Normalization Test of Data On Extension Mechanisms of the Knowledge Discovery Metamodel

October 14, 2017

1 Normalization Test in Development Activities on Time

In order to use multifactorial ANOVA for analyzing the data, a precondition imposed for this analysis is the assumption of normality. Therefore, our first analysis is checking whether there is not a violation of normality by using Shapiro-Wilk test.

Firstly, we load the data and perform the analysis of variance:

```
dtExtensionDevelopment = read.csv("extension_development.csv")
# convert to nominal factor
dtExtensionDevelopment$Subject = factor(dtExtensionDevelopment$Subject)
# convert to nominal factor
dtExtensionDevelopment$Activity = factor(dtExtensionDevelopment$Activity)
# fit model
m = aov(Time ~ Technique*Activity, data=dtExtensionDevelopment)
```

Secondly, we create the Q-Q (quantile-quantile) plot, for comparing two probability distributions:

```
# plot residuals
qqnorm(residuals(m)); qqline(residuals(m))
```

Thirdly, we perform the Shapiro-Wilk test of normality:

```
##Check Normality
# but really what matters most is the residuals
shapiro.test(residuals(m))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(m)
## W = 0.89755, p-value = 2.129e-09
```

The Shapiro-Wilk test shows a significant p-value ($< .05$) meaning the rejection of the null hypothesis that these data (residuals) were independently drawn from a common normal distribution. This is a problem if we want to run our lmm test or even a 2-way rANOVA test. Also, in Figure 1 we present a Normal Q-Q plot for visually inspecting whether our data (time variable) plausibly came from normal distribution. We can see that the spread of data does not conform with the line and consequently in accordance with the result of Shapiro-Wilk analysis.

2 Normalization Test in Development Activities on Error

As in the previous section, we check whether there is not a violation of normality by using ShapiroWilk test.

Firstly, we load the data and perform the analysis of variance:

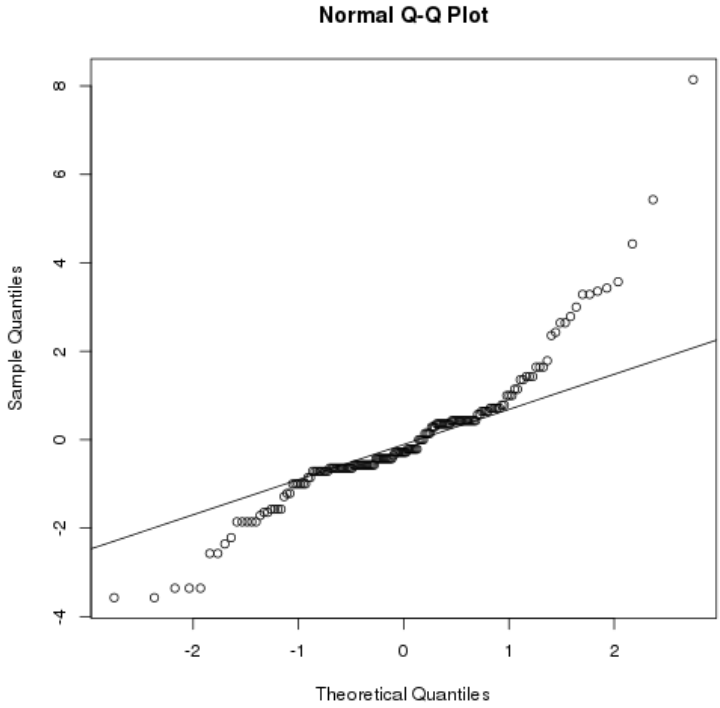

Figure 1: Normal Q-Q Plot of Development Activities on Time Variable

```
dtExtensionDevelopment = read.csv("extension_development.csv")
# convert to nominal factor
dtExtensionDevelopment$Subject = factor(dtExtensionDevelopment$Subject)
# convert to nominal factor
dtExtensionDevelopment$Activity = factor(dtExtensionDevelopment$Activity)
# fit model
m = aov(Error ~ Technique*Activity, data=dtExtensionDevelopment)
```

Secondly, we create the Q-Q (quantile-quantile) plot, for comparing two probability distributions:

```
# plot residuals
qqnorm(residuals(m)); qqline(residuals(m))
```

Thirdly, we perform the Shapiro-Wilk test of normality:

```
##Check Normality
# but really what matters most is the residuals
shapiro.test(residuals(m))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(m)
## W = 0.47312, p-value < 2.2e-16
```

The Shapiro-Wilk test shows a significant p-value ($< .05$) meaning the rejection of the null hypothesis that these data (residuals) were independently drawn from a common normal distribution. Also, in Figure 2 we present a Normal Q-Q plot for visually inspecting whether our data (error variable) plausibly came from normal distribution. We can see that the spread of data does not conform with the line and consequently in accordance with the result of Shapiro-Wilk analysis.

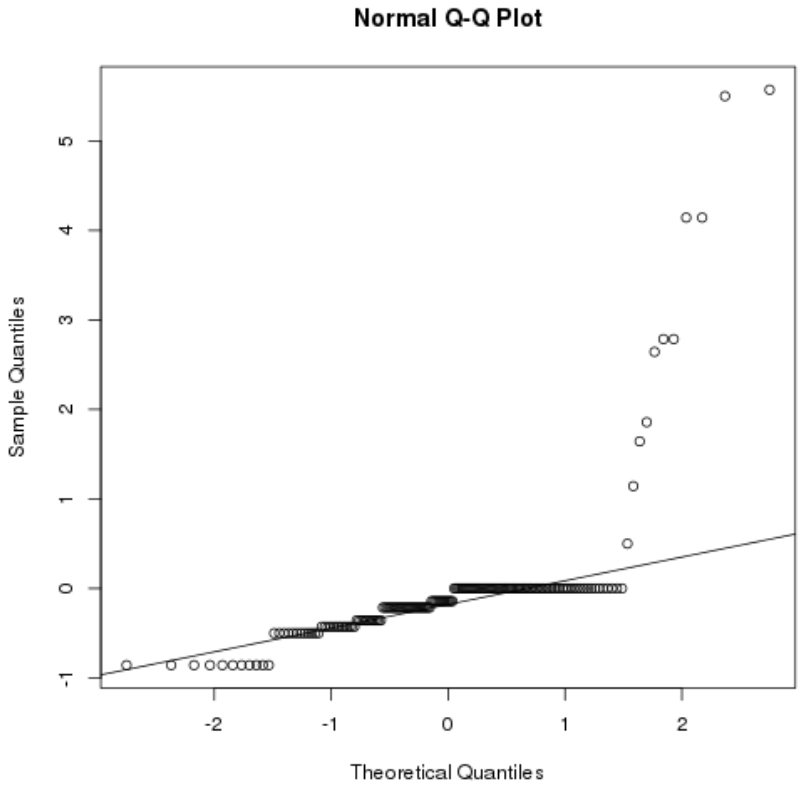

Figure 2: Normal Q-Q Plot of Development Activities on Error Variable

3 Normalization Test in Maintenance Activities on Time

In this section, we check whether there is not a violation of normality by using ShapiroWilk test. Firstly, we load the data and perform the analysis of variance:

```
dtExtensionMaintenance = read.csv("extension_maintenance.csv")
# convert to nominal factor
dtExtensionMaintenance$Subject = factor(dtExtensionMaintenance$Subject)
# convert to nominal factor
dtExtensionMaintenance$Activity = factor(dtExtensionMaintenance$Activity)
# fit model
m = aov(Time ~ Technique*Activity, data=dtExtensionMaintenance)
```

Secondly, we create the Q-Q (quantile-quantile) plot, for comparing two probability distributions:

```
# plot residuals
qqnorm(residuals(m)); qqline(residuals(m))
```

Thirdly, we perform the Shapiro-Wilk test of normality:

```
##Check Normality
# but really what matters most is the residuals
shapiro.test(residuals(m))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(m)
## W = 0.97532, p-value = 0.3042
```

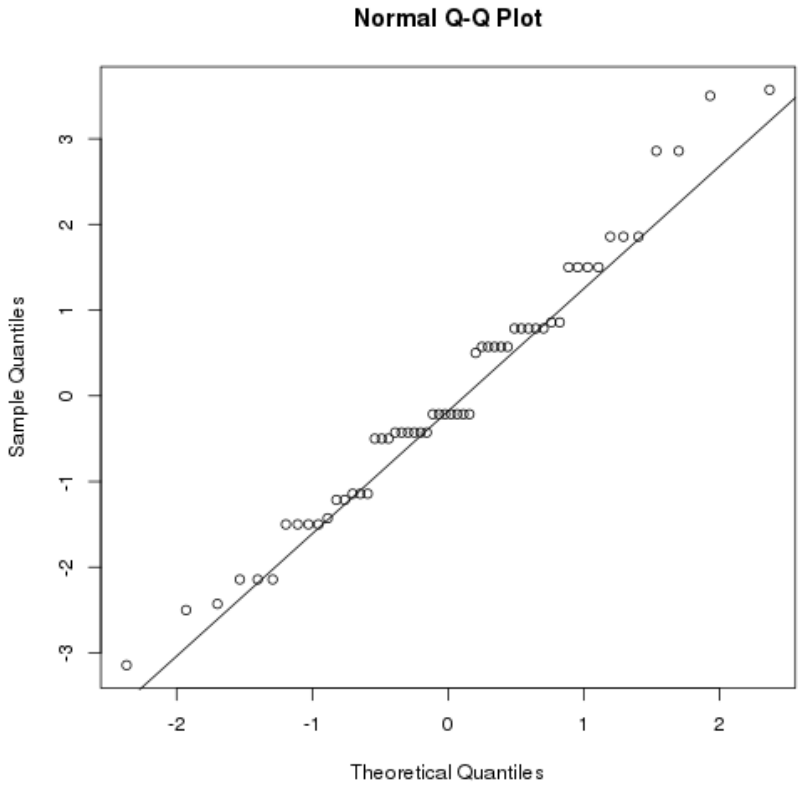

Figure 3: Normal Q-Q Plot of Maintenance Activities on Time Variable

As p-values $> .05$ we accept the null hypothesis that residuals of the data are normally distributed. Indeed, in Figure 3 we see that data conformed with the sloped line and consequently in accordance with the result of Shapiro-Wilk analysis.

4 Normalization Test in Maintenance Activities on Error

In this section, we check whether there is not a violation of normality by using ShapiroWilk test. Firstly, we load the data and perform the analysis of variance:

```
dtExtensionMaintenance = read.csv("extension_maintenance.csv")
# convert to nominal factor
dtExtensionMaintenance$Subject = factor(dtExtensionMaintenance$Subject)
# convert to nominal factor
dtExtensionMaintenance$Activity = factor(dtExtensionMaintenance$Activity)
# fit model
m = aov(Error ~ Technique*Activity, data=dtExtensionMaintenance)
```

Secondly, we create the Q-Q (quantile-quantile) plot, for comparing two probability distributions:

```
# plot residuals
qqnorm(residuals(m)); qqline(residuals(m))
```

Thirdly, we perform the Shapiro-Wilk test of normality:

```
##Check Normality
# but really what matters most is the residuals
shapiro.test(residuals(m))
```

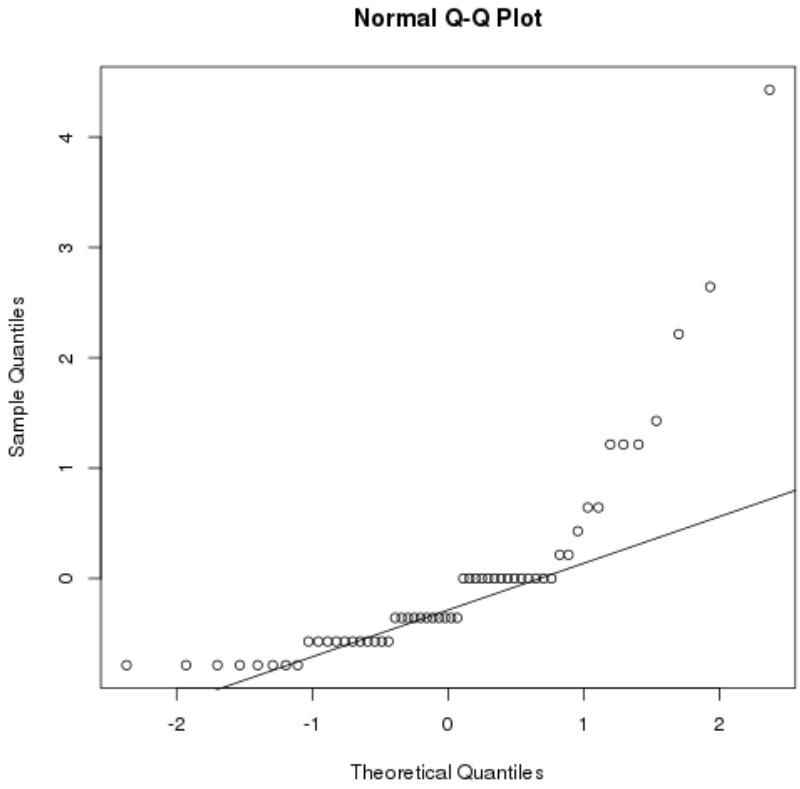

Figure 4: Normal Q-Q Plot of Maintenance Activities on Error Variable

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m)
## W = 0.70834, p-value = 3.043e-09
```

The result shows a significant p-value ($< .05$), meaning the rejection of the hypothesis of data distributed normally. In Figure 4 we present a Normal Q-Q plot for visually inspecting whether our data (error variable) plausibly came from normal distributions. We can see that the spread of data does not conform with the line and consequently in accordance with the result of Shapiro-Wilk analysis.