Hospital Infection Rates

Cody Frisby

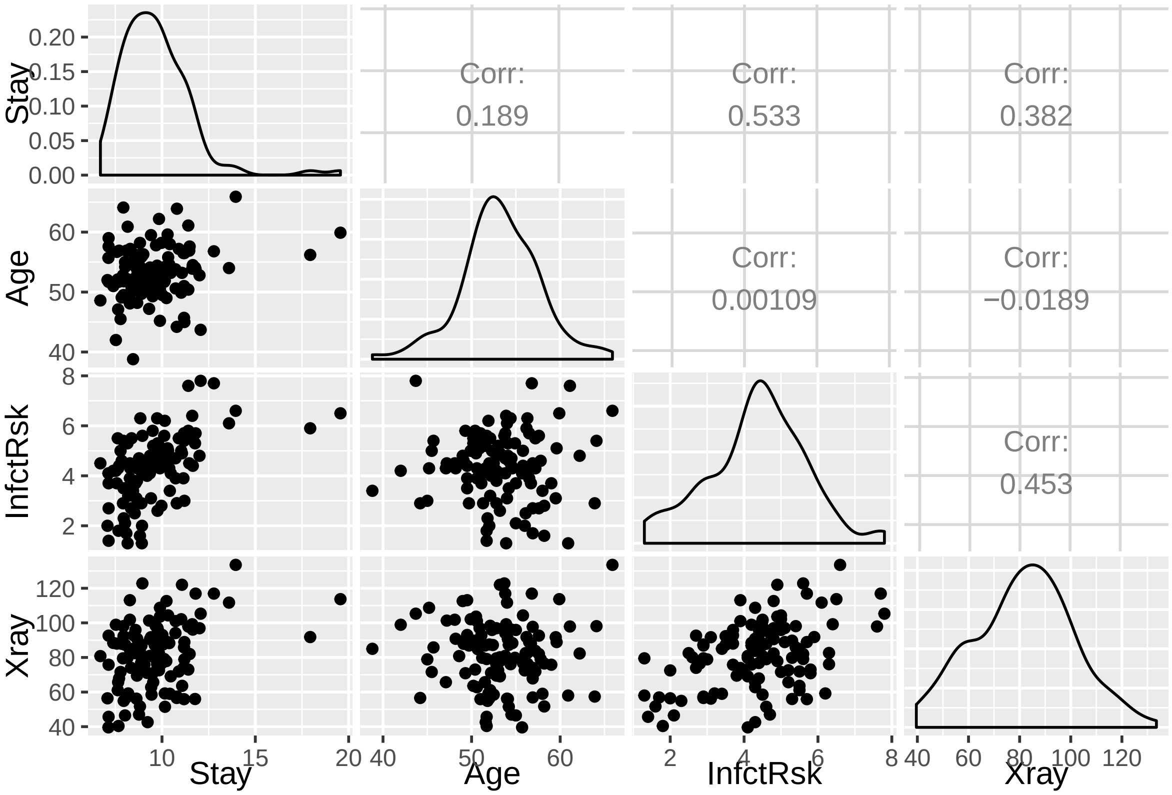
February 26, 2016

1) Problem Statement

Infection developed during one's stay in a hospital is a problem. The hospital should be a place where one gets well, not worse. Using data from hospitals we may be able to identify what the main factors are that contribute to infection. Since there is bound to be variation, developing a statistical model that has little bias may allow us to prevent infections from happening while someone is admitted to the hospital by identifying the most important factors. Also, we may want to know how well we can predict infection rate.

2) Exploratory Analysis

We would like to assess whether or not a multiple linear regression model is appropriate using the given data. First I'd like to look at a scatter plot matrix and the correlation between all the variables.



It appears there is some positive correlation between infection risk (InfctRsk) and Stay & Xray. Meaning, as a person's stay increases their infection risk increases. Additionally, the more xrays the higher the infection rate. There is very little correlation between infection risk and age. We may want to leave age out of the model since it may not have any effect on infection rates. The only concern for MLR model is the correlation between Xray and Stay at 0.382. Although, this may not be a problem since it is not extremely high. MLR may be an appropriate model decision in understanding significant factors for increased infection rate.

3) MLR model

Here we write out our model.

$$Infection\ Rate = \beta_0 + \beta_1 Stay + \beta_2 Age + \beta_3 Xray + \varepsilon_i$$

Where we assume ε are independent identically distributed and $\sim N(0, \sigma^2)$. Also, β_1 is interpreted as, holding all the other variables constant, for every unit change in Stay infection risk will go up by this factor. We would want to fit a model using these data and assumptions to know just was this factor, or β_1, is equal to.

4) Fit the MLR model

Here we fit the above model. Below are the estimated β_is from above.

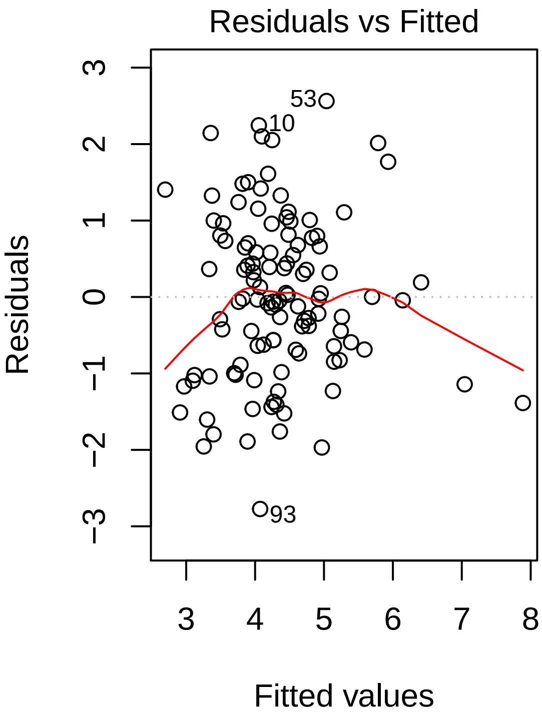
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0011617	1.3147238	0.7614996	0.4480031
Stay	0.3081809	0.0593956	5.1886107	0.0000010
Age	-0.0230052	0.0235158	-0.9782886	0.3300979
Xray	0.0196609	0.0057586	3.4142112	0.0008992

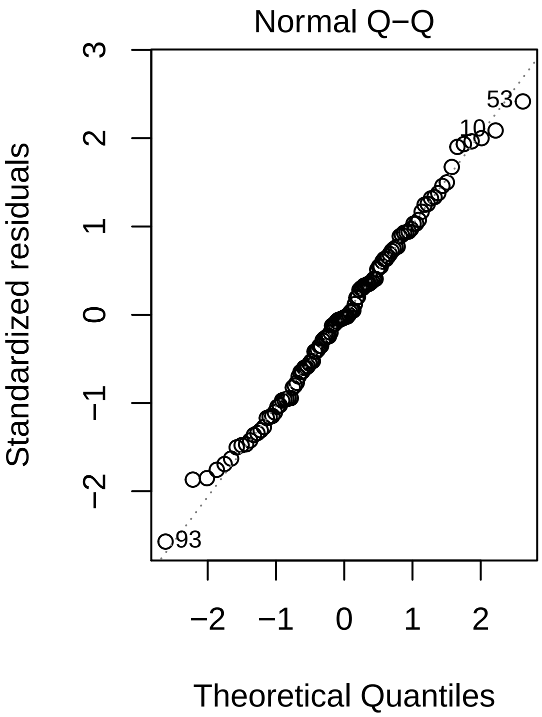
From this table, we can determine that variables *Stay* and *Xray* have a significant affect on infection rate while *Age* does not. $\hat{\beta}_0$, intercept, is also not significant, but in this context that is OK since we are not concerned what hospital infection rates are when Stay, Xray, and Age are all zero.

$$Inf\hat{ct}Rsk = 1.001 + 0.308(Stay) + -0.023(Age) + 0.02(Xray)$$

5) Model Adequacy

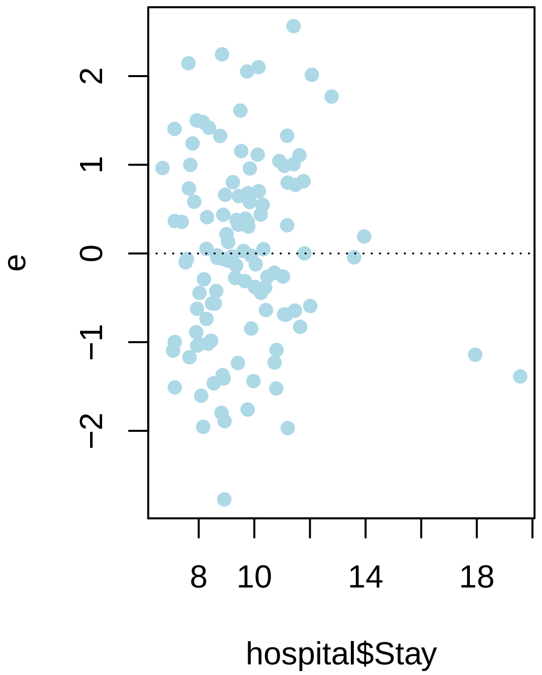
Here, we'd like to take a look at the adequacy of our model assumptions.

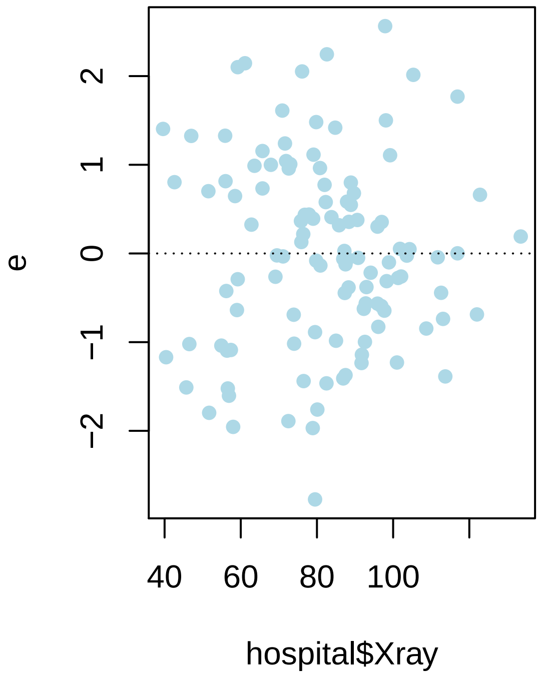


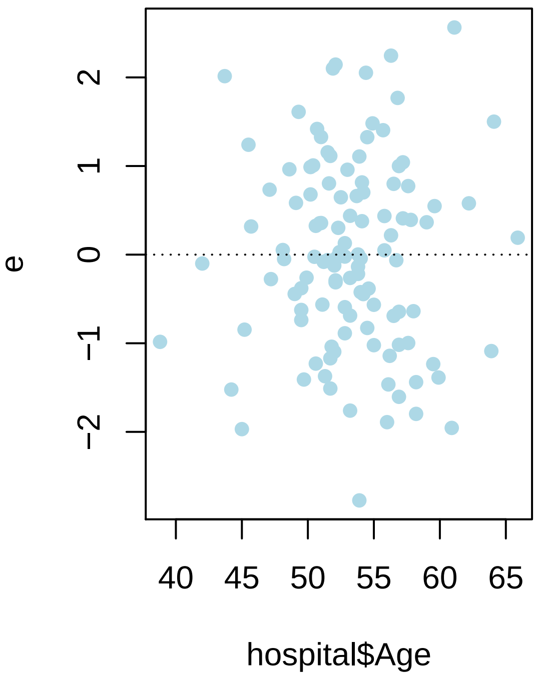


We need to check for linearity between our predictors and our response, *InfctRsk*. This assumption looks valid for *Stay* and *Xray* but not so much *Age*. There doesn't appear to much linearity between *InfctRsk* and *Age*. The QQplot looks great for normality of the errors from our model. This means that the errors from the model follow a "bell curve" or normal distribution. This is what we want. The equality of variances looks fine as well when our predicted values are lower. Equality of variances means that for every sub population of x_i the values for infection rate are also normally distributed. There isn't a lot of higher predicted values which may be why the residual plot looks a little concerning as $\hat{y}_i$ increases. The independence assumption is a little suspect. The longer one stays in the hospital, the more likely they are to get xrays. There may be a relationship here so indepdence assumption may not be fully satisfied.

We also want to plot the residuals by each predictor variable.







All three plots look OK, only a few outliers with *Stay* variable and the residuals. Otherwise, everything looks fine, and our model assumptions hold up.

6) Prediction

Using our above model to predict when $Stay = 8$, $Age = 50$, and $Xray = 82$:

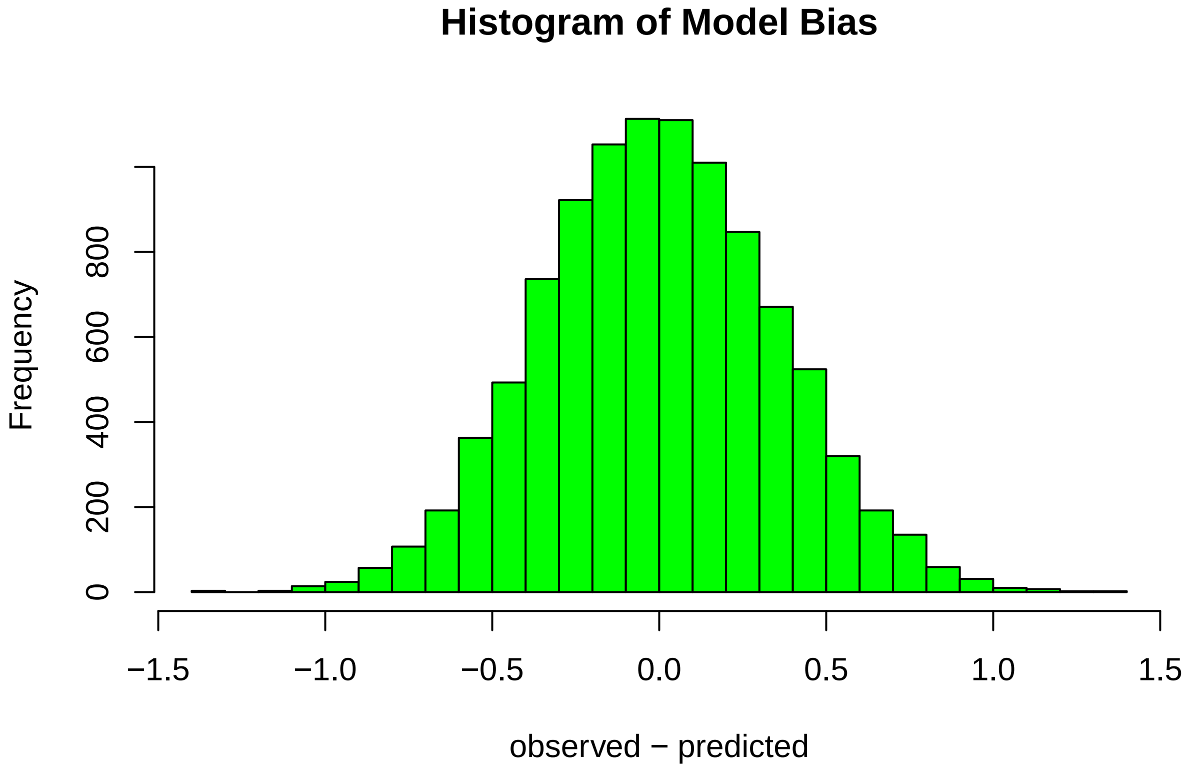
$$InfctRsk = \hat{\beta}_0 + \hat{\beta}_1(8) + \hat{\beta}_2(50) + \hat{\beta}_3(82)$$

InfctRsk	95% lower	95% upper
3.928544	3.629581	4.227508

We predict infection rate will be 3.9285442 ± 0.2989636, on average.

7) Cross Validation

Here, we test the model by taking a 10 percent simple random sample, fitting the model above using the remaining $113 - 11 = 102$ observations. We then test the model using the sampled 11 observations. We do this 10,000 times. This gives us an idea what the bias of our model will be, on average. I also display a historgram of the 10,000 simulations of the bias.



predicted bias	-0.0029598
predicted mse	1.2112948
predicted rmse	1.0798370

Bias is calculated using the difference of the observed *InfectRsk* (n=11) and the predicted *InfectRsk* (n=11). The simulation we performed calculates this value 10,000 times. A value for the predicted bias that is close to zero is very good. This means that on average our model predicts infection rates with very little bias. There is, however, variation around this value so we'd like to understand just how far we can be off on average as well. The histogram above displays this visually but we can also just take a 95% interval from this simulation. Here we display that quantile:

2.5%	-0.6785662
97.5%	0.6948670

So, we expect, on average, to be between -0.6785662 and 0.694867.

8) Conclusions

In understanding which factors influence infection rate the most we found that *Age* was not significant but *Stay* and *Xray* were very significant factors, *stay* being the most significant.

Additionaly we found that we have a model that can predict very well. The bias is very low, close to zero. How well can we predict? And is it good enough? We can answer the first question by calculating a confidence interval for $\hat{\sigma}$. The formula for this is as follows

$$C\left[\sqrt{\frac{SSE}{\chi^2_{1-\alpha/2:df}}} \leq \sigma \leq \sqrt{\frac{SSE}{\chi^2_{\alpha/2:df}}}\right] = 1 - \alpha$$

where *df* is equal to n-k-1.

$$C\left[\sqrt{\frac{128.2808541}{139.7838975}} \leq \sigma \leq \sqrt{\frac{128.2808541}{81.9996832}}\right] = 1 - 0.05$$

So [0.957971, 1.2507624] is the 95% confidence interval for $\hat{\sigma}$, and the answer to how well we can predict infection rate. In this instance we may only care about the upper bound since we are interested in how well we can predict infection rates.

R Code:

```r
# read the data from csv file cuz excel files are awful.
hospital1 <- read.csv("~/Documents/MATH3710/ProblemSets/problem4/hospitalinfectiondata.csv")
hospital <- hospital1[complete.cases(hospital1),]
hospital <- hospital[,2:5]
rm(hospital1)
y <- hospital$InfctRsk; x1 <- hospital$Age; x2 <- hospital$Stay
x3 <- hospital$Xray
# note, lm(y ~ xray + age) may be the best model here, although
# there may not be a significant diff between leaving in xray or stay
library(GGally)
ggpairs(hospital)
fit.all <- lm(InfctRsk ~ ., hospital)
fit.omit.age <- lm(InfctRsk ~ Stay + Xray, hospital) # model omits age
betas <- summary(fit.all)$coeff
b0 <- betas[1,1]; b1 <- betas[2,1]; b2 <- betas[3,1]; b3 <- betas[4,1]
```

```r
knitr::kable(summary(fit.all)$coeff)
par(mfrow=c(1,2))
plot(fit.all, which = c(1,2))
par(mfrow=c(1,2))
e <- fit.all$residuals
plot(hospital$Stay, e, col = "lightblue", pch = 16); abline(h=0, lty = 3)
plot(hospital$Xray, e, col = "lightblue", pch = 16); abline(h=0, lty = 3)
plot(hospital$Age, e, col = "lightblue", pch = 16); abline(h=0, lty = 3)
test <- data.frame(Stay = 8, Age = 50, Xray = 82)
hospital.pred <- predict.lm(fit.all, newdata = test, interval =
                           "confidence", se.fit = TRUE)
preds <- hospital.pred$fit
colnames(preds) <- c("InfctRsk", "95% lower", "95% upper")
# cross validation code below
bias <- vector()
hospital$index <- 1:dim(hospital)[1]
for (i in 1:10000) {
  ind <- sample(1:dim(hospital)[1], round(0.1*dim(hospital)[1])) # 10 percent
  test <- hospital[hospital$index %in% ind, ]
  training <- hospital[!(hospital$index %in% ind), ]
  # now to fit a new model with our subsetted training data.
  fit.train <- lm(InfctRsk ~ ., data = training)
  testing <- predict.lm(fit.train, newdata = test)
  e <- test$InfctRsk - testing
  e.squared <- e^2
  p.i <- c(mean(e), mean(e.squared), sqrt(mean(e.squared)))
  bias <- rbind(bias, p.i)
}
colnames(bias) <- c("predicted bias", "predicted mse", "predicted rmse")
results <- apply(bias, 2, mean)
hist(bias[,1], col = "green", breaks = 20, main = "Histogram of Model Bias",
     xlab = "observed - predicted")
# end cross validation code
# to calculate a CI for sigma
n <- dim(hospital)[1]; k <- 3
# note, n-k-1*sigmahat = sse
sse <- anova(fit.all)[4,2]
chi1 <- qchisq(0.975, df=n-k-1)
chi2 <- qchisq(0.025, df=n-k-1)
```