

STAT 3119

Week4: 9/17/2019 @GWU

Outline Week 4

- A1. Review one factor ANOVA model and Analysis of factor levels means for Quiz #2
- A2. Sample size planning based on CI (chap 17.8)
- A3. Analysis of Factor Effects when Factor Is Quantitative (chap 17.8)
- A4. ANOVA Diagnostics (chapter 18.1)
- B1. Thurs: Quiz#2, review HW#1, HW#2, Quiz#1 (solutions will be posted online. No new materials this Thurs.)

Review: One way ANOVA

We studies in the last lesson about the one factor (one-way) ANOVA model (**Cell means model**) :

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- We illustrate how to perform an ANOVA analysis.

TABLE 16.3 ANOVA Table for Single-Factor Study.

Source of Variation	SS	df	MS
Between treatments	$SSTR = \sum n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$r - 1$	$MSTR = \frac{SSTR}{r - 1}$
Error (within treatments)	$SSE = \sum \sum (Y_{ij} - \bar{Y}_{i.})^2$	$n_T - r$	$MSE = \frac{SSE}{n_T - r}$
Total	$SSTO = \sum \sum (Y_{ij} - \bar{Y}_{..})^2$	$n_T - 1$	

The test statistic

$$\text{Test statistic } F^* = MSTR/MSE$$

When H_0 holds. F^* is distributed as $F(r - 1, n_T - r)$, where r = number of factor levels, n_T the totla sample size.

Review: Inference of factor level means

For each parameter of interest that we discussed, e.g. θ , we need to know what is the point estimator $\hat{\theta}$ and the estimator of the standard deviation $s(\hat{\theta})$, then we can make inference: $(1 - \alpha)$ two-sided CI is

$$\hat{\theta} \pm t(1 - \alpha/2, n_T - r)s(\hat{\theta})$$

The test statistic for the $H_0 : \theta = \theta_0$ would be

$$t^* = (\hat{\theta} - \theta_0)/s(\hat{\theta}) \sim t(n_T - r) \text{ distribution}$$

Summary table: find estimator and est SD to plug-in.

Parameter θ	Estimator $\hat{\theta}$	Estimated SD $s(\hat{\theta})$	(1- α) confidence interval $\hat{\theta} \pm t(1 - \alpha/2, n_T - r)s(\hat{\theta})$	Test statistic for $H_0 : \theta = \theta_0$ $(\hat{\theta} - \theta_0)/s(\hat{\theta}) \sim t(n_T - r)$
μ_i	$\hat{\mu}_i = \bar{Y}_{i.}$	$s^2\{\bar{Y}_{i.}\} = \frac{MSE}{n_i}$	$\bar{Y}_{i.} \pm t(1 - \alpha/2; n_T - r)s\{\bar{Y}_{i.}\}$	$H_0: \mu_i = c \quad t^* = \frac{\bar{Y}_{i.} - c}{s\{\bar{Y}_{i.}\}}$
$D = \mu_i - \mu_{i'}$	$\hat{D} = \bar{Y}_{i.} - \bar{Y}_{i'}$	$s^2\{\hat{D}\} = MSE \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)$	$\hat{D} \pm t(1 - \alpha/2; n_T - r)s\{\hat{D}\}$	$H_0: \mu_i - \mu_{i'} = 0 \quad t^* = \frac{\hat{D}}{s\{\hat{D}\}}$
$L = \sum_{i=1}^r c_i \mu_i$	$\hat{L} = \sum_{i=1}^r c_i \bar{Y}_{i.}$	$s^2\{\hat{L}\} = MSE \sum_{i=1}^r \frac{c_i^2}{n_i}$	$\hat{L} \pm t(1 - \alpha/2; n_T - r)s\{\hat{L}\}$	$H_0: \sum c_i \mu_i = c \quad t^* = \frac{\hat{L} - c}{s(\hat{L})}$

The linear combination L is the most general form, the single/pair difference/contrast of the factor level means μ_i are its special case.

Review: Simultaneous Inference Procedures for factor level means

- Tukey procedure applies for all the pairwise comparisons; Scheffe procedure applies for the set of all possible contrasts and Bonferroni procedure applies for any sets of comparisons.
- If all pairwise comparisons are of interest, the Tukey procedure is superior to the Bonferroni or scheffe procedure. All three procedures are of the form “estimator +/- multiplier×SE”.

Bonferroni multiple (for g comparisons) $t(1 - \alpha/2g; n_T - r)$

Tukey multiple $1/\sqrt{2} * q(1 - \alpha; r, n_T - r)$

Scheffé multiple $\sqrt{(r - 1)F(1 - \alpha; r - 1, n_T - r)}$

Note: In any given problem, one may compute the Bonferroni multiple as well as the Scheffé multiple and, when appropriate, the Tukey multiple, and select the one that is smallest. This choice is proper since it does not depend on the observed data.

Sample size planning With Estimation Approach (Ch 17.8)

- Sample size justification may be based on the width of the CI (precision of the estimation) or in conjunction with the control of type I/II errors, e.g. getting initial sample size based on the required type I error and power.
- If the anticipated widths of the confidence intervals based on the initial sample sizes are satisfactory, then it is done. *Iterations:* If one or more widths are too great, larger sample sizes need to be tried next. If the widths are narrower than they need be, smaller sample sizes should be tried next.
- When there are several major comparisons of interest or multiple CIs to consider, then a simultaneous inference procedure may be used to adjust for multiple comparisons to get the proper family confidence coefficient and determine the expected widths of the confidence intervals.

Example 1 (page 759)- Equal sample sizes

Still consider the Snow tires example in Chapter 16.10 : A company owning a large fleet of trucks wishes to determine whether or not four different brands of snow tires have the same mean tread life (in thousands of miles). Assume the sample size for each tire brand are to be equal, i.e. , $n_i = n$.

Management wishes three types of estimates:

1. A comparison of the mean tread lives for each pair of brands: $L_1 = \mu_i - \mu'_i$ (this is a group of $\binom{4}{2}=6$ comparisons)
2. A comparison of the mean tread lives for the two high-priced brands (1 and 4) and the two low-priced brands (2 and 3)

$$L_2 = \frac{\mu_1 + \mu_4}{2} - \frac{\mu_2 + \mu_3}{2}$$

3. A comparison of the mean tread lives for the national brands (1, 2, and 4) and the local brand (3):

$$L_3 = \frac{\mu_1 + \mu_2 + \mu_4}{3} - \mu_3$$

Management further has indicated that it wishes a family confidence coefficient of .95 for the entire set of comparisons.

Example 1 (2) Sample size calculation

- From past experience, they can assume $\sigma = 2$, also they consider $n = 10$ as starting sample size
- All these comparisons are contrasts or linear combinations of factor levels means:

We know from (17.21) that the variance of an estimated contrast \hat{L} when $n_i \equiv n$ is:

$$\sigma^2\{\hat{L}\} = \frac{\sigma^2}{n} \sum c_i^2 \quad \text{when } n_i \equiv n$$

Hence, given $\sigma = 2$ and $n = 10$, the anticipated values of the standard deviations of the required estimators are:

Contrast	Anticipated Variance	Anticipated Standard Deviation
Pairwise comparisons	$\frac{(2)^2}{10} [(1)^2 + (-1)^2] = .80$.89
High- and low-priced brands	$\frac{(2)^2}{10} \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 \right] = .40$.63
National and local brands	$\frac{(2)^2}{10} \left[\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + (-1)^2 \right] = .53$.73

- Here, there are 8 comparisons involving all contrasts (but not all are pairwise differences), we can either use Bonferroni or Scheffe procedure to adjust for multiple comparisons.

```
#Bonferroni multiple, df=n_T- r = 40 -4 =36
g = 8 ; df= 36
qt(1- 0.05/(2*g), df)
```

```
## [1] 2.904552
```

```
# Scheffe multiple r-1=3, n_T-r=36
sqrt(3*qt(.95, 3, 36))
```

```
## [1] 2.93237
```

The textbook showed using the Scheffe procedure, the anticipated widths of these CIs are:

Contrast	Anticipated Width of Confidence Interval = $\pm S\sigma\{\hat{L}\}$
Pairwise comparisons	$\pm 2.93(.89) = \pm 2.61$ (thousand miles)
High- and low-priced brands	$\pm 2.93(.63) = \pm 1.85$ (thousand miles)
National and local brands	$\pm 2.93(.73) = \pm 2.14$ (thousand miles)

Management was satisfied with these anticipated widths. However, it was decided to increase the sample sizes from 10 to 15 in case the actual standard deviation of the tread lives of tires is somewhat greater than the anticipated value $\sigma = 2$ (thousand miles).

Also note: Scheffe procedure controls for all the contrasts. Bonferroni multiple may be smaller if the number of test is not too many. In this case, if there are more than 10 comparisons, then Bonferroni multiple, increasing with the number of test g , will be greater than Scheffe multiple.

```
#Bonferroni multiple, df=n_T- r = 40 -4 =36
g = 10 ;df= 36
qt(1- 0.05/(2*g), 36)
```

```
## [1] 2.990487
```

Example 2 (page 761) —Unequal Sample Sizes

- There are three comparisons of interest to compare brand 1-3 vs. brand 4.

$$L_1 = \mu_1 - \mu_4, \quad L_2 = \mu_2 - \mu_4, \quad L_3 = \mu_3 - \mu_4$$

- Set a family confidence coefficient of .90 and the desired precision to be ± 1 .
- The sample size for brand 4 is to be twice as large as for the other brands in order to improve the precision of the three pairwise comparisons. We can get the variance of these pairwise comparison $\sigma(\hat{L})$ as follows

We know from (17.13) that the variance of an estimated difference $\hat{L}_i = \bar{Y}_{i.} - \bar{Y}_{4.}$ (the difference is now denoted more generally by \hat{L}) is for $i = 1, 2, 3$:

$$\sigma^2\{\hat{L}_i\} = \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_4} \right)$$

We shall denote the sample sizes for brands 1, 2, and 3 by n and for brand 4 by $2n$. Hence, the variance of \hat{L}_i becomes:

$$\sigma^2\{\hat{L}_i\} = \sigma^2 \left(\frac{1}{n} + \frac{1}{2n} \right) = \frac{3\sigma^2}{2n}$$

- Then the estimated CI width will be "constant* $\sigma(\hat{L})$ ", then compare CI width with the pre-specified limits to proceed with the modification of the sample size in next step.

- In this case, all three procedures can be used.

```
#Bonferroni multiple, df=n_T- r = 30 +20 -4 = 46
g = 3 ; df= 46
qt(1- 0.1/(2*g), df)
```

```
## [1] 2.193893
```

```
# Scheffe multiple r-1=3, n_T-r=46
sqrt(3*qt(.90, 3, 46))
```

```
## [1] 2.573066
```

```
# Tukey multiple
1/sqrt(2)* qtukey(.90, nm=4, df=46)
```

```
## [1] 2.357556
```

- Therefore, in this example, Bonferroni CI has narrower CI and better precision than using the Tukey or Scheffe procedure.
- **Note:** In general, in sample size planning, Since one cannot be certain that the planning value for the standard deviation is correct, it is advisable to study a range of values for the standard deviation before making a final decision on sample size.

Analysis of Factor Effects when Factor Is Quantitative (Ch 17.9)

- When the factor under investigation is quantitative or ordinal, the analysis of factor effects can be carried beyond the point of multiple comparisons to include a study of the nature of the response function.
- For one factor study, we can also graph the data, and use the regression analysis that you learned in previous course to estimate the response as a function of the factor levels.
- We illustrate the regression analysis using the example in Chapter 17.9.

Example

In a study to reduce raw material costs in a glassworks firm, an operations analyst collected the experimental data in Table 17.4 on the **number of acceptable units produced** from equal amounts of raw material by **28 entry-level piecework employees** who had received special training as part of the experiment. **Four training levels were used (6, 8, 10, and 12 hours)**, with **seven** of the employees being assigned at *random* to each level. The higher the number of acceptable pieces, the more efficient is the employee in utilizing the raw material. This study is a *single-factor completely randomized design with four factor levels*.

TABLE 17.4
Data—
Piecework
Trainees
Example.

Treatment (hours of training)		Employee (<i>j</i>)						
	<i>i</i>	1	2	3	4	5	6	7
1	6 hours	40	39	39	36	42	43	41
2	8 hours	53	48	49	50	51	50	48
3	10 hours	53	58	56	59	53	59	58
4	12 hours	63	62	59	61	62	62	61

ANOVA Analysis in R

- We already how to analyze this data by ANOVA analysis and estimate/plot the main effects of the training on the outcome.

1. read and relabel data

```
# read data from week4 folder online
Ex17 = read.table(
  url("https://raw.githubusercontent.com/npmladabook/Stat3119/master/Week4/CH17_TA04.txt"))
names(Ex17) = c("response", "training", "units")

# make another categorical variable
Ex17$training2 = as.factor(Ex17$training)

# relabel level
levels(Ex17$training2) <- c("6h", "8h", "10h", "12h")
str(Ex17)
```

```
## 'data.frame': 28 obs. of 4 variables:
## $ response : int 40 39 39 36 42 43 41 53 48 49 ...
## $ training : int 1 1 1 1 1 1 1 2 2 2 ...
## $ units : int 1 2 3 4 5 6 7 1 2 3 ...
## $ training2: Factor w/ 4 levels "6h","8h","10h",...: 1 1 1 1 1 1 1 2 2 2 ...
```

```
head(Ex17, 3)
```

```
## response training units training2
## 1      40         1      1         6h
## 2      39         1      2         6h
## 3      39         1      3         6h
```

2. ANOVA analysis (similar to Page 763)

```
fit = aov(response ~ training2, data = Ex17)
summary(fit)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## training2   3 1808.7   602.9   141.5 2.17e-15 ***
## Residuals  24  102.3     4.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

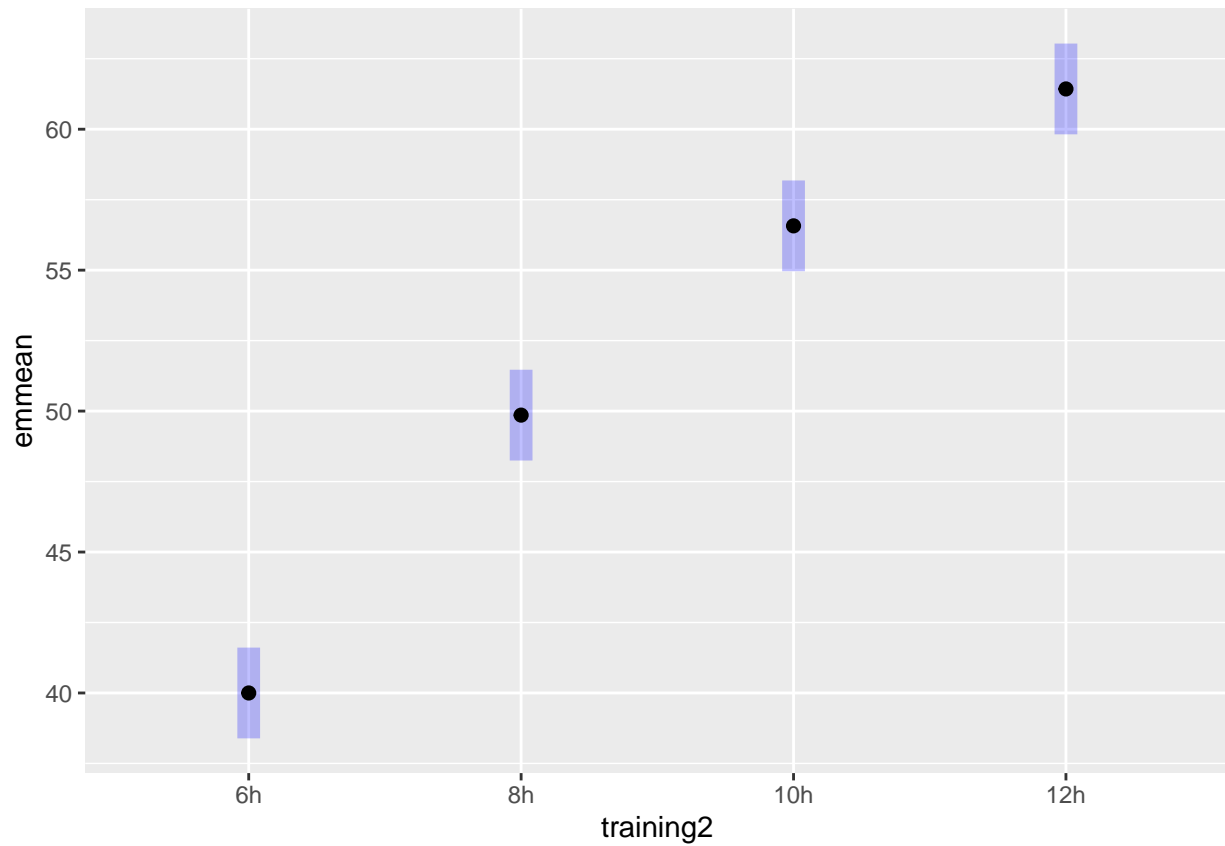
3. Main effects analysis

```
library(emmeans)

# get the estimate, SE, df and CI
(Est.mean <- emmeans(fit, ~ training2))
```

```
## training2 emmean SE df lower.CL upper.CL
## 6h 40.0 0.78 24 38.4 41.6
## 8h 49.9 0.78 24 48.2 51.5
## 10h 56.6 0.78 24 55.0 58.2
## 12h 61.4 0.78 24 59.8 63.0
##
## Confidence level used: 0.95
```

```
# plot the main effects with 95% CIs
plot(Est.mean, horizontal=F)
```



4. Tukey's HSD procedure

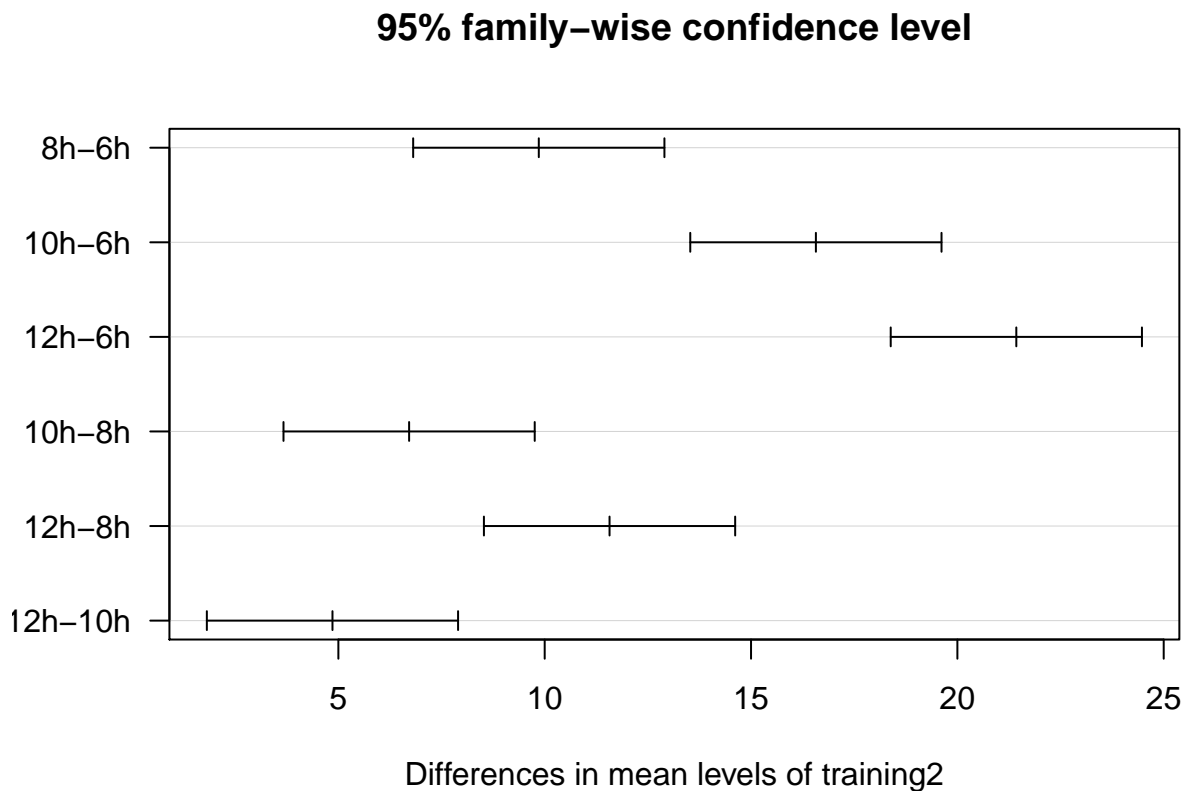
```
## Tukey HSD with built-in function, including plots:
TukeyHSD(fit)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = response ~ training2, data = Ex17)
##
## $training2
## diff lwr upr p adj
## 8h-6h 9.857143 6.813050 12.901236 0.0000000
```



```
## 10h-6h 16.571429 13.527335 19.615522 0.0000000
## 12h-6h 21.428571 18.384478 24.472665 0.0000000
## 10h-8h 6.714286 3.670192 9.758379 0.0000157
## 12h-8h 11.571429 8.527335 14.615522 0.0000000
## 12h-10h 4.857143 1.813050 7.901236 0.0010237
```

```
plot(TukeyHSD(fit), las=1)
```



Results: All pairwise comparisons showed significant differences. 1) Adjusted P-value all < 0.01 ; 2) For Tukey's HSD plot, nonpaired difference and its CI intercept with the vertical line at 0.

Regression analysis in R

The main effect plot shows an increasing trend of response with the longer training hours. We also note 1) The factor level: training hour is quantitative in nature 2) the trend is not linear: From 6hr to 12 hrs, the increment is slower with every 2 hr of training. In practice, we can fit both linear and quadratic model if the trend is not clear from visual inspection.

```
# Linear regression fit
Ex17$hr[Ex17$training==1] <- 6
Ex17$hr[Ex17$training==2] <- 8
Ex17$hr[Ex17$training==3] <- 10
Ex17$hr[Ex17$training==4] <- 12
```

```

# We center the covariate X by mean(X)=9
# so intercept corresponding to mean response at X=9
Ex17$X <- Ex17$hr -9
Ex17$X.Squared <-(Ex17$X )^2

Lmfit <- lm(response ~ X +X.Squared , data = Ex17 )
summary(Lmfit)

```

```

##
## Call:
## lm(formula = response ~ X + X.Squared, data = Ex17)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.0643 -1.0643  0.3357  1.2607  3.3357
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.52679     0.61364   87.228  <2e-16 ***
## X             3.55000     0.17143   20.708  <2e-16 ***
## X.Squared    -0.31250     0.09583   -3.261   0.0032 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.028 on 25 degrees of freedom
## Multiple R-squared:  0.9462, Adjusted R-squared:  0.9419
## F-statistic: 219.7 on 2 and 25 DF,  p-value: < 2.2e-16

```

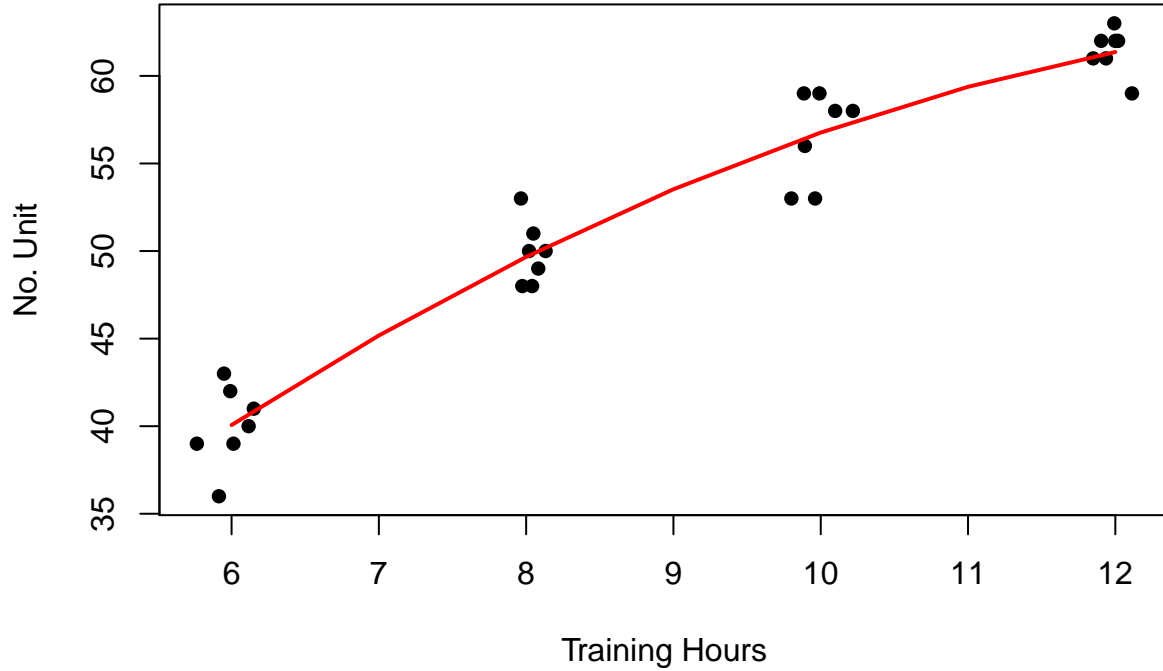
results: We can obtain the regression coefficient for the regression equation:

$$\hat{Y} = 53.52679 + 3.55000x - 0.31250x^2$$

```

plot ( Ex17$hr+rnorm(28, 0, 0.1), Ex17$response, pch=16,
       xlab= "Training Hours", ylab="No. Unit")
newX= 6:12-9
lines(6:12, predict.lm(Lmfit, newdata=data.frame(X=newX, X.Squared= newX^2 )),col=2,lwd=2 )

```



ANOVA Diagnostics: Introduction (Chapter 18)

We have gone through the data structure, ANOVA model and analysis for one-factor study. Next, we will discuss the use of residual plots for diagnosing the appropriateness of ANOVA models, as well as tests for model assumptions and some remedial measure to improve the model fit.

The actual sequence of developing and using any statistical model is, **before apply the model**,

1. Examine whether the proposed model is appropriate for the set of data at hand and whether there are serious departures from the conditions assumed by the model.
2. If the proposed model is not appropriate, consider remedial measures, such as transformation of the data or modification of the model.
3. After review of the appropriateness of the model and completion of any necessary remedial measures, inferences based on the model can be undertaken

Residual Analysis for ANOVA (Ch 18.1)

-The residuals e_{ij} for the ANOVA cell means model are defined by

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_i.$$

- Like in regression, we can define the semistudentized residuals as

$$e_{ij}^* = e_{ij} / \sqrt{MSE}$$

- The studentized residuals as

$$r_{ij} = e_{ij} / s\{e_{ij}\}$$

where

$$s\{e_{ij}\} = \sqrt{MSE(n_i - 1)/n_i}$$

.

Note that the semistudentized residuals e_{ij}^* and the studentized residuals r_{ij} provide essentially the same information, differing only by a constant factor.

Residual Plots

As you have learned before, residual plots useful for analysis of variance models include: (1) plots against the fitted values and (2) normal probability plots.

We consider now how residual plots can be helpful in diagnosing the following departures from ANOVA model (16.2):

1. Nonconstancy of error variance (across all factor levels) : we can use plot of residuals vs. fitted values. When the sample sizes differ greatly, studentized residuals should be used in these plots.
2. Nonnormality of error terms
3. Outliers: outlying values in the graphs
4. Omission of important explanatory variables (in the residual vs. fitted plot, we can stratify the data according to other categorical variables to check other pattern or trend)

Rust inhibitor example

This example is used to illustrate the use of residual plots for evaluating the appropriateness of analysis of variance models.

In a study of the effectiveness of different rust inhibitors, four brands (A, B, C, D) were tested. Altogether, 40 experimental units were randomly assigned to the four brands, with 10 units assigned to each brand. The higher the coded value, the more effective is the rust inhibitor. This study is a completely randomized design, where the levels of the single factor correspond to the four rust inhibitor brands.

TABLE 17.2
Data and
Analysis of
Variance
Results—Rust
Inhibitor
Example (data
are coded).

(a) Data Rust Inhibitor Brand				
	A	B	C	D
j	$i = 1$	$i = 2$	$i = 3$	$i = 4$
1	43.9	89.8	68.4	36.2
2	39.0	87.1	69.3	45.2
3	46.7	92.7	68.5	40.7
...
8	38.9	88.1	65.2	38.7
9	43.6	90.8	63.8	40.9
10	40.0	89.1	69.2	39.7
$\bar{Y}_{i.}$	43.14	89.44	67.95	40.47
$\bar{Y}_{..} = 60.25$				
(b) Analysis of Variance				
Source of Variation	SS	df	MS	
Between brands	15,953.47	3	5,317.82	
Error	221.03	36	6.140	
Total	16,174.50	39		

We can use R to run the ANOVA model to fit this model fit.

```
# read data from week3 folder online
Ex17 =read.table(
  url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week3/CH17_TA02.txt"))
names(Ex17) = c("response", "brand", "units")
Ex17$brand = as.factor(Ex17$brand)

# relabel level from 1:4 to A to D
levels(Ex17$brand)<- LETTERS[1:4]

fit<- aov(response~ brand, data=Ex17)
summary(fit)

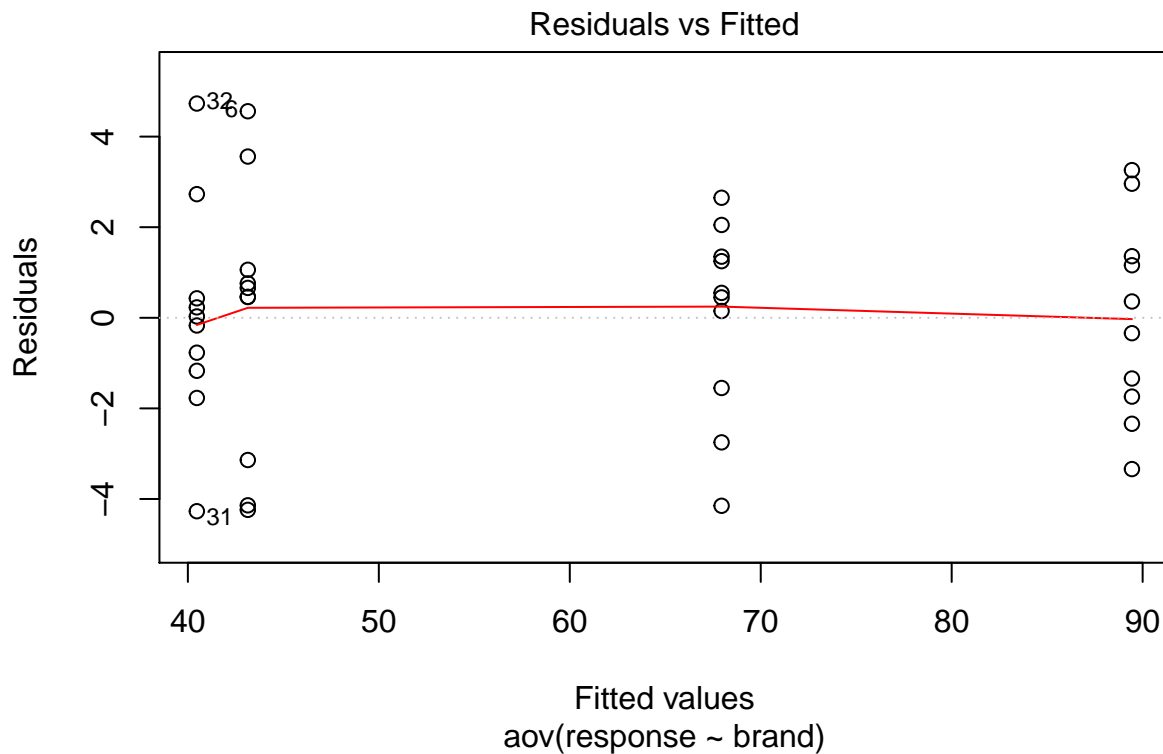
##           Df Sum Sq Mean Sq F value Pr(>F)
## brand      3  15953    5318   866.1 <2e-16 ***
## Residuals 36    221      6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both plots are easier to generate from R directly from ANOVA model fit.

1. Check the homogeneity of variance assumption

The residuals versus fitted plot can be used to check the **homogeneity of variances**.

```
plot(fit,1)
```



```
# if no smoother red line  
#plot(fit,1, add.smooth = FALSE)
```

Note:

This plot suggests the constancy of the error variance is likely to be true. We can see the same extent of scatter or spread of the residuals around zero for each factor level. There is no evident relationships between residuals and fitted values (the mean of each groups).

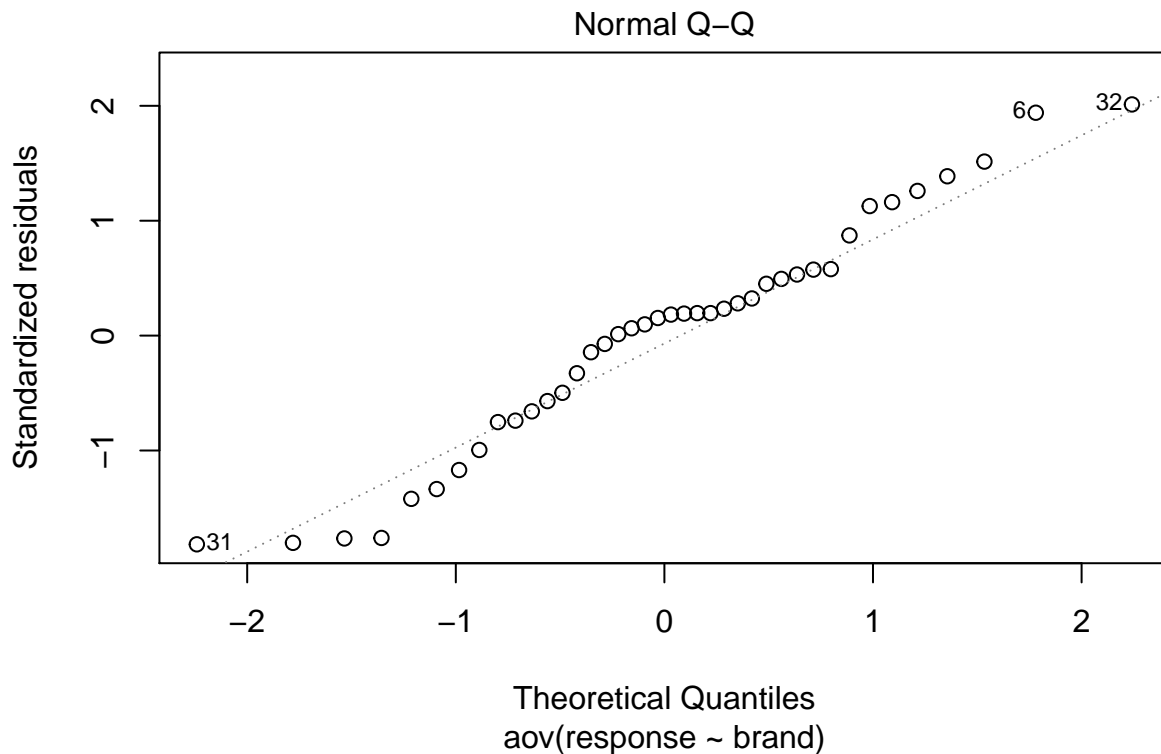
The red line is very close to the horizontal line at 0 suggesting the residuals likely to have a mean of 0. It is a smoother (a local average) of the residuals. We could get rid of it by using the function call `plot(fit, 1, add.smooth = FALSE)`.

This plot gives a rough visual inspection for series deviation and we will use formal statistical tests for the constancy of error variance.

2. Check the normality assumption

To check if the residuals follow a normal distribution, we can get the normal probability plot (normal Q-Q plot). In the plot below, the quantiles of the residuals are plotted against the quantiles of the normal distribution. A 45-degree reference line is also plotted. The normal probability plot of residuals is used to check the assumption that the residuals are normally distributed. It should approximately follow a straight line.

```
plot(fit,2)
```



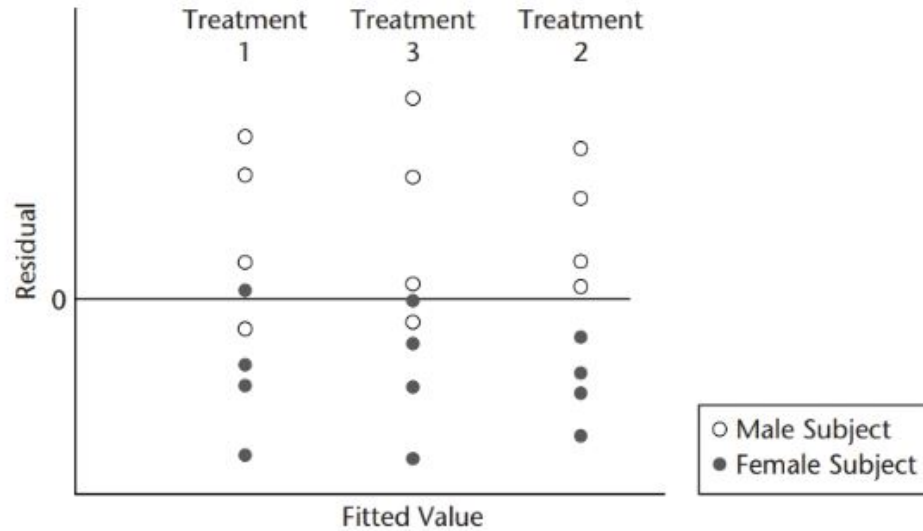
Note:

This figure does not indicate any serious departures from normality. The pattern of the points is reasonably linear except possibly in the tails.

Residual plot to detect Omission of important explanatory variables

- This does not directly relate to the ANOVA model assumptions, but more on the response prediction.
- We can use the residual vs. fitted plot to detect whether there is omission of important explanatory variables, by stratifying the data according to other categorical variables to check for other patterns or trends.

FIGURE 18.5
Residual Plot
against Fitted
Values
Illustrating
Omission of
Important
Explanatory
Variable.



The results in Figure 18.5 suggest strongly that for each of the motivational treatments studied, the treatment effects do differ according to gender. So the research may need to include gender in the model to explain the response better.

Summary this week

- Reminder: Hw#2 Last week's homework assignment: **17.8, 17.11** (due 9/19 by 6 pm before class; submit online from blackboard).
- Reading: Chapter 17.8, 17.9, 18.1
- Quiz #2 : Thursday
- Hw#3: This week's homework assignment will be posted on Blackboard (due 9/26 by 6 pm before class; submit online from blackboard).