MSIS 5503 – Statistics for Data Science – Fall 2021 - Assignment 10 (10 Points)

This is an individual assignment. Problems like these are likely to appear on your exam. Do not consult other students or otherwise plagiarize. Violations will be subject to Academic Integrity actions

Type your answers and submit it to Canvas by
Sunday, October 31, 11:59 pm

Important: Your submission must be a Word or pdf document with your name clearly stated in the document.
You should copy and paste all computer output into this document appropriately.

The SENIC Data Set

These data were obtained as part of the Study on the Efficacy of Nosocomial Infection Control (SENIC) to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in US hospitals. This data set consists of a random sample of n=113 hospitals selected from the original N=338 hospitals surveyed. Each hospital is given an ID number, and is measured on 11 other variables.

Variable Name	Variable Label	Description
ID	Identification Number	1 – 113
Stay	Length of Stay	Average length of stay of all patients in the hospital (measured in days)
Age	Age	Average age of patients (in years)
InfctRsk	Infection Risk	Average estimated probability of acquiring infection in hospital (in percent)
Culture	Routine Culturing Ratio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
Xray	Routine Chest X-ray Ratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
Beds	Number of Beds	Average number of beds in hospital during study period
MedSchool	Medical School Affiliation	1=Yes, 2=No
Region	Region	Geographic Region, where 1=NE 2=NC 3=S 4=W
Census	Average Daily Census	Average number of patients in hospital per day during study period
Nurses	Number of Nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number of full-time + ½ number of part-time)

Facilities	Available Facilities & Services	Percent of 35 potential facilities and services that are provided by the hospital

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

All analysis done in R

Question 1 (1 point):

a) (**0.25 points**) Produce the Pairwise Correlation Matrix with variables in the order: InfctRsk Stay Age Culture Xray Beds Nurses

```
> M <- cbind(infctRsk, stay, age, culture, xray, beds, nurses)
> print(round(cor(M),4))
          infctRsk   stay     age culture    xray    beds  nurses
infctRsk    1.0000 0.5334  0.0011  0.5592  0.4534  0.3598  0.3940
stay        0.5334 1.0000  0.1889  0.3267  0.3825  0.4093  0.3404
age         0.0011 0.1889  1.0000 -0.2258 -0.0189 -0.0588 -0.0829
culture     0.5592 0.3267 -0.2258  1.0000  0.4250  0.1397  0.1989
xray        0.4534 0.3825 -0.0189  0.4250  1.0000  0.0458  0.0774
beds        0.3598 0.4093 -0.0588  0.1397  0.0458  1.0000  0.9155
nurses      0.3940 0.3404 -0.0829  0.1989  0.0774  0.9155  1.0000
```

b) (**0.25 points**) Produce a Partial Correlation Matrix with variables in the same order.

```
> library(ppcor)
> parcor <- pcor(M)
> print(parcor)
$estimate
            infctRsk       stay         age    culture        xray        beds       nurses
infctRsk  1.00000000  0.2815894  0.07331893  0.3974645  0.21701481 -0.02610794  0.14755166
stay      0.28158936  1.0000000  0.27583363  0.1124571  0.23035028  0.32561174 -0.19954940
age       0.07331893  0.2758336  1.00000000 -0.2929705 -0.03267718 -0.08089941  0.00678254
culture   0.39746446  0.1124571 -0.29297051  1.0000000  0.19105205 -0.13608901  0.10420822
xray      0.21701481  0.2303503 -0.03267718  0.1910520  1.00000000 -0.11350668  0.03374808
beds     -0.02610794  0.3256117 -0.08089941 -0.1360890 -0.11350668  1.00000000  0.89977869
nurses    0.14755166 -0.1995494  0.00678254  0.1042082  0.03374808  0.89977869  1.00000000
```

c) (**0.5 points**) Choose any one variable among Stay Age Culture Xray Beds Nurses and explain the difference between the Pairwise Correlation of that variable with InfctRsk and its Partial Correlation with InfctRsk. *In other words, (in your own words) what does partial correlation convey that is different from pairwise correlation*?

Consider Stay and InfctRsk. The partial correlation between them (0.2816) "controls for" (or accounts for the effect of the pairwise correlation of all other variables with InfctRsk and *with Stay*.

Question 2 (2 points):

a) (**0.5 points**) Run a simple regression model between InfctRsk and Beds. Then run a simple regression model between InfctRsk and Nurses. What is your conclusion for each of these models (*based on the hypothesis tests for the betas*?)

```
> mod1 <- lm(infctRsk ~ beds)
> summary(mod1)

Call:
lm(formula = infctRsk ~ beds)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6117 -0.8142  0.0831  0.7259  3.6832

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.7240393  0.1951667  19.081  < 2e-16 ***
beds        0.0025016  0.0006158   4.062 9.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.257 on 111 degrees of freedom
Multiple R-squared:  0.1294,    Adjusted R-squared:  0.1216
F-statistic:  16.5 on 1 and 111 DF,  p-value: 9.087e-05

> mod2 <- lm(infctRsk ~ nurses)
> summary (mod2)

Call:
lm(formula = infctRsk ~ nurses)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4773 -0.8315  0.0315  0.7316  3.8140

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.697665   0.186387  19.839  < 2e-16 ***
nurses      0.003793   0.000840   4.516 1.58e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.238 on 111 degrees of freedom
Multiple R-squared:  0.1552,    Adjusted R-squared:  0.1476
F-statistic:  20.4 on 1 and 111 DF,  p-value: 1.578e-05
```

When used individually in a simple regression model to predict InfctRsk, both Nurses and Beds are statistically significant predictors of InfctRsk.

b) (**0.5 points**) Run a multiple regression model that predicts InfctRsk using both Nurses and Beds. What is your conclusion about Beds and Nurses (*based on the hypothesis tests*?)

```
> mod3 <- lm(infctRsk ~ beds + nurses)
> summary (mod3)

Call:
lm(formula = infctRsk ~ beds + nurses)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4768 -0.8302  0.0260  0.7305  3.8151

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.699e+00  1.936e-01  19.105   <2e-16 ***
beds        -3.959e-05  1.515e-03  -0.026   0.9792
nurses       3.844e-03  2.097e-03   1.833   0.0696 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.244 on 110 degrees of freedom
Multiple R-squared:  0.1552,    Adjusted R-squared:  0.1399
F-statistic: 10.11 on 2 and 110 DF,  p-value: 9.348e-05
```

Compared with the conclusion in (a), we see that neither variable is a significant predictor of InfctRsk, when included together in the model.

c) (**0.5 points**) *Explain* the results that you see in (b) in light of the results in (a). **Hint:** Look at the correlation matrix as well.

The correlation between Nurses and Beds is 0.9155. When we use both in the regression model as predictors, the slope coefficients are calculated "controlling for the other variable", i.e., accounting for the effect of the correlation of each predictor on the other. This makes neither variable statistically significant (Later, we will see this is called "Multicollinearity").

d) (**0.5 points**) What do the results from (a) and (b) say about choosing predictors for regression models, in general.

This tells us that choosing predictors on the basis of statistical significance is complicated because the significance of the slope of a predictor depends on what other predictors are already in the model. This is due to inter-correlation among predictors.

Question 3 (4 points):

a) (**0.5 points**) Run a Stepwise regression (backward selection) that predicts InfctRsk using Stay Age Culture Xray Beds Nurses with the probability of retaining a predictor in the model = 0.05. (i.e., **prem** = 0.05)

```
> df1 <- data.frame(infctRsk, stay, age, culture, xray, beds, nurses)
> step_model <- lm(infctRsk ~ ., data=df1)
> ols_step_backward_p(step_model, prem=0.05, details=FALSE)
Backward Elimination Method

Final Model Output
------------------

                        Model Summary
---------------------------------------------------------------
R                       0.714       RMSE                 0.956
R-Squared               0.510       Coef. Var           21.944
Adj. R-Squared          0.492       MSE                  0.913
Pred R-Squared          0.456       MAE                  0.753
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression    102.751          4         25.688    28.128    0.0000
Residual       98.629        108          0.913
Total         201.380        112
--------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     0.370         0.524                 0.707    0.481    -0.669     1.409
       stay     0.194         0.055        0.276    3.526    0.001     0.085     0.302
    culture     0.046         0.010        0.348    4.551    0.000     0.026     0.065
       xray     0.013         0.005        0.183    2.348    0.021     0.002     0.023
     nurses     0.002         0.001        0.217    2.995    0.003     0.001     0.003
-----------------------------------------------------------------------------------------
```

b) (**1.5 points**) Develop a final regression model based on that results of (a) that also shows the standardized coefficients. For this final model:

i. Write out the regression prediction equation.

(Predicted or Expected) InfctRsk = 0.370 + 0.194Stay + 0.002Nurses + 0.046Culture + 0.013Xray

ii. Interpret the slope coefficient in the context of the problem for any one of the predictors, in the context of the problem.

The expected Infection Risk increases by 0.194 (in Infection Risk Units) for each additional day stay in the hospital, controlling for the effect of Nurses, Culture and Xray on Infection Risk and Number of days of stay in the hospital.

iii. Interpret the F-test (i.e., what null hypothesis does it test, what is the value of the test statistic and what is the conclusion of the hypothesis test)

The F-test tells us whether the model with all the predictors included is significant (or whether all the predictors taken together provide a statistically significant explanation of the variability in Infection Risk.). It tests the null hypothesis that every slope coefficient is zero against the alternate hypothesis that *at least one of the slope coefficients* is significantly different from zero. In this case, it tells us that at least one of the slope coefficients of Stay, Nurses, Culture and Xray is significantly different from zero.

iv. Interpret the R^2

$R^2 = 0.510$ tells us that the together the four predictors Stay, Nurses, Culture and Xray explain 51.0% of the variability in Infection Risk.

v. Interpret the Adjusted R^2

Adjusted $R^2 = 0.492$ tells us that the together the four predictors Stay, Nurses, Culture and Xray explain 49.2% of the variability in Infection Risk, taking into account that we have 4 predictors. Since the difference with R^2 is small, it tells us that we have not used too many predictors.

vi. Predict the Infection Risk for any one choice of values for the predictors.

Student will have to put in values for the predictors and predict expected Infection Risk.

c) (**1 point**) Run a *standardized regression model* and interpret the standardized coefficients *relative to each other.*

```
> z_infctRsk <- (infctRsk - mean(infctRsk))/sd(infctRsk)
> z_stay <- (stay - mean(stay))/sd(stay)
> z_culture <- (culture - mean(culture))/sd(culture)
> z_xray <- (xray - mean(xray))/sd(xray)
> z_nurses <- (nurses - mean(nurses))/sd(nurses)
> #
> z_mod3 <- lm(z_infctRsk ~ z_stay+z_culture+z_xray+z_nurses)
> summary(z_mod3)

Call:
lm(formula = z_infctRsk ~ z_stay + z_culture + z_xray + z_nurses)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45986 -0.52894  0.02208  0.40819  1.82924

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.613e-16  6.704e-02   0.000  1.00000
z_stay      2.759e-01  7.825e-02   3.526  0.00062 ***
z_culture   3.481e-01  7.649e-02   4.551 1.41e-05 ***
z_xray      1.832e-01  7.801e-02   2.348  0.02069 *
z_nurses    2.167e-01  7.234e-02   2.995  0.00340 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7127 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
```

The standardized model says that Culture has the largest effect on Infection Risk followed by Stay, Nurses and Xrays, based on the standardized coefficients.

d) (**0.5 points**) Run a Stepwise regression (forward selection) that predicts InfctRsk using Stay Age Culture Xray Beds Nurses with the probability of entering a predictor in the model = 0.05. (i.e., **penter** = 0.05)

```
> library(olsrr)
> df1 <- data.frame(infctRsk, stay, age, culture, xray, beds, nurses)
> step_model <- lm(infctRsk ~ ., data=df1)
> #ols_step_backward_p(step_model, prem=0.05, details=FALSE)
> ols_step_forward_p(step_model, penter=0.05, details=FALSE)
```

```
Final Model Output
------------------

                        Model Summary
----------------------------------------------------------------
R                       0.714       RMSE                 0.956
R-Squared               0.510       Coef. Var           21.944
Adj. R-Squared          0.492       MSE                  0.913
Pred R-Squared          0.456       MAE                  0.753
----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression    102.751          4         25.688    28.128    0.0000
Residual       98.629        108          0.913
Total         201.380        112
--------------------------------------------------------------------

                                 Parameter Estimates
-----------------------------------------------------------------------------------------
      model     Beta    Std. Error    Std. Beta      t       Sig      lower    upper
-----------------------------------------------------------------------------------------
(Intercept)    0.370         0.524                 0.707    0.481    -0.669    1.409
    culture    0.046         0.010         0.348   4.551    0.000     0.026    0.065
       stay    0.194         0.055         0.276   3.526    0.001     0.085    0.302
     nurses    0.002         0.001         0.217   2.995    0.003     0.001    0.003
       xray    0.013         0.005         0.183   2.348    0.021     0.002    0.023
-----------------------------------------------------------------------------------------
```

e) (**0.5 points**) Explain the differences in result (if any) between forward and backward selections i.e., *why did they occur*.

In this case, there were no differences in the final model between forward and backward selection based on the specified prem and penter parameters.

Question 4 (3 points):

a) (**0.5 points**) Generate a dummy variable d_medschool which takes on a value of 1 if MedSchool is 2, else 0 if not. Then predict InfctRsk using d_medschool. Now interpret the beta coefficient and the test for the beta coefficient for this model.

```
> d_medschool <- ifelse(medschool == 2, 1, 0)
> mod4 <- lm(infctRsk~d_medschool)
> summary(mod4)

Call:
lm(formula = infctRsk ~ d_medschool)

Residuals:
   Min     1Q Median     3Q    Max
-2.924 -0.824  0.076  0.776  3.576

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0941     0.3177  16.035   <2e-16 ***
d_medschool  -0.8702     0.3447  -2.525    0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.31 on 111 degrees of freedom
Multiple R-squared:  0.0543,    Adjusted R-squared:  0.04578
F-statistic: 6.374 on 1 and 111 DF,  p-value: 0.013

> print(anova(mod4))
Analysis of Variance Table

Response: infctRsk
             Df  Sum Sq Mean Sq F value Pr(>F)
d_medschool   1  10.936 10.9355  6.3737  0.013 *
Residuals   111 190.444  1.7157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The slope coefficient of -0.8702 tells us that the Expected (or Average or Mean) Infection Risk decreases by -0.87 units (of Infection Risk) when the hospital is in a location with a Medical School. The test for the slope tells us that the estimated slope coefficient is significantly different from zero at $\alpha = 0.05$, indicating that the expected (mean) Infection risk is different (less) for Hospitals having a Medical School in the same location. This is also confirmed by the F-test in the ANOVA output of the regression model.

b) (**1 point**) Run the same analysis in (a) using two-sample t-test (comparison of means). What is the null hypothesis that is tested, and what is the conclusion?

```
> medschool_infctRsk <- subset(df$InfctRsk, df$MedSchool == 2)
> nomedschool_infctRsk <- subset(df$InfctRsk, df$MedSchool == 1)
> ttest <- t.test(medschool_infctRsk, nomedschool_infctRsk, alternative = c("two.sided"),
+                 mu = 0, paired = FALSE, conf.level = 0.95)
> print(ttest)

        Welch Two Sample t-test

data:  medschool_infctRsk and nomedschool_infctRsk
t = -2.8772, df = 25.01, p-value = 0.008093
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4930247 -0.2472939
sample estimates:
mean of x mean of y
 4.223958  5.094118
```

The t-test tests the null hypothesis that the Mean Infection Risk is the same for hospitals whether there is a Medical School in the same location or not. The alternate hypothesis is that the difference is not zero, i.e., that there is a difference in mean Infection Risk. The conclusion, on the basis of the t-test is that there is a significant difference at $\alpha = 0.05$ (in the Mean Infection Risk depending on whether a Medical School is in the same location as the hospital).

c) (**0.5 points**) Compare the tests and the results in (a) and (b) and indicate whether they are the same or different.

Comparing the results of the regression ANOVA and the t-test, results from both analyses lead to the same conclusions.

d) (**1 point**) Run a multiple regression model which predicts InfctRsk using beds and d_medschool.

```
> mod5 <- lm(infctRsk~d_medschool+beds)
> summary(mod5)

Call:
lm(formula = infctRsk ~ d_medschool + beds)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6121 -0.8192  0.0987  0.7187  3.6885

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.8556509  0.5037629   7.654 8.04e-12 ***
d_medschool -0.1167642  0.4117248  -0.284  0.77725
beds         0.0023731  0.0007667   3.095  0.00249 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.262 on 110 degrees of freedom
Multiple R-squared:  0.1301,    Adjusted R-squared:  0.1143
F-statistic: 8.224 on 2 and 110 DF,  p-value: 0.0004695

> print(anova(mod5))
Analysis of Variance Table

Response: infctRsk
             Df  Sum Sq Mean Sq F value   Pr(>F)
d_medschool   1  10.936 10.9355  6.8664 0.010024 *
beds          1  15.258 15.2581  9.5806 0.002494 **
Residuals   110 175.186  1.5926
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i. Interpret the results of this model.

We see that with the inclusion of the number of Beds in the model to predict Infection Risk, there is no significant difference in mean (or expected) Infection Risk whether there is a medical school in the same location as the hospital or not. This result is different from the previous t-test and regression result because of the inter-correlation between Beds and d_medschool. That is, once we account for the number of Beds in a hospital, the presence of medical school (or absence) makes no difference.

ii. Write out the regression equations for predicting InfctRsk for models in areas with a Medical School and Without a Medical School (identify each model clearly).

Model where there is a Medical School:

Expected Infection Risk = 3.738 + 0.0024Beds

Model where there is NO Medical School:

Expected Infection Risk = 3.856 + 0.0024Beds