Rita Qifan Yang
McGill ID: 260893989
qifan.yang@mail.mcgill.ca
PSYC 560 Assignment 1

[Q1]

1.1 Apply MLR

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2.080496   0.138731  14.997  < 2e-16 ***
genderfemale                 -0.008398   0.077691  -0.108 0.913948
age                           0.012834   0.003013   4.259 2.25e-05 ***
ses                          -0.224968   0.076184  -2.953 0.003223 **
maritalmarried               -0.246915   0.133551  -1.849 0.064784 .
familystep or foster families -0.334680   0.088767  -3.770 0.000173 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.162 on 971 degrees of freedom
Multiple R-squared:  0.07338,   Adjusted R-squared:  0.06861
F-statistic: 15.38 on 5 and 971 DF,  p-value: 1.422e-14
```

1.1.1 R-squared: The multiple R_2 is 0.07338, the adjusted R_2 is 0.06861. The coefficient of determination (R_2) indicates the proportion of the total variation in Y accounted for by the regression model. The larger it is, the more the variance of Y is explained, with 0 indicating no explanation and 1 indicating perfect explanation. However, since R_2 never decreases when new predictors are added to the model, here we will access the adjusted R-squared value. In this case, 6.9% of the variation in cigarette consumption is explained by the effects of gender, age, SES(average of parental income and education level), marital, and family.

1.1.2 Coefficient estimates and their statistical significance: The coefficient estimate of the intercept is 1.411136, which means that on average, the participants from American northwest urban areas are 6-month to lifetime abstainers.

Gender: The coefficient estimate of the female dummy variable (1= female, 0 = otherwise) is -0.008398, which means that cigarette consumption decreases by 0.008398 on average for each when the gender is female compared to male. The p-value for gender is 0.913948. Since this value is greater than .05, gender does not have a statistically significant relationship with cigarette consumption.

Age: The coefficient estimate of age is 0.012834, which means that cigarette consumption increases by 0.012834 on average for each one unit increase in age. The p-value for age is 2.25e-05. Since this value is less than .05, age has a statistically significant relationship with cigarette consumption.

SES: The coefficient estimate of SES(average parental income and education level) is -0.224968, which means that cigarette consumption decreases by 0.224968 on average for each one unit increase in average parental income and education level. The p-value for the SES is 0.003223. Since this value is less than .05, parental income and education level has a statistically significant relationship with cigarette consumption.

Marital: The coefficient estimate of married dummy variable (1=married or living in a committed relationship, 0 = otherwise) is -0.246915, which means that cigarette consumption decreases by 0.246915 on average when married/living in a committed relationship compared to single. The p-value for marital status is 0.064784. Since this value is greater than .05, marital status does not have a statistically significant relationship with cigarette consumption.

Family: The coefficient estimate of step/foster family dummy variable (1 = step or foster family, 0 = otherwise) is -0.334680, which means that cigarette consumption decreases by 0.334680 on average when the participant has a step or foster family as compared to others. The p-value for age is 0.000173. Since this value is less than .05, whether or not the participant lives in a step or foster family has a statistically significant relationship with cigarette consumption.

1.2 Add quadratic term of age to (1)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6298394  0.2395539   2.629 0.008694 **
gender      -0.0565649  0.0779613  -0.726 0.468289
age          0.0945630  0.0201864   4.684 3.21e-06 ***
I(age^2)    -0.0014960  0.0003654  -4.094 4.60e-05 ***
ses         -0.2025341  0.0757719  -2.673 0.007645 **
marital     -0.2777004  0.1326934  -2.093 0.036627 *
family       0.2997649  0.0884678   3.388 0.000731 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.153 on 970 degrees of freedom
Multiple R-squared:  0.08912,   Adjusted R-squared:  0.08348
F-statistic: 15.82 on 6 and 970 DF,  p-value: < 2.2e-16
```

The quadratic term of age has a p-value <0.001, since this value is less than 0.05, the quadratic term of age has a statistically significant relationship with cigarette consumption.

1.3 KNN regression at K = 5, Calculate and report R-squared

The R-squared value of KNN regression at K = 5 is 0.2779032.

[Q2]

Below is the respective mean squared error using three different models to predict new 227 participants’ cigarette consumption:

MSE_knn: 1.486892 (KNN regression)

MSE_lm1: 1.451811 (Multiple Linear Regression)

MSE_lm2: 1.419193 (Polynomial regression with the quadratic term of age)

The smaller the MSE, the better the model performance. Therefore, the Polynomial regression model in (2) with the quadratic term of age performs the best.