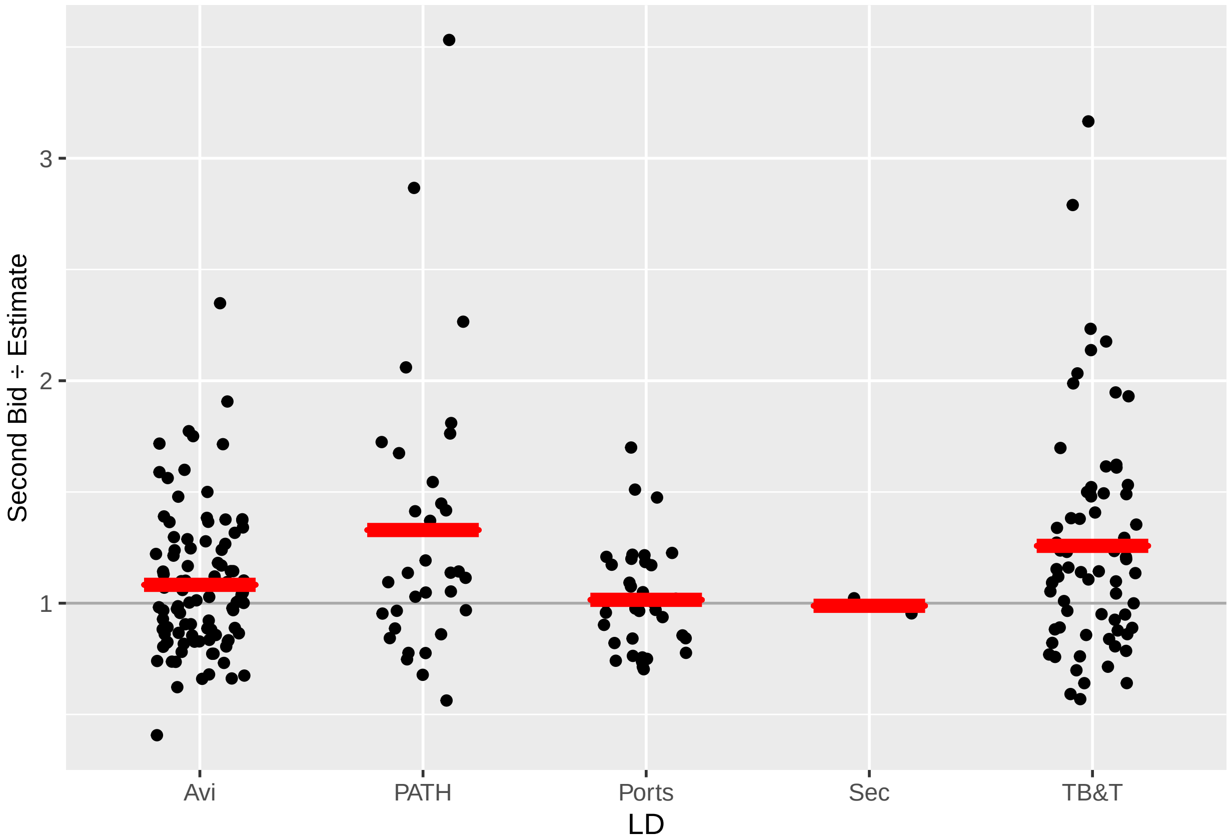


The agency can predict paving projects fairly well.

By line 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  < 2e-16 ***
## 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 **
## ---
```

```
## 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
```

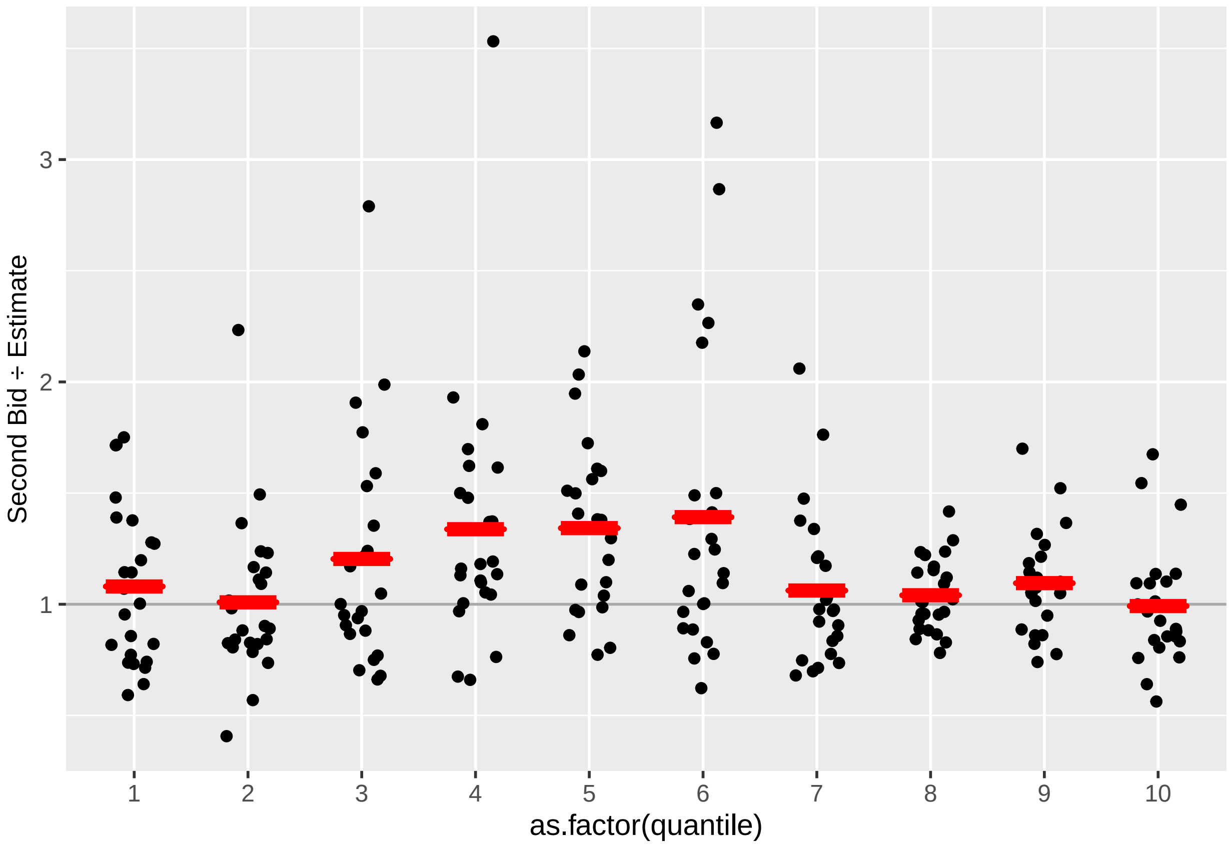


Four obvious outliers, two apiece for PATH and TB&T, push the average accuracy metric up for each, otherwise there's little apparent difference across departments.

We see signals that projects of medium size (in dollar terms) may be less evasive than much larger or smaller projects. What's the relationship between low bids and accuracy, when we start considering the size of low bids?

```
##
## 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 Pr(>|t|)
## (Intercept)                 1.15535    0.02694  42.888  < 2e-16 ***
## as.factor(bids$quantile).L -0.09813    0.08546  -1.148  0.25203
## as.factor(bids$quantile).Q -0.32559    0.08551  -3.808  0.00018 ***
## as.factor(bids$quantile).C  0.09462    0.08538   1.108  0.26892
## as.factor(bids$quantile)^4  0.17352    0.08521   2.036  0.04287 *
## as.factor(bids$quantile)^5 -0.16602    0.08509  -1.951  0.05227 .
```

```
## as.factor(bids$quantile)^6 -0.07635    0.08503  -0.898  0.37018
## as.factor(bids$quantile)^7 -0.05900    0.08501  -0.694  0.48834
## as.factor(bids$quantile)^8  0.06234    0.08500   0.733  0.46411
## as.factor(bids$quantile)^9  0.10264    0.08500   1.208  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
```



These are the most surprising results. I would 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.

Visualizing and testing the data iteratively like this has offered some initial insight into what might be accounting for variation in accuracy.

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 and 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.

A. Base model (manual selection)

Interpretation: specification was manual and intuitive.

Note: ensure the "accuracy" variable we 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.

First remove accuracy.

```
train$accuracy=NULL
test$accuracy=NULL
```

Run model.

```
##
## 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
```

The equation above throws extra information at the engineering estimate and tries to predict the low qualifying bid. If the result is noticably closer to the low qualifying bid than the original estiamte, you can use the delta as a post-estimation fudge factor to adjust your 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 low bids?

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##  -1569853   1720892   3478279  14925516   9814439 425157238
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    142000   1556557   3185420  14925516   9143818 479645000
```

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. 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.

B. Machine learning

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

The output below represents the algorithm's effort to look for variatbles that might be dependable in adding predictive power to the original engineering estimate.

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

```
##  [1] "Q"
##  [2] "Proj"
##  [3] "Proj.Name"
##  [4] "Date"
##  [5] "Format"
##  [6] "Engr.Est"
##  [7] "Low.Bid"
##  [8] "Var"
##  [9] "Result"
## [10] "Bids"
## [11] "Loc"
## [12] "Qtr"
## [13] "LD"
## [14] "Lead"
## [15] "Typeology"
## [16] "Second.Bid"
## [17] "Consumer.price.index"
## [18] "Employment.in.communications"
## [19] "Employment.in.construction"
## [20] "Employment.in.education.and.health"
## [21] "Employment.in.financial.and.business.services"
## [22] "Employment.in.financial.services"
## [23] "Employment.in.government"
## [24] "Employment.in.other.services"
## [25] "Employment.in.production.industries"
## [26] "Employment.in.professional.services"
## [27] "Employment.in.real.estate"
## [28] "Employment.in.retail"
## [29] "Employment.in.transport.services"
## [30] "Employment.in.wholesale"
## [31] "Output.in.communications"
## [32] "Output.in.construction"
## [33] "Output.in.financial.services"
## [34] "Output.in.government"
## [35] "Output.in.retail"
## [36] "Output.in.education.and.health"
## [37] "Output.in.financial.and.business.services"
## [38] "Output.in.other.services"
## [39] "Output.in.production.industries"
## [40] "Output.in.professional.services"
## [41] "Output.in.real.estate"
## [42] "Output.in.transport.services"
## [43] "Output.in.wholesale"
## [44] "Personal.disposable.income..nominal"
## [45] "Personal.disposable.income..real"
## [46] "Personal.income..nominal"
## [47] "Retail.sales..nominal"
## [48] "Retail.sales..real"
## [49] "Total.employment"
```

```
## [50] "Total.office.based.employment"
## [51] "Total.output"
## [52] "Total.population"
## [53] "permits"
## [54] "permits_1"
## [55] "cci"
## [56] "Format2"
## [57] "quantile"
```

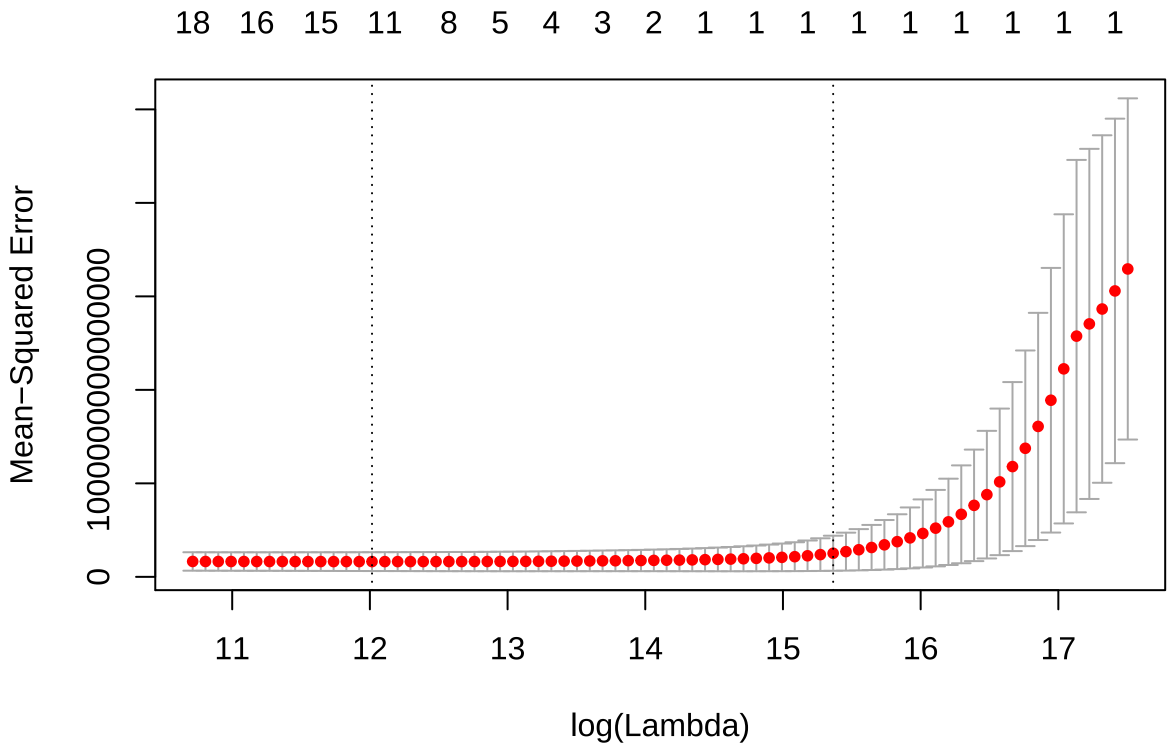
```
##         Q        Proj
## 1 2015-Q1 LGA-124.166
## 2 2015-Q1 LGA-124.231
## 3 2015-Q1 JFK-134.025
## 4 2015-Q1  PN-654.004
## 5 2015-Q1 PAT-084.057
## 6 2015-Q2  PJ-924.624
##                                                                                  Proj.Name
## 1                 Laguardia Airport-Rehabilitation of Runway 13-31 and Associated Taxiways
## 2                       Laguardia Airport-Rehabilitation of Taxiways West of Runway 4-22
## 3 John F. Kennedy International Airport-Unmanned AOA Gates and Perimeter Fence Enhancement-Phase II
## 4                        Port Newark-Berths 30, 32, and 34 Fender Systems Reconstruction
## 5                     PATH-Access Control and CCTV at Substation and Communications Rooms
## 6              Port Jersey Marine Terminal-Paving and Utility Rehabilitation via Work Order
##         Date Format Engr.Est  Low.Bid         Var Result Bids Loc Qtr
## 1 0015-02-03  Other 25770000 28747550  0.11554327   FAIL    3  NY   1
## 2 0015-02-22  Other  9700000  8742268 -0.09873526   GOOD    3  NY   1
## 3 0015-02-25  M/WBE   640000   574000 -0.10312500   GOOD    4  NY   1
## 4 0015-03-09 Public  9600000  6748000 -0.29708333   GOOD   12  NJ   1
## 5 0015-03-17  Other  4450000  3620000 -0.18651685   GOOD    3  NJ   1
## 6 0015-04-08  M/WBE  1300000   878175 -0.32448077   GOOD    7  NJ   2
##      LD Lead Typeology Second.Bid Consumer.price.index
## 1   Avi    Z    Paving 29484532.0              259.312
## 2   Avi    Z    Paving  9825550.0              259.312
## 3   Sec    Z     Infra   610750.9              259.312
## 4 Ports    Z     Infra  7064400.0              259.312
## 5  PATH    Z  Building  4298317.0              259.312
## 6 Ports    Z    Paving  1321680.0              260.360
##   Employment.in.communications Employment.in.construction
## 1                     276.0735                   331.6689
## 2                     276.0735                   331.6689
## 3                     276.0735                   331.6689
## 4                     276.0735                   331.6689
## 5                     276.0735                   331.6689
## 6                     274.6349                   340.2549
##   Employment.in.education.and.health
## 1                           1666.401
## 2                           1666.401
## 3                           1666.401
## 4                           1666.401
## 5                           1666.401
## 6                           1677.252
##   Employment.in.financial.and.business.services
## 1                                      2149.166
## 2                                      2149.166
```

```
## 3                                   2149.166
## 4                                   2149.166
## 5                                   2149.166
## 6                                   2161.926
##   Employment.in.financial.services Employment.in.government
## 1                         549.0220                 1191.476
## 2                         549.0220                 1191.476
## 3                         549.0220                 1191.476
## 4                         549.0220                 1191.476
## 5                         549.0220                 1191.476
## 6                         549.8766                 1194.775
##   Employment.in.other.services Employment.in.production.industries
## 1                     1190.153                            339.2611
## 2                     1190.153                            339.2611
## 3                     1190.153                            339.2611
## 4                     1190.153                            339.2611
## 5                     1190.153                            339.2611
## 6                     1195.202                            339.4585
##   Employment.in.professional.services Employment.in.real.estate
## 1                            1410.922                  189.2218
## 2                            1410.922                  189.2218
## 3                            1410.922                  189.2218
## 4                            1410.922                  189.2218
## 5                            1410.922                  189.2218
## 6                            1422.188                  189.8612
##   Employment.in.retail Employment.in.transport.services
## 1             872.4502                         321.4322
## 2             872.4502                         321.4322
## 3             872.4502                         321.4322
## 4             872.4502                         321.4322
## 5             872.4502                         321.4322
## 6             870.8953                         324.0663
##   Employment.in.wholesale Output.in.communications Output.in.construction
## 1                400.9382                 31940.55               9182.419
## 2                400.9382                 31940.55               9182.419
## 3                400.9382                 31940.55               9182.419
## 4                400.9382                 31940.55               9182.419
## 5                400.9382                 31940.55               9182.419
## 6                401.1779                 32623.61               9690.083
##   Output.in.financial.services Output.in.government Output.in.retail
## 1                     59006.30             30193.99         15904.98
## 2                     59006.30             30193.99         15904.98
## 3                     59006.30             30193.99         15904.98
## 4                     59006.30             30193.99         15904.98
## 5                     59006.30             30193.99         15904.98
## 6                     60937.42             30189.85         16158.21
##   Output.in.education.and.health Output.in.financial.and.business.services
## 1                       30386.12                                  172889.8
## 2                       30386.12                                  172889.8
## 3                       30386.12                                  172889.8
## 4                       30386.12                                  172889.8
## 5                       30386.12                                  172889.8
## 6                       30592.22                                  174773.0
##   Output.in.other.services Output.in.production.industries
```

```
## 1                19872.47                        15694.17
## 2                19872.47                        15694.17
## 3                19872.47                        15694.17
## 4                19872.47                        15694.17
## 5                19872.47                        15694.17
## 6                19995.72                        15495.74
##   Output.in.professional.services Output.in.real.estate
## 1                        54670.04              59213.50
## 2                        54670.04              59213.50
## 3                        54670.04              59213.50
## 4                        54670.04              59213.50
## 5                        54670.04              59213.50
## 6                        54499.78              59335.79
##   Output.in.transport.services Output.in.wholesale
## 1                     13251.22            22804.42
## 2                     13251.22            22804.42
## 3                     13251.22            22804.42
## 4                     13251.22            22804.42
## 5                     13251.22            22804.42
## 6                     13618.77            23273.89
##   Personal.disposable.income..nominal Personal.disposable.income..real
## 1                            251601.9                         245211.6
## 2                            251601.9                         245211.6
## 3                            251601.9                         245211.6
## 4                            251601.9                         245211.6
## 5                            251601.9                         245211.6
## 6                            256008.6                         248296.5
##   Personal.income..nominal Retail.sales..nominal Retail.sales..real
## 1                 302741.9              69862.11           73297.36
## 2                 302741.9              69862.11           73297.36
## 3                 302741.9              69862.11           73297.36
## 4                 302741.9              69862.11           73297.36
## 5                 302741.9              69862.11           73297.36
## 6                 308350.7              71230.83           74279.39
##   Total.employment Total.office.based.employment Total.output
## 1         8739.020                      2149.166     361598.3
## 2         8739.020                      2149.166     361598.3
## 3         8739.020                      2149.166     361598.3
## 4         8739.020                      2149.166     361598.3
## 5         8739.020                      2149.166     361598.3
## 6         8779.643                      2161.926     365783.7
##   Total.population permits permits_1  cci Format2 quantile
## 1         18358.81   38139     43358 9962   Other        9
## 2         18358.81   38139     43358 9962   Other        8
## 3         18358.81   38139     43358 9962   Other        1
## 4         18358.81   38139     43358 9962  Public        7
## 5         18358.81   38139     43358 9962   Other        6
## 6         18371.72   54098     38139 9975   Other        2
```

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

```r
plot(lasso)
```



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

```
coef(lasso, s=lasso$lambda.1se)
```

```
## 45 x 1 sparse Matrix of class "dgCMatrix"
##                                                               1
## (Intercept)                                     2588278.5414532
## Engr.Est                                              0.7812806
## 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                                                .
```

A little fine-tuning identifies the two parameters that survive as the model gets further from ordinary:

```
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
```

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.

```
bids = fastDummies::dummy_cols(bids, select_columns = "quantile")
bids$quantile = as.factor(bids$quantile)
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
```