Lecture 6

Multiple Linear Regression

Last week we learned

Elements of multivariate normal distribution

Matrix approach

Multiple linear regression

Multivariate normal distribution

- Let $z_i \sim N(\mu_i, \sigma_{ii}^2)$ and $\sigma_{ij}^2 = \mathrm{Cov}(z_i, z_j)$ for $1 \le i, j \le n$.

- The joint distribution of the z_i's is the *multivariate normal distribution*

$$\boldsymbol{z} \sim N_n(\boldsymbol{\mu}, V)$$

where

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \qquad V = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix}$$

- Let $\boldsymbol{w} = A\boldsymbol{z} + \boldsymbol{b}$ where A and $\boldsymbol{b}$ are a matrix and a vector of constants. Then

$$\mathbb{E}(\boldsymbol{w}) = A\,\mathbb{E}(\boldsymbol{z}) + \boldsymbol{b} = A\boldsymbol{\mu} + \boldsymbol{b} \qquad \mathrm{Var}(\boldsymbol{w}) = A\,\mathrm{Var}(\boldsymbol{z})A^T = A\,V\,A^T$$

In particular,

$$\boldsymbol{w} \sim N_n(A\boldsymbol{\mu} + \boldsymbol{b}, AVA^T)$$

Matrix form of the SLR model

- The SLR model in the matrix form

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The matrix X is called the *design matrix*

- The least squares estimate of $\boldsymbol{\beta}$ minimising the sum of squared errors, $\sum_{i=1}^{n} \varepsilon_i^2$, is

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

- The fitted values on the estimated regression line are

$$\hat{y} = X\hat{\boldsymbol{\beta}}$$

Lecture 6

Multiple Linear Regression

Aim: to understand the basics of MLR models

1. Least squares estimation
2. Properties of the least squares estimator
3. Estimating variance of the random error term
4. Confidence intervals

Multiple linear regression (MLR)

- A MLR model with p predictors $X_1, \ldots, X_p$ at levels $x_1, \ldots, x_p$ is

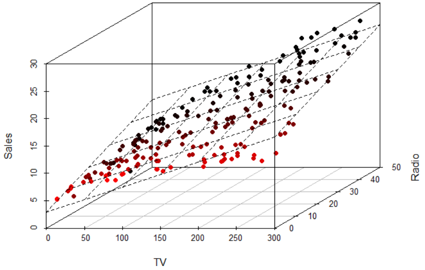
$$\mu = \mathbb{E}(Y|X = \boldsymbol{x}) = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p$$

- For each response y_i of Y we introduce a random error ε_i and write

$$y_i = \mathbb{E}(Y|X = \boldsymbol{x}_i) + \varepsilon_i = \mu_i + \varepsilon_i$$

- The standard assumption on the sampling distribution of ε_i is

$$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2) \qquad \Longrightarrow \qquad Y_i = \mu_i + \varepsilon_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$$



- Suppose that $y_1, y_2, \ldots, y_n$ are responses of Y at the values $x_{11}, x_{21}, \ldots, x_{n1}$ of the variable X_1, and at the values $x_{12}, x_{22}, \ldots, x_{n2}$ of the variable X_2, and so on...

		Predictors			
Observation, i	Response, y_i	X_1	X_2	$\ldots$	X_p
1	y_1	x_{11}	x_{12}	$\ldots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\ldots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\ldots$	x_{np}

- We then write

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \varepsilon_i$$

where ε_i are unknown random errors and $\beta_0, \beta_1, \ldots, \beta_p$ are (partial) regression coefficients that need to be estimated.

Least squares estimation

- The least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are the values of $\beta_0, \beta_1, \ldots, \beta_p$ which minimize the error sum of squares

$$SS_{\varepsilon} = \sum_{1 \le i \le n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p)^2$$

- We need to differentiate SS_{ε} with respect to each $\beta_0, \beta_1, \ldots, \beta_p$:

$$\begin{aligned}
\frac{\partial SS_{\varepsilon}}{\partial \beta_0} &= -2 \sum_{1 \le i \le n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p) \\
\frac{\partial SS_{\varepsilon}}{\partial \beta_1} &= -2 \sum_{1 \le i \le n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p) x_{i1} \\
\vdots \quad & \quad \vdots \\
\frac{\partial SS_{\varepsilon}}{\partial \beta_p} &= -2 \sum_{1 \le i \le n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p) x_{ip}
\end{aligned}$$

and require

$$\left.\frac{\partial SS_{\varepsilon}}{\partial \beta_0}\right|_{\beta_i = \hat{\beta}_i} = \left.\frac{\partial SS_{\varepsilon}}{\partial \beta_1}\right|_{\beta_i = \hat{\beta}_i} = \ldots = \left.\frac{\partial SS_{\varepsilon}}{\partial \beta_p}\right|_{\beta_i = \hat{\beta}_i} = 0$$

for all i at the same time.

- In practice, it is more convenient to use the matrix approach to MLR:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The matrix X is called the design matrix.

- The sum of errors squared is then

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

- The least squares estimate of $\boldsymbol{\beta}$ minimising the SS_{ε} is given by

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

Lecture 6

Multiple Linear Regression

Aim: to understand the basics of MLR models

1. Least squares estimation
2. Properties of the least squares estimator
3. Estimating variance of the random error term
4. Confidence intervals

Properties of the least squares estimator

- The standard assumptions on the sampling distribution of the random errors are

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$$

- Since $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the sampling distribution of responses is

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$$

- We want to determine the sampling distribution of the least squares estimator

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

- Perhaps a better notation for $\hat{\boldsymbol{\beta}}$, viewed as an estimator, is $\hat{\boldsymbol{B}}$, but we shall stick to the standard notation used in most textbooks.

Claim. The sampling distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 C)$$

where $C = (X^T X)^{-1}$.

Proof. We know that

$$\hat{\boldsymbol{\beta}} = C X^T \boldsymbol{Y} \quad \text{where} \quad \boldsymbol{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$$

Hence $\hat{\boldsymbol{\beta}}$ is also a multivariate normally distributed random variable with mean

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(CX^T\boldsymbol{Y}) = CX^T\,\mathbb{E}(\boldsymbol{Y}) = CX^T\,\mathbb{E}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = CX^TX\boldsymbol{\beta} = \boldsymbol{\beta}$$

and variance

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(CX^T\boldsymbol{Y}) &= CX^T\,\mathrm{Var}(\boldsymbol{Y})\,(CX^T)^T \\ &= CX^T\sigma^2 XC^T = \sigma^2 CC^{-1}C = \sigma^2 C \end{aligned}$$

Properties of the least squares estimator

- When $p = 1$, we should recover the results we found for the SLR model. In this case

$$\hat{\boldsymbol{\beta}} \sim N_2(\boldsymbol{\beta}, \sigma^2 C)$$

where

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad C = (X^T X)^{-1} = \frac{1}{s_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

- Using

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_1) \end{bmatrix} = \sigma^2 C$$

and $\frac{1}{n}\sum_{i=1}^n x_i^2 = \frac{1}{n} s_{xx} + \bar{x}^2$ we find

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right) \qquad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$$

which agrees with our earlier results.

Example

- A soft drink bottler is analysing the vending machine service routes in their distribution system:

  – the bottler is interested in **predicting the amount of time** required by the route driver to service the vending machines in an outlet;

  – the activity includes **stocking the machine** with beverage products and minor **maintenance or housekeeping**.

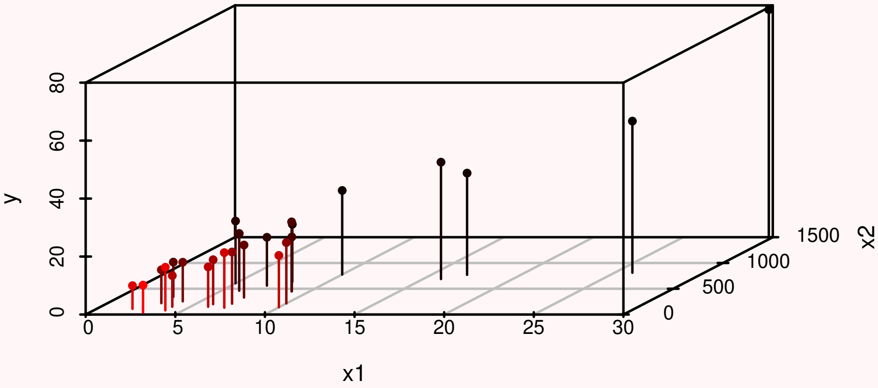
- The analyst responsible for the study has suggested that the two most important variables affecting the **delivery time** (y) are

  – the **number of cases** of product stocked (x_1), and

  – the **distance walked** by the route driver (x_2).

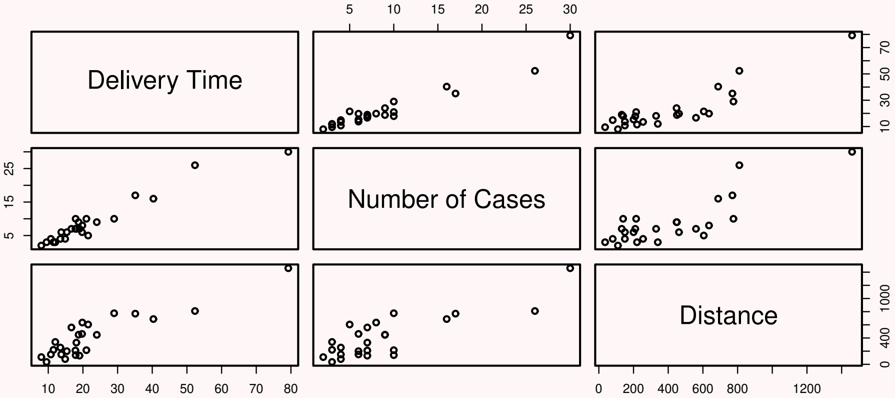
- The analyst has collected 25 observations on delivery time.

- 3D plot is often not the best way of displaying the data




```
> library(scatterplot3d)
> scatterplot3d( x1, x2, y, highlight.3d=TRUE, type="h" )
```

- Scatterplot matrix is the preferred way of visually inspecting the data




```
> pairs( y~x1+x2, labels = c("Delivery Time",
                           "Number of Cases", "Distance") )
```

- The model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \;\;\Rightarrow\;\; \boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{25} \end{bmatrix}$$

- We want to find the least squares estimate $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$:

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{bmatrix} \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10,232 & 133,899 & 6,725,688 \end{bmatrix}$$

$$X^T \boldsymbol{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{bmatrix} \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix} = \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix}$$

- The least-squares estimate $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ is then:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10,232 & 133,899 & 6,725,688 \end{bmatrix}^{-1} \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix}$$

$$= \begin{bmatrix} 0.11321518 & -0.00444859 & -0.00008367 \\ -0.00444859 & 0.00274378 & -0.00004786 \\ -0.00008367 & -0.00004786 & 0.00000123 \end{bmatrix} \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix}$$

$$= \begin{bmatrix} 2.34123115 \\ 1.61590712 \\ 0.01438483 \end{bmatrix}$$

- The fitted model is:

$$\hat{y} = 2.341 + 1.616\,x_1 + 0.014\,x_2$$

- $\boldsymbol{y}$ and X in R:

```
> y  <- c(16.68,11.50,12.03,14.88,...,10.75)  # delivery times
> x0 <- rep(1,length(y))                      # vector of 1's
> x1 <- c(7,3,3,4,6,7,...,4)                  # number of cases
> x2 <- c(560,220,340,80,150,330,...,150)     # distance walked
> X  <- cbind(x0, x1, x2)                     # design matrix
```

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

```
> ( hb <- solve( t(X) %*% X ) %*% t(X) %*% y ) # estimates
         [,1]
x0 2.34123115
x1 1.61590721
x2 0.01438483
```

Here `solve()` gives the inverse of a matrix, `t()` gives the transpose, and `%*%` is the matrix multiplication.

- The same result using the built-in linear model function `lm()`:

```
> lm( y ~ x1 + x2 ) # MLR model
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
    2.34123      1.61591      0.01438
```

Lecture 6

Multiple Linear Regression

Aim: to understand the basics of MLR models

1. Least squares estimation
2. Properties of the least squares estimator
3. Estimating variance of the random error term
4. Confidence intervals

Estimating variance of the random error term

- Recall that the sum of squared residuals

$$SS_E = \sum_{i=1}^{n} e_i^2$$

measures how closely the SLR model fits the data and can be used to estimate σ^2. The same is true for multiple linear regression.

- Let $\boldsymbol{e}$ denote the vector of residuals

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

where $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$ is the vector of fitted values (points on the fitted regression plane).

- Then the least squares estimator $\boldsymbol{E}$ of $\boldsymbol{e}$ is then given by

$$\boldsymbol{E} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$$

where $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is the random variable of responses and $\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}}$ is its least squares estimator.

Estimating variance of the random error term

- It can be shown that

$$\mathbb{E}(\boldsymbol{E}^T\boldsymbol{E}) = \sigma^2(n-p-1)$$

  Thus

$$\hat{\sigma}^2 \equiv MS_E = \frac{SS_E}{n-p-1} = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p-1}$$

  is an unbiased estimate of σ^2. The number

$$\nu_E = n-p-1$$

  is referred to as the residual degree of freedom. It is the number of linearly independent residuals e_i.

- The estimated variance $\hat{\sigma}^2$ is model dependent, therefore when comparing two models for the same data, we would usually choose the model with a smaller $\hat{\sigma}^2$.

Lecture 6

Multiple Linear Regression

Aim: to understand the basics of MLR models

1. Least squares estimation
2. Properties of the least squares estimator
3. Estimating variance of the random error term
4. Confidence intervals

Confidence interval on regression parameters

- We showed above that

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 C)$$

  This means that the marginal distribution of any estimator $\hat{\beta}_j$ of β_j is

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$$

  where c_{jj} is the jth diagonal element of the matrix $C = (X^T X)^{-1}$.

- Consequently,

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \sim t_{n-p-1}$$

  where

$$\mathrm{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 c_{jj}}, \qquad \hat{\sigma}^2 = MS_E = SS_E/(n-p-1)$$

- The $100(1-\alpha)\%$ confidence interval for any regression coefficient, β_j, is given by

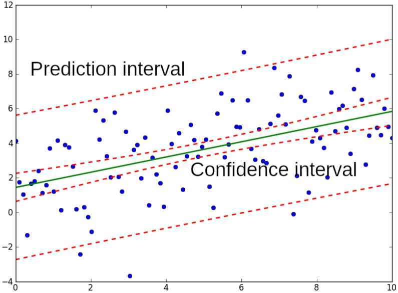
$$\mathrm{CI}(\beta_j) = \left[\hat{\beta}_j - t_{\alpha/2,n-p-1} \cdot \mathrm{se}(\hat{\beta}_j),\ \hat{\beta}_j + t_{\alpha/2,n-p-1} \cdot \mathrm{se}(\hat{\beta}_j)\right]$$

Confidence interval on mean response

- We want to construct a CI on the mean response

$$\mu_0 = \mathbb{E}(Y|X = \boldsymbol{x}_0) = \boldsymbol{x}_0^T \boldsymbol{\beta}$$

at a particular level $\boldsymbol{x}_0^T = (1, x_{01}, x_{02}, \ldots, x_{0p})$ of predictors $X_1, X_2, \ldots, X_p$.



Confidence interval on mean response

- The least squares estimator of μ_0 is

$$\hat{\mu}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + x_{01}\hat{\beta}_1 + \ldots + x_{0p}\hat{\beta}_p.$$

> **Claim.** The sampling distribution of $\hat{\mu}_0$ is
>
> $$\hat{\mu}_0 \sim N(\mu_0, \sigma^2 \boldsymbol{x}_0^T C \boldsymbol{x}_0).$$

Proof. Since $\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 C)$, we only need to compute the mean and variance:

$$\mathbb{E}(\hat{\mu}_0) = \mathbb{E}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T\, \mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T \boldsymbol{\beta} = \mu_0$$

$$\mathrm{Var}(\hat{\mu}_0) = \mathrm{Var}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T\, \mathrm{Var}(\hat{\boldsymbol{\beta}})\, \boldsymbol{x}_0 = \sigma^2 \boldsymbol{x}_0^T C \boldsymbol{x}_0$$

- The $100(1-\alpha)\%$ CI on the mean response μ_0 at the level $\boldsymbol{x}_0$ is given by

$$\mathrm{CI}(\mu_0) = \left[\hat{\mu}_0 - t_{\alpha/2,\, n-p-1} \cdot \mathrm{se}(\hat{\mu}_0),\ \hat{\mu}_0 + t_{\alpha/2,\, n-p-1} \cdot \mathrm{se}(\hat{\mu}_0)\right]$$

where $se(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0^T C \boldsymbol{x}_0}$.

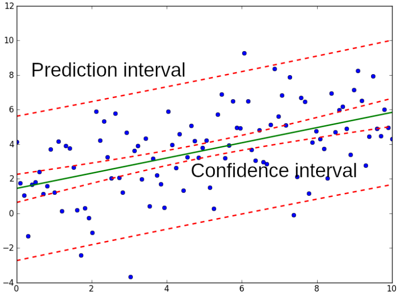
Prediction interval on a future observation

- Future observations are sampled from

$$\hat{Y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} + \varepsilon_0$$

where $\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2 C)$ is the least squares estimator of $\boldsymbol{\beta}$, and $\varepsilon_0 \sim N(0, \sigma^2)$ is the sampling distribution of the random errors.

- Prediction interval takes into account both the error from the estimated model and the error associated with a new observation.



Claim. The sampling distribution of $\hat{Y}_0$ is

$$\hat{Y}_0 \sim N(y_0, \sigma^2(1 + \boldsymbol{x}_0^T C \boldsymbol{x}_0))$$

Proof. Since $\hat{Y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} + \varepsilon_0$, we only need to compute the mean and variance:

$$\mathbb{E}(\hat{Y}_0) = \mathbb{E}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) + \mathbb{E}(\varepsilon_0) = \boldsymbol{x}_0^T \mathbb{E}(\hat{\boldsymbol{\beta}}) + 0 = \boldsymbol{x}_0^T \boldsymbol{\beta} = y_0$$

$$\mathrm{Var}(\hat{Y}_0) = \mathrm{Var}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) + \mathrm{Var}(\varepsilon_0) = \boldsymbol{x}_0^T \sigma^2 C \boldsymbol{x}_0 + \sigma^2 = \sigma^2(1 + \boldsymbol{x}_0^T C \boldsymbol{x}_0)$$

- The $100(1-\alpha)\%$ PI on a new observation y_0 at $\boldsymbol{x}_0$ is

$$\mathrm{PI}(y_0) = \left[\hat{y}_0 - t_{\alpha/2, n-p-1} \cdot \mathrm{se}(\hat{y}_0),\ \hat{y}_0 + t_{\alpha/2, n-p-1} \cdot \mathrm{se}(\hat{y}_0)\right]$$

where $\hat{y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ and $\mathrm{se}(\hat{y}_0) = \sqrt{\hat{\sigma}^2(1 + \boldsymbol{x}_0^T C \boldsymbol{x}_0)}$.

Summary

- The matrix form of the MLR model is

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The matrix X is called the design matrix.

- The least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = CX^T\boldsymbol{y} \quad \text{where} \quad C = (X^TX)^{-1}$$

- The least squares estimate of σ^2 is

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-p-1} = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p-1}$$

where $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ is the vector of residuals and $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$ is the vector of fitted values.

Summary

- The $100(1-\alpha)\%$ confidence interval for any regression coefficient, β_j, is given by

$$\text{CI}(\beta_j) = \left[\hat{\beta}_j - t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{\beta}_j),\ \hat{\beta}_j + t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{\beta}_j)\right]$$

where $\hat{\beta}_j = (CX^T\boldsymbol{y})_{jj}$ and $\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 c_{jj}}$

- A $100(1-\alpha)\%$ CI on the mean response μ_0 at the level $\boldsymbol{x}_0$ is given by

$$\text{CI}(\mu_0) = \left[\hat{\mu}_0 - t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{\mu}_0),\ \hat{\mu}_0 + t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{\mu}_0)\right]$$

where $\hat{\mu}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$ and $\text{se}(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2\boldsymbol{x}_0^T C\boldsymbol{x}_0}$.

- A $100(1-\alpha)\%$ PI on a new observation y_0 at $\boldsymbol{x}_0$ is

$$\text{PI}(y_0) = \left[\hat{y}_0 - t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{y}_0),\ \hat{y}_0 + t_{\alpha/2,n-p-1} \cdot \text{se}(\hat{y}_0)\right]$$

where $\hat{y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$ and $\text{se}(\hat{y}_0) = \sqrt{\hat{\sigma}^2(1 + \boldsymbol{x}_0^T C\boldsymbol{x}_0)}$.

Next week

Hypothesis testing