Introduction to One-Way ANOVA

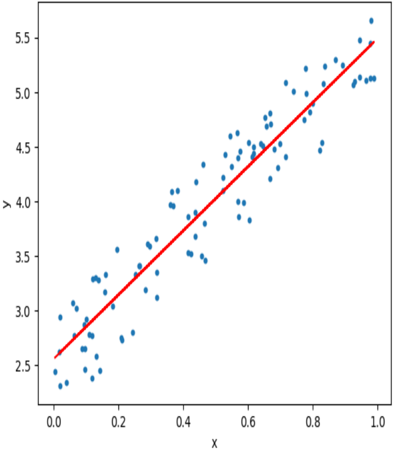
Mohammed Quazi

PhD Candidate – Statistics
Department of Mathematics & Statistics
University of New Mexico
mquazi@unm.edu
https://math.unm.edu/~mquazi/
https://github.com/mquazi

Overview

Regression

- Statistical modeling technique
- Simplest form is to predict one variable using another



Simple Linear Regression (SLR)

- Main question: What is the predicted value for a given point?
- Use only one variable to predict – predictor variable
- Predicted variable is called the response variable
- Model form:

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 x_t + \epsilon_t \\
x_t &= \text{values of predictor} \\
t &= 1, \ldots, n \text{ observations} \\
\epsilon_t &\text{ iid Normal}(0, \sigma^2)
\end{aligned}
\tag{1}
$$

One-Way ANOVA

- Main question: Does this factor affect the response?
- Factor is an explanatory variable which may be related to the response
- Factor levels are "values" of the factor
- Factor-level combinations are called treatments
- Model form:

$$
\begin{aligned}
Y_{it} &= \mu_i + \epsilon_{it} \\
&\text{factor/treatment (trt) } i = 1, 2 \\
&t = 1, \ldots, r_i \text{ observations for trt } i \\
\epsilon_{it} &\text{ iid Normal}(0, \sigma^2)
\end{aligned}
\tag{2}
$$

Comparison

Compare	**Regression**	**ANOVA**
Response	Quant	Quant
Factors	Quant	Quant or Qual
Data	Observational	Experimental
Relation	Linear/quadratic...	Not specified
Model (Simple)	$Y_t = \beta_0 + \beta_1 x_t + \epsilon_t$	$Y_{it} = \mu_i + \epsilon_{it}$
True mean	true mean wt $= \beta_0 + \beta_1 x$	$\mu_1 =$ true mean life length for nonsmokers $\mu_2 =$ true mean life length for smokers

ANOVA models

- ANOVA model seen so far is the Cell Means model: $Y_{it} = \mu_i + \epsilon_{it}$

$$\begin{aligned} i &= 1, \ldots, \nu \text{ trts} \\ r_i &= \# \text{ of obsns for trt } i \\ n &= \text{total } \# \text{ of obsns} \\ y_{it} &= t^{th} \text{ response observed for trt } i \\ \mu_i &= \text{true mean for trt } i \\ \epsilon_{it} &= \text{ random errors iid Normal}(0, \sigma^2) \end{aligned}$$

- Factor Effects model: $Y_{it} = \mu_{..} + \alpha_i + \epsilon_{it}$

$$\begin{aligned} \mu_{..} &= \text{ overall constant}, \quad (\mu_i = \mu_{..} + \alpha_i) \\ \alpha_i &= \text{ effect due to trt } i \end{aligned}$$

- Fitted values: $\hat{y}_{it} = \hat{\mu}_i = \hat{\mu}_{..} + \hat{\alpha}_i = \bar{y}_{i.}$
- Residuals: $e_{it} = y_{it} - \hat{y}_{it}$, for each trt i, $\sum_{t=1}^{r_i} e_i = 0$

ANOVA table for one factor

Source	DF	SS	MS=SS/DF	F	p-value
Trt	$\nu - 1$	ssTr	msTr	F^*	$P[F_{\nu-1,n-\nu} >$ observed $F^*]$
Error	$n - \nu$	ssE	msE		
Total	$n - 1$	ssTot			

- *F*-test

H_0 : No difference in the trts $\leftrightarrow$ $H_0 : \mu_1 = \mu_2 = \ldots = \mu_\nu = \mu_{..}$ $\leftrightarrow$ $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_\nu = 0$

vs.

H_1 : At least two trts differ $\leftrightarrow$ H_1 : Not all μ_i are equal $\leftrightarrow$ H_1 : Not all $\alpha_i = 0$

- Model under H_0 is the reduced model: $Y_{it} = \mu_{..} + \epsilon_{it}$ (all have same mean)
- Model under H_1 is the full model: $Y_{it} = \mu_i + \epsilon_{it} = \mu_{..} + \alpha_i + \epsilon_{it}$
- Reject H_0 if $F^* > F_{\nu-1,n-\nu,\alpha}$ or if p-value $< \alpha$
- If we conclude H_1, then the next step is analysis of factor level effects

Inferences for factor level effects

We may conduct inferences for

- trt means $\mu_i = \mu_{..} + \alpha_i$
- differences between trts $D = \mu_i - \mu_j = \mu_{..} + \alpha_i - \mu_{..} - \alpha_j = \alpha_i - \alpha_j$ (nonsmokers vs. smokers)
- contrasts $L = \frac{\mu_2 + \mu_3}{2} - \mu_1$ (smokers of any kind vs. nonsmokers)

Residual analysis (model diagnostics)

Recall the residuals: $e_{it} = y_{it} - \hat{y}_{it}$

The purpose of residual analysis is to verify the model assumptions:

- No outliers
- Normally distributed errors
- Uncorrelated errors
- Constant error variance $\mathrm{Var}(\epsilon_{it}) = \sigma^2$
- Other factors

VMD experiment

- Files to execute this study are here: github.com/mquazi/Intro_OW_ANOVA
- An experiment to compare the performance of different hole sizes ($5\mu m, 10\mu m, 20\mu m, 30\mu m$) is investigated based on the volumetric median diameter (VMD) of the droplets for atomizers. 8 observations of VMD are recorded for each hole size.
- VMD is the response variable
- Hole size is the factor at 4 levels (or 4 trts)
- Main questions: (1) Does the hole size affect the VMD? (2) Which hole size levels stand out?

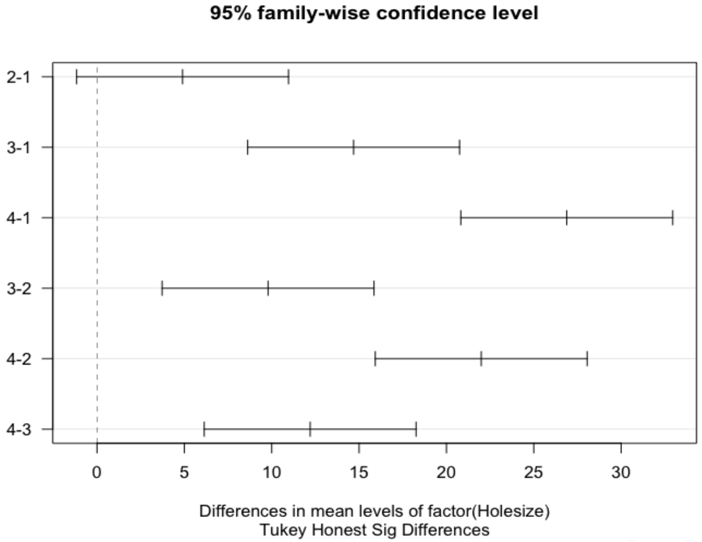
Conclusion (1/2)

Table: ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Holesize)	3	3,381.914	1,127.305	57.132	0
Residuals	28	552.480	19.731		

- p-value< 0.05, reject H_0. There is a statistically significant relationship between the VMD and the hole size.
- Now we can analyze effects of hole size levels

Conclusion(2/2)



Thank You!