Florida Blue Interview Project

Andrew S. Cistola, MPH

Please complete all items in the Questions to Complete section and provide a brief description of your results. You may use any tool to complete the tasks that you wish (R, Python, SAS, Excel, etc.)

### Statistical Analysis of Emergency Room Data

Question 1: Using the ER.csv file, please describe the relationship between cost and the number of ER visits.

In order to describe the relationship between the total cost by patient in the previous year (cost) the number of visits to the ER by patient in the past year (ER visits), an ordinary least squares (OLS) regression model was developed and tested. A regression model can describe the magnitude and significance of a possible relationship of the predictor variable (ER Visits) on the outcome (cost). In order for the regression model to provide information appropriate for this interpretation, a number of assumptions need to be met concerning the nature of the data.

1. It can be assumed that the outcome is dependent on the predictor and the predictor is independent of the outcome.
2. The data is randomly sampled among a population and the number of samples exceed the number of predictors.
3. The predictor and the outcome show evidence of association.
4. There is a linear relationship between the outcome and predictor.
5. The error between the linear model and the actual data follows a normal distribution with a mean of 0.
6. The error between the linear model and the actual data is not associated with each other and does not show clear patterns.
7. The error between the linear model and the actual data is randomly and evenly distributed across the data[[1]](#footnote-1)

*Assumption 1-3*

Since the high cost of care experienced in emergency settings is not likely to induce patients to select to receive emergency care and the average yearly cost occurs after the visits occur, it is reasonable assume that ER visits are independent of cost and cost is dependent on the number of ER visits. In the sample provided, there are no missing values and the number of observations (N = 500) well exceed the possible predictors variables.[[2]](#footnote-2) Spearman’s rank Correlation with an r = -0.669618362402279

Figure 1. Original OLS Model

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7548.0 871.1 8.665 <2e-16

ER\_Visits 1815.2 102.8 17.655 <2e-16

Very\_High 27965.6 545.5 51.262 <2e-16

High 19283.5 454.6 42.422 <2e-16

Moderate 14341.1 392.7 36.518 <2e-16

Low 6999.0 340.0 20.583 <2e-16

Very\_Low NA NA NA NA

Residual standard error: 2270 on 494 degrees of freedom

Multiple R-squared: 0.91,

Adjusted R-squared: 0.9091

F-statistic: 999.1 on 5 and 494 DF,

p-value: < 2.2e-16

####OLS Assumption 1: Sampling (Random sample, observations > predictors, predictor is associated with outcome)

Number of observations: 500 Number of missing values: Cost ER\_Visits Severity\_Level Very\_High High 0 0 0 0 0 Moderate Low Very\_Low 0 0 0

Spearman Correlation : -0.669618362402279

####OLS Assumption 2: Specification (Relationship between predictor and outcome is linear)

Rainbow test

data: OLS Rain = 0.87188, df1 = 250, df2 = 243, p-value = 0.859

Significant = Non-linearity

####OLS Assumption 3: Normality (Errors are normal with a mean = 0) [“\_fig/ER\_QQ\_plot.png”]

Robust Jarque Bera Test

data: resid(OLS) X-squared = 0.40872, df = 2, p-value = 0.8152

Signficiant = Non-normal

Anderson-Darling test of goodness-of-fit  
Null hypothesis: uniform distribution

data: resid(OLS) An = Inf, p-value = 1.2e-06

Signficiant = Non-normal

####OLS Assumption 4: No Autocorrelation (Error terms are not correlated with each other)

Durbin-Watson test

data: OLS DW = 1.8981, p-value = 0.1705 alternative hypothesis: true autocorrelation is greater than 0

Signficiant = Autocorrelation

####OLS Assumption 5: Homoskedasticity (Error is even across observations) [“\_fig/ER\_residuals\_plot.png”]

studentized Breusch-Pagan test

data: OLS BP = 8.2348, df = 5, p-value = 0.1438

Signficiant = Homoscedastic

Goldfeld-Quandt test

data: OLS GQ = 0.84236, df1 = 243, df2 = 243, p-value = 0.9091 alternative hypothesis: variance increases from segment 1 to 2

Significant = Heteroscedastic

####OLS Assumption 6: No Colinearity (Predictors are not correlated with each other) [“\_fig/ER\_correlation\_plot.png”] # Check for Multicollinearity

Low Correlation

Term VIF Increased SE Tolerance

ER\_Visits 3.32 1.82 0.30 Very\_High 4.62 2.15 0.22 High 3.21 1.79 0.31 Moderate 2.39 1.55 0.42 Low 1.79 1.34 0.56

###Generalized Linear Models

Log link [Y = ln(DV)] (aka: Log-Linear) with poisson error (aka: Poisson regression)

Call: glm(formula = F, family = G, data = D)

Deviance Residuals: Min 1Q Median 3Q Max  
-36.022 -7.963 -0.587 8.373 42.943

Coefficients: (1 not defined because of singularities) Estimate Std. Error z value Pr(>|z|)  
(Intercept) 9.5479001 0.0021959 4348.0 <2e-16  *ER\_Visits 0.0570751 0.0002537 225.0 <2e-16*  Very\_High 0.8875676 0.0013684 648.6 <2e-16  *High 0.6537686 0.0011673 560.1 <2e-16*  Moderate 0.5055522 0.0010320 489.9 <2e-16  *Low 0.2648308 0.0009431 280.8 <2e-16*  Very\_Low NA NA NA NA  
— Signif. codes: 0 ‘*’ 0.001 ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 885568 on 499 degrees of freedom

Residual deviance: 78440 on 494 degrees of freedom AIC: Inf

Number of Fisher Scoring iterations: 3

Log link [Y = ln(DV)] (aka: Log-Linear) with gamma error (aka: Negative binomial regression)

Call: glm.nb(formula = F, data = D, init.theta = 194.59593, link = log)

Deviance Residuals: Min 1Q Median 3Q Max  
-3.1634 -0.6554 -0.0460 0.6374 3.6692

Coefficients: (1 not defined because of singularities) Estimate Std. Error z value Pr(>|z|)  
(Intercept) 9.545116 0.027603 345.81 <2e-16  *ER\_Visits 0.057416 0.003257 17.63 <2e-16*  Very\_High 0.889583 0.017285 51.47 <2e-16  *High 0.654558 0.014404 45.44 <2e-16*  Moderate 0.505880 0.012446 40.65 <2e-16  *Low 0.265213 0.010780 24.60 <2e-16*  Very\_Low NA NA NA NA  
— Signif. codes: 0 ‘*’ 0.001 ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(194.5959) family taken to be 1)

Null deviance: 5468.27 on 499 degrees of freedom

Residual deviance: 500.62 on 494 degrees of freedom AIC: 9154

Number of Fisher Scoring iterations: 1

Theta: 194.6   
 Std. Err.: 12.4

2 x log-likelihood: -9139.965

###Hierarchical Linear Models

Hierarchical Linear Model with Random Intercepts

Linear mixed model fit by REML. t-tests use Satterthwaite’s method [ lmerModLmerTest] Formula: F Data: D

REML criterion at convergence: 9153.2

Scaled residuals: Min 1Q Median 3Q Max -3.1809 -0.6733 0.0126 0.6507 3.2434

Random effects: Groups Name Variance Std.Dev. Severity\_Level (Intercept) 116661601 10801  
Residual 5152978 2270  
Number of obs: 500, groups: Severity\_Level, 5

Fixed effects: Estimate Std. Error df t value Pr(>|t|)  
(Intercept) 21305.505 4871.102 4.124 4.374 0.0111 \*  
ER\_Visits 1808.623 102.764 494.942 17.600 <2e-16 \*\*\* — Signif. codes: 0 ‘*’ 0.001 ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects: (Intr) ER\_Visits -0.127 One Way ANOVA for MLE ANOVA-like table for random-effects: Single term deletions

Model: Cost ~ ER\_Visits + (1 | Severity\_Level) npar logLik AIC LRT Df Pr(>Chisq)  
 4 -4576.6 9161.2  
(1 | Severity\_Level) 3 -5028.3 10062.6 903.39 1 < 2.2e-16 \*\*\* — Signif. codes: 0 ‘*’ 0.001 ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Intraclass Correlation Coefficient

Adjusted ICC: 0.958

Conditional ICC: 0.881

Is model singular? (overly-complex): FALSE

###Model Performance and Comparison

[“\_fig/ ER \_performance\_plot.png”] # Comparison of Model Performance Indices

## Name | Model | AIC | BIC | RMSE | Sigma | Nagelkerke’s R2 | Score\_log | Score\_spherical | R2 | R2 (adj.) | R2 (cond.) | R2 (marg.) | ICC

OLS | lm | 9154.442 | 9183.944 | 2256.349 | 2270.010 | | | | 0.910 | 0.909 | | |  
POI | glm | Inf | Inf | 2238.316 | 12.601 | 1.000 | -Inf | 0.002 | | | | |  
NB | negbin | 9153.965 | 9183.467 | 2238.565 | 1.007 | 1.000 | -Inf | 0.003 | | | | |  
RAND | lmerModLmerTest | 9161.201 | 9178.060 | 2256.362 | 2270.017 | | | | | | 0.961 | 0.080 | 0.958

Question 2: Using the healthDat.csv and fl\_zips.csv files, please bring in the city and county that each patient lives in and show how you did it. Additionally, if you do not have access to a SQsupported tool, please write SQL pseudo-code that will accomplish this as well.

Question 3: Using the healthDat.csv and fl\_zips.csv files, please provide 2-3 observations or recommendations as they relate to costs. Keep in mind that the goal is to reduce overall costs for this population.

Question 4: Using the healthDat.csv file, please list the 3 most prevalent chronic conditions as well as describe any relationships between conditions that you may have found.

Question 5: Please produce your best dashboard using the healthDat.csv and/or the fl\_zips.csv files.Hint: cost is important!

Updated : 1628985418.33307 by [andrewcistola@pm.me](mailto:andrewcistola@pm.me)

Disclaimer: Please note that all the health data is entirely fictional and made up by the Analytics team.

© Andrew S. Cistola, MPH

1. These statements reflect in general assumptions for OLS regression in order to obtain the best linear unbiased estimators (BLUE) but are not directly quoting the original theorem or derivative quotations or explanations of OLS assumptions. [↑](#footnote-ref-1)
2. Since this is an exercise, the sample is assumed to be random or reasonably reflective of the population. However, in a true scenario, the sampling strategy would need to be evaluated for systematic error. [↑](#footnote-ref-2)