Stat 230 HW 5

Name: Victor Huang

worked with: Eric, Katie, Nina, Jeanny

Homework 5 is due **by 3pm Thursday, Oct. 21**. Please complete the assignment in this Markdown document, filling in your answers and R code below. I didn't create answer and R chunk fields like I did with homework 1, but please fill in your answers and R code in the same manner as hw 1. Submit a hard copy of the **compiled pdf or word doc** either

- in class
- in drop-in office hours
- in the paper holder outside my CMC 222 office door

Tips for using Markdown with homework sets:

- Work through a problem by putting your R code into R chunks in this .Rmd. Run the R code to make sure it works, then knit the .Rmd to verify they work in that environment.
  - Make sure you load your data in the .Rmd and include any needed `library` commands.
- Feel free to edit or delete questions, instructions, or code provided in this file when producing your homework solution.
- For your final document, you can change the output type from `html_document` to `word_document` or `pdf_document`. These two to output types are better formatted for printing.
  - on maize: you may need to allow for pop-ups from this site
- If you want to knit to pdf while running Rstudio from your computer (*not* from maize), you will need a LaTeX compiler installed on your computer. This could be MiKTeX, MacTeX (mac), or TinyTex. The latter is installed in R: first install the R package `tinytex`, then run the command `tinytex::install_tinytex()` to install this software.
  - If you are using maize, you don't need to install anything to knit to pdf!

Problem 1: ANOVA

Below is the ANOVA table for the regression of percent bodyfat (%) on midarm, triceps, and thigh skinfold measurements (cm): $\mu(\text{bodyfat}|X) = \beta_0 + \beta_1\text{midarm} + \beta_2\text{triceps} + \beta_3\text{thigh}$. Use this output to answer the questions (a)-(f) that follow along with the fact that the total sum of squares SST for bodyfat is $SST = 495.3895$.

```
> bodyfat.lm<- lm(bodyfat ~ midarm + triceps + thigh)
> anova(bodyfat.lm)
Analysis of Variance Table
```

```
Response: bodyfat
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
   midarm  1   10.0516  10.0516  1.63433 0.2193400
  triceps  1  379.4037 379.4037 61.68860 0.0000007
    thigh  1        A?   7.5293  1.22421 0.2848944
Residuals 16   98.4049       B?
```

(1a)

Fill in the ?'s (A and B).

$A = 495.3895 - 10.0516 - 379.4037 - 98.4049 = 7.5293$ $B = \frac{98.4049}{16} = 6.15030625$

(1b)

How many observations (n) are in the data set?

$n = (16 + 1 + 1 + 1) + 1 = 20$

(1c)

What is the estimated model standard deviation $\hat{\sigma}$ for the full model?

$\hat{\sigma} = \sqrt{\frac{SSR}{n-(p+1)}} = \sqrt{\frac{98.4049}{20-(3+1)}} = \sqrt{\frac{98.4049}{16}} = 2.479981$

(1d)

What is the SSreg(midarm, triceps, thigh), the regression sum of squares for the regression of bodyfat on midarm, triceps, and thigh?

$SSreg_{midarm} = 495.3895 - 10.0516 = 485.3379, SSreg_{triceps} = 495.3895 - 379.4037 = 115.9858, SSreg_{thigh} = 495.3895 - 7.5293 = 487.8602$

(1e)

Suppose you want to test the significance of thigh in the model that already includes midarm and triceps. Use the information above to test this with an F test. State your null and alternative models, give the F test stat and p-value for this comparison, and give your conclusion.

$H_0 : \beta_{thigh} = 0; H_A : \beta_{thigh} \neq 0$

Looking at the table above we get a F-stat of 1.22421 for thighs. The corresponding p-value for this F-stat is 0.2848944. Since our p-value is significantly higher than 0.05, we fail to reject the null hypothesis. As such, we can conclude that the thigh's skin fold thickness is not associated with total body fat percentage.

(1f)

Suppose you want to test the significance of thigh and triceps in the model that already includes midarm. Use the information above to test this with an F test. State your null and alternative models, compute the F test stat and p-value for this comparison, and give your conclusion. (Note that the test stat/p-value are **not given in the table** - you will need to compute them from the info given.)

$H_0 : \beta_{thigh} = \beta_{tricep} = 0; H_A : \beta_{thigh}, \beta_{tricep} \neq 0$

$SSR(thigh, tricep) = 379.4037 + 7.5293 = 386.933$ $MSR(thigh, tricep) = \frac{98.4049}{16} = 6.150306$ $F = \frac{386.933/16}{6.150306} = \frac{24.18331}{6.150306} = 3.93205$

```
1- pf(3.93205, 2, 20)
[1] 0.03629525
```

$p - value = 0.03629525$

As we can see above, we calculate the F-value to be 3.93205. The corresponding p-value is shown to be 0.03629525. Since the p-value is less than 0.05, we can reject the null hypothesis and accept the alternative. As such we can conclude that at least one variable between thigh and tricep skinfold thickness is correlated with body fat percentage

Problem 2: Crab Claws ch.10 exercise 10

Show all work ("by hand") but use R to get the p-value.

$ModelDFforLarger = 32, RSS_{larger} = 5.99713$ $ModelDFforSmaller = 34, RSS_{smaller} = 8.38155$ $DF = 34 - 32 = 2$ $ExtraSS = 8.38155 - 5.99713 = 2.38442$ $F - stat = \frac{2.38442/2}{5.99713/32} = \frac{1.19221}{0.1874103} = 6.361497$

```
1-pf(6.361497, 2, 32)
[1] 0.004719606
```

$p - value = 0.004719606$

As we can see, we get a f-stat of 6.361497 and a corresponding p-value of 0.004719606. Since the p-value is less than 0.05, we are able to reject the null hypothesis and conclude that the slopes are different for the three species.

Problem 3: Brain Weights ch.10 exercise 12

Recall that the model fit in section 9.1.2 is `log(BrainWt)` on `log(BodyWt)`, `log(Litter)`, and `log(Gestation)`. (Use any log-base that you like). The data for this is found in `case0902`.

```
case0902 <- case0902
case0902_lm <- lm(log(Brain) ~ log(Body) + log(Litter) + log(Gestation), data = case0902)
anova(case0902_lm)
Analysis of Variance Table

Response: log(Brain)
               Df Sum Sq Mean Sq   F value    Pr(>F)
log(Body)       1 416.40  416.40 1847.4486 < 2.2e-16 ***
log(Litter)     1   8.69    8.69   38.5493  1.52e-08 ***
log(Gestation)  1   1.99    1.99    8.8132  0.003813 **
Residuals      92  20.74    0.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_{litter} = \beta_{gestation} = 0; H_A : \beta_{litter}, \beta_{gestation} \neq 0$

$SSR(litter, gestation) = 416.40 + 8.69 = 425.09$ $MSR(litter, gestation) = \frac{20.74}{92} = 0.2254348$ $F = \frac{425.09/92}{0.2254348} = \frac{4.620543}{0.2254348} = 20.49614$

```
1- pf(20.49614, 2, 92)
[1] 4.347346e-08
```

$p - value = 4.347346e - 08$

As we can see above, we calculate the F-value to be 20.49614. The corresponding p-value is shown to be 4.347346e-08. Since the p-value is less than 0.05, we can reject the null hypothesis and accept the alternative. As such we can conclude that at least one variable between gestation period and litter size is correlated with brain weight.

Problem 4: Wages and Race revisited

Refer back to the wages and race problem in homework 4.

(4a)

In R, fit the interaction model described below:

$$\begin{aligned}\mu(\log(WeeklyEarnings)) = \beta_0 &+ \beta_1 Educ + \beta_2 Exper + \beta_3 RaceNotBlack + \beta_4 MetStatus + \beta_5 regionNE \\ &+ \beta_6 regionS + \beta_7 regionW + \beta_8 regionNE \times RaceNotBlack \\ &+ \beta_9 regionS \times RaceNotBlack + \beta_{10} regionW \times RaceNotBlack\end{aligned}$$

Use an F test to test whether the effect of race (black/nonblack) on earnings of males differs by region, after controlling for race, region, education, experience, and metropolitan status. Write down the null and alternative hypotheses in terms of a mean function for log(earnings) (e.g. Null: $\mu(\log(WeeklyEarnings)) =$ ... vs. Alt: $\mu(\log(WeeklyEarnings)) =$ ...), then use R to do the F test of these hypotheses. State your conclusion, in context, for this test.

$$\begin{aligned}H_0 : \mu(\log(WeeklyEarnings)) = \beta_0 &+ \beta_1 Educ + \beta_2 Exper + \beta_3 RaceNotBlack + \beta_4 MetStatus + \beta_5 regionNE \\ &+ \beta_6 regionS + \beta_7 regionW\end{aligned}$$

$$\begin{aligned}H_a : \mu(\log(WeeklyEarnings)) = \beta_0 &+ \beta_1 Educ + \beta_2 Exper + \beta_3 RaceNotBlack + \beta_4 MetStatus + \beta_5 regionNE \\ &+ \beta_6 regionS + \beta_7 regionW + \beta_8 regionNE \times RaceNotBlack \\ &+ \beta_9 regionS \times RaceNotBlack + \beta_{10} regionW \times RaceNotBlack\end{aligned}$$

```
ex1029_lm <- lm(log(WeeklyEarnings) ~ MetropolitanStatus + Exper + Educ + Race + Region, data = ex1029)
ex1029_lm_new <- lm(log(WeeklyEarnings) ~ MetropolitanStatus + Exper + Educ + Race*Region, data = ex102
anova(ex1029_lm_new)
Analysis of Variance Table

Response: log(WeeklyEarnings)
                     Df Sum Sq Mean Sq   F value Pr(>F)
MetropolitanStatus    1  191.9  191.89  686.7350 <2e-16 ***
Exper                 1  505.6  505.61 1809.4958 <2e-16 ***
Educ                  1 1954.7 1954.70 6995.5971 <2e-16 ***
Race                  1  116.1  116.15  415.6831 <2e-16 ***
```

```
Region                3   32.4   10.80   38.6579 <2e-16 ***
Race:Region           3    0.3    0.11    0.3968 0.7553
Residuals         25426 7104.5    0.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the data above, we get a F-value of 0.3968 and a p-value of 0.7553. Since our p-value is larger than 0.05, we fail to reject the null hypothesis. As such, we conclude that region does not affect the income gap by race.

(4b)

Fit the no interaction model (below) and use it to interpret the effect that race (black vs. nonblack) has on earnings (original scale, not logged scale) after controlling for all other predictors, and give a confidence interval for this effect too.

$$\mu(\log(WeeklyEarnings)) = \beta_0 + \beta_1 Educ + \beta_2 Exper + \beta_3 RaceNotBlack + \beta_4 MetStatus + \beta_5 regionNE + \beta_6 regionS + \beta_7 regionW$$

```
ex1029_lm_new2 <- lm(log(WeeklyEarnings) ~ MetropolitanStatus + Exper + Educ + Region + Race, data = ex
summary(ex1029_lm_new2)

Call:
lm(formula = log(WeeklyEarnings) ~ MetropolitanStatus + Exper +
    Educ + Region + Race, data = ex1029)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6881 -0.2997  0.0432  0.3425  3.7050

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         4.5138725  0.0221711 203.593  < 2e-16
MetropolitanStatusNotMetropolitanArea -0.1592142  0.0076933 -20.695  < 2e-16
Exper                                0.0178191  0.0002798  63.677  < 2e-16
Educ                                 0.0971174  0.0011958  81.213  < 2e-16
RegionNortheast                      0.0370482  0.0097032   3.818 0.000135
RegionSouth                         -0.0597360  0.0090537  -6.598 4.25e-11
RegionWest                          -0.0026049  0.0098397  -0.265 0.791215
RaceNotBlack                         0.2331872  0.0126232  18.473  < 2e-16

(Intercept)                         ***
MetropolitanStatusNotMetropolitanArea ***
Exper                               ***
Educ                                ***
RegionNortheast                     ***
RegionSouth                         ***
RegionWest
RaceNotBlack                        ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5286 on 25429 degrees of freedom
```

```
Multiple R-squared:  0.2827,    Adjusted R-squared:  0.2825
F-statistic:  1432 on 7 and 25429 DF,  p-value: < 2.2e-16
confint(ex1029_lm_new2)
                                           2.5 %      97.5 %
(Intercept)                             4.47041596  4.55732904
MetropolitanStatusNotMetropolitanArea -0.17429355 -0.14413476
Exper                                   0.01727066  0.01836764
Educ                                    0.09477345  0.09946129
RegionNortheast                         0.01802931  0.05606708
RegionSouth                            -0.07748174 -0.04199023
RegionWest                             -0.02189121  0.01668138
RaceNotBlack                            0.20844504  0.25792926
```

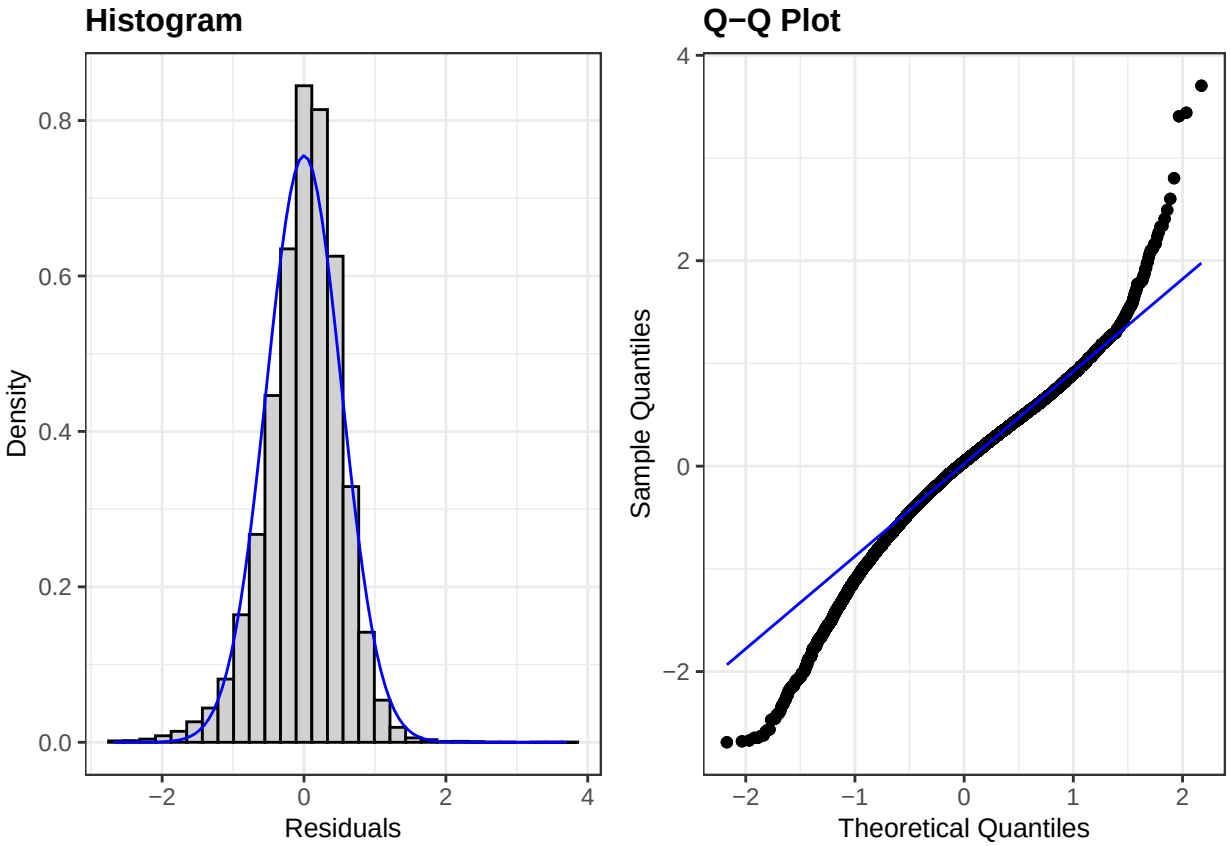
$e^{\mu(\log(WeeklyEarnings))=\beta_0+\beta_1 Educ+\beta_2 Exper+\beta_3 RaceNotBlack+\beta_4 MetStatus+\beta_5 regionNE+\beta_6 regionS+\beta_7 regionW} = e^{0.2436034}$

As we can see, if race is not black, we get a multiplicative increase of $e^{0.2436034}$ on the median of weekly earnings which corresponds to a 26.26% increase an median weekly earnings. Our CI for the multiplicative change of not being black is from $e^{0.20844504}$ to $e^{0.25792926}$

(4c)

Describe the distribution of the residuals for the model given in part (b) with both a histogram and normal qq plot. Our modeling goal is to explore the effect of race (black/notblack) on mean earnings of males after controlling for region, education, experience, and MetStatus. With this in mind, is the distribution of these residuals concerning? Explain.

```
resid_panel(ex1029_lm_new, plots = c("hist","qq"))
```

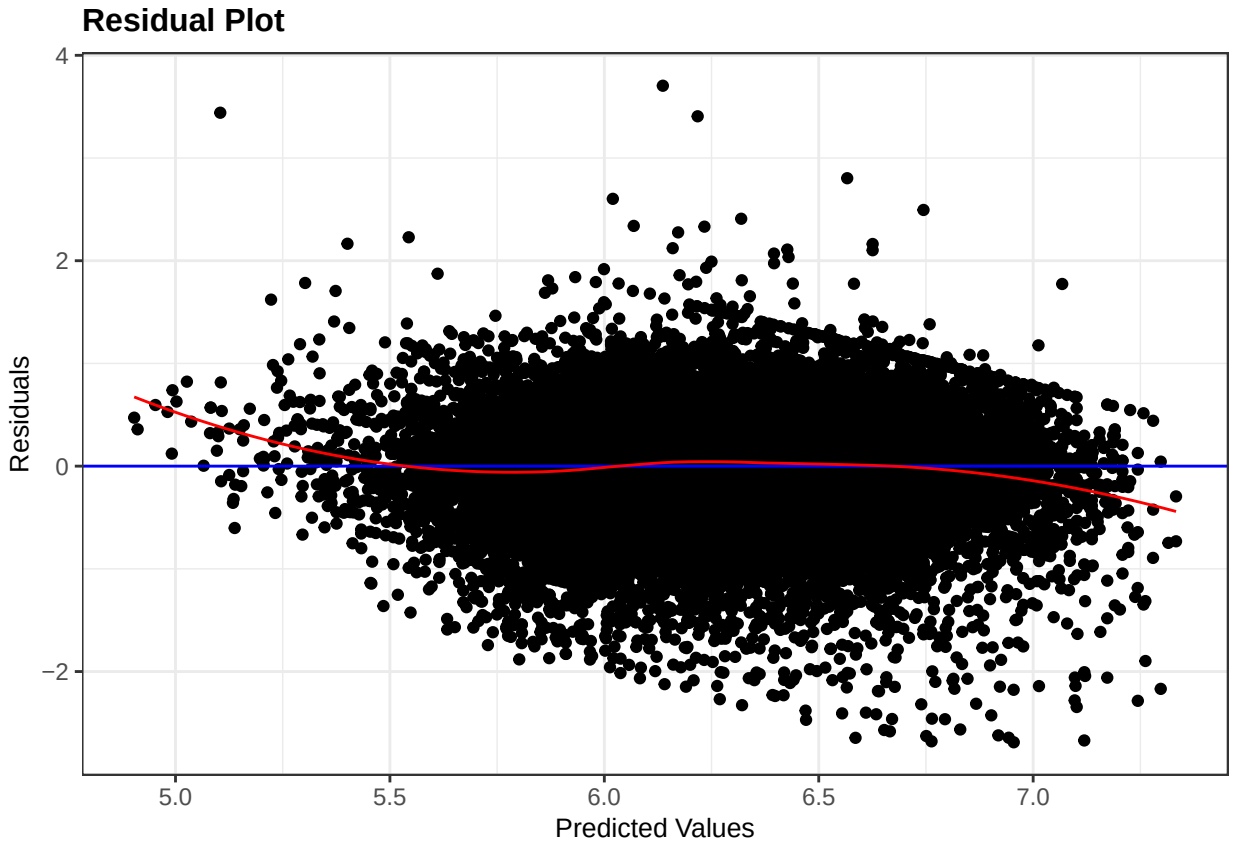


The residuals are not concerning, even though the qq plot if not normally distributed, this not important as we will only be finding the mean and not making a prediction.

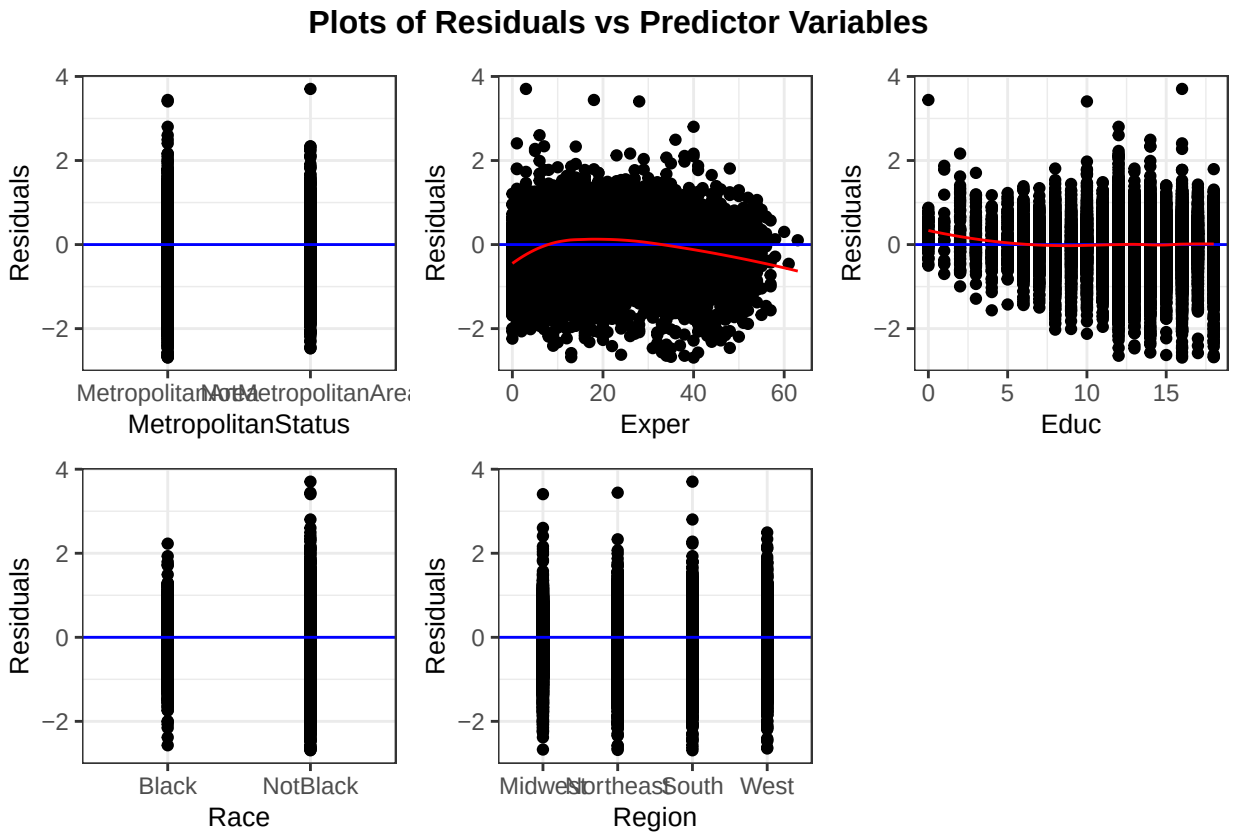
(4d)

Using the `ggResidpanel` package, create the six possible residual plots for this model (fitted + 5 predictors). For each, comment on the linearity and constant variance assumptions made for this model. This is a large data set (with lots of residuals), so add `smoother = TRUE` to add a smoother line to help detect nonlinearity in the overlapping points in your residual plots.

```
resid_panel(ex1029_lm_new, plots = "resid", smoother = TRUE)
`geom_smooth()` using formula 'y ~ x'
```




```
resid_xpanel(ex1029_lm_new, smoother = TRUE)
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
```



1. Linearity is up held, looking at the redline we see curvature, as such, there is deviance in the constant variance. 2. Linearity is held, constant variance changes due to metstatus. 3. Linearity is held, looking at the redline we see curvature, as such, there is deviance. 4. Linearity is held, but since we can observe curvature in the redline, we see that constant variance as deviations. 5. Linearity is upheld, constant variance changes due to race status. 6. Linearity is upheld, constant variance changes due to region status.

(4e)

One way to "test" for curvature is to add a quadratic term to your model. Since part (d) suggest a nonlinear effect of experience, add a quadratic term for experience to the linear model above and fit this model to the data. Use the t-test results to determine whether the nonlinear effect of experience is significant.

```
ex1029_new_lm2 <- lm(WeeklyEarnings ~ MetropolitanStatus + Exper + Educ + Region + Race + I(Exper^2), da
summary(ex1029_new_lm2)

Call:
lm(formula = WeeklyEarnings ~ MetropolitanStatus + Exper + Educ +
    Region + Race + I(Exper^2), data = ex1029)

Residuals:
    Min      1Q  Median      3Q     Max
-1047.5  -199.8   -48.1   133.4 18312.7

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -530.8632    16.5591 -32.059  < 2e-16 ***
MetropolitanStatusNotMetropolitanArea -105.8196     5.6572 -18.705  < 2e-16 ***
```

```
Exper                                29.2193     0.6935  42.131  < 2e-16 ***
Educ                                 56.8221     0.8922  63.689  < 2e-16 ***
RegionNortheast                      22.7718     7.1348   3.192 0.001416 **
RegionSouth                         -24.2219     6.6566  -3.639 0.000274 ***
RegionWest                           11.2414     7.2360   1.554 0.120309
RaceNotBlack                        132.3940     9.2809  14.265  < 2e-16 ***
I(Exper^2)                           -0.4196     0.0149 -28.159  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 388.6 on 25428 degrees of freedom
Multiple R-squared:  0.2342,    Adjusted R-squared:  0.234
F-statistic: 972.3 on 8 and 25428 DF,  p-value: < 2.2e-16
```

Since the p-value for the quadratic experience term is 2e-16 and less than the 0.05 treshold. We can conclude that the n quadratic term is indeed significant.

(4f)

Using your model from (e), report the case numbers of the cases with the highest leverage, studentized residuals and Cook's distance values. Use the data for these cases and basic EDA to explain why their respective case influence stat is high. Then explain why none of these cases need to be removed from our data to adequately model earnings.

If you'd like to use the `augment`ed data frame to find the row number of these "max" case influence stats, I suggest that you do the following

- Add **row numbers** to your data set, e.g. like this using `dplyr`:

```
ex1029 <- ex1029 %>% mutate(case = row_number())
ex1029_aug <- augment(ex1029_lm, data = ex1029)
```

Largest cook's distance is 0.0687246720 at case 17962, highest leverage is 0.006503881 at case 22486, largest studentized residuals is 46.155157 case 17962,

```
ex1029_newnew <- ex1029 %>% slice (-17962, -22486, -17962)
ex1029_newnew_lm <- lm(WeeklyEarnings ~ MetropolitanStatus + Exper + Educ + Region + Race, data = ex102
summary(ex1029_lm)

Call:
lm(formula = log(WeeklyEarnings) ~ MetropolitanStatus + Exper +
    Educ + Race + Region, data = ex1029)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6881 -0.2997  0.0432  0.3425  3.7050

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        4.5138725  0.0221711 203.593  < 2e-16
MetropolitanStatusNotMetropolitanArea -0.1592142  0.0076933 -20.695  < 2e-16
Exper                              0.0178191  0.0002798  63.677  < 2e-16
```

```
Educ                                 0.0971174  0.0011958  81.213  < 2e-16
RaceNotBlack                         0.2331872  0.0126232  18.473  < 2e-16
RegionNortheast                      0.0370482  0.0097032   3.818 0.000135
RegionSouth                         -0.0597360  0.0090537  -6.598 4.25e-11
RegionWest                          -0.0026049  0.0098397  -0.265 0.791215

(Intercept)                         ***
MetropolitanStatusNotMetropolitanArea ***
Exper                               ***
Educ                                ***
RaceNotBlack                        ***
RegionNortheast                     ***
RegionSouth                         ***
RegionWest
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5286 on 25429 degrees of freedom
Multiple R-squared:  0.2827,    Adjusted R-squared:  0.2825
F-statistic:  1432 on 7 and 25429 DF,  p-value: < 2.2e-16
summary(ex1029_newnew_lm)

Call:
lm(formula = WeeklyEarnings ~ MetropolitanStatus + Exper + Educ +
    Region + Race, data = ex1029)

Residuals:
    Min      1Q  Median      3Q     Max
-1087.6  -209.6   -50.4   141.7 18211.9

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -448.8142    16.5526 -27.114  < 2e-16 ***
MetropolitanStatusNotMetropolitanArea -103.0315     5.7437 -17.938  < 2e-16 ***
Exper                                 10.5689     0.2089  50.588  < 2e-16 ***
Educ                                  61.0902     0.8928  68.425  < 2e-16 ***
RegionNortheast                       19.7959     7.2443   2.733 0.006288 **
RegionSouth                          -23.3496     6.7594  -3.454 0.000552 ***
RegionWest                            15.6392     7.3461   2.129 0.033272 *
RaceNotBlack                         131.3797     9.4243  13.941  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 394.6 on 25429 degrees of freedom
Multiple R-squared:  0.2104,    Adjusted R-squared:  0.2101
F-statistic: 967.8 on 7 and 25429 DF,  p-value: < 2.2e-16
```

Looking at the the two r-squared values, we do not see significant difference. Since there s no chagne in r-squared and experience, there s no need to remove these data points.

Problem 5: Agstrat revisited

Revisit Homework 4 problem 3 that has you fitting a parallel line models. Use an ANOVA F test to determine if `region` is a statistically significant predictor of `acres92` after accounting for `acres82`.

```
agstrat <- read.csv("http://people.carleton.edu/~kstclair/data/agstrat.csv")
agstrat_new <- agstrat %>% slice(-119,-168)
agstrat_lm <- lm(log(acres92) ~ log(acres82) + region, data = agstrat_new)
anova(agstrat_lm)
Analysis of Variance Table

Response: log(acres92)
             Df Sum Sq Mean Sq    F value  Pr(>F)
log(acres82)   1 363.18  363.18 18637.3370 < 2e-16 ***
region         3   0.14    0.05     2.4229 0.06602 .
Residuals    293   5.71    0.02
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the code and results above, we get a F-value of 2.4229 and a p-value of 0.06602. Since our p-values is larger than 0.05, we fail to reject the null hypothesis and conclude that region is not a statistically significant predictor of acres in 92 after accounting for acres in 82.

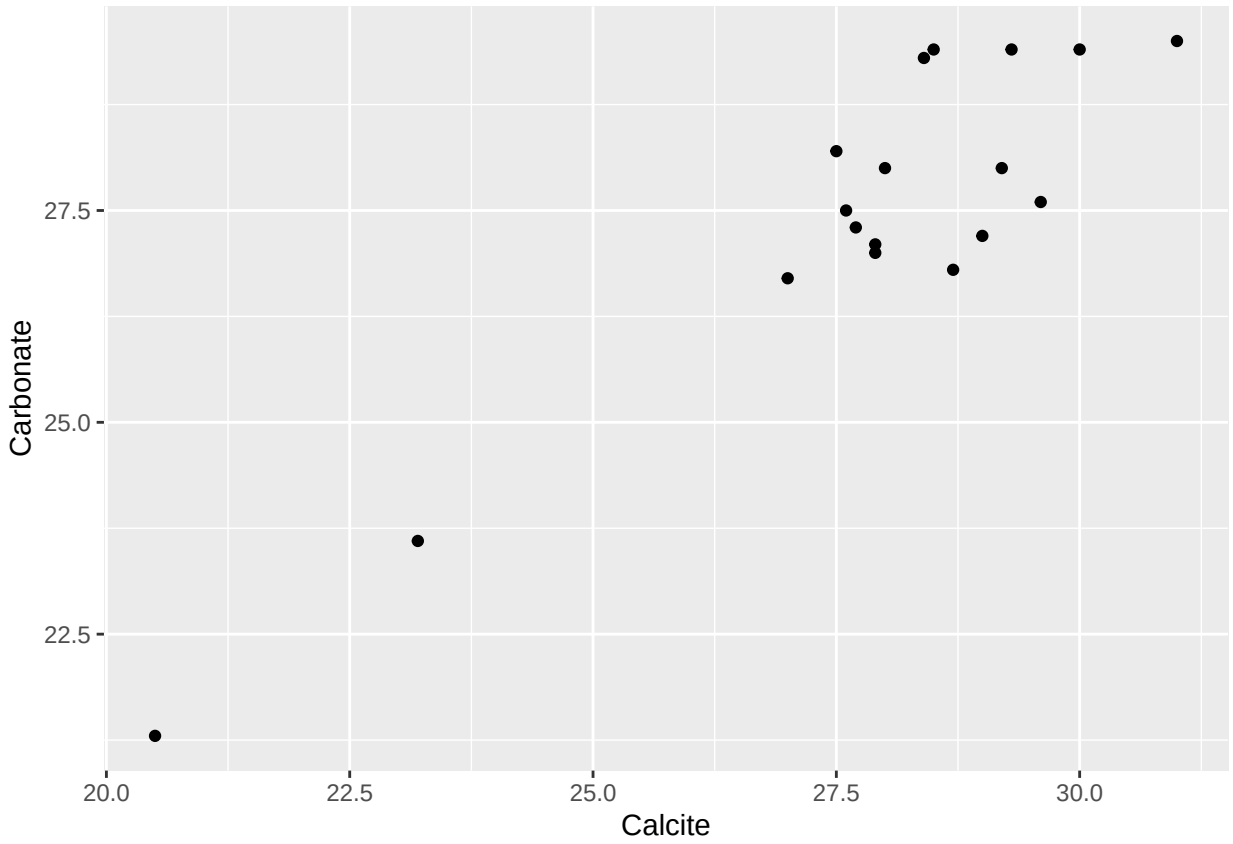
Problem 6: Warm-bloded T. Rex?

Consider the data in ch. 11 exercise 20 (`ex1120`). Review the background info provided for this exercise but answer the questions below.

(6a)

Create a scatterplot of `Calcite` against `Carbonate`. Identify by row number the two cases that are obvious outliers.

```
ex1120 <- ex1120
ggplot(ex1120, aes(x=Calcite, y=Carbonate)) + geom_point()
```



The row numbers for the two outliers are rows 1 and 2.

(6b)

Using all 18 data cases, fit the regression of `Calcite` on `Carbonate`. Report the following info:

- Slope, SE for slope
- p-value for slope
- 95% CI for slope
- R^2

```
ex1120_lm <- lm(Calcite ~ Carbonate, data = ex1120)
summary(ex1120_lm)

Call:
lm(formula = Calcite ~ Carbonate, data = ex1120)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46796 -0.64104 -0.04927  0.67301  1.55856

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4984     3.1766  -0.472    0.644
Carbonate     1.0703     0.1156   9.259 7.93e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.9959 on 16 degrees of freedom
Multiple R-squared:  0.8427,    Adjusted R-squared:  0.8329
F-statistic: 85.73 on 1 and 16 DF,  p-value: 7.929e-08
confint(ex1120_lm)
                2.5 %   97.5 %
(Intercept) -8.2324764 5.235629
Carbonate    0.8252384 1.315332
```

Slope = 1.0702, SE for Slope = 0.1156, p-value = 7.93e-08, 95% CI = 0.8252384, 1.315332 ### (6c) Repeat (b) but omit the case with the smallest `Carbonate` value.

```
data_c <- filter(ex1120, Carbonate > 23)
ex1120_lm_c <- lm(Calcite ~ Carbonate, data = data_c)
summary(ex1120_lm_c)

Call:
lm(formula = Calcite ~ Carbonate, data = data_c)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2799 -0.4816 -0.1364  0.7184  1.4871

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6727     4.6247   0.578    0.572
Carbonate     0.9217     0.1663   5.541 5.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9807 on 15 degrees of freedom
Multiple R-squared:  0.6718,    Adjusted R-squared:  0.6499
F-statistic:  30.7 on 1 and 15 DF,  p-value: 5.653e-05
confint(ex1120_lm_c)
                2.5 %    97.5 %
(Intercept) -7.1845567 12.529928
Carbonate    0.5671886  1.276304
```

Slope = 0.9217, SE of Slope: 0.1663, p-value = 5.65e-05, 95% CI = 0.5671886, 1.276304

(6d)

Repeat (b) but omit the cases with the smallest two `Carbonate` values.

```
data_d <- filter(ex1120, Carbonate > 24)
ex1120_lm_d <- lm(Calcite ~ Carbonate, data = data_d)
summary(ex1120_lm_d)

Call:
lm(formula = Calcite ~ Carbonate, data = data_d)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.1844 -0.7038 -0.1139  0.6854  1.5492

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.0589     6.1592   1.958   0.0705 .
Carbonate     0.5896     0.2196   2.684   0.0178 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8875 on 14 degrees of freedom
Multiple R-squared:  0.3398,    Adjusted R-squared:  0.2926
F-statistic: 7.205 on 1 and 14 DF,  p-value: 0.0178
confint(ex1120_lm_d)
                 2.5 %    97.5 %
(Intercept) -1.1513854 25.269098
Carbonate    0.1184916  1.060627
```

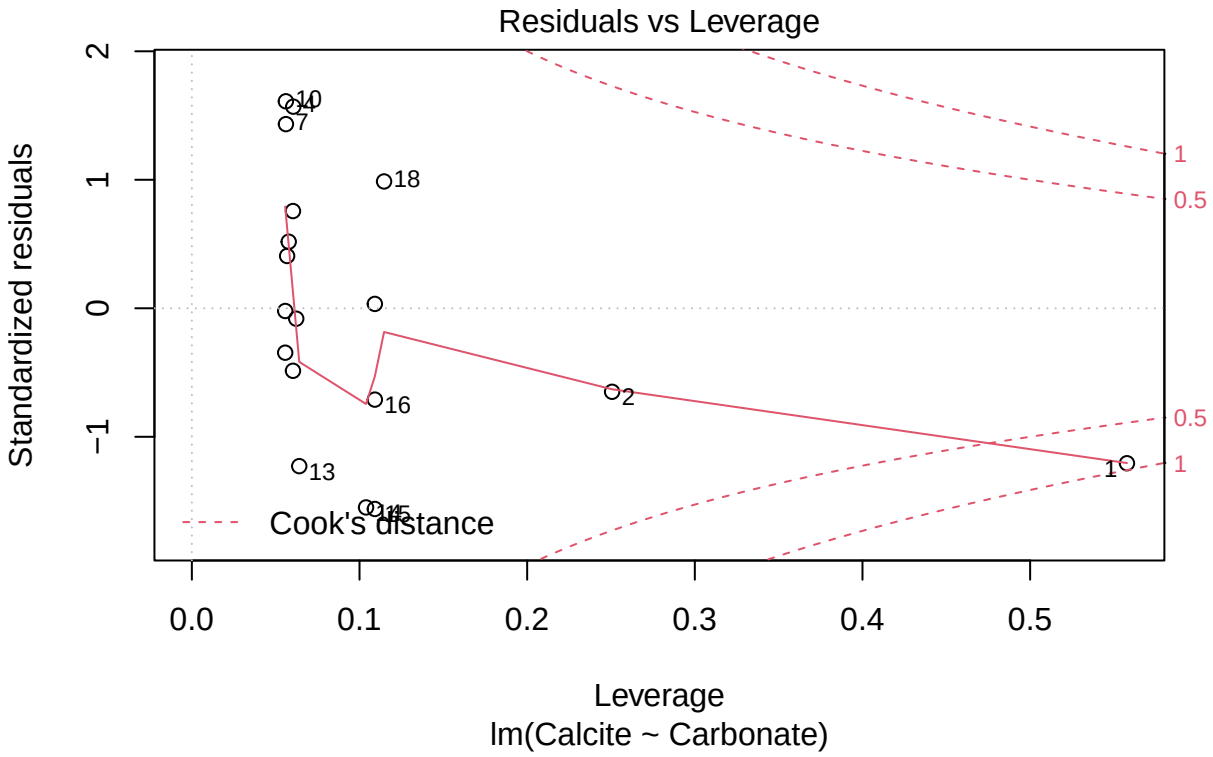
Slope = 0.5896, SE of Slope: 0.2196, p-value = 0.0178, 95% CI = 0.1184916, 1.060627

(6e)

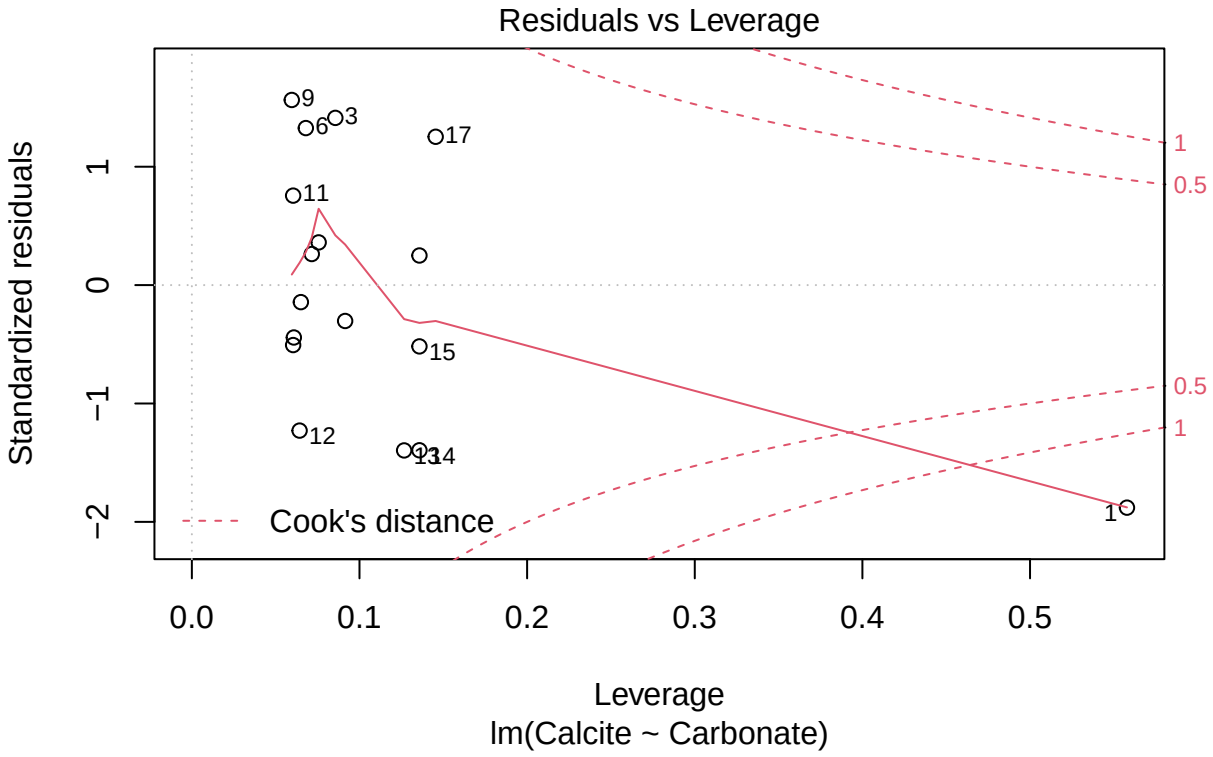
For your three models in (b), (c) and (d), provide the case-influence plot: `plot(my_lm, which=5, id.n=10)` and describe which case, or cases, have values of leverage, studentized residual, or Cook's distance beyond the usual threshold used to flag potentially unusual cases. Note that the row numbers shown in `plot` will be the row number of the *full* dataset with 18 data points, even if used `subset` to remove case(s) in the `lm` command.

Case 1 and 2 (the plots for the dataset with 2 and 1 outlier(s)) have Cook's distances beyond the usual threshold.

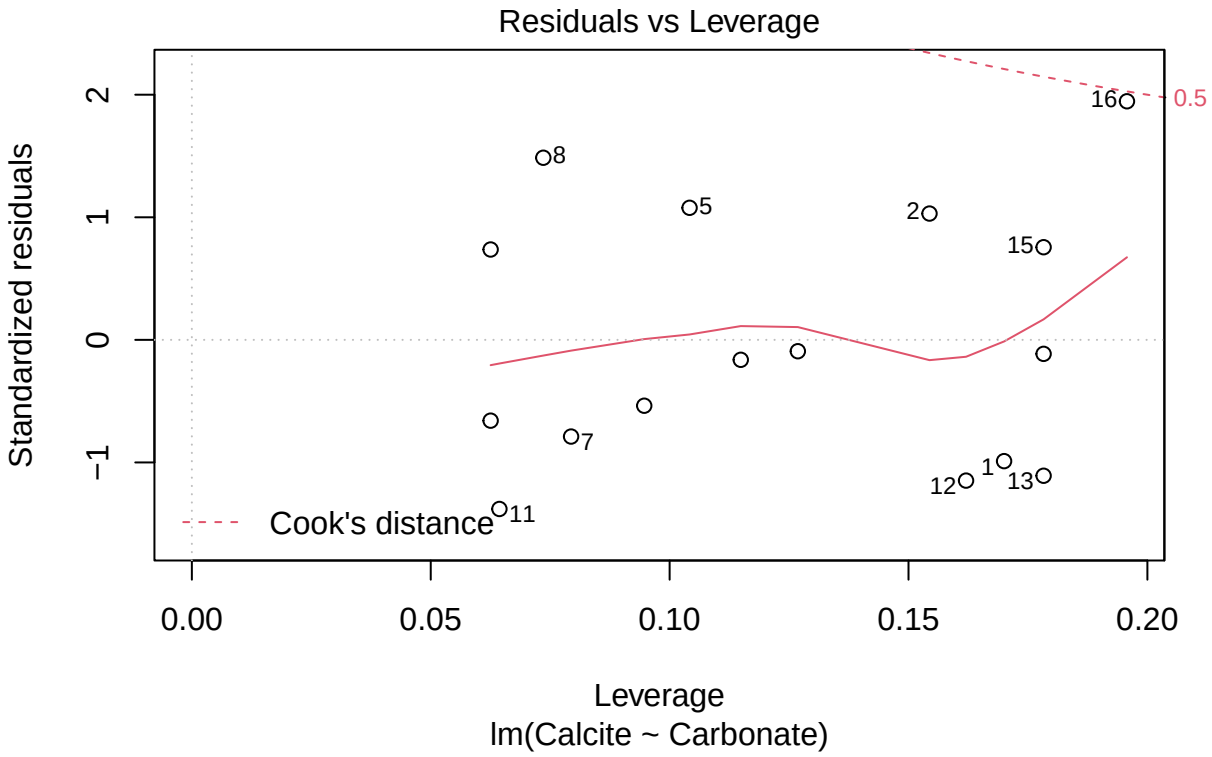
```
plot(ex1120_lm, which=5, id.n=10)
```




```
plot(ex1120_lm_c, which=5, id.n=10)
```




```
plot(ex1120_lm_d, which=5, id.n=10)
```



(6f)

Explain why the case influence stats for case 2 change so much between model (b) and model (c). How might pairs of influential cases not be found with the usual case influence stats?

Because for case 1 there are two outliers, as such, the "weight" of each outlier becomes less significant as both show their influence on the total. With one of the outliers removed, all the influence falls onto the remaining point, making it much more influential and the stas change so much between the two cases.

(6g)

Use your results reported in (b) and (d) to justify why cases 1 and 2 are influential. How is the slope for `Carbonate` values between 25 and 30 different from the slope when we don't restrict the range of `Carbonate`?

The r-squared value dropped significantly, the two points are overly influential. When we do restrict the range of Carbonate, there is a weaker response, as such, our slop drops.