A Methodological Framework for Two-Stage Residual Analysis

Team Sharks Project

November 5, 2025

Abstract

This document details the statistical theory behind the two-stage estimation procedure proposed for our project. It begins with the simple OLS (Ordinary Least Squares) case, explains why the mean-zero property of OLS residuals is not a requirement for our bias analysis, and then expands the framework to include more flexible, non-linear machine learning models. Finally, it introduces a "double-residual" method as a more principled and robust alternative for the second stage.

1 The Core Two-Stage Framework

Our project's goal is to determine if a set of contextual variables, $\mathbf{Z}$, can explain part of a player's salary, Y, *after* we have already accounted for the player's on-court performance, $\mathbf{X}$.

Let Y_i be the (log) salary for player i, $\mathbf{x}_i$ be the vector of p performance statistics for that player, and $\mathbf{z}_i$ be the vector of m contextual (bias) factors.

Our analysis consists of two distinct stages:

1. **Stage 1: The Performance Model.** We model salary as a function of performance. This creates a "performance-expected" salary, $\hat{Y}_i$.

$$Y_i = f(\mathbf{x}_i) + \epsilon_i$$

   We then compute the residual, $\hat{\epsilon}_i$, which represents the portion of salary *unexplained* by performance.

$$\hat{\epsilon}_i = Y_i - \hat{f}(\mathbf{x}_i)$$

2. **Stage 2: The Bias Model.** We model this residual (our "mispricing" metric) as a function of the contextual factors.

$$\hat{\epsilon}_i = g(\mathbf{z}_i) + \nu_i$$

If we find that $\mathbf{Z}$ has a statistically significant ability to explain $\hat{\epsilon}_i$, we have found evidence of systematic bias.

2 The "Original Idea": OLS in Both Stages

The simplest implementation of this framework uses Ordinary Least Squares (OLS) for both stages.

2.1 Stage 1: OLS Performance Model

We assume the function f is linear. We fit the model (within a specific role k):

$$Y_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

This gives us the fitted values $\hat{Y}_i = \hat{\beta}_0 + \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, and the OLS residuals:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

2.2 Stage 2: OLS Bias Model

We then regress these residuals onto the contextual factors $\mathbf{Z}$, again using a linear model:

$$\hat{\epsilon}_i = \gamma_0 + \mathbf{z}_i^T \boldsymbol{\gamma} + \nu_i$$

Our object of interest is the vector of coefficients $\hat{\boldsymbol{\gamma}}$. If any coefficient $\hat{\gamma}_j$ is statistically significant, it implies that the j-th contextual factor (e.g., 'Draft Status') has an explanatory relationship with salary, even after controlling for performance.

3 The "Mean-Zero" Property: A Clarification

It is correct to note that OLS has properties related to a zero mean. However, it's crucial to distinguish between a **model assumption** and a **mechanical property** of the fitting procedure.

- **Model Assumption:** The OLS model *assumes* that the *true* error term ϵ_i has a conditional mean of zero: $E[\epsilon_i | \mathbf{x}_i] = 0$. This is a statement about the world.
- **Mechanical Property:** When you fit an OLS model *with an intercept* (like $\hat{\beta}_0$), the resulting set of *residuals* $\hat{\epsilon}_i$ is mathematically guaranteed to have a sample mean of exactly zero.

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

Does this mechanical property matter for our Stage 2 regression?
Answer: No, it is not necessary and does not affect our results.
Here's why: Our Stage 2 model, $\hat{\epsilon}_i = \gamma_0 + \mathbf{z}_i^T \boldsymbol{\gamma} + \nu_i$, *also has an intercept* (γ_0).

- **Case 1 (OLS Residuals):** If we use OLS residuals, we know mean$(\hat{\epsilon}) = 0$. When we fit the Stage 2 model, the intercept $\hat{\gamma}_0$ will be fitted around zero (after accounting for $\mathbf{Z}$).
- **Case 2 (Other Model):** Let's say we use a different model (like a Random Forest) and our residuals $\hat{\epsilon}_{RF}$ have a mean of 0.5. When we fit the Stage 2 model, the OLS procedure will simply adjust the intercept $\hat{\gamma}_0$ to account for this. The intercept will be fitted around 0.5.

In both cases, the intercept γ_0 absorbs the mean of the outcome variable ($\hat{\epsilon}$). The values of our *actual interest*, the slope coefficients $\boldsymbol{\gamma}$, are calculated based on the *variance* and *covariance* of the variables, not their means. Therefore, the mean-zero property is a non-issue.

4 Beyond OLS: Using Better Models for Stage 1

The definition of a residual is generic: $\hat{\epsilon}_i = Y_i - \hat{f}(\mathbf{x}_i)$. The function f does not have to be linear. In fact, using a more powerful, non-linear model in Stage 1 is a **major improvement** to our methodology.

4.1 The "Polluted Residual" Problem

If we use a simple OLS model, but the true relationship between performance $\mathbf{X}$ and salary Y is complex and non-linear (e.g., diminishing returns for "points per game"), our OLS model $\hat{f}_{OLS}$ will be a poor fit.

The resulting OLS residual will be "polluted":

$$\hat{\epsilon}_{OLS} = (Y - \hat{f}_{OLS}) = \underbrace{\text{(True Unexplained Salary)}}_{\text{What we want}} + \underbrace{\text{(Non-linearity OLS missed)}}_{\text{Pollution}}$$

When we run our Stage 2 regression, we are regressing on this polluted signal, and our bias estimates $\hat{\boldsymbol{\gamma}}$ will themselves be biased and unreliable.

4.2 The "Clean Residual" Solution

If we use a more flexible model like a **Random Forest** or **Gradient Boosting (XGBoost)** for Stage 1, it is much better at capturing complex non-linearities and interactions.

$$\hat{f}_{RF} \approx f(\mathbf{x}_i)$$

Because $\hat{f}_{RF}$ is a much more accurate estimate of the "expected salary," its residual is a much "cleaner" measure of the true unexplained salary.

$$\hat{\epsilon}_{RF} = (Y - \hat{f}_{RF}) \approx \text{(True Unexplained Salary)}$$

Using this clean residual as the outcome variable in Stage 2 will give us a much more accurate and trustworthy estimate of the bias coefficients $\boldsymbol{\gamma}$.

5 A More Principled Method: Double-Residual Regression (DML)

The two-stage "clean residual" approach from Section 4 is a major improvement. However, a more advanced econometric method, known as **Double-Residual Regression** or **Double/Debiased Machine Learning (DML)**, is the modern standard for this problem. It is an extension of the Frisch-Waugh-Lovell theorem to non-linear and machine learning models.

This method "cleans" the confounding influence of $\mathbf{X}$ from *both* the outcome Y and the variable-of-interest Z before the final regression.

Let's say we want to find the unbiased effect of a single contextual factor z_j (e.g., 'Draft Status') on salary Y, after controlling for the high-dimensional vector of performance stats $\mathbf{x}$.

The DML procedure is as follows:

1. **Model 1 (Outcome Residuals):** Model salary Y as a function of performance $\mathbf{X}$ using a flexible machine learning model (e.g., XGBoost, Random Forest).

$$Y_i = f(\mathbf{x}_i) + \epsilon_{Y,i}$$

We then compute the "Y-residuals," which represent the portion of salary unexplained by performance:

$$\hat{\epsilon}_{Y,i} = Y_i - \hat{f}(\mathbf{x}_i)$$

2. **Model 2 (Treatment Residuals):** Model the contextual factor z_{ij} as a function of performance $\mathbf{X}$ using a flexible model (this can be a different model type).

   $$z_{ij} = h_j(\mathbf{x}_i) + \epsilon_{Z,ij}$$

   We then compute the "Z-residuals," which represent the portion of 'Draft Status' (or 'Age', etc.) that is *uncorrelated* with on-court performance:

   $$\hat{\epsilon}_{Z,ij} = z_{ij} - \hat{h}_j(\mathbf{x}_i)$$

   We must repeat this step for *each* contextual factor j that we wish to test.

3. **Final Stage (The Debiased Regression):** Run a simple OLS regression using the residuals from the first two steps.

   $$\hat{\epsilon}_{Y,i} = \gamma_0 + \gamma_j \hat{\epsilon}_{Z,ij} + \nu_i$$

The resulting coefficient $\hat{\gamma}_j$ is a "debiased" estimate of the effect of z_j on Y. It represents the relationship between the part of salary unexplained by performance and the part of 'Draft Status' unexplained by performance.

5.1 Interpreting the Final OLS Coefficient

The coefficient $\hat{\gamma}_j$ from the final stage requires careful interpretation. It is **not** simply "the effect of draft status on salary" – it is something more specific and more meaningful.

What $\hat{\gamma}_j$ Measures

The coefficient $\hat{\gamma}_j$ quantifies the relationship between salary and a contextual factor *after removing the portion of both that can be explained by performance.* In practical terms:

> "For every unit increase in draft status that is independent of performance, salary increases by $\hat{\gamma}_j$ dollars, where this salary difference is also independent of performance."

A Concrete Example

Suppose we analyze draft status and obtain $\hat{\gamma}_{\text{draft}} = 2.5$ (in millions of dollars). Consider two players with *identical* performance statistics:

- **Player A:** First-round pick (Draft Status = 1), earns $15M

- **Player B:** Undrafted (Draft Status = 0), earns $10M

The coefficient $\hat{\gamma}_{\text{draft}} = 2.5$ captures this $2.5M premium that Player A receives, which cannot be attributed to superior performance (since their stats are identical). This is the "draft status premium" – evidence of systematic bias in salary determination.

Why the Double-Residual Approach Matters

One might ask: why not simply regress the Stage 1 residuals $\hat{\epsilon}_{Y,i}$ on the *original* contextual factor z_{ij}? The answer lies in a subtle but critical confounding issue.

If we regress $\hat{\epsilon}_{Y,i}$ directly on z_{ij}, we are asking: "Does draft status correlate with overpayment?" However, this relationship is confounded. High draft picks may be systematically overpaid not because of bias, but because they possess intangible qualities (leadership, work ethic, basketball IQ) that:

1. Are difficult to capture in our performance metrics $\mathbf{X}$,
2. Legitimately justify higher salaries, and
3. Are correlated with draft status (scouts consider these when drafting).

By *also* residualizing z_{ij} in Step 2, we ensure that $\hat{\epsilon}_{Z,ij}$ represents the component of draft status that is truly orthogonal to performance. The final regression then isolates the "pure" draft status effect – the premium paid for draft pedigree alone, stripped of any performance-related justification.

Why OLS in the Final Stage?

After the complex machine learning procedures in Steps 1 and 2 have "cleaned" both variables, the final stage uses simple OLS for several reasons:

- **Sufficiency:** The residuals $\hat{\epsilon}_{Y,i}$ and $\hat{\epsilon}_{Z,ij}$ are already purged of confounding. We now simply need to measure their linear relationship.
- **Interpretability:** OLS provides clean, interpretable coefficients with standard errors and p-values.
- **Theoretical Foundation:** The statistical theory guaranteeing that $\hat{\gamma}_j$ is "debiased" (asymptotically unbiased and normally distributed) relies on using OLS in this final stage.
- **Computational Simplicity:** After the computational expense of training ML models, OLS is trivial and instantaneous.

In summary, the final OLS regression is not an arbitrary choice – it is the theoretically justified and practically optimal method for extracting the causal parameter of interest from the purged residuals.

6 A Critical Implementation Detail: Cross-Fitting

To properly prevent the overfitting bias we just discussed, we cannot use the same data to *train* the models ($\hat{f}$ and $\hat{h}$) and to *generate the residuals* for the final OLS. Doing so would "bake in" the overfitting.

The standard solution is **cross-fitting** (or **sample-splitting**). Here is the practical algorithm for the project:

1. **Split Data:** Randomly partition your dataset of n players into K "folds" (e.g., $K = 5$).
2. **Iterate through Folds:** For each fold k from 1 to K:

- Let the "training" data be all folds *except* fold k.
- Let the "prediction" data be *only* fold k.
- **Train Models:** On the "training" data, fit your machine learning models $\hat{f}_k$ and $\hat{h}_{jk}$ (one h for each z_j).

$$\hat{f}_k \leftarrow \text{Model}(Y_{\text{train}} \sim \mathbf{X}_{\text{train}})$$
$$\hat{h}_{jk} \leftarrow \text{Model}(Z_{j,\text{train}} \sim \mathbf{X}_{\text{train}})$$

- **Generate Residuals:** On the "prediction" data (fold k), use the models you just trained to predict $\hat{Y}$ and $\hat{Z}_j$. Calculate and *store* the residuals for this fold.

$$\hat{\epsilon}_{Y,k} = Y_{\text{predict}} - \hat{f}_k(\mathbf{X}_{\text{predict}})$$

$$\hat{\epsilon}_{Z,jk} = Z_{j,\text{predict}} - \hat{h}_{jk}(\mathbf{X}_{\text{predict}})$$

3. **Concatenate Residuals:** After the loop finishes, you will have K sets of residuals. Stack them together to form two complete vectors, $\hat{\boldsymbol{\epsilon}}_Y$ and $\hat{\boldsymbol{\epsilon}}_{Z,j}$, both of length n.

4. **Run Final OLS:** Run final, simple OLS regression on these "out-of-sample" residuals.

$$\hat{\boldsymbol{\epsilon}}_Y = \gamma_0 + \gamma_j \hat{\boldsymbol{\epsilon}}_{Z,j} + \boldsymbol{\nu}$$

6.1 The Final Model

Run Step 4 in one of two ways:

- **One-by-One:** Run a separate simple regression for each contextual factor j, as shown above. This gives you the isolated, debiased effect $\hat{\gamma}_j$ for each factor.

- **All-at-Once:** Compute the residuals $\hat{\boldsymbol{\epsilon}}_{Z,j}$ for *all* your contextual factors (Age, Draft Status, etc.). Then, run a single *multiple* regression:

$$\hat{\boldsymbol{\epsilon}}_Y = \gamma_0 + \gamma_1 \hat{\boldsymbol{\epsilon}}_{Z,1} + \gamma_2 \hat{\boldsymbol{\epsilon}}_{Z,2} + \cdots + \gamma_m \hat{\boldsymbol{\epsilon}}_{Z,m} + \boldsymbol{\nu}$$

This "all-at-once" model is likely what we want. The resulting coefficients $\hat{\boldsymbol{\gamma}}$ and their p-values are the final output of our analysis, giving us the debiased effect of each contextual factor while controlling for all performance metrics *and* all other contextual factors