Multiple Linear Regression using Matrices - II I

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

Multiple Linear Regression using Matrices - II II

- The sum of squares

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

  is the squared length of the vector $\boldsymbol{y} - \boldsymbol{X}\beta$, or the squared distance between the observed $\boldsymbol{y}$ and the point $\boldsymbol{X}\beta$.
- The method of least squares finds the value of β, called $\hat{\beta}$, for which the point $\boldsymbol{X}\hat{\beta}$ on the expectation surface is closest to the point $\boldsymbol{y}$.
- From geometric considerations, $\boldsymbol{X}\hat{\beta}$ must be the perpendicular projection of $\boldsymbol{y}$ on the expectation surface.

Hat matrix

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

Sampling distribution of y

We assume that:

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

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

equivalently

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

Sampling distribution of $\hat{\beta}$

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

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

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

Sampling distribution of $\hat{y}$

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

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

$$Cov(\hat{\boldsymbol{y}}) = Cov(\boldsymbol{H}\boldsymbol{y}) = \boldsymbol{H}Cov(\boldsymbol{y})\boldsymbol{H}^T = \boldsymbol{H}\sigma^2\boldsymbol{I}\boldsymbol{H}^T$$
$$= \sigma^2\boldsymbol{H}\boldsymbol{H}^T = \sigma^2\boldsymbol{H} = \sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

Sampling distribution of residuals

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

- elements of $\boldsymbol{r}$ are normally distributed, because linear combinations of normal random variables are normally distributed, with
- $$\boldsymbol{E}(\boldsymbol{r}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{E}(\boldsymbol{y}p) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\beta = \boldsymbol{0}$$
- $$Cov(\boldsymbol{r}) = Cov((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}) = (\boldsymbol{I} - \boldsymbol{H})\sigma^2\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{H})$$
$$= \sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^T = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$
- The variance of the i'th residual is $\sigma^2(1 - h_{ii})$ where h_{ii} is the i'th diagonal element of $\boldsymbol{H}$. This means that the residuals will typically have different variances.

Covariance of residuals and fitted values

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

  (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).

Estimation of σ^2

- The residuals are contained in the vector $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$.
- The residual, or error, sum of squares is $SSE = \sum e_i^2 = \boldsymbol{e}^T \boldsymbol{e}$
- 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

$$\boldsymbol{E}(SSE) = \boldsymbol{E}(\boldsymbol{e}^T\boldsymbol{e}) = \boldsymbol{0}^T\boldsymbol{0} + \sigma^2 tr(\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{H}) = \sigma^2(tr\boldsymbol{I} - tr(\boldsymbol{H}))$$

$$= \sigma^2(n-p)$$

(Why?)

Confidence intervals I

- We will see that SSE is independent of $\hat{\beta}$, and
- where C_{jj} is the $\boldsymbol{j+1}$'st diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, the standard error of $\hat{\beta}_j$ is $\hat{\sigma}\sqrt{C_{jj}}$, and
- $t = \frac{\hat{\beta}_j - \beta j}{\hat{\sigma}\sqrt{C_{jj}}}$ has a t distribution with $n-(k+1) = n-p$ degrees of freedom.
- It follows that a $100(1-\alpha)\%$ confidence interval for β_j is given by

$$\hat{\beta}_j \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{C_{jj}}$$

- where $\boldsymbol{x}_0$ is a vector of covariate values, a $100(1-\alpha)\%$ confidence interval for the mean of y when $\boldsymbol{x} = \boldsymbol{x}_0$, $\mu_{y|\boldsymbol{x}_0} = E(y|\boldsymbol{x}_0)$, is given by

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

Prediction intervals (not responsible)

- In future, you plan to measure the value of y when the predictor variables equal $\boldsymbol{x}_0$.
- A point estimate of the future y is given by $\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.
- The goal is to find an interval in which the future y will lie, with probability $1 - \alpha$.
- A $100(1-\alpha)\%$ prediction interval for a future value of y when $\boldsymbol{x} = \boldsymbol{x}_0$ is given by

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

Simultaneous confidence intervals for β_j's

- When making more than one confidence interval, it is important to ensure that the joint coverage of all of the intervals constructed is $\geq 100(1-\alpha)$.
- If m intervals are constructed, with each interval having coverage probability $100(1-\alpha/m)$, it follows from Bonferroni's inequality that the probability that the m intervals will simulataneously cover their associated parameters is at least $1-\alpha$.
- For example, the $k+1$ intervals

$$\hat{\beta}_j \pm t_{\alpha/(2(k+1)),n-1-k}\sqrt{MSE}\sqrt{C_{j,j}},\ j=0,1,2,\ldots,k$$

have joint coverage of at least $100(1-\alpha)$.
In general, if constructing m simultaneous confidence intervals, just replace α by α/m, and use the usual procedure for the CI.

Bonferroni's inequality (Not responsbile for this material)

- Where E_1 and E_2 are events, and E^c is the complement of E, because the probability of the union is less than or equal to the sum of the probabilities, we know that

$$P(E_1^c \bigcup E_2^c) \leq P(E_1^c) + P(E_2^c)$$

- then use the fact that $P(E) = 1 - P(E^c)$, similarly $P(E_1^c) = 1 - P(E_1)$, and let $E = E_1^c \bigcup E_2^c$, from which $E^c = E_1 \bigcap E2$. Putting all of this together it follows that

$$P(E_1 \bigcap E_2) = 1 - P(E_1^c \bigcup E_2^c) \geq 1 - (P(E_1^c) + P(E_2^c))$$

- Letting $P(E_1) = 1 - \alpha/2$ and $P(E_2) = 1 - \alpha/2$, it follows that

$$P(E_1 \bigcap E_2) \geq 1 - (\alpha/2 + \alpha/2) = 1 - \alpha$$

Application to simultaneous inference - take this as a given.

- Let E_1 be the event that β_j is contained in the interval with endpoints $\hat{\beta}_j \pm t_{(\alpha/2)/2,n-p}\sqrt{MSE}\sqrt{C_{j+1,j+1}}$. The probability of this is $1-\alpha/2$.
- Let E_2 be the event that β_m is contained in the interval with endpoints $\hat{\beta}_m \pm t_{(\alpha/2)/2,n-p}\sqrt{MSE}\sqrt{C_{m+1,m+1}}$. The probability of this is $1-\alpha/2$.
- It follows from Bonferroni's inequality that the probability that EACH of β_j and β_m is simultaneously contained in their associated intervals is at least $1-\alpha$.
- We say that the two intervals form a joint $100(1-\alpha)\%$ confidence region for (β_j, β_m).
- What works for two the the β_j's, works for an arbitrary number, hence the form for the $k+1$ simultanous intervals given above, where α is replaced by $\alpha/(k+1)$.

Simultaneous confidence region for $\boldsymbol{\beta}$

- A joint $100(1-\alpha)\%$ confidence region for the vector $\boldsymbol{\beta}$ is given by those points $\boldsymbol{\beta}$ for which

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq p\,\text{MSE}\; F_{\alpha,p,n-p}$$

- This is a p dimensional ellipse centred at $\hat{\boldsymbol{\beta}}$.