PM592: Regression Analysis for Health Data Science
Lab 7 – Advanced Variable Coding
Data Needed: *ex_sos*

This lab is devoted entirely to the exercise.

Lab 7 Exercises

Objective(s):	Determine the functional form of independent variables, build an association model from start to finish
Datasets Required:	ex_sos

This data set re-creates the findings from the Sources of Strength suicide prevention intervention. Baseline data was acquired on students participating in one of the intervention schools (n=1000). These students answered surveys that included questions such as:

- Trusted adults (TATOT): Name up to 7 adults at school who you trust and feel that you can talk to openly. This variable is the sum of the number of adults they named.
- Maladaptive help-seeking (M_HELP): A series of questions asking them about their help-seeking attitudes. This survey consisted of 6 questions that were combined into a composite score. Higher scores reflected students were more maladaptive with regard to help-seeking for problems—that is, they viewed seeking help for themselves and/or their friends more unfavorably.
- Age, in years (AGE)
- Ethnicity (ETH): classified as white (W), black (B), or other (O).
- Gender (GENDER): classified as 1=female, 0=male

Researchers were interested in whether more adult connections at school improved students' attitudes toward seeking help. Additionally, they were interested in whether this effect was even stronger for students who named no adults vs. naming any adults. They wanted to adjust for age, ethnicity, and gender.

1) Begin this analysis by producing the following descriptive information:
   a) To make the scale of m_help more interpretable, create a variable called m_help.z which is the z-score for m_help.

```
> sos <- sos %>%
+   mutate(m_help.z = scale(m_help)[,1])
```

   b) Summary statistics for all variables in the data set. Ensure that the ranges make sense and examine any outliers.

```
> skim(ex_sos)
── Data Summary ────────────────────────────────────────
                           Values
Name                       ex_sos
Number of rows             1000
Number of columns          6
_____________________
Column type frequency:
  character                1
  numeric                  5
_____________________
Group variables            None

── Variable type: character ────────────────────────────
```

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
1 eth	0	1	1	1	0	3	0

Variable type: numeric

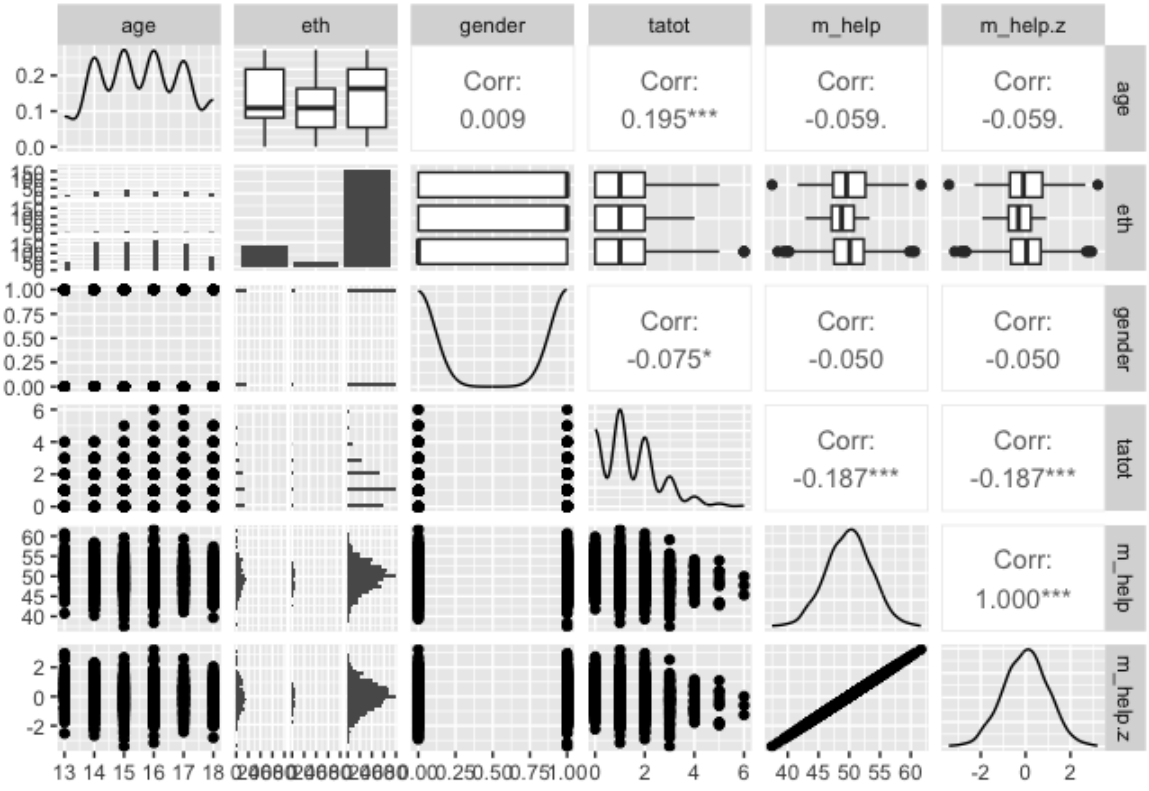
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1 age	0	1	1.56e+ 1	1.44	13	14	16	17	18	
2 gender	0	1	5.04e- 1	0.500	0	0	1	1	1	
3 tatot	0	1	1.34e+ 0	1.17	0	0	1	2	6	
4 m_help	0	1	4.99e+ 1	3.64	37.5	47.5	50.0	52.3	61.6	
5 m_help.z	0	1	-5.54e-16	1	-3.42	-0.668	0.0143	0.661	3.22	

Yes, the variables make sense. The average age is 15.6, about there is about an even split on males and females. The average maladaptive help-seeking score is 5, and there are a few very high values with the maximum being 61.6.

c) Does anything concern you about the distributions of these variables?

Trusted adults variable is skewed right, but this is not surprising. Nothing is reallt concerning about the distributions of the variables.

2) Run ggpairs() on all variables in the data set.



a) Does anything concern you about the associations among all these variables?

No, nothing sticks out as concerning. m_help is correlated with m_help.z but this is because they measure the same thing.

b) Do any variables appear to be collinear?

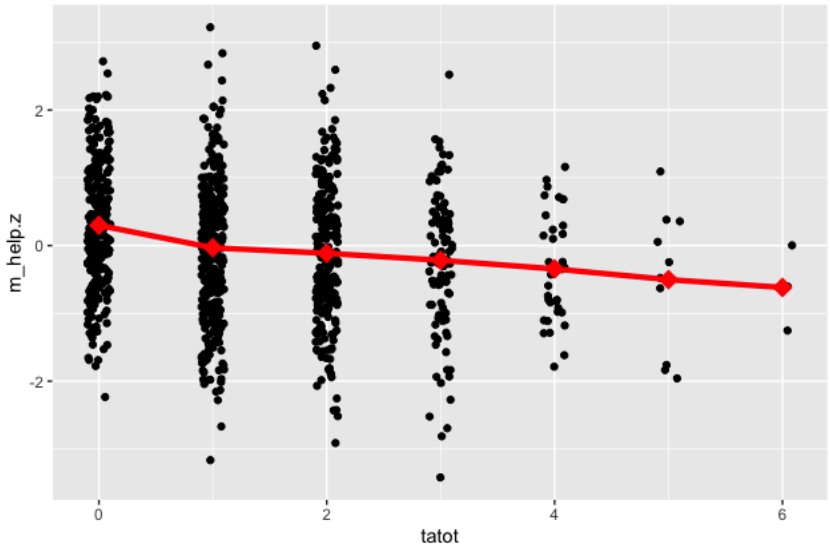
No, none of the variables are correlated at a high level, the highest correlation is between age and trusted adults at R = 0.195.

3) Examine the effect of trusted adults – keep in mind the research question: whether maladaptive attitudes are even higher for those who named no adults.

a) Examine the mean of m_help.z for each value of tatot.

```
> sos %>%
+  group_by(tatot) %>%
+  summarize(mean(m_help.z))
# A tibble: 7 × 2
  tatot `mean(m_help.z)`
  <dbl>            <dbl>
1     0            0.299
2     1          -0.0325
3     2           -0.114
4     3           -0.217
5     4           -0.344
6     5           -0.502
7     6           -0.618
```

b) Produce a graph of m_help.z by tatot that includes the mean of m_help.z for each value of tatot (this is in the lab R file).



c) Does the difference in mean m_help.z associated with 1 vs 0 trusted adults appear greater than the difference associated with 2 vs 1 trusted adults?

Yes it does. The slope is steeper going from 0 to 1 than 1 to 2.

(Questions 3d-3j are optional for advanced modeling practice)

d) Create the following two variables as a set to describe the effect of trusted adults:

$$X_{NOAD} = \begin{cases} 1, \text{if tatot} == 0 \\ 0, \text{otherwise} \end{cases} \qquad X_{TATOT1} = \begin{cases} (\text{tatot} - 1), \text{if tatot} > 0 \\ 0, \text{otherwise} \end{cases}$$

e) What is the baseline value for this variable set?

f) What will β_{NOAD} represent?

g) What will β_{TATOT1} represent?

h) Fit the regression model of m_help.z on noad and tatot1. What is your output?

i) Use the Extra Sums of Squares test to determine whether this coding scheme fits better than a model in which trusted adults is treated completely linear (tatot).

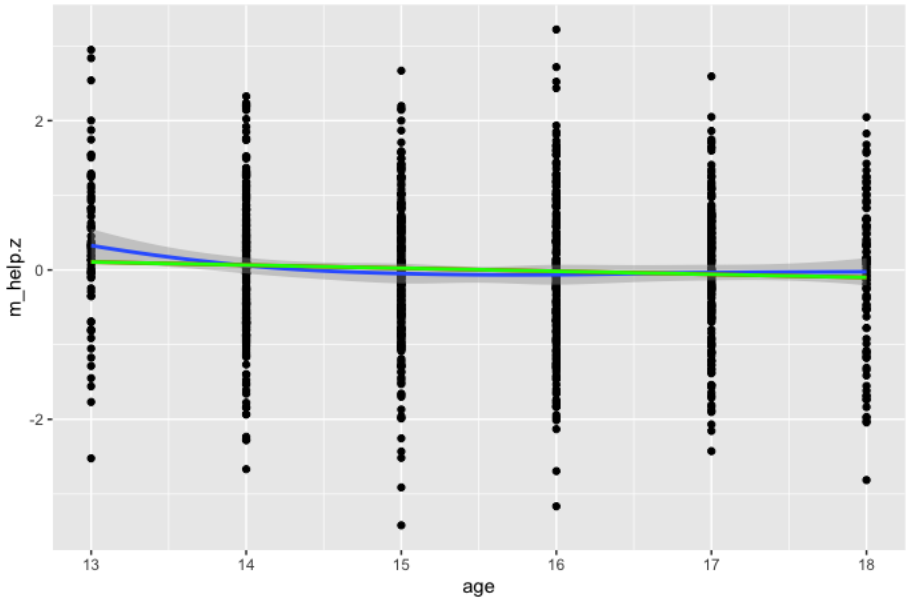
j) Given this information, how will you model trusted adults?

4) Examine the effect of age

a) Examine the mean of m_help.z for each value of age.

```
> sos %>%
+   group_by(age) %>%
+   summarise(mean(m_help.z))
# A tibble: 6 × 2
    age `mean(m_help.z)`
  <dbl>            <dbl>
1    13           0.340
2    14           0.0420
3    15          -0.0479
4    16          -0.0637
5    17          -0.0192
6    18          -0.0317
```

b) Produce a graph of m_help.z by age that includes a locally-weighted "loess" smoothed line.



c) Does age appear to have a linear effect on m_help.z? If not, explain what the relationship might be.

The locally-weighted line shows that age has a slightly non-linear effect on m_help.z. However, the linear and quadratic functions overlayed on the graph are almost the same.

d) Fit the regression model of m_help.z on however you decided to code age. What is your output?

```
> lm(m_help.z ~ age.c, data = sos) %>% summary()


Call:
lm(formula = m_help.z ~ age.c, data = sos)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4437 -0.6631  0.0196  0.6747  3.2376

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.783e-16  3.158e-02   0.000   1.0000
age.c       -4.131e-02  2.199e-02  -1.879   0.0606 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9987 on 998 degrees of freedom
Multiple R-squared:  0.003524,   Adjusted R-squared:  0.002525
F-statistic: 3.529 on 1 and 998 DF,  p-value: 0.06059
```

e) Given this information, how will you model age?

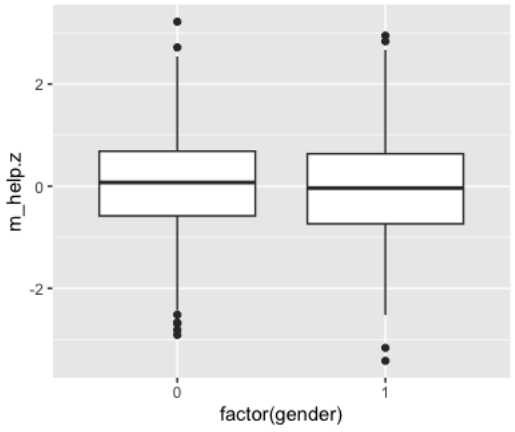
Since the hierarchical and fractional polynomials approach suggested quadratic and linear respectively, age will be modeled linearly for the sake of parsimony.

5) Examine the effect of gender

a) Examine the mean of m_help.z for each value of gender.

```
> sos %>%
+   group_by(gender) %>%
+   summarise(mean(m_help.z))
# A tibble: 2 × 2
  gender `mean(m_help.z)`
   <dbl>            <dbl>
1      0           0.0507
2      1          -0.0499
```

b) Produce a graph or boxplot of m_help.z by gender.



c) Determine whether m_help.z varies by gender using whichever method you like.

```
> lm(m_help.z ~ gender, data = sos) %>% summary()

Call:
lm(formula = m_help.z ~ gender, data = sos)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3698 -0.6542  0.0173  0.6500  3.1697

Coefficients:
            Estimate Std. Error t value
(Intercept)  0.05072    0.04487   1.130
gender      -0.10063    0.06320  -1.592
            Pr(>|t|)
(Intercept)    0.259
gender         0.112

Residual standard error: 0.9992 on 998 degrees of freedom
Multiple R-squared:  0.002534,   Adjusted R-squared:  0.001535
F-statistic: 2.535 on 1 and 998 DF,  p-value: 0.1116
```

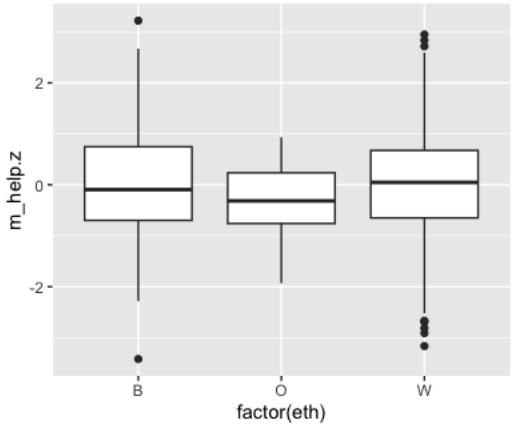
No, the p-value for the gender coefficient is > 0.05 so m_help.z does not vary significantly by gender.

6) Examine the effect of ethnicity

a) Examine the mean of m_help.z for each value of ethnicity.

```
# A tibble: 3 × 2
  eth   `mean(m_help.z)`
  <chr>            <dbl>
1 B              -0.0138
2 O              -0.287
3 W               0.0166
```

b) Produce a graph or boxplot of m_help.z by ethnicity.



c) Do you think that the "other" ethnicity category should be combined with any other group? Consider the sample size in that category, and how similar it is to other groups.

Although the n is is still large enough to have some meaningful results (38), and the mean m_help.z value is very different from B or W, so I think that it should be kept secret.

d) Determine whether m_help.z varies by ethnicity using whichever method you like.

```
> lm(m_help.z ~ as.factor(eth), data = sos) %>% summary()

Call:
lm(formula = m_help.z ~ as.factor(eth), data = sos)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4059 -0.6592  0.0087  0.6545  3.2342

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.01379    0.07733  -0.178    0.859
as.factor(eth)O -0.27277    0.17961  -1.519    0.129
as.factor(eth)W  0.03038    0.08506   0.357    0.721

Residual standard error: 0.9993 on 997 degrees of freedom
Multiple R-squared:  0.003374,   Adjusted R-squared:  0.001375
F-statistic: 1.688 on 2 and 997 DF,  p-value: 0.1855
```

No, the p-value for the variable set as a whole is 0.1855, which is above the alpha level of 0.05.

7) Combined model

a) Include all covariates of interest with the effect for trusted adults in one model.

```
> lm(m_help.z ~ tatot + age + gender + eth, data = sos) %>% summary()

Call:
lm(formula = m_help.z ~ tatot + age + gender + eth, data = sos)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1676 -0.6780 -0.0059  0.6558  3.1231

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.51875    0.34793   1.491   0.1363
tatot       -0.16045    0.02719  -5.902 4.92e-09 ***
age         -0.01632    0.02203  -0.741   0.4591
gender      -0.12493    0.06226  -2.007   0.0451 *
```

ethO -0.26181 0.17639 -1.484 0.1381

ethW 0.03003 0.08349 0.360 0.7192

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9808 on 994 degrees of freedom

Multiple R-squared: 0.04289, Adjusted R-squared: 0.03807

F-statistic: 8.908 on 5 and 994 DF, p-value: 2.747e-08

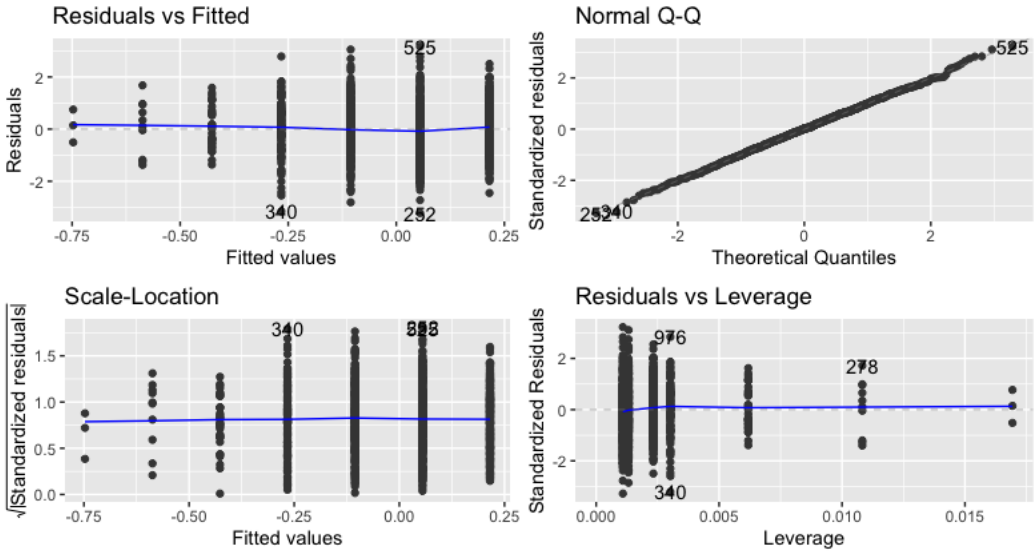
(-0.16052 - -0.16045) / (-0.16052) = 0.04% change → none of the covariates changed the tatot coefficient estimate by much.

b) Establish your "preliminary final model" – a model you're content with evaluating

$$\hat{Y}_{mhelp} = 0.21558 - 0.16052X_{tatot}$$

8) Model diagnostics

a) Examine the assumptions of linear regression. Are these assumptions met?

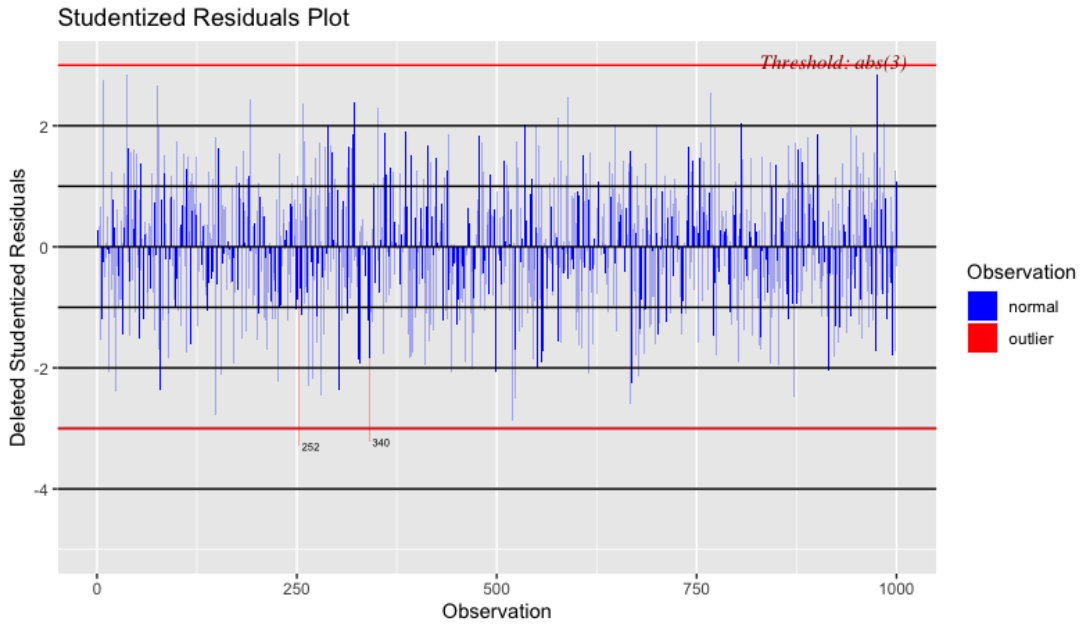


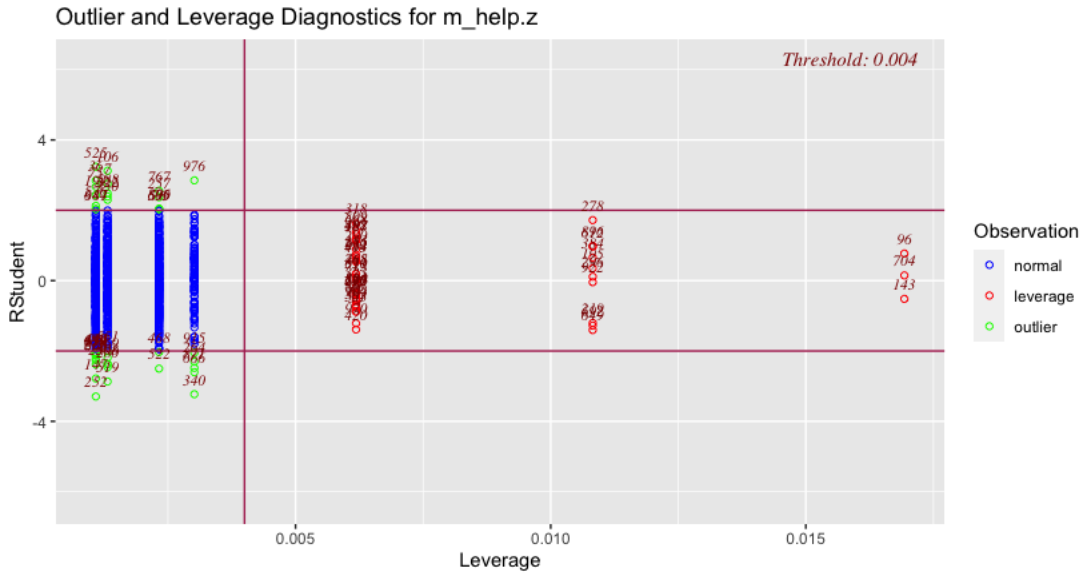
Linearity: the residuals vs fitted values plot shows a random scatter around midline of 0, the assumption of linearity is met

Normality: the Normal Q-Q plot shows the standardized residuals following a straight line, the assumption of normality is met

Homoscedasticity: the scale-location plot shows that the variance of the residuals is more or less consistent across the fitted values. The variances are slightly lower for lower fitted values, but overall I say that the assumption of homoscedasticity is met.

b) Examine influential observations. Which observations stand out as being influential?





According to the jackknife residuals plot, points 252 and 340 are highly influential points.

c) Perform a sensitivity analysis by excluding the influential observations. How different is your model and do your conclusions change?

Yes, the model does change when excluding influential points 252 and 340. The equation of the regression line of total adults on maladaptive score goes from: $\hat{Y}_{mhelp.z} = 0.21558 - 0.16052X_{tatot}$ (p=2.36e-09) to: $\hat{Y}_{mhelp.z} = 0.21788 - 0.15748X_{tatot}$(p=3.43e-09). The intercept, slope magnitude, and p-value all slightly increase with the exclusion of the influential points, but the conclusions remain the same.

9) Wrap-Up

a) Report the means of variables, unadjusted (univariable) parameter estimates, and adjusted (multivariable) parameter estimates in a table.

```
> data.frame(
+   unadjusted = c(m$coefficients, NA, NA, NA, NA),
+   adjusted = adj_m$coefficients,
+   row.names = names(adj_m$coefficients)
+ )
```

	unadjusted	adjusted
(Intercept)	0.2155761	0.51875393
tatot	-0.1605183	-0.16045413
age	NA	-0.01631523
gender	NA	-0.12493025
ethO	NA	-0.26181276
ethW	NA	0.03002977

b) Write a concluding sentence or two about the effect of trusted adults on maladaptive help-seeking.

The number of trusted adults is positively associated with a lower maladaptive score in high schoolers, with a coefficient of -0.16 (p=2.36e-09), and is not confounded by age, gender, or ethnicity.