Question 1

Part A

i)

The regression model equation is $Y_i = \beta_0 + \beta_1 X_i + e_i$, where:

- $i = 1, \ldots, 527$
- Y_i is the total cholesterol of the i-th subject in mmol/L
- X_i is the age of the i-th subject in years
- β_0 is the intercept corresponding to the average of Y in mmol/L when X is 0 years
- β_1 is the slope or expected change in Y for a 1 year increase in X
- e_i is the random error with $\mathbb{E}(e_i) = 0$, $\text{Var}(e_i) = \sigma^2$, $\text{Cor}(e_i, e_j) = 0$, and $e_i \sim N(0, \sigma^2)$
- $\mathbb{E}(Y|X)$ is the expected value of Y for a given X.

ii)

According to this data, a 10 year increase in the subject's age is associated with a 0.262 mmol/L increase in their total cholesterol on average, with a 95% confidence interval of $[0.204, 0.321]$ and a p-value of $< 2 \cdot 10^{-16}$.

```
>
> Call:
> lm(formula = tc ~ age, data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.6589 -0.6383 -0.0557  0.5009  4.3216
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept) 4.376191   0.135848  32.214   <2e-16 ***
> age         0.026243   0.002966   8.849   <2e-16 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.9801 on 525 degrees of freedom
> Multiple R-squared:  0.1298,  Adjusted R-squared:  0.1281
> F-statistic: 78.31 on 1 and 525 DF,  p-value: < 2.2e-16

>                 2.5 %     97.5 %
> (Intercept) 4.1093176 4.64306405
> age         0.0204172 0.03206905
```

iii)

The coefficient between age and total cholesterol is 0.360.

```
> Pearson coefficient: 0.360273
```

iv)

The simple linear regression model has a very significant p-value for the association between age and total cholesterol, suggesting that we can be confident there is a relationship between the variables. Similarly, the Pearson correlation coefficient indicates a positive correlation between the variables; however, it is somewhat low, suggesting that the strength of the linear relationship is not particularly strong. This indicates that outliers could be present or that simple linear regression might be insufficient for modeling the relationship between the variables.

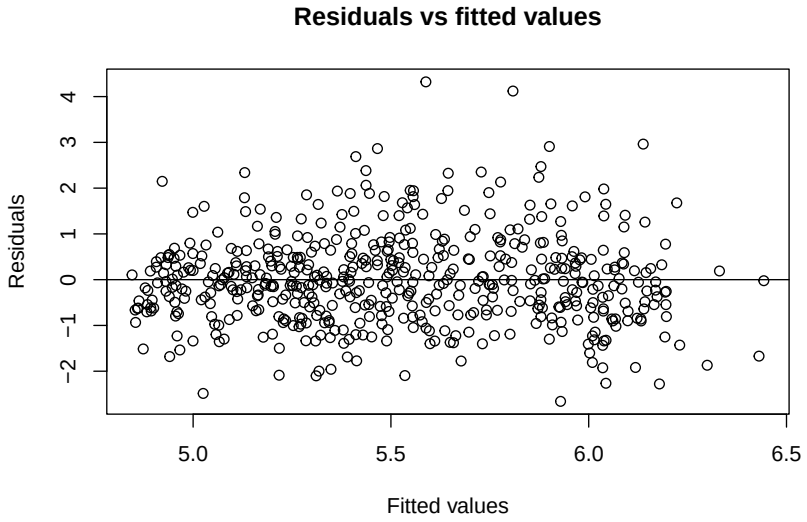
Part B

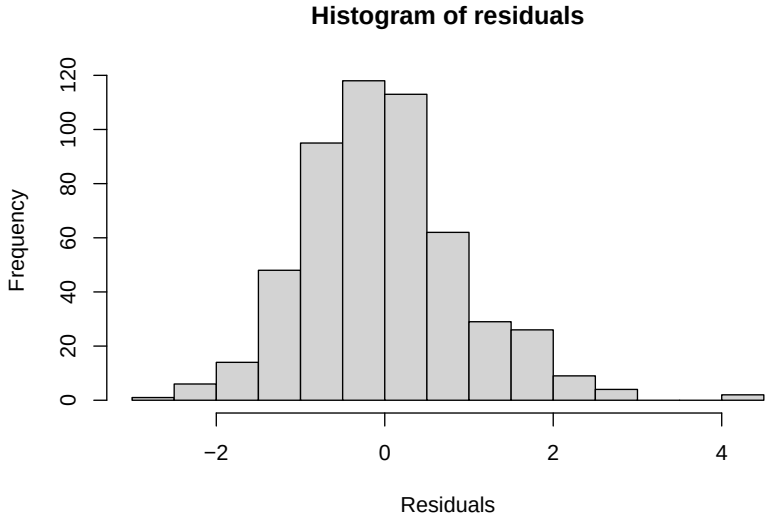
ii)

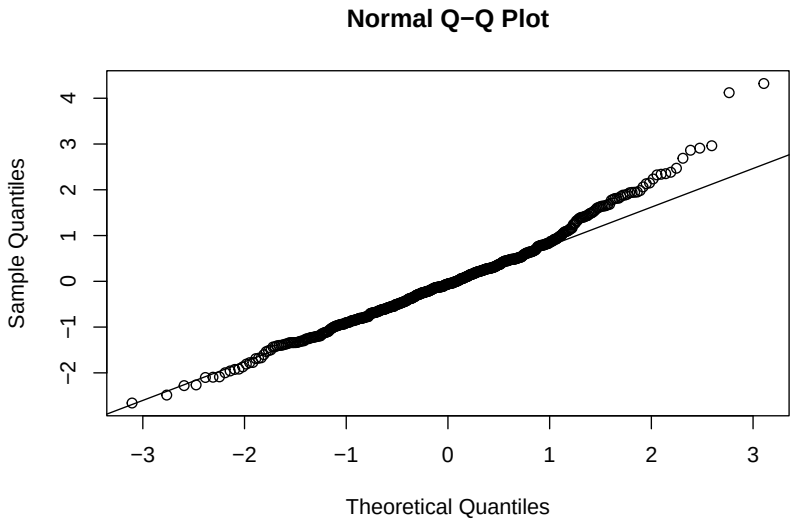
In the scatterplot of the residuals vs fitted values, there does not seem to be any pattern in the residuals, which appear homoscedastic. In the histogram, the residuals appear to be mostly normally distributed with a slight right skew. In the QQ plot, the residuals also look normally distributed since they mostly follow the diagonal straight line with a slight deviation on the right side of the plot.

iii)

In the histogram, there appears to be two outliers with residuals greater than 4. This can also be seen in the scatterplot and the QQ plot, the latter of which shows the outliers as the two rightmost points.

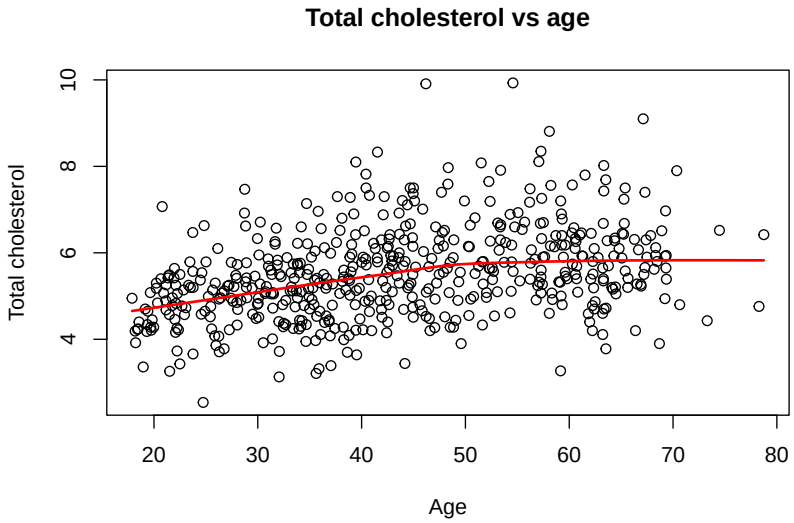






Part C

i)



ii)

```
>
> Call:
> lm(formula = tc ~ age + I(age^2), data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.6542 -0.6410 -0.0461  0.5151  4.1698
>
> Coefficients:
>               Estimate Std. Error t value Pr(>|t|)
> (Intercept)  3.0658041  0.3907005   7.847 2.41e-14 ***
> age          0.0920305  0.0186508   4.934 1.08e-06 ***
```

```
> I(age^2)    -0.0007389  0.0002069  -3.572 0.000387 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.9693 on 524 degrees of freedom
> Multiple R-squared:  0.1505,  Adjusted R-squared:  0.1472
> F-statistic: 46.41 on 2 and 524 DF,  p-value: < 2.2e-16
```

iii)

The fitted model equation is $\hat{Y}_i = 3.066 + 0.092X_i - 0.0007X_i^2$, where $i = 1, \ldots, 527$, $\hat{Y}_i$ is the estimated total cholesterol of the i-th subject in mmol/L, X_i is the age of the i-th subject in years, and X_i^2 is the squared age of the i-th subject in squared years.

iv)

Yes, I do see statistical evidence of a nonlinear relationship between age and total cholesterol according to this data. This is since the p-value for the coefficient of Age_i^2 is $0.0004 < 0.001$, which is statistically significant. This is consistent with the scatterplot of total cholesterol vs age, where the LOWESS curve is not a straight line.

Question 2

Part A

i)

The null hypothesis (or H_0) is $\mu_1 = \mu_2$ and the alternative hypothesis (or H_1) is $\mu_1 \neq \mu_2$, where μ_1 is the mean total cholesterol of the female subjects and μ_2 is the mean total cholesterol of the male subjects. The t-test yields a p-value of 0.523, which is not significant. Thus, there is insufficient evidence to conclude that the mean total cholesterol is statistically different between sexes.

```
>
>   Two Sample t-test
>
> data:  df$tc by df$gender
> t = -0.63874, df = 525, p-value = 0.5233
> alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
> 95 percent confidence interval:
>  -0.2392340  0.1218354
> sample estimates:
> mean in group 0 mean in group 1
>        5.490799        5.549498
```

ii)

The results for linear model and the equal variance t-test have the following in common:

- In the t-test results, the difference in the mean total cholesterol between the biological sexes is 0.0587, which is the same as the estimated slope for biological sex in the linear model.
- The absolute value of the t-statistic is 0.639 for both the t-test and the estimated slope for biological sex in the linear model. In addition, the degrees of freedom and p-value corresponding to this t statistic are 525 and 0.523 respectively for both the t-test and the linear model.
- In the t-test results, the 95% confidence interval is $[-0.239, 0.122]$, which is the same (but with opposite signs) as the 95% confidence interval for the estimated slope for biological sex in the linear model.

```
>
> Call:
> lm(formula = tc ~ gender, data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.9508 -0.6908 -0.0908  0.6242  4.4392
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  5.49080    0.06189  88.722   <2e-16 ***
> gender       0.05870    0.09190   0.639    0.523
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 1.05 on 525 degrees of freedom
> Multiple R-squared:  0.0007765,   Adjusted R-squared:  -0.001127
> F-statistic: 0.408 on 1 and 525 DF,  p-value: 0.5233
```

```
>                 2.5 %   97.5 %
> (Intercept)  5.3692208 5.612376
> gender      -0.1218354 0.239234
```

Part B

i)

No, since biological sex is not associated with age. This violates one of the conditions for the classical definition of a confounder.

ii)

The fitted regression model equation is $\hat{Y}_i = 3.039 + 0.091X_{1i} - 0.0007X_{1i}^2 + 0.094X_{2i}$, where $i = 1, \ldots, 527$, $\hat{Y}_i$ is the estimated total cholesterol of the i-th subject in mmol/L, X_{1i} is the age of the i-th subject in years, X_{1i}^2 is the squared age of the i-th subject in squared years, and X_{2i} is the biological sex of the i-th subject.

```
>
> Call:
> lm(formula = tc ~ age + I(age^2) + gender, data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.6152 -0.6114 -0.0555  0.5250  4.1205
>
> Coefficients:
>               Estimate Std. Error t value Pr(>|t|)
> (Intercept)  3.0390448  0.3913591   7.765 4.32e-14 ***
> age          0.0909455  0.0186723   4.871 1.48e-06 ***
> I(age^2)    -0.0007241  0.0002073  -3.494 0.000517 ***
> gender       0.0944912  0.0851936   1.109 0.267882
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.9691 on 523 degrees of freedom
> Multiple R-squared:  0.1525,  Adjusted R-squared:  0.1476
> F-statistic: 31.36 on 3 and 523 DF,  p-value: < 2.2e-16
```

iii)

The estimated slope for age is 0.091 in the model with biological sex added and 0.092 in the model without biological sex added, so the change between the estimated slope is $\frac{0.092-0.091}{0.091} \times 100 = 1.19\%$. The estimated slope for squared age is -0.00072 in the model with biological sex added and -0.00074 in the model without biological sex added, so the change between the estimated slope is $\frac{-0.00074+0.00072}{-0.00072} \times 100 = 2.04\%$. Thus, the estimated slopes for both age and squared age change less than 10% when biological sex is added to the model, meaning that biological sex does not meet the criteria for operational confounding.

iv)

Biological sex is not a confounder of the effect of linear and quadratic age on total cholesterol. This is since it does not meet either the classical nor the operational definitions of a confounder.

v)

No, since in part A ii), the equal variance t-test did not find a statistically significant difference in the mean total cholesterol between the biological sexes.

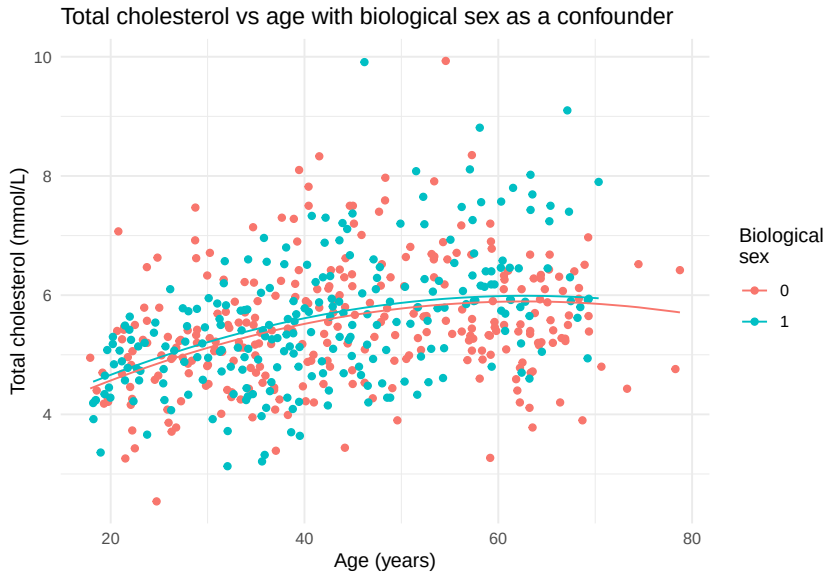
Part C

i)

For males, the fitted model equation is $\hat{Y}_i = 3.039 + 0.091X_i - 0.0007X_i^2$, where $i = 1, \ldots, 527$, $\hat{Y}_i$ is the estimated total cholesterol of the i-th subject in mmol/L, X_i is the age of the i-th subject in years, and X_i^2 is the squared age of the i-th subject in squared years. For females, the fitted model equation is $\hat{Y}_i = 3.134 + 0.091X_i - 0.0007X_i^2$ for i, $\hat{Y}_i$, X_i, and X_i^2 defined similarly.

ii)

I notice that the fitted values follow a downwards parabola, which makes sense given that the models have squared age as a covariate. In addition, the difference in the fitted values between the biological sexes is small, which is consistent with the t-test comparing total cholesterol by biological sex in part A i).



Part D

i)

The model equation is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i*3i} + \beta_5 X_{2i*3i} + e_i$, where:

- $i = 1, \ldots, 527$
- Y_i is the total cholesterol of the i-th subject in mmol/L
- X_{1i} is the age of the i-th subject in years
- X_{2i} is the squared age of the i-th subject in squared years
- X_{3i} is the biological sex of the i-th subject
- X_{1i*3i} is the interaction term between age and biological sex
- X_{2i*3i} is the interaction term between squared age and biological sex
- When the subject is male ($X_{3i} = 0$):
  - β_0 is the intercept corresponding to the average of Y in mmol/L when X_1 is 0 years
  - β_1 is the expected change in Y for a 1 year increase in X_1, holding all other covariates constant
  - β_2 is the expected change in Y for a 1 squared year increase in X_2, holding all other covariates constant
- When the subject is female ($X_{3i} = 1$):
  - $\beta_0 + \beta_3$ is the intercept corresponding to the average of Y in mmol/L when X_1 is 0 years
  - $\beta_1 + \beta_4$ is the expected change in Y for a 1 year increase in X_1, holding all other covariates constant
  - $\beta_2 + \beta_5$ is the expected change in Y for a 1 squared year increase in X_2, holding all other covariates constant
- e_i is the random error with $\mathbb{E}(e_i) = 0$, $\text{Var}(e_i) = \sigma^2$, $\text{Cor}(e_i, e_j) = 0$, and $e_i \sim N(0, \sigma^2)$.

Biological sex does appear to be an effect modifier of the effect of linear and quadratic age on total cholesterol, since the interaction terms Age$*$Gender and Age^{2*}Gender are statistically significant at the $\alpha = 0.05$ level.

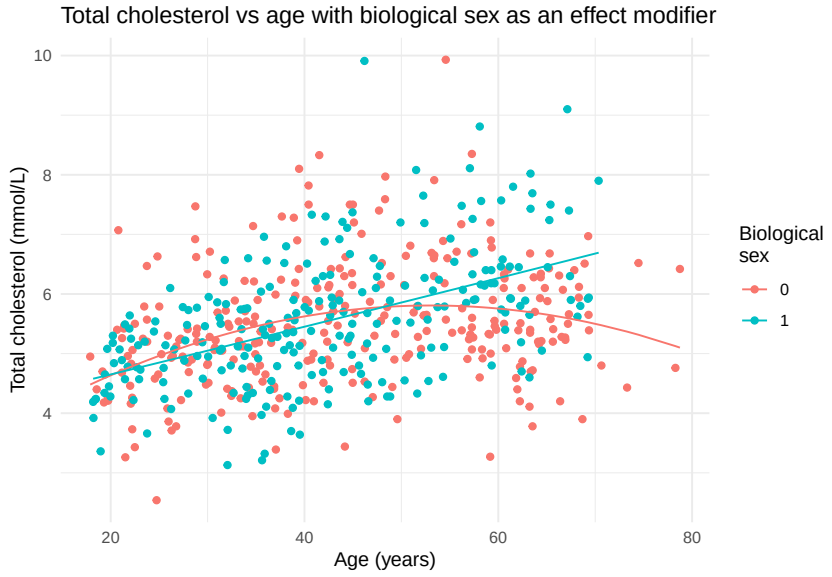
```
>
> Call:
> lm(formula = tc ~ age + I(age^2) + gender + age:gender + I(age^2):gender,
>     data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.4971 -0.6287 -0.0584  0.5450  4.2082
>
> Coefficients:
>                   Estimate Std. Error t value Pr(>|t|)
> (Intercept)      2.7876252  0.5259996   5.300 1.72e-07 ***
> age              0.1137966  0.0247662   4.595 5.44e-06 ***
> I(age^2)        -0.0010722  0.0002699  -3.973 8.10e-05 ***
> gender           1.0753560  0.7772306   1.384  0.16708
> age:gender      -0.0747839  0.0373402  -2.003  0.04572 *
> I(age^2):gender  0.0010894  0.0004176   2.609  0.00935 **
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.953 on 521 degrees of freedom
> Multiple R-squared:  0.1836,  Adjusted R-squared:  0.1757
> F-statistic: 23.43 on 5 and 521 DF,  p-value: < 2.2e-16
```

ii)

For males, the fitted model equation is $\hat{Y}_i = 2.788 + 0.114 X_i - 0.001 X_i^2$, where $i = 1, \ldots, 527$, $\hat{Y}_i$ is the estimated total cholesterol of the i-th subject in mmol/L, X_i is the age of the i-th subject in years, and X_i^2 is the squared age of the i-th subject in squared years. For females, the fitted model equation is $\hat{Y}_i = 3.863 - 0.039 X_i + (1.72 \cdot 10^{-5}) X_i^2$ for i, $\hat{Y}_i$, X_i, and X_i^2 defined similarly.

iii)

I notice that the lines of the fitted values are not parallel, which further suggests that biological sex is an effect modifier of the effect of linear and quadratic age on total cholesterol. Furthermore, the plot shows that after around 50 years of age, female subjects have more total cholesterol over time on average compared to male subjects.



Part E

The best model in terms of R^2, R^2_{adj}, and RMSE is the model with age, squared age, sex, and interaction terms as the covariates. This model has the highest R^2 (0.1836); while this could be because R^2 is non-decreasing as more covariates are added, the model also has the highest adjusted R^2 (0.1757), which has a penalty for any additional covariates. Finally, this model has the lowest RMSE (0.953), indicating that it has the lowest estimated standard deviation.

Part F

i)

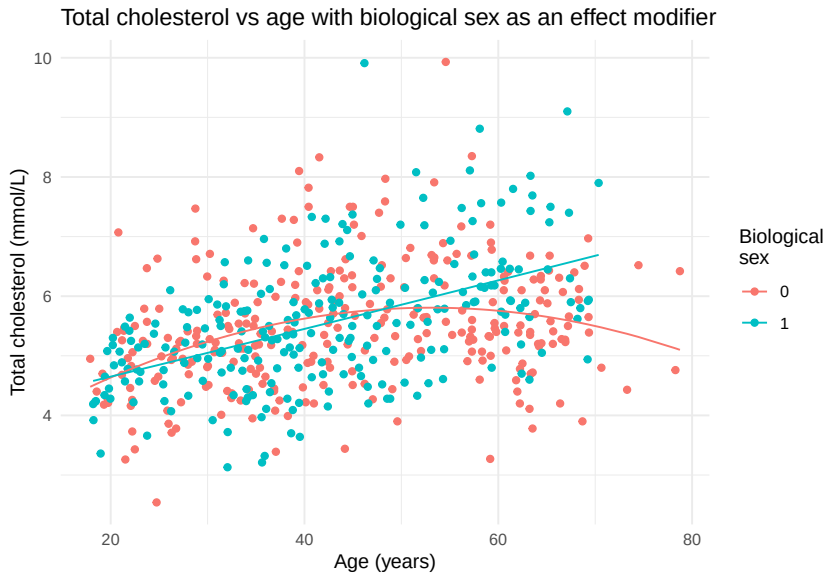
For males, the fitted model equation is $\hat{Y}_i = 2.788 + 0.114X_i - 0.001X_i^2$, where $i = 1, \ldots, 527$, $\hat{Y}_i$ is the estimated total cholesterol of the i-th subject in mmol/L, X_i is the age of the i-th subject in years, and X_i^2 is the squared age of the i-th subject in squared years. For females, the fitted model equation is $\hat{Y}_i = 3.863 - 0.039X_i + (1.72 \cdot 10^{-5})X_i^2$ for i, $\hat{Y}_i$, X_i, and X_i^2 defined similarly.

```
>
> Call:
> lm(formula = tc ~ age + I(age^2), data = df[df$gender == 0, ])
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.4971 -0.5673 -0.1069  0.5668  4.1255
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  2.787625   0.530065   5.259 2.85e-07 ***
> age          0.113797   0.024958   4.560 7.62e-06 ***
> I(age^2)    -0.001072   0.000272  -3.942 0.000102 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
>
> Residual standard error: 0.9604 on 285 degrees of freedom
> Multiple R-squared:  0.1077,  Adjusted R-squared:  0.1014
> F-statistic:  17.2 on 2 and 285 DF,  p-value: 8.861e-08


>
> Call:
> lm(formula = tc ~ age + I(age^2), data = df[df$gender == 1, ])
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.0646 -0.6323 -0.0462  0.5394  4.2082
>
> Coefficients:
>              Estimate Std. Error t value Pr(>|t|)
> (Intercept) 3.863e+00  5.668e-01   6.815 7.78e-11 ***
> age         3.901e-02  2.768e-02   1.409    0.160
> I(age^2)    1.723e-05  3.157e-04   0.055    0.957
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.944 on 236 degrees of freedom
> Multiple R-squared:  0.2608,  Adjusted R-squared:  0.2545
> F-statistic: 41.63 on 2 and 236 DF,  p-value: 3.266e-16
```

ii)



iii)

Both approaches yield the same fitted coefficients and scatterplots. In the first approach in part D, all the coefficients are contained in one model output, meaning the fitted values for either biological sex can be computed by looking at just that output. In the second approach in part F, the coefficients are contained in two separate model outputs, meaning the fitted values for females must be computed by looking at both outputs.

Another difference is that the model metrics (e.g., R^2, R^2_{adj}, F-statistic and its p-value) are calculated separately for each biological sex in the second approach but together in the first approach. In short, an advantage of the first approach is that it is compact and shows all coefficients in one model output, but with the disadvantage of losing information about the model metrics for each biological sex.

Question 3

Part A

i)

For the model with continuous BMI, the regression model equation is $Y_i = \beta_0 + \beta_1 X_i + e_i$, where:

- $i = 1, \ldots, 527$
- Y_i is the total cholesterol of the i-th subject in mmol/L
- X_i is the BMI of the i-th subject in kg/m^2
- β_0 is the intercept corresponding to the average of Y in mmol/L when X is 0 kg/m^2
- β_1 is the expected change in Y for a 1 kg/m^2 increase in X
- e_i is the random error with $E(e_i) = 0$, $\text{Var}(e_i) = \sigma^2$, $\text{Cor}(e_i, e_j) = 0$, and $e_i \sim N(0, \sigma^2)$.

For the model with categorical BMI, the regression model equation is $Y_i = \beta_0 + \beta_1 I_{1i} + \beta_2 I_{2i} + \beta_3 I_{3i} + e_i$, where:

- i, Y_i, and e_i are defined in the same way as above
- $j = 1, 2, 3$ for the BMI categories of underweight, overweight, and obese respectively
- $I_{ji} = 1$ if the i-th subject has the BMI category j and 0 otherwise
- β_0 is the intercept corresponding to the average of Y in mmol/L when the subject has the BMI category of normal weight
- β_j is the coefficient corresponding to the average change in total cholesterol in mmol/L if the subject has the BMI category j vs if the subject has the BMI category of normal weight, holding all other covariates constant.

The categorical BMI model has more coefficients than the other model. Moreover, there are only three possible changes in the total cholesterol for the categorical BMI model, whereas for the other model, there are infinitely many possible changes in the total cholesterol for a 1 kg/m^2 increase in the BMI. Note that the indicator variables do not overlap.

ii)

The categorical BMI model is the same as a one-way ANOVA model since both make comparisons between different groups of a variable.

iii)

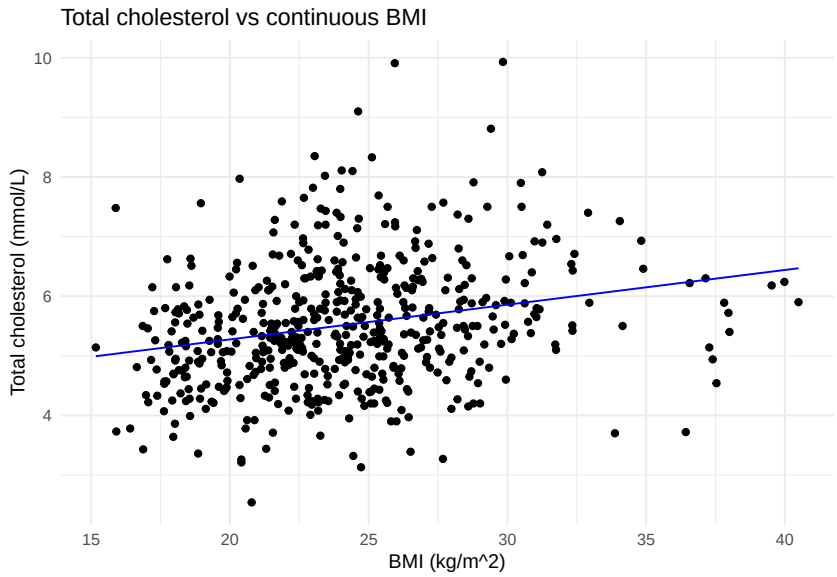
For the continuous BMI model, the fitted model equation is $\hat{Y}_i = 4.110 + 0.058 X_i$, where $i = 1, \ldots, 527$, $\hat{Y}_i$ is the estimated total cholesterol of the i-th subject in mmol/L, and X_i is the BMI of the i-th subject in kg/m^2. For the categorical BMI model, the fitted model equation is $\hat{Y}_i = 5.450 - 0.454 I_{1i} + 0.193 I_{2i} + 0.592 I_{3i}$ where i and $\hat{Y}_i$ is defined as previously, $I_{1i} = 1$ if the i-th subject is underweight, $I_{2i} = 1$ if the i-th subject is overweight, and $I_{3i} = 1$ if the i-th subject is obese.

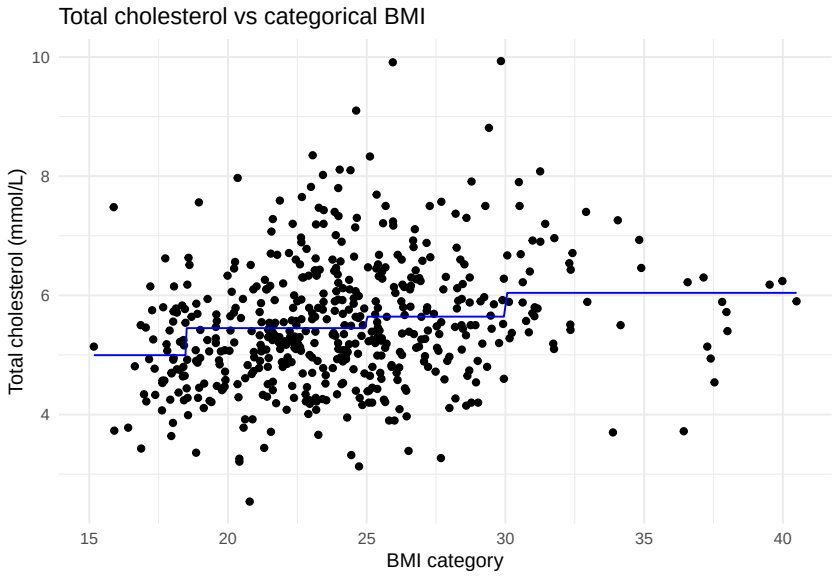
```
>
> Call:
> lm(formula = tc ~ bmi_cts, data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.7806 -0.6333 -0.1273  0.5735  4.2889
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)  4.11005    0.25007  16.435  < 2e-16 ***
> bmi_cts      0.05825    0.01019   5.719  1.8e-08 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

```
> Residual standard error: 1.019 on 525 degrees of freedom
> Multiple R-squared:  0.05864, Adjusted R-squared:  0.05685
> F-statistic:  32.7 on 1 and 525 DF,  p-value: 1.803e-08


>
> Call:
> lm(formula = tc ~ bmi_cat, data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.9102 -0.6548 -0.1419  0.6233  4.2868
>
> Coefficients:
>                    Estimate Std. Error t value Pr(>|t|)
> (Intercept)         5.45022    0.06203  87.867  < 2e-16 ***
> bmi_catUnderweight -0.45382    0.15765  -2.879 0.004158 **
> bmi_catOverweight   0.19296    0.10265   1.880 0.060694 .
> bmi_catObese        0.59170    0.16185   3.656 0.000282 ***
> ---
> Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
>
> Residual standard error: 1.025 on 523 degrees of freedom
> Multiple R-squared:  0.05214, Adjusted R-squared:  0.0467
> F-statistic:  9.59 on 3 and 523 DF,  p-value: 3.575e-06
```

iv)





v)

I prefer the continuous BMI model. Although both models have statistically significant fitted coefficients, the continuous BMI model fits the data better with $R^2_{adj} = 0.0569$ compared to 0.0467 for the other model; this is also reflected by the model's p-value for the F-statistic and the residual standard error. Moreover, converting a continuous variable to a categorical one could cause some information about the data distribution to be lost.

Part B

Including a quadratic BMI term marginally improves the continuous BMI model, since in the new model the R^2_{adj} is higher and the p-value associated with the F-statistic is lower, but the fitted intercept is no longer statistically significant.

```
>
> Call:
> lm(formula = tc ~ bmi_cts + I(bmi_cts^2), data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.7664 -0.6523 -0.1061  0.5651  4.2049
>
> Coefficients:
>              Estimate Std. Error t value Pr(>|t|)
> (Intercept)   1.619210   0.996621   1.625  0.10483
> bmi_cts       0.257627   0.077911   3.307  0.00101 **
> I(bmi_cts^2) -0.003859   0.001495  -2.581  0.01012 *
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 1.014 on 524 degrees of freedom
> Multiple R-squared:  0.07046, Adjusted R-squared:  0.06691
> F-statistic: 19.86 on 2 and 524 DF,  p-value: 4.859e-09
```

Part C

i)

The best model is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i*3i} + \beta_5 X_{2i*3i} + \beta_6 X_{6i} + \beta_7 X_{7i} + e_i$, where:

- i, Y_i, X_{1i}, X_{2i}, X_{3i}, X_{1i*3i}, X_{2i*3i}, β_0 to β_5, and e_i are defined in the same way as in question 2 part D i).
- X_{6i} is the BMI of the i-th subject in kg/m^2
- X_{7i} is the squared BMI of the i-th subject in kg^2/m^4
- β_6 is the expected change in Y for a 1 kg/m^2 increase in X_6, holding all other covariates constant
- β_7 is the expected change in Y for a 1 kg^2/m^4 increase in X_7, holding all other covariates constant.

ii)

I fitted three candidate models based on the model with biological sex as an effect modifier from question 2 part D. The candidate models are as follows:

- Model 1: age, squared age, biological sex, age∗biological sex, squared age∗biological sex, continuous BMI
- Model 2: age, squared age, biological sex, age∗biological sex, squared age∗biological sex, categorical BMI
- Model 3: age, squared age, biological sex, age∗biological sex, squared age∗biological sex, continuous BMI, squared BMI.

I selected the best model by finding the model that satisfies the most of the following criteria: having the highest R^2_{adj}, the most significant F-statistic, the lowest RMSE, and the lowest AIC. The model with continuous BMI and squared BMI as covariates satisfied all four criteria, and I subsequently determined it to be the best model.

```
> Adjusted R^2 for models 1, 2, and 3 respectively: 0.1843731 0.1842259 0.1896014

> F-statistic p-value for models 1, 2, and 3 respectively: <2.2e-16 <2.2e-16 <2.2e-16

> RMSE for models 1, 2, and 3 respectively: 0.948 0.9481 0.9449

> AIC value for models 1, 2, and 3 respectively: 1448.213 1450.277 1445.809
```

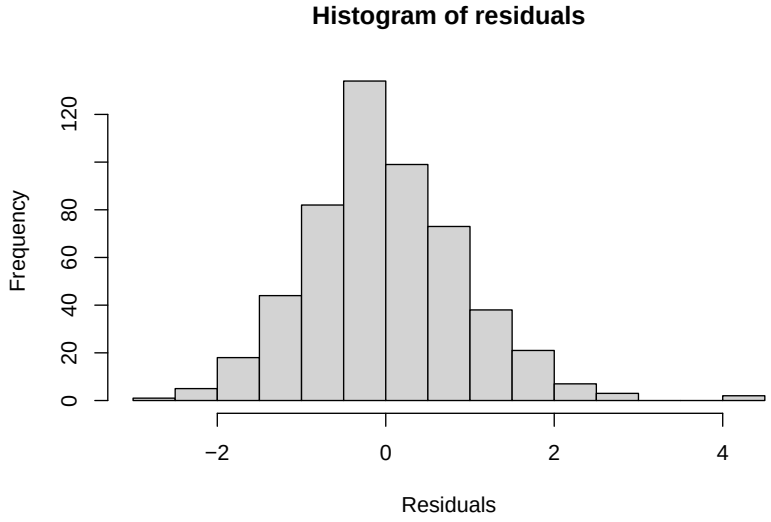
iii)

R^2_{adj} does improve by adding the effects of BMI; it is equal to 0.1757 without them and 0.1896 with them.

```
> Adjusted R^2 for the model without and with the effects of BMI respectively: 0.1757208 0.1896014
```

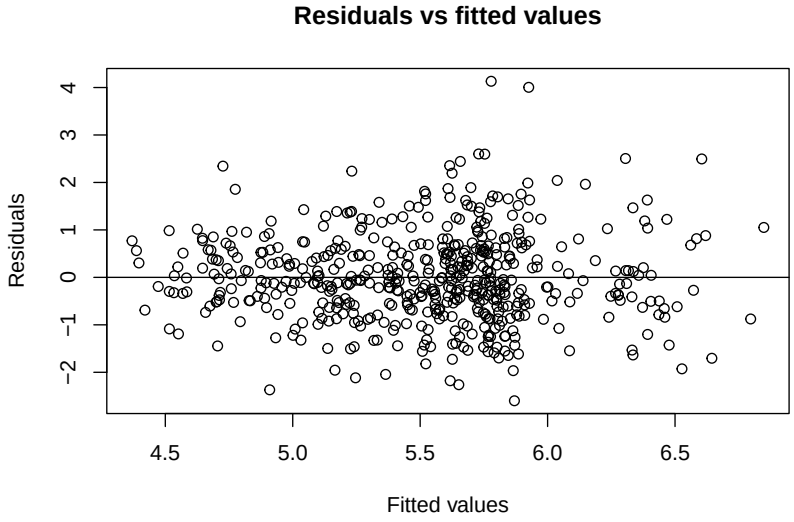
iv)

The residuals do look approximately normally distributed, although their distribution is slightly right skewed and there are two outliers with residuals greater than 4.



v)

The scatterplot does not show any evidence of heteroscedasticity or other model violations, although there does appear to be a slight clustering when the fitted value is around 5.8.



Part D

i)

Yes, there are points that have high leverage, as determined by the threshold on the diagonal elements on the hat matrix, and there are points that have high influence, as determined by the thresholds on the Cooks distance, DFFITS, and DFBETA of each point. In fact, there are 7 points which have high leverage and high influence according to all of these criteria.

```
> [1]  25 242 262 468 477 498 514
```

ii)

The 7 points found in the previous part are dropped, which does not change the model appreciably since the R^2_{adj} is 0.1916 vs 0.1896 before and the RMSE is 0.9293 vs 0.9449 before. However, the AIC of the new model has a noticeable decrease from 1446 to 1409, suggesting that the fit of the model does improve without the points.

```
>
> Call:
> lm(formula = tc ~ age + I(age^2) + gender + age:gender + I(age^2):gender +
>     bmi_cts + I(bmi_cts^2), data = df[-pts, ])
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.5936 -0.5677 -0.0884  0.5421  4.1530
>
> Coefficients:
>                   Estimate Std. Error t value Pr(>|t|)
> (Intercept)      0.7108179  1.0612743   0.670  0.50330
> age              0.1135159  0.0255215   4.448 1.06e-05 ***
> I(age^2)        -0.0011082  0.0002797  -3.962 8.50e-05 ***
> gender           1.4708814  0.7756658   1.896  0.05849 .
```

```
> bmi_cts          0.1422683  0.0763971   1.862  0.06314 .
> I(bmi_cts^2)    -0.0021168  0.0014610  -1.449  0.14797
> age:gender      -0.0922288  0.0375054  -2.459  0.01426 *
> I(age^2):gender  0.0012610  0.0004225   2.984  0.00298 **
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.9293 on 512 degrees of freedom
> Multiple R-squared:  0.2025,  Adjusted R-squared:  0.1916
> F-statistic: 18.57 on 7 and 512 DF,  p-value: < 2.2e-16

> [1] 1409.374
```

Appendix

```
fit_1a <- lm(tc ~ age, data = df)
summary(fit_1a)
confint(fit_1a)
cat("Pearson coefficient:", cor(df$age, df$tc, method = "pearson"))
plot(fitted(fit_1a), residuals(fit_1a), main = "Residuals vs fitted values", xlab = "Fitted values", ylab = "
hist(residuals(fit_1a), main = "Histogram of residuals", xlab = "Residuals")
qqnorm(residuals(fit_1a)); qqline(residuals(fit_1a))
plot(df$age, df$tc, main = "Total cholesterol vs age", xlab = "Age", ylab = "Total cholesterol"); lines(lowes
fit_1c <- lm(tc ~ age + I(age^2), data = df)
summary(fit_1c)
t.test(df$tc ~ df$gender, var.equal = T)
fit_2a <- lm(tc ~ gender, data = df)
summary(fit_2a)
confint(fit_2a)
fit_2b <- lm(tc ~ age + I(age^2) + gender, data = df)
summary(fit_2b)
df |> ggplot(aes(age, tc, color = factor(gender))) + geom_point() +
  geom_line(mapping = aes(y = predict(fit_2b))) +
  labs(title = "Total cholesterol vs age with biological sex as a confounder",
       x = "Age (years)", y = "Total cholesterol (mmol/L)",
       color = "Biological\nsex")
fit_2d <- lm(tc ~ age + I(age^2) + gender + age:gender + I(age^2):gender, data = df)
summary(fit_2d)
df |> ggplot(aes(age, tc, color = factor(gender))) + geom_point() +
  geom_line(mapping = aes(y = predict(fit_2d))) +
  labs(title = "Total cholesterol vs age with biological sex as an effect modifier",
       x = "Age (years)", y = "Total cholesterol (mmol/L)",
       color = "Biological\nsex")
fit_2f1 <- lm(tc ~ age + I(age^2), data = df[df$gender == 0,])
fit_2f2 <- lm(tc ~ age + I(age^2), data = df[df$gender == 1,])
summary(fit_2f1)
summary(fit_2f2)
predictions <- rep(0, nrow(df))
predictions[which(df$gender == 0)] <- predict(fit_2f1)
predictions[which(df$gender == 1)] <- predict(fit_2f2)
df$predictions <- predictions

df |> ggplot(aes(age, tc, color = factor(gender))) + geom_point() +
  geom_line(mapping = aes(y = predictions, color = factor(gender))) +
  labs(title = "Total cholesterol vs age with biological sex as an effect modifier",
       x = "Age (years)", y = "Total cholesterol (mmol/L)",
```

```
       color = "Biological\nsex")
df$bmi_cts <- df$weight / (df$height / 100)^2
df$bmi_cat <- cut(df$bmi_cts, breaks = c(-Inf, 18.5, 25.0, 30.0, Inf), labels = c("Underweight", "Normal weig
df$bmi_cat <- relevel(df$bmi_cat, ref = "Normal weight")
fit_3a1 <- lm(tc ~ bmi_cts, data = df)
fit_3a2 <- lm(tc ~ bmi_cat, data = df)
summary(fit_3a1)
summary(fit_3a2)
df |> ggplot(aes(bmi_cts, tc)) + geom_point() +
  geom_line(mapping = aes(y = predict(fit_3a1)), color = "blue") +
  labs(title = "Total cholesterol vs continuous BMI",
       x = "BMI (kg/m^2)", y = "Total cholesterol (mmol/L)")
df |> ggplot(aes(bmi_cts, tc)) + geom_point() +
  geom_line(mapping = aes(y = predict(fit_3a2)), color = "blue") +
  labs(title = "Total cholesterol vs categorical BMI",
       x = "BMI category", y = "Total cholesterol (mmol/L)")
fit_3b <- lm(tc ~ bmi_cts + I(bmi_cts^2), data = df)
summary(fit_3b)
# cat("Adjusted R^2 with and without a quadratic BMI term:", summary(fit_3b)$r.squared, summary(fit_3a2)$r.sq
fit_3c1 <- lm(tc ~ age + I(age^2) + gender + age:gender + I(age^2):gender + bmi_cts, data = df)
fit_3c2 <- lm(tc ~ age + I(age^2) + gender + age:gender + I(age^2):gender + bmi_cat, data = df)
fit_3c3 <- lm(tc ~ age + I(age^2) + gender + age:gender + I(age^2):gender + bmi_cts + I(bmi_cts^2), data = df
cat("Adjusted R^2 for models 1, 2, and 3 respectively:", summary(fit_3c1)$adj.r.squared, summary(fit_3c2)$adj
cat("F-statistic p-value for models 1, 2, and 3 respectively:", "<2.2e-16", "<2.2e-16", "<2.2e-16", "\n")
cat("RMSE for models 1, 2, and 3 respectively:", 0.9480, 0.9481, 0.9449, "\n")
cat("AIC value for models 1, 2, and 3 respectively:", AIC(fit_3c1), AIC(fit_3c2), AIC(fit_3c3))
cat("Adjusted R^2 for the model without and with the effects of BMI respectively:", summary(fit_2d)$adj.r.squ
hist(residuals(fit_3c3), main = "Histogram of residuals", xlab = "Residuals")
plot(fitted(fit_3c3), residuals(fit_3c3), main = "Residuals vs fitted values", xlab = "Fitted values", ylab =
pts <- Reduce(intersect, list(
  which(hatvalues(fit_3c3) > 16 / nrow(df)),
  which(cooks.distance(fit_3c3) > 4 / (nrow(df) - 2)),
  which(abs(dffits(fit_3c3)) > 2*sqrt(8 / nrow(df))),
  which(abs(dfbeta(fit_3c3)) > 2 / sqrt(nrow(df)))
))
pts
fit_3d <- lm(tc ~ age + I(age^2) + gender + age:gender + I(age^2):gender + bmi_cts + I(bmi_cts^2), data = df[
summary(fit_3d)
AIC(fit_3d)
```