Recitation 9

Rachel Lee
Nov. 10, 2021

Introduction to the GLM

The general linear model is a general framework for linear regression that assumes one (or more, in the case of multivariate analysis) outcome(s) may be modeled by a linear combination of known predictors plus a random error term.

- data = fit + residual (Tukey, 1977)
-

$$\begin{array}{ccccc} \text{observed value on dependent variable} & = & \text{sum of effects of 'allowed for' factors} & + & \text{sum of effects from other factors} \end{array}$$

- The 'allowed for' factors are variables that are explicitly accounted for in the statistical model.
- The 'other' factors are assumed to be random (i.e., not systematically related to the outcome).

Notation for the GLM

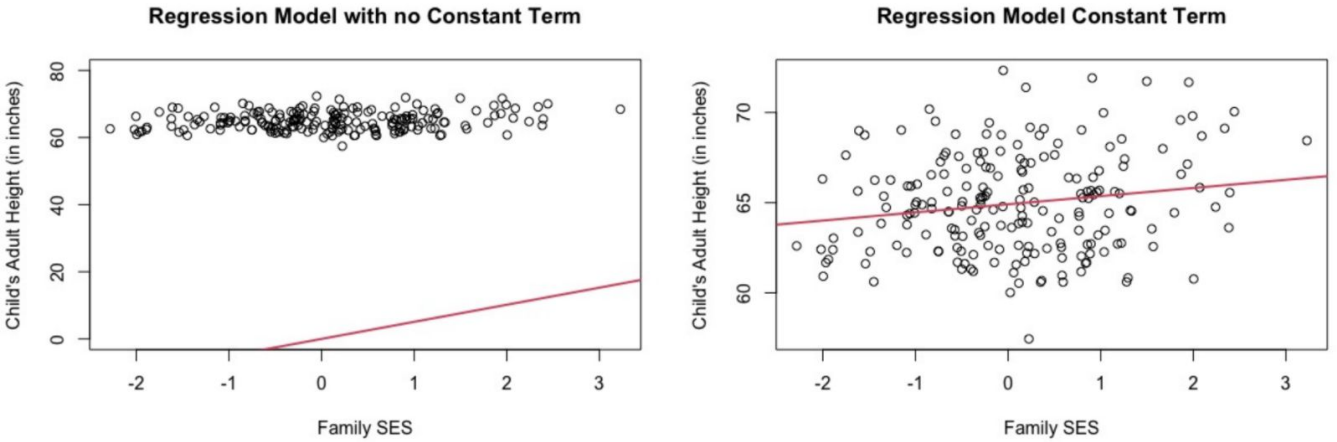
- There are three primary ways that GLMs may be written.
  1. in scalar notation with an index for participant,
  2. in vector notation, with a vector for each variable, and
  3. in matrix notation.
- Scalars are single numbers; vectors are ordered lists of numbers; matrices are ordered two-dimensional arrays.

An Example to Demonstrate Notation

- Consider using a GLM to model a child's adult height based on their mother's height.
  - Suppose a sample of data on $n = 200$ children was collected. The data can be organized in vectors, where $\mathbf{y}$ is child's adult height (in.), $\mathbf{x_1}$ is mother's height, $\mathbf{x_2}$ is family socioeconomic status (on a scale of about -3 to 3), etc.
  - Data values:

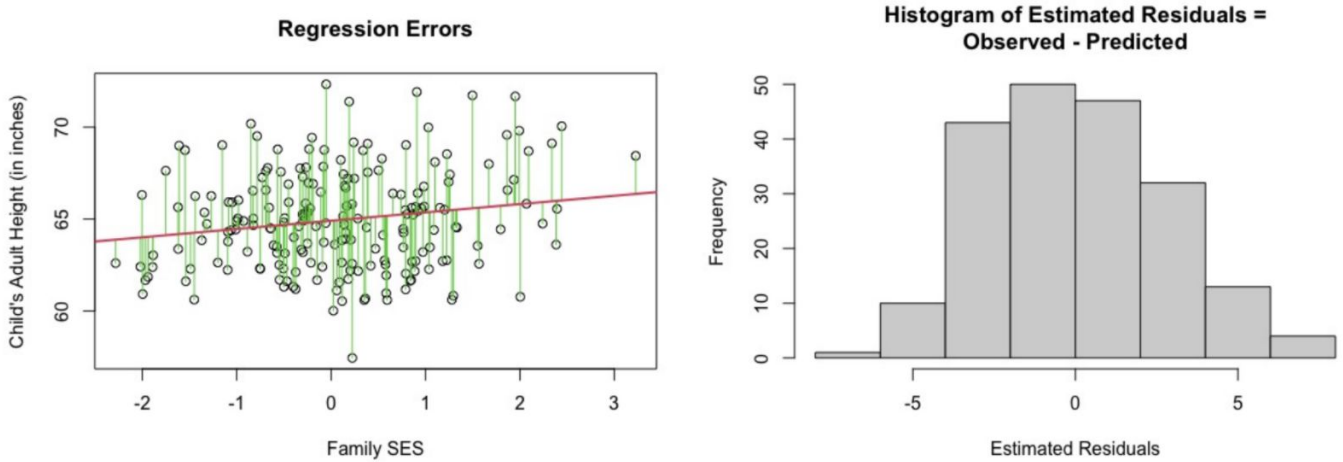
$$\mathbf{y} = \begin{bmatrix} Y_1 = 63 \\ Y_2 = 72 \\ \vdots \\ Y_{200} = 66 \end{bmatrix}, \mathbf{x_1} = \begin{bmatrix} X_{1,1} = 62 \\ X_{2,1} = 66 \\ \vdots \\ X_{200,1} = 63 \end{bmatrix}, \mathbf{x_2} = \begin{bmatrix} X_{1,2} = -0.54 \\ X_{2,2} = 1.20 \\ \vdots \\ X_{200,2} = 0.21 \end{bmatrix}$$

The Need for a Constant Term



Both plots show the same data on different scales. The plot on the left includes the regression line fit with no constant term; note that it is forced to go through the meaningless point at (0,0). The plot on the right includes the regression line with a constant term; this line is not forced to go through the origin; instead, the intercept is allowed to be freely estimated by the data.

An Example to Demonstrate Notation



In the plot on the left, the green vertical lines represent the estimates for ϵ_i for each participant i based on the estimated regression model in red. The plot on the right shows the estimated residuals in a histogram.

An Example to Demonstrate Notation

- The final model here is $Y_i = \beta_0 + \beta_1 X_{2i} + \epsilon_i$, which is called a *simple* linear regression model, because it has only one predictor.
- Consider adding additional predictors such as mother's height (inches), father's height (inches), yes/no exposure to chronic illness or disease, yes/no exposure to drugs or medications that can alter growth, and stress level. The *multiple* linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \beta_1 X_{3i} + \beta_1 X_{4i} + \beta_1 X_{5i} + \epsilon_i$$

- The model above involves a constant term, β_0, a *linear combination* of the 'allowed for' factors, and a random error term to pick up the remaining slack in the variation in the data that is not explained by the 'allowed for' factors.

An Example to Demonstrate Notation

Three ways to write the same model:

- scalar notation:
  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \beta_1 X_{3i} + \beta_1 X_{4i} + \beta_1 X_{5i} + \epsilon_i$
- vector notation: $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \beta_3 \mathbf{x_3} + \beta_4 \mathbf{x_4} + \beta_5 \mathbf{x_5} + \boldsymbol{\epsilon}$
- matrix notation: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & X_{1,3} & X_{1,4} & X_{1,5} \\ 1 & X_{2,1} & X_{2,2} & X_{2,3} & X_{2,4} & X_{2,5} \\ 1 & X_{3,1} & X_{3,2} & X_{3,3} & X_{3,4} & X_{3,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{200,1} & X_{200,2} & X_{200,3} & X_{200,4} & X_{200,5} \end{bmatrix} \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

	Model	Least-Squares Estimates	Errors
Full:	$Y_i = \mu + \varepsilon_{i_F}$	$\hat{\mu} = \bar{Y}$	$\sum e_{i_F}^2 = \sum (Y_i - \bar{Y})^2$
Restricted:	$Y_i = \mu_0 + \varepsilon_{i_R}$	No parameters estimated	$\sum e_{i_R}^2 = \sum (Y_i - \mu_0)^2$

One-Way Between-Subjects Design

- Sometimes, we want to look at more than two groups of data and compare them. We want to see if more than two groups of data are different. While we could use T-tests to compare the means from two different groups of data, but we need a different kind of test when comparing three or more groups.
- We can use a 1-Way ANOVA test to compare three or more groups or conditions in an experiment. A 1-Way ANOVA can help you find out if the means for each group / condition are significantly different from one another or if they are relatively the same. If the means are significantly different, you can say that the variable being **manipulated**, your Independent Variable (IV), had an effect on the variable being **measured**, your Dependent Variable (DV).
- It's called one-way because we use this test to analyze data from experiments that have only **one IV**. If we were analyzing data from experiments with more than one IV, we would need to use a different test.

One-Way Between-Subjects Design

- The independent samples t test uses the difference between the two group means as a measure of variability between groups and uses the standard error of the difference between means as a measure of error variability.

$$t = \frac{\text{Difference Between the Two Group Means}}{\text{Standard Error of the Difference Between Means (Error)}}$$

What is ANOVA?

- **AN**alysis **O**f **Va**riance. With ANOVA,
- We analyze and compare the variability of scores between conditions and within conditions. This helps us find out if the IV had a significant effect on the DV.
- Data could be analyzed with an independent samples ANOVA. The results based on a t-test will always match those based on an ANOVA. The logic of the ANOVA is very similar to that of the t test. Again, a ratio of variability between the groups to error variability is calculated. The resulting statistic is referred to as the F-ratio.

$$F\text{-ratio} = \frac{\text{Variability Between Group (Mean Square}_{\text{between groups}})}{\text{Error Variability (Mean Square}_{\text{error}})}$$

$$F\text{-ratio} = \frac{\text{Treatment Effect + Systematic Error}}{\text{Random Error}}$$

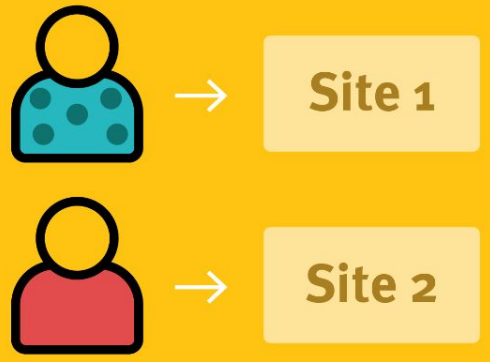
$$F = \frac{(E_R - E_F)/(df_R - df_F)}{E_F / df_F}$$

T-test vs. ANOVA

- Similarities:
  - Both tests are used to determine if there are statistically significant group differences, based on **averages**.
  - Both require a categorical IV with a continuous DV
- Example: Does caffeine consumption (IV) effect attention (DV) in young adults?

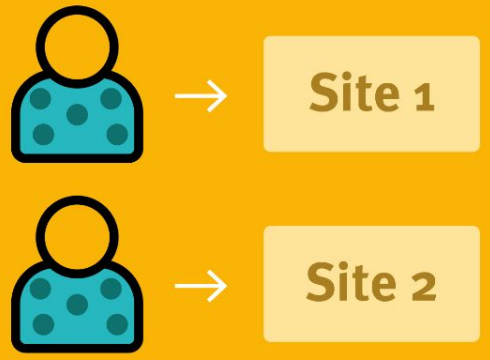
t-Test		ANOVA	
2 levels	1 categorical IV	At least 3 levels	At least 1 categorical IV
Example IV: Caffeine consumption *Levels*: • 0 oz. coffee consumption • 4 oz. coffee consumption		**Example IV: Caffeine consumption** *Levels*: • 0 oz. coffee consumption • 4 oz. coffee consumption • 8 oz. coffee consumption	

Between-Subjects



vs.

Within-Subjects



Numeric Example

Assume that you work in the research office of a large school system. For the last several years, the mean score on the WISC-R, which is administered to all elementary school children in your district, has been holding fairly steady at about 98. A parent of a hyperactive child in one of your special education programs maintains that the hyperactive children in the district are actually brighter than this average. To investigate this assertion, you randomly select the files of six hyperactive children and examine their WISC-R scores.

TABLE 3.1
HYPERACTIVE CHILDREN'S WISC-R SCORES

Full-Model Analysis

IQ Scores Y_i	*Prediction Equations*	*Parameter Term* $\hat{\mu}$	*Error Scores* $e_{i_F} = Y_i - \hat{\mu}$	*Squared Errors* $e_{i_F}^2$
96	$= \hat{\mu} + e_1$	104	−8	64
102	$= \hat{\mu} + e_2$	104	−2	4
104	$= \hat{\mu} + e_3$	104	0	0
104	$= \hat{\mu} + e_4$	104	0	0
108	$= \hat{\mu} + e_5$	104	+4	16
110	$= \hat{\mu} + e_6$	104	+6	36
$\sum = 624$			$\sum = 0$	$E_F = 120$
$\bar{Y} = 104$				

Restricted-Model Analysis

IQ Scores Y_i	*Prediction Equations*	*Parameter Term* μ_0	*Error Scores* $e_{i_R} = Y_i - \mu_0$	*Squared Errors* $e_{i_R}^2$
96	$= \mu_0 + e_1$	98	−2	4
102	$= \mu_0 + e_2$	98	4	16
104	$= \mu_0 + e_3$	98	6	36
104	$= \mu_0 + e_4$	98	6	36
108	$= \mu_0 + e_5$	98	10	100
110	$= \mu_0 + e_6$	98	12	144
				$E_R = 336$