STAT 308 – In Class Exercise 2

For the problems in which calculations are needed, please include your `R` code with your answers, otherwise you would not be given full credit on the exam.

The `state.x77` data matrix is automatically available in base `R`. The dataset provides information on information from all 50 US states from the 1970s about the following categories:

- `Population`: population estimate as of July 1, 1975
- `Income`: per capita income (1974)
- `Illiteracy`: illiteracy (1970, percent of population)
- `Life Exp`: life expectancy in years (1969-71)
- `Murder`: murder and non-negligent manslaughter rate per 100,000 population (1976)
- `HS Grad`: percent high-school graduates (1970)
- `Frost`: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- `Area`: land area in square miles

1. Suppose we wish to know if state's life expectancy is related linearly to at least one of the other variables in the dataset.

a. State the least squares regression line with life expectancy as the response variable with all the other variables as explanatory variables.

```
state.x77 <- as.data.frame(state.x77)
mod <- lm(`Life Exp` ~ .,state.x77)
summary(mod)
```

```
##
## Call:
## lm(formula = `Life Exp` ~ ., data = state.x77)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
## Population   5.180e-05  2.919e-05   1.775   0.0832 .
## Income      -2.180e-05  2.444e-04  -0.089   0.9293
## Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
```

```
## Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## `HS Grad`    4.893e-02  2.332e-02   2.098   0.0420 *
## Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
## Area        -7.383e-08  1.668e-06  -0.044   0.9649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

$\hat{LifeExp} = 70.94 + 0.0000518Population - 0.0000218Income + 0.003382Illiteracy - 0.3011Murder + 0.0489HSGrad - 0.005735Frost - 7.383 \times 10^{-8}Area$

b. Report and interpret the r^2 value for the full linear model in context of the problem.

```
summary(mod)$r.squared
```

```
## [1] 0.7361563
```

73.62% of the variation in life expectancy is explained through a linear model with population, income, illiteracy, murder rate, high school graduation rate, days below freezing, and area as explanatory variables.

c. Perform a formal hypothesis test to answer if life expentancy is related linearly to at least one other variable in the dataset. Please report all the information needed to perform a hypothesis test. Assume $\alpha = 0.05$.

$$H_0 : \beta_1 = \cdots = \beta_7 = 0$$

$$H_a : \text{At least one } \beta \neq 0$$

```
summary(mod)
```

```
##
## Call:
## lm(formula = `Life Exp` ~ ., data = state.x77)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
## Population   5.180e-05  2.919e-05   1.775   0.0832 .
## Income      -2.180e-05  2.444e-04  -0.089   0.9293
## Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
## Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## `HS Grad`    4.893e-02  2.332e-02   2.098   0.0420 *
## Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
## Area        -7.383e-08  1.668e-06  -0.044   0.9649
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

$F = 16.74$, $p-value = 2.534e-10$. $p-value < \alpha$, reject H_0 and can conclude that at least one of population, income, illiteracy, murder rate, high school graduation rate, days below freezing, and area is linearly related to life expectancy.

```
mod2 <- lm(`Life Exp` ~ 1,state.x77)
anova(mod2,mod)
```

```
## Analysis of Variance Table
##
## Model 1: `Life Exp` ~ 1
## Model 2: `Life Exp` ~ Population + Income + Illiteracy + Murder + `HS Grad` +
##     Frost + Area
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     49 88.299
## 2     42 23.297  7    65.002 16.741 2.534e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F = \frac{65.002/7}{23.297/42} = \frac{MSM}{MSE} = \frac{\text{Explained Variance}}{\text{Unexplained Variance}}$ and under H_0 F follows an F distribution with 7 and 42 degrees of freedom.

2. Suppose we know that murder rate is related linearly to life expectancy. Perform a hypothesis test to determine if adding illiteracy rate to our linear model for life expectancy with murder rate included significantly improves the predictive ability of our model. Assume $\alpha = 0.05$.

Reduced Model: $\hat{LifeExp} = \beta_0 + \beta_1 Murder$

Full Model: $\hat{LifeExp} = \beta_0 + \beta_1 Murder + \beta_2 Illiteracy$

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

```
mod_red <- lm(`Life Exp` ~ Murder,state.x77)
mod_full <- lm(`Life Exp` ~ Murder + Illiteracy,state.x77)
anova(mod_red,mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: `Life Exp` ~ Murder
## Model 2: `Life Exp` ~ Murder + Illiteracy
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     48 34.461
## 2     47 34.188  1   0.27323 0.3756 0.5429
```

```
summary(mod_full)
```

```
##
## Call:
## lm(formula = `Life Exp` ~ Murder + Illiteracy, data = state.x77)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80433 -0.47593  0.06604  0.42339  2.53621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.02758    0.28568 255.623  < 2e-16 ***
## Murder      -0.26395    0.04641  -5.688 7.96e-07 ***
## Illiteracy  -0.17225    0.28106  -0.613    0.543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8529 on 47 degrees of freedom
## Multiple R-squared:  0.6128, Adjusted R-squared:  0.5963
## F-statistic: 37.19 on 2 and 47 DF,  p-value: 2.071e-10
```

$F = 0.3756$ or $t = -0.613$, $p - value = 0.5429$. $p - value \geq \alpha$, so we fail to reject H_0 and can conclude that adding illteracy to our linear model which already includes murder rate does not signficantly improve the predictive ability of life expectancy.

3. Suppose we know that murder rate is related linearly to life expectancy. Perform a hypothesis test to determine if adding high school graduation rate to our linear model for life expectancy with murder rate included significantly improves the predictive ability of our model. Assume $\alpha = 0.05$.

Reduced Model: $\hat{LifeExp} = \beta_0 + \beta_1 Murder$

Full Model: $\hat{LifeExp} = \beta_0 + \beta_1 Murder + \beta_2 HSGrad$

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

```
mod_red <- lm(`Life Exp` ~ Murder,state.x77)
mod_full <- lm(`Life Exp` ~ Murder + `HS Grad`,state.x77)
anova(mod_red,mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: `Life Exp` ~ Murder
## Model 2: `Life Exp` ~ Murder + `HS Grad`
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     48 34.461
## 2     47 29.770  1     4.691 7.4059 0.009088 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod_full)
```

```
##
## Call:
## lm(formula = `Life Exp` ~ Murder + `HS Grad`, data = state.x77)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66758 -0.41801  0.05602  0.55913  2.05625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.29708    1.01567  69.213  < 2e-16 ***
## Murder      -0.23709    0.03529  -6.719 2.18e-08 ***
## `HS Grad`    0.04389    0.01613   2.721  0.00909 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7959 on 47 degrees of freedom
## Multiple R-squared:  0.6628, Adjusted R-squared:  0.6485
## F-statistic:  46.2 on 2 and 47 DF,  p-value: 8.016e-12
```

$F = 7.406$ or $t = 2.721$, $p - value = 0.00909$. $p - value < \alpha$, so we reject H_0 and can conclude that adding high school graduation rate to our linear model that already includes murder rate significant improves the predictive ability of life expectancy.

4. Perform a hypothesis test to determine if adding all the additional variables to our linear model for life expectancy with both murder rate and high school graduation rate already included significantly improves the predictive ability of our model. Assume $\alpha = 0.05$.

Reduced Model: $\hat{LifeExp} = \beta_0 + \beta_1 Murder + \beta_2 HSGrad$

Full Model: $\hat{LifeExp} = \beta_0 + \beta_1 Murder + \beta_2 HSGrad + \beta_3 Illiteracy + \beta_4 Population + \beta_5 Income + \beta_6 Frost + \beta_7 Area$

$$H_0 : \beta_3 = \cdots = \beta_7 = 0$$

$$H_a : \text{At least one } \beta \neq 0$$

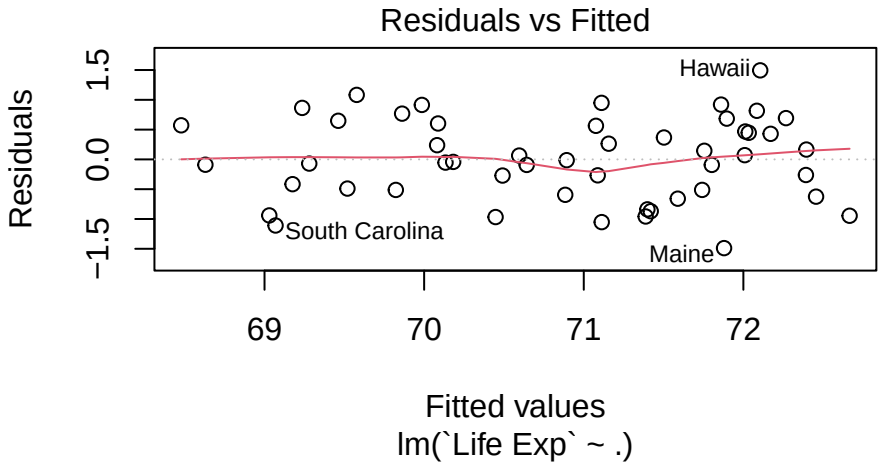
```
mod_red <- lm(`Life Exp` ~ Murder + `HS Grad`,state.x77)
mod_full <- lm(`Life Exp` ~ .,state.x77)
anova(mod_red,mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: `Life Exp` ~ Murder + `HS Grad`
## Model 2: `Life Exp` ~ Population + Income + Illiteracy + Murder + `HS Grad` +
##     Frost + Area
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     47 29.770
## 2     42 23.297  5    6.4732 2.334 0.05871 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

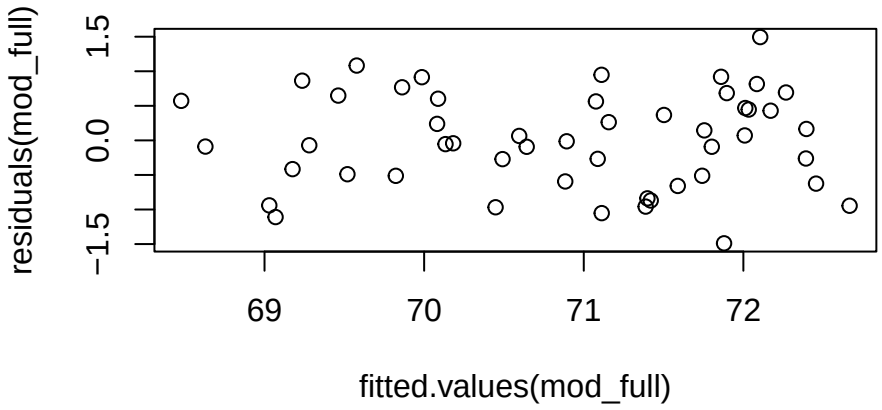
$F = 2.334$, $p - value = 0.0587$. $p - value > \alpha$, so we fail to reject H_0 and can conclude that adding all the other explanatory variables to a linear model that includes murder rate and high school graduation rate does not significantly improve to predictive ability for life expectancy.

5. Determine if the assumptions for homoscedasticity and normally distributed residuals are violated for the full linear model in (1). (Hint: the method of checking these assumptions is the same for multiple linear regression as in simple linear regression.)

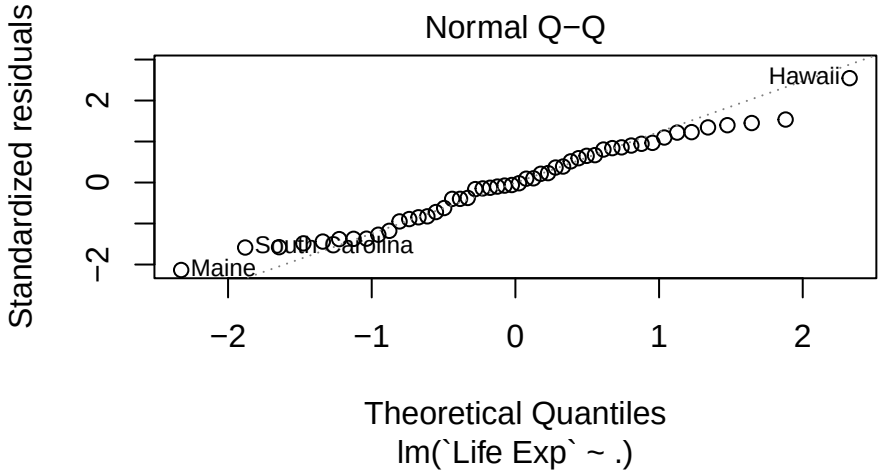
```
plot(mod_full,1)
```



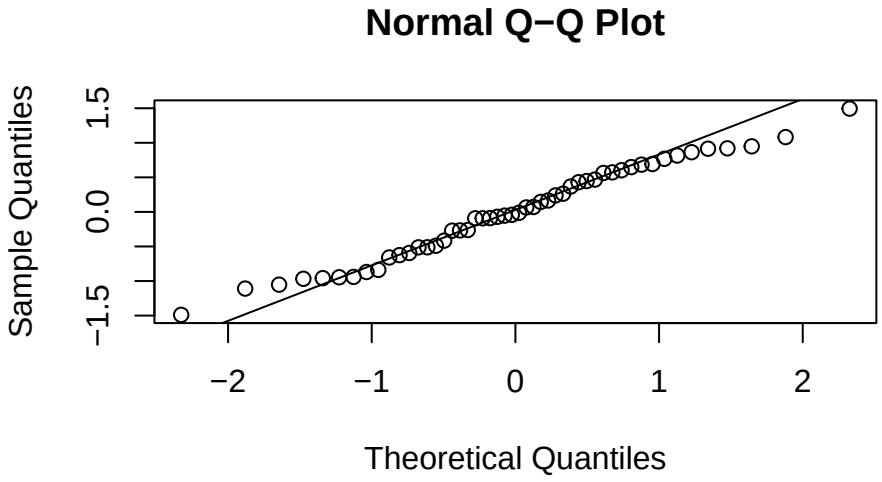

```
plot(fitted.values(mod_full),residuals(mod_full))
```



```
plot(mod_full,2)
```




```
qqnorm(residuals(mod_full))
qqline(residuals(mod_full))
```



The residuals plots do not seem to have any discernable pattern, and they appear to be spread evenly across all fitted values and are centered at zero. The QQ plots seem to deviate a little bit from the 45 degree line in the tails, but the majority of the points seem to fit directly on the line. Therefore, the assumptions of homoscedasticity and normality do not appear to be violated.