

Unit 6 – Analysis of Variance
Practice Problems(2 of 2)
Solutions

Before you begin. Download from the course website
lbw.xlsx

(Source: Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied Logistic Regression: Third Edition*. These data are copyrighted by John Wiley & Sons Inc. and must be acknowledged and used accordingly. Data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986.)

Low birth weight is an outcome of concern because of its links to infant mortality and birth defects. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight. The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of whom had low birth weight babies and 130 of which had normal birth weight babies.

In this homework, we will use three variables to gain practice in performing a two-way analysis of variance: lbw.xlsx has 189 observations on 3 variables.

Data dictionary/Codebook

Position	Variable	Label	Type	Codings
1	id	Identification code		Range: 4, 226
2	race	Race	numeric	1 = white 2 = african american 3 = other
3	ftv	Number of visits to physician during 1 st trimester	numeric	Range: 0, 6
4	btw	Birthweight (grams)	numeric	Range: 709, 4990

Outcome Variable

Y = btw

Factor I

racef, coded: 1, 2 or 3

Note: you will create this from race in exercise #2

Factor II

no_trimester1, coded: 0, 1

Note: you will create this from ftv in exercise #2

Preliminaries

```
import excel data: lbw_anova
library(readxl)
lbw_anova <- read_excel("lbw.xlsx")
lbw_anova <- as.data.frame(lbw_anova)
str(lbw_anova)

## 'data.frame': 189 obs. of 4 variables:
## $ id : num 85 86 87 88 89 91 92 93 94 95 ...
## $ race: num 2 3 1 1 1 3 1 3 1 1 ...
## $ bwt : num 2523 2551 2557 2594 2600 ...
## $ ftv : num 0 3 1 2 0 0 1 1 1 0 ...
```

#1.

State the analysis of variance model using notation μ , α_i , β_j , $(\alpha\beta)_{ij}$ and σ^2 as appropriate. Define all terms and constraints on the parameters.

Answer:

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \text{ where } \epsilon_{ijk} \sim \text{Normal}(0, \sigma_{\text{error}})$$

with $i = 1, 2, 3$ indexing race group

$j = 1, 2$ indexing visits to MD in 1st trimester (zero or at least one)

μ = population mean birthweight (g), over all groups

α_i = deviation from mean effect of race = i , with $\sum_{i=1}^3 \alpha_i = 0$

β_j = deviation from mean effect of visits to MD = j with $\sum_{j=1}^2 \beta_j = 0$

$(\alpha\beta)_{ij}$ = extra deviation from mean (beyond main effects) with

$$\sum_{i=1}^3 (\alpha\beta)_{ij} = 0 \quad \text{and} \quad \sum_{i=j}^2 (\alpha\beta)_{ij} = 0$$

#2.

By any means you like, create the following three new variables

(1) **racef** = factor version of race

(2) **no_trimester1** that is a 0/1 indicator of “no visits in the first trimester and defined as follows:

$$\begin{aligned} \text{no_trimester1} &= 1 \text{ if } \text{ftv}=0 \\ &0 \text{ for all other values of } \text{ftv} \end{aligned}$$

(3) **no_trimester1f** = factor version of **no_trimester1**

```

library(tidyverse)

ready <- lbw %>%
  mutate(racef = recode_factor(race,
    "1" = "White",
    "2" = "African American",
    "3" = "Other")) %>% # create factor var racef

  mutate(no_trimester1 = ifelse(ftv==0,1,0)) %>% # numeric version (0/1)
  mutate(no_trimester1f= recode_factor(no_trimester1,
    "0" = "Visits",
    "1" = "No visits")) # factor version

glimpse(ready)

## Rows: 189
## Columns: 7
## $ race      <dbl> 2, 3, 1, 1, 1, 3, 1, 1, 3, 3, 3, 3, 1, 1, 2, 1, 3...
## $ bwt       <dbl> 2523, 2551, 2557, 2594, 2600, 2622, 2637, 2637, 2663, 2...
## $ ftv       <dbl> 0, 3, 1, 2, 0, 0, 1, 1, 1, 0, 0, 1, 0, 2, 0, 0, 0, 3, 0...
## $ id        <dbl> 85, 86, 87, 88, 89, 91, 92, 93, 94, 95, 96, 97, 98, 99, ...
## $ racef     <fct> African American, Other, White, White, White, Other, Wh...
## $ no_trimester1 <dbl> 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1...
## $ no_trimester1f <fct> No visits, Visits, Visits, Visits, Visits, No visits...

```

#3.

By any means you like, produce descriptive statistics of $Y=bwt$, separately for groups defined by `racef` and `no_trimester1f`.

```

library(tidyverse)

ready %>%                                     # Summary by group using dplyr
  group_by(racef,no_trimester1f) %>%
  summarize(n=sum(!is.na(bwt)),                # sum(!is.na()) to get number of complete obs
            mean=mean(bwt),
            sd=sd(bwt),
            min = min(bwt),
            P25 = quantile(bwt, 0.25),
            median = median(bwt),
            P75 = quantile(bwt, 0.75),
            max = max(bwt))

## # A tibble: 6 × 10
## # Groups:   racef [3]
##   racef      no_trimester1f     n   mean    sd   min    P25 median    P75   max
##   <fct>      <fct>     <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 White       Visits       53  3176.  756.  1021  2663  3080  3770  4990
## 2 White       No visits    43  3013.  690.  1818  2417  3076  3622. 4238
## 3 African Amer... Visits     12  2719.  620.  1701  2307. 2920  3047  3860
## 4 African Amer... No visits   14  2720.  677.  1135  2381  2650. 3275. 3790
## 5 Other        Visits      24  2877.  651.  1588  2448  2792. 3308. 3997
## 6 Other        No visits    43  2763.  762.  709  2261  2863  3228. 4054

```

#4.

Fit the two-way analysis of variance. Show the analysis of variance table.

```
fit_anova <- aov(bwt ~ racef + no_trimester1f + racef*no_trimester1f, data=ready)
anova(fit_anova)

## Analysis of Variance Table
##
## Response: bwt
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## racef            2 5048361 2524181  4.9123 0.008354 **
## no_trimester1f   1  692771  692771  1.3482 0.247104
## racef:no_trimester1f 2 140324  70162  0.1365 0.872458
## Residuals      183 94033843 513846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

KEY: With respect to the variability in birthweight (Y=bwt),

- (1) There is NO statistically significant evidence of an interaction race x visits to MD (p-value = .87)
- (2) There is NO statistically significant evidence of variations by whether or not the mother visited her Ob/GYN in the first trimester (p-value = .25)
- (3) However, we do see statistically significant evidence of race group differences in mean birthweight (p-value = .008)

#5.

This time, perform the two way analysis of variance as a regression.

Preliminaries: Create 0/1 indicator vars and interaction vars.

```
library(tidyverse)

ready <- ready %>%
  mutate(African_American = ifelse(race==2,1,0)) %>%          # 0/1 indicator African American
  mutate(Race_Other = ifelse(race==3,1,0)) %>%                  # 0/1 indicator African American

  mutate(AfrAmer_novisits1 = African_American*no_trimester1) %>%  # interaction
  mutate(Other_novisits1 = Race_Other*no_trimester1)                 # interaction

str(ready)

## 'data.frame': 189 obs. of 11 variables:
## $ race      : num  2 3 1 1 1 3 1 3 1 ...
## $ bwt       : num  2523 2551 2557 2594 2600 ...
## $ ftv       : num  0 3 1 2 0 0 1 1 1 0 ...
## $ id        : num  85 86 87 88 89 91 92 93 94 95 ...
## $ racef     : Factor w/ 3 levels "White","African American",...: 2 3 1 1 1 3 1 3 1 1 ...
## $ no_trimester1 : num  1 0 0 0 1 1 0 0 0 1 ...
## $ no_trimester1f: Factor w/ 2 levels "Visits","No visits": 2 1 1 1 2 2 1 1 1 2 ...
## $ African_American: num  1 0 0 0 0 0 0 0 0 0 ...
## $ Race_Other : num  0 1 0 0 0 1 0 1 0 0 ...
## $ AfrAmer_novisits1: num  1 0 0 0 0 0 0 0 0 0 ...
## $ Other_novisits1 : num  0 0 0 0 0 1 0 0 0 0 ...
```

#5.

This time, perform the two way analysis of variance as a regression. Show

```

fit_lm <- lm(bwt ~ African_American + Race_Other + no_trimester1 + AfrAmer_novisits1 + Other_novisits1,
               data=ready)
anova(fit_lm)

## Analysis of Variance Table
##
## Response: bwt
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## African_American   1 1520693 1520693  2.9594 0.087068 .
## Race_Other          1 3527668 3527668  6.8652 0.009527 **
## no_trimester1       1 692771  692771  1.3482 0.247104
## AfrAmer_novisits1   1 116469  116469  0.2267 0.634579
## Other_novisits1     1 23855   23855  0.0464 0.829646
## Residuals         183 94033843 513846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(fit_lm)

##
## Call:
## lm(formula = bwt ~ African_American + Race_Other + no_trimester1 +
##      AfrAmer_novisits1 + Other_novisits1, data = ready)
##
## Residuals:
##    Min      1Q      Median      3Q      Max 
## -2155.32 -513.32   -13.49   551.68  1813.68 
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)    
## (Intercept) 3176.32     98.46  32.259 <0.000000000000002 *** 
## African_American -457.24    229.16  -1.995     0.0475 *    
## Race_Other   -299.70    176.37  -1.699     0.0910 .    
## no_trimester1 -163.67    147.12  -1.112     0.2674    
## AfrAmer_novisits1 164.80    318.07   0.518     0.6050    
## Other_novisits1  50.53    234.53   0.215     0.8296    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 716.8 on 183 degrees of freedom
## Multiple R-squared:  0.05886,    Adjusted R-squared:  0.03315 
## F-statistic: 2.289 on 5 and 183 DF,  p-value: 0.04769

```

With respect to the variability in birthweight (Y=bwt), we learn a bit more with the regression approach:

- (1) There is NO statistically significant evidence of an interaction race x visits to MD (p-values = .61 and .83)
- (2) Compared to mothers of White race, mothers with race=OTHER have a mean birthweight that is estimated to be 299.70 grams lower (beta = -299.70) and the is marginally statistically significant (p-value = .09).
- (3) Compared to mothers of White race, African American mothers have a mean birthweight that is estimated to be 457.24 grams lower (beta = -457.24) and the is statistically significant (p-value = .0475).

#6.

By any means you like, perform a partial F-test of the null hypothesis that, controlling for `racef` and `no_trimester1f`, the extra predictive significance of the interaction of `racef` and `no_trimester1f` is zero.

```
Q6. partial F test of interactions - aov
reduced1 <- aov(bwt ~ racef + no_trimester1f, data=ready)
full1 <- aov(bwt ~ racef + no_trimester1f + racef*no_trimester1f, data=ready)
anova(reduced1, full1)

## Analysis of Variance Table
##
## Model 1: bwt ~ racef + no_trimester1f
## Model 2: bwt ~ racef + no_trimester1f + racef * no_trimester1f
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     185 94174167
## 2     183 94033843  2   140324 0.1365 0.8725
```

Controlling for the main effects of race and any visits to MD in the first trimester, there is no additional predictive significance of their interaction (Partial F test p-value = .87).

```
Q6. partial F test of interactions - lm
reduced2 <- lm(bwt ~ African_American + Race_Other + no_trimester1, data=ready)
full2 <- lm(bwt ~ African_American + Race_Other + no_trimester1 + AfrAmer_noVisits1 + Other_noVisits1,
            data=ready)
anova(reduced2, full2)

## Analysis of Variance Table
##
## Model 1: bwt ~ African_American + Race_Other + no_trimester1
## Model 2: bwt ~ African_American + Race_Other + no_trimester1 + AfrAmer_noVisits1 +
##           Other_noVisits1
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     185 94174167
## 2     183 94033843  2   140324 0.1365 0.8725
```

Controlling for the main effects of race and any visits to MD in the first trimester, there is no additional predictive significance of their interaction (Partial F test p-value = .87).

#7.

Obtain the predicted means of `bwt` for each group defined by `racef` and `no_trimester1f`, in two ways: (1) from the analysis of variance; and (2) from the regression. Verify that they are identical.

Q7. predicted means - aov()

```
library(emmeans)
# means over no_trimester1f, separately by racef
predicted_means7a = emmeans::emmeans(fit_anova, specs = "no_trimester1f", by="racef")
predicted_means7a

## racef = White:
## no_trimester1f emmean      SE df lower.CL upper.CL
## Visits          3176  98.5 183     2982    3371
## No visits       3013 109.3 183     2797    3228
##
## racef = African American:
## no_trimester1f emmean      SE df lower.CL upper.CL
## Visits          2719 206.9 183     2311    3127
## No visits       2720 191.6 183     2342    3098
##
## racef = Other:
## no_trimester1f emmean      SE df lower.CL upper.CL
## Visits          2877 146.3 183     2588    3165
## No visits       2763 109.3 183     2548    2979
##
## Confidence level used: 0.95
```

Q7. predicted means - lm()

```

library(tidyverse)

# Simplest is to now fit regression model with predictors as factor vars
fit_lm2 <- lm(bwt ~ factor(racef) + factor(no_trimester1f) + factor(racef)*factor(no_trimester1f),
               data=ready)

# Data frame of 6 groups: Factor I (racef) at 3 Levels x Factor II (no_trimester1f) at 2 Levels
mygroups <- data.frame(
  racef = rep(factor(c("White", "African American","Other")),2),
  no_trimester1f = rep(factor(c("Visits", "No visits")),3)
)

# Data frame of predicted means for the 6 groups
myhats <- as.data.frame(predict(fit_lm2, newdata = mygroups, interval = "confidence"))

# Combine group identification with predicted means
predicted_means7b <- cbind(mygroups,myhats)

# sort, rename variables for legibility and print
predicted_means7b <- predicted_means7b %>%
  arrange(racef,no_trimester1f) %>%
  rename(mean=fit,
         'Lower 95% CI' = lwr,
         'Upper 95% CI' = upr)
predicted_means7b

##           racef no_trimester1f      mean Lower 95% CI Upper 95% CI
## 2 African American     No visits 2720.214    2342.223    3098.206
## 5 African American     Visits   2719.083    2310.806    3127.361
## 6          Other       No visits 2763.488    2547.807    2979.169
## 3          Other       Visits   2876.625    2587.929    3165.321
## 4          White      No visits 3012.651    2796.970    3228.332
## 1          White      Visits   3176.321    2982.050    3370.592

```

Supplement

Plot of predicted means w 95% CI: From aov()

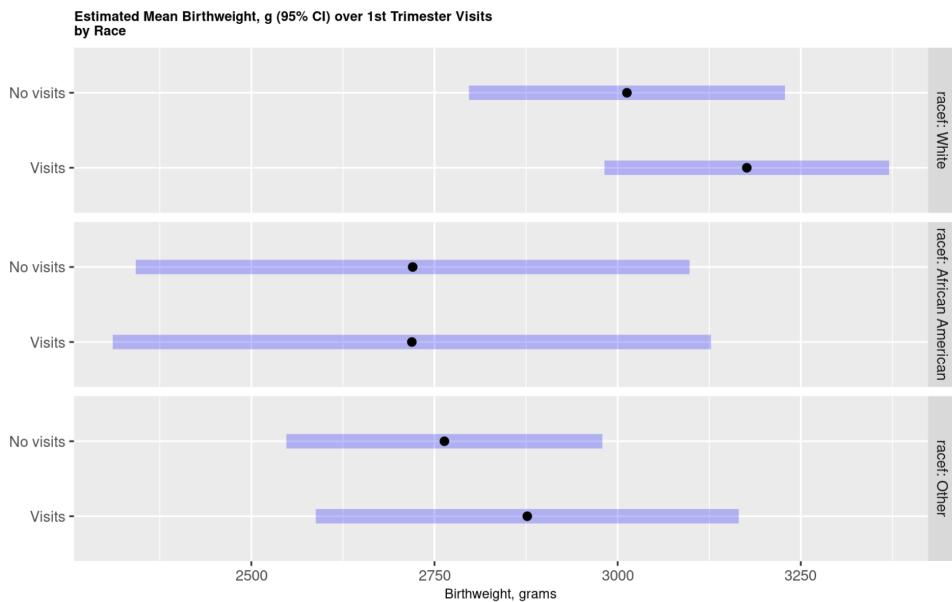
```

library(emmeans)
library(ggplot2)

# Get means over racef, separately by trimester1f
yhat_anova = emmeans::emmeans(fit_anova, specs = "no_trimester1f", by="racef")

# (extra). Plot of means over no_trimester1f, stratified by racef
plot(yhat_anova) +
  labs(x="Birthweight, grams") +
  labs(y="") +
  ggtitle("Estimated Mean Birthweight, g (95% CI) over 1st Trimester Visits\nby Race") +
  theme(axis.title = element_text(size = 8),
        plot.title = element_text(size = 8, face = "bold"))

```

**Supplement**

Plot of predicted means w 95% CI: From lm()

```

library(tidyverse)
library(ggplot2)
library(grid)
library(gridExtra)

# Simplest is to now fit regression model with predictors as factor vars
fit_lm2 <- lm(bwt ~ factor(racef) + factor(no_trimester1f) + factor(racef)*factor(no_trimester1f),
               data=ready)

# Data frame of 6 groups: Factor I (racef) at 3 Levels x Factor II (no_trimester1f) at 2 Levels
mygroups <- data.frame(
  racef = rep(factor(c("White", "African American", "Other")), 2),
  no_trimester1f = rep(factor(c("Visits", "No visits")), 3)
)

# Data frame of predicted means for the 6 groups
myhats <- as.data.frame(predict(fit_lm2, newdata = mygroups, interval = "confidence"))

# Combine group identification with predicted means
mydata <- cbind(mygroups, myhats)

# (extra). Plot of means over no_trimester1f, separately by racef
panel1 <- mydata %>% filter(racef=="White")
p1 <- ggplot(data=panel1) +
  aes(x=no_trimester1f, y=fit) +
  geom_errorbar(aes(ymin=lwr,
                     ymax=upr),
                width=.05,
                color="blue") +
  geom_point(color="blue") +
  scale_y_continuous(limits=c(2000,3500), breaks=seq(2000,3500,500)) +      # Be sure to use SAME y-axis scale for all
  ggtitle("White") +
  xlab(" ") +
  ylab("Mean Birthweight, grams (95% CI)") +                                # Y-axis Label for Left most panel only
  theme(axis.text.x = element_text(size = 8, angle=45, vjust=1, hjust=1),
        axis.title = element_text(size = 9),
        plot.title = element_text(size = 9))

```

```

panel2 <- mydata %>% filter(racef=="African American")
p2 <- ggplot(data=panel2) +
  aes(x=no_trimester1f, y=fit) +
  geom_errorbar(aes(ymin=lwr,
                     ymax=upr),
                 width=.05,
                 color="blue") +
  geom_point(color="blue") +
  scale_y_continuous(limits=c(2000,3500), breaks=seq(2000,3500,500)) +      # Be sure to use SAME y-axis scale for all
  ggtitle("African American") +
  xlab(" ") +
  ylab(" ") +                                # Y-axis label here would be too busy
  theme(axis.text.x = element_text(size = 8, angle=45, vjust=1, hjust=1),
        axis.title = element_text(size = 9),
        plot.title = element_text(size = 9))

panel3 <- mydata %>% filter(racef=="Other")
p3 <- ggplot(data=panel3) +
  aes(x=no_trimester1f, y=fit) +
  geom_errorbar(aes(ymin=lwr,
                     ymax=upr),
                 width=.05,
                 color="blue") +
  geom_point(color="blue") +
  scale_y_continuous(limits=c(2000,3500), breaks=seq(2000,3500,500)) +      # Be sure to use SAME y-axis scale for all
  ggtitle("Other") +
  xlab(" ") +
  ylab(" ") +                                # Y-axis label here would be too busy
  theme(axis.text.x = element_text(size = 8, angle=45, vjust=1, hjust=1),
        axis.title = element_text(size = 9),
        plot.title = element_text(size = 9))

gridExtra::grid.arrange(p1, p2, p3, ncol=3,
                       top=textGrob("Estimated Mean Birthweight, g (95% CI) over 1st Trimester Visits: by Race"))

```

