Multiple Regression by Matrix Algebra.

STA 211: The Mathematics of Regression

Miheer Dewaskar

Slides adapted from lectures by Prof. Jerry Reiter and Prof. Yue Jiang.

Multiple Regression by Matrix Algebra

- For simple linear regression, we showed
  - how to compute MLE $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$.
  - how to prove unbiasedness, $E(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta}$, and to derive $Var(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$.
  - expressions for predicted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and residuals $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{y}$.
- But most regressions use more than one explanatory variable. How can matrix algebra help for multiple regression?

Multiple Linear Regression Model

- Model for multiple linear regression with independent observations:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

- We can write this as

$$y_i = \begin{pmatrix} 1 & x_{i1} & x_{i2} & . & . & . & x_{ip} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

Multiple Linear Regression in Matrix Form

- Let $\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{ip} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}.$
- Then, $y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$.

A More Compact Version of the Model

- We can write this even more compactly. Let

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2p} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{np} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

- Then, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where each $\epsilon_i \sim N(0, \sigma^2)$ independently.

Writing Model as a Probability Density Function

- We can write the linear model as the p.d.f.,

$$f(y \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) = N(\mathbf{x}^t \boldsymbol{\beta}, \sigma^2)$$

.

- Writing out the pdf, we have

$$f(y \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y - \mathbf{x}^t \boldsymbol{\beta})^2 / 2\sigma^2).$$

- This matches the expression that we used for simple linear regression!

Maximum Likelihood Estimates, Predicted Values, Residuals

- MLE: could take $p + 2$ partial derivatives of likelihood (with respect to $\beta_0, \ldots, \beta_p, \sigma^2$). But that would be awful!
- We can use the same logic and derivations as done for simple linear regression, but with $(p + 1) > 2$ explanatory variables!
- Thus, the matrix expressions for the MLE, the predicted values, and the residuals are EXACTLY the same as what we derived previously and summarized on the first slide.

Variances of Maximum Likelihood Estimates, Predicted Values, Residuals

- So are the matrix expressions for theoretical variances!
  - $Var(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$.
  - $Var(\hat{y}_{new,avg} \mid \mathbf{X}, \mathbf{x}_{new}) = \sigma^2\mathbf{x}_{new}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_{new}$.
  - $Var(\hat{y}_{new,ind} \mid \mathbf{X}, \mathbf{x}_{new}) = \sigma^2(1 + \mathbf{x}_{new}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_{new})$.
- Estimate σ^2 using an unbiased estimator, $RSE^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})^t(\mathbf{y} - \mathbf{X}\hat{\beta})/(n - (p + 1))$.
- We will show that we match the answers from the *lm* function in *R*.

Properties of MLE, Predicted Values, Residuals

- Expected values using matrix expressions are identical to what we did previously!
  - $E(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta}$, so MLE for $\boldsymbol{\beta}$ is unbiased.
  - $E(\hat{\mathbf{y}} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$.
  - $E(\mathbf{r} \mid \mathbf{X}) = \mathbf{0}$.
- As are variances!
  - $Var(\mathbf{r} \mid \mathbf{X}) = \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)$.

What actually *is* a "linear model" anyway?

Which of the following (if any) depict a relationship that can be considered a "linear regression model"?

1. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \epsilon_i$
2. $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$
3. $y_i = \beta_0 + \frac{\beta_1 x_{i1}}{\beta_2 x_{i2} + \beta_3 x_{i3}} + \epsilon_i$
4. $y_i = \beta_0 + \beta_1 x_{i1}^{(x_{i2} + x_{i3})} + \epsilon_i$
5. $y_i = \beta_0 + \beta_1 \sin(x_{i1} + \beta_2 x_{i2}) + \beta_3 x_{i3} + \epsilon_i$
6. $y_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$
7. $y_i = \beta_1 x_{i1}^{(x_{i2} + \beta_2 x_{i3})} + \epsilon_i$
8. $y_i = \beta_0 + \beta_1 \cos(x_{i1}) + \beta_2 \sin(x_{i2}) + \beta_3 x_{i3}^{1/2} + \epsilon_i$
9. $y_i = \beta_1 e^{x_{i1}} + \beta_2 e^{x_{i2}} + \epsilon_i$
10. $y_i = \beta_0 + \beta_1 e^{\beta_2 x_{i1}} + \beta_3 x_{i2} + \epsilon_i$

Linear regression models

Linear regression models are linear *in the parameters*. That is, for a given observation Y_i:

$$Y_i = \beta_0 + \beta_1 f_1(X_{i1}) + \beta_2 f_2(X_{i2}) + \cdots + \beta_p f_p(X_{ip}) + \epsilon_i$$

The functions $f_1, \cdots f_p$ may themselves be non-linear, but as long as the $\boldsymbol{\beta}$ are linear in $\mathbf{y}$, we have a linear regression model.

- Why would we want to use any function such that $f_k(u) \neq u$?
- What about $y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$?

Transforming predictors

Still technically a "linear regression model":

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \epsilon_i$$

Example: linear model interpretations

Let's consider some various regression functions (most of them linear). What happens when x_1 changes in some various models?

$$\frac{\partial}{\partial x_1}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\right) = \beta_1$$

$$\frac{\partial}{\partial x_1}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2\right) = \beta_1 + 2\beta_2 x_1$$

$$\frac{\partial}{\partial x_1}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2\right) = \beta_1 + \beta_3 x_2$$

$$\frac{\partial}{\partial x_2}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2\right) = \beta_2 + \beta_3 x_1$$

That other model

$$y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$
$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Although the RHS is a linear function of β, we do not have a linear model *for* y; it is linear in $\log(y)$.

This type of model (i.e., a linear relationship between $\boldsymbol{\beta}$ and an invertible function of y) is known as a *generalized* linear model (well, technically with the conditional expectation of Y, but more on that later), and we will study this class of model in a few lectures.

Example: linear model interpretations

Let's focus on the "interaction model"

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i3} + \epsilon_i$$

- What might we expect from a change in either of the predictor variables?
- How might we use this intuition to interpret a model with these so-called "interaction terms"?

Unpacking the design matrix

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\boldsymbol{\epsilon}}$$

- $\mathbf{X}$ is sometimes called the design matrix.
- How might you include a *categorical* predictor with k levels in a design matrix?

Dummy coding of categorical variables

Suppose we are trying to predict the amount of sleep a Duke student gets based on whether they are in Pratt (vs. non-Pratt; these are the only two options). Consider the following model:

$$Sleep_i = \beta_0 + \beta_1 \mathbf{1}(Pratt_i = \text{“}Yes\text{”}) + \beta_2 \mathbf{1}(Pratt_i = \text{“}No\text{”})$$

In-class exercise:

- Write out the design matrix for this hypothesized linear model.
- Demonstrate that the design matrix is not of full column rank (that is, affirmatively provide one of the columns in terms of the others).
- Use this intuition to explain why when we include categorical predictors, we cannot include both indicators for every level of the variable *and* an intercept.

R, RStudio, etc.

Matrix operations in R

- `as.matrix()` function sets an object as a matrix object in R
- `%*%` is the matrix multiplication operation (e.g., `A %*% B` for two matrices A and B)
- `t()` function takes the transpose of a matrix
- `solve()` function inverts a matrix

Matrix operations in R

- `install.packages("palmerpenguins")`
- `library(palmerpenguins)`
- `head(penguins)`

A basic regression model

$$(\text{Body mass})_i = \beta_0 + \beta_1(\text{Flipper length})_i + \beta_2(\text{Bill length})_i + \epsilon_i$$

```
> summary(lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins))

Call:
lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
    data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-1090.5  -285.7   -32.1   244.2  1287.5

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
bill_length_mm        6.047      5.180   1.168    0.244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 394.1 on 339 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:   0.76,     Adjusted R-squared:  0.7585
F-statistic: 536.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_0 =$ -5736.897; $\hat{\beta}_1 =$ 48.145; $\hat{\beta}_2 =$ 6.047
- Recover these estimates *from the dataset directly* using matrix operations.