Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

February 5, 2026

Publisher: Online
Version 0.1.2

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

Predictive Inference with Linear Regression in Moderately High Dimensions

1

"Infer: to form an opinion or guess that something is true because of the information that you have."

– Cambridge Dictionary [1].

Least squares—particularly in the form of linear regression—is among the most widely used and intuitive statistical methods, useful for both predictive inference and identifying associations. Introduced in the early 1800s by L. Legendre and C. F. Gauss, it has become a foundational tool in data analysis. Here, we review the properties of least squares estimation for linear models in moderately high-dimensional settings, focusing on its utility in prediction and association analysis. This discussion sets the stage for our subsequent review of modern statistical (machine) learning approaches, which relax dimensionality assumptions and incorporate nonlinear models.

1.1 Foundation of Linear Regression

Regression and the Best Linear Prediction Problem

We consider a scalar random variable Y, an outcome of interest, and a p-vector of covariates

$$X = (X_1, \ldots, X_p)'.$$

We assume that a constant of 1 is included as the first component in X; that is, $X_1 = 1$.

For theoretical purposes, we first consider linear regression in the population. Working in the population means that we have access to unlimited amounts of data to compute population moments – such as $\mathrm{E}[Y]$, $\mathrm{E}[YX]$, and $\mathrm{E}[XX']$ – and that we can define "ideal" quantities. After defining these ideal quantities, we then turn to estimation with real data, which we will take to be a sample of observations drawn from the population.

Our first goal is to construct the best linear prediction rule for Y using X. That is, the predicted value of Y given X will be of the linear form:

$$\sum_{j=1}^{p} \beta_j X_j = \beta' X, \text{ for } \beta = (\beta_1, \ldots, \beta_p)',$$

where β's are called the regression parameters or coefficients.

We define β as any solution to the *Best Linear Prediction (BLP) Problem*,

$$\min_{b \in \mathbb{R}^p} \mathrm{E}\left[(Y - b'X)^2\right],$$

where we minimize the Expected or Mean Squared Error (MSE) for predicting Y using the linear rule

$$b'X = \sum_{j=1}^{p} b_j X_j, \quad b = (b_1, \ldots, b_p)'.$$

The solution to this optimization problem, $\beta'X$, is called the *Best Linear Predictor* (BLP) of Y using X. This jargon refers to the fact that $\beta'X$ is the best, according to MSE, linear prediction rule for Y among all possible linear prediction rules.

We can compute β by solving the first order conditions for the BLP problem:

$$\mathrm{E}\left[(Y - \beta'X)X\right] = 0.$$

Figure 1.1: The only known portrait of Legendre (a friendly caricature) by Julien Léopold Boilly. Source: Wikipedia. The hairstyle is amazing.

Figure 1.2: AI-based reconstruction of a more realistic portrait of Legendere based on Boilly's caricature. Source: generated by authors using ChatGPT-4, who noted that "this image captures the intense and expressive nature that was suggested by the caricature, with a dignified and classical appearance."

These equations are also referred to as the Normal Equations and are obtained by setting the derivative of the objective function $b \mapsto \mathrm{E}\left[(Y - b'X)^2\right]$ with respect to b equal to zero. Thus, any solution to the BLP problems satisfies the Normal Equations.

Defining the regression error or residual as

$$\varepsilon := (Y - \beta'X),$$

we can write the Normal Equations as[1]

$$\mathrm{E}[\varepsilon X] = 0, \quad \text{or equivalently} \quad \varepsilon \perp X.$$

Therefore, the BLP problem provides a simple decomposition of Y:

$$Y = \beta'X + \varepsilon, \quad \varepsilon \perp X,$$

where $\beta'X$ is the part of Y that can be linearly predicted or explained with X, and ε is whatever remains – the so-called unexplained or residual part of Y.

1: Note that we use $\perp$ to denote orthogonality between random variables, and $\perp\!\!\!\perp$ to denote full statistical independence. That is, for random variables U and V, $U \perp V$ means $\mathrm{E}[UV] = 0$. Further, if U is a *centered random variable*, then $U \perp\!\!\!\perp V$ implies $U \perp V$, but the reverse implication is not true in general. Indeed, let $U \sim N(0,1)$ and $V = U^2 - 1$, then $U \perp V$ but $U \not\perp\!\!\!\perp V$.

Best Linear Approximation Property

The normal equation $\mathrm{E}[(Y - \beta'X)X] = 0$ implies by the law of iterated expectations that

$$\mathrm{E}[(\mathrm{E}[Y \mid X] - \beta'X)X] = 0.$$

Therefore, the BLP of Y is also the BLP for the conditional expectation of Y given X. This observation is important and motivates the use of various transformations of regressors to form X.

From Best Linear Predictor to Best Predictor

Here we explain the use of constructed features or regressors. If W are "raw" regressors/features, *technical (constructed) regressors* are of the form

$$X = T(W) = (T_1(W), ..., T_p(W))',$$

where the set of transformations $T(W)$ is sometimes called the *dictionary* of transformations. Example transformations include polynomials, interactions between variables, and applying functions such as the logarithm or exponential. In the wage analysis

reported below, for example, we use quadratic and cubic transformations of experience, as well as interactions (products) of these regressors with education and geographic indicators.

The main motivation for the use of constructed regressors is to build *more flexible and potentially better* prediction rules. The potential for improved prediction arises because we are using prediction rules $\beta' X = \beta' T(W)$ that are *nonlinear* in the original raw regressors W and may thus capture more complex patterns that exist in the data. Conveniently, the prediction rule $\beta' X$ is still linear with respect to the parameters, β, and with respect to the constructed regressors $X = T(W)$.

In the population, the *best predictor* of Y given W is

$$g(W) = \mathrm{E}[Y \mid W],$$

the *conditional expectation* of Y given W. The *conditional expectation function* $g(W)$ is also called the *regression function* of Y on W. Specifically, the conditional expectation function $g(W)$ solves the best prediction problem[2]

$$\min_{m(W)} \mathrm{E}\left[(Y - m(W))^2\right].$$

Here we minimize the MSE among all prediction rules $m(W)$ (linear or nonlinear in W).

As the conditional expectation solves the same problem as the best linear prediction rule among a larger class of candidate rules, the conditional expectation generally provides better predictions than the best linear prediction rule.[3]

By using $\beta' T(W)$, we are implicitly approximating the best predictor $g(W) = \mathrm{E}[Y|W]$. Indeed, for any parameter b,

$$\mathrm{E}\left[(Y - b'T(W))^2\right] = \mathrm{E}\left[(g(W) - b'T(W))^2\right] + \mathrm{E}\left[(Y - g(W))^2\right],$$

That is, the mean squared prediction error is equal to the mean squared approximation error of $b'T(W)$ to $g(W)$ plus a constant that does not depend on b. Therefore, minimizing the mean squared prediction error is the same as minimizing the mean squared approximation error. Thus, the BLP $\beta' T(W)$ is the *Best Linear Approximation* (BLA) to the best predictor, which is the regression function $g(W)$. Finally, as the dictionary of transformations $T(W)$ becomes richer, the quality of the approximation of the BLA $\beta' T(W)$ to the best predictor $g(W)$ improves.

2: This result follows by rewriting the objective function as

$$\min_{m(W)} \mathrm{E}[\mathrm{E}[(Y - m(W))^2 \mid W]],$$

noting that it is equivalent to

$$\mathrm{E}[\min_{\mu \in \mathbb{R}} \mathrm{E}[(Y - \mu)^2 \mid W]],$$

and deriving the first order conditions for the inner minimization: $E(Y \mid W) - \mu = 0$.

3: Unless the conditional expectation function turns out to be linear, in which case the conditional expectation and best linear prediction rule coincide.

Example 1.1.1 (Approximating a Smooth Function with a Polynomial Dictionary) Suppose $W \sim U(0,1)$ where U denotes the continuous uniform distribution, and

$$g(W) = \exp(4 \cdot W).$$

We use

$$T(W) = \underbrace{(1, W, W^2, \ldots, W^{p-1})'}_{p \text{ terms}}$$

to form the BLA/BLP, $\beta' T(W)$. Figure 1.3 provides a sequence of panels that illustrate the approximation properties of the BLA/BLP corresponding to p =2, 3, and 4:

- With $p = 2$ we get a linear in W approximation to $g(W)$. As the figure shows, the quality of this approximation is poor.
- With $p = 3$ we get a quadratic-in-W approximation to $g(W)$. Here, the approximation quality is markedly improved relative to $p = 2$ though approximation errors are still clearly visible.
- With $p = 4$ we get a cubic-in-W approximation to $g(W)$, and the quality of approximation appears to be excellent.

This simple example highlights the motivation for using nonlinear transformations of raw regressors in linear regression analysis. What this example does not yet reveal are the *statistical* challenges of dealing with higher and higher dimension p when learning from a finite sample.

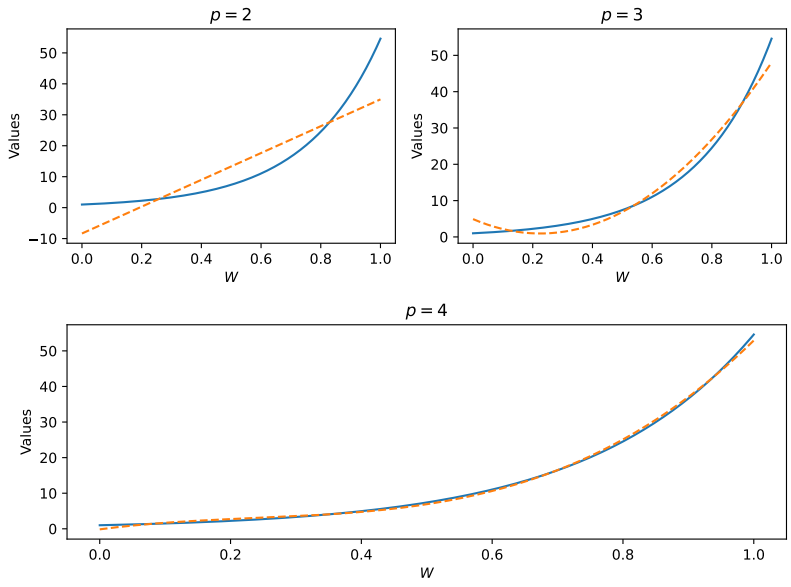

Figure 1.3: Refinements of Approximation to Regression Function $g(W)$ by using polynomials of W.

There are many ways of generating flexible approximations, which are studied by approximation theory and nonparametric statistical learning theory.[4]

4: See, e.g., Tsybakov [2]. We will also consider nonlinear approximations using trees and neural networks in Chapter 8.

When we have multiple variables, we may generate transformations of each of the variables and employ interactions – products involving these terms. As a simple concrete example, consider a case with two raw regressors, W_1 and W_2. We could build polynomials of second order in each of the raw regressors – $(1, W_1, W_1^2)$, $(1, W_2, W_2^2)$. We may then collect these variables along with the interaction in the raw regressors, $W_1 W_2$ in a vector

$$(1, W_1, W_2, W_1^2, W_2^2, W_1 W_2)$$

for use in a regression model. There are, of course, many other possibilities such as considering higher order polynomial terms, e.g. W_1^3; higher order interactions, e.g. $W_1^2 W_2$; and other nonlinear transformations, e.g. $\log(W_1)$.

1.2 Statistical Properties of Least Squares

The Best Linear Prediction Problem in Finite Samples

In practice, the researcher does not have access to the entire population, but observes only a sample

$$\{(Y_i, X_i)\}_{i=1}^n = ((Y_1, X_1), ..., (Y_n, X_n)).$$

We assume that this sample is a random sample from the distribution of (Y, X). Formally, this condition means that the observations were obtained as realizations of independently and identically distributed (iid) copies of the random variable (Y, X).[5]

5: By treating the observations as iid, we are modeling the data as independent random draws with replacement from a population. Other possible models include sampling without replacement from a finite population, stratified sampling, observations of a process over time, and other schemes or scenarios that induce dependence between the data points. For the most part, we focus on the iid model throughout this book, as it typically conveys the key aspects of the inferential problem sufficiently well.

We construct the best in-sample linear prediction rule for Y using X analogously to the population case by replacing theoretical expected values, E, with empirical averages, $\mathbb{E}_n$.[6] Specifically, given X, our predicted value of Y will be

$$\sum_{j=1}^p \hat{\beta}_j X_j = \hat{\beta}' X, \text{ for } \hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)',$$

where $\hat{\beta}$ is any solution to the *Best Linear Prediction Problem in the Sample*, also known as Ordinary Least Squares (OLS):[7]

6: $\mathbb{E}_n$ abbreviates the notation $\frac{1}{n}\sum_{i=1}^n$. For example,

$$\mathbb{E}_n[f(Y, X)] := \frac{1}{n}\sum_{i=1}^n f(Y_i, X_i).$$

7: The hat notation $\hat{}$ is commonly used to denote estimators—quantities derived from a sample. For instance, β represents the best linear predictor (BLP) in the population (the estimand), whereas $\hat{\beta}$ is the corresponding estimator computed from the sample.

$$\min_{b \in \mathbb{R}^p} \mathbb{E}_n[(Y - b'X)^2].$$

That is, $\hat{\beta}$ minimizes the sample MSE for predicting Y using the linear rule $b'X$. The $\hat{\beta}$'s are called the sample regression coefficients.

We can compute $\hat{\beta}$ as any solution to the Sample Normal Equations,

$$\mathbb{E}_n[X(Y - X'\hat{\beta})] = 0,$$

which are obtained as the first order conditions to the Best Linear Prediction Problem in the Sample. Further, defining the residuals (or, in-sample regression errors) as

$$\hat{\varepsilon}_i := (Y_i - \hat{\beta}'X_i),$$

we obtain the decomposition

$$Y_i = X_i'\hat{\beta} + \hat{\varepsilon}_i, \quad \mathbb{E}_n[X\hat{\varepsilon}] = 0,$$

where $X_i'\hat{\beta}$ is the predicted or explained part of Y_i, and $\hat{\varepsilon}_i$ is the unexplained or residual part.

Properties of Sample Linear Regression

The best linear prediction rule in the population is $\beta'X$, and a key question is whether $\hat{\beta}'X$ estimates (that is, approximates using data) $\beta'X$ well.

The best linear prediction rule is also the best linear rule for predicting future values of Y given a new draw X, when new (Y, X) are sampled from the same population. Therefore, if we can approximate the best linear prediction rule in the population, we can also approximate the best linear prediction rule for predicting outcomes given future X's sampled from the population.

The fundamental statistical issue is that we are trying to estimate p parameters, $\beta_1, ..., \beta_p$, without imposing any assumptions on these parameters. Intuitively, to estimate each parameter well, we need many observations per parameter. This intuition suggests that n/p should be large, or, equivalently that p/n should be small, in order for estimation error to be small. The following result captures this intuition more formally.

The following theorem bounds the root mean square approximation error (RMSE) defined as:

$$\sqrt{\mathrm{E}_X[(\beta'X - \hat{\beta}'X)^2]} = \sqrt{(\hat{\beta} - \beta)'\mathrm{E}_X[XX'](\hat{\beta} - \beta)},$$

where E_X is the expectation with respect to X alone.

Theorem 1.2.1 (Approximation of BLP by OLS) *Under regularity conditions,*[a] *the RMSE is bounded by:*

$$const_{\mathrm{P},\alpha} \cdot \sqrt{\mathrm{E}\varepsilon^2}\sqrt{p/n},$$

the inequality holds with probability approaching $1 - \alpha$ *as* $n \to \infty$*, and* $const_{\mathrm{P},\alpha}$ *is a constant that depends on the distribution of* (Y, X) *and* α*.*

[a] See Notes (Section 1.6) for references.

Theorem 1.2.1 says that, for nearly all realizations of data, the sample linear regression is close to the population linear regression if n is large and p is much smaller than n:[8]

$$\sqrt{\mathrm{E}_X[(\beta'X - \hat{\beta}'X)^2]} \approx 0.$$

In other words, under our requirement of p/n small, the sample BLP approximates the population BLP well.

8: Given indexed random variables (vectors, elements) A_n and B_n in a metric space equipped with metric d, the notation $A_n \approx B_n$ means that the distance between A_n and B_n concentrates around 0 – formally, that $\lim_{n\to\infty} \mathrm{P}(d(A_n, B_n) \leq \varepsilon) = 1$ for each $\varepsilon > 0$.

Analysis of Variance

Analysis of variance involves the decomposition of the variation of Y into explained and unexplained parts. Explained variation is a measure of the predictive performance of a model. This decomposition can be conducted both in the population and in the sample.

The main idea is to use the previous decomposition of Y,

$$Y = \beta'X + \varepsilon, \quad \mathrm{E}[\varepsilon X] = 0,$$

to decompose the variation in Y into the sum of *explained variation* and *residual variation*:

$$\mathrm{E}[Y^2] = \mathrm{E}[(\beta'X)^2] + \mathrm{E}[\varepsilon^2].$$

The quantity

$$\mathrm{MSE}_{pop} = \mathrm{E}[\varepsilon^2]$$

is the population MSE. The ratio of the explained variation to the total variation is the population R^2:

$$R^2_{pop} := \frac{\mathrm{E}[(\beta' X)^2]}{\mathrm{E}[Y^2]} = 1 - \frac{\mathrm{E}[\varepsilon^2]}{\mathrm{E}[Y^2]} \in [0, 1].$$

That is, R^2_{pop} is the proportion of variation of Y explained by the BLP.

Remark 1.2.1 The "standard" definition of R^2 assumes that we work with a *centered random variable* Y, that is, we recenter Y such that $\mathrm{E}[Y] = 0$. (However, our definition above does not require this property).

The decomposition of the variance in the sample proceeds analogously. Using the representation

$$Y_i = \hat{\beta}' X_i + \hat{\varepsilon}_i$$

and the orthogonality condition $\mathbb{E}_n[X\hat{\varepsilon}] = 0$ provided by the sample Normal Equations, we obtain the decomposition

$$\mathbb{E}_n[Y^2] = \mathbb{E}_n[(\hat{\beta}' X)^2] + \mathbb{E}_n[\hat{\varepsilon}^2].$$

Thus, we can define the sample MSE,

$$\mathrm{MSE}_{sample} = \mathbb{E}_n[\hat{\varepsilon}^2],$$

and the sample R^2,

$$R^2_{sample} := \frac{\mathbb{E}_n[(\hat{\beta}' X)^2]}{\mathbb{E}_n[Y^2]} = 1 - \frac{\mathbb{E}_n[\hat{\varepsilon}^2]}{\mathbb{E}_n[Y^2]} \in [0, 1].$$

By the law of large numbers and Theorem 1.2.1, when p/n is small, we have the following approximations:

$$\mathbb{E}_n[Y^2] \approx \mathrm{E}[Y^2], \quad \mathbb{E}_n[(\hat{\beta}' X)^2] \approx \mathrm{E}[(\beta' X)^2], \quad \mathbb{E}_n[\hat{\varepsilon}^2] \approx \mathrm{E}[\varepsilon^2].$$

Thus, when p/n is small and n is large, the sample fit measures are good approximations to population fit measures:

$$\mathrm{MSE}_{sample} \approx \mathrm{MSE}_{pop} \text{ and } R^2_{sample} \approx R^2_{pop}.$$

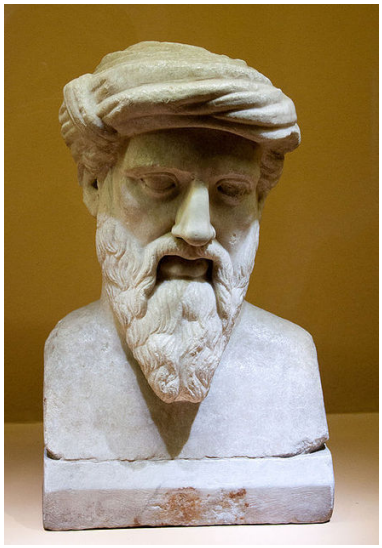

Figure 1.4: Pythagoras of Samos invented least squares and analysis of variance for the case of $n = 2$ and $p \leq 2$ around 570 BC. He was therefore the first known machine learner.

Overfitting: What Happens When p/n Is Not Small

When p/n is not small, the picture about predictive performance of the in-sample BLP becomes inaccurate and possibly misleading. In this setting, the in-sample linear predictor can be substantially different from the population BLP.

Consider an extreme example where $p = n$ and all variables in X are linearly independent. In this case, we have

$$\text{MSE}_{sample} = 0 \text{ and } R^2_{sample} = 1$$

no matter what MSE_{pop} and R^2_{pop} are. E.g. we could have $R^2_{sample} = 1$ even if $R^2_{pop} = 0$. Therefore, here we have an extreme example of *overfitting*, where the in-sample predictive performance overstates the out-of-sample predictive performance of the linear model. The following example illustrates less extreme cases.

The Notebooks 1.7.3 contain code for the numerical experiment.

Example 1.2.1 (Overfitting Example) Suppose $X \sim N(0, I_p)$ and $Y \sim N(0, 1)$ are statistically independent. It follows that the best linear predictor of Y is $\beta' X = 0$ and that $R^2_{pop} = 0$.

- If $p = n$, then the typical R^2_{sample} is $1 \gg 0$.
- If $p = n/2$, then the typical R^2_{sample} is about $.5 \gg 0$.
- If $p = n/20$, then the typical R^2_{sample} is about $.05 > 0$.

These results can be deduced by simulation or analytically.

Provided $p < n$, better measures of out-of-sample predictive ability are the "adjusted" R^2 and MSE:[9]

$$\text{MSE}_{adjusted} = \frac{n}{n-p}\mathbb{E}_n[\hat{\varepsilon}^2], \quad R^2_{adjusted} := 1 - \frac{n}{n-p}\frac{\mathbb{E}_n[\hat{\varepsilon}^2]}{\mathbb{E}_n[Y^2]}.$$

9: The adjustment factor $\frac{n}{n-p}$ is derived in a homoskedastic model, so that $\text{E}[\text{MSE}_{adjusted}] = \text{MSE}_{pop}$, see e.g., p. 8 in [3] for the derivation.

The adjustment by $\frac{n}{n-p}$ corrects for overfitting and provides a more accurate assessment of predictive ability of the linear model in Example 1.2.1 and more generally under the assumption of homogeneous ε. The intuition is that models with many parameters increase the in-sample fit and potentially cause overfitting. Hence, the number of parameters is incorporated in the definition of $\text{MSE}_{adjusted}$ and $R^2_{adjusted}$ in an attempt to account for this phenomenon.

Measuring Predictive Ability by Sample Splitting

How should we measure the predictive ability of the linear model (or other nonlinear models that we will discuss) more reliably, even in cases when p/n is not small?

A general way to measure predictive performance is to perform *data splitting*. The idea can be summarized in two parts:

1. Use a random part of a dataset, called the training sample, for estimating/training the prediction rule.
2. Use the other part, called the testing sample, to evaluate the quality of the prediction rule, recording out-of-sample mean squared error and R^2.

Generally, a predictive model is trained on a sample and the real test of its predictive ability happens when "new, unseen" observations arrive. With new observations in hand, we learn how far off our predictions are, when compared to the realized values. By partitioning the data set into two parts, we preserve an "unseen" set of observations on which to test our model, mimicking this process of ex-post performance assessment.[10]

10: If the "test set" is used many times to evaluate models, it becomes a "validation" set. The term "test set" is often reserved for the final evaluations of very few models.

The data splitting procedure can be described more formally as follows:

Generic Evaluation of Prediction Rules by Sample-Splitting

1. Randomly partition the data into training and testing samples. Suppose we use n observations for training and m for testing/validation.
2. Use the training sample to compute a prediction rule $\hat{f}(X)$. For example, $\hat{f}(X) = \hat{\beta}'X$ in the linear model.
3. Let $\mathcal{I}$ denote the indexes of the observations in the test sample. Then the out-of-sample/test mean squared error is

$$\text{MSE}_{test} = \frac{1}{m}\sum_{k\in\mathcal{I}}(Y_k - \hat{f}(X_k))^2,$$

and the out-of-sample/test R^2 is

$$R^2_{test} = 1 - \frac{\text{MSE}_{test}}{\frac{1}{m}\sum_{k\in\mathcal{I}} Y_k^2}.$$

In Section 3.B, we consider a more data-efficient evaluation procedure called cross-validation. In brief, we split the data into K folds of about equal size. For each fold, we repeat the evaluation procedure by designating that fold as the "test" set and using the remaining folds for training. We then average the values of MSE_{test} computed in each fold. Moreover, we can also average these results over different Monte Carlo seeds.

There is an important variation on the sample splitting procedure, called *stratified splitting* that provides guarantees that the training and test samples are similar across key subgroups called "strata".[11] In large samples, training and test samples will be similar across the subgroups by virtue of the laws of large numbers, but similarity is not guaranteed in moderate-sized samples. For more discussion, please see this blog on Data Splitting [4].

11: For example, we can make sure that the proportions of college-graduates and non-college-graduates are the same in both training and test samples. These issues are important in moderate-sized samples.

1.3 Inference about Predictive Effects or Association

Here we examine inference on *predictive effects*, which describe how our (population best linear) predictions change if the value of a regressor changes by a unit, while the other regressors remain unchanged.

Specifically, we partition the vector of regressors X into two components:

$$X = (D, W')',$$

where D represents the "target" regressor of interest, and W represents the other regressors, sometimes called the controls. We can therefore write

$$Y = \underbrace{\beta_1 D + \beta_2' W}_{\text{predicted value}} + \underbrace{\varepsilon}_{\text{error}}, \tag{1.3.1}$$

and ask the question: [12]

> How does the predicted value of Y change if D increases by a unit while W remains unchanged?

12: Note that this question is purely about the properties of the prediction rule and generally has nothing to do with causality.

The answer is the predicted value of Y changes by

$$\beta_1.$$

Example 1.3.1 (Wage Differences) In the analysis of wages, which we will discuss later in more detail, an interesting question can be formulated as:

- "What is the difference in predicted wages between female and non-female workers with the same job-relevant characteristics?"

Let D represent the female indicator and W represent experience, educational, occupational, and geographic characteristics. The answer to the question is then the population regression coefficient

$$\beta_1$$

corresponding to D.

Understanding β_1 via "Partialling-Out"

"Partialling-out" is an important tool that provides conceptual understanding of the regression coefficient β_1.

In the *population*, we define the partialling-out operation as a procedure that takes a random variable V and creates the "residualized" variable $\tilde{V}$ by subtracting the part of V that is linearly predicted by W:

$$\tilde{V} = V - \gamma_{VW}' W, \quad \gamma_{VW} \in \arg\min_{\gamma} \mathrm{E}\left[(V - \gamma' W)^2\right].$$

When V is a vector, we apply the operation to each component. It can be shown that the partialling-out operation is linear in the sense that[13]

$$Y = \nu V + \mu U \implies \tilde{Y} = \nu \tilde{V} + \mu \tilde{U}.$$

13: Verify this as a reading exercise. Use the definition of the BLP decompositions of U and V with respect to regressors W, to derive a BLP decomposition of Y with respect to W.

Formally, this operation is well defined on the space of random variables with finite second moments.

We apply the partialling-out operation to both sides of our regression equation $Y = \beta_1 D + \beta_2' W + \varepsilon$ to get

$$\tilde{Y} = \beta_1 \tilde{D} + \beta_2' \tilde{W} + \tilde{\varepsilon},$$

which simplifies to the decomposition:

$$\tilde{Y} = \beta_1 \tilde{D} + \varepsilon, \quad \mathrm{E}\left[\varepsilon \tilde{D}\right] = 0. \tag{1.3.2}$$

Decomposition (1.3.2) follows because partialling-out eliminates $\beta_2' W$, since $\tilde{W} = 0$, and leaves ε untouched, $\tilde{\varepsilon} = \varepsilon$, since ε is

linearly unpredictable by X and therefore by W. Moreover, $\mathrm{E}[\varepsilon\tilde{D}] = 0$ since $\tilde{D}$ is a linear function of $X = (D, W')'$ and ε is orthogonal to X and therefore to any linear function of X.

The decomposition (1.3.2) implies that $\mathrm{E}\varepsilon\tilde{D} = 0$ is the Normal Equation for the population regression of $\tilde{Y}$ on $\tilde{D}$. Therefore, we just rediscovered the following result.

Theorem 1.3.1 (Frisch-Waugh-Lovell) *Assume that Y, D, W have finite second moments and that D is not perfectly predictable by W, i.e., $\mathrm{E}[\tilde{D}^2] > 0$. The population linear regression coefficient β_1 can be recovered from the population linear regression of $\tilde{Y}$ on $\tilde{D}$:*

$$\beta_1 = \arg\min_{b_1} \mathrm{E}[(\tilde{Y} - b_1\tilde{D})^2] = (\mathrm{E}[\tilde{D}^2])^{-1}\mathrm{E}[\tilde{D}\tilde{Y}].$$

In other words, β_1 can be interpreted as a (univariate) linear regression coefficient in the linear regression of *residualized* Y on *residualized* D, where the residuals[14] are defined by partialling-out the linear effect of W from Y and D.

14: Technically, these are regression *errors*, not residuals, as we are here working with the population, whereas residuals refer to errors to the sample regression fit. However, we will not adhere strictly to this distinction as it will be convenient to apply analogous logic to partialling-out in the population and the sample.

When we work with the *sample*, we simply mimic the partialling-out operation in the population in the sample. In what follows, we assume p/n is small, so sample linear regression provides high-quality partialling-out. By the FWL Theorem applied to the sample instead of in the population, the sample linear regression of Y on D and W gives us the estimator $\hat{\beta}_1$ which is identical to the estimator obtained via sample partialling-out.

It is useful to give the formula for $\hat{\beta}_1$ in terms of sample partialling-out:

$$\hat{\beta}_1 = \arg\min_{b} \mathbb{E}_n[(\check{Y} - b\check{D})^2] = (\mathbb{E}_n[\check{D}^2])^{-1}\mathbb{E}_n[\check{D}\check{Y}], \quad (1.3.3)$$

where $\check{V}_i$ is the residual left after predicting V_i with controls W_i in the sample and we assume $\mathbb{E}_n[\check{D}^2] > 0$. That is,

$$\check{V}_i = V_i - \hat{\gamma}_{VW}'W_i, \quad \hat{\gamma}_{VW} \in \arg\min_{\gamma} \mathbb{E}_n[(V - \gamma'W)^2].$$

From Theorem 1.2.1, we know that using sample linear regression for partialling-out will provide high-quality estimates of the residuals when p/n is small. When p/n is not small, using sample linear regression for partialling-out won't be such a good idea and an alternative is to use penalized regression or dimension reduction. We will cover this in Chapter 3, but we can definitely try it out in the empirical example that concludes this chapter before we even attempt to understand it.[15]

15: Why not?

Adaptive Statistical Inference

We next consider the large sample properties of the estimator $\hat{\beta}_1$.

Theorem 1.3.2 (Adaptive Statistical Inference) *Under regularity conditions and if $p/n \approx 0$, the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on the stochastic behavior of $\hat{\beta}_1$. Namely,*

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \approx \sqrt{n}\mathbb{E}_n[\tilde{D}\varepsilon]/\mathbb{E}_n[\tilde{D}^2] \tag{1.3.4}$$

and consequently,

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \overset{a}{\sim} N(0, \mathsf{V})$$

where

$$\mathsf{V} = (\mathrm{E}[\tilde{D}^2])^{-1}\mathrm{E}[\tilde{D}^2\varepsilon^2](\mathrm{E}[\tilde{D}^2])^{-1}.$$

We can equivalently write

$$\hat{\beta}_1 \overset{a}{\sim} N(\beta_1, \mathsf{V}/n).$$

That is, $\hat{\beta}_1$ is approximately normally distributed with mean β_1 and variance V/n. Thus, $\hat{\beta}_1$ concentrates in a $\sqrt{\mathsf{V}/n}$- neighborhood of β_1 with deviations controlled by the normal law.

The first result in Theorem 1.3.2, equation (1.3.4), states the estimator minus the estimand is an approximate centered average. The remaining properties stated in the theorem then follow from the central limit theorem.

The *adaptivity* refers to the fact that estimation of residuals $\check{D}$ has a negligible impact on the large sample behavior of the OLS estimator – the approximate behavior is the same as if we had used true residuals $\tilde{D}$ instead. This adaptivity property will be derived later as a consequence of a more general phenomenon which we shall call *Neyman orthogonality*.[16]

The estimated standard error of $\hat{\beta}_1$ is $\sqrt{\hat{\mathsf{V}}/n}$, where $\hat{\mathsf{V}}$ is any estimator of V based on the plug-in principle such that $\hat{\mathsf{V}} \approx V$. The standard estimator for independent data is called the Eicker-Huber-White robust variance estimator ([5], [6],[7], [8]):

$$\hat{\mathsf{V}} = (\mathbb{E}_n[\check{D}^2])^{-1}\mathbb{E}_n[\check{D}^2\hat{\varepsilon}^2](\mathbb{E}_n[\check{D}^2])^{-1}.$$

This standard error estimator formally works when $p/n \approx 0$, but fails in settings where p/n is not small; see, e.g., [9].

Consider the set, called the $(1-\mathsf{a})\%$ *confidence interval*, [17]

The notation $A_n \overset{a}{\sim} N(0, \mathsf{V})$ reads as A_n is approximately distributed as $N(0, \mathsf{V})$. Approximate distribution formally means that $\sup_{R\in\mathscr{R}} |\mathrm{P}(A_n \in R) - \mathrm{P}(N(0, \mathsf{V}) \in R)| \approx 0$, where $\mathscr{R}$ is the collection of rectangular sets (intervals for the case of A_n being a scalar random variable).

16: We'll defer the formal defintion of Neyman orthogonality for a bit. See Section 4.3.

17: An alternative is to use $\frac{n}{n-p}(\mathbb{E}_n[\check{D}^2])^{-1}\mathbb{E}_n[\widehat{\varepsilon}^2]$ instead of $\hat{V}$, which is a "classical" choice. However, this choice is not robust to the presence of *heteroscedasticity* – a term that denotes the conditional variance of ε depending on X – and is therefore not valid in general. We never use the "classical" choice in this book, even though it is still a default reporting option in some statistical software.

$$[\hat{l}, \hat{u}] := \left[\hat{\beta}_1 - z_{1-a/2}\sqrt{\hat{V}/n}, \hat{\beta}_1 + z_{1-a/2}\sqrt{\hat{V}/n}\right],$$

where $z_{1-a/2}$ denotes the $(1-a/2)-$quantile of the standard normal distribution. We say that a $(1-a)\times 100\%$ confidence interval contains the true value β_1 "$(1-a)\times 100\%$ of the time," approximately. For example, the 95% confidence interval is given by

$$\left[\hat{\beta}_1 - 1.96\sqrt{\hat{V}/n}, \hat{\beta}_1 + 1.96\sqrt{\hat{V}/n}\right],$$

and contains β_1 approximately 95% of the time.

Remark 1.3.1 (What does "of the time" mean?) If we imagine drawing samples of size n repeatedly from the same population, a $(1-a)\times 100\%$ confidence interval would contain β_1 in approximately $(1-a)\times 100\%$ of those samples:

$$P\big([\hat{l}, \hat{u}] \text{ contains } \beta_1\big) \approx 1-a.$$

In other words, aside from "atypical" samples—occurring with (small) probability $\approx a$—the confidence interval contains the population value of the best linear predictor coefficient β_1. In practice, of course, we do not repeatedly redraw samples; we work with a single fixed sample. Nevertheless, we hope that our sample is not one of those atypical ones, so that we do indeed achieve the desired coverage (or containment) of the true coefficient. Thus, the idea of repeated sampling, also known as frequentist inference, is only used to define the exact meaning of the confidence.

1.4 Application: Wage Prediction and Gaps

In labor economics, an important question is what determines the wage of workers. Interest in this question goes back at least to the work of Jacob Mincer (see [10]). While determining the factors that lead to a worker's wage is a causal question, we can begin to investigate it from a predictive perspective. We aim to answer two main questions:

- The Prediction Question: How can we use job-relevant characteristics, such as education and experience, to best predict wages?

- The Predictive Effect or Association Question: What is the difference in predicted wages between male and female workers with the same job-relevant characteristics?

We illustrate using data from the 2015 March Supplement of the U.S. Current Population Survey (CPS 2015). As outcome, Y, we use the log hourly wage, and we let X denote various characteristics of workers, including sex.

We focus on a (sub) sample of single (never married) workers, which is of size $n = 5,150$. Table 1.1 provides mean characteristics of some key variables.

	Sample Mean
Log Wage	2.97
Female	0.44
Some High School	0.02
High School Graduate	0.24
Some College	0.28
College Graduate	0.32
Advanced Degree	0.14
Experience	13.76

Table 1.1: Descriptive statistics for sample of never married workers.

We will estimate a linear predictive (regression) model for log hourly wage using job-relevant characteristics

$$Y = \beta' X + \varepsilon, \quad \varepsilon \perp X,$$

and assess the quality of the empirical prediction rule $\hat{\beta}' X$ using out-of-sample prediction performance.

We will also analyze if there is a gap (difference) in pay for male and female workers.[18] Any such gap may partly reflect discrimination in the labor market. We will discuss the potential to learn about discrimination as a causal mechanism in more detail in Chapter 6.

18: More precisely, we will analyze the sex pay gap using data from the CPS 2015, where `sex` is recorded as binary and self-reported variable. In labor economics, such analyses are often referred to as "gender wage gap" studies; see, for example, the Nobel lecture by Claudia Goldin [11] for an overview of this area of research.

Prediction of Wages

Our goal here is to predict (log) wages using various characteristics of workers, and assess the predictive performance of two linear models using adjusted MSE and R^2 and out-of-sample MSE and R^2.

We start with two different specifications of covariates to use as predictors: [19]

19: The Notebooks 1.7.1 contain the predictive exercise for wages.

- In the **Basic Model** X consists of a set of raw regressors (e.g. sex, experience, education indicators, occupation and industry indicators, and regional indicators), for a total of $p = 51$ regressors. Our basic specification is inspired by the famous Mincer equation from labor economics; see, e.g., [10] for a review.

- In the **Flexible Model**, X consists of all raw regressors from the basic model as well as *technical regressors*, which are transformations of the raw regressors, namely, polynomials in experience (experience^2, experience^3, and experience^4) and additional two-way interactions of the polynomials in experience with all other raw regressors except for sex. An example of a regressor created through a two-way interaction is *experience* times the indicator of having a *college degree*. In total, we have $p = 246$ regressors.

	p	R^2_{sample}	MSE_{sample}	R^2_{adj}	MSE_{adj}
basic	51	0.30	0.23	0.30	0.23
flexible	246	0.35	0.22	0.31	0.23
flexible Lasso	246	0.32	0.23	0.31	0.23

Table 1.2: Assessment of predictive performance with in-sample R^2 and MSE.

To enable both in- and out-of-sample performance evaluation. We start by randomly selecting 80% of the observations as the training sample and keep the other 20% for use as a test sample.

Table 1.2 shows measures of predictive performance in the training data. That is, the table reports predictive performance on the same data that were used to estimate the model parameters. The flexible regression model performs slightly better than the basic model (higher R^2_{adj} and lower MSE_{adj}). Note also that the discrepancy between the unadjusted and adjusted measures is not large, which is expected given that

$$p/n \text{ is small.}$$

We report results for evaluating the prediction rules in the test data in Table 1.3. That is, the table reports predictive performance on *new* data that were not used to estimate the models.

Based on this exercise, it appears that the basic regression model works slightly better than the flexible regression at predicting log wages for new observations. That is, we see that the test (out-of-sample) MSE and R^2 for the basic regression model

	MSE_{test}	R^2_{test}
basic	0.197	0.328
flexible	0.206	0.296
flexible Lasso	0.200	0.317

Table 1.3: Assessment of predictive performance on a 20% validation sample.

are respectively slightly lower and higher than those of the flexible regression model, indicating slightly superior out-of-sample predictive performance. This behavior is different from that obtained when looking at the within sample fit statistics reported in Table 1.2.

Tables 1.2 and 1.3 also provide the test MSE of the flexible model estimated via Lasso regression.[20] Lasso is a penalized regression method used to reduce the complexity of a regression model when the ratio p/n is not small. It achieves this by penalizing the sum of the absolute values of the regression coefficients, thereby shrinking some coefficients to exactly zero and effectively performing variable selection.

20: Lasso stands for "Least Absolute Shrinkage and Selection Operator".

We introduce this method in more detail in Chapter 3, but it is applied here despite appearing as a "black box" at this stage. The out-of-sample MSE can also be computed for any other black-box prediction method. In this example, Lasso performs similarly to the basic and flexible regression models estimated using OLS. This result is not surprising, given the modest dimensionality of the problem and the similarity in performance between the two OLS-estimated models.

Finally, to highlight the potential of estimating the linear model via OLS to overfit, we consider one more model.

- In the **Extra Flexible Model**, X consists of sex and all two-way interactions between experience, experience^2, experience^3, experience^4, and all other raw regressors except for sex. In total, we have $p = 979$ regressors in this specification.

	OLS	Lasso
MSE_{sample}	0.178	0.210
MSE_{adj}	0.235	0.223
MSE_{test}	0.250	0.199
R^2_{sample}	0.467	0.368
R^2_{adj}	0.345	0.331
R^2_{test}	0.148	0.322

Table 1.4: Assessment of predictive performance in the extra flexible model with $p = 979$ regressors.

We report measures of predictive performance in the training and test data from OLS and Lasso estimates of our "extra flexible" model in Table 1.4. Here, we see that the model estimated by OLS appears to be overfitting. The in-sample statistics substantially overstate predictive performance relative to the performance we see in the test data. For example, the R^2 and adjusted R^2 in the training data are 0.467 and 0.345, both of which substantially overstate the R^2 obtained in the test data, 0.148. We also see that the performance on the test data for the extra flexible model is substantially worse than the performance of the much simpler basic and flexible models. That is, it looks like the OLS estimates of the extra flexible model have specialized to fitting aspects of the training data that do not generalize to the test data and lead to a deterioration in predictive performance relative to the simpler models.

The performance of the Lasso contrasts sharply with this behavior. We see that the in-sample and out-of-sample predictive performance measures for the Lasso based estimates of the extra flexible model are similar to each other. They are also similar to the performance of the simpler models. It seems that Lasso is finding a competitive predictive model without overfitting even in the extra flexible model. We will return to this behavior in Chapter 3 where we will show that Lasso and related methods are able to find good prediction rules in even extremely high-dimensional settings, where for example $p \gg n$, where OLS breaks down theoretically and in practice.

Wage Gap

The Notebooks 1.7.2 contain the code for this section.

An important question is whether there is a gap (i.e., difference) in predicted wages between male and female workers with the same job-relevant characteristics. To answer this question, we estimate the log-linear regression model:

$$Y = \beta_1 D + \beta_2' W + \varepsilon, \tag{1.4.1}$$

where Y is log-wage, D is the sex indicator (1 if female and 0 otherwise) and the W's are other determinants of wages. W includes a constant of 1, education, polynomials in experience, region, and occupation and industry indicators plus all two-way interactions of polynomial in experience with region, occupation, and industry indicators. This gives us $p = 246$ regressors.

As we have log-transformed wages, we are analyzing the relative difference in pay for male and female workers. Table 1.5 tabulates

	All	Male	Female
Log Wage	2.9708	2.9878	2.9495
Less then High School	0.0233	0.0318	0.0127
High School Graduate	0.2439	0.2943	0.1809
Some College	0.2781	0.2733	0.2840
College Graduate	0.3177	0.2940	0.3473
Advanced Degree	0.1371	0.1066	0.1752
Experience	13.7606	13.7840	13.7313

Table 1.5: Empirical means for the groups defined by the sex variable for never-married workers.

mean characteristics given sex. It shows that the difference in average log-wage between never married male and never married female workers is equal to 0.038 with male workers earning more. Thus, in this group, male average wage is about 3.8% higher than female average wage.[21] We also observe that never married female workers are relatively more educated than never married male workers.

21: This interpretation relies on the approximation $\log(a) - \log(b) \approx (a-b)/b$, which is accurate whenever $(a-b)/b$ is small and $a, b > 0$.

Table 1.6 summarizes the regression results. Overall, we see that the unconditional wage gap of size 3.8% for female workers increases to about 7% after controlling for worker characteristics. This means we would predict a female worker's wage to be about 7% less per hour on average than the wage of a male worker who had the same experience, education, geographical region and occupation.

The partialling-out approach provides a numerically identical estimate for the coefficient β_1 ($\beta_1 \approx 7\%$), numerically confirming the FWL theorem. Using Lasso for partialling-out (*p-out w/Lasso*) gives similar results to using OLS. This similarity is expected here, since

$$p/n \text{ is small,}$$

and partialling out by least squares should work well.

	Estimate	Std. Error
reg without controls	−0.038	0.016
reg with controls	−0.070	0.015
partial out reg w/ controls	−0.070	0.015
Double Lasso (p-out w/ Lasso)	−0.072	0.015

Table 1.6: Estimated conditional wage gaps for never married workers.

To sum up, our estimate of the conditional wage gap for never-married workers using OLS is about −7% and the 95% confidence interval is about [−10%, −4%].

Kitagawa-Oaxaca-Blinder Decomposition

One way to understand the estimate with controls (−0.070) is as the part of the total gap (−0.038) that cannot be explained by differences in group characteristics. Namely, take Eq. (1.4.1) and average it in the male and female groups[22] to obtain the decomposition:[23]

$$\underbrace{\mathbb{E}_n[Y \mid D=1] - \mathbb{E}_n[Y \mid D=0]}_{-0.038} = \underbrace{\hat{\beta}_1}_{-0.070} + \underbrace{\hat{\beta}_2'(\mathbb{E}_n[W \mid D=1] - \mathbb{E}_n[W \mid D=0])}_{0.032}.$$

Here, the 0.032 difference in average log wages, predicted based on differences in observed characteristics (W) and slopes ($\hat{\beta}_2$), suggests higher average log wages for female workers compared to male workers. However, this positive difference is counteracted by a negative difference of −0.070 that remains unexplained by the characteristics. This unexplained difference arises from the different pay predicted for female and male workers possessing the same characteristics.[24]

1.5 Inference on Predictive Effect when $p/n < 1$ is not small*

In order to wrap up and provide a stylized illustration of the impact of dimensionality p on inference, we revisit the extra-flexible model from the prediction exercise, which used $p = 979$ controls. To further "stress-test" the inference, we reduce the sample size to $n = 1000$ by selecting a random subset of the original observations.[25]

In this reduced-sample setting, we have $p/n \approx 1$, meaning the usual theory for estimating linear model coefficients using OLS no longer applies. [15] provide more refined results for OLS estimates of regression coefficients in the case where $p/n \to C < 1$. They show that OLS estimates of individual coefficients can still be consistent in this regime and also provide an estimator for the asymptotic variance that is consistent when $p/n < 1/2$, as long as certain regularity conditions are satisfied. Furthermore, they note that the usual Eicker-Huber-White robust variance estimator is inconsistent in this high-dimensional setting, while the jackknife variance estimator,[26] although not consistent, remains conservative.

22: $\mathbb{E}_n[\cdot \mid D = d]$ abbreviates $\mathbb{E}_n$ for the subsample of the data where $D = d$, for $d = 0, 1$.

23: Decompositions of this sort and the one given below are called *Kitagawa-Oaxaca-Blinder decomposition* introduced in [12], [13], and [14], in different contexts.

24: Differences of this sort potentially reveal discrimination in the pay structure, a question we will return to later in the book.

25: Using the full sample yields results very similar to those reported in the previous section. In this case, $p/n \approx 1/5$, which is small enough for conventional OLS inference to perform reasonably well, demonstrating its substantial resilience. Please verify this as an exercise using the Wage Gap Notebook.

26: The jackknife variance estimator is a resampling-based method. First, an estimate $\hat{\theta}$ is computed using all n observations. Then, each observation is omitted one at a time, and the model is refitted to obtain a new estimate $\hat{\theta}_{(-i)}$. The variance estimate is calculated as

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(-i)} - \hat{\theta}\right)^2,$$

which measures the sensitivity of $\hat{\theta}$ to individual data points by comparing leave-one-out estimates to the original estimate.

We report estimates of the conditional wage gap in this setup in Table 1.7. Specifically, we report point estimates from OLS applied to the full set of variables, providing both the Eicker-Huber-White standard error (HC0) and the jackknife standard error (HC3).[27] These are provided mainly for illustration, but we note that HC0 is known to be inconsistent and to behave very poorly—generally being far too small—in the high-dimensional setting. HC3 is more reliable, but one should also be skeptical given that $p/n \approx 1$ in this example. Finally, we report point estimates and standard errors for the Double Lasso procedure. The resulting estimator is consistent, asymptotically normal, and has estimable standard errors under the structure outlined in Chapter 4 even when $p \gg n$. For now, we can think of it as a point of comparison.

27: The Eicker-Huber-White variance estimator is often referred to as "HC0" and the jackknife as "HC3."

	Estimate	HC0	HC3
Regression	-0.067	0.039	0.073
Double Lasso (p-out w/ Lasso)	-0.054	0.034	0.034

Table 1.7: The estimated conditional wage gaps for never married workers with approximately 1000 controls in a sample of 1000 observations.

Comparing to the case with the full data set, we see that point estimates are not wildly different but that standard errors are larger. Part of the standard error difference is predicted simply by the difference in sample sizes. Specifically, $\sqrt{5150/1000} \approx 2.27$, so we would expect standard errors to be about 2.27 times larger with $n = 1000$ observations than with $n = 5150$. This inflation holds almost exactly for the Double Lasso estimates.

More interestingly, now that $p/n \not\approx 0$, we start seeing substantial differences in standard errors between unregularized partialling out (OLS) and partialling out with Lasso (also known as Double Lasso). While we do not want to take the OLS standard errors too seriously—given that the Huber-Eicker-White standard error does not work in this setting and we are also skeptical of the jackknife here—the comparison between the OLS and Double Lasso standard errors, as well as the comparison to the full-sample results, is revealing. Relative to the full sample results, the jackknife standard error (HC3) is much larger than would be expected simply due to the decrease in sample size. The difference from this expectation (partially) reflects the impact of dimensionality on the OLS estimate of the regression coefficient. In contrast, the Double Lasso appears to be roughly insensitive to the dimensionality of the control variables and scales exactly as one would expect given the difference in sample size.

The punchline of this final example is that OLS is no longer adaptive in the "p/n not small" regime. The lack of adaptivity

means that conventional properties of OLS may fail, and that other procedures may become preferable to OLS.

1.6 Notes

Least squares were invented by Legendre ([16]) and Gauss ([17]) around 1800. Frisch, Waugh, and Lovell ([18],[19],[20]) discovered the partialling-out interpretation of the least squares coefficients in the 1930s. The asymptotic theory mentioned in Theorems 1.2.1 and 1.3.2 is more recent and has been developed since early work of Huber in the 70s on m-estimators (estimators that minimize objective functions that correspond empirical averages of losses) under moderately high dimensions; see e.g. [21] and the textbook [22].

For a good, concise treatment of classical least squares, see for example, Chapter 1 in Amemiya's classical graduate econometrics text [3]; Bruce Hansen's new textbook [23] is an excellent up-to-date reference.

Regularity conditions under which Theorem 1.2.1 and Theorem 1.3.2 hold under $p \to \infty$ and $p/n \to 0$ asymptotics can be found in [24] and [15]. The results of the latter reference allow for $p/n \to c < 1$, which introduces an additional asymptotic variance term when $c > 0$; the case with $c = 0$ recovers Theorem 1.3.2. See also review [25] for some recent understanding of properties of least squares estimators.

1.7 Notebooks

Notebook 1.7.1 (Predicting Wages) Predicting Wages R Notebook and Predicting Wages Python Notebook contain a simple predictive exercise for wages. We will return to this dataset and prediction problem repeatedly in future chapters, re-estimating it using a broad range of ML estimators and providing a means of comparing their performance.

Notebook 1.7.2 (Analyzing Wage Gaps) Wage Gaps R Notebook and Wage Gaps Python Notebook contain a simple analysis of wage gaps.

Notebook 1.7.3 (Exploring Overfitting Notebook) The Linear Model Overfitting R Notebook and the Linear Model Over-

fitting Python Notebook contain a set of simple simulations that show how measures of fit perform in a high p/n setting.

1.8 Exercises

Exercise 1.8.1 (Sample Splitting) Write a notebook (R, Python, etc.) where you briefly explain the idea of sample splitting to evaluate the performance of prediction rules to a fellow student, and show how to use it on the wage data. The explanation should be clear and concise (one paragraph suffices) so that a fellow student can understand. You can take our notebooks as a starting point, but provide a bit more explanation and modify them by exploring different specifications of the models (or looking at an interesting subset of the data or even other data – for example, the data you use for your research or thesis work).

Exercise 1.8.2 (Least Squares and Partialling-out) Write a notebook (R, Python, etc), where you carry out a wage gap analysis, focusing on the subset of college-educated workers. The analysis should be analogous to what we've presented – explaining "partialling out," generating point estimates and standard errors – but don't hesitate to experiment and explain more. Exploring other data-sets or similar questions, e.g. wage gaps by race, is always welcome.

Exercise 1.8.3 (Discovering Heterogeneity in Wage Gap) Write a notebook (R, Python, etc.) where you carry out a wage gap analysis and decomposition for each major education group separately. In essence, this amounts to performing linear wage gap regressions and decompositions for each group separately. Report the findings and any patterns of heterogeneity you observe. Is the heterogeneity you see economically and statistically significant? What if you perform the analysis by occupation groups instead? How do these group-wise decompositions contribute to the overall wage gap?

Exercise 1.8.4 (Machine Learning in Ancient Greece) The half-serious link to Pythagoras was serious in its half. Consider sample linear regression with $n = 2$ and just one regressor, so that $Y_i = \hat{\beta} X_i + \hat{\varepsilon}_i$ for $i = 1, 2$, where $\hat{\beta}$ is the ordinary least squares estimator, a scalar quantity in this case. Let $\mathsf{Y} = (Y_1, Y_2)'$, $\mathsf{X} = (X_1, X_2)'$, $\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2)'$, and let $\hat{\mathsf{Y}} = \hat{\beta}\mathsf{X}$. Find

the connection between the decomposition $\mathsf{Y}'\mathsf{Y}/n = \hat{\mathsf{Y}}'\hat{\mathsf{Y}}/n + \hat{\varepsilon}'\hat{\varepsilon}/n$ and the Pythagorean theorem. Find the geometric interpretation for $\hat{Y}$, and write the explicit formula for $\hat{\beta}$ in this case. If you get stuck, google the "geometric interpretation of least squares."

1.A Central Limit Theorem*

Univariate

Consider the scaled sum $W = \sum_{i=1}^{n} X_i/\sqrt{n}$ of independent and identically distributed variables X_i such that $\mathrm{E}[X] = 0$ and $\mathrm{Var}(X) = 1$. The classical CLT states that W is approximately Gaussian provided that none of the summands are too large, namely

$$\sup_{x\in\mathbb{R}} |\mathrm{P}(W \leq x) - \mathrm{P}(N(0,1) \leq x)| \approx 0.$$

This result is reassuring, but the theorem does not inform us how small the error is in a given setting.

The Berry-Esseen theorem provides a quantitative characterization of the error.

Theorem 1.A.1 (Berry-Esseen's Central Limit Theorem)

$$\sup_{x\in\mathbb{R}} |\mathrm{P}(W \leq x) - \mathrm{P}(N(0,1) \leq x)| \leq K\mathrm{E}[|X|^3]/\sqrt{n},$$

for a numerical constant $K < .5$.

The result asserts that the Gaussian approximation error rate declines like $1/\sqrt{n}$. It also states that given n, the approximation quality improves as the third absolute moment $\mathrm{E}[|X|^3]$ decreases. This results gives a good guide regarding when the Gaussian approximation gives accurate results.[28] Of course, one can also check the approximation quality via simulation experiments that mimic the practical situation.

28: Consider, for instance, the case when X_i are centered and standardized Bernoulli random variables with success probability p, i.e., $X_i = \frac{Z_i - p}{\sqrt{p(1-p)}}$ and Z_i is Bernoulli with success probability p. The error in the Berry-Esseen theorem, in this case, becomes $\approx 1/\sqrt{p(1-p)n}$. Thus, the error in the Gaussian approximation is guaranteed to be small by the Berry-Esseen theorem only if $p(1-p)n$ is large. Thus, for extreme probabilities, where either success or failure events are extremely rare for the given sample size, i.e., when $p \cdot n$ or $(1-p) \cdot n$ is small, the use of the Gaussian approximation is not advisable.

Multivariate

Later in the book, we will use multivariate central limit theorems as well. To this end, we are going to state the following more general result due to [26], which refines earlier results by [27] and [28].

Let $\mathcal{I}$ be a countable set (either finite or infinite) and let $X_i, i \in \mathcal{I}$, be independent $\mathbb{R}^d$-valued random vectors. Assume that $\mathrm{E}[X_i] = 0$ for all i and that $\sum_{i\in\mathcal{I}} \mathrm{Var}(X_i) = I_d$. It is well known that in this case, the sum $W := \sum_{i\in\mathcal{I}} X_i$ exists almost surely and that $\mathrm{E}W = 0$ and $\mathrm{Var}(W) = I_d$.

Theorem 1.A.2 (Multivariate CLT; [26]) *For X_i and W as above and all measurable convex sets $A \subseteq \mathbb{R}^d$, we have*

$$|\mathrm{P}(W \in A) - \mathrm{P}(N(0, I_d) \in A)| \leq \left(42d^{1/4} + 16\right) \sum_{i\in I} \mathrm{E}\left[\|X_i\|^3\right].$$

Bibliography

[1] Cambridge Dictionary, *Infer* (cited on page 11).

[2] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009 (cited on page 16).

[3] Takeshi Amemiya. *Advanced Econometrics*. Cambridge, MA: Harvard University Press, 1985 (cited on pages 20, 34).

[4] *Data Splitting | R-bloggers*. https://www.r-bloggers.com/2016/08/data-splitting/, Accessed: 2022-25-02 (cited on page 22).

[5] Friedhelm Eicker. 'Limit theorems for regressions with unequal and dependent errors'. In: ed. by Lucien M. Le Cam and Jerzy Neyman. 1967 (cited on page 25).

[6] Peter J. Huber. *The behavior of maximum likelihood estimates under nonstandard conditions*. English. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 221-233 (1967). 1967 (cited on page 25).

[7] Halbert White. 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity'. In: *Econometrica* (1980), pp. 817–838 (cited on page 25).

[8] Friedhelm Eicker. 'Asymptotic normality and consistency of the least squares estimators for families of linear regressions'. In: *Annals of Mathematical Statistics* 34.2 (1963), pp. 447–456 (cited on page 25).

[9] Matias Cattaneo, Michael Jansson, and Whitney Newey. 'Alternative Asymptotics and the Partially Linear Model with Many Regressors'. In: *Working Paper, http://econ-www.mit.edu/files/6204* (2010) (cited on page 25).

[10] Thomas Lemieux. 'The "Mincer equation" thirty years after schooling, experience, and earnings'. In: *Jacob Mincer a Pioneer of Modern Labor Economics*. Springer, 2006, pp. 127–145 (cited on pages 26, 28).

[11] Claudia Goldin. 'Nobel Lecture: An Evolving Economic Force'. In: *American Economic Review* 114.6 (2024), pp. 1515–1539 (cited on page 27).

[12] Evelyn M Kitagawa. 'Components of a difference between two rates'. In: *Journal of the American Statistical Association* 50.272 (1955), pp. 1168–1194 (cited on page 32).

[13] Ronald Oaxaca. 'Male-Female Wage Differentials in Urban Labor Markets'. In: *International Economic Review* 14.3 (1973), pp. 693–709. (Visited on 10/10/2023) (cited on page 32).

[14] Alan S. Blinder. 'Wage Discrimination: Reduced Form and Structural Estimates'. In: *Journal of Human Resources* 8.4 (1973), pp. 436–455. (Visited on 10/10/2023) (cited on page 32).

[15] Matias D. Cattaneo, Michael Jansson, and Whitney K. Newey. 'Inference in linear regression models with many covariates and heteroscedasticity'. In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1350–1361 (cited on pages 32, 34).

[16] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805*. Courcier, 1806 (cited on page 34).

[17] Carl-Friedrich Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae*. Henricus Dieterich, 1823 (cited on page 34).

[18] Ragnar Frisch and Frederick V Waugh. 'Partial time regressions as compared with individual trends'. In: *Econometrica* (1933), pp. 387–401 (cited on page 34).

[19] Michael C Lovell. 'Seasonal adjustment of economic time series and multiple regression analysis'. In: *Journal of the American Statistical Association* 58.304 (1963), pp. 993–1010 (cited on page 34).

[20] Michael C Lovell. 'A simple proof of the FWL theorem'. In: *Journal of Economic Education* 39.1 (2008), pp. 88–91 (cited on page 34).

[21] Peter J Huber. 'Robust regression: asymptotics, conjectures and Monte Carlo'. In: *Annals of Statistics* (1973), pp. 799–821 (cited on page 34).

[22] Elvezio M Ronchetti and Peter J Huber. *Robust Statistics*. John Wiley & Sons Hoboken, NJ, USA, 2009 (cited on page 34).

[23] Bruce E. Hansen. *Econometrics*. Princeton University Press, 2022 (cited on page 34).

[24] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. 'Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results'. In: *Journal of Econometrics* 186.2 (2015), pp. 345–366 (cited on page 34).

[25] Arun K. Kuchibhotla, Lawrence D. Brown, and Andreas Buja. 'Model-free study of ordinary least squares linear regression'. In: *arXiv preprint arXiv:1809.10538* (2018) (cited on page 34).

[26] Martin Raič. 'A multivariate Berry–Esseen theorem with explicit constants'. In: *Bernoulli* 25.4A (2019), pp. 2824–2853 (cited on pages 36, 37).

[27] Vidmantas Bentkus. 'On the dependence of the Berry–Esseen bound on dimension'. In: *Journal of Statistical Planning and Inference* 113.2 (2003), pp. 385–402 (cited on page 36).

[28] F Goetze. 'On the rate of convergence in the multivariate CLT'. In: *Annals of Probability* (1991), pp. 724–739 (cited on page 36).