MSIS 5503 – Statistics for Data Science – Fall 2021 - Assignment 11 (15 Points)

Data File: hospinfct.csv (available in Canvas under Assignments).

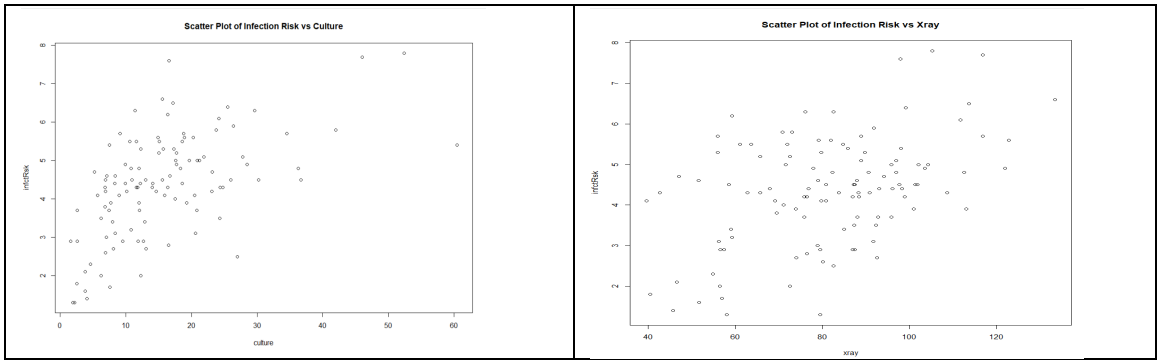
All analysis done in R

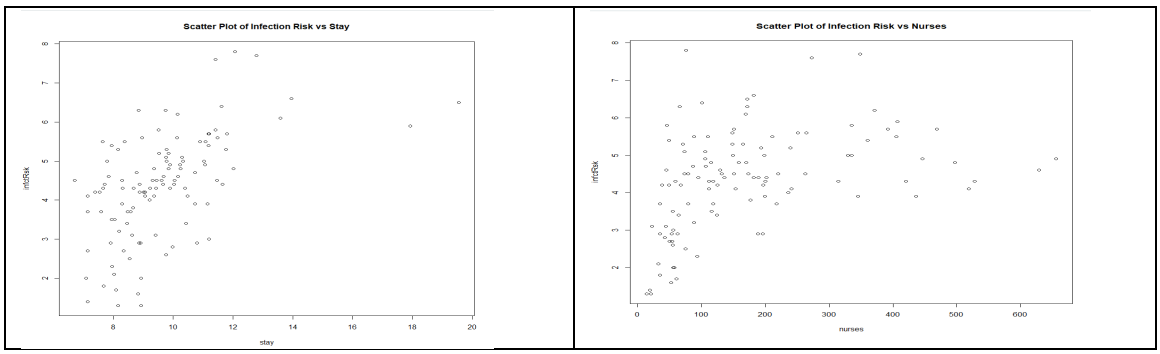
Question 1 (7 points):

Consider the regression model (Model 1) that predicts InfctRsk using Culture, Xray, Stay and Nurses as predictors.

```
> mod1 <- lm(infctRsk ~ culture+xray+stay+nurses)
> summary(mod1)

Call:
lm(formula = infctRsk ~ culture + xray + stay + nurses)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95753 -0.70926  0.02961  0.54734  2.45284

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3702661  0.5240788   0.707  0.48139
culture     0.0456048  0.0100219   4.551 1.41e-05 ***
xray        0.0126838  0.0054019   2.348  0.02069 *
stay        0.1935679  0.0548928   3.526  0.00062 ***
nurses      0.0020861  0.0006965   2.995  0.00340 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9556 on 108 degrees of freedom
Multiple R-squared:  0.5102,    Adjusted R-squared:  0.4921
F-statistic: 28.13 on 4 and 108 DF,  p-value: 5.189e-16
```

a) Produce Plots of each predictor against InfctRsk and **comment on the linearity** between the InfctRsk and each predictor.



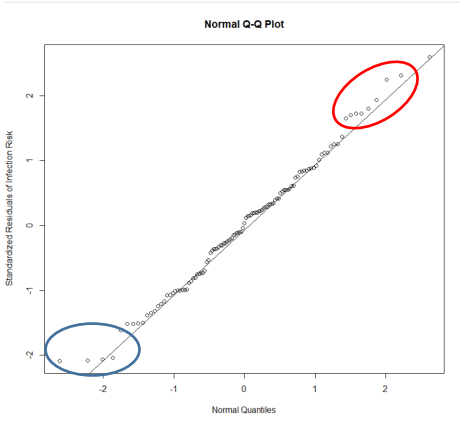


Infection Risk vs Culture: Data Indicates Some Curvature and Some Heteroscedasticity
Infection Risk vs Xray: Mostly Linear, Some Heteroscedasticity, may be outliers
Infection Risk vs Stay: Mostly Linear, may be outliers
Infection Risk vs Nurses: Data Indicates Some Curvature and Some Heteroscedasticity

b) For each plot, **indicate whether there are potential outliers** based on visual examination
(see above) – All interpretations are somewhat subjective

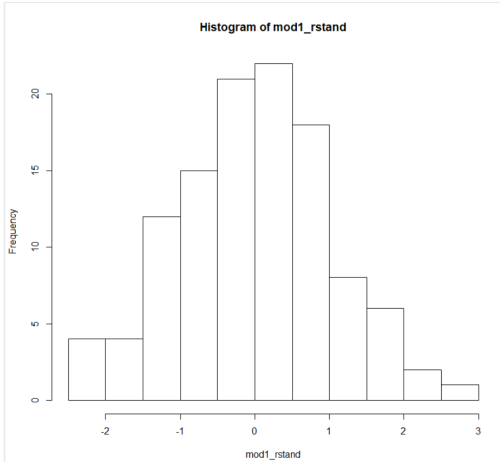
c) **Write your conclusion** about the normality of the Standardized residuals from this model based on Q-Q plots, histogram, skewness, kurtosis, Kolmogorov-Smirnov Test and Shapiro-Wilk Test. **You must have a comment for each of these checks**.

```
> library(moments)
> mod1_rstand <-rstandard(mod1)
> qqnorm(mod1_rstand, ylab="Standardized Residuals of Infection Risk", xlab="Normal Quantiles")
> qqline(mod1_rstand)
> hist(mod1_rstand)
> print(skewness(mod1_rstand))
[1] 0.1401317
> print(kurtosis(mod1_rstand))
[1] 2.721159
```



Q-Q Plot indicates that there are fewer lower valued residuals (blue ellipse) and fewer higher valued residuals (red ellipse) than would be expected from a normal distribution. This means that the residuals are more concentrated around the middle than a normal

distribution. This means that the distribution of the residuals has “thin tails” indicating that the distribution is platykurtic.
Further, both sets of extreme values are on the same side of the qqline, indicating some bow-shape and therefore some skewness as well. Since they are above the line at the ends, the skewness could be positive.
The presence of extreme residuals at either end means that these could be outliers and need close attention. Note that all of these apply to the standardized residuals from the current model, without any variable transformation.
We conclude from our visual inspection that normality of residuals *may* be an issue.



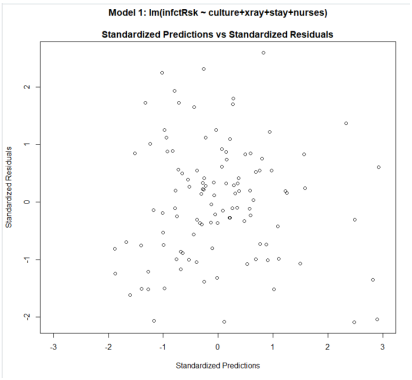
Visually, the histogram indicates a that the left side of the distribution is heavier with somewhat longer right tail (positive skew). It also appears somewhat flattened indicating that it is platykurtic.

The skewness of 0.141 indicates positive or right-skew and the kurtosis (2.72) is less than 3 which indicates it is platykurtic.

```
> # Residuals Normality Tests
> #
> # Kolomogorov-Smirnov Test
> print(ks.test(mod1_rstand,"pnorm"))

        One-sample Kolmogorov-Smirnov test

data:  mod1_rstand
D = 0.0498, p-value = 0.942
alternative hypothesis: two-sided

> # Shapiro-Wilk Test
> print(shapiro.test(mod1_rstand))

        Shapiro-Wilk normality test

data:  mod1_rstand
W = 0.99001, p-value = 0.5794
```

The p-values form both the K-S test and the S-W test indicate that the residuals from our model are not significantly different (at $\alpha = 0.05$) from that of a normal distribution.

d) Plot the standardized Residuals against the Standardized Predictions from the model, and **comment on the homoscedasticity (or otherwise) of the residuals).**



There is no distinct pattern in the plot, which shows the standardized residuals somewhat evenly distributed around the zero line. Heteroscedasticity of the residuals does not appear to be an issue.

e) Produce the Multicollinearity Diagnostics and for the predictors in the model and **indicate your conclusions.**

```
> # Multicollinearity Diagnostics
> #
> library(olsrr)
> #
> ols_vif_tol(mod1)
# A tibble: 4 x 3
  Variables Tolerance   VIF
  <chr>         <dbl> <dbl>
1 culture       0.775  1.29
2 xray          0.745  1.34
3 stay          0.741  1.35
4 nurses        0.867  1.15
```

The tolerances are values are greater than 0.2 and the VIFs are less than 5 indicating that multicollinearity among the predictors is not present in this model.

f) Produce the Outlier Diagnostics (Leverage, Studentized Residuals and Cook's D) for the all observations in the model. *Identify the observation* (i.e., its ID) *with highest Cook's D value*. Interpret its Leverage (**what it means**), its Studentized Residual (**what it means**) and its Cook's D (**what it means**).

```
> # Outlier Diagnostics
> #
> leverage <- hatvalues(mod1)
> stud_res <- rstudent(mod1)
> cook_dist <- cooks.distance(mod1)
> #
> #
> df_out1 <-data.frame(id, culture, xray, stay, nurses, infctRsk, leverage, stud_res, cook_dist)
> print(round(df_out1,4))
    id culture  xray  stay nurses infctRsk leverage stud_res cook_dist
1    1     9.0  39.6  7.13    241      4.1   0.0611   1.0089    0.0133
2    2     3.8  51.7  8.82     52      1.6   0.0394  -1.5199    0.0187
3    3     8.1  74.0  8.34     54      2.7   0.0192  -0.7436    0.0022
4    4    18.9 122.8  8.95    148      5.6   0.0655   0.8313    0.0097
5    5    34.5  88.9 11.20    151      5.7   0.0459   0.1555    0.0002
6    6    21.9  97.0  9.76    106      5.1   0.0187   0.4107    0.0006
7    7    16.7  79.0  9.68    129      4.6   0.0105   0.3387    0.0002
8    8    60.5  85.8 11.18    360      5.4   0.2114  -2.0726    0.2235
9    9    24.4  90.8  8.67    118      4.3   0.0235  -0.2732    0.0004
10  10    29.6  82.6  8.84     66      6.3   0.0402   1.8170    0.0271
11  11    28.5 122.0 11.07    656      4.9   0.1597  -2.1211    0.1657
12  12     6.8  83.8  8.30     59      4.3   0.0239   0.8750    0.0037
13  13    46.0 116.9 12.78    349      7.7   0.0956   0.6005    0.0077
14  14    20.8  88.0  7.58     79      3.7   0.0301  -0.3885    0.0009
15  15    14.6  76.4  9.00     38      4.2   0.0179   0.3928    0.0006
```

Observation 8 has the highest Cook's D value of 0.2235. Since the value is relatively small, it is not influential and therefore there are no influential outliers. The leverage for

observation 8 is 0.2114 which means that it is fairly unremarkable in terms of its value on the "x-axis" (it is not an extreme value of the combination of predictors). Its studentized residual is -2.027 which means that it is an extreme value on the "y-axis" i.e., it is an extreme value relative to the model prediction of Infection Risk.

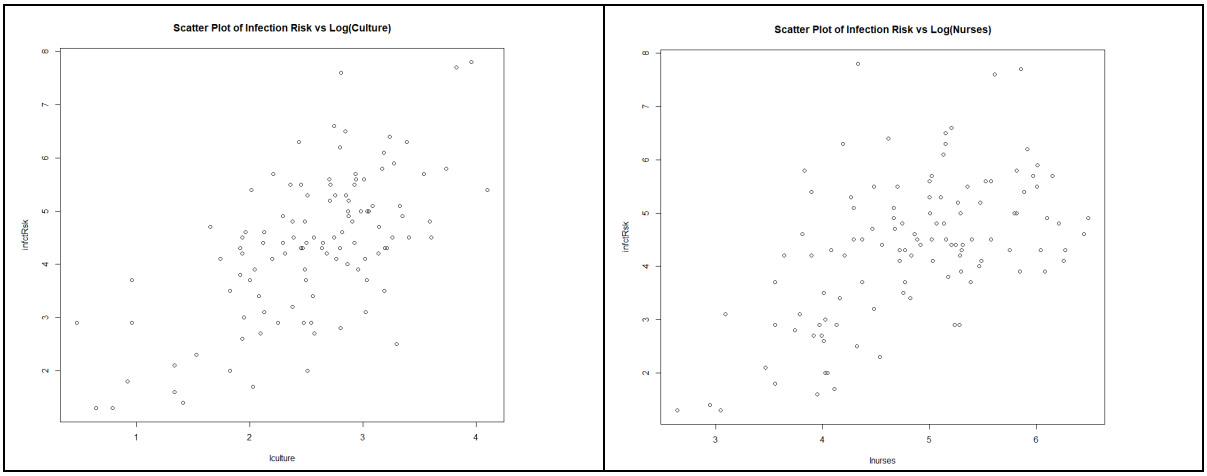
Question 2 (7 points):

Transform the Culture and Nurses variable using **log transformations** to produce lculture and lnurses as predictors.

Consider the regression model (Model 2) that predicts InfctRsk using lculture, Xray, Stay and lnurses as predictors.

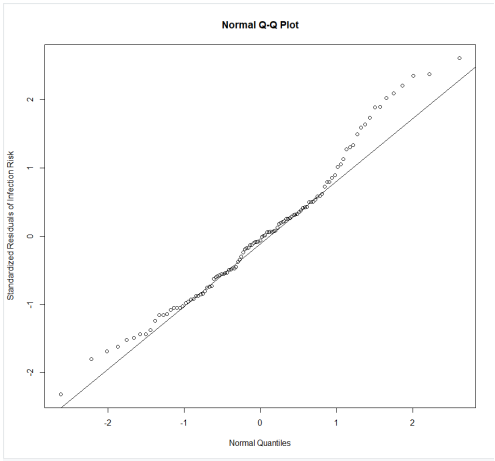
```
> mod2 <- lm(infctRsk ~ lculture+xray+stay+lnurses)
> summary(mod2)

Call:
lm(formula = infctRsk ~ lculture + xray + stay + lnurses)

Residuals:
     Min       1Q   Median       3Q      Max
-1.99589 -0.62880 -0.05365  0.42856  2.26145

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.082035   0.583025  -3.571 0.000532 ***
lculture     0.751300   0.146996   5.111 1.39e-06 ***
xray         0.008171   0.005178   1.578 0.117483
stay         0.179924   0.050730   3.547 0.000578 ***
lnurses      0.438012   0.115206   3.802 0.000238 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8791 on 108 degrees of freedom
Multiple R-squared:  0.5856,    Adjusted R-squared:  0.5702
F-statistic: 38.15 on 4 and 108 DF,  p-value: < 2.2e-16
```

Repeat steps (a) and (b) from question 1 for lculture and lnurses only.



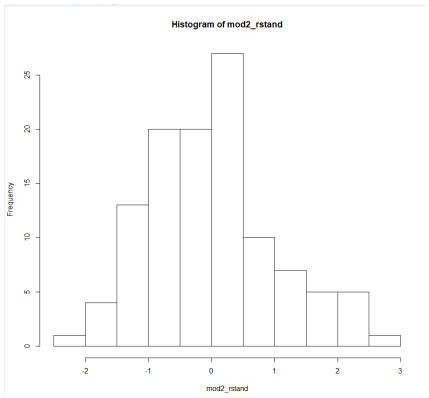
There is a definite reduction in the curvature in both plots compared to before transformation. That is, both plots appear to show a linear relationship between Infection Risk and the transformed culture and nurses predictors.

Repeat steps c) through f) for Model 2.



The Q-Q plot shows a sharp departure from normality. The tails are no longer thin nor are they heavy since the values are spread out more and not concentrated towards the middle of the line. So, it is neither platykurtic nor leptokurtic. However, there is a distinct bow-shaped pattern indicating substantial skewness (positive since the bow is predominantly above the line). Overall, the Q-Q plot indicates substantial departure from normality.

The histogram bellow indicates that the distribution is definitely right-skewed.

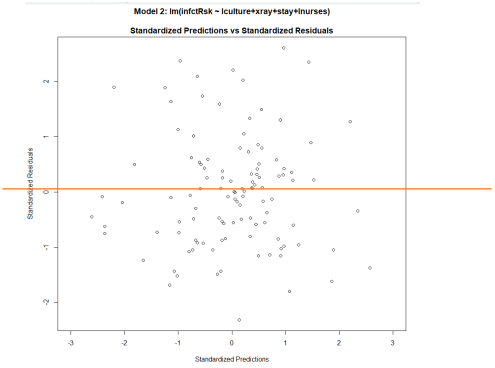



```
> print(skewness(mod2_rstand))
[1] 0.4661255
> print(kurtosis(mod2_rstand))
[1] 2.954274
```

```
> # Residuals Normality Tests
> #
> # Kolomogorov-Smirnov Test
> print(ks.test(mod2_rstand,"pnorm"))

        One-sample Kolmogorov-Smirnov test

data:  mod2_rstand
D = 0.068367, p-value = 0.6664
alternative hypothesis: two-sided

> # Shapiro-Wilk Test
> print(shapiro.test(mod2_rstand))

        Shapiro-Wilk normality test

data:  mod2_rstand
W = 0.97693, p-value = 0.04769
```

The skewness value shows that the distribution of the standardized residuals has a positive skew. The kurtosis of the distribution of the standardized residuals indicates that it is pretty close to that of a normal distribution (close to kurtosis of 3)

The p-values form both the K-S test indicates that the residuals from our model are not significantly different (at $\alpha = 0.05$) from that of a normal distribution. But, the S-W test rejects the null hypothesis that the residuals are from a normal distribution.



There is no distinct pattern in the plot, which shows the standardized residuals somewhat evenly distributed around the zero line. Heteroscedasticity of the residuals does not appear to be an issue.

```
> ols_vif_tol(mod2)
# A tibble: 4 x 3
  Variables Tolerance   VIF
  <chr>         <dbl> <dbl>
1 lculture      0.661  1.51
2 xray          0.686  1.46
3 stay          0.734  1.36
4 lnurses       0.748  1.34
> #
```

The tolerances are values are greater than 0.2 and the VIFs are less than 5 indicating that multicollinearity among the predictors is not present in this model.

```
> #
> # Outlier Diagnostics
> #
> leverage <- hatvalues(mod2)
> stud_res <- rstudent(mod2)
> cook_dist <- cooks.distance(mod2)
> #
> #
> df_out2 <-data.frame(id, culture, xray, stay, nurses, infctRsk, leverage, stud_res, cook_dist)
> print(round(df_out2,4))
     id culture  xray  stay nurses infctRsk leverage stud_res cook_dist
```

53	53	16.6	97.9	11.41	273	7.6	0.0233	2.6763	0.0323
54	54	52.4	105.3	12.07	76	7.8	0.0794	2.3970	0.0949
55	55	8.4	56.2	8.63	44	3.1	0.0374	-0.0997	0.0001
56	56	7.7	73.9	11.15	199	3.9	0.0275	-0.5520	0.0017
57	57	2.6	75.8	7.14	35	3.7	0.0729	1.9170	0.0564
58	58	16.4	65.7	7.65	314	4.3	0.0474	-0.1753	0.0003
59	59	19.3	101.0	10.73	345	3.9	0.0288	-1.8167	0.0192
60	60	15.6	97.7	11.46	132	4.5	0.0205	-0.5510	0.0013
61	61	8.0	59.0	10.42	64	3.4	0.0385	-0.2990	0.0007
62	62	18.8	55.9	11.18	392	5.7	0.0610	0.5782	0.0044

Observation 54 now has the highest Cook's D value of 0.0949 and it is not influential (so none of the other observations are influential outliers as well). Its leverage of 0.0794 makes it unremarkable on the "x-axis" in that it is not an extreme value on the "x-axis" when all the predictors lculture, xray, stay, lnurses are considered. It's studentized residual (2.3970) is an extreme value on the "Y-axis". However, Cook's D takes into account both leverage and the studentized residual and in combination observation 54 is not an influential outlier. So, there are no influential outliers.

Question 3 (1 point):

Compare Model 1 and Model 2. Write a summary on which model you would prefer and why. In your answer, comment on differences (if any) in terms of the linearity, normality, heteroscedasticity, multicollinearity and outliers between the two models, in addition to model fit statistics such as R-squared.

```
> mod1 <- lm(infctRsk ~ culture+xray+stay+nurses)
> summary(mod1)

Call:
lm(formula = infctRsk ~ culture + xray + stay + nurses)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95753 -0.70926  0.02961  0.54734  2.45284

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3702661  0.5240788   0.707  0.48139
culture     0.0456048  0.0100219   4.551 1.41e-05 ***
xray        0.0126838  0.0054019   2.348  0.02069 *
stay        0.1935679  0.0548928   3.526  0.00062 ***
nurses      0.0020861  0.0006965   2.995  0.00340 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9556 on 108 degrees of freedom
Multiple R-squared:  0.5102,    Adjusted R-squared:  0.4921
F-statistic: 28.13 on 4 and 108 DF,  p-value: 5.189e-16
```

```
> mod2 <- lm(infctRsk ~ lculture+xray+stay+lnurses)
> summary(mod2)

Call:
lm(formula = infctRsk ~ lculture + xray + stay + lnurses)

Residuals:
     Min       1Q   Median       3Q      Max
-1.99589 -0.62880 -0.05365  0.42856  2.26145

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.082035   0.583025  -3.571 0.000532 ***
lculture     0.751300   0.146996   5.111 1.39e-06 ***
xray         0.008171   0.005178   1.578 0.117483
stay         0.179924   0.050730   3.547 0.000578 ***
lnurses      0.438012   0.115206   3.802 0.000238 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8791 on 108 degrees of freedom
Multiple R-squared:  0.5856,    Adjusted R-squared:  0.5702
F-statistic: 38.15 on 4 and 108 DF,  p-value: < 2.2e-16
```

Neither model had influential outliers, were both homoscedastic in their residuals, and did not exhibit any multicollinearity.

In terms of R^2, Model 2 with transformed predictors is clearly superior. Thus, if the goal was better *prediction, Model 2 would be preferred over Model 1*. The improved R^2 in Model 2 is a result of improved linearity of the relationship of the transformed predictors lculture and lnurses with Infection Risk.

However, Model 2 appears to violate an important assumption of regression, namely normality of the noise term. The residuals were distinctly more non-normal than Model 2. While this does not affect predictions from the model, the significance tests for the predictors in Model 2 may be unreliable. It was seen in Model 2 that Xray was no longer a significant predictor in the presence

of the other predictors. But this conclusion may not be correct in light of the potential non-normality of the residuals. Thus, for purposes of *explanation* (i.e., which predictors are important) *Model 1 may be preferred over Model 2*.