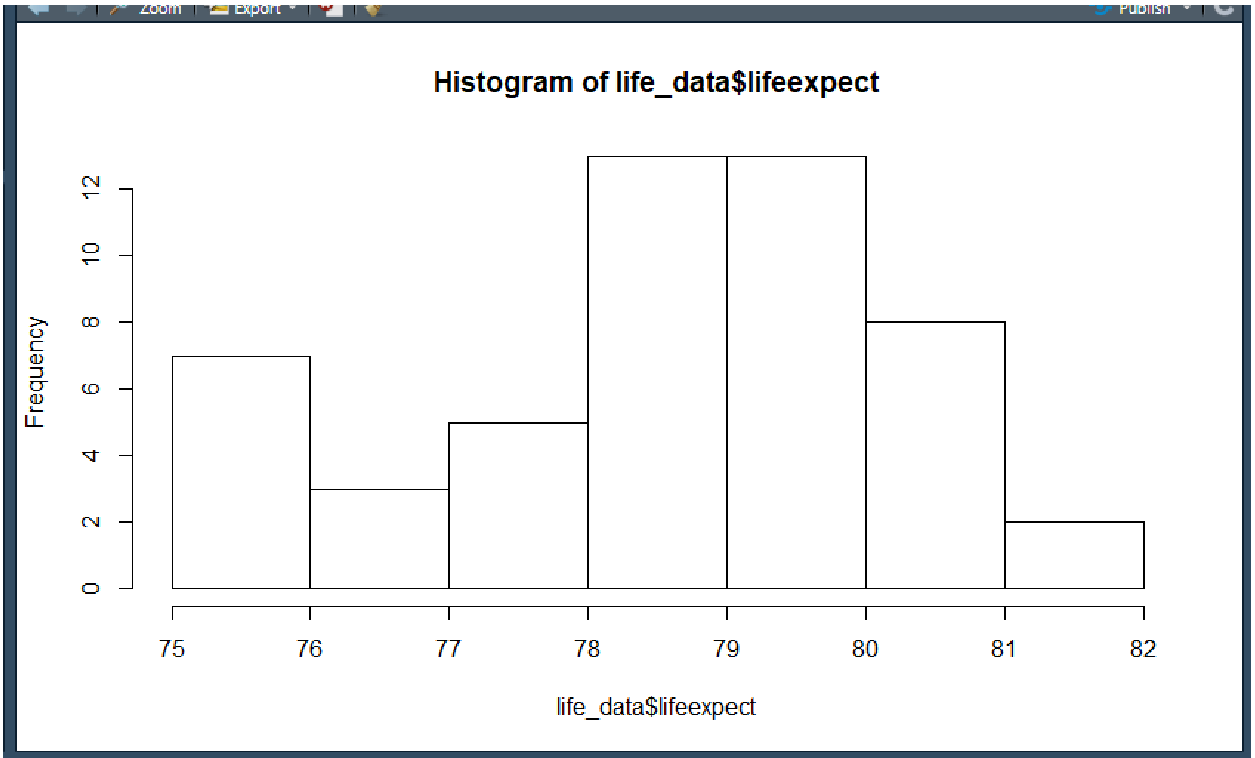
Douglas Stigler GEB 6895

Assignment 3 Due 10.09.2019

## Question 1:

**A)**




11:08 PM

```
> # Inspect the data.
> summary(life_data)
        state      lifeexpect        medinc        uninsured         smoke          obesity        teenbirth    
 Alabama   : 1   Min.   :75.00   Min.   :38.16   Min.   : 5.00   Min.   : 9.40   Min.   :21.00   Min.   :15.70  
 Alaska    : 1   1st Qu.:77.70   1st Qu.:44.63   1st Qu.:11.70   1st Qu.:16.45   1st Qu.:26.00   1st Qu.:27.35  
 Arizona   : 1   Median :78.80   Median :49.02   Median :13.80   Median :18.10   Median :27.00   Median :33.00  
 Arkansas  : 1   Mean   :78.62   Mean   :50.02   Mean   :14.25   Mean   :18.49   Mean   :27.65   Mean   :34.23  
 California: 1   3rd Qu.:79.90   3rd Qu.:55.57   3rd Qu.:16.65   3rd Qu.:20.55   3rd Qu.:30.00   3rd Qu.:41.65  
 Colorado  : 1   Max.   :81.30   Max.   :66.63   Max.   :24.80   Max.   :26.00   Max.   :35.00   Max.   :55.00  
 (Other)   :45                                                                                                  
     gunlaw           metro       
 Min.   :0.0000   Min.   : 29.66  
 1st Qu.:0.0000   1st Qu.: 63.20  
 Median :0.0000   Median : 74.88  
 Mean   :0.4118   Mean   : 74.01  
 3rd Qu.:1.0000   3rd Qu.: 87.78  
 Max.   :1.0000   Max.   :100.00  
```

| | state | lifeexpect | medinc | uninsured | smoke | obesity | teenbirth | gunlaw | metro |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Alabama | 75.4 | 40.933 | 14.4 | 21.9 | 33 | 43.6 | 0 | 71.46 |
| 2 | Alaska | 78.3 | 57.848 | 18.3 | 20.8 | 27 | 38.3 | 0 | 67.36 |
| 3 | Arizona | 79.6 | 46.896 | 19.1 | 16.6 | 26 | 41.9 | 0 | 92.53 |
| 4 | Arkansas | 76.0 | 38.587 | 18.5 | 22.4 | 31 | 52.5 | 0 | 60.27 |
| 5 | California | 80.8 | 54.283 | 18.9 | 12.9 | 25 | 31.5 | 1 | 97.73 |
| 6 | Colorado | 80.0 | 60.233 | 14.3 | 16.9 | 21 | 33.4 | 0 | 86.33 |
| 7 | Connecticut | 80.8 | 65.998 | 10.5 | 14.9 | 22 | 18.7 | 1 | 91.37 |
| 8 | Delaware | 78.4 | 55.214 | 11.7 | 18.0 | 28 | 30.5 | 1 | 78.04 |
| 9 | D.C. | 76.5 | 56.928 | 11.4 | 15.7 | 22 | 45.4 | 0 | 100.00 |

There does not appear to be any issues with the data for the purpose of building a linear regression model.

**B)**

```
Call:
lm(formula = lifeexpect ~ medinc + uninsured + smoke + obesity +
    teenbirth + gunlaw + metro, data = life_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7711 -0.3769 -0.1080  0.4822  1.3171

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 89.735528   2.251922  39.848  < 2e-16 ***
medinc      -0.010854   0.022245  -0.488 0.628090
uninsured    0.045937   0.036861   1.246 0.219422
smoke       -0.221999   0.050253  -4.418 6.64e-05 ***
obesity     -0.126588   0.050311  -2.516 0.015679 *
teenbirth   -0.078177   0.018433  -4.241 0.000116 ***
gunlaw       0.484511   0.250156   1.937 0.059353 .
metro       -0.015507   0.006564  -2.363 0.022747 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6722 on 43 degrees of freedom
Multiple R-squared:  0.8602,    Adjusted R-squared:  0.8375
F-statistic: 37.81 on 7 and 43 DF,  p-value: 2.315e-16
```

The variables that appear to have explanitory power are smoke, obesity, teenbirth, and metro because of the values of significance within 5%. The other variables (medinc, unisured, and gunlaw) would be good canadiates for omissin in the model.

**C)**

```
####################################################
# Removing one variable at a time from full model
####################################################

#Model without medinc variables
life_a_model <- lm(data = life_data,
                   formula = lifeexpect ~ uninsured + smoke + obesity + teenbirth + gunlaw + metro)
#Summarize a model
summary(life_a_model)

####################

#Model without uninsured variables
life_b_model <- lm(data = life_data,
                   formula = lifeexpect ~ medinc + smoke + obesity + teenbirth + gunlaw + metro)

#Summarize b model
summary(life_b_model)

####################


#Model without gunlaw variables
life_c_model <- lm(data = life_data,
                   formula = lifeexpect ~ medinc + uninsured + smoke + obesity + teenbirth  + metro)

#Summarize c model
summary(life_c_model)

####################
```

Removing one variable at a time all have different inpacts in the adjusted R^2, dpeending on which variable. Medinc being removed moves the adjusted R^2 closer to 1, while removing uninsured and gunlaw each move adjusted R^2 closer to 0.

**D)**

```
####################################################
# Removing variables from full model
####################################################

#Model with -1 variables
life_d_model <- lm(data = life_data,
                       formula = lifeexpect ~ uninsured + smoke + obesity + teenbirth + gunlaw + metro)
#Summarize -1 model
summary(life_d_model)

####################

#Model with -2 variables
life_e_model <- lm(data = life_data,
                   formula = lifeexpect ~ smoke + obesity + teenbirth + gunlaw + metro)

#Summarize -2 model
summary(life_e_model)

####################


#Model with -3 variables
life_f_model <- lm(data = life_data,
                   formula = lifeexpect ~ smoke + obesity + teenbirth  + metro)

#Summarize -3 model
summary(life_f_model)

####################


#Model with -4 variables
life_g_model <- lm(data = life_data,
                   formula = lifeexpect ~ smoke + obesity + teenbirth)

#Summarize -4 model
summary(life_g_model)

####################
```

Removing one non-significant to 5% variable at time causes the metro variable to become non-significant to 5% during the process. After removing the non-significant variables all estimates were like the original model with reduced std. deviations in some cases.

Original:

```
Call:
lm(formula = lifeexpect ~ medinc + uninsured + smoke + obesity +
    teenbirth + gunlaw + metro, data = life_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7711 -0.3769 -0.1080  0.4822  1.3171

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 89.735528   2.251922  39.848  < 2e-16 ***
medinc      -0.010854   0.022245  -0.488 0.628090
uninsured    0.045937   0.036861   1.246 0.219422
smoke       -0.221999   0.050253  -4.418 6.64e-05 ***
obesity     -0.126588   0.050311  -2.516 0.015679 *
teenbirth   -0.078177   0.018433  -4.241 0.000116 ***
gunlaw       0.484511   0.250156   1.937 0.059353 .
metro       -0.015507   0.006564  -2.363 0.022747 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6722 on 43 degrees of freedom
Multiple R-squared:  0.8602,    Adjusted R-squared:  0.8375
F-statistic: 37.81 on 7 and 43 DF,  p-value: 2.315e-16

>
```

With variables removed:

```
Call:
lm(formula = lifeexpect ~ smoke + obesity + teenbirth, data = life_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5115 -0.3408 -0.0091  0.3743  1.3860 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 87.95101    0.89483  98.287  < 2e-16 ***
smoke       -0.20643    0.04932  -4.186 0.000124 ***
obesity     -0.11023    0.04855  -2.271 0.027802 *  
teenbirth   -0.07210    0.01279  -5.638 9.47e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7051 on 47 degrees of freedom
Multiple R-squared:  0.8319,	Adjusted R-squared:  0.8212 
F-statistic: 77.52 on 3 and 47 DF,  p-value: < 2.2e-16

>
```

**E)** Using another approach by adding variables one at a time, I concluded that the full model would be the best. This is because the adjusted R^2 was the greatest in that model compared to my others. This was likely due to the way I chose to add variables and I could have arrived at a different outcome had I added the variables in a different order. Comparing the full model recommendation derived from adding one variable at a time versus the variable removal method, I choose to base my recommendation off the adjusted R^2 value, so I would recommend the model with only the medinc variable removed. If you were choosing the recommendation off T values, P values, or only having variables within a 5% level of significance than a different model may be recommended.

Recommended:

```
Call:
lm(formula = lifeexpect ~ uninsured + smoke + obesity + teenbirth +
    gunlaw + metro, data = life_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.78062 -0.42901 -0.06467  0.45527  1.30810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.78329    1.11369  79.720  < 2e-16 ***
uninsured    0.04979    0.03569   1.395 0.170069
smoke       -0.21763    0.04902  -4.440 5.98e-05 ***
obesity     -0.11707    0.04597  -2.547 0.014453 *
teenbirth   -0.07676    0.01804  -4.254 0.000108 ***
gunlaw       0.46849    0.24583   1.906 0.063235 .
metro       -0.01593    0.00645  -2.470 0.017477 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6663 on 44 degrees of freedom
Multiple R-squared:  0.8595,    Adjusted R-squared:  0.8403
F-statistic: 44.85 on 6 and 44 DF,  p-value: < 2.2e-16
```

Full Model:

```
Call:
lm(formula = lifeexpect ~ medinc + uninsured + smoke + obesity +
    teenbirth + gunlaw + metro, data = life_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7711 -0.3769 -0.1080  0.4822  1.3171

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 89.735528   2.251922  39.848  < 2e-16 ***
medinc      -0.010854   0.022245  -0.488 0.628090
uninsured    0.045937   0.036861   1.246 0.219422
smoke       -0.221999   0.050253  -4.418 6.64e-05 ***
obesity     -0.126588   0.050311  -2.516 0.015679 *
teenbirth   -0.078177   0.018433  -4.241 0.000116 ***
gunlaw       0.484511   0.250156   1.937 0.059353 .
metro       -0.015507   0.006564  -2.363 0.022747 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6722 on 43 degrees of freedom
Multiple R-squared:  0.8602,    Adjusted R-squared:  0.8375
F-statistic: 37.81 on 7 and 43 DF,  p-value: 2.315e-16
```

Only significant variables:

```
Call:
lm(formula = lifeexpect ~ smoke + obesity + teenbirth, data = life_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5115 -0.3408 -0.0091  0.3743  1.3860 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 87.95101    0.89483  98.287  < 2e-16 ***
smoke       -0.20643    0.04932  -4.186 0.000124 ***
obesity     -0.11023    0.04855  -2.271 0.027802 *  
teenbirth   -0.07210    0.01279  -5.638 9.47e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7051 on 47 degrees of freedom
Multiple R-squared:  0.8319,	Adjusted R-squared:  0.8212 
F-statistic: 77.52 on 3 and 47 DF,  p-value: < 2.2e-16

>
```

## Question 2:

**A)** There do not appear to be any problems with this data. The dependent variable is binary representing each of the hospitals with a 1 or 0.

```
> # Inspect the data.
> summary(hospital_data)
      OBS              D             DISTANCE           INCOME            OLD
 Min.   :  1.0   Min.   :0.0000   Min.   :-3.746   Min.   :41.12   Min.   :0.0000
 1st Qu.:125.5   1st Qu.:0.0000   1st Qu.:-3.412   1st Qu.:43.40   1st Qu.:1.0000
 Median :250.0   Median :1.0000   Median :-1.970   Median :44.39   Median :1.0000
 Mean   :250.0   Mean   :0.7295   Mean   :-1.011   Mean   :45.71   Mean   :0.8377
 3rd Qu.:374.5   3rd Qu.:1.0000   3rd Qu.: 1.570   3rd Qu.:47.93   3rd Qu.:1.0000
 Max.   :499.0   Max.   :1.0000   Max.   : 3.765   Max.   :55.17   Max.   :1.0000
>
```

| | OBS | D | DISTANCE | INCOME | OLD |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 3.00000 | 42.24318 | 0 |
| 2 | 2 | 0 | -3.74610 | 43.41784 | 1 |
| 3 | 3 | 0 | -3.74610 | 43.41784 | 1 |
| 4 | 4 | 0 | -3.73690 | 43.82044 | 0 |
| 5 | 5 | 1 | -3.73690 | 43.43426 | 1 |
| 6 | 6 | 1 | -3.71620 | 42.68955 | 1 |
| 7 | 7 | 1 | -3.70920 | 44.05817 | 1 |
| 8 | 8 | 1 | -3.70920 | 44.27197 | 0 |
| 9 | 9 | 1 | -3.70920 | 44.05817 | 1 |
| 10 | 10 | 1 | -3.70920 | 44.05817 | 1 |
| 11 | 11 | 1 | -3.70920 | 44.05817 | 1 |
| 12 | 12 | 1 | -3.70920 | 44.05817 | 1 |

**B)** After building the initial linear model, it would appear that DISTANCE is the only significant variable within 5%.

```
##################################################
# Full Linear Model
##################################################

#Model with all variables
life_linear_model <- lm(data = hospital_data,
                        formula = D ~ DISTANCE + INCOME + OLD)

#Summarize full model
summary(life_linear_model)
```

```
Call:
lm(formula = D ~ DISTANCE + INCOME + OLD, data = hospital_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.98891 -0.34933  0.07475  0.19036  0.66563 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.193452   0.297168   4.016 6.84e-05 ***
DISTANCE    -0.071995   0.007601  -9.471  < 2e-16 ***
INCOME      -0.010807   0.006257  -1.727   0.0848 .  
OLD         -0.051046   0.048009  -1.063   0.2882    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3922 on 495 degrees of freedom
Multiple R-squared:  0.2267,	Adjusted R-squared:  0.222 
F-statistic: 48.36 on 3 and 495 DF,  p-value: < 2.2e-16
```

**C)** I would recommend the logistic model because the result is binary with 1 or 0 being the only option.

```
##################################################
# Full Logistic Model
##################################################

#Model with all variables
life_log_model <- glm(data = hospital_data,
                      formula = D ~ DISTANCE + INCOME + OLD)

#Summarize full model
summary(life_log_model)
```


```
Call:
lm(formula = D ~ DISTANCE + INCOME + OLD, data = hospital_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.98891 -0.34933  0.07475  0.19036  0.66563 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.193452   0.297168   4.016 6.84e-05 ***
DISTANCE    -0.071995   0.007601  -9.471  < 2e-16 ***
INCOME      -0.010807   0.006257  -1.727   0.0848 .  
OLD         -0.051046   0.048009  -1.063   0.2882    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3922 on 495 degrees of freedom
Multiple R-squared:  0.2267,	Adjusted R-squared:  0.222 
F-statistic: 48.36 on 3 and 495 DF,  p-value: < 2.2e-16
```

**D)** Distance does have the sign I would expect. Since it is calculated by subtracting the distance from one hospital from another distance can be negative. If the number is negative it would be closer to the hospital represented by zero versus if it was positive than the patient would be closer to the hospital represented by 1.

**E)** This variable is statiscally significant within 5%. It tells me that older people are more senestive to distance, while combining the variables appears to be more significant than either on their own.

```
####################################################
# Logistic Model #2
####################################################

#Model with all variables
life_log2_model <- glm(data = hospital_data,
                       formula = D ~ DISTANCE + INCOME + OLD + OLD*DISTANCE)

#Summarize full model
summary(life_log2_model)

####################################################
# End
```

```
> summary(life_log2_model)

Call:
glm(formula = D ~ DISTANCE + INCOME + OLD + OLD * DISTANCE, data = hospital_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.96745  -0.30391   0.05647   0.22165   0.71061

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.179614   0.294222   4.009 7.03e-05 ***
DISTANCE     -0.022671   0.016628  -1.363 0.173378
INCOME       -0.009937   0.006200  -1.603 0.109635
OLD          -0.087841   0.048799  -1.800 0.072459 .
DISTANCE:OLD -0.059310   0.017831  -3.326 0.000946 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1507826)

    Null deviance: 98.477  on 498  degrees of freedom
Residual deviance: 74.487  on 494  degrees of freedom
AIC: 479.01

Number of Fisher Scoring iterations: 2

>
> ####################################################
```