Review of Multivariate Linear Regressions

Teaching team of STA314

Department of Statistical Sciences
University of Toronto

Monday September 9th, 2024

Multivariate Linear Regression

Assume $(y_i, \mathbf{x}_i)$, for $1 \leq i \leq n$, are independent and satisfy

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where

(a) <u>uncorrelatedness</u>: $\mathbf{x}_i$ is uncorrelated with ϵ_i

(b) <u>linearity</u>: $\mathbb{E}[\epsilon_i] = 0$

(c) <u>homoscedasticity</u>: $\text{Var}(\epsilon_i) = \sigma^2$

(d) <u>normality</u>: $\epsilon_i \sim N(0, \sigma^2)$.

Notation

Using the matrix notation,

- $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$,
- $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times(p+1)}$ with $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{ip}]^\top \in \mathbb{R}^{p+1}$ for $1 \leq i \leq n$,
- $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \cdots, \hat{\beta}_p]^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]^\top \in \mathbb{R}^{p+1}$.

we know

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2.$$

Inference on β

- The *unknown* variance σ^2 may be estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\right)^2.$$

- A 95% asymptotic[1] confidence interval of β_j, for $1 \le j \le p$, has the form of

$$\left[\hat{\beta}_j - 1.96 \cdot SE(\hat{\beta}_j),\ \ \hat{\beta}_j + 1.96 \cdot SE(\hat{\beta}_j)\right],$$

  where

$$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}$$

- Hypothesis testing

$$H_0 : \beta_j = 0 \quad (\text{There is no linear relationship between } Y \text{ and } X_j)$$

  vs

$$H_1 : \beta_j \neq 0 \quad (\text{There is linear relationship between } Y \text{ and } X_j)$$

[1] This means the limit where the sample size tends to infinity.

Inference on β

The test-statistic under the null hypothesis:

$$t = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

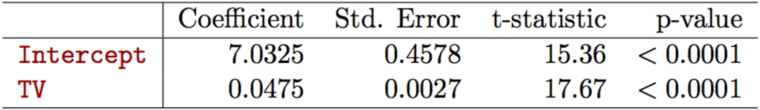
- which has a t-distribution with $n - p - 1$ degrees of freedom, when $\beta_j = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the ***p*-value**.
- In most applications, we reject the null hypothesis if the p-value ≤ 0.05.
- Can be generalized to

  $$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

  via F-statistics. (c.f. pp 75-78 of the textbook.)

Results for Advertising Data

Y : **Sales**, X : **TV** budget, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 7.0325 + 0.0475 x_i$.



	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

- The *p*-value for **TV budget** is smaller than 0.05, so that we reject the null hypothesis $\beta_1 = 0$.
- This indicates that **TV budget** is significant for predicting **Sales**.

Other considerations in linear regression models

- The coefficient of determination: R^2
- Qualitative Predictors
- Extend to non-linearity
  - Adding interaction terms
  - Adding transformed predictors
- Model diagnosis for (a) – (d)

The coefficient of determination: R-squared (R^2)

Meaning of R-squared

R^2 is the proportion of the variation in the outcome (Y) that can be explained from the predictors (X).

Recall that for each training point $(\mathbf{x}_i, y_i)$, its fitted value is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}.$$

with its residual defined as $y_i - \hat{y}_i$. The residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

is nothing but the training MSE of the model given $(\hat{\beta}_0, \ldots, \hat{\beta}_p)$.

The coefficient of determination: R^2

- The total sum of squares is

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \text{with} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

  quantifying the total variance of Y in the sample $(y_1, \ldots, y_n)$.

- R^2 measures the proportion of variability in Y that can be explained by regressing Y onto X.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

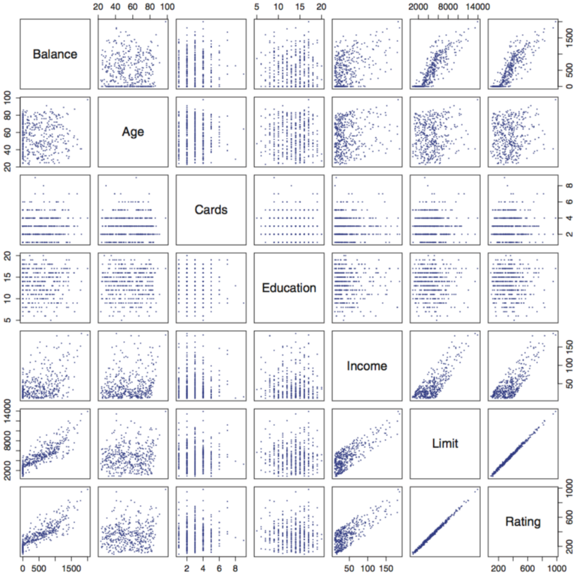
- $0 \leq R^2 \leq 1$. R^2 *close to* 1 indicates a large proportion of the variability in the response that is explained by the predictors.

- However, a large value of R^2 does **NOT** imply that the model fits the data well. It always favors more flexible models, which may overfit the data! (Adjusted R^2 later.)

Qualitative Predictors

- Some predictors are not quantitative but qualitative, taking a discrete set of values.
- These are also called **categorical predictors** or **factor variables**.
- See for example the scatterplot matrix of the credit card data.

Credit Card Data — Quantitative variable



Credit Card Data — Qualitative variable

In addition to the 7 quantitative variables, there are four qualitative variables:

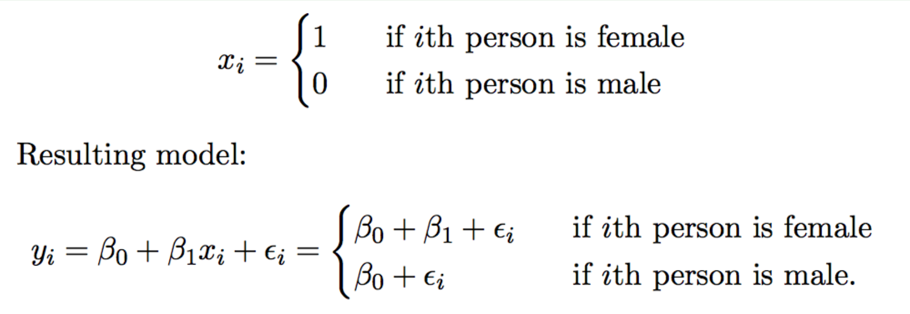
- gender (Male/Female)
- student (Student /Not student)
- status (Married/Not Married)
- ethnicity (Caucasian/ African American (AA)/Asian).

For more information on the data set, feel free to check https://rdrr.io/cran/ISLR/man/Credit.html.

Qualitative predictors with two levels

Example (study the difference in credit card balance between males and females, ignoring the other variables)

We create a new **dummy variable** of the predictor (gender):



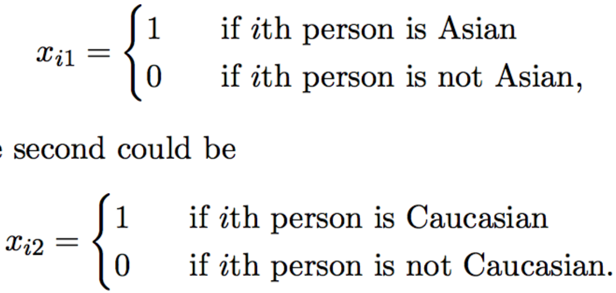
$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Qualitative predictors with more than two levels

With more than two levels, we create additional dummy variables. For example, for the ethnicity variable we create two dummy variables. The first could be



$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

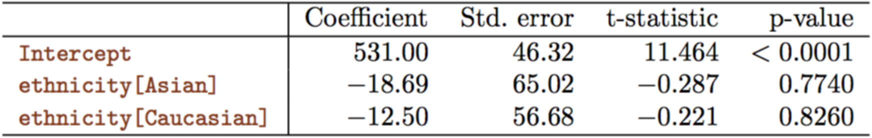
Qualitative Predictors with More Than Two Levels

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

- There are always one fewer dummy variables than the number of levels.
- The level when all dummy variables are 0 – African American in this example – is known as the baseline.

Credit Card Data



	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

Interpretation: The Asian category tends to have 18.69 less debt than the AA category, and that the Caucasian category tends to have 12.50 less debt than the AA category.

Exercise: What does the p-value tell us about the coefficient?

Extension to non-linearity: adding **interaction** terms

- Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

  Regardless of the value of X_2, one-unit increase in X_1 will lead to β_1-unit increase in Y.

- Consider the model with **interaction** terms

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \\ &= \beta_0 + (\underbrace{\beta_1 + \beta_3 X_2}_{\tilde{\beta}_1}) X_1 + \beta_2 X_2 + \epsilon. \end{aligned}$$

  Since $\tilde{\beta}_1$ changes with X_2, the effect of X_1 on Y is no longer constant: adjusting X_2 will change the impact of X_1 on Y.

- β_1 and β_2 are the coefficients of **main effects** while β_3 is that of the **interaction**.

Interaction

Example (Gender + Education)

Consider

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

where

$$x_{i1} = \begin{cases} 1 & \text{if the } i\text{the person is female} \\ 0 & \text{if the } i\text{the person is male} \end{cases}$$

$$x_{i2} = \text{the number of years of education.}$$

Interpretation of β_2: one more year of education leads to a β_2 unit change in the credit card balance with gender held fixed.

Example (Gender + Education)

Now consider

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i \\
&= \begin{cases} \beta_0 + \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i2} + \epsilon_i, & \text{if the } i\text{the person is female} \\ \beta_0 + \beta_2 x_{i2} + \epsilon_i, & \text{if the } i\text{the person is male} \end{cases}
\end{aligned}
$$

where

$$
x_{i1} = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}, \quad x_{i2} = \text{the number of years of education.}
$$

- **Interpretation of β_2**: one more year education leads to β_2-unit change in credit card balance with for male.
- **Interpretation of β_3**: for one more year education, the difference in credit card balance between female and male is β_3.
- **How about $\beta_3 + \beta_2$?**

Interactions

Read pages 89-90 of the textbook for more examples.

Extension to non-linearity: adding transformed predictors

This is another way to increase model complexity by introducing more predictors / parameters. For instance,

$$f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^3$$

We will come back to this later in Week 4.

Diagnosis of Linear Models

- Non-linearity of the response-predictor relationships.
- Correlation of the error terms among training samples.
- Non-constant variance of error terms.
- Outliers.
- Collinearity.

Quick Memory Test

Are the following statement true?

- The smaller the p-value for a coefficient in the model, the larger (in absolute value) the coefficient.
- The higher the coefficient of determination, R^2, the better the model.
- In a group of people, the average income is a qualitative variable.
- For a qualitative variable with more than two levels, we should use one dummy variable with multiple levels to represent the variable accurately.