Overview:

Introduction

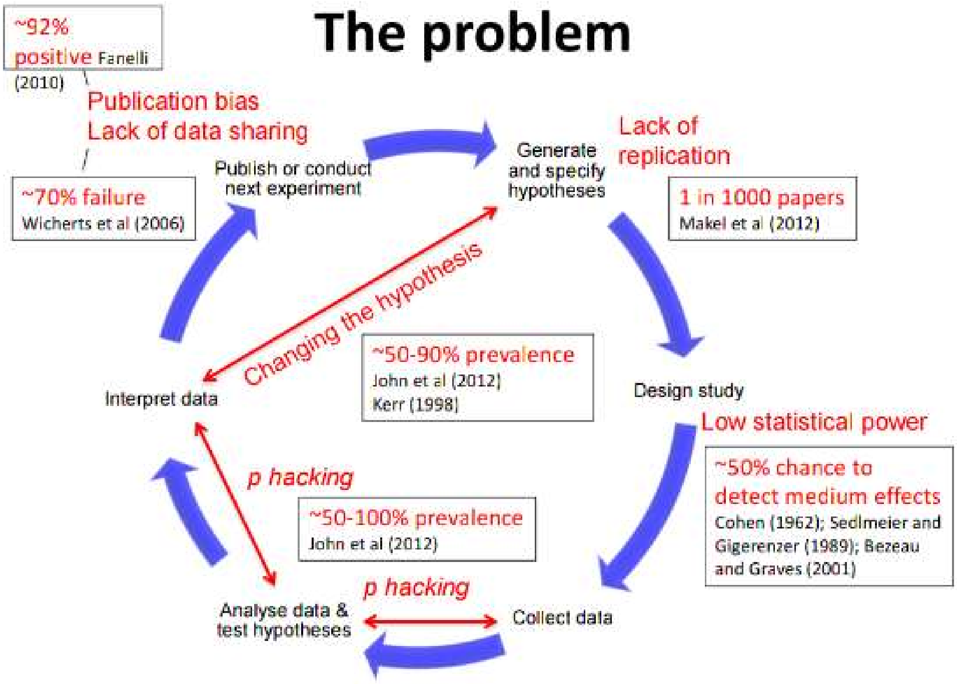
How do we systematically investigate the process or variables that affect outcome of the process, such as product quality? Design of experiments (also known as experiment design or experimental design) is a field of study which looks into this type of investigations through planning and design to ensure that the data collected from the experiment will be sufficient and erroneous conclusions can be prevented.

The following video gives an introduction on design of experiments for data science.

Data Scientist's Toolbox: Experimental Design

https://youtu.be/vSXOJnGNtM4

Source: (The John Hopkins Data Science Lab, 2016)



The problem
~92% positive Fanelli (2010)
Publication bias
Lack of data sharing
~70% failure Wicherts et al (2006)
Publish or conduct next experiment
Generate and specify hypotheses
Lack of replication
1 in 1000 papers Makel et al (2012)
Changing the hypothesis
~50-90% prevalence John et al (2012) Kerr (1998)
Interpret data
Design study
Low statistical power
~50% chance to detect medium effects Cohen (1962); Sedlmeier and Gigerenzer (1989); Bezeau and Graves (2001)
p hacking
~50-100% prevalence John et al (2012)
p hacking
Analyse data & test hypotheses
Collect data

Source: Data Science for Experimental Design, (Alex Morley, 2018)

Design of experiments is important for a systematic generation of data when the effect of noisy factors has to be identified. In the data generation process, it is inevitable for variation to exist hence controlled and well-designed experiments are fundamental for robust process engineering to produce reliable products. Whilst even controllable factors contain a certain amount of uncontrollable variation that affects the response and some factors cannot be controlled at all (such as environmental factors), care should be exercised to minimize and control the effect of such noisy influencing factors through design of experiments.

Design of experiments can be utilized for:

- Systematic generation of new data
- Systematic reduction of data bases
- Optimization and tuning of algorithm parameters for improving data analysis and modelling

Experimental design is crucial for reliability, validity and replicability of results.

6.1 Analysis of Variance (ANOVA)

In some data science problems (such an comparison between different COVID-19 treatments), we need to perform a comparison of multiple groups with numeric data. The statistical procedure that tests for a statistically significant difference for a certain measure of interest among the groups is called analysis of variance, or ANOVA. For example, one may be interested to determine if there is a significant difference in the mean salary for different population groups. The population groups are defined based on objectives of analysis - some possibilities include categorization based on industry type or education level. This is an extension of the two-sample t-test which we have seen in Week 2. A naïve approach would be to make pairwise (pair-by-pair) comparisons for all the groups involved but this would inflate the error. ANOVA is the statistical procedure to conduct a single overall test (also known as omnibus test) for multi-group comparisons. The following video (up to 2:11) gives a brief introduction on ANOVA.

What is an ANOVA?

https://www.youtube.com/watch?v=uzcqMeNK7Kw

Source: (U of G Library, 2019)

Within the ANOVA framework, explanatory (or independent) categorical variables are referred to as factors and the number of groups within each variable are called levels. In this topic, we consider one-way ANOVA, which implies that only one factor (i.e. explanatory variable) is involved.

Example: We are interested in testing for significance difference in average household income between states in Malaysia. In this problem, the numeric (response) variable is household income and state is the independent variable or factor. Since Malaysia comprises 13 states and 3 federal territories, potentially there are 16 groups or levels within the variable.

Before we proceed to interpretation of ANOVA results, we need to know that the following assumptions underlying ANOVA:

- Population distributions are normal
- Samples have equal variances
- Observations are independent

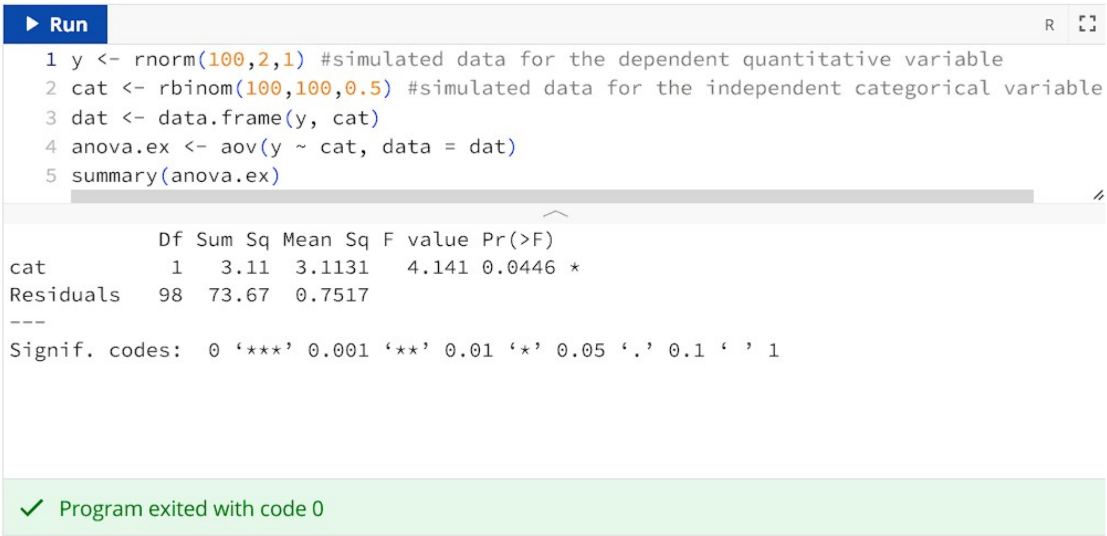
If the assumptions are not met, non-parametric alternatives such as the Kruskal-Wallis test may be more appropriate.

In a one-way ANOVA in which there are k levels within the factor, the hypothesis takes the form:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1: \text{At least one of the group means differ}$$

The statistical test for ANOVA is based on the F-statistic, which is based on the ratio of the variance across group means to the variance due to residual error. The higher this ratio, the more statistically significant the result. If the data follows a normal distribution, the statistic has an *F-distribution* thus its *p-value* can be computed. In general, if the p-value is less than 0.05, then the null hypothesis is rejected and we can conclude that at least one of the group means are not equal.

In R, we can perform ANOVA by using the aov() function. To illustrate the use of this function in R, we use a dummy data set simulated for 100 observations on two variables: the dependent quantitative variable y and independent categorical variable cat.



```r
y <- rnorm(100,2,1) #simulated data for the dependent quantitative variable
cat <- rbinom(100,100,0.5) #simulated data for the independent categorical variable
dat <- data.frame(y, cat)
anova.ex <- aov(y ~ cat, data = dat)
summary(anova.ex)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)  
cat          1   3.11  3.1131   4.141 0.0446 *
Residuals   98  73.67  0.7517                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

✓ Program exited with code 0

From the results output for the example with simulated data above, the F-test statistic is 0.413 with corresponding p-value 0.522. At the 5% significance level, there is not sufficient evidence to conclude that there is a significant difference between the groups (or levels) in the categorical variable.

The F test statistic value is read from the F value output and its corresponding p-value is read from Pr(>F) in the R output. These values may differ from your run of the same code above due to different values simulated for y and cat.

6.1.1 Post-hoc test

If the null hypothesis is rejected in ANOVA, i.e. we find that there is a statistically significant difference between the groups, we would then proceed to find out where the difference is. In other words, we conduct post-hoc multiple comparison tests to determine which groups are different from each other.

In a nutshell, post-hoc multiple comparison tests are t-tests which examine mean differences between the groups in ANOVA. There are several types of post-hoc tests that will control for Type I error, including:

- Tukey's multiple comparison test
- Fisher's Least Significant Difference (LSD) method

- Bonferroni's correction method
- Scheffe test's procedure
- Dunnett's multiple comparison test

In this course, we will explore further on the Tukey's and Dunnett's multiple comparison tests. The interested student may explore other tests listed above at your own pace.

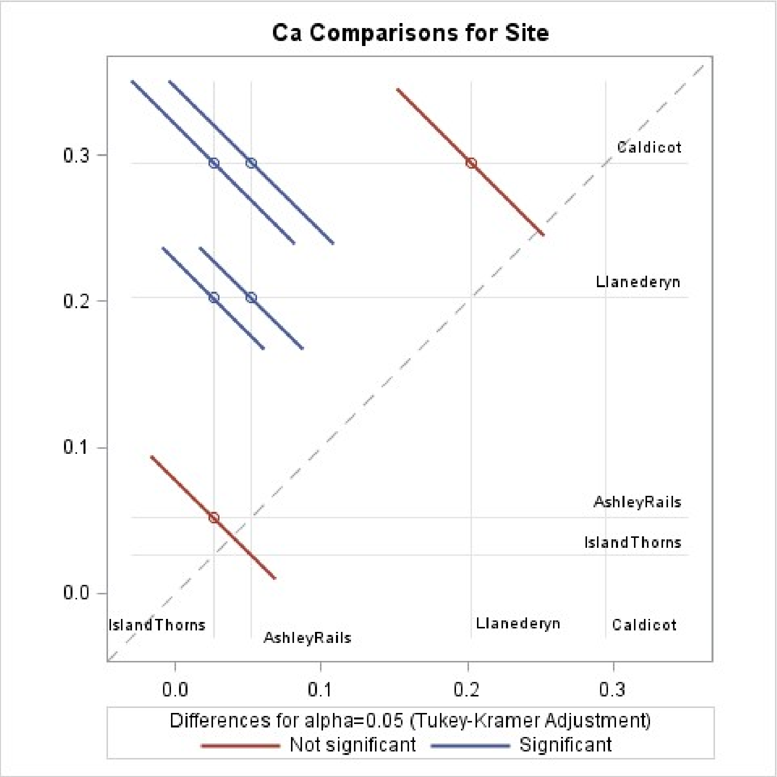
Tukey's Multiple Comparison Test

Tukey's multiple comparison method is also known as the honestly significant difference (HSD) test. This method is appropriate when considering pairwise comparisons. The experimentwise error rate is:

- equal to α when all pairwise comparisons (all possible combinations of two treatment means) are considered
- less than α when fewer than all pairwise comparisons are considered

The test works by constructing confidence intervals for the differences between all possible pairs of population means. If the difference between the two means in a pair does not contain 0, these two means are said to be significantly different.

A diffogram is often useful in interpreting Tukey's multiple comparison test results.



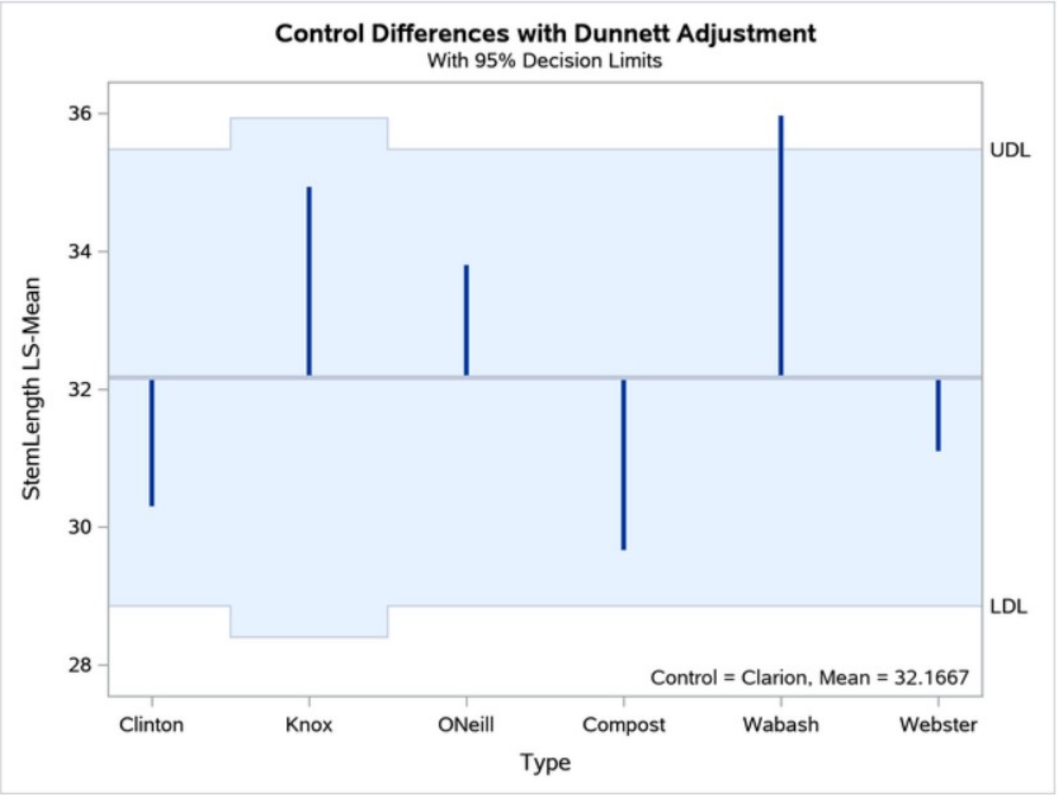
Source: The diffogram and other graphs for multiple comparisons of means, (Rick Wicklin, 2017)

In the diffogram above, the pairs which are represented by blue lines are those which do not cross the diagonal threshold. These pairs are said to have statistically significant differences, e.g. IslandThorns and Llanederyn.

Dunnett's Multiple Comparison Test

Dunnett's multiple comparison method is used when comparing to a (true) control group, such as a placebo group. The experiment-wise error rate is less than the stated α and the test takes into account the correlations among tests. This method computes and tests k - 1 groupwise differences, where k is the number of levels of the classification variable.
LS-means control plots can be used to graphically present the results of Dunnett's multiple comparison method.

Figure 49.31: LS-Means Plot of Differences against a Control



Source: Graphics for LS-Mean Comparisons, (SAS Institute, 2019)

In the control plot above, the group (or level) corresponding to the line which exceeds the shaded area is the group (Wabash) which is said to have a statistically significant difference from the control group (Clarion).

Question 1

Which one of the following is useful for making pairwise-comparisons with a control group?

ANS: Dunnett's Multiple Comparison Test

Question 2

Post-hoc multiple comparison tests are for: (select all that applies)

ANS:

- examining mean differences between groups for pairwise comparisons
- controlling Type I error in multiple comparison tests in ANOVA

6.1.2 [Discussion Activity] Explore Activity: One-Way ANOVA

In Week 4, we used linear regression analysis to model the response variable SalePrice in AmesHouseNormal.csv data set. Of particular interest in this topic is if there is a significant difference in the mean sale price for certain categorical variable(s), such as LandContour (land contour) or BldgType (building type). Before conducting a one-way analysis of variance (ANOVA), we can examine the distribution of the data points first to determine if an ANOVA is worth pursuing. This can be done through appropriate visualization such as box-and-whisker plots which show the data distribution by groups.

Box-and-whisker plots or boxplots were discussed in Week 1.

After examining the distribution and characteristics of the data set, we may then consider conducting ANOVA.

Time: 30 minutes

Purpose: The purpose of this activity is to perform one-way ANOVA using R.

Task: Based on the box-and-whisker plots or any other suitable descriptive statistics, identify and discuss one categorical variable of interest to determine if there is a significant difference in the mean sale price for different levels of the variable. If the difference is statistically significant, perform a post-hoc test.

6.2 Factorial ANOVA

Introduction

In the preceding topic, we discussed one-way ANOVA whereby we have only one factor (group) that is varying. In real life scenarios, we could have a second factor involved with data collected on each combination. In this case, we have two-way ANOVA. Two-way ANOVA is similar to one-way ANOVA but with consideration towards the interaction effect. Interaction effects indicate presence of differences that are not uniform across all combinations of the independent variables. Interaction is discussed in the next subtopic.

There can of course be more than two independent variables involved in the analysis, and this results in n-way ANOVA (with n indicating the number of independent variables). This is also known as factorial ANOVA.

Question

In the previous topic, we attempted to find out if there is a significant difference in the mean sale price of houses for an identified categorical (or independent) variable in the AmesHouseNormal.csv data set. Name one other explanatory (independent) variable that could be used to explain house sale price.

6.2.1 Interaction

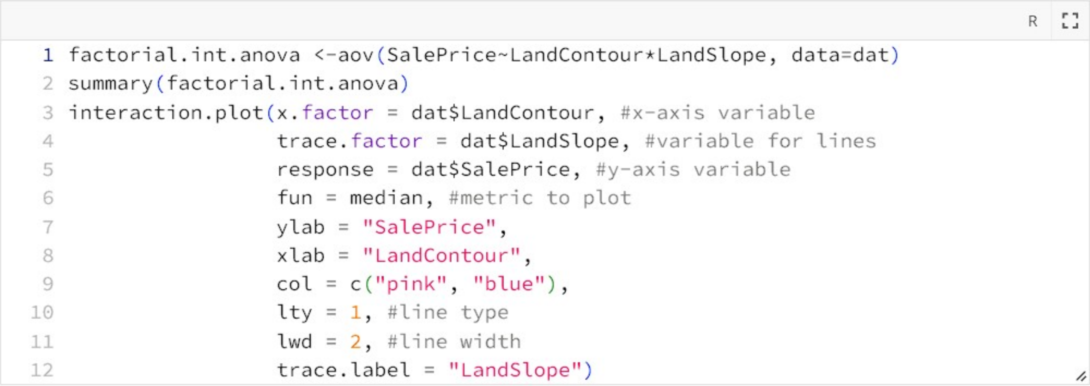
When two or more explanatory variables are involved in a statistical model or analysis, interaction is a statistical concept which needs to be considered.

Interaction is said to exist when the effect of a change in the level (for categorical variable) or value (for quantitative variable) of one explanatory variable on the mean outcome depends on the level or value of another explanatory variable. The lack of interaction is called additivity.

An interaction variable is created as the product of two (or more) explanatory variables, denoted as A*B in some software programs or textbooks for explanatory variables A and B. Interaction variables can be created manually or automatically, depending on the program.

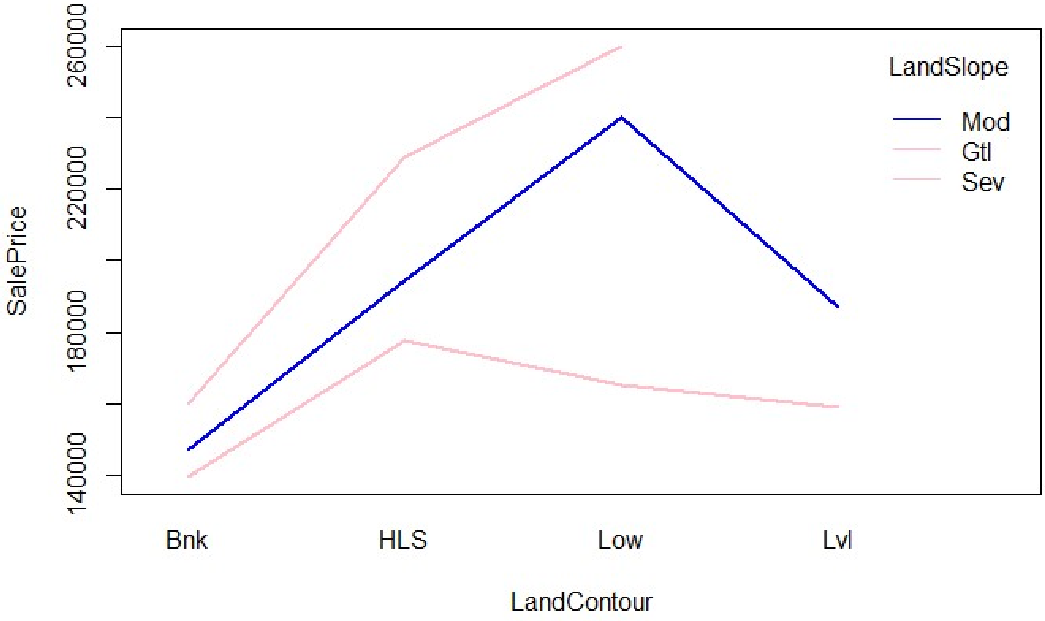
A *profile* plot, also known as an interaction plot, can be used to look at outcome means for two factors simultaneously. A plot drawn with parallel lines suggests an additive model, while non-parallel lines suggests an interaction model.

As an example, the following R code produces a two-way ANOVA for LandContour and LandSlope as the independent variables for SalePrice in the AmesHouseNormal.csv data set. The results imply that the interaction effect is not significant (p-value = 0.7127).



```
factorial.int.anova <-aov(SalePrice~LandContour*LandSlope, data=dat)
summary(factorial.int.anova)
interaction.plot(x.factor = dat$LandContour, #x-axis variable
                 trace.factor = dat$LandSlope, #variable for lines
                 response = dat$SalePrice, #y-axis variable
                 fun = median, #metric to plot
                 ylab = "SalePrice",
                 xlab = "LandContour",
                 col = c("pink", "blue"),
                 lty = 1, #line type
                 lwd = 2, #line width
                 trace.label = "LandSlope")
```

The following figure is an interaction plot created.



6.2.2 [Discussion Activity] Application Activity: Factorial ANOVA

This activity is a continuation of the ANOVA in Discussion Activity 6.1.2 using the AmesHouseNormal.csv data set.

Time: 30 minutes

Purpose: The purpose of this activity is to perform factorial ANOVA using R.

Task: Based on the box-and-whisker plots or any other suitable descriptive statistics, identify and discuss another one or more categorical variable(s) of interest to determine if there is a significant difference in the mean sale price for different levels of the variable(s). Discuss the presence of interaction effects, if any. Refer to the code in 6.2.1 to perform factorial ANOVA with interaction effects and produce the interaction plot.

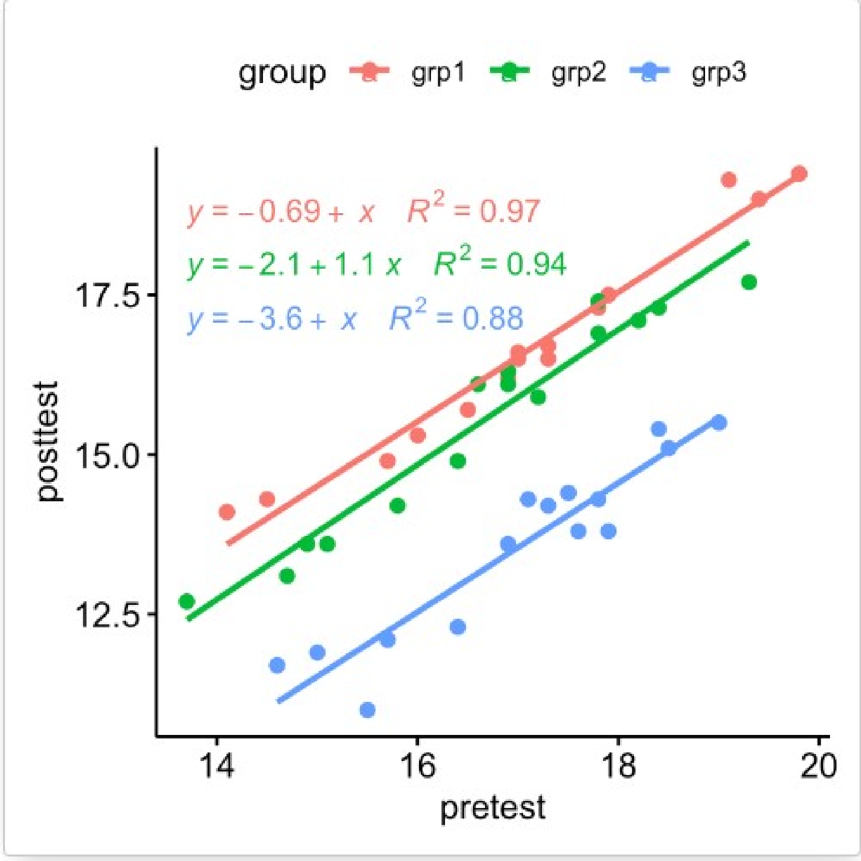
6.3 Analysis of Covariance (ANCOVA)

Besides categorical or qualitative independent variables, quantitative variables can be included into the analysis and this is known as analysis of covariance (ANCOVA). Interpretation of ANCOVA results is slightly different from ANOVA results due to the nature of the independent variables involved.

Analysis of Covariance (ANCOVA) easily explained

https://www.youtube.com/watch?v=a61mkzQRf6c

Source: (Working On My Aura, 2017)



Source: ANCOVA in R, (Data Novia, n.d.)

Application Activity: Analysis of Covariance

This activity is a continuation of the ANOVA in Discussion Activity 6.1.2 using the AmesHouseNormal.csv data set.

Time: 30 minutes

Purpose: The purpose of this activity is to perform ANCOVA using R.

Task: In addition to the categorical variable from Discussion Activity 6.1.2, identify and discuss one or more quantitative variable(s) of interest to determine if there is a significant difference in the mean sale price. Discuss the presence of interaction effects, if any.

The R functions for this task is the same as 6.2.2.