Linear Regression

Definition 1. A **linear regression** of dependent variable y_t on independent variables $(x_{t_1}, \ldots, x_{t_k})$ is

$$y_t = \beta_1 x_{t_1} + \ldots + \beta_k x_{t_k} + w_t,$$

where $\beta_1, \ldots, \beta_k$ are unknown and fixed regression coefficients and w_t is an iid random error process with zero mean and variance σ_w^2. (Note that often $x_t = 1$ so that β_1 is the intercept of the line.)

Let $\mathbf{x} \equiv (x_{t_1}, \ldots, x_{t_k})'$ and $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_k)'$. The regression can then be written more compactly as

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + w_t.$$

Definition 2. The **error** of a regression is defined as

$$w_t \equiv y_t - \boldsymbol{\beta}' \mathbf{x}_t.$$

In words, the error is the difference between the actual value and that predicted by the model.

Definition 3. The **sum of squared errors (SSE)** is

$$\text{SSE} \equiv \sum_{t=1}^{n} w_t^2 = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2,$$

which gives an overall measure of the difference between the data and the regression line.

Remark 1. A natural way of estimating β coefficients is by choosing values that minimize SSE (the best estimate makes the fewest aggregate errors), which is called **ordinary least squares (OLS)**. Ergo

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{n} w_t^2 = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{n} (y_t - \boldsymbol{\beta}' \mathbf{x}_t)^2.$$

Remark 2. We have n data points. Let x_{tk} denote the tth observation for the kth regressor. We have the system of n equations

$$\begin{aligned}
y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + w_1, \\
y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + w_2, \\
&\vdots \\
y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + w_n.
\end{aligned}$$

This system can be compactly expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w},$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}.$$

Remark 3. The sum of squared residuals can be expressed

$$\text{SSE} \equiv \underbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'}_{1 \times n} \underbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}_{n \times 1}.$$

When using matrices, it is helpful to consider the dimensionality of the object. SSE is a number, and here we can see that we end up with a 1×1 matrix. If we instead tried $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'$, we'd end up with an $n \times n$ object and therefore would not have SSE.

Remark 4. The OLS problem becomes

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &\equiv \arg\min_{\boldsymbol{\beta}} \ \underbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'}_{1 \times n} \underbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}_{n \times 1} \\
&= \arg\min_{\boldsymbol{\beta}} \ \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.
\end{aligned}$$

which gives the (1×1) objective function. Differentiation with respect to $\boldsymbol{\beta}$ and allows us to find the critical values that minimize SSE. But differentiating with respect to matrices requires our attention.

Matrix Differentiation

Definition 4. For column vector $\mathbf{y}$ of length n and column vector $\mathbf{x}$ of length k, we define

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \equiv \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \ldots & \frac{\partial y_1}{\partial x_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \ldots & \frac{\partial y_n}{\partial x_k} \end{bmatrix}.$$

Remark 5. The following proofs will assume that $n = k = 2$ just to make things easier. Therefore

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Showing more general results is a straightforward extension of what follows.

Proposition 1. *For* $\mathbf{x}$, $\mathbf{y}$, *and* $\mathbf{A}$ *as previously defined,*

$$\frac{d}{d\mathbf{x}}[\mathbf{y}'\mathbf{A}\mathbf{x}] = \mathbf{y}'\mathbf{A}.$$

Proof. Expanding $\mathbf{y}'\mathbf{A}\mathbf{x}$ yields

$$\mathbf{y}'\mathbf{A}\mathbf{x} = y_1 a_{11} x_1 + y_1 a_{12} x_2 + y_2 a_{21} x_1 + y_2 a_{22} x_2.$$

Because this object has only one row, we know from appealing to definition (4) that our final object will be a row vector of $k = 2$ derivatives, specifically

$$\frac{\partial \mathbf{y}'\mathbf{A}\mathbf{x}}{\partial x_1} = y_1 a_{11} + y_2 a_{21},$$
$$\frac{\partial \mathbf{y}'\mathbf{A}\mathbf{x}}{\partial x_2} = y_1 a_{12} + y_2 a_{22}.$$

Therefore with matrices, we have

$$\begin{aligned} \frac{d}{d\mathbf{x}}[\mathbf{y}'\mathbf{A}\mathbf{x}] &= \begin{bmatrix} y_1 a_{11} + y_2 a_{21} & y_1 a_{12} + y_2 a_{22} \end{bmatrix} \\ &= \begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \\ &= \mathbf{y}'\mathbf{A}. \end{aligned}$$

□

Proposition 2. *For* $\mathbf{x}$, $\mathbf{y}$, *and* $\mathbf{A}$ *as previously defined,*

$$\frac{d}{d\mathbf{y}}[\mathbf{y}'\mathbf{A}\mathbf{x}] = \mathbf{x}'\mathbf{A}'$$

Proof. Expanding $\mathbf{y}'\mathbf{A}\mathbf{x}$ yields

$$\mathbf{y}'\mathbf{A}\mathbf{x} = y_1 a_{11} x_1 + y_1 a_{12} x_2 + y_2 a_{21} x_1 + y_2 a_{22} x_2.$$

Now we differente with respect to y_1 and y_2, giving

$$\frac{\partial \mathbf{y}'\mathbf{A}\mathbf{x}}{\partial y_1} = a_{11} x_1 + a_{12} x_2,$$
$$\frac{\partial \mathbf{y}'\mathbf{A}\mathbf{x}}{\partial y_2} = a_{21} x_1 + a_{22} x_2.$$

Therefore with matrices, we have

$$\begin{aligned} \frac{d}{d\mathbf{x}}[\mathbf{y}'\mathbf{A}\mathbf{x}] &= \begin{bmatrix} a_{11} x_1 + a_{12} x_2 & a_{21} x_1 + a_{22} x_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \\ &= \mathbf{x}'\mathbf{A}' \end{aligned}$$

□

Proposition 3. *For* $\mathbf{x}$, $\mathbf{y}$, *and* $\mathbf{A}$ *as previously defined,*

$$\frac{d}{d\mathbf{x}}[\mathbf{x}'\mathbf{A}\mathbf{x}] = \mathbf{x}'(\mathbf{A} + \mathbf{A}').$$

Proof. Expanding $\mathbf{x}'\mathbf{A}\mathbf{x}$ yields

$$\mathbf{x}'\mathbf{A}\mathbf{x} = x_1^2 a_{11} + x_1 x_2 a_{12} + x_1 x_2 a_{21} + x_2^2 a_{22}.$$

Differentiate with respect to x_1 and x_2, which gives

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_1} = 2x_1 a_{11} + x_2(a_{12} + a_{21})$$
$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_2} = x_1(a_{12} + a_{21}) + 2x_2 a_{22}.$$

Therefore with matrices, we have $d[\mathbf{x}'\mathbf{A}\mathbf{x}]/d\mathbf{x}$ equal to

$$\begin{aligned} &\begin{bmatrix} 2x_1 a_{11} + x_2(a_{12} + a_{21}) & x_1(a_{12} + a_{21}) + 2x_2 a_{22} \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2a_{11} & a_{12} + a_{21} \\ a_{12} + a_{21} & 2a_{22} \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \right) \\ &= \mathbf{x}'(\mathbf{A} + \mathbf{A}'). \end{aligned}$$

□

OLS Solution

Remark 6. Keeping in mind that $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, the preceding properties of matrix differentiation can be applied respectively to yield

$$\begin{aligned} \frac{d}{d\boldsymbol{\beta}}[-\mathbf{y}'\mathbf{X}\boldsymbol{\beta}] &= -\mathbf{y}'\mathbf{X}, \\ \frac{d}{d\boldsymbol{\beta}}[-\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}] &= -\mathbf{y}'\mathbf{X}, \\ \frac{d}{d\boldsymbol{\beta}}[\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}] &= \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + [\mathbf{X}'\mathbf{X}]') = 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}. \end{aligned}$$

Remark 7. Ergo the first-order condition (after a transpose) is

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.$$

Assuming that $\mathbf{X}'\mathbf{X}$ is nonsingular, we can premultiply both sides by $(\mathbf{X}'\mathbf{X})^{-1}$ to get

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Note that because $\mathbf{X}$ is not in general a square matrix, we cannot simplify further by distributing the inverse.

Remark 8. The minimized sum of squared errors can therefore be expressed as

$$\text{SSE}^* = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Proposition 4. *Supposing that the model is correctly specified and the error term is uncorrelated with* $\mathbf{X}$ *(i.e. zero conditional mean), the OLS estimates will be unbiased, i.e.* $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Proof. This is because

$$\begin{aligned}
E[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{w})|\mathbf{X}] \\
&= \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w})|\mathbf{X}] \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{w}|\mathbf{X}] \\
&= \boldsymbol{\beta},
\end{aligned}$$

recalling that $E[\mathbf{w}|\mathbf{X}] = 0$ as the zero conditional mean assumption. Since $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$ clearly does not depend on $\mathbf{X}$, we conclude that

$$E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

and unbiasedness is established. □

Remark 9. If errors are homoskedastic and independent, then OLS estimates will the best linear unbiased estimators (BLUE).

Proposition 5. *The variance-covariance matrix of* $\hat{\boldsymbol{\beta}}$ *is given by*

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma_w^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Proof. Because $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, covariance directly gives

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = E[\underbrace{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}_{k\times 1}\underbrace{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'}_{1\times k}].$$

Note the dimensionality: we have k regressors, so we want a $k \times k$ covariance matrix. Let's use $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}$ to instead write

$$\begin{aligned}
\text{Cov}(\hat{\boldsymbol{\beta}}) &= E\big[\big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}\big)\big((\mathbf{X}'\mathbf{X})^{-1})\mathbf{X}'\mathbf{w}\big)'\big] \\
&= E\big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}\mathbf{w}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\big].
\end{aligned}$$

Treating $\mathbf{X}$ as data (i.e. a bunch of non-stochastic numbers), we can write

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\big[\mathbf{w}\mathbf{w}'\big]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

The error w_t has zero mean. Ergo the variance of error is $\text{Var}(\mathbf{w}) = E[\mathbf{w}\mathbf{w}']$, which looks like

$$E\begin{bmatrix} w_1^2 & w_1w_2 & \dots & w_1w_n \\ w_2w_1 & w_2^2 & \dots & w_2w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_nw_1 & w_nw_2 & \dots & w_n^2 \end{bmatrix}.$$

Because errors are uncorrelated, it follows that $E[w_sw_t] = 0$ when $s \neq t$, and $E[w_t^2] = \sigma_w^2$. Thus we can write simply $E[\mathbf{w}\mathbf{w}'] = \sigma_w^2\mathbf{I}_n$. Therefore

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma_w^2(\mathbf{X}'\mathbf{X})^{-1}$$ □

Remark 10. The error variance σ_w^2 has unbiased estimator given by the **mean squared error (MSE)**, i.e.,

$$s_w^2 \equiv \text{MSE} \equiv \frac{\text{SSE}}{n-k}.$$

Remark 11. Let $\mathbf{C} \equiv (\mathbf{X}'\mathbf{X})^{-1}$ and c_{ij} be the i, jth element of $\mathbf{C}$. Define the ***t*-statistic** for β_i to be

$$t_{n-k} \equiv \frac{\hat{\beta}_i - \beta_i}{s_w\sqrt{c_{ii}}}.$$

If w_t has normal distribution, then $t_{n-k} \sim t(n-k)$ distribution. If w_t does not have normal distribution, then the result is approximately true for large n.

Remark 12. We can jointly test the significance of several regressors by comparing the SSE of the full *unrestricted* model, call it SSE_u; to the SSE of a *restricted* model with q fewer regressors, call it SSR_r; according to the ***F*-statistic** defined as

$$F_{q,n-k} \equiv \frac{(\text{SSE}_\text{r} - \text{SSE}_\text{u})/q}{\text{SSE}_\text{u}/(n-k)},$$

because $F_{q,n-k} \sim F(q, n-k)$.

Remark 13. It can be shown that the maximum likelihood estimator for the variance of a regression with k variables is

$$\hat{\sigma}_k^2 = \frac{\text{SSE}_k}{n}$$

Note that because that every time an additional regressor is thrown in, SSE will (weakly) decrease, and this is true even if the additional regressor is junk. So a simple reduction in errors is not a good measure of whether an additional regressor is useful.

Instead, we might add another element: adding a regressor will reduce SSE, but does it reduce SSE by enough to reasonably conclude that it was helpful?