m1-peer-reviewed

February 27, 2024

1 Module 1 - Peer reviewed

1.0.1 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.

[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()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()
```

1.0.2 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

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

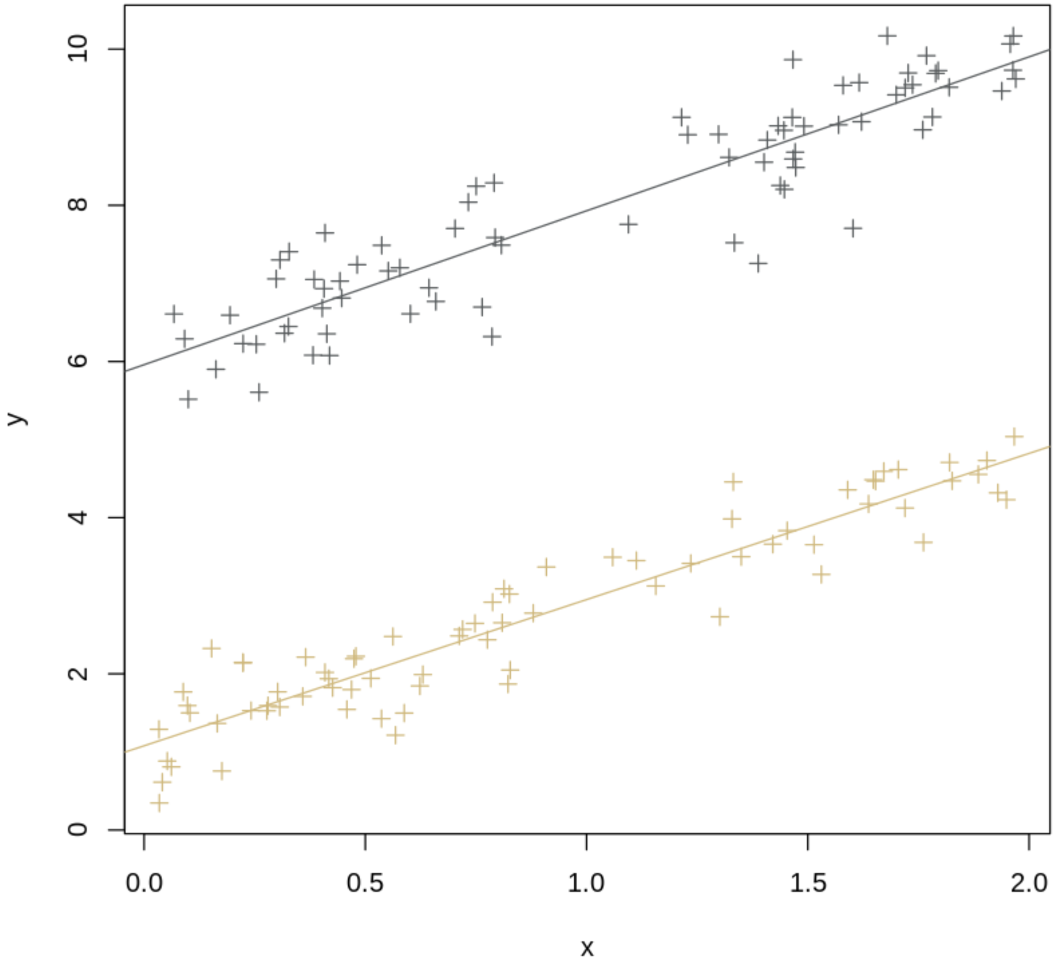
where X is a continuous covariate, Z is a dummy variable coding the levels of a two level factor, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We choose values for the parameters below (b0,...,b2).

[2]:

```r
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 <dbl>	z <fct>	y <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?

The regression lines have nearly the same slope and thus are almost parallel, which would imply that any interaction effect between the variables is not significant. Or, in this case, there is no interaction.

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:

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

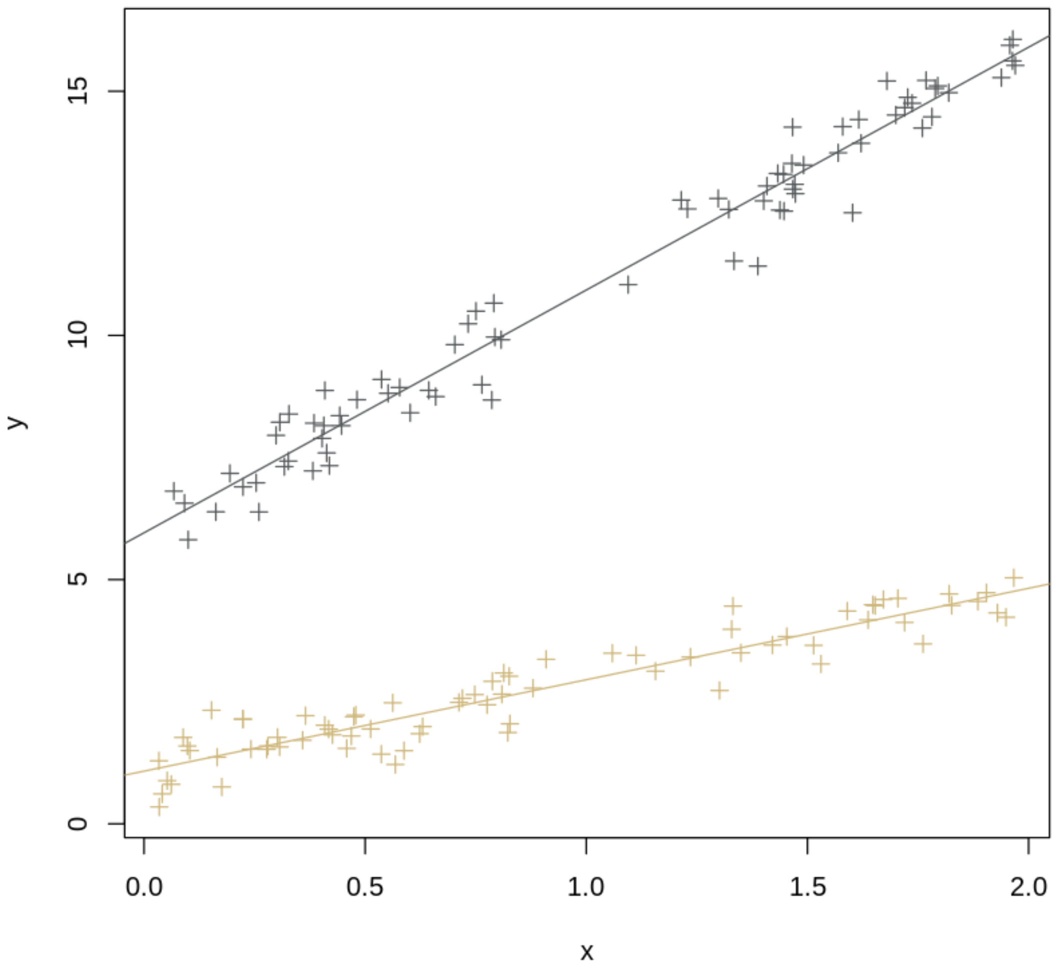
where X is a continuous covariate, Z is a dummy variable coding the levels of a two level factor, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We choose values for the parameters below (b0,...,b3).

[3]:

```r
#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?**

Now that we have added the interaction term, the slopes of the regressions lines are quite different and so the lines are not parallel, which means that there is significant effect from the interactions between the variables.

1.1 Problem #2

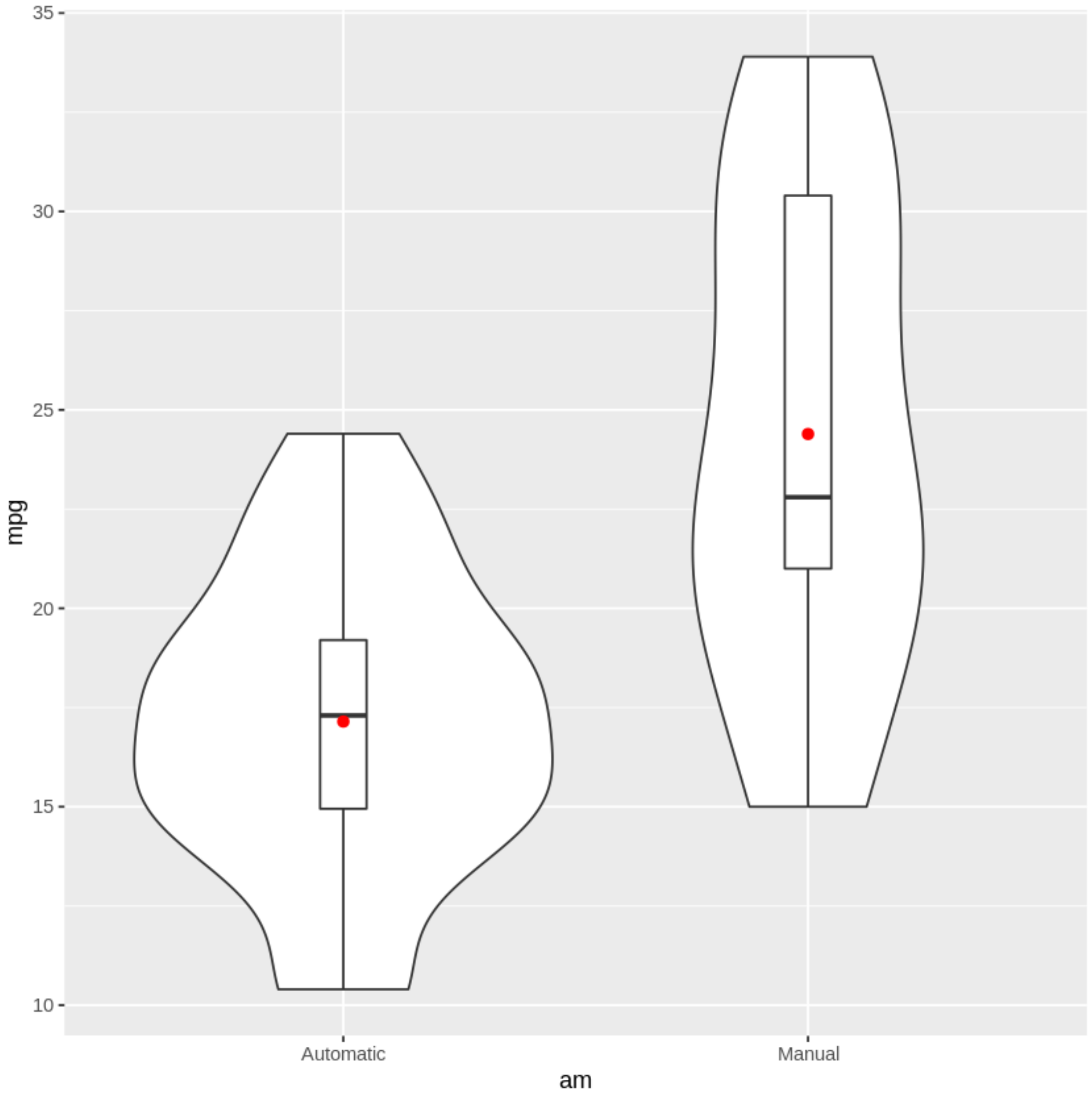
In this question, we ask you to analyze the `mtcars` dataset. The goal if this question will be to try to explain the variability in miles per gallon (mpg) using transmission type (am), while adjusting for horsepower (hp).

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

[4]:

```r
data(mtcars)

# your code here
mtcars = mutate(mtcars, am = case_when(am == 0 ~ "Automatic", am == 1 ~␣
 ↪"Manual"))

p <- ggplot(mtcars, aes(x=am, y=mpg)) +
        geom_violin() +
        geom_boxplot(width=0.1) +
        stat_summary(fun=mean, geom="point", size=2, color="red")
p
```



The violin plots above show miles per gallon against transmission type. The graphs show that the manual transmission data has a larger interquartile range and the values are not closely gathered around the median; it is rather long, skinny, and evenly distributed. If anything, the manual transmission data is gathered around the first quartile. On the otherhand, the data for automatic transmissions are more tightly grouped around the median with the violin being extremely wide in the middle. It is also visually evident that manual transmissions are more fuel efficient than automatic transmissions being that the median car with a manual transmission achieves around 24mpg. This opposes automatic transmission cars which have a median of about 17 mpg. Overall, the graph leads us to believe that manual cars have increased MPG.

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

[5]:

```
# your code here
am0 = mtcars[mtcars$am=='Automatic',]
am1 = mtcars[mtcars$am=='Manual',]

mean(am0$mpg) - mean(am1$mpg)
```

-7.24493927125506

The mean difference in mpg for the Automatic group compared to the Manual group is approximately -7 miles per gallon, which means that automatic transmissions perform worse, at least in this data set.

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.

[6]:

```
# your code here
diff_mpg_trans <- lm(mpg ~ am, data = mtcars)
diff_mpg_trans

diff_mpg_trans_horses <- lm(mpg ~ am + hp, data = mtcars)
diff_mpg_trans_horses

diff_mpg_trans_horses_interaction <- lm(mpg ~ am + hp + am:hp, data = mtcars)
diff_mpg_trans_horses_interaction

summary(diff_mpg_trans_horses_interaction)
```

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

Coefficients:
(Intercept)     amManual
     17.147        7.245



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

Coefficients:
```

```
(Intercept)     amManual           hp 
   26.58491      5.27709     -0.05889 
```

```

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

Coefficients:
(Intercept)     amManual           hp  amManual:hp  
 26.6248479    5.2176534   -0.0591370    0.0004029  
```

```

Call:
lm(formula = mpg ~ am + hp + 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
```

These results are consistant with the visual exploration of the data via violin plots. It is interesting to note that hp has a negative effect on mpg according to the models that consider mpg. A t-test with the interaction term shows that the interaction is not significant.

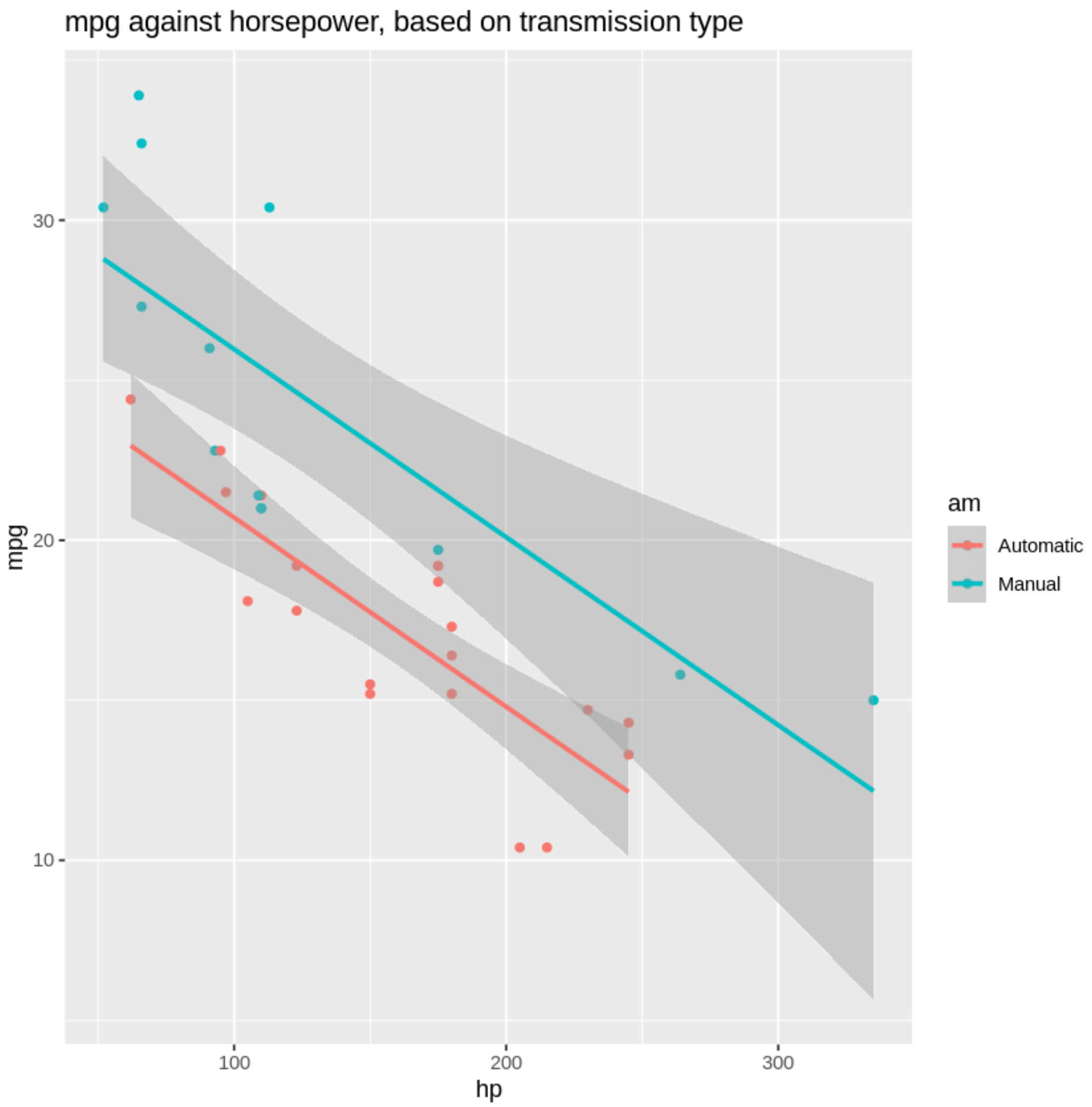
2. (d) Construct a plot of mpg against horsepower, and color points based in transmission type. Then, overlay the regression lines with the interaction term, and the lines without. How are these lines consistent with your answer in (b) and (c)?

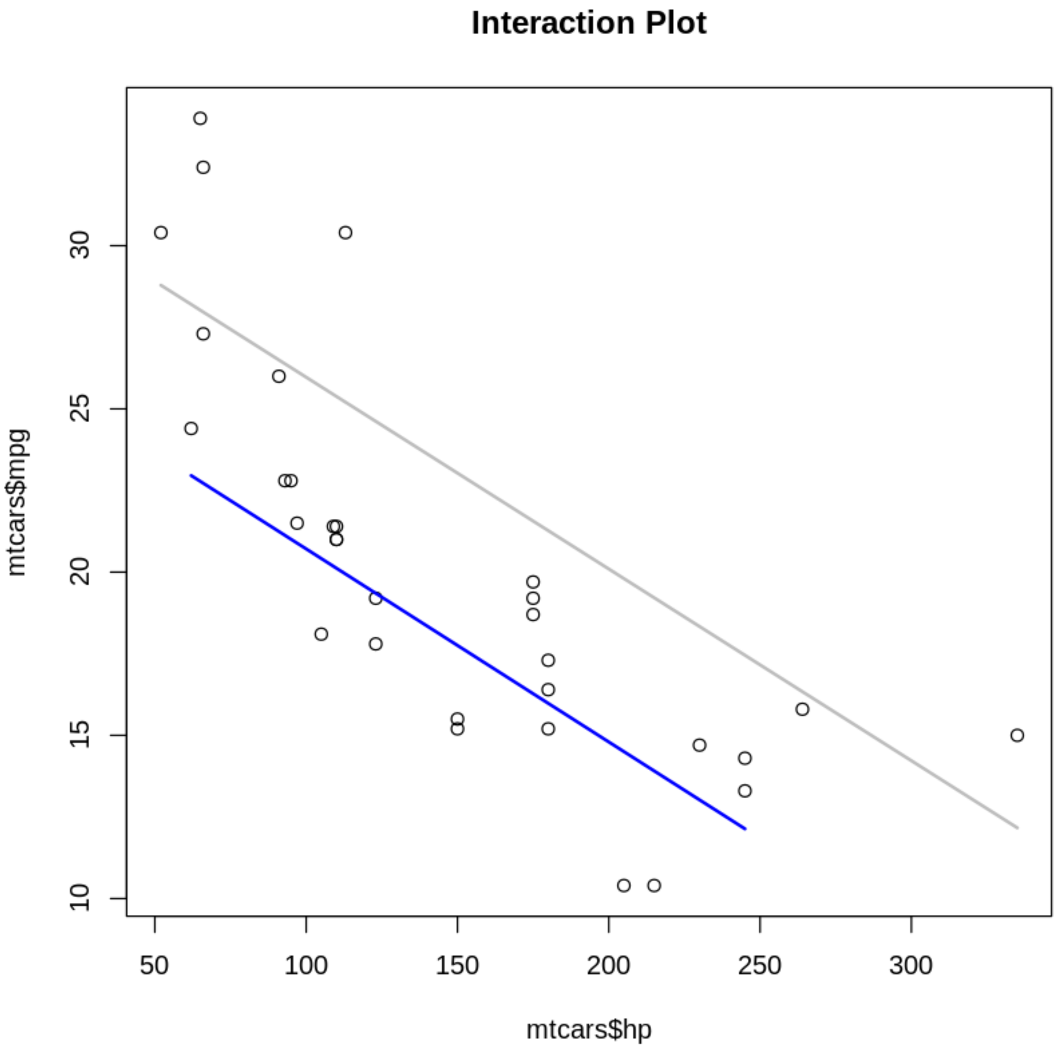
[7]:

```r
# your code here
p = ggplot(mtcars, aes(x=hp, y=mpg, color=am)) +
    geom_point() +
    geom_smooth(method='lm') +
    labs(title="mpg against horsepower, based on transmission type")
```

```
p


pd0 <- predict(diff_mpg_trans_horses_interaction,newdata=am0)
pd1 <- predict(diff_mpg_trans_horses_interaction,newdata=am1)
plot(mtcars$hp,
    mtcars$mpg,
     main="Interaction Plot",
     bg=ifelse(mtcars$am == "Automatic", "steelblue",'gray'))
lines(am0$hp,pd0,col="blue",lwd=2)
lines(am1$hp,pd1,col="gray",lwd=2)
```

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





The regression lines superimposed over the scatterplot showing mpg against hp for each transmission type are consistent with the conclusions drawn in each of the previous parts. The regression lines for both the model with the interaction term and the model without both show the same general trend: they are parallel. This is consistent with the t-test conducted in the previous section to determine that the interaction is not significant.

[ ]: