Moderation

Moderation occurs when the relationship between two variables depends on a third variable.

- The third variable is referred to as the moderating variable or simply the moderator.
- The moderator affects the direction and/or strength of the relation between your explanatory and response variable.
- When testing a potential moderator, we are asking the question whether there is an association between two constructs, **but separately for different subgroups within the sample.**
  - This is also called a *stratified* model, or a *subgroup analysis.*

How to determine if a variable is a moderator

- What you are looking for first depends on which of the 3 scenarios below describe your original analysis (i.e., your original ANOVA test).
- Whether your third variable is a moderator depends on what you see happening in your moderator analysis (i.e., the second ANOVA test split by your third variable).

- If ANY of the 3 scenarios explained below occur in your analysis then your Third Variable IS a Moderator of the Bivariate Relationship.

Scenario 1 - Significant relationship at bivariate level (i.e., ANOVA, Chi-Square, Correlation) (saying expect the effect to exist in the entire population) then when test for moderation the third variable is a moderator if the strength (i.e., p-value is Non-Significant) of the relationship changes. Could just change strength for one level of third variable, not necessarily all levels of the third variable.

Scenario 2 - Non-significant relationship at bivariate level (i.e., ANOVA, Chi-Square, Correlation) (saying do not expect the effect to exist in the entire population) then when test for moderation the third variable is a moderator if the relationship becomes significant (saying expect to see it in at least one of the sub-groups or levels of third variable, but not in entire population because was not significant before tested for moderation). Could just become significant in one level of the third variable, not necessarily all levels of the third variable.

Scenario 3 - Significant relationship at bivariate level (i.e., ANOVA, Chi-Square, Correlation) (saying expect the effect to exist in the entire population) then when test for moderation the third variable is a moderator if the direction (i.e., means change order/direction) of the relationship changes. Could just change direction for one level of third variable, not necessarily all levels of the third variable.

What to look for in each type of analysis.

ANOVA - look at the p-value, r-squared, means, and the graph of the ANOVA and compare to those values in the Moderation (i.e., each level of third variable) output to determine if third variable is a moderator or not.

Chi-Square - look at the p-value, the percents for the columns in the cross tabs table, and the graph for the Chi-Square and compare to those values in the Moderation (i.e., each level of third variable) output to determine if third variable is a moderator or not.

Correlation - look at the correlation coefficient (r), p-value, calculate the r-squared, and the graph for the Correlation and compare to those values in the Moderation (i.e., each level of third variable) output to determine if third variable is a moderator or not.

Examples

```
library(knitr)
opts_chunk$set(warning=FALSE, message=FALSE)
library(dplyr)
library(ggplot2); library(gridExtra)
library(pander) # Used for printing nice linear model tables
panderOptions("digits", 3)
iris<-iris # only used to run sample code
load("C:/Box Sync/Data/AddHealth/addhealth_clean.Rdata")
```

ANOVA

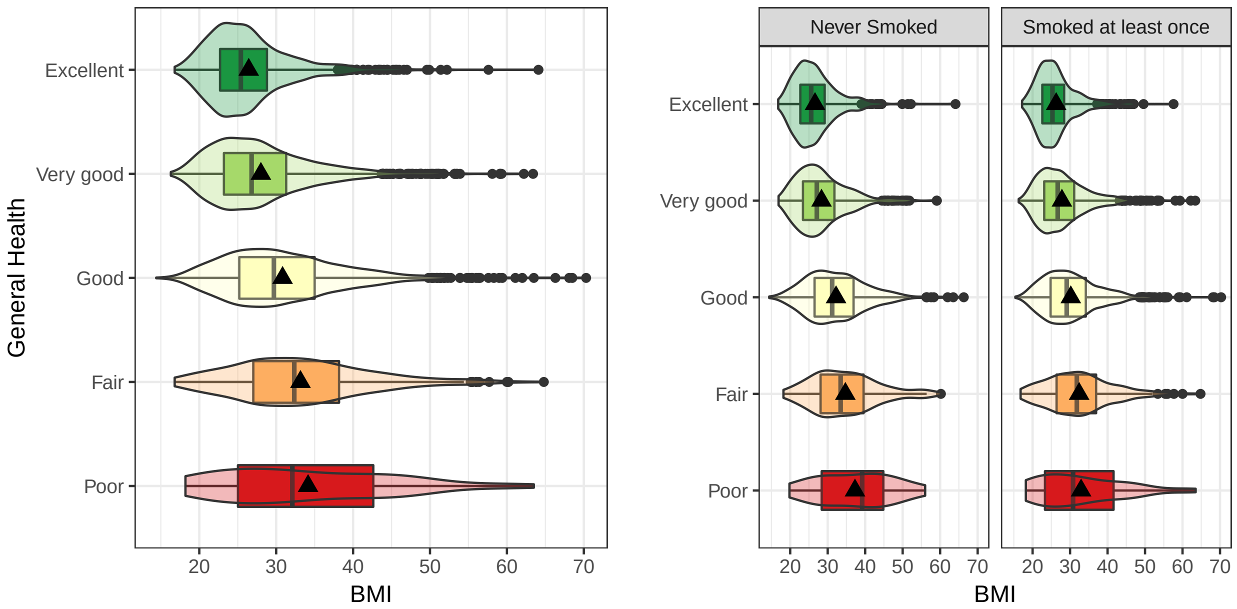
1. **Identify response, explanatory, and moderating variables**

- Categorical explanatory variable = Perceived General Health (variable `genhealth`)
- Quantitative response variable = Body Mass Index (variable `BMI`)
- Categorical Potential Moderator = Ever smoked (variable `eversmoke_c`)

2. **Visualize and summarise the potential effect of the moderator**

These plots are modified for readability in several ways. These modifications may not apply to your situation. When running into trouble start taking off the fancy layers. 1. `scale_x_discrete` is used to reverse the display order of the limits on general health. 2. `coord_flip` is used to turn the boxplots horizontal, thereby removing the overlapping category labels. 3. `scale_fill_brewer` is used to apply a red-yellow-green color palette to the categories of general health. This is applicable because general health is an ordinal variable that has an interpretation of "good" to "bad".

```
bmi.plot <- addhealth %>% select(genhealth, BMI, eversmoke_c) %>% na.omit()

twoway.boxplot <- ggplot(bmi.plot, aes(x=genhealth, y=BMI, fill=genhealth)) +
                  geom_boxplot(width=.4) + geom_violin(alpha=.3) +
                  stat_summary(fun.y="mean", geom="point", size=3, pch=17,
                  position=position_dodge(width=0.75)) + theme_bw() +
                  labs(x="General Health", y="BMI") +
                  scale_x_discrete(limits=rev(levels(bmi.plot$genhealth)))+
                  scale_fill_brewer(guide=FALSE, palette="RdYlGn", direction=-1) +
                  theme(legend.position = "top")+ coord_flip()

threeway.boxplot <- ggplot(bmi.plot, aes(x=genhealth, y=BMI, fill=genhealth)) +
                  geom_boxplot(width=.4) + geom_violin(alpha=.3) +
                  stat_summary(fun.y="mean", geom="point", size=3, pch=17,
                  position=position_dodge(width=0.75)) + theme_bw() +
                  facet_wrap(~eversmoke_c)+
                  labs(x="", y="BMI") +
                  scale_x_discrete(limits=rev(levels(bmi.plot$genhealth)))+
                  scale_fill_brewer(guide=FALSE, palette="RdYlGn", direction=-1) +
                  theme(legend.position = "top")+ coord_flip()

grid.arrange(twoway.boxplot, threeway.boxplot, ncol=2)
```




```
bmi.plot %>% group_by(eversmoke_c, genhealth) %>%
             summarise(mean=mean(BMI)) %>%
             tidyr::spread(genhealth, mean) %>%
              kable(digits=1, caption = "Average BMI per general health category, for smokers and non sm
```

Table 1: Average BMI per general health category, for smokers and non smokers separately

eversmoke_c	Excellent	Very good	Good	Fair	Poor
Never Smoked	26.6	28.4	32.2	34.7	37.3
Smoked at least once	26.3	27.8	30.2	32.4	32.9

note the code above uses the `spread` function in `tidyr` package to turn the table of means on it's side

The average BMI for those reporting excellent health is basically the same between smokers and non-smokers. As perceived general health decreases, the average BMI increases, but it seems to do so at a faster rate in non-smokers compare to smokers. There seems to be more high end outlying points that are smokers compared to non-smokers. Otherwise the relationship between BMI and general health appears the same within each smoking group.

3. **Write the relationship you want to examine in the form of a research question - including a statement about the modifier**

Does ever smoking change the relationship between perceived general health and BMI? Is the distribution of BMI the same across perceived general health status, for both smokers and non-smokers?

4. **Fit both the original, and stratified models.**

The `pander()` function prints these as nice tables. This is not required, just recommended.

```
twoway.anova <- summary(aov(BMI ~ genhealth, data=addhealth))
pander(twoway.anova, caption="Original model")
```

Table 2: Original model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genhealth	4	22563	5641	109	2.05e-89
Residuals	5037	260096	51.6	NA	NA

```
pander(summary(aov(BMI ~ genhealth,
                   data=filter(addhealth, eversmoke_c=="Smoker"))),
       caption="Model for Smokers")
```

Quitting from lines 126-133 (lec04_moderation.Rmd) Error in `contrasts<-(*tmp*`, value = contr.funs[1 + isOF[nn]]) : contrasts can be applied only to factors with 2 or more levels Calls: … model.matrix -> model.matrix.default -> contrasts<-

```
pander(summary(aov(BMI ~ genhealth,
                   data=filter(addhealth, eversmoke_c=="Non Smoker"))),
       caption="Model for Non-Smokers")
```

Quitting from lines 126-133 (lec04_moderation.Rmd) Error in `contrasts<-(*tmp*`, value = contr.funs[1 + isOF[nn]]) : contrasts can be applied only to factors with 2 or more levels Calls: … model.matrix -> model.matrix.default -> contrasts<-

5. **Determine if the Third Variable is a moderator or not.**

Both the original ANOVA and the stratified ANOVA models for smokers and non-smokers separately are highly significant. There is not a clear difference in the relationship between BMI and general health status between smokers and non-smokers, so ever being a smoker is not a moderating variable for this relationship.

Chi-Square

1. **Identify response and explanatory variables**

- Categorical explanatory variable = Gender (variable `female_c`)
- Categorical response variable = General Health (variable `genhealth`)
- Categorical Potential Moderator = Ever smoked (variable `eversmoke_c`)

2. **Visualize and summarise the potential effect of the moderator**

Set up the two and three way tables for printing, and as a data frame for plotting.

For the two way table, I put `female_c` as rows and `genhealth` as columns. So when calculating proportions of females within each general health category, I want the column margins (`margin=2`).

```
crosstab.gender.genhealth <- table(addhealth$female_c, addhealth$genhealth)
proptab.gender.genhealth  <- prop.table(crosstab.gender.genhealth, margin=2)
plot.crostab <- data.frame(proptab.gender.genhealth)
kable(proptab.gender.genhealth, digits=2)
```

	Excellent	Very good	Good	Fair	Poor
Male	0.52	0.46	0.45	0.4	0.31
Female	0.48	0.54	0.55	0.6	0.69

Here I create a three way table: `female_c` as rows (1), `genhealth` as columns (2), and then stratified by eversmoke_c (3).

```
freq.gend.health.smoke <- table(addhealth$female_c, addhealth$genhealth, addhealth$eversmoke_c)
```

To get the correct column proportions within each level of eversmoke_c, I put c(3,2) for the margin argument: calculate the proportions within variable (3), and then across (2).

```
prop.gend.health.smoke <- prop.table(freq.gend.health.smoke, margin=c(3,2))
plot.threeway <- data.frame(prop.gend.health.smoke)
kable(prop.gend.health.smoke[,,1], digits=2, caption="General Health by Gender - for Non Smokers")
```

Table 4: General Health by Gender - for Non Smokers

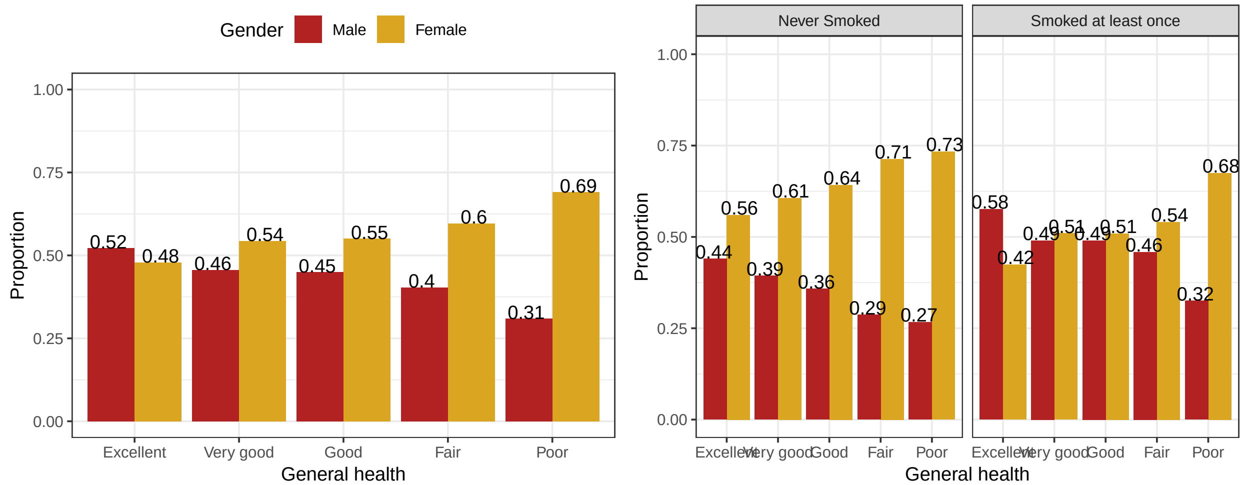
	Excellent	Very good	Good	Fair	Poor
Male	0.44	0.39	0.36	0.29	0.27
Female	0.56	0.61	0.64	0.71	0.73

```
kable(prop.gend.health.smoke[,,2], digits=2, caption="General Health by Gender - for Smokers")
```

Table 5: General Health by Gender - for Smokers

	Excellent	Very good	Good	Fair	Poor
Male	0.58	0.49	0.49	0.46	0.32
Female	0.42	0.51	0.51	0.54	0.68

```
biv.crosstab <- ggplot(plot.crostab, aes(x=Var2, y=Freq, fill=Var1)) +
              geom_col(position = position_dodge()) + theme_bw() +
              geom_text(aes(y=Freq+.02, label=round(Freq,2)),
                        position = position_dodge(width=1)) +
              labs(x="General health", y="Proportion") +
               scale_fill_manual(name="Gender",
                                 values=c("firebrick", "goldenrod")) +
              theme(legend.position = "top") +
              scale_y_continuous(limits=c(0,1))

threeway.crosstab <- ggplot(plot.threeway, aes(x=Var2, y=Freq, fill=Var1)) +
              geom_col(position = position_dodge()) + theme_bw() + facet_wrap(~Var3) +
              geom_text(aes(y=Freq+.02, label=round(Freq,2)), position = position_dodge(width=1)) +
              labs(x="General health", y="Proportion") +
               scale_fill_manual(guide=FALSE, values=c("firebrick", "goldenrod")) +
              scale_y_continuous(limits=c(0,1))

grid.arrange(biv.crosstab,threeway.crosstab, ncol=2)
```

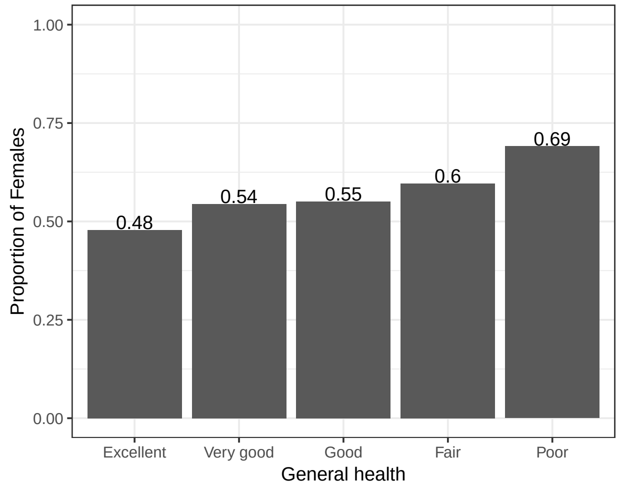


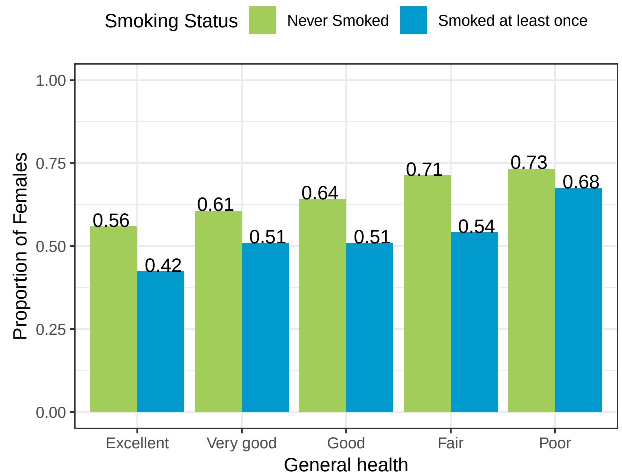
The gender difference across levels of general health seems to be due to the non-smokers. With the exception of those reporting excellent, or poor health, the proportions of male and females seem equal within each general health category for the smokers. For the non-smokers, as perceived general health worsens, the proportion of females in that category increases.

There is likely reason to believe that smokers have a different view on their general health compared to non-smokers, and this viewpoint is pretty similar between males and females.

Since this our explanatory variable is only two levels we can plot the proportion of "yes" values - here it would be females only. This makes for a more easily understood plot. Additional color names can be found here: http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

```
female.only_2 <- filter(plot.crostab, Var1=="Female")

plot2 <- ggplot(female.only_2, aes(x=Var2, y=Freq)) +
              geom_col(position = position_dodge()) + theme_bw() +
              geom_text(aes(y=Freq+.02, label=round(Freq,2)),
                    position = position_dodge(width=1)) +
              labs(x="General health", y="Proportion of Females") +
              scale_y_continuous(limits=c(0,1))

female.only_3 <- filter(plot.threeway, Var1=="Female")

plot3 <- ggplot(female.only_3, aes(x=Var2, y=Freq, fill=Var3)) +
              geom_col(position = position_dodge()) + theme_bw() +
              geom_text(aes(y=Freq+.02, label=round(Freq,2)), position = position_dodge(width=1)) +
              labs(x="General health", y="Proportion of Females") +
              theme(legend.position = "top") +
              scale_fill_manual(name="Smoking Status", values=c("darkolivegreen3", "deepskyblue3")) +
              scale_y_continuous(limits=c(0,1))

grid.arrange(plot2,plot3, ncol=2)
```





3. **Write the relationship you want to examine in the form of a research question - including a statement about the modifier**

Does ever smoking change the relationship between gender and general health? Is the distribution of gender (proportion of females) equal across all levels of general health, for both smokers and non smokers?

4. **Fit both the original, and stratified models.**

```
chisq.test(addhealth$female_c, addhealth$genhealth)
```

```
##
##  Pearson's Chi-squared test
##
## data:  addhealth$female_c and addhealth$genhealth
## X-squared = 26.471, df = 4, p-value = 2.543e-05
```

```
by(addhealth, addhealth$eversmoke_c, function(x) chisq.test(x$female_c, x$genhealth))
```

```
## addhealth$eversmoke_c: Never Smoked
##
##  Pearson's Chi-squared test
##
## data:  x$female_c and x$genhealth
## X-squared = 13.438, df = 4, p-value = 0.009324
##
## --------------------------------------------------------
## addhealth$eversmoke_c: Smoked at least once
##
##  Pearson's Chi-squared test
##
## data:  x$female_c and x$genhealth
## X-squared = 20.994, df = 4, p-value = 0.0003175
```

5. **Determine if the Third Variable is a moderator or not.**

The relationship between gender and general health is significant in both the main effects and the stratified model. The distribution of females across general health categories differs greatly between smokers and non-smokers. This fits into **Scenario 3**, so Smoking is a significant moderator.

Linear Regression

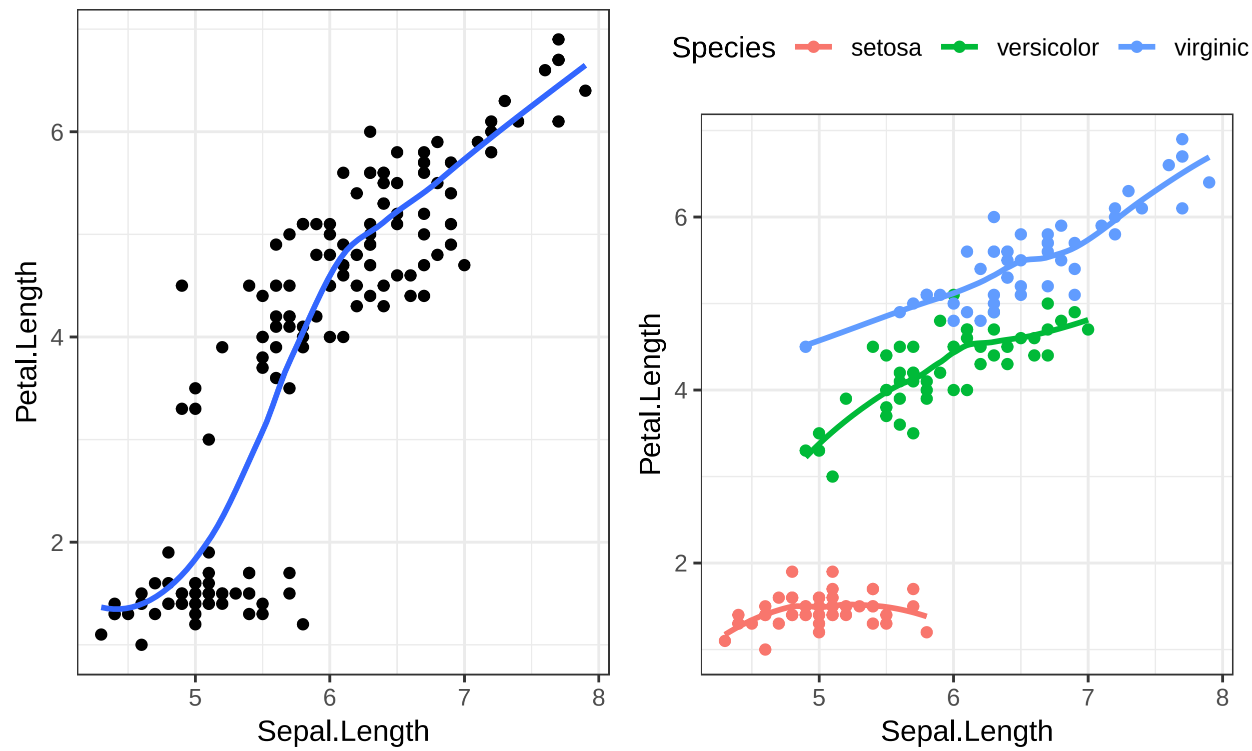
1. **Identify response, explanatory, and moderating variables**

- Quantitative explanatory variable = Sepal Length (variable `Sepal.Length`)
- Quantitative response variable = Petal Length (variable `Petal.Length`)
- Categorical Potential Moderator = Species (variable `Species`)

2. **Visualize and summarise the potential effect of the moderator**

Plot your original bivariate plot on the left,

```
linreg.plot <- ggplot(iris, aes(x=Sepal.Length, y=Petal.Length)) +
                geom_point() + geom_smooth(se=FALSE) +
                theme_bw()

linreg.modifier.plot <- ggplot(iris,
                            aes(x=Sepal.Length, y=Petal.Length, col=Species)) +
                          geom_point() + geom_smooth(se=FALSE) +
                          theme_bw() + theme(legend.position="top")

grid.arrange(linreg.plot, linreg.modifier.plot , ncol=2)
```




```
cor(iris$Sepal.Length, iris$Petal.Length)
```

```
## [1] 0.8717538
```

```
by(iris, iris$Species, function(x) cor(x$Sepal.Length, x$Petal.Length))
```

```
## iris$Species: setosa
## [1] 0.2671758
## --------------------------------------------------------
```

```
## iris$Species: versicolor
## [1] 0.754049
## ------------------------------------------------------------
## iris$Species: virginica
## [1] 0.8642247
```

There is a strong, positive, linear relationship between the sepal length of the flower and the petal length when ignoring the species. The correlation coefficient r for *virginica* and *veriscolor* are similar to the overall r value, 0.86 and 0.75 respectively compared to 0.87. However the correlation between sepal and petal length for species *setosa* is only 0.26.

The points are clearly clustered by species, the slope of the lowess line between *virginica* and *versicolor* appear similar in strength, whereas the slope of the line for *setosa* is closer to zero. This would imply that petal length for *Setosa* may not be affected by the length of the sepal.

3. **Write the relationship you want to examine in the form of a research question - including a statement about the modifier**

Does the species of the flower change or modify the linear relationship between the length of the flower's sepal and it's petal?

4. **Fit both the original, and stratified models.**

Fit the Main effects model, print the regression table of coefficients. The `pander()` function prints these as nice tables. This is not required, just recommended.

```r
iris.linear.model <- lm(Petal.Length ~ Sepal.Length, data=iris)
pander(iris.linear.model)
```

Table 6: Fitting linear model: Petal.Length ~ Sepal.Length

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.1	0.507	-14	6.13e-29
Sepal.Length	1.86	0.0859	21.6	1.04e-47

Fit the stratified model, print the regression table of coefficients.

```r
setosa.model     <- lm(Petal.Length ~ Sepal.Length, data=filter(iris, Species=="setosa"))
veriscolor.model <- lm(Petal.Length ~ Sepal.Length, data=filter(iris, Species=="versicolor"))
virginica.model  <- lm(Petal.Length ~ Sepal.Length, data=filter(iris, Species=="virginica"))
pander(setosa.model)
```

Table 7: Fitting linear model: Petal.Length ~ Sepal.Length

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.803	0.344	2.34	0.0238
Sepal.Length	0.132	0.0685	1.92	0.0607

```r
pander(veriscolor.model)
```

Table 8: Fitting linear model: Petal.Length ~ Sepal.Length

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.185	0.514	0.36	0.72
Sepal.Length	0.686	0.0863	7.95	2.59e-10

	Estimate	Std. Error	t value	Pr(>|t|)

```r
pander(virginica.model)
```

Table 9: Fitting linear model: Petal.Length ~ Sepal.Length

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.61	0.417	1.46	0.15
Sepal.Length	0.75	0.063	11.9	6.3e-16

5. **Determine if the Third Variable is a moderator or not.**

The estimate for sepal length in the original model is 1.85, p-value <.0001. - For *setosa* the estimate is 0.13, pvalue of 0.06 - For *versicolor* the estimate is 0.69, pvalue <.0001 - For *virginica* the estimate is 0.75, pvalue <.0001

Within the *setosa* species, there is little to no relationship between sepal and petal length. For *veriscolor* and *virginica* the relationship is still signficantly positive. This is **Scenario 1**, so Species moderates the effect of sepal length on petal length.