# Getting Started

1. The General Linear Model
2. The Design Matrix
3. One-way ANOVA as Regression
4. anova() vs Anova() in R

<u>Examples (For Illustration!):</u>

1. Design Matrix Example
2. ANOVA as Regression Example

# Some perspective

- In STAT511, we presented simple linear regression and one-way ANOVA as separate models.
- The general linear model framework can be used for both multiple regression and one-way ANOVA.
- Most of STAT512 will focus on specific examples of the general linear model: multiple regression, polynomial regression, ANCOVA, factorial ANOVA.
- Note that general linear models are different from generalized linear models. Logistic regression is an example of a generalized linear model.

# Some Review

- Simple Linear Regression (numerical response, numerical predictor)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- One-way ANOVA (numerical response, categorical predictor or factor)

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (\text{Default, Effects Model})$$

$$\mu = \text{baseline or intercept, } \alpha_i = \text{effects or adjustments}$$

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad (\text{Alternate, No Intercept, Means Model})$$

$$\mu_i = \text{population group/trt means}$$

# 1. The General Linear Model

- A single general model can be used for multiple regression models in which a response (Y) is related to a set of numerical predictors and for models that relate Y to a set of categorical predictors.
- The general linear model has the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

- For multiple regression models, the X's represent numerical predictors (ex: weight).
- When Y is related to a set of categorical predictors (ex: treatment A, B or C) the X's represent indicator (or dummy) variables (coded as 0 or 1).

- When a categorical predictor (factor) is included in the model, R creates indicator variables automatically.
- The general linear model can also include numerical variables raised to powers (polynomial regression) and products of either numerical or indicator variables (interaction terms).
- The general linear model can also be used for the case where Y is related to both numerical and categorical variables. A particular example of this is ANCOVA.

# 2. The Design Matrix

- First let's discuss the matrix form of the simple linear regression model.
- **Design Matrix Example (Corn Data):** Response of Y = corn yield (bu/plot) to X = fertilizer (lbs/plot). Total of n = 10 observations. Assuming a linear relationship is reasonable, then simple linear regression is appropriate:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

# Matrix form of Simple Linear Regression

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_9 \\ y_{10} \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_9 \\ \beta_0 + \beta_1 x_{10} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_9 \\ \varepsilon_{10} \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_9 \\ y_{10} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_9 \\ 1 & x_{10} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_9 \\ \varepsilon_{10} \end{pmatrix}$$

$$\mathbf{y}_{10\times 1} = \mathbf{X}_{10\times 2}\beta_{2\times 1} + \varepsilon_{10\times 1}$$

1. The matrix "**X**" is called the design (or model) matrix.
2. The design matrix has n rows and (k+1) columns. For multiple regression, k is the number of predictors. So # rows = # observations and # columns = # parameters (or coefficients) in the model.
3. The columns of the design matrix define the predictor variables, and hence the model. We sometimes write the design matrix as a way of describing the design.
4. If two models have the same design matrix, then they are the same model.
5. Very complicated formulas for parameter estimates, residuals, sums of squares, etc. can all be written in a simple form using matrix notation. All these formulas involve the design matrix, e.g.:

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

# 3. One-way ANOVA as Regression

- **ANOVA as Regression Example**: We have t = 3 treatments with 2 observations per treatment. Total of n = 6 observations.
- Since the goal is to compare mean response for the 3 treatments using one-way ANOVA, we want to consider treatment as a factor (or categorical predictor).

| Y | trt |
|---|---|
| 6.3 | 1 |
| 5.9 | 1 |
| 4.3 | 2 |
| 4.8 | 2 |
| 3.7 | 3 |
| 3.9 | 3 |

- When a categorical predictor (factor) is included in the model, R creates indicator variables automatically.
- We will create indicator variables "by hand" for illustration. We will call the indicator variables X1, X2, X3.

Let $X_1$ = 1 if observation is trt 1.
= 0 if not trt 1.

$X_2$ = 1 if observation is trt 2.
= 0 if not trt 2.

$X_3$ = 1 if observation is trt 3.
= 0 if not trt 3.

| Y | trt | X1 | X2 | X3 |
|---|---|---|---|---|
| 6.3 | 1 | 1 | 0 | 0 |
| 5.9 | 1 | 1 | 0 | 0 |
| 4.3 | 2 | 0 | 1 | 0 |
| 4.8 | 2 | 0 | 1 | 0 |
| 3.7 | 3 | 0 | 0 | 1 |
| 3.9 | 3 | 0 | 0 | 1 |

- Recall the one-way ANOVA model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- But **one-way ANOVA is equivalent to a regression with indicator variables** X1, X2, X3:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- For an observation from trt 1 ($x_{i1} = 1$, $x_{i2} = 0$, $x_{i3} = 0$)

$$Y_i = \beta_0 + \beta_1 + \varepsilon_i$$

- For an observation from trt 2 ($x_{i1} = 0$, $x_{i2} = 1$, $x_{i3} = 0$)

$$Y_i = \beta_0 + \beta_2 + \varepsilon_i$$

- For an observation from trt 3 ($x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 1$)

$$Y_i = \beta_0 + \beta_3 + \varepsilon_i$$

# Overparameterization

- It is impossible to uniquely estimate four parameters (coefficients) when there are only three trt groups (and hence three means). We must restrict the parameter estimates in some way.
- Example: suppose $\mu_1=1$, $\mu_2=2$, $\mu_3=3$.

  If $\beta_0=0$, then $\beta_1=1$, $\beta_2=2$, $\beta_3=3$.

  But we could also have $\beta_0=1$, then $\beta_1=0$, $\beta_2=1$, $\beta_3=2$.
- To avoid overparameterization, R omits the first indicator variable. This is similar to setting $\alpha_1=0$ (or $\beta_1=0$).
- The intercept ($\mu$ or $\beta_0$) is the mean for the first level of the categorical predictor. Hence the first group acts as a baseline or reference group.

# Parameter Interpretation

- The ANOVA model and Regression on indicator variables are really the same model (but with different names for the parameters/coefficients and different subscripting of the y's).
- For our example, there are three means, so a "saturated" model requires three parameters/coefficients (not four!). R omits the first indicator variable.
- Parameter/coefficient interpretation:

  $\beta_0 = \mu =$ Trt 1 mean

  $\beta_2 = \alpha_2 =$ Trt 2 mean minus Trt 1 mean

  $\beta_3 = \alpha_3 =$ Trt 3 mean minus Trt 1 mean

# ANOVA as Regression Example

- We consider fitting the ANOVA model 4 different ways.
- Model1: Fit the default “effects” model using the lm() function. This will be our typical approach! Look at parameter (coefficient) estimates, model matrix, ANOVA table and emmeans.
- Model2: Fit the alternate “no intercept” or “means” model using the lm() function.
- Model3: Fit the “effects” model “by hand” by creating the 3 indicator variables. This model is overparameterized, for illustration only.
- Model4: Fit the “effects” model “by hand” using just 2 indicator variables. This is equivalent to Model1.

- When a factor variable (categorical predictors) is included in the model, R creates indicator variables.
- When one-way ANOVA is the analysis of interest, typical research questions would be addressed by estimating, testing and comparing means. So we will typically focus on the `anova()` and `emmeans()` output for the one-way ANOVA analysis.
- Notice that the parameter/coefficient estimates do not directly address these standard research questions! In other words, the `summary()` information is not of primary interest for the one-way ANOVA analysis.
- By reparameterizing (with the no intercept approach) the coefficients are easier to interpret. One reason to consider different parameterizations is for convenience of interpretation (and testing).

# Matrix form of one-way ANOVA (Regression Notation)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 0 + \beta_2 0 \\ \beta_0 + \beta_1 0 + \beta_2 0 \\ \beta_0 + \beta_1 1 + \beta_2 0 \\ \beta_0 + \beta_1 1 + \beta_2 0 \\ \beta_0 + \beta_1 0 + \beta_2 1 \\ \beta_0 + \beta_1 0 + \beta_2 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

$$\mathbf{y}_{\mathbf{6\times 1}} = \mathbf{X}_{\mathbf{6\times 3}} \beta_{3\times 1} + \varepsilon_{6\times 1}$$

# Matrix form of one-way ANOVA
# (ANOVA Notation)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} \mu & +\alpha_2 & 0+\alpha_3 & 0 \\ \mu & +\alpha_2 & 0+\alpha_3 & 0 \\ \mu & +\alpha_2 & 1+\alpha_3 & 0 \\ \mu & +\alpha_2 & 1+\alpha_3 & 0 \\ \mu & +\alpha_2 & 0+\alpha_3 & 1 \\ \mu & +\alpha_2 & 0+\alpha_3 & 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}$$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}$$

$$\mathbf{y}_{\mathbf{6\times 1}} = \mathbf{X}_{\mathbf{6}\times 3}\beta_{3\times 1} + \varepsilon_{6\times 1}$$

# 4. anova() versus Anova() in R

- This topic does not really fit with the rest of these notes, but because it is such a point of confusion I want to mention it at the beginning of the course.
- The `anova()` function is part of the `stats` package (base R).
- The `Anova()` function is part of the `car` package.
- There is also an `aov()` function (part of the stats package) but this is just a wrapper for `lm()` and should not be used with unbalanced designs. We will not use `aov()` in STAT512.

- With only a single predictor variable or factor, there is no difference between `anova()` and `Anova()`.
- However, when there are multiple predictors or factors, there can be differences. **We are generally interested in the `Anova()` results!**
- When applied to an individual lm object, the `anova()` function will produce a sequential (or Type I) ANOVA table. The resulting table will show tests produced by fitting a sequence of models to the data. The results depend on the order the variables are listed.
- The `Anova()` function will produce unique or marginal (or Type III) ANOVA table. The resulting table will show tests for adding one of the predictors to a model that includes all the others. The results do NOT depend on the order the variables are listed.