PM592: Regression Analysis for Health Data Science
Lab 6 – Confounding & Interactions

Data Needed: *subjdata*

This lab is devoted entirely to the exercise.

Lab 6 Exercises

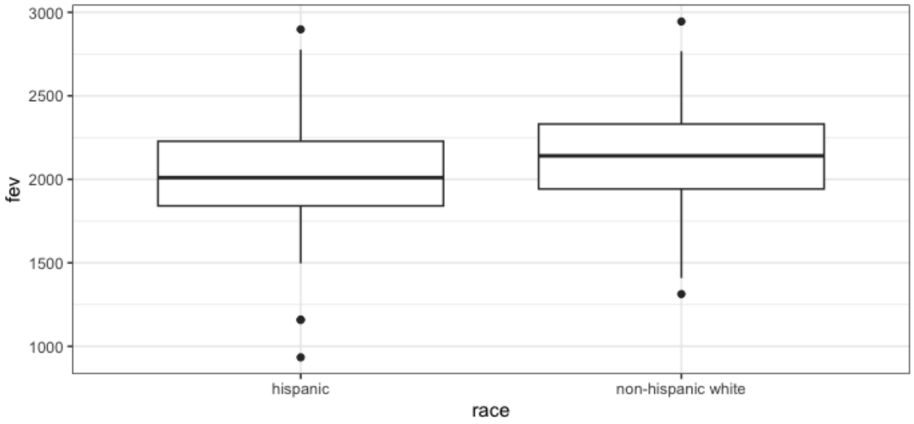
Objective(s):	Determine the presence of confounding and/or interactions in the data set, report the extent of confounding (if present), explain an interaction in terms of stratum-specific effects
Datasets Required:	`subjdata`

The data we are using for this lab includes children from three of the 12 CHS communities. One question we are interested in is whether certain racial/ethnic groups have differences in lung function, and why. We will examine the two largest ethnic groups: Hispanic and non-Hispanic white. Create a new data set that includes only Hispanic and non-Hispanic white subjects. The existing coding of the race variable in this data set is:

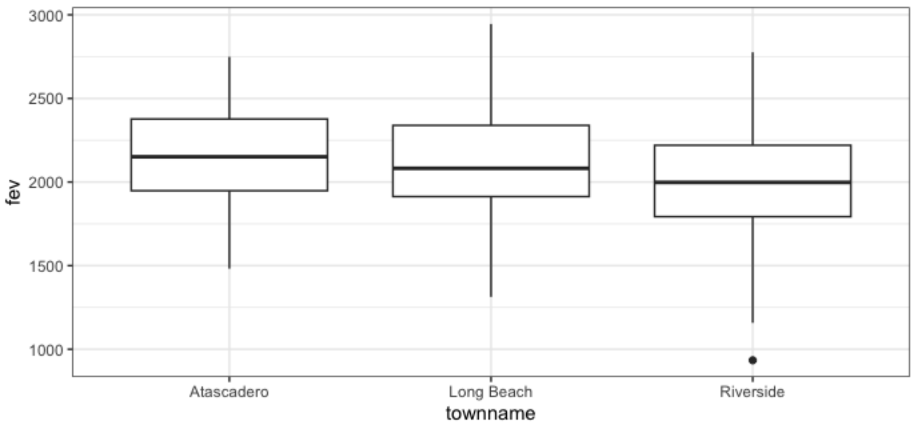
$$X_{RACE} = \begin{cases} 1, \text{Asian} \\ 2, \text{African American} \\ 3, \text{Hispanic} \\ 4, \text{non} - \text{Hispanic white} \\ 5, \text{Other} \end{cases}$$

1) Produce the following boxplots:

a) The association between FEV and race category (Hispanic vs. non-Hispanic white)



b) The association between FEV and community



c) Describe your impressions of these boxplots

It looks like Hispanics have lower mean FEV (2000) than non-hispanic whites (2150), and then mean FEV is highest for Atascadero, slightly lower than Long Beach, and lowest for Riverside.

2) Perform an analysis of the effect of race on lung function.

a) Is race associated with FEV in an unadjusted model? Explain the association.

```
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept)  2034.01      31.05  65.507   <2e-16 ***
race4        81.09      38.72   2.094   0.0371 *
---
```

Yes, race is associated with FEV in an unadjusted model (p=0.037). Mean FEV is expected to be 81.09 units higher for non-hispanic whites compared to Hispanics (the reference group).

b) Is race associated with FEV in a model adjusting for community?

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)          2117.25      48.08  44.039  < 2e-16 ***
race.f4              36.85      41.63   0.885  0.37685
townname.fLong Beach   -17.36      47.63  -0.365  0.71572
townname.fRiverside   -139.59      47.05  -2.967  0.00327 **
---
```

Adjusting for community, it looks like race is no longer associated with FEV (p=0.38).

c) How much did the slope estimate for race change in 2b? (Express this as a percent change.)

The slope estimate for race changed by (36.85 – 81.09) / 81.09 = 54.6%

d) Explain why you believe this parameter estimate did or did not change. (Just saying "it is/isn't a confounder" is not sufficient; explain what you think may be happening with the relationships under examination.)

I believe this parameter estimate changed because the community variable can sensibly be a cause of FEV. Communities can be located near major ports or freeways, which can have an effect on its population's FEV measurements. Additionally, there is probably a correlation between community and FEV, communities tend to have larger populations of certain races.

e) Calculate the percent of the sample that is Hispanic by town. What are your impressions?

Atascadero: 11%

Long Beach: 42%

Riverside: 55%

Riverside has the highest Hispanic population out of the three towns, which confirms the suspicions that community has an effect on the relationship between race and fev.

f) Based on your analysis in this question, write the equation of the model you would use to explain the association between race and lung function.

Since community has an effect on the relationship between race and fev, it should be included in the model:

$$\hat{Y} = \beta_0 + \beta_{NHW} X_{NHW} + \beta_{LB} X_{LB} + \beta_{RIV} X_{RIV}$$

3) Examine asthma as a potential confounder by including it in your model from 2f.

a) Does asthma appear to confound the relationship between race and FEV?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2018.55      31.76  63.565   <2e-16 ***
race.f4        69.72      38.44   1.814   0.0708 .
asthma        104.39      48.84   2.137   0.0334 *
---
```

(69.72 – 81.09) / 81.09 = 14% change. Asthma does seem to change the relationship between race and FEV when accounted for in the model.

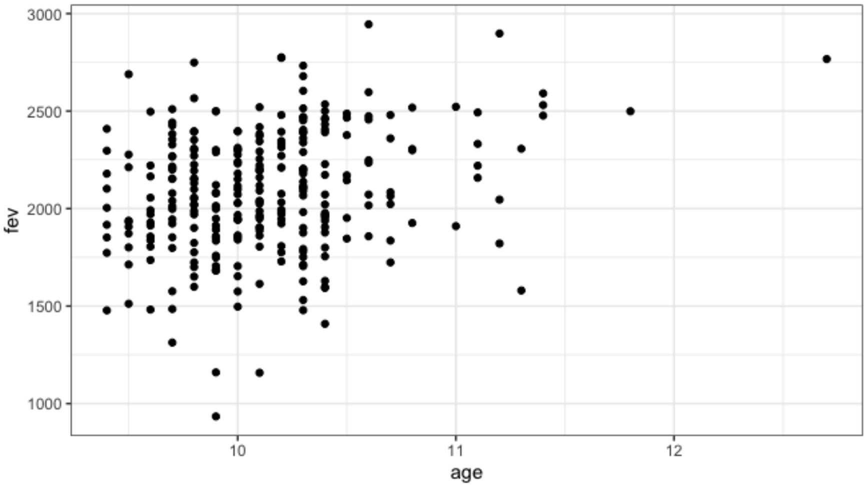
b) Can asthma sensibly be a confounder of this relationship?

Asthma can cause FEV to vary. However, asthma cannot sensibly be a cause of race. Therefore it cannot be a confounder in this relationship.

Another question we may want to examine is whether pollution inhibits the development of lung function over time. Levels of the pollutants ozone (measured in ppb), nitrogen dioxide (NO2, in ppb), and particulate matter (PM10, small bits of dust with diameter 10 microns or less, measured in micrograms per cubic meter, or µg/m3) were measured at a single central site within each community. The values stored in SUBJDATA represent the average level of each pollutant over a 1-year period. Since there was only one monitoring site for each town, every child in the same town was assigned the same exposure level for each pollutant.

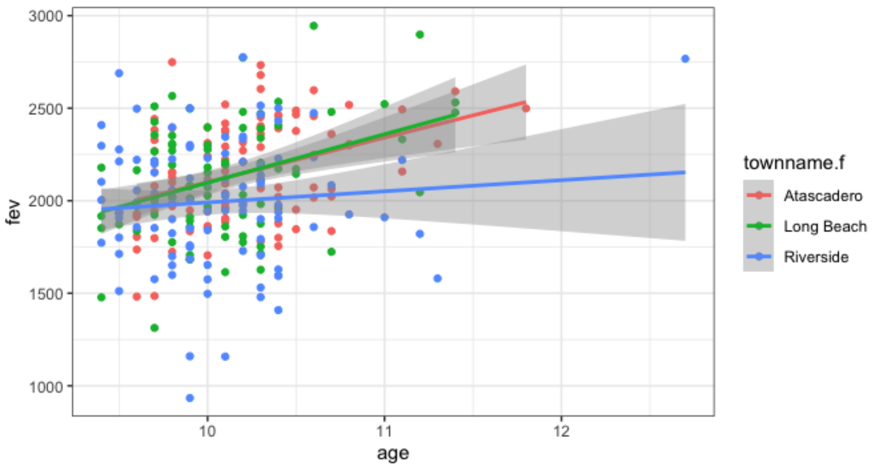
4) Analyze the overall and community-specific relationships between age and lung function.

   a) Provide a scatterplot of the overall relationship between age and FEV. What are your impressions?



   There does seem to be a slight positive relationship between age and fev.

   b) Provide a scatterplot of the relationship between age and FEV, specifying “color = factor(townname)” in the aesthetic. Based on this graph, do you believe there is an interaction between town and age on their relationship with FEV?



Based on this graph, it appears that Atascadero and Long Beach have similar relationships in age vs FEV, and Riverside has less of a strong relationship between age and fev than the other two.

c) Create Model 0: the regression of FEV on age and community. Is age related to FEV, adjusting for community? Provide this p-value, along with the best-fit equation for this model.

```
age            170.378    39.896   4.271 2.68e-05 ***
```

Yes, age is related to FEV (p=2.68e-05). The equation is:

$$\hat{Y} = 410.89 + 170.38(X_{age}) + -8.96(X_{LB}) + \ -130.37(X_R)$$

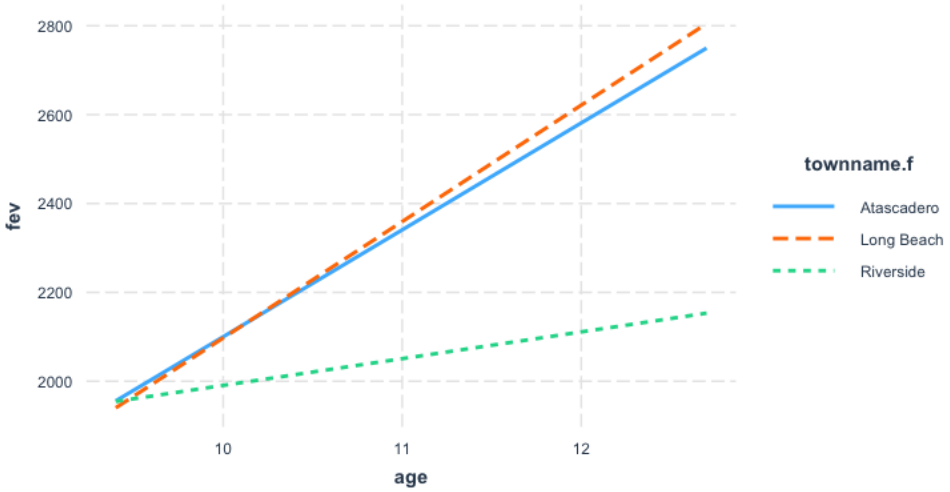
d) Create Model 1: the regression of FEV on age, community, and the age X community interaction. Provide the best-fit equation for this model.

$$\hat{Y} = -306 + 240(X_{age}) + \ -216(X_{LB}) + 1694(X_R) + 21(X_{age}X_{LB}) + \ -180(X_{age}X_R)$$

e) Compute a p-value for the overall effect of the interaction by using the anova() function on Model 0 vs. Model 1. Is the overall interaction statistically significant? (Remember: the alpha for interaction terms can sometimes be relaxed as interactions are typically underpowered.)

p=0.06, the p-value is below the relaxed alpha value of 0.15, so the interaction is worth investigating.

f) Use the interact_plot() function (pred = age, modx = townname) and the sim_slopes() function to determine the community-specific effects of age on FEV. Report the β_{AGE} terms for each community along with their p-values. Verify that these are the community-specific β_{AGE} terms you can get from your equation in 4d.



The slope of the regression of fev on age is lowest for Riverside, higher for Atascadero, and highest for Long Beach.

Slope of age when townname.f = Riverside:

```
  Est.    S.E.   t val.      p

------- ------- -------- ------

 60.20   61.01     0.99   0.32
```

Slope of age when townname.f = Long Beach:

```
   Est.    S.E.   t val.      p

-------- ------- -------- ------

 261.99   75.47     3.47   0.00
```

Slope of age when townname.f = Atascadero:

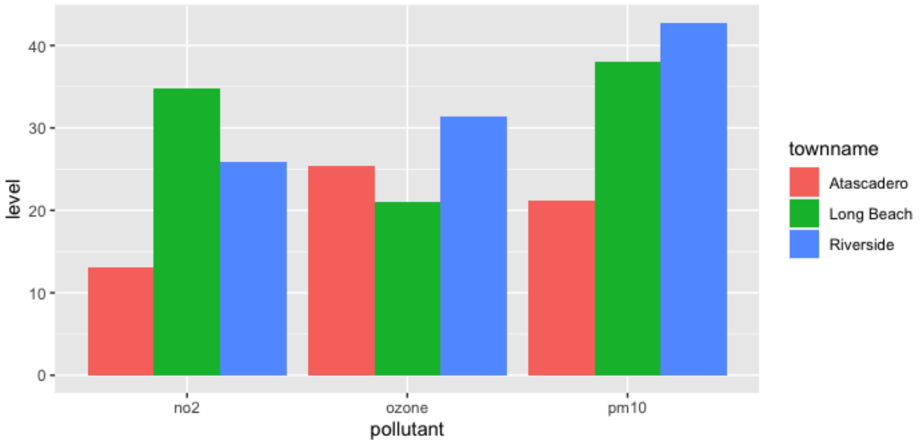
```
   Est.    S.E.   t val.      p

-------- ------- -------- ------
```

```
240.62   72.11    3.34   0.00
```

Let’s look at this question in a different way. Because all individuals in the same community were assigned the same values for pollution, there is no way that we can disentangle the effect of community from the effect of pollution. Instead of looking at community as an effect modifier (interaction), let’s look at each pollutant as an effect modifier.

5) Examine how pollution varies based on community.

   a) Use the code in the lab R document to produce a bar chart of pollution level by pollutant and town.

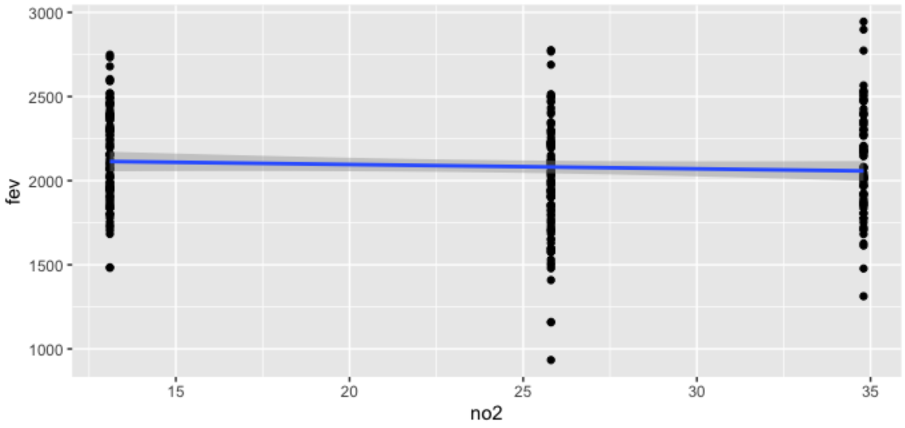


   b) How do the levels of each pollutant vary based on town?

   The ozone and pm10 pollutants seem to be highest for Riverside. Long Beach is highest in no2 pollution.

6) Choose one of the pollutants and perform a basic analysis. –No2

   a) Provide a scatter plot with a best-fit regression line of the relationship between FEV and the pollutant.



   b) Provide the results of a regression of FEV on pollutant – including the estimate of the slope, t-value, and p-value.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2147.945     54.009  39.770   <2e-16 ***
no2           -2.591      2.126  -1.219    0.224
---
```

Slope: -2.591

t-value: -1.219

p-value: 0.224

c) Provide a written interpretation of the slope.

The slope indicates that a one-unit increase in no2 levels is predicted to decrease FEV in an individual by 2.6.

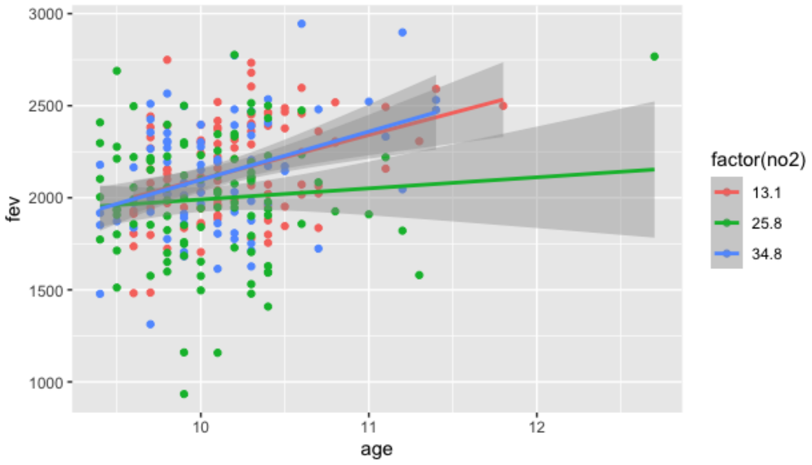
d) What is the estimated difference in mean FEV between the most and least polluted towns?

$$
\begin{aligned}
&2147.945 + \ -2.591(34.8) \\
&-\left(2147.945 + \ -2.591(13.1)\right) \\
&= \ -90.1668 \ - (-33.9421) \\
&= \ -56.2247
\end{aligned}
$$

The estimated difference in mean FEV between the most and least polluted towns is 56.2.

7) Assess effect modification.

a) Provide a scatterplot of the relationship between age and FEV, specifying “color = [your pollutant]” in the aesthetic. Based on this graph, do you believe there is an interaction between pollution and age on their relationship with FEV?



From this graph, it looks like the slope of the regression line for age vs fev is higher for no2 levels of 13.1 and 34.8. The slope of the line looks lower for no2 levels of 25.8. There may be some interaction between pollution and age on their relationship with FEV.

b) Create Model 0: the regression of FEV on age and pollutant. Is age related to FEV, adjusting for pollutant? Provide this p-value, along with the best-fit equation for this model.

```
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
      (Intercept)  280.090    418.572   0.669    0.504
      age         181.933     40.451   4.498 1.01e-05 ***
        no2         -1.478      2.072  -0.713    0.476
```

$\hat{Y} = 280.1 + 182.0X_{age} + -1.5X_{no2}$

p = 1.01e-05

c) Create Model 1: the regression of FEV on age, pollutant, and the age X pollutant interaction. Provide the best-fit equation for this model.

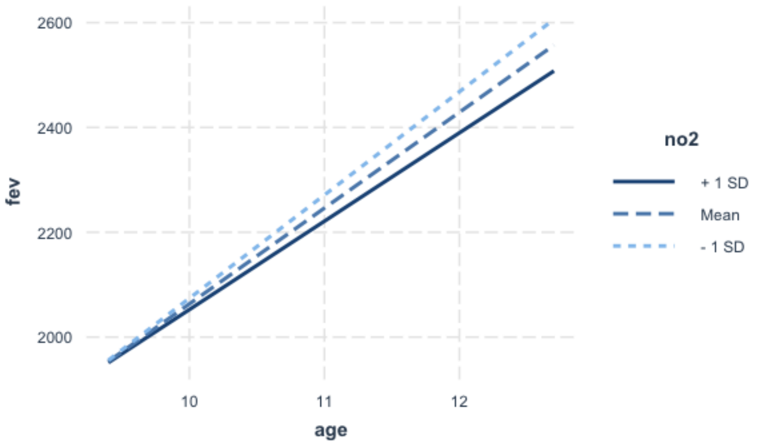
```
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept) -122.276   1281.421  -0.095   0.9240
age         221.477    125.714   1.762   0.0792 .
no2          14.983     49.583   0.302   0.7627
age:no2      -1.621      4.878  -0.332   0.7399
```

$\hat{Y} = -122.3 + 221.5X_{age} + 15.0X_{no2} + -1.62X_{age}X_{no2}$

d) Because this model only has one variable for the interaction term, it is not necessary to perform the Extra SS F-test. Instead, we can look directly at the p-value for the Wald test in the output. Is the overall interaction statistically significant? (Remember: the alpha for interaction terms can sometimes be relaxed as interactions are typically underpowered.)

The overall interaction is not statistically significant (p=0.74), even with the higher alpha level.

e) Use the interact_plot() function (pred = age, modx = [your pollutant]) and the sim_slopes() function to determine the effect of age on FEV at differing values of pollutant. Report the β_{AGE} terms for the mean, -1SD, and +1SD values of the pollutant along with p-values. Verify that these are the β_{AGE} terms you get from your equation in 7c.



Slope increases a bit at -1 SD below the mean no2 level, and decreases a bit at +1 SD above the mean no2 level.

```
SIMPLE SLOPES ANALYSIS

Slope of age when no2 = 15.05760 (- 1 SD):

   Est.    S.E.   t val.      p
-------- ------- -------- ------
  197.07   60.97     3.23   0.00

Slope of age when no2 = 23.84346 (Mean):

   Est.    S.E.   t val.      p
-------- ------- -------- ------
  182.83   40.61     4.50   0.00

Slope of age when no2 = 32.62932 (+ 1 SD):

   Est.    S.E.   t val.      p
-------- ------- -------- ------
        168.59   57.04     2.96   0.00
```

We will stop here. Note that the final model you arrived at in 4c should still be checked for the assumptions of linear regression, outliers, and influential values. Also note that your continuous-by-continuous interaction has two important assumptions:

1) For each value of $X_{POLLUTANT}$, the relationship between X_{age} and FEV is linear. (You can check this

assumption with the linearity.check = T option in interact_plot).

2) There is a linear change in the β_{AGE} coefficient as $X_{POLLUTANT}$ increases.