

# Small Data, Big Problems

## Penalized-Constrained Regression for Cost Estimation

Kevin Joy

Max Watstein

2026-02-17

# Table of contents

<b>Abstract</b>	<b>1</b>
<b>1 Introduction</b>	<b>2</b>
<b>2 Motivating Example</b>	<b>3</b>
2.1 The Learning Curve Problem . . . . .	3
2.2 Example Scenario . . . . .	3
<b>3 Theoretical Foundations</b>	<b>5</b>
3.1 Ridge Regularization . . . . .	5
3.2 Lasso Regularization and Constrained Optimization . . . . .	5
<b>4 PCReg Framework</b>	<b>6</b>
4.1 PCReg Model . . . . .	6
4.1.1 Objective Function . . . . .	6
4.1.2 Model Specification . . . . .	7
4.1.3 Cross Validation (CV) for Hyperparameter Selection . . . . .	7
4.2 Model Diagnostics . . . . .	8
4.2.1 Optimization Diagnostics . . . . .	8
4.2.2 In-Sample Fit Statistics . . . . .	8
4.2.3 Generalized Degrees of Freedom (GDF)-Adjusted Fit Statistics . . . . .	8
4.2.4 Bootstrap Confidence Intervals . . . . .	9
4.3 Software Implementation . . . . .	9
4.3.1 Python Package . . . . .	9
4.3.2 Excel Template . . . . .	9
4.4 Example Scenario . . . . .	9
4.4.1 PCReg Model Specification . . . . .	9
4.4.2 PCReg Model Results . . . . .	10
4.4.3 PCReg Model Diagnostics . . . . .	11
<b>5 Simulation Study</b>	<b>13</b>
5.1 Simulation Design . . . . .	13
5.2 Example Scenario . . . . .	13
<b>6 Key Findings</b>	<b>17</b>
6.1 PCReg vs OLS . . . . .	17
6.1.1 PCReg Significantly Outperforms OLS When Coefficients Are Unreasonable . . . . .	17
6.1.2 PCReg Performs Comparably When OLS Coefficients Are Reasonable . . . . .	18
6.2 OLS-LearnOnly Performs Poorly Outside Training Range . . . . .	18
6.3 Coefficient Bias Analysis . . . . .	18
6.4 Parameter Effects on Model Performance . . . . .	19
<b>7 Practical Guidance</b>	<b>21</b>
<b>8 Limitations</b>	<b>22</b>
<b>9 Summary</b>	<b>23</b>

<b>References</b>	<b>24</b>
<b>Acronyms</b>	<b>25</b>
<b>A PCReg Mathematical Notation</b>	<b>26</b>
A.1 Complete Notation Table . . . . .	26
A.2 Prediction Function Forms . . . . .	26
A.3 Cost Estimating Terminology Mapping . . . . .	26
<b>B Other Models to Consider for Future Methods</b>	<b>27</b>
B.1 Model Descriptions . . . . .	27
B.2 Preliminary Results . . . . .	27
B.3 Future Work . . . . .	28
<b>C Software Documentation</b>	<b>29</b>
C.1 Installation . . . . .	29
C.1.1 Installing from GitHub . . . . .	29
C.1.2 Cloning the Repository . . . . .	29
C.1.3 Downloading the Excel Template . . . . .	29
C.2 Basic Usage . . . . .	29
C.3 Citation . . . . .	30
<b>D Detailed Diagnostic Report</b>	<b>31</b>
D.1 Model Specification . . . . .	31
D.2 Coefficient Estimates . . . . .	31
D.3 Fit Statistics . . . . .	31
D.4 Constraint Summary . . . . .	31
D.5 Bootstrap Results . . . . .	32
D.5.1 Bootstrap Summary Statistics . . . . .	32
D.6 Data Summary . . . . .	32
D.7 Model Equation . . . . .	32

# Abstract

Small datasets with intercorrelation pose serious challenges to the stability of coefficients generated by Ordinary Least Squares (OLS) regression. A motivating example in cost estimating is Learning Curve with Rate Effect analysis, where datasets are typically small, the lot midpoint (Learning) is correlated to lot quantity (Rate) as production ramps up, and slopes are expected to be  $\leq 100\%$ . Lasso, Ridge, and Elastic Net regularization methods address multicollinearity by penalizing coefficients. Separately, constrained optimization methods can impose explicit restrictions on coefficient values when prior knowledge about their behavior is known, such as bounding slopes within a known range. This paper investigates the combined effects of penalized regularization methods and constrained optimization. We explore how to assess model stability and goodness-of-fit using likelihood-free diagnostic techniques suited to optimization-based regressions, such as cross-validation for generalization error.

# 1 Introduction

While developing Cost Estimating Relationships (CERs) for small datasets (5-30 data points), a recurring pattern emerged. The regression models often showed strong fit statistics with high  $R^2$  and low standard error, but nonsensical coefficients. Coefficients often had wrong signs, implausible magnitudes, and poor p-values. As Department of Defense (DoD) analysts, this story may feel all too familiar.

Intercorrelated datasets are a frequent presence in cost analysis, causing regression models to misbehave. This is especially true for Learning Curve with Rate Effect analysis, where datasets are typically small and the lot midpoint is usually correlated to lot quantity. Traditional remedies for multicollinearity can be insufficient for learning curve datasets. Increasing sample size is infeasible as it requires waiting years for additional production lots and cost data to become available. Dropping explanatory variables, such as rate effect, leads to model misspecification that ignores a fundamental cost driver. This issue is compounded when predicting outside the relevant range, which is very common. Even when these remedies are employed, OLS can still produce implausible coefficients with the wrong signs, such as learning curve slopes (LCSs)  $>100\%$ . Regularization techniques like Ridge, Lasso, and Elastic Net can help stabilize estimates, but they don't guarantee plausible coefficients. Constrained optimization can enforce domain knowledge, but without regularization, it may not solve the underlying multicollinearity problem. This paper explores the combination of penalized regularization and constrained optimization to address both challenges.

For our research, we reviewed existing literature on multicollinearity, regularization, constrained optimization, and learning theory. Drawing on these foundations, we developed the Penalized-Constrained Regression (PCReg) framework, which integrates Elastic Net penalties with domain-knowledge coefficient constraints. We then used Monte Carlo simulation to evaluate PCReg's accuracy estimating the true parameters and compared its performance against OLS across a range of sample sizes and collinearity conditions, measuring both coefficient recovery and out-of-sample prediction accuracy.

This paper develops and validates PCReg for cost estimation, contributing the following:

1. **Model:** Integration of Elastic Net penalties and constrained optimization
2. **Diagnostic:** Bootstrap confidence intervals and Generalized Degrees of Freedom (GDF)-adjusted fit statistics for optimization-based regression (see Section 4.2.3 for discussion of GDF)
3. **Practical Guidance:** When to use PCReg and how to assess model fit

Additionally, we have developed the following tools for the cost estimating community to apply PCReg:

1. **Python Package:** PCReg application in Python
2. **Excel Template:** PCReg application in Excel with limited implementation

## 2 Motivating Example

### 2.1 The Learning Curve Problem

The motivating example for this research is the learning curve with rate effect:

$$\text{Average Unit Cost} = T_1 \cdot (\text{Lot Midpoint})^b \cdot (\text{Lot Quantity})^c \cdot \varepsilon$$

$$\begin{aligned} \text{where: } T_1 &= \text{theoretical first unit cost} \\ b &= \text{learning slope in log space} \\ c &= \text{rate slope in log space} \\ \varepsilon &= \text{multiplicative error} \end{aligned} \tag{1}$$

$b$  and  $c$  are both slopes in log space. When transformed to unit space, they become the LCS ( $\text{LCS} = 2^b$ ) and the rate effect ( $2^c$ ), both typically ranging from 70-100%. As DoD programs ramp up production from Engineering and Manufacturing Development (EMD) to Full Rate Production (FRP), lot midpoint (the learning predictor) becomes highly correlated with lot quantity (the rate predictor). By definition, LCS and Rate Effect are  $\leq 100\%$ , meaning costs strictly decrease with cumulative production or lot quantity.

If estimated LCS or Rate Effect are  $>100\%$ , this is often an indicator of multicollinearity, or it may suggest that other explanatory variables, such as supply chain disruptions or diseconomies of scale, are missing and need to be modeled. This does not indicate a violation of the definition itself. Once these factors are properly modeled as separate explanatory variables, estimated LCS and Rate Effect should be  $\leq 100\%$ , consistent with their definition. It is the analyst's responsibility to ensure the model is properly specified and includes all relevant cost drivers.

### 2.2 Example Scenario

We demonstrate the “Learning Curve Problem” using the following dataset which was generated by our Simulation Study (Section 5).

Table 1: Example Scenario Dataset

lot_midpoint	lot_quantity	observed_cost
11.694	30.000	44.822
44.592	30.000	43.961
79.598	40.000	42.172
129.251	60.000	39.040
226.613	140.000	38.229
414.709	241.000	26.349
713.347	360.000	22.446
1076.104	360.000	21.463
1437.463	360.000	24.394
1812.743	390.000	24.922

Using OLS, we get the following results:

Table 2: Example Scenario OLS Results

Metric	OLS
$T_1$	114
LCS	100.8%
Rate Effect	82.5%
$R^2$	0.91
F-statistic	36.7 (p=0.00019)

We find that by using OLS, this example dataset yields invalid coefficients, specifically an LCS of 100.8% which violates our definition of LCS being  $\leq 100\%$ . But otherwise this is a good fit with strong  $R^2$  (0.91) and significant F-statistic (p=0.00019). This is a common pattern in cost estimating where OLS produces good fit statistics but implausible coefficients.

## 3 Theoretical Foundations

### 3.1 Ridge Regularization

C. M. Theobald, in *Generalizations of Mean Square Error Applied to Ridge Regression* (Theobald 1974), showed that for any OLS problem there exists a ridge parameter  $\lambda^* > 0$  such that the ridge estimator has strictly lower Mean Squared Error (MSE) than OLS. R. W. Farebrother, in *Further Results on the Mean Square Error of Ridge Regression* (Farebrother 1976), extended Theobald’s result to broader classes of weighted MSE criteria. Together, these results establish that ridge shrinkage can always improve on OLS in terms of population MSE, not merely training error. The optimal  $\lambda^*$  depends on unknown population parameters, which cross-validation estimates empirically.

### 3.2 Lasso Regularization and Constrained Optimization

Our paper is not the first to explore combining regularization (L1 penalties) with constraint optimization. James, Paulson, and Rusmevichientong (2020) developed the *Penalized and Constrained* optimization method (PAC) “to compute the solution path for high-dimensional, linearly-constrained criteria”. The paper demonstrated that PAC methods remain superior to unconstrained alternatives even when constraints are imperfectly specified:

*“PAC and relaxed PAC are surprisingly robust to random violations in the constraints... they were still both superior to their unconstrained counterparts for all values [of constraint error] and all settings.”*



## 4 PCReg Framework

The PCReg framework addresses the Learning Curve Problem, highlighted by our Section 2.2, by applying the Section 3 research findings to the cost estimation domain. Because PCReg uses optimization rather than closed-form estimation, traditional OLS diagnostics are not directly available, so PCReg also provides specialized model diagnostics based on GDF, bootstrap inference, and cross-validation.

### 4.1 PCReg Model

The PCReg Model integrates regularization penalties (Lasso [L1] and Ridge [L2]) with domain-knowledge constraints ( $\theta_{\text{lower}} \leq \theta \leq \theta_{\text{upper}}$ ) to stabilize coefficient estimates, while ensuring they remain within plausible ranges.

**Notation:** We use  $\theta$  (rather than  $\beta$ ) to emphasize that PCReg handles both linear regression and nonlinear optimization problems. For standard linear models,  $\theta$  represents regression coefficients; for nonlinear models (e.g., learning curves),  $\theta$  represents the optimized parameters. See Section A in the Appendix for complete notation details.

#### 4.1.1 Objective Function

The objective function minimized by PCReg combines a loss function with regularization penalties subject to coefficient bounds:

$$\text{Objective: } \arg \min_{\theta} \underbrace{\sum_{i=1}^n L(y_i, \hat{y}_i)}_{\text{Loss}} + \underbrace{\lambda \left[ \alpha \|\theta\|_1 + \frac{1-\alpha}{2} \|\theta\|_2^2 \right]}_{\text{Elastic Net Penalty}}$$

$$\text{Subject to: } \theta_{\text{lower}} \leq \theta \leq \theta_{\text{upper}} \quad (\text{parameter bounds})$$

$$\begin{aligned} \text{where: } \quad \hat{y}_i &= f(X_i, \theta) &&= \text{predicted value} \\ L(y_i, \hat{y}_i) &&&= \text{loss function (e.g., SSE, SSPE)} \\ \lambda &\geq 0 &&= \text{regularization penalty} \\ \alpha &\in [0, 1] &&= \text{L1/L2 ratio} \\ \|\theta\|_1 &= \sum_j |\theta_j| &&= \text{Lasso penalty} \\ \|\theta\|_2^2 &= \sum_j \theta_j^2 &&= \text{Ridge penalty} \end{aligned}$$

The lower and upper bounds ( $\theta_{\text{lower}}$  and  $\theta_{\text{upper}}$ ) can be set to achieve loose bounds or tight bounds for the coefficient:

- **Loose bounds** (e.g., 70%-100%): Wide range which allows flexibility while preventing egregious coefficient violations
- **Tight bounds** (e.g., 85%-95%): Narrower range which incorporates strong prior knowledge but risks over-constraining coefficients

### 4.1.2 Model Specification

Users can specify the following components of the PCReg model:

- **Functional Form:** Defaults to linear (`X @ coef + intercept`), but any function can be entered
- **Loss function:** Defaults to Sum of Squared Percentage Errors (SSPE), but alternatives include Sum of Squared Errors (SSE), MSE, or any custom loss function
- **Coefficient bounds:** Componentwise constraints that encode domain knowledge (e.g., LCS between 70% and 100%)
- **Penalty exclusion:** Option to exclude specific coefficients from regularization while still applying bounds
- **Optimization Method:** Defaults to Sequential Least Squares Programming (SLSQP) for constrained optimization, but any optimization method can be used

The framework maximizes flexibility and collapses to known regression techniques in special cases:

- OLS when functional form is linear,  $\alpha = 0$ , and bounds are not specified
- Ridge when functional form is linear,  $\alpha > 0$  and `l1_ratio` = 0
- Lasso when functional form is linear,  $\alpha > 0$  and `l1_ratio` = 1
- Elastic Net when functional form is linear,  $\alpha > 0$  and  $0 < \text{`l1\_ratio`} < 1$
- Constrained regression when bounds are specified but  $\alpha = 0$

### 4.1.3 Cross Validation (CV) for Hyperparameter Selection

Finding the optimal penalty parameters for PCReg requires testing different values rather than using a formula. CV is a method that evaluates how well a model predicts new data by testing it on observations not used during training. The CV process evaluates all combinations of user-specified penalties ( $\lambda$ ) and L1 ratio values ( $\alpha$ ) to find the hyperparameters that produce the best out-of-sample predictions.

Traditional K-Fold CV (Stone 1974) splits the training data into K subsets, repeatedly training on K-1 subsets and validating on the remaining one. For small datasets, this approach further reduces the already limited training sample size.

Generalized Cross Validation (GCV) approximates leave-one-out cross validation using a mathematical formula rather than actually splitting data (Golub, Heath, and Wahba 1979). It estimates prediction error based on training residuals adjusted by model complexity (degrees of freedom), making it efficient for small samples.

PCReg implements both K-Fold CV and GCV, with GCV as the default for small datasets. For each combination of penalty parameters, the GCV score is calculated as:

$$\text{GCV} = \frac{n \cdot \text{RSS}}{(n - \text{df})^2}$$

where:    `RSS`    = Residual Sum of Squares  
             `n`        = number of observations  
             `df`       = effective degrees of freedom

The optimal penalty parameters are those that minimize the GCV score. See Section 4.2.3 for how PCReg calculates effective degrees of freedom for constrained and penalized models.

## 4.2 Model Diagnostics

The flexibility of PCReg means that classical inference (p-values and F-statistics) cannot be guaranteed, because the framework supports coefficient bounds, regularization penalties, custom loss functions, and custom functional forms. No single inference framework covers all scenarios. When coefficients are at bounds (from constraints or L1 penalties), classical theory is no longer directly applicable. Additionally, introducing custom loss and functional forms requires specifying their corresponding distributional assumptions as in Generalized Linear Models (GLMs). We therefore focus on fit statistics that are robust and informative across model specifications.

### 4.2.1 Optimization Diagnostics

We report optimization diagnostics such as convergence status, number of iterations, and whether a parameter is at a bound. When coefficients are at bounds, this indicates that constraints are actively shaping the solution. We have found that applying constraints can change parameters without forcing to a bound. Just because a coefficient is not at a bound does not mean it is unaffected by the constraints; constraints can still significantly impact the optimization process.

### 4.2.2 In-Sample Fit Statistics

Standard in-sample fit measures remain valid across all model configurations, including Coefficient of Determination ( $R^2$ ), SSE, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and residual diagnostics.

### 4.2.3 Generalized Degrees of Freedom (GDF)-Adjusted Fit Statistics

Generalized Cross Validation (GCV), Standard Error of Estimate (SEE), and Adjusted  $R^2$  all rely on degrees of freedom. However, classical degrees of freedom ( $df = n - p$ ) assumes unconstrained estimation. GDF suggests that optimization-based models should lose degrees of freedom to take into account the constraints. We implement both Gaines, Kim, and Zhou (2018) and Hu (2010) GDF formulas to properly adjust fit statistics (SEE, Standard Percentage Error (SPE), Adjusted  $R^2$ ) and account for the effective model complexity.

**Gaines et al. (2018) Formula (Default):** Gaines, Kim, and Zhou (2018) derive degrees of freedom for constrained Lasso by counting only constraints that are actually binding at the solution:

$$\text{GDF}_{\text{Gaines}} = n - |\text{Active predictors}| - |\text{Equality constraints}| - |\text{Binding inequality constraints}|$$

where  $n$  is the sample size and active predictors are non-zero coefficients. This reflects the intuition that non-binding constraints don't reduce effective degrees of freedom.

**Hu (2010) Alternative Formula:** Hu (2010) proposes a more conservative approach where *all specified constraints* count against degrees of freedom, regardless of whether they bind:

$$\text{GDF}_{\text{Hu}} = n - p - (\# \text{ Constraints}) + (\# \text{ Redundancies})$$

where  $p$  is the total number of estimated parameters (including intercept if fitted). Hu’s method always reduces GDF when constraints are specified, making it appropriate for Minimum-Percentage Error under the Zero-Percentage Bias (ZMPE)-type CERs where constraints fundamentally shape the solution.

#### 4.2.4 Bootstrap Confidence Intervals

Standard bootstrap confidence intervals can be misleading for penalized regression (Goeman 2025), as they assume interior solutions and no regularization bias. Our framework addresses this by computing **both** constrained and unconstrained bootstrap resampling:

- **Constrained bootstrap:** Resamples data and refits the model *with* the specified bounds and regularization (penalized-constrained estimation), quantifying coefficient stability under the imposed modeling constraints
- **Unconstrained bootstrap:** Resamples data and fits OLS *without* bounds or penalization (pure unconstrained estimation), showing coefficient variability in the absence of constraints and regularization. This serves as a baseline for comparison.

### 4.3 Software Implementation

The DoD cost estimating community can apply PCReg using either a Python Package or Excel Template.

#### 4.3.1 Python Package

We developed a full feature Python package, [penalized\\_constrained](#), following scikit-learn conventions and API design principles. The package provides a comprehensive implementation of the PCReg framework with support for model fitting, hyperparameter selection via cross-validation, diagnostic reporting with GDF-adjusted fit statistics, and bootstrap confidence intervals. See Appendix Section C for additional information.

#### 4.3.2 Excel Template

To make the methodology accessible to practitioners without programming expertise, we developed an [Excel Template](#) that implements a simplified version of the PCReg algorithm. The template provides an intuitive interface for applying PCReg to cost estimation problems. Visit the [GitHub repository](#) for more information.

### 4.4 Example Scenario

Now we will apply PCReg to the example scenario introduced earlier.

#### 4.4.1 PCReg Model Specification

We apply the following PCReg Model Specification to the example scenario:

```
# Define the nonlinear functional form
def prediction_fn(X, params):
    """Learning curve with rate effect: T1 * (midpoint^b) * (quantity^c)"""
    T1, b, c = params
```

```

    return T1 * (X[:, 0] ** b) * (X[:, 1] ** c)

# Configure PCReg model
pc_gcv = pcreg.PenalizedConstrainedCV(
    coef_names=['T1', 'b', 'c'],
    bounds={'T1': (0, None), 'b': (-0.5, 0), 'c': (-0.5, 0)},
    prediction_fn=prediction_fn,
    fit_intercept=False,
    x0=[100, -0.1, -0.1],
    alphas=np.linspace(0, 5, 10), # Range of lambda values to test
    l1_ratios=[0.0, 0.5, 1.0], # Test Ridge, Elastic Net, and Lasso penalties
    selection='gcv',           # Generalized Cross-Validation
    loss='sspe',               # Sum of Squared Percentage Errors
    penalty_exclude=['T1'],    # Don't penalize the intercept
    n_jobs=1
)

# Fit the model
pc_gcv.fit(X_train, y_train)

```

*Note:* Due to scikit-learn API conventions,  $\lambda$  is labeled **alpha** and  $\alpha$  is labeled **l1\_ratio** in the code. These correspond to regularization strength and L1/L2 mixing ratio in the theoretical formulation.

#### Key Specification Details:

- **Coefficients:**  $T_1$  (first unit cost),  $b$  (learning slope),  $c$  (rate slope)
- **Bounds:**  $T_1 > 0$ ,  $-0.5 \leq b \leq 0$ ,  $-0.5 \leq c \leq 0$
- **alphas (penalty):** Range between 0 and 5 using linear spacing to explore Lasso, Ridge, and Elastic Net penalties
- **l1\_ratios:** Range between 0 and 1 to explore different mixes of L1 and L2 regularization
- **Penalty Selection:** GCV automatically selects optimal  $\lambda$  without data splitting
- **Loss Function:** SSPE minimizes percentage errors (appropriate for cost data)
- **Penalty Exclusion:**  $T_1$  is not penalized (only slopes are regularized)

Below is a comparison of OLS and PCReg specifications used for this example scenario:

Table 3: Comparison of OLS and PCReg coefficient Specification

Specification	OLS	PCReg
loss function	SSE	SSPE
functional form	Log-Linear	Nonlinear
lambda (penalty)	N/A	1.6667
alpha (L1 ratio)	N/A	0.00

#### 4.4.2 PCReg Model Results

Using OLS and PCReg, we get the following results:

Table 4: Comparison of OLS and PCReg coefficient estimates

Metric	OLS	PCReg
$T_1$	114	100
LCS	100.8%	96.4%
Rate Effect	82.5%	88.3%
$R^2$ (Train)	0.91	0.90

We see that the PCReg model yields a plausible LCS (96.4%), however the  $R^2$  is slightly lower than the OLS regression.

#### 4.4.3 PCReg Model Diagnostics

We were able to find a reasonable solution with PCReg that satisfies our domain knowledge constraints (LCS and Rate Effect  $\leq 100\%$ ) and we still have a reasonable  $R^2$ . Now we must assess the stability of the coefficients. The PCReg model includes bootstrap diagnostics to assess coefficient stability and the impact of constraints. Comparing constrained and unconstrained bootstrap distributions reveals the impact of constraints on coefficient stability. When the constrained and unconstrained distributions differ substantially, constraints are actively shaping the solution. When they are similar, the data naturally satisfies the constraints. The constrained and unconstrained bootstrap results for each coefficient are summarized below.

Table 5: Constrained Bootstrap (with bounds and regularization)

Coefficient	Mean	Std	95% CI
$T_1$	99.48	7.85	[82.69, 120.57]
$b$	-0.0650	0.0285	[-0.1378, -0.0284]
$c$	-0.1633	0.0329	[-0.2076, -0.0932]

Table 6: Unconstrained Bootstrap (no bounds,  $\alpha=0$ )

Coefficient	Mean	Std	95% CI
$T_1$	109.74	15.40	[100.00, 143.55]
$b$	-0.0050	0.0709	[-0.1781, 0.1119]
$c$	-0.2510	0.0916	[-0.3988, -0.0546]

Figure 1 shows the Bootstrap coefficient distributions comparing constrained (blue) vs unconstrained (red) estimation. Vertical lines show fitted values. Constraints reduce variance and keep estimates within economically plausible ranges.

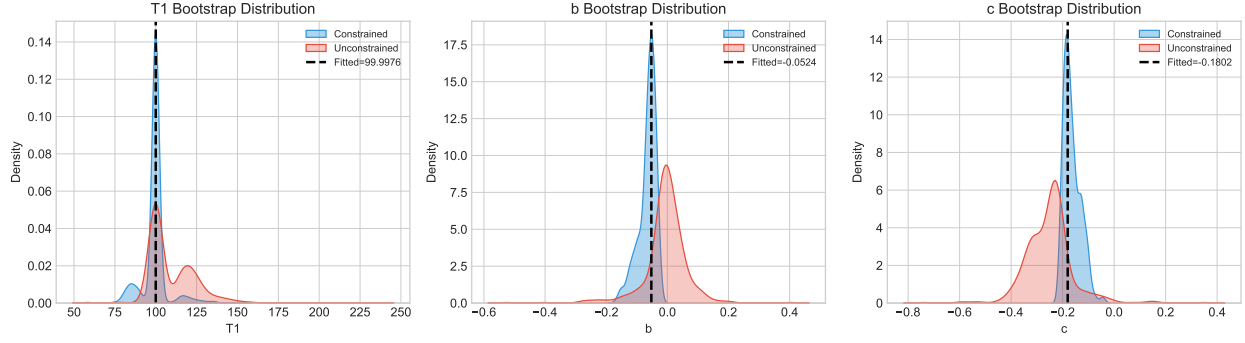


Figure 1: Bootstrap Coefficient Distributions

The bootstrap results above reveal three key diagnostic insights:

1. **Constraints at bounds:** When bootstrap samples frequently hit constraint boundaries, the data is pulling towards implausible values, which is exactly when constraints help most.
2. **Variance reduction:** The constrained bootstrap shows tighter distributions than the unconstrained, reducing coefficient uncertainty at the cost of some bias.
3. **Divergence indicates constraint impact:** The differences between constrained and unconstrained means indicate how much the constraints are actively shaping the solution.

A comprehensive diagnostic report with full model specifications and bootstrap distributions is provided in Appendix Section [D](#).

## 5 Simulation Study

We performed a Monte Carlo simulation study to evaluate PCReg and compare with traditional OLS.

### 5.1 Simulation Design

Our simulation study specifies four factors and generates 8,100 scenarios (81 factor combinations  $\times$  100 replications). The four factors and their corresponding levels are provided below:

Table 7: Simulation study factorial design

Factor	Levels
Number of Training Lots	5, 10, 30
CoV error	0.01, 0.1, 0.2
LCS	85%, 90%, 95%
Rate effect	80%, 85%, 90%

For each scenario, we:

1. **Randomly selected a quantity profile** from the Selected Acquisition Report (SAR) database (actual defense program procurement histories)
2. **Generated simulated average unit costs** using the learning curve model (Equation 1) with the scenario’s true parameters
3. **Added multiplicative lognormal noise** with the specified Coefficient of Variation (CoV)
4. **Split data** with first  $n$  training lots used for model fitting and remaining lots reserved for testing
5. **Fit models** to the training data using the following approaches:
  1. **OLS**: Learning curve with rate effect
  2. **OLS-LearnOnly**: Learning curve without rate effect, representing the common remedy of dropping a correlated predictor
  3. **PCReg**: Employed default PCReg configuration
6. **Evaluated Models** comparing the estimated coefficients against the true parameters, measuring coefficient recovery and out-of-sample prediction accuracy

This approach ensures realistic lot structures (varying quantities, realistic ramp-up patterns) while controlling the true underlying parameters. The number of test lots varies by scenario depending on the program’s total lot history. Some programs have many available lots beyond training, others have few or none. This variability is acceptable as it reflects real-world conditions.

### 5.2 Example Scenario

The following tables describe the example scenario. Table 8 shows the simulation factor levels selected for this scenario from the factorial design (Table 7), and Table 9 shows the derived characteristics specific to this scenario.



Table 8: Example Scenario Simulation Factors

Factor	Value
Number of Training Lots	10
CoV Error	0.1
True LCS	95.0%
True Rate Effect	90.0%

Table 9: Example Scenario Derived Characteristics

Characteristic	Value
Total lots (from SAR)	30
Test lots	20
True $T_1$	100
Predictor Correlation	0.95

Below is the example scenario simulation dataset with training and test lots. The training data matches the dataset shown earlier, but now we also see the test lots with their true underlying costs.

Table 10: Motivating example dataset with training and test lots

lot_type	lot_midpoint	lot_quantity	observed_cost	true_cost
train	11.694	30	44.822	49.71
train	44.592	30	43.961	45.022
train	79.598	40	42.172	41.287
train	129.251	60	39.040	37.451
train	226.613	140	38.229	31.586
train	414.709	241	26.349	27.811
train	713.347	360	22.446	25.136
train	1076.1	360	21.463	24.382
train	1437.46	360	24.394	23.865
train	1812.74	390	24.922	23.176
test	2208.26	400	—	22.752
test	2669.38	525	—	21.526
test	3231.51	600	—	20.798
test	3789.4	512	—	21.056
test	4320.37	550	—	20.627
test	4866.8	542	—	20.491
test	5408.09	540	—	20.343
test	5955.19	554	—	20.12
test	6509.39	554	—	19.988
test	7063.56	554	—	19.868
test	7615.72	550	—	19.779
test	8165.84	550	—	19.677
test	8715.95	550	—	19.583

lot_type	lot_midpoint	lot_quantity	observed_cost	true_cost
test	9266.04	550	—	19.494
test	9816.12	550	—	19.411
test	10366.2	550	—	19.333
test	10916.3	550	—	19.259
test	11466.3	550	—	19.189
test	12016.4	550	—	19.123
test	12556.5	530	—	19.168

Figure 2 shows the learning curve data with 10 training lots (blue) and the true underlying cost distribution (dashed line). Point sizes reflect lot quantities. The goal is to predict the true relationship, not just fit the noisy observations.

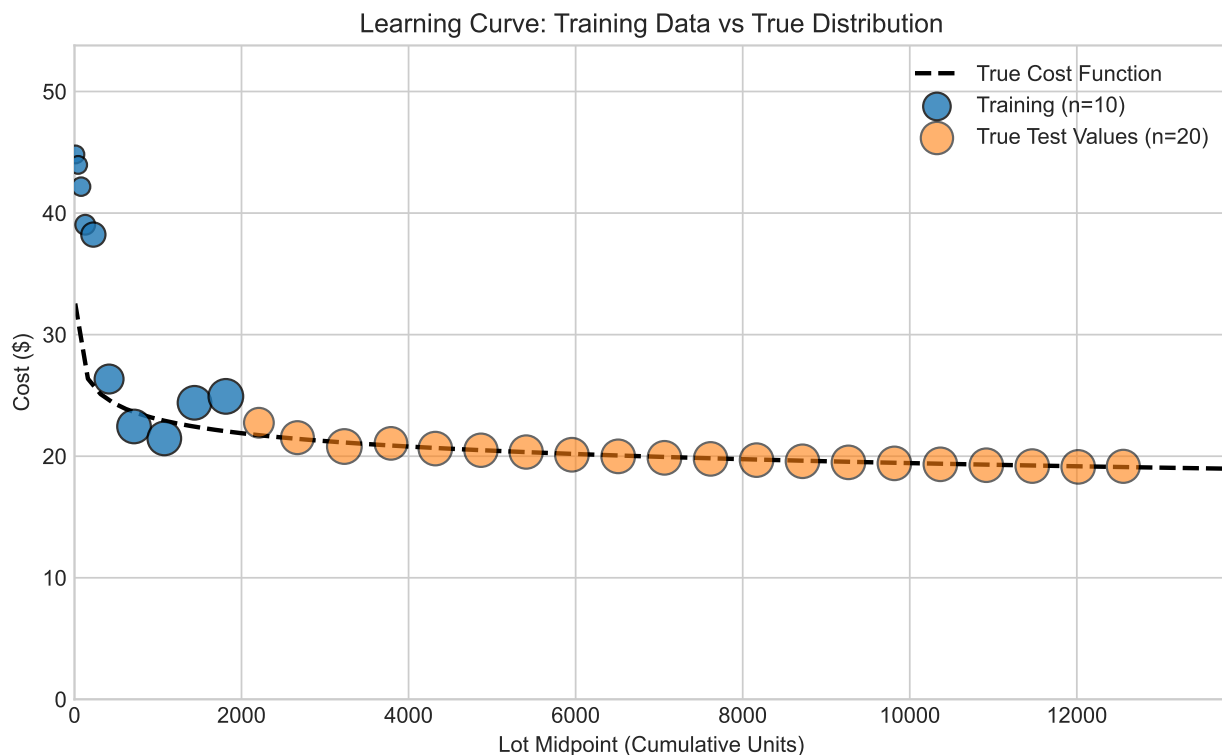


Figure 2: Example Scenario Learning Curve

OLS results in a *higher* Train  $R^2$  (0.912) than PCReg (0.907). This is expected because penalties and constraints add bias to the training fit. However, PCReg results in a lower (better) MAPE on on out-of-sample Test data, meaning its predictions are closer to the true underlying costs. For the remainder of the paper we will refer to MAPE on Test data as Test MAPE.

Table 11: Comparison of estimation methods.

Metric	True	OLS	OLS-LearnOnly	PCReg
$T_1$	100	114	79	100
LCS	95.0%	100.8%	89.2%	97.4%
Rate Effect	90.0%	82.5%	—	87.3%
Valid Coefficients	Yes	NO	Yes	Yes
Train $R^2$	—	0.912	0.796	0.907
Test MAPE (vs True)	—	9.8%	8.4%	3.9%

Figure 3 shows the Residuals for each model on training data (left) and against true test values (right). PCReg shows larger training residuals but smaller errors when predicting the true underlying relationship.

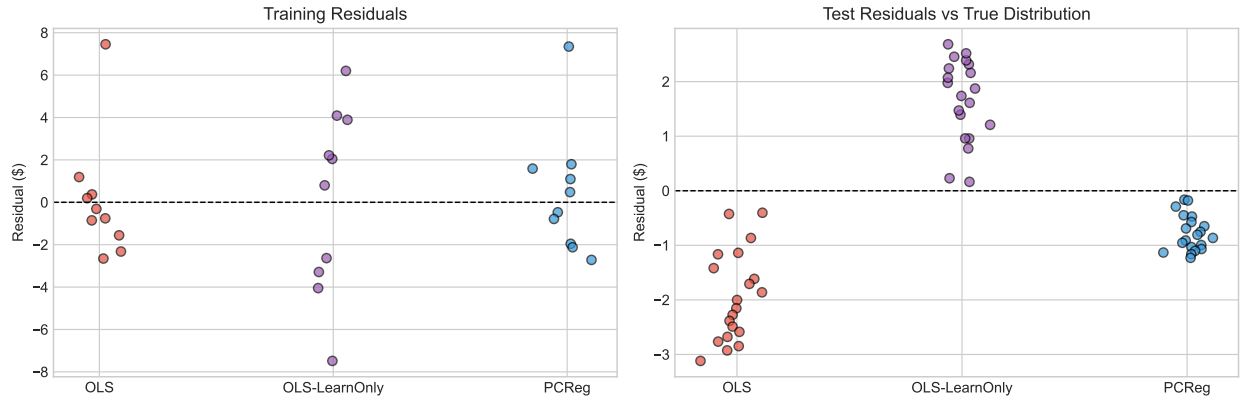


Figure 3: Example Scenario Residuals

## 6 Key Findings

### 6.1 PCReg vs OLS

Across 8,100 simulation scenarios, OLS produced economically unreasonable coefficients, as measured by LCS or rate effect outside 70-100%, in 1,355 scenarios which is **16.7%** of cases. Win rates are determined by Test MAPE score, where the model with the lower Test MAPE wins.

Table 12: PCReg vs OLS

OLS Coefficients	Scenarios	% of Scenarios	OLS Win Rate	PCReg Win Rate
Reasonable (70-100%)	6,745	83.3%	66.3%	33.7%
Unreasonable (<70% or >100%)	1,355	16.7%	21.3%	78.7%
<b>Overall</b>	8,100	100.0%	58.8%	41.2%

Figure 4 shows the distribution of Test MAPE for OLS vs PCReg, stratified by whether OLS produced reasonable coefficients. When unreasonable (right), OLS shows a heavy right tail of large errors that PCReg avoids.

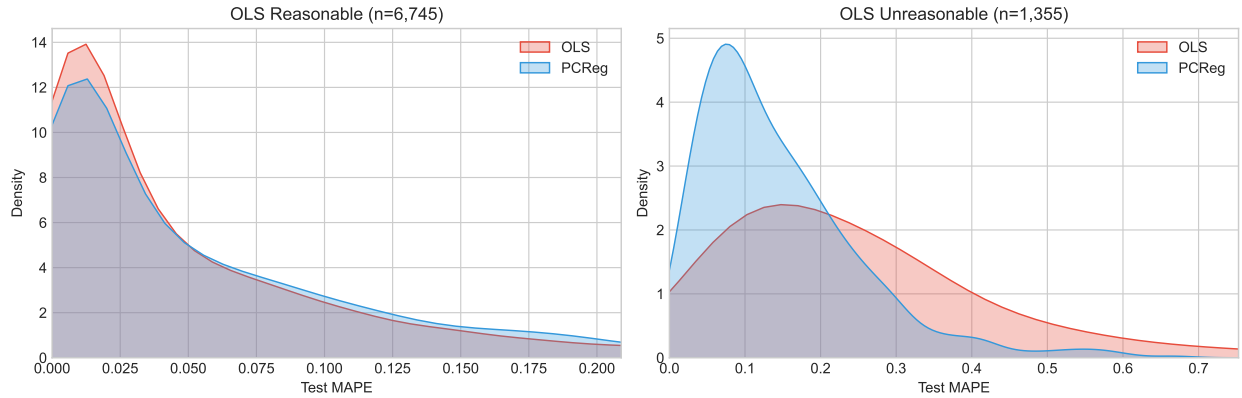


Figure 4: Distribution of Test MAPE

#### 6.1.1 PCReg Significantly Outperforms OLS When Coefficients Are Unreasonable

When OLS produces unreasonable coefficients, PCReg wins **78.7%** of scenarios on Test MAPE.

Table 13: Performance comparison when OLS produces unreasonable coefficients

Metric	OLS Mean	PCReg Mean	OLS Median	PCReg Median
Test MAPE	0.2862	0.1456	0.2094	0.1179
T1 APE	66.5658	0.5108	0.5694	0.3418
LC Abs Error	0.1182	0.0483	0.0943	0.0489

Metric	OLS Mean	PCReg Mean	OLS Median	PCReg Median
RC Abs Error	0.3548	0.1219	0.2347	0.1

### 6.1.2 PCReg Performs Comparably When OLS Coefficients Are Reasonable

When OLS produces reasonable coefficients (n=6,745), OLS “wins” on Test MAPE in **66.3%** of scenarios. However, the performance differences are negligible in practical terms.

Table 14: Performance comparison when OLS produces reasonable coefficients

Metric	OLS Mean	PCReg Mean	OLS Median	PCReg Median
Test MAPE	0.0549	0.061	0.0253	0.0298
T1 APE	0.1628	0.1562	0.0623	0.0709
LC Abs Error	0.0156	0.0154	0.007	0.007
RC Abs Error	0.0304	0.0353	0.0153	0.0172

## 6.2 OLS-LearnOnly Performs Poorly Outside Training Range

OLS-LearnOnly, which ignores the rate effect, exhibits poor extrapolation:

Table 15: Mean Test MAPE by model

Model	Mean Test MAPE
OLS-LearnOnly	15.9%
OLS	9.9%
PCReg	7.7%

OLS-LearnOnly has worse Test MAPE than full OLS in **63.9%** of scenarios. While simplifying the model may seem appealing, omitting the rate effect leads to systematic prediction errors outside the training range.

PCReg outperforms OLS-LearnOnly on Test MAPE in **67.9%** of all scenarios:

- When OLS coefficients are **reasonable**: PCReg wins **70.9%**
- When OLS coefficients are **unreasonable**: PCReg wins **52.8%**

## 6.3 Coefficient Bias Analysis

Constraints introduce bias in coefficient estimates. We analyze whether this bias is systematic and how it affects prediction below. A Mean/Median near 0 indicates unbiased estimation; lower Std indicates more stable estimates.

Table 16: Coefficient bias statistics

Coefficient	OLS	OLS	OLS Std	PCReg	PCReg	PCReg Std
	Mean	Median		Mean	Median	
$T_1$	1108.37	-0.2233	69135.6	-3.5793	-2.7313	46.1247
$b$	-0.0002	-0.0002	0.1059	-0.0078	-0.0017	0.0536
$c$	-0.0023	0.0003	0.292	0.0348	0.0111	0.1174

Figure 5 shows the distribution of coefficient errors by model (1st-99th percentile). The vertical line at 0 indicates no bias. PCReg shows tighter distributions despite some bias, leading to lower variance in predictions.

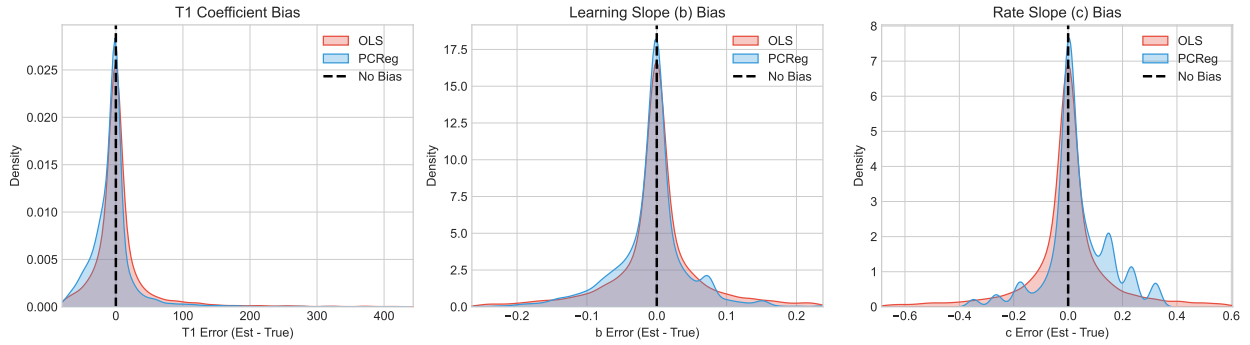


Figure 5: Distribution of coefficient errors by model

While OLS is theoretically unbiased with mean errors near 0, it has high variance and wide distribution. PCReg may have slight bias but much lower variance, leading to better overall prediction accuracy as explained by the classic bias-variance tradeoff.

#### 6.4 Parameter Effects on Model Performance

Beyond coefficient reasonableness, several design factors systematically influence when PCReg outperforms OLS. We examine sample size, predictor correlation, and noise level.

Table 17: Win Rate by Training Size

Sample Size	N	PCReg Win	% OLS	OLS Mean	PCReg Mean
		Rate	Unreasonable	MAPE	MAPE
5	2700	56.9%	32.0%	16.2%	10.7%
10	2700	46.5%	16.8%	7.4%	6.4%
30	2700	20.2%	1.4%	3.7%	4.8%

**Sample Size Effect:** As sample size, or training size, decreases, OLS becomes less stable. With  $n=5$  lots, PCReg wins **56.9%** of scenarios compared to **20.2%** at  $n=30$ . The rate of unreasonable OLS coefficients also increases from **1.4%** to **32.0%** as sample size decreases.

Table 18: Win Rate by Correlation

Correlation	N	PCReg Win Rate	% OLS Unreasonable	OLS Mean MAPE	PCReg Mean MAPE
<0.80	3187	31.8%	8.5%	6.7%	6.5%
0.80-0.90	873	52.9%	22.1%	10.3%	7.7%
0.90-0.95	2204	38.9%	15.2%	9.8%	7.6%
>0.95	1416	56.5%	35.0%	14.5%	8.7%

**Correlation Effect:** Higher predictor correlation destabilizes OLS. At correlation >0.95, PCReg wins **56.5%** of scenarios versus **31.8%** at lower correlations. This is the multicollinearity effect. When predictors are highly correlated, OLS coefficient estimates become unstable and constraints provide crucial stability.

Table 19: Win Rate by CV (Noise)

CoV Error	N	PCReg Win Rate	% OLS Unreasonable	OLS Mean MAPE	PCReg Mean MAPE
0.01	2700	36.7%	0.1%	0.9%	1.0%
0.1	2700	43.3%	16.2%	9.4%	8.1%
0.2	2700	43.7%	33.9%	19.6%	14.1%

**Noise Level Effect:** Higher CoV error increases the advantage of PCReg. With more noise, OLS has greater difficulty separating learning and rate effects, leading to more unreasonable coefficient estimates.

**Compounding Effects:** These factors interact—small samples with high correlation and high noise represent the most challenging scenarios for OLS, where PCReg provides the greatest benefit.

## 7 Practical Guidance

Based on our simulation results, we recommend the following decision rules for practitioners:

1. **If OLS produces unreasonable coefficients** (LCS or Rate Effect outside 70-100%): **Use PCReg**, it significantly outperforms OLS in these scenarios winning **79%%** of the time
2. **For small samples ( $n = 5$  or  $10$  lots) with noisy data ( $\text{CoV} = .20$ ): Prefer PCReg**, it wins **57%-67%** of scenarios
3. **For large samples ( $n = 30$  lots): OLS is generally preferred**, it wins **~78%-84%** of scenarios
4. **For intermediate cases:** Either method is acceptable; PCReg provides insurance against unreasonable coefficients with minimal downside

**Bottom line:** The difference between OLS and PCReg is typically small and can be controlled by explicitly defining loose constraints and arbitrarily small penalties.



## 8 Limitations

Several limitations of this study should be acknowledged:

**Loss Function Differences.** OLS and PCReg optimize different objective functions: OLS minimizes SSE in log-space, while PCReg minimizes SSPE in unit-space. Both are evaluated on unit-space metrics (MAPE). Ultimately we chose to measure against unit-space metrics, which directly determine prediction accuracy on future programs. OLS’s reliance on log-transforms means it estimates the conditional *median* in unit space, not the mean. This is a known limitation of log-linear regression that PCReg addresses by optimizing directly in unit space.

**Penalty Selection.** Our simulation tested a limited grid of penalty values ( $\lambda$  and  $\alpha$ ). Optimal penalty selection remains an active area of research, and the current grid may not capture the best-performing penalty configurations in all scenarios. Additionally, there is not a universally accepted method for selecting penalties in constrained regression, and GCV is one of several possible approaches. Future work could explore alternative penalty selection methods and a wider range of penalty values. A common heuristic is to choose the smallest penalty term that stabilizes coefficients without significantly increasing training error, which can be implemented with a lasso and ridge trace plot.

**Simulation-Based Evaluation.** All results are based on Monte Carlo simulation with known data-generating parameters. While the simulation uses realistic lot structures drawn from SAR data, no real-world cost datasets were used for validation. Our simulation evaluates models on their ability to recover the true underlying cost function, using noiseless test data at quantities beyond the training range. This design reflects the real-world cost estimation problem of predicting future production lots, where the analyst’s goal is to estimate the true population relationship. However, in the real world, population parameters are unknown and the size of data is limited, which affects the ability to create test and training splits. Future work will apply PCReg to historical SAR data to evaluate the method’s performance on real cost estimation problems.

## 9 Summary

Our simulation study demonstrates that PCReg provides meaningful advantages in small-sample, high-noise scenarios where OLS is most likely to produce unreasonable coefficients, while large samples favor standard OLS. The constraints introduce beneficial bias that reduces prediction variance. This is a classic bias-variance tradeoff that works in the analyst’s favor when data are limited. GCV enables reliable hyperparameter selection with as few as 5 observations. Furthermore, OLS-LearnOnly’s poor extrapolation performance reinforces that omitting the rate effect oversimplifies the problem. The guidance is clear: match the method to the sample size and data quality, with PCReg serving as effective insurance when uncertainty is high.

## References

- Farebrother, R. W. 1976. “Further Results on the Mean Square Error of Ridge Regression.” *Journal of the Royal Statistical Society: Series B* 38 (2): 248–50.
- Gaines, Brian R., Jeongyoun Kim, and Hua Zhou. 2018. “Algorithms for Fitting the Constrained Lasso.” *Journal of Computational and Graphical Statistics* 27 (4): 861–71. <https://doi.org/10.1080/10618600.2018.1473777>.
- Goeman, Jelle J. 2025. *Penalized: L1 (Lasso and Fused Lasso) and L2 (Ridge) Penalized Regression Models*. <https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf>.
- Golub, Gene H., Michael Heath, and Grace Wahba. 1979. “Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter.” *Technometrics* 21 (2): 215–23. <https://doi.org/10.1080/00401706.1979.10489751>.
- Hu, Shu-Ping. 2010. “Generalized Degrees of Freedom for Constrained CERs.” PRT-191. Tecolote Research.
- James, Gareth M., Courtney Paulson, and Paat Rusmevichientong. 2020. “Penalized and Constrained Optimization: An Application to High-Dimensional Website Advertising.” *Journal of the American Statistical Association* 115 (529): 107–22. <https://doi.org/10.1080/01621459.2019.1609970>.
- Stone, Mervyn. 1974. “Cross-Validatory Choice and Assessment of Statistical Predictions.” *Journal of the Royal Statistical Society: Series B (Methodological)* 36 (2): 111–47.
- Theobald, C. M. 1974. “Generalizations of Mean Square Error Applied to Ridge Regression.” *Journal of the Royal Statistical Society: Series B* 36 (1): 103–6.

## Acronyms

Acronym	Definition
CER	Cost Estimating Relationship
CoV	Coefficient of Variation
CV	Cross Validation
DoD	Department of Defense
EMD	Engineering and Manufacturing Development
FRP	Full Rate Production
GCV	Generalized Cross Validation
GDF	Generalized Degrees of Freedom
GLM	Generalized Linear Model
LCS	Learning Curve Slope
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
OLS	Ordinary Least Squares
PAC	Penalized and Constrained
PCReg	Penalized-Constrained Regression
RMSE	Root Mean Squared Error
RSS	Residual Sum of Squares
SAR	Selected Acquisition Report
SEE	Standard Error of Estimate
SLSQP	Sequential Least Squares Programming
SPE	Standard Percentage Error
SSE	Sum of Squared Errors
SSPE	Sum of Squared Percentage Errors
ZMPE	Minimum-Percentage Error under the Zero-Percentage Bias

## A PCReg Mathematical Notation

### A.1 Complete Notation Table

Table 21: Complete mathematical notation for PCReg framework

Symbol	Definition	Example
$X$	Feature matrix ( $n \times p$ )	Lot midpoint, quantity, complexity
$X_i$	i-th row of X	Predictor values for observation i
$y_i$	Observed response	Actual cost for observation i
$\theta$	Parameter vector ( $p \times 1$ )	Coefficients being estimated
$\hat{y}_i = f(X_i, \theta)$	Predicted value	Model prediction
$\lambda \geq 0$	Regularization strength	Overall penalty weight
$\alpha \in [0, 1]$	L1/L2 mixing parameter	0=Ridge, 1=Lasso
$L(y_i, \hat{y}_i)$	Loss function	SSE, SSPE, MSE, etc.

### A.2 Prediction Function Forms

#### Linear Model (Default):

$$\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_p x_{ip} = X_i \theta + \theta_0$$

For linear models,  $\theta = [\beta_1, \dots, \beta_p]$  represents regression coefficients (slopes), and  $\theta_0$  is the intercept.

#### Nonlinear Model (Learning Curve Example):

$$\hat{y}_i = T_1 \cdot (\text{LotMidpoint}_i)^b \cdot (\text{LotQuantity}_i)^c$$

Here  $\theta = [T_1, b, c]$  with domain-specific meanings:

- $T_1$  = Base cost parameter
- $b$  = LCS exponent
- $c$  = Rate effect exponent

### A.3 Cost Estimating Terminology Mapping

For the cost estimating community, PCReg notation maps to familiar CER concepts:

Table 22: Mapping between mathematical and cost estimating terminology

Math Notation	CER Terminology	Example
$X$	Cost drivers	Lot size, quantity, weight
$\theta$	CER parameters	Slopes, intercepts, exponents
$y$	Historical costs	Actual program costs
$\hat{y}$	Cost estimates	CER predictions
Bounds on $\theta$	Domain constraints	LCS: 70-100%, b: -0.5 to 0
$\lambda$	Regularization	Shrinkage for small datasets

PCReg addresses the typical CER challenge: small samples (5-30 points), correlated predictors, and enforcing domain knowledge bounds on parameters.

## B Other Models to Consider for Future Methods

During our simulation study, we evaluated several alternative modeling approaches beyond the primary comparison between PCReg, OLS, and OLS-LearnOnly presented in the main paper. These preliminary results suggest promising directions for future methodological development. However, we require additional time to complete a thorough analysis before making definitive recommendations.

### B.1 Model Descriptions

#### PCReg Variants:

- **PCReg**: Our primary model using GCV penalty selection with SSPE loss and domain constraints (default configuration)
- **PCReg\_Tight**: PCReg with tighter coefficient bounds (narrower feasible region for domain constraints)
- **PCReg\_ConstrainOnly**: Constrained optimization without regularization penalty ( $\lambda = 0$ , constraints only)
- **PCRegGCV\_LogMSE**: PCReg using log-transformed MSE loss instead of SSPE
- **PCReg\_CV**: PCReg using K-Fold cross validation for penalty selection instead of GCV
- **PCReg\_AICc**: PCReg using corrected Akaike Information Criterion (AICc) for penalty selection

#### Baseline Methods:

- **OLS**: Standard ordinary least squares regression (log-log linear model)
- **OLS\_LearnOnly**: OLS using only the learning curve variable (excludes quantity/rate effect)
- **RidgeCV**: Ridge regression (L2 penalty only) with cross validation, no domain constraints
- **LassoCV**: Lasso regression (L1 penalty only) with cross validation, no domain constraints
- **BayesianRidge**: Bayesian Ridge regression with automatic relevance determination

### B.2 Preliminary Results

Table 23: Preliminary model performance across all 8,100 scenarios. Lower Test MAPE is better.

Model	Test MAPE	Test SSPE	b Error	c Error	R2
PCReg_Tight	0.0561	0.0861	0.0200	0.0337	0.852
PCReg_ConstrainOnly	0.0764	0.2052	0.0354	0.0791	0.877
PCReg	0.0773	0.2074	0.0338	0.0827	0.871
PCRegGCV_LogMSE	0.0778	0.2832	0.0341	0.0776	0.877
PCReg_CV	0.0794	0.2538	0.0353	0.0814	0.862
BayesianRidge	0.0801	0.4595	0.0360	0.0886	0.882
PCReg_AICc	0.0808	0.2171	0.0349	0.0913	0.863
RidgeCV	0.0952	0.9362	0.0481	0.1216	0.896
LassoCV	0.0957	0.9767	0.0428	0.1082	0.858
OLS	0.0994	0.9727	0.0520	0.1319	0.897
OLS_LearnOnly	0.1592	0.9543	0.0827	0.2361	0.789

### B.3 Future Work

These results are preliminary and require further investigation to understand:

- The performance advantages of tighter constraint bounds (PCReg\_Tight)
- Trade-offs between different penalty selection criteria (GCV vs. CV vs. AICc)
- The role of loss function choice (SSPE vs. log-transformed MSE)
- Whether constraints alone (without regularization) provide sufficient regularization
- How standard penalized regression methods (Ridge, Lasso, Bayesian Ridge) compare when adapted to nonlinear cost models

Additionally, we plan to pursue the following:

- **Real-world validation:** Fit PCReg on actual SAR data to establish empirical benchmarks and validate the simulation findings against real cost programs
- **Penalty selection research:** Conduct a more comprehensive study of penalty parameter selection, exploring finer grids and adaptive methods for identifying optimal regularization strength
- **Trace plots:** Develop coefficient trace plots that visualize how estimates change across penalty values, enabling practitioners to select the optimal penalty for their individual case and better understand the bias-variance tradeoff in their specific context

We plan to conduct additional analysis to provide comprehensive guidance on selecting among these alternatives for cost estimation applications.

## C Software Documentation

### C.1 Installation

**Note:** The package is currently available on [GitHub](#) and will be released to PyPI in the future. Please check back for pip installation instructions.

#### C.1.1 Installing from GitHub

You can install the package directly from GitHub using pip:

```
pip install git+https://github.com/frankij11/Penalized-Constrained-Regression.git
```

#### C.1.2 Cloning the Repository

Alternatively, you can clone the repository and install locally:

```
git clone https://github.com/frankij11/Penalized-Constrained-Regression.git
cd Penalized-Constrained-Regression
pip install -e .
```

#### C.1.3 Downloading the Excel Template

For users preferring Excel, download the [PCReg Template](#) directly, or visit the repository to access all files.

### C.2 Basic Usage

```
import numpy as np
import penalized_constrained as pcreg

# Generate sample data
np.random.seed(42)
X = np.random.randn(100, 2)
y = X @ np.array([-0.15, -0.07]) + 4.5 + 0.1 * np.random.randn(100)

# Fit with cross-validated hyperparameter selection
model = pcreg.PenalizedConstrainedCV(
    bounds=[(-1, 0), (-1, 0)], # Both coefficients must be 0
    loss='sspe',               # Sum of Squared Percentage Error
    cv=5
)
model.fit(X, y)

# Results
print(f"Best alpha: {model.alpha_:.4f}")
print(f"Best l1_ratio: {model.l1_ratio:.2f}")
print(f"Coefficients: {model.coef_}")
print(f"Active constraints: {model.active_constraints_}")
```



### C.3 Citation

```
@inproceedings{joy2026pcreg,  
  title={Penalized-Constrained Regression: Combining Regularization  
    and Domain Constraints for Cost Estimation},  
  author={Joy, Kevin and Watstein, Max},  
  booktitle={ICEAA Professional Development \& Training Workshop},  
  year={2026}  
}
```

## D Detailed Diagnostic Report

This appendix provides comprehensive diagnostic information for the PCReg model fit from the motivating example.

### D.1 Model Specification

Parameter	Value
Model Type	PenalizedConstrainedCV
Loss Function	sspe
Alpha ( $\lambda$ )	1.666667
L1 Ratio ( $\alpha$ )	0.0000
Fit Intercept	False
Optimization Method	SLSQP
Convergence Status	Yes
GDF Method	gaines

### D.2 Coefficient Estimates

Parameter	Estimate	Bootstrap SE	95% CI Lower	95% CI Upper	Lower Bound	Upper Bound	Status
T1	99.9976	7.84584	82.6878	120.569	0	None	free
b	-0.052357	0.028546	-0.13784	-0.028375	-0.5	0.0	free
c	-0.180243	0.032862	-0.207598	-0.093203	-0.5	0.0	free

### D.3 Fit Statistics

Statistic	Value
R <sup>2</sup>	0.900604
Adjusted R <sup>2</sup>	0.872206
SEE	3.44416
SPE	0.103307
MAPE	0.0693
RMSE	2.88159
CV	0.087908
GDF	7

### D.4 Constraint Summary

Number of specified bounds: 5

Number of active constraints: 0

No constraints are active (all parameters are in the interior).

## D.5 Bootstrap Results

Number of bootstrap samples: 500

Confidence level: 95%

Successful constrained fits: 500

Successful unconstrained fits: 500

### D.5.1 Bootstrap Summary Statistics

Parameter	Constrained Mean	Constrained Std	Constrained CI Lower	Constrained CI Upper	Unconstrained Mean	Unconstrained Std	Unconstrained CI Lower	Unconstrained CI Upper
T1	99.477643	7.845841	82.687838	120.569444	109.736319	15.398629	99.996917	143.553038
b	-0.064994	0.028546	-0.137840	-0.028375	-0.005025	0.070858	-0.178079	0.111875
c	-0.163312	0.032862	-0.207598	-0.093203	-0.250956	0.091588	-0.398782	-0.054577
Intercept	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

## D.6 Data Summary

Property	Value
Number of samples	10
Number of features	2
Target mean	32.7798
Target std	9.1401
Target min	21.4632
Target max	44.8221

## D.7 Model Equation

Custom functional form:

```
def prediction_fn(X, params):  
    T1, b, c = params  
    return T1 * (X[:, 0] ** b) * (X[:, 1] ** c)
```

Model Equation: Custom prediction function Parsed:  $T1 * X[:, 0] ** b * X[:, 1] ** c$