Data Task: Stanford University Pre-Doctoral Application

Marcelo Ortiz Villavicencio

November 15, 2020

0. Summarize the data. Do some observations appear to be data entry errors (accounting for fact that phenomena in #1 are legitimate)?

Descriptive Statistics						
	Min	**Max**	**Median**			
Duration Shifts	2 hours	10 hours	9 hours			
	Unique					
Shift Id	398					
Physicians	43					
	mean	**sd**	**p0**	**p25**	**p75**	**p100**
Log Length of Stay	1.12	0.382	-0.276	1.01	1.38	2.12

1. Some patients may arrive before their physician's shift starts and therefore would have to wait. Other patients may be discharged after their physician's shift ends (and the physician would have to stay past the end of shift). What percentages of visits fall in these categories?

Before I could answer the question, I had to work on the date and time format in order to make my comparisons. So, I changed the format of the variables **ed_tc** and **dcord_tc**. Then with the variable **shiftid**, I create two new variables named **shift_start** and **shift_end**. With these 4 variables, and with the same format, I can make the following comparisons:

Patient arriving before shift starts = **ed_tc** < **shift_start**

Patient discharged after shift ends = **dcord_tc** > **shift_end**

The number of patients who meet the first conditions is 607 (6.87% of total). The number of patients who meet the second condition is 2448 (27.72% of total).

2. Describe hourly patterns of patient arrivals and the average severity of these patients. Extra credit: How might one formally test whether patient severity is or is not predicted by hour of the day?

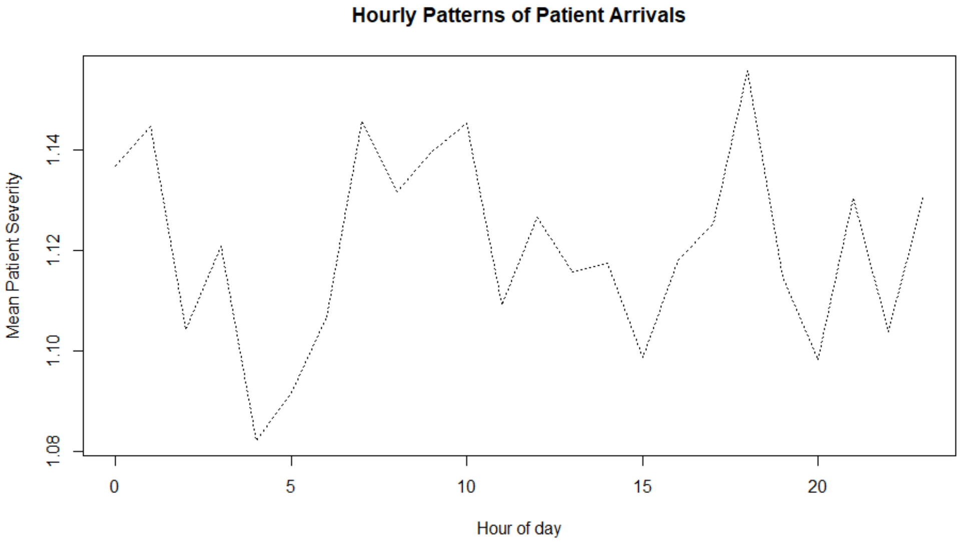
In Figure 1, You can observe the patterns of severity of patients according to their time of arrival. The highest peak of severity occurs at 6 p.m., although there are also important peaks at 1 a.m. and between 8 and 10 a.m. Regarding extra credit question, I would try a regression to prove that. For instance, I aware that 6 p.m. is problematic, so I would create a dummy variable that takes value 1 if the patient arrivals at 6 p.m. and zero otherwise. Hence, I would run a regression of patient severity on the dummy variable like this:

$Severity_i = \beta_0 + \beta_1 ProblematicTime_i + \varepsilon_i$

So, β_1 captures the average difference of patients severity between patients that arrive in a problematic time and patients that arrive in non-problematic times. That is,

$\beta_1 = E[Severity_i|ProblematicTime = 1] - E[Severity_i|ProblematicTime = 0]$

Figure 1: Hourly Patterns of Patient Arrivals



3. Create and include with your solutions a dataset recording the "census," or number of patients under a physician's care (patients who have arrived and

have not yet been discharged), during each hour of a physician's shift from beginning to 4 hours past the end of shift. The observations in this dataset should correspond to the shift (shiftid), physician (phys_name), and the hour of shift (index). index should be defined as follows, so that it should mostly negative values and have a maximum of 3 in the dataset: The hour ending at the same time as shift end is indexed -1, the hour beginning at shift end is indexed 0, and the hour beginning one hour after shift end is indexed 1, etc. Hint: You will need to transform the text in shiftid into numerical shift beginning and end times capturing both date and hour; you should ignore patient hour observations falling outside of the shift times of interest. How does the census vary with time relative to end of shift? Discuss conceptually how you construct censuses, and note (and for extra credit, address) issues with discrete time. For extra credit, you may want to produce a "lower bound" census, "upper bound" census, and "exact" census for each observation.

First, I obtained a list of unique shifts for each of the physicians. Then I did an iteration with the elements of this unique shift list. In each iteration, I saved the name of the physician and the **shiftid**. I also calculated the duration of their shift, with the variables that I had already built in question 1. Then, inside the loop, I built two temporary dataframes that were updated in each iteration. The first dataframe named **filter_df** captures all patients who meet an appointment with a specific date and doctor. The second dataframe named **temp_df** builds a small version of the large census. I think so because it was easier for me to debug.

In this second dataframe, I create hour, index and census variables. In hour I save the current time as the physician's shift passes by until the end (considering the 4 additional hours). Then I create the index variable according to the statement of the question. To my knowledge, most doctors work 9-10 hour shifts. Therefore, this index variable ranges from -9 or -10 to 3, for each doctor's shift. Then, through a filter I use the information of **temp_df** with **filter_df**. I count for each time stamp, how many patients the doctor has on the current time and I update my table **temp_df**. Then, I save the table **temp_df** of each iteration to finally obtain the complete census data with 6929 observations. The census vary negatively with time relative to end of shift. That is, less patient is left as the end of the shift approaches.

4. Which physician appears to be the fastest at discharging patients? You should answer this with a regression of log length of stay. You may also show results graphically. Discuss how you thought about which variables you control for. What are potential threats (and any evidence of them) to your assessment (you do not need to actually address these threats)? How robust are your estimates of physician effects to various specifications? Are there any concerns you might have with a finite number of patients per physician?

According to figure 2, the physician who seems to be the fastest in discharging patients is Benjamin. I would also propose the following regression model:

$$LogLengthStay_{ijt} = \alpha Census_{jt} + \beta Time2EndShift_{jt} + \varepsilon_{ijt}$$

where $LogLengthStay_{ijt}$ is the log length to stay of patient i with physician j and t is time of patient arrival to ED, $Census_{jt}$ denotes the number of patients under the care of physician j at time t; and $Time2EndShift_{jt}$ indicates how many hours are left in the physician's shift j relative to time t. The results of the model are presented in table 1.

Regression analysis indicates a negative relationship between Log Length of Stay and Census variables, although the effect is significant at 10%. In addition, there is a negative relationship between Log Length of Stay and Time To End Shift variables. This would indicate that an increase in the number of clients in a physician's care is related to a decrease in the severity of the patient. Probably the most burdened physicians are looking for faster discharge.

I would like to try other patient level controls and physician related features. Regarding patient's characteristics, it would be interesting to add patient clinical information and demographics. With respect to the physician's characteristics, it would be interesting to evaluate the workload, number of available beds, number of nurses assisting him, etc. It is likely that with this information you can get an idea if indeed the specification of this straightforward model is robust.

Figure 2: Which physician appears to be the fastest at discharging patients?

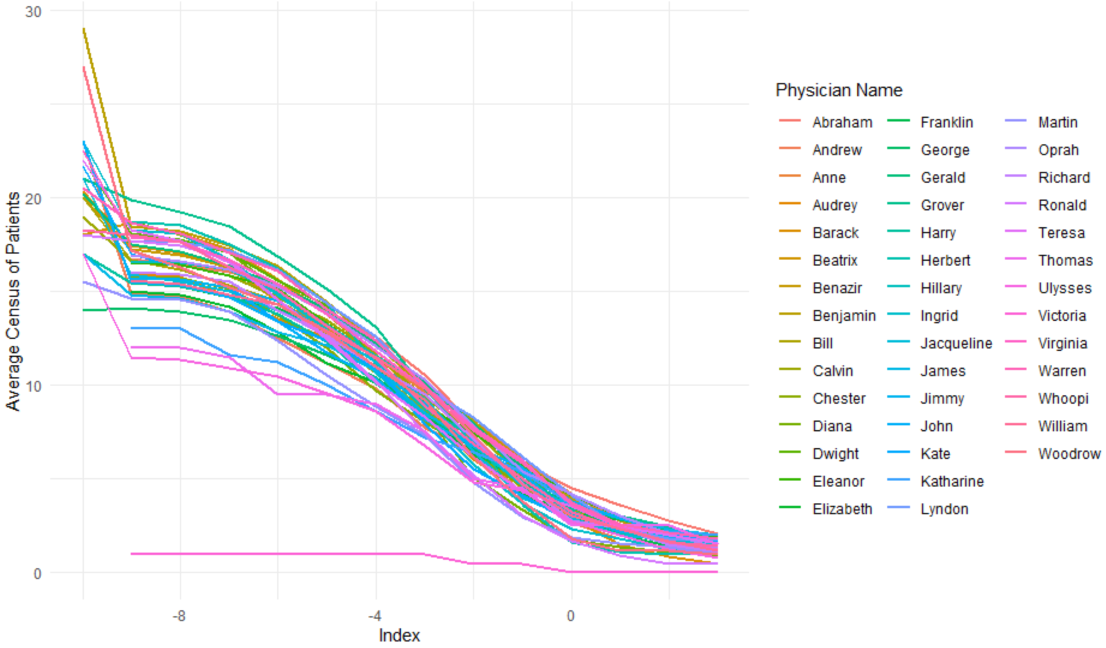


Table 1: Which physician appears to be the fastest at discharging patients?

	Dependent variable:	
	Log Length of Stay	
	OLS	OLS
	(1)	(2)
Census	0.0013	-0.0025*
	(0.0009)	(0.0014)
Time 2 End Shift		-0.0120***
		(0.0031)
Constant	1.1029***	1.0892***
	(0.0140)	(0.0147)
Note: Standard errors clustered by physician	*p<0.1; **p<0.05; ***p<0.01	