Module 1 - Peer reviewed

Outline:

In this homework assignment, there are four objectives.

1. To assess your knowledge of ANOVA/ANCOVA models
2. To apply your understanding of these models to a real-world datasets

General tips:

1. Read the questions carefully to understand what is being asked.
2. This work will be reviewed by another human, so make sure that you are clear and concise in what you are attempting to explain or answer.

In [1]:
```
# Load Required Packages
library(tidyverse)
library(ggplot2)
library(dplyr)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

✓ ggplot2 3.3.0     ✓ purrr   0.3.4
✓ tibble  3.0.1     ✓ dplyr   0.8.5
✓ tidyr   1.0.2     ✓ stringr 1.4.0
✓ readr   1.3.1     ✓ forcats 0.5.0

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Problem #1: Simulate ANCOVA Interactions

In this problem, we will work up to analyzing the following model to show how interaction terms work in an ANCOVA model.

$$Y_i = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon_i$$

This question is designed to enrich understanding of interactions in ANCOVA models. There is no additional coding required for this question, however we recommend messing around with the coefficents and plot as you see fit. Ultimately, this problem is graded based on written responses to questions asked in part **(a)** and **(b)**.

To demonstrate how interaction terms work in an ANCOVA model, let's generate some data. First, we consider the model

$$Y_i = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon_i$$

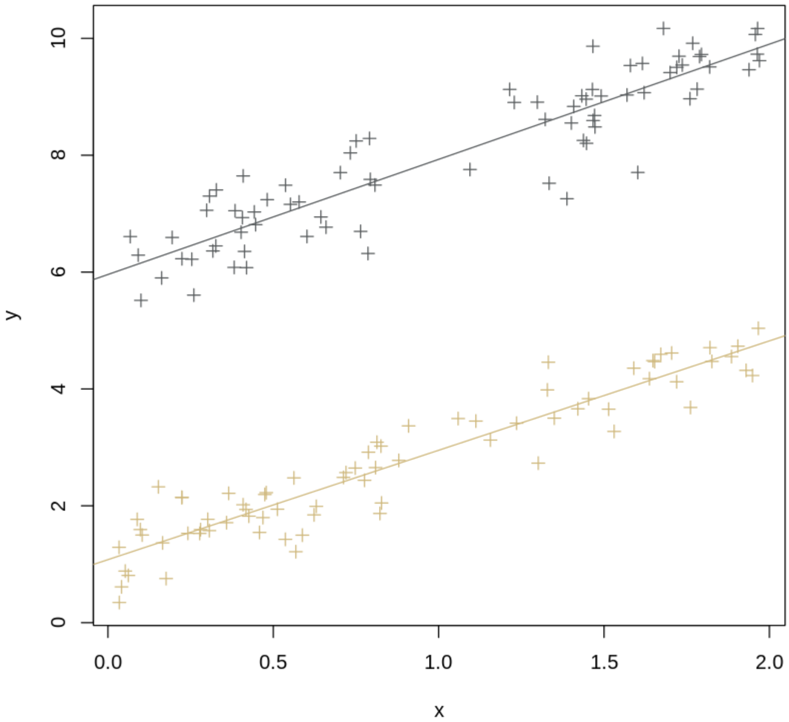
where X is a continuous covariate, Z is a dummy variable coding the levels of a two level factor, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We choose values for the parameters below (b0,...,b2).

In [2]:

```
rm(list = ls())
set.seed(99)

#simulate data
n = 150
# choose these betas
b0 = 1; b1 = 2; b2 = 5; eps = rnorm(n, 0, 0.5);
x = runif(n,0,2); z = runif(n,-2,2);
z = ifelse(z > 0,1,0);
# create the model:
y = b0 + b1*x + b2*z + eps
df = data.frame(x = x,z = as.factor(z),y = y)
head(df)

#plot separate regression lines
with(df, plot(x,y, pch = 3, col = c("#CFB87C","#565A5C")[z]))
abline(coef(lm(y[z == 0] ~ x[z == 0], data = df)), col = "#CFB87C")
abline(coef(lm(y[z == 1] ~ x[z == 1], data = df)), col = "#565A5C")
```

A data.frame: 6 × 3

	x	z	y
	<dbl>	<fct>	<dbl>
1	0.09159879	1	6.290179
2	1.96439135	1	10.168612
3	0.57805656	1	7.200027
4	0.03370108	0	1.289331
5	1.82614045	0	4.470862
6	0.71220319	0	2.485743



1. (a) What happens with the slope and intercept of each of these lines?

In this case, we can think about having two separate regression lines--one for Y against X when the unit is in group $Z = 0$ and another for Y against X when the unit is in group $Z = 1$. What do we notice about the slope of each of these lines?

In this ANCOVA model without an interaction term (XZ), the covariate X affects the outcome Y consistently across the levels of Z. The intercept shift between the lines for different Z levels indicates the main effect of the factor Z, while the slopes indicate the effect of the continuous covariate X on Y.

1. (b) Now, let's add the interaction term (let $\beta_3 = 3$). What happens to the slopes of each line now?

The model now is of the form:

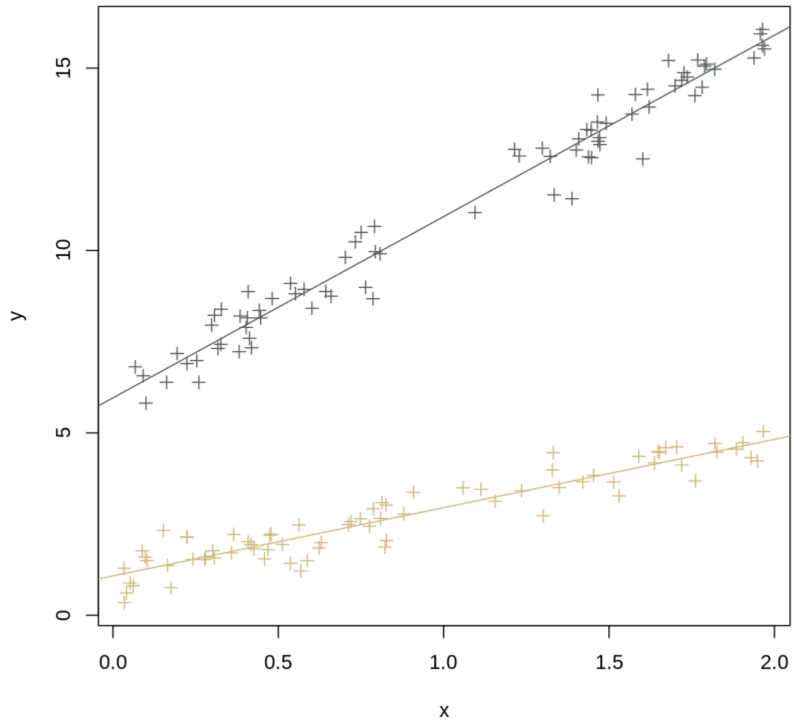
$$Y_i = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon_i$$

where X is a continuous covariate, Z is a dummy variable coding the levels of a two level factor, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We choose values for the parameters below (b0,...,b3).

In [3]:
```
#simulate data
set.seed(99)
n = 150
# pick the betas
b0 = 1; b1 = 2; b2 = 5; b3 = 3; eps = rnorm(n, 0, 0.5);

#create the model
y = b0 + b1*x + b2*z + b3*(x*z) + eps
df = data.frame(x = x,z = as.factor(z),y = y)
head(df)

lmod = lm(y ~ x + z, data = df)
lmodz0 = lm(y[z == 0] ~ x[z == 0], data = df)
lmodz1 = lm(y[z == 1] ~ x[z == 1], data = df)
# summary(lmod)
# summary(lmodz0)
# summary(lmodz1)

# lmodInt = lm(y ~ x + z + x*z, data = df)
# summary(lmodInt)

#plot separate regression lines
with(df, plot(x,y, pch = 3, col = c("#CFB87C","#565A5C")[z]))
abline(coef(lm(y[z == 0] ~ x[z == 0], data = df)), col = "#CFB87C")
abline(coef(lm(y[z == 1] ~ x[z == 1], data = df)), col = "#565A5C")
```

A data.frame: 6 × 3

	x	z	y
	<dbl>	<fct>	<dbl>
1	0.09159879	1	6.564975
2	1.96439135	1	16.061786
3	0.57805656	1	8.934197
4	0.03370108	0	1.289331
5	1.82614045	0	4.470862
6	0.71220319	0	2.485743



In this case, we can think about having two separate regression lines--one for Y against X when the unit is in group $Z = 0$ and another for Y against X when the unit is in group $Z = 1$. **What do you notice about the slope of each of these lines?**

In this ANCOVA model with an interaction term (XZ), both the covariate X and the interaction between X and Z affect the outcome Y differently across the levels of Z. The slope for Z=1 is greater than the slope for Z=0, indicating that the effect of X on Y is stronger when Z=1. The intercepts indicate the main effect of the factor Z, while the slopes indicate the effect of the continuous covariate X on Y and how it changes with Z.

Problem #2

In this question, we ask you to analyze the `mtcars` dataset. The goal if this question will be to try to explain the variability in miles per gallon (mpg) using transmission type (am), while adjusting for horsepower (hp).

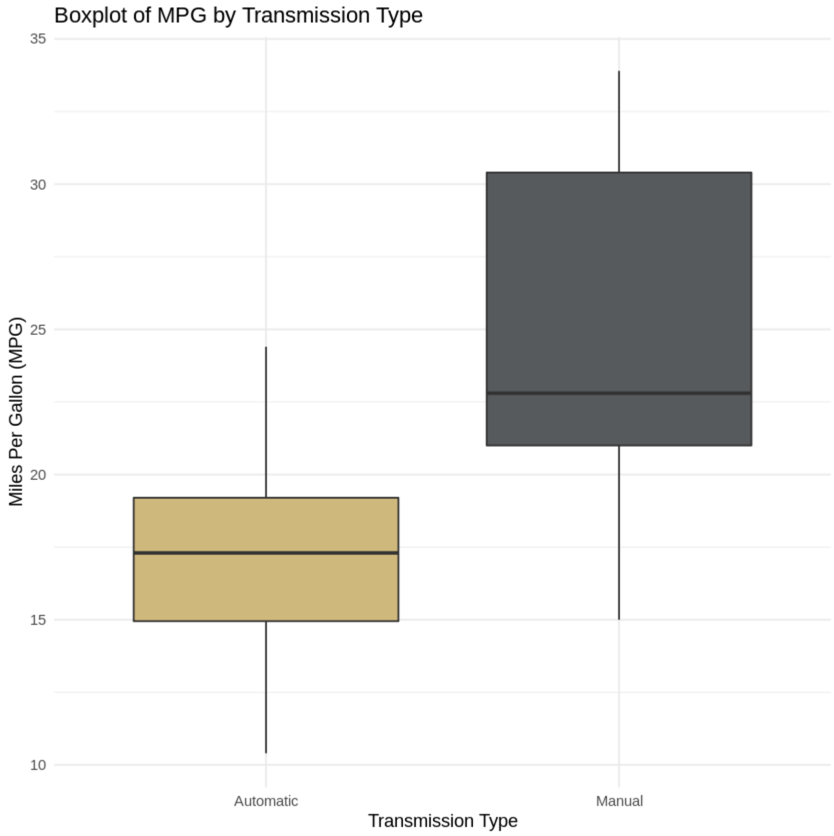
To load the data, use `data(mtcars)`

2. (a) Rename the levels of am from 0 and 1 to "Automatic" and "Manual" (one option for this is to use the revalue() function in the plyr package). Then, create a boxplot (or violin plot) of mpg against am. What do you notice? Comment on the plot

In [4]:

```
library(plyr)
library(dplyr)

# Load the mtcars dataset
data(mtcars)

# Rename levels of am
mtcars$am <- revalue(as.factor(mtcars$am), c("0" = "Automatic", "1" = "Manual"))

# Create a boxplot of mpg against am
ggplot(mtcars, aes(x = am, y = mpg)) +
  geom_boxplot(fill = c("#CFB87C", "#565A5C")) +
  labs(title = "Boxplot of MPG by Transmission Type",
       x = "Transmission Type",
       y = "Miles Per Gallon (MPG)") +
  theme_minimal()


# your code here
```

```
------------------------------------------------------------------------------

You have loaded plyr after dplyr - this is likely to cause problems.
If you need functions from both plyr and dplyr, please load plyr first, then dpl
yr:
library(plyr); library(dplyr)

------------------------------------------------------------------------------


Attaching package: ‘plyr’


The following objects are masked from ‘package:dplyr’:

    arrange, count, desc, failwith, id, mutate, rename, summarise,
    summarize


The following object is masked from ‘package:purrr’:

    compact
```



From the boxplot, we can observe that cars with manual transmission generally have higher mpg compared to cars with automatic transmission. This suggests that manual transmission may be associated with better fuel efficiency.

2. (b) Calculate the mean difference in mpg for the Automatic group compared to the Manual group.

In [5]:

```
# Calculate mean mpg for each transmission type
mean_mpg_auto <- mean(mtcars$mpg[mtcars$am == "Automatic"])
mean_mpg_manual <- mean(mtcars$mpg[mtcars$am == "Manual"])

# Calculate the mean difference
mean_diff <- mean_mpg_manual - mean_mpg_auto
mean_diff
```

7.24493927125506

The mean difference in mpg between the Manual and Automatic groups indicates that manual transmission cars have a higher average mpg compared to automatic transmission cars.

2. (c) Construct three models:

1. An ANOVA model that checks for differences in mean mpg across different transmission types.
2. An ANCOVA model that checks for differences in mean mpg across different transmission types, adjusting for horsepower.
3. An ANCOVA model that checks for differences in mean mpg across different transmission types, adjusting for horsepower and for interaction effects between horsepower and transmission type.

Using these three models, determine whether or not the interaction term between transmission type and horsepower is significant.

In [6]:
```
# ANOVA model
anova_model <- aov(mpg ~ am, data = mtcars)
summary(anova_model)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
am           1  405.2   405.2   16.86 0.000285 ***
Residuals   30  720.9    24.0                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In [7]:
```
# ANCOVA model without interaction
ancova_model <- lm(mpg ~ am + hp, data = mtcars)
summary(ancova_model)
```

```
Call:
lm(formula = mpg ~ am + hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3843 -2.2642  0.1366  1.6968  5.8657 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 26.584914   1.425094  18.655  < 2e-16 ***
amManual     5.277085   1.079541   4.888 3.46e-05 ***
hp          -0.058888   0.007857  -7.495 2.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.909 on 29 degrees of freedom
Multiple R-squared:  0.782,	Adjusted R-squared:  0.767 
F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

In [8]:
```
# ANCOVA model with interaction
ancova_interaction_model <- lm(mpg ~ am * hp, data = mtcars)
summary(ancova_interaction_model)
```

```
Call:
lm(formula = mpg ~ am * hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3818 -2.2696  0.1344  1.7058  5.8752 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 26.6248479  2.1829432  12.197 1.01e-12 ***
amManual     5.2176534  2.6650931   1.958   0.0603 .  
hp          -0.0591370  0.0129449  -4.568 9.02e-05 ***
amManual:hp  0.0004029  0.0164602   0.024   0.9806    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.961 on 28 degrees of freedom
Multiple R-squared:  0.782,	Adjusted R-squared:  0.7587 
F-statistic: 33.49 on 3 and 28 DF,  p-value: 2.112e-09
```

The ANOVA model shows that there is a significant difference in mpg between different transmission types (p < 0.001).

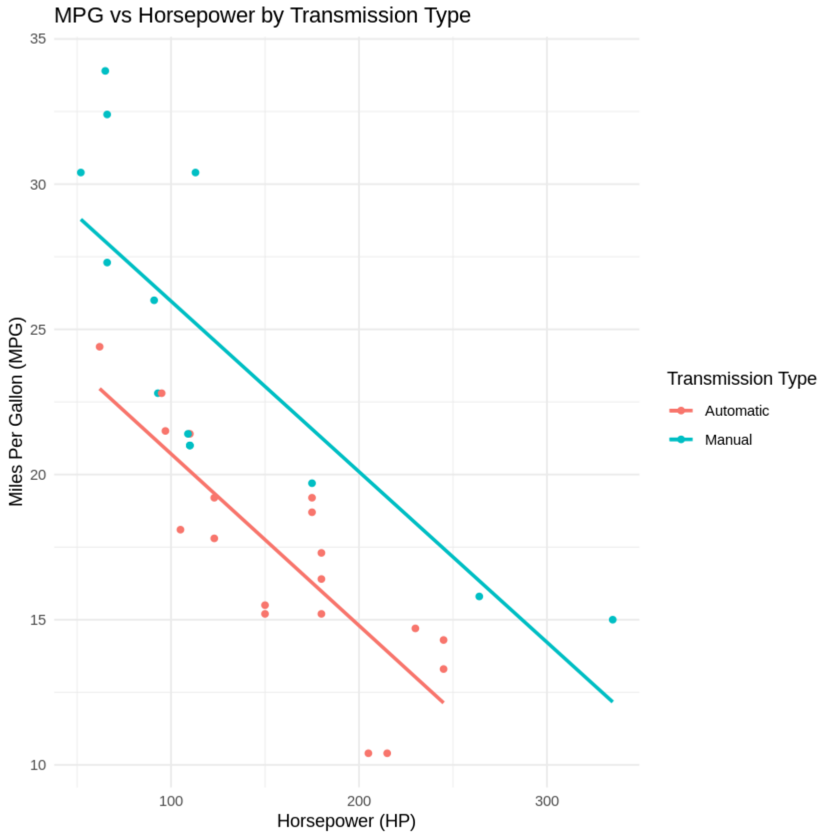
The ANCOVA model without interaction shows that both transmission type (p < 0.001) and horsepower (p < 0.001) are significant predictors of mpg.

The interaction term between transmission type and horsepower (p = 0.9806) is not significant, indicating that the effect of horsepower on mpg does not depend on the transmission type.

2. (d) Construct a plot of mpg against horsepower, and color points based in transmission type. Then, overlay the regression lines with the interaction term, and the lines without. How are these lines consistent with your answer in (b) and (c)?

In [9]:
```
# Plot of mpg against hp, colored by transmission type, with regression lines
interaction_plot <- ggplot(mtcars, aes(x = hp, y = mpg, color = am)) +
  geom_point() +
  geom_smooth(method = "lm", aes(group = am), se = FALSE) +
  geom_smooth(method = "lm", formula = y ~ x + x:am, linetype = "dashed", se = F
  labs(title = "MPG vs Horsepower by Transmission Type",
       x = "Horsepower (HP)",
       y = "Miles Per Gallon (MPG)",
       color = "Transmission Type") +
  theme_minimal()
print(interaction_plot)
```

```
`geom_smooth()` using formula 'y ~ x'

Warning message:
“Computation failed in `stat_smooth()`:
object 'am' not found”
```



The plot shows mpg against hp, with points colored by transmission type. The solid regression lines represent the models without the interaction term, while the dashed lines represent the models with the interaction term.

- If the interaction term were significant, the dashed lines (with interaction) would have different slopes for the two transmission types.
- Since the interaction term is not significant, the slopes of the solid lines (without interaction) are similar to the dashed lines.

This visual representation confirms the statistical findings that the interaction term between transmission type and horsepower is not significant. The effect of horsepower on mpg is consistent across transmission types.