02 - Regression Analysis

Alec Stashevsky

12/05/2021

Build Regression Data

To simplify the analysis, each variable of interest has been coded as either DX or LX to indicate either a categorical demographic variable or a ordinal Likert-scale variable, respectively. Please refer to the `Variable Keynames.xlsx` file to find the specific question encoded by the variable name. Delta is a continuous variable which measures the difference between $L3$ and $L2$, discussed more in the second regression section.

A count summary of regression data is provided in the output below:

```
##  L3     L2     L1     D1        D4     D5     D6         Delta
##  0: 2   0: 3   0: 2   0: 5   1   :29   0:26   0:36   Min.   :-1.0000
##  1: 6   1: 7   1: 7   1:46   2   :14   1:28   1:41   1st Qu.: 0.0000
##  2: 3   2:10   2:10   2:27   3   :18   2: 9   2: 1   Median : 0.0000
##  3:12   3:17   3:25          4   : 5   3: 8          Mean   : 0.5385
##  4:30   4:32   4:34          5   :11   4: 7          3rd Qu.: 1.0000
##  5:25   5: 9                 NA's: 1                 Max.   : 2.0000
```

There is one missing value for respondent 37 on Question D4. This respondent may be dropped from the regression, lets make sure to check.

Regression Analysis - Importance of Teaching Climate Change

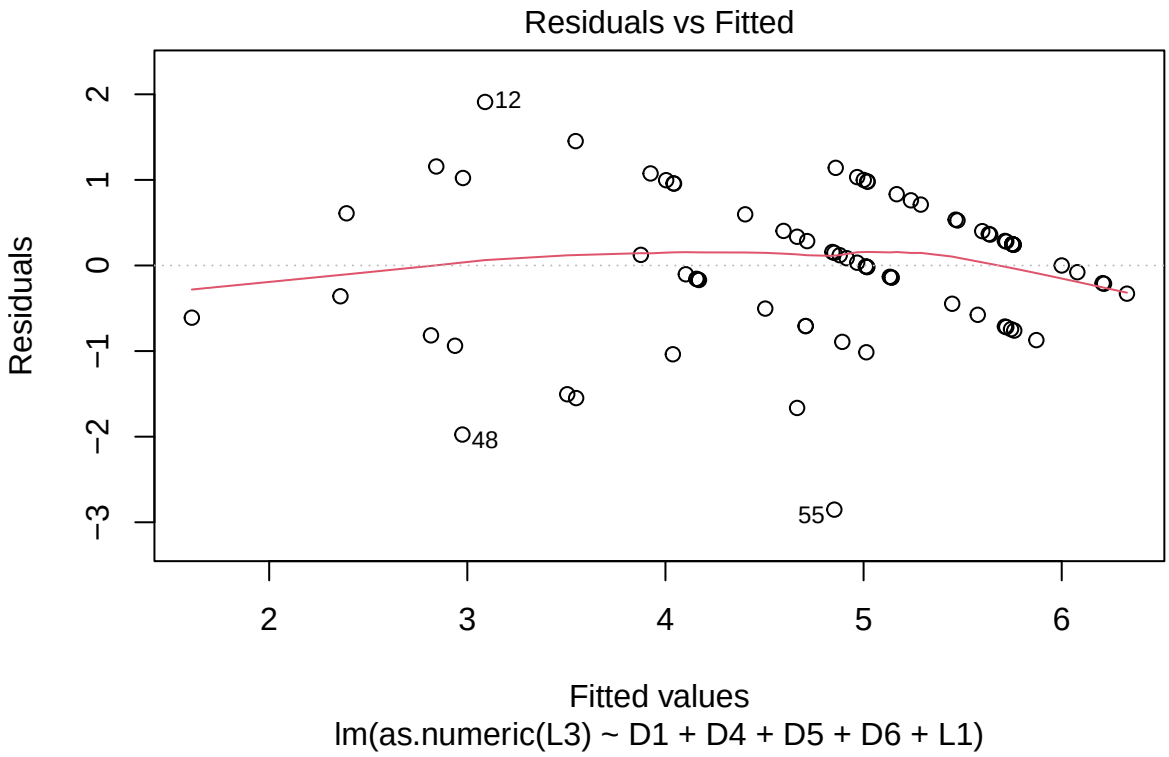
Linear Regression

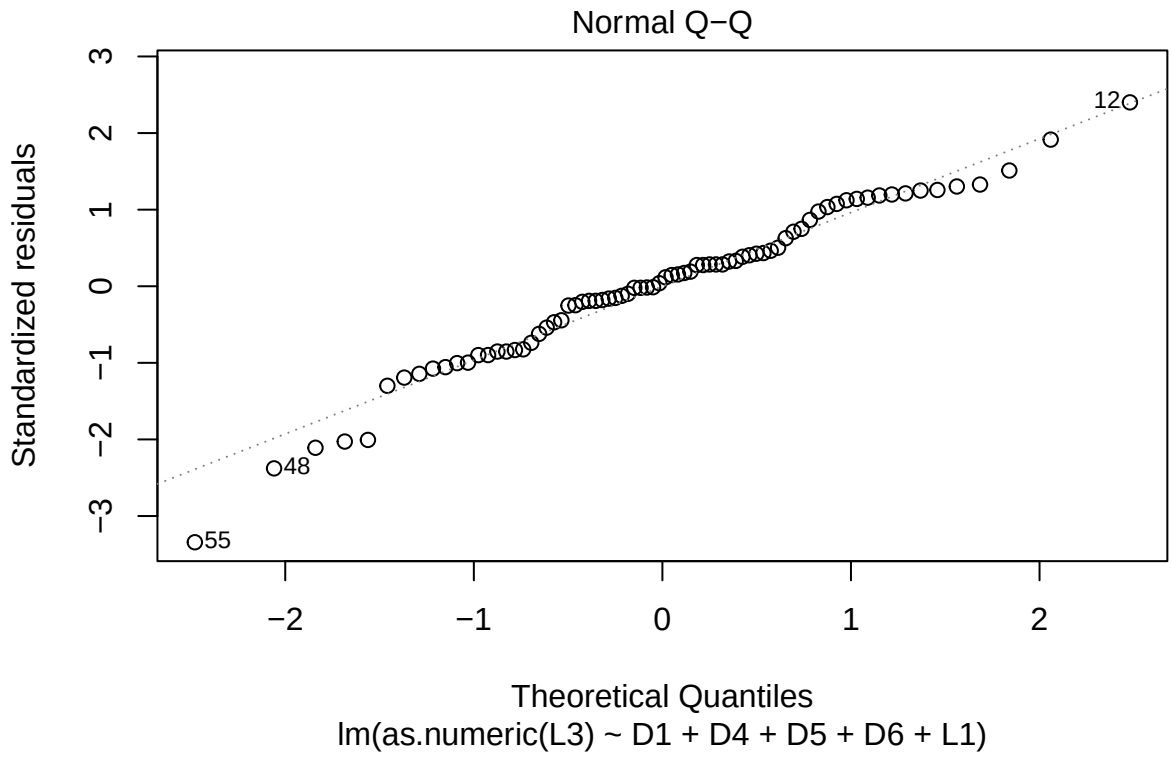
We will first test a multivariate linear regression where we treat the independent variable as continuous. Our first regression will be of the form,

$$L3 = \beta_0 + \beta_1 D1 + \beta_2 D4 + \beta_3 D5 + \beta_4 D6 + \beta_5 L1 + \epsilon$$

where DX indicates categorical demographic variables and $L1$ is an ordinal Likert scale variable. Note: $L3$ must be continuous for to perform a linear regression.

The output of the regression is summarized below.


Residuals vs Fitted
Residuals
Fitted values
lm(as.numeric(L3) ~ D1 + D4 + D5 + D6 + L1)
12
48
55
2
1
0
−1
−2
−3
2
3
4
5
6




```
##
## Call:
## lm(formula = as.numeric(L3) ~ D1 + D4 + D5 + D6 + L1, data = reg.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85233 -0.50379  0.03249  0.52603  1.91045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.761439   0.616920   6.097 8.45e-08 ***
## D11          0.295777   0.547764   0.540    0.591
## D12         -0.128991   0.481548  -0.268    0.790
## D42         -0.118515   0.331210  -0.358    0.722
## D43         -0.158640   0.290014  -0.547    0.586
## D44         -0.615535   0.468412  -1.314    0.194
## D45          0.334547   0.358522   0.933    0.354
## D51         -0.007011   0.293092  -0.024    0.981
## D52          0.411491   0.572116   0.719    0.475
## D53         -0.453252   0.556065  -0.815    0.418
## D54         -0.549863   0.606765  -0.906    0.368
## D61          0.121441   0.229831   0.528    0.599
## D62          0.242449   0.970472   0.250    0.804
## L1.L         2.714093   0.482005   5.631 5.03e-07 ***
## L1.Q         0.107345   0.461424   0.233    0.817
## L1.C        -0.257982   0.359610  -0.717    0.476
```

```
## L1^4         0.066886   0.314968   0.212    0.833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9246 on 60 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6063, Adjusted R-squared:  0.5014
## F-statistic: 5.776 on 16 and 60 DF,  p-value: 2.582e-07
```

Here we can see the linear (.L) contrast of our Likert predictor is significant at the 1% level and indicates a *positive linear* trend on the response to question L3.

The above output provides β coefficients for *each level* of the categorical and ordinal predictors. For example, $D11$ indicates the output for respondents who are coded as *1* on question $D1$. This group includes respondents who identified as Residents/Fellows. The first level of each factor is omitted as a base comparison. So, the coefficient estimate for $D11 = 0.296$ indicates that a respondent who identified as Resident/Fellow will have an average 0.296 increase in their response to question $L3$ as compared to the respondents who are categorized as "Other" for question $D1$. This group is the omitted level, $D10$.

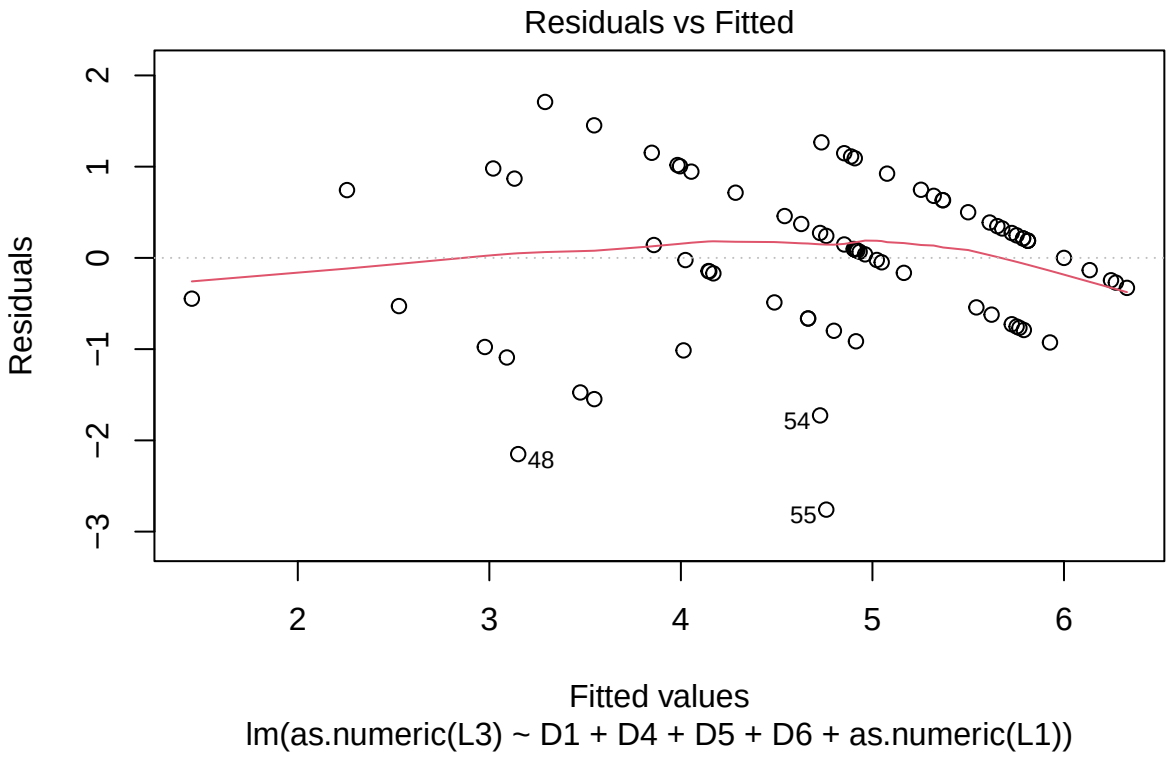
To test the effect of a predictor at the factor-level, an analysis of variance (ANOVA) is run on the model and summarized below.

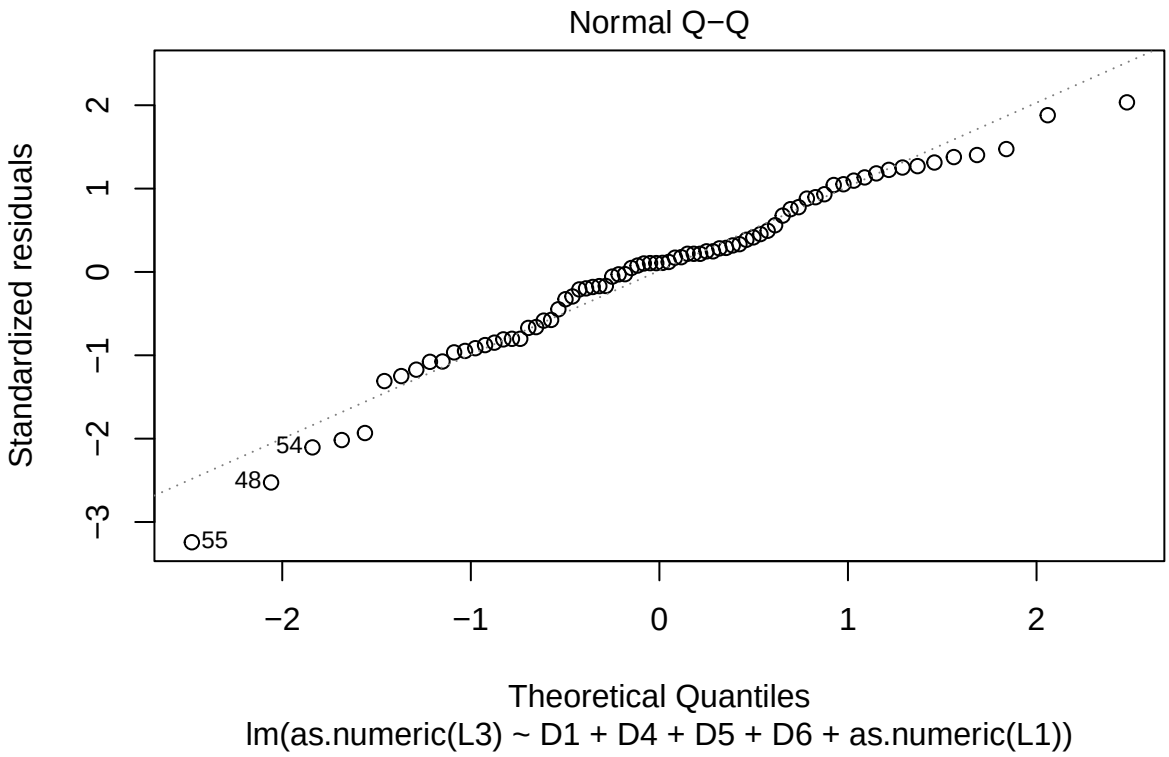
```
## Analysis of Variance Table
##
## Response: as.numeric(L3)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## D1         2  6.149  3.0744  3.5959    0.0335 *
## D4         4  6.678  1.6694  1.9526    0.1134
## D5         4  2.176  0.5440  0.6362    0.6386
## D6         2  1.888  0.9441  1.1042    0.3381
## L1         4 62.124 15.5309 18.1657 8.006e-10 ***
## Residuals 60 51.298  0.8550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA estimates using type I SS tell us how much of the variability in $L3$ can be explained by our predictor variables *in sequential order.* This means we first measure the variability of $L3$ that is explained by $D1$, then how much of the *residual variability* can be explained by $D4$, how much of the *remaining residual variability* can be explained by $D5$, and so on, in order of their specification in the model.

The output above indicates that the variability of $D1$ explains a significant portion of the variability in $L3$ at the 95% confidence level. Additionally, the *remaining residual variability* of $L3$ attributable to $L1$ after accounting for variability explained by the previous variables ($D1$, $D4$, $D5$, and $D5$) is significant at the 99% confidence level.

We run an additional specification identical to the model above except $L1$ is treated as a continuous variable to be consistent with the treatment of $L3$, which is required to be continuous in the context of linear regression. That output is summarized below.


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
2
3
4
5
6
54
48
55
Fitted values
lm(as.numeric(L3) ~ D1 + D4 + D5 + D6 + as.numeric(L1))




```
##
## Call:
## lm(formula = as.numeric(L3) ~ D1 + D4 + D5 + D6 + as.numeric(L1),
##     data = reg.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75906 -0.52822  0.09215  0.50031  1.70900
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.08858    0.73086   1.489    0.141
## D11             0.30480    0.52456   0.581    0.563
## D12            -0.07943    0.46789  -0.170    0.866
## D42            -0.13434    0.32247  -0.417    0.678
## D43            -0.17414    0.28300  -0.615    0.541
## D44            -0.62240    0.45887  -1.356    0.180
## D45             0.34384    0.34386   1.000    0.321
## D51             0.02535    0.27989   0.091    0.928
## D52             0.38816    0.53138   0.730    0.468
## D53            -0.44085    0.53878  -0.818    0.416
## D54            -0.54121    0.58624  -0.923    0.359
## D61             0.11504    0.21756   0.529    0.599
## D62             0.21354    0.94854   0.225    0.823
## as.numeric(L1)  0.87862    0.10161   8.647 2.66e-12 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9074 on 63 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.602,  Adjusted R-squared:  0.5198
## F-statistic: 7.329 on 13 and 63 DF,  p-value: 1.774e-08
```

Here we see that when the Likert variable $L1$ is specified as continuous, it's effect is consistent with the ordinal specification in the first regression.

We build an ANOVA table for this second model, summarized below.

```
## Analysis of Variance Table
##
## Response: as.numeric(L3)
##                Df Sum Sq Mean Sq F value    Pr(>F)
## D1              2  6.149   3.074  3.7342   0.02934 *
## D4              4  6.678   1.669  2.0277   0.10124
## D5              4  2.176   0.544  0.6607   0.62160
## D6              2  1.888   0.944  1.1467   0.32423
## as.numeric(L1)  1 61.553  61.553 74.7626 2.657e-12 ***
## Residuals      63 51.869   0.823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA of the alternative specification yields quantitatively similar results with $D1$ and $L1$ explaining a significant portion of the variation of $L3$ at the 95% and 99% confidence levels, respectively. This is consistent with the first specification.

Ordinal Logistic Regression

Now, we will proceed with an ordinal logistic regression of the form,

$$logit(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - \cdots - \eta_p x_p$$

where Y is an ordinal Likert variable with J categories. $P(Y \leq j)$ represents the cumulative probability of Y being less than or equal to a specific category $j = 1, \ldots, J-1$. In our case $J = 5$ and the response and predictor variables are the same as our linear regression specification. Note that we longer have β coefficients, but rather η coefficients which are of opposite sign. Thus, a positive η coefficient actually indicates a lower log-odds probability and vice versa.

```
## Call:
## polr(formula = L3 ~ D1 + D4 + D5 + D6 + L1, data = reg.data,
##     Hess = TRUE)
##
## Coefficients:
##        Value Std. Error    t value
## D11   1.0576  1.218e+00  8.684e-01
## D12  -0.5168  1.113e+00 -4.642e-01
## D42  -0.4778  7.499e-01 -6.371e-01
## D43   0.1154  6.469e-01  1.783e-01
## D44  -1.2533  1.027e+00 -1.220e+00
## D45   2.0300  9.201e-01  2.206e+00
```

```
## D51   0.0258  6.973e-01  3.699e-02
## D52   1.5792  1.351e+00  1.169e+00
## D53  -0.9633  1.215e+00 -7.928e-01
## D54  -0.7626  1.329e+00 -5.736e-01
## D61  -0.2089  5.143e-01 -4.062e-01
## D62  13.9155  2.696e-07  5.162e+07
## L1.L  5.8806  1.189e+00  4.945e+00
## L1.Q  0.9322  1.023e+00  9.110e-01
## L1.C -0.2054  7.979e-01 -2.575e-01
## L1^4 -0.3481  6.746e-01 -5.160e-01
##
## Intercepts:
##     Value         Std. Error    t value
## 0|1       -3.6118        1.5905       -2.2709
## 1|2       -1.4147        1.4052       -1.0068
## 2|3       -0.8339        1.3910       -0.5995
## 3|4        0.8686        1.3823        0.6284
## 4|5        4.0724        1.4439        2.8205
##
## Residual Deviance: 150.0491
## AIC: 192.0491
## (1 observation deleted due to missingness)
```

Ordinal logistic regression does not operate on the same normality assumptions as linear regression, rendering p-values spurious. In the interpretation that follow, I focus on variables which have a relatively large test statistic (and therefore relatively small p-value).

The results from the ordinal logistic regression show that for respondents who are coded as *2* on question $D6$ ($D62$), the log odds of answering *0* versus answering *1* or *2* and so on for question $L3$ are about *13.9 points lower* as compared to those who are coded as *0* for question $D6$. That is, if a respondent did not identify as either Male or Female, they are *less likely* to attribute no importance to the inclusion of education on climate change and mental health (after watching the video). Additionally, for respondents who are coded as *5* from question $D4$ (represented as $D45$ in the above table), the log odds of answering *0* as compared to answering *1*, *2*, and so on for question $L3$, are about *2 points lower* than respondents from unknown region ($D5 = 0$). That is, respondents based in the Western Region, are *less likely* to attribute no importance to the inclusion of education on climate change and mental health (after watching the video) as compared to those with no known region. Lastly, we notice a significant effect on $L1.L$. Similarly, this shows us that the more worried a respondent is about climate change, the less likely they are to attribute no importance to the inclusion of education on climate change and mental health (after watching the video). Moreover, this follows a linear trend as worriedness increases.

As with our linear regression models, I perform an ANOVA to test the effects agnostic of the factor level.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: L3
##    LR Chisq Df Pr(>Chisq)
## D1    2.721  2    0.25652
## D4    9.086  4    0.05897 .
## D5    6.434  4    0.16901
## D6    0.750  2    0.68732
## L1   57.951  4  7.812e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above ANOVA table uses type II SS to attribute variability explained by the predictors in order of their specification in the model. Here we see $L1$ explains a significant portion of *remaining residual variability* of $L3$ (at the 99% confidence level) after controlling for variability explained the previous predictors.

The interpretation of ordinal logistic regression is quite complex. To aid in the interpretation I have provided an additional packet of plots which illustrate the effects of predictor variables on each level of the independent variable in the `OLR Diagnostics.pdf` file. These effects plots are built by holding all variables constant at their most probable values (usually means) except one predictor of choice. This predictor is then "wiggled" and impact on the independent variable $L3$ is measured for each level separately. Additionally, I provide one *interaction effects* plot, which "wiggles" both $D1$ and $L1$. I have refrained from providing every permutation of interaction effects to keep it simpler, but I can provide additional, more complex interaction effects plot which you may be interested in.

Regression Analysis - Effect of Video

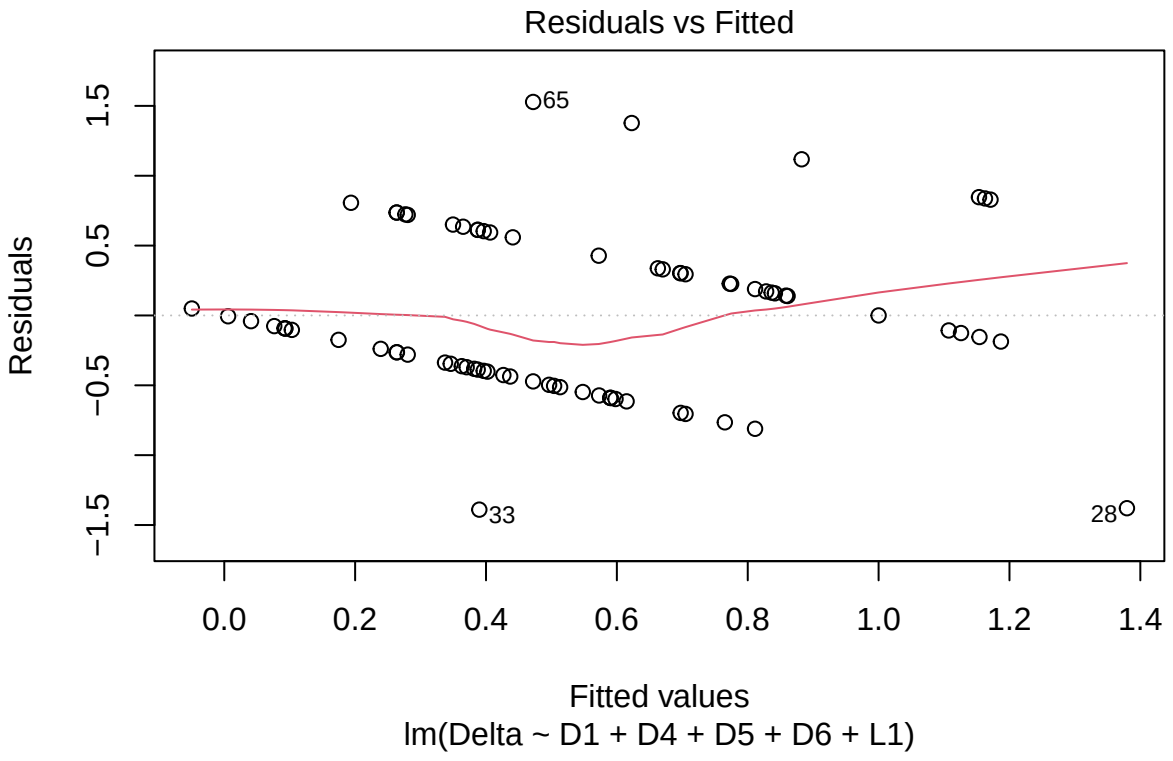
Linear Regression

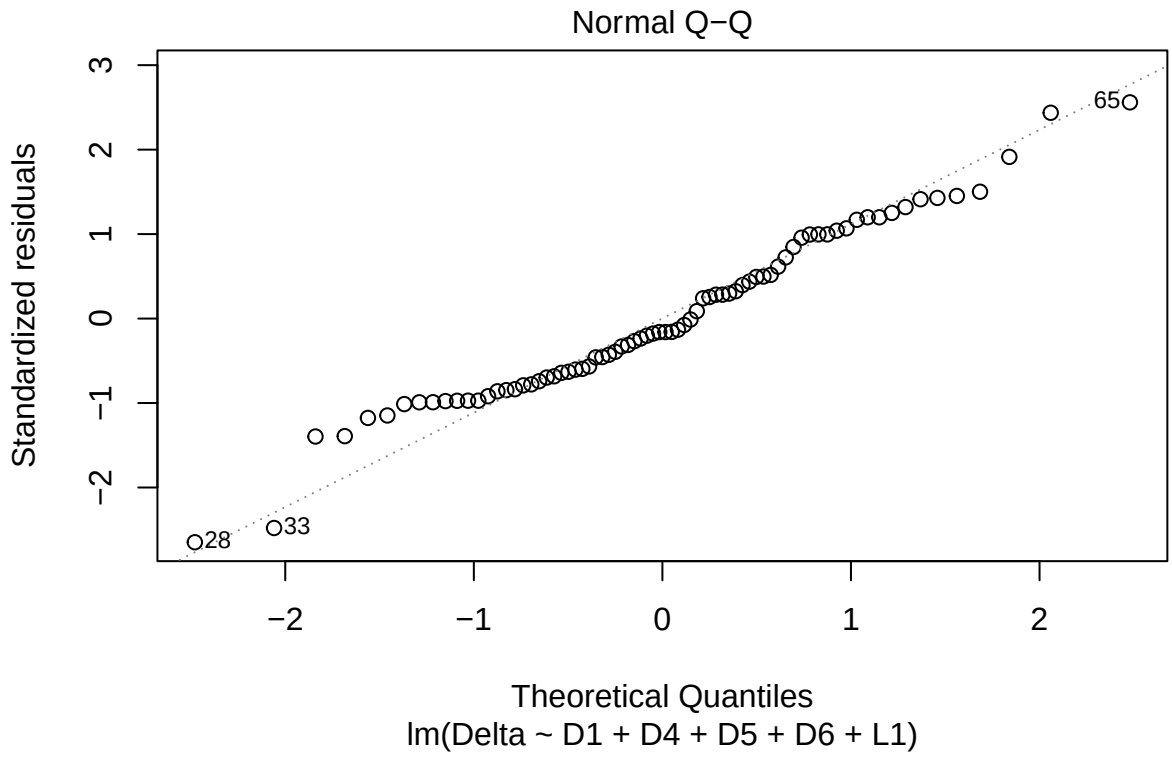
To assess the affect of the video on respondent's view on including climate change curricula, we run a multivariate linear regression where the independent variable is the change (Δ) between the response before and after watching the accompanying video. We must treat this as continuous because $\Delta = L3 - L2$. Thus, our regression will be of the form,

$$\Delta = D1 + D4 + D5 + D6 + L1 + \epsilon$$

where DX indicates categorical demographic variables and LX indicates ordinal Likert scale variables as previously specified.

The results from the model are output below:






```
##
## Call:
## lm(formula = Delta ~ D1 + D4 + D5 + D6 + L1, data = reg.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38977 -0.40218 -0.09326  0.33736  1.52797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.57361    0.43316   1.324   0.1904
## D11         -0.40200    0.38460  -1.045   0.3001
## D12          0.06309    0.33811   0.187   0.8526
## D42          0.12593    0.23255   0.541   0.5902
## D43          0.33448    0.20363   1.643   0.1057
## D44          0.57802    0.32889   1.757   0.0839 .
## D45          0.33373    0.25173   1.326   0.1900
## D51          0.10697    0.20579   0.520   0.6051
## D52         -0.64219    0.40170  -1.599   0.1151
## D53          0.11543    0.39043   0.296   0.7685
## D54         -0.77812    0.42603  -1.826   0.0728 .
## D61         -0.02435    0.16137  -0.151   0.8806
## D62          0.71963    0.68140   1.056   0.2952
## L1.L         0.38531    0.33843   1.139   0.2594
## L1.Q        -0.32384    0.32398  -1.000   0.3215
## L1.C         0.05713    0.25249   0.226   0.8218
```

```
## L1^4         0.16816    0.22115   0.760   0.4500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6492 on 60 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2358, Adjusted R-squared:  0.03198
## F-statistic: 1.157 on 16 and 60 DF,  p-value: 0.3279
```

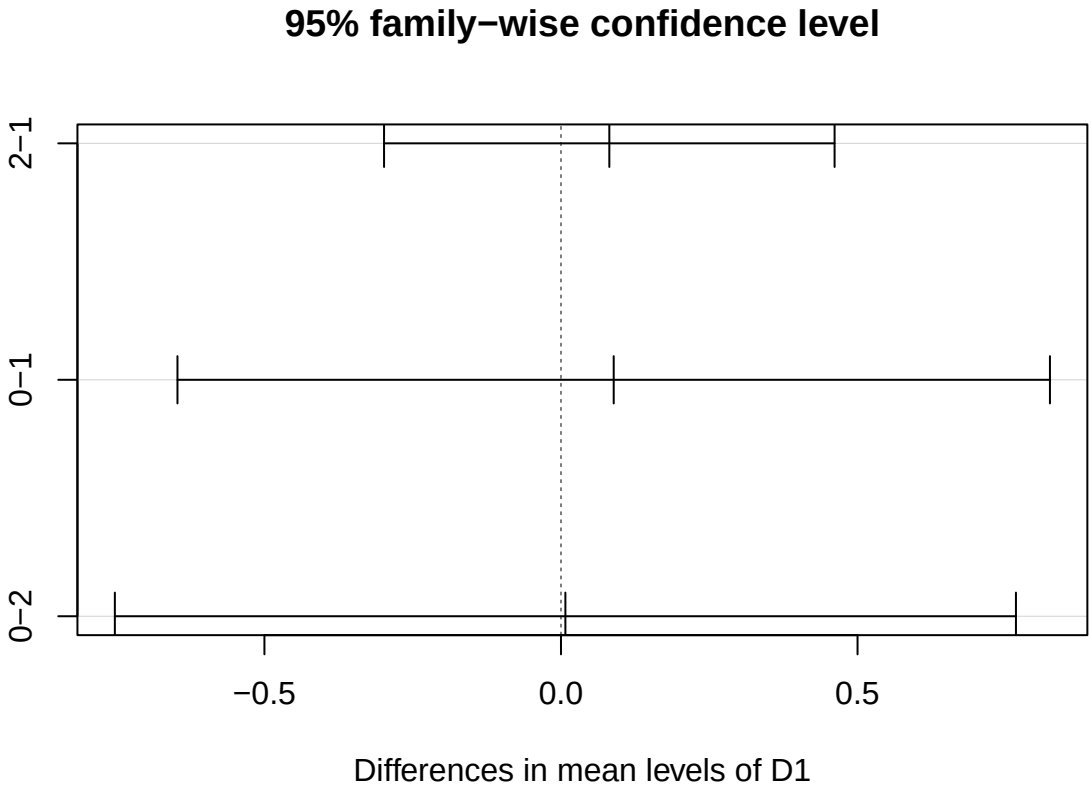
Here we see that respondents who are from the Southwest ($D44$ in the above table) attribute more importance on average to the inclusion of education on climate change and mental health after watching the video then respondents of unknown region. This result is significant at the 90% confidence level. Respondents older than 60 years ($D54$ in the above table) show they are less likely to attribute more importance to inclusion of education on climate change and mental health after watching the video. In fact, compared to respondents 30 or younger, they attribute 0.77 less import to climate change curricula with respect to our Likert 5-scale. This result is significant at the 90% confidence level.

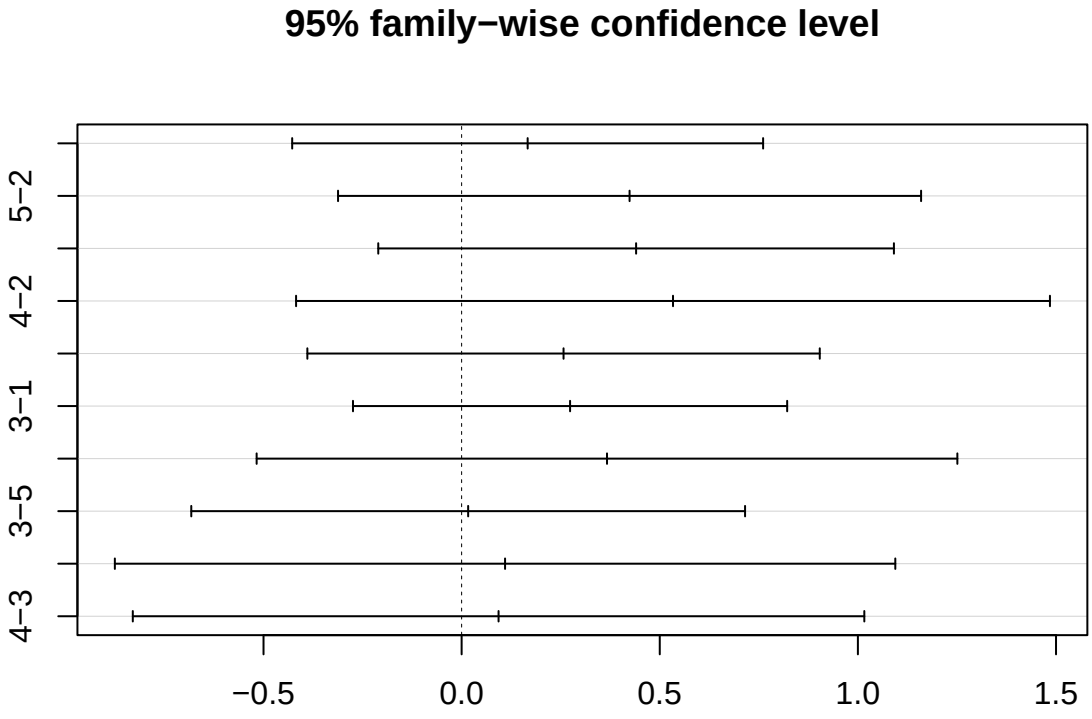
The output below summarizes the ANOVA of the above model.

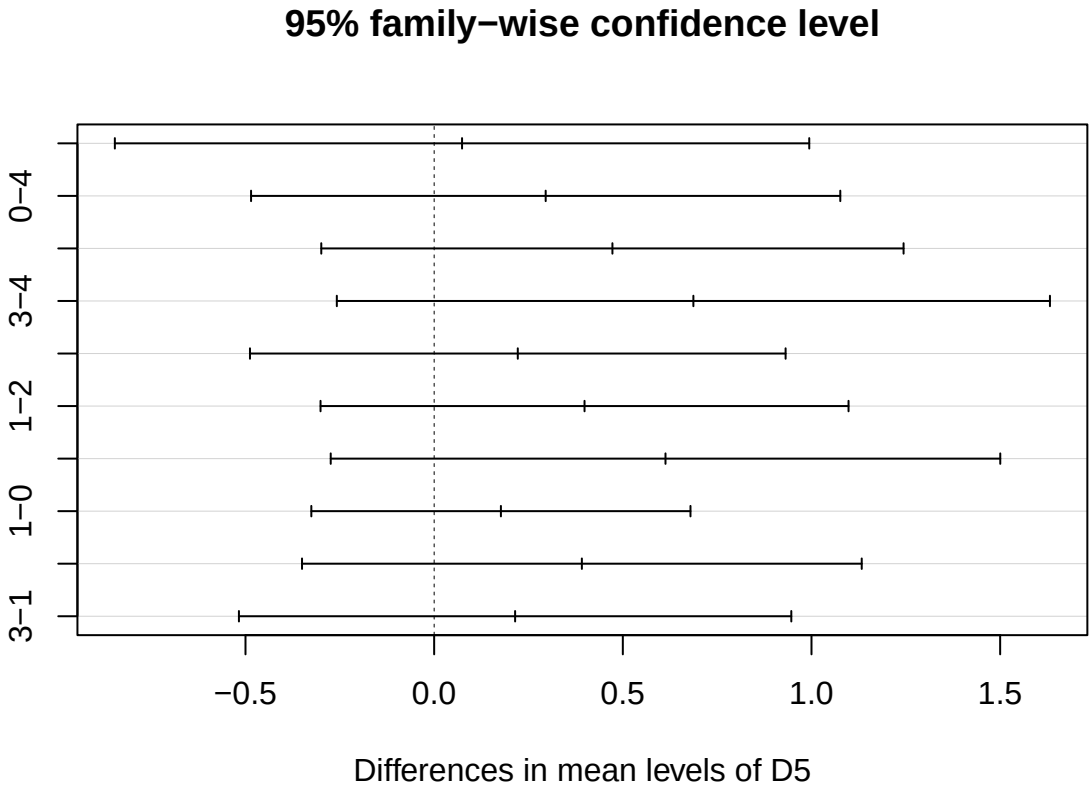
```
## Analysis of Variance Table
##
## Response: Delta
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## D1         2  0.1279 0.06397  0.1518 0.85951
## D4         4  2.4579 0.61448  1.4579 0.22630
## D5         4  3.6962 0.92405  2.1924 0.08058 .
## D6         2  0.5473 0.27363  0.6492 0.52609
## L1         4  0.9725 0.24313  0.5769 0.68050
## Residuals 60 25.2890 0.42148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

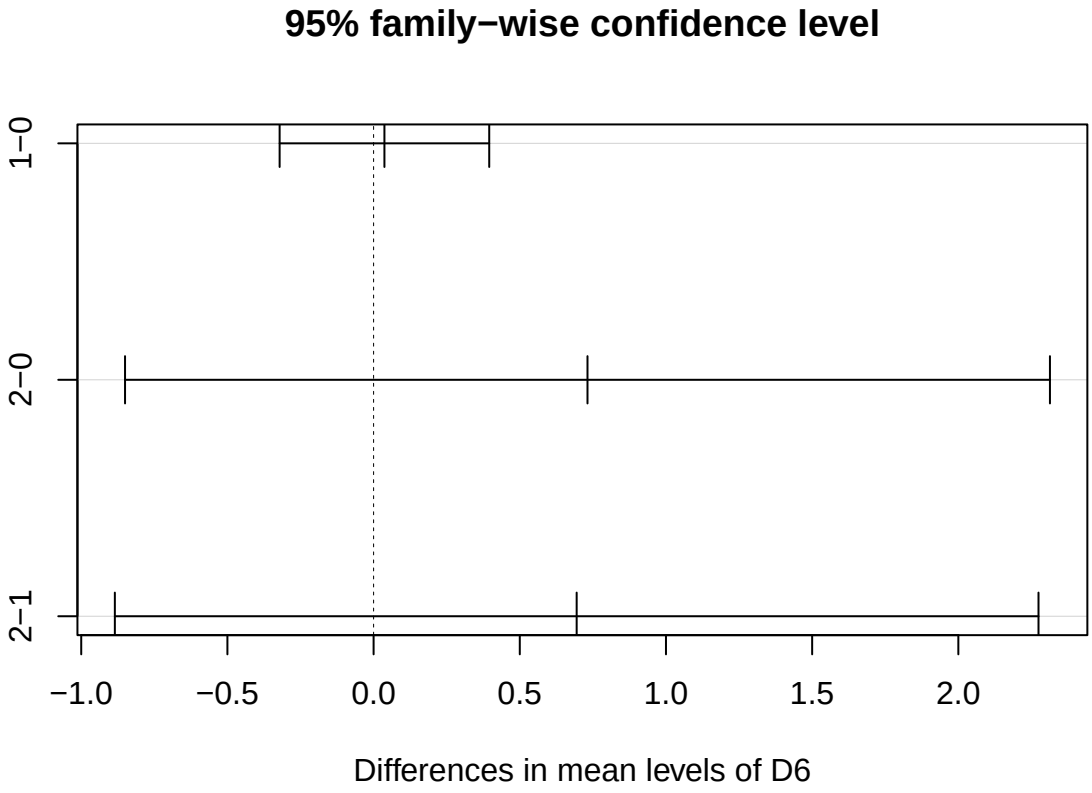
Here we see that a respondents age range ($D5$) is the only predictor which significantly explains any variability in Δ, even after accounting for the variability explained by the previous variables ($D1$ and $D4$).

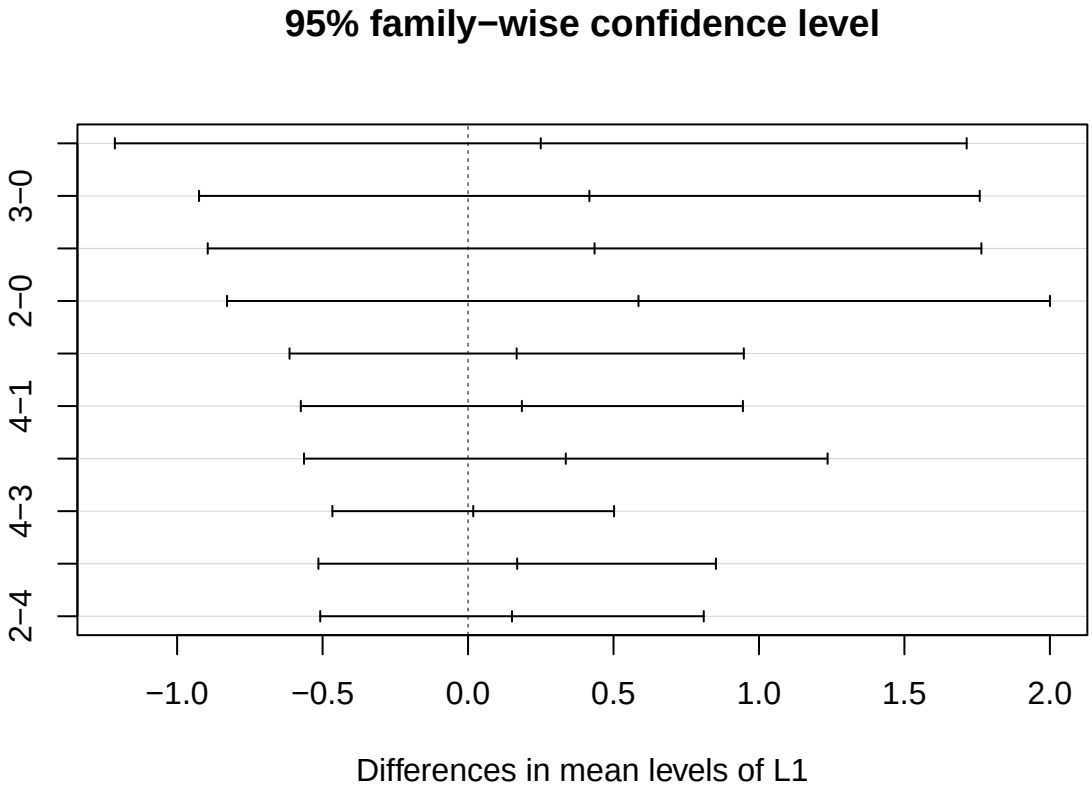
Finally, I compute compute Tukey Honest Significant Differences (HSD) to create a set of confidence intervals testing the differences between the means of the levels of each factor used as a predictor in our models.


95% family−wise confidence level
2−1
0−1
0−2
−0.5
0.0
0.5
Differences in mean levels of D1


95% family−wise confidence level
5−2
4−2
3−1
3−5
4−3
−0.5
0.0
0.5
1.0
1.5
Differences in mean levels of D4


95% family−wise confidence level
0−4
3−4
1−2
1−0
3−1
−0.5
0.0
0.5
1.0
1.5
Differences in mean levels of D5


95% family−wise confidence level
1−0
2−0
2−1
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Differences in mean levels of D6



Here we see that there are no significant differences between the means in any of the predictor variables.