Regression, Design and Analysis of Single-Factor Experiments

Lecture #8, #9

Adi Sarid
Tel-Aviv University
updated: 2019-12-21

Reminder from Previous Lecture

We focused on multiple linear regression, which assumes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

- $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$
- For hypothesis testing on β we also require homoscedastity
- We used the relationship $SS_T = SS_R + SS_E$ to define:
- The coefficient of determination $R^2 = \frac{SS_R}{SS_T}$
- When $0 \leq R^2 \leq 1$. High values correspond to a strong relationship
- For the simple linear regression, we saw that $R^2 = \hat{\rho}^2$, the correlation of X and Y

Reminder from Previous Lecture (2) - Solution of the Multiple Linear

We represented the multiple linear regression as a matrix equation:

$$y = X\beta + \epsilon$$

Where:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We are looking for β which minimizes $L = \epsilon^t \epsilon = (y - X\beta)(y - X\beta)$

$$\frac{\partial L}{\partial \beta} = 0$$

Reminder (3) - Least Squares Estimation of the Parameters

The resulting equations are given by:

$$X^t X \hat{\beta} = X^t y$$

In case that $X^t X$ is a non-singular matrix (i.e., invertible), the solution is unique and equals

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

Once $\hat{\beta}$ is found, we can use it to predict our values:

$$\hat{y} = X\hat{\beta}$$

We can also compute the residuals:

$$e = y - \hat{y}$$

Let $p = k + 1$ (the number of parameters including the constant β_0), then:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - p} = \frac{SS_E}{n - p}$$

Is an unbiased estimate of σ_ϵ^2

Reminder (4) - Hypothesis Tests on Individual Coefficients

We used the fact that:

- $\hat{\beta}$ is unbiased and
- $\mathrm{Var}(\hat{\beta}_j) = \sigma^2[(X^tX)^{-1}]_{jj}$

To devise a hypothesis test for the individual $\hat{\beta}_j$:

- $H_0 : \beta_j = 0$
- $H_1 : \beta_j \neq 0$

With the statistic

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}$$

Reminder (5) - Hypothesis Tests on All Coefficients

We used the F-statistic

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$$

To devise a broader test:

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
- $H_1 : \exists i \text{ such that } \beta_i \neq 0$

We examined the corresponding ANOVA table

Source of Variation	**Sum of Squares**	**df**	**Mean Squares**	F_0
Regression	SS_R	k	MS_R	$\frac{MS_R}{MS_E}$
Error	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

The adjusted R^2

We defined the adjusted R^2_{adj}:

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

Which provides a penalty for increasing the number of parameters, however it does not necessarily help us avoid over-fitting.

Confidence and prediction intervals

We talked about confidence and prediction intervals:

Confidence interval:

$$\hat{\mu}_{Y|x_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0} \leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0}$$

Prediction interval:

$$\hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 (1 + x_0^t (X^t X)^{-1} x_0)} \leq Y_0 \leq \hat{y}_0 + t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 (1 + x_0^t (X^t X)^{-1} x_0)}$$

Let's see this in action, in the following R Shiny app: https://sarid.shinyapps.io/intervals_demo/

Note About Extrapolation - Thought Experiment

What do you think is the problem with trying to provide an extrapolation (fit) and intervals (confidence for mean and prediction for a new observation) for the number of bird strikes with the following parameters:

- Flight height = 22 thousand feet
- Flight speed = 42 kts
- Sky = "No Cloud"
- Number of engines = 2

Can you think of a similar example but from a different domain?

03:00

Example - Outliers' Influence

Another "danger" in linear regression is what happens when the data contains outliers. Linear regression is very sensitive in this sense.

```
# wildlife_impacts <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-07-23/wildlife_impacts.csv")
# write_csv(wildlife_impacts %>% count(height), "lectures/data/wildlife_impacts_small.csv")
wildlife_small <- read_csv("data/wildlife_impacts_small.csv", col_types = cols()) %>%
  mutate(rounded_height = round(height/1000)) %>%
   group_by(rounded_height) %>%
   summarize(n = sum(n)) %>%
   filter(!is.na(rounded_height))

wildlife_err <- wildlife_small
wildlife_err[19, 2] <- 600000 # instead of 6 we multiplied this observation by 1000

p1 <- ggplot(wildlife_small, aes(x = rounded_height, y = log10(n))) +
   geom_point() +
   stat_smooth(method = "lm") + coord_cartesian(ylim = c(-1, 5)) + theme_bw()
p2 <- ggplot(wildlife_err, aes(x = rounded_height, y = log10(n))) +
   geom_point() +
   stat_smooth(method = "lm") + coord_cartesian(ylim = c(-1, 5)) + theme_bw()
cowplot::plot_grid(p1, p2)
```

How are Types of Variables Used in Regression?

As you have probably noticed, in the bird-planes example, we used a `sky` variable which has three values (factor). The regression model is linear, if so, how are factor variables treated?

- Factors are turned into dummy variables (0/1).
  - How many dummies are needed for a 3-level factor? why?
- Characters are treated the same
- Ordinals - depending on definition, might be entered as polynomials, factors, or continuous
- Logicals - as a 0/1 variable

Questions (hint: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$):

- What is the meaning of the coefficient β of a logical variable?
- What is the meaning of the coefficient β of a factor?
- How would you consider a date type variable?

03:00

Variable Selection via Stepwise Regression

In cases we have many independent variables (i.e., X_is) we want to reduce the complexity of the regression model, and focus only on the most important or most influential variables.

For that, we can use the **stepwise regression** algorithm.

- Backward elimination
- Forward selection
- Backward-forward

All algorithms are greedy algorithms: look one step ahead, and choose the best course of action. Hence, the might not reach the optimal solution.

The Backward Elimination Algorithm

1. Set current model = the "full" model (all independent variables)
2. Use some accuracy measure to examine the influence of the removal of each of the variables. E.g., the Akaike Information Criteria (AIC = $-2\log(L) + 2p$)
3. Are there a variabled which contributes to the improvement of the measure (e.g. decrease of the AIC)?

   a. **Yes**: Remove the variable which its removal contributes the most (to the improvement of the measure), Update the current model accordingly, and go to step 2.

   b. **No**: Algorithm stops and outputs the current model.

The forward selection

A similar algorithm, instead of removing, add variables one at a time

The backward-forward

Combination, at each iteration check both removal and addition.

Intuition for the AIC

AIC = $-2\log(L) + 2p$

The AIC uses the likelihood L:

$$L = \prod_{i=1}^{n} f(\hat{y}_i)$$

(Higher liklihood value is better, hence lower $-2\log(L)$ is better)

As the number of parameters in the model increases, the model is more prone to overfitting, hence we a penalty of $2p$ (a high p, i.e., a lot of parameters is bad for us).

Hence, we want to minimize the AIC.

The value of AIC has no meaning, except for the ability to compare two models.

Other measures also exist, e.g.:

- Montgomery uses the F-Statistic
- Another measure is the BIC = $-2\log(L) + \log(n)p$

Stepwise Regression - Example: Car Efficiency.

```r
mtcars_lm <- lm(formula = mpg ~ ., data = mtcars)
step(mtcars_lm, direction = "backward")
```

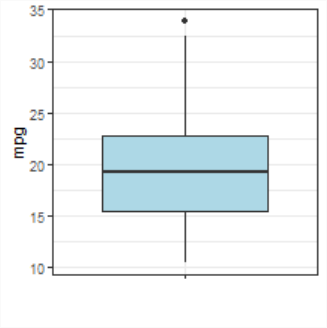
```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
```

Refresher and Recap - Test-like Exercises: Descriptive Statistics

Now we are going to solve four exercises which might resemble exercises which will appear in the test.

Explain the different elements of a boxplot, i.e., in the following chart:

1. What does the line in the middle of the box (the blue area) stands for?
2. What do the top and bottom boundaries of the chart's box stand for?
3. What are the whiskers and how do they help us?
4. What does the dot at the upper side of the chart stands for?
5. How would you use a boxplot to recognize a normal distribution?



Test-like Exercises: Confidence and Prediction Intervals

The following exercise is similar to Walpole, *et al.* (Chapter 9, page 245, ex. 7).

A random sample of $n = 100$ car owners show that in the state of Virginia, a car drives 23,500 km per year with a sample standard deviation of 3900 km.

a. Construct a 99% confidence interval for the average number of kilometers a car is driven annually in Virginia.

b. What can we assert with a 99% confidence interval about the possible size of our error if we estimate the average number of kilometers driven by car owners in Virgina to be 23,500 km?

c. Danny has just moved to Virginia, provide an upper bound (one sided prediction interval) at a 95% confidence interval to the amount of kilometers Danny will drive in the next year.

Test-like Exercises: Two Sample Tests, p-value

The following exercise is from Walpole, *et al.*, (chapter 9, page 319, ex. 9).

A study at the University of Colorado shows that running increases the percent resting metabolic rate (RMR) in older women. The average RMR of 30 elderly women runners was 34.0% higher than the average RMR of 30 sedentary elderly women. The standard deviations were 10.5 and 10.2 respectively.

Was there a significant increase in RMR of the women runnders over the sedentary women?

Assume the populations are normally distributed with equal variances. Use a p-value to report you conclusions.

Test-like Exercises: Interpretation of a Model's Output

The `ChickWeight` dataset contains the results of a feeding experiment of 50 chicks' (`Chick`) with their tracked weight (`weight`), over a period of 21 days (`Time`), each chick was subjected to a different type of diet (`Diet`).

In the following model (see bottom of slide), we are using the interaction of `Time*factor(Diet)` as one of the explanatory variables, along with `Time` as another explanatory variable. The dependent variable is the chick's `weight`.

Questions:

1. The original `Diet` variable is numeric. Why are we using it in the regression model as `factor(Diet)`?
2. How many levels does the `factor(Diet)` variable has, explain.
3. Why do we need the interaction of `Time*factor(Diet)` in the model? (why is `weight ~ Time + factor(Diet)` not enough)
4. Which dietary method helps increase the chick's weight the most? Explain how you deduced this from the model's output.

See the next slide for an additional question.

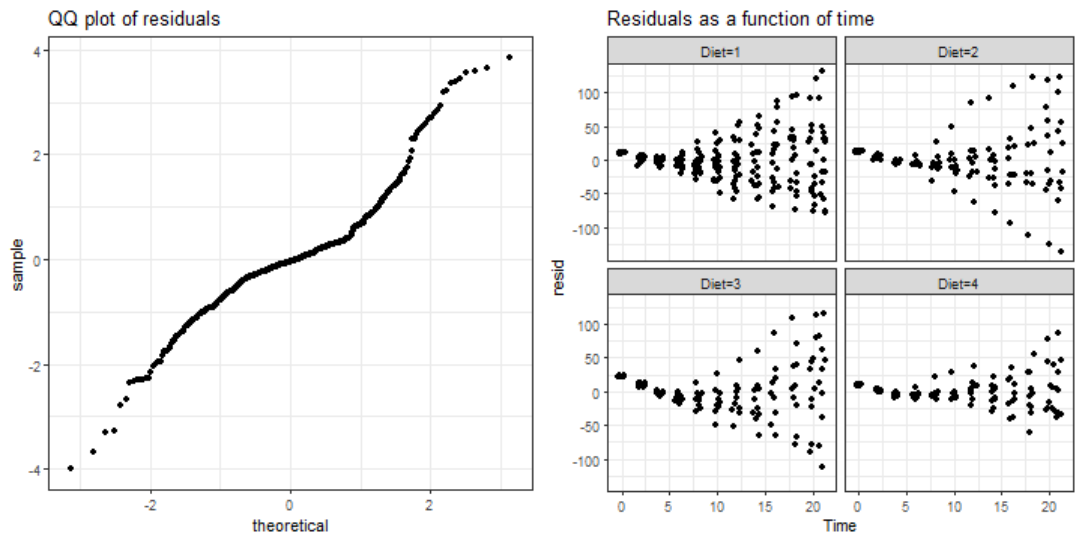
```
chick_lm <- lm(formula = weight ~ Time + Time*factor(Diet), data = ChickWeight)
summary(chick_lm)
```

```
##
## Call:
## lm(formula = weight ~ Time + Time * factor(Diet), data = ChickWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -135.425  -13.757   -1.311   11.069  130.391
##
```

Test-like Exercises: Interpretation of a Model's Output - continued

Look at the following qqplot of residuals and plot of the residuals as a function of time.

1. Explain what are the underlying assumptions of the linear regression model.
2. Observing the residuals' plots below, would you say that any of the linear regressions assumptions are violated? which one?



Experiment Design - Motivation

With methods of hypothesis testing, we were able to discern the differences between two groups (i.e., unpaird two-sample test).

For example trying to see if cars with 4 cylinders are more efficient than cars with 8 cylinders.

```
mtcars %>%
   filter(cyl != 6) %>%
   t.test(formula = mpg ~ cyl, data = .)
```

```
##
##     Welch Two Sample t-test
##
## data:  mpg by cyl
## t = 7.5967, df = 14.967, p-value = 1.641e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.318518 14.808755
## sample estimates:
## mean in group 4 mean in group 8
##        26.66364        15.10000
```

Experiment Design - Motivation (2)

Sometimes, we have more than two levels, in which case, we would like to devise a test which will compare all levels.

In the case of a single variable with multiple levels, this is called *single factor experiments* (in the next lecture we will also discuss *multi factor experiments*, when there are multiple levels).

The oricess invlove:

- Conjecture - The hypothesis that motivates the experiment
- Experiment - The actual test performed to investigate the conjecture
- Analysis - Statistical analysis of the collected data
- Conclusions - What has been learned

The process is iterated: to improve the experiment (e.g., add new variables or change the methods) and learn more.

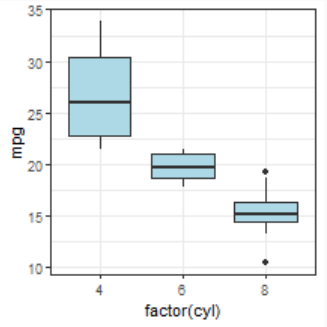
The following material is covered in Montgomery chapter 13.

Experiment Design - Motivation (3)

We would like a statistical that would highlight the efficiency of cars as a function of the number of cylinders.

A boxplot can visually illustrate what we are looking for, but we cannot yield strength or significance from it.

```
ggplot(mtcars, aes(y = mpg, x = factor(cyl))) +
   geom_boxplot(fill = "lightblue") +
   theme_bw()
```



The Completely Randomized Single-Factor Experiment

Each factor level is called a *treatment* (i.e., for different treatments administered in an experiment).

The experimenter randomely samples subjects (either of equally sized groups or varying sized groups).

We describe the observations for a **completely randomized design** by a linear statistical model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, \; j = 1, \ldots n$$

The value $\mu_i = \mu + \tau_i$ is the mean value of the ith treatment.

We assume that the errors $\epsilon_{i,j}$ are normally and independently distributed $\mathcal{N}(0, \sigma^2)$.

Two methods to select the treatments:

- **Fixed-effects model**: the a treatments were chosen specifically.
- **Random-effects model** (components of variance): the a treatments were a random sample from a larger population of treatments. We would like to extend the conclusions for additional treatments.

For now, we are going to focus on the **fixed effects model**.

The Fixed-Effects Model

The treatments τ_i are defined as deviations from the overall mean μ such that:

$$\sum_{i=1}^{a} \tau_i = 0$$

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij}, \quad \bar{y}_{i\cdot} = y_{i\cdot}/n$$

$$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}, \quad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N, \quad N = a \cdot n$$

We are interested in the following test:

- $H_0 : \tau_1 = \ldots = \tau_a = 0$
- $H_1 : \exists i \mid \tau_i \neq 0$

The Sum of Squares

Again, we use the sum of squares equation: $SS_T = SS_{\text{Treatments}} + SS_E$:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

The expected value of each of the errors can be computed (out of scope):

$$E(SS_{\text{Treatments}}) = (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2$$

$$E(SS_E) = a(n-1)\sigma^2$$

The degrees of freedom of each error are:

- $an - 1$ for SS_T,
- $a - 1$ for $SS_{\text{Treatments}}$
- $a(n-1)$ for SS_E

The ANOVA Test

If the null hypothesis holds, then $\tau_1 = \ldots = \tau_a = 0$, and

$MS_{\text{Treatments}} = SS_{\text{Treatments}}/(a-1)$ is an unbiased estimator of σ^2

The error mean square $MS_E = SS_E/[a(n-1)]$ is also an unbiased estimator of σ^2

Hence the following statistic has an F-distribution

$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/[a(n-1)]} = \frac{MS_{\text{Treatments}}}{MS_E}$$

With $a-1$ and $a(n-1)$ degrees of freedom.

We use an upper-tail, one-sided critical region and reject H_0 if $f_0 > f_{\alpha,a-1,a(n-1)}$.

The ANOVA Table

The corresponding ANOVA table, for a Single-Factor Experiment, with a Fixed-Effects Model:

Source of Variation	**Sum of Squares**	**df**	**Mean Squares**	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n-1)$	MS_E	
Total	SS_T	$an-1$		

For an unbalanced experiment (each treatment has varying group size, n_i) we would have:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SS_{\text{Treatments}} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

And $SS_E = SS_T - SS_{\text{Treatment}}$

ANOVA Example

```
mtcars2 <- mtcars %>%
   mutate(cyl_fct = factor(cyl))
mtcars_anova <- aov(formula = mpg ~ cyl_fct, data = mtcars2)
summary(mtcars_anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl_fct      2  824.8   412.4    39.7 4.98e-09 ***
## Residuals   29  301.3    10.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Usually, ANOVA will be followed by a multiple comparisons procedures, that will help identify which factors contribute to the variation.

There are many such multiple comparison tests. See this book: Bretz F., Hothorn T., and Westfall P., Multiple Comparisons Using R, CRC Press, 2011.

Dunnett's Test

The one-sided Dunnett's test takes the minimum (or maximum depending on direction) of the a pairwise t tests:

$$t_i = \frac{\bar{y}_{i\cdot} - \bar{y}_0}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_0}}}$$

Where y_0 relates to the control group (to which we are comparing the treatments), and s^2 is the pooled variance, i.e.:

$$s^2 = \sum_{i=0}^{m} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_{i\cdot})^2 / v$$

With $v = \sum_{i=0}^{m} n_i - (a+1)$ degrees of freedom.

The Dunnett's test is availble in `multcomp` like many other procedures for multiple comparisons.

Dunnett's Test Example

```r
suppressWarnings(suppressMessages(library(multcomp)))
glht(mtcars_anova,
     linfct = mcp(cyl_fct = "Dunnett"),
     alternative = "less") %>%
   summary()
```

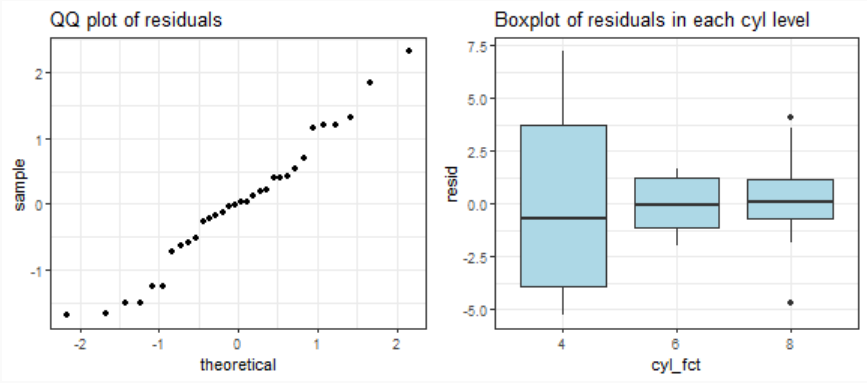
```
##
##      Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = mpg ~ cyl_fct, data = mtcars2)
##
## Linear Hypotheses:
##            Estimate Std. Error t value   Pr(<t)
## 6 - 4 >= 0   -6.921      1.558  -4.441 0.000117 ***
## 8 - 4 >= 0  -11.564      1.299  -8.905 8.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Verifying ANOVA assumptions

The ANOVA model assumes that observations are normally and independently distributed, with the same variance for each treatment.

This can be verified using hypothesis tests on the residuals, or viewing a proper plot (e.g., qqplot). The following illustrates that some assumptions are invalid in the previous example (which?)

```
mtcars2_resid <- mtcars2 %>%
   mutate(resid = mtcars_anova$residuals)
boxplot_resid <- ggplot(mtcars2_resid, aes(x = cyl_fct, y = resid)) +
   geom_boxplot(fill = "lightblue") +
   theme_bw() + ggtitle("Boxplot of residuals in each cyl level")
qqplot_resid <- ggplot(mtcars2_resid, aes(sample = (resid - mean(resid))/sd(resid))) +
   geom_qq() +
   theme_bw() + ggtitle("QQ plot of residuals")
cowplot::plot_grid(qqplot_resid, boxplot_resid)
```



Determining the Sample Size

The selection of the sample size is based on the difference we want to detect, and at what test power:

$$1 - \beta = P(\text{Reject } H_0 | H_1) = P(F_0 > f_{1-\alpha, a-1, a(n-1)} | H_1)$$

The effect size represents the differences in means between groups:

$$\text{ES} = \frac{\mu_{\text{experiment}} - \mu_{\text{control}}}{s}$$

```
pwr::pwr.anova.test(k = 3, n = NULL, f = 0.25, sig.level = 0.05, power = 0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 52.3966
##               f = 0.25
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```