The Classical Linear Regression Model and Least Squares

Greene Chapters: 2, 3
R script `mod1_2a, mod1_2b, mod1_2c, mod1_2d`

Regression Analysis

Much work in applied econometrics is based on regression analysis.

Definition

A *regression* is a stipulated relationship between the *expectation* of a dependent variable and a combination of explanatory data and unknown parameters.
Generically:

$$E(y_i \mid \mathbf{x_i}, \boldsymbol{\beta}) = f(\mathbf{x_i}, \boldsymbol{\beta}) \quad (1)$$

We can perform regression analysis in fully parametric and semi-parametric frameworks. We usually write the regression model as:

$$y_i = E(y_i \mid \mathbf{x_i}, \boldsymbol{\beta}) + \left(y_i - E(y_i \mid \mathbf{x_i}, \boldsymbol{\beta})\right) = E(y_i \mid \mathbf{x_i}, \boldsymbol{\beta}) + \varepsilon_i = f(\mathbf{x_i}, \boldsymbol{\beta}) + \varepsilon_i \quad (2)$$

where ε_i is an "error term" that includes everything we can't explain about y_i even knowing $\mathbf{x_i}$ and (hypothetically) population parameter vector $\boldsymbol{\beta}$.

What's in the error term?

The error term is also called the "disturbance term". It can include some or all of the following components:

1. *Specification error:* There are other factors than $\mathbf{x_i}$ that drive the observed variability in y_i, and / or the functional relationship between y_i and $\mathbf{x_i}$ is incorrect. The former is essentially unavoidable in most applications, and does not necessarily jeopardize our ability to estimate the effect of $\mathbf{x_i}$ on y_i. The latter is a bigger problem. It can produce inefficient and, in some cases, misleading estimates.

2. *Measurement error* (also known as "*errors in variables*"): Usually a minor problem if we mis-measure y_i, but potentially a serious issue if we mis-measure elements of $\mathbf{x_i}$.

3. *Human erratic behavior:* The stipulated relationship between y_i and $\mathbf{x_i}$ is generally correct, but units of observation (usually people) slightly change their behavior from case to case, introducing deviations from the stipulated relationship. This is a truly random part of the error term, and in most cases does not jeopardize estimation results.

In a *parametric* framework, we would assign a density to ε_i (such as $n(0, \sigma^2)$), where *n* stands for "normally distributed". In that case, σ^2 becomes an additional parameter that has to be estimated. In a semi-parametric framework we would simply state that $E(\varepsilon_i) = 0$ to preserve the relationship in (1).

A special case of a regression model is a model that is linear in $\boldsymbol{\beta}$, i.e. one that can be written as

$$y_i = \mathbf{x_i}'\boldsymbol{\beta} + \varepsilon_i \quad (3)$$

Such models can be estimated via a technique called "Least Squares (LS)". If certain assumption on ε_i hold, the model is called "Classical Linear Regression Model" (CLRM), and estimation can proceed via "Ordinary Least Squares" (OLS), the topic of the next section.

R practice: Building a regression model for study time

R script `mod1_2a` illustrates how to build a regression relationship with simulated data. The script also shows the gain in accuracy as the regression model chosen by the researcher approaches the true model.

Notation

Assume you have a sample of wages and explanatory variables for *n* individuals.
For a single observation (= person, firm, household, etc), the CLRM can be written as:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x_i'\beta} + \varepsilon_i \tag{4}$$

The β - terms are unknown population parameters. In a regression context, they are often referred to as "coefficients". In practice, $\mathbf{x_{i1}}$ is often simply the number "1", which then makes β_1 the intercept or "constant term" of the regression model. The remaining β - terms are "slope coefficients". By convention, I will index the β - terms from 1 through *k*. (In some texts, the index is from 0 to *k*, which implies that the total number of coefficients is *k*+1).

We can stack these equations for all *i=1..n* individuals:

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{aligned} \tag{5}$$

Next, we note that the left hand side and the error terms can be compactly expressed as a vectors $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}'$ and $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix}'$. For a given individual, the right hand side (minus the error term) can be written as an inner product of vectors:

$$\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \mathbf{x_i'} \times \boldsymbol{\beta} \quad \text{where}$$

$$\mathbf{x_i'} = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ik} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \tag{6}$$

We can now write the entire system of linear equations as:

$$\mathbf{y}=\begin{bmatrix}\mathbf{x_1'\beta}\\ \mathbf{x_2'\beta}\\ \vdots\\ \mathbf{x_n'\beta}\end{bmatrix}+\boldsymbol{\varepsilon}=\begin{bmatrix}\mathbf{x_1'}\\ \mathbf{x_2'}\\ \vdots\\ \mathbf{x_n'}\end{bmatrix}\cdot\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{X}\cdot\boldsymbol{\beta}+\boldsymbol{\varepsilon}\qquad\text{where}\quad\mathbf{X}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1k}\\ x_{21}&x_{22}&\cdots&x_{2k}\\ \vdots&\vdots&\ddots&\vdots\\ x_{n1}&x_{n2}&\cdots&x_{nk}\end{bmatrix}\tag{7}$$

Assumptions

The CLRM rests on 4 key assumptions. An optional 5th assumption (normality of the error term) can be added in cases where a fully parametric approach is desired or needed.

Assumption 1:

The stipulated relationship is linear in parameters (i.e. in $\boldsymbol{\beta}$). Note: This does not mean that the elements of data matrix **X** also have to be linear. It's OK to have log-terms, squared terms, etc in the **X** matrix. Certain nonlinear forms of y are also permissible, as long as they can be transformed to linearity. For example, the following common model is a permissible CLRM:

$$y_i=\exp\left(\mathbf{x_i'\beta}+\varepsilon_i\right)\leftrightarrow\ln\left(y_i\right)=\mathbf{x_i'\beta}+\varepsilon_i\tag{8}$$

The log-transformation on both sides preserves linearity. If all of the x_i's are logged as well, the model is called "log-log" or "log-linear". If all of the x_i's are linear, the model is called "semi-log". Naturally, a mix of logged and linear regressors is also possible. As we will see below, logging either side changes the interpretation of marginal effects.

Assumption 2:

The data matrix **X** has to be "full rank". **X** has dimensions *n x k*, *n>k*, so "full rank" implies that **X** is rank "*k*". In words: All variables in **X** are linearly independent. If this wasn't the case, there would be an infinite number of solutions for the estimate of $\boldsymbol{\beta}$.

Example: see R `script mod1_2b`

Assumption 3:

The expectation of all error terms, conditional on **X**, is zero, i.e

$$E\left(\boldsymbol{\varepsilon}\,|\,\mathbf{X}\right)=\begin{bmatrix}E\left(\varepsilon_1\,|\,\mathbf{X}\right)\\ E\left(\varepsilon_2\,|\,\mathbf{X}\right)\\ \vdots\\ E\left(\varepsilon_n\,|\,\mathbf{X}\right)\end{bmatrix}=\mathbf{0}\tag{9}$$

In words, this implies that no element of the data matrix contains any information about the expectation of any error term, so $\boldsymbol{\varepsilon}$ is a true vector of "unknown" or "unobservable" effects. If this assumption is violated, our estimated parameters will "pick up" undesired effects from "the stuff in the error term", thus producing misleading results. This is the infamous "omitted variable (OV)" problem, which comes in many sizes & shapes. We will talk about OV's at length later in this course.

As discussed in Greene, ch. 2, Assumption 3 also implies the following relationships:

$$E[\varepsilon_i] = 0$$
$$Cov(\varepsilon_i, \mathbf{X}) = 0 \quad (10)$$
$$E[\mathbf{y} \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$$

Assumption 3 is often mis-interpreted and can be confusing. Here is a closer look.

First, we need to distinguish between the distribution of the error term and the distribution of the explanatory variables. In the CLRM, we assume that each observation (say a given individual *i*) is assigned her own error term, ε_i. In addition, error terms are independent from each other. Thus, we can focus on one error term at the time when examining the underpinnings of Assumption 3.

In the CLRM, by Assumption 4, this error term follows a distribution with mean zero and variance σ^2, and *all* error terms, ε_i, i=1…n, follow the *exact same distribution.* Thus we can think of all error terms as realizations from the *same* distribution. However, we will later encounter models where each ε_i has its own distribution (i.e. a violation of Assumption 4), but Assumption 3 may still hold. Thus, for this discussion of Assumption 3, we will make no further assumption regarding the exact distribution of ε_i, and how it relates to the distribution of $\varepsilon_j, j \neq i$, other than invoking independence. In other words, it's OK to think of each ε_i as following its own distribution.

Now consider some explanatory variable x_k (example: "SAT score"). Generally, we consider each explanatory variable as a true random variable in the statistical sense, following some underlying distribution, potentially jointly with other regressors. Individual *i*'s realization of this variable is x_{ik}. The set of realizations of x_k for the sample at hand is vector $\mathbf{x_k}$. For cross-sectional regression, where each observation corresponds to a different individual (or "cross-sectional unit") it makes sense to think of each $x_{ik}, i = 1 \cdots n$ in $\mathbf{x_k}$ as realization from the population distribution $f(x_k)$. This distribution is shared by all individuals in the sample. In time series analysis this is not as clear-cut, since x_{tk} (example: water use in month *t*) could feasibly follow a different distribution than $x_{(t-1)k}$ (water use in month *t-1*). Thus, as with the error term, we can assume without loss of generality that each $x_{ik}, i = 1 \cdots n, k = 1 \cdots K$ follows its own (unspecified) distribution within the context of this discussion.

So let's first focus on a single draw of the error term, ε_i, and a single draw of some regressor for person *i*, x_{ik}. For simplicity, let's also assume that there are no other explanatory variables in the model (so we don't have to worry about potential correlation of x_{ik} with $x_{il}, l \neq k$). This means we can drop the "*k*" subscript for now.

To talk about conditional expectations in a meaningful way, we must assume that ε_i and x_i (potentially) follow a joint distribution, say $f(\varepsilon_i, x_i)$. This joint density can always be written as a product of a marginal and a conditional density, i.e.

$$f(\varepsilon_i, x_i) = f(\varepsilon_i \mid x_i) f(x_i) = f(x_i \mid \varepsilon_i) f(\varepsilon_i) \quad (11)$$

The marginal density, marginal (unconditional) expectation, and conditional expectation of ε_i are given as follows:

$$\begin{aligned} f(\varepsilon_i) &= \int_{x_i} f(\varepsilon_i, x_i)\, dx_i \\ E(\varepsilon_i) &= \int_{\varepsilon_i} \varepsilon_i f(\varepsilon_i)\, d\varepsilon_i \\ E(\varepsilon_i \mid x_i) &= \int_{\varepsilon_i} \varepsilon_i f(\varepsilon_i \mid x_i)\, d\varepsilon_i \end{aligned} \tag{12}$$

We can now derive an important relationship, called the "Law of Iterated Expectations" (see Greene p. 16 and p. 1007):

$$\begin{aligned} &E(\varepsilon_i) = \int_{\varepsilon_i} \varepsilon_i f(\varepsilon_i)\, d\varepsilon_i = \int_{\varepsilon_i} \varepsilon_i \left(\int_{x_i} f(\varepsilon_i, x_i)\, dx_i \right) d\varepsilon_i = \int_{\varepsilon_i} \varepsilon_i \left(\int_{x_i} f(\varepsilon_i \mid x_i) f(x_i)\, dx_i \right) d\varepsilon_i = \\ &\int_{x_i} \left(\int_{\varepsilon_i} \varepsilon_i f(\varepsilon_i \mid x_i)\, d\varepsilon_i \right) f(x_i)\, dx_i = E_{x_i}\left(E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) \end{aligned} \tag{13}$$

This says that the unconditional expectation of ε_i can be interpreted as the expectation, over x_i, of the conditional expectation of $\varepsilon_i \mid x_i$.

Now back to Assumption 3: It implies that $E_{\varepsilon_i}(\varepsilon_i \mid x_i) = 0$. By (13), this automatically implies that the unconditional expectation is zero as well, since $E_{x_i}\left(E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) = E_{x_i}(0) = 0$.

Assumption 3 is often interpreted as "ε_i and x_i are uncorrelated", i.e. have a covariance of zero. Let's check if this is correct, i.e. if $E_{\varepsilon_i}(\varepsilon_i \mid x_i) = 0 \Rightarrow \operatorname{cov}(\varepsilon_i, x_i) = 0$:

$$\begin{aligned} &\operatorname{cov}(\varepsilon_i, x_i) = E_{\varepsilon_i, x_i}\left((\varepsilon_i - E(\varepsilon_i)) * (x_i - E(x_i)) \right) = \\ &E_{\varepsilon_i, x_i}\left(\varepsilon_i x_i - E(\varepsilon_i) x_i - \varepsilon_i E(x_i) + E(\varepsilon_i) E(x_i) \right) = \\ &E_{\varepsilon_i, x_i}(\varepsilon_i x_i) - E_{\varepsilon_i, x_i}\left(E(\varepsilon_i) x_i \right) - E_{\varepsilon_i, x_i}\left(\varepsilon_i E(x_i) \right) + E(\varepsilon_i) E(x_i) = \\ &E_{x_i}\left(E_{\varepsilon_i}(\varepsilon_i x_i \mid x_i) \right) - E_{x_i}\left(E_{\varepsilon_i}\left(E(\varepsilon_i) x_i \mid x_i \right) \right) - E_{x_i}\left(E_{\varepsilon_i}\left(\varepsilon_i E(x_i) \mid x_i \right) \right) + E(\varepsilon_i) E(x_i) = \\ &E_{x_i}\left(x_i E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) - E_{x_i}\left(E(\varepsilon_i) x_i \right) - E_{x_i}\left(E(x_i) E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) + E(\varepsilon_i) E(x_i) = \\ &E_{x_i}\left(x_i E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) - E(\varepsilon_i) E(x_i) - E_{x_i}\left(E(x_i) E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) + E(\varepsilon_i) E(x_i) = \\ &E_{x_i}\left(x_i E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) - E_{x_i}\left(E(x_i) E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) = \\ &E_{x_i}\left(x_i E_{\varepsilon_i}(\varepsilon_i \mid x_i) - E(x_i) E_{\varepsilon_i}(\varepsilon_i \mid x_i) \right) = E_{x_i}\left(E_{\varepsilon_i}(\varepsilon_i \mid x_i)(x_i - E(x_i)) \right) \end{aligned} \tag{14}$$

Naturally, if $E_{\varepsilon_i}(\varepsilon_i \mid x_i) = 0$ the last line in (14) will always equal zero, so indeed we have $E_{\varepsilon_i}(\varepsilon_i \mid x_i) = 0 \Rightarrow \text{cov}(\varepsilon_i, x_i) = 0$. Note that the last line also corresponds to Greene's Theorem B.2 on p. 1007.

Is the reverse true, i.e. does $\text{cov}(\varepsilon_i, x_i) = 0 \Rightarrow E_{\varepsilon_i}(\varepsilon_i \mid x_i) = 0$ hold? It will as long as $x_i \neq E(x_i) \forall i$, else we have

$$\text{cov}(\varepsilon_i, x_i) = E_{x_i}\left(E_{\varepsilon_i}(\varepsilon_i \mid x_i)\left(E(x_i) - E(x_i)\right)\right) = E_{x_i}\left(E_{\varepsilon_i}(\varepsilon_i \mid x_i) * 0\right) = 0 \tag{15}$$

regardless of the value of $E_{\varepsilon_i}(\varepsilon_i \mid x_i)$. Naturally, this can only hold for the constant term in a given regression model.

From Assumption 3 it also follows that

$$E(y_i \mid x_i) = E(x_i\beta + \varepsilon_i \mid x_i) = E(x_i\beta) + E(\varepsilon_i \mid x_i) = E(x_i\beta) + 0 = x_i\beta \tag{16}$$

i.e. we have indeed a true regression – relationship. For this reason Assumption 3 is often referred to as the "regression" assumption.

Now let's go a step further and assume that there are two explanatory variables in the model , x_{ik} and x_{il}, and that these two regressors are potentially correlated with each other and with the error term. (For example, x_{il} could be "first year GPA" x_{ik} could be "SAT score", and both could, in theory, be correlated with the unobserved variable "ability", or "spunk"). Thus, we are now considering the tri-variate density $f(\varepsilon_i, x_{ik}, x_{il})$. This does not pose any additional complications.

Assumption 3 simply requires that $E(\varepsilon_i \mid x_{ik}, x_{il}) = \int_{\varepsilon_i} \varepsilon_i f(\varepsilon_i \mid x_{ik}, x_{il}) d\varepsilon_i = 0$

As before, this implies $E(\varepsilon_i) = 0$ since

$$E(\varepsilon_i) = E_{x_{ik}, x_{il}}\left(E_{\varepsilon_i}(\varepsilon_i \mid x_{ik}, x_{il})\right) = E_{x_{ik}, x_{il}}(0) = 0 \tag{17}$$

By analogy to our derivation above, the covariance between ε_i and each of the two regressors must be zero as well.

Pushing this one last step further, we now extend Assumption 3 also to observations on regressors for other individuals in the sample. Consider a second individual, *j*, for whom we observe realizations for the same two regressors, x_{jk} and x_{jl}. This simply extends the exposition to a five-variate joint density, i.e. $f(\varepsilon_i, x_{ik}, x_{il}, x_{jk}, x_{jl})$, and Assumption 3 requires that

$E(\varepsilon_i \mid x_{ik}, x_{il}, x_{jk}, x_{jl}) = \int_{\varepsilon_i} \varepsilon_i f(\varepsilon_i \mid x_{ik}, x_{il}, x_{jk}, x_{jl}) d\varepsilon_i = 0$

All other results follow by analogy. Thus, we can compactly express Assumption 3 for any generic CLRM as shown in (9).

R script `mod1_2c` illustrates the implications of Assumption 3.

Assumption 4

Homoskedasticity (= equal variance) of error terms:

$$V\left(\varepsilon_i\right)=\sigma^2 \qquad \forall i=1\cdots n \tag{18}$$

So we assume all *n* error terms are "drawn" from the same distribution with mean 0 and variance σ^2.

Error terms are uncorrelated ("non-autocorrelated"):

$$Cov\left(\varepsilon_i,\varepsilon_j\right)=E\left(\left(\varepsilon_i-E\left(\varepsilon_i\right)\right)\cdot\left(\varepsilon_j-E\left(\varepsilon_j\right)\right)\right)=E\left(\varepsilon_i\cdot\varepsilon_j\right)=0, \qquad \forall i\neq j \tag{19}$$

Violations of these assumptions is called "heteroskedasticity" and "autocorrelation", respectively. We'll address these issues later in this course.

For the full model, Assumption 4 can be written as

$$V\left(\boldsymbol{\varepsilon}\right)=E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)=E\begin{bmatrix}\varepsilon_1\varepsilon_1 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n\\ \varepsilon_2\varepsilon_1 & \varepsilon_2\varepsilon_2 & \cdots & \varepsilon_2\varepsilon_n\\ \vdots & \vdots & \ddots & \vdots\\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & \varepsilon_n\varepsilon_n\end{bmatrix}=\begin{bmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{bmatrix}=\sigma^2\mathbf{I} \tag{20}$$

Disturbances that satisfy this property are called "spherical".

For the dependent variable, this implies:

$$V\left(\mathbf{y}\mid\mathbf{X}\right)=V\left(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right)=V\left(\mathbf{X}\boldsymbol{\beta}\right)+V\left(\boldsymbol{\varepsilon}\right)+2Cov\left(\mathbf{X}\boldsymbol{\beta},\boldsymbol{\varepsilon}\right)=V\left(\boldsymbol{\varepsilon}\right)=\sigma^2\mathbf{I} \tag{21}$$

Assumption 5

The error terms follow a normal distribution, i.e.

$$\varepsilon_i\sim n\left(0,\sigma^2\right) \quad \forall i \qquad \text{or} \qquad \boldsymbol{\varepsilon}\sim n\left(\mathbf{0},\sigma^2\mathbf{I}\right) \tag{22}$$

For the dependent variable, this implies:

$$\mathbf{y}\sim n\left(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}\right) \tag{23}$$

This normality assumption is not necessary to derive parameter estimates for $\boldsymbol{\beta}$, but it is needed to derive some exact statistical results for estimators, and to construct certain test statistics.

*The statistical properties of **X***

In some applications, the elements of the observed data **X** can be viewed as fixed (e.g. in experimental settings where all "data" are perfectly controlled), but in most cases **X** itself will follow some distribution in the underlying population. If so, we view the entire analysis as "conditional on **X**", which means we essentially "freeze" **X** at the observed values. We understand that a different **X** might produce different results, but we hope that these differences would vanish with (hypothetical) collection of more and more data.

What's important is Assumption 3: Whichever mechanism generates **X** is unrelated to the mechanism that generates the error terms.

Estimation

The population model at the observation level is given by (4). We seek an estimator for $\boldsymbol{\beta}$ with desirable statistical properties. We will denote this estimator generically as **b**, and the resulting estimate for $E[y_i|\mathbf{x_i}]$ as $\hat{y}_i$, i.e.

$$\hat{E}\left[y_i \mid \mathbf{x_i}\right] = \hat{y}_i = \mathbf{x_i'b} \tag{24}$$

The term $\hat{y}_i$ is often called "**fitted value**". The difference between actually observed y_i and estimated (or "predicted") $\hat{y}_i$ is called "residual" e_i, i.e.

$$e_i = y_i - \hat{y}_i = y_i - \mathbf{x_i'b} \tag{25}$$

In passing, we note that this implies the following equality

$$\mathbf{x_i'\beta} + \varepsilon_i = \mathbf{x_i'b} + e_i \qquad \text{and} \qquad \mathbf{X\beta} + \boldsymbol{\varepsilon} = \mathbf{Xb} + \mathbf{e}\,. \tag{26}$$

To find **b**, a natural strategy might be to get $\hat{y}_i$ as close to y_i as possible for all observations. By (25) this implies "making the residuals as small as possible". A more popular criterion is to **minimize the sum of squared residuals**, which penalizes larger residuals relatively more.

See slide 2 of the PowerPoint slides "OLSregression" on our course web page, and also Greene Fig. 3.1 (p. 22). Denoting a candidate for **b** as $\mathbf{b_0}$, our optimization goal is thus given as

$$\begin{aligned} &\min_{\mathbf{b_0}} \mathbf{e'e} = \left(\mathbf{y} - \mathbf{Xb_0}\right)'\left(\mathbf{y} - \mathbf{Xb_0}\right) = \mathbf{y'y} - \mathbf{b_0'X'y} - \mathbf{y'Xb_0} + \mathbf{b_0'X'Xb_0} = \\ &\mathbf{y'y} - 2\mathbf{y'Xb_0} + \mathbf{b_0'X'Xb_0} \end{aligned} \tag{27}$$

We solve for **b** by deriving the First Order Conditions (FOC), using "matrix calculus" (see Matrix Algebra notes)

$$\begin{aligned} &\frac{\partial \mathbf{e'e}}{\partial \mathbf{b_0}} = -2\mathbf{X'y} + 2\mathbf{X'Xb_0} \\ &-2\mathbf{X'y} + 2\mathbf{X'Xb} = \mathbf{0} \end{aligned} \tag{28}$$

The second line is often referred to as the Least Squares "**Normal Equation**". If **X** is full rank, **X'X** will be symmetric and full rank as well, thus its inverse will exist, and we can solve (28) for $\mathbf{b_0}$ to obtain

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{29}$$

As shown in Greene p. 21, this solution is indeed a minimum and unique, since the second derivative yields a positive definite matrix. The solution is called the **Ordinary Least Squares** (OLS) estimator.

See script *mod1_2b* for an example of OLS using wage data.

Orthogonality and Projection

In Euclidean space, "orthogonaliy" between vectors **a** and **b** (some **b**, not our OLS estimator) arises if two vectors form a right angle. Mathematically, this implies that their inner product is zero, i.e. **a'b** = 0. Orthogonality is also possible between a matrix **X** and a conformable vector **a**. In that case, it implies that every column of **X** is orthogonal to **a**, and that **X'a** = **0**.

In contrast, *population orthogonality* between two variables implies that the two variables are perfectly *un*correlated for the entire population, in the sense that the *expectation* of their inner product is zero (essentially our Assumption 3 for the CLRM for any regressor **x** and the error term). However, for any finite set of draws for these variables (i.e. for a given sample), orthogonality in the Euclidean sense is unlikely to hold (though we would expect the inner product to be relatively close to zero).

Perfect (Euclidean) orthogonality between explanatory variables is generally a very desirable property in regression analysis. In reality, it hardly ever occurs (the best we can hope for is "low correlation"), but in experimental settings researches can 'design" orthogonal explanatory variables. We will visit the topic of "orthogonality" many times in this course.

For the OLS model, we have orthogonality between data matrix **X** and the residuals by construction. Starting with the normal equation, we have:

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = 0 \Rightarrow -\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right) = -\mathbf{X}'\mathbf{e} = \mathbf{0} \tag{30}$$

Intuitively, this makes a lot of sense - the residuals should only contain what **X** couldn't explain about **y**.

If the first column of **X** is a vector of "1's" (i.e. we are estimating a regression model with a constant term - the standard case, we gain a few more interesting insights from the normal equation and relationships that flow from it:

(i) The residuals sum to zero:

$$\mathbf{X}'\mathbf{e} = \begin{bmatrix}\mathbf{i} & \mathbf{c}_2 & \cdots & \mathbf{c}_\mathbf{k}\end{bmatrix}'\mathbf{e} = \begin{bmatrix}\mathbf{i}'\mathbf{e} \\ \mathbf{c}_2'\mathbf{e} \\ \vdots \\ \mathbf{c}_\mathbf{k}'\mathbf{e}\end{bmatrix} = \mathbf{0} \tag{31}$$

(ii) The regression hyperplane passes through the data means:

$$\mathbf{X'Xb} = \mathbf{X'y} \Leftrightarrow \begin{bmatrix} \mathbf{i}' \\ \mathbf{c}_2' \\ \vdots \\ \mathbf{c}_k' \end{bmatrix} \begin{bmatrix} \mathbf{i} & \mathbf{c_1} & \cdots & \mathbf{c_k} \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{i}' \\ \mathbf{c}_2' \\ \vdots \\ \mathbf{c}_k' \end{bmatrix} \mathbf{y}$$

1st equation:

$$\mathbf{i}' \begin{bmatrix} \mathbf{i} & \mathbf{c_1} & \cdots & \mathbf{c_k} \end{bmatrix} \mathbf{b} = \mathbf{i}'\mathbf{y} \Leftrightarrow \begin{bmatrix} n & \sum_{i=1}^{n} c_{i1} & \cdots & \sum_{i=1}^{n} c_{ik} \end{bmatrix} \mathbf{b} = \sum_{i=1}^{n} y_i \tag{32}$$

$$\Rightarrow \begin{bmatrix} 1 & \sum_{i=1}^{n} c_{i1} \Big/ n & \cdots & \sum_{i=1}^{n} c_{ik} \Big/ n \end{bmatrix} \mathbf{b} = \sum_{i=1}^{n} y_i \Big/ n \quad \Rightarrow \bar{\mathbf{x}}'\mathbf{b} = \bar{y} \quad \text{where } \bar{\mathbf{x}} = \begin{bmatrix} 1 & \sum_{i=1}^{n} c_{i1} \Big/ n & \cdots & \sum_{i=1}^{n} c_{ik} \Big/ n \end{bmatrix}'$$

Next, let's derive two important matrices in regression analysis:

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \mathbf{Xb} = \mathbf{y} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y} = \left(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\right)\mathbf{y} = \mathbf{My} \quad \text{and} \\ \hat{\mathbf{y}} &= \mathbf{y} - \mathbf{e} = \mathbf{y} - \mathbf{My} = \left(\mathbf{I} - \mathbf{M}\right)\mathbf{y} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y} = \mathbf{Py} \end{aligned} \tag{33}$$

In regression jargon, **M** is known as the "residual maker" and **P** is the "projection matrix". Note that both matrices are a function of **X**. So with each new **X**, we get a new **M** and **P**. Also, both matrices are symmetric and idempotent (means **M*****M** = **M**).

Intuitively, just remember that **M** turns **y** into "everything **X** couldn't explain", i.e. the residuals, and **P** turn **y** into "everything **X** can explain", i.e. the fitted values. Some useful relationships that follow:

$$\begin{aligned} &\mathbf{MX} = \mathbf{0} \\ &\mathbf{PM} = \mathbf{MP} = \mathbf{0} \\ &\mathbf{PX} = \mathbf{X} \\ &\mathbf{y} = \mathbf{Py} + \mathbf{My} \\ &\mathbf{e}'\mathbf{e} = \mathbf{e}'\mathbf{y} \\ &\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e} \end{aligned} \tag{34}$$

Partitioned Regression and Partial Regression

R script `mod1_2d`

Question: What computations are involved in obtaining, in isolation, a subset of the coefficients of a multiple regression model (while still controlling for all other effects)? In the early days of regression analysis, the resulting strategy of "partitioned regression" was an important "shortcut" to reduce computation time. Today, we show this primarily to gain further insights into the mechanics of OLS.

First, assume we have data matrix **X** and corresponding coefficient vector $\boldsymbol{\beta}$, but we are primarily interested in the effects of a subset of variables (columns) of **X**. Denote this subset as $\mathbf{X_2}$ and the remaining variables as $\mathbf{X_1}$. Conformably, also split the coefficient vector into 2 parts $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_1$. We can then write the CLRM as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{35}$$

Our objective is to find the OLS solution for $\mathbf{b_2}$. The normal equations can now be written as

$$\begin{bmatrix} \mathbf{X_1'X_1} & \mathbf{X_1'X_2} \\ \mathbf{X_2'X_1} & \mathbf{X_2'X_2} \end{bmatrix} \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1'y} \\ \mathbf{X_2'y} \end{bmatrix} \tag{36}$$

First, use the first set of equations to solve for $\mathbf{b_1}$ in terms of $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{y}$, and $\mathbf{b_2}$:

$$\mathbf{X_1'X_1b_1} + \mathbf{X_1'X_2b_2} = \mathbf{X_1'y} \Leftrightarrow \mathbf{b_1} = \left(\mathbf{X_1'X_1}\right)^{-1}\mathbf{X_1'}\left(\mathbf{y} - \mathbf{X_2b_2}\right) \tag{37}$$

Note that if all columns of $\mathbf{X_1}$ are orthogonal to all columns of $\mathbf{X_2}$ (a rare occurrence to say the least...), $\mathbf{X'_1X_2} = \mathbf{0}$, and we have the usual OLS normal equations, and $\mathbf{b_1} = \left(\mathbf{X_1'X_1}\right)^{-1}\mathbf{X_1'y}$. This implies that we could then estimate $\mathbf{b_1}$ simply by regressing $\mathbf{y}$ on $\mathbf{X_1}$. Intuitively, this further implies that there is no correlation between $\mathbf{X_1}$ and $\mathbf{X_2}$, so we don't have to "control" for the effects of $\mathbf{X_2}$ when estimating the effects of $\mathbf{X_1}$ on $\mathbf{y}$. (Note: However, we would still lose efficiency and explanatory power for our model if $\mathbf{X_2}$ has something to say about $\mathbf{y}$).

Combining (37) with the second set of normal equations from (36) yields the solution for $\mathbf{b_2}$:

$$\begin{aligned} &\left(\mathbf{X_2'X_1}\right)\mathbf{b_1} + \left(\mathbf{X_2'X_2}\right)\mathbf{b_2} = \mathbf{X_2'y} \quad \Rightarrow \\ &\left(\mathbf{X_2'X_1}\right)\left(\left(\mathbf{X_1'X_1}\right)^{-1}\mathbf{X_1'}\left(\mathbf{y} - \mathbf{X_2b_2}\right)\right) + \left(\mathbf{X_2'X_2}\right)\mathbf{b_2} = \mathbf{X_2'y} \quad \Rightarrow \\ &\mathbf{b_2} = \left(\mathbf{X_2'M_1X_2}\right)^{-1}\mathbf{X_2'M_1y} \qquad \text{where} \\ &\mathbf{M_1} = \mathbf{I} - \mathbf{X_1}\left(\mathbf{X_1'X_1}\right)^{-1}\mathbf{X_1'} \end{aligned} \tag{38}$$

$\mathbf{M_1}$ is the residual maker matrix in a regression of anything on $\mathbf{X_1}$. Note again that this reduces to the standard OLS formula if $\mathbf{X'_1X_2} = \mathbf{0}$. We can alternatively write the partitioned regression as

$$\begin{aligned} &\mathbf{b_2} = \left(\mathbf{X_2^{*\prime}X_2^*}\right)^{-1}\mathbf{X_2^{*\prime}y^*} \qquad \text{where} \\ &\mathbf{X_2^*} = \mathbf{M_1X_2}, \quad \mathbf{y^*} = \mathbf{M_1y} \end{aligned} \tag{39}$$

Thus, we can think of this as a two-step process: In step 1 we regress all columns of $\mathbf{X_2}$ on $\mathbf{X_1}$ and collect the residuals (i.e. get $\mathbf{M_1X_2}$). These residuals will then contain what's left of the information in $\mathbf{X_2}$ after netting out the effect of $\mathbf{X_1}$. We do the same for $\mathbf{y}$ by regressing it on $\mathbf{X_1}$ and collecting residuals. In step 2 we then regress the purged-of-$\mathbf{X_1}$-effects version of $\mathbf{y}$ on the purged-of-$\mathbf{X_1}$-effects version of $\mathbf{X_2}$ to get the pure effect of $\mathbf{X_2}$ on $\mathbf{y}$, i.e. $\mathbf{b_2}$**.** This is the famous "Frisch-Waugh Theorem" (Greene p. 28).

The process can easily extended to derive the partitioned result for an individual coefficient. In that case, just think of $\mathbf{X_2}$ as a single column (say "$\mathbf{z}$"), and of $\mathbf{b_2}$ as a scalar, say "c". We then get

$$c = \left(\mathbf{z^{*\prime}z^*}\right)^{-1}\mathbf{z^{*\prime}y^*} \qquad \text{where} \quad \mathbf{z^*} = \mathbf{M_1z}, \quad \mathbf{y^*} = \mathbf{M_1y} \tag{40}$$

If you consider $\mathbf{z}$ to be a new variable to be added to $\mathbf{X}$, you get the result in Corollary 3.2.1, p. 34.

R script `mod1_2d` illustrates the Partitioned Regression approach. Note that the estimated standard deviation for the error term, and thus the standard errors and t-values for the partitioned regression results are a bit different from those obtained in the full model. This difference lies solely in the (*n*-*k*) correction in the denominator for the *s2*-formula, since "*k*" differs between the full and partitioned model. Naturally,

this difference diminishes under increasing sample size, as *n* will vastly outweigh *k* in either case. (You can verify this by dropping the "minus *k*" correction in the *s2* formula for both models – the s.e.'s and t-values will now be identical)

Goodness of Fit and Analysis of Variance (ANOVA)

A special case of the residual-maker matrix $\mathbf{M_1}$ is the (n by n) *deviation-from-the-mean* matrix $\mathbf{M_0}$:

$$\mathbf{M_0} = \mathbf{I} - \mathbf{i}\left(\mathbf{i}'\mathbf{i}\right)^{-1}\mathbf{i}' \tag{41}$$

where **I** is the n by n identity matrix and **i** is an n by 1 vector of ones. Intuitively, this matrix "regresses" any vector, say **x**, or matrix, say **X**, against a column of ones and captures the resulting residuals. These residuals will simply be the deviation of each observation in **x** or **X** from its respective column mean. Example:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \mathbf{M_0 x} = \begin{bmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{bmatrix} = \mathbf{x} - \mathbf{i}\overline{x}$$
$$\mathbf{X} = \begin{bmatrix} \mathbf{c_1} & \mathbf{c_2} & \cdots & \mathbf{c_k} \end{bmatrix} \quad \mathbf{M_0 X} = \begin{bmatrix} \mathbf{c_1} - \mathbf{i}\overline{c_1} & \mathbf{c_2} - \mathbf{i}\overline{c_2} & \cdots & \mathbf{c_k} - \mathbf{i}\overline{c_k} \end{bmatrix} \tag{42}$$

We can now compactly write our CLRM in "deviation from the mean" form as

$$\mathbf{M_0 y} = \mathbf{M_0 X b} + \mathbf{M_0 e} = \mathbf{M_0 X b} + \mathbf{e} \tag{43}$$

The second equality follows from the fact that the mean of the residual vector is zero, i.e. $\mathbf{M_0 e} = \mathbf{e}$ (IFF the model includes an intercept).

Noting that it then follows that $\mathbf{e'M_0 X} = \mathbf{e}'\mathbf{X} = \mathbf{0}$, we can derive the sum of squared differences of the elements of **y** from their mean, known as "Total Sum of Squares" (SST) as

$$\left(\mathbf{M_0 y}\right)'\left(\mathbf{M_0 y}\right) = \mathbf{y}'\mathbf{M_0 y} = \mathbf{b}'\mathbf{X}'\mathbf{M_0 X b} + \mathbf{e}'\mathbf{M_0 X b} + \mathbf{b}'\mathbf{X}'\mathbf{M_0 e} + \mathbf{e}'\mathbf{M_0 e} =$$
$$\mathbf{b}'\mathbf{X}'\mathbf{M_0 X b} + \mathbf{e}'\mathbf{e} = \hat{\mathbf{y}}'\mathbf{M_0}\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e} \tag{44}$$

If the regression includes an intercept (= constant term), we have $\overline{\hat{y}} = \overline{y}$ and $\hat{\mathbf{y}} - \mathbf{i}\overline{\hat{y}} = \hat{\mathbf{y}} - \mathbf{i}\overline{y}$, and we can interpret (44) as Total Sum of Squares = Regression Sum of Squares + Error Sum of Squares, or *SST* = *SSR* + *SSE*. Note: The last term is also known as "sum of squared residuals", but since the resulting acronym would also be SSR, we'll stick with Greene's terminology of "Error Sum of Squares".

Intuition: Recall that the fundamental aim of regression analysis is to explain observed variation in **y** via observed variation in **X**. Equation (44) says that we can break down the variation in **y** into two components: A part that's explained by the variation of **X** (SSR), and a part that our regression cannot explain (SSE). It is thus natural to use the ratio of SSR / SST as Goodness-of-Fit measure for our regression model. This is the "*Coefficient of Determination*", better known as "R^2".

$$R^2 = SSR / SST = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M_0}\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{M_0}\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M_0}\mathbf{y}} \tag{45}$$

This statistic lies between 0 and 1. A value of "0" implies that **X** has nothing to say about **y**. In a 2-dimensional model, this would be a flat regression line through the mean of **y**. A value of 1 implies a "perfect fit" - **y** actually lies in the hyperplane described by the columns of **X**.

As discussed in more detail in Greene, Ch. 3, a problem with this fit measure is that it never declines as more explanatory variables, even meaningless ones, are added to the model. In other words, the conventional R^2 measure doesn't penalize for superfluous regressors, or, alternatively, doesn't reward for parsimony. We therefore prefer the "Adjusted R^2" given as

$$\bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e} / (n-k)}{\mathbf{y}'\mathbf{M_0}\mathbf{y} / (n-1)} = 1 - \frac{\mathbf{e}'\mathbf{e} \cdot (n-1)}{\mathbf{y}'\mathbf{M_0}\mathbf{y} \cdot (n-k)} \tag{46}$$

Thus, as the number of regressors (including intercept) *k* increases relative to sample size *n*, the last term increases, and the adjusted fit deteriorates. The ordinary and adjusted R^2 are related as follows:

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k}\right)\left(1 - R^2\right) \tag{47}$$

Note that the preceding derivations break down when the regression model does not include a constant term. For models without constant term, the interpretation of R^2 becomes ambiguous, and this measure should not be used.

Instead, you may want to consider the following alternative goodness-of-fit measures, the *Akaike Information Criterion (AIC)* and the Schwartz or *Bayesian Information Criterion (BIC)*. Both work for models without constant term and, for that matter, also for nonlinear regression models. They also reward parsimony. They are usually given in log-form as follows:

$$\log AIC = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2K}{n} \qquad \log BIC = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{K\log(n)}{n} \tag{48}$$

(see Greene ch. 5 for more details). Note that both statistics DECLINE as model fit improves. (So a smaller number implies a better fit).

Some software packages produce a summary of squared deviations in a table accompanying the main regression results. This table is often referred to as "Analysis of Variance" (ANOVA) Table. In STATA for example, "Model" SS means SSR, "Residual" SS means SSE, and "Total" SS means SST.

See script `mod1_2d` for basic goodness-of-fit derivations.

Finally, if you wish to use $\bar{R}^2$ to choose between two models, the following must hold:

1. The vector of dependent observations must be identical for both models. (So you can't compare a model that is linear in **y** to one that uses, say, log **y**. Also, sample sizes must be identical)

2. The models have to be linear-in-parameters. For nonlinear regression models use AIC, BIC or related measures.

3. Both models must have a constant term, and the mean of the error term in the population model must be zero.