Chapter 11

Regression

11.1 Linear Regression

Linear regression is one of the core topics of statistics. It is even more central to econometrics, where a shortage of controlled experiments often leaves econometricians trying to factor heterogeneity out ex post. We begin with a nontraditional view of linear regression based on minimal assumptions.

11.1.1 The Setup

Let's start with the kind of prediction problem discussed in §8.2.2. We study a system with vector input $\mathbf{x}_n \in \mathbb{R}^K$ followed by scalar output y_n. For example,

- $\mathbf{x}_n$ is a description of a lottery (probabilities, possible outcomes, etc.) in a controlled experiment and y_n is willingness to pay in order to participate (see, e.g., Peysakhovich and Naecker 2015).
- $\mathbf{x}_n$ is a set of household characteristics (ethnicity, age, location, etc.) and y_n is household wealth at some later date (see, e.g., McKernan et al. 2014).
- $\mathbf{x}_n$ is price of electricity, prices of alternatives, temperature, household income, and measurements of the regional income distribution, while y_n is regional electricity consumption (see, e.g., Auffhammer and Wolfram 2014).

Although we don't exclude the possibility that y is categorical (i.e., discrete), our loss function will be oriented toward regression (where y takes values in $\mathbb{R}$).

Suppose that we have N observations $\mathbf{z}_n := (\mathbf{x}_n, y_n)$, all of which are draws from some fixed joint distribution P. Since P is fixed, we are assuming that the system is

stationary across the set of draws. Our aim is to predict new output values from input values on the basis of this data. In particular, our problem is to

$$\text{choose a function } f\colon \mathbb{R}^K \to \mathbb{R} \text{ such that } f(\mathbf{x}) \text{ is a good predictor of } y \tag{11.1}$$

To define "good predictor" mathematically, we need a loss function. Throughout this chapter we will be using quadratic loss. Thus, in the language of §8.2.2, our aim is to minimize the prediction risk

$$R(f) := \mathbb{E}_P (y - f(\mathbf{x}))^2 \tag{11.2}$$

As we saw in §5.2.5, the minimizer of (11.2) over the set of all $\mathscr{B}$-measurable functions is the regression function $f^*(\mathbf{x}) := \mathbb{E}_P[y \,|\, \mathbf{x}]$. If we could compute this, then all our problems would be solved. But we cannot compute it because P is not known. Instead we apply the principle of empirical risk minimization (see §8.2.2), which leads to the problem

$$\min_{f \in \mathcal{H}} R_{\text{emp}}(f) \quad \text{where} \quad R_{\text{emp}}(f) := \frac{1}{N} \sum_{n=1}^{N} (y_n - f(\mathbf{x}_n))^2 \tag{11.3}$$

Here $\mathcal{H}$ is the hypothesis space, a set of candidate functions mapping $\mathbb{R}^K$ into $\mathbb{R}$. For obvious reasons, the problem (11.3) is called a **least squares** problem.

If we take $\mathcal{H}$ to be the set of all functions from $\mathbb{R}^K$ to $\mathbb{R}$, then, provided the input vectors are all distinct, we can set the empirical risk $R_{\text{emp}}(f)$ to zero by choosing any function f satisfying $y_n = f(\mathbf{x}_n)$ for all n. However, as discussed at length in §8.2.3, minimizing empirical risk is different from minimizing the prediction risk $R(f)$. The latter is what we actually want to minimize. Thus $\mathcal{H}$ must be restricted.

In this chapter we consider the case $\mathcal{H} = \mathcal{H}_\ell$, where $\mathcal{H}_\ell$ is all linear functions from $\mathbb{R}^K$ to $\mathbb{R}$. Recalling theorem 3.1.1 on page 48, we can write

$$\mathcal{H}_\ell = \left\{ \text{all } f\colon \mathbb{R}^K \to \mathbb{R} \text{ such that } f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{b} \text{ for some } \mathbf{b} \in \mathbb{R}^K \right\} \tag{11.4}$$

The problem (11.3) then reduces to

$$\min_{\mathbf{b} \in \mathbb{R}^K} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\mathsf{T}\mathbf{b})^2 \tag{11.5}$$

The term $\frac{1}{N}$ has been dropped because it does not affect the minimizer. This is the multivariate version of (8.22) on page 227.

The idea of choosing $\mathbf{b}$ to minimize (11.5) is intuitive: We are choosing a "line of best fit" to minimize in-sample prediction error. This optimization problem has a long

tradition. It dates back at least as far as Carl Gauss's work on the orbital position of Ceres, published in 1801.

You might be wondering whether the choice $\mathcal{H} = \mathcal{H}_\ell$ is a suitable one. This is an excellent question. It might not be. However, there are good reasons to start with $\mathcal{H}_\ell$, even in this setting where no linearity assumptions are imposed. First, $\mathcal{H}_\ell$ is a natural starting point when seeking a class of simple, well-behaved functions. Second, as we'll see, setting $\mathcal{H} = \mathcal{H}_\ell$ allows us to obtain an analytical expression for the minimizer, which simplifies both analysis and computation. Third, the technique has an extension from $\mathcal{H}_\ell$ to broader classes of functions, as described in §11.2.1.

11.1.2 The Least Squares Estimator

Now let's solve (11.5). With our knowledge of overdetermined systems (see §3.3.2), we already have all the necessary tools. This will be more obvious after we switch to matrix notation. To do this, let

$$\mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{x}_n := \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nK} \end{pmatrix} = n\text{th observation of all regressors} \tag{11.6}$$

and

$$\mathbf{X} := \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_N^\mathsf{T} \end{pmatrix} :=: \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} \tag{11.7}$$

Sometimes $\mathbf{X}$ is called the **design matrix**. By construction, $\operatorname{col}_k \mathbf{X}$ = all observations on the kth regressor. Also, for any $\mathbf{b} \in \mathbb{R}^K$, we have

$$\mathbf{X}\mathbf{b} = \begin{pmatrix} \mathbf{x}_1^\mathsf{T}\mathbf{b} \\ \mathbf{x}_2^\mathsf{T}\mathbf{b} \\ \vdots \\ \mathbf{x}_N^\mathsf{T}\mathbf{b} \end{pmatrix}$$

It follows that the objective function in (11.5) can be written as

$$\sum_{n=1}^{N} (y_n - \mathbf{x}_n^\mathsf{T}\mathbf{b})^2 = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$$

Since strictly increasing transforms preserve the set of minimizers (see §15.4),

$$\operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^K} \|\mathbf{y}-\mathbf{X}\mathbf{b}\|^2 = \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^K} \|\mathbf{y}-\mathbf{X}\mathbf{b}\| \tag{11.8}$$

We already know how to solve for the minimizer on the right-hand side of (11.8). By theorem 3.3.2 (page 63), the solution is

$$\hat{\boldsymbol{\beta}} := (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \tag{11.9}$$

Traditionally, this random vector $\hat{\boldsymbol{\beta}}$ is called the **least squares estimator**. Once we move to more classical assumptions it will be an estimator of a particular parameter vector. At this stage it just defines our answer to the problem posed in (11.1). That is,

$$\text{given } \mathbf{x}\in\mathbb{R}^K, \text{ our prediction of } y \text{ is } f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\hat{\boldsymbol{\beta}}$$

In terms of geometric interpretation, since $\mathbf{X}\hat{\boldsymbol{\beta}}$ solves (11.8), it is the closest point in colspace $\mathbf{X}$ to $\mathbf{y}$. In particular,

$$\mathbf{P}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} \quad \text{when} \quad \mathbf{P} := \operatorname{proj}(\operatorname{colspace}\mathbf{X})$$

(See (3.13) on page 63.) In what follows, $\mathbf{M}$ is the residual projection, as defined in (2.11) on page 32.

11.1.2.1 Assumptions

Theorem 3.3.2 and our definition of $\hat{\boldsymbol{\beta}}$ in 11.9 require that $\mathbf{X}$ has full column rank (or, equivalently, that the columns of $\mathbf{X}$ are linearly independent—see page 50).

Assumption 11.1.1 $\mathbf{X}$ has full column rank with probability one.

By theorem 2.1.3 on page 20, $N \geqslant K$ is a necessary condition for assumption 11.1.1 to hold. (If $N < K$, then $\mathbb{R}^N$, which is necessarily spanned by N vectors, cannot contain K linearly independent vectors.)

If assumption 11.1.1 fails, then a minimizer of (11.8) still exists but is no longer unique (see ex. 3.5.34). While we can treat this case, it rarely occurs in well-designed regression problems.

Let's also put some mild regularity conditions on the common joint distribution P of each data point $\mathbf{z}_n := (\mathbf{x}_n, y_n)$.

Assumption 11.1.2 P is such that all elements of $\mathbb{E}_P[\mathbf{z}_n\mathbf{z}_n^\mathsf{T}]$ are finite. Moreover

$$\boldsymbol{\Sigma}_\mathbf{x} := \mathbb{E}_P[\mathbf{x}_n\mathbf{x}_n^\mathsf{T}] \text{ is finite and positive definite} \tag{11.10}$$

Finite second moments are imposed because we want to evaluate expected squared errors. This assumption cannot be weakened unless we are willing to work with a different loss function. Positive definiteness of $\boldsymbol{\Sigma}_{\mathbf{x}}$ ensures that the asymptotic limit of our estimator is well defined.[1]

11.1.2.2 Notation

There's a range of standard notation associated with linear least squares estimation. Let's collect it in one place. First, the projection

$$\hat{\mathbf{y}} := \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{y}$$

is called the **vector of fitted values**. The nth fitted value $\hat{y}_n$ is the prediction $\mathbf{x}_n^\intercal \hat{\boldsymbol{\beta}}$ associated with least squares estimate and the nth observation $\mathbf{x}_n$ of the input vector. The vector $\mathbf{M}\mathbf{y}$ is often denoted $\hat{\mathbf{u}}$, and called the **vector of residuals**:

$$\hat{\mathbf{u}} := \mathbf{M}\mathbf{y} = \mathbf{y} - \hat{\mathbf{y}}$$

The vector of residuals corresponds to the error that occurs when $\mathbf{y}$ is approximated by $\mathbf{P}\mathbf{y}$. From fact 2.2.8 on page 33 we have

$$\mathbf{M}\mathbf{y} \perp \mathbf{P}\mathbf{y} \quad \text{and} \quad \mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y} \tag{11.11}$$

In other words, $\mathbf{y}$ can be decomposed into two orthogonal vectors $\mathbf{P}\mathbf{y}$ and $\mathbf{M}\mathbf{y}$, where the first represents the best approximation to $\mathbf{y}$ in colspace $\mathbf{X}$, and the second represents the residual.

Related to the fitted values and residuals, we have some standard definitions:

- **Total sum of squares** $:=$ TSS $:= \|\mathbf{y}\|^2$.
- **Residual sum of squares** $:=$ RSS $:= \|\mathbf{M}\mathbf{y}\|^2$.
- **Explained sum of squares** $:=$ ESS $:= \|\mathbf{P}\mathbf{y}\|^2$.

By (11.11) and the Pythagorean law (page 27),

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{11.12}$$

When running regressions it is conventional to report the **coefficient of determination**, or R^2. The plain vanilla definition of R^2 is

$$R^2 := \frac{\text{ESS}}{\text{TSS}} \tag{11.13}$$

1. In essence, positive definiteness of $\boldsymbol{\Sigma}_{\mathbf{x}}$ requires that no random variable in $\mathbf{x}$ can be written as a linear combination of other variables in $\mathbf{x}$. See exercise 11.4.1.

Many regression packages report an alternative definition of R^2. See §11.2.3 below.

11.1.3 Out-of-Sample Fit

We have stressed a number of times that learning from data (statistics) means generalization from current observations to new ones. As such, the most important measure of success for a statistical procedure is out-of-sample fit. So how does linear least squares perform out-of-sample? We start with a general observation about linear predictors.

Theorem 11.1.1 *If ℓ is the linear function $\ell(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{b}$, then*

$$R(\ell) = \mathbb{E}(y - f^*(\mathbf{x}))^2 + \mathbb{E}(f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*)^2 + (\mathbf{b}^* - \mathbf{b})^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{x}(\mathbf{b}^* - \mathbf{b})$$

Here f^* is the regression function and $\mathbf{b}^* = {\boldsymbol{\Sigma}_\mathbf{x}}^{-1}\,\mathbb{E}[\mathbf{x}\,y]$ is the vector of coefficients in the best linear predictor (see page 147). $R(f)$ is the prediction risk of f and expectations are taken under the unknown joint distribution P of the pairs $(\mathbf{x}, y)$.

The proof of theorem 11.1.1 is given in §11.1.3.1. For now let's look at interpretation. The general question is how well we can generalize (i.e., reduce prediction risk) using linear functions. Theorem 11.1.1 decomposes the prediction risk of an arbitrary linear predictor $\ell(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{b}$ into three terms:

(i) The **intrinsic risk** $\mathbb{E}(y - f^*(\mathbf{x}))^2$.

(ii) The **approximation error** $\mathbb{E}(f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*)^2$.

(iii) The **estimation error** $(\mathbf{b}^* - \mathbf{b})^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{x}(\mathbf{b}^* - \mathbf{b})$.

The intrinsic risk is also called Bayes risk (see example 8.2.3 on page 226). It is the residual error after y is approximated with the best possible predictor (i.e., the regression function). It is large to the extent that y is hard to predict using $\mathbf{x}$.

The approximation error or *bias* is the deviation between the best predictor and the best linear predictor. It reflects the cost of our decision to approximate the regression function using a linear architecture. If this architecture is held fixed, the approximation error is also fixed and cannot be reduced in the estimation process.

The estimation error is caused by the deviation of our estimator from the best linear predictor $\mathbf{b}^*$. This deviation occurs because we are predicting using finite sample information on the joint distribution of $(\mathbf{x}, y)$.

Theorem 11.1.1 tells us that once $\mathcal{H}$ is set to the class of linear functions, the best we can do is find an estimation method (a learning algorithm) that produces an estimate that is close to $\mathbf{b}^*$ on average when the sample size is sufficiently large. The next result states that the least squares estimator $\hat{\boldsymbol{\beta}}$ has this property.

Theorem 11.1.2 *Let assumptions 11.1.2–11.1.1 hold and let $\hat{\boldsymbol{\beta}}_N$ be the least squares estimator given sample size N. If the observations $\{\mathbf{z}_n\}$ are independent, then*

$$\hat{\boldsymbol{\beta}}_N \xrightarrow{p} \mathbf{b}^* \quad as \quad N \to \infty \tag{11.14}$$

The proof is below. Independence is required only for the LLN to function. We could weaken this to ergodicity (see §7.1.1) and obtain the same conclusion.[2]

On one level, theorems 11.1.1 and 11.1.2 are reassuring. They tell us that if the underlying process is relatively linear, then we will attain small risk asymptotically. On the other hand, the Glivenko–Cantelli theorem tells us that we can learn everything about the underlying distribution in the limit. Here we are bounded away from this kind of consistency whenever the approximation error is positive.

As will be discussed in chapter 14, a general principle of induction is that we should introduce bias in finite samples to avoid overfitting, while at the same time reducing bias asymptotically, as the empirical distribution converges to the true distribution. By comparison, in standard linear regression the bias is held fixed by the linearity assumption.

11.1.3.1 Proofs

Proof of theorem 11.1.1. Fix $\mathbf{b} \in \mathbb{R}^K$ and let $\ell(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{b}$. In view of (8.17) on page 226, we have the prediction risk

$$R(\ell) = \mathbb{E}[(y - f^*(\mathbf{x}))^2] + \mathbb{E}[(f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b})^2]$$

Hence the result will established if we can show that

$$\mathbb{E}[(f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b})^2] = \mathbb{E}[(f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*)^2] + \mathbb{E}[(\mathbf{b}^* - \mathbf{b})^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}(\mathbf{b}^* - \mathbf{b})] \tag{11.15}$$

To see that (11.15) holds, observe that

$$f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b} = f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^* + \mathbf{x}^\mathsf{T}(\mathbf{b}^* - \mathbf{b}) \tag{11.16}$$

The terms $f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*$ and $\mathbf{x}^\mathsf{T}(\mathbf{b}^* - \mathbf{b})$ are orthogonal. The reason is that $\mathbf{x}^\mathsf{T}\mathbf{b}^*$ is the orthogonal projection of $f^*(\mathbf{x})$ onto $S = \text{span}\{\mathbf{x}\}$, the linear subspace of L_2 spanned by all linear combinations of the form $\mathbf{a}^\mathsf{T}\mathbf{x}$. (See ex. 5.4.18 on page 156.) As such, $f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*$ is orthogonal to every element of the target subspace $\text{span}\{\mathbf{x}\}$. This includes $\mathbf{x}^\mathsf{T}(\mathbf{b}^* - \mathbf{b})$.

For any orthogonal elements u and v of L_2 we have $\mathbb{E}[(u+v)^2] = \mathbb{E}[u^2] + \mathbb{E}[v^2]$.

2. Later, in §13.1, we'll do something similar (i.e., weaken independence to ergodicity) in a setting with some additional structure.

(This is the Pythagorean law in L_2.) Squaring both sides of (11.16), taking expectations and applying this law gives (11.15). The proof of theorem 11.1.1 is done. □

Proof of theorem 11.1.2. The proof of theorem 11.1.2 is not hard if we express $\hat{\boldsymbol{\beta}}_N$ in a slightly different way. Multiplying and dividing by N in the definition of $\hat{\boldsymbol{\beta}}_N$ and then expanding out the matrix products (see ex. 11.4.9) gives

$$\hat{\boldsymbol{\beta}}_N = \left[\frac{1}{N}\mathbf{X}^\mathsf{T}\mathbf{X}\right]^{-1} \cdot \frac{1}{N}\mathbf{X}^\mathsf{T}\mathbf{y} = \left[\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\mathsf{T}\right]^{-1} \cdot \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n y_n \tag{11.17}$$

By the matrix LLN in fact 6.2.3 (page 173), we have

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\mathsf{T} \xrightarrow{p} \boldsymbol{\Sigma}_\mathbf{x} \quad \text{and} \quad \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n y_n \xrightarrow{p} \mathbb{E}[\mathbf{x}y] \quad \text{as} \quad N \to \infty$$

By fact 6.2.1 on page 170, convergence in probability is preserved over the taking of inverses and products. Hence $\hat{\boldsymbol{\beta}}_N \xrightarrow{p} \boldsymbol{\Sigma}_\mathbf{x}^{-1}\mathbb{E}[\mathbf{x}\,y] = \mathbf{b}^*$, as was to be shown. □

11.1.4 In-Sample Fit

In-sample fit measures how well a given model fits the same data set that it was estimated on. The difference between in-sample fit (empirical risk) and out-of-sample fit (risk) was discussed in §8.2.3. In-sample fit of a regression is often measured with R^2 (see (11.13)). Let's make some further comments on R^2 and then discuss how R^2 relates to in-sample fit.

Fact 11.1.1 $0 \leqslant R^2 \leqslant 1$ with $R^2 = 1$ if and only if $\mathbf{y} \in \operatorname{colspace}\mathbf{X}$.

That $R^2 \leqslant 1$ is immediate from $\|\mathbf{P}\mathbf{y}\| \leqslant \|\mathbf{y}\|$ (cf. theorem 2.2.2 on page 31). Exercise 11.4.17 asks you to prove the second claim. More generally, a high R^2 indicates $\mathbf{y}$ is relatively close to $\operatorname{colspace}\mathbf{X}$. This fact suggests that we can increase R^2 at least weakly by adding regressors. As we do so the column space of $\mathbf{X}$ expands, pushing out towards $\mathbf{y}$. Here's a formal statement:

Fact 11.1.2 Let $\mathbf{X}_a$ and $\mathbf{X}_b$ be two design matrices. If R_a^2 and R_b^2 are the respective coefficients of determination, then

$$\operatorname{colspace}\mathbf{X}_a \subset \operatorname{colspace}\mathbf{X}_b \quad \Longrightarrow \quad R_a^2 \leqslant R_b^2$$

For a proof, see exercise 11.4.8 and its solution.

High R^2 is sometimes equated with successful regression. This is a misunderstanding of the aim of statistics. The correct definition of statistical learning is effective generalization from existing data. In the present context this means that the linear predictor produced by regression attains low risk.

So how does R^2 relate to risk? What R^2 actually measures is the degree to which empirical risk is minimized. To see this, note that

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - N\frac{R_{\text{emp}}(\hat{f})}{\text{TSS}}$$

where R_{emp} is as defined in (11.3) and $\hat{f}$ is our linear predictor $\hat{f}(\mathbf{x}) = \mathbf{x}^\intercal\hat{\boldsymbol{\beta}}$. Thus high R^2 means low empirical risk and good in-sample fit. But low empirical risk is no guarantee of low prediction risk, as was emphasized in §8.2.3.

Here's a simulation that shows how we can produce high R^2 without estimating anything meaningful. We take x_n and y_n as independent draws from a uniform distribution on $[0, 1]$. By construction, there is no relationship between these two variables. For the regressors we take the powers $1, x, x^2, \ldots, x^K$, where K is a positive integer. The R code below runs these regressions for different values of K. At $K = 25$ the value of R^2 is around 0.95. This is despite the fact that no relationship exists between x and y.

```
set.seed(1234)
N <- 25
y <- runif(N)
x <- runif(N)
X <- rep(1, N)

Kmax <- 25
for (K in 1:Kmax) {
    X <- cbind(X, x^K)
    results <- lm(y ~ 0 + X)
    Py2 <- sum(results$fitted.values^2)
    y2 <- sum(y^2)
    cat("K =", K, "R^2 =", Py2 / y2, "\n")
}
```

(You can obtain all code from the text at `johnstachurski.net/emet.html`.)

To finish this section, let's draw a connection between fact 11.1.2, which says that the value of R^2 is at least weakly increasing in the number of right-hand side variables, and fact 8.2.1 on page 229. Suppose that $\mathbf{x}$ lists a large number of possible regressors.

Let the hypothesis space be

$$\mathcal{H}_j := \left\{ \text{all } f\colon \mathbb{R}^j \to \mathbb{R} \text{ s.t. } f(\mathbf{x}) = \mathbf{x}^\intercal \mathbf{b} \text{ for some } \mathbf{b} \in \mathbb{R}^j \right\} \tag{11.18}$$

Here $1 \leqslant j \leqslant K$. Empirical risk minimization over $\mathcal{H}_j$ is equivalent to linear regression over the first j regressors. Empirical risk falls as j increases by fact 8.2.1 on page 229. Hence R^2 increases. This is the same conclusion as fact 11.1.2.

11.2 The Geometry of Least Squares

In this section we cover transformations of the data and an important theorem on subsets of the least squares estimator.

11.2.1 Transformations and Basis Functions

In discussing the decision to set $\mathcal{H} = \mathcal{H}_\ell$ in §11.1.1, we mentioned that we can use many of the same ideas when extending $\mathcal{H}$ to a broader class of functions. The idea is to first transform the data using some arbitrary function $\boldsymbol{\phi}\colon \mathbb{R}^K \to \mathbb{R}^J$. The action of $\boldsymbol{\phi}$ on $\mathbf{x} \in \mathbb{R}^K$ is

$$\mathbf{x} \mapsto \boldsymbol{\phi}(\mathbf{x}) = \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \vdots \\ \phi_J(\mathbf{x}) \end{pmatrix} \in \mathbb{R}^J$$

The individual functions $\phi_1, \ldots, \phi_J$ mapping $\mathbb{R}^K$ into $\mathbb{R}$ are sometimes called **basis functions**. In machine learning texts, the range of $\boldsymbol{\phi}$ is called **feature space**. Linear least squares is now applied in feature space. That is, we solve the empirical risk minimization problem when the hypothesis space is

$$\mathcal{H}_{\boldsymbol{\phi}} := \{\text{all functions } \ell \circ \boldsymbol{\phi}, \text{ where } \ell \text{ is a linear function from } \mathbb{R}^J \text{ to } \mathbb{R}\}$$

The empirical risk minimization problem is then

$$\min_{\ell} \sum_{n=1}^{N} \{y_n - \ell(\boldsymbol{\phi}(\mathbf{x}_n))\}^2 = \min_{\boldsymbol{\gamma} \in \mathbb{R}^J} \sum_{n=1}^{N} (y_n - \boldsymbol{\gamma}^\intercal \boldsymbol{\phi}(\mathbf{x}_n))^2 \tag{11.19}$$

Switching to matrix notation, if

$$\boldsymbol{\Phi} := \begin{pmatrix} \phi_1(\mathbf{x}_1) & \cdots & \phi_J(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \cdots & \phi_J(\mathbf{x}_2) \\ \vdots & \cdots & \vdots \\ \phi_1(\mathbf{x}_N) & \cdots & \phi_J(\mathbf{x}_N) \end{pmatrix} \in \mathbb{R}^{N \times J} \tag{11.20}$$

then the objective in (11.19) can be expressed as $\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\gamma}\|^2$. Since increasing functions don't affect minimizers, the problem becomes

$$\operatorname*{argmin}_{\boldsymbol{\gamma} \in \mathbb{R}^J} \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\gamma}\| \tag{11.21}$$

Assuming that $\boldsymbol{\Phi}$ is full column rank, the solution is

$$\hat{\boldsymbol{\gamma}} := (\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\mathsf{T}\mathbf{y}$$

Example 11.2.1 Adding an intercept to a regression can be regarded as a transformation of the data. Indeed adding an intercept is equivalent to applying the transformation

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_K \end{pmatrix}$$

In practice, adding an intercept means fitting an extra parameter, and this extra degree of freedom allows a more flexible fit in our regression.

Example 11.2.2 Let $K = 1$, so that $x_n \in \mathbb{R}$. Consider the mononomial basis functions $\phi_j(x) := x^{j-1}$, so that

$$\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{\phi}(x_n) = \boldsymbol{\gamma}^\mathsf{T} \begin{pmatrix} x_n^0 \\ x_n^1 \\ \vdots \\ x_n^{J-1} \end{pmatrix} = \sum_{j=1}^{J} \gamma_j x_n^{j-1} \tag{11.22}$$

The mononomial basis transformation applied to scalar x corresponds to univariate polynomial regression, as discussed in §8.2.3. Under this transformation, the matrix $\boldsymbol{\Phi}$ in (11.20) is called the **Vandermonde matrix**. By the Weierstrass approximation theorem, polynomials of sufficiently high order can effectively approximate any one-dimensional continuous nonlinear relationship.

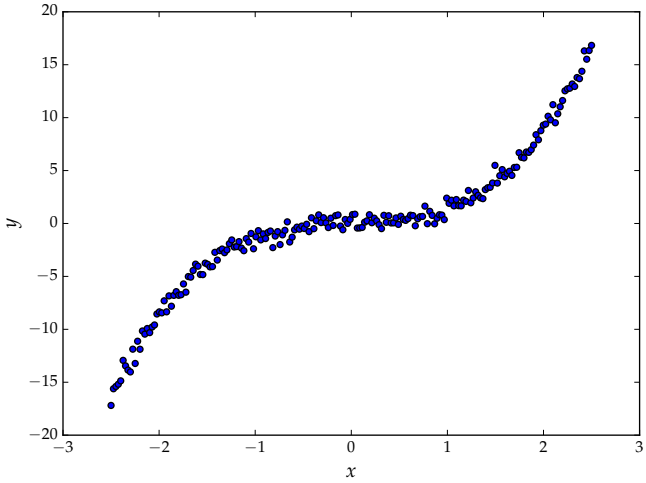

Figure 11.1 Nonlinear relationship between x and y

Example 11.2.3 Example 11.2.2 uses mononomials as the basis functions. A common alternative is to use orthogonal polynomials such as Chebychev polynomials or Hermite polynomials. Other alternatives include wavelets and splines. In econometrics this procedure is often referred to as nonparametric series regression. A key topic is the optimal number of basis functions.[3]

Figures 11.1 and 11.2 help illustrate how transformations can reduce approximation error. In figure 11.1 it is clear that no linear function mapping x to y can produce small approximation error. Figure 11.2 shows the data after applying the transformation $\mathbb{R} \ni x \mapsto \boldsymbol{\phi}(x) := (x, x^3)' \in \mathbb{R}^2$. The plane drawn in figure 11.2 represents a linear function $\ell \colon \mathbb{R}^2 \to \mathbb{R}$. The composition $\ell \circ \boldsymbol{\phi}$ has low approximation error. The two figures illustrate how nonlinear data can become linear when projected into higher dimensions.

Below, when y is regressed on $\mathbf{x}$, we can imagine that the data have already been transformed, and $\mathbf{x}$ is the result. Hence we use $\mathbf{X}$ to denote the design matrix instead of Φ without loss of generality.

3. See, for example, Hong and White (1995), Sun (2011), or Chen and Christensen (2015).

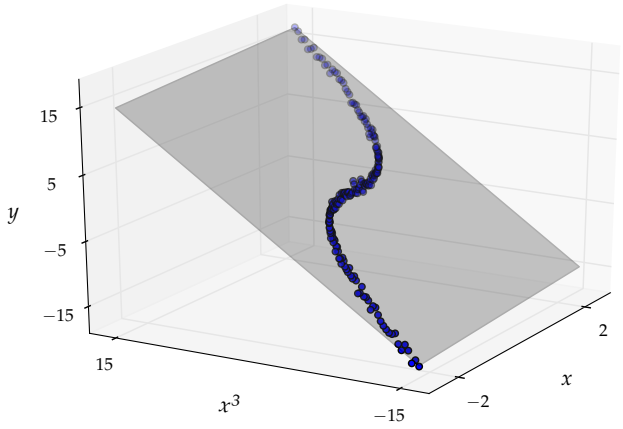

Figure 11.2 Approximate linearity after projecting the data to $\mathbb{R}^2$

11.2.2 The Frisch–Waugh–Lovell Theorem

The Frisch–Waugh–Lovell (FWL) theorem yields an expression for an arbitrary subvector of the least squares estimator $\hat{\boldsymbol{\beta}}$ obtained by regressing $\mathbf{y}$ on $\mathbf{X}$. In stating the theorem, we continue with the assumptions of §11.1. Let $\mathbf{y}$ and $\mathbf{X}$ be given and let $\hat{\boldsymbol{\beta}}$ be as in (11.9). In addition, let K_1 be an integer with $1 \leqslant K_1 < K$, and let

- $\mathbf{X}_1$ be a matrix consisting of the first K_1 columns of $\mathbf{X}$,
- $\mathbf{X}_2$ be a matrix consisting of the remaining $K_2 := K - K_1$ columns,
- $\hat{\boldsymbol{\beta}}_1$ be the $K_1 \times 1$ vector consisting of the first K_1 elements of $\hat{\boldsymbol{\beta}}$.
- $\hat{\boldsymbol{\beta}}_2$ be the $K_2 \times 1$ vector consisting of the remaining K_2 elements of $\hat{\boldsymbol{\beta}}$,
- $\mathbf{P}_1 := \text{proj}(\text{colspace}\, \mathbf{X}_1)$, and
- $\mathbf{M}_1 := \mathbf{I} - \mathbf{P}_1 =$ the corresponding residual projection (see page 32).

Theorem 11.2.1 (FWL theorem) *The vector* $\hat{\boldsymbol{\beta}}_2$ *satisfies*

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{y} \tag{11.23}$$

Proof. From (11.11) and the definitions above we have

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{M}\mathbf{y} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \mathbf{M}\mathbf{y}$$

Premultiplying both sides of this expression by $\mathbf{X}_2^\intercal \mathbf{M}_1$, we obtain

$$\mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{y} = \mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 + \mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{M} \mathbf{y} \tag{11.24}$$

The first term on the right-hand side is zero by fact 3.3.1 on page 61. The last term is also zero because

$$(\mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{M} \mathbf{y})^\intercal = \mathbf{y}^\intercal \mathbf{M}^\intercal \mathbf{M}_1^\intercal \mathbf{X}_2 = \mathbf{y}^\intercal \mathbf{M} \mathbf{M}_1 \mathbf{X}_2 = \mathbf{y}^\intercal \mathbf{M} \mathbf{X}_2 = \mathbf{0}$$

In the first equality, we used the usual property of transposes (fact 3.2.4); in the second, we used symmetry of $\mathbf{M}$ and $\mathbf{M}_1$ (see page 61); in the third, we used fact 2.2.9 on page 34; and in the fourth, we used fact 3.3.1 again (which tells us that $\mathbf{M}$ maps all columns of $\mathbf{X}$, and hence all columns of $\mathbf{X}_2$, to the zero vector).

In light of the above, (11.24) becomes $\mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{y} = \mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2$. To go from this equation to (11.23), we just need to check that the matrix premultiplying $\hat{\boldsymbol{\beta}}_2$ is nonsingular. The proof is left as an exercise (ex. 11.4.23). □

As exercise 11.4.22 asks you to show, the expression for $\hat{\boldsymbol{\beta}}_2$ in theorem 11.2.1 can be rewritten as

$$\hat{\boldsymbol{\beta}}_2 = [(\mathbf{M}_1 \mathbf{X}_2)^\intercal \mathbf{M}_1 \mathbf{X}_2]^{-1} (\mathbf{M}_1 \mathbf{X}_2)^\intercal \mathbf{M}_1 \mathbf{y} \tag{11.25}$$

Close inspection of this formula confirms the following claim: There is another way to obtain $\hat{\boldsymbol{\beta}}_2$ besides just regressing $\mathbf{y}$ on $\mathbf{X}$ and then extracting the last K_2 elements: we can also regress $\mathbf{M}_1 \mathbf{y}$ on $\mathbf{M}_1 \mathbf{X}_2$ to produce the same result.

To get some feeling for what this means, let's look at a special case, where $\mathbf{X}_2$ is the single column $\operatorname{col}_K \mathbf{X}$, containing the observations on the Kth regressor. To tie into this notation let's write $\mathbf{X}_1$ as $\mathbf{X}_{-K}$ to remind us that it stands for all columns of $\mathbf{X}$ expect the Kth one, and similarly for $\mathbf{M}_1$. In view of the preceding discussion, the least squares estimate $\hat{\beta}_K$ can be found by regressing

$$\tilde{\mathbf{y}} := \mathbf{M}_{-K} \mathbf{y} = \text{residuals of regressing } \mathbf{y} \text{ on } \mathbf{X}_{-K} \tag{11.26}$$

on

$$\tilde{\mathbf{x}}_K := \mathbf{M}_{-K} \operatorname{col}_K \mathbf{X} = \text{residuals of regressing } \operatorname{col}_K \mathbf{X} \text{ on } \mathbf{X}_{-K} \tag{11.27}$$

Loosely speaking, these two residual terms $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}_K$ can be thought of as the parts of $\mathbf{y}$ and $\operatorname{col}_K \mathbf{X}$ that are "not explained by" $\mathbf{X}_{-K}$. Thus, on an intuitive level, the process for obtaining the least squares estimate $\hat{\beta}_K$ is as follows:

(i) Remove effects of all other regressors from $\mathbf{y}$ and $\operatorname{col}_K \mathbf{X}$, producing $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}_K$.

(ii) Regress $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{x}}_K$.

This is obviously different from the process for obtaining the coefficient of the vector

$\text{col}_K \mathbf{X}$ in a simple univariate regression, the latter being just

(i) Regress $\mathbf{y}$ on $\text{col}_K \mathbf{X}$.

In words, the difference between the univariate least squares estimated coefficient of the Kth regressor and the multiple regression least squares coefficient is that the multiple regression coefficient $\hat{\beta}_K$ measures the *isolated relationship* between x_K and y, without taking into account indirect channels involving other variables.

We can illustrate this idea further with a small simulation. Suppose that

$$y = x_1 + x_2 + u \quad \text{where} \quad u \overset{\text{IID}}{\sim} \text{N}(0, 1)$$

If we generate N independent observations from this model and regress y on the observations of (x_1, x_2), then, provided that N is sufficiently large, the coefficients for x_1 and x_2 will both be close to unity (see theorem 11.1.2). However, if we regress y on x_1 alone, then the coefficient for x_1 will depend on the relationship between x_1 and x_2. For example:

```
> N <- 1000
> x1 <- runif(N)
> x2 = 10 * exp(x1) + rnorm(N)
> y <- x1 + x2 + rnorm(N)
> results <- lm(y ~ 0 + x1)
> results$coefficients
      x1
      30.83076
```

Here the coefficient for x_1 is much larger than unity, because an increase in x_1 tends to have a large positive effect on x_2, which in turn increases y. The coefficient in the univariate regression reflects this total effect.

11.2.2.1 Application: Simple Regression

Here's an easy application of the FWL theorem: deriving the familiar expression for the slope coefficient in simple linear regression (see §8.2.1.1) from the multivariate expression. Simple linear regression is a special case of multivariate regression when $\mathbf{1}$ is the first column of $\mathbf{X}$ and $K = 2$. In this section, the second column of $\mathbf{X}$ will be denoted by $\mathbf{x}$. As we saw in (8.14) on page 224, the least squares estimates are

$$\hat{\beta}_2 = \frac{\sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^N (x_n - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

where $\bar{x}$ is the sample mean of $\mathbf{x}$ and $\bar{y}$ is the sample mean of $\mathbf{y}$. We can rewrite the slope coefficient $\hat{\beta}_2$ more succinctly as

$$\hat{\beta}_2 = [(\mathbf{x} - \bar{x}\mathbf{1})^\mathsf{T}(\mathbf{x} - \bar{x}\mathbf{1})]^{-1}(\mathbf{x} - \bar{x}\mathbf{1})^\mathsf{T}(\mathbf{y} - \bar{y}\mathbf{1}) \tag{11.28}$$

By the FWL theorem (equation 11.25), we also have

$$\hat{\beta}_2 = [(\mathbf{M}_c\mathbf{x})^\mathsf{T}\mathbf{M}_c\mathbf{x}]^{-1}(\mathbf{M}_c\mathbf{x})^\mathsf{T}\mathbf{M}_c\mathbf{y} \tag{11.29}$$

where $\mathbf{M}_c$ is the residual projection associated with the linear subspace $S = \text{span}\{\mathbf{1}\}$, as defined in (3.10) on page 61. For this residual projection $\mathbf{M}_c$ and any $\mathbf{z}$, we have $\mathbf{M}_c\mathbf{z} = \mathbf{z} - \bar{z}\mathbf{1}$. Hence the right-hand sides of (11.28) and (11.29) coincide.

11.2.3 Centered Observations

Let's generalize the preceding discussion to the case where there are multiple nonconstant regressors. The only difference is that instead of one column $\mathbf{x}$ of observations on a single nonconstant regressor, we have a matrix $\mathbf{X}_2$ containing multiple columns, each a vector of observations on a nonconstant regressor. If the least squares estimate $\hat{\boldsymbol{\beta}}$ is partitioned into $(\hat{\beta}_1, \hat{\boldsymbol{\beta}}_2)$, then we can write

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{1}\beta_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2$$

Applying the FWL theorem (equation 11.25) once more, we can write $\hat{\boldsymbol{\beta}}_2$ as

$$\hat{\boldsymbol{\beta}}_2 = [(\mathbf{M}_c\mathbf{X}_2)^\mathsf{T}\mathbf{M}_c\mathbf{X}_2]^{-1}(\mathbf{M}_c\mathbf{X}_2)^\mathsf{T}\mathbf{M}_c\mathbf{y}$$

where $\mathbf{M}_c$ is the residual projection in (3.10). As we saw in the last section, $\mathbf{M}_c\mathbf{y}$ is $\mathbf{y}$ centered around its mean. Similarly, $\mathbf{M}_c\mathbf{X}_2$ is a matrix formed by taking each column of $\mathbf{X}_2$ and centering it around its mean. It follows that in a least squares regression with an intercept, the estimated coefficients of the nonconstant regressors are equal to the estimated coefficients of a zero-intercept regression performed after all variables have been centered around their mean.

Let's use some related ideas to discuss an alternative to the coefficient of determination introduced in (11.13). There are several versions of R^2 reported in common regression packages. One of these is so called centered R^2. The version in (11.13) will henceforth be called the uncentered R^2 for clarity.

One motivation for introducing an alternative to uncentered R^2 is that it fails to be invariant to certain changes of units. While it is invariant to changes of units that involve rescaling of the regressand $\mathbf{y}$ (see ex. 11.4.2), it is not invariant to changes of units that involve addition or subtraction (actual inflation versus inflation in excess

of a certain level, income versus income over a certain threshold, etc.) whenever $\mathbf{X}$ contains an intercept. Exercise 11.4.3 asks you to prove this.

This is one reason many econometricians use the **centered** R^2 rather than R^2, at least when the regression contains an intercept. For the purposes of this section, let's assume that this is the case (or, more generally, that $\mathbf{1} \in \operatorname{colspace} \mathbf{X}$). Centered R^2 is defined as

$$R_c^2 := \frac{\|\mathbf{P}\mathbf{M}_c\mathbf{y}\|^2}{\|\mathbf{M}_c\mathbf{y}\|^2} = \frac{\|\mathbf{M}_c\mathbf{P}\mathbf{y}\|^2}{\|\mathbf{M}_c\mathbf{y}\|^2} \tag{11.30}$$

$\mathbf{M}_c$ is as defined in (3.10) on page 61. The equality of the two expressions for R_c^2 is left as an exercise (ex. 11.4.6). Adding a constant to each element of $\mathbf{y}$ will have no effect on R_c^2 because $\mathbf{M}_c$ maps constant vectors to $\mathbf{0}$ (see example 3.3.1).

Centered R^2 can be rewritten (ex. 11.4.7) as

$$R_c^2 = \frac{\sum_{n=1}^N (\hat{y}_n - \bar{y})^2}{\sum_{n=1}^N (y_n - \bar{y})^2} \tag{11.31}$$

It is a further exercise (ex. 11.4.5) to show that, in the case of the simple regression, the R_c^2 is equal to the square of the sample correlation between the regressor and regressand, as defined in (8.5) on page 218. Thus R_c^2 is a measure of correlation.

11.3 Further Reading

For additional references on the material covered in this chapter see, for example, Friedman et al. (2009), Ruud (2000), Cameron and Trivedi (2005), or Davidson and MacKinnon (2004).

11.4 Exercises

Ex. 11.4.1 Let $\mathbf{x} := (z, az)^\top$, where $a \in \mathbb{R}$ and z is any scalar random variable. Show that $\mathbb{E}\mathbf{x}\mathbf{x}^\top$ is nonnegative definite but fails to be positive definite.

Ex. 11.4.2 Show that uncentered R^2 is invariant to changes of units that involve rescaling of the regressand $\mathbf{y}$ (dollars versus cents, kilometers versus miles, etc.)

Ex. 11.4.3 Fix $\mathbf{X}, \mathbf{y}$ and consider regressing $\mathbf{y}$ on $\mathbf{X}$. Suppose that $\mathbf{X}$ contains the intercept, in the sense that $\mathbf{1} \in \operatorname{colspace} \mathbf{X}$. Let R^2 represent the uncentered coefficient of determination. Let R_α^2 represent the same when $\mathbf{y}$ is replaced by $\mathbf{y} + \alpha\mathbf{1}$. Show that $R_\alpha^2 \to 1$ as $\alpha \to \infty$.

Ex. 11.4.4 Show that R^2 is invariant to a rescaling of the regressors.

Ex. 11.4.5 Show that, in the case of the simple regression model in §11.2.2.1, R_c^2 is equal to the square of the sample correlation between $\mathbf{x}$ and $\mathbf{y}$.

Ex. 11.4.6 Confirm the equality of the two alternative expressions for R_c^2 in (11.30).

Ex. 11.4.7 Verify the expression for R_c^2 in (11.31).

Ex. 11.4.8 Prove fact 11.1.2 on page 306.

Ex. 11.4.9 Verify expression (11.17).

Ex. 11.4.10 Let's show that $\hat{\boldsymbol{\beta}}$ solves the least squares problem in a slightly different way: Let $\mathbf{b}$ be any $K \times 1$ vector, and let $\hat{\boldsymbol{\beta}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.

(i) Show that $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})\|^2$.

(ii) Using (i), argue that $\hat{\boldsymbol{\beta}}$ is the minimizer of $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$ over all $K \times 1$ vectors $\mathbf{b}$.

Ex. 11.4.11 Verify that $\sum_{n=1}^N (y_n - \mathbf{b}^\top \mathbf{x}_n)^2 = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$.

Ex. 11.4.12 Show carefully that any solution to $\min_{\mathbf{b} \in \mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$ is also a solution to $\min_{\mathbf{b} \in \mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|$, and vice versa.

Ex. 11.4.13 Confirm that $\mathbf{P}\mathbf{1} = \mathbf{1}$ whenever $\mathbf{1} \in \operatorname{colspace} \mathbf{X}$.

Ex. 11.4.14 Show that, for any regression containing the intercept, the vector of residuals must sum to zero.

Ex. 11.4.15 Show that, for any regression containing the intercept, the mean of the fitted values $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$ is equal to the mean of $\mathbf{y}$.

Ex. 11.4.16 Show that $\mathbf{P}\mathbf{M} = \mathbf{M}\mathbf{P} = \mathbf{0}$. Using this fact (instead of the orthogonal projection theorem), show that the vector of fitted values and the vector of residuals are orthogonal.

Ex. 11.4.17 Show that if $R^2 = 1$, then the vector of residuals is identically zero, $\mathbf{P}\mathbf{y} = \mathbf{y}$, and $\mathbf{y} \in \operatorname{colspace} \mathbf{X}$.

Ex. 11.4.18 Suppose that the regression contains an intercept. Let $\bar{y}$ be the sample mean of $\mathbf{y}$, and let $\bar{\mathbf{x}}$ be a $1 \times K$ row vector such that the kth element of $\bar{\mathbf{x}}$ is the sample mean of the kth column of $\mathbf{X}$. Show that $\bar{y} = \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}$.

Ex. 11.4.19 Suppose the regression contains an intercept. Let $\mathbf{M}_c$ be as defined in (3.10). Show that the following identity always holds:

$$\|\mathbf{M}\mathbf{y}\|^2 = \|\mathbf{M}_c \mathbf{y}\|^2 - \|\mathbf{P}\mathbf{M}_c \mathbf{y}\|^2 \tag{11.32}$$

Ex. 11.4.20 Let $\mathbf{X}_a$ and $\mathbf{X}_b$ be $N \times K_a$ and $N \times K_b$ respectively. Suppose that every column of $\mathbf{X}_a$ is also a column of $\mathbf{X}_b$. Show that colspace $\mathbf{X}_a \subset$ colspace $\mathbf{X}_b$.

Ex. 11.4.21 Let $\mathbf{x} := (x_1, \ldots, x_N)$ and $\mathbf{y} := (y_1, \ldots, y_N)$ be sequences of scalar random variables. Show that the sample correlation $\hat{\varrho}$ between $\mathbf{x}$ and $\mathbf{y}$ (defined in (8.5) on page 218) can be written as

$$\hat{\varrho} = \frac{(\mathbf{M}_c \mathbf{x})^\top (\mathbf{M}_c \mathbf{y})}{\|\mathbf{M}_c \mathbf{x}\| \|\mathbf{M}_c \mathbf{y}\|}$$

Ex. 11.4.22 Show that the two expressions for $\hat{\boldsymbol{\beta}}_2$ in (11.23) and (11.25) are equal.[4]

Ex. 11.4.23 At the end of the proof of theorem 11.2.1, it was claimed that the matrix $\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2$ is nonsingular. Verify this claim.

Ex. 11.4.24 (Computational) Build an arbitrary data set $\mathbf{X}$, $\mathbf{y}$ by simulation. Run a regression with the intercept, and record the values of the estimated coefficients of the nonconstant (i.e., $k \geqslant 2$) regressors. Confirm that these values are equal to the estimated coefficients of the no-intercept regression after all variables have been centered around their mean.

11.4.1 Solutions to Selected Exercises

Solution to Ex. 11.4.1. The expression $\mathbb{E}\,\mathbf{x}\mathbf{x}^\top$ is nonnegative definite for any $\mathbf{x}$ because, given $\mathbf{c} \in \mathbb{R}^K$, we have

$$\mathbf{c}^\top \mathbb{E}\,\mathbf{x}\mathbf{x}^\top \mathbf{c} = \mathbb{E}(\mathbf{c}^\top \mathbf{x})(\mathbf{x}^\top \mathbf{c}) = \mathbb{E}(\mathbf{x}^\top \mathbf{c})^2 \geqslant 0$$

Here we used the fact that the transpose of a scalar is equal to the scalar. However, if $\mathbf{x} := (z, az)^\top$, then

$$\mathbb{E}\,\mathbf{x}\mathbf{x}^\top = \mathbb{E}z^2 \begin{pmatrix} 1 & a \\ a & a^2 \end{pmatrix}$$

The second column is a multiple of the first, so the matrix is singular and the determinant is zero. As such it cannot be positive definite (see fact 3.2.9 on page 60). □

Solution to Ex. 11.4.2. If $\mathbf{y}$ is scaled by $\alpha \in \mathbb{R}$, then

$$\frac{\|\mathbf{P}\alpha\mathbf{y}\|^2}{\|\alpha\mathbf{y}\|^2} = \frac{\|\alpha\mathbf{P}\mathbf{y}\|^2}{\|\alpha\mathbf{y}\|^2} = \frac{\alpha^2\|\mathbf{P}\mathbf{y}\|^2}{\alpha^2\|\mathbf{y}\|^2} = \frac{\|\mathbf{P}\mathbf{y}\|^2}{\|\mathbf{y}\|^2}$$

Hence the uncentered R^2 is unchanged. □

4. Hint: Use the symmetry and idempotence of the matrix $\mathbf{M}_1$.

Solution to Ex. 11.4.3. Fix $\alpha \in \mathbb{R}$. By definition, R^2_α is

$$\frac{\|\mathbf{P}(\mathbf{y}+\alpha\mathbf{1})\|^2}{\|\mathbf{y}+\alpha\mathbf{1}\|^2} = \frac{\|\mathbf{P}\mathbf{y}+\alpha\mathbf{P}\mathbf{1}\|^2}{\|\mathbf{y}+\alpha\mathbf{1}\|^2} = \frac{\|\mathbf{P}\mathbf{y}+\alpha\mathbf{1}\|^2}{\|\mathbf{y}+\alpha\mathbf{1}\|^2} = \frac{\alpha^2\|\mathbf{P}\mathbf{y}/\alpha+\mathbf{1}\|^2}{\alpha^2\|\mathbf{y}/\alpha+\mathbf{1}\|^2} = \frac{\|\mathbf{P}\mathbf{y}/\alpha+\mathbf{1}\|^2}{\|\mathbf{y}/\alpha+\mathbf{1}\|^2}$$

where the second inequality follows from the fact that $\mathbf{1} \in \text{colspace}\,\mathbf{X}$. Taking the limit as $\alpha \to \infty$, we find that $R^2_\alpha \to 1$ as $\alpha \to \infty$. □

Solution to Ex. 11.4.4. This follows immediately from the definition of R^2, and the fact that, for any $\alpha \neq 0$,

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} = \frac{\alpha^2}{\alpha^2}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} = (\alpha\mathbf{X})((\alpha\mathbf{X})^\mathsf{T}(\alpha\mathbf{X}))^{-1}(\alpha\mathbf{X}^\mathsf{T})$$ □

Solution to Ex. 11.4.5. From exercise 11.4.21, the squared sample correlation between $\mathbf{x}$ and $\mathbf{y}$ can be written as

$$\hat{\varrho}^2 = \frac{[(\mathbf{M}_c\mathbf{x})^\mathsf{T}(\mathbf{M}_c\mathbf{y})]^2}{\|\mathbf{M}_c\mathbf{x}\|^2\|\mathbf{M}_c\mathbf{y}\|^2}$$

Also, $R^2_c = \|\mathbf{M}_c\mathbf{P}\mathbf{y}\|^2/\|\mathbf{M}_c\mathbf{y}\|^2$. Hence it suffices to show that, for the simple linear regression model in §11.2.2.1, we have

$$\|\mathbf{M}_c\mathbf{P}\mathbf{y}\| = \frac{|(\mathbf{M}_c\mathbf{x})^\mathsf{T}(\mathbf{M}_c\mathbf{y})|}{\|\mathbf{M}_c\mathbf{x}\|} \tag{11.33}$$

Let $\mathbf{X} = (\mathbf{1}, \mathbf{x})$ be the design matrix, where the first column is $\mathbf{1}$ and the second column is $\mathbf{x}$. Let

$$\hat{\beta}_1 := \bar{y} - \hat{\beta}_2\bar{x} \quad \text{and} \quad \hat{\beta}_2 := \frac{(\mathbf{M}_c\mathbf{x})^\mathsf{T}(\mathbf{M}_c\mathbf{y})}{\|\mathbf{M}_c\mathbf{x}\|^2}$$

be the least squares estimators of β_1 and β_2 respectively (see §11.2.2.1). We then have

$$\mathbf{P}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{1}\hat{\beta}_1 + \mathbf{x}\hat{\beta}_2$$

$$\therefore \quad \mathbf{M}_c\mathbf{P}\mathbf{y} = \mathbf{M}_c\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{M}_c\mathbf{x}\hat{\beta}_2$$

$$\therefore \quad \|\mathbf{M}_c\mathbf{P}\mathbf{y}\| = \|\mathbf{M}_c\mathbf{x}\hat{\beta}_2\| = |\hat{\beta}_2|\|\mathbf{M}_c\mathbf{x}\| = \frac{|(\mathbf{M}_c\mathbf{x})^\mathsf{T}(\mathbf{M}_c\mathbf{y})|}{\|\mathbf{M}_c\mathbf{x}\|^2}\|\mathbf{M}_c\mathbf{x}\|$$

Canceling $\|\mathbf{M}_c\mathbf{x}\|$ we get (11.33). This completes the proof. □

Solution to Ex. 11.4.6. It is sufficient to show that $\mathbf{P}\mathbf{M}_c = \mathbf{M}_c\mathbf{P}$. Since $\mathbf{1} \in \text{colspace}\,\mathbf{X}$ by assumption, we have $\mathbf{P}\mathbf{1} = \mathbf{1}$, and $\mathbf{1}^\mathsf{T}\mathbf{P} = (\mathbf{P}^\mathsf{T}\mathbf{1})^\mathsf{T} = (\mathbf{P}\mathbf{1})^\mathsf{T} = \mathbf{1}^\mathsf{T}$. Therefore $\mathbf{P}\mathbf{1}\mathbf{1}^\mathsf{T} =$

$\mathbf{1}\mathbf{1}^\top\mathbf{P}$, and

$$\mathbf{P}\mathbf{M}_c = \mathbf{P} - \frac{1}{N}\mathbf{P}\mathbf{1}\mathbf{1}^\top = \mathbf{P} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{P} = \mathbf{M}_c\mathbf{P} \qquad \square$$

Solution to Ex. 11.4.7. It suffices to show that $\|\mathbf{P}\mathbf{M}_c\mathbf{y}\|^2 = \sum_{n=1}^N (\hat{y}_n - \bar{y})^2$. This holds because

$$\sum_{n=1}^N (\hat{y}_n - \bar{y})^2 = \|\mathbf{P}\mathbf{y} - \mathbf{1}\bar{y}\|^2 = \|\mathbf{P}\mathbf{y} - \mathbf{P}\mathbf{1}\bar{y}\|^2 = \|\mathbf{P}(\mathbf{y} - \mathbf{1}\bar{y})\|^2 = \|\mathbf{P}\mathbf{M}_c\mathbf{y}\|^2 \qquad \square$$

Solution to Solution 11.4.8. Adopt the notation and assumptions of fact 11.1.2 on page 306. Let $\mathbf{y}$ be given. Let $\mathbf{P}_a$ and $\mathbf{P}_b$ be the projections onto the column spaces of $\mathbf{X}_a$ and $\mathbf{X}_b$ respectively, so that $R_i^2 = \|\mathbf{P}_i\mathbf{y}\|^2 / \|\mathbf{y}\|^2$ for $i \in \{a, b\}$. From fact 2.2.7 on page 32 we have $\mathbf{P}_a\mathbf{P}_b\mathbf{y} = \mathbf{P}_a\mathbf{y}$. Using this fact, and setting $\mathbf{y}_b := \mathbf{P}_b\mathbf{y}$, gives

$$\frac{R_a^2}{R_b^2} = \left(\frac{\|\mathbf{P}_a\mathbf{y}\|}{\|\mathbf{P}_b\mathbf{y}\|}\right)^2 = \left(\frac{\|\mathbf{P}_a\mathbf{P}_b\mathbf{y}\|}{\|\mathbf{P}_b\mathbf{y}\|}\right)^2 = \left(\frac{\|\mathbf{P}_a\mathbf{y}_b\|}{\|\mathbf{y}_b\|}\right)^2 \leqslant 1$$

where the final inequality follows from theorem 2.2.2 on page 31. Hence $R_b^2 \geqslant R_a^2$, and regressing with $\mathbf{X}_b$ produces (weakly) larger R^2. $\square$

Solution to Ex. 11.4.10. Part (ii) follows immediately from part (i). Regarding part (i), observe that

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})\|^2$$

By the Pythagorean law, the claim

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})\|^2$$

will be confirmed if $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \perp \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})$. This follows from the definition of $\hat{\boldsymbol{\beta}}$, because for arbitrary $\mathbf{a} \in \mathbb{R}^K$ we have

$$(\mathbf{X}\mathbf{a})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}) = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{y}) = 0 \qquad \square$$

Solution to Ex. 11.4.12. Let $\hat{\boldsymbol{\beta}}$ be a solution to $\min_{\mathbf{b}\in\mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$, which is to say that

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \leqslant \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \quad \text{for any } \mathbf{b} \in \mathbb{R}^K$$

If a and b are nonnegative constants with $a \leqslant b$, then $\sqrt{a} \leqslant \sqrt{b}$, and hence

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\| \leqslant \|\mathbf{y} - \mathbf{X}\mathbf{b}\| \quad \text{for any } \mathbf{b} \in \mathbb{R}^K$$

In other words, $\hat{\boldsymbol{\beta}}$ is a solution to $\min_{\mathbf{b}\in\mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|$. The "vice versa" argument

follows along similar lines. □

Solution to Ex. 11.4.14. To see why the vector of residuals sums to zero, observe that

$$\mathbf{1}^\top \mathbf{M}\mathbf{y} = \mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{P}\mathbf{y} = \mathbf{1}^\top \mathbf{y} - (\mathbf{P}^\top \mathbf{1})^\top \mathbf{y} = \mathbf{1}^\top \mathbf{y} - (\mathbf{P}\mathbf{1})^\top \mathbf{y}$$

where the last equality uses the fact the $\mathbf{P}$ is symmetric. Moreover, by exercise 11.4.13, we have $\mathbf{P}\mathbf{1} = \mathbf{1}$ whenever $\mathbf{1}$ is a column of $\mathbf{X}$. Therefore

$$\mathbf{1}^\top \mathbf{M}\mathbf{y} = \mathbf{1}^\top \mathbf{y} - (\mathbf{P}\mathbf{1})^\top \mathbf{y} = \mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{y} = 0 \qquad \square$$

Solution to Ex. 11.4.15. Since $\mathbf{1} \in \operatorname{colspace} \mathbf{X}$, we have $\mathbf{P}\mathbf{1} = \mathbf{1}$. It follows that

$$\frac{1}{N}\sum_{n=1}^{N} \hat{y}_n = \frac{1}{N}\mathbf{1}^\top \hat{\mathbf{y}} = \frac{1}{N}\mathbf{1}^\top \mathbf{P}\mathbf{y} = \frac{1}{N}\mathbf{1}^\top \mathbf{y} = \frac{1}{N}\sum_{n=1}^{N} y_n \qquad \square$$

Solution to Ex. 11.4.17. If $R^2 = 1$, then, by (11.12) on page 303, we have $\|\mathbf{M}\mathbf{y}\|^2 = 0$, and hence have $\hat{\mathbf{u}} = \mathbf{M}\mathbf{y} = \mathbf{0}$. Since $\mathbf{M}\mathbf{y} + \mathbf{P}\mathbf{y} = \mathbf{y}$, this implies that $\mathbf{P}\mathbf{y} = \mathbf{y}$. But then $\mathbf{y} \in \operatorname{colspace} \mathbf{X}$ by (vi) of theorem 2.2.2. □

Solution to Ex. 11.4.19. Fact 2.2.8 tells us that for any conformable vector $\mathbf{z}$ we have $\mathbf{z} = \mathbf{P}\mathbf{z} + \mathbf{M}\mathbf{z}$, where the two vectors on the right-hand side are orthogonal. Letting $\mathbf{z} = \mathbf{M}_c\mathbf{y}$, we obtain

$$\mathbf{M}_c\mathbf{y} = \mathbf{P}\mathbf{M}_c\mathbf{y} + \mathbf{M}\mathbf{M}_c\mathbf{y}$$

From fact 2.2.9 we have $\mathbf{M}\mathbf{M}_c\mathbf{y} = \mathbf{M}\mathbf{y}$. Using this result, orthogonality, and the Pythagorean law, we obtain

$$\|\mathbf{M}_c\mathbf{y}\|^2 = \|\mathbf{P}\mathbf{M}_c\mathbf{y}\|^2 + \|\mathbf{M}\mathbf{y}\|^2$$

Rearranging gives (11.32) □

Solution to Ex. 11.4.20. This is just fact 2.1.3 on page 16. □

Solution to Ex. 11.4.23. By fact 3.2.9, to show that $\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2$ is nonsingular, it suffices to show that the matrix is positive definite. By idempotence and symmetry of $\mathbf{M}_1$,

$$\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 = \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{M}_1 \mathbf{X}_2 = \mathbf{X}_2^\top \mathbf{M}_1^\top \mathbf{M}_1 \mathbf{X}_2 = (\mathbf{M}_1 \mathbf{X}_2)^\top \mathbf{M}_1 \mathbf{X}_2$$

Take any $\mathbf{a} \neq \mathbf{0}$. We need to show that

$$\mathbf{a}^\top (\mathbf{M}_1 \mathbf{X}_2)^\top \mathbf{M}_1 \mathbf{X}_2 \mathbf{a} = (\mathbf{M}_1 \mathbf{X}_2 \mathbf{a})^\top \mathbf{M}_1 \mathbf{X}_2 \mathbf{a} = \|\mathbf{M}_1 \mathbf{X}_2 \mathbf{a}\|^2 > 0$$

Since the only vector with zero norm is the zero vector, it now suffices to show that $\mathbf{M}_1\mathbf{X}_2\mathbf{a}$ is nonzero. From fact 2.2.8 on page 33, we see that $\mathbf{M}_1\mathbf{X}_2\mathbf{a} = \mathbf{0}$ only when $\mathbf{X}_2\mathbf{a}$ is in the column span of $\mathbf{X}_1$. Thus, the proof will be complete if we can show that $\mathbf{X}_2\mathbf{a}$ is not in the column span of $\mathbf{X}_1$.

Indeed, $\mathbf{X}_2\mathbf{a}$ is not in the column span of $\mathbf{X}_1$. If it were, then we could write $\mathbf{X}_1\mathbf{b} = \mathbf{X}_2\mathbf{a}$ for some $\mathbf{b} \in \mathbb{R}^{K_1}$. Rearranging, we get $\mathbf{X}\mathbf{c} = \mathbf{0}$ for some nonzero $\mathbf{c}$ (recall $\mathbf{a} \neq \mathbf{0}$). This contradicts linear independence of the columns of $\mathbf{X}$. □