Homework 6

Qianbo Wang
uni: qw2180

Consider the ChickWeight data in R. The body weights of the chicks were measured at birth (i.e., time=0) and every second day thereafter until day 20. They were also measured on day 21. There were four groups of chicks on different protein diets.

Problem 1

Determine whether there is a significant difference in the mean weights of the four groups on Day 18.

- **Without adjusting for Birth Weight.**

Since without adjusting for Birth Weight, the model is just simple one-way anova. The anova table is as follows:

Anova without adjusting for Birth Weight

```
Call:
   aov(formula = weight ~ Diet, data = sub_day18)

Terms:
                     Diet Residuals
Sum of Squares   36690.44 114840.84
Deg. of Freedom         3        43

Residual standard error: 51.67898
Estimated effects may be unbalanced
            Df Sum Sq Mean Sq F value Pr(>F)
Diet         3  36690   12230   4.579 0.0072 **
Residuals   43 114841    2671
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Since for F-test, the null and alternative hypothesis are:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 : \mu_i \neq \mu_j \text{ for at least one pair (i , j), } i \neq j$$

And the $p-value = 0.0072 < 0.05$. Then we should reject the null hypothesis, i.e. there is a significant difference in the mean weights of the four groups on Day 18.

- **Adjusting for Birth Weight. Give the LS Means (i.e., adjusted for Birth Weight).**

Since we need do anova analysis based on adjusting for Birth Weight, then we simply add a new variable as Birth Weight, i.e. Give the LS means. Then this time the model is anovca. The anovca table is as follows:

Anova with adjusting for Birth Weight

```
Call:
   aov(formula = weight ~ birthweight + Diet, data = sub_day18)

Terms:
                birthweight      Diet Residuals
Sum of Squares     15728.23  27190.80 108612.25
Deg. of Freedom           1         3        42

Residual standard error: 50.85279
Estimated effects may be unbalanced
            Df Sum Sq Mean Sq F value Pr(>F)
birthweight  1  15728   15728   6.082 0.0178 *
Diet         3  27191    9064   3.505 0.0234 *
Residuals   42 108612    2586
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

And the Birth Weight means of the four groups is as follows:

Table 1: Birth Weight LS Means

Diet	Birth Weight mean	LS mean
1	41.58824	164.2240
2	40.70000	183.2445
3	40.70000	229.7409
4	41.00000	201.7337

Since this time for F-test, the null and alternative hypothesis are:

$$H_0 : \hat{\mu_1} = \hat{\mu_2} = \hat{\mu_3} = \hat{\mu_4}$$

$$H_1 : \hat{\mu_i} \neq \hat{\mu_j} \text{ for at least one pair (i , j), } i \neq j$$

And the $p - value$ for Birth Weight is $p - value = 0.0178 < 0.05$ and $p - value$ for Diet is $p - value = 0.0234 < 0.05$. Then we should reject the null hypothesis, i.e. there is a significant difference in the mean weights of the four groups on Day 18 regarding the Birth Weight differences.

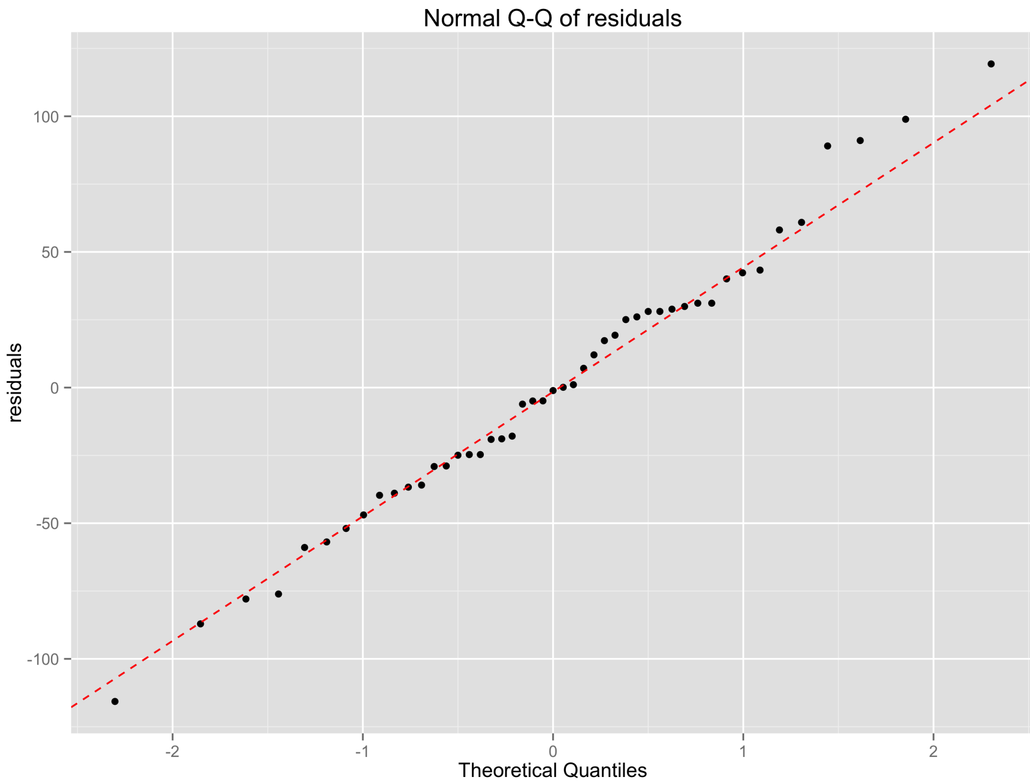
- **Check the validity of your assumptions, including parallelism. Suggest measures that you would take if the assumptions are not satisfied.**

Since the assumptions are:

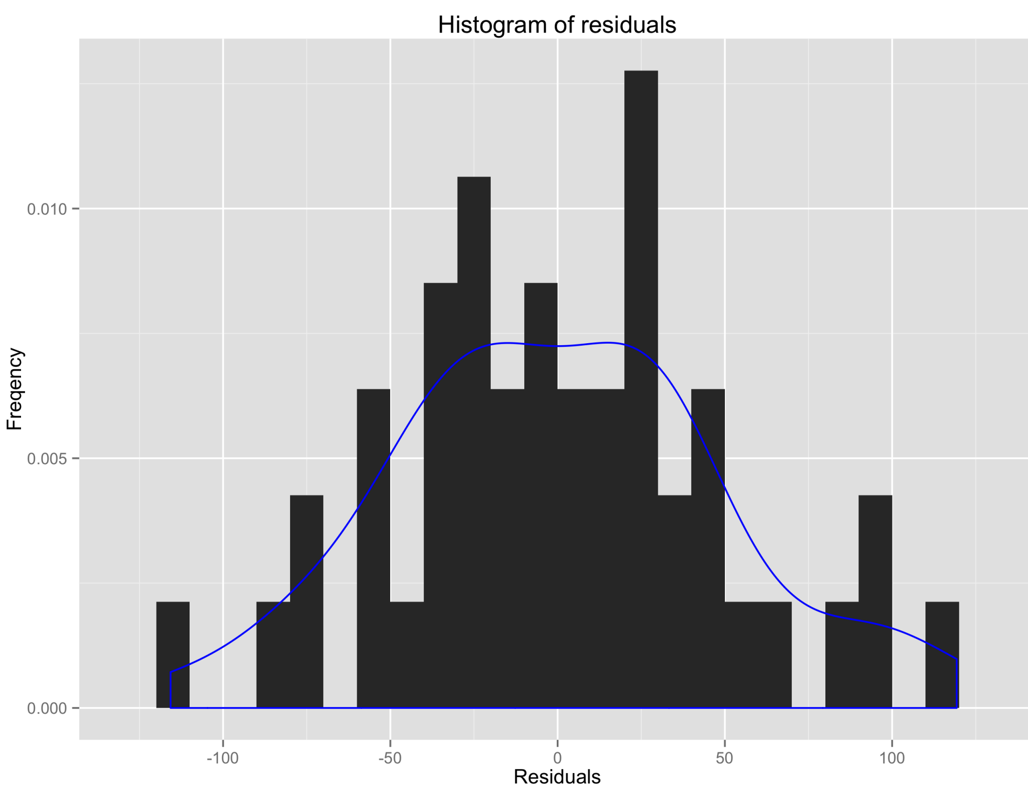
- i.i.d. normal
- constant variance

1. Check i.i.d. Normality

First, plot Q-Q plot on residuals, the plot is as follows:



Then, plot histogram with density on residuals, the plot is as follows:



Since from the plot, the residuals seem to satisfy normal distribution. Then do Shapiro-Wilk test on residuals, The null and alternative hypothesis of Shapiro-Wilk test are:

$$H_0 : \epsilon_i \sim \text{Normal}$$

$$H_1 : \epsilon_i \text{ not from Normal}$$

And the Shapiro-Wilk test result is as follows:

Shapiro-Wilk Test on residuals

```
Shapiro-Wilk normality test

data:  residual$residuals
W = 0.9886, p-value = 0.9242
```

Since $p - value = 0.9242$, then we should not reject the null, i.e. the residuals satisfy normal distribution.

Then, check the i.i.d. assumption, use nonparametric Kruskal-Wallis test on raw data. The null and alternative hypothesis of Kruskal-Wallis test are:

$$H_0 : w_1, w_2, w_3, w_4 \text{ i.i.d.}$$

$$H_1 : w_1, w_2, w_3, w_4 \text{ not i.i.d.}$$

The result is as follows:

Kruskal-Wallis test on raw data

```
Kruskal-Wallis rank sum test

data:  weight by Diet
Kruskal-Wallis chi-squared = 10.6234, df = 3, p-value = 0.01395
```

Since $p - value = 0.01395$, then we should reject the null, i.e. the weight data in the four Diet group are not from the same distribution.

2. Check Constant Variance

First, use Bartlett's on raw data to check the constant variance. The null and alternative hypothesis of Bartlett's test are:

$$H_0 : {\sigma_1}^2 = ... = {\sigma_4}^2$$

$$H_1 : {\sigma_i}^2 \neq {\sigma_j}^2 \text{ for at least one pair (i , j), } i \neq j$$

And the Bartlett's test is highly dependent on the normal assumption, since we have check the validation of normality, then we can use Bartlett's test. The result is as follows:

Bartlett's test on raw data

```
Bartlett test of homogeneity of variances

data:  sub_day18$weight and sub_day18$Diet
Bartlett's K-squared = 3.576, df = 3, p-value = 0.311
```

Since $p - value = 0.3111 > 0.05$, then we should not reject the null, i.e. the constant variance assumption satisfies.

Then use Levene's Test on raw data to check the constant variance. The null and alternative hypothesis of Levene's test are:

$$H_0 : {\sigma_1}^2 = ... = {\sigma_4}^2$$

$$H_1 : {\sigma_i}^2 \neq {\sigma_j}^2 \text{ for at least one pair (i , j), } i \neq j$$

The result is as follows:

Levene's test on raw data

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3   1.059 0.3764
      43
```

Since $p - value = 0.3764 > 0.05$, then we should not reject the null, i.e. the constant variance assumption satisfies.

2. Check parallelism

Use two-way anova to check the parallelism, the null and alternative hypothesis are:

$$H_0 : \text{ there is joint effect with Diet and Birth Weight}$$

$$H_1 : \text{ there is no joint effect with Diet and Birth Weight}$$

The result is as follows:

Two-way Anova for parallelism check

```
                 Df Sum Sq Mean Sq F value Pr(>F)
birthweight       1  15728   15728   6.134 0.0177 *
Diet              3  27191    9064   3.535 0.0234 *
birthweight:Diet  3   8610    2870   1.119 0.3530
Residuals        39 100002    2564
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

And the $p - value = 0.3530 > 0.05$. Then we should not reject the null hypothesis, i.e. there is no interaction effect between Diet group and the Birth Weight. So we do not need do marginal inference on BirthWeight and Diet.

Problem 2

Perform an appropriate repeated measures ANOVA to determine whether there is a significant difference in the mean weights of the four groups using the measurements on Days 10, 18, and 21.

- **Do the analyses assuming compound symmetry and unstructured covariance structures and compare the results.**

1. Compound Symmetry

For this part of the problem, I simply use SAS to do the repeat measures ANOVA because SAS has a easily-used implementation of repeated measures ANOVA based on different structure assumptions on covariance. Since for the assumption, it assumes that the covariance between the data is compound symmetry or unstructured. Then the result is in the following page.

2. Unstructured Covariance

Since the assumption is unstructured covariance, then the result is in the following page.

Repeated Measures on Chick Weight
compound symmetry assumption

23:48 Wednesday, October 21, 2015 **1**

The Mixed Procedure

Model Information	
Data Set	REPEAT.REPEATDATA
Dependent Variable	weight
Covariance Structure	Compound Symmetry
Subject Effect	Chick(Diet)
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Class Level Information		
Class	**Levels**	**Values**
Diet	4	1 2 3 4
Chick	49	1 10 11 12 13 14 15 16 17 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 43 44 45 46 47 48 49 5 50 6 7 8 9
Time	3	10 18 21

Dimensions	
Covariance Parameters	2
Columns in X	20
Columns in Z	0
Subjects	49
Max Obs Per Subject	3

Number of Observations	
Number of Observations Read	141
Number of Observations Used	141
Number of Observations Not Used	0

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	1396.07743029	
1	2	1338.99956090	0.00000228
2	1	1338.99828947	0.00000000

Convergence criteria met.

Repeated Measures on Chick Weight
compound symmetry assumption

23:48 Wednesday, October 21, 2015 **2**

The Mixed Procedure

Estimated R Correlation Matrix for Chick(Diet) 1 1

Row	Col1	Col2	Col3
1	1.0000	0.6626	0.6626
2	0.6626	1.0000	0.6626
3	0.6626	0.6626	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	Chick(Diet)	1534.18
Residual		781.13

Fit Statistics

-2 Res Log Likelihood	1339.0
AIC (smaller is better)	1343.0
AICC (smaller is better)	1343.1
BIC (smaller is better)	1346.8

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	57.08	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Diet	3	45	5.95	0.0016
Time	2	84	192.58	<.0001
Diet*Time	6	84	3.76	0.0023

Repeated Measures on Chick Weight
unstructured correlation assumption

23:48 Wednesday, October 21, 2015 **1**

The Mixed Procedure

Model Information	
Data Set	REPEAT.REPEATDATA
Dependent Variable	weight
Covariance Structure	Unstructured
Subject Effect	Chick(Diet)
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Class Level Information		
Class	**Levels**	**Values**
Diet	4	1 2 3 4
Chick	49	1 10 11 12 13 14 15 16 17 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 43 44 45 46 47 48 49 5 50 6 7 8 9
Time	3	10 18 21

Dimensions	
Covariance Parameters	6
Columns in X	20
Columns in Z	0
Subjects	49
Max Obs Per Subject	3

Number of Observations	
Number of Observations Read	141
Number of Observations Used	141
Number of Observations Not Used	0

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	1396.07743029	
1	3	1177.78560780	0.00103860
2	1	1177.29205471	0.00004402
3	1	1177.27077572	0.00000009
4	1	1177.27073492	0.00000000

Repeated Measures on Chick Weight
unstructured correlation assumption

The Mixed Procedure

Convergence criteria met.

Estimated R Correlation Matrix for Chick(Diet) 1 1			
Row	Col1	Col2	Col3
1	1.0000	0.7988	0.6615
2	0.7988	1.0000	0.9632
3	0.6615	0.9632	1.0000

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
UN(1,1)	Chick(Diet)	428.98
UN(2,1)	Chick(Diet)	882.94
UN(2,2)	Chick(Diet)	2848.34
UN(3,1)	Chick(Diet)	895.29
UN(3,2)	Chick(Diet)	3359.26
UN(3,3)	Chick(Diet)	4270.54

Fit Statistics	
-2 Res Log Likelihood	1177.3
AIC (smaller is better)	1189.3
AICC (smaller is better)	1190.0
BIC (smaller is better)	1200.6

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
5	218.81	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Diet	3	45	6.03	0.0015
Time	2	45	100.88	<.0001
Diet*Time	6	45	2.47	0.0375

Compare the result:
Since both under compound symmetry covariance and unstructured covariance assumption, all of the p-values are significant, then we should conclude that the chicken have significant different of weight on at least two Diet group and between Day 10, Day18 and Day 21.

- **Check the validity of your assumptions.**

1. Normality

Since the assumptions for repeated measures are normality and constant variance. Then under the compound symmetry assumption and unstructured covariance assumption, the plots of the residuals given by SAS are as follows: Since from the plots, we see that the data is approximately normally distributed. Then do normal test on residuals, the result is also on the following pages.
Since under compound symmetry covariance assumption, for Shapiro-Wilk test on residuals gives the $p-value = 0.9778 > 0.05$, then we should not reject the null, i.e. the normality assumption satisfies.
And under unstructured covariance assumption, for Shapiro-Wilk test on residuals gives the $p-value = 0.9753 > 0.05$, then we should not reject the null, i.e. the normality assumption satisfies.

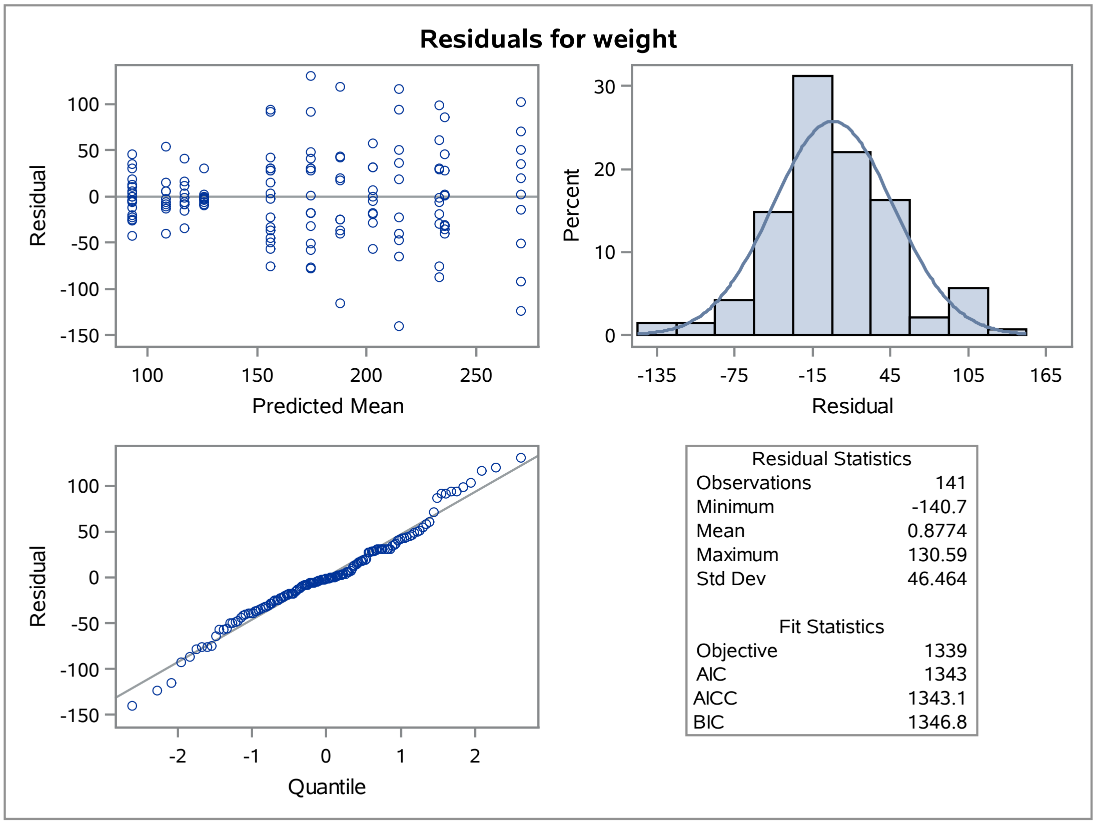
2. Constant Variance

Since normality assumption satisfies then use Bartlett's test on raw data gives the $p-value = 0.0561 > 0.05$, then we should not reject the null, i.e. the Constant Variance assumption satisfies.

Repeated Measures on Chick Weight
compound symmetry assumption

23:48 Wednesday, October 21, 2015 **3**

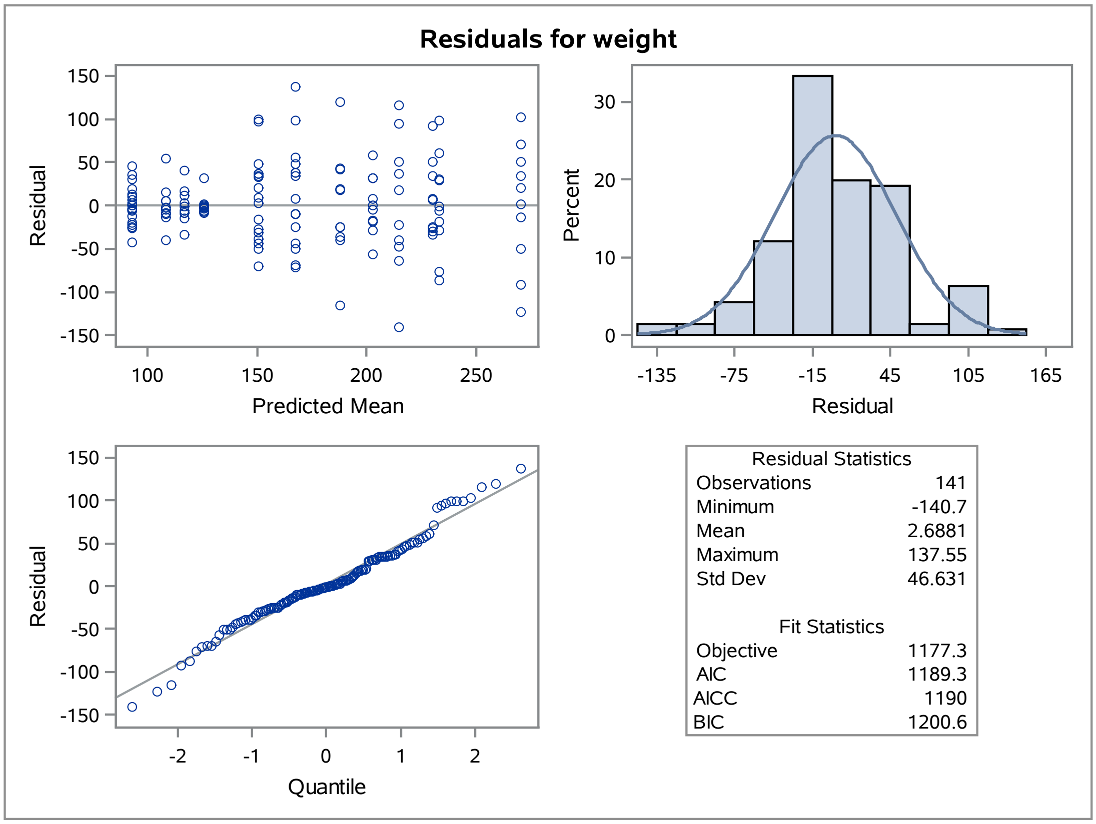
The Mixed Procedure



Repeated Measures on Chick Weight
unstructured correlation assumption

23:48 Wednesday, October 21, 2015 3

The Mixed Procedure



Noramlity test on residuals
compound symmetry assumption

23:48 Wednesday, October 21, 2015 1

The UNIVARIATE Procedure
Variable: Resid (Residual)

Moments			
N	141	Sum Weights	141
Mean	0.87741402	Sum Observations	123.715377
Std Deviation	46.4638738	Variance	2158.89157
Skewness	0.09585474	Kurtosis	0.97181334
Uncorrected SS	302353.37	Corrected SS	302244.82
Coeff Variation	5295.5472	Std Error Mean	3.91296406

Basic Statistical Measures			
Location		Variability	
Mean	0.8774	Std Deviation	46.46387
Median	-1.1000	Variance	2159
Mode	-17.4135	Range	271.28652
		Interquartile Range	54.60000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0.224233	Pr > |t|	0.8229
Sign	M	-4.5	Pr >= |M|	0.5006
Signed Rank	S	-51.5	Pr >= |S|	0.9160

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.977752	Pr < W	0.0212
Kolmogorov-Smirnov	D	0.080669	Pr > D	0.0235
Cramer-von Mises	W-Sq	0.187511	Pr > W-Sq	0.0078
Anderson-Darling	A-Sq	1.136853	Pr > A-Sq	0.0056

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	130.5865
99%	119.3000
95%	91.6745
90%	50.7000
75% Q3	29.9000
50% Median	-1.1000

Noramlity test on residuals
unstructured correlation assumption

23:48 Wednesday, October 21, 2015 1

The UNIVARIATE Procedure
Variable: Resid (Residual)

Moments			
N	141	Sum Weights	141
Mean	2.68808255	Sum Observations	379.019639
Std Deviation	46.6314288	Variance	2174.49016
Skewness	0.11916932	Kurtosis	1.05114803
Uncorrected SS	305447.458	Corrected SS	304428.622
Coeff Variation	1734.74691	Std Error Mean	3.92707474

Basic Statistical Measures			
Location		Variability	
Mean	2.6881	Std Deviation	46.63143
Median	-1.1000	Variance	2174
Mode	-10.4541	Range	278.24589
		Interquartile Range	55.80000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0.6845	Pr > |t|	0.4948
Sign	M	-4.5	Pr >= |M|	0.5006
Signed Rank	S	183	Pr >= |S|	0.7079

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.975347	Pr < W	0.0119
Kolmogorov-Smirnov	D	0.079551	Pr > D	0.0273
Cramer-von Mises	W-Sq	0.203494	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.25304	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	137.5459
99%	119.3000
95%	97.1833
90%	54.5000
75% Q3	31.1000
50% Median	-1.1000

Constant Variance test on raw data

23:48 Wednesday, October 21, 2015 3

The GLM Procedure

Bartlett's Test for Homogeneity of weight Variance			
Source	DF	Chi-Square	Pr > ChiSq
Diet	3	7.5574	0.0561

Constant Variance test on raw data

The GLM Procedure

Levene's Test for Homogeneity of weight Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Diet	3	3.2505E8	1.0835E8	3.90	0.0103
Error	137	3.804E9	27766281		

R Code:

```
rm(list=ls())
data('ChickWeight')
#avova on day 18
sub_day18<-subset(ChickWeight,Time==18)
anova_day18<-aov(data = sub_day18, weight ~ Diet)
summary(anova_day18)

sink('/Users/raymond/Drive/STAT W4201/HW6/anova.txt')
anova_day18
summary(anova_day18)
sink()

#anova adjust by LS mean on day 18
sub_day0<-subset(ChickWeight,Time==0)
sub_day18[,'birthweight']<-sub_day0$weight[match(sub_day18$Chick,sub_day0$Chick)]
group_mean<-aggregate(sub_day18$birthweight,list(sub_day18$Diet),mean)
colnames(group_mean)<-c('Diet','birthweight_mean')
sub_day18<-merge(sub_day18,group_mean,by.y = 'Diet')
anova_adjust_day18<-aov(data=sub_day18, weight ~ birthweight + Diet)

sink('/Users/raymond/Drive/STAT W4201/HW6/anova_adjust.txt')
anova_adjust_day18
summary(anova_adjust_day18)
sink()

#LS means
coef<-anova_adjust_day18$coefficients[2]
y_hat<-anova_adjust_day18$fitted.values
y_frame<-data.frame(yhat=y_hat,Diet=sub_day18$Diet)
yhat_mean<-aggregate(y_frame$yhat, list(y_frame$Diet),mean)
colnames(yhat_mean)<-c('Diet','yhat')
mu<-yhat_mean$yhat-coef*(group_mean$birthweight_mean-mean(sub_day18$birthweight))
lsmean_result<-cbind(group_mean,lsmean=mu)
sink('/Users/raymond/Drive/STAT W4201/HW6/LSmeans.txt')
lsmean_result
sink()
#Or simply use lsmeans function in library lsmeans
library (lsmeans)
lsmean<-lsmeans(anova_adjust_day18,'Diet')
sink('/Users/raymond/Drive/STAT W4201/HW6/LSmeans1.txt')
lsmean
sink()

library (ggplot2)
#check assumptions
#Normality
#Q-Q plot for residuals
residual<-data.frame(anova_day18['residuals'],sub_day18['Diet'])
normal_plot<-ggplot(data=residual,aes(sample=residuals))+
  stat_qq()+ggtitle('Normal Q-Q of residuals')+
  ylab('residuals')+xlab('Theoretical Quantiles')
resid_quantile<-quantile(residual$residuals,c(0.25, 0.75))
norm_quantile<-qnorm(c(0.25, 0.75))
slope<-diff(resid_quantile)/ diff (norm_quantile)
inter<-resid_quantile[1]-slope*norm_quantile[1]
normal_plot+geom_abline(slope=slope,intercept=inter,linetype=2,color='red')
ggsave(filename = '/Users/raymond/Drive/STAT W4201/HW6/qqplot.png')

#histogram for residuals
histogram_plot<-ggplot(data=residual,aes(residuals))+
  geom_bar(binwidth=10,aes(y=..density..))+ggtitle("Histogram of residuals")+
  ylab('Freqency')+xlab('Residuals')
histogram_plot+geom_density(color='blue')
ggsave(filename = '/Users/raymond/Drive/STAT W4201/HW6/histogram.png')

#shapiro test on residuals
```

```
shapiro.test(residual$residuals)
sink('/Users/raymond/Drive/STAT W4201/HW6/normalcheck.txt')
shapiro.test(residual$residuals)
sink()

#check constant variance
#bartlett test on data
bartlett.test(x=sub_day18$weight,g=sub_day18$Diet)
sink('/Users/raymond/Drive/STAT W4201/HW6/constantcheck.txt')
bartlett.test(x=sub_day18$weight,g=sub_day18$Diet)
sink()

#levene test on data
library(car)
leveneTest(y=sub_day18$weight,group=as.factor(sub_day18$Diet))

#nonparametric check iid assumptions
#kruskal test on data
kruskal.test(weight~Diet,data=sub_day18)
sink('/Users/raymond/Drive/STAT W4201/HW6/identicalcheck.txt')
kruskal.test(weight~Diet,data=sub_day18)
sink()

#check for parallelism
#two-way anova
anova_inter<-aov(weight ~ birthweight*Diet,data = sub_day18)
summary(anova_inter)
sink('/Users/raymond/Drive/STAT W4201/HW6/paracheck.txt')
summary(anova_inter)
sink()

#repeated measures
#anova on day 10, 18, 21
sub_repeat<-subset(ChickWeight,Time %in% c(10,18,21))
repeated<-aov(weight ~ Diet*Time+Error(Chick),data=sub_repeat)

write.table(sub_repeat, file = '/Users/raymond/Drive/STAT W4201/HW6/ChickWeight.csv',col.names=TRUE,row.names=
    FALSE,sep=",")
```

SAS code:

```
libname repeat 'C:\Users\QIANBO\Desktop\HW6';
*import data from csv;
proc import out=repeat.repeatdata
    Datafile = 'C:\Users\QIANBO\Desktop\HW6\ChickWeight.csv'
    Dbms =csv replace;
run;
proc print data = repeat.repeatdata;
run;

ods listing close;
ods graphics on;
options papersize=letter;
ods pdf file ='C:\Users\QIANBO\Desktop\HW6\compoundsy.pdf';
 title 'Repeated Measures on Chick Weight';
 title2 'compound symmetry assumption';

*correlation compound symmetry assumption;
proc mixed data = repeat.repeatdata;
 class Diet Chick Time;
model weight= Diet Time Diet*Time / outp=repeat.csparam residual;
repeated/type=cs sub=Chick(Diet) rcorr ;
run;
ods pdf close;

ods pdf file ='C:\Users\QIANBO\Desktop\HW6\unstructured.pdf';
 title 'Repeated Measures on Chick Weight';
 title2 'unstructured correlation assumption';
*correlation unstructured assumption;
proc mixed data = repeat.repeatdata;
 class Diet Chick Time;
model weight = Diet Time Diet*Time / outp=repeat.unparam residual;
repeated/type=un sub=Chick(Diet) rcorr;
run;
ods pdf close;
ods graphics off;

ods pdf file ='C:\Users\QIANBO\Desktop\HW6\normalcs.pdf';
 title 'Noramlity test on residuals';
 title2 'compound symmetry assumption';
*test normality;
proc univariate data=repeat.csparam normal;
var resid;
run;
ods pdf close;

ods pdf file ='C:\Users\QIANBO\Desktop\HW6\normalun.pdf';
 title 'Noramlity test on residuals';
 title2 'unstructured correlation assumption';
*test normality;
proc univariate data=repeat.unparam normal;
var resid;
run;
ods pdf close;

ods pdf file ='C:\Users\QIANBO\Desktop\HW6\homo.pdf';
 title 'Constant Variance test on raw data';
 title2 ;
proc glm data=repeat.repeatdata;
 class Diet Chick Time;
model weight = Diet;
means Diet / HOVTEST=BARTLETT;
means Diet / HOVTEST=LEVENE;
run;
ods pdf close;
ods listing;
 title;
```