

Geometry of least squares I

- Geometrically, the vector $\mathbf{y}$ is a point in the n dimensional sample space.
- The vector $\boldsymbol{\beta}$ is a point in the p dimensional *parameter space*., where $p = k + 1$.
- For each $\boldsymbol{\beta}$, $\mathbf{X}\boldsymbol{\beta}$ is a point in the sample space (i.e. the $n \times p$ matrix $\mathbf{X}$ is a mapping from the parameter space to the sample space).
- The values $\mathbf{X}\boldsymbol{\beta}$ form a p dimensional linear subspace or plane within the sample space. It is also assumed that $\mathbf{X}$ is of rank p .
- The sum of squares

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

is the squared length of the vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, or the squared distance between the observed $\mathbf{y}$ and the point $\mathbf{X}\boldsymbol{\beta}$.

Geometry of least squares II

- The method of least squares finds the value of β , called $\hat{\beta}$, for which the point $\mathbf{X}\hat{\beta}$ on the expectation surface is closest to the point $\mathbf{y}$.
- From geometric considerations, $\mathbf{X}\hat{\beta}$ must be the perpendicular projection of $\mathbf{y}$ on the expectation surface.

- The 'hat' matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, projects $\mathbf{y}$ onto the expectation surface at a point $\hat{\mathbf{y}}$ closest to $\mathbf{y}$.
- So $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \mathbf{X}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}] = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ is the least squares estimate of $\boldsymbol{\beta}$.
- Note that $\mathbf{H}\mathbf{X} = \mathbf{X}$. This means, for example, that any vector $\mathbf{v}$ in the column space of $\mathbf{X}$, $\mathbf{H}\mathbf{v} = \mathbf{v}$. As an example, $\mathbf{H}\mathbf{1} = \mathbf{1}$.
- Note that $\mathbf{H}$ is symmetric, and that $\mathbf{H}^2 = \mathbf{H}$.

Sampling distribution of $\mathbf{y}$

We assume that:

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

- elements of $\boldsymbol{\epsilon}$ are normally distributed with
- $\mathbf{E}(\boldsymbol{\epsilon}) = \mathbf{0}$
- $\mathbf{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$

equivalently

- elements of $\mathbf{y}$ are normally distributed with
- $\mathbf{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and
- $\mathbf{Cov}(\mathbf{y}) = \sigma^2 \mathbf{I}$

Sampling distribution of $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- elements of $\hat{\beta}$ are normally distributed, because linear combinations of normal random variables are normally distributed, with
- $E(\hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta$ and
-

$$\begin{aligned} Cov(\hat{\beta}) &= Cov((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Cov(\mathbf{y}) (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \end{aligned}$$

Sampling distribution of $\hat{\mathbf{y}}$

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

- elements of $\hat{\mathbf{y}}$ are normally distributed, because linear combinations of normal random variables are normally distributed, with

- $E(\hat{\mathbf{y}}) = \mathbf{H}E(\mathbf{y}) = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$

-

$$\begin{aligned} \text{Cov}(\hat{\mathbf{y}}) &= \text{Cov}(\mathbf{H}\mathbf{y}) = \mathbf{H}\text{Cov}(\mathbf{y})\mathbf{H}^T = \mathbf{H}\sigma^2\mathbf{I}\mathbf{H}^T \\ &= \sigma^2\mathbf{H}\mathbf{H}^T = \sigma^2\mathbf{H} = \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \end{aligned}$$

Sampling distribution of residuals

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

- elements of $\mathbf{r}$ are normally distributed, because linear combinations of normal random variables are normally distributed, with



$$\mathbf{E}(\mathbf{r}) = (\mathbf{I} - \mathbf{H})\mathbf{E}(\mathbf{y}) = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$



$$\begin{aligned}\text{Cov}(\mathbf{r}) &= \text{Cov}((\mathbf{I} - \mathbf{H})\mathbf{y}) = (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H}) \\ &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{I} - \mathbf{H})\end{aligned}$$

- The variance of the i 'th residual is $\sigma^2(1 - h_{ii})$ where h_{ii} is the i 'th diagonal element of $\mathbf{H}$. This means that the residuals will typically have different variances.

Covariance of residuals and fitted values



$$\begin{aligned}\text{Cov}(\mathbf{r}, \hat{\mathbf{y}}) &= \text{Cov}((\mathbf{I} - \mathbf{H})\mathbf{y}, \mathbf{H}\mathbf{y}) \\ &= \sigma^2(\mathbf{I} - \mathbf{H})\mathbf{H} = \mathbf{0}\end{aligned}$$

(Why?)

- which means that residuals and fitted values are uncorrelated
- which means that residuals and fitted values are independent (as uncorrelated normal random variables are independent).
- Using similar steps, it follows that the elements of $\hat{\beta}$ and the residuals are independent.

Estimation of σ^2

- The residuals are contained in the vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.
- The residual, or error, sum of squares is $SSE = \sum e_i^2 = \mathbf{e}^T \mathbf{e}$
- With k predictor variables, the error mean square is

$$MSE = \frac{SSE}{n - (k + 1)} = \frac{SSE}{n - p}$$

- $\hat{\sigma}^2 = MSE$ is an unbiased estimate of σ^2 because

$$\begin{aligned} E(SSE) &= E(\mathbf{e}^T \mathbf{e}) = \mathbf{0}^T \mathbf{0} + \sigma^2 \text{tr}(\mathbf{I}(\mathbf{I} - \mathbf{H})) = \sigma^2(\text{tr} \mathbf{I} - \text{tr}(\mathbf{H})) \\ &= \sigma^2(n - p) \end{aligned}$$

(Why?)

Confidence interval for $\mathbf{v}'\boldsymbol{\beta}$ I

- For a fixed vector $\mathbf{v}$, $\mathbf{v}'\boldsymbol{\beta}$ is estimated as $\mathbf{v}'\hat{\boldsymbol{\beta}}$.
- $\mathbf{v}'\hat{\boldsymbol{\beta}}$ is normally distributed with mean $\mathbf{v}'\boldsymbol{\beta}$ and variance

$$\sigma^2 \mathbf{v}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}$$

- Standardizing, it follows that

$$Z = \frac{\mathbf{v}'\hat{\boldsymbol{\beta}} - \mathbf{v}'\boldsymbol{\beta}}{\sigma \sqrt{\mathbf{v}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}}}$$

has a standard normal distribution.

Confidence interval for $\mathbf{v}'\beta$ II

- If Z has a standard normal distribution and is independent of W , which has a χ^2 distribution with m degrees of freedom, then

$$t = \frac{Z}{\sqrt{W/m}}$$

is DEFINED to have a t-distribution with m degrees of freedom, denoted as t_m .

- In notes on Cochran's theorem, we will see that $\frac{SSE}{\sigma^2}$ has a χ^2 distribution with $(n - p)$ degrees of freedom.
- It follows that

$$t = \frac{Z}{\sqrt{\frac{SSE}{\sigma^2(n-p)}}}$$

Confidence interval for $\mathbf{v}'\beta$ III

has a t distribution with $(n - p)$ degrees of freedom, denoted t_{n-p} .

- substituting for Z and noting that the estimate of σ is $\hat{\sigma} = \sqrt{MSE}$ it follows that

$$t = \frac{\mathbf{v}^T \hat{\beta} - \mathbf{v}^T \beta}{\hat{\sigma} \sqrt{\mathbf{v}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{v}}}$$

is distributed as t_{n-p} .

- Where $\pm t_{n-p, \alpha/2}$ are the upper and lower $\alpha/2$ 'th percentiles of t_{n-p} , by definition of a percentile,

$$P(-t_{n-p, \alpha/2} \leq t \leq t_{n-p, \alpha/2}) = 1 - \alpha$$

Confidence interval for $\mathbf{v}'\beta$ IV

- Rearranging the inequalities it follows that the interval with upper and lower endpoints

$$\mathbf{v}^T \hat{\beta} \pm t_{n-p, \alpha/2} \hat{\sigma} \sqrt{\mathbf{v}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{v}}$$

is a $100(1 - \alpha)\%$ confidence interval for $\mathbf{v}^T \beta$.

- Examples:
 - Let $\mathbf{v} = (1, 0, 0, \dots, 0)$ to form a $100(1 - \alpha)\%$ confidence interval for β_0 .
 - Let $\mathbf{v} = (0, 1, 0, \dots, 0)$ to form a $100(1 - \alpha)\%$ confidence interval for β_1 .
 - Let $\mathbf{v} = (0, 0, \dots, 0, 1)$ to form a $100(1 - \alpha)\%$ confidence interval for β_k .
 - Let $\mathbf{x}_0^T = (1, x_1, x_2, \dots, x_k)$. The mean of y when $\mathbf{x} = \mathbf{x}_0$, is $\mathbf{x}_0^T \beta$. A $100(1 - \alpha)\%$ confidence interval for the mean of y when $\mathbf{x} = \mathbf{x}_0$ is given by letting $\mathbf{v} = \mathbf{x}_0$.
 - By an appropriate choice of $\mathbf{v}$, the formula gives a confidence interval for any linear combination of the elements of β .