Regression Analysis of Infection Risk

Professor Xinyue Liao, WINTER 2018

Glynnis Foley, B.H. Friday, 9 a.m.
Katherine Goren, M.E. Friday, 12 p.m.
Daniel Mindlin, Z.Y. Friday, 10 a.m.
Dennis Shen, Z.Y. Friday, 10 a.m.

I. INTRODUCTION

The dataset, Infection Risk, contains information from the Study on the Efficacy of Nosocomial Infection Control conducted in the United States, between 1975-1976. This data regards 113 hospitals and 12 variables: ID number, length of stay, age, infection risk, routine culturing ratio, routine chest x-ray ratio, number of beds, medical school affiliation, geographic region, average daily census, number of nurses, and availability facilities and services. The ultimate goal of the study was to determine whether infection surveillance and control programs have reduced the rates of hospital acquired infection. In the report that follows, we discuss the regression analysis we use in order to answer the following research questions:

1. Can we utilize the dataset in order to find predictors for infection control?
2. What is the best model for predicting infection control?
3. Do we need to modify the data in any way in order to help find this best model?

II. QUESTIONS OF INTEREST

1. Can we identify particular predictor-response relationships between any of the 12 variables and infection risk?
2. Do any of the predictors we found in question 1 display non-linear trends and consequently require transformations to analyze the data?
3. Are there any outliers or highly influential points that could be impacting our response in our dataset?
4. Are there any interaction terms?

III. REGRESSION MODEL USED PER QUESTION OF INTEREST

1. To begin, we used stepwise regression in order to select which variables have the lowest AIC values; AIC values display quality of each model, relative to each of the other models where the smaller the value, the higher the quality of the model. AIC is effective in predicting model quality as it utilizes information about the sum of squares error, the number of parameters, and the sample size of the data set. As mentioned the stepwise function provides us with the five predictor variables which qualify as adequate for infection risk; its goal is to find the middle ground between an underspecified model and a model with extraneous variables through the use of multiple F-tests using the variables in the dataset.
2. After selecting the appropriate model using the method prescribed above, we used the box-cox function in order to determine whether or not we needed to transform the data by plotting log-likelihood of the model versus λ. The function shows which transformations should be used on the inputted data set; e.g. a plot centered at $\lambda=0$ indicates that a logarithmic transformation should be used.

3. To test for influential points in our model, we checked hat values for high leverage points. High leverage points have abnormal predictor values, e.g. too extreme x-values. In addition, we checked studentized values, calculated via dividing the residuals by approximating their standard deviation to check for outliers in the dataset. An outlier is a data point whose response does not follow the general trend of the entirety of the data, aka a y-value that shows to be a trend discrepancy. In order to help pinpoint highly influential points, we used both Cook's distance and DFFIT values. Influential points impact any part of regression analysis, namely the predicted responses, the estimated slope coefficients, or the hypothesis test results. DFFIT values quantify the number of standard deviations that the fitted value changes when the i-th data point is omitted. In regards to the interpretation of DFFIT values, if the value is high, it indicates an influential data point. Similarly to DFFIT values, Cook's distance summarizes how much all of the fitted values change when the i-th observation is deleted. A data point having a large Cook's distance indicates that the data point strongly influences the fitted values. Finally, we check the partial residuals vs. fit plot to confirm equal variances which implies a constant trend.
4. In order to complete our analysis, we checked for interaction terms in our model. To do so, we utilized summaries of anova tables for each possible interaction term. We checked each summary for each possible interaction for statistical significance.

IV. RESULTS AND INTERPRETATION

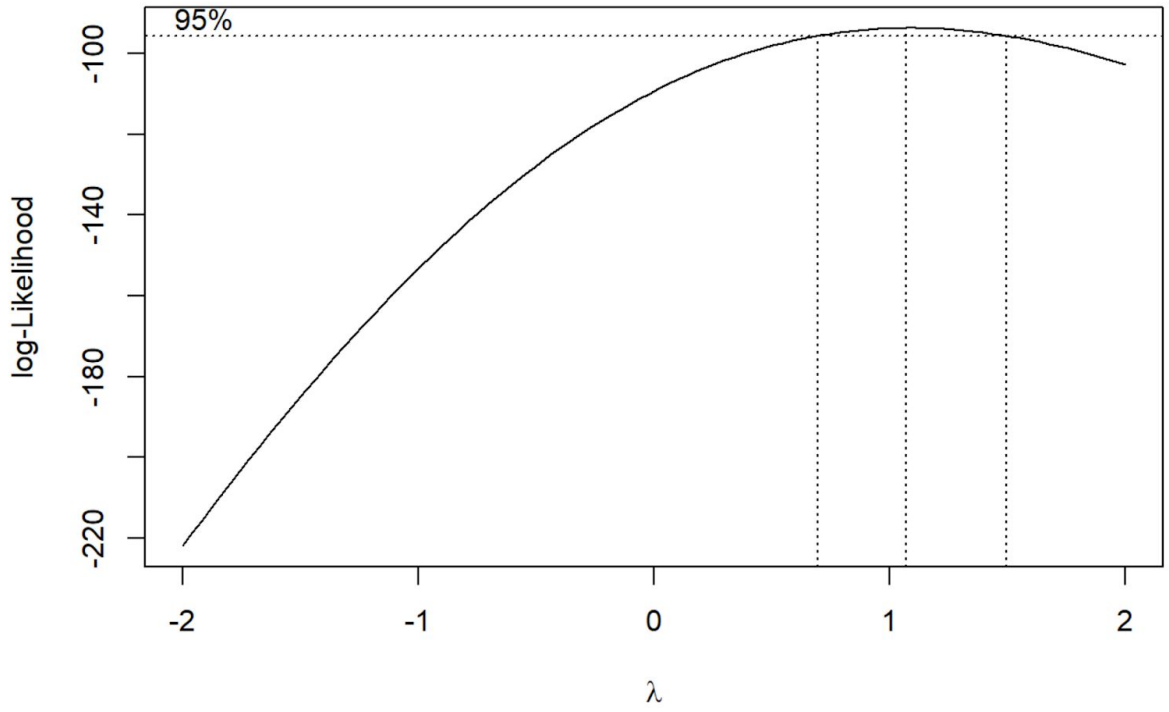
1. The following variables qualify as good predictors for infection risk: routine culturing ratio, length of stay, geographic region, availability of facilities and services, and routine chest x-ray ratio. These 5 predictors account for 55.9% of the variability in the response. According to diagnostics, the QQ-normal, Residual vs Fit, and Residual vs Order indicate the correlation to have LINE distribution (See "r residual check for question 1").

```
summary(mod1)
```

```
##
## Call:
## lm(formula = InfctRsk ~ Culture + Stay + Facilities + Region +
##     Xray)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89739 -0.60475  0.05352  0.64507  2.40506
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.089797   0.678339  -1.607  0.11115
## Culture      0.049345   0.009737   5.068 1.74e-06 ***
## Stay         0.258449   0.057992   4.457 2.09e-05 ***
## Facilities   0.020056   0.006203   3.233  0.00164 **
## Region2      0.279098   0.251996   1.108  0.27059
## Region3      0.310702   0.259516   1.197  0.23391
## Region4      1.028169   0.331121   3.105  0.00245 **
## Xray         0.012003   0.005245   2.288  0.02412 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9197 on 105 degrees of freedom
## Multiple R-squared:  0.559,  Adjusted R-squared:  0.5296
## F-statistic: 19.01 on 7 and 105 DF,  p-value: 3.41e-16
```

2. The boxcox function used displayed a plot with a curve centered at λ=1, indicating that no transformation need be used. According to diagnostics, the QQ-normal, Residual vs Fit, and Residual vs Order indicate the correlation to have LINE distribution (See “r residual check for question 2”)

```
boxcox(mod1)
```



3. According to the model described in part 3- answer 3, we removed the three points that we determined to most likely to be negatively impacting our model which were 8,47 and 53. With our updated data set, we developed a new model in which our r-squared value was improved by approximately 2%, but our MSE increased substantially. We therefore decided to continue further with our original model stated in part 4 – answer 1. According to diagnostics, the QQ-normal, Residual vs Fit, and Residual vs Order indicate the correlation to have LINE distribution (See "r residual check for question 3")

```
hv=hatvalues(mod1)
hv.max=which(hv==max(hv))
hv.max
```

```
## 47
## 47
```

```
rs=rstudent(mod1)
rs.max=which(rs==max(rs))
rs.max
```

```
## 53
## 53
```

```
dffit=dffits(mod1)
dffit.max=which(dffit==max(dffit))
dffit.max
```

```
## 53
## 53
```

```
which(ck.dist > .2)
```

```
## 8
## 8
```

```
summary(modn)
```

```
##
## Call:
## lm(formula = InfctRskn ~ Culturen + Stayn + +Regionn + Facilitiesn +
##     Xrayn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94417 -0.49945  0.07033  0.54748  1.98527
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.107509   0.706305  -1.568 0.119971
## Culturen     0.060736   0.010537   5.764 8.80e-08 ***
## Stayn        0.271896   0.064416   4.221 5.29e-05 ***
## Regionn2     0.381869   0.244758   1.560 0.121812
## Regionn3     0.277501   0.251413   1.104 0.272291
## Regionn4     1.085464   0.319424   3.398 0.000969 ***
## Facilitiesn  0.019204   0.005949   3.228 0.001676 **
## Xrayn        0.008633   0.005074   1.702 0.091889 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8723 on 102 degrees of freedom
## Multiple R-squared:   0.58,  Adjusted R-squared:  0.5512
## F-statistic: 20.12 on 7 and 102 DF,  p-value: < 2.2e-16
```

4. From the method explained in part 3-answer 4, we determined that routine culturing ratio and availability of facilities and services were interaction terms. We consequently came up with a new model utilizing these terms and from the ANOVA table conducted on the new model, it can be seen that R^2 increased by 3.7% and the MSE increased by 3.6 %. Therefore, the updated model is our best model at predicting infection risk. From these results, we can infer that monitoring bacterial irregularities may guide facilities to better allocate their resources to prevent infection in patients. According to diagnostics, the QQ-normal, Residual vs Fit, and Residual vs Order indicate the correlation to have LINE distribution (See “r residual check for question 4”)

```
a3<-aov(InfctRsk~Culture*Facilities, data=infectionrisk)
summary(a3)
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## Culture              1  62.96   62.96  62.080 2.67e-12 ***
## Facilities           1  19.92   19.92  19.642 2.23e-05 ***
## Culture:Facilities   1   7.95    7.95   7.834  0.00606 **
## Residuals          109 110.55    1.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V. CONCLUSION

We set out on this project with essentially nothing more than raw data. We were given a dataset: hospital figures regarding the virulent infections which rampaged down their halls. Our task was to use these numbers to shed light on the ways that these hospitals tried to tackle this issue, to show the hospitals how effective their attempted solutions to infection spread are, and most importantly, to construct a linear regression which took this raw lump of data and transfigured it into a sophisticated, informative model. We elected to use four questions of interest to address this:

Firstly, whether any particular predictor-response relationships between any of the 12 variables and infection risk existed. To answer this, we employed a number of regression techniques (outlined above) to find the variables which most affected variability in the response.

Next, whether these predictors displayed non-linear trends which necessitated transformation. Put simply, we found that they did not. Next, if we encountered any outliers or highly influential points that impacted our findings, and whether any of the terms we used interacted in meaningful ways. We used Cook's distance and DFFIT values as well as high hat values to find influential points as well. To complete our analysis, we checked for interaction terms via an interaction plot and ANOVA table and determined that routine culturing ratio: availability of facilities and services as an interaction term improved our model. The R^2 value for our model including the interaction term is larger than the model without the interaction term, and even though the MSE is larger with the interaction, we end up choosing our model including the interaction term because we value the additional accuracy and the increase in variance is not significant.

We conclude that routine culturing ratio, length of stay, geographic region, availability of facilities and services, and routine chest x-ray ratio, namely with routine culturing ratio and availability of facilities and services as interaction terms, are the aspects of medical procedure that hospitals ought to focus on when attempted to improve their performance in preventing disease in their patients.

VI. APPENDIX[CODE]

```{r}
library(alr3)
library(leaps)
library(MASS)
#data('infectionrisk.txt')
infectionrisk=read.table("infectionrisk.txt", header=T)

ID=infectionrisk$ID
Stay=infectionrisk$Stay
Age=infectionrisk$Age
InfctRsk=infectionrisk$InfctRsk
Culture=infectionrisk$Culture
Xray=infectionrisk$Xray
Beds=infectionrisk$Beds
MedSchool=infectionrisk$MedSchool
Region=infectionrisk$Region
Region=factor(Region)
Census=infectionrisk$Census
Nurses=infectionrisk$Nurses
Facilities=infectionrisk$Facilities
pairs(infectionrisk)
n=dim(infectionrisk)[1]
mod0=lm(InfctRsk~1)
mod.all=lm(InfctRsk~.,data=infectionrisk)
step(mod0, scope = list(lower = mod0, upper = mod.all), k = log(n))
mod1=lm(InfctRsk~Culture + Stay + Facilities+Region+ Xray)
bs=coef(mod1)
summary(mod1)
boxcox(mod1)
#boxcox output shows a center around 1 ==> no transformation should be used
p=12

m.s=lm(InfctRsk~1,data=infectionrisk)
add1(m.s, ~.+ID+Stay+Age+Culture+Xray+Beds+MedSchool+Region+ Census+
Nurses+Facilities, data = infectionrisk, test = 'F')
```

```
m.s.1=lm(InfctRsk~Stay, data=infectionrisk)
add1(m.s.1, ~.+ID+Age+Culture+Xray+Beds+MedSchool+Region+ Census+ Nurses+Facilities, data = infectionrisk, test = 'F')
m.s.2=lm(InfctRsk~Stay+Culture, data=infectionrisk)
add1(m.s.2, ~.+ID+Age+Xray+Beds+MedSchool+Region+ Census+ Nurses+Facilities, data = infectionrisk, test = 'F')
m.s.3=lm(InfctRsk~Stay+Culture+Facilities, data=infectionrisk)
add1(m.s.3, ~.+ID+Age+Xray+Beds+MedSchool+Region+ Census+ Nurses, data = infectionrisk, test = 'F')
m.s.4=lm(InfctRsk~Stay+Culture+Facilities+Region, data=infectionrisk)
add1(m.s.4, ~.+ID+Age+Xray+Beds+MedSchool+ Census+ Nurses, data = infectionrisk, test = 'F')
m.s.5=lm(InfctRsk~Stay+Culture+Facilities+Region+Xray, data=infectionrisk)
add1(m.s.5, ~.+ID+Age+Beds+MedSchool+ Census+ Nurses, data = infectionrisk, test = 'F')

#mod1=m.s.5
#we start with this model
```
```{r outliers check}
hv=hatvalues(mod1)
hv.max=which(hv==max(hv))
hv.max
plot(hv, ylab = 'Hat Values', main = 'Hat Values')
text(1:n, hv, cex= .7, pos=4)
avg.hat = p/n
#cutoff: 2p/n
abline(h=2*avg.hat, col=4)
#cutoff: 3p/n
abline(h=3*avg.hat, col=2)
text(2, y=2*avg.hat, expression(2 %*% p/n), pos=3)
text(2, y=3*avg.hat, expression(3 %*% p/n), pos=3)
#The case with the highest leverage is 47
rs=rstudent(mod1)
rs.max=which(rs==max(rs))
rs.max
#The case with the largest externally studentized value is 53
dffit=dffits(mod1)
dffit.max=which(dffit==max(dffit))
dffit.max
```

```
#The case with the the largest difference of fit value is 53
ck.dist=cooks.distance(mod1)
which(ck.dist >1)
which(ck.dist > .2)
#Using Cook's distance we can determine that neither of these are influential and there's likely no reason to remove them from the data set
inf.index(mod1)
par.res=resid(mod1, type='partial')
plot(Stay, par.res[,1])
which(Stay>16)
plot(Culture, par.res[,2])
which(Culture >58)
#plot(Facilities, par.res[,3])
#plot(Region, par.res[,4])
#plot(Xray, par.res[,5])
#plots of residuals of Facilities, Region, and Xray are random
```

```{r outlier removal}
#using the possible influential points that were found using hat values, externally studentized residuals, difference of fits, and residual plot checks we remove three data points to determine whether or not they are negatively impacting our model
new.ir=infectionrisk[-c(8,47,53),]
dim(new.ir)
IDn=new.ir$ID
Stayn=new.ir$Stay
Agen=new.ir$Age
InfctRskn=new.ir$InfctRsk
Culturen=new.ir$Culture
Xrayn=new.ir$Xray
Bedsn=new.ir$Beds
MedSchooln=new.ir$MedSchool
Regionn=new.ir$Region
Regionn=factor(Regionn)
Censusn=new.ir$Census
Nursesn=new.ir$Nurses
Facilitiesn=new.ir$Facilities
n
n=dim(new.ir)[1]
n
```

```
modn0=lm(InfctRsk~1)
modnall=lm(InfctRsk~.,data=new.ir)
step(modn0, scope = list(lower = modn0, upper = modnall), k = log(n))
modn=lm(InfctRskn ~ Culturen + Stayn + +Regionn + Facilitiesn + Xrayn)
summary(mod1)
summary(modn)

#We removed the three data points that were most likely negatively impacting our model to develop modn
#modn vs. mod1 (original model)
#modn has a larger r^2 however it also has a larger mse so mod1 is preferred to mod1
```

```{r}
boxcox(mod1)
#boxcox output shows a center around 1 ==> no transformation should be used
#no need to logarithmically transform the data
```

```{r interaction term check}
a1<-aov(InfctRsk~Culture*Stay, data=infectionrisk)
summary(a1)
a2<-aov(InfctRsk~Culture*Region, data=infectionrisk)
summary(a2)
a3<-aov(InfctRsk~Culture*Facilities, data=infectionrisk)
summary(a3)
a4<-aov(InfctRsk~Culture*Xray, data=infectionrisk)
summary(a4)
a5<-aov(InfctRsk~Stay*Region, data=infectionrisk)
summary(a5)
a6<-aov(InfctRsk~Stay*Facilities, data=infectionrisk)
summary(a6)
a7<-aov(InfctRsk~Stay*Xray, data=infectionrisk)
summary(a7)
a8<-aov(InfctRsk~Region*Facilities, data=infectionrisk)
summary(a8)
a9<-aov(InfctRsk~Region*Xray, data=infectionrisk)
summary(a9)
a0<-aov(InfctRsk~Facilities*Xray, data=infectionrisk)
summary(a0)
```

#Culture and Facilities look like the only two predictors that could be interacting
mod.i=lm(InfctRsk~Culture+Stay+Region+Facilities+Xray+Culture:Facilities)
summary(mod1)
summary(mod.i)
#The R^2 value for our model including the interaction term is 3.7% larger than the model without the interaction term. Are MSE is larger with the interaction by 3.6% so. We end up choosing our model including the interaction term because we value the additional accuracy and the increase in variance is not significant.
```

#Diagnostics
```{r residual check for question 1}

resid1= resid(mod1)
qqnorm(resid1)
qqline(resid1)
#Q-Q plot is normal
lag.plot(resid1, lags = 1)
fr=fitted(mod1)

plot(fr,resid1,xlab = "Fitted value",ylab = "Residual", main = 'Residual Vs Fit')
abline(h=0,lty=2)

plot(ID,resid1,xlab = "Order",ylab = "Residual", main = 'Residual Vs Order')
abline(h=0,lty=2)

```
```{r residual check for question 3}

residn= resid(modn)
qqnorm(residn)
qqline(residn)
#Q-Q plot is normal
lag.plot(residn, lags = 1)
fr=fitted(modn)

plot(fr,residn,xlab = "Fitted value",ylab = "Residual", main = 'Residual Vs Fit')
abline(h=0,lty=2)

```
plot(IDn,residn,xlab = "Order",ylab = "Residual", main = 'Residual Vs Order')
abline(h=0,lty=2)
```

```{r residual check for question 4}

resid= resid(mod.i)
qqnorm(resid)
qqline(resid)

lag.plot(resid, lags = 1)
fr=fitted(mod.i)

plot(fr,resid,xlab = "Fitted value",ylab = "Residual", main = 'Residual Vs Fit')
abline(h=0,lty=2)

plot(ID,resid,xlab = "Order",ylab = "Residual", main = 'Residual Vs Order')
abline(h=0,lty=2)


```