Estimating Non-Linear Models for Cost Prediction

11/24/24

Overall Objectives

The overall objective was to evaluate four non-linear models to accurately predict cost based on trade size and volatility. To achieve this, I estimated optimal parameters using a cross-validation procedure with an MAE metric, as further described below. Although I chose model ”c” to compute the test estimates, the computed fit metrics were broadly similar, despite significantly different cost estimate profiles. I conclude that much more work is necessary before any model could be used with confidence. Additional data to reduce the apparent noise in realized costs would be helpful.

$$(a)\ Cost_i = \beta_0 + \beta_1 TradeSize_i^{0.5} + \epsilon_i$$

$$(b)\ Cost_i = \beta_0 + \beta_1 TradeSize_i + \epsilon_i$$

$$(c)\ Cost_i = \beta_0 + \beta_1 TradeSize_i^{0.5} + \beta_2 Volatility_i + \beta_3 Volatility_i \times TradeSize_i^{0.5} + \epsilon_i$$

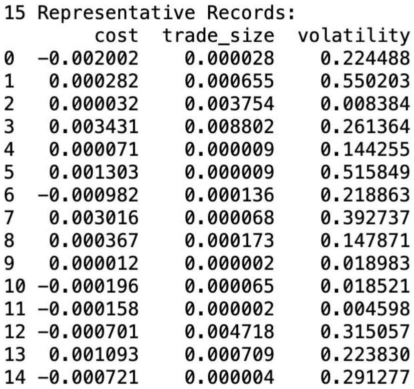
$$(d)\ Cost_i = \beta_0 + \beta_1 TradeSize_i + \beta_2 Volatility_i + \beta_3 Volatility_i \times TradeSize_i + \epsilon_i$$

Note: the prompt pdf posed four specific questions. Explicit answers are provided on pages 14 and 15 of this report.

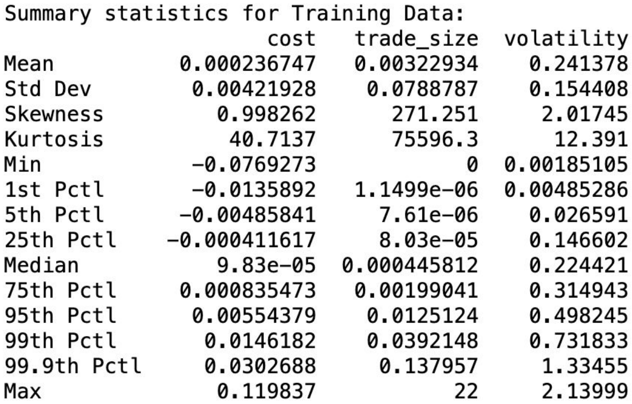
Data Overview: Representative records and Summary Statistics

Here are 15 representative records, as well as summary statistics for the training data.

- All of the variables have large outliers (discussed further in next slide)
- Many costs are counter-intuitively negative
- The data shows a wide range of volatility values; are these from different asset classes?



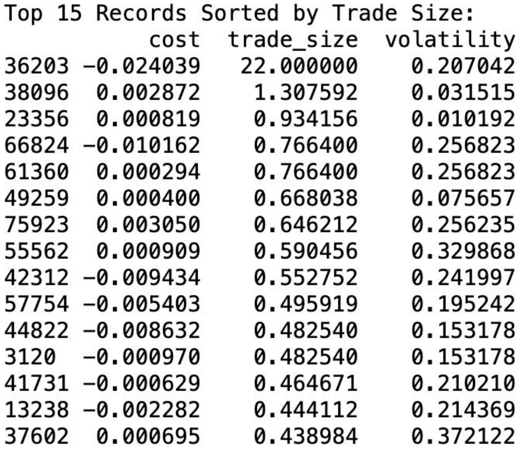
```
15 Representative Records:
        cost  trade_size  volatility
0  -0.002002    0.000028    0.224488
1   0.000282    0.000655    0.550203
2   0.000032    0.003754    0.008384
3   0.003431    0.008802    0.261364
4   0.000071    0.000009    0.144255
5   0.001303    0.000009    0.515849
6  -0.000982    0.000136    0.218863
7   0.003016    0.000068    0.392737
8   0.000367    0.000173    0.147871
9   0.000012    0.000002    0.018983
10 -0.000196    0.000065    0.018521
11 -0.000158    0.000002    0.004598
12 -0.000701    0.004718    0.315057
13  0.001093    0.000709    0.223830
14 -0.000721    0.000004    0.291277
```



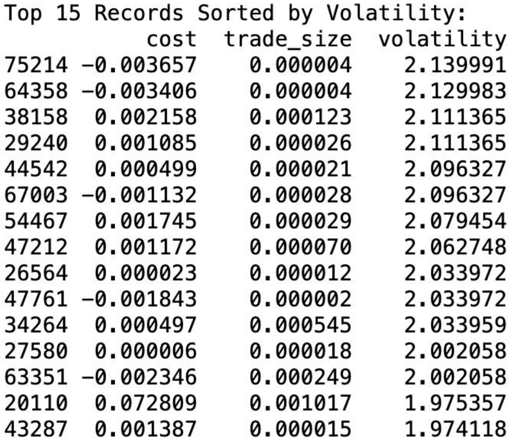
```
Summary statistics for Training Data:
                    cost   trade_size  volatility
Mean         0.000236747   0.00322934    0.241378
Std Dev       0.00421928    0.0788787    0.154408
Skewness        0.998262      271.251     2.01745
Kurtosis         40.7137      75596.3      12.391
Min           -0.0769273            0  0.00185105
1st Pctl      -0.0135892   1.1499e-06  0.00485286
5th Pctl     -0.00485841     7.61e-06    0.026591
25th Pctl   -0.000411617     8.03e-05    0.146602
Median          9.83e-05  0.000445812    0.224421
75th Pctl    0.000835473   0.00199041    0.314943
95th Pctl     0.00554379    0.0125124    0.498245
99th Pctl      0.0146182    0.0392148    0.731833
99.9th Pctl    0.0302688     0.137957     1.33455
Max             0.119837           22     2.13999
```

Data Overview: Outliers

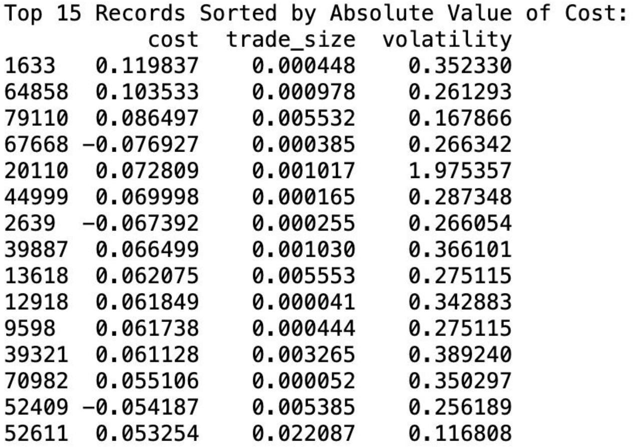
Here are sorts of the training data records by trade size, volatility, and the absolute value of cost. Outliers are extreme, especially for trade size and cost.



```
Top 15 Records Sorted by Trade Size:
           cost  trade_size  volatility
36203 -0.024039   22.000000    0.207042
38096  0.002872    1.307592    0.031515
23356  0.000819    0.934156    0.010192
66824 -0.010162    0.766400    0.256823
61360  0.000294    0.766400    0.256823
49259  0.000400    0.668038    0.075657
75923  0.003050    0.646212    0.256235
55562  0.000909    0.590456    0.329868
42312 -0.009434    0.552752    0.241997
57754 -0.005403    0.495919    0.195242
44822 -0.008632    0.482540    0.153178
3120  -0.000970    0.482540    0.153178
41731 -0.000629    0.464671    0.210210
13238 -0.002282    0.444112    0.214369
37602  0.000695    0.438984    0.372122
```



```
Top 15 Records Sorted by Volatility:
           cost  trade_size  volatility
75214 -0.003657    0.000004    2.139991
64358 -0.003406    0.000004    2.129983
38158  0.002158    0.000123    2.111365
29240  0.001085    0.000026    2.111365
44542  0.000499    0.000021    2.096327
67003 -0.001132    0.000028    2.096327
54467  0.001745    0.000029    2.079454
47212  0.001172    0.000070    2.062748
26564  0.000023    0.000012    2.033972
47761 -0.001843    0.000002    2.033972
34264  0.000497    0.000545    2.033959
27580  0.000006    0.000018    2.002058
63351 -0.002346    0.000249    2.002058
20110  0.072809    0.001017    1.975357
43287  0.001387    0.000015    1.974118
```



```
Top 15 Records Sorted by Absolute Value of Cost:
           cost  trade_size  volatility
1633   0.119837    0.000448    0.352330
64858  0.103533    0.000978    0.261293
79110  0.086497    0.005532    0.167866
67668 -0.076927    0.000385    0.266342
20110  0.072809    0.001017    1.975357
44999  0.069998    0.000165    0.287348
2639  -0.067392    0.000255    0.266054
39887  0.066499    0.001030    0.366101
13618  0.062075    0.005553    0.275115
12918  0.061849    0.000041    0.342883
9598   0.061738    0.000444    0.275115
39321  0.061128    0.003265    0.389240
70982  0.055106    0.000052    0.350297
52409 -0.054187    0.005385    0.256189
52611  0.053254    0.022087    0.116808
```

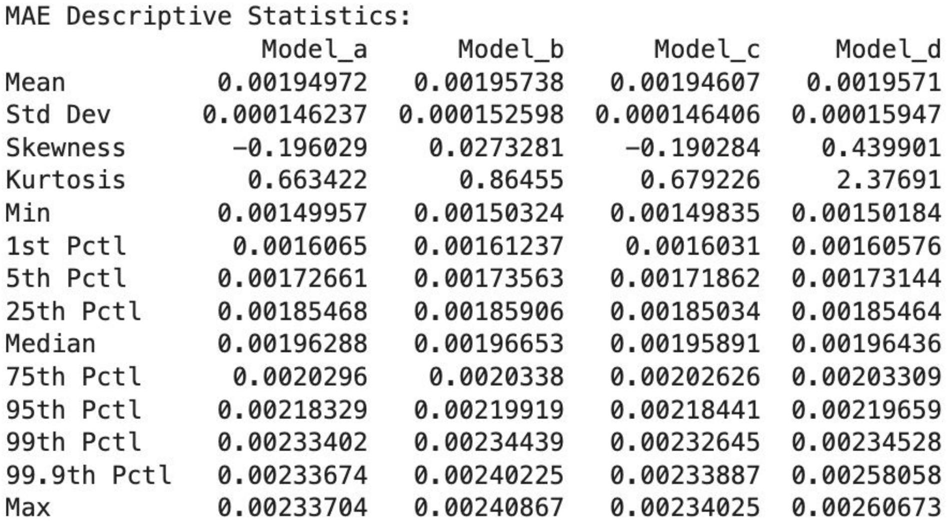
Methodology

K-Fold Cross Validation

- Outliers detected earlier may have been due to errors in the data, or to significant noise. Regardless, this argues for use of a method that reduces the influence of extreme observations. There are several options but I chose use of an MAE metric (as opposed to the perhaps more common MSE metric) in the estimation since this purportedly reduces the influence of outliers.
- To further reduce noise, I used K-fold cross validation, a common AI technique. I used 100 folds, which generated 100 estimates for both the parameters of each model and the associated MAE.
- Using these empirical distributions, I simply took the average parameter value across all folds as my final parameter estimate for each model, and similarly for the MAEs.

Estimation Results: MAE Summary by Model

To my eye, the mean MAE as well as the percentiles are broadly similar across the four models. This suggests other criteria should be used to decide among the models. I elected to base my choice of model on the cost estimates that resulted from each of the models, as described on the pages that follow.

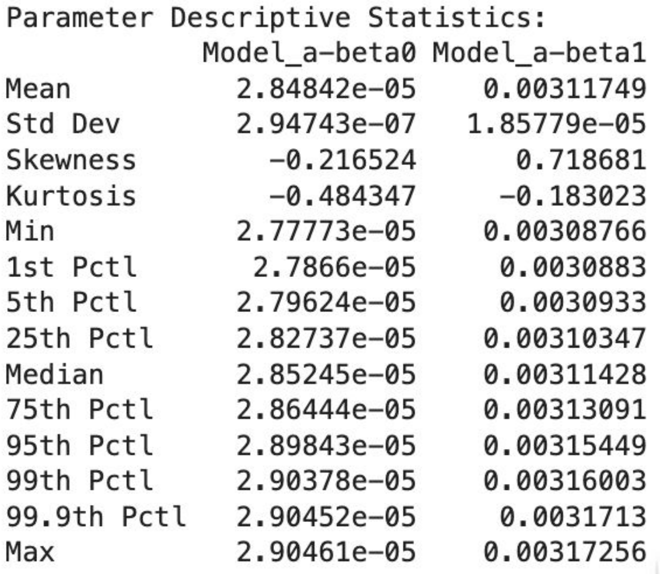


```
MAE Descriptive Statistics:
                  Model_a      Model_b      Model_c      Model_d
Mean           0.00194972   0.00195738   0.00194607   0.0019571
Std Dev       0.000146237  0.000152598  0.000146406  0.00015947
Skewness        -0.196029    0.0273281    -0.190284    0.439901
Kurtosis         0.663422      0.86455     0.679226     2.37691
Min            0.00149957   0.00150324   0.00149835   0.00150184
1st Pctl        0.0016065   0.00161237    0.0016031   0.00160576
5th Pctl       0.00172661   0.00173563   0.00171862   0.00173144
25th Pctl      0.00185468   0.00185906   0.00185034   0.00185464
Median         0.00196288   0.00196653   0.00195891   0.00196436
75th Pctl       0.0020296    0.0020338   0.00202626   0.00203309
95th Pctl      0.00218329   0.00219919   0.00218441   0.00219659
99th Pctl      0.00233402   0.00234439   0.00232645   0.00234528
99.9th Pctl    0.00233674   0.00240225   0.00233887   0.00258058
Max            0.00233704   0.00240867   0.00234025   0.00260673
```

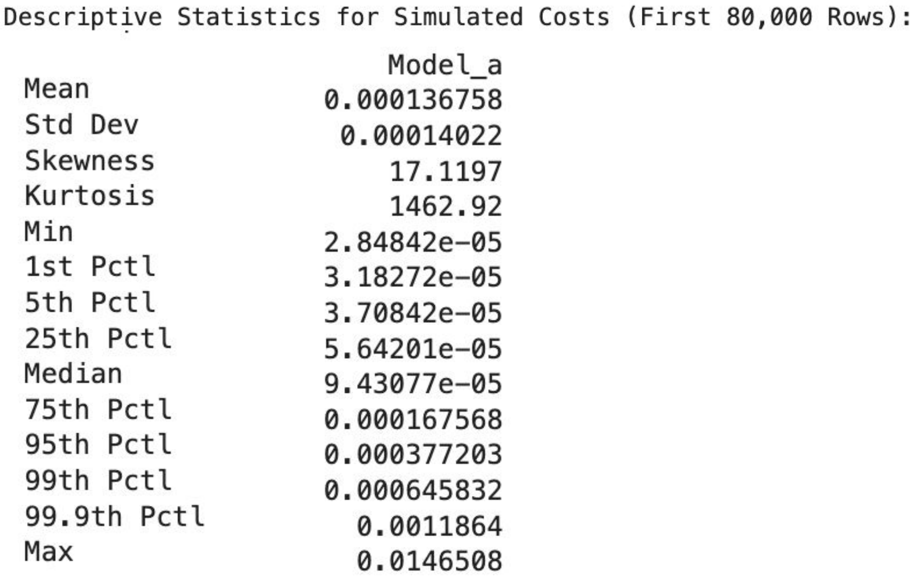
Estimation Results: Parameter and Cost Stats for Model (a)

The mean trade size in our training sample is 0.32%. The mean cost estimate is 1.36758 bps.

$$(a)\ Cost_i = \beta_0 + \beta_1 TradeSize_i^{0.5} + \epsilon_i$$



```
Parameter Descriptive Statistics:
             Model_a-beta0 Model_a-beta1
Mean           2.84842e-05    0.00311749
Std Dev        2.94743e-07   1.85779e-05
Skewness         -0.216524      0.718681
Kurtosis         -0.484347     -0.183023
Min            2.77773e-05    0.00308766
1st Pctl        2.7866e-05     0.0030883
5th Pctl       2.79624e-05     0.0030933
25th Pctl      2.82737e-05    0.00310347
Median         2.85245e-05    0.00311428
75th Pctl      2.86444e-05    0.00313091
95th Pctl      2.89843e-05    0.00315449
99th Pctl      2.90378e-05    0.00316003
99.9th Pctl    2.90452e-05     0.0031713
Max            2.90461e-05    0.00317256
```

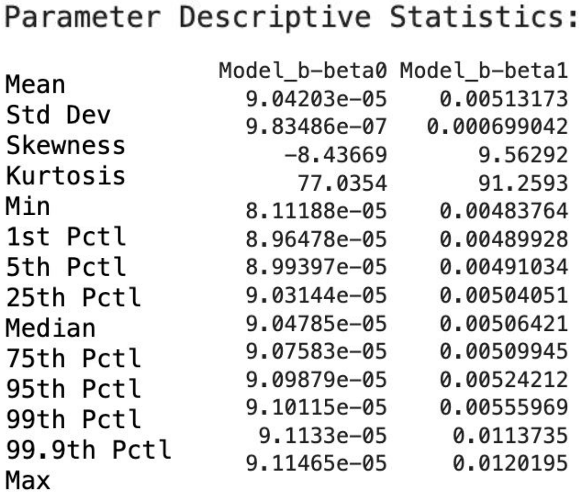


```
Descriptive Statistics for Simulated Costs (First 80,000 Rows):

                    Model_a
Mean            0.000136758
Std Dev          0.00014022
Skewness            17.1197
Kurtosis            1462.92
Min             2.84842e-05
1st Pctl        3.18272e-05
5th Pctl        3.70842e-05
25th Pctl       5.64201e-05
Median          9.43077e-05
75th Pctl       0.000167568
95th Pctl       0.000377203
99th Pctl       0.000645832
99.9th Pctl       0.0011864
Max               0.0146508
```

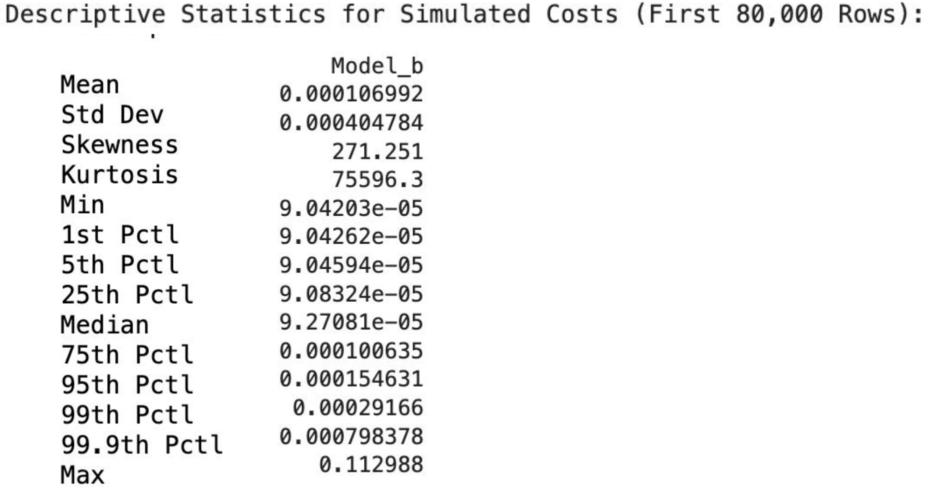
Estimation Results: Parameter and Cost Stats for Model (b)

The mean trade size in our training sample is 0.32%. The mean cost estimate is 1.06992 bps.

$$(b)\ Cost_i = \beta_0 + \beta_1 TradeSize_i + \epsilon_i$$



```
Parameter Descriptive Statistics:

               Model_b-beta0 Model_b-beta1
Mean             9.04203e-05    0.00513173
Std Dev          9.83486e-07   0.000699042
Skewness            -8.43669       9.56292
Kurtosis             77.0354       91.2593
Min              8.11188e-05    0.00483764
1st Pctl         8.96478e-05    0.00489928
5th Pctl         8.99397e-05    0.00491034
25th Pctl        9.03144e-05    0.00504051
Median           9.04785e-05    0.00506421
75th Pctl        9.07583e-05    0.00509945
95th Pctl        9.09879e-05    0.00524212
99th Pctl        9.10115e-05    0.00555969
99.9th Pctl       9.1133e-05     0.0113735
Max              9.11465e-05     0.0120195
```

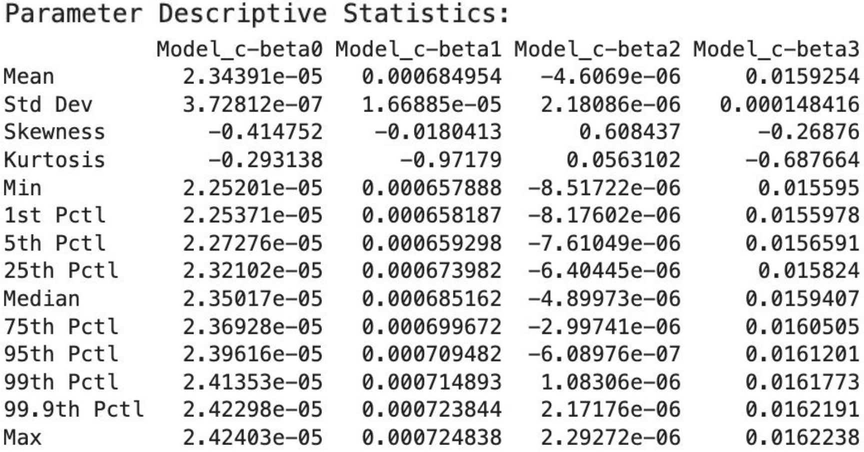


```
Descriptive Statistics for Simulated Costs (First 80,000 Rows):

                  Model_b
Mean          0.000106992
Std Dev       0.000404784
Skewness          271.251
Kurtosis          75596.3
Min           9.04203e-05
1st Pctl      9.04262e-05
5th Pctl      9.04594e-05
25th Pctl     9.08324e-05
Median        9.27081e-05
75th Pctl     0.000100635
95th Pctl     0.000154631
99th Pctl      0.00029166
99.9th Pctl   0.000798378
Max              0.112988
```

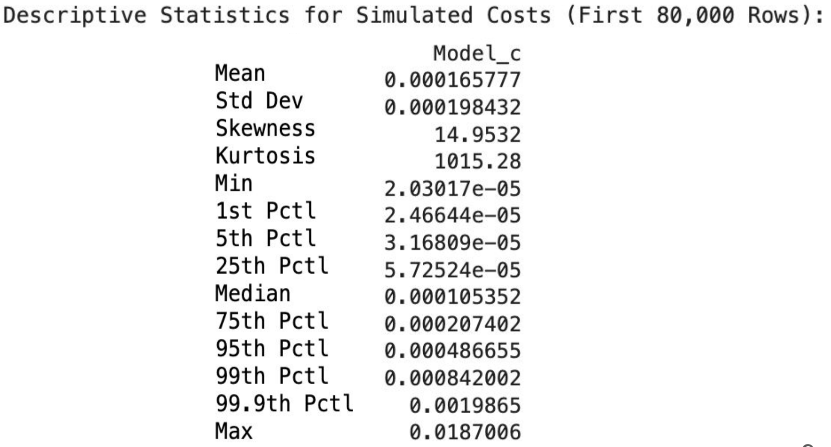
Estimation Results: Parameter and Cost Stats for Model (c)

The mean trade size in our training sample is 0.32%. The mean volatility is 24.1%. The mean cost estimate is 1.65777 bps.

$$(c)\ Cost_i = \beta_0 + \beta_1 TradeSize_i^{0.5} + \beta_2 Volatility_i + \beta_3 Volatility_i \times TradeSize_i^{0.5} + \epsilon_i$$



```
Parameter Descriptive Statistics:
             Model_c-beta0 Model_c-beta1 Model_c-beta2 Model_c-beta3
Mean          2.34391e-05   0.000684954   -4.6069e-06     0.0159254
Std Dev       3.72812e-07   1.66885e-05   2.18086e-06   0.000148416
Skewness        -0.414752    -0.0180413      0.608437      -0.26876
Kurtosis        -0.293138      -0.97179     0.0563102     -0.687664
Min           2.25201e-05   0.000657888  -8.51722e-06      0.015595
1st Pctl      2.25371e-05   0.000658187  -8.17602e-06     0.0155978
5th Pctl      2.27276e-05   0.000659298  -7.61049e-06     0.0156591
25th Pctl     2.32102e-05   0.000673982  -6.40445e-06      0.015824
Median        2.35017e-05   0.000685162  -4.89973e-06     0.0159407
75th Pctl     2.36928e-05   0.000699672  -2.99741e-06     0.0160505
95th Pctl     2.39616e-05   0.000709482  -6.08976e-07     0.0161201
99th Pctl     2.41353e-05   0.000714893   1.08306e-06     0.0161773
99.9th Pctl   2.42298e-05   0.000723844   2.17176e-06     0.0162191
Max           2.42403e-05   0.000724838   2.29272e-06     0.0162238
```

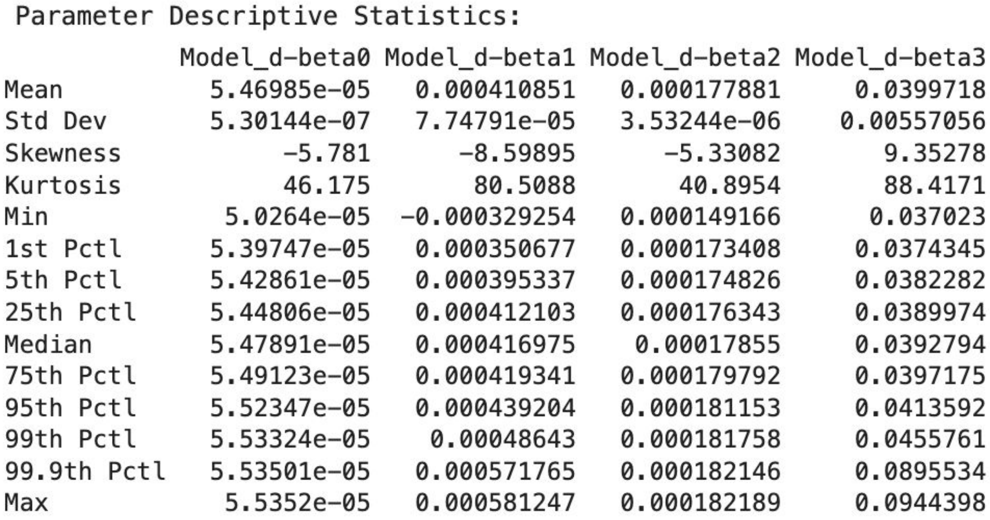


```
Descriptive Statistics for Simulated Costs (First 80,000 Rows):
                  Model_c
Mean          0.000165777
Std Dev       0.000198432
Skewness          14.9532
Kurtosis          1015.28
Min           2.03017e-05
1st Pctl      2.46644e-05
5th Pctl      3.16809e-05
25th Pctl     5.72524e-05
Median        0.000105352
75th Pctl     0.000207402
95th Pctl     0.000486655
99th Pctl     0.000842002
99.9th Pctl     0.0019865
Max             0.0187006
```

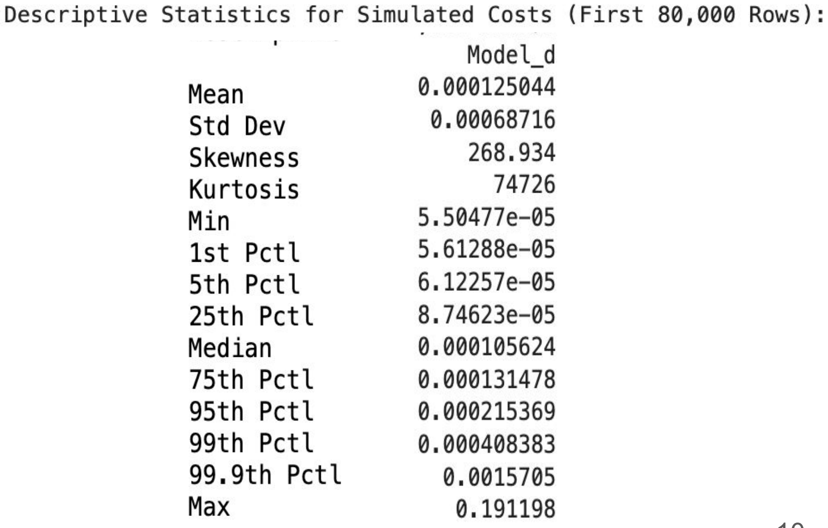
Estimation Results: Parameter and Cost Stats for Model (d)

The mean trade size in our training sample is 0.32%. The mean volatility is 24.1%. The mean cost estimate is 1.25044 bps.

$$(d)\ Cost_i = \beta_0 + \beta_1 TradeSize_i + \beta_2 Volatility_i + \beta_3 Volatility_i \times TradeSize_i + \epsilon_i$$



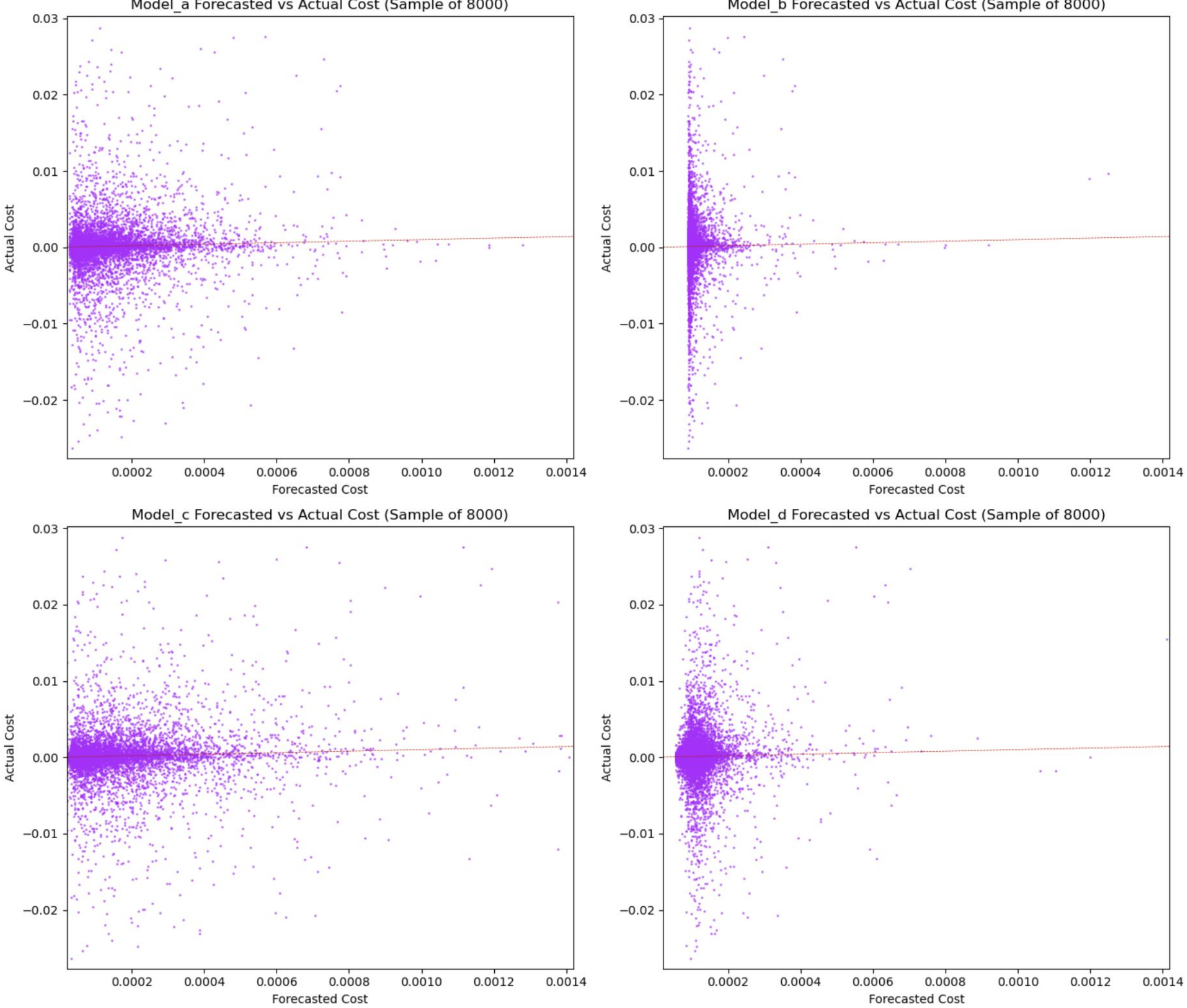
```
 Parameter Descriptive Statistics:

            Model_d-beta0 Model_d-beta1 Model_d-beta2 Model_d-beta3
Mean          5.46985e-05   0.000410851   0.000177881     0.0399718
Std Dev       5.30144e-07   7.74791e-05   3.53244e-06    0.00557056
Skewness           -5.781      -8.59895      -5.33082       9.35278
Kurtosis           46.175       80.5088       40.8954       88.4171
Min            5.0264e-05  -0.000329254   0.000149166      0.037023
1st Pctl      5.39747e-05   0.000350677   0.000173408     0.0374345
5th Pctl      5.42861e-05   0.000395337   0.000174826     0.0382282
25th Pctl     5.44806e-05   0.000412103   0.000176343     0.0389974
Median        5.47891e-05   0.000416975    0.00017855     0.0392794
75th Pctl     5.49123e-05   0.000419341   0.000179792     0.0397175
95th Pctl     5.52347e-05   0.000439204   0.000181153     0.0413592
99th Pctl     5.53324e-05    0.00048643   0.000181758     0.0455761
99.9th Pctl   5.53501e-05   0.000571765   0.000182146     0.0895534
Max            5.5352e-05   0.000581247   0.000182189     0.0944398
```



```
Descriptive Statistics for Simulated Costs (First 80,000 Rows):
                      Model_d
Mean              0.000125044
Std Dev            0.00068716
Skewness              268.934
Kurtosis                74726
Min               5.50477e-05
1st Pctl          5.61288e-05
5th Pctl          6.12257e-05
25th Pctl         8.74623e-05
Median            0.000105624
75th Pctl         0.000131478
95th Pctl         0.000215369
99th Pctl         0.000408383
99.9th Pctl         0.0015705
Max                  0.191198
```

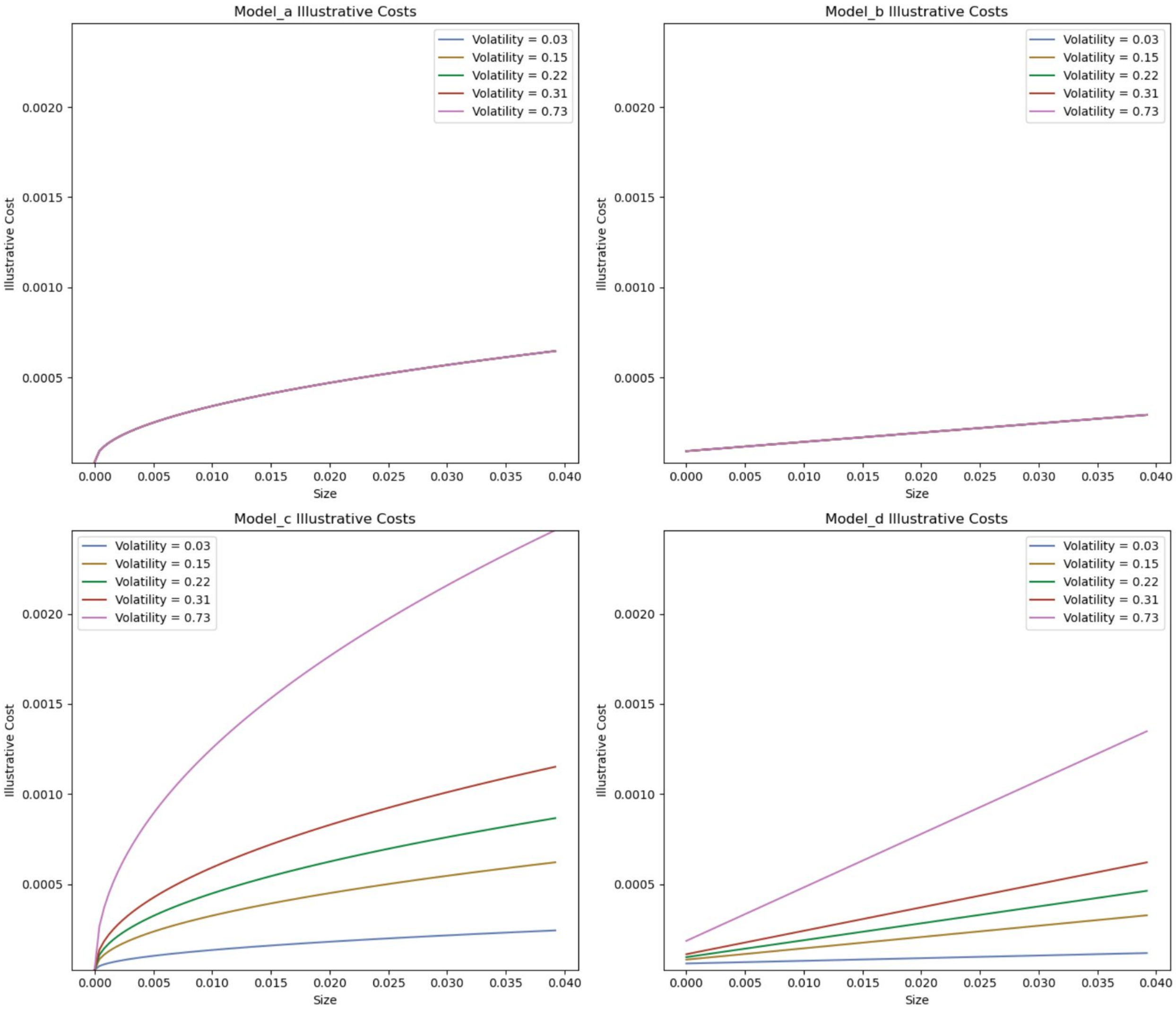
Estimation Results: Scatter plots

The scatter plots show the forecasted vs actual cost for each model.



Estimation Results: Illustrative Costs Distributions

The plots illustrate costs for a range of sizes (1st to 99th percentile) and select volatilities (1st, 25th, 50th, 75th and 99th percentiles).



Model_a Illustrative Costs

Model_b Illustrative Costs

Model_c Illustrative Costs

Model_d Illustrative Costs

Which model should I choose?

While none stand out, and our results argue primarily for more research, if forced I would choose (c) as the most suitable.

- It achieved the lowest MAE (though all four models were very close).
- It yields the largest costs for large trades with high volatility, which seems prudent.

$$(c)\ Cost_i = \beta_0 + \beta_1 TradeSize_i^{0.5} + \beta_2 Volatility_i + \beta_3 Volatility_i \times TradeSize_i^{0.5} + \epsilon_i$$

Prompt Questions 1 and 2

1) Which value of p (0.5 or 1.0) fits the data better? How would you quantify the difference between the two?

Models (a) and (c) have *p* set to 0.5, and their MAEs are consistently lower than those of (b) and (d), which have *p* set to 1. Therefore, according to our chosen metric, they better fit the data.

(2) Is including volatility in the model helpful? If so, how would you quantify the improvement?

Including volatility in the model (in (c) and (d)) appears to have a negligible (though slightly positive) effect on MAE. Therefore including volatility in the model is helpful, albeit only very marginally.

Prompt Questions 3 and 4

3) Using one of the above four models, or any model of your choosing, predict Cost values for the test set. Describe your modeling process. For example, explain how you chose this model and detail any preprocessing or other modeling steps and decisions. How confident are you in its predictions?

My estimation methodology is as previously described. While our results do not appear to be on average unreasonable, we believe more research is required before we could confidently recommended one of the model.   In particular, more data would be helpful.

4) Are there additional models, techniques, or data you would like to try if you had more time and access to data?

1. The dominant feature of the data is that cost is measured with significant noise in the form of market moves. Removing contemporaneous market moves for the duration of each trade would be helpful.
2. Are data from different asset classes? Separating them might be helpful.
3. Presumably bid / offer could be measured directly because it is observable. This could help refine cost estimates.
4. Should some gross outliers (e.g., a record with a trade size of 22x ADV) be removed?
5. Consider using other error metrics (e.g., MSE)