Stat 21 Homework 5

Person 1, Person 3, etc

Due: Sunday, March 13th by midnight

Contents

Use this file as the template for your submission. Do not delete anything from this template unless you are prompted to do so (e.g. where to write your name above, where to write your solutions or code below). Make sure you have installed the following packages in your version of RStudio: `tidyverse`, `knitr` **before** you attempt to knit this document.

Your completed assignment should be submitted as a single **PDF** using the link under Week 7 titled "Submit HW 5 to Gradescope". You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and a written answer interpreting the output within the context of the problem. You are allowed to work with your classmates on this homework assignment but you are expected to write up your own solutions. Every answer must be supported by a written statement unless otherwise specified. *A good rule of thumb is to make sure your answer is understandable to someone who hasn't read the problem question (or code output associated with it).*

Additionally, make sure that when you upload your solutions to Gradescope, you select which pages go correspond with which questions. Also, check to make sure that your knitted homework document is not uploaded as an extra-long single page document. Failure to do these things will result in a penalty on your homework grade. Finally, I strongly recommend that you address and resolve any knitting or R coding issues before Saturday as solutions to any R-coding questions that are not knitted properly will not receive any credit.

Part I: Non-R Problems

Problem 1

Decide if the following statements are true or false and explain why:

(a) For a multiple regression problem, the adjusted coefficient of determination will always be smaller than the regular, unadjusted coefficient of determination.

(b) If we fit a multiple regression model and then add a new predictor to the model, the (unadjusted) coefficient of determination will never decrease.

Solution:

(a) True

(b) False

Problem 2

Caterpillars go through free growth periods during each stage of their life. However, these periods end as the animal prepares to mold and then moves into the next stage of life. A biologist is interested in checking to see whether two different regression lines are needed to model the relationship between metabolic rates and body size of caterpillars for free growth and no free growth periods.

(a) Identify the MLR model for predicting metabolic rate ($Mrate$) from size ($BodySize$) and an indicator variable for free growth ($Ifgp = 1$ for free growth, 0 otherwise) that would allow for two different regression lines (slopes and/or intercepts) depending on the free growth status.

(b) Identify the MLR model for predicting $Mrate$ from $BodySize$ and $Ifgp$, when the rate of change in the mean $Mrate$ with respect to size is the same during free growth and no free growth periods.

(c) Identify the full and reduced models that would be used in a nested F-test to check if one or two regression lines are needed to model metabolic rates.

Solution:

In each of the following, let $x_1 = BodySize$, $y = Mrate$, and $x_2 = \begin{cases} 1, & \text{if in free growth period} \\ 0, & \text{otherwise} \end{cases}$. Also, assume ϵ are identically distributed random noise centered at zero with a constant variance.

(a) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$

(b) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

(c) model in (a) is full, model in (b) is reduced

Rubric: For full credit, students must define any variables they use in their model besides the ones listed in the problem. Also, for full credit, student answers must:

(a) contain an interaction effect between the two predictors

(b) contain main effect terms only

(c) correct answer

Problem 3

Suppose the biologist in Problem 2 gives you data on 53 caterpillars. Identify the degrees of freedom for error in each of the models for parts (a) and (b).

Solution:

(a) $53 - 3 - 1 = 49$

(b) $53 - 2 - 1 = 50$

Rubric: For full credit, students must answer both (a) and (b) correctly

Problem 4

Suppose that X_1 and X_2 are positively related with $X_1 = 2X_2 - 4$. Let $Y = 0.5X_1 + 5$ summarize a positive linear relationship between Y and X_1.

(a) Substitute the first equation into the second to show a linear relationship between Y and X_2. Comment on the direction of the association between Y and X_2 in the new equation.

(b) Now add the original two equations and rearrange terms to give an equation in the form $Y = aX_1 + bX_2 + c$. Are the coefficients of X_1 and X_2 both in the direction you would expect based on the signs in the separate equations?

Solution:

(a) $Y = 0.5X_1 + 5 = 0.5(2X_2 - 4) + 5 = (0.5 \times 2)X_2 - (0.5 \times 4) + 5 = X_2 + 3$ is again an equation for a line with respect to X_2. The coefficient of X_2 is also positive

(b) $X_1 + Y = 2X_2 - 4 + 0.5X_1 + 5$ which implies that $Y = 2X_2 - 4 + 0.5X_1 + 5 - X_1 = -0.5X_1 + 2X_2 + 1$ so now it appears that X_1 is *negatively associated* with Y (this demonstrates potential confusing effects in the presence of multicollinearity)

Rubric: For full credit, students must answer both (a) and (b) correctly with math and provide accurate comments

Part II: R Problems

Problem 5

```
library(Stat2Data)
data(ElephantsFB)
head(ElephantsFB)
```

```
##     Age Height Firstborn
## 1  1.40    120         0
## 2 17.50    227         0
## 3 12.75    235         0
## 4 11.17    210         0
## 5 12.67    220         1
## 6 12.67    189         1
```

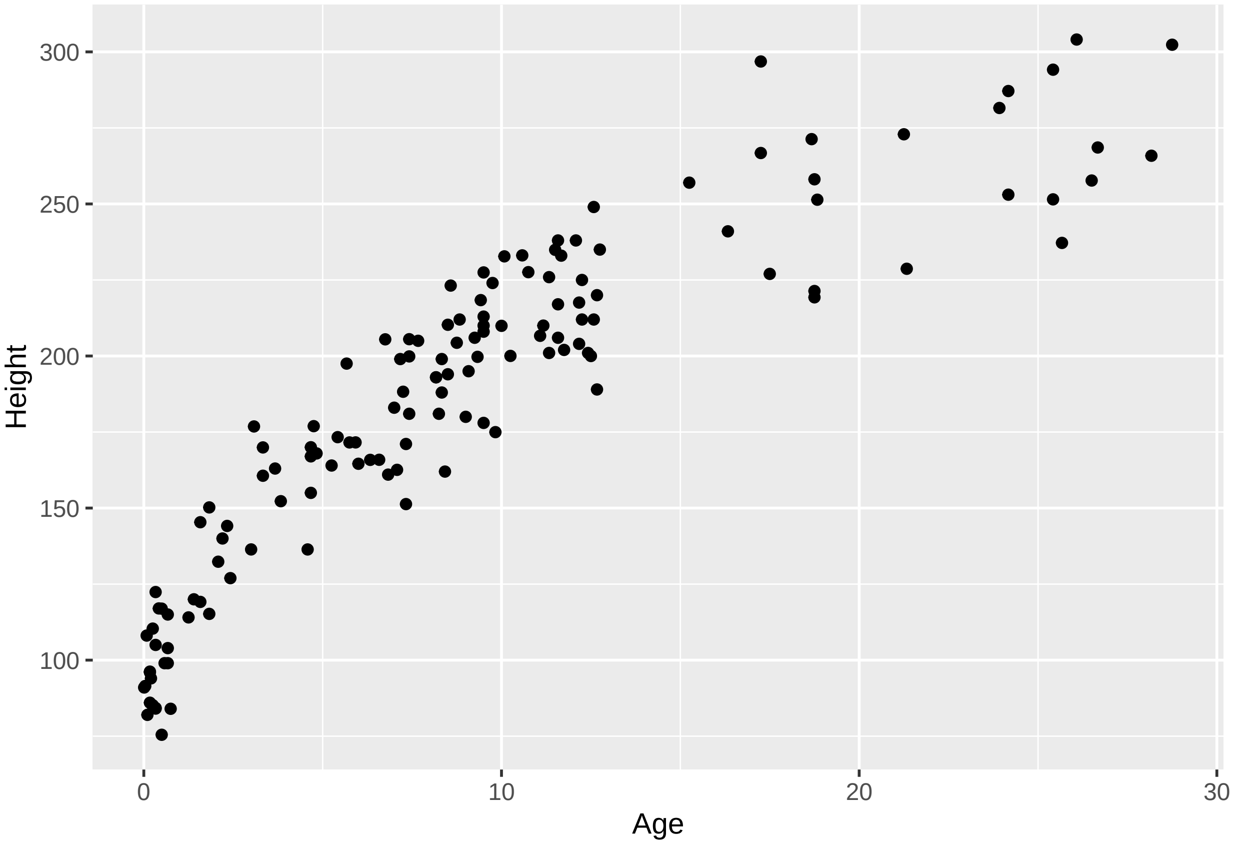
Elephants are worse off if there is a drought during the first two years of their life. Suppose we are interested in determining if maternal experience matters as well. That is, we want to determine if it is helpful or harmful if an elephant is firstborn. The data set `ElephantsFB` has data on 138 male African elephants that lived through droughts in the first two years of life. The variable `Height` records shoulder height in cm, `Age` is the age in years, and `Firstborn` is 1 for firstborn and 0 for non-firstborn.

(a) Plot `Height` against `Age` and comment on the pattern.

(b) What is the fitted quadratic regression model for using `Age` to predict `Height`?

(c) Use the fitted model from part (b) to predict the height of a 15-year old elephant. What does the model predict the height will be?

```
ggplot(ElephantsFB, aes(x=Age, y=Height)) +
  geom_point()
```



```r
elephant_mod <- lm(Height ~ Age + Age:Age, ElephantsFB)
elephant_mod %>% summary
```

```
##
## Call:
## lm(formula = Height ~ Age + Age:Age, data = ElephantsFB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.893 -16.119   2.531  17.969  54.132
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.1155     3.3255   37.02   <2e-16 ***
## Age           6.9329     0.2898   23.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.46 on 136 degrees of freedom
## Multiple R-squared:  0.8079, Adjusted R-squared:  0.8065
## F-statistic: 572.1 on 1 and 136 DF,  p-value: < 2.2e-16
```

Solution:

(a) there is a curve in this scatterplot

(b) $ExpectedHeight = 100.2 + 13.383(Age) - 0.2643(Age^2)$

(c) $100.2 + 13.383(15) - 0.2643(15^2) = 241.5$ cm

Rubric: For full credit students must answer all parts correctly (including the units in part (c))

Animals that are stressed might increase their oxygen consumption. Biologists measured oxygen consumption of shore crabs that were either exposed to 7.5 minutes of ship-noise or 7.5 minutes of ambient harbor noise. They noticed two things: (1) the greater the mass of the crab, the greater the rate of oxygen consumption and (2) ship-noise affected the crabs differently from ambient noise. The data set `CrabShip` includes the variable `Noise`, which has two levels: "ambient" and "ship". The variable `Mass` (g) is the mass of the crab. The variable `Oxygen` ($\mu molesh^{-1}$) is the rate of oxygen consumption. Use this information to answer problems 6-8.

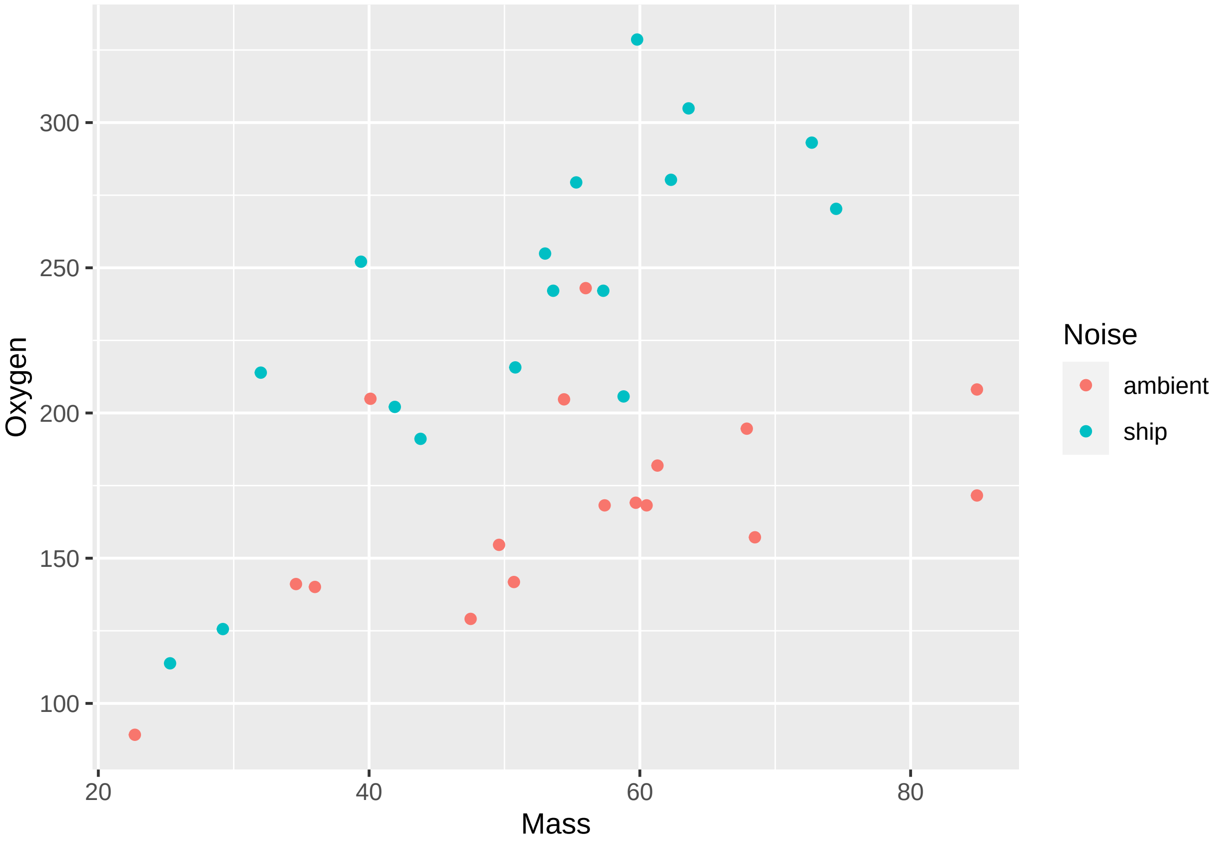
```
library(Stat2Data)
data(CrabShip)
head(CrabShip)
```

```
##   Mass Oxygen   Noise
## 1 22.7   89.2 ambient
## 2 34.6  141.1 ambient
## 3 36.0  140.1 ambient
## 4 40.1  204.9 ambient
## 5 47.5  129.1 ambient
## 6 49.6  154.6 ambient
```

Problem 6

Make a scatter plot of $Y = Oxygen$ vs $X = Mass$ and with $Treatment$ as a grouping variable (i.e. use different colors or different plotting symbols for the two levels of `Noise`.) Comment on the plot.

```
ggplot(CrabShip, aes(x=Mass, y=Oxygen, color=Noise)) +   ## Note, this will only work if the variable No
  geom_point()
```



Solution: looks like the slope for treatment level ship is steeper (higher) than that for treatment level ambient. in other words, higher oxygen levels for different mass values for ship group than for ambient group. more noticeable difference for moderate to large crabs.
Rubric: for full credit must produce a valid plot and provide a valid interpretation

Problem 7

(a) Fig the regression of `Oxygen` on `Mass` and test whether there is a linear association between the two variables.

(b) Fit a model that produces parallel regression lines for the two levels of `Noise`.

(c) Fit the general model that produces nonparallel regression lines for the two levels of `Noise`.

```
## a
moda <- lm(Oxygen~Mass, CrabShip)
moda %>% summary
```

```
##
## Call:
## lm(formula = Oxygen ~ Mass, data = CrabShip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -86.785 -41.557  -3.188  38.177 114.558
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 108.3950    33.2650   3.259  0.00265 **
## Mass          1.7667     0.6011   2.939  0.00606 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.02 on 32 degrees of freedom
## Multiple R-squared:  0.2126, Adjusted R-squared:  0.188
## F-statistic: 8.639 on 1 and 32 DF,  p-value: 0.006063
```

```
## b
modb <- lm(Oxygen~Mass+Noise, CrabShip)
modb %>% summary
```

```
##
## Call:
## lm(formula = Oxygen ~ Mass + Noise, data = CrabShip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.364 -18.938  -3.953  17.210  74.905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.4279    24.9816   2.179   0.0371 *
## Mass          2.0734     0.4231   4.901 2.85e-05 ***
## Noiseship    75.2795    12.8000   5.881 1.72e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.03 on 31 degrees of freedom
## Multiple R-squared:  0.6278, Adjusted R-squared:  0.6038
## F-statistic: 26.15 on 2 and 31 DF,  p-value: 2.221e-07
```

```
## c
modc <- lm(Oxygen~Mass+Noise+Mass:Noise, CrabShip)
modc %>% summary
```

```
##
## Call:
## lm(formula = Oxygen ~ Mass + Noise + Mass:Noise, data = CrabShip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.719 -21.350  -4.149  12.715  73.261
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    103.2703    29.3894   3.514  0.00142 **
## Mass             1.1869     0.5121   2.318  0.02746 *
## Noiseship      -34.3904    43.0782  -0.798  0.43096
## Mass:Noiseship   2.0705     0.7826   2.646  0.01286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.9 on 30 degrees of freedom
## Multiple R-squared:  0.6982, Adjusted R-squared:  0.6681
## F-statistic: 23.14 on 3 and 30 DF,  p-value: 5.942e-08
```

Solution:

In each model below, let $y = Oxygen$, $x_1 = Mass$, and $x_2 = \begin{cases} 1, & \text{if ship noise} \\ 0, & \text{otherwise} \end{cases}$ (Note, by default, R chooses the reference category for a set of indicator variables alphanumerically)

(a) There are three equivalent tests one could conduct here: a test for the significance of the single predictor, a test for the significance of the correlation between the predictor and the response, and the ANOVA overall F-test for the fit of the SLR model. The p-value for any of these tests is 0.006063 which is small enough to conclude that there is a statistically significant linear relationship between Mass and Oxygen (this is only if the residuals plots show no clear indication of non-linearity however!).

$$\hat{y} = 108.3950 + 1.7667x_1$$

(b)

$$\hat{y} = 54.4279 + 2.0734x_1 + 75.2795x_2$$

(c)

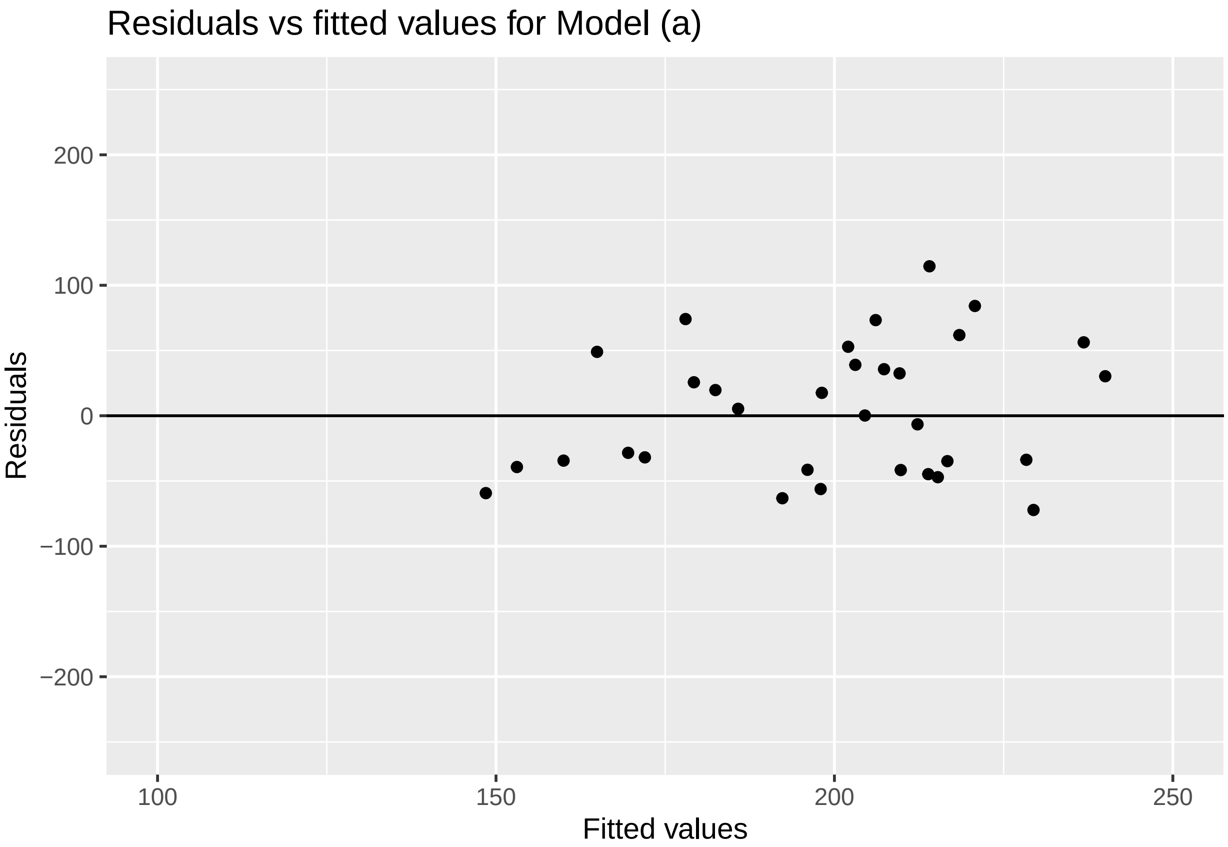
$$\hat{y} = 103.2703 + 1.1869x_1 - 34.3904x_2 + 2.0705x_1x_2$$

Rubric: For full credit, the models should be appropriately specified and include all necessary terms. Don't deduct points for students not stating the hypotheses of the test in part (a), just make sure they specify the correct p-value.

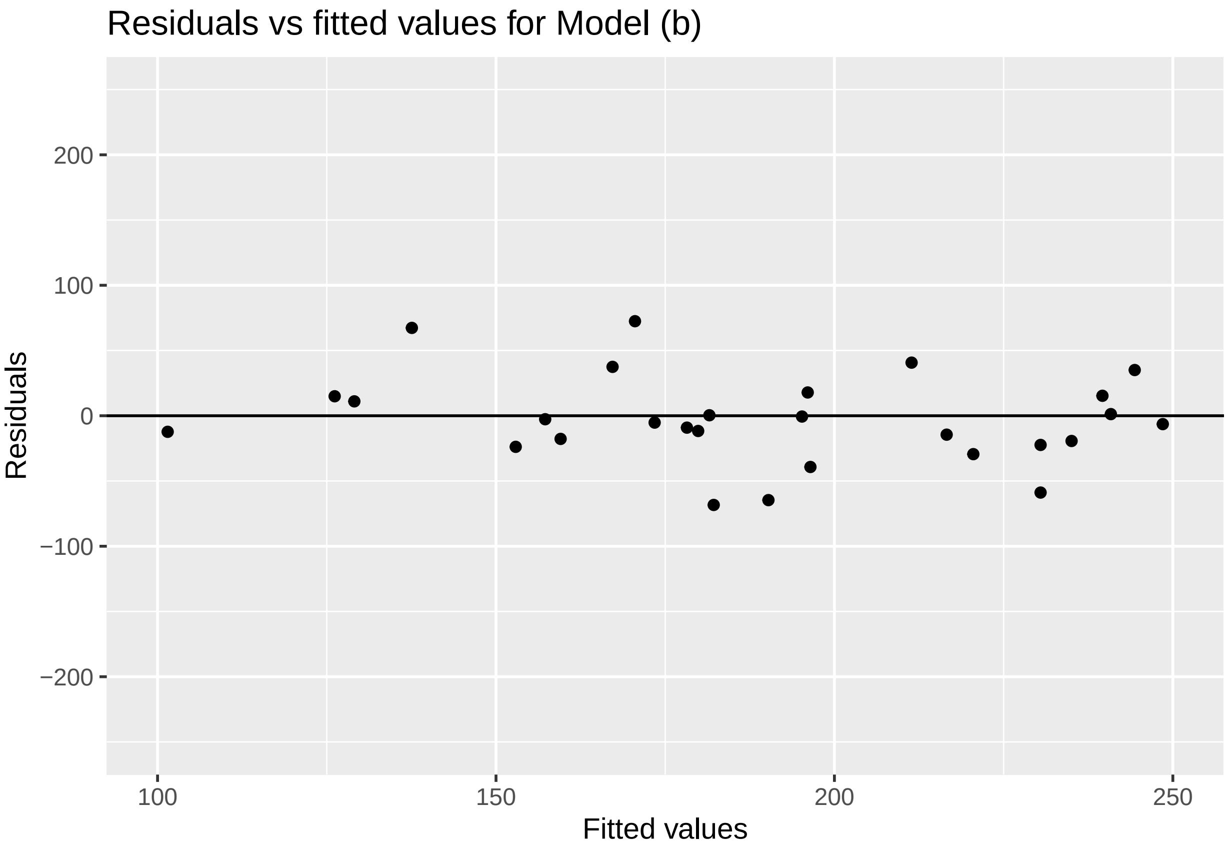
Problem 8

Which model from Problem 7 is the best choice and why? Write down the fitted prediction equation for each level of `Noise` for your final choice. (Hint: Assess the regression model conditions in each of the models from Problem 7 to help inform your choice.)

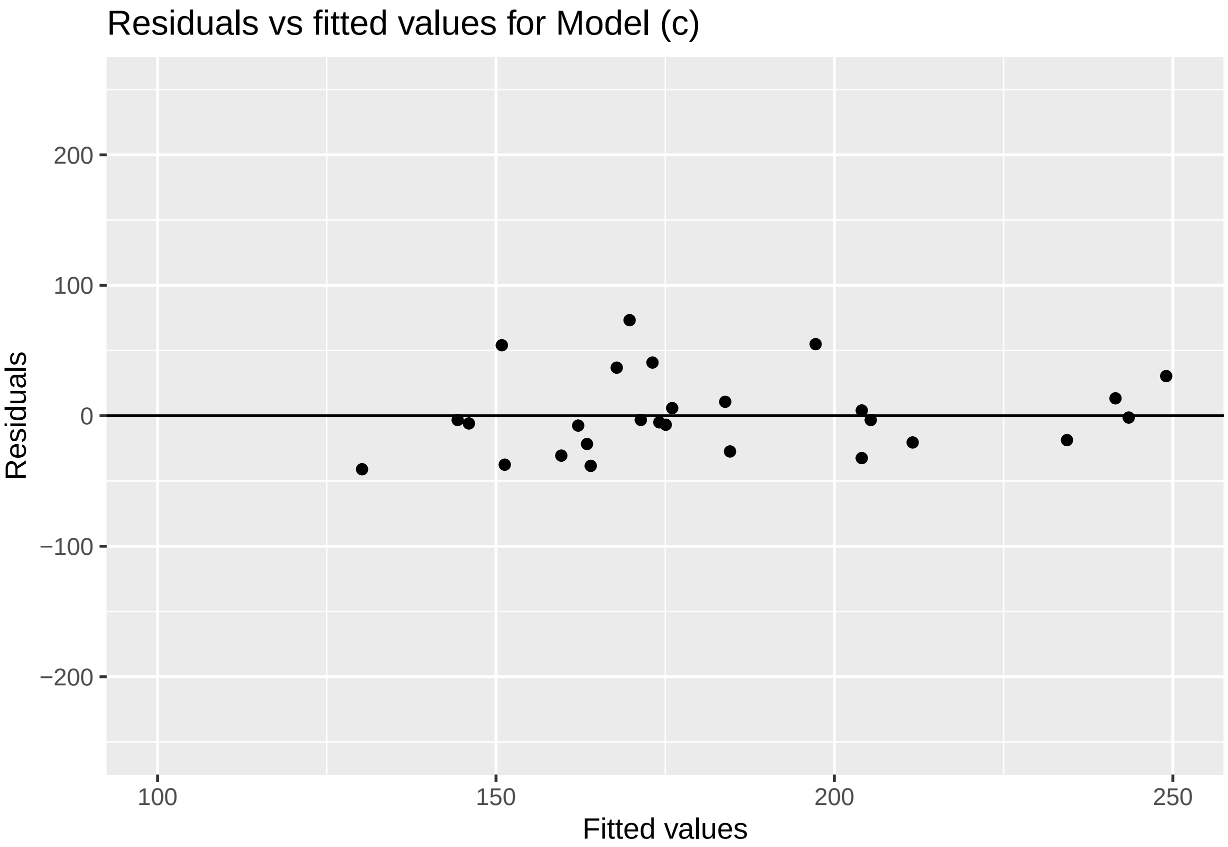
```r
crab_data_all <- CrabShip %>% mutate(resids1 = moda$residuals,
                                     fits1 = moda$fitted.values,
                                     resids2 = modb$residuals,
                                     fits2 = modb$fitted.values,
                                     resids3 = modc$residuals,
                                     fits3 = modc$fitted.values)

ggplot(crab_data_all, aes(x=fits1, y=resids1)) +
  geom_point() +
  geom_abline(slope = 0, intercept = 0) +
  xlim(100, 250) + ylim(-250, 250) +
  labs(title="Residuals vs fitted values for Model (a)", x="Fitted values", y="Residuals")
```



```
ggplot(crab_data_all, aes(x=fits2, y=resids2)) +
  geom_point() +
  geom_abline(slope = 0, intercept = 0) +
  xlim(100, 250) + ylim(-250, 250) +
  labs(title="Residuals vs fitted values for Model (b)", x="Fitted values", y="Residuals")
```



```
ggplot(crab_data_all, aes(x=fits3, y=resids3)) +
  geom_point() +
  geom_abline(slope = 0, intercept = 0) +
  xlim(100, 250) + ylim(-250, 250) +
  labs(title="Residuals vs fitted values for Model (c)", x="Fitted values", y="Residuals")
```



Solution: Based on the adjusted R^2 values, the model allowing for nonparallel regression lines fits best. The p-values for overall F-tests of model fit don't provide any discriminating evidence since there are all small enough to reject each null hypothesis at most typical α levels. The residual plot for model (a) might show some curvature but otherwise, the residual plots for the other models are comparable and don't provide much discriminating information either.

Regression equation for crabs with ship noise treatment:

$$\hat{y} = (103.2703 - 34.3904) + (1.1869 + 2.0705)x_1$$

Regression equation for crabs with ambient noise treatment:

$$\hat{y} = 103.2703 + 1.1869x_1$$

Rubric: Full credit for any valid attempts to discern among the three models and for writing out the corresponding regression equations for each treatment level.

The data set `MathEnrollment` contains data on total enrollments in mathematics courses at a small liberal arts college where the academic year consists of two semesters, one in the fall and another in the spring. Use this data set that spans from Fall 2001 to Spring 2012 to answer Problems 9 and 10.

```
library(Stat2Data)
data(MathEnrollment)
head(MathEnrollment)
```

```
##   AYear Fall Spring
## 1  2001  259    246
## 2  2002  301     206
## 3  2003  343     288
## 4  2004  307     215
## 5  2005  286     230
## 6  2006  273     247
```

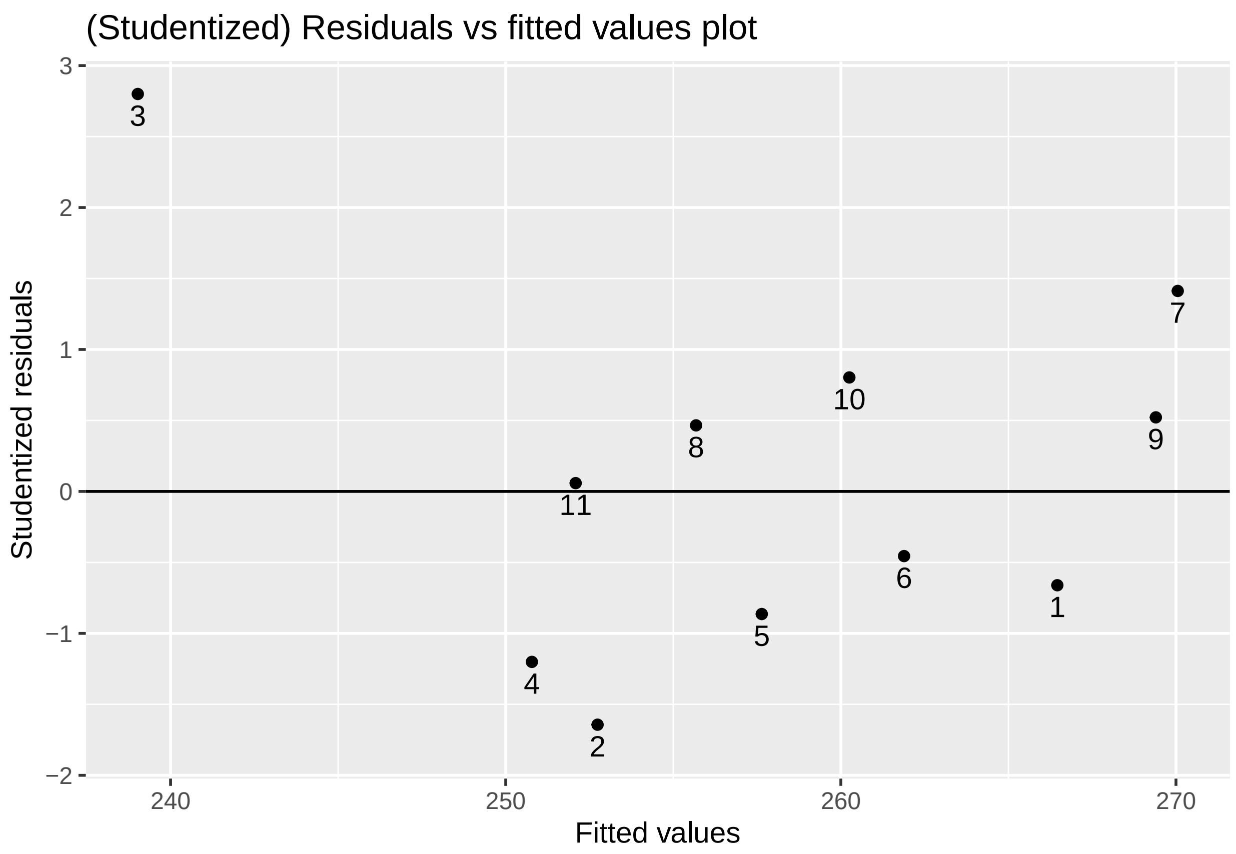
Problem 9

(a) Fit a regression model for predicting spring enrollment (`Spring`) from fall enrollment (`Fall`). Identify which year provides unusual data and determine how influential this data point is.

(b) Create a new data set by removing the influential data point you found in part (a) and then fit the same regression model. Prepare the appropriate residual plots and comment on the slight problems with the conditions for inference in this model. In particular, make sure that you plot the residuals against order (or `AYear`) and comment on the trend.

(c) What percent of the variability in spring enrollment is explained by this simple linear model and what is the conclusion of a test for the significance of a linear association between spring and fall enrollments?

```
## a
enroll_mod <- lm(Spring ~ Fall, MathEnrollment)

MathEnrollment_all <- MathEnrollment %>% mutate(studresids = rstudent(enroll_mod),
                                                fits = enroll_mod$fitted.values)

ggplot(MathEnrollment_all, aes(x=fits, y=studresids)) +
    geom_point() +
    labs(title="(Studentized) Residuals vs fitted values plot", x="Fitted values", y="Studentized resid
    geom_hline(yintercept=0) +
    geom_text(label=rownames(MathEnrollment_all), nudge_y=-0.15) ## this adds a label just below each d
```



```
MathEnrollment_all[3,]
```

```
##   AYear Fall Spring studresids     fits
## 3  2003  343    288   2.799773 239.0216
```

```
## b
MathEnrollment %>% dim
```

```
## [1] 11  3
```

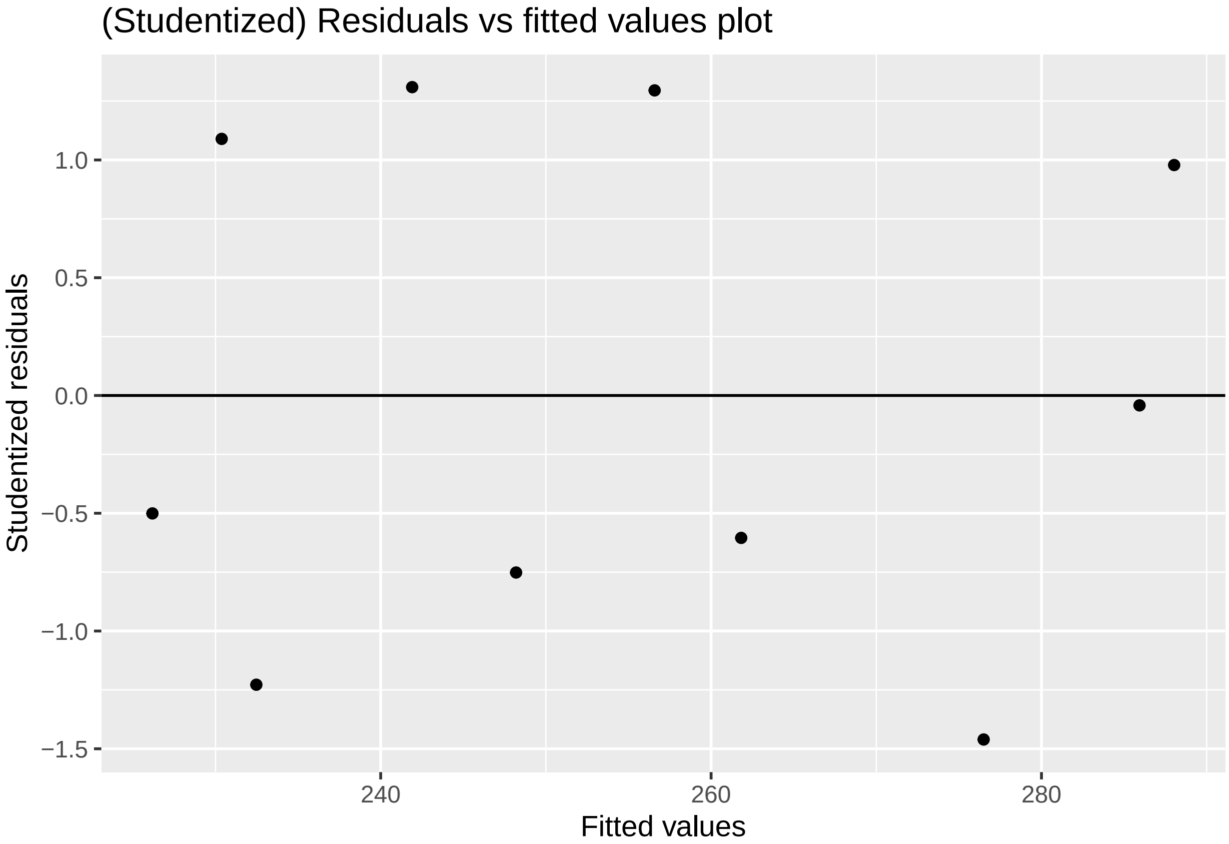
```
MathEnrollment_subset <- MathEnrollment %>% filter(AYear!=2003)
MathEnrollment_subset %>% dim
```

```
## [1] 10  3
```

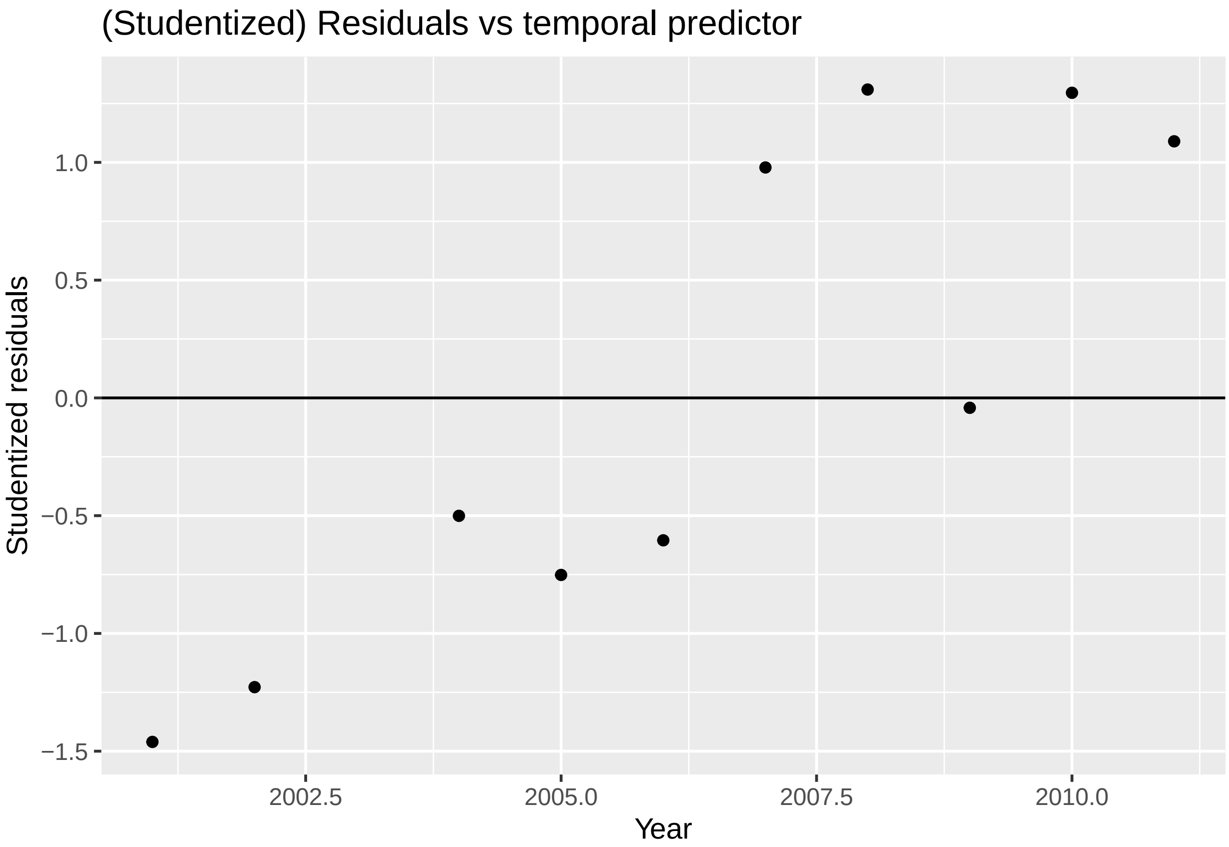
```
enroll_mod_subset <- lm(Spring ~ Fall, MathEnrollment_subset)
enroll_mod_subset %>% summary
```

```
##
## Call:
## lm(formula = Spring ~ Fall, data = MathEnrollment_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.500 -17.353  -6.058  22.711  29.418
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 548.0094   106.7236   5.135 0.000891 ***
## Fall         -1.0483     0.3805  -2.755 0.024870 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.94 on 8 degrees of freedom
## Multiple R-squared:  0.4868, Adjusted R-squared:  0.4227
## F-statistic: 7.589 on 1 and 8 DF,  p-value: 0.02487
```

```
MathEnrollment_subset_all <- MathEnrollment_subset %>% mutate(studresids = rstudent(enroll_mod_subset),
                                                              fits = enroll_mod_subset$fitted.values)

ggplot(MathEnrollment_subset_all, aes(x=fits, y=studresids)) +
    geom_point() +
    labs(title="(Studentized) Residuals vs fitted values plot", x="Fitted values", y="Studentized resid
    geom_hline(yintercept=0)
```



```
ggplot(MathEnrollment_subset_all, aes(x=AYear, y=studresids)) +
    geom_point() +
    labs(title="(Studentized) Residuals vs temporal predictor", x="Year", y="Studentized residuals") +
    geom_hline(yintercept=0)
```



Solution:

(a) The data point corresponding to year 2003 has an unexpectedly high (studentized) residual. we can tell how influential it is by assessing the same regression model applied to the full data set and then applied to the subset that does not include this data point.

(b) The residual vs fitted values plot looks great however, enrollment data is temporal in nature which can be seen in the residuals vs year plot. This latter plot indicates that the residuals increase with time. To meet the independent noise assumption, the residuals should not show any clear trends or patterns like this.

(c) model describes about 43% of the variability in Spring enrollment and the model fit is adequate at an $\alpha = 0.05$ level (all in all, this looks like a decent model **if** we ignore the problem of independent random noise)

Rubric: Only check to make sure students attempted each part and that they correctly reduced the size of the data set for parts (b) and (c) by one point. Everything else about this problem please grade for completion.

Problem 10

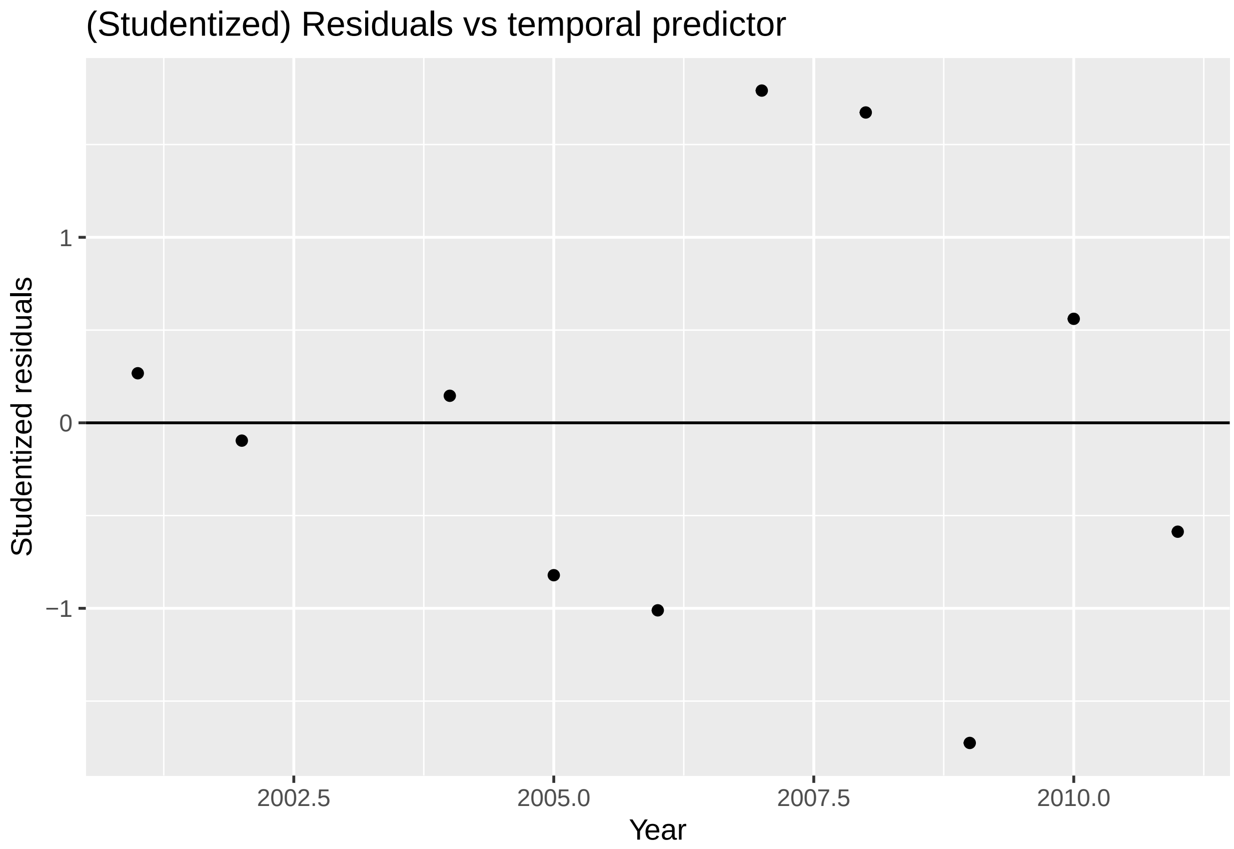
(a) Using the data set with the influential data point (from Problem 9 (a)) removed, now fit a MLR model for predicting spring enrollment (`Spring`) from fall enrollment (`Fall`) and another predictor academic year (`AYear`). Report the fitted regression equation.

(b) Create appropriate residual plots and comment on the conditions for inference. Did the slight problems with the residual plots that you noticed in Problem 9 (b) disappear?

```
enroll_mod_subset2 <- lm(Spring ~ Fall + AYear, MathEnrollment_subset)
enroll_mod_subset2 %>% summary
```

```
##
## Call:
## lm(formula = Spring ~ Fall + AYear, data = MathEnrollment_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1945  -9.3982   0.3212   5.8503  18.2036
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.172e+04  2.686e+03  -4.361  0.00331 **
## Fall        -1.007e+00  2.041e-01  -4.933  0.00169 **
## AYear        6.107e+00  1.337e+00   4.566  0.00258 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.37 on 7 degrees of freedom
## Multiple R-squared:  0.871,  Adjusted R-squared:  0.8342
## F-statistic: 23.64 on 2 and 7 DF,  p-value: 0.0007704
```

```
MathEnrollment_subset_all2 <- MathEnrollment_subset %>% mutate(studresids = rstudent(enroll_mod_subset2
                                                               fits = enroll_mod_subset2$fitted.values)

ggplot(MathEnrollment_subset_all2, aes(x=AYear, y=studresids)) +
    geom_point() +
    labs(title="(Studentized) Residuals vs temporal predictor", x="Year", y="Studentized residuals") +
    geom_hline(yintercept=0)
```



Solution:

(a) $\hat{Spring} = -11720 - 1.007(Fall) + 6.107(AYear)$

(b) yes, including AYear as a predictor seems to eliminate the trend between year and the residuals that we saw in the previous problem.

Rubric: Please grade this problem for completion.