Model Diagnostics and Remedial Measures

One-Way ANOVA – Stat 230

In this class, you will work with your group to explore the use of normal Q-Q plots to assess the normality of residuals in ANOVA models. You will also learn how to apply transformations to the response variable in order to remedy violations to the model assumptions.

Reading Normal Q-Q Plots

In the reading you learned how normal Q-Q plots are constructed and why they are useful. Today, we'll start off by creating a number of these plots to practice reading them.

Activity A. Go to https://shiny.mathcs.carleton.edu/users/aloy/qqnorm/ or https://loya.shinyapps.io/qqplots/. You'll see a web app where you can generate a histogram and normal Q-Q plot for a variety of sample sizes and distributional shapes. Looking at multiple examples will help you develop your intuition about what constitute a substantial deviation from a straight line on a normal Q-Q plot. Do the following:

(a) The default settings are a sample size of 15 observations drawn from a normal population distribution. Generate a few of these plots by pressing "Plot it!". Do the points always fall on the line?

(b) Now, increase the sample size. How does the Q-Q plot change as the sample size increases?

(c) Now, change the shape of the population distribution to "right skewed". Generate a few Q-Q plots at a few sample sizes. Describe the form/shape of the Q-Q plot for right-skewed data.

(d) Now, change the shape of the population distribution to "left skewed". Generate a few Q-Q plots at a few sample sizes. Describe the form/shape of the Q-Q plot for left-skewed data.

(e) Now, change the shape of the population distribution to "heavy tailed". Generate a few Q-Q plots at a few sample sizes. Describe the form/shape of the Q-Q plot for heavy-tailed data.

Now that you have calibrated your intuition, let's create a normal Q-Q plot in R.

Creating Normal Q-Q plots

Now that you have had more practice reading normal Q-Q plots, let's use the understanding you've gained to diagnose models!

In Activity 26 you calculated the residuals from the one-way ANOVA model for the `Games1` data set using the following code:

```
games1 <- read.csv("https://aloy.rbind.io/kuiper_data/Games1.csv")
games_anova <- aov(Time ~ Type, data = games1)
games_resid <- resid(games_anova)
```

Exercise E.10

(a) Create a normal Q-Q plot and histogram of the residuals. Comment on the normality assumption for the random error terms. To create a normal Q-Q plot in R, we use the `gf_qq()` and `gf_qqline()` functions (ignore the warning message printed).

```
gf_qq(~games_resid) |>
  gf_qqline()
```

(b) Create a normal Q-Q plot and histogram of the observed response variable, `Time`.

(c) Explain why residuals should be used instead of the observed responses to test the normality assumption.

Applying Transformations

Transforming the response variable can often help remedy non-normal error terms or subgroups with dramatically different variances. The textbook has a nice discussion on when the square root, log, reciprocal, and logit transformations are often useful, so be sure to review this.

Extended Activity 31. To practice working with transformations, let's work through an adapted version of extended activity 31 from the textbook. You can load the data via

```
emissions <- read.csv("https://aloy.rbind.io/kuiper_data/Emissions.csv")
```

For a description of the data, see page 47 of the textbook. In the daily prep, you worked through parts (a)-(c).

(d) Fit a one-way ANOVA model to the transformed version of emissions, `log(Emissions)`. To do this, use `log(Emissions)` as the response variable in your `aov()` formula.

(e) Calculate the residuals from your ANOVA model and check whether the necessary conditions/assumptions are met for the ANOVA F-test.

The textbook as a great wrap-up discussion of this activity, be sure to read the last paragraph of extended activity 31.

Back transformations

In activity 31, you applied a transformation to remedy violations so that you could conduct an appropriate F-test for a one-way ANOVA model. Now, let's consider transforming the response for the simple linear regression formulation of a two-sample t-test.

The below code extracts the rows of the data set for the pre-63 and 70-71 cases.

```
sub_emissions <- emissions |>
  filter(Year == "1963-1967" | Year == "1970-1971")
```

Now, let's fit a simple linear regression model with an indicator variable that flags 1970-1971 (i.e., the indicator is 1 when for cases from 1970-1971). In this model, a natural log transformation was applied to the response variable.

```
emissions_lm <- lm(log(Emissions) ~ Year, data = sub_emissions)
```

As seen in the book, a 95% confidence interval for the difference in mean emissions between the years is given by

```r
confint(emissions_lm, parm = 2) # CI only for the second parameter
```

```
                  2.5 %     97.5 %
Year1970-1971 -1.295767 -0.3771826
```

However, this interval is on the natural log scale, which isn't as meaningful/interpretable as the original emissions scale. To get back to the original scale, we need to undue the transformation, which can be done in R (no need for a calculator!)

```r
exp(confint(emissions_lm, parm = 2))
```

```
                  2.5 %    97.5 %
Year1970-1971 0.2736878 0.6857909
```

> **Danger**
>
> Back transforming a logarithmic transformation does not make the results have the "same meaning" as the original data. Caution is needed when interpreting parameter estimates and confidence intervals with log-transformed data.

Activity B. To discover how this back transformation can be interpreted, let's work with the two samples directly:

```r
y1 <- sub_emissions |>
  filter(Year == "1963-1967") |>
  pull(Emissions)
y2 <- sub_emissions |>
  filter(Year == "1970-1971") |>
  pull(Emissions)
```

(a) Compute the sample means and medians for the untransformed samples, `y1` and `y2`.

(b) Compute the sample means and medians for the transformed samples, `log(y1)` and `log(y2)`.

(c) Is the natural log of the difference in sample means approximately equal to the difference in means of the logged data?

(d) Is the natural log of the ratio of sample means from the untransformed samples approximately equal to the difference in means of the logged data?

(e) Is the natural log of the ratio of sample medians from the untransformed samples approximately equal to the difference in medians of the logged data?

(f) Based on (c)-(e), how would you suggest interpreting the back transformation given by:

```r
exp(mean(log(y1)) - mean(log(y2)))
```

```
[1] 2.308216
```

Recall that if a distribution is normal, then the mean and median will be equal.

(g) Based on your suggestion, write a one-sentence interpretation of the back-transformed confidence interval given by

```
exp(confint(emissions_lm, parm = 2))
```

```
                  2.5 %    97.5 %
Year1970-1971 0.2736878 0.6857909
```

If the logged data are roughly symmetric within groups…

- means of the logged data ≈ medians of the logged data
- difference in medians of the logged data ≈ log ratio of the medians of the original data
- if we "undo" the log, then we can interpret the back-transformed value as the ratio of medians

Thus, we are 95% confident that the median emissions from 1970-1971 were between 0.274 to 0.686 times the median emissions from 1963-1967.

Another way to say this is that we are 95% confident that the median emissions from 1970-1971 were lower than the median emissions from 1963-1967 by a factor of between 0.274 to 0.686.