Assignment 4 - Identifying outliers and influential cases

Let us work with the same dataset HERS. We would like to check the influential cases. This time we will build a model to investigate the relationship between BMI, exercise at least 3 times per week, age, and race (raceth).

# Q1. First, we will build the model

Carry out a regression with BMI as a dependent variable and the other variables listed as predictor variables. Cut and paste your output in the box below. Provide the syntax. What is the value of the R2 for the model (report R2 in percentage with two digits decimal).

Hint: Make sure that you add “**i.**” in front of the ***categorical*** variables like race. For instance the code for race should be added as ***i.raceth***.

# Q2. interpret the relationship between the exercise, age, race and BMI at both the model and the predictor levels.

# Q3. Let us check the outliers of Y, first.

Remember, when we are looking at outliers on Y, we aren’t merely looking at whether a case has an unusually high or unusually low value (that’s actually expected). What we really want to know is when their observed outcome (yi) is very different than the outcome that the model would predict ().

Consider the following scenarios.

**Scenario 1:** We could have a participant whose BMI is three standard deviations above the mean (e.g., someone who is morbidly obese), but our model may predict their BMI very well.

**Scenario 2:** On the other hand, we could have someone whose had a slightly elevated or lowered BMI, but our model predicted something WAY different.

What we are worried about is when our model breaks down: so we are interested in scenario 2

Remember that we sill use the studentized residuals to identify cases (participants) for which our model does not predict the outcome well.

## Data Prep

Before we just jump into diagnosing our model, we need to generate case identifiers—i.e., numbers that uniquely identify each participant. Many data sets will already include case identifiers, but HERS does not. So, let’s create this:

gen case\_id = \_n

Next, let’s generate the studentized residual (the difference between yi and ).

(Hint: look back at the lecture to see how we do this)

## Diagnostics on Y

Now, you might have noticed that when you created the residual, Stata gave you the following message: “(5 missing values generated)”.

### Q3.1. Why might we have five missing values? (select One)

* This is an error in Stata, so we should go back and check our syntax
* Not all cases had complete data for all included variables. Cases without complete data have missing residuals.
* Some cases have erroneous data, and this caused the missing values.
* Gremlins. It’s always gremlins.

Now, we have a few ways to approach figuring out whether we have any problematic Y residuals.

First, let’s generate a histogram of the residuals to get a global sense of the distribution of the Y residuals. Remember, we start to get concerned if the residuals are ±2.5 standard deviations and really worried if they are greater than ±3 standard deviations. Here is my histogram. (BTW, if you wonder why my histogram may look different than yours, it’s because I’m using the “set scheme” command—Stata has a ton of different visual schemes. Check out <https://github.com/asjadnaqvi/Stata-schemes> and find the one that speaks to you. My style is “white tableau”. There is even a scheme in honor of Taylor Swift’s Red album [huh?].)

Chart, histogram

Description automatically generated

### Q3.2. So, do we have any outliers on Y that we are concerned about? (select one)

* No, everything looks fine.
* It’s difficult to tell without further investigation.
* Only in the lower tail.
* Yes, especially in the upper tail.

Let’s dig a little deeper.

We’ll use the hilo command to examine the values that concern us. In the lecture, we examined the highest and lowest 5 cases. Looking at the histogram, we may want to examine more than just five. So, examine the hilo output examining the highest 25 value.

### Q3.3. How many cases have residuals ≥3.0?

# Q4. Without knowing anything else about the performance of the model, what might we start to suspect about our model predicting BMI?

Select the best answer.

1. The model appears to be less accurate for those who fall in the center of the BMI distribution.
2. The model appears to accurately predict the BMI across its distribution.
3. The model cannot accurately predict the BMI in any part of the distribution.
4. The model performs particularly poorly in patients with very high BMI.
5. The model appears to be less accurate for those who have very low BMI.

# Q5. Now, lets look for outliers on the predictors.

Generate leverage values and then calculate a threshold value for leverage.

### Q5.1. What is the threshold value for leverage values? (Round to three digits)

## Now, that we have our leverage threshold, let us see how many of our cases have outliers above the threshold value.

### Q5.2. Paste the syntax you’d use to see how many cases have a leverage value above the threshold

Wow, that’s a lot of possible multivariate outliers on our predictors. How many precisely? Let’s count:

count if lev > 0.003

Ok, we have n=331 multivariate outliers on X!

### Q5.3. What should we be thinking at this point? Select one

1. There are so many! This clearly means that our model is going to be worthless.
2. We don’t really have to be worried about outliers until we’ve looked at influence.
3. Leverage estimation is wrong.

# Q6. identify any particularly problematic cases.

Create a plot of both Y and X outliers together in order to identify potentially problematic cases (be sure and label the cases by their ID).

### Q6.1. Paste the plot box below.

### Q6.2. Provide the syntax.

### Q6.3. We are looking for values in the upper right quadrant (or nearby). What is the case ID of the case the appears to be a likely outlier on both Y and X?

# Q7. The next step is to find out which cases are influential on *all* predicted Y values.

Which statistic should we choose?

1. DFOTS
2. Cooks distance
3. DFBETAS
4. F-test

# Q8. Create a list of all case ids that are above Cook’s value. How many are there?

### Q8.1 Paste the syntax you’d use to list the cases above the influence threshold and the syntax you use to count the number of cases above this value

### Q8.2 How many cases were above the threshold?

Hmm. What should we do? Well, we have more diagnostic tests to run, so we don’t have to do anything right now, but we log on that we have many influential cases.

# Q9. Now let’s turn our attention to influence on the coefficients of our predictor variables.

We are most interested in how the measures of age, exercise and age, (all together) are associated with BMI, so we really want to know which cases are having a strong influence on the coefficients of the predictive variables.

### Q9.1. In this case, which statistic is preferable?

1. DFOTS
2. Cooks distance
3. DFBETAS
4. F-test

Note: Let’s rerun our model and not treat raceth as a categorical variable (If you ran the regression model with race as an indicator variable (e.g., regress BMI exercise age i.raceth), then you will have four dfbetas (two for race—one for black and one for “other” race).

So, why would we do this? Think about it: for people of race “other” we would expect their race to have a high influence on their race category coefficient. So, that’s not terribly informative. So, let’s re-run the model with out race as an indicator variable: regress BMI exercise age raceth

Then we’ll compute our influence measure and its threshold.

### Q9.2. Compute a threshold value for the statistic that tells us about the influence on our predictor coefficients. What is the threshold value?

## Now let us draw a scatterplot that combines the DFBETAS scores of all three variables. Influential points can be in either direction.

Now, we could simply list off all the influential cases (using the list command) as we did above, but it may be easier to get a sense of the cases with high influence on the different predictor variables if we create a composite scatter plot.

When you do this, you want to add Y reference lines at the positive and negative threshold values (this will help you see the influential cases better—especially important in large data sets). Label the influential cases by case-id and ensure that the Y axis scales in a way that you can view the cases separately.

### Q9.3. Provide the syntax for the combined scatterplot

### Q9.4. Paste the plot below.

# Q10. Wow! So many dots! What does all of this mean?

There is something strange going on here. While we would normally expect to see is a handful of influential cases (both on predicted values of the outcomes and on the model coefficients). Having so many influential points is a concern. Let us check the collinearity now since we know that high collinearity can cause problems, so let’s check to see whether collinearity is creating this mess!

**Note:** The threshold values that we are using in this course **are not absolute**. There are other ways (in my opinion, more conservative) for calculating the threshold values. (e.g., see <https://online.stat.psu.edu/stat462/node/173/>). In real life, for large datasets (where the influence of any one point is likely to be washed out by all the other data points), I’d probably use more conservative methods. In smaller datasets (where the influence of any one case is going to be proportionally larger), I’d use the more liberal methods we have used here.

### Q10.1. What is the highest variance inflation factor value for the predictors in this model?

### Q10.2. Does the variance inflation factor value indicate any multicollinearity problem in our model?

### Q10.3. Do you find any variables on the right-hand side of the model (predictors) to be conceptually colinear which each other?

The next step is to set up the stage for other diagnostic models. We will do this over the next module.