Multivariable Analysis with R

Humayun Kabir, BScN, MPH, MSc (Student)
MSc in Health Research Methodology

Department of Health Research Methods, Evidence, and Impact, McMaster University, Canada

DASH: Data Analysis Support Hub Workshop Series

Date: February 20, 2024

Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

McMaster University Library

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Laslovarga, “Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area,” 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.jpg

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information: scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Attendance that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

Verify your participation at a session: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

DASH: Data Analysis Support Hub Workshops

Register for upcoming DASH events:

March 22: Machine Learning with R: Logistic Regression – Humayun Kabir

March 28: Intermediate Python Programming – Seyed Amirreza Mousavi

April 30: Survival Analysis with R – Humayun Kabir

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- Creating data visualizations, including charts, graphs, and scatter plots
- Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- Choosing which software package to use, including free and open-source software
- Troubleshooting problems related to file formats, data retrieval, and download
- Selecting methodology and type of data analysis to use in a thesis project

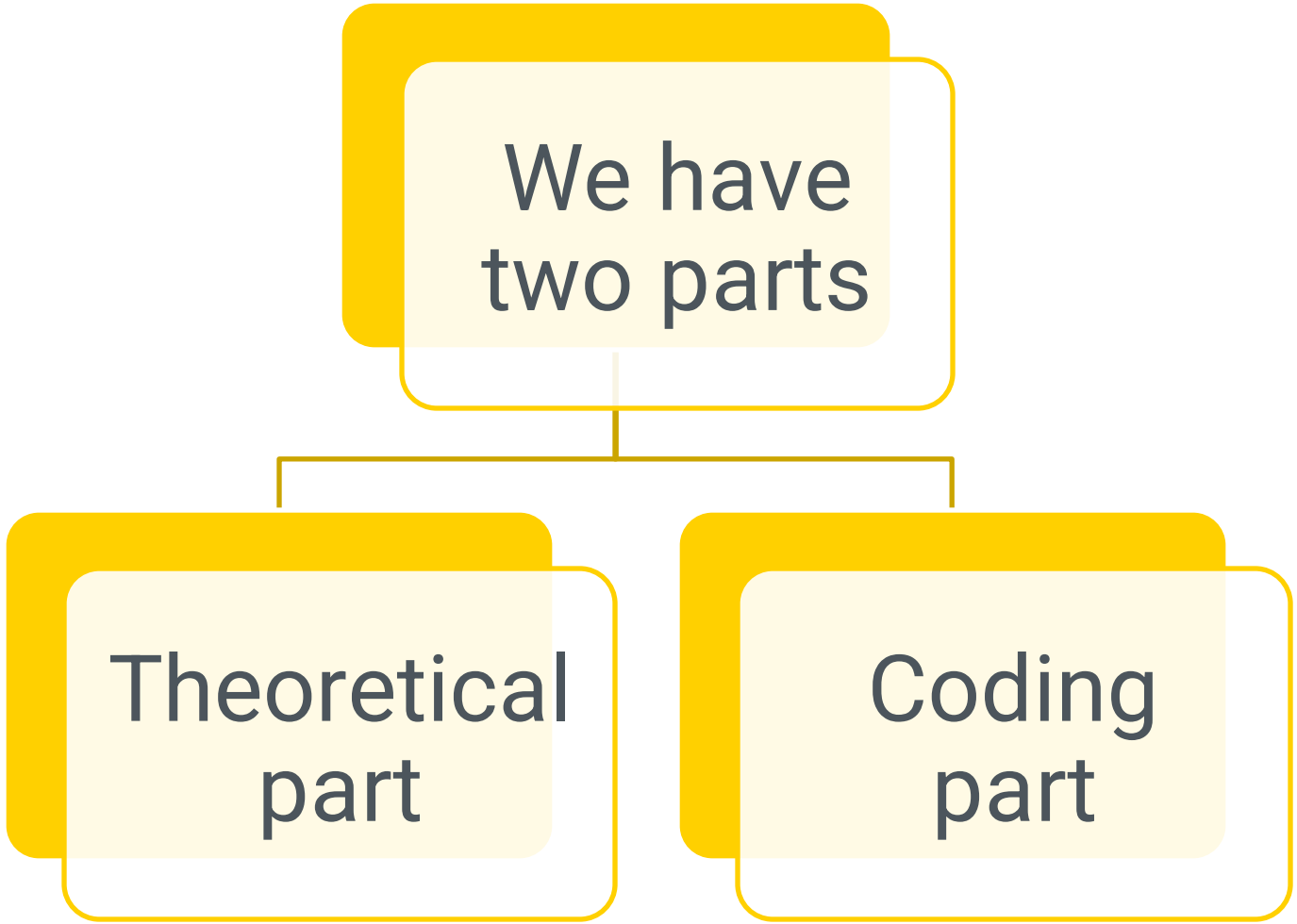
Book an appointment: https://library.mcmaster.ca/services/dash

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences. In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

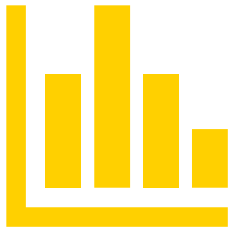
Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Multivariable Analysis with R



Multivariable Analysis with R

Objective



1. Learn what is multivariable analysis

2. Learn common multivariable analysis

3. Fit common multivariable model in R

Multivariate or Multivariable Regression?

Multivariate analysis refers to statistical models that have 2 or more dependent or outcome variables

Multivariable analysis refers to statistical models in which there are multiple independent or response variables.

Bertha Hidalgo, Melody Goodman, “Multivariate or Multivariable Regression?”, American Journal of Public Health 103, no. 1 (January 1, 2013): pp. 39-40.

Reasons for multivariable analysis

- **Exploring associations, hypothesis generation –** no *a priori* hypotheses
  - Describe associations, avoid the use of the term 'significance'
- **Test associations** – Requires *a priori* thinking
  - Have to have a Hypothesis in mind
  - Driven by the goal of obtaining a meaningful estimate of the exposure-outcome relationship, accounting for
    - Confounding – Mediation
    - Effect modification –Moderation –Interaction
  - Depending on the study design
    - Associations can be considered 'directional', i.e. "causal" hypothesis
- **Prediction**- Driven by statistical metrics

Source: Dr. Shofiqul Islam, McMaster University

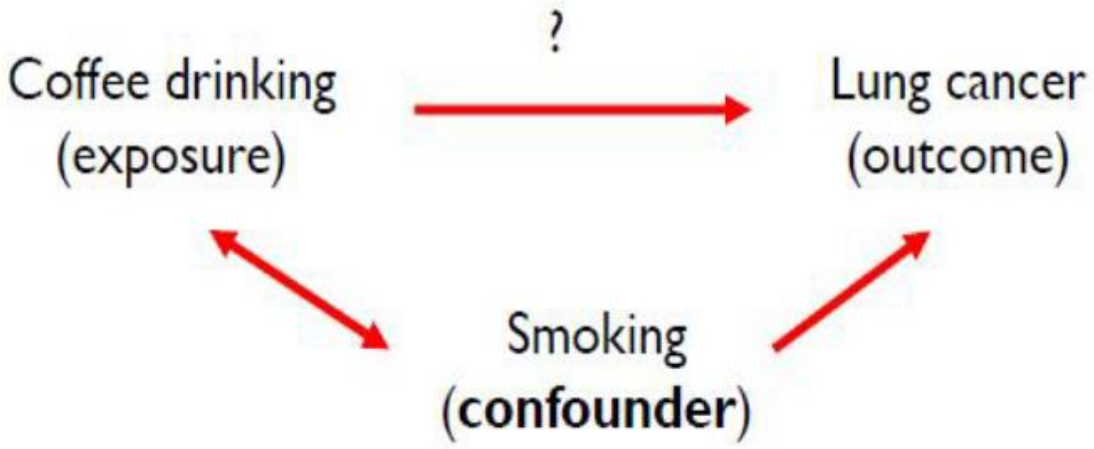
Common multivariable models

- Linear regression
- Logistic regression
- Ordinal logistic regression
- Nominal logistic regression
- Poisson regression
- Negative binomial regression
- Cox proportional hazard model (separate workshop)
- Etc.

Why do we Need Separate Regression Model?

- Different Assumptions
- Type of Dependent Variable
- Data Structure
- Research Objectives
- Model Performance
- Model Complexity

Confounding



- ✓ Confounder must be an independent risk factor for the outcome either a causal factor or a surrogate.
- ✓ Confounder must be associated with the exposure (e.g. smoking and coffee drinking).
- ✓ Confounder cannot be an intermediate variable between the exposure and the outcome (e.g. smoking is not caused by coffee drinking)
- ✓ A confounder can strengthen or weaken a relationship in the model
- ✓ We need to adjust the confounder in our model to find the true association.

Source: Dr. Shofiqul Islam, McMaster University

Confounding

Statistically, do 2 models:

- $Y = \beta_0 + \beta_1 X$
- $Y = \beta_0 + \beta_1 X + \beta_2 C$

'Compare' β_1 with β_1

We can include the variable in the model along with the exposer variable

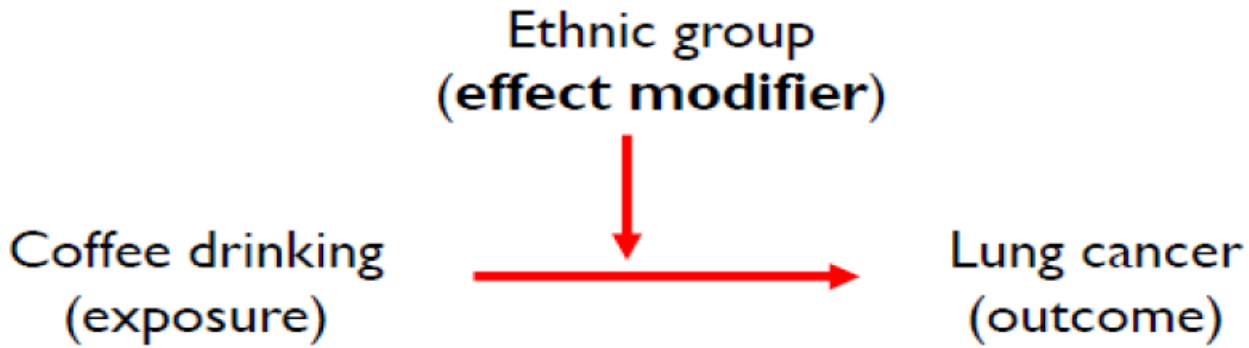
Look for 10% change in coefficient as a rule of thumb can be a useful to find out the confounder.

Source: Dr. Shofiqul Islam, McMaster University

Effect modifiers

- The relationship between the exposure and the outcome is different in different levels of a 3rd variable ('interaction' in statistics)

- Example:



- Question: Is the relationship between coffee drinking and lung cancer different in different ethnic groups?

- How do we check that?
  - Test of interaction – If significant - Stratified analysis

Source: Dr. Shofiqul Islam, McMaster University

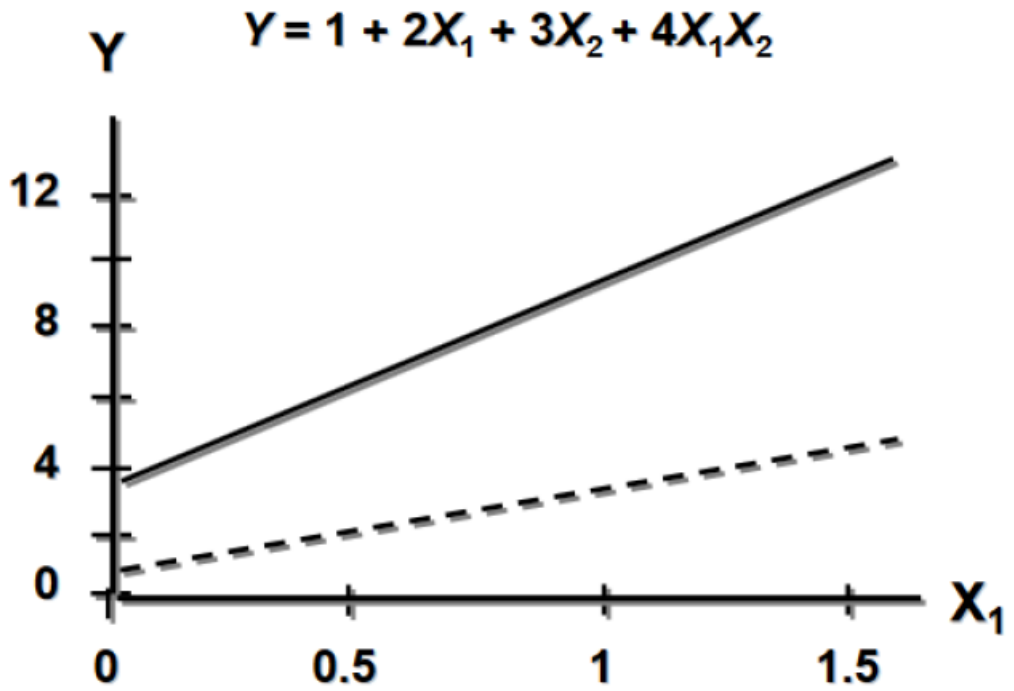
Effect modification = interaction

Interaction

= Effect modifier

= Moderator

- An interaction means that the effect of X on Y depends on the level of a third variable.
- No causal sequence is implied by interaction.
- Also known as modification or moderation



Source: Dr. Shofiqul Islam, McMaster University

Multicollinearity

- Refers to a situation when two independent(X) variables are highly correlated?
- Regression coefficients can change dramatically according to whether other variables are included or excluded
- Hard to interpret the effect of X1 when X2 held constant if they are highly correlated
- Reduces the precision of the estimated coefficients, which weakens the statistical power: May lead to an unreliable p

Source: Dr. Shofiqul Islam, McMaster University

Multicollinearity before fitting the model

Make the correlationmatrix between exposures (X) variables

Look for large correlation coefficients between exposures (X) variables

Multicollinearity after fitting the model

Collinearity – Metrics

- **Variance Inflation Factor (VIF)** –how correlated is an independent variable with other variables in the model
  - VIF=1 - there is no correlation between this independent variable and any others.
  - 1<VIF<5 - there is a moderate correlation, but it is not severe enough to warrant corrective measures.
  - VIF>5~10 - critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable
- **Condition index** – looking at the correlation matrix –square root of the ratio of 1st eigenvalue (principal component) to the others
  - If CI >= 30, Moderate or severe collinearity [see KKNR Sec. 14.5]

Source: Dr. Shofiqul Islam, McMaster University

Linear regression

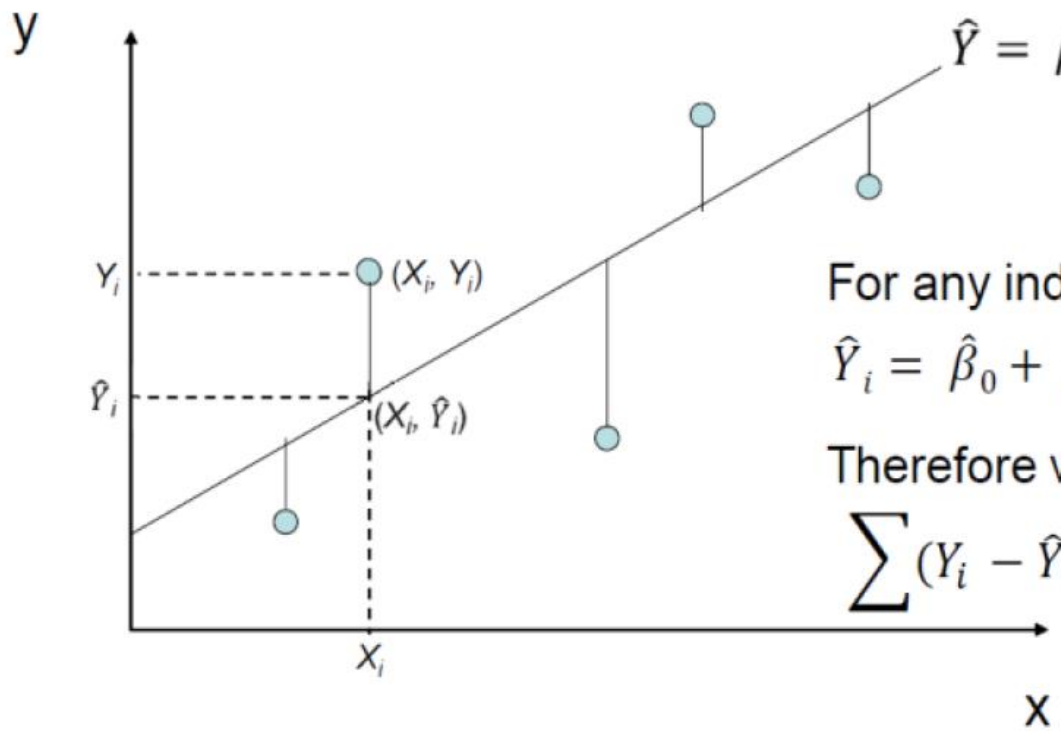
Assumptions of linear regression

Linear regression assumes that…

1. Linearity: True mean of y (contentious) is a linear function of x
2. Y is distributed normally at each value of X
3. The variance of Y at every value of X is constant (homogeneity of variances)
4. The observations are independent

Source: Dr. Dipak Kumar Mitra, North South University

Geometry of Least Square



For any individual Y_i,

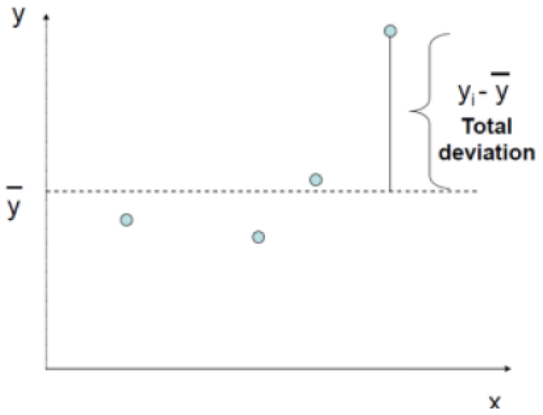
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Therefore we minimize:

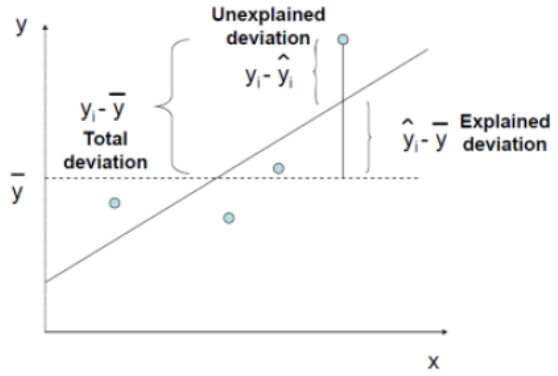
$$\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

Source: Dr. Shofiqul Islam, McMaster University

Variance decomposition



$$Var\ Y = \frac{\Sigma(y_i - \bar{y})^2}{n-1}$$



$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2$$

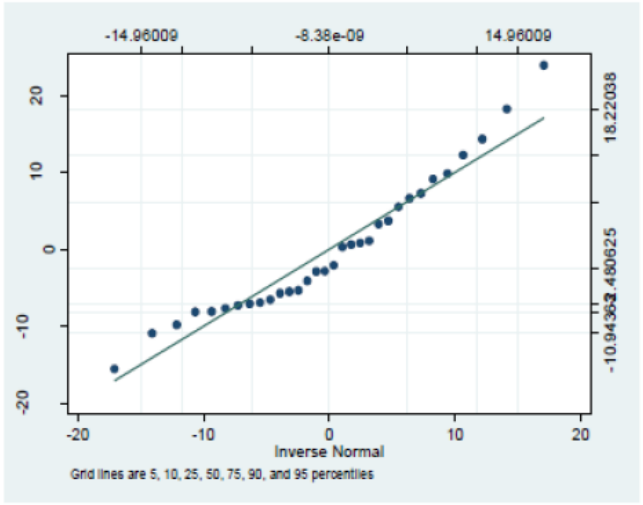
Total SS Residual SS Regression SS

Source: Dr. Shofiqul Islam, McMaster University

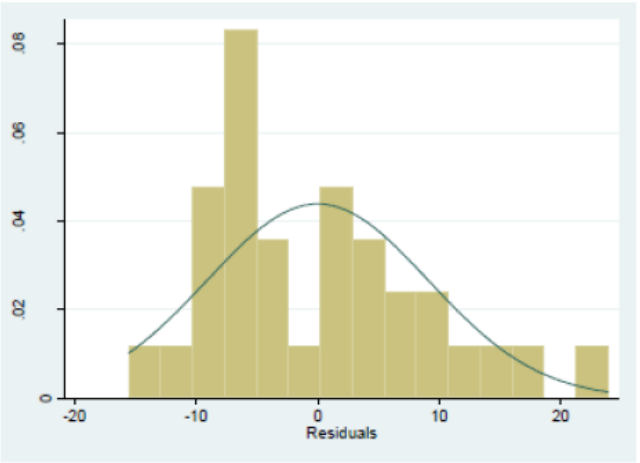
Assessing ‘goodness of fit’ assumptions

Normality – of residuals or of Y?

- Q-Q plot, histogram
- Shapiro-Wilk test



. qnorm residuals, grid



. histogram residuals, normal bin(15)

Source: Dr. Shofiqul Islam, McMaster University

What is Logistic Regression?

- So far, we have focused on continuous dependent variables and conducted simple or multivariable linear regression.
- Many situations where the dependent variable is binary or categorical.

✓ Dead vs. Alive

✓ CVD vs No CVD

✓ Employed vs. Unemployed

✓ Guilty vs. Not guilty

- Requires to consider a special type of model called logistic regression

Source: Dr. Shofiqul Islam, McMaster University

Linear vs Logistic Regression

- Linear Regression:
  - $E(Y|X) = \beta_0 + \beta_1 X$
- Leads to a linear probability model:
  - $E(Y|X) = Prob(Y=1|X) = p \quad 0 < p < 1$
  - Probabilities ranges between 0 and 1

- Problems with a linear relationship
  - A linear probability model can produce results outside of this range
  - Additive probabilities may not be suitable to quantify associations or effects with binary outcomes in biomedical and health research

Source: Dr. Shofiqul Islam, McMaster University

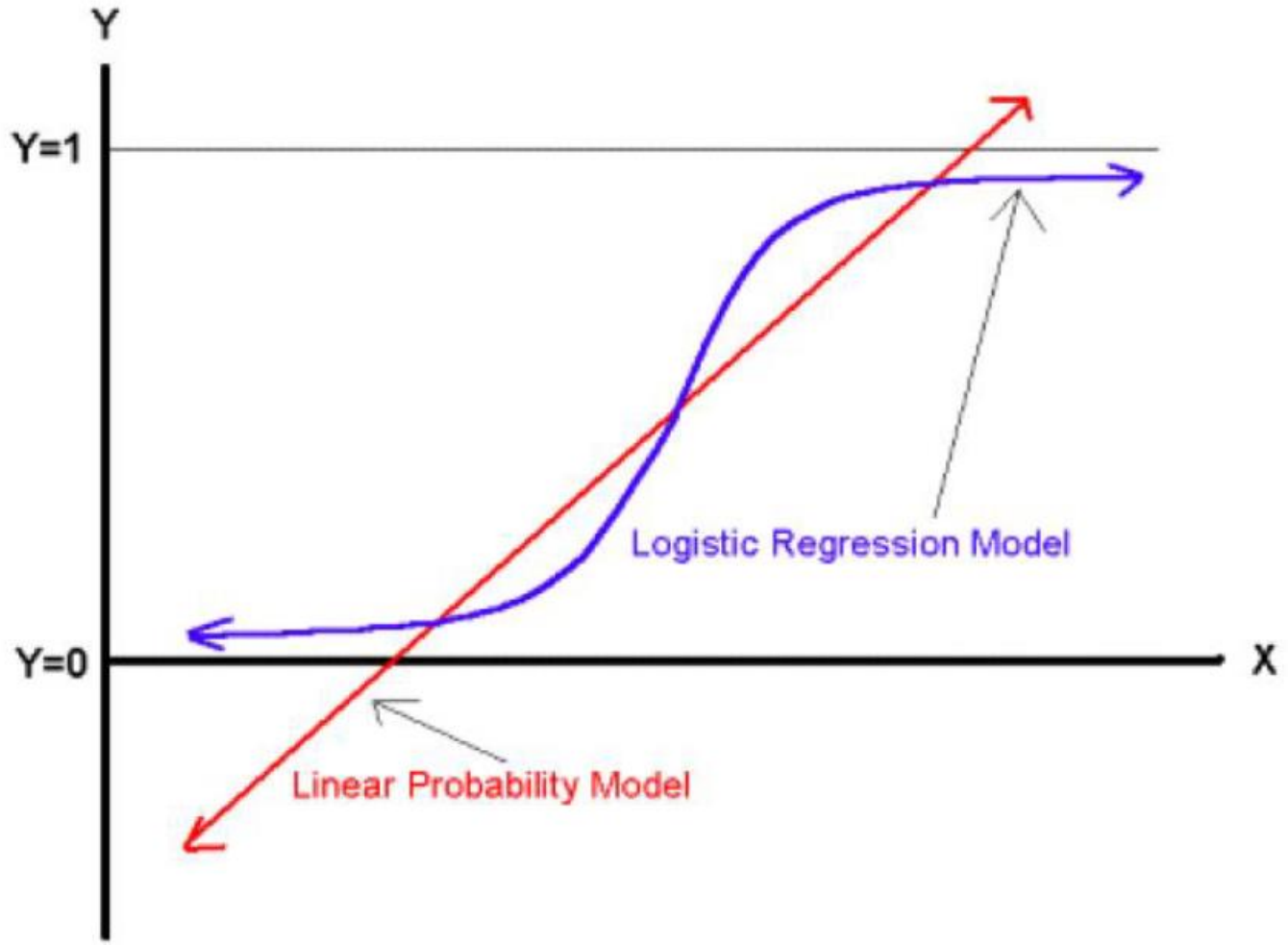
A new Regression Model

- Need to modify our regression equation so that:
  1. Predictions lie between 0 and 1
  2. Effects of covariates can be interpreted on a relative (multiplicative) scale.

- Solution
  - Use the log odds or logit of p to represent the
  - Outcome: $\text{logit } p = \ln\left(\frac{p}{1-p}\right)$
- Where p is the probability of having the outcome
  - When $p \rightarrow 0$, logit $p \rightarrow -\infty$
  - When $p \rightarrow 1$, logit $p \rightarrow \infty$

Source: Dr. Shofiqul Islam, McMaster University

Comparing the Linear vs Logistic fit



Source: Dr. Shofiqul Islam, McMaster University

A new Regression Model

- Logit transformation of the binary Outcome leads to Logistic Regression
- The logistic function: $\frac{1}{1+e^{-x}}$
- Inverse is called the logit: $\ln(\frac{x}{1-x})$
- Logistic regression models:
  - The log-odds (or logit) of a binary outcome as a straight-line function of covariates:

$$E[\ln(\frac{p}{1-p})|X] = \beta_0 + \beta_1 * X_1 + \ldots + \beta_k * X_k$$

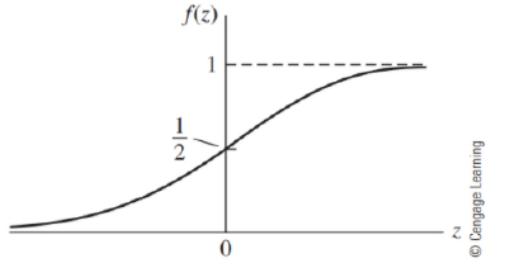


FIGURE 22.1 The logistic function $f(z) = \frac{1}{1+e^{-z}}$

Source: Dr. Shofiqul Islam, McMaster University

Logistic Function

- How do we transform the logit back to p ?

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$e^{ln\left(\frac{p}{1-p}\right)} = e^{(\beta_0+\beta_1 X)}$$

$$\frac{p}{1-p} = e^{(\beta_0+\beta_1 X)}$$

$$p = (1-p)\, e^{(\beta_0+\beta_1 X)}$$

$$p = e^{(\beta_0+\beta_1 X)} - pe^{(\beta_0+\beta_1 X)}$$

$$p + pe^{(\beta_0+\beta_1 X)} = e^{(\beta_0+\beta_1 X)}$$

$$p\left(1 + e^{(\beta_0+\beta_1 X)}\right) = e^{(\beta_0+\beta_1 X)}$$

$$p = \frac{e^{(\beta_0+\beta_1 X)}}{1 + e^{(\beta_0+\beta_1 X)}}$$

Source: Dr. Shofiqul Islam, McMaster University

Assumptions of Logistic Regression

- The logistic regression model assumes:
  - Outcome is a binary or dichotomous variable
  - There is a linear relationship between the logit of the outcome and each predictor variables
  - There is no influential values (extreme values or outliers) in the continuous predictors
  - There are no high correlations (multi-collinearity) among the predictors

Source: Dr. Shofiqul Islam, McMaster University

Logistic Regression

- Goodness of fit is examined using
  - Measures of predictive ability:
    - Pseudo R^2 by McFadden (1974)
    - Generalized R^2 by Cox-Snell (1989)
    - Tjur (2009) coefficient of discrimination
    - Diagnostic test criteria - sensitivity/specificity, area under the ROC curve
- Goodness of fit statistics –
  - Deviance and Pearson chi-squared statistics
  - Hosmer-Lemeshow (1980) test
- Information criteria – Akaike (AIC) & Bayesian (BIC)

Source: Dr. Shofiqul Islam, McMaster University

Nominal Logistic Regression

- Nominal (polytomous) logistic regression
- The coefficients of the model give the probability of being in a particular category relative to the reference category
  - Calculate OR

$$\Pr(y = 1) = \frac{1}{1 + e^{X\beta^{(2)}} + e^{X\beta^{(3)}}}$$

$$\Pr(y = 2) = \frac{e^{X\beta^{(2)}}}{1 + e^{X\beta^{(2)}} + e^{X\beta^{(3)}}}$$

$$\Pr(y = 3) = \frac{e^{X\beta^{(3)}}}{1 + e^{X\beta^{(2)}} + e^{X\beta^{(3)}}}$$

Source: Dr. Shofiqul Islam, McMaster University

Logistic Regression with >2 Categories

- **Ordinal** logistic regression
  - Model assumes proportional odds – equal 'distance' between categories
  - The assumption of proportional odds implies the ORs of predictors will be the same for all 3 dichotomous divisions of the outcome
    - Example: 4 ordered categories None, Mild, Mod, Sev have 3 ways to divide the outcome into two categories while preserving the order:
      - None vs Mild, Mod, Sev
      - None, Mild vs Mod, Sev
      - None, Mild, Mod vs Sev
- The probability of the j^{th} observation being in an i^{th} category is modeled – you obtain the accumulated probabilities

$$\Pr(\text{outcome}_j = i) = \Pr(\kappa_{i-1} < \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + u_j \leq \kappa_i)$$

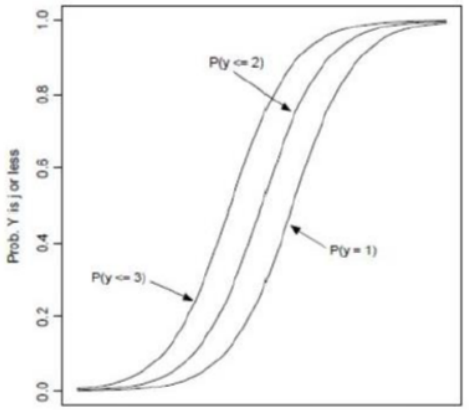
Source: Dr. Shofiqul Islam, McMaster University

Proportional Odds Model

- With k = 1, 2, 3, 4 ordinal categories

$$ln\left[\frac{\Pr(Y \geq k)}{1-\Pr(Y<k)}\right] = \beta_{01} + \beta_{02} + \beta_{03} + \beta_1 X_1,$$

- Have 3 intercepts but only one coefficient for each predictor
- Sometimes called ‘cumulative logit’



Source: Dr. Shofiqul Islam, McMaster University

Contact

Book an appointment with DASH: https://library.mcmaster.ca/services/dash

Contact DASH: Data Analysis Support Hub: libdash@mcmaster.ca

Let’s move to the coding part