Test Exercise 1

Jader Martins

January 5, 2021

(a) Use all data to estimate the coefficients a and b in a simple regression model, where expenditures is the dependent variable and age is the explanatory factor. Also compute the standard error and the t-value of b.

```
holiday.lm <- lm("Expenditures ~ Age", data)
summary(holiday.lm)
```

```
##
## Call:
## lm(formula = "Expenditures ~ Age", data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8965 -4.2301 -0.8984  4.3525  7.7739
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.24111    3.88208  29.428  < 2e-16 ***
## Age          -0.33360    0.09537  -3.498  0.00185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.073 on 24 degrees of freedom
## Multiple R-squared:  0.3377, Adjusted R-squared:  0.3101
## F-statistic: 12.24 on 1 and 24 DF,  p-value: 0.001852
```

Our a is 114.24 and our b is -0.33.

(b) Make the scatter diagram of expenditures against age and add the regression line y = a + bx of part (a) in this diagram. What conclusion do you draw from this diagram?

```
ggplot(data, aes(x=Age, y=Expenditures)) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x, se=F)
```

It seems that we have a Simpson's paradox problem, so a linear regression in the entire data does not fits well.

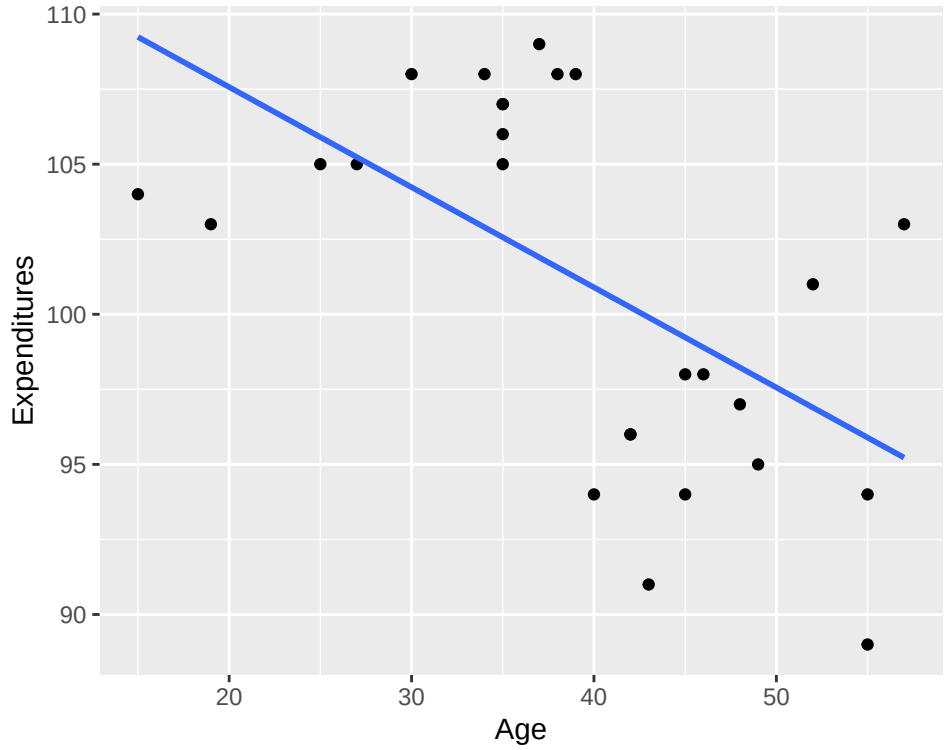

Figure 1: Regression diagnosis plot.

(c) It seems there are two sets of observations in the scatter diagram, one for clients aged 40 or higher and another for clients aged below 40. Divide the sample into these two clusters, and for each cluster estimate the coefficients a and b and determine the standard error and t-value of b.

```
b <- data[data$Age < 40,]
holiday.lm <- lm("Expenditures ~ Age", b)
summary(holiday.lm)
```

```
##
## Call:
## lm(formula = "Expenditures ~ Age", data = b)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1613 -0.5775 -0.1613  0.7982  1.8286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 100.23228    1.41590   70.79 5.55e-16 ***
## Age           0.19797    0.04438    4.46 0.000962 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.153 on 11 degrees of freedom
```

```
## Multiple R-squared:  0.644,  Adjusted R-squared:  0.6116
## F-statistic:  19.9 on 1 and 11 DF,  p-value: 0.0009619
```

```
a <- data[data$Age >= 40,]
holiday.lm <- lm("Expenditures ~ Age", a)
summary(holiday.lm)
```

```
##
## Call:
## lm(formula = "Expenditures ~ Age", data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9278 -1.4631  0.9763  2.3905  5.7793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  88.8719     9.4585   9.396 1.37e-06 ***
## Age           0.1465     0.1974   0.742    0.474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.833 on 11 degrees of freedom
## Multiple R-squared:  0.04767,    Adjusted R-squared:  -0.0389
## F-statistic: 0.5507 on 1 and 11 DF,  p-value: 0.4736
```

(d) Discuss and explain the main differences between the outcomes in parts (a) and (c). Describe in words what you have learned from these results.

Below 40 years, the Age explains well the variability in Expenditures, but above 40 year, it does not seems to be correlated. Another interesting aspect is that an Simpson's paradox occur in the entire data and as a result we have a negative correlation between Age and Expenditure, which is counterintuitive, but if we split the data we obtain the expected positive correlation.