hw9

Jiayi

2024-11-10

Load data

```
data("chickwts", package = "faraway")
```

```
## Warning in data("chickwts", package = "faraway"): data set 'chickwts' not found
```

```
head(chickwts)
```

```
##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

1. Fit One-Way ANOVA Model

```
anova_model <- aov(weight ~ feed, data = chickwts)
summary(anova_model)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## feed         5 231129   46226   15.37 5.94e-10 ***
## Residuals   65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 5.94 * 10^−10, much less than 0.05. Therefore, we have strong evidence to reject the hypothesis of the same mean weights. Thus, the mean weights of chicks differ significantly across different feed types.

2. 95% Confidence Intervals for Mean Weight by Feed

```
# Calculate group means and 95% CIs
ci_results <- confint(lm(weight ~ feed - 1, data = chickwts))
ci_results
```

```
##                   2.5 %   97.5 %
## feedcasein    291.9608 355.2058
## feedhorsebean 125.5593 194.8407
## feedlinseed   187.1275 250.3725
## feedmeatmeal  243.8805 309.9377
## feedsoybean   217.1518 275.7053
## feedsunflower 297.2942 360.5392
```

3. Tukey's Pairwise Comparisons

```
tukey_results <- TukeyHSD(anova_model)
tukey_results
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ feed, data = chickwts)
##
## $feed
##                             diff         lwr       upr     p adj
## horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
## linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
## meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
## soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
## sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
## linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
## meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
## soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
## sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
## meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
## soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
## sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
## soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
## sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
## sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

Based on the Tukey's HSD test with a 0.05 significance level,

Some of the **p-value in the p adj column is below 0.05**, it indicates a statistically significant difference in mean weights between the two feeds.Some examples may include horsebean-casein, linseed - casein, soybean-casein, meatmeal-horsebean, soybean-horsebean, sunflower-horsebean, sunflower-linseed, and sunflower-soybean.

Some of the **p-value is above 0.05**, there is no significant difference in mean weights between those feeds.This includes meatmeal-casein, sunflower-casein, linseed-horsebean, meatmeal-linseed, soybean-linseed, soybean-meatmeal, and sunflower-meatmeal.

We can also tell the difference from the **confidence level**. If the confidence interval (lwr and upr) **does not include zero**, it suggests a statistically significant difference in means. If it includes zero, it means that the difference is not statistically significant.

e.g. **horsebean - casein**: The interval [-232.35, -94.42] **does not include zero**, reinforcing the significant difference. **sunflower - casein**: The interval [-60.42, 71.09] includes zero, meaning there's no significant difference.

4. Check for Outliers Using Bonferroni Correction

```
# Identify potential outliers using Bonferroni correction
outliers <- chickwts %>%
  mutate(studentized_residuals = rstudent(anova_model)) %>%
  filter(abs(studentized_residuals) > qt(1 - 0.05 / n(), df.residual(anova_model)))
outliers
```

```
## [1] weight                feed                  studentized_residuals
## <0 rows> (or 0-length row.names)
```

This means no outliers here.

5. Test for Constant Error Variance

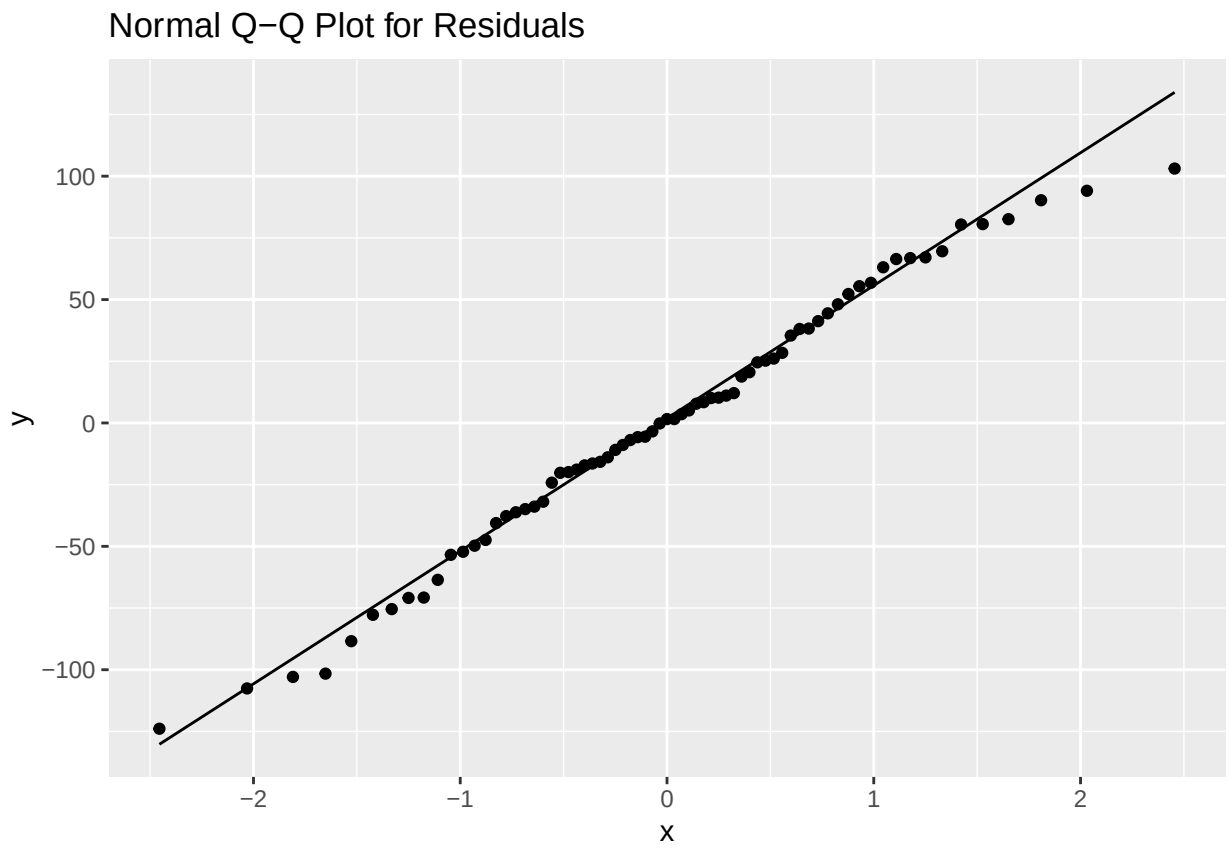
```
# Perform levene test for homogeneity of variances
levene_test <- anova(lm(abs(residuals(anova_model)) ~ feed, data = chickwts))
levene_test
```

```
## Analysis of Variance Table
##
## Response: abs(residuals(anova_model))
##           Df Sum Sq Mean Sq F value Pr(>F)
## feed       5   4985  997.03  0.9873 0.4324
## Residuals 65  65639 1009.83
```

The p-value is 0.4324, which is much above 0.05, we have no sufficent evidence to reject the null hypothesis of constant variance.

6. Checking Normality of Errors

```
# QQ Plot
ggplot(data.frame(residuals = residuals(anova_model)), aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal Q-Q Plot for Residuals")
```




```
# Shapiro-Wilk test for normality
shapiro_test <- shapiro.test(residuals(anova_model))
shapiro_test$p.value
```

```
## [1] 0.6272233
```

Based on the QQ-plot, we can see that the residuals appear to follow a roughly normal distribution, especially in the main range of values, with only minor deviations in the tails.

This suggests that the normality assumption for the residuals is reasonably satisfied for the purposes of ANOVA. Minor deviations at the tails are generally acceptable, as ANOVA is robust to slight non-normality.