Lecture 38: Estimating Variance

We know that $\hat{q} = (X^T X)^{-1} X^T Y$ gives an unbiased estimator for q_0, where Y comes from the linear model

$$Y = X q_0 + \varepsilon,$$

with $\mathbb{E}(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma_0^2 \cdot I$.

Our goals for this lecture are:
- Compute $\text{Var}(\hat{q})$
- estimate σ_0^2.

Warm-up #1

Later it will prove helpful to work with the trace of a matrix, the sum of the diagonal entries.

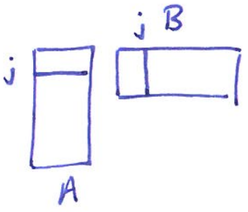
If $A \in \mathbb{R}^{n \times n}$, then

$$\text{trace}(A) = a_{1,1} + a_{2,2} + \cdots + a_{n,n}.$$

One of the most useful properties of the trace is that you can commute matrices under the trace:

If $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times n}$, then

$$\begin{aligned}\text{trace}(AB) &= \sum_{j=1}^{n} (AB)_{j,j} \\ &= \sum_{j=1}^{n} \sum_{k=1}^{p} a_{jk} b_{kj} \\ &= \sum_{k=1}^{p} \sum_{j=1}^{n} b_{kj} a_{jk} \\ &= \sum_{k=1}^{p} (BA)_{k,k} = \text{trace}(BA).\end{aligned}$$



So $\text{trace}(AB) = \text{trace}(BA)$.

Warm-up #2

$\hat{q} = (X^TX)^{-1}X^TY$: Let us focus on the matrix $(X^TX)^{-1}X^T$.

$$X^+ = (X^TX)^{-1}X^T = \left(\boxed{X^T}\,\boxed{X}\right)^{-1}\boxed{X^T}$$

$$= \boxed{(X^TX)^{-1}}\,\boxed{X^T} = p\,\overset{n}{\boxed{X^+}}$$

$X^+ \in \mathbb{R}^{p \times n}$ is the pseudoinverse of $X \in \mathbb{R}^{n \times p}$.

X^+ is a left-inverse of X: $\left(X^+X = (X^TX)^{-1}X^TX = I \in \mathbb{R}^{p \times p}\right)$

but it is not generally a right inverse (unless $n = p$):

$$H = \underset{\substack{\uparrow \\ \text{rank } p}}{X}\,\underset{\substack{\nwarrow \\ \text{rank } p}}{X^+} = X(X^TX)^{-1}X^T \in \mathbb{R}^{n \times n}$$

The product of two rank-p matrices cannot have rank larger than p. Since $H \in \mathbb{R}^{n \times n}$ is an $n \times n$ matrix of rank $\le p$, it cannot equal $I \in \mathbb{R}^{n \times n}$ (a rank n matrix) unless $n = p$. (If $n = p$, then X is square and $X^+ = X^{-1}$. Usually we have $n >> p$: many more observations than quantities of interest ("QoIs").)

H still has some nice properties:

$$H^2 = X\underbrace{X^+X}_{=I}X^+ = XX^+ = H$$

So H is a projector. It is also symmetric:

$$H^T = (X(X^TX)^{-1}X^T)^T = X^{TT}((X^TX)^{-1})^TX^T = X(X^TX)^{-1}X^T = H.$$

Aside....

Here we have used the fact that X^TX is symmetric and the inverse of a symmetric (invertible) matrix is also symmetric. (Here's a proof using "the spectral method" from the first part of the semester. Let $\lambda_1, .., \lambda_p$ be the eigenvalues of X^TX, with orthonormal eigenvectors $v_1, \dots, v_p$:

$$X^TX = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_p \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \begin{bmatrix} - v_1^T - \\ \vdots \\ - v_p^T - \end{bmatrix} := V\Lambda V^T$$

Then $$(V\Lambda V^T)(V\Lambda^{-1}V^T) = V\Lambda \underbrace{V^TV}_{=I \text{ by orthonormality}} \Lambda^{-1}V^T = V\underbrace{\Lambda\Lambda^{-1}}_{=I}V^T$$
$$= \underset{\substack{\uparrow \\ \text{by orthonormality.}}}{VV^T = I}$$

Hence $V\Lambda^{-1}V^T = (X^TX)^{-1}$. Now

$$[(X^TX)^{-1}]^T = (V\Lambda^{-1}V^T)^T = V(\Lambda^{-1})^TV^T \underset{\substack{\uparrow \\ \Lambda^{-1} \text{ is} \\ \text{diagonal}}}{=} V\Lambda^{-1}V^T = (X^TX)^{-1},$$

So $(X^TX)^{-1}$ is symmetric.

Variance of $\hat{q}$

Now we can determine how the noise ε that pollutes the observations Y filters through to affect the unbiased estimator $\hat{q}$.

$$\text{Var}(\hat{q}) = \mathbb{E}((\hat{q} - \mathbb{E}(\hat{q}))(\hat{q} - \mathbb{E}(\hat{q}))^T)$$ Using the definition of Variance of a vector variable

$$= \mathbb{E}((\hat{q} - q_0)(\hat{q} - q_0)^T)$$

Since $\mathbb{E}(\hat{q}) = q_0$.

Substitute in our formula for $\hat{q}$:

$$
\begin{aligned}
\hat{q} - q_0 &= X^+ Y - q_0 \\
&= X^+ (X q_0 + \varepsilon) - q_0 \\
&= \underbrace{X^+ X}_{= I} q_0 + X^+ \varepsilon - q_0 \\
&= q_0 + X^+ \varepsilon - q_0 = X^+ \varepsilon .
\end{aligned}
$$

$$Y = X q_0 + \varepsilon$$

Thus

$$
\begin{aligned}
\mathrm{Var}(\hat{q}) &= \mathbb{E}\left((\hat{q} - q_0)(\hat{q} - q_0)^T \right) \\
&= \mathbb{E}\left((X^+ \varepsilon)(X^+ \varepsilon)^T \right) \\
&= \mathbb{E}\left(X^+ \varepsilon \varepsilon^T (X^+)^T \right) \\
&= X^+ \, \mathbb{E}(\varepsilon \varepsilon^T) (X^+)^T
\end{aligned}
$$

Where this last step used the linearity of the expected value and the fact that X is a constant matrix.

Now $\mathbb{E}(\varepsilon \varepsilon^T) = \mathbb{E}\left((\varepsilon - \underset{=0}{\mathbb{E}(\varepsilon)})(\varepsilon - \underset{=0}{\mathbb{E}(\varepsilon)})^T \right) = \mathrm{Var}(\varepsilon)$

$= \sigma_0^2 I.$

Thus

$$
\begin{aligned}
\mathrm{Var}(\hat{q}) &= X^+ (\sigma_0^2 I)(X^+)^T \\
&= \sigma_0^2 X^+ (X^+)^T
\end{aligned}
$$

Look at

$$
\begin{aligned}
X^+ (X^+)^T &= (X^T X)^{-1} X^T \left((X^T X)^{-1} X^T \right)^T \\
&= (X^T X)^{-1} X^T \underbrace{X^{T^T}}_{X} \Big(\underbrace{(X^T X)^{-1}}_{\text{symmetric}} \Big)^T \\
&= (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= (X^T X)^{-1} .
\end{aligned}
$$

Thus we can conclude

$$\mathrm{Var}(\hat{q}) = \underset{\substack{\uparrow \\ \mathrm{Var}(\varepsilon) = \sigma_0^2 I}}{\sigma_0^2} \; \underset{\substack{\uparrow \\ \text{Scaling} \\ \text{factor.}}}{(X^T X)^{-1}}$$

So $(X^T X)^{-1}$ is a "magnification factor" that describes how much the Variance of the noise ε affects $\hat{q}$.

Unbiased estimate for σ_0^2.

Now we would like to estimate this variance factor σ_0^2. Our route to this estimate travels through the <u>residual</u> <u>vector</u> $\hat{R} = Y - X\hat{q}$, which is the mismatch between our observations Y and our unbiased estimate $X\hat{q}$ for those observations.

$$\begin{aligned} \hat{R} = Y - X\hat{q} &= Y - XX^{+}Y \\ &= (I - XX^{+})Y \\ &= (I-H)Y \\ &= (I-H)(Xq_0 + \varepsilon) \\ &= (I-H)Xq_0 + (I-H)\varepsilon \end{aligned}$$

Note that

$$\begin{aligned} (I-H)Xq_0 = Xq_0 - HXq_0 &= Xq_0 - X\underbrace{(X^TX)^{-1}X^TX}_{=I}q_0 \\ &= Xq_0 - Xq_0 = 0. \end{aligned}$$

Hence

$$\hat{R} = \underbrace{(I-H)Xq_0}_{=0} + (I-H)\varepsilon = (I-H)\varepsilon.$$

Thus, we can compute an expected value for the square of the norm of the mismatch:

$$\begin{aligned}
\mathbb{E}(\|\hat{R}\|^2) &= \mathbb{E}(\hat{R}^T\hat{R}) \\
&= \mathbb{E}(((I-H)\varepsilon)^T((I-H)\varepsilon)) \\
&= \mathbb{E}(\varepsilon^T \underbrace{(I-H)^T}_{= I-H \quad (H=H^T)} (I-H)\varepsilon) \\
&= \mathbb{E}(\varepsilon^T \underbrace{(I-H)(I-H)}\varepsilon) \\
&\qquad (I-H)(I-H) = I-H-H+\underbrace{H^2}_{=H} = I-H-H+H = I-H \\
&= \mathbb{E}(\varepsilon^T(I-H)\varepsilon)
\end{aligned}$$

Let $B = I - H \in \mathbb{R}^{n\times n}$. Then

$$\varepsilon^T(I-H)\varepsilon = \varepsilon^T B \varepsilon = \sum_{j=1}^{n}\sum_{k=1}^{n} \varepsilon_j b_{j,k} \varepsilon_k$$

So by linearity of the expected value,

$$\begin{aligned}
\mathbb{E}(\|\hat{R}\|^2) &= \mathbb{E}(\varepsilon^T B\varepsilon) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} b_{jk}\,\mathbb{E}(\varepsilon_j\varepsilon_k) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} b_{jk}\,\mathbb{E}(\varepsilon_j\varepsilon_k - \underbrace{\mathbb{E}(\varepsilon_j)}_{=0}\underbrace{\mathbb{E}(\varepsilon_k)}_{=0}) \quad \text{since } \mathbb{E}(\varepsilon)=0 \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} b_{j,k}\,\mathrm{Cov}(\varepsilon_j,\varepsilon_k)
\end{aligned}$$

Recall that $\mathrm{Var}(\varepsilon) = \sigma_0^2 I$

which means $\mathrm{Cov}(\varepsilon_j, \varepsilon_k) = \begin{cases} \sigma_0^2, & j = k \\ 0, & j \neq k \end{cases}$

Hence
$$\mathbb{E}(\|\hat{R}\|^2) = \sum_{j=1}^{n} \sum_{k=1}^{n} b_{j,k} \mathrm{Cov}(\varepsilon_j, \varepsilon_k)$$
$$= \sum_{j=1}^{n} b_{j,j}\, \sigma_0^2$$
$$= \sigma_0^2 \,\mathrm{trace}(B).$$

Now
$$\mathrm{trace}(B) = \mathrm{trace}(I - H)$$
$$= \underbrace{\mathrm{trace}(I)}_{I \in \mathbb{R}^{n\times n}} - \underbrace{\mathrm{trace}(H)}_{H = XX^+}$$

} trace = sum of diagonals is linear

$$= n - \mathrm{trace}(XX^+)$$

KEY: use Warm-up #1

$$= n - \mathrm{trace}(\underbrace{X^+X}_{= I \in \mathbb{R}^{p\times p}})$$
$$= n - p$$

$${}_{p}\overset{n}{\boxed{\;X^+\;}}\;{}_{n}\overset{p}{\boxed{X}} = X^T X \in \mathbb{R}^{p\times p}$$

Thus
$$\mathbb{E}(\|\hat{R}\|^2) = \sigma_0^2 \,\mathrm{trace}(B)$$
$$= \sigma_0^2 (n-p).$$

Define $\hat{\sigma}_0^2 = \dfrac{\|\hat{R}\|^2}{n-p}$.

Then
$$\mathbb{E}(\hat{\sigma}_0^2) = \frac{\mathbb{E}(\|\hat{R}\|^2)}{n-p} = \frac{\sigma_0^2(n-p)}{n-p} = \sigma_0^2.$$

Thus $\hat{\sigma}_0^2$ is an unbiased estimator for σ_0^2.

In summary:

① $\hat{q} = X^+ Y = (X^TX)^{-1}X^TY$ = unbiased estimator for q_0.

② $\text{Var}(\hat{q}) = \sigma_0^2 (X^TX)^{-1}$

③ $\hat{\sigma}_0^2 = \frac{\|\hat{R}\|^2}{n-p} = \frac{\| Y - X\hat{q} \|^2}{n-p}$ = unbiased estimator for σ_0^2.

See ls_demo1.m and ls_demo2.m on the website for examples with

$$X = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 \end{bmatrix} \in \mathbb{R}^{n\times 3}.$$

One last question: What if $n=p$? Does ③ break down?

In this case, $X^+ = X^{-1}$, so $\hat{q} = X^{-1}Y$, and $\hat{R} = Y - XX^{-1}Y = 0$, so ③ reduces to a 0/0 indeterminate form; we learn nothing about σ_0^2 Since $\mathbb{E}(\|\hat{R}\|^2) = \mathbb{E}(0) = 0$.