2024 Spring Statistical Method Final Project by Meng Hsuan Ho & Pin Tzu Tseng

Data Problem 6

The data given in the file depression.dat contains information on age (in years), sex (0=male, 1=female), the work problems index (WP), the marital conflict index (MC), and a depression index (DEP) for a sample of 39 new admissions to a psychiatric clinic at a large university hospital. A higher value on the work problems, marital conflict, and depression indices indicates a more negative outcome on the given index. **Researchers want to characterize how the depression index is associated with work and marital problems, on average, after adjusting for sex and age.** They also are interested in assessing **whether the impact of the MC and WP indices each depend on sex** (i.e. is there evidence of interactions between sex and MC and sex and WP?). Build a regression model or models to address the researchers' questions and interpret the results.

Instructions

1. (2 points) Identify the response variable of interest in the study.

- depression index (DEP)

2. (4 points) Identify the explanatory variables measured which are needed to address the study goals. Note that some variables may have been measured but are not needed to address the study goals. Do not include those.

- Sex, age, the marital conflict index (MC), work problems index (WP)

3. (4 points) State whether each of the explanatory variables are factor (i.e. categorical) or quantitative.

- The explanatory variable of sex is a factor, and others including age, MC, and WP are all quantitative.

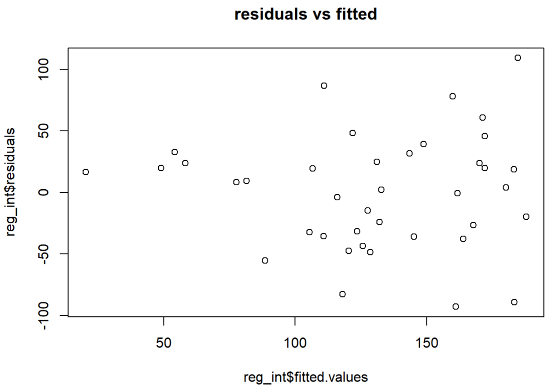
4. (5 points) Give the initial model needed to address the research questions of interest.

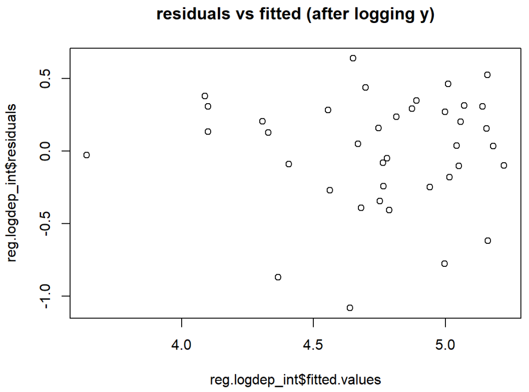
- Multiple Linear Regression
  DEP = b0 + b1*AGE + b2*SEX + b3*WP + b4*MC + b5*SEX*WP + b6*SEX*MC

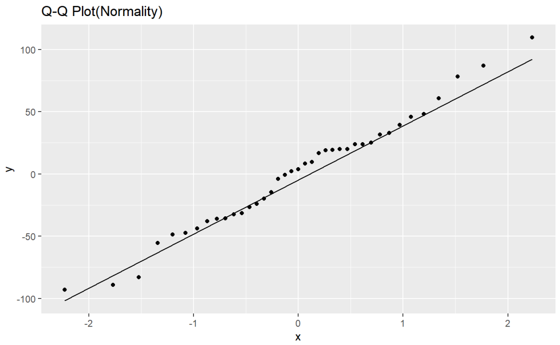
5. (20 points) State and assess the assumptions of the model in (4). Provide supporting evidence for each assumption. Do not include output unless you use it to address a particular assumption but be thorough in including any tool that you used. Be clear about which tool is being used to assess which assumption. As part of your assessment, address the following:

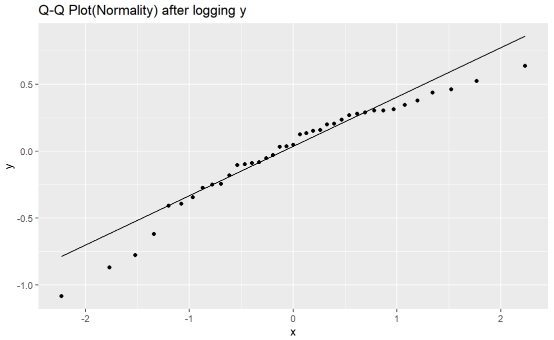
(a) **Is there evidence that the response variable should be transformed? If so, what transformation do you recommend and why?**

- There is no evidence we should transform the response variable, the depression index (DEP). Even though the variance looks not so equal, the plots that have a transformation of DER are getting worse.



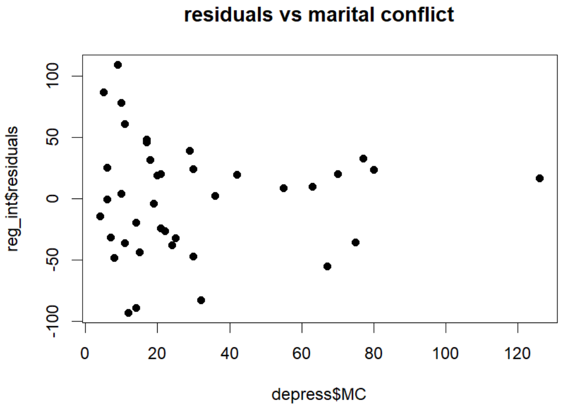


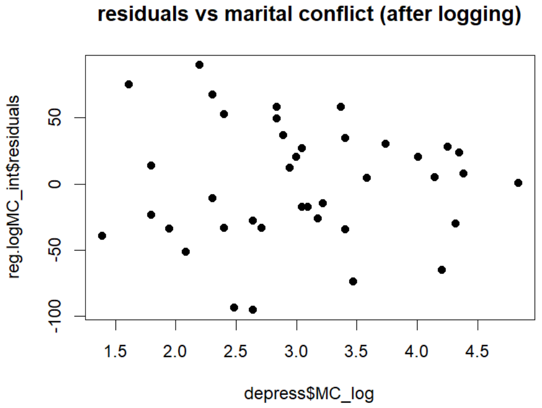


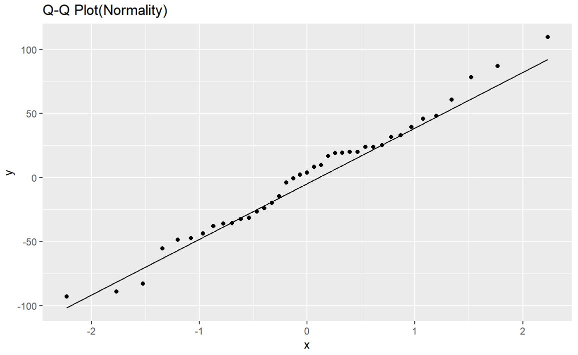


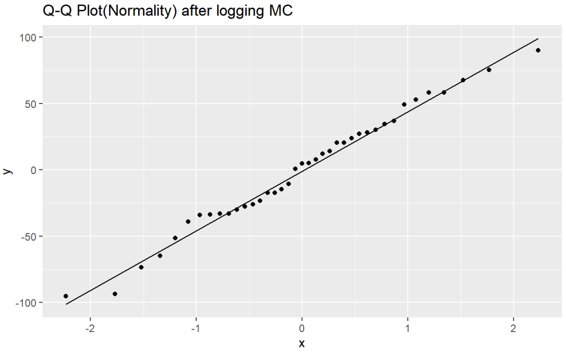
(b) Is there evidence that any of the quantitative explanatory variables should be transformed? If so, what transformation do you recommend and why?

- Yes, we think we should transform one of the quantitative explanatory variables, the marital conflict (MC) because the variance does not look equal on the plot.
  After transforming the marital conflict data, its variance looks equal and the distribution of the whole data looks more normal.



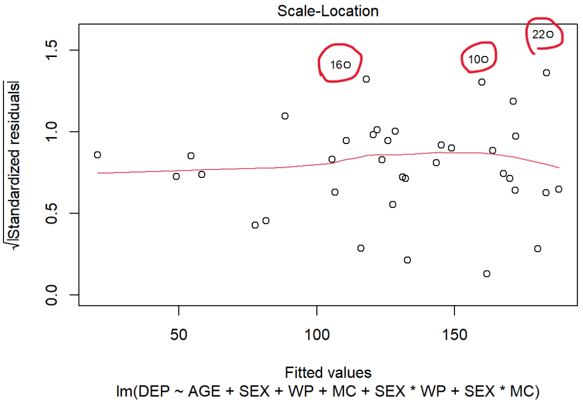


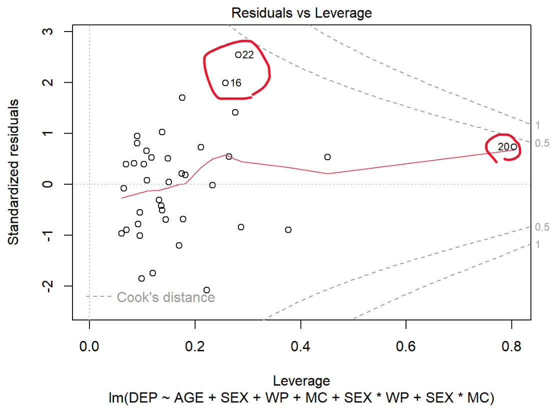




(c) Identify potential outliers based only on the residual plots (you will look at leverage, Cook's distance, and DFFIT values later).

- Yes, it has some outliers.

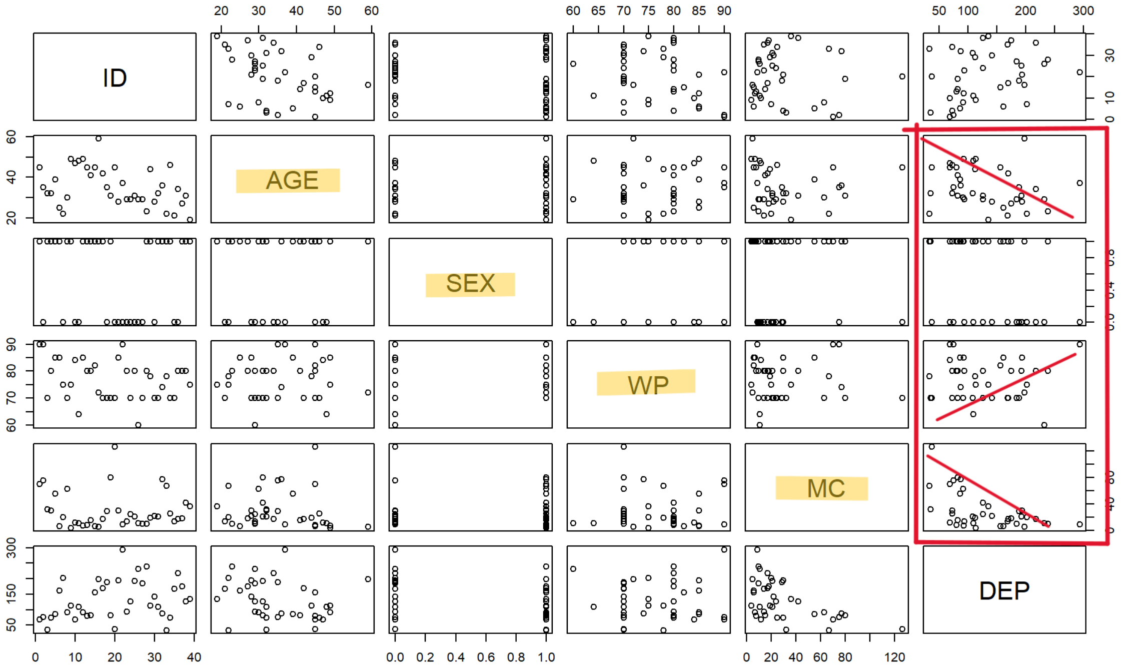


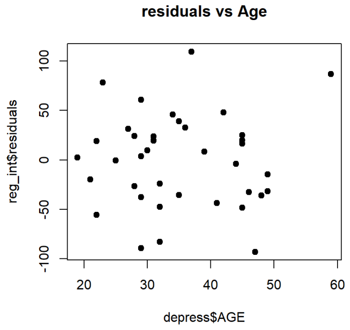


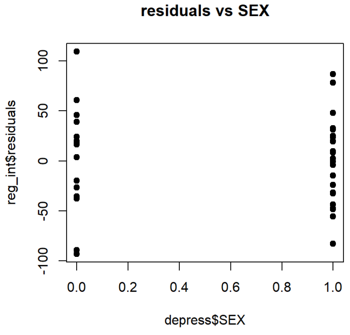
(d) Be sure to describe any other assumptions you investigated.

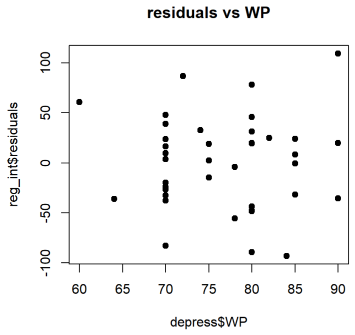
- This study has Independent observations.
- Linear association between (mean) y and each explanatory variable

Besides the association between (mean) y and marital conflict index (MC), the plots show the relationship with other explanatory variables has equal variance.









6. (15 points) Using the full data set (that is, do not remove any potential outliers from the data set) but incorporating any transformations recommended in (5), address the specific questions of interest in the chosen Data Problem. Clearly state which question you are addressing and give supporting evidence. Do not provide any output unless you refer to it in your response.

- After including the transformation of the marital conflict index (MC), we can know the association between the depression index (DEP), marital conflict index (MC), and work problems index (WP).
  It has enough evidence to prove that there is an association between the depression index and the marital conflict index on average.
  It doesn't have significant evidence to prove that there is an association between the depression index and the work problem index.
- Methodology:

  Null Ho: b1=b2=b3=b4=b5=b6=0
  Alternative Ha: b1≠b2≠b3≠b4≠b5≠b6≠0

1. About testing the association between the depression index (DEP) and the work problems index (WP).
   The p-value is 0.45733, we fail to reject the null hypothesis
   There is not enough evidence to prove that there is an association between the depression index and the work problems index on average.
2. About testing the association between the depression index (DEP) and the marital conflict index (MC).
   The p-value is 0.00971, we reject the null hypothesis
   There is significant evidence to prove that there is an association between the depression index and the marital conflict index on average.

```
> summary(reg.logMC_int)
Call:
lm(formula = DEP ~ AGE + SEX + WP + MC_log + SEX * WP + SEX * MC_log, data= depress)

Residuals:
   Min      1Q      Median     3Q      Max
-95.048  -31.404    4.863    29.164   90.082

Coefficients:
                  Estimate    Std. Error   t value     Pr(>|t|)
(Intercept)       294.8758     116.4661     2.532     0.01645 *
AGE                -2.1821       0.9275    -2.353     0.02494 *
SEX                49.6449     188.3836     0.264     0.79383
WP                  1.0774       1.4319     0.752     0.45733
MC_log            -48.7802      17.7365    -2.750     0.00971 **
SEX:WP             -1.4813       2.2869    -0.648     0.52178
SEX:MC_log          8.9800      21.6608     0.415     0.68122
—
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.07 on 32 degrees of freedom
Multiple R-squared:  0.4811,  Adjusted R-squared:  0.3838
F-statistic: 4.945 on 6 and 32 DF,  p-value: 0.00111
```

- After including the transformation of the marital conflict index (MC), we know the impact of the marital conflict index (MC) and work problems index (WP) indices each depend on sex. There is no significant evidence to prove that those contain interactions between sex and WP and sex and MC.
- Methodology:

  Null Ho: DEP = b0 + b1*AGE + b2*SEX + b3*WP + b4*MC
  Alternative Ha: DEP = b0 + b1*AGE + b2*SEX + b3*WP + b4*MC + b5*SEX*WP + b6*SEX*MC

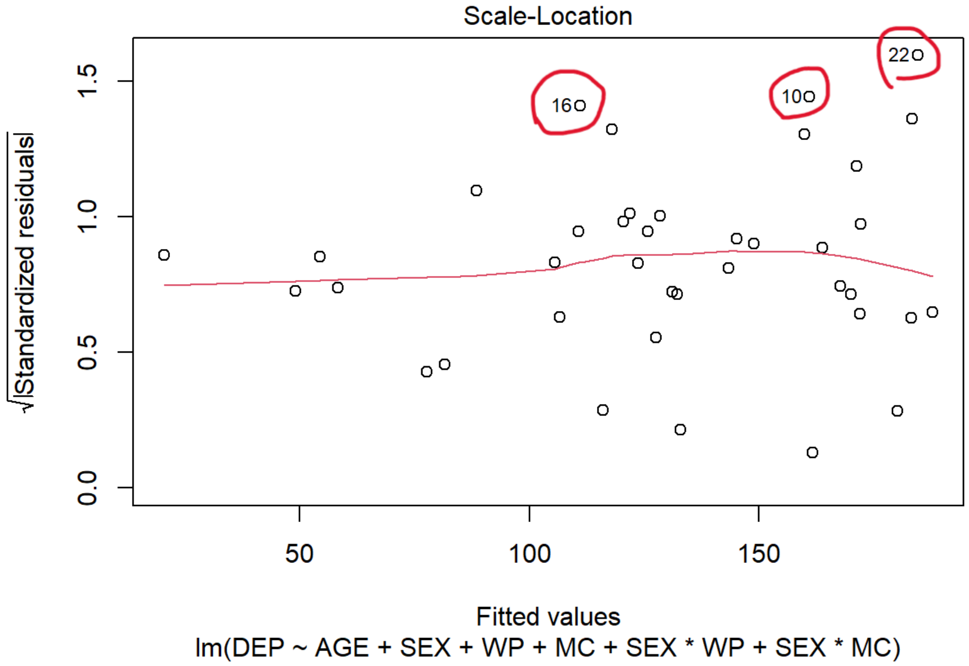
  We got a p-value of 0.746, which is greater than sigma=0.05.

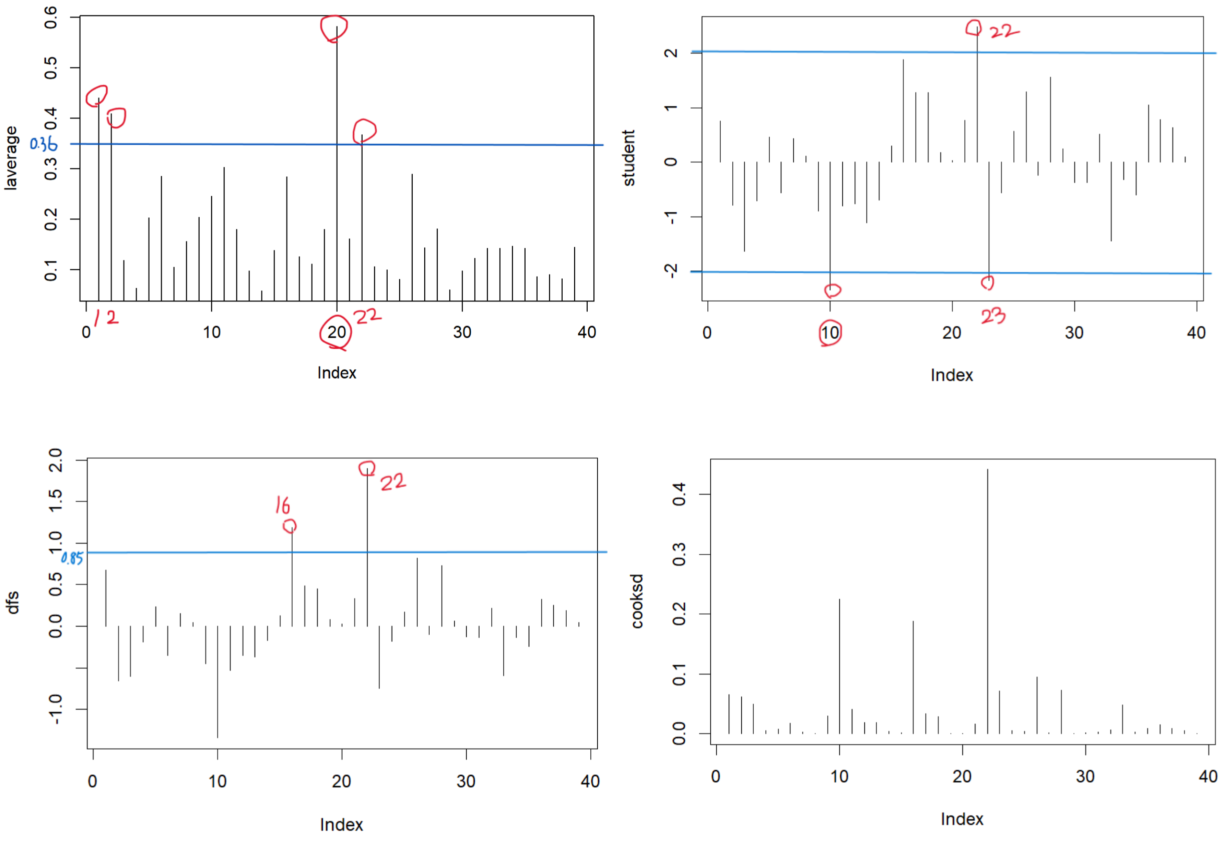
  Thus, the result fails to reject Ho, so there is no significant evidence to prove that there are interactions between sex and WP and sex and MC

```
> anova(reg,reg.logMC_int)
Analysis of Variance Table

Model 1: DEP ~ AGE + SEX + WP + MC_log

Model 2: DEP ~ AGE + SEX + WP + MC_log + SEX * WP + SEX * MC_log

   Res.Df   RSS     Df    Sum of Sq       F        Pr(>F)

1    34      78464

2    32      77040    2     1423.8        0.2957      0.746
```

7. (15 points) Examine the diagnostic measures for outliers from the model you fit in (6). Which observations have extreme studentized residuals? Leverages? Cook's distance? DFFIT values? Describe these observations. Investigate the impact of these observations and revisit your analysis from the previous question. Summarize your overall conclusions and address the research questions.

Observations 10, 16, and 22 highlighted that they have more extreme values on one or more of these diagnostics. (studentized residuals, DFFITs, cook's distance)





the examination of outliers:

No.1.2.20.22 leverage greater than 0.358974

No.10.22.23 studentized residuals (student) greater than 2 or smaller than -2

No.16.22 DFFIT (dffits) greater than 0.8473185

None Cook's Distance (cooksd) greater than 1

	DEP	fitted	residuals	laverage	student	dffits	cooksd
1	69	40.87533	28.12467	0.440413	0.761199	0.675297	0.066014
2	75	104.855	-29.855	0.409854	-0.78734	-0.65614	0.062241
3	35	108.4767	-73.4767	0.119372	-1.63713	-0.60275	0.049312
4	73	107.0055	-34.0055	0.064234	-0.71089	-0.18625	0.005033
5	86	65.58634	20.41366	0.203593	0.460423	0.232794	0.007937
6	161	184.3166	-23.3166	0.285267	-0.556	-0.35126	0.018015
7	202	181.538	20.46204	0.105076	0.435213	0.149129	0.00326
8	91	85.8803	5.119697	0.15698	0.111876	0.048277	0.000344
9	113	152.1228	-39.1228	0.204635	-0.89118	-0.45203	0.029379

10	68	161.599	-93.599	0.246365	**-2.34711**	-1.34196	0.225494
11	109	142.1142	-33.1142	0.30313	-0.80398	-0.53025	0.040615
12	92	125.8102	-33.8102	0.18066	-0.75615	-0.35506	0.018254
13	80	131.244	-51.244	0.098629	-1.10379	-0.36512	0.018916
14	82	114.9538	-32.9538	0.059143	-0.68666	-0.17216	0.004305
15	156	141.8859	14.11411	0.139328	0.305641	0.122974	0.002223
16	198	122.6323	75.36775	0.284723	1.887537	**1.190883**	0.187578
17	170	111.8299	58.17009	0.126188	1.280894	0.486759	0.033183
18	188	129.6585	58.34152	0.111811	1.273879	0.451978	0.028626
19	82	74.19021	7.809792	0.181004	0.173194	0.081421	0.000977
20	37	36.17963	0.820368	**0.582363**	0.025465	0.03007	0.000133
21	194	159.44	34.55997	0.162091	0.764459	0.336229	0.016363
22	294	203.9177	90.08231	**0.367472**	**2.488728**	**1.896925**	0.442264
23	94	189.0485	-95.0485	0.106728	**-2.1644**	-0.74814	0.071704
24	126	151.9826	-25.9826	0.100085	-0.55211	-0.18412	0.004951
25	192	164.9055	27.09446	0.08125	0.569993	0.169505	0.004193
26	232	179.2653	52.73468	0.28953	1.28816	0.822325	0.094652
27	184	194.6881	-10.6881	0.143816	-0.23191	-0.09505	0.00133
28	238	170.3698	67.63019	0.181206	1.556761	0.732355	0.073357
29	112	99.807	12.193	0.061125	0.252684	0.064474	0.000612
30	141	158.4091	-17.4091	0.098505	-0.36861	-0.12185	0.00218
31	108	125.2411	-17.2411	0.123745	-0.37028	-0.13915	0.002843
32	87	63.18484	23.81516	0.143027	0.518283	0.211735	0.006554
33	33	97.65552	-64.6555	0.142586	-1.44721	-0.59016	0.048111
34	73	87.7519	-14.7519	0.147471	-0.32102	-0.13352	0.00262
35	168	195.732	-27.732	0.14342	-0.6046	-0.24739	0.00892
36	218	168.6668	49.33316	0.086337	1.053684	0.323903	0.014936
37	175	138.2472	36.75279	0.091106	0.780885	0.247232	0.00884
38	126	95.79599	30.20401	0.0829	0.636798	0.191456	0.005336
39	135	130.1367	4.863264	0.144831	0.105512	0.043422	0.000278

- After excluding each and gathering the observations of No.1.2.10.16.20.22.23

model term Coefficients:									
Estimate	Full	no.1	no.2	no.10	no.16	no.20	no.22	no.23	no.1.2.10.16.20.22.23
(Intercept)	294.8758	301.1123	235.7802	242.6002	308.3183	295.2756	382.5676	287.1455	204.863
WP	1.0774	1.0928	1.6075	2.0306	1.1106	1.0921	**-0.8304**	1.5487	1.885
MC_log	-48.7802	-48.3509	-41.2065	-59.7532	-47.8548	-49.2509	-30.4792	-54.3307	-30.804
SEX:WP	-1.4813	-2.4188	-2.0122	-2.4055	-0.6178	-1.4963	**0.416**	-1.9589	-1.955
SEX:MC_log	8.98	3.4827	1.3261	23.0476	13.5778	9.4187	**-10.437**	13.8682	**-4.92**

After excluding observation 22, there is a significant impact on the slope between the depression index and the work problem index, and the slope between the depression index and transformed marital conflict index depends on sex, and the slope between the depression index and the work problem index depends on sex in the model.

There is no significant impact on the transformed marital conflict index after excluding outliers.

model term P-value									
Pr(>|t|)	Full	no.1	no.2	no.10	no.16	no.20	no.22	no.23	no.1.2.10.16.20.22.23
(Intercept)	0.01645	0.0155	0.1002	0.03698	0.00996	0.0191	0.00204	0.01408	0.141
WP	0.45733	0.4542	0.3199	0.1572	0.42639	0.4906	0.59202	0.26813	0.3944
MC_log	0.00971	0.011	**0.0507**	0.00158	0.00866	**0.0657**	**0.10079**	0.00318	**0.4526**
SEX:WP	0.52178	0.3614	0.4077	0.27768	0.78524	0.537	0.8548	0.37513	0.4956
SEX:MC_log	0.68122	0.8805	0.956	0.28415	0.52236	0.7384	0.63162	0.50654	0.9074

For p-value, there are some significant impacts on the marital conflict after logging in the model without outliers, so marital conflict is more important to the model.

The explanatory variables of WP, WP*SEX, and MC*SEX do not significantly affect the model, so it is not necessary to involve those variables in the model.

- To determine the final model by excluding the extreme outliers.

Interaction			
Model 1: DEP ~ AGE + SEX + WP + MC_log Model 2: DEP ~ AGE + SEX + WP + MC_log + SEX*WP+SEX * MC_log			
Pr(>F)	Full	no.22	no.1.2.10.16.20.22.23
2	0.746	0.8828	0.7273

Model: DEP ~ AGE + SEX + WP + MC_log			
Pr(>F)	Full	no.22	no.1.2.10.16.20.22.23
WP	0.66909	0.504329	0.138366
MC_log	9.29E-05	0.000122	0.003953

- Stepwise AIC method

```
> step(reg.logMC_int)
Start:  AIC=309.95
DEP ~ AGE + SEX + WP + MC_log + SEX * WP + SEX * MC_log
                Df    Sum of Sq     RSS      AIC
- SEX:MC_log  1        413.8       77454   308.16
- SEX:WP        1       1010.1       78050   308.46
<none>                                77040   309.95
- AGE              1      13327.3      90367    314.18

Step:  AIC=308.16
DEP ~ AGE + SEX + WP + MC_log + SEX:WP
             Df   Sum of Sq    RSS      AIC
- SEX:WP    1      1010        78464    306.67
<none>                         77454    308.16
- AGE         1     15875       93329    313.43
- MC_log     1      46305      123758   324.44

Step:  AIC=306.67
DEP ~ AGE + SEX + WP + MC_log
          Df    Sum of Sq    RSS     AIC
- WP        1        374       78838  304.85
<none>                         78464  306.67
- SEX       1      11762      90226  310.11
- AGE       1      15556       94019  311.72
- MC_log  1      45295      123758 322.44

Step:  AIC=304.85
DEP ~ AGE + SEX + MC_log
              Df Sum of Sq   RSS        AIC
<none>                78838      304.85
- SEX           1     11438       90276   308.14
- AGE           1     15344       94182    309.79
- MC_log      1     45349      124187   320.57

Call: lm(formula = DEP ~ AGE + SEX + MC_log, data = depress)
Coefficients:
(Intercept)        AGE          SEX        MC_log
    357.042      -2.245      -35.624      -41.721
```

We did the AIC to find the most significant model which is DEP = b0+ b1*AGE + b2*SEX + b3*MC_log (AIC is 304.85).

- **Restate**

  According to the stepwise AIC method, the model only includes the explanatory variables of age, sex, and marital conflict index (MC) has a more significant association with the depression index (DEP) compared to other models that include the explanatory variables of age, sex, marital conflict index (MC), and work problems (WP).

  The final model is DEP = b0 + b1*AGE + b2*SEX + b3*MC

  Individuals with a 50% more marital conflict index, but the same level of sex and age, have -41.721*log(1.5) = -16.91641 worse depression index, on average.

  There is not enough evidence to support that there is an association between the depression index and the work problems index on average. However, there is evidence to support that there is an association between the depression index and the marital problems index on average. Otherwise, there is no significant impact on marital conflicts and work problems depending on sex on average.