Session 5

Devesh Tiwari

June 6, 2016

Introduction

Today we will spend some time on Steps 3 and 4 of the model building process: Model Execution and Evaluation, and Model Diagnostics.

After formulating a question for which you want to build a model, and conducting an EDA to get a sense of what you will include in your model or models, you need to actually execute the model in r or your software package of choice. At this point, it is important to think about the following questions:

(1) With these results in hand, do I find statistical evidence in favor of my initial hypothesis?

(2) What questions can I answer and which questions are not answerable?

(3) What additional data or analysis do I need to answer questions fully?

(4) If I have more than one model, which model do I like the best and why?

We talked about model diagnostics last time, but Steps 3 and 4 are directly related to each other and are pretty iterative. The results of your residuals analysis directly impact whether or not you can interpret the results of your statistical tests (or if you can even run thos tests).

You may also find yourself cycling back and forth between running a model, examining its residuals, and then re-running different models. Note that as you run more and more models, you should be come more and more skeptical of its results (multiple comparisons problem). This does not mean that you throw away your answers or insight, but that you should note whatever skepticism you have in your final report.

Breakout sessions 1 and 2

In this exercise, you will be examining some output of someone else's data analysis output. This particular analysis analyzes US county level data with information about counties' average life expectancy, obesity rate, and the proportion of the population who are physically active. These data are found at the Institute for Health Metrics and Evaluation at the University of Washington. These data have been slightly modified for this exercise.

This analysis examines the question, "Do people who are classified as being obese have shorter life spans than people who are not obese?"

Examine the following R-output closely, as it contains the only information you will have to answer these questions. If you feel that you do not have enough information to answer a given question, then make whatever assumptions you need to make in order to complete this exercise. If you feel that you do not have enough information to answer any of these questions, then please feel free to state that in your answers.

Please break up into groups in order to answer these questions. All groups are required to answer number 1, after that choose one of the remaining questions to answer.

Questions

1. In Model 2, formally test the hypothesis that counties that have a higher proportion of obese people have lower life expectancies.

2. Note that Model 2 has a higher R^2 than Model 1. What does this mean? Does this mean that Model 2 is a better model than Model 1? In general, do you think models with a higher R^2 are better than those with a loewr R^2?

3. Imagine you had added a third regressor to model 2. How would determine whether the inclusion of this variable improves the explanatory power of your model? What is the name of the test and what information would you need to conduct such a test? Describe what this test does in words, if you can.

4. These data also contain county level data separated by gender. In other words, we can examine the relationship between life expectancy, obesity, and physical activity for men and women separately. With the information presented here, how would you examine whether the relationship between physical activity and life expectancy is lower for women than it is for men? If you had access to the underlying data and R, what equation would you estimate in order to determine whether the relationship between physical activity and life expectancy?

5. Please interpret the scatter-plot that was produced from the output in Model 2. What information does this scatter-plot display? What does this scatterplot tell you about Model 2? Which of the underlying assumptions of classical linear models can you test by examining this plot? Which of the assumptions can't you test?

6. Suppose you were worried about heterogeneity.How would the presence of heterogeneity change your interpretation of these results? For each statistical test you conducted thus far, what would you do to account for the presence of heterogeneity?

Output

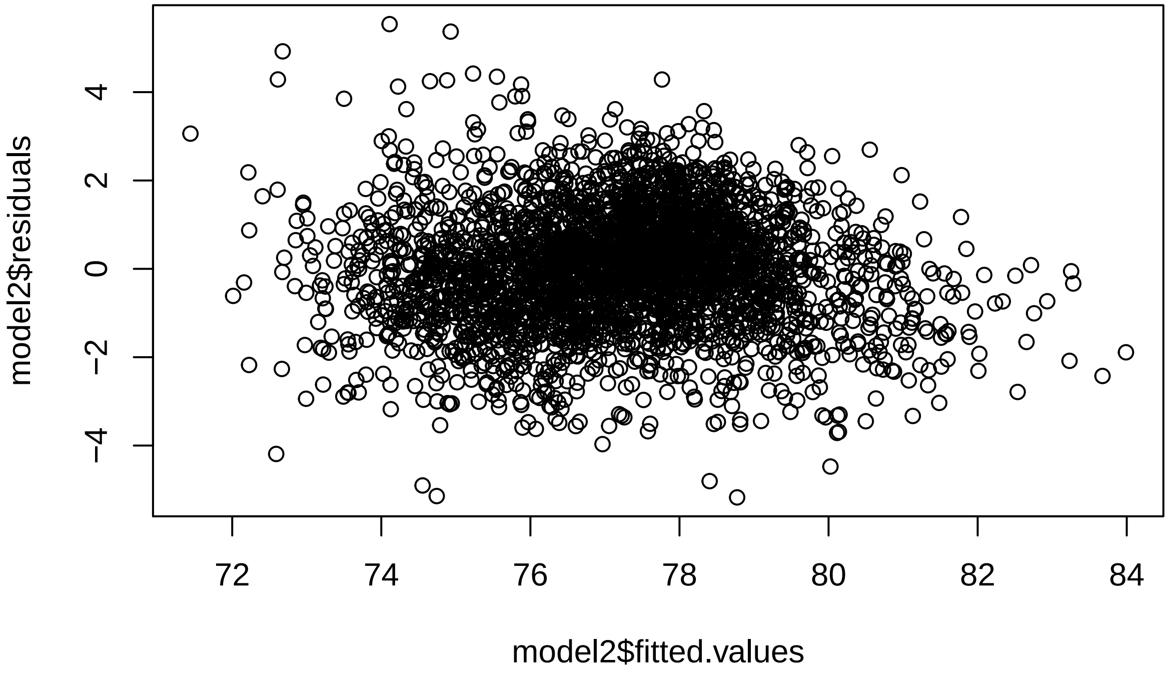
```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
##
## =======================================================
## Statistic             N    Mean  St. Dev.  Min    Max
## -------------------------------------------------------
## LifeExpetAll        3,142 77.203  2.161   68.400 83.250
## ObesityAll          3,142 38.037  4.331   18.550 53.000
## PhysicalActivityAll 3,142 51.456  6.455   31.500 75.100
## -------------------------------------------------------
```

```
##
## ===========================================================================
##                                       Dependent variable:
##                     -------------------------------------------------------
##                                          LifeExpetAll
##                               model1                      model2
## ---------------------------------------------------------------------------
## ObesityAll                   -0.367***                    -0.143***
##                               (0.006)                      (0.010)
##
## PhysicalActivityAll                                         0.181***
##                                                             (0.006)
##
## Constant                     91.179***                    73.318***
##                               (0.231)                      (0.670)
##
## ---------------------------------------------------------------------------
## Observations                   3,142                        3,142
## R2                             0.542                        0.634
## Adjusted R2                    0.542                        0.633
## Residual Std. Error      1.463 (df = 3140)           1.309 (df = 3139)
## F Statistic         3,717.289*** (df = 1; 3140) 2,715.501*** (df = 2; 3139)
## ===========================================================================
## Note:                                           *p<0.1; **p<0.05; ***p<0.01
```



```
##
## ===============================================================
##                                       Dependent variable:
##                                  -------------------------------
##                                  MaleLifeExpect FemaleLifeExpect
##                                       male           female
## ---------------------------------------------------------------
## MaleObesity                        -0.140***
##                                     (0.013)
##
## MalePhysical                        0.231***
##                                     (0.007)
##
## FemaleObesity                                     -0.103***
##                                                    (0.007)
##
## FemalePhysical                                     0.136***
##                                                    (0.005)
##
## Constant                          67.621***       76.937***
##                                    (0.798)         (0.490)
##
## ---------------------------------------------------------------
## Observations                        3,142           3,142
## R2                                  0.530           0.596
## Adjusted R2                         0.529           0.596
## Residual Std. Error (df = 3139)     1.710           1.221
## F Statistic (df = 2; 3139)       1,767.205***    2,319.528***
## ===============================================================
## Note:                               *p<0.1; **p<0.05; ***p<0.01
```

Data Activity

Load the dataset called *twoyear.Rdata*. This dataset has individual level information about respondents wages, education, job experience, and other demographic data. We are interested in answering the questions: Do men and women have the same wage? Do they realize the same rate of return on their prior level of experiences? Is there a non-linear relationship between experience and wage?

(1) Based on your EDA and the question(s) at hand, outline more than one modeling strategy and implement it. Be sure that you can justify why you chose the models you chose.

(2) Examine the residuals and check to see if you should be worried about any of the CLM assumptions being violated. Be sure to conduct any statistical tests that are needed.

(3) Based on your post-regression diagnostics, go ahead and make some adjustments to your models if needed.

(4) Answer the questions that were posed at the begining of this exericse. Give substantive interpretations as well.

(5) Which model do you like best? Why? What criteria are you using when evaluating model quality?