A Primer in Econometric Theory

Lecture 10: Regression

John Stachurski
Lectures by Akshay Shanker

March 26, 2017

Linear Regression

Start with the prediction problem discussed in §8.2.2 — a system with vector input $\mathbf{x}_n \in \mathbb{R}^K$ followed by scalar output y_n

Examples:

- $\mathbf{x}_n$ 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
- $\mathbf{x}_n$ is a set of household characteristics (ethnicity, age, location, etc.) and y_n is household wealth at some later date
- $\mathbf{x}_n$ is price of electricity, prices of alternatives, temperature, household income, and measurements of the regional income distribution, while y_n is regional electricity consumption

Suppose we have N observations $\mathbf{z}_n := (\mathbf{x}_n, y_n)$, all draws from fixed joint distribution P

Since P is fixed, we are assuming the system is stationary across the set of draws

Our problem:

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

To define "good predictor" mathematically, we need a loss function

We will be using quadratic loss, thus minimize the prediction risk given by

$$R(f) := \mathbb{E}_P (y - f(\mathbf{x}))^2 \tag{1}$$

Minimizer of (1) over the set of all $\mathscr{B}$-measurable functions is the regression function $f^*(\mathbf{x}) := \mathbb{E}_P [y \mid \mathbf{x}]$

Recall we cannot compute the regression function because P is not known

Instead we apply the principle of empirical risk minimization, which leads to the problem

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

Here $\mathcal{H}$ is the hypothesis space, a set of candidate functions mapping $\mathbb{R}^K$ into $\mathbb{R}$

The problem (2) is called a **least squares** problem

As discussed at length in §8.2.3, minimizing empirical risk is different from minimizing the prediction risk $R(f)$ — thus $\mathcal{H}$ must be restricted

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, write

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

Problem (2) reduces to

$$\min_{\mathbf{b} \in \mathbb{R}^K} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\mathsf{T}\mathbf{b})^2 \tag{3}$$

Intuition: "line of best fit" to minimize in-sample prediction error

Good reasons to start with $\mathcal{H}_\ell$, even where no linearity assumptions are imposed:

1. $\mathcal{H}_\ell$ is a natural starting point when seeking a class of simple, well-behaved functions
2. setting $\mathcal{H} = \mathcal{H}_\ell$ allows us to obtain an analytical expression for the minimizer, which simplifies both analysis and computation
3. the technique has an extension from $\mathcal{H}_\ell$ to broader classes of functions

Least Squares Estimator

Now let's solve (3). Let

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

and

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

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

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

The objective function in (3) 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

$$\operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\| \tag{4}$$

Using the orthogonal projection theorem (recall theorem 3.3.2 in ET), the solution is

$$\hat{\boldsymbol{\beta}} := (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \tag{5}$$

Traditionally, $\hat{\boldsymbol{\beta}}$ 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 (3). 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 (4), it is the closest point in $\operatorname{colspace}\mathbf{X}$ to $\mathbf{y}$:

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

In what follows, $\mathbf{M}$ is the residual projection, as defined in (2.11) in ET

Assumptions

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

By theorem 2.1.3, $N \geq K$ is a necessary condition for the assumption to hold

(If $N < K$, then $\mathbb{R}^N$, which is necessarily spanned by N vectors, cannot contain K linearly independent vectors)

If this assumption does not hold, then minimizer of (4) still exists but is no longer unique (see ex. 3.5.34)

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

$$\boldsymbol{\Sigma}_{\mathbf{x}} := \mathbb{E}_P[\mathbf{x}_n \mathbf{x}_n^\intercal] \text{ is finite and positive definite} \tag{6}$$

Finite second moments imposed to evaluate expected squared errors

Assumption cannot be weakened unless we are willing to work with a different loss function

Notation

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 $\mathbf{x}_n^\mathsf{T}\hat{\boldsymbol{\beta}}$ associated with least squares estimate and the nth observation $\mathbf{x}_n$ of the input vector

The vector $\mathbf{My}$ is often denoted $\hat{\mathbf{u}}$, and called the **vector of residuals**:

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

The vector of residuals corresponds to the error that occurs when $\mathbf{y}$ is approximated by $\mathbf{Py}$

From fact 2.2.8

$$\mathbf{My} \perp \mathbf{Py} \quad \text{and} \quad \mathbf{y} = \mathbf{Py} + \mathbf{My} \tag{7}$$

In other words, $\mathbf{y}$ can be decomposed into two orthogonal vectors $\mathbf{Py}$ and $\mathbf{My}$:

- first represents the best approximation to $\mathbf{y}$ in $\operatorname{colspace} \mathbf{X}$
- second represents the residual

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

- **Total sum of squares** $:= \text{TSS} := \|\mathbf{y}\|^2$
- **Residual sum of squares** $:= \text{RSS} := \|\mathbf{My}\|^2$
- **Explained sum of squares** $:= \text{ESS} := \|\mathbf{Py}\|^2$

By (7) and the Pythagorean law

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{8}$$

When running regressions it is conventional to report the **coefficient of determination**, or R^2:

$$R^2 := \frac{\text{ESS}}{\text{TSS}} \tag{9}$$

Out of Sample Fit

How does linear least squares perform out-of-sample? Start with a general observation:

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

$R(f)$ is the prediction risk of f and expectations are taken under the unknown joint distribution P of the pairs $(\mathbf{x}, y)$

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:

1. The **intrinsic risk** $\mathbb{E}\left(y - f^*(\mathbf{x})\right)^2$
2. The **approximation error** $\mathbb{E}\left(f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*\right)^2$
3. The **estimation error** $\left(\mathbf{b}^* - \mathbf{b}\right)^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{x}\left(\mathbf{b}^* - \mathbf{b}\right)$

The intrinsic risk is also called Bayes risk, it is the residual error after y is approximated with the best possible predictor

- 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

The estimation error is caused by the deviation of our estimator from the best linear predictor $\mathbf{b}^*$

- deviation occurs because we are predicting using finite sample information on the joint distribution of $(\mathbf{x}, y)$

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

$$\hat{\boldsymbol{\beta}}_N \xrightarrow{p} \mathbf{b}^* \quad \text{as} \quad N \to \infty \tag{10}$$

Independence required only for the LLN to function — can weaken to ergodicity

Proofs

Proof. [Proof of theorem 11.1.1] Fix $\mathbf{b} \in \mathbb{R}^K$ and let $\ell(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{b}$

Recall from (8.17) in ET that we can write the the prediction risk as

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

To establish result, we show

$$\mathbb{E}\left[(f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b})^2\right]$$

$$= \mathbb{E}\left[(f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*)^2\right] + \mathbb{E}\left[(\mathbf{b}^* - \mathbf{b})^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}(\mathbf{b}^* - \mathbf{b})\right] \quad (11)$$

Proof. [Proof of theorem 11.1.1](cont.)

To see (11) holds, observe

$$f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b} = f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^* + \mathbf{x}^\mathsf{T}(\mathbf{b}^* - \mathbf{b}) \tag{12}$$

The terms $f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*$ and $\mathbf{x}^\mathsf{T}(\mathbf{b}^* - \mathbf{b})$ are orthogonal because:

- $\mathbf{x}^\mathsf{T}\mathbf{b}^*$ 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 $\mathbf{a}^\mathsf{T}\mathbf{x}$
- as such, $f^*(\mathbf{x}) - \mathbf{x}^\mathsf{T}\mathbf{b}^*$ is orthogonal to every element of the target subspace $\text{span}\{\mathbf{x}\}$, including $\mathbf{x}^\mathsf{T}(\mathbf{b}^* - \mathbf{b})$

Proof.[Proof of theorem 11.1.1](cont.)

For any orthogonal elements u and v of L_2 we have

$$\mathbb{E}\left[(u+v)^2\right] = \mathbb{E}\left[u^2\right] + \mathbb{E}\left[v^2\right]$$

(This is the Pythagorean law in L_2.)

Squaring both sides of (12), taking expectations and applying this law gives (11) □

Proof. [Proof of theorem 11.1.2]

First 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

$$\hat{\boldsymbol{\beta}}_N = \left[\frac{1}{N}\mathbf{X}^\intercal\mathbf{X}\right]^{-1} \cdot \frac{1}{N}\mathbf{X}^\intercal\mathbf{y}$$

$$= \left[\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\intercal\right]^{-1} \cdot \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n y_n \quad (13)$$

Proof. [Proof of theorem 11.1.2](cont.)

By the matrix LLN in fact 6.2.3, we have

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

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 □

In-Sample Fit

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 Equation (9) above)

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

That $R^2 \leq 1$ is immediate from $\|\mathbf{P}\mathbf{y}\| \leq \|\mathbf{y}\|$

Exercise 11.4.17 asks you to prove the second claim

More generally, a high R^2 indicates $\mathbf{y}$ is relatively close to $\operatorname{colspace} \mathbf{X}$

We can increase R^2 at least weakly by adding regressors

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

$$\operatorname{colspace} \mathbf{X}_a \subset \operatorname{colspace} \mathbf{X}_b \implies R_a^2 \leq R_b^2$$

For a proof, see exercise 11.4.8

Misleading to equate high R^2 with a successful regression

Note

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

where R_{emp} is as defined in (2) and $\hat{f}$ is our linear predictor $\hat{f}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\hat{\boldsymbol{\beta}}$

High R^2 means low empirical risk and good in-sample fit

But low empirical risk no guarantee of low prediction risk, as emphasized in §8.2.3

Let's link fact 11.1.2 with fact 8.2.1

- fact 8.2.1 says we can always decrease empirical risk by increasing the hypothesis space

Suppose $\mathbf{x}$ lists a large number of possible regressors. Let the hypothesis space be

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

where $1 \leq j \leq K$

Empirical risk minimization over $\mathcal{H}_j$ equivalent to linear regression over the first j regressors

Empirical risk falls as j increases by fact 8.2.1 — hence R^2 increases; same conclusion as fact 11.1.2

Transformations and Basis Functions

In discussing the decision to set $\mathcal{H} = \mathcal{H}_\ell$, we mentioned we can use many of the same ideas when extending $\mathcal{H}$ to a broader class of functions

First transform the data using some arbitrary function $\boldsymbol{\phi}: \mathbb{R}^K \to \mathbb{R}^J$

The action of $\boldsymbol{\phi}$ on $\mathbf{x} \in \mathbb{R}^K$

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

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**

We solve the empirical risk minimization problem when the hypothesis space is

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

The empirical risk minimization problem is then

$$\min_{\ell} \sum_{n=1}^{N} \{y_n - \ell(\boldsymbol{\phi}(\mathbf{x}_n))\}^2 = \min_{\gamma \in \mathbb{R}^J} \sum_{n=1}^{N} (y_n - \gamma^\intercal \boldsymbol{\phi}(\mathbf{x}_n))^2 \quad (14)$$

Switching to matrix notation, if

$$\boldsymbol{\Phi} := \begin{pmatrix} \phi_1(\mathbf{x}_1) & \cdots & \phi_J(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \cdots & \phi_J(\mathbf{x}_2) \\ \vdots & \cdots & \vdots \\ \phi_1(\mathbf{x}_N) & \cdots & \phi_J(\mathbf{x}_N) \end{pmatrix} \in \mathbb{R}^{N \times J} \tag{15}$$

Then the objective in (14) can be expressed as $\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\gamma}\|^2$. Since increasing functions don't affect minimizers, the problem becomes

$$\operatorname*{argmin}_{\boldsymbol{\gamma} \in \mathbb{R}^J} \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\gamma}\| \tag{16}$$

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

$$\gamma^\mathsf{T} \boldsymbol{\phi}(x_n) = \gamma^\mathsf{T} \begin{pmatrix} x_n^0 \\ x_n^1 \\ \vdots \\ x_n^{J-1} \end{pmatrix} = \sum_{j=1}^{J} \gamma_j x_n^{j-1} \tag{17}$$

The mononomial basis transformation applied to scalar x corresponds to univariate polynomial regression, as discussed in §8.2.3 of ET

Under this transformation, the matrix $\boldsymbol{\Phi}$ in (15) is called the **Vandermonde matrix**

Weierstrass approximation theorem: polynomials of sufficiently high order can effectively approximate any one-dimensional continuous nonlinear relationship

Example. 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

In this figure, clear no linear function mapping x to y can produce small approximation error

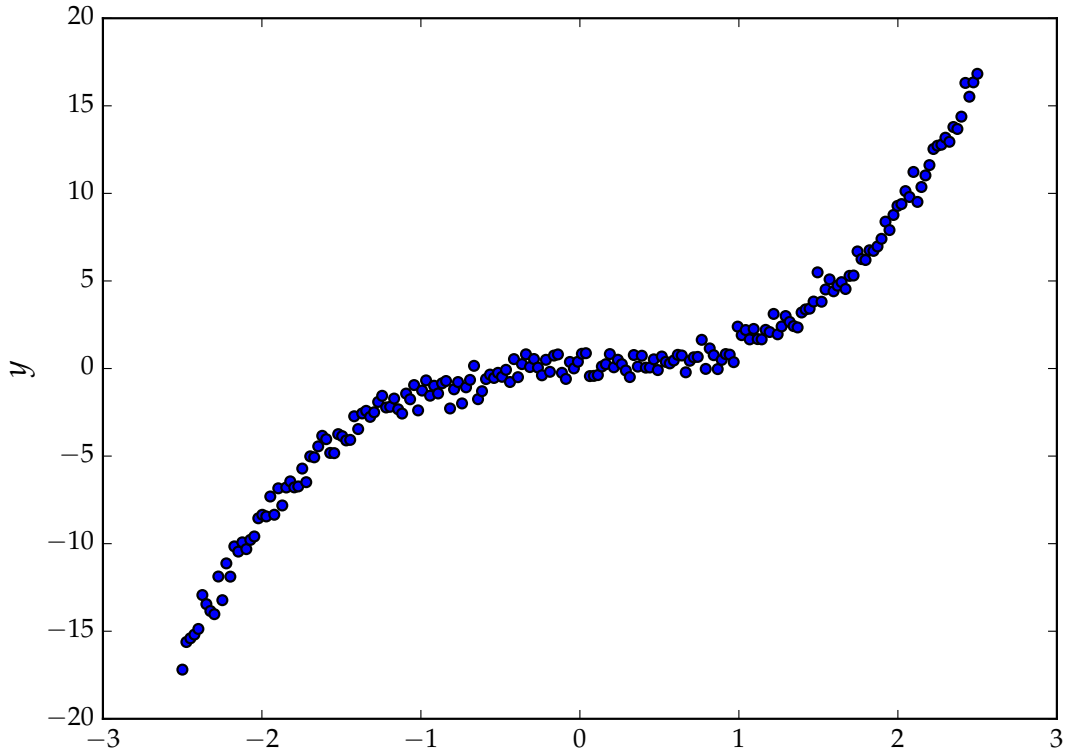

Figure: Nonlinear relationship between x and y

Figure on following slide shows data after applying the transformation $\mathbb{R} \ni x \mapsto \boldsymbol{\phi}(x) := (x, x^3)^\mathsf{T} \in \mathbb{R}^2$

The plane drawn in the figure 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

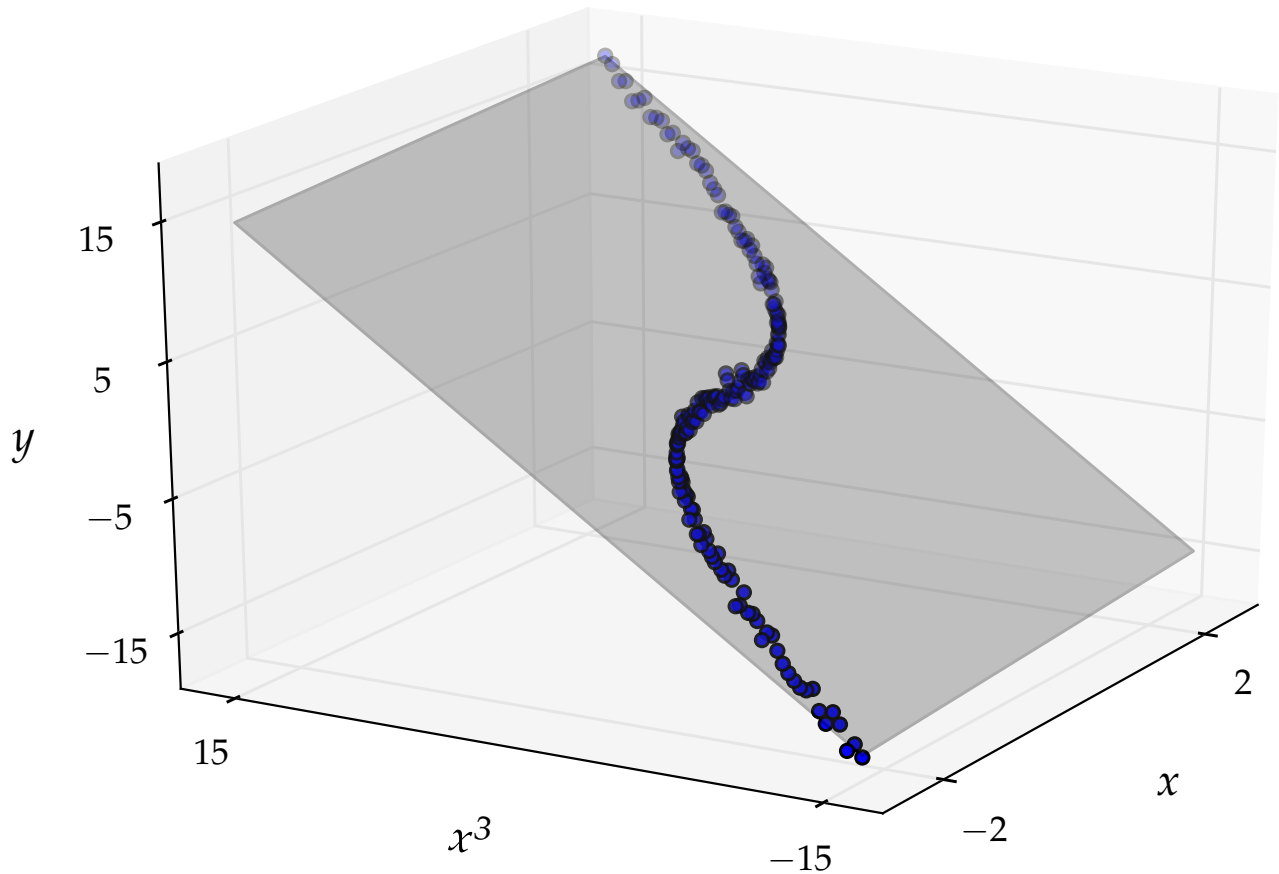

Figure: Approximate linearity after projecting the data to $\mathbb{R}^2$

The Frisch–Waugh–Lovell Theorem

The Frisch–Waugh–Lovell (FWL) theorem yields an expression for an arbitrary sub-vector of the least squares estimator $\hat{\boldsymbol{\beta}}$ obtained by regressing $\mathbf{y}$ on $\mathbf{X}$

Continue with assumptions made already in the lecture

Let $\mathbf{y}$ and $\mathbf{X}$ be given and let $\hat{\boldsymbol{\beta}}$ be the least squares estimator, as given by equation (5) above

In addition, let K_1 be an integer with $1 \leq K_1 < K$, 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

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

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

For a proof, see ET page 311

The expression for $\hat{\boldsymbol{\beta}}_2$ in theorem 11.2.1 can be rewritten as

$$\hat{\boldsymbol{\beta}}_2 = [(\mathbf{M}_1\mathbf{X}_2)^\intercal \mathbf{M}_1\mathbf{X}_2]^{-1}(\mathbf{M}_1\mathbf{X}_2)^\intercal \mathbf{M}_1\mathbf{y} \tag{18}$$

(see exercise 11.4.22)

The above formula gives us 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

For intuition: consider the case where $\mathbf{X}_2$ is the single column $\mathrm{col}_K \mathbf{X}$, containing the observations on the Kth regressor

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 $\mathbf{M}_1$

The least squares estimate $\hat{\beta}_K$ can be found by regressing

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

on

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

The 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" $\mathbf{X}_{-K}$

Intuitively, the process for obtaining the least squares estimate $\hat{\beta}_K$ is:

1. remove effects of all other regressors from $\mathbf{y}$ and $\operatorname{col}_K \mathbf{X}$, producing $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}_K$
2. regress $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{x}}_K$

Different from the process for obtaining the coefficient of the vector $\operatorname{col}_K \mathbf{X}$ in a simple univariate regression:

1. regress $\mathbf{y}$ on $\operatorname{col}_K \mathbf{X}$

Difference between the univariate least squares estimated coefficient of the Kth regressor and the multiple regression least squares coefficient:

- the multiple regression coefficient $\hat{\beta}_K$ measures the *isolated relationship* between x_K and y
- does not take into account indirect channels involving other variables

We can illustrate further with simulation. Suppose

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

Generate N independent observations from this model

Regress y on the observations of (x_1, x_2)

- coefficients for x_1 and x_2 will both be close to unity, provided N is sufficiently large

Regress y on x_1 alone

- 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

- an increase in x_1 tends to have a large positive effect on x_2, which in turn increases y

Simple Regression

Application of FWL Theorem: derive expression for the slope coefficient in simple linear regression from the multivariate expression

Simple linear regression as special case of multivariate regression

- $\mathbf{1}$ is the first column of $\mathbf{X}$ and $K = 2$

The second column of $\mathbf{X}$ will be denoted by $\mathbf{x}$

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 $\hat{\beta}_2$

$$\hat{\beta}_2 = [(\mathbf{x} - \bar{x}\mathbf{1})^\mathsf{T}(\mathbf{x} - \bar{x}\mathbf{1})]^{-1}(\mathbf{x} - \bar{x}\mathbf{1})^\mathsf{T}(\mathbf{y} - \bar{y}\mathbf{1}) \tag{19}$$

By the FWL theorem (equation 18)

$$\hat{\beta}_2 = [(\mathbf{M}_c\mathbf{x})^\mathsf{T}\mathbf{M}_c\mathbf{x}]^{-1}(\mathbf{M}_c\mathbf{x})^\mathsf{T}\mathbf{M}_c\mathbf{y} \tag{20}$$

here $\mathbf{M}_c$ is the residual projection associated with the linear subspace $S = \text{span}\{\mathbf{1}\}$

For this residual projection $\mathbf{M}_c$ and any $\mathbf{z}$, we have $\mathbf{M}_c\mathbf{z} = \mathbf{z} - \bar{z}\mathbf{1}$
— RHS of (19) and (20) coincide

Generalize to the case where there are multiple nonconstant regressors

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

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

Applying the FWL theorem, we can write $\hat{\boldsymbol{\beta}}_2$ as

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

where $\mathbf{M}_c$ is the residual projection (Equation (3.10) in ET)

$\mathbf{M}_c\mathbf{y}$ is $\mathbf{y}$ centered around its mean

$\mathbf{M}_c\mathbf{X}_2$ is a matrix formed by taking each column of $\mathbf{X}_2$ and centering it around its

.....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

Centered R^2

Several versions of R^2 reported in common regression packages

One of these is so called centered R^2

The version we discussed so far will now be called uncentered R^2

Why introduce alternative to uncentered R^2?

- fails to be invariant to certain changes of units that involve addition or subtraction whenever $\mathbf{X}$ contains an intercept
- for e.g. actual inflation versus inflation in excess of a certain level, income versus income over a certain threshold, etc.

Define centered R^2

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

(See ex. 11.4.6 to prove equality)

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)

Rewrite R^2 (ex. 11.4.7) as

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

In the case simple regression, R_c^2 is a measure of correlation

- R_c^2 is equal to the square of the sample correlation between the regressor and regressand, as defined by Equation 8.5 (shown in ex. (11.4.5))