Question 1

Part A

i)

No, continuous BMI does not confound the effect of sex in the relationship with total cholesterol. It does not meet the classical definition of confounding since continuous BMI is a risk factor for total cholesterol and a consequence of sex.

ii)

No, sex does not confound the effect of continuous BMI in the relationship with total cholesterol. It meets the classical definition of confounding since sex is a risk factor for total cholesterol, is associated with continuous BMI, and is not a consequence of continuous BMI. However, it does not meet the operational definition of confounding: the estimated slope for continuous BMI is 0.0299 and 0.0305 in the models with and without sex added respectively, so the change in the estimate is 2.16% as shown below, which is smaller than the threshold of 10%.

```
> Estimated coefficient for the crude analysis: 0.03052549
```

```
> Estimated coefficient for the adjusted analysis: 0.02988095
```

```
> Change in estimated coefficient: 2.157016 %
```

Part B

i)

No, continuous BMI does not modify the effect of sex in the relationship with total cholesterol, since the interaction term between continuous BMI and sex has a p-value of 0.52, which is not significant at the $\alpha = 0.05$ level (shown in the output below). My interpretation is that the relationship between sex and total cholesterol does not vary by the value of continuous BMI.

```
>
> Call:
> lm(formula = tc ~ age + I(age^2) + gender + bmi_cts + gender:bmi_cts,
>     data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.7144 -0.6025 -0.0793  0.5501  4.1128
>
> Coefficients:
>                  Estimate Std. Error t value Pr(>|t|)
> (Intercept)     2.4450100  0.4983415   4.906 1.24e-06 ***
> age             0.0790801  0.0189820   4.166 3.63e-05 ***
> I(age^2)       -0.0006209  0.0002088  -2.973  0.00308 **
> gender          0.3810542  0.4818589   0.791  0.42942
> bmi_cts         0.0373623  0.0155395   2.404  0.01655 *
> gender:bmi_cts -0.0125739  0.0196427  -0.640  0.52237
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.9628 on 521 degrees of freedom
> Multiple R-squared:  0.1667,  Adjusted R-squared:  0.1587
> F-statistic: 20.85 on 5 and 521 DF,  p-value: < 2.2e-16
```

ii)

No, since effect modification works in both directions, and the previous part indicated that continuous BMI does not modify the effect of sex in the relationship with total cholesterol. My interpretation is that the relationship between continuous BMI and total cholesterol does not vary by sex.

iii)

The non-fitted model equation is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{3i*4i} + e_i$, 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 (0 if male and 1 if female)
- X_{4i} is the continuous BMI of the i-th subject in kg/m^2
- X_{3i*4i} is the interaction term between continuous BMI and biological sex
- e_i is the error term for the i-th observation
- β_0 is the intercept corresponding to the average of Y in mmol/L when X_1 is 0 years, the subject is male, and X_4 is 0 kg/m^2
- 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
- β_3 is the expected change in Y when the subject is female compared to male, holding all other covariates constant
- β_4 is the expected change in Y for a 1 kg/m^2 increase in X_4, holding all other covariates constant
- β_5 is the expected change in Y for a 1 kg/m^2 increase in X_4 when the subject is female, holding all other covariates constant.

iv)

The non-fitted equation is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_4 X_{4i} + e_i$ for males and $Y_i = (\beta_0 + \beta_3) + \beta_1 X_{1i} + \beta_2 X_{2i} + (\beta_4 + \beta_5) X_{4i} + e_i$, where i, Y_i, X_{1i} to X_{4i}, X_{3i*4i}, and e_i are defined in the same way as in the previous part.

For males, the fitted equation is $\hat{Y}_i = 2.45 + 0.079 X_{1i} - 0.0006 X_{2i} + 0.037 X_{4i}$, where $\hat{Y}_i$ is the estimated total cholesterol of the i-th subject in mmol/L and i, X_{1i}, X_{3i}, and X_{4i} are defined as before, according to this data. For females, the fitted equation is $\hat{Y}_i = (2.45 + 0.38) + 0.079 X_{1i} - 0.0006 X_{2i} + (0.037 - 0.013) X_{4i} = 2.83 + 0.079 X_{1i} - 0.0006 X_{2i} + 0.024 X_{4i}$, where $\hat{Y}_i$ is the estimated total cholesterol of the i-th subject in mmol/L and i, X_{1i}, X_{2i}, and X_{4i} are defined the same as before, according to this data.

v)

The new model equation would be $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + (\beta_4 + \beta_7 X_{3i}) I_{4i} + (\beta_5 + \beta_8 X_{3i}) I_{5i} + (\beta_6 + \beta_9 X_{3i}) I_{6i} + e_i$, 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 (0 if male and 1 if female)
- I_{4i} is the indicator function for the BMI category underweight
- I_{5i} is the indicator function for the BMI category overweight
- I_{6i} is the indicator function for the BMI category obese
- β_0 is the intercept corresponding to the average of Y in mmol/L when X_1 is 0 years, the subject is male, and the patient's BMI is in the normal weight category
- 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
- β_3 is the expected change in Y when the subject is female compared to male and her BMI is in the normal weight category, holding all other covariates constant

- $\beta_4 + \beta_0$, $\beta_5 + \beta_0$, and $\beta_6 + \beta_0$ are the intercepts corresponding to the average of Y in mmol/L when X_1 is 0 years, the subject is male, and the patient's BMI is in the underweight, overweight, and obese categories respectively, holding all other covariates constant
- $\beta_7 + \beta_3$, $\beta_8 + \beta_3$, and $\beta_9 + \beta_3$ are the expected changes in Y when the patient is female and her BMI is in the underweight, overweight, and obese categories respectively, holding all other covariates constant.

The terms in this model that need to be statistically significant are $\beta_7 X_{3i} I_{4i}$, $\beta_8 X_{3i} I_{5i}$, and $\beta_9 X_{3i} I_{6i}$, which are the interactions between the variable for sex and the indicator variables for the BMI categories of underweight, overweight, and obese respectively.

Question 2

Part A

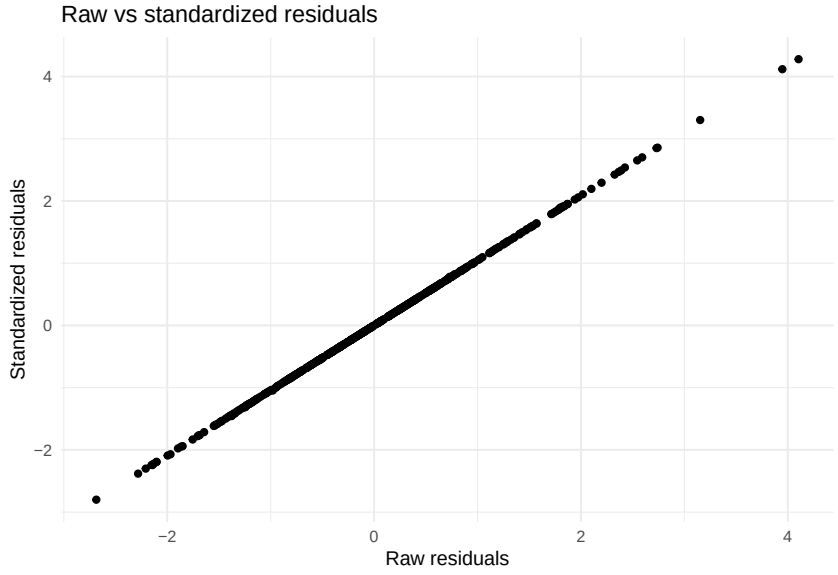
The total cholesterol value is 5.84 mmol/L or 225.8 mg/dL for a 30 year old male with BMI 30 kg/m^2.

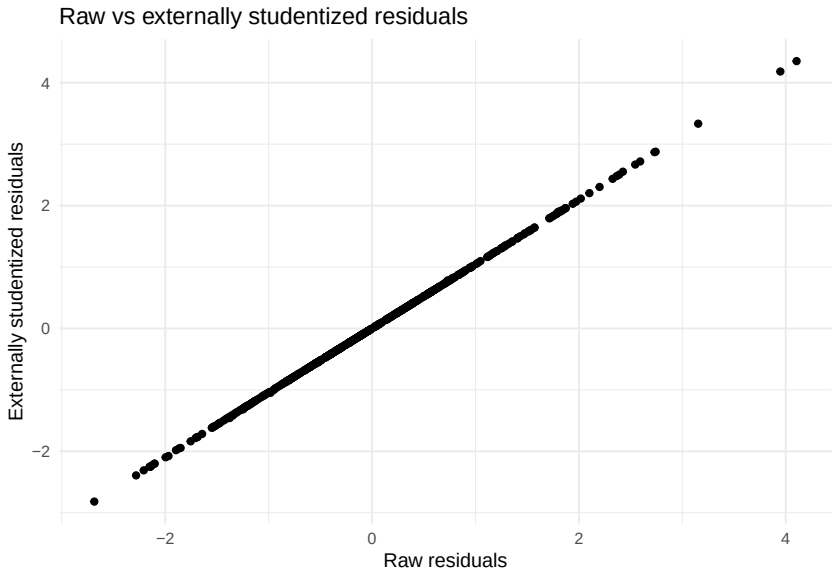
Part B

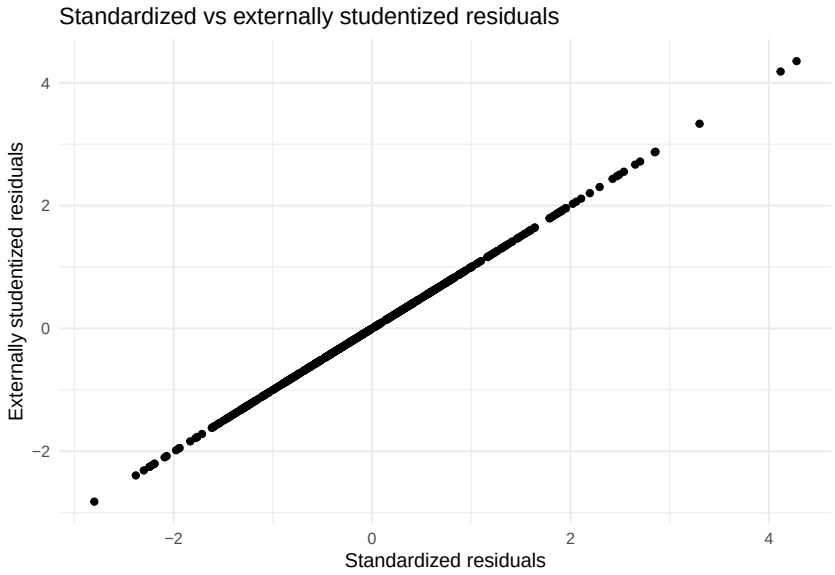
The R language provides functions for both standardized and studentized residuals; specific details can be found at https://stat.ethz.ch/R-manual/R-devel/library/stats/html/influence.measures.html. The function `rstandard()` uses the formula $z_i = \frac{e_i}{\sqrt{SSE/(n-(p+1))}}$ to re-normalize the residuals to have unit variance. The function `rstudent()` calculates externally studentized residuals using the formula $r_{(i)} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$, or what the documentation calls a "leave-one-out measure of the error variance".

Part C

iii)







iv)

I find that the correlations for all three comparisons to be very close to 1. I interpret this as showing that the adjustments to the raw residuals in the standardized and studentized formulas are more or less constant for each data point in this SCCS2 example. This means that using any of the three measures would yield similar results, and determining which residual to use would not have a large impact.

```
> Correlation for raw vs standardized residuals: 0.9999964
```

```
> Correlation for raw vs externally studentized residuals: 0.9999888
```

```
> Correlation for standardized vs externally studentized residuals: 0.9999911
```

Part D

i)

I find that the mean of the raw residuals is extremely close to zero. Thus, the claim is true in this example.

```
> Mean of raw residuals: 8.360186e-18
```

ii)

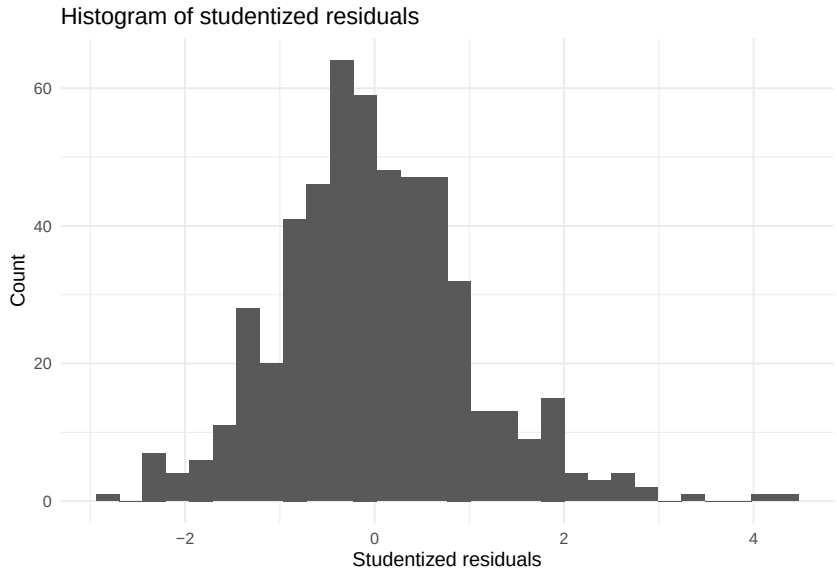
I find that both the standardized and the externally studentized residuals have means that are very close to zero. Thus, they can be said to have mean zero.

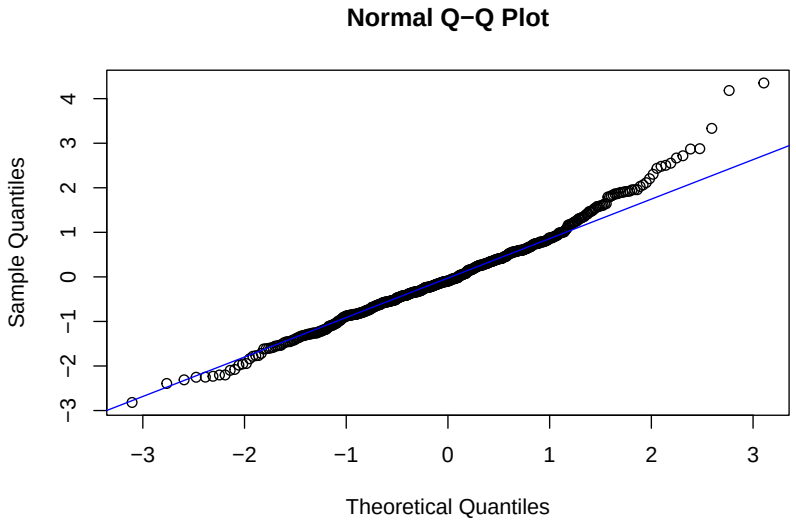
```
> Mean of standardized residuals: -8.860029e-05
```

```
> Mean of externally studentized residuals: 0.0004110206
```

Part E

Using the externally studentized residuals, the histogram and QQ plot below show that the normality assumption generally appears to hold for these data, despite the histogram being slightly assymmetrical and the QQ plot having a departure from the diagonal line on the right end.





Part F

i)

```
>          Age Sex Continuous BMI Total cholesterol Studentized residuals
> 43  20.79398   0       21.56679              7.07               2.551752
> 66  48.34497   0       20.34894              7.97               2.480455
> 91  37.04860   0       26.50933              3.39              -2.201778
> 92  48.31485   0       21.87004              7.59               2.029638
> 93  24.73922   0       20.78217              2.54              -2.392314
> 96  40.42163   0       22.99086              7.82               2.435924
> 99  68.68720   0       25.80639              3.90              -2.097338
> 103 54.57906   0       29.83798              9.93               4.182635
> 143 44.16975   0       21.30395              3.44              -2.201108
> 149 28.74196   0       23.27265              7.47               2.503601
> 191 39.44422   0       24.41728              8.10               2.718052
> 204 59.16496   0       27.67561              3.27              -2.818656
> 243 53.38809   0       28.77855              7.91               2.064331
> 256 57.27584   0       23.05327              8.35               2.667724
> 284 41.52772   0       25.11651              8.33               2.877129
> 305 32.05750   1       24.72425              3.13              -2.310396
> 387 63.32649   1       23.42319              8.02               2.203052
> 409 57.07050   1       24.02658              8.11               2.303288
> 451 46.19028   1       25.94075              9.91               4.351831
> 459 38.62560   1       33.87477              3.70              -2.252506
> 468 32.07666   1       36.42918              3.72              -2.077117
> 492 35.89596   1       24.44444              3.32              -2.250280
> 505 35.62218   1       20.41259              3.21              -2.231002
> 512 58.08898   1       29.40531              8.81               2.869683
> 514 67.12663   1       24.61937              9.10               3.333163
> 516 51.52088   1       31.25248              8.08               2.114362
```

ii)

```
>          Age Sex Continuous BMI Total cholesterol Studentized residuals
> 103 54.57906   0       29.83798              9.93               4.182635
> 451 46.19028   1       25.94075              9.91               4.351831
> 514 67.12663   1       24.61937              9.10               3.333163
```

iii)

These data points stick out as outliers because they have the largest total cholesterol values in the dataset. Their total cholesterol values are above 9 mmol/L, while the mean total cholesterol value in the dataset is 5.52 mmol/L.

iv)

These data points exhibit high outlying-ness due to their high total cholesterol values as discussed above. However, they do not exhibit high leverage since their hat values are below the threshold of $\frac{2(p+1)}{n} = \frac{2(5)}{527} = 0.019$.

```
> Hat values for the outliers: 0.007421154 0.006016941 0.01459026
```

Question 3

Part A

ii)

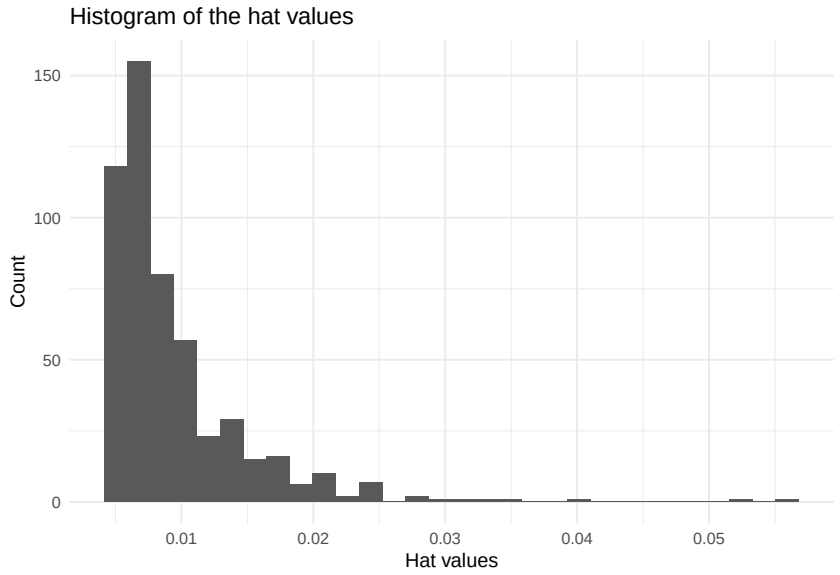
Indeed, the hat values in this example are all positive and their average is equal to $\frac{p+1}{n} = \frac{5}{527} = 0.0095$.

```
> Number of non-positive hat values: 0
```

```
> Average of hat values: 0.009487666
```

Part B

i)



ii)

```
>          Age Sex Continuous BMI Total cholesterol  Hat value
> 22  18.55989   0       24.61002              4.40 0.02130464
> 44  17.88912   0       18.99863              4.95 0.02088954
> 101 70.64203   0       21.43580              4.80 0.02089109
> 163 78.28063   0       24.52592              4.76 0.05294147
> 171 54.45585   0       39.52477              6.18 0.02806296
> 177 69.31691   0       31.04451              5.65 0.02068107
> 229 68.91171   0       18.61224              6.51 0.01993954
> 242 78.74332   0       23.12740              6.42 0.05594461
> 262 74.46954   0       22.59336              6.52 0.03321720
> 273 73.29227   0       19.81879              4.43 0.03062362
> 339 19.40862   1       25.11617              4.65 0.01973520
> 341 18.96509   1       18.85106              3.36 0.01919320
> 342 18.40931   1       19.33900              4.23 0.02044650
> 343 18.19028   1       21.73541              4.19 0.02104435
> 346 68.46817   1       17.66506              5.80 0.02406201
> 348 69.34155   1       19.26583              5.94 0.02342059
> 362 68.17796   1       20.46731              5.62 0.01939245
> 363 53.87269   1       39.98462              6.24 0.02860173
```

```
> 404 18.47502   1       22.94011              4.24 0.02070129
> 405 18.22861   1       20.88418              3.92 0.02078579
> 422 36.58864   1       37.53626              4.54 0.02498135
> 425 48.82409   1       37.81213              5.89 0.02284115
> 429 69.24298   1       29.89477              5.92 0.02111978
> 441 40.44627   1       37.96829              5.72 0.02513655
> 446 41.15263   1       36.56713              6.22 0.02104752
> 455 70.35729   1       30.48458              7.90 0.02472936
> 457 67.40041   1       37.14214              6.30 0.03012543
> 460 38.48597   1       37.27653              5.14 0.02363497
> 468 32.07666   1       36.42918              3.72 0.02379351
> 475 67.48528   1       40.49367              5.90 0.04034805
> 477 69.21835   1       37.40035              4.94 0.03499262
> 527 60.70910   1       38.00430              5.40 0.02421347
```

iii)

```
>          Age Sex Continuous BMI Total cholesterol  Hat value
> 163 78.28063   0       24.52592              4.76 0.05294147
> 242 78.74332   0       23.12740              6.42 0.05594461
> 475 67.48528   1       40.49367              5.90 0.04034805
```

iv)

These data points stick out as having high leverage because observations 163 and 242 have the two largest age values in the dataset, while observation 475 has the largest continuous BMI value in the dataset. Since leverage measures how much a covariate(s) deviate from their means, the fact that these points have high leverages makes sense.

v)

Leverage is driven by values of a covariate or covariates. This is because the hat matrix is calculated without using the outcomes.

Question 4

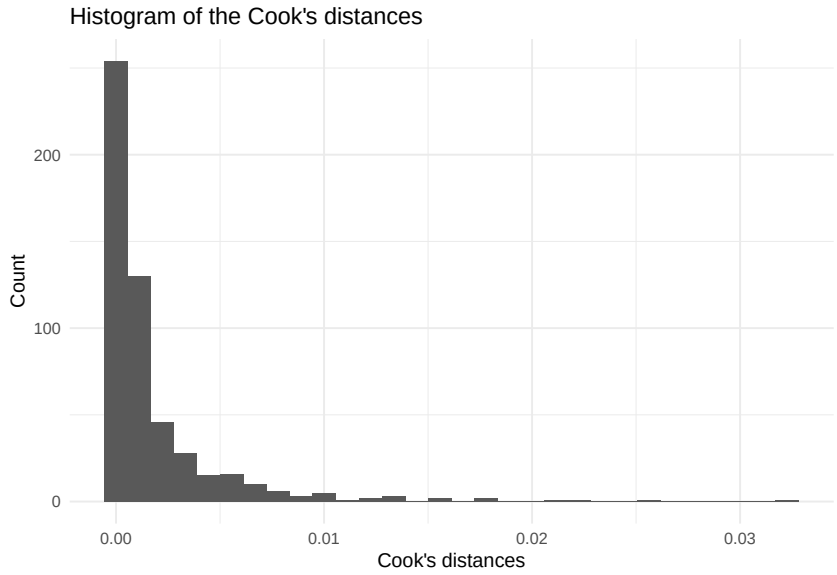
Part A

ii)

This is true in this example; all Cook's distances are positive here.

```
> Number of non-positive Cook's distances: 0
```

iii)



Part B

i)

```
>          Age Sex      BMI Total chol. Residuals   Hat value Cook's distance
> 25  63.53456   0 16.40226        3.78 -1.945399 0.016521840     0.012648252
> 43  20.79398   0 21.56679        7.07  2.551752 0.014025273     0.018331201
> 66  48.34497   0 20.34894        7.97  2.480455 0.007989882     0.009814110
> 93  24.73922   0 20.78217        2.54 -2.392314 0.008426749     0.009640282
> 99  68.68720   0 25.80639        3.90 -2.097338 0.015090199     0.013392039
> 103 54.57906   0 29.83798        9.93  4.182635 0.007421154     0.025358620
> 163 78.28063   0 24.52592        4.76 -1.046604 0.052941469     0.012244316
> 191 39.44422   0 24.41728        8.10  2.718052 0.005445438     0.007992222
> 204 59.16496   0 27.67561        3.27 -2.818656 0.006109172     0.009638689
> 256 57.27584   0 23.05327        8.35  2.667724 0.005713123     0.008083777
> 273 73.29227   0 19.81879        4.43 -1.315102 0.030623622     0.010912047
> 284 41.52772   0 25.11651        8.33  2.877129 0.005654165     0.009284668
> 387 63.32649   1 23.42319        8.02  2.203052 0.010049572     0.009781814
> 422 36.58864   1 37.53626        4.54 -1.418059 0.024981349     0.010284456
> 430 58.23135   1 18.95034        7.56  1.879107 0.011937687     0.008491187
> 451 46.19028   1 25.94075        9.91  4.351831 0.006016941     0.022166501
> 455 70.35729   1 30.48458        7.90  1.897684 0.024729356     0.018172161
> 459 38.62560   1 33.87477        3.70 -2.252506 0.015163740     0.015503443
> 468 32.07666   1 36.42918        3.72 -2.077117 0.023793513     0.020898743
> 477 69.21835   1 37.40035        4.94 -1.457973 0.034992622     0.015382934
> 498 56.22450   1 15.88121        7.48  1.913593 0.017388586     0.012894443
> 512 58.08898   1 29.40531        8.81  2.869683 0.007872502     0.012890361
> 514 67.12663   1 24.61937        9.10  3.333163 0.014590264     0.032274418
> 516 51.52088   1 31.25248        8.08  2.114362 0.009344782     0.008378333
```

ii)

```
>          Age Sex Continuous BMI Total cholesterol Studentized residuals
> 103 54.57906   0       29.83798              9.93              4.182635
> 514 67.12663   1       24.61937              9.10              3.333163
>       Hat value Cook's distance
```

```
> 103 0.007421154       0.02535862
> 514 0.014590264       0.03227442
```

iii)

Observation 103 has the maximum total cholesterol value in the dataset, contributing to a high Studentized residual value and thus a high Cook's distance. Observation 514 has a relatively high age value (dataset mean is 43.5 years) and high total cholesterol value (dataset mean is 24.2 mmol/L), contributing to a Studentized residual value and a high hat value respectively, which in turn contribute to a high Cook's distance.

Part C

i)

I choose DFFITS as the measure of influence.

Part C

ii)

The threshold $|\text{DFFITS}_i| > 2\sqrt{\frac{p+1}{n}} = 2\sqrt{\frac{5}{527}}$ is chosen for $i = 1, \ldots, 527$.

```
>          Age Sex Continuous BMI Total cholesterol DFFITS value
> 25  63.53456   0       16.40226              3.78   -0.2521480
> 43  20.79398   0       21.56679              7.07    0.3043415
> 66  48.34497   0       20.34894              7.97    0.2226094
> 93  24.73922   0       20.78217              2.54   -0.2205392
> 99  68.68720   0       25.80639              3.90   -0.2596078
> 103 54.57906   0       29.83798              9.93    0.3616622
> 163 78.28063   0       24.52592              4.76   -0.2474526
> 191 39.44422   0       24.41728              8.10    0.2011222
> 204 59.16496   0       27.67561              3.27   -0.2209856
> 256 57.27584   0       23.05327              8.35    0.2022190
> 273 73.29227   0       19.81879              4.43   -0.2337445
> 284 41.52772   0       25.11651              8.33    0.2169576
> 387 63.32649   1       23.42319              8.02    0.2219687
> 422 36.58864   1       37.53626              4.54   -0.2269843
> 430 58.23135   1       18.95034              7.56    0.2065473
> 451 46.19028   1       25.94075              9.91    0.3385871
> 455 70.35729   1       30.48458              7.90    0.3021814
> 459 38.62560   1       33.87477              3.70   -0.2795034
> 468 32.07666   1       36.42918              3.72   -0.3242795
> 477 69.21835   1       37.40035              4.94   -0.2776338
> 498 56.22450   1       15.88121              7.48    0.2545604
> 512 58.08898   1       29.40531              8.81    0.2556269
> 514 67.12663   1       24.61937              9.10    0.4055829
> 516 51.52088   1       31.25248              8.08    0.2053538
```

Question 5

Part A

The observations satisfying one of the three mentioned criteria are shown below. I observe that two observations (103 and 514) have both very large Studentized residual absolute values and very large Cook's distances. Interestingly, there are no observations that have both very large hat values and very large Cook's distance.

```
> Observation IDs with the absolute value of the Studentized residuals > 3: 103 451 514

> Observation IDs with hat value > 4(p+1)/n: 163 242 475

> Observation IDs with Cook's distance > 12/n: 103 514
```

Part B

My overall findings have not changed much. The new model has a slightly lower R^2_{adj} (0.1595 vs 0.1597), a slightly lower residual standard error (0.9221 vs 0.9622), and slightly different coefficient estimates and p-values (e.g., 2.72 vs 2.62 for the intercept, which is a 3% relative change). In my opinion, the models' results are not significantly different.

```
> [1] "Model run on all observations:"

>
> Call:
> lm(formula = tc ~ age + I(age^2) + as.factor(gender) + bmi_cts,
>     data = df)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -2.6861 -0.5977 -0.0903  0.5483  4.1049
>
> Coefficients:
>                      Estimate Std. Error t value Pr(>|t|)
> (Intercept)         2.6223982  0.4139550   6.335 5.13e-10 ***
> age                 0.0792126  0.0189701   4.176 3.48e-05 ***
> I(age^2)           -0.0006227  0.0002087  -2.984  0.00298 **
> as.factor(gender)1  0.0774204  0.0847893   0.913  0.36162
> bmi_cts             0.0298810  0.0102355   2.919  0.00366 **
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.9622 on 522 degrees of freedom
> Multiple R-squared:  0.1661,  Adjusted R-squared:  0.1597
> F-statistic: 25.99 on 4 and 522 DF,  p-value: < 2.2e-16

> [1] "Model run after eliminating the observations in question:"

>
> Call:
> lm(formula = tc ~ age + I(age^2) + as.factor(gender) + bmi_cts,
>     data = new_df)
>
> Residuals:
>      Min       1Q   Median       3Q      Max
> -2.65432 -0.57138 -0.07401  0.56057  2.78131
>
> Coefficients:
>                      Estimate Std. Error t value Pr(>|t|)
> (Intercept)         2.7164836  0.4078872   6.660 7.03e-11 ***
> age                 0.0762966  0.0189559   4.025 6.55e-05 ***
> I(age^2)           -0.0006011  0.0002105  -2.856  0.00447 **
> as.factor(gender)1  0.0607624  0.0816723   0.744  0.45723
> bmi_cts             0.0288282  0.0099503   2.897  0.00392 **
> ---
```

```
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.9221 on 516 degrees of freedom
> Multiple R-squared:  0.166,   Adjusted R-squared:  0.1595
> F-statistic: 25.68 on 4 and 516 DF,  p-value: < 2.2e-16
```

Appendix

```
fit_1a_1 <- lm(tc ~ bmi_cts + age + I(age^2), data = df)
fit_1a_2 <- lm(tc ~ bmi_cts + age + I(age^2) + gender, data = df)
coef_1 <- summary(fit_1a_1)$coef[2, 1]
coef_2 <- summary(fit_1a_2)$coef[2, 1]
cat("Estimated coefficient for the crude analysis:", coef_1, "\n")
cat("Estimated coefficient for the adjusted analysis:", coef_2, "\n")
cat("Change in estimated coefficient:", (coef_1 - coef_2) / coef_2 * 100, "%")
summary(lm(tc ~ age + I(age^2) + gender + bmi_cts + gender:bmi_cts, data = df))
fit_2a <- lm(tc ~ age + I(age^2) + as.factor(gender) + bmi_cts, data = df)
raw_res <- df$tc - fit_2a$fitted.values
std_res <- rstandard(fit_2a)
stu_res <- rstudent(fit_2a)
ggplot(df, aes(raw_res, std_res)) + geom_point() +
  labs(title = "Raw vs standardized residuals",
       x = "Raw residuals", y = "Standardized residuals")
ggplot(df, aes(raw_res, stu_res)) + geom_point() +
  labs(title = "Raw vs externally studentized residuals",
       x = "Raw residuals", y = "Externally studentized residuals")
ggplot(df, aes(std_res, stu_res)) + geom_point() +
  labs(title = "Standardized vs externally studentized residuals",
       x = "Standardized residuals", y = "Externally studentized residuals")
cat("Correlation for raw vs standardized residuals:", cor(raw_res, std_res, method = "pearson"), "\n")
cat("Correlation for raw vs externally studentized residuals:", cor(raw_res, stu_res, method = "pearson"), "\
cat("Correlation for standardized vs externally studentized residuals:", cor(std_res, stu_res, method = "pear
cat("Mean of raw residuals:", mean(raw_res))
cat("Mean of standardized residuals:", mean(std_res), "\n")
cat("Mean of externally studentized residuals:", mean(stu_res))
data.frame(stu_res = stu_res) |> ggplot(aes(stu_res)) + geom_histogram() +
  labs(title = "Histogram of studentized residuals",
       x = "Studentized residuals", y = "Count")
qqnorm(stu_res); qqline(stu_res, col = "blue")
tbl_2f1 <- df[abs(stu_res) > 2, c("age", "gender", "bmi_cts", "tc")]
tbl_2f1$stu_res <- stu_res[abs(stu_res) > 2]
colnames(tbl_2f1) <- c("Age", "Sex", "Continuous BMI", "Total cholesterol", "Studentized residuals")
tbl_2f1
tbl_2f2 <- df[abs(stu_res) > 3, c("age", "gender", "bmi_cts", "tc")]
tbl_2f2$stu_res <- stu_res[abs(stu_res) > 3]
colnames(tbl_2f2) <- c("Age", "Sex", "Continuous BMI", "Total cholesterol", "Studentized residuals")
tbl_2f2
hv <- hatvalues(fit_2a)
cat("Hat values for the outliers:", hv[rownames(tbl_2f2) |> as.integer()], "\n")
cat("Number of non-positive hat values:", sum(hv <= 0), "\n")
cat("Average of hat values:", mean(hv))
data.frame(h = hv) |> ggplot(aes(h)) + geom_histogram() +
  labs(title = "Histogram of the hat values", x = "Hat values", y = "Count")
tbl_3b1 <- df[hv > 2*5/527, c("age", "gender", "bmi_cts", "tc")]
tbl_3b1$h <- hv[hv > 2*5/527]
colnames(tbl_3b1) <- c("Age", "Sex", "Continuous BMI", "Total cholesterol", "Hat value")
```

```r
tbl_3b1
tbl_3b2 <- df[hv > 4*5/527, c("age", "gender", "bmi_cts", "tc")]
tbl_3b2$h <- hv[hv > 4*5/527]
colnames(tbl_3b2) <- c("Age", "Sex", "Continuous BMI", "Total cholesterol", "Hat value")
tbl_3b2
cd <- cooks.distance(fit_2a)
cat("Number of non-positive Cook's distances:", sum(cd <= 0))
data.frame(cd = cd) |> ggplot(aes(cd)) + geom_histogram() +
  labs(title = "Histogram of the Cook's distances",
       x = "Cook's distances", y = "Count")
tbl_4bi <- df[cd > 4/527, c("age", "gender", "bmi_cts", "tc")]
tbl_4bi$r <- stu_res[cd > 4/527]
tbl_4bi$h <- hv[cd > 4/527]
tbl_4bi$c <- cd[cd > 4/527]
colnames(tbl_4bi) <- c("Age", "Sex", "BMI", "Total chol.", "Residuals", "Hat value", "Cook's distance")
tbl_4bi
tbl_4bii <- df[cd > 12/527, c("age", "gender", "bmi_cts", "tc")]
tbl_4bii$r <- stu_res[cd > 12/527]
tbl_4bii$h <- hv[cd > 12/527]
tbl_4bii$c <- cd[cd > 12/527]
colnames(tbl_4bii) <- c("Age", "Sex", "Continuous BMI", "Total cholesterol", "Studentized residuals", "Hat va
tbl_4bii
ind <- abs(dffits(fit_2a)) > 2*sqrt(5/527)
tbl_4cii <- df[ind, c("age", "gender", "bmi_cts", "tc")]
tbl_4cii$d <- dffits(fit_2a)[ind]
colnames(tbl_4cii) <- c("Age", "Sex", "Continuous BMI", "Total cholesterol", "DFFITS value")
tbl_4cii
cat("Observation IDs with the absolute value of the Studentized residuals > 3:", rownames(tbl_2f2), "\n")
cat("Observation IDs with hat value > 4(p+1)/n:", rownames(tbl_3b2), "\n")
cat("Observation IDs with Cook's distance > 12/n:", rownames(tbl_4bii))
ind <- c(rownames(tbl_2f2), rownames(tbl_3b2), rownames(tbl_4bii)) |> as.integer() |> unique()
new_df <- df[-ind,]
print("Model run on all observations:")
summary(fit_2a)
print("Model run after eliminating the observations in question:")
summary(lm(tc ~ age + I(age^2) + as.factor(gender) + bmi_cts, data = new_df))
```