Capstone 01: the variability of the medical practice (updated on 12/19/2019)

1-) Introduction to the problem.

In the health economics literature, moral hazard and supplier-induced demand are well-known issues that threaten the stability of health insurance schemes. According to moral hazard, insured consumers tend to be less cost-conscious and, as a result, overspend on health care. According to supplier-induced demand, health care providers can stimulate the moral hazard behavior of their patients in order to increase their income or protecting themselves against medical malpractice lawsuits. Because of these two issues, health insurance companies need to monitor and evaluate the medical practice of doctors serving their clients. However, other factors such as specialty, locality and time may also affect the medical practice. The present study analyzes a sample of the database of a Colombian health insurance company. The sample contains the number of patients, visits and ambulatory surgeries performed for each doctor working with the company during the years between 2008 and 2012. The sample also informs the doctor's specialty and locality (more specifically, the Colombian department). There is also information regarding the cost of a visit and a surgery, but unfortunately that information is not disaggregated: the sample contains the mean cost of a visit and a surgery per specialty in each year.

The main purpose of the present study is to understand the role of specialty, locality and time in predicting and explaining the variability of the medical practice in terms of number of patients, visits and ambulatory surgeries. If the role of specialty, locality and time is weak or unusual (unexpected), then other factors such as moral hazard and supplier-induced demand may need to be investigated.

2-) Data.

As explained, the data comes from a sample of the database of a Colombian health insurance company. Please, visit https://github.com/tmivanus/Springboard/tree/master/Capstone_01 to see all data and codes. The data was originally divided into seven csv files:

- *datos_ambulatorios_2008*: doctor, specialty, patients, visits and ambulatory surgeries in 2008.
- *datos_ambulatorios_2009*: doctor, specialty, patients, visits and ambulatory surgeries in 2009.
- *datos_ambulatorios_2010*: doctor, specialty, patients, visits and ambulatory surgeries in 2010.
- *datos_ambulatorios_2011*: doctor, specialty, patients, visits and ambulatory surgeries in 2011.
- *datos_ambulatorios_2012*: doctor, specialty, patients, visits and ambulatory surgeries in 2012.
- *datos_departamentos*: doctor, year (2008-2012) and locality (Colombian department).
- *costos_promedios_ambulatorios*: specialty, year (2008-2012) and mean cost of visit and ambulatory surgery (in USD, United States Dollar).

00_erase_doc_name_datos_ambulatorios_ano code file removes the names of the doctors from the first five csv files, substituting them for identification codes in order to preserve data confidentiality. It creates five new csv files to replace those with confidentiality issues:

- *datos_ambulatorios_2008_new*.
- *datos_ambulatorios_2009_new*.
- *datos_ambulatorios_2010_new*.
- *datos_ambulatorios_2011_new*.
- *datos_ambulatorios_2012_new*.

The new csv files are saved in the original_data directory. The csv files with confidentiality issues are not made available precisely for that reason.

01_data_outpatient_years code file organize the newly created five csv files into python pandas dataframes, which are then saved in the data directory as the following csv files (please, check the data directory):

- *data_outpatient_2008*.
- *data_outpatient_2009*.
- *data_outpatient_2010*.
- *data_outpatient_2011*.
- *data_outpatient_2012*.

01_optional_data_outpatient_year_details code file is optional in case the reader wishes to see the tasks that are being performed in 01_data_outpatient_years code file step-by-step (just choose a year in the second cell). Otherwise, the reader can skip it.

02_mean_outpatient_costs_usd code file deals with the *costos_promedios_ambulatorios* csv file, organizing it into a python pandas dataframe and saving it as *mean_outpatient_costs_usd* csv file in the data directory. That code file also translates the specialty of each doctor from Spanish to English using Google Translate.

03_data_departments code file organizes the *datos_departamentos* csv file into a python pandas dataframe, saving it as *data_departments* csv file in the data directory.

04_panel code file put together all the previously created csv files in the data directory to form a panel data. It concatenates all the *data_outpatient* files from 2008 to 2012 and then merges it with *mean_outpatient_costs_usd* and *data_departments* files. The resulting panel data is saved in the data directory as *data_panel* csv file. It contains 8294 rows and 12 columns of information that is going to be analyzed in the present study. Table 2.1 identifies the 12 columns. It is worth noting that 10 rows were discarded because of errors involving the doctor's specialty name or specialty identification code.

Table 2.1

doc_code	doctor identification code.
spec_code	specialty identification code.
spec_es	specialty name in Spanish.
spec_en	specialty name in English.
year	year from 2008 to 2012.
n_visits	number of visits
n_surgeries	number of ambulatory surgeries.
n_patients	number of patients.
c_visit	mean cost of a visit per specialty in each year.
c_surgery	mean cost of an ambulatory surgery per specialty in each year.
dep_code	department identification code.
department	department name.

3-) First analyses.

05_first_analyses code file contains the first analyses made using the resulting *data_panel* csv file in the data directory. Because many medical specialties and departments (localities) don't have enough data, I decided to focus on 8 specialties and 9 departments that do have significant amount of data for all years between 2008 and 2012. Table 3.1 shows the 8 specialties and 9 departments:

Table 3.1

Medical specialties (identification code and name)	Departments (identification code and name)
- 137, 'general surgery' - 143, 'plastic surgery' - 200, 'dermatology' - 341, 'gynecology and obstetrics' - 480, 'ophthalmology' - 514, 'orthopedics and traumatology' - 521, 'otorhinolaryngology' - 750, 'urology'	- 5, 'antioquia' - 8, 'atlantico' - 11, 'bogota' - 13, 'bolivar' - 17, 'caldas' - 54, 'n. de santander' - 66, 'risaralda' - 68, 'santander' - 76, 'valle del cauca'

Figure 3.1 shows the 8 specialties. They all have more than 50 doctors observed per year for all five years as represented by doc_count.

Figure 3.1

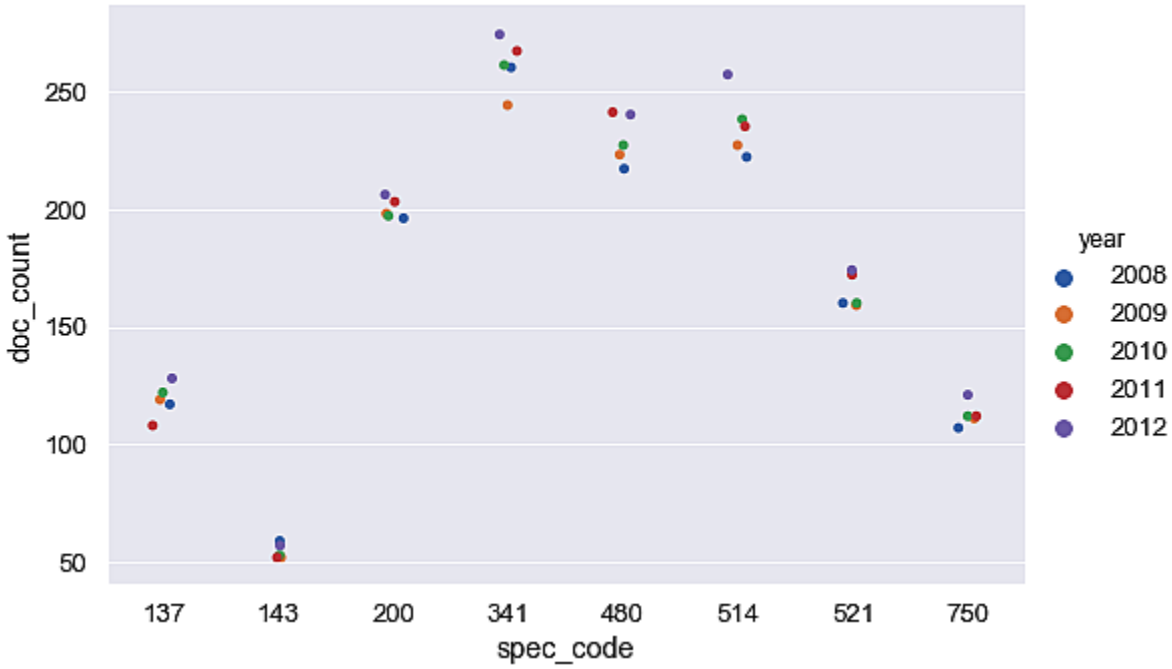
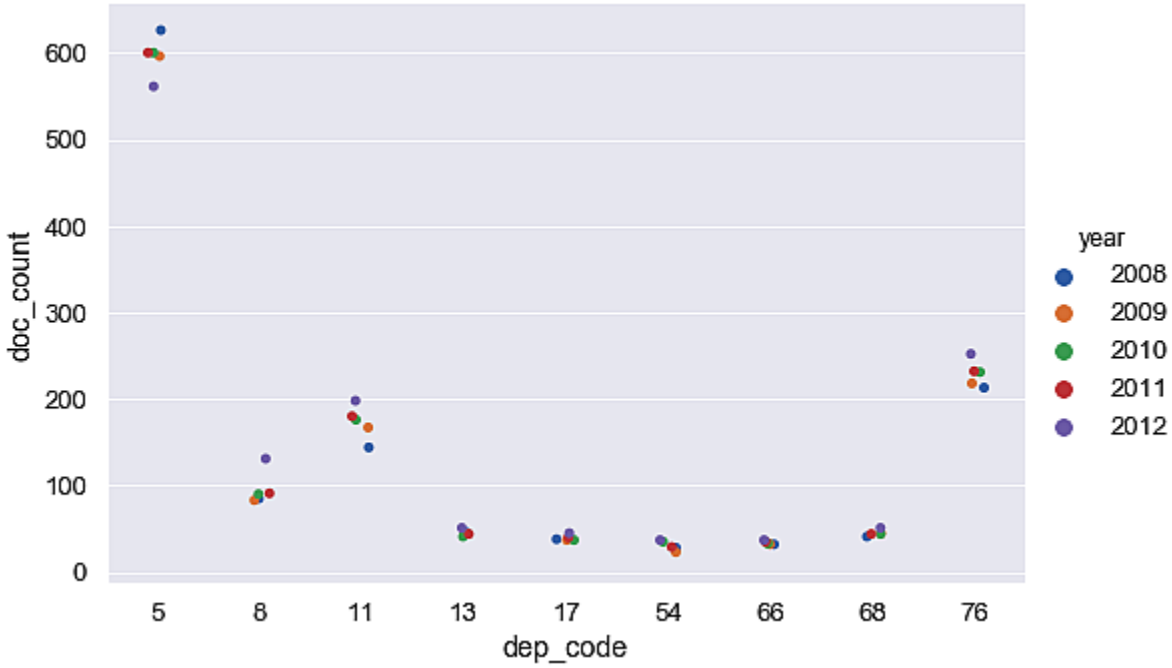


Figure 3.2 shows the 9 departments. They all have more than 22 doctors observed per year for all five years as represented by doc_count.

Figure 3.2



The subset of *data_panel* csv file with only those 8 specialties and 9 departments contains 6444 entries. Based on that subset, Figure 3.3 shows that the variables n_patients and n_visits are not the same but are closely related, so I decided to use only one of them (n_visits) in the present study.

Figure 3.3

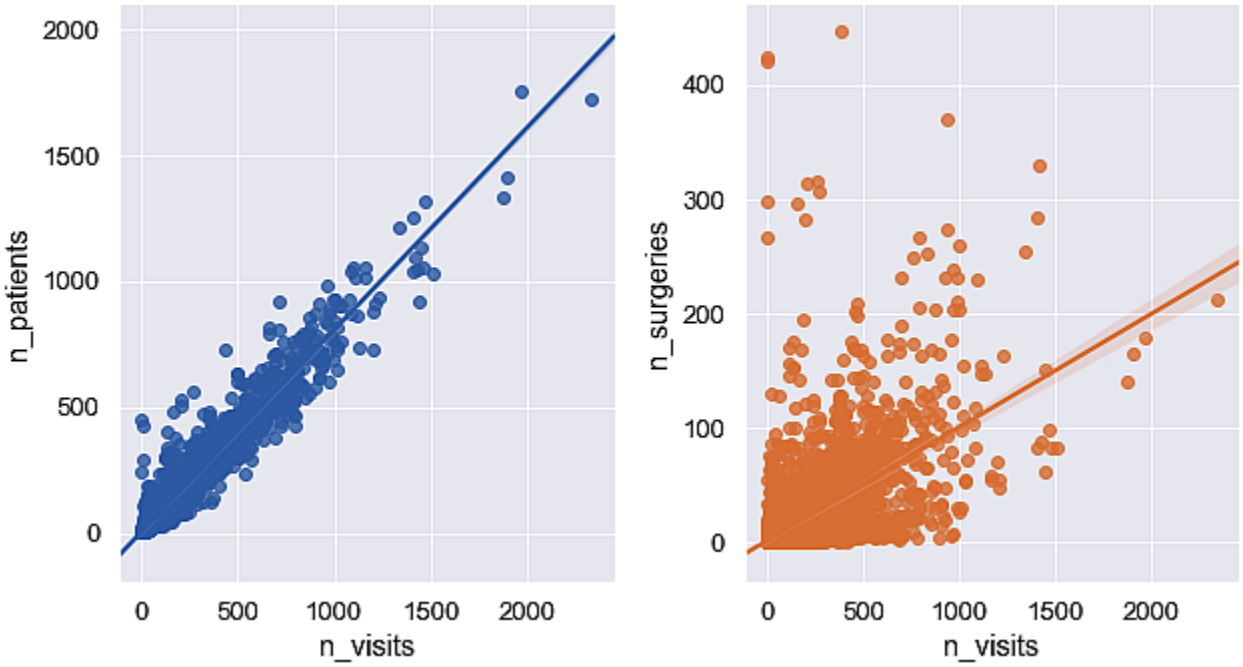


Figure 3.3 also shows that n_visits by itself may not be as good a predictor of n_surgeries as might be expected, meaning that other factors such as specialty, locality and time (or perhaps moral hazard and supplier-induced demand) may be interfering with that relationship.

One problem still remains: when the number of doctors in the 8 specialties and 9 departments are cross-checked (i.e., when the information in figures 3.1 and 3.2 are cross-checked), it becomes clear that some specialties in some departments in some years still don't have enough data or don't have data at all. For example, there is no data for plastic surgery in Bolivar department in 2010 or for general surgery in Santander department in 2011. It is easy to spot the problem with a simple calculation. If there are 8 specialties, 9 departments and 5 years, and each specialty in each department in each year should have at least 20 entries, then the total number of entries should be at least $8 \cdot 9 \cdot 5 \cdot 20 = 7200$, but the total number of entries for the current subset is 6444.

4-) Correlations.

06_correlations code file contain analyses involving correlations and statistical tests using bootstrap techniques. Because of the remaining problem abovementioned, I decided to narrow the focus a bit more and work with a smaller subpanel (i.e., a smaller subset of *data_panel* csv file in the data directory) containing 4 specialties, 2 departments and 5 years that obeys the following condition: having more than 30 observations (doctors) for each specialty in each department in each year. The resulting subpanel has 2552 entries, more than the minimum required to obey the condition ($4 \cdot 2 \cdot 5 \cdot 30 = 1200$). Table 4.1 shows the 4 specialties and 2 departments.

Table 4.1

Medical specialties (identification code and name)	Departments (identification code and name)
- 200, 'dermatology' - 341, 'gynecology and obstetrics' - 480, 'ophthalmology' - 521, 'otorhinolaryngology'	- 5, 'antioquia' - 76, 'valle del cauca'

In addition, the variable surgeries2visits is created to represent the ratio between number of surgeries and number of visits (to be used if necessary). Dummy variables for each specialty and department are also created. Table 4.2 identifies the columns in subpanel (it updates Table 2.1).

Table 4.2

doc_code	doctor identification code.
spec_code	specialty identification code.
spec_es	specialty name in Spanish.
spec_en	specialty name in English.
year	year from 2008 to 2012.
n_visits	number of visits
n_surgeries	number of ambulatory surgeries.
n_patients	number of patients.
c_visit	mean cost of a visit per specialty in each year.
c_surgery	mean cost of an ambulatory surgery per specialty in each year.
dep_code	department identification code.
department	department name.
surgeries2visits	ratio between n_surgeries and n_visits.

spec_der	dummy variable for 200, 'dermatology'.
spec_gyn	dummy variable for 341, 'gynecology and obstetrics'.
spec_oph	dummy variable for 480, 'ophthalmology'.
spec_oto	dummy variable for 521, 'otorhinolaryngology'.
dep_ant	dummy variable for 5, 'antioquia'.
dep_val	dummy variable for 76, 'valle del cauca'.

Figure 4.1

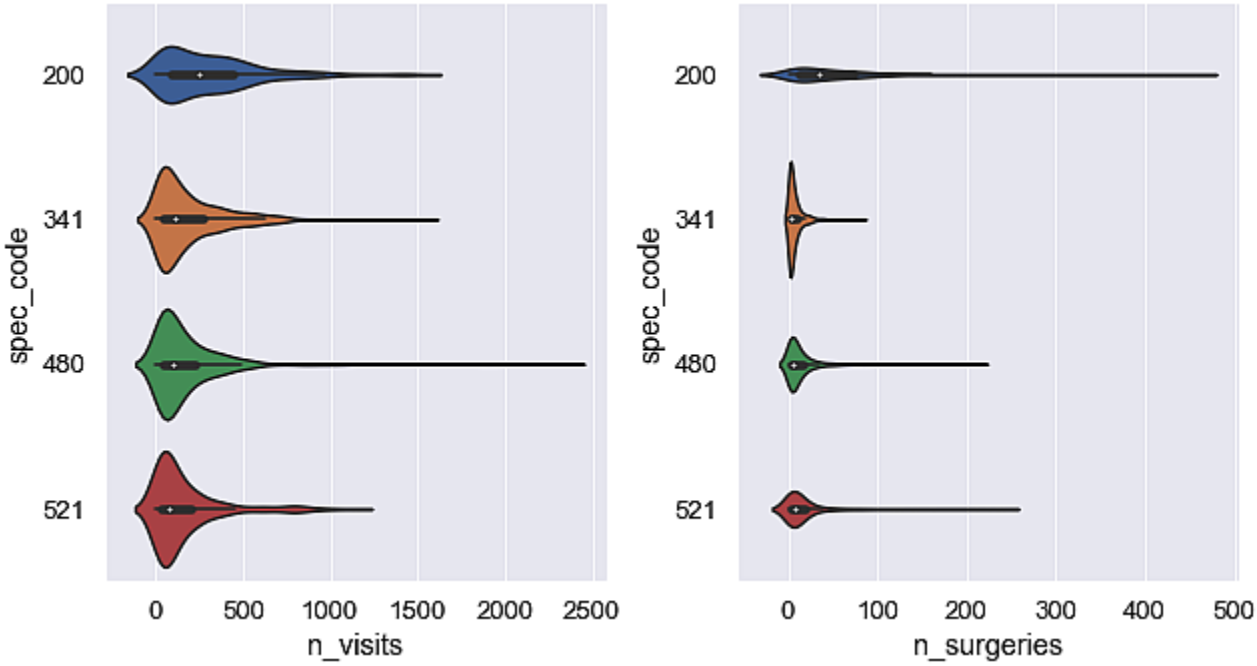


Figure 4.2

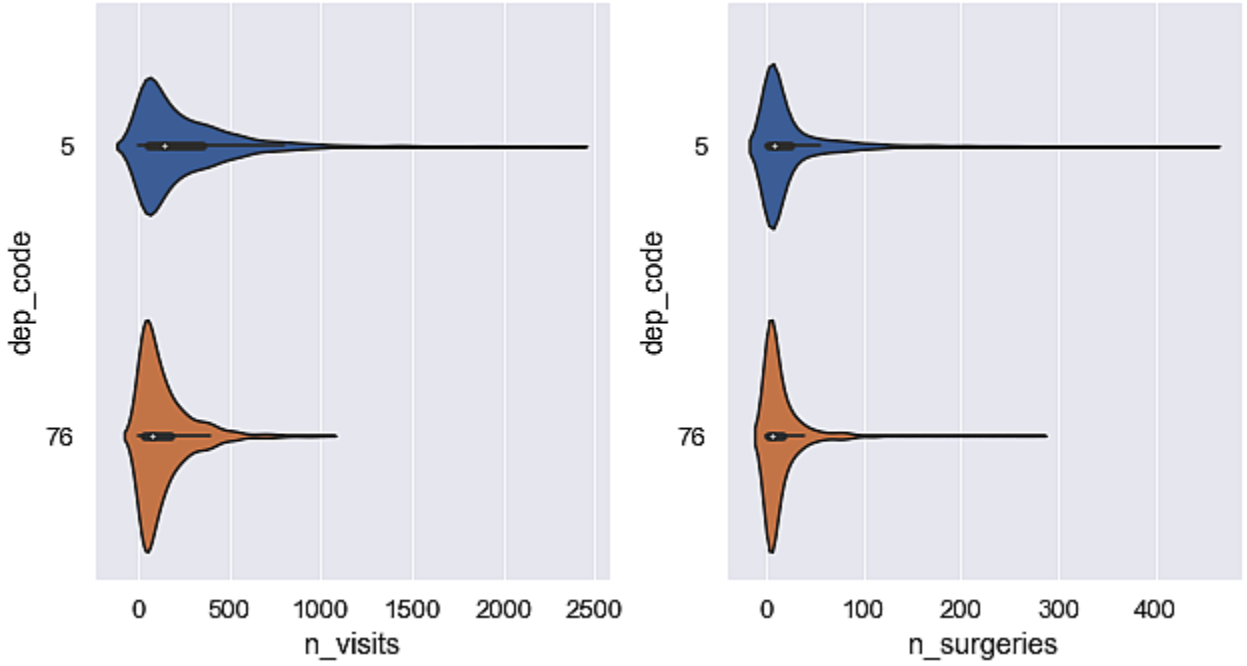
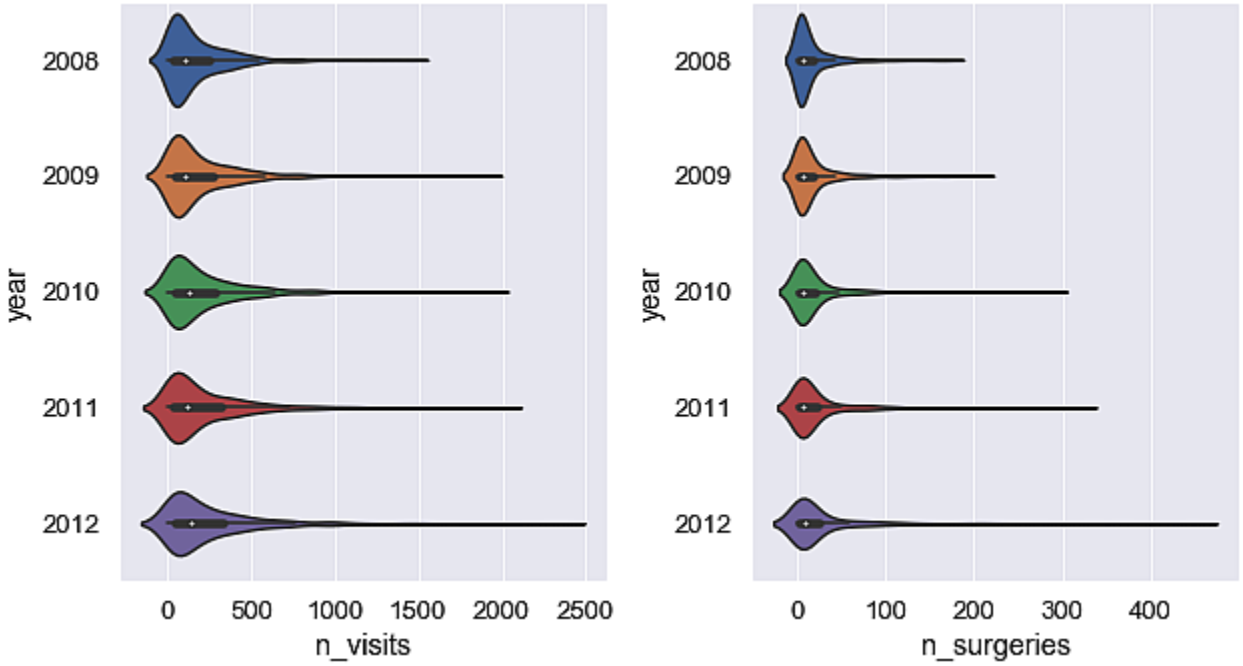


Figure 4.3

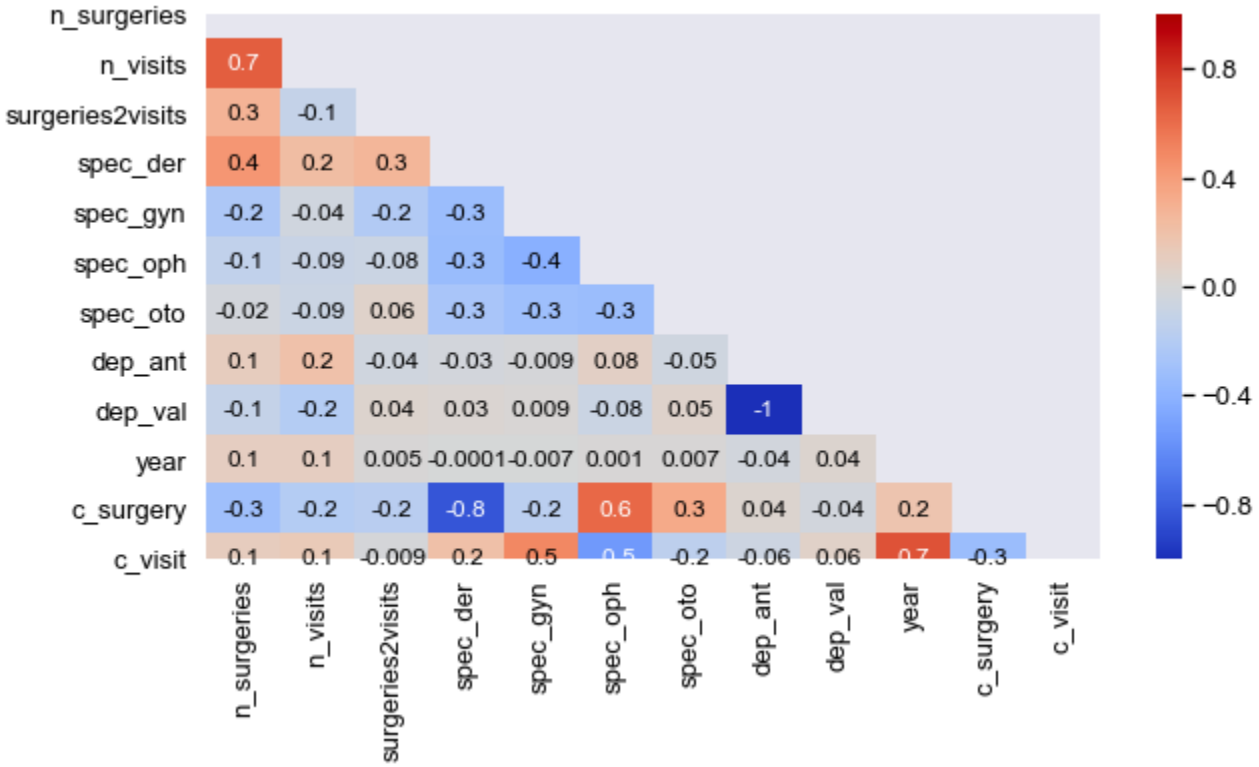


Based on that subpanel, figures 4.1, 4.2 and 4.3 show the distributions of n_visits and n_surgeries across respectively specialty, department and year. In each violin image, width illustrates the range of observations for n_visits or n_surgeries while height illustrates the frequency of each observation. Between specialty, department and year, specialty seems to be the main factor affecting n_visits and n_surgeries, followed by department and then year.

Figure 4.4 is a heatmap plot illustrating correlations between some selected variables. The numbers inside each box are linear correlation coefficients. They assume values among -1 for perfect negative linear correlation, 0 for no linear correlation and 1 for perfect positive linear correlation. The variables spec_code and dep_code are not included because their identification codes represent an arbitrary scale that may cause random correlations with other variables. Instead, the dummy variables for specialty and department are included. However, it is important to keep in mind that correlations between dummy variables for the same category are to be expected and do not mean anything special (e.g., linear correlation coefficient of -1 between dep_ant and dep_val).

Figure 4.4 indicates a strong positive linear correlation between n_visits and n_surgeries. However, Figure 3.3 has showed that relationship may need to take into account other factors like specialty, locality and time. No strong correlations seem to exist between specialty, locality and time, which is good if they are going to be used as predictors in regressions. As previously explained, the variables c_visit and c_surgery are not disaggregated: they inform the mean cost of a visit and a surgery per specialty in each year. Because of that, c_visit and c_surgery seem to present substantial correlations with some of the dummy variables for specialty as well as year. It may not be advisable to put these variables together as predictors in regressions.

Figure 4.4



In order to better understand the influence of specialty, locality and time on n_visits and n_surgeries, a series of statistical tests were performed using bootstrap techniques.

Table 4.3

H_0: difference in means of visits is by chance		p-value:	result:
specialty	dermatology vs. gynecology and obstetrics	0.0	H_0 is rejected
	dermatology vs. ophthalmology	0.0	H_0 is rejected
	dermatology vs. otorhinolaryngology	0.0	H_0 is rejected
	gynecology and obstetrics vs. ophthalmology	0.05	H_0 is not rejected
	gynecology and obstetrics vs. otorhinolaryngology	0.02	H_0 is rejected
	ophthalmology vs. otorhinolaryngology	0.27	H_0 is not rejected
locality	antioquia vs. valle del cauca	0.0	H_0 is rejected
year	2008 vs. 2009	0.034	H_0 is not rejected
	2008 vs. 2010	0.001	H_0 is rejected
	2008 vs. 2011	0.001	H_0 is rejected
	2008 vs. 2012	0.0	H_0 is rejected
	2009 vs. 2010	0.12	H_0 is not rejected
	2009 vs. 2011	0.058	H_0 is not rejected
	2009 vs. 2012	0.003	H_0 is rejected
	2010 vs. 2011	0.333	H_0 is not rejected
	2010 vs. 2012	0.044	H_0 is not rejected
	2011 vs. 2012	0.066	H_0 is not rejected

Table 4.4

H_0: difference in means of ambulatory surgeries is by chance		p-value:	result:
specialty	dermatology vs. gynecology and obstetrics	0.0	H_0 is rejected
	dermatology vs. ophthalmology	0.0	H_0 is rejected
	dermatology vs. otorhinolaryngology	0.0	H_0 is rejected
	gynecology and obstetrics vs. ophthalmology	0.0	H_0 is rejected
	gynecology and obstetrics vs. otorhinolaryngology	0.0	H_0 is rejected
	ophthalmology vs. otorhinolaryngology	0.0	H_0 is rejected
locality	antioquia vs. valle del cauca	0.0	H_0 is rejected
year	2008 vs. 2009	0.049	H_0 is not rejected
	2008 vs. 2010	0.006	H_0 is rejected
	2008 vs. 2011	0.0	H_0 is rejected
	2008 vs. 2012	0.0	H_0 is rejected
	2009 vs. 2010	0.156	H_0 is not rejected
	2009 vs. 2011	0.027	H_0 is not rejected
	2009 vs. 2012	0.0	H_0 is rejected
	2010 vs. 2011	0.181	H_0 is not rejected
	2010 vs. 2012	0.025	H_0 is rejected
	2011 vs. 2012	0.114	H_0 is not rejected

The first test had the following H_0 hypothesis: difference in means of visits or ambulatory surgeries for different categories is by chance (i.e., the means actually belong to the same distribution). The test requires a p-value equal to or smaller than 0.025 to reject H_0.

Tables 4.3 and 4.4 show the results. Most comparisons involving specialty and locality reject H_0. There is significant probability that the difference in means of visits could be by chance only when comparing gynecology and obstetrics vs. ophthalmology and ophthalmology vs. otorhinolaryngology. In the category year, H_0 is not rejected in the short run when comparing close years, but it is rejected in the long run when comparing distant years.

The second test had the following H_0 hypothesis: there is no difference in means of visits or ambulatory surgeries for different categories. For this test, 95% confidence intervals for the means were used. If intervals overlap, H_0 is not rejected. If intervals do not overlap, H_0 is rejected.

Figures 4.5 and 4.6 show the results, which are in line with those saw in the first test. There is significant probability that the means of visits are the same for otorhinolaryngology, ophthalmology and gynecology & obstetrics. Those are the three specialties whose difference in means of visits could be by chance according to the first test. There is also significant probability that the means are the same in the short run, when observing close years, but not in the long run, when observing distant years. In all other situations, H_0 is rejected, meaning that the means are most likely different. Thus, there is considerable evidence showing that specialty, locality and year have some influence on n_visits and n_surgeries, at least in the subpanel being analyzed, although that influence seems to be clearer on n_surgeries than on n_visits.

Figure 4.5

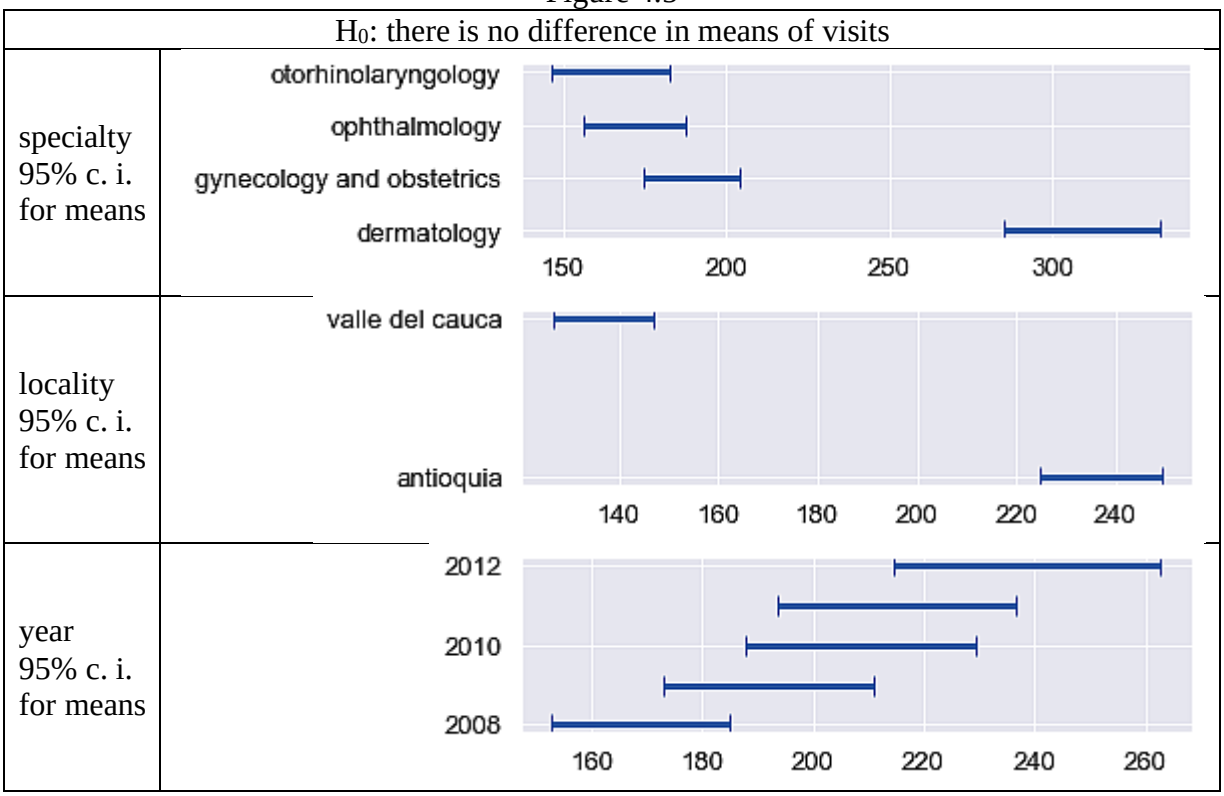
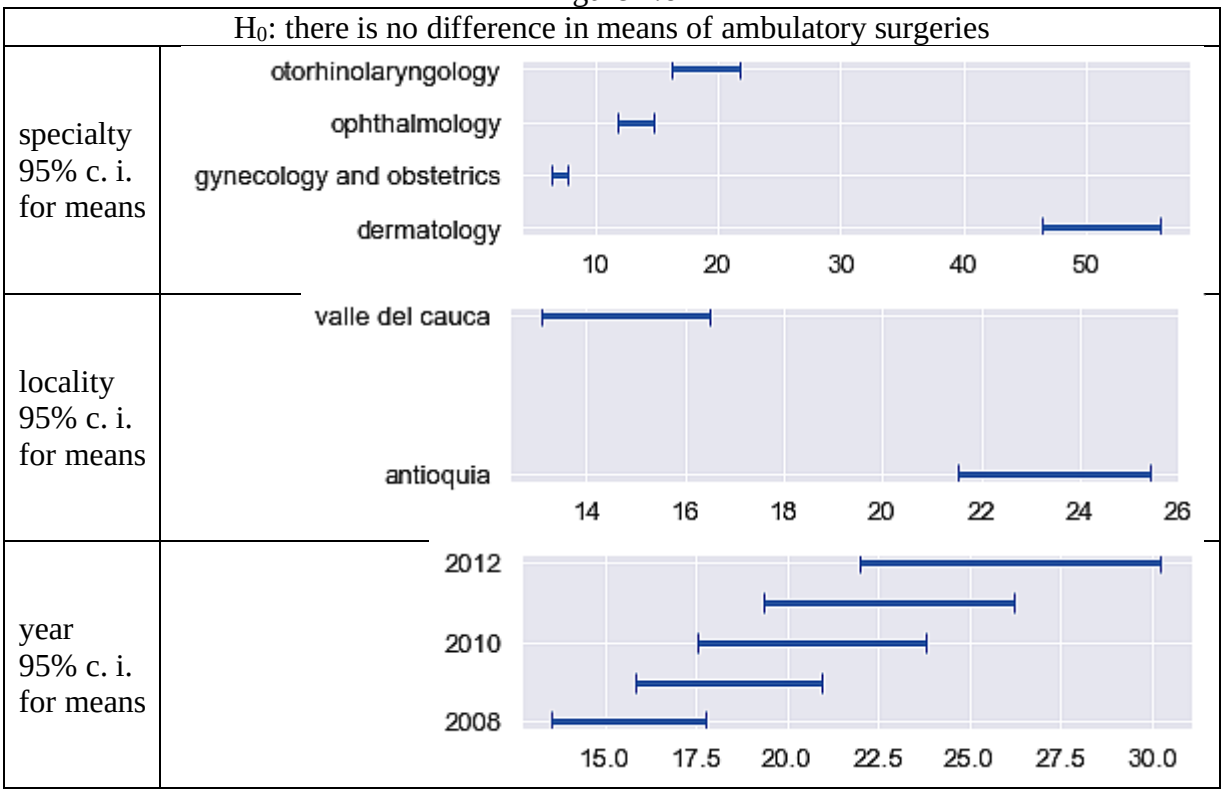
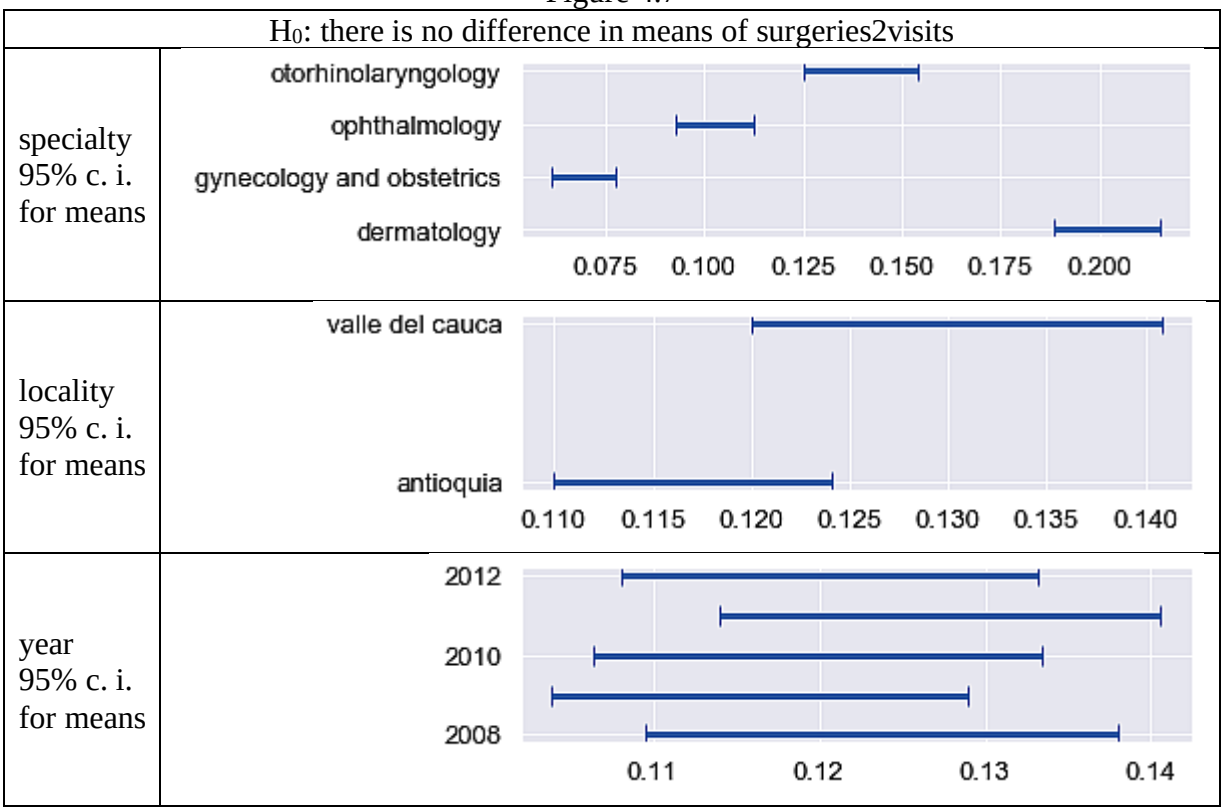


Figure 4.6



The application of the second test on the variable surgeries2visits provides interesting information. Remember: surgeries2visits is the ratio between n_surgeries and n_visits. In other words, it is merely n_surgeries divided by n_visits. Figure 4.7 shows the results. Where the means of n_surgeries and n_visits present analogous patterns, as in the categories locality and year in tables 4.5 and 4.6, the means of surgeries2visits are not significantly different. Where the means of n_surgeries and n_visits present distinct patterns, as in the category specialty in tables 4.5 and 4.6, the means of surgeries2visits are indeed significantly different. Consequently, it is quite possible to have the test on n_surgeries and n_visits rejecting H_0 but the test on surgeries2visits not rejecting it for the same category. Analyzing the behavior of surgeries2visits against different categories represents a first approach to looking at how those categories may affect the relationship between n_visits and n_surgeries. The results suggest an important role for specialty.

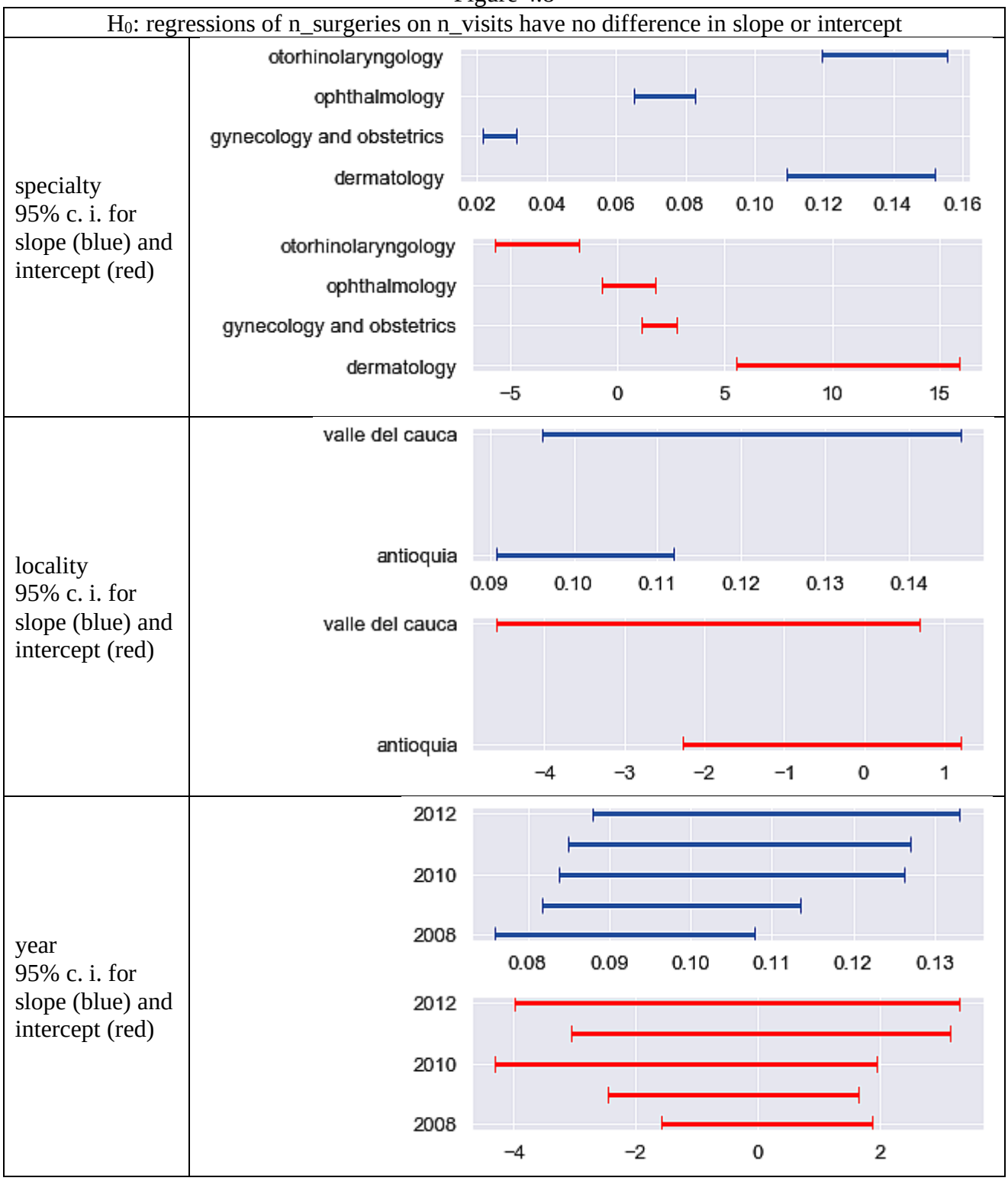
Figure 4.7



Finally, the third test was specifically intended to capture the effect of specialty, department and year on the relationship between n_visits and n_surgeries seen in Figure 3.3. The test had the following H_0 hypothesis: regressions of n_surgeries on n_visits have no difference in slope or intercept for different categories. As before, for this test, 95% confidence intervals for slope and intercept were used. If intervals overlap, H_0 is not rejected. If intervals do not overlap, H_0 is rejected.

Figure 4.8 shows the results. There is a significant chance that otorhinolaryngology and dermatology have the same slope, but not the same intercept. There is also a significant chance that ophthalmology and gynecology & obstetrics have the same intercept, but not the same slope. H_0 is not rejected when comparing different localities and years.

Figure 4.8



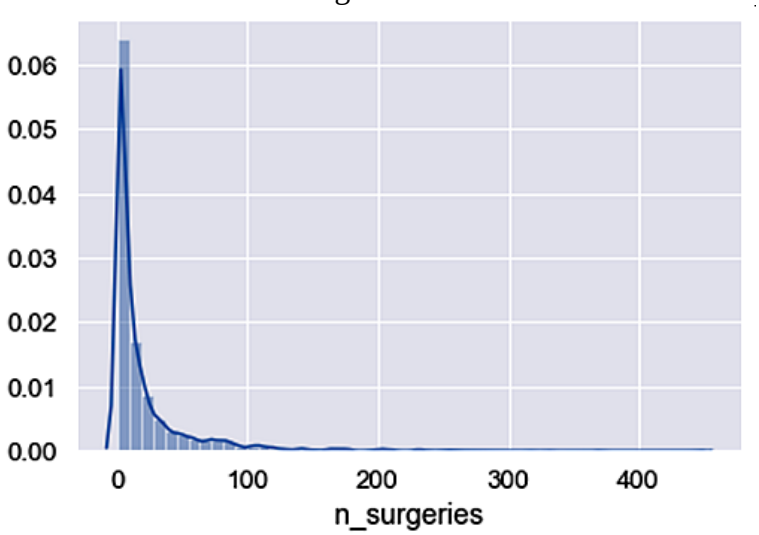
It is worth mentioning that uncertainty, as measured by the size of the confidence interval, seems to be greater for Valle del Cauca than for Antioquia. Uncertainty also seems to increase over the years, and the confidence intervals for the slope tends to move to the right.

5-) Supervised classification and regression.

07_supervised_classification_and_regression code file performs supervised classification to predict excessive number of ambulatory surgeries and regression to predict the number of ambulatory surgeries in general. The tasks are performed using the subpanel cited in section 4 and detailed in tables 4.1 and 4.2. It has 2552 entries containing 4 specialties, 2 departments and 5 years. There are more than 30 observations (doctors) for each specialty in each department in each year.

The definition of excessive number of ambulatory surgeries is subjective and can be easily changed in the code file. Because n_surgeries is not normally distributed, as showed in Figure 5.1, instead of using mean and standard deviation, I use a certain percentile to define the threshold for what constitutes excessive number of ambulatory surgeries. More specifically, I use the 70th percentile, but that threshold can be changed in the very beginning of cell 4 in the code file by altering the value of 'v_percentile'.

Figure 5.1



Having defined excessive number of ambulatory surgeries as n_surgeries equal to or greater than the 70th percentile of the data, the following supervised classification algorithms were implemented in order to predict it: k-nearest neighbors (k-NN), logistic regression (LogR), decision tree with randomized search of parameters (DTree) and support vector machine (SVM). They all use the same list of attributes: n_visits, year, spec_gyn, vis_spec_gyn, spec_oph, vis_spec_oph, spec_oto, vis_spec_oto, dep_ant and vis_dep_ant. Please, refer to Table 4.1 to identify the variables in that list. Variables starting with "vis" are interaction variables between n_visits and dummies (e.g., vis_spec_gyn is n_visits multiplied by spec_gyn). They are important to capture joint effects on the target variable (i.e., on the possible occurrence of excessive number of ambulatory surgeries). It is worth noting that dummies for specialty dermatology (spec_der) and department valle del cauca (dep_val) were left out of the list to avoid perfect collinearity with the other dummies. The dataset used in all algorithms was standardized using StandardScaler from the sklearn.preprocessing package. Moreover, 5-fold cross-validation procedure was applied.

Table 5.1

question: excessive number of ambulatory surgeries?						
k-NN n_neighbors: 7	accuracy	answer	precision	recall	f1-score	support
	0.89	0 (no)	0.91	0.95	0.93	449
		1 (yes)	0.86	0.77	0.81	189
		average	0.88	0.86	0.87	638
		weighted avg.	0.89	0.89	0.89	638
LogR C: 0.413838383838384	accuracy	answer	precision	recall	f1-score	support
	0.88	0 (no)	0.89	0.94	0.92	449
		1 (yes)	0.84	0.74	0.78	189
		average	0.87	0.84	0.85	638
		weighted avg.	0.88	0.88	0.88	638
DTree min_samples_leaf: 0.05 max_features: None max_depth: 2 criterion: entropy	accuracy	answer	precision	recall	f1-score	support
	0.85	0 (no)	0.88	0.92	0.90	449
		1 (yes)	0.78	0.70	0.74	189
		average	0.83	0.81	0.82	638
		weighted avg.	0.85	0.85	0.85	638
SVM (SVC) C: 100 gamma: 0.01	accuracy	answer	precision	recall	f1-score	support
	0.89	0 (no)	0.90	0.95	0.92	449
		1 (yes)	0.86	0.74	0.79	189
		average	0.88	0.84	0.86	638
		weighted avg.	0.89	0.89	0.88	638

Table 5.2

Confusion matrix k-NN n_neighbors: 7		predicted 0	predicted 1
	actual 0	425	24
	actual 1	44	145

Figure 5.2

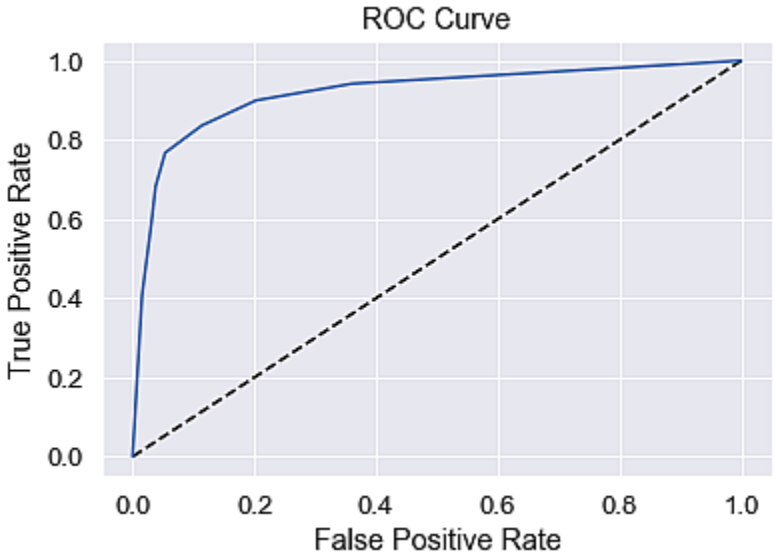


Table 5.1 shows the optimized parameters for each method as well as the reported results in terms of accuracy (the percentage of elements that was correctly classified), precision (the percentage of selected elements that is relevant) and recall (the percentage of relevant elements that was selected). F1-score, which is the harmonic mean of precision and recall, is also presented. Based on both accuracy and f1-score, the best supervised classification algorithm to use with the current dataset is k-NN with 7 neighbors.

Table 5.2 shows the confusion matrix for the k-NN classifier. Out of 449 actual "0" elements, it correctly predicted 425, and out of 189 actual "1" elements, it correctly predicted 145.

Finally, Figure 5.2 shows the Receiver Operating Characteristic curve (ROC curve) for the k-NN classifier. The Area Under the Receiver Operating Characteristics curve (AUROC) is 0.9182, very close to 1, meaning that the k-NN classifier applied to the current dataset has indeed a high predictive capacity.

Can the number of ambulatory surgeries in general be predicted? In order to answer that question, a regression model of n_surgeries on a list of variables was estimated. The list of variables is the same list of attributes previously used in the supervised classification algorithms. However, this time the dataset was not standardized to keep the interpretation of the regression as clear as possible. First, an Ordinary Least Square (OLS) estimation using the whole dataset was performed to observe the behavior of the residuals and look for outliers using the outlier_test from the statsmodels package. Once the outliers were identified and removed, the regression model was reestimated using natural logarithm on n_surgeries and n_visits as well as the 5-fold cross validation procedure. The transformation of n_surgeries and n_visits was necessary for the model to generate well-behaved residuals. The logarithmic transformation was possible because their minimum values in the dataset are 1, not zero.

Table 5.3

Variables	Coefficients	Standard Errors	t values	Probabilites
Intercept	-41.2329	44.557	-0.925	0.355
ln_n_visits	0.8051	0.067	12.098	0.000
year	0.0201	0.022	0.907	0.365
spec_gyn	-0.1667	0.351	-0.474	0.635
ln_vis_spec_gyn	-0.2870	0.070	-4.129	0.000
spec_oph	-0.5237	0.351	-1.493	0.136
ln_vis_spec_oph	-0.0952	0.070	-1.370	0.171
spec_oto	-0.4755	0.345	-1.379	0.168
ln_vis_spec_oto	-0.0422	0.071	-0.597	0.550
dep_ant	-0.1704	0.257	-0.663	0.508
ln_vis_dep_ant	0.0307	0.055	0.554	0.580

Table 5.3 shows the results of the reestimated regression model without outliers. The natural logarithms of n_surgeries and n_visits are represented by ln_n_surgeries and ln_n_visits, keeping in mind that ln_n_surgeries is the model's dependent variable. The variables starting with "ln_vis" are interaction variables between ln_n_visits and dummies (e.g., ln_vis_spec_gyn is ln_n_visits multiplied by spec_gyn). Again, the dummies spec_der and dep_val were left out of the regression to avoid perfect collinearity with the other dummies. As a result, the estimated coeficients for intercept and ln_n_visits assume that specialty is dermatology and department is valle del cauca. If dummies included in the

regression are activated (i.e., if they are set to 1), they will modify the coeficients for intercept and ln_n_visits in order to represent the activated specialty and/or department.

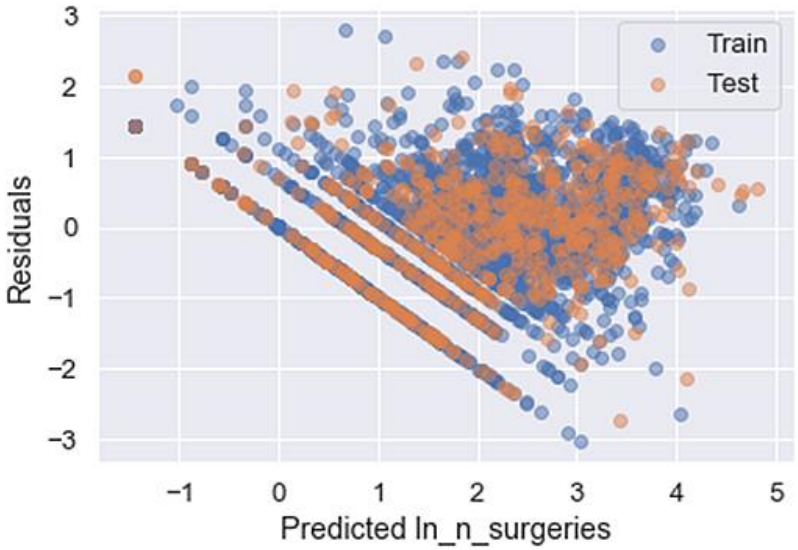
Surprisingly, most of the estimated coefficients are not statistically significant. Only ln_n_visits and ln_vis_spec_gyn are statistically significant, meaning that ln_n_surgeries may be explained by only these two regressors. Table 5.4 illustrates that possibility, maintaining the inclusion of the intercept.

Table 5.4

Variables	Coefficients	Standard Errors	t values	Probabilites
Intercept	-1.4520	0.119	-12.205	0.0
ln_n_visits	0.8068	0.025	32.510	0.0
ln_vis_spec_gyn	-0.2007	0.015	-13.764	0.0

The intercept -1.4520 is the expected arithmetic mean of ln_n_surgeries when ln_n_visits is 0. Thus, the exponentiated value exp(-1.4520) = 0.2341 is the expected geometric mean of n_surgeries when n_visits is exp(0) = 1. The coefficient 0.8068 of ln_n_visits means that for any 1% increase in n_visits, the expected increase in n_surgeries will be $1.01^{0.8068} - 1 = 0.8060\%$ if specialty is not gynecology & obstretics. The coefficient -0.2007 of ln_vis_spec_gyn means that for any 1% increase in n_visits, the expected increase in n_surgeries will actually be $1.01^{0.8068-0.2007} - 1 = 0.6049\%$ if specialty is gynecology & obstretics.

Figure 5.3



The model illustrated by Table 5.4 has a very parsimonious number of regressors and reached a good R^2 of 0.6569, but confidence in the results ultimately depends on the behavior of the residuals. Figure 5.3 shows a plot of the residuals against the predicted ln_n_surgeries. The residuals seem random around zero with no noticeable heteroscedasticity. The parallel lines to the left are more an oddity than a problem. They are caused by the fact that n_surgeries is composed of natural (not real) numbers with many low value observations (e.g., doctors performing only 1, 2 or 3 ambulatory surgeries in a year). The logarithmic transformation of those natural low value numbers produces jumps as seen in Figure 5.4, which plots ln_n_surgeries against ln_n_visits. The parallel lines in Figure 5.3 reflect those jumps in

Figure 5.4. The real issue is the possible existence of an upward trend in Figure 5.3 among the residuals, which would suggest that the model's specification needs adjustment. Despite that issue, the charts in figures 5.5 and 5.6 show that the residuals are indeed normally distributed around zero. The charts in Figure 5.5 compare the distribution of the train and test residuals against the expected normal distribution while the charts in Figure 5.6 do the same thing using Normal Q-Q plots.

Figure 5.4

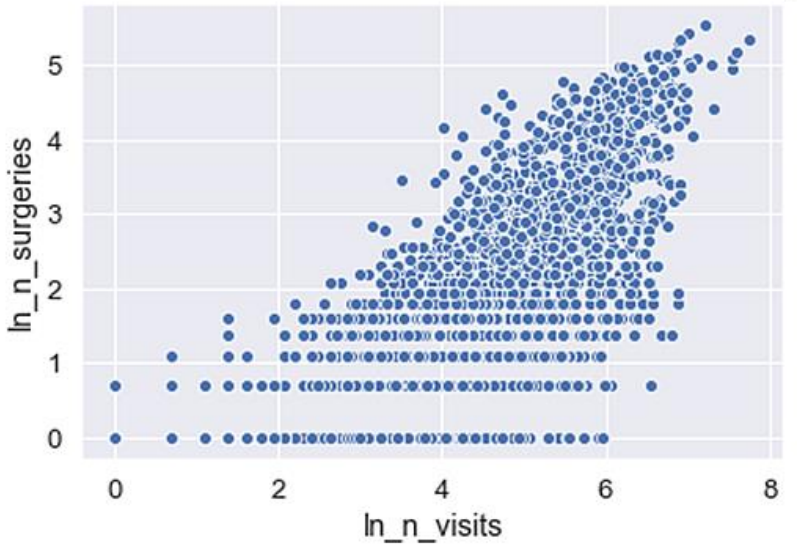


Figure 5.5

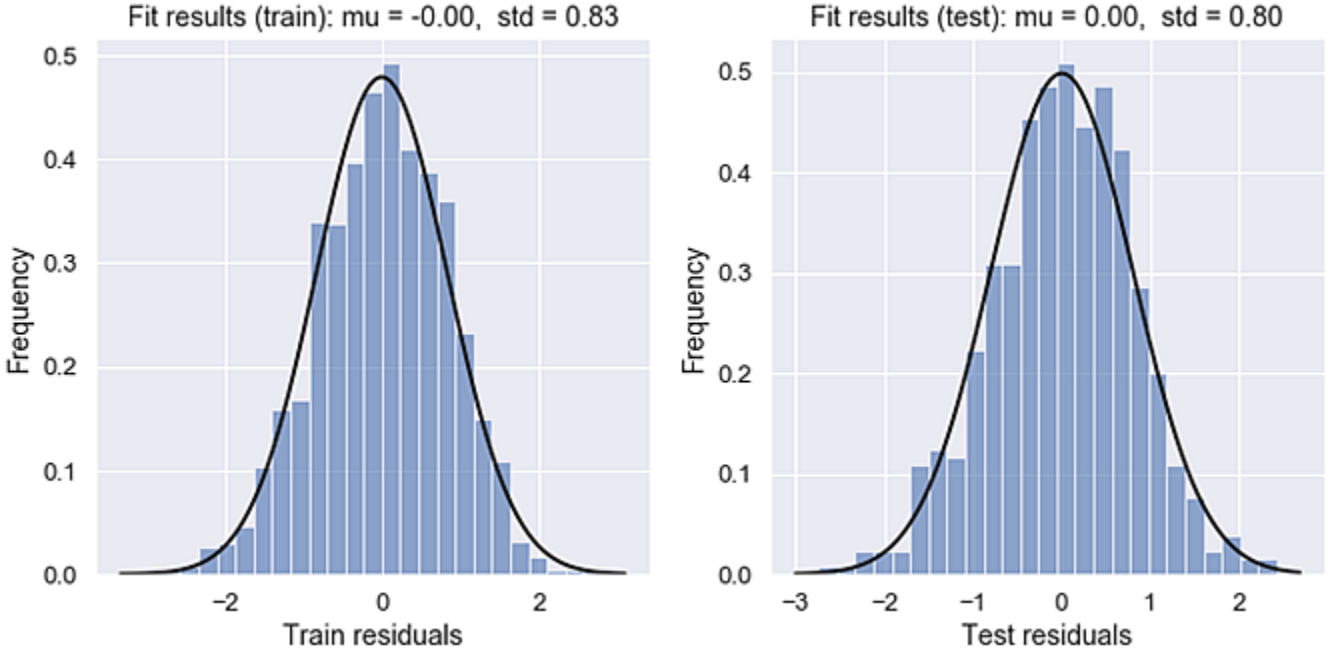
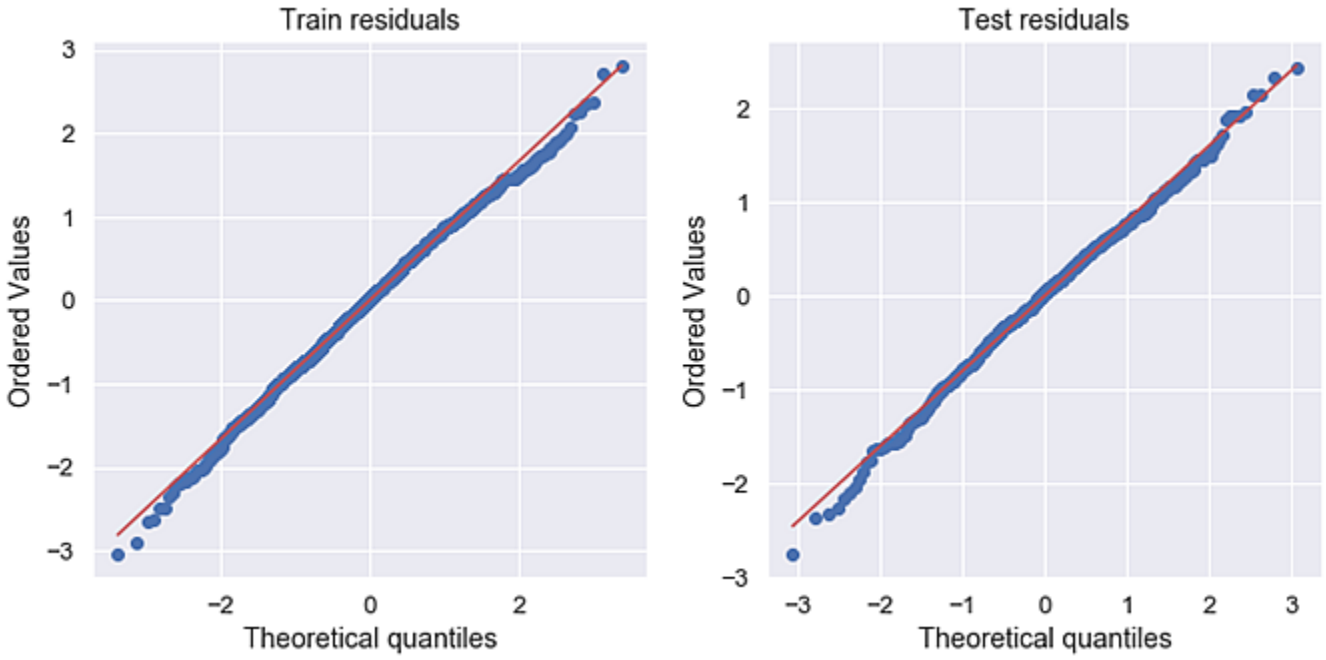
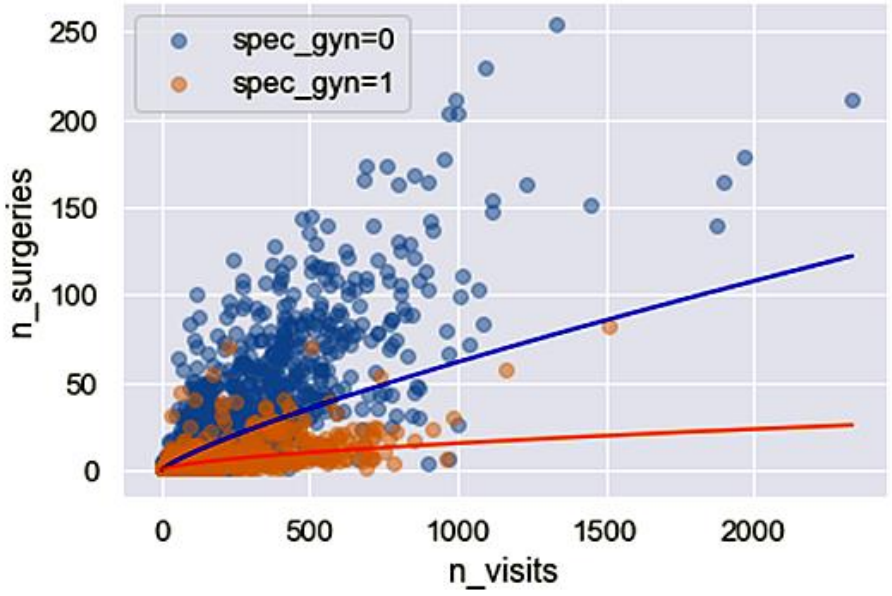


Figure 5.6



Finally, Figure 5.7 shows the estimated curves according to Table 5.4 in the plot of n_surgeries against n_visits. In particular, the curve for spec_gyn = 0 seems to suffer from biased errors for high values of n_visits. That could explain the possible upward trend among the residuals in Figure 5.3. There is enough evidence in the residuals that the relationship between n_surgeries and n_visits is nonlinear, but the nonlinearity needs to be better specified in the regression model.

Figure 5.7



6-) Final thoughts.

As mentioned in the beginning of this report, the main purpose of the present study is to understand the role of specialty, locality and time in predicting and explaining the variability of the medical practice in terms of number of patients, visits and ambulatory surgeries. By analyzing the role of those factors, it is also possible to have some insight about the possible influence of other factors such as moral hazard and supplier-induced demand.

The results are based on a dataset with 2552 entries containing 4 medical specialties, 2 Colombian departments and 5 years, and having more than 30 observations (doctors) for each specialty in each department in each year. Thus, the results are limited to a smaller number of specialties and localities. Among those three factors, specialty seems to be the most important, although year may be important for long run analysis.

The study shows that it is possible to perform supervised classification to predict excessive number of ambulatory surgeries successfully. Good scores were obtained using k-NN with 7 neighbors and having number of visits plus categorical dummies for specialty, department and year as attributes.

The task to predict and perhaps explain the number of ambulatory surgeries in general is trickier. The relationship between number of surgeries and number of visits does not seem to be linear, but a good nonlinear regression model has not been achieved yet. On the other hand, the nonlinearity between those variables is a possible symptom indicating that factors like moral hazard and supplier-induced demand should be examined. That’s because in a scenario where those factors were not important, everything else like specialty, locality and time being considered, it would be reasonable to expect a linear relationship between number of surgeries and number of visits, which is not the case.