Multiple Linear Regression: Notation and Estimation

Author: Nicholas G Reich, Jeff Goldsmith

This material is part of the ***statsTeachR*** *project*

Multiple linear regression model

- Observe data $(y_i, x_{i1}, \ldots, x_{ip})$ for subjects $1, \ldots, n$. Want to estimate $\beta_0, \beta_1, \ldots, \beta_p$ in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_1 x_{ip} + \epsilon_i; \; \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

- Assumptions (residuals have mean zero, constant variance, are independent) are as in SLR
- Impose linearity which (as in the SLR) is a big assumption
- Our primary interest will be $E(y|\mathbf{x})$
- Eventually estimate model parameters using least squares

Omitted variable bias

What happens if the true regression model is

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

but we ignore x_2 and fit the simple linear regression

$$y_i = \beta_0^* + \beta_1^* x_{i,1} + \epsilon_i^*$$

Does $\beta_1^* = \beta_1$?

Omitted variable bias

When should you be concerned?

If both of the following conditions are met, then $\beta_1^* = \beta_1$:

- The omitted variable is unrelated to the outcome
- The omitted variable is uncorrelated with the retained variable

Extra credit for problem set 1: create a simulation where you show an example of omitted variable bias.

Matrix notation

- Observe data $(y_i, x_{i1}, \ldots, x_{ip})$ for subjects $1, \ldots, n$. Want to estimate $\beta_0, \beta_1, \ldots, \beta_p$ in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_1 x_{ip} + \epsilon_i; \;\; \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

- Notation is cumbersome. To fix this, let
  - $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{ip}]$
  - $\boldsymbol{\beta}^T = [\beta_0, \beta_1, \ldots, \boldsymbol{\beta}_p]$
  - Then $y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$

Multiple linear regressoion

- Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & x_{ij} & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Then we can write the model in a more compact form:

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

- $\mathbf{X}$ is called the *design matrix*

Matrix notation

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

- ϵ is a random vector rather than a random variable
- $E(\boldsymbol{\epsilon}) = 0$ and $Cov(\boldsymbol{\epsilon}) = \sigma^2 I$
- Note that *Cov* means the "variance-covariance matrix"

Mean, variance and covariance of a random vector

- Let $\mathbf{y}^T = [y_1, \ldots, y_n]$ be an n-component random vector. Then its mean and variance are defined as

$$\begin{aligned} E(\mathbf{y})^T &= [E(y_1), \ldots, E(y_n)] \\ Var(\mathbf{y}) &= E\left[(\mathbf{y} - E\mathbf{y})(\mathbf{y} - E\mathbf{y})^T\right] = E(\mathbf{y}\mathbf{y}^T) - (E\mathbf{y})(E\mathbf{y})^T \end{aligned}$$

- Let $\mathbf{y}$ and $\mathbf{z}$ be an n-component and an m-component random vector respectively. Then their covariance is an $n \times m$ matrix defined by

$$Cov(\mathbf{y}, \mathbf{z}) = E\left[(\mathbf{y} - E\mathbf{y})(\mathbf{z} - \mathbf{z})^T\right]$$

Least squares

As in simple linear regression, we want to find the $\boldsymbol{\beta}$ that minimizes the residual sum of squares.

$$RSS(\boldsymbol{\beta}) = \sum_i \epsilon_i^2 = \epsilon^T \epsilon$$

After taking the derivative, setting equal to zero, we obtain:

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

Sampling distribution of $\hat{\boldsymbol{\beta}}$

If our usual assumptions are satisfied and $\epsilon \overset{iid}{\sim} \mathsf{N}\left[0, \sigma^2\right]$ then

$$\hat{\boldsymbol{\beta}} \sim \mathsf{N}\left[\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right].$$

- This will be used later for inference.
- Even without Normal errors, asymptotic Normality of LSEs is possible under reasonable assumptions.

Definitions

- *Fitted values*: $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$
- *Residuals / estimated errors*: $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$
- *Residual sum of squares*: $\sum_{i=1}^{n} \hat{\epsilon_i}^2 = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}$
- *Residual variance*: $\hat{\sigma^2} = \frac{RSS}{n-p-1}$
- *Degrees of freedom*: $n - p - 1$

R^2 and sums of squares

- Regression sum of squares $SS_{reg} = \sum(\hat{y}_i - \bar{y})^2$
- Residual sum of squares $SS_{res} = \sum(y_i - \hat{y}_i)^2$
- Total sum of squares $SS_{tot} = \sum(y_i - \bar{y})^2$
- Coefficient of determination

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

Hat matrix

Some properties of the hat matrix:

- It is a projection matrix: $\mathbf{HH} = \mathbf{H}$
- It is symmetric: $\mathbf{H}^T = \mathbf{H}$
- The residuals are $\hat{\epsilon} = (\mathbf{I} - \mathbf{H})\mathbf{y}$
- The inner product of $(\mathbf{I} - \mathbf{H})\mathbf{y}$ and $\mathbf{Hy}$ is zero (predicted values and residuals are uncorrelated).

Projection space interpretation

The hat matrix projects $\mathbf{y}$ onto the column space of $\mathbf{X}$.
Alternatively, minimizing the $RSS(\beta)$ is equivalent to minimizing the Euclidean distance between $\mathbf{y}$ and the column space of $\mathbf{X}$.

Lung Data Example (con't from last clas)

```
mlr2 <- lm(disease ~ crowding + education + airqual,
           data=dat, x=TRUE, y=TRUE)
X = mlr2$x
y = mlr2$y
(betaHat = solve( t(X) %*% X) %*% t(X) %*% y )

##                [,1]
## (Intercept) -7.7505
## crowding     1.3128
## education    1.4377
## airqual      0.2881

coef(mlr2)

## (Intercept)    crowding   education     airqual
##     -7.7505      1.3128      1.4377      0.2881
```

Today's big ideas

- Multiple linear regression models, interpretation, notation, biases