STAT 425

Multiple Linear Regression. Part 1

Multiple Linear Regression Model

- In most applications we will want to use several predictors, instead of a single predictor as in simple linear regression (SLR).
- Now we have data of the form: $(y_i, \mathbf{x}_i)_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^t$ with $x_{i1} = 1$
- Assume the model:

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

  - $(\beta_1, \ldots, \beta_p, \sigma^2)$: the unknown but true parameters.
  - e_i's: random errors

Main model assumptions:

1. The mean function $E[y_i] = x_{i1}\beta_1 + x_{i2}\beta_2 \ldots + x_{ip}\beta_p$ is linear in the p predictors.
2. The errors e_i's are uncorrelated with mean 0 and constant variance σ^2. This is equivalent to: $E[e_i] = 0$ and $Cov(e_i, e_j) = \sigma^2\delta_{ij}$, with $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$.
3. For hypothesis testing we further assume that e_i are i.i.d and $e_i \sim N(0, \sigma^2)$

Matrix representation of the MLR

MLR is valid for all observations for $i = 1, \ldots, n$. We can write:

$$
\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1p}\beta_p + e_1 \\ x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2p}\beta_p + e_2 \\ \ldots \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \ldots + x_{np}\beta_p + e_n \end{pmatrix}
$$

$$
= \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}
$$

$$
\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}
$$

Matrix $\mathbf{X}$ is normally called the design matrix

Least Square Estimation

- Using matrix representation, we can express the MLR model as[1]

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- The Least-Squares estimate of $\boldsymbol{\beta}$ is the vector that minimizes the Residual Sum of Squares (RSS):

$$RSS = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

[1] By default the intercept is included in the model, then the 1st column of the design matrix $\mathbf{X}$ is a vector of 1's. We further assume that the rank of $\mathbf{X}$ is p, i.e., no columns of $\mathbf{X}$ is a linear combination of the other columns and $\mathbf{X}$ is a *tall and skinny matrix* $(n > p)$.

Differentiating RSS with respect to $\boldsymbol{\beta}$ and setting to zero, we have

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}^t_{p\times n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})_{n\times 1} = \mathbf{0}_{p\times 1} \\
&\Longrightarrow \mathbf{X}^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \quad \text{normal equation} \\
&\Longrightarrow (\mathbf{X}^t\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y} \\
&\Longrightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \quad (*)
\end{aligned}$$

Note that the inverse of the $p \times p$ matrix $(\mathbf{X}^t\mathbf{X})$ exists since we assume the rank of $\mathbf{X}$ is p.

Next let's check the equation (*) for SLR.

$$
\begin{aligned}
\mathbf{X}^t\mathbf{X} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix} \\
\left(\mathbf{X}^t\mathbf{X}\right)^{-1} &= \frac{1}{n\sum x_i^2 - (n\bar{x})^2} \begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \\
\mathbf{X}^t\mathbf{y} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}
\end{aligned}
$$

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{y} \\ &= \frac{1}{n\sum x_i^2 - (n\bar{x})^2}\begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}\begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}\end{aligned}$$

So $\hat{\beta}_1$ is given by [a]

$$\hat{\beta}_1 = \frac{-n^2\bar{x}\bar{y} + n\sum x_i y_i}{n\sum x_i^2 - (n\bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\mathsf{Sxy}}{\mathsf{Sxx}}$$

Similarly we can check the calculation for $\hat{\beta}_0$.

[a] $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$ and $\sum(x_i - \bar{x})(x_i - \bar{x}) = \sum x_i^2 - n\bar{x}^2$.

Fitted values and Residuals

- Fitted values

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{t}}\mathbf{y} = \mathbf{H}_{n\times n}\mathbf{y}_{n\times 1}$$

  $\mathbf{H}_{n\times n}$: is called the hat matrix, since it returns *y-hat*.

- Residuals
  The estimated residuals are given by:

$$\mathbf{r}_{n\times 1} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

- The residuals $\mathbf{r}$ are used to estimate the error variance:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 = \frac{RSS}{n-p}$$

Note that the LS estimate $\hat{\boldsymbol{\beta}}$ satisfies the normal equations:

$$\mathbf{X}^t(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

From this equation we can say that the residuals $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ satisfy:

- $\mathbf{X^t r} = \mathbf{0}$. This implies that when we calculate the inner product of each column of matrix $\mathbf{X}$ with the residual vector $\mathbf{r}$, this product is zero.
- In particular, when we include the intercept, the first equation implies that $\mathbf{1}^t\mathbf{r} = \sum_{i=1}^n r_i = 0$.
- The inner product $\hat{\mathbf{y}}^{\mathbf{t}}\mathbf{r} = \hat{\boldsymbol{\beta}}^t\mathbf{X^t r} = 0$. This means that the residual vector is orthogonal to each column of $\mathbf{X}$ and $\hat{\mathbf{y}}^{\mathbf{t}}$

The Hat Matrix

The hat matrix is defined as:

$$\mathbf{H}_{n\times n} = \mathbf{X}(\mathbf{X^tX})^{-\mathbf{1}}\mathbf{X^t}$$

- Consider a linear combination of the columns of $\mathbf{X}$ of the form $\mathbf{v} = \mathbf{X}\mathbf{a}_{p\times 1}$. The $\mathbf{Hv} = \mathbf{v}$, since:

$$\mathbf{HX} = \mathbf{X}(\mathbf{X^tX})^{-\mathbf{1}}\mathbf{X^t}\mathbf{X} = \mathbf{X}$$

Properties of matrix $\mathbf{H}$

- Symmetric: $\mathbf{H^t} = [\mathbf{X}(\mathbf{X^tX})^{-1}\mathbf{X^t}]^{\mathbf{t}} = \mathbf{X}(\mathbf{X^tX})^{-1}\mathbf{X^t} = \mathbf{H}$
- Idempotent[2]: $\mathbf{HH} = \mathbf{HH^t} = \mathbf{H}$
  $\mathbf{HH} = \mathbf{X}(\mathbf{X^tX})^{-1}\mathbf{X^t}\mathbf{X}(\mathbf{X^tX})^{-1}\mathbf{X^t} = \mathbf{X}(\mathbf{X^tX})^{-1}\mathbf{X^t} = \mathbf{H}$
- $trace(\mathbf{H}) = p$, the number of LS coefficients to be estimated.

[2] This property also implies that $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$

Goodness of Fit: R-square

We use the R^2 to measure how well the model fits the data. R^2 is the fraction of the total variance explained by the model:

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

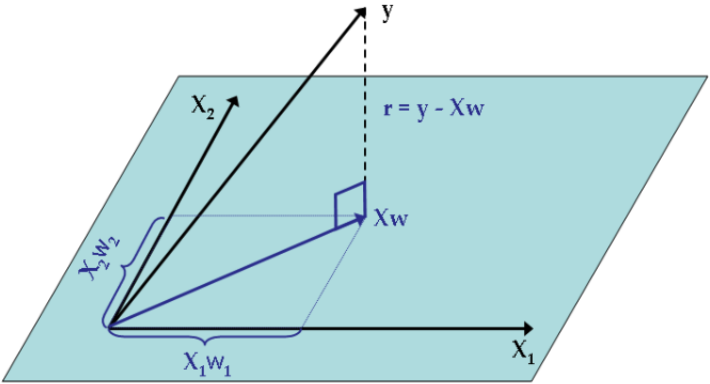
$$0 \leq R^2 \leq 1$$

This can also be written as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

Geometrical interpretation of the LS estimation

In $\Re^3$:



- All linear combinations $\mathbf{Xw}$ ($\mathbf{w} \in \Re^p$) of the columns of matrix $\mathbf{X}$ form a sub-space of dimension p in $\Re^n$ (denoted by $C(\mathbf{X})$). In the previous figure think about all the linear combinations of X_1 and X_2.
- Finding $\hat{\boldsymbol{\beta}}$ that minimizes $||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$ is equivalent to finding a vector $\hat{\mathbf{y}}$ from the estimation space that minimizes $||\mathbf{y} - \hat{\mathbf{y}}||^2$. From the figure it is intuitive that the fitted value is the projection of $\mathbf{y}$ onto the estimation space.
- Matrix $\mathbf{H}_{n\times n}$ is the projection matrix:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X^tX})^{-\mathbf{1}}\mathbf{X^ty} = \mathbf{H_{n\times n}y}$$

  $\mathbf{H}$ is symmetric, unique and idempotent, and the $trace(\mathbf{H}) = p$, which is the dimension of vector space $C(\mathbf{X})$.

- Error space: this sub-space of dimension $(n-p)$ is denoted by $C(\mathbf{X})^T$, and it is orthogonal to the estimation space. The matrix $(\mathbf{I_n} - \mathbf{H})$ is the projection matrix of the error space.
- Residuals: The estimated residuals can be calculated as:

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

  If the intercept is included in the model $\sum_i r_i = 0$. Due to the normal equation $\mathbf{X^t}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$:

$$\sum_{i=1}^{n} r_i X_{ij} = 0 \;\; for \quad j = 1, \ldots, p$$

Geometric Interpretation: $\mathbf{r}$ is the projection of $\mathbf{y}$ onto the error space orthogonal to $C(\mathbf{X})$. So $\mathbf{r}$ is orthogonal to any vector in $C(\mathbf{X})$. Especially, $\mathbf{r}$ is orthogonal to each column of $\mathbf{X}$.

An example

Savings Data set

##Savings rates in 50 countries The savings data frame has 50 rows and 5 columns. The data is averaged over the period 1960-1970. This data frame contains the following columns:

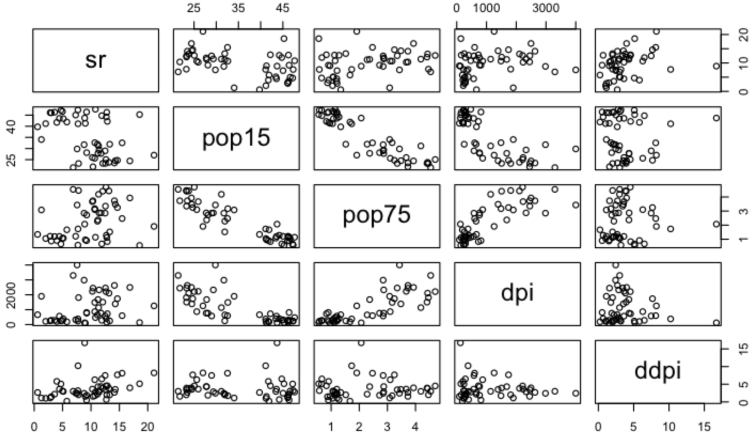
- sr – personal saving divided by disposable income
- pop15 – percent population under age of 15
- pop75 – percent population over age of 75
- dpi – per-capita disposable income in dollars
- ddpi – percent growth rate of dpi

```r
library(faraway)
?savings
head(savings)
```

```
##              sr pop15 pop75     dpi ddpi
## Australia 11.43 29.35  2.87 2329.68 2.87
## Austria   12.07 23.32  4.41 1507.99 3.93
## Belgium   13.17 23.80  4.43 2108.47 3.82
## Bolivia    5.75 41.89  1.67  189.13 0.22
## Brazil    12.88 42.19  0.83  728.47 4.56
## Canada     8.79 31.72  2.85 2982.88 2.43
```

Plotting the data using the function **pairs(.)**

```
>pairs(saving)
```



Simple Linear Regression using function **lm**: $sr \sim pop75$

```
summary(lm(sr ~ pop75,data=savings))
```

```
##
## Call:
## lm(formula = sr ~ pop75, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2657 -3.2295  0.0543  2.3336 11.8498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.1517     1.2475   5.733  6.4e-07 ***
## pop75         1.0987     0.4753   2.312   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.294 on 48 degrees of freedom
## Multiple R-squared:  0.1002, Adjusted R-squared:  0.08144
## F-statistic: 5.344 on 1 and 48 DF,  p-value: 0.02513
```

Multiple Linear Regression: $sr \sim pop15 + pop75 + dpi + ddpi$

```
fullmodel=lm(sr~pop15+pop75+dpi+ddpi, data=savings)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Contradictory results for the estimated $\hat{\beta}_{pop75}$? Some predictors might be highly correlated:

```
# Lets look at the correlation matrix
cor(savings[,-1])
```

```
##            pop15       pop75        dpi        ddpi
## pop15  1.00000000 -0.90847871 -0.7561881 -0.04782569
## pop75 -0.90847871  1.00000000  0.7869995  0.02532138
## dpi   -0.75618810  0.78699951  1.0000000 -0.12948552
## ddpi  -0.04782569  0.02532138 -0.1294855  1.00000000
```

This correlation might cause contradictory results, with some regression coefficients having an unexpected sign.

Rank deficiency

- The design matrix $\mathbf{X}$ is an $n \times p$ matrix[3]. If this matrix is not of full rank (i.e., its columns are not linearly independent), the matrix $\mathbf{X^tX}$ can not be inverted (singular matrix).
- If the matrix $\mathbf{X^tX}$ is singular the LS solutions is not unique (identifiability problem)
- **R** can cope well with this problem. To solve the LS equations **R** uses the QR decomposition. You can read more on this in the supplemental material.

[3]You can use function *model.matrix(.)* in **R** to extract the model matrix of a fitted model