Performance Assessment D208 – Predictive Modeling  
Task I

Doug Haunsperger

College of IT, Western Governors University

October 31, 2023

# Part I. Rese**arch Questi**on

## A1. Question Description

I am choosing to examine the question, “What might cause longer hospital stays?”. In other words, I will determine if any of the other data collected in our dataset have a correlation and possible causal relationship with the number of days spent in the hospital (Initial\_days).

## A2. Goals

The answer to this question could lead to improvements in patient treatment. For example, if it is shown that there are controllable factors that contribute to lengthier stays, the hospital can try to ameliorate those factors from the beginning of the stay, or recommend standards of care to primary care physicians prior to patients landing in the hospital. Alternatively, if it is shown that there are non-controllable factors (i.e. income, gender, etc.) the hospital can model lengths of stay for incoming patients and plan appropriate staffing levels over the coming weeks.

# Part II. Method Justification

## B1. Summary of Assumptions

A multiple linear regression model tries to predict a response variable Y based on different predictor variables Xn with different weights βn, and can be expressed as an equation:

Y = β0 + β1X1+ β2X2+ β2X3+… + βkXk  (Bobbit, 2020)

Some sources, such as Larose & Larose, additionally add a “+ ε” term to account for random error (2019, p. 151).

Liu et al. (2016) give four factors assumed to hold for linear regression models to be valid:

1. Linearity – there exists a linear relationship between each predictor variable and the response variable.
2. Homoscedasticity – the residual variances are the same for any value of the predictor variables.
3. Independence – each observation in the dataset is unrelated to the others.
4. Normality – for any fixed set of predictor variables, the response variable output is normally distributed.

## B2. Tool Benefits

Python, via the numpy, statsmodels, and scikit-learn packages, has several easy-to-use functions for fitting multiple linear regression models. Via matplotlib and seaborn, there are multiple way to visualize model fit, residuals, errors, etc.

## B3. Appropriate Technique

Multiple Linear Regression is an appropriate technique to use for this research question because the task is to “understand the relationship between multiple predictor variables and a response variable” (emphasis omitted, Bobbit, 2020). Additionally, the response variable Initial\_days is continuous, satisfying the “single continuous target” requirement (Larose & Larose, 2019, p. 151)

# Part III. Data Preparation

## C1. Data Cleaning

To clean my data in preparation for the linear regression analysis, I set out to do the following tasks:

1. Verify no null or missing data
2. Verify no duplicate records
3. Check for and remove outliers

Code and explanation of methods is in the attached Jupyter notebook, “PredMod-LinReg.ipynb” – see section C1.

## C2. Summary Statistics

Summary statistics given by the built-in pandas describe() method are shown in the attached Jupyter notebook, section C2.

## C3. Visualizations

See attached Jupyter notebook, section C3.

Univariate visualizations are shown as histograms for quantitative variables, and as bar charts (Seaborn countplot()) for categorical variables. Bivariate visualizations are shown as distribution plots (Seaborn displot()) for quantitative variables and violin plots for categorical variables.

For those explanatory variables with large right-skew (Population and Income), I also show visualizations of their log transformation.

## C4. Data Transformation

I have two goals in further transforming the data prior to running the regression analysis. First, I need to transform all categorical values into numeric data. Yes/No values can easily be encoded as 1/0. Complication\_risk is ordinal and can be encoded as 1/2/3 for Low/Medium/High. Marital is nominal data, not ordinal, so needs to be transformed to a one-hot encoding. Code and references are given in the attached Jupyter notebook, section C4.

## C5. Prepared Data Set

The cleaned and transformed data set used for the regression analysis is attached as ‘clean\_xform\_medical\_data.csv’.

# Part IV. Model Comparison & Analysis

## D1. Initial Model

See attached Jupyter notebook, section D1. I created a “kitchen sink” model with the entire dataset prepared in Part III. The statsmodels.OLS summary() reports a possibility of “strong multicollinearity or other numerical problems”. Checking the variance inflation factor with code provided by Dr. Sewell does show features with very high VIFs, but there are no apparent strong pairwise correlations shown in the correlation matrix or heatmap.

## D2. Justification of Model Reduction

I have chosen to use Backward Stepwise Elimination to reduce the feature set of my initial model. At each step, I “remove[] the least significant feature (based on p-value) … which improves the performance of the model.” (Middleton, 2022) This process terminates when all remaining feature have a p-value less than some chosen alpha value. In my case, I have chosen a slightly higher-than-standard alpha value of 0.07 in order to keep at least 3 explanatory variables in my reduced model. Even so, the adjusted R-squared value of the model is quite low at 0.001. See attached Jupyter notebook, section D2.

## D3. Reduced Linear Regression Model

See attached Jupyter notebook, section D3. Three explanatory variables remain in the model – Age, Arthritis, and Asthma. These variables are not correlated with each other and satisfy the non-multicollinearity requirement of linear regression.

## E1. Model Comparison

My original “kitchen sink” model had 22 potential explanatory variables, including the 4 dummy one-hot variables for the Marital column. In the summary stats, the adjusted R-squared was 0.001, which indicates that the model explanatory variables do not do well at explaining the variance in the response variable (Bobbit, 2020). The F-test p-value, however, is 0.08, which is close to statistical significance at an alpha level of 0.05. This indicates that the model may “provide[] a better fit to the data than a model that contains no independent variables” (Frost, n.d.).

My reduced model contains only three explanatory variables. The adjusted R-squared value is unchanged at 0.001, but the F-test p-value has improved to a statistically significant 0.009.

## E2. Output and Calculations

The residual standard error (RSE) of both models is nearly identical at 26.3.

Refer to section E2 of the attached Jupyter notebook for residual plots. I have included a residual plot of Age by itself, and also on a FacetGrid with all four combinations of values of Asthma and Arthritis. The Lowess line on all plots is near 0, indicating the predictor variables are homoscedastic.

I have also provided a Q-Q plot of the reduced model. The S-shape indicates a non-normal distribution of residuals, which could invalidate the model.

## E3. Code

Refer to attached Jupyter notebook.

# Part V. Data Summary & Implications

## F1. Results

Regression equation for the reduced model:

Initial\_days = 32.88 + (0.026 \* Age) + (1.159 \* Arthritis) – (1.088 \* Asthma)

This means that the hospital can expect a patient without asthma or arthritis to stay for 32.88 days, plus an extra 0.026 days for each year in their age. A patient having arthritis would be expected to stay an extra 1.159 days, all else equal, and a patient having asthma would be expected to stay 1.088 days less, again holding all else equal.

Each of these explanatory variables is statistically significant, given my choice of alpha=0.07. Using a standard value of 0.05 for alpha, Asthma would fall outside that threshold.

Taken as a whole, the model shows statistical significance due to the F-test p-value of 0.009. However, practically, the residuals are so large that the model has very little predictive value. For a target variable with a mean value of 34.4, an RSE of 26.3 means that the expected prediction error is of the same order of magnitude as the target variable itself. This is likely to be due in part to the non-normal distribution of the residuals.

## F2. Recommendations

For future analysis, given the V-shaped distribution of the Initial\_days target variable, I would recommend splitting the dataset about the mean, and do a logistic regression to attempt to find a difference in the ‘high’ vs. ‘low’ populations. This could then be combined with a new linear regression on the two sub-samples to attempt to model stay lengths within them.

# Part VI. Demonstration & Supporting Documentation

## G. Demonstration Video

## A video describing my methods and code can be found at: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4da39f78-4581-4d1a-bca9-b0ac00ee3f27>

## H. Third-party Code Sources

Badole, M. (2023, September 15). Mastering Multiple Linear Regression: A Comprehensive Guide. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/#h-step-2-handling-categorical-variables>

Brydon, M. (2021). *Multiple Linear Regression.* Basic Analytics in Python. <https://www.sfu.ca/~mjbrydon/tutorials/BAinPy/10_multiple_regression.html>

Kassies, R. (2013, May 6). *Make more than one chart in same IPython Notebook cell.* StackOverflow. <https://stackoverflow.com/questions/16392921/make-more-than-one-chart-in-same-ipython-notebook-cell>

Sewell, W. (n.d.). *D208 Predictive Modeling Webinar Episode 1t.* Western Governors University. <https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/ER_vJMbYtxJGpxImpZ0DUQcBoVcORYKanFVKNKFcEXkRow?rtime=_ZkGUN_W2kg>

Sudheer, S. (2023, July 21). *12 Univariate Data Visualizations With Illustrations in Python.* Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/07/univariate-analysis-visualization-with-illustrations-in-python/>

Van den Broeck, M. (n.d). Lesson 3: Assessing Model Fit. *Introduction to Regression with statsmodels in Python* [MOOC]. DataCamp. <https://campus.datacamp.com/courses/introduction-to-regression-with-statsmodels-in-python/assessing-model-fit-e78fd9fe-6303-4048-8748-33b19c4222fe?ex=3>

Van den Broeck, M. (n.d). Lesson 3: Multiple Linear Regression. *Intermediate Regression with statsmodels in Python* [MOOC]. DataCamp. <https://campus.datacamp.com/courses/intermediate-regression-with-statsmodels-in-python/multiple-linear-regression-3?ex=8>

## I. References

Bobbit, Z. (October 27, 2020). *Introduction to Multiple Linear Regression.* Statology. <https://www.statology.org/multiple-linear-regression/>

Frost, J. (n.d.). *How to Interpret the F-test of Overall Significance in Regression Analysis.* Statistics by Jim. <https://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/>

Larose, C., & Larose, D. (2019). *Data Science Using Python and R.* Wiley.

Liu, C., Milton, J., & McIntosh, A. (January 6, 2016). Simple Linear Regression. *Correlation and Regression with R.* Boston University School of Public Health. <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html>

Middleton, K. (November 2022). *Getting Started with D208, Part I.* Western Governors University. <https://westerngovernorsuniversity.sharepoint.com/:b:/r/sites/DataScienceTeam/Shared%20Documents/Graduate%20Team/D208/Student%20Facing%20Resources/Dr.%20Middleton%20Getting%20Started%20with%20D208(Part%20I)COIT.pdf?csf=1&web=1&e=CLRFMI>

*pandas.DataFrame.describe — pandas 2.1.2 documentation*. (n.d.). Pandas.pydata.org. <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html>

*statsmodels.regression.linear\_model.RegressionResults – statsmodels 0.15.0 documentation.* (n.d.).Statsmodels.org. <https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.RegressionResults.html>

‌