Lecture 1

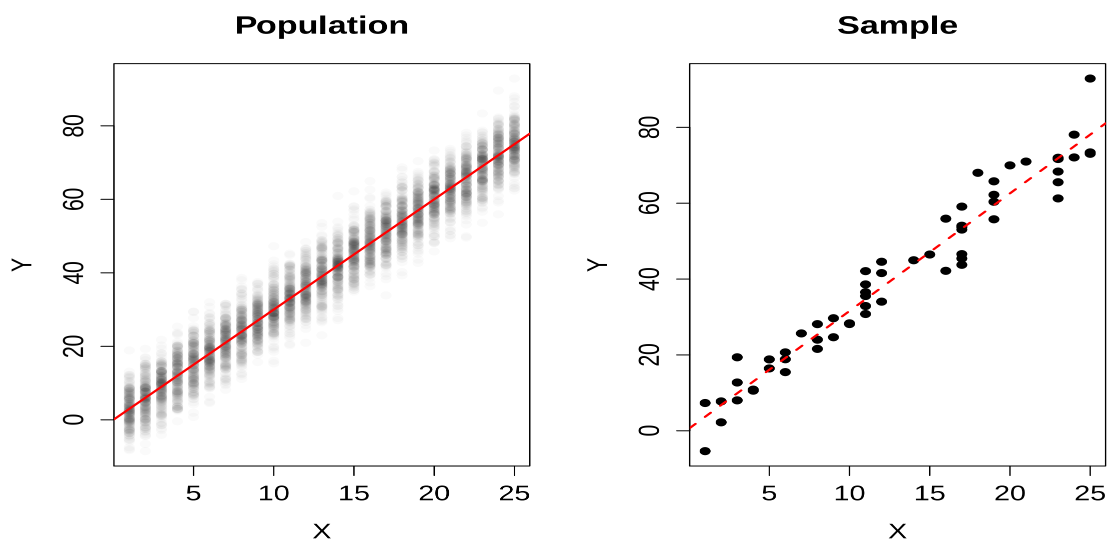
Review of Regression
Textbook Sections: NA

Simple Linear Regression

Recall the simple linear regression setting: $Y = \beta_0 + \beta_1 X + \epsilon$, where the errors ϵ are uncorrelated, and $E[\epsilon] = 0$ and $Var[\epsilon] = \sigma^2$. There is a distribution of Y values for each value of X, and the means of these distributions fall on a line. At any fixed value of X, Y has expected value $\beta_0 + \beta_1 X$, and variance σ^2.

Population and Sample

Population:	all possible (X, Y) pairs
True Regression Line:	$Y = \beta_0 + \beta_1 X$ contains the expected values of Y at each value of X
Sample:	n pairs: $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$
Estimated Regression Line:	$\hat{Y} = b_0 + b_1 X$



The goal is to come up with the best line based on the sample. First off, we should clarify what is meant by "best."

Fitted Values

The fitted values are the values of the estimated regression equation at the sample X values.
The fitted values are denoted by $\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_n$, where $\hat{Y}_i = b_0 + b_1 X_i$.

Residuals

The residuals are the deviations of the observed response values Y_i from the fitted values $\hat{Y}_i$.
The residuals are denoted by $e_1, e_2, \cdots, e_n$, where $e_i = Y_i - \hat{Y}_i$.

The residuals are very important in regression analysis. If the model is appropriate for the data, then we expect the residuals to exhibit certain properties. Many model diagnostic procedures are based on analysis of residuals.

Least Squares Estimation

The Least Squares Estimation (LSE) approach aims to minimize the sum of squared residuals of the regression line: $\sum_{i=1}^{n}(Y_i - [b_0 + b_1 X_i])^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

The line obtained with this approach is the "best" in the following sense. Of all possible lines, it has the lowest sum of squared residuals for this sample.

Summary Statistics

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

$S_{\text{xx}} = \sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad S_{\text{yy}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad S_{\text{xy}} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$

Least Squares Equations

The least squares regression line is the line $\hat{Y} = b_0 + b_1 X$, where b_0 and b_1 are calculated as follows.

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{\text{xy}}}{S_{\text{xx}}} \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

Correlation

A measure of linear relationship between X and Y (in the population) is called the correlation coefficient which is defined as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

In order to know ρ we will need to know $Cov(X,Y)$, $Var(X)$ and $Var(Y)$, and we typically do not know these quantities. However we can estimate each of these using our data set as follows:

$$\widehat{Cov(X,Y)} = S_{XY}/(n-1), \qquad \widehat{Var(X)} = S_{XX}/(n-1), \qquad \widehat{Var(Y)} = S_{YY}/(n-1).$$

Plugging in these estimates we get the sample correlation coefficient

$$\hat{\rho} = \frac{S_{XY}/(n-1)}{\sqrt{[S_{XX}/(n-1)][S_{YY}/(n-1)]}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}.$$

SSE and MSE

Another name for the sum of squared residuals is Sum of Squared Errors (SSE).

SSE $= \sum_{i=1}^{n}(Y_i - \hat{Y})^2$ SSE $= S_{\text{yy}} - b_1^2 S_{\text{xx}}$ SSE $= \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i Y_i$

The Mean Squared Error (MSE) is defined as $MSE = \frac{1}{n-p} SSE$, where p is the number of coefficients (βs) estimated. In simple regression we estimate β_0 and β_1, so $MSE = \frac{1}{n-2} SSE$.

The least squares estimate of the population variance σ^2 is MSE.

Multiple Regression

We will now consider multiple linear regression, or linear regression with multiple predictors.
In this scenario the model has the following form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i \text{ for } i = 1, \cdots, n.$$

There are $(p-1) \geq 1$ predictor variables, and p regression coefficients to be estimated.

The assumptions on the errors ϵ_i are the same (they are iid $N(0, \sigma^2)$ random variables).

The same model can be re-expressed in matrix form:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1(p-1)} \\ 1 & X_{21} & X_{22} & \cdots & X_{2(p-1)} \\ \vdots & & & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n(p-1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Even though the dimensions of $\mathbf{X}$ and β change, the model is still be described by the same matrix equation.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

In matrix form, the whole $p \times 1$ vector of estimated regression coefficients is computed as

$$\mathbf{b} = (\mathbf{X^T X})^{-\mathbf{1}} \mathbf{X^T Y}$$

Note that $\mathbf{b}$ is a random vector. The expectation and variance-covariance matrix of $\mathbf{b}$ are below.

$$\mathrm{E}[\mathbf{b}] = \beta \qquad \mathrm{Var}[\mathbf{b}] = \sigma^{\mathbf{2}} (\mathbf{X^T X})^{-\mathbf{1}}$$

The $n \times 1$ vector of fitted values is $\hat{\mathbf{Y}} = \mathbf{Xb}$.
The $n \times 1$ vector of residuals $\mathbf{e}$ is found by subtracting the fitted from the observed values of the response.

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb}$$

As before, the MSE is computed as $SSE/(n-p)$, where

$$SSE = \mathbf{e^T e} = (\mathbf{Y} - \mathbf{Xb})^{\mathbf{T}} (\mathbf{Y} - \mathbf{Xb})$$

Below are matrix expressions for the variances and estimated variances used for confidence intervals and hypothesis tests.

$$\boldsymbol{\sigma}^{\mathbf{2}}[\boldsymbol{b}] = \sigma^2 \cdot (\mathbf{X^T X})^{-\mathbf{1}} \qquad \mathbf{s^2}[\mathbf{b}] = MSE \cdot (\mathbf{X^T X})^{-\mathbf{1}}$$

Keep in mind that $\sigma^{\mathbf{2}}[\mathbf{b}]$ and $\mathbf{s^2}[\mathbf{b}]$ are both $p \times p$ matrices. The variances (or estimates of variances) of b_0 and b_1 are on the diagonal of the matrix. Check page 207 of the textbook for details.

Three Testing Problems

Suppose we have three predictors, and fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$. There are three types of questions that can be answered with hypothesis tests.

1. Can all predictors be dropped from the model?

   Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ $\quad H_1$: at least one of $\beta_1, \beta_2, \beta_3$ is nonzero

   Method: F test

2. Can one predictor be dropped from the model?

   Example: Can X_1 can be dropped from the model? This is equivalent to asking "does adding variable X_1 to the model $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ improve prediction of Y?"

   Hypotheses: $H_0 : \beta_1 = 0$ $\quad H_1 : \beta_1 \neq 0$

   Method: (partial) F test and t test are equivalent

3. Can several predictors be dropped from the model?

   Example: Can X_1 and X_3 be dropped from the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$? An equivalent question is: "does the addition of variables X_1 and X_3 to the model $Y = \beta_0 + \beta_2 X_2 + \varepsilon$ significantly improve the prediction of Y?"

   Hypotheses: $H_0 : \beta_1 = \beta_3 = 0$ $\quad H_1$: not both of β_1 and β_3 are zero

   Method: (partial) F test

In each case failing to reject H_0 is equivalent to concluding that knowing certain predictors most likely does not help in predicting Y.