STA 5207: Homework 4

Due: Friday, February 9 by 11:59 PM

Include your R code in an R chunks as part of your answer. In addition, your written answer to each exercise should be self-contained so that the grader can determine your solution without reading your code or deciphering its output

Exercise 1 (Average Treatment Effect) [25 Points]

For this exercise we will use a subset of the `hips` data set from the `faraway` package, which can be found in `hips_subset.csv` on Canvas. This data set contains data used to study a new treatment for Ankylosing spondylitis (AS), which is a chronic form of arthritis. A study was conducted to determine whether daily stretching of the hip tissues would improve mobility. There are 75 AS patients who where randomly allocated to a control (standard treatment) or a treatment (the new treatment) group. The data set contains three variables:

- `fbef` : flexion angle before
- `faft`: flexion angle after.
- `grp`: treatment group. A factor with levels `control` (individuals received the standard treatment) and `treat` (individuals received a new treatment).

In this exercise, we will determine if there is a statistically significant *average treatment effect*, that is, whether there is a difference in the average value of `faft` between individuals in the `treat` and `control` group who started with the same value of `fbef`.

1. (2 points) Load the data and check its structure using `str()`. Verify that `grp` is a factor. If not, coerce it to be a factor. Include your code and its output below. What is the default reference level chosen by R?

```r
data = read.csv("hips_subset.csv")
str(data)
```

```
## 'data.frame':    75 obs. of  3 variables:
##  $ fbef: int  125 120 135 135 100 110 122 122 124 124 ...
##  $ faft: int  126 127 135 135 113 115 123 125 126 135 ...
##  $ grp : chr  "treat" "treat" "treat" "treat" ...
```

```r
is.factor(data$grp)
```

```
## [1] FALSE
```

```r
data$grp =  as.factor(data$grp)
is.factor(data$grp)
```

```
## [1] TRUE
```

```r
str(data)
```

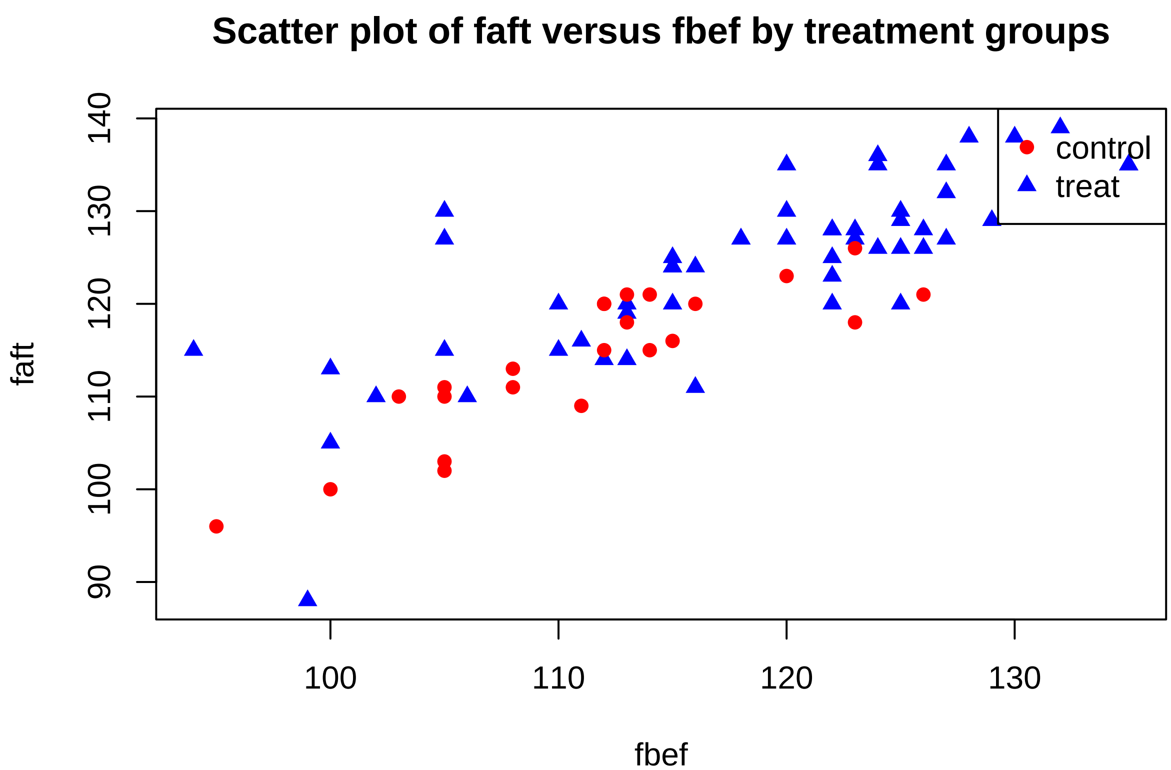
```
## 'data.frame':    75 obs. of  3 variables:
##  $ fbef: int  125 120 135 135 100 110 122 122 124 124 ...
##  $ faft: int  126 127 135 135 113 115 123 125 126 135 ...
##  $ grp : Factor w/ 2 levels "control","treat": 2 2 2 2 2 2 2 2 2 2 ...
```

2. (2 points) Using the plotting functions discussed in class, make a scatter plot of `faft` versus `fbef` . Use a different color point and shape for each level of `grp`. Also be sure to label the axes appropriately and include a legend. Based on the scatter plot, does the linear relationship between `faft` and `fbef` seem to differ between treatment groups? Briefly explain.

```r
# Define colors and shapes for each group
colors <- c('treat' = 'red', 'control' = 'blue')
shapes <- c('treat' = 16, 'control' = 17)

# Create the scatter plot
plot(data$fbef, data$faft, col = colors[data$grp], pch = shapes[data$grp],
     xlab = "fbef", ylab = "faft", main = "Scatter plot of faft versus fbef by treatment groups")

# Add legend
legend("topright", legend = levels(factor(data$grp)),
       col = unique(colors), pch = unique(shapes))
```



3. (4 points) Fit a simple linear regression model with `faft` as the response and `fbef` as the predictor. Give the estimated regression equation.

```r
model = lm(faft ~ fbef, data = data)
summary(model)$coefficients
```

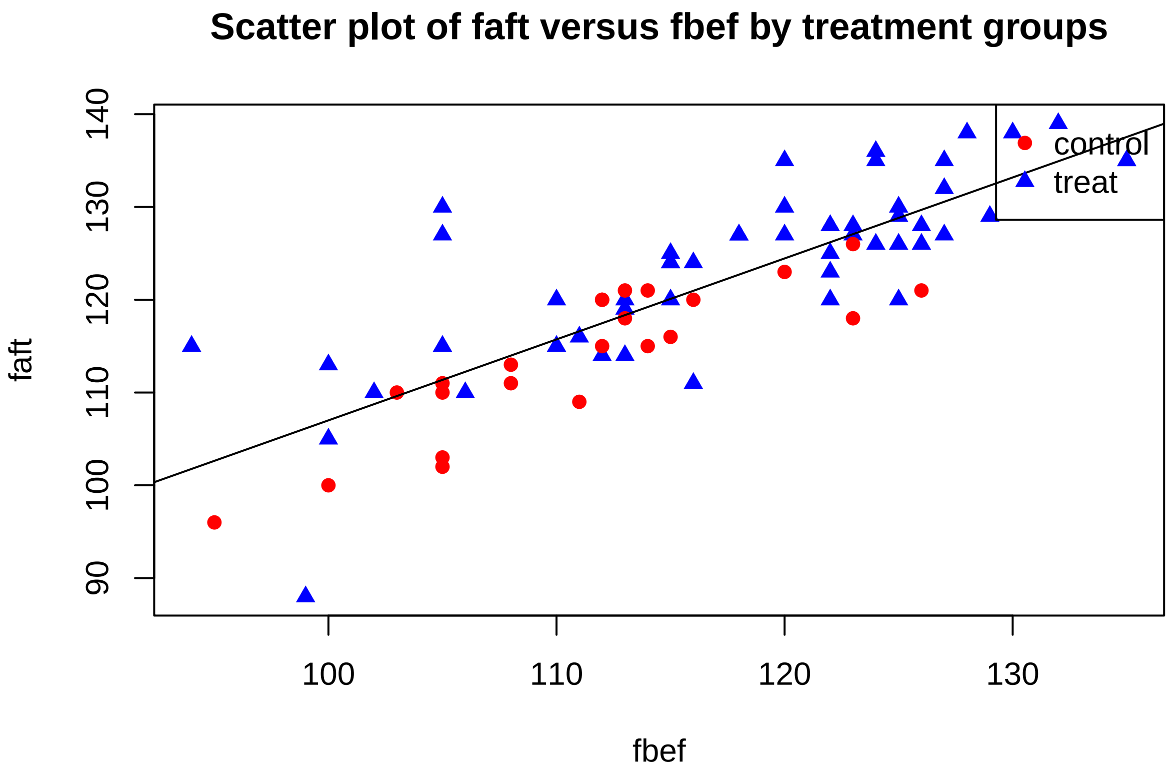
```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 19.7321226 8.30619880  2.37559 2.014778e-02
## fbef         0.8726813 0.07141478 12.21990 2.503684e-19
```

4. (3 points) Using the plotting functions discussed in class, make a scatter plot of `faft` versus `fbef`. Use a different color point and shape for each level of `grp`. Also be sure to label the axes appropriately and include a legend. Add the fitted regression line from the SLR model you estimated in Question 3 to the scatter plot. Comment on how well this line models the data.

```r
# Create the scatter plot
plot(data$fbef, data$faft, col = colors[data$grp], pch = shapes[data$grp],
     xlab = "fbef", ylab = "faft", main = "Scatter plot of faft versus fbef by treatment groups")

abline(model)

# Add legend
legend("topright", legend = levels(factor(data$grp)),
       col = unique(colors), pch = unique(shapes))
```



5. (5 points) Fit an additive multiple regression model with `faft` as the response and `fbef` and `grp` as the predictors. Give the two separate estimated regression equations for the `control` and `treat` groups.

```r
model1 = lm(faft ~ fbef + grp, data = data)
coef(model1)
```

```
## (Intercept)        fbef    grptreat
##  25.1914193   0.7973406   4.7223852
```

```r
int_control = coef(model1)[1]
int_treat = coef(model1)[1] + coef(model1)[3]
slope_all_species = coef(model1)[2]
```

6. (3 points) Using the plotting functions discussed in class, make a scatter plot of `faft` versus `fbef`. Use a different color point and shape for each level of `grp`. Also be sure to label the axes appropriately and include a legend. Add the two fitted regression lines from the additive model to the scatter plot with the same colors as their respective points (one line for each level of `grp`). Comment on how well these lines model the data.

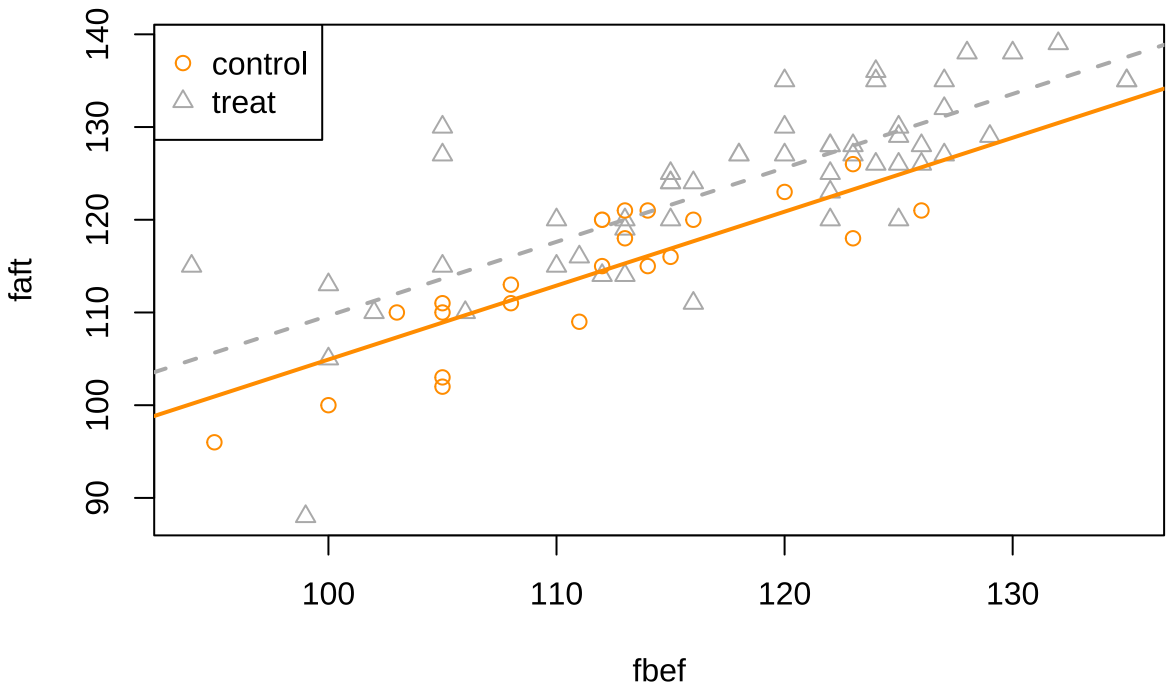
```r
plot_colors = c("Darkorange", "Darkgrey")

plot(faft ~ fbef, data = data,
```

```r
     col = plot_colors[grp],  pch = as.numeric(grp),
     xlab = "fbef", ylab = "faft")


# NEW: ablines take the intercept and slope of each regression line calculated above
abline(int_control, slope_all_species, col= plot_colors[1], lty = 1, lwd = 2)
abline(int_treat, slope_all_species, col = plot_colors[2], lty = 2, lwd = 2)

legend("topleft", levels(data$grp), col=plot_colors,  pch = c(1, 2, 3))
```



7. (6 points) Use an appropriate test to determine whether there is a significant average treatment effect, that is, perform a test to compare the model from Question 3 to the model from Question 5. Report the following:

- The null and alternative hypotheses.
- The value of the test statistic.
- The p-value of the test.
- A statistical decision at $\alpha = 0.01$.
- A conclusion in the context of the problem.

```r
anova(model, model1)
```

```
## Analysis of Variance Table
##
## Model 1: faft ~ fbef
## Model 2: faft ~ fbef + grp
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     73 2524.1
## 2     72 2206.9  1    317.14 10.347 0.001944 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 2 (The Iris Data Set) [35 Points]

For this exercise we will use the `iris` data set. This is a default data set in `R`. You can also find the data in `iris.csv` on Canvas. This data set gives measurements of 50 flowers from 3 species of iris. The data set contains the following 4 variables:

- `Sepal.Length`: sepal length in cm.
- `Sepal.Width`: sepal width in cm.
- `Petal.Length`: petal length in cm.
- `Petal.Width`: petal width in cm.
- `Species`: The three species of iris flowers: "setosa", "versicolor", and "virginica".

For this exercise, we will model `Sepal.Width` as a function of `Sepal.Length` and `Species`.

1. (2 points) Load the data and check its structure using `str()`. Verify that `species` is a factor. If not, coerce it to be a factor. Include your code and its output below. What is the default reference level chosen by `R`?

```
iris = read.csv("iris.csv")
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : chr  "setosa" "setosa" "setosa" "setosa" ...
```

```
is.factor(iris$Species)
```

```
## [1] FALSE
```

```
iris$Species =  as.factor(iris$Species)
is.factor(iris$Species)
```

```
## [1] TRUE
```

```
str(iris)
```

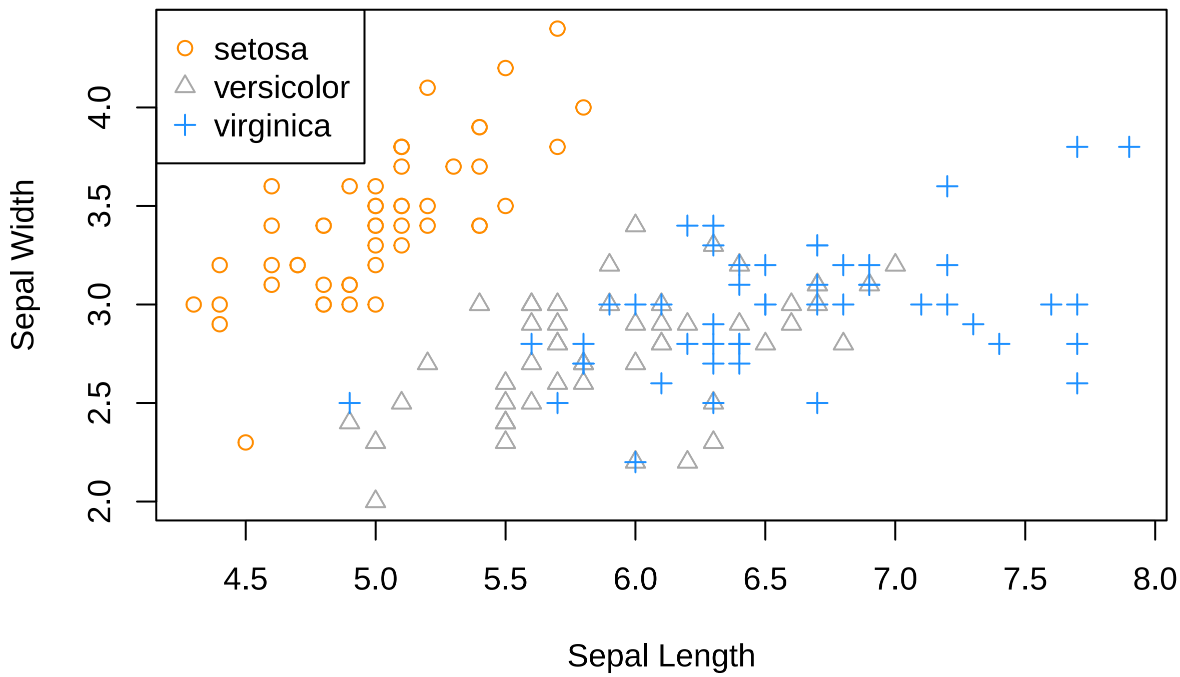
```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

2. (2 points) Using the plotting functions discussed in class, make a scatter plot of `Sepal.Width` versus `Sepal.Length`. Use a different color point and shape for each `Species`. Also be sure to label the axes appropriately and include a legend. Based on the scatter plot, does the linear relationship between `Sepal.Width` and `Sepal.Length` seem to differ between flower species? Briefly explain.

```
plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")

plot(Sepal.Width ~ Sepal.Length, data = iris,
     col = plot_colors[Species],  pch = as.numeric(Species),
     xlab = "Sepal Length", ylab = "Sepal Width")
```

```r
legend("topleft", levels(iris$Species), col=plot_colors,  pch = c(1, 2, 3))
```



3. (4 points) Estimate a simple linear regression model with `Sepal.Width` as the response and only `Sepal.Length` as the predictor. Give the estimated regression equation and an estimate for the average change in `Sepal.Width` for a 1 cm increase in `Sepal.Length` for `setosa` flowers.

```r
iris_model = lm(Sepal.Width ~ Sepal.Length, data = iris)
coef(iris_model)
```

```
##  (Intercept) Sepal.Length 
##    3.4189468   -0.0618848
```

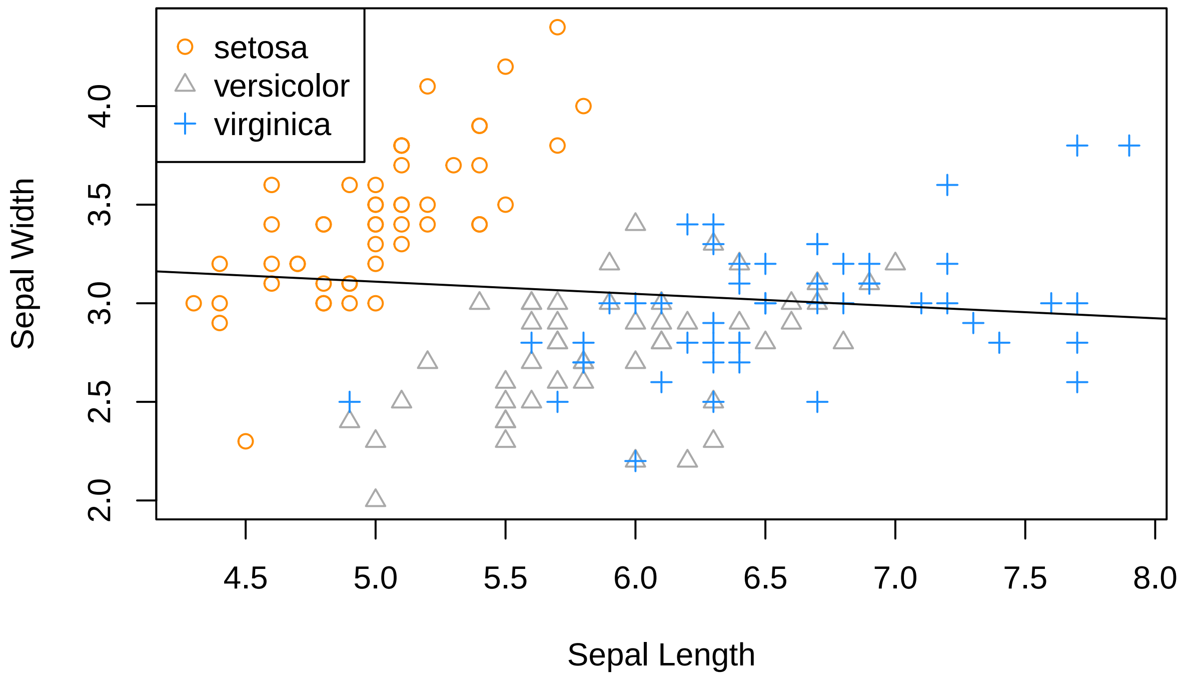
4. (3 points) Using the plotting functions discussed in class, make a scatter plot of `Sepal.Width` versus `Sepla.Length`. Use a different color point and shape for each `Species`. Also be sure to label the axes appropriately and include a legend. Add the fitted regression line for the SLR model you estimated in Question 3 to the scatter plot. Comment on how well this line models the data.

```r
plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")

plot(Sepal.Width ~ Sepal.Length, data = iris,
     col = plot_colors[Species],  pch = as.numeric(Species),
     xlab = "Sepal Length", ylab = "Sepal Width")

abline(iris_model)


legend("topleft", levels(iris$Species), col=plot_colors,  pch = c(1, 2, 3))
```



5. (4 points) Fit an additive multiple regression model with `Sepal.Width` as the response and `Sepal.Length` and `Species` as the predictors. Give the three separate estimated regression equations for **`setosa`**, **`versicolor`**, and **`virginica`** flowers. Also, give an estimate for the average change in `Sepal.Width` for a 1 cm increase in `Sepal.Length` for **`setosa`** flowers.

```
iris_model_add = lm(Sepal.Width ~ Sepal.Length + Species, data = iris)
coef(iris_model_add)
```

```
##       (Intercept)      Sepal.Length Speciesversicolor  Speciesvirginica
##         1.6765001         0.3498801        -0.9833885        -1.0075104
```

```
int_setosa = coef(iris_model_add)[1]
int_versicolor = coef(iris_model_add)[1] + coef(iris_model_add)[3]
int_virginica = coef(iris_model_add)[1] + coef(iris_model_add)[4]
slope_all_species = coef(iris_model_add)[2]
```

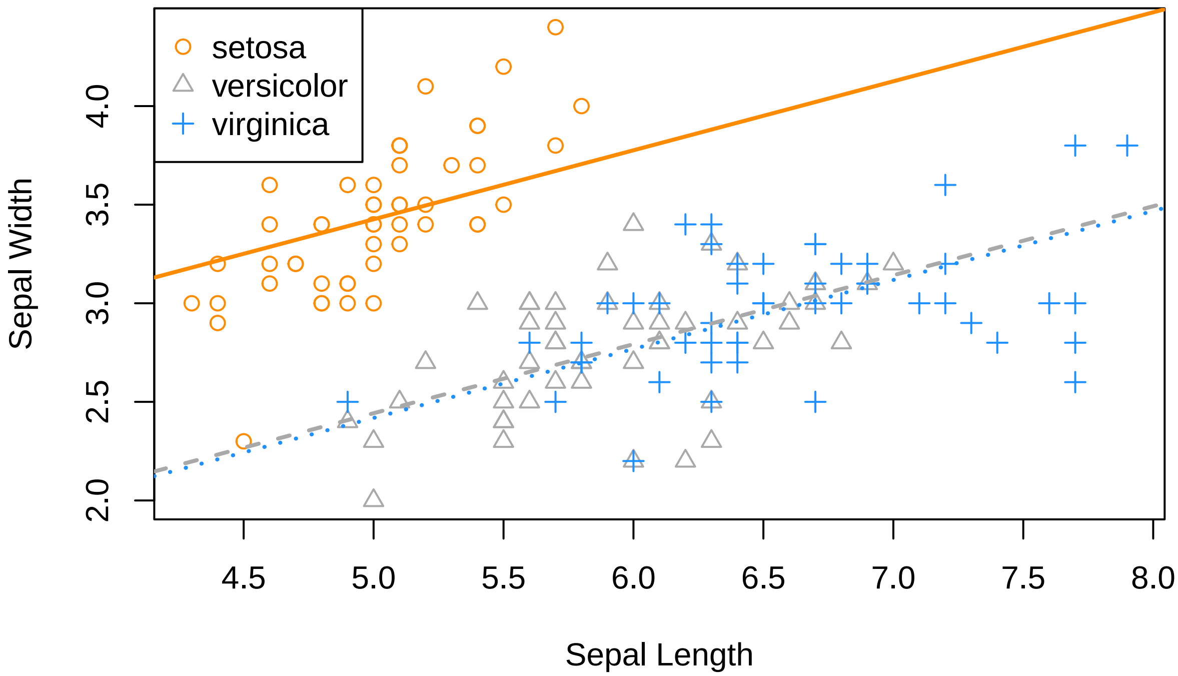
6. (3 points) Using the plotting functions discussed in class, make a scatter plot of `Sepal.Width` versus `Sepal.Length`. Use a different color point and shape for each `Species`. Also be sure to label the axes appropriately and include a legend. Add the three fitted regression lines from the additive model to the scatter plot with the same colors as their respective points (one line for each species type). Comment on how well these lines model the data.

```
plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")

plot(Sepal.Width ~ Sepal.Length, data = iris,
     col = plot_colors[Species],  pch = as.numeric(Species),
     xlab = "Sepal Length", ylab = "Sepal Width")

# NEW: ablines take the intercept and slope of each regression line calculated above
abline(int_setosa, slope_all_species, col= plot_colors[1], lty = 1, lwd = 2)
abline(int_versicolor, slope_all_species, col = plot_colors[2], lty = 2, lwd = 2)
abline(int_virginica, slope_all_species, col = plot_colors[3], lty = 3, lwd = 2)

legend("topleft", levels(iris$Species), col=plot_colors,  pch = c(1, 2, 3))
```



7. (5 points) Use an appropriate test to compare the SLR model from Question 3 to the additive model in Question 5 at an $\alpha = 0.01$ significance level. Report the following:

- The null and alternative hypotheses.
- The value of the test statistic.
- The p-value of the test.
- The model you prefer based on the results of the test

8. (4 points) Fit an interaction MLR model with `Sepal.Width` as the response and `Sepal.Length` and `Species` as the predictors. Give the three separate estimated regression equations for `setosa`, `versicolor`, and `virginica` flowers. Also, give an estimate for the average change in `Sepal.Width` for a 1 cm increase in `Sepal.Length` for `setosa` flowers.

```
model_int = lm(Sepal.Width ~ Sepal.Length * Species, data = iris)
coef(model_int)
```

```
##                     (Intercept)                    Sepal.Length
##                      -0.5694327                       0.7985283
##               Speciesversicolor                Speciesvirginica
##                       1.4415786                       2.0157381
## Sepal.Length:Speciesversicolor  Sepal.Length:Speciesvirginica
##                      -0.4788090                      -0.5666378
```

```
int_setosa = coef(model_int)[1]
int_versicolor = coef(model_int)[1] + coef(model_int)[3]
int_virginica = coef(model_int)[1] + coef(model_int)[4]
slope_setosa =  coef(model_int)[2]
slope_versicolor = coef(model_int)[2] + coef(model_int)[5]
slope_virginica = coef(model_int)[2] + coef(model_int)[6]
```

9. (3 points) Using the plotting functions discussed in class, make a scatter plot of `Sepal.Width` versus `Sepal.Length`. Use a different color point and shape for each `Species`. Also be sure to label the axes appropriately and include a legend. Add the three fitted regression lines from the interaction model

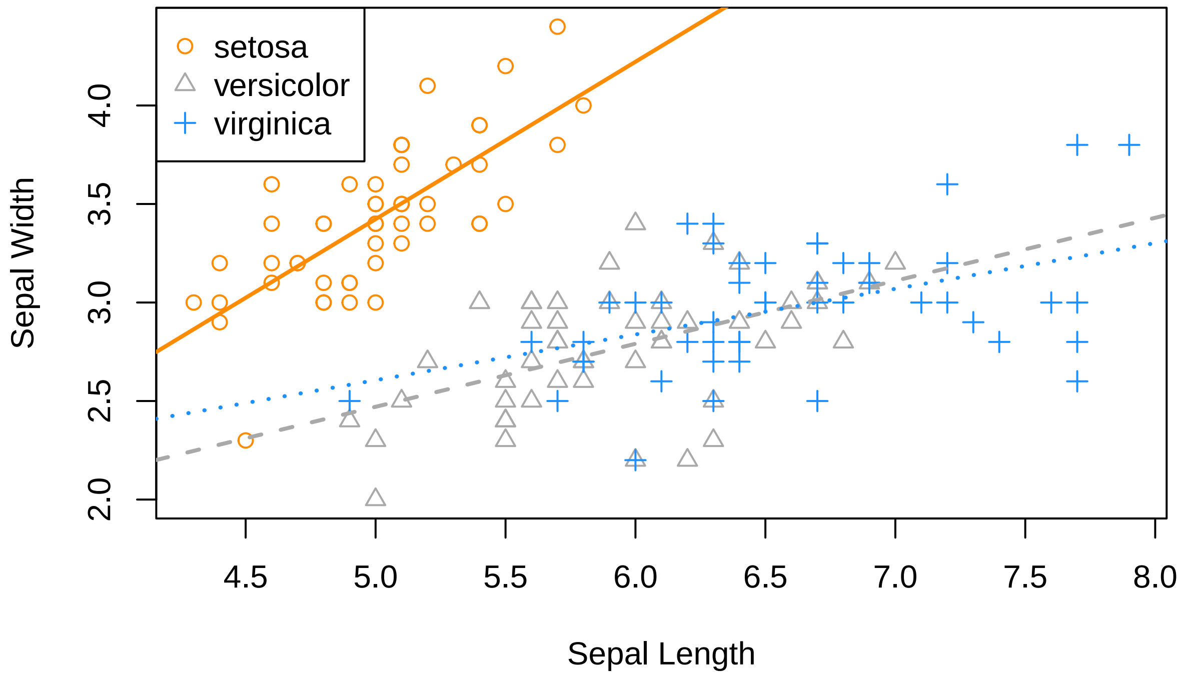
to the scatter plot with the same colors as their respective points (one line for each species type). Comment on how well these lines model the data.

```
plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")

plot(Sepal.Width ~ Sepal.Length, data = iris,
     col = plot_colors[Species],  pch = as.numeric(Species),
     xlab = "Sepal Length", ylab = "Sepal Width")

# NEW: ablines take the intercept and slope of each regression line calculated above
abline(int_setosa, slope_setosa, col= plot_colors[1], lty = 1, lwd = 2)
abline(int_versicolor, slope_versicolor, col = plot_colors[2], lty = 2, lwd = 2)
abline(int_virginica, slope_virginica, col = plot_colors[3], lty = 3, lwd = 2)

legend("topleft", levels(iris$Species), col=plot_colors,  pch = c(1, 2, 3))
```



10. (5 points) Use an appropriate test to compare the additive model from Question 5 to the interaction model in Question 8 at an $\alpha = 0.01$ significance level. Report the following:
    - The null and alternative hypotheses.
    - The value of the test statistic.
    - The p-value of the test.
    - The model you prefer based on the results of the test.

Exercise 3 (2015 EPA Emissions Data Set) [40 Points]

For this exercise we will use the `epa` data set, which can be found in the `epa.csv` file on Canvas. This data set contains detailed descriptions of 4,411 vehicles manufactured in 2015 that were used for fuel economy testing as performed by the Environmental Projection Agency. The variables in the dataset are:

- `CO2`: carbon dioxide (the primary byproduct of all fossil fuel combustion), in g/mi.
- `horse`: rated horsepower, in foot-pounds per second.

- `type`: vehicle type: Car, Truck, or Both (for vehicles that meet specifications of both car and truck, like smaller SUVs or crossovers).

In this exercise, we will model `CO2` a a function of `horse` and `type`.

1. (2 points) Load the data and check its structure using `str()`. Verify that `type` is a factor. If not, coerce it to be a factor. Include your code and its output below. What is the default reference level chosen by `R`?

```r
epa = read.csv("epa.csv")
str(epa)
```

```
## 'data.frame':    4411 obs. of  3 variables:
##  $ CO2  : num  550 344 512 297 603 ...
##  $ horse: int  510 510 552 552 565 565 420 420 430 430 ...
##  $ type : chr  "Car" "Car" "Car" "Car" ...
```

```r
is.factor(epa$type)
```

```
## [1] FALSE
```

```r
epa$type =  as.factor(epa$type)
is.factor(epa$type)
```

```
## [1] TRUE
```

```r
str(epa)
```

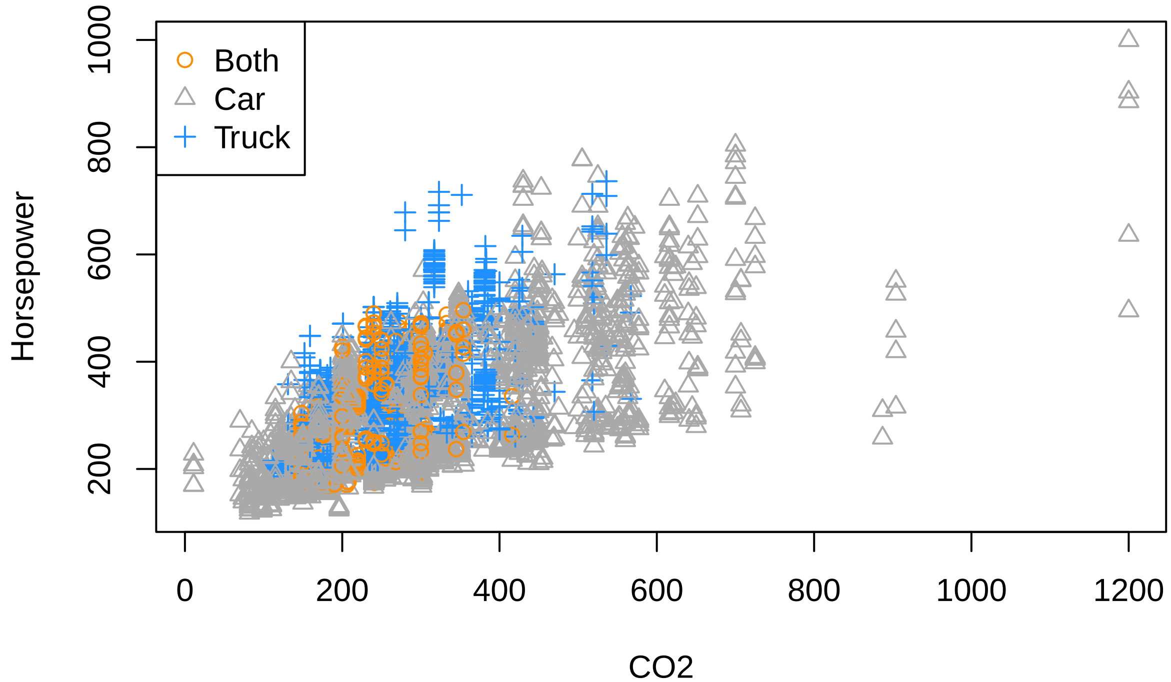
```
## 'data.frame':    4411 obs. of  3 variables:
##  $ CO2  : num  550 344 512 297 603 ...
##  $ horse: int  510 510 552 552 565 565 420 420 430 430 ...
##  $ type : Factor w/ 3 levels "Both","Car","Truck": 2 2 2 2 2 2 2 2 2 2 ...
```

2. (2 points) Using the plotting functions discussed in class, make a scatter plot of `CO2` versus `horse`. Use a different color point and shape for each vehicle `type`. Also be sure to label the axes appropriately and include a legend. Based on the scatter plot, does the linear relationship between `CO2` and `horse` seem to differ between vehicle type? Briefly explain.

```r
plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")

plot(CO2 ~ horse, data = epa,
     col = plot_colors[type],  pch = as.numeric(type),
     xlab = "CO2", ylab = "Horsepower")


legend("topleft", levels(epa$type), col=plot_colors,  pch = c(1, 2, 3))
```



3. (4 points) Estimate a simple linear regression model with `CO2` as the response and only `horse` as the predictor. Give the estimated regression equation and an estimate for the average change in `CO2` for a one foot-pound per second increase in `horse` for a vehicle of type `Truck`.

```
epa_model = lm(CO2 ~ horse, data = epa)
coef(epa_model)
```

```
## (Intercept)       horse
## 154.7178499   0.5498996
```

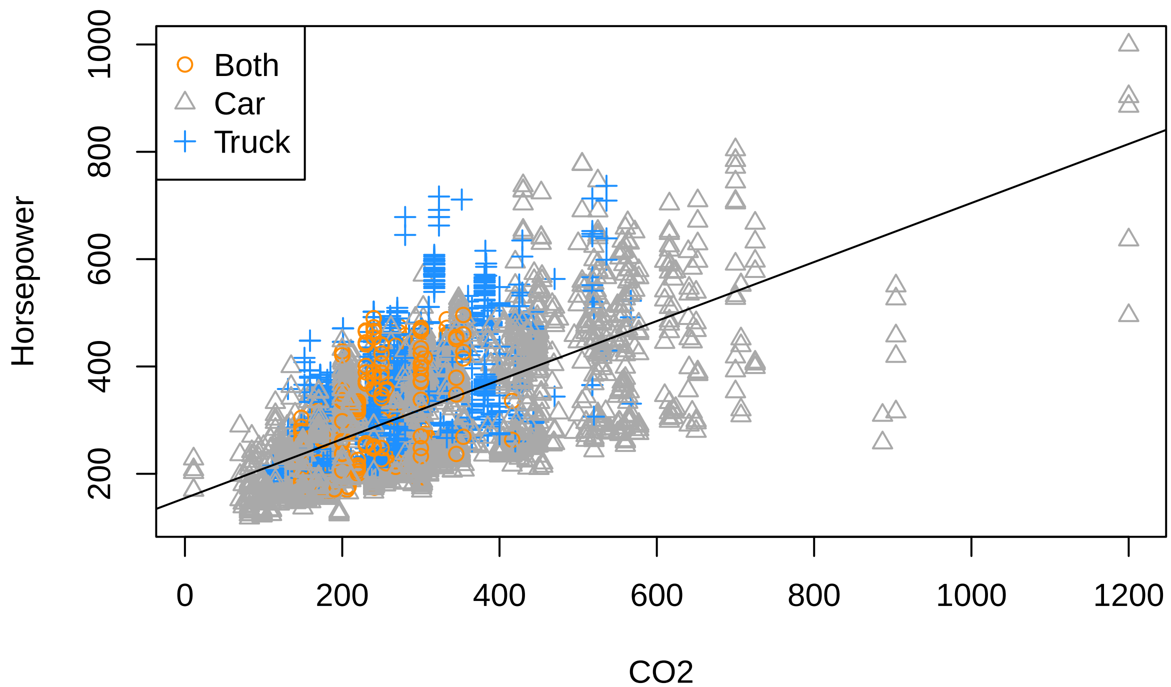
4. (3 points) Using the plotting functions discussed in class, make a scatter plot of `CO2` versus `horse`. Use a different color point and shape for each vehicle `type`. Also be sure to label the axes appropriately and include a legend. Add the fitted regression line for the SLR model you estimated in Question 3 to the scatter plot. Comment on how well this line models the data.

```
plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")

plot(CO2 ~ horse, data = epa,
     col = plot_colors[type],  pch = as.numeric(type),
     xlab = "CO2", ylab = "Horsepower")

abline(epa_model)

legend("topleft", levels(epa$type), col=plot_colors,  pch = c(1, 2, 3))
```



5. (5 points) Fit an additive multiple regression model with `CO2` as the response and `horse` and `type` as the predictors. Give the three separate estimated regression equations for `Car`, `Truck`, and `Both` vehicles. Also, give an estimate for the average change in `CO2` for a one foot-pound per second increase in `horse` for a vehicle of type `Truck`.

```
epa_model_add = lm(CO2 ~ horse + type, data = epa)
coef(epa_model_add)
```

```
## (Intercept)       horse     typeCar   typeTruck
## 155.9821483   0.5611008 -22.4250731  40.0744433
```

```
int_both = coef(epa_model_add)[1]
int_car = coef(epa_model_add)[1] + coef(epa_model_add)[3]
int_truck = coef(epa_model_add)[1] + coef(epa_model_add)[4]
slope_all_types = coef(epa_model_add)[2]
```

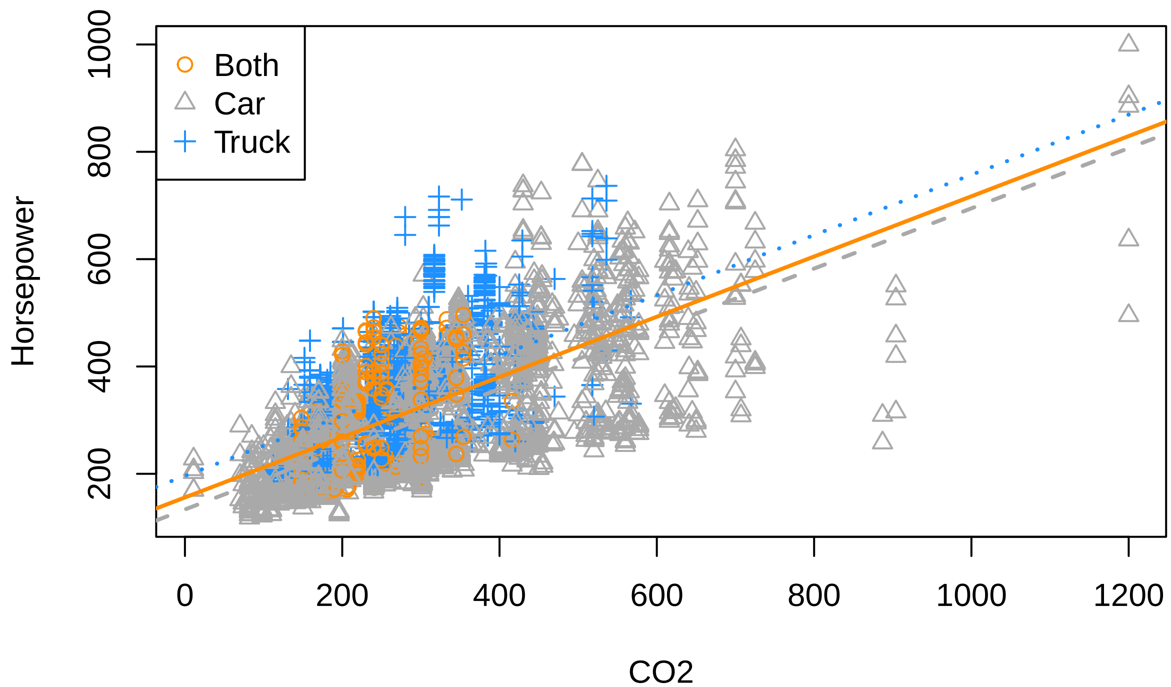
6. (3 points) Using the plotting functions discussed in class, make a scatter plot of `CO2` versus `horse`. Use a different color point and shape for each vehicle `type`. Also be sure to label the axes appropriately and include a legend. Add the three fitted regression lines from the additive model to the scatter plot with the same colors as their respective points (one line for each species type). Comment on how well these lines model the data.

```
plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")

plot(CO2 ~ horse, data = epa,
     col = plot_colors[type],  pch = as.numeric(type),
     xlab = "CO2", ylab = "Horsepower")

# NEW: ablines take the intercept and slope of each regression line calculated above
abline(int_both, slope_all_types, col= plot_colors[1], lty = 1, lwd = 2)
abline(int_car, slope_all_types, col = plot_colors[2], lty = 2, lwd = 2)
abline(int_truck, slope_all_types, col = plot_colors[3], lty = 3, lwd = 2)

legend("topleft", levels(epa$type), col=plot_colors,  pch = c(1, 2, 3))
```



7. (5 points) Use an appropriate test to compare the SLR model from Question 3 to the additive model in Question 5 at an $\alpha = 0.05$ significance level. Report the following:

   - The null and alternative hypotheses.
   - The value of the test statistic.
   - The p-value of the test.
   - The model you prefer based on the results of the test

8. (5 points) Fit an interaction MLR model with `CO2` as the response and `horse` and `type` as the predictors. Give the three separate estimated regression equations for `Car`, `Truck`, and `Both` vehicles. Also, give an estimate for the average change in `CO2` for a one foot-pound per second increase in `horse` for a vehicle of type `Truck`.

```
epa_model_int = lm(CO2 ~ horse*type, data = epa)
coef(epa_model_int)
```

```
##      (Intercept)           horse         typeCar       typeTruck   horse:typeCar
##    149.89711799      0.58605967    -11.17957646      7.66403763     -0.04285633
## horse:typeTruck
##      0.11532862
```

```
int_both = coef(epa_model_int)[1]
int_car = coef(epa_model_int)[1] + coef(epa_model_int)[3]
int_truck = coef(epa_model_int)[1] + coef(epa_model_int)[4]
slope_both =  coef(epa_model_int)[2]
slope_car = coef(epa_model_int)[2] + coef(epa_model_int)[5]
slope_truck = coef(epa_model_int)[2] + coef(epa_model_int)[6]
```

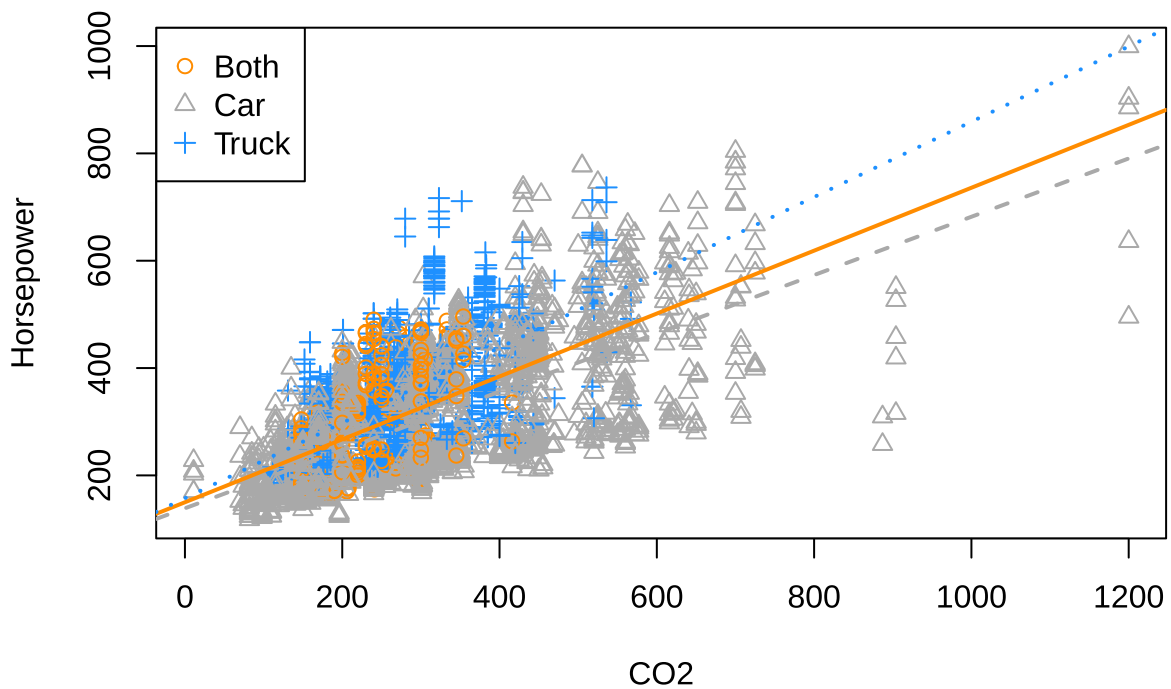
9. (3 points) Using the plotting functions discussed in class, make a scatter plot of `CO2` versus `horse`. Use a different color point and shape for each vehicle `type`. Also be sure to label the axes appropriately and include a legend. Add the three fitted regression lines from the interaction model to the scatter plot with the same colors as their respective points (one line for each species type). Comment on how well these lines model the data.

```r
plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")

plot(CO2 ~ horse, data = epa,
     col = plot_colors[type],  pch = as.numeric(type),
     xlab = "CO2", ylab = "Horsepower")

# NEW: ablines take the intercept and slope of each regression line calculated above
abline(int_both, slope_both, col= plot_colors[1], lty = 1, lwd = 2)
abline(int_car, slope_car, col = plot_colors[2], lty = 2, lwd = 2)
abline(int_truck, slope_truck, col = plot_colors[3], lty = 3, lwd = 2)


legend("topleft", levels(epa$type), col=plot_colors,  pch = c(1, 2, 3))
```



10. (5 points) Use an appropriate test to compare the additive model from Question 5 to the interaction model in Question 8 at an $\alpha = 0.05$ significance level. Report the following:

    - The null and alternative hypotheses.
    - The value of the test statistic.
    - The p-value of the test.
    - The model you prefer based on the results of the test.

11. (3 points) Give a 95% prediction interval using the model you chose in Question 10 for a 2015 BMW M4, which is a vehicle with 425 horse power and considered type `Car`.