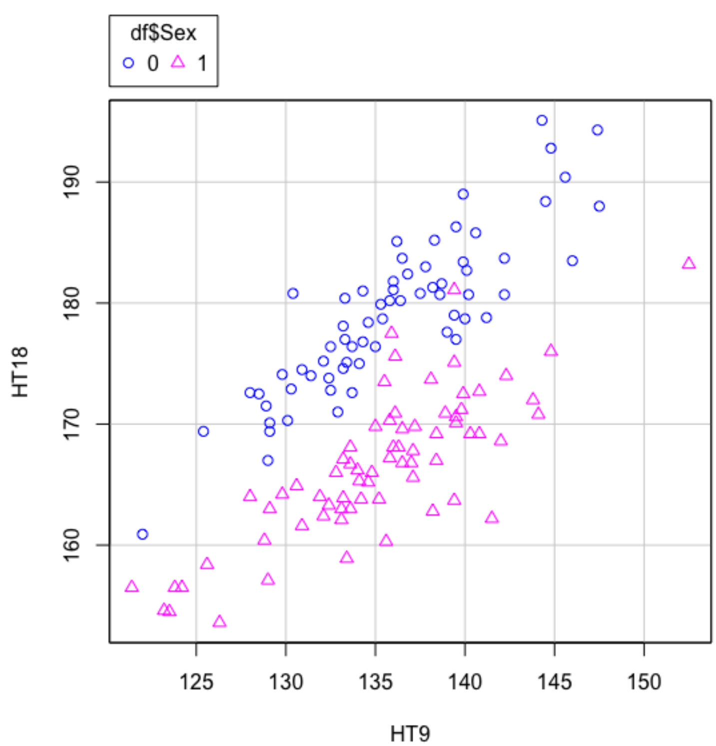
Exercise 1: Do problem 5.14. For part 1 use scatterplot() fumnction with the groups agrumet to get different plotting symbols for males and females, as described in this week's lab. You will also need to turn BGsall$sex ino a factor variable before you do part to 2 and 3. For part 2 testing the parallel regerssion model consists of testing the interaction term, since the interaction allows for non parallel slopers in HT9. For part 3, remember that the difference between males and females is represented by a particular model coefficient hence you are asked to simple find a confidence interval ona coeficient.

Refer to the Berkely Guidence study decrived in Problem 3.3, Using data BGSall, consider the regression HT18 on HT9 and grouping factor Sex.

5.14.1 Draw the scatterplot of HT18 versus HT9, using a different symbol for males and females. Comment on the information in the graph about the apprpriate mean function for these data.

Solution:
Using scatterplot we get the following plot,

Figure 1: Boys vs Girls Predicted Height in Centimeters



Code:

```
df = BGSall
scatterplot(HT18 ~ HT9 , groups = df$Sex,
                         data = df ,
                         regLine = FALSE,
                         smooth = FALSE)
```

As expected we can see that if we fitted the straight line mean function to the data, the boys average height would be greater than the girls. Looking at the data I'd imagine that fitting an SLR to each data we would get very similar slope coefficients and a significant difference in intercepts.

5.14.2 Obtain the appropriate test for a parallel regression model.

Solution:
To see if a parallel regression model is sufficient for this data we need to test the significance of the interaction term of the general model. Using the Type-2 Partial F test(Anova) we get that the interaction term is significant on the $alpha = .05$ level. Looking at the data it seems that a parallel model would be sufficient but it doesn't hurt to include the interaction term.
Code:

```
df$Sex <- factor(df$Sex)
Anova(lm(HT18 ~ HT9 * Sex, data = df))
Anova Table (Type II tests)

Response: HT18
          Sum Sq  Df  F value   Pr(>F)
HT9       3740.5   1 322.1883 < 2e-16 ***
Sex       4624.0   1 398.2872 < 2e-16 ***
HT9:Sex     34.4   1   2.9638 0.08749 .
Residuals 1532.5 132
```

5.14.3 Assuming the parallel regression model is adequate, estimate a 95 percent confidence interval for the difference between males and females. For the parallel repression model, this is the difference in the intercepts of the two groups.

Solution:
As stated previosly the difference between males and females in the data is encoded in the Sex coefficient of the parallel regression model. We can see this by setting all other predictors to zero and computing the intercept for both males and females, as expected the difference is the coefficient of the Sex predictor. To compute the difference we simply need to find the confidence interval for that regresison coefficient.
Code:

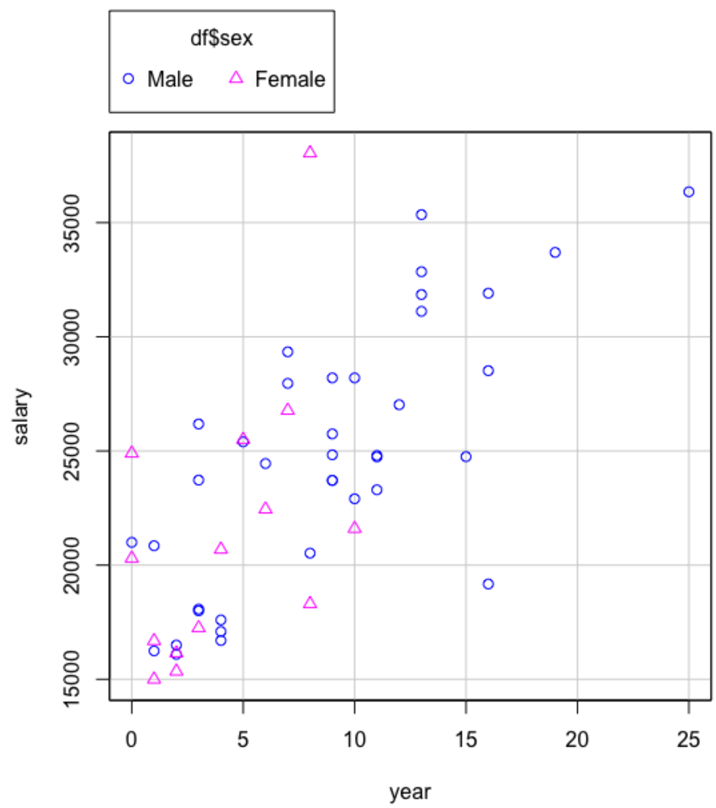
```
confint(lm(HT18 ~ HT9 + Sex, data = df), 'Sex', level = .95)
          2.5 %     97.5 %
Sex  -12.86355  -10.52813
```

Exercise 2: do Proble, 5.17. In part 1 all you need to do is get the scatterplot between salary and year, with different plotting symbol for the levels of sex. In part 2, use a simple two-sample t-test. In part 3 use the parallel regression model. Skip part 4.

5.17.1 Get appropriate graphical summaries of the data and discuss the graphs.

Solution:
Plotting salary as the response with year and sex as predictors we get the following,

Figure 2: Salaries vs Tenure for Both Male and Female Employees



Code:

```
df <- salary
scatterplot(salary ~ year , groups = df$sex,
                            data = df ,
                            regLine = FALSE,
                            smooth = FALSE)
```

From the scatterplot we can see that generally there are fewer female employees. The female employees also have a significant earnings ceiling, when compared to male earnings. A majority of female employees are below the 27,000 dollar earnings, while there seems to be a a significant proportion of male employees which have higher earnings that that.

5.17.2 Test the hypothesis that the mean salary for men and women is the same. What alternative hypothesis do you think is appropriate.

Solution:
Performing a two-sample t-test using the data, we suppose that the null hypothesis is that there is no difference in the mean salaries of male and female employees, and the alternative hypothesis is that the mean of male employee salaries are greater than female employees. Subsetting the data and performing the simple two-sample t-test we get that we reject the null will a p-value of .0353 and conclude that on the $\alpha = .05$ significance level the mean male salary is greater than the mean female salary.
Code:

```
> femaleSalaries <- df[(df$sex == 'Female'),]$salary
> maleSalaries <- df[(df$sex == 'Male'),]$salary

> t.test(maleSalaries, femaleSalaries, alternative = "greater",
                                       var.equal = TRUE)

        Two Sample t-test

data:  maleSalaries and femaleSalaries
t = 1.8474, df = 50, p-value = 0.0353
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 310.087      Inf
sample estimates:
mean of x mean of y
 24696.79  21357.14
```

5.17.3 Assuming no interactione between sex and other predictors, obtain a 95 percent confidence interval for the difference in salary between males and females.

Solution:
Proceeding similary to the previous problem, we need to find a confidence interval for the sex predictor coefficient of the parallel regression.
Code:

```
df$sex <- factor(df$sex, ordered = FALSE)
confint(lm(salary ~ year + sex, data = df), 'sexFemale', level = .95)
              2.5 %   97.5 %
sexFemale -2722.757 3125.69
```

Exercise 3: Use the Wool data from 5.19. Turn the three predictors len, amp, and load into factors and use log(cycles) as the response insted of cycles. Do the following:

a. Fit the model for log(cycles) using the three main effects and the three two-way interactions; report the type-II sums of squares Anova table. Which main effects and which interactons would you keep in the model based on $\alpha = .05$

Solution:
Fitting the model in r we get the following,
Code:

```
df <- Wool
df$len <- factor(df$len, order = FALSE)
df$load <- factor(df$load, order = FALSE)
df$amp <- factor(df$amp, order = FALSE)
df$cycles <- log(df$cycles)

summary(lm(cycles ~ len + amp + load +
                    len:amp + len:load + amp:load,
                    data = df))

Call:
lm(formula = cycles ~ len + amp + load +
                      len:amp + len:load + amp:load,
                      data = df)

Residuals:
     Min        1Q    Median        3Q       Max
-0.12779  -0.05537  -0.01802   0.06325   0.15780

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.362917   0.120807  52.670 1.87e-11 ***
len300          0.913780   0.151801   6.020 0.000316 ***
len350          1.963516   0.151801  12.935 1.21e-06 ***
amp9           -0.413379   0.151801  -2.723 0.026121 *
amp10          -1.203298   0.151801  -7.927 4.67e-05 ***
load45         -0.375588   0.151801  -2.474 0.038457 *
load50         -0.609676   0.151801  -4.016 0.003861 **
len300:amp9    -0.001114   0.166290  -0.007 0.994817
len350:amp9    -0.614678   0.166290  -3.696 0.006074 **
len300:amp10    0.064964   0.166290   0.391 0.706242
len350:amp10   -0.152966   0.166290  -0.920 0.384537
len300:load45   0.083463   0.166290   0.502 0.629248
len350:load45   0.145059   0.166290   0.872 0.408448
len300:load50  -0.133655   0.166290  -0.804 0.444766
```

```
len350:load50  -0.273658   0.166290  -1.646 0.138450
amp9:load45    -0.074416   0.166290  -0.448 0.666379
amp10:load45   -0.003211   0.166290  -0.019 0.985067
amp9:load50    -0.035285   0.166290  -0.212 0.837264
amp10:load50   -0.084089   0.166290  -0.506 0.626717
---------------------------------------------------------
Residual standard error: 0.144 on 8 degrees of freedom
Multiple R-squared:  0.9928,    Adjusted R-squared:  0.9768
F-statistic: 61.71 on 18 and 8 DF,  p-value: 1.236e-06
```

Generating the type-II Anova table we can see that based on an $\alpha = .05$ significance level we might want to consider dropping the len:load interaction as well as the amp:load interaction.

Code:

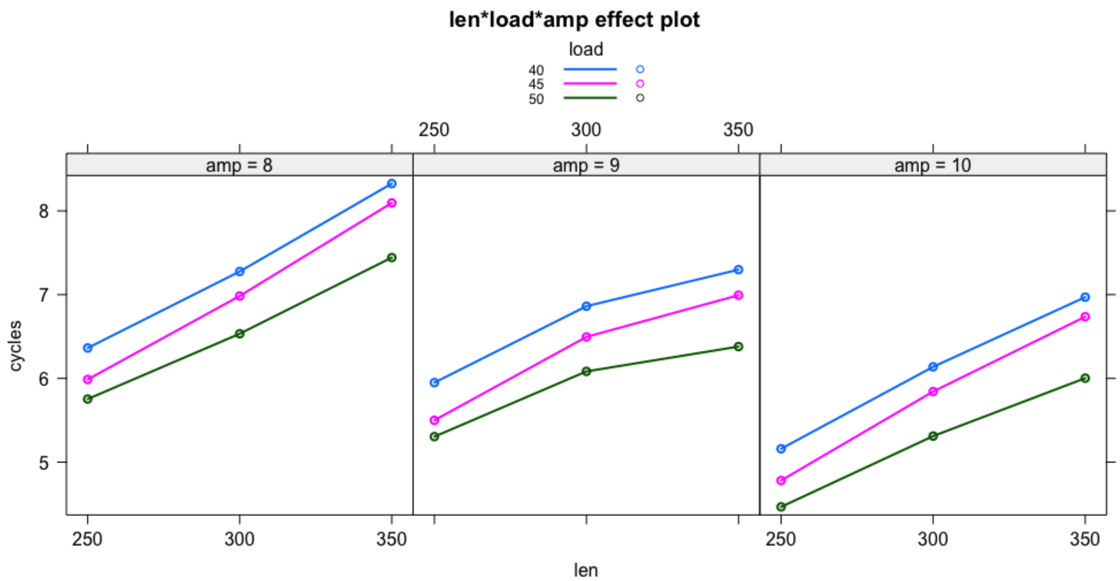
```
Anova(lm(cycles ~ len + amp + load +
                  len:amp + len:load + amp:load,
                  data = df))

Anova Table (Type II tests)
Response: cycles
           Sum Sq Df  F value    Pr(>F)
len       12.5159  2 301.7441 2.930e-08 ***
amp        7.1674  2 172.7986 2.620e-07 ***
load       2.8019  2  67.5509 9.767e-06 ***
len:amp    0.4012  4   4.8357   0.02806 *
len:load   0.1358  4   1.6364   0.25620
amp:load   0.0146  4   0.1760   0.94456
Residuals  0.1659  8
------------------------------------------------
```

b. Produce the effects plot gor the full second-order model fit in part a.

Solution:
Using r we can produce the effect plot for the full second-order model. Doing so we get,

Figure 3: len, amp, and load Second Order Effect Plot



Code:

```
Model <- lm(cycles ~ len + amp + load +
                     len:amp + len:load + amp:load,
                     data = df)

plot(Effect(c('len', 'load', 'amp'), Model), multiline = TRUE)
```

c. Obtain estimates fo the level means of amp in the model that only contains main effects using emmeans().

Solution:
Fitting the effects only model, and using the emmeans function we get,
Code:

```
model <- lm(cycles ~ len + amp + load, data = df)
emmeans(model, 'amp')

 amp  emmean     SE df lower.CL upper.CL
 8      6.97 0.0631 20     6.84     7.11
 9      6.32 0.0631 20     6.19     6.45
 10     5.71 0.0631 20     5.58     5.84

Results are averaged over the levels of: len, load
Confidence level used: 0.95
```