**Workshop on Assumptions & Outliers and influential values**

So far we have built simple linear regression and multiple regression models. This is not the actual first step of linear regression. To conduct linear regression you first check assumptions to ensure that your data are appropriate for linear regression modeling. The full regression modeling process has 4 steps:

1. Check the assumptions for your intended model(s)

* If assumptions fail, try:
* transforming or recoding variable(s) and check assumptions again
* using a different type of statistical model
* reporting descriptive and visual statistics instead

(2) If assumptions pass, estimate the model(s)

(3) Conduct review of outliers and influential values

* If you find data entry errors or other obvious mistakes/problems in the data, correct or remove the problem observations and re-estimate the model without the problematic observations

(4) Interpret and report the results

* Value, direction, significance, confidence interval of slope (b, t, p, 95% CI)
* Significance of the model (F, p)
* Model fit (adjusted R2)
* Results of assumption checking and diagnostics

**Before you begin: DETERMINE THE MODEL YOU INTEND TO BUILD**

Use the larger model estimated during Exercise 5:

*distance to syringe program = opioid prescriptions + metro + uninsured*

1. Step 1: CHECK ASSUMPTIONS

Assumptions of multiple regression:

* Observations are independent
* The outcome is continuous
* Continuous outcome and predictor variables are normally distributed\*
* The relationship between continuous predictors and the outcome is linear (linearity)
* The variance is constant with the points distributed equally around the line (homoscedasticity)
* The residuals are independent
* The residuals are normally distributed
* No perfect multicollinearity

*Observations are independent*

There is nothing to test statistically, this assumption is met (or failed) during data collection. Meeting this assumption requires the observations in your data to be *unrelated* to each other. Examples of data that would fail this assumption are data where siblings or spouses or neighbors are surveyed, data where people may be included more than once.

Think about the syringe data, do you believe this assumption is (circle one):

MET or NOT MET

*The outcome is continuous*

Think about the syringe data and the model above, do you believe this assumption is (circle one):

MET or NOT MET

*Continuous outcome and predictor variables are normally distributed*

This “assumption”\* of linear regression is not often included in lists of official assumptions, however, the calculation of the slopes relies on means and means are not a good representation of a variable when it is skewed. Although it doesn’t have to be checked as a formal part of linear regression, checking normality of the outcome and any continuous predictors can help you identify potential problems before you check the more complicated assumptions (e.g., linearity and homoscedasticity). This can be checked by creating histograms of the continuous outcome and any continuous predictors. Open your week 5 exercise document and start a new R chunk at the bottom by putting the cursor at the end of the document, click on “Insert” at the top of the code window, and choose “R.”

In this new R chunk, write and run the code to create histograms for the 3 variables in the model that are continuous. Loosely sketch the histograms below. Circle whether the assumption is MET or NOT MET based on the graphs.

*The relationship between continuous predictors and the outcome is linear (linearity)*

This assumption is the same as the linearity assumption checked for the correlation coefficient in week 3. The assumption must be met for the relationship between the outcome and **each of the continuous predictors (but not the categorical predictors).** So, you will check the outcome of **distance** to see if it is linearly related to **percent uninsured** and **opioid prescriptions.** Check your notes for week 3 about testing this assumption for correlation coefficients. It will be the same for linear regression. Copy/paste and edit the code from the week 3 workshop to make a scatterplot with a Loess curve. Update the code to include the correct data set and variable names. Run the code and loosely sketch the results here. Indicate whether the assumption appears MET or NOT MET.

*The variance is constant with the points distributed equally around the line (homoscedasticity)*

This assumption is the same as for the correlation coefficient, however, it is difficult to check visually when there are so many predictors. Instead, the homoscedasticity assumption is checked with the **Breusch-Pagan** test from the *lmtest* package in R. Install the *lmtest* package, open it with the *library* function, and use the *bptest* function like this:

bptest(formula = name.regression.model)

Where you put the name of your regression model in place of the highlighted text after the equal sign.

The Breusch-Pagan, or BP, test tests the null hypothesis that the variance is constant. If the null hypothesis is rejected due to a small p-value, the assumption is not met. If the null hypothesis is retained, the assumption is met.

Determine whether the assumption is MET or NOT MET based on the BP test.

*The residuals are independent*

The **Durbin-Watson** test can be used to determine whether the model violates the assumption of independent residuals. When residuals are independent that means that there is no pattern among the residuals where it looks like they may be related, for example, if the residuals all get bigger at at larger values of one predictor, this pattern might indicate something strange is going on with the measurement or the model. The null hypothesis for the Durbin-Watson test is that the residuals are independent. The alternative hypothesis is that the residuals are not independent. The R code to use this test is also available in the *lmtest* package:

dwtest(formula = name.regression.model)

Replace the highlighted text with the name of your regression model. Examine the output to find the p-value and determine whether the assumption is MET or NOT MET.

*The residuals are normally distributed*

Normally distributed residuals indicate that the regression line is far above a few points, far below a few others, and relatively near most of the points. If the residuals are skewed, that would mean that the regression line is pretty far away from some of the higher or lower values of the outcome. The residuals are saved with your regression model and can be accessed by using name.regression.model$residuals. Plot a simple histogram using the base R histogram code, replacing the highlighted text with your regression model name:

hist(x = name.regression.model$residuals)

Loosely sketch the histogram and indicate whether the assumption is MET or NOT MET.

*No perfect multicollinearity*

Multicollinearity occurs when two predictors are too highly correlated with each other and so they are redundant. To find multicollinearity, either use a correlation coefficient (anything above r = .7 is too high) or use a VIF, which determines how much variance in each predictor is explained by the other predictors. Values of VIF that are above 5 are problematic for multicollinearity.

Use cor() to get correlations or use the vif() function in the car package to get VIFs:

library(package = “car”)

vif(name.regression.model)

*Which assumptions did the model meet? Which did the model fail?*

**What happens when you meet or fail assumptions?**

* If the model fails any\* of the assumptions (\*except the optional normality one), it would be considered *biased* and not reported (or reported with the caveat that it is biased and should not be generalized outside the sample). If this is the case, there are several things to try:
  + Transform non-normal continuous variables to be more normally distributed and try the assumption checking again
  + Recode non-normal continuous predictor variables into categorical variables and try the assumption checking again
  + Recode the non-normal outcome variable into a categorical variable and develop a different type of model
  + Rely on visuals and descriptive statistics in reporting
* If the model meets all of the assumptions it would be considered *unbiased* and you would go ahead to Step 2, estimating the model. Then, once the model was estimated, you would check to make sure it doesn’t have suspicious outliers or influential values to worry about (Step 3):

Step 2: ESTIMATE THE MODEL (you already did this in your exercise 5 work)

Step 3: FINDING OUTLIERS AND INFLUENTIAL VALUES

Even if the assumptions are all met, there may still be problematic observations in the data that are making the regression results unreliable. There are three types of problematic observations to review before reporting the model:

* An *outlier* is an observation with unusual values.
* A *regression outlier* has an unusual value of the outcome given its value(s) of predictor(s).
* An *influential observation* changes the slope of the regression line.

There are many ways to check for outlying and influential values. Here are three useful measures:

1. **Standardized residuals:** The difference between the observed and predicted values, standardized into z-scores so that observations with standardized residuals > 1.96 are not being predicted accurately by the model and should be examined.
2. **Leverage values:** Leverage is the amount the predicted value of the outcome would change if the observed value of the outcome was changed by one unit. Leverage values range between 0 and 1. Leverage values above > 2k/n, where k is the number of parameters and n is the sample size, are considered problematic and should be reviewed.
3. **Cook’s distance:** Each observation is removed and the model is re-estimated. The sum of squared differences between predicted values of the outcome with the particular observation and without that observation is Cook’s D. The cutoff for a high Cook’s D is 4/n, where n is the sample size. Observations with high Cook’s D should be reviewed.

Keep working in your Exercise 5 file and use the code below to identify outliers and influential values. Wherever you see name.regression.model replace it with the name of your regression model. After you identify the problematic observations, answer some multiple choice questions about them.

**Standardized residuals**

Code to add a variable containing standardized residuals to the data. The new variable is called “standardres” (you can name it whatever you want by changing this name in the code) and naming the data frame that includes the standardized residuals something new, in this case it is named “dist.ssp.diag” (you can name it whatever you want):

dist.ssp.diag <- dist.ssp %>%

mutate(standardres = rstandard(model = name.regression.model))

Code to check for standardized residuals bigger than 1.96:

dist.ssp.diag %>%

filter(abs(x = standardres) > 1.96)

How many counties had big standardized residuals?

**Leverage**

Sample code to add standardized residuals and leverage values to the new data that has the standardized residuals in it too. Leverage values are often called “hat values”:

dist.ssp.diag <- dist.ssp %>%

mutate(standardres = rstandard(model = name.regression.model)) %>%

mutate(lever = hatvalues(model = name.regression.model))

Code to check for hat values bigger than 2\*(p+1)/n where p is the number of predictors in the model:

dist.ssp.diag %>%

filter(abs(x = lever) > 2\*(3+1)/n())

How many counties had big leverage values?

**Cook’s D**

Sample code to add all three diagnostic measures, standardized residuals and leverage values and Cook’s D, to the new dist.ssp.diag data:

dist.ssp.diag <- dist.ssp %>%

mutate(standardres = rstandard(model = name.regression.model)) %>%

mutate(lever = hatvalues(model = name.regression.model)) %>%

mutate(cooks.dist = cooks.distance(model = name.regression.model))

Code to check for Cook’s D values bigger than 4/n:

dist.ssp.diag %>%

filter(cooks.dist > 4/n())

How many counties had big Cook’s D values?

**Identifying the truly problematic**

First add the predicted values to the data so that you can examine how well the model did predicting the outcome for each county. Add it as a line in the code you’ve been working on:

dist.ssp.diag <- dist.ssp %>%

mutate(standardres = rstandard(model = name.regression.model)) %>%

mutate(lever = hatvalues(model = name.regression.model)) %>%

mutate(cooks.dist = cooks.distance(model = name.regression.model)) %>%

mutate(predict.distance = predict(object = name.regression.model))

Then, counties that are trouble by more than one measure are considered a potential problem and should be examined. There are many ways to find these counties. Using filter with the “and” symbol **&** and the “or” symbol **|** is one way. In this code, the filter finds counties that are above the cutoff for at least 2 of the 3 measures (standardres, lever, cooks.dist):

dist.ssp.diag %>%

filter(cooks.dist > 4/n() & abs(x = lever) > 2\*(3+1)/n() |

cooks.dist > 4/n() & abs(x = standardres) > 1.96 |

abs(x = standardres) > 1.96 & abs(x = lever) > 2\*(3+1)/n() )

How many counties were truly problematic?

*Copy and run the code from the electronic version of this document in GitHub to examine visually which of the observations are truly problematic.*

dist.ssp.diag %>%

drop\_na() %>%

mutate(problem.obs = if\_else(condition = (cooks.dist > 4/n() & abs(x = lever) > 2\*(3+1)/n() |

cooks.dist > 4/n() & abs(x = standardres) > 1.96 |

abs(x = standardres) > 1.96 & abs(x = lever) > 2\*(3+1)/n()),

# if this is TRUE, then it is an outlier, otherwise it's not

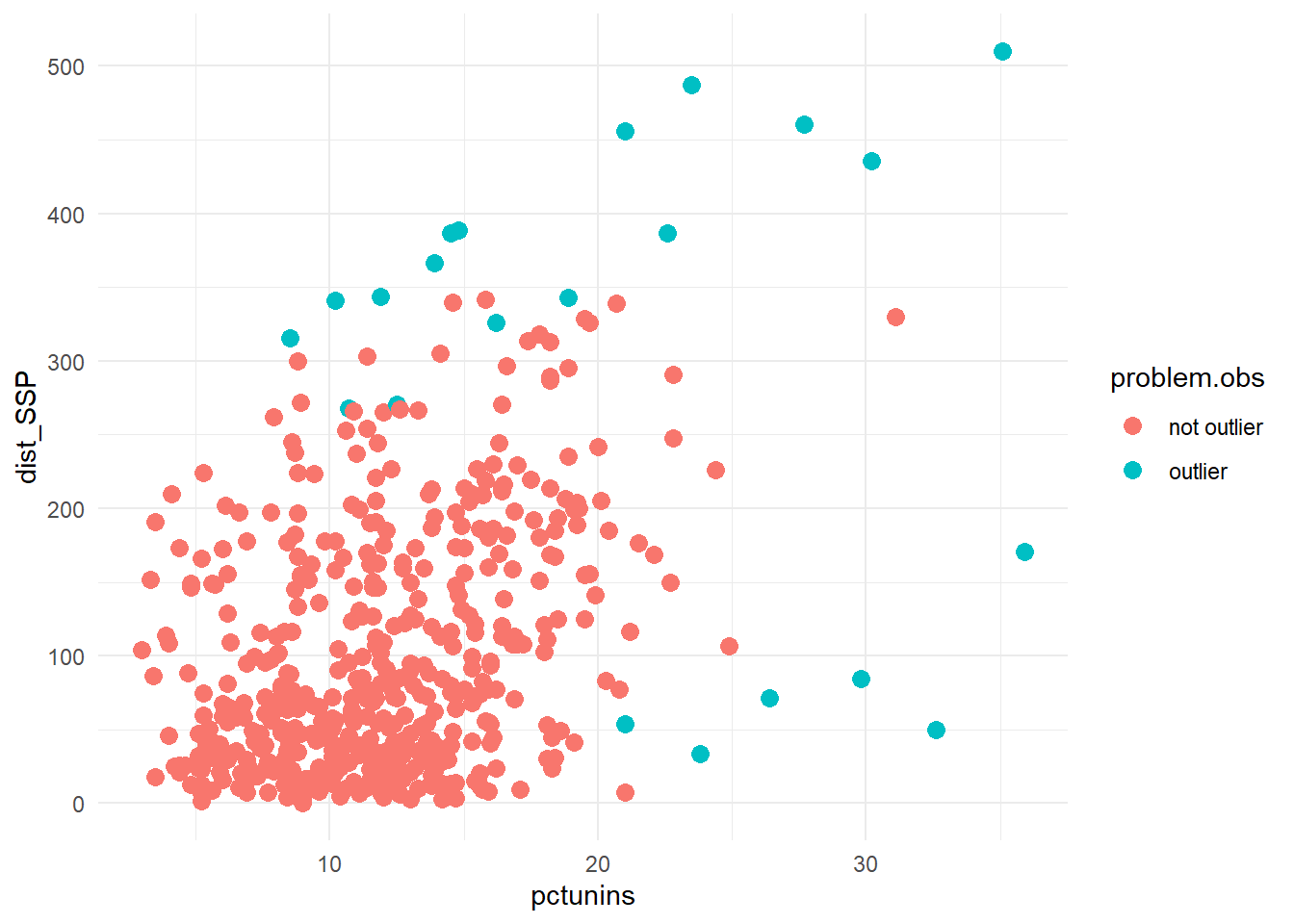
true = "outlier", false = "not outlier")) %>%

# pipe the new variable into a graph

ggplot(aes(x = pctunins, y = dist\_SSP, color = problem.obs)) +

geom\_point(size = 3) + theme\_minimal()

*Your graph should look similar to this:*



The graph shows that the outliers had very high and very low values of distance and many were at higher ends of of percent uninsured. Review the values of the outcome, predictors, and predicted values in the output you got after you filtered for the **truly problematic.** See if you can figure out why these counties were identified as outliers.

Remember:

* Mean distance to syringe program: 107.74
* Mean opioid prescription rate: 68.33
* Mean percent uninsured: 12.18

**Exploring the outliers and influential values**

**Work with a partner to figure out why some of the counties may have been outliers or influential values. Answer the 10 multiple choice questions on one of the scratch off sheets. Put both your names on the scratch off and submit to the done jar for your participation today! If you don’t complete the scratch off, put your names on whatever you’ve done and submit before you leave today.**

1. Hendry County, Florida was one of the problem counties. Which of the following is true for Hendry County?
2. The predicted distance was a lot further than the actual distance to a syringe program of 84.33 miles.
3. It has a low HIV prevalence rate compared to the mean HIV prevalence in the sample.
4. The predicted distance of 84.33 miles was a lot closer than the actual distance to a syringe program.
5. It has a low opioid prescription rate compared to the mean for the sample.
6. El Paso County, TX is one of the problem counties. Which of the following is NOT TRUE according to your results?
7. It has a very high HIV prevalence rate compared to other counties in the sample.
8. The predicted distance to a syringe program is 33.47 miles
9. El Paso is closer to a syringe program compared to many of the other counties in the sample.
10. The predicted distance to a syringe program from the county was 187.45 miles.
11. The county with the fewest opioid prescriptions per 100 people is how far from a syringe program?
12. 315.87 miles
13. 267.67 miles
14. 54.00 miles
15. 79.56 miles
16. The county closest to a syringe program is in which state?
17. FL
18. SD
19. ME
20. TX
21. How many of the 22 problem counties are metro?
22. 22
23. 9
24. 11
25. 13
26. The county predicted to be the closest to a syringe program was...
27. Coryell County, TX
28. El Paso County, TX
29. Concordia Parish, LA
30. Baker County, FL
31. The lowest uninsurance percent is what percent and what county?
32. 71.33% in Concordia Parish, LA
33. 3.3% in Presidio County, TX
34. 10.2% in Coryell County, TX
35. 8.5% in Kennebec County, ME
36. The farthest predicted distance from a syringe program was...
37. 510.00 miles
38. 300.40 miles
39. 315.87 miles
40. 456.00 miles
41. How many different states had at least one county on the list?
42. 6
43. 5
44. 3
45. 4
46. Which of the following is true about Lamb County, TX and is much different from the mean value for the counties in the sample?
47. It has a very low distance to syringe program
48. It has a very low HIV prevalence rate
49. It has a very high opioid prescription rate
50. It has a very low uninsurance percent

STEP 4: Interpret and report the results

Now that you have checked assumptions, estimated the model, and identified outliers and influential values, it is finally time to report the model. This should include the following:

* Value, direction, significance, confidence interval of slope (b, t, p, 95% CI)
* Significance of the model (F, p)
* Model fit (adjusted R2)
* Results of assumption checking and diagnostics

*An example of a full linear interpretation for this model (given failed assumptions, we would do something different, this is just to guide you in how to construct and word a full interpretation):*

A linear regression model predicting distance to the nearest syringe program was statistically significantly better than the mean of the outcome at explaining the outcome [F(3, 496) = 43.07; p < .05]. The model explained 20.19% of the variation in distance (adjusted R2 = .2019). The three predictors were all statistically significantly different from 0 (p < .05). For every additional opioid prescription per 100 people, the distance to a syringe program decreased by .32 miles (b = -.32; t = -3.07; p = .002). The confidence interval suggested that, in the population, for each additional opioid prescription per 100 people, the distance from a county decreased by .11 to .52 miles (95% CI: -.11 to -.52).

Non-metro counties were 29.63 miles further from syringe programs compared to metro counties (b = 29.63; t = 3.84; p = .0001). The confidence interval suggested that, in the population, non-metro counties were between 14.48 and 44.79 miles further from a syringe program (95% CI: 14.48-44.79). For every additional percent of county residents uninsured, the nearest syringe program was 7.35 miles further away (b = 7.35; t = 9.53; p < .05). The confidence interval suggested that, in the population, for every additional percent uninsured in a county, the nearest syringe program was 5.83 to 8.86 miles further away (95% CI: 5.83-8.86).

The model met the independence of observations, continuous outcome, and independent residuals assumption but failed the other assumptions (normally distributed variables, linearity, homoscedasticity, constant variance, normally distributed residuals). There were 22 values that were outliers or influential, most of which had high or low values for the outcome or one or more predictor. There were no obvious data entry or other errors, so all the observations were kept in the model.

That’s all the new stuff for today!

Turn in your multiple choice sheet and pick up a mock midterm before you leave.