**Part I: Research Question**

Research Question: *Which variables are significant in predicting the number of days a patient stays in the hospital during their initial visit?*

The objectives of the data analysis are to learn how we can predict what factors about a patient lead to a longer hospital stay. Based on our previous analysis of this dataset, we learned that there is a definite correlation between readmission rates and length of initial visit. If we can predict which patients will have a longer visit, we can detect which patients have a higher risk of readmission. Thus, we can adjust our care accordingly. The number of different patient characteristics we know will likely allow us to create a model that can predict, with a fair degree of accuracy, the length of an initial hospital visit. Aside from the possibility of readmission, knowing the estimated length of a hospital visit will be useful in administering care to a patient.

**Part II: Method Justification**

A multiple regression model assumes several stipulations. First, it assumes a linear relationship between multiple independent (predictor) variables and a dependent (target) variable, which can be verified by creating a scatterplot and visually inspecting the distribution of the points. A multiple regression model also assumes that each independent variable is not correlated with or dependent on any other independent variables. If this assumption is violated, it can create uncertainty regarding which independent variable has collinearity and/or causation with the dependent variable. Multiple regression models also assume normally distributed residuals (i.e. differences between the observed value and the value that the model predicts for that observation), and also that there are no major outliers that are powerfully influencing the model.

I’ll be using Python programming language; specifically, I’ll employ and utilize the pandas, NumPy, Matplotlib, Seaborn, and sklearn libraries. I’m familiar with Python and prefer to use it in a Jupyter notebook. As a general-purpose language, I know I will be able to do what I need to do. The code and syntax in Python and pandas is readable and comprehensible. Sklearn will allow me to use functions to build a regression model, while Seaborn and Matplotlib will allow visualization techniques that will be helpful for further understanding the data. Most importantly, I built regression models using Python and the rest of these libraries/packages before, so I will be comfortable doing so again.

Multiple regression is an appropriate technique to analyze my research question because it will allow us to take multiple variables into account when predicting a patient’s length of initial hospital visit. Thus, we will not be constrained to analyzing the relationship between initial days and only one single predictor variable – and instead, we can use multiple predictor variables. In the next performance assessment for this course, we will further build on this analysis to predict hospital readmission rates based on initial days. This analysis and multiple regression model will allow us to gather several factors when ultimately predicting readmission status. Furthermore, because our response variable is continuous and quantitative in nature, multiple linear regression is suitable.

**Part III: Data Preparation**

My data preparation goals begin with confirming the data is complete and accurate (i.e. clean), and if not, cleaning it as necessary. I will reduce the data to only the variables that I plan to use for my regression model. At that point, I will want to check for any outliers that will excessively influence my model and remove the row that holds the outlier. Since I know I’ll be working with several categorical values, I will need to convert all of those to dummy variables so they will work with the model. This will create several additional columns - I will also need to remove the original converted column. These manipulations will allow for data that is conducive to creating a multiple regression model.

Summary statistics for all variables are below. The blue highlighted variable is our response (dependent) variable, and the others are our predictor (independent) variables. The first table shows continuous variables, while the following table (on the next page) show categorical variables.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Lat** | 10000 | 38.75 | 5.40 | 17.97 | 35.26 | 39.42 | 42.04 | 70.56 |
| **Lng** | 10000 | -91.24 | 15.21 | -174.21 | -97.35 | -88.40 | -80.44 | -65.29 |
| **Population** | 10000 | 9965.25 | 14824.76 | 0 | 694.75 | 2769 | 13945 | 122814 |
| **Children** | 10000 | 2.10 | 2.16 | 0 | 0 | 1 | 3 | 10 |
| **Age** | 10000 | 53.51 | 20.64 | 18 | 36 | 53 | 71 | 89 |
| **Income** | 10000 | 40490.50 | 28521.15 | 154.08 | 19598.78 | 33768.42 | 54296.40 | 207249.10 |
| **VitD\_ levels** | 10000 | 17.96 | 2.02 | 9.81 | 16.63 | 17.95 | 19.35 | 26.39 |
| **Doc\_ visits** | 10000 | 5.01 | 1.05 | 1 | 4 | 5 | 6 | 9 |
| **Full\_meals eaten** | 10000 | 1.00 | 1.01 | 0 | 0 | 1 | 2 | 7 |
| **vitD\_supp** | 10000 | 0.40 | 0.63 | 0 | 0 | 0 | 1 | 5 |
| **Initial\_days** | 10000 | 34.46 | 26.31 | 1.00 | 7.90 | 35.84 | 61.16 | 71.98 |
| **TotalCharge** | 10000 | 5312.17 | 2180.39 | 1938.31 | 3179.37 | 5213.95 | 7459.70 | 9180.73 |
| **Additional charges** | 10000 | 12934.53 | 6542.60 | 3125.70 | 7986.49 | 11573.98 | 15626.49 | 30566.07 |
| **Timely admission** | 10000 | 3.52 | 1.03 | 1 | 3 | 4 | 4 | 8 |
| **Timely treatment** | 10000 | 3.51 | 1.03 | 1 | 3 | 3 | 4 | 7 |
| **Timely visits** | 10000 | 3.51 | 1.03 | 1 | 3 | 4 | 4 | 8 |
| **Reliability** | 10000 | 3.52 | 1.04 | 1 | 3 | 4 | 4 | 7 |
| **Options** | 10000 | 3.50 | 1.03 | 1 | 3 | 3 | 4 | 7 |
| **Hours\_of treatment** | 10000 | 3.52 | 1.03 | 1 | 3 | 4 | 4 | 7 |
| **Courteous staff** | 10000 | 3.49 | 1.02 | 1 | 3 | 3 | 4 | 7 |
| **Active listening** | 10000 | 3.51 | 1.04 | 1 | 3 | 3 | 4 | 7 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **count** | **unique** | **top** | **freq** |
| **Area** | 10000 | 3 | Rural | 3369 |
| **Marital** | 10000 | 5 | Widowed | 2045 |
| **Gender** | 10000 | 3 | Female | 5018 |
| **Soft\_drink** | 10000 | 2 | No | 7425 |
| **Initial\_admin** | 10000 | 3 | Emergency Admission | 5060 |
| **HighBlood** | 10000 | 2 | No | 5910 |
| **Stroke** | 10000 | 2 | No | 8007 |
| **Complication\_risk** | 10000 | 3 | Medium | 4517 |
| **Overweight** | 10000 | 2 | Yes | 7094 |
| **Arthritis** | 10000 | 2 | No | 6426 |
| **Diabetes** | 10000 | 2 | No | 7262 |
| **Hyperlipidemia** | 10000 | 2 | No | 6628 |
| **BackPain** | 10000 | 2 | No | 5886 |
| **Anxiety** | 10000 | 2 | No | 6785 |
| **Allergic\_rhinitis** | 10000 | 2 | No | 6059 |
| **Reflux\_esophagitis** | 10000 | 2 | No | 5865 |
| **Asthma** | 10000 | 2 | No | 7107 |
| **Services** | 10000 | 4 | Blood Work | 5265 |

See attached Jupyter notebook for data preparation steps and annotated code.

**Univariate Exploration:**

Chart, histogram

Description automatically generatedChart

Description automatically generated

Graphical user interface, application

Description automatically generated

A picture containing text, crossword puzzle, scoreboard, vector graphics

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

**Bivariate Exploration:**

A picture containing scatter chart

Description automatically generated

A picture containing text, sky, clouds, cloudy

Description automatically generated Chart

Description automatically generated

A picture containing text, sky, clouds, cloudy

Description automatically generatedChart, scatter chart

Description automatically generated

A picture containing text

Description automatically generated A picture containing text, sky, clouds, cloudy

Description automatically generated

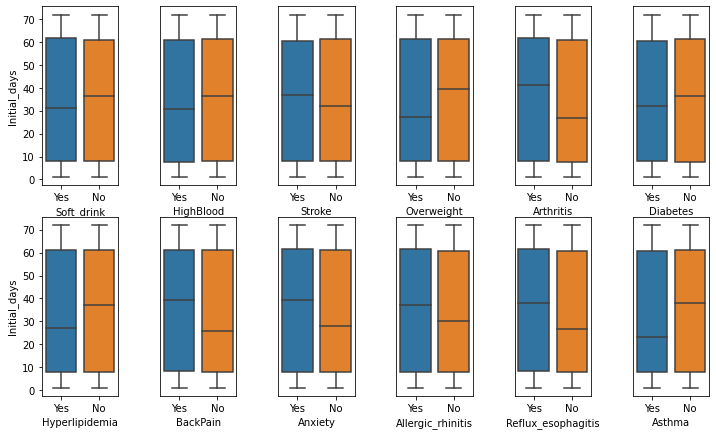
A picture containing text, sky, clouds, screen

Description automatically generated A picture containing text, sky, clouds, cloudy

Description automatically generated

A screenshot of a video game

Description automatically generated with low confidence



Prepared data set attached separately.

**Part IV: Model Comparison and Analysis**

**Initial Model:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dep. Variable:** | Initial\_days | **R-squared:** | 0.006 |  |  |  |
| **Model:** | OLS | **Adj. R-squared:** | 0.001 |  |  |  |
| **Method:** | Least Squares | **F-statistic:** | 1.284 |  |  |  |
| **Date:** | Thu, 16 Sep 2021 | **Prob (F-statistic):** | 0.0944 |  |  |  |
| **Time:** | 15:14:36 | **Log-Likelihood:** | -46859 |  |  |  |
| **No. Observations:** | 10000 | **AIC:** | 9.38E+04 |  |  |  |
| **Df Residuals:** | 9953 | **BIC:** | 9.42E+04 |  |  |  |
| **Df Model:** | 46 |  |  |  |  |  |
| **Covariance Type:** | nonrobust |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | 34.3678 | 4.534 | 7.579 | 0 | 25.48 | 43.256 |
| **Lat** | -0.0355 | 0.05 | -0.706 | 0.48 | -0.134 | 0.063 |
| **Lng** | -0.0162 | 0.017 | -0.929 | 0.353 | -0.05 | 0.018 |
| **Population** | 2.80E-05 | 1.82E-05 | 1.54 | 0.124 | -7.65E-06 | 6.37E-05 |
| **Children** | 0.2834 | 0.122 | 2.327 | 0.02 | 0.045 | 0.522 |
| **Age** | 0.0533 | 0.039 | 1.372 | 0.17 | -0.023 | 0.129 |
| **Income** | -1.23E-05 | 9.24E-06 | -1.335 | 0.182 | -3.04E-05 | 5.78E-06 |
| **Doc\_visits** | -0.1707 | 0.252 | -0.677 | 0.498 | -0.665 | 0.323 |
| **Full\_meals\_eaten** | -0.4253 | 0.262 | -1.626 | 0.104 | -0.938 | 0.087 |
| **vitD\_supp** | 0.6345 | 0.419 | 1.514 | 0.13 | -0.187 | 1.456 |
| **Timely\_admission** | -0.8292 | 0.379 | -2.187 | 0.029 | -1.572 | -0.086 |
| **VitD\_levels** | -0.0357 | 0.131 | -0.273 | 0.785 | -0.292 | 0.221 |
| **Additional\_charges** | -0.0001 | 0 | -0.885 | 0.376 | 0 | 0 |
| **Timely\_treatment** | 0.2377 | 0.35 | 0.679 | 0.497 | -0.448 | 0.924 |
| **Timely\_visits** | 0.0295 | 0.323 | 0.091 | 0.927 | -0.604 | 0.663 |
| **Reliability** | -0.3468 | 0.288 | -1.206 | 0.228 | -0.911 | 0.217 |
| **Options** | 0.026 | 0.303 | 0.086 | 0.932 | -0.568 | 0.62 |
| **Hours\_of\_treatment** | -0.0352 | 0.313 | -0.112 | 0.911 | -0.648 | 0.578 |
| **Courteous\_staff** | 0.3513 | 0.295 | 1.192 | 0.233 | -0.226 | 0.929 |
| **Active\_listening** | -0.0486 | 0.277 | -0.175 | 0.861 | -0.592 | 0.495 |
| **Area\_Suburban** | 0.324 | 0.644 | 0.503 | 0.615 | -0.938 | 1.586 |
| **Area\_Urban** | 0.7907 | 0.646 | 1.224 | 0.221 | -0.475 | 2.057 |
| **Marital\_Married** | 1.206 | 0.835 | 1.444 | 0.149 | -0.432 | 2.844 |
| **Marital\_Never Married** | 1.8454 | 0.839 | 2.199 | 0.028 | 0.201 | 3.49 |
| **Marital\_Separated** | 1.7674 | 0.838 | 2.109 | 0.035 | 0.124 | 3.41 |
| **Marital\_Widowed** | 1.72 | 0.833 | 2.065 | 0.039 | 0.087 | 3.353 |
| **Gender\_Male** | 0.4625 | 0.534 | 0.867 | 0.386 | -0.583 | 1.508 |
| **Gender\_Nonbinary** | 0.9316 | 1.839 | 0.507 | 0.612 | -2.673 | 4.536 |
| **Soft\_drink\_Yes** | 0.1849 | 0.603 | 0.307 | 0.759 | -0.997 | 1.366 |
| **Initial\_admin\_Emergency Admission** | -0.6761 | 0.648 | -1.043 | 0.297 | -1.947 | 0.595 |
| **Initial\_admin\_Observation Admission** | -0.2097 | 0.751 | -0.279 | 0.78 | -1.681 | 1.262 |
| **HighBlood\_Yes** | 0.8994 | 1.501 | 0.599 | 0.549 | -2.042 | 3.841 |
| **Stroke\_Yes** | -0.0807 | 0.662 | -0.122 | 0.903 | -1.378 | 1.216 |
| **Overweight\_Yes** | -0.6139 | 0.58 | -1.058 | 0.29 | -1.751 | 0.523 |
| **Arthritis\_Yes** | 0.9849 | 0.55 | 1.79 | 0.073 | -0.093 | 2.063 |
| **Diabetes\_Yes** | -0.0922 | 0.592 | -0.156 | 0.876 | -1.252 | 1.067 |
| **Hyperlipidemia\_Yes** | -0.2104 | 0.558 | -0.377 | 0.706 | -1.304 | 0.883 |
| **Anxiety\_Yes** | 0.665 | 0.564 | 1.179 | 0.238 | -0.441 | 1.771 |
| **Allergic\_rhinitis\_Yes** | 0.2061 | 0.539 | 0.382 | 0.702 | -0.851 | 1.263 |
| **Reflux\_esophagitis\_Yes** | 0.6447 | 0.535 | 1.204 | 0.228 | -0.405 | 1.694 |
| **Asthma\_Yes** | -0.8093 | 0.581 | -1.393 | 0.164 | -1.948 | 0.33 |
| **Services\_CT Scan** | 0.402 | 0.836 | 0.481 | 0.631 | -1.236 | 2.04 |
| **Services\_Intravenous** | -0.7728 | 0.595 | -1.299 | 0.194 | -1.939 | 0.393 |
| **Services\_MRI** | 0.8622 | 1.399 | 0.616 | 0.538 | -1.881 | 3.605 |
| **Complication\_risk\_Low** | 1.0494 | 0.735 | 1.428 | 0.153 | -0.392 | 2.49 |
| **Complication\_risk\_Medium** | -0.1151 | 0.604 | -0.191 | 0.849 | -1.299 | 1.068 |
| **BackPain\_Yes** | 0.9099 | 0.536 | 1.697 | 0.09 | -0.141 | 1.961 |
|  |  |  |  |  |  |  |
| **Omnibus:** | 41963.946 | **Durbin-Watson:** | 0.17 |  |  |  |
| **Prob(Omnibus):** | 0 | **Jarque-Bera (JB):** | 1259.998 |  |  |  |
| **Skew:** | 0.071 | **Prob(JB):** | 2.48E-274 |  |  |  |
| **Kurtosis:** | 1.267 | **Cond. No.** | 8.90E+05 |  |  |  |

An R-squared value of .006 says that even with all the variables we included, only 0.6% of our model fit the observed data. When adjusted for the number of variables, the R=squared is only 0.01. Additionally, our model violates several assumptions of a working multiple regression model, so we'll need to reduce it. We want to check for multicollinearity and take out any independent variables that are correlated with other variables. We can check for multicollinearity by calculating variance inflation factors (VIF’s), and take out variables with VIF’s above 10. For the model below, I will leave the constant/intercept in regardless of its VIF because much of the variables are binary and/or categorical, leading to a small variance, and thus increasing the VIF for the constant/intercept. Furthermore, I’m not worried about multicollinearity with something that is not an actual variable in my dataset. We should also take out variables with p-values above 0.05. These variables with high p-values are not statistically significant to the changes in our dependent variable, and are therefore not helping our model.

**Reduced model:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dep. Variable:** | Initial\_days | **R-squared:** | 0.001 |  |  |  |
| **Model:** | OLS | **Adj. R-squared:** | 0.001 |  |  |  |
| **Method:** | Least Squares | **F-statistic:** | 2.906 |  |  |  |
| **Date:** | Thu, 16 Sep 2021 | **Prob (F-statistic):** | 0.0126 |  |  |  |
| **Time:** | 15:19:18 | **Log-Likelihood:** | -46881 |  |  |  |
| **No. Observations:** | 10000 | **AIC:** | 9.38E+04 |  |  |  |
| **Df Residuals:** | 9994 | **BIC:** | 9.38E+04 |  |  |  |
| **Df Model:** | 5 |  |  |  |  |  |
| **Covariance Type:** | nonrobust |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 35.2005 | 1.021 | 34.481 | 0 | 33.199 | 37.202 |
| **Children** | 0.2811 | 0.122 | 2.312 | 0.021 | 0.043 | 0.519 |
| **Timely\_admission** | -0.5731 | 0.255 | -2.249 | 0.025 | -1.073 | -0.074 |
| **Marital\_Never\_Married** | 1.164 | 0.723 | 1.61 | 0.107 | -0.253 | 2.581 |
| **Marital\_Separated** | 1.2101 | 0.722 | 1.675 | 0.094 | -0.206 | 2.626 |
| **Marital\_Widowed** | 1.0295 | 0.715 | 1.439 | 0.15 | -0.373 | 2.432 |
|  |  |  |  |  |  |  |
| **Omnibus:** | 41311.402 | **Durbin-Watson:** | 0.161 |  |  |  |
| **Prob(Omnibus):** | 0 | **Jarque-Bera (JB):** | 1283.07 |  |  |  |
| **Skew:** | 0.071 | **Prob(JB):** | 2.43E-279 |  |  |  |
| **Kurtosis:** | 1.251 | **Cond. No.** | 19.3 |  |  |  |

Our reduced model looks a lot different, and no better. Since we removed so many variables, our R-squared value decreased, but our adjusted r-squared stayed at 0.01. Our model still does not fit well. Also, three of our p-values are now above 0.05, so if we want to further reduce this model, we can remove those variables.

Further reduced model:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dep. Variable:** | Initial\_days | **R-squared:** | 0.001 |  |  |  |
| **Model:** | OLS | **Adj. R-squared:** | 0.001 |  |  |  |
| **Method:** | Least Squares | **F-statistic:** | 5.015 |  |  |  |
| **Date:** | Thu, 16 Sep 2021 | **Prob (F-statistic):** | 0.00665 |  |  |  |
| **Time:** | 15:20:06 | **Log-Likelihood:** | -46883 |  |  |  |
| **No. Observations:** | 10000 | **AIC:** | 9.38E+04 |  |  |  |
| **Df Residuals:** | 9997 | **BIC:** | 9.38E+04 |  |  |  |
| **Df Model:** | 2 |  |  |  |  |  |
| **Covariance Type:** | nonrobust |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 35.8822 | 0.968 | 37.061 | 0 | 33.984 | 37.78 |
| **Children** | 0.2738 | 0.122 | 2.252 | 0.024 | 0.035 | 0.512 |
| **Timely\_admission** | -0.5687 | 0.255 | -2.231 | 0.026 | -1.068 | -0.069 |
|  |  |  |  |  |  |  |
| **Omnibus:** | 41245.176 | **Durbin-Watson:** | 0.16 |  |  |  |
| **Prob(Omnibus):** | 0 | **Jarque-Bera (JB):** | 1285.499 |  |  |  |
| **Skew:** | 0.071 | **Prob(JB):** | 7.20E-280 |  |  |  |
| **Kurtosis:** | 1.249 | **Cond. No.** | 16.9 |  |  |  |

My initial research question was related to *all* variables, so I included all variables that contained any patient information that could possibly be a predictor of days of initial hospital stay. I left out variables that were nonsensical codes (UID, customer\_id, etc.), or had too many categories to quantify with dummy variables (job), and/or extra categories related to location (county, zip, etc.). The reduced model kept all variables that were both statistically significant (based on a p-value of 0.05) and not correlated with each other (based on VIF values of 10). Variables with p-values above 0.05 or VIF values of above 10 were removed from the initial model.

Neither our initial nor reduced models are strong, based on our adjusted R-squared values of 0.01. This means that 0.1% of our data can be explained by our model, and that is a very, very weak model that doesn’t fit our data well.

A picture containing icon

Description automatically generated

Chart, scatter chart

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

Graphical user interface, diagram, table, Excel

Description automatically generated

Below is a screenshot of the first 10 and last 10 values of our model (Predicted value) vs. the actual values in the data (Observed value), and the residual (Observed value – Predicted value).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Predicted value** | **Observed value** | **Residual** |  |  | **Predicted value** | **Observed value** | **Residual** |
| **0** | 34.449955 | 10.58577 | -23.864185 |  | **9990** | 34.176189 | 62.73599 | 28.559801 |
| **1** | 34.997487 | 15.129562 | -19.867925 |  | **9991** | 34.176189 | 56.87419 | 22.698001 |
| **2** | 35.566169 | 4.772177 | -30.793992 |  | **9992** | 34.997487 | 56.61571 | 21.618223 |
| **3** | 34.176189 | 1.714879 | -32.46131 |  | **9993** | 32.74391 | 40.3553 | 7.61139 |
| **4** | 35.018637 | 1.254807 | -33.76383 |  | **9994** | 36.387467 | 37.93212 | 1.544653 |
| **5** | 34.428806 | 5.95725 | -28.471556 |  | **9995** | 34.723721 | 51.56122 | 16.837499 |
| **6** | 33.607507 | 9.05821 | -24.549297 |  | **9996** | 35.271254 | 68.66824 | 33.396986 |
| **7** | 37.229915 | 14.228019 | -23.001896 |  | **9997** | 34.997487 | 70.15418 | 35.156693 |
| **8** | 34.176189 | 6.180339 | -27.99585 |  | **9998** | 33.860124 | 63.3569 | 29.496776 |
| **9** | 33.586358 | 1.632554 | -31.953804 |  | **9999** | 35.797636 | 70.85059 | 35.052954 |

Supporting code attached in Jupyter notebook.

**Part V: Data Summary and Implications**

Equation of reduced regression model:

*Initial\_days* = 35.8822 + Children (\*0.2738) + Timely\_admission (\* -0.5687)

This equation means that the number of initial days a person stays in the hospital is: 35.822 plus the number of children they have (multiplied by .2738) plus their response to the importance of timely admission on a scale of 1-8, where 1 is the highest importance (multiplied by -0.5687). Thus, more children means more initial\_days, and a higher answer to timely admission importance (meaning a lower importance) means less initial\_days. These are both somewhat counterintuitive, because I would guess that a patient may want to leave the hospital sooner if they have more children (or maybe not). I would also think that, in general, a person who places less importance on timely admission might have a longer initial visit because they come in late more often.

Statically, this equation means next to nothing. An R-squared value of 0.01 implies that we have a model that explains 0.1% of our data, which is not enough to extrapolate any useful conclusions.

The limitations of any multiple regression model are significant, but our specific model does not even give us confidence to say any of this data is legitimate, at most, and at least, inconsequential. We also can’t ascertain the casual apparatus of any of these relationships (or lack thereof), but only the statistical magnitude of the relationship itself.

My recommendation at this point would be to investigate this data to be sure it is accurate. The fact that the distribution of the Initial\_days variable is bimodal is cause for skepticism, as is the lack of any intuitive presumptive correlation (for example, children do not increase with age in our dataset). If the data is indeed accurate, we need to collect more and different kinds of data, because what we have does not give us the information that we need to build a useful model. Our current model doesn’t give us enough information to accurately predict a person’s initial hospital stay.