Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

October 10, 2019

Topics

Objectives

Objectives:

- State the group means version of the one-factor ANOVA model and the treatment effects version.
- Interpret residuals and fitted values.
- Use residuals to check normality and constant standard deviation assumptions.

The One-Factor ANOVA Model (Two Versions)

- A ***statistical model*** is a mathematical representation of data, with components corresponding to the **two sources** of **variation**:

The One-Factor ANOVA Model (Two Versions)

- A ***statistical model*** is a mathematical representation of data, with components corresponding to the **two sources of variation**:
  - Deterministic variation in the data

The One-Factor ANOVA Model (Two Versions)

- A ***statistical model*** is a mathematical representation of data, with components corresponding to the **two sources of variation**:
    - Deterministic variation in the data
    - Random variation in the data

Example

Suppose $X \sim N(\mu, \sigma)$.

Example

Suppose $X \sim N(\mu, \sigma)$.

Let

$$\epsilon = X - \mu.$$

Example

Suppose $X \sim N(\mu, \sigma)$.

Let

$$\epsilon = X - \mu.$$

Then ϵ is a linear function of X, so ϵ is normally distributed (Slides 1). More precisely,

$$\epsilon \sim N(0, \sigma).$$

Example

Suppose $X \sim N(\mu, \sigma)$.

Let

$$\epsilon = X - \mu.$$

Then ϵ is a linear function of X, so ϵ is normally distributed (Slides 1). More precisely,

$$\epsilon \sim N(0, \sigma).$$

We can write X in the form of a **statistical model** as

$$X = \mu + \epsilon,$$

where

μ is the ***true mean*** of X (i.e. ***expected value***).
ϵ is a $N(0, \sigma)$ **random error** term.

- Recall that the *assumptions* for the **ANOVA F test** are that the I samples are drawn independently from $N(\mu_1, \sigma)$, $N(\mu_2, \sigma)$ …, $N(\mu_I, \sigma)$ populations.

- Recall that the *assumptions* for the **ANOVA $\boldsymbol{F}$ test** are that the $\boldsymbol{I}$ samples are drawn independently from $N(\mu_1, \sigma)$, $N(\mu_2, \sigma)$ ..., $N(\mu_I, \sigma)$ populations.
- We can write the assumptions in the form of a **statistical model** (next slide).

One-factor ANOVA Model (Group Means Version):

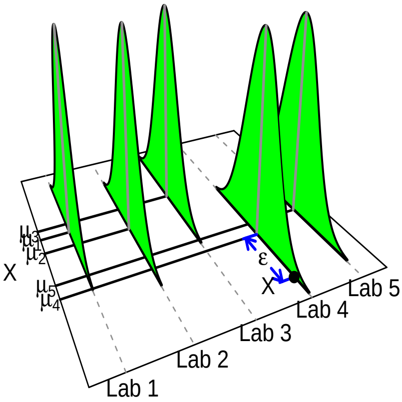
$$X_{ij} \;=\; \mu_i + \epsilon_{ij}\;, \tag{1}$$

where

μ_i is the ***true mean*** response for the ith treatment group

ϵ_{ij} are iid $N(0, \sigma)$ ***random errors***

One−Factor Analysis of Variance Model



- Sometimes a different (but equivalent) **statistical model** is used to describe data in a **one-factor ANOVA** context (next slide).

One-factor ANOVA Model (Treatment Effects Version):

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij} , \qquad (2)$$

where

μ is a constant called the ***true grand mean***
α_i is the ***treatment effect*** for ith treatment
ϵ_{ij} are iid $N(0, \sigma)$ ***random errors***

Proposition

If we define

$$\mu = \frac{\sum_{i=1}^{I} \mu_i}{I} \qquad \text{and} \qquad \alpha_i = \mu_i - \mu,$$

then

1. The two models (1) and (2) are equivalent.
2. $\sum \alpha_i = 0$.

Proposition

If we define

$$\mu \;=\; \frac{\sum_{i=1}^{I}\mu_i}{I} \qquad \text{and} \qquad \alpha_i \;=\; \mu_i - \mu,$$

then

1. The two models (1) and (2) are equivalent.
2. $\sum \alpha_i \;=\; 0$.

The first result holds because we can write μ_i in (1) as

$$\mu_i \;=\; \mu + (\mu_i - \mu) \;=\; \mu + \alpha_i\,.$$

Proposition

If we define

$$\mu = \frac{\sum_{i=1}^{I} \mu_i}{I} \qquad \text{and} \qquad \alpha_i = \mu_i - \mu,$$

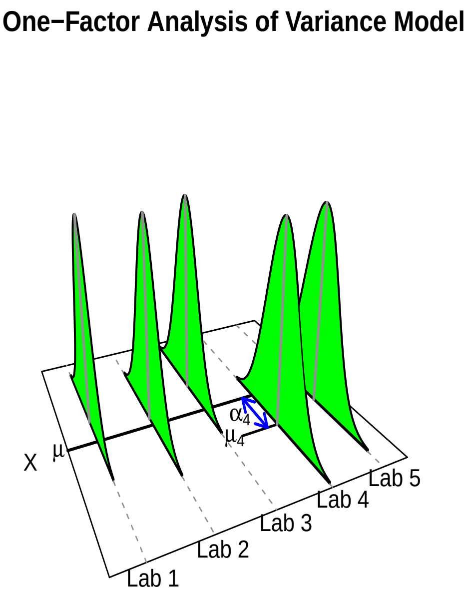
then

1. The two models (1) and (2) are equivalent.
2. $\sum \alpha_i = 0$.

The first result holds because we can write μ_i in (1) as

$$\mu_i = \mu + (\mu_i - \mu) = \mu + \alpha_i \,.$$

The second holds because deviations away from a mean always sum to zero.


One−Factor Analysis of Variance Model
X
μ
α4
μ4
Lab 1
Lab 2
Lab 3
Lab 4
Lab 5

- In terms of the **treatment effects version** of the **ANOVA model**, the hypotheses are:

$$H_0\colon \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$$
$$H_a\colon \text{Not all } \alpha_i\text{'s equal zero} \tag{3}$$

which are equivalent to the hypotheses about the μ_i's stated previously.

Estimating Model Parameters

- Recall that the **group means** version of the ANOVA model is

$$X_{ij} \;=\; \mu_i + \epsilon_{ij}\,,$$

and the **treatment effects** version is

$$X_{ij} \;=\; \mu + \alpha_i + \epsilon_{ij}\,.$$

Model Parameter Estimators: The unknown model parameters μ_i, μ, α_i, and σ are estimated by $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}}$, and $\hat{\boldsymbol{\sigma}}$ defined as:

Model Parameter	Estimator
μ_i	$\hat{\mu}_i = \bar{X}_{i.}$
μ	$\hat{\mu} = \bar{X}_{..}$
$\alpha_i = \mu_i - \mu$	$\hat{\alpha}_i = \bar{X}_{i.} - \bar{X}_{..}$
σ	$\hat{\sigma} = \sqrt{\text{MSE}}$

Predicted Values and Residuals

- The ***fitted value*** (or ***predicted value***) for the jth individual in the ith treatment group, $\hat{X}_{ij}$, is defined as:

$$\hat{X}_{ij} = \hat{\mu} + \hat{\alpha}_i$$

Predicted Values and Residuals

- The ***fitted value*** (or ***predicted value***) for the jth individual in the ith treatment group, $\hat{X}_{ij}$, is defined as:

$$\begin{aligned} \hat{X}_{ij} &= \hat{\mu} + \hat{\alpha}_i \\ &= \bar{X}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) \end{aligned}$$

Predicted Values and Residuals

- The ***fitted value*** (or ***predicted value***) for the jth individual in the ith treatment group, $\hat{X}_{ij}$, is defined as:

$$\begin{aligned} \hat{X}_{ij} &= \hat{\mu} + \hat{\alpha}_i \\ &= \bar{X}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) \\ &= \bar{X}_{i.} \end{aligned}$$

Predicted Values and Residuals

- The ***fitted value*** (or ***predicted value***) for the jth individual in the ith treatment group, $\hat{X}_{ij}$, is defined as:

$$\begin{aligned} \hat{X}_{ij} &= \hat{\mu} + \hat{\alpha}_i \\ &= \bar{X}_{\cdot\cdot} + (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) \\ &= \bar{X}_{i\cdot} \end{aligned}$$

  This is the value we'd predict, based on the data, for the response of the jth individual in the ith treatment group.

- The ***residual*** for the jth observation in the ith group, e_{ij}, is defined as

$$e_{ij} = X_{ij} - \hat{X}_{ij}$$

- The ***residual*** for the jth observation in the ith group, e_{ij}, is defined as

$$\begin{aligned} e_{ij} &= X_{ij} - \hat{X}_{ij} \\ &= X_{ij} - \bar{X}_{i.} \end{aligned}$$

- The ***residual*** for the jth observation in the ith group, e_{ij}, is defined as

$$\begin{aligned} e_{ij} &= X_{ij} - \hat{X}_{ij} \\ &= X_{ij} - \bar{X}_{i.} \end{aligned}$$

This is the deviation of the observed response X_{ij} away from the value model-predicted value.

Proposition

The residuals sum to zero within each treatment group, i.e.

$$\sum_j e_{ij} \;=\; 0 \qquad \text{for each } i = 1, 2, \ldots, I\,.$$

Therefore they sum to zero across all groups:

$$\sum_i \sum_j e_{ij} \;=\; 0\,.$$

Proposition

The residuals sum to zero within each treatment group, i.e.

$$\sum_j e_{ij} \;=\; 0 \qquad \text{for each } i = 1, 2, \ldots, I\,.$$

Therefore they sum to zero across all groups:

$$\sum_i \sum_j e_{ij} \;=\; 0\,.$$

This is because the **residuals** are just **deviations** away from the **group means.**

- **Comment**: We can write

$$X_{ij} \;=\; \hat{\mu} + \hat{\alpha}_i + e_{ij},$$

which resembles the **one-factorANOVA model.**

- **Comment**: We can write

$$X_{ij} \;=\; \hat{\mu} + \hat{\alpha}_i + e_{ij},$$

which resembles the **one-factorANOVA model.**

In particular, the **residual** e_{ij} corresponds to the **random error** term ϵ_{ij} in the model.

Checking the Model Assumptions

- For the **ANOVA F test**, we assume the ϵ_{ij}'s are iid $N(0, \sigma)$.

Checking the Model Assumptions

- For the **ANOVA F test**, we assume the ϵ_{ij}'s are iid $N(0, \sigma)$.

  Note that σ is assumed to be **constant** from one group to the next.

Checking the Model Assumptions

- For the **ANOVA F test**, we assume the ϵ_{ij}'s are iid $N(0, \sigma)$.

  Note that σ is assumed to be **constant** from one group to the next.

- **Checking the Normality Assumption**: Use a **histogram** or **normal probability plot** of the **residuals**.

Checking the Model Assumptions

- For the **ANOVA F test**, we assume the ϵ_{ij}'s are iid $N(0, \sigma)$.

  Note that σ is assumed to be **constant** from one group to the next.

- **Checking the Normality Assumption**: Use a **histogram** or **normal probability plot** of the **residuals**.

- **Checking the Constant σ Assumption**: Plot the **residuals** versus the **fitted values**.

Checking the Model Assumptions

- For the **ANOVA F test**, we assume the ϵ_{ij}'s are iid $N(0, \sigma)$.

  Note that σ is assumed to be **constant** from one group to the next.

- **Checking the Normality Assumption**: Use a **histogram** or **normal probability plot** of the **residuals**.

- **Checking the Constant σ Assumption**: Plot the **residuals** versus the **fitted values**.

  Usually, when σ *isn't* constant, it increases with the group mean (second plot, next slide).

Checking the Model Assumptions

- For the **ANOVA F test**, we assume the ϵ_{ij}'s are iid $N(0, \sigma)$.

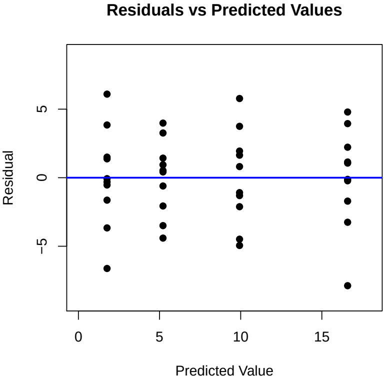
  Note that σ is assumed to be **constant** from one group to the next.

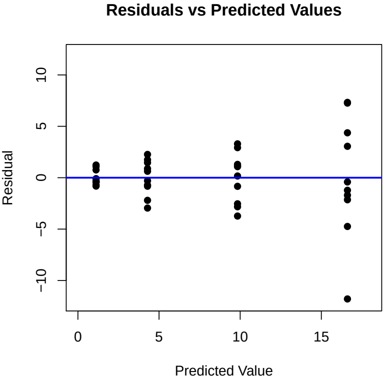
- **Checking the Normality Assumption**: Use a **histogram** or **normal probability plot** of the **residuals**.

- **Checking the Constant σ Assumption**: Plot the **residuals** versus the **fitted values**.

  Usually, when σ *isn't* constant, it increases with the group mean (second plot, next slide).

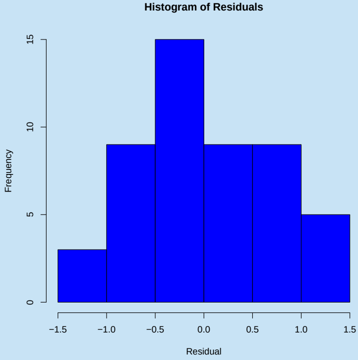
  Poisson data are an example.

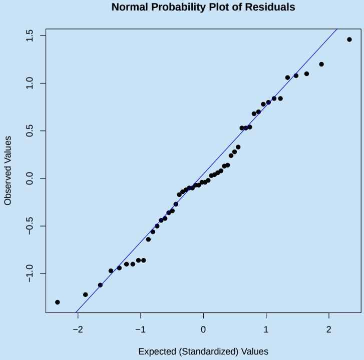

Residuals vs Predicted Values
Residual
Predicted Value
5
0
−5
0
5
10
15


Residuals vs Predicted Values
Residual
Predicted Value
10
5
0
−5
−10
0
5
10
15

Example

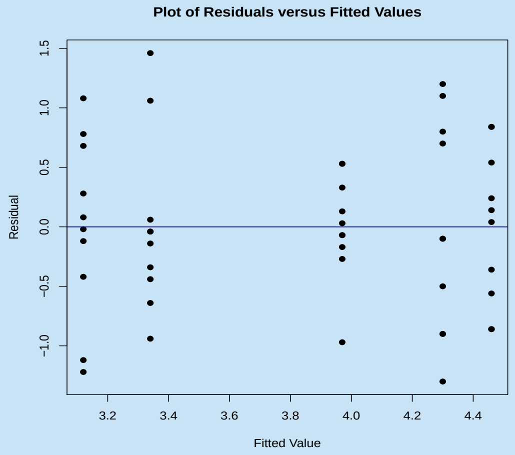
For the data on lead measurements from five labs, a **histogram** and **normal probability plot** of the **residuals** are below.





The plots suggest that the **normality assumption** appears to be met.

A scatterplot of the **residuals** versus the **fitted values** is on the next slide.


Plot of Residuals versus Fitted Values
Residual
1.5
1.0
0.5
0.0
−0.5
−1.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
Fitted Value

The plot supports the **constant σ assumption.**

The plot supports the **constant σ assumption.**

Therefore, the results of the **ANOVA F test** are valid.