STAT 452 – Statistical Modeling I

Case Study 3

Paolo Furlanetto Ferrari

05/08/2023

Summary

I. Introduction

Nosocomial infections are infections acquired during hospitalization in a health-care facility.[1] Various types of bacteria, viruses and fungi can cause nosocomial infections. These can be transmitted through contaminated equipment, personal contact with other patients or health-care professionals, or even through the air ventilation. Negative consequences of nosocomial infections are various: increased patient morbidity, increased risk of death, increased length-of-stay and increased cost of treatment. Thus, it is important to make hospitals and clinics safe of such infections.

In an effort to identify the impacts of control measures in the rate of nosocomial infections on hospitals in the US, the Study on the Efficacy of Nosocomial Infection Control (SENIC) Project conducted data analysis of patient stays across 338 hospitals during the years 1975-1976. One of the specific goals of the SENIC was to identify if the length-of-stay of patients in hospitals depended on a number of different variables.

In this Case Study, we analyze the dependence of the length-of-stay versus 3 factor variables: (i) the geographical region of the hospital, (ii) a factor representing the age of the patient and (iii) a factor representing the availability of facilities and services offered by the hospital. The analyzed data is a sample of the original SENIC data containing 113 records.

For the purpose of the analysis, we split this Case Study into three parts:

- In Part A, we will look whether the geographical region of a hospital has influence on the average length-of-stay.
- In Part B, we will look to see if the region and the age of a patient has an influence on the average length-of-stay, including if the combination between these two variables has a significant effect.
- In Part C, we will look whether the region, age and availability of services and facilities have an influence on the average length-of-stay, including their combined actions.

To achieve these goals, we fit various ANOVA models of the different factor variables and test the statistical significance of each. Next, we apply diagnostic and remedial measures to check the validity of the ANOVA assumptions. Finally, we report the results of the models encountered for each of the cases A, B and C.

II. Part A

II.a) Data overview and ANOVA models

In this first part, we test whether the length-of-stay depends on the geographical region. Figure 1 shows a box-plot of these two variables, with the Region variable taking 4 possible values.

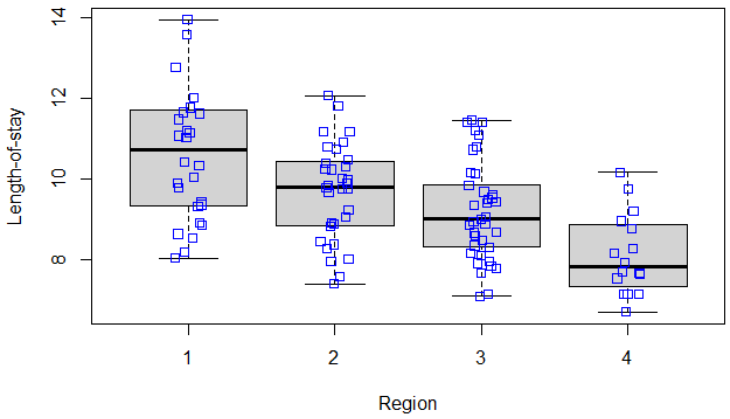

Figure 1: Box-plot of Length-of-stay vs Region. Blue squares are randomly sampled points.

Figure 1 indicates that each Region has a different mean Length-of-stay.

A full ANOVA model is run on the dataset. Using a confidence level of 95%, the results indicated that all 4 Regions have statistically significant different means from the base level (Region 1). Next, we diagnosed the results by inspecting the quantile-quantile plots shown in Figure 2. The results indicated: (i) the presence of two potential outliers points and (ii) the residuals do not have equal variance.

To remedy these two observations, we checked for the presence of high leverage points, high influential points and outliers. The results identified the two data points seen in the QQ-plot Figure 2 as outliers, using an outlier test with significance of 95%. Thus, we excluded these points from subsequent analysis.

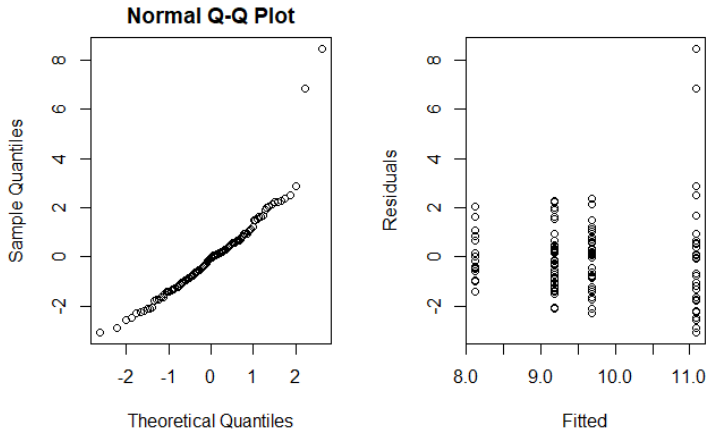

Figure 2: Quantile-quantile plot and residuals of the fitted ANOVA model.

Next, to fix the unequal variance of the residuals, we performed a Box-Cox transformation of the data using an exponent λ of -1, identified by a minimum 95% likelihood criterion. Then, we checked the good validity of the new model by inspecting the residuals (using the QQ and residuals plots) as well as its overall p-value (2e-7). Finally, we performed Tukey and Bonferroni tests to check the significance between the differences of the mean lengths-of-stay for each Region. The Tukey intervals are graphically shown in Figure 3.

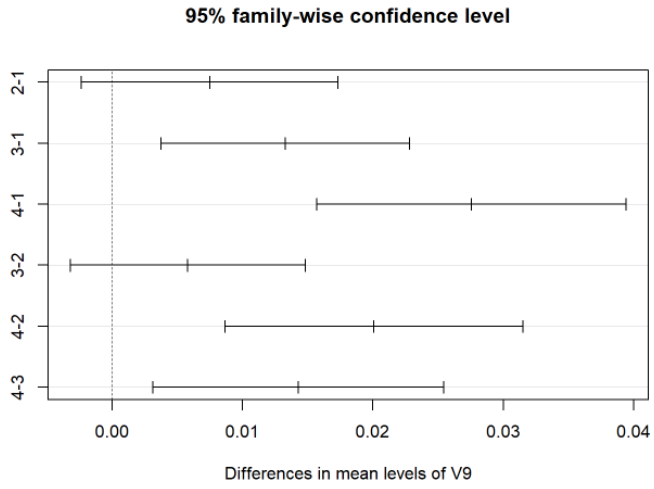

Figure 3: Confidence interval for the differences between the transformed length-of-stay versus Region (V9) for final model of Part A.

II.c) Conclusions

Based on the results of the Tukey and Bonferroni tests on the final model of Part A, we can say with 95% confidence that:

- Region 4 (W) has different mean length-of-stay than Regions 1, 2 and 3 (NE, NC and S).
- Region 1 (NE) has a different mean length-of-stay than Region 3 (S).

Moreover, we cannot affirm that Regions 1 and 2 have statistically significant different lengths-of-stay. The similar holds for Regions 2 vs Region 3.

III. Part B

III.a) Data overview and ANOVA models

In Part B, we test whether the age of the patients is statistically significant factor in the length-of-stay for a provided sub-set of the data. To see the effect of Age, we partitioned the patients into two groups: (i) Patients aged 54 years or more and (ii) patients aged 53.9 years or less. Figure 4 show the mean length-of-stay for each of these groups and for each Region of the study.

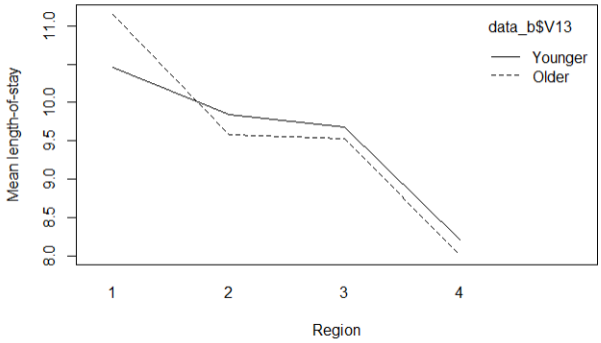

Figure 4 Interaction plot of the effect of Region and Age on the mean Length-of-stay. In the legend, "Younger" and "Older" refers to the age of 54 years.

The first model attempted was an ANOVA model that accounted for both Region and Age factors as well as their interaction. However, it was found that the Age factor was not statistically significant. In fact, the p-value of the model containing Age only was 0.87. Thus, we excluded this variable in the subsequent analysis of this part.

Since the data for Part B is a subset of the data for Part A, the results of the ANOVA regression models could be different. After fitting a simple model and checking for normality assumptions, we had to perform the same remedial steps for Part A. The data was again transformed using a Box-Cox transformation with λ of -1. The final step was checking for significance of the differences using Tukey and Bonferroni tests. The Tukey intervals are graphically shown in Figure 5.

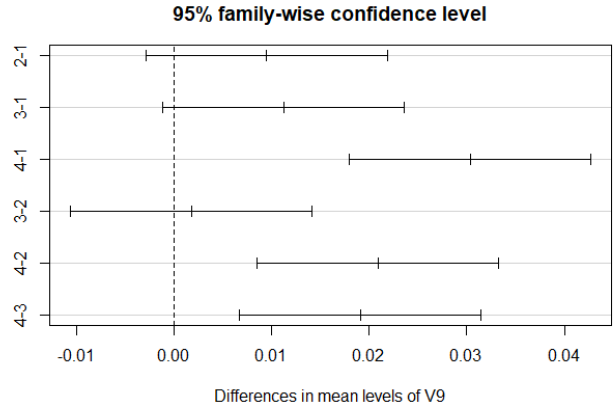

Figure 5: Confidence interval for the differences between the transformed length-of-stay versus different Regions (V9) for final model of Part B.

III.b) Conclusions

Based on the results of the Tukey tests on the final model of Part B, we can say with 95% confidence that:

- Patients older or younger than 54 years do not have different mean lengths-of-stay.
- Region 4 (N) has different mean length-of-stay than Regions 1, 2 and 3 (NE, NC and S).
- Regions 1, 2 and 3 (NE, NC and S) do not have different lengths-of-stay.

Note that the last conclusion is slightly different than Part A because the datasets are different. Also note that the conclusion from the Bonferroni test is slightly different than this one [2], but because one of the p-value is too close to the critical value of 0.05, we adopted the results from the Tukey test.

IV. Part C

IV.a) Data overview and ANOVA models

In this final part, we evaluated the impact of Region, Age and Availability factors on the mean lengths-of-stay. This time, the patients are grouped by older or younger than 53 years (Age Factor). Also, the hospitals are grouped if their potential facilities and services are larger or smaller than 40.2% of the total 35 types specified by the SENIC project (Availability Factor). Figure 6 shows the interaction plots for the Region Factor with the other two factor variables just described:

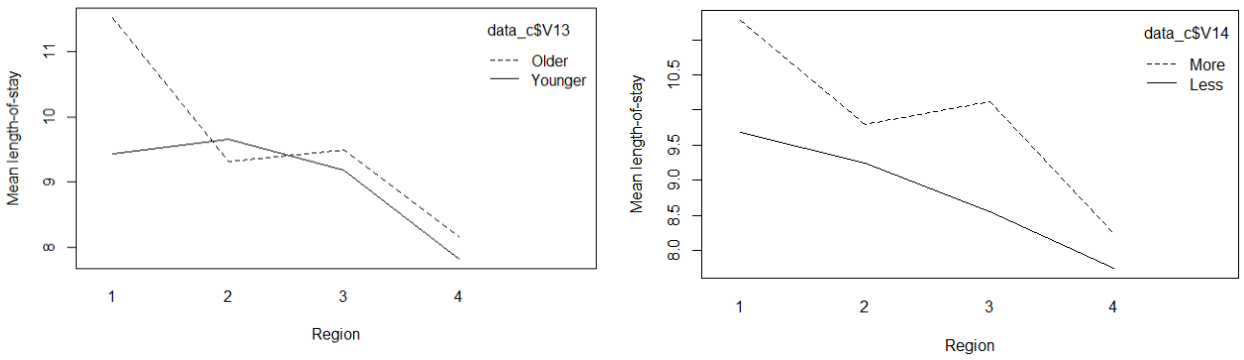

Figure 6: Interaction plots of the effect of Region and Age (left) and Region and Availability (right) on the mean Length-of-stay. In the left plot, "Younger" and "Older" refers to 53 years. In the right plot, "More" and "Less" refers to 40.2%.

Again, the first model attempted was an ANOVA model that accounted for both Region, Age and Availability factors as well as their 3 interaction terms. However, it was found that the Age factor was not statistically significant similar to Part B (even though both used different data sets and different thresholds for the classification). Thus, we excluded this variable in the subsequent analysis of this part.

Next, we fitted an ANOVA model with Region and Availability and their interaction term. It was found that the interaction was insignificant (p-value = 0.44). Thus, we fitted a model accounting only for additive effects between Age and Avalability factors.

Not surprisingly, after checking for normality assumptions, we had to perform the same remedial steps for Parts A and B. The data was again transformed using a Box-Cox transformation with λ of -1. The final step was checking for significance of the differences using Tukey and Bonferroni tests. The Tukey intervals are graphically shown in Figure 7.

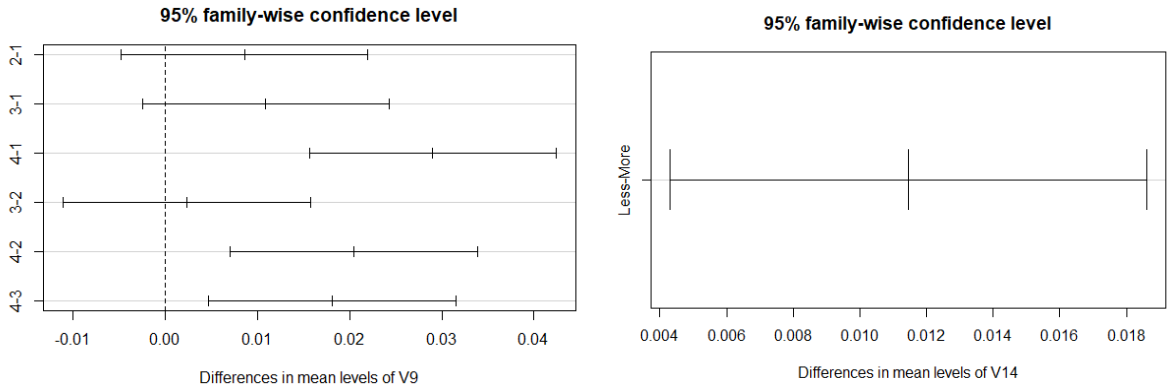

Figure 7: Confidence interval for the differences between the transformed length-of-stay versus different Regions (V9, left) and versus different Availabilities (V14, right) for final model of Part C.

IV.b) Conclusions

Based on the results of the Tukey and Bonferroni tests on the final model of Part B, we can say with 95% confidence that:

- Patients older or younger than 53 years do not have different mean lengths-of-stay.
- Hospitals with availabilities larger than 40.2% have larger mean lengths-of-stay than hospitals with less than 40.2%.
- Region 4 has different mean length-of-stay than Regions 1, 2 and 3 (NE, NC and S).
- Regions 1, 2 and 3 (NE, NC and S) do not have different mean lengths-of-stay.

V. Conclusion

In this study, we analyzed the length-of-stay for patients treated across 118 hospitals in the years 1975 to 1976. We compared the length-of-stay for patients hospitalized in different Regions, with different Ages and with different availabilities of services and facilities. In general terms, it was found that: (i) The North Region have different length-of-stay compared to South, Northwest and North-central Regions; (ii) that patients younger or older than 53 or 54 years have the same length of stay on average; and that (iii) the average length of stay is larger for hospitals with available facilities which are 40.2% or more of the total 35 potential facilities/services of US hospitals.

References

[1] Mehta Y, Gupta A, Todi S, et al. Guidelines for prevention of hospital acquired infections. *Indian J Crit Care Med*. 2014;18(3):149-163.

[2] Attached R Markdown and HTML files.

[3] Stat 452 Lecture Notes (Spring 2023).