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   1:46   1:28   0:26   0:35   Min.   :-1.0000
##  1: 6   1: 7   1: 7   2:27   2:13   1:26   1:37   1st Qu.: 0.0000
##  2: 3   2:10   2:10          3:18   2: 8   2: 1   Median : 0.0000
##  3:12   3:16   3:24          4: 5   3: 7          Mean   : 0.5342
##  4:27   4:28   4:30          5: 9   4: 6          3rd Qu.: 1.0000
##  5:23   5: 9                                      Max.   : 2.0000
```

```
##  L2     L1     D1     D4     D5     D6
##  0: 3   0: 3   1:53   1:31   0:30   0:36
##  1: 7   1: 7   2:29   0: 2   1:30   1:45
##  2:14   2:12          2:14   2: 8   2: 1
##  3:18   3:28          3:20   3: 8
##  4:29   4:32          4: 5   4: 6
##  5:11                 5:10
```

Regression Analysis - Importance of Teaching Climate Change

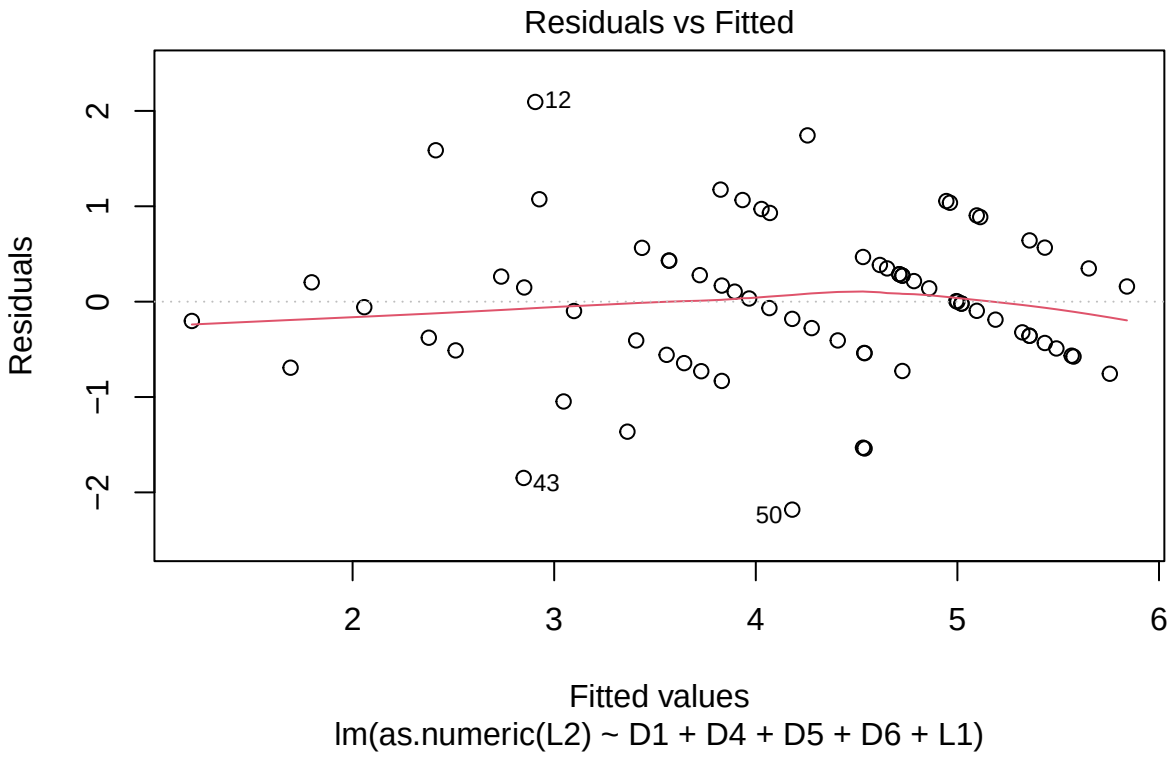
Linear Regression

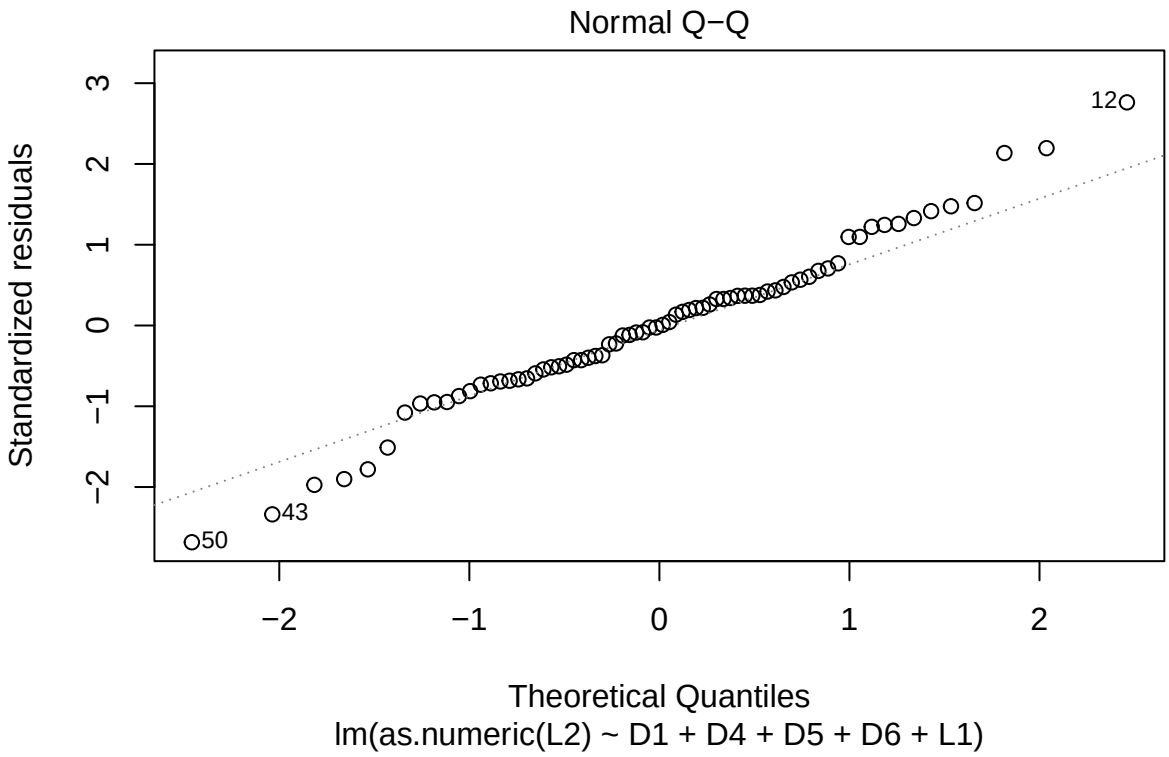
We will first test a multivariate linear regression where we treat the independent variable as continuous. Our first regression will be of the form,

$$L2 = \beta_0 + \beta_1 D1 + \beta_2 D4 + \beta_3 D5 + \beta_4 D6 + \beta_5 L1 + \epsilon$$

where DX indicates categorical demographic variables and $L1$ is an ordinal Likert scale variable. Note: $L2$ 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(L2) ~ D1 + D4 + D5 + D6 + L1)
12
43
50
2
1
0
−1
−2
2
3
4
5
6




```
##
## Call:
## lm(formula = as.numeric(L2) ~ D1 + D4 + D5 + D6 + L1, data = reg.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1808 -0.4910  0.0000  0.3846  2.0939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.85238    0.27537  13.990  < 2e-16 ***
## D12         -0.95301    0.43972  -2.167   0.0344 *
## D42         -0.24473    0.31657  -0.773   0.4427
## D43         -0.47057    0.27466  -1.713   0.0921 .
## D44         -1.15625    0.44609  -2.592   0.0121 *
## D45          0.08498    0.35620   0.239   0.8123
## D51         -0.07569    0.27599  -0.274   0.7849
## D52          1.27544    0.59187   2.155   0.0354 *
## D53         -0.58537    0.56682  -1.033   0.3061
## D54          0.29588    0.62210   0.476   0.6362
## D61          0.13307    0.22310   0.596   0.5532
## D62         -0.43362    0.92241  -0.470   0.6401
## L1.L         2.29693    0.46406   4.950 6.95e-06 ***
## L1.Q         0.46788    0.44676   1.047   0.2994
## L1.C        -0.33654    0.34412  -0.978   0.3322
## L1^4        -0.12667    0.30241  -0.419   0.6769
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8811 on 57 degrees of freedom
## Multiple R-squared:  0.6507, Adjusted R-squared:  0.5588
## F-statistic:  7.08 on 15 and 57 DF,  p-value: 2.234e-08
```

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 L2.

The above output provides β coefficients for *each level* of the categorical and ordinal predictors. For example, $D12$ indicates the output for respondents who are coded as *2* on question $D1$. This group includes respondents who identified as Program Directors. The first level of each factor is omitted as a base comparison. So, the coefficient estimate for $D12 = -0.95$ indicates that a respondent who identified as Program Director will have an average 0.95 *decrease* in their response to question $L2$ as compared to the respondents who identified as Residents/Fellows for question $D1$. This group is the omitted level, $D11$.

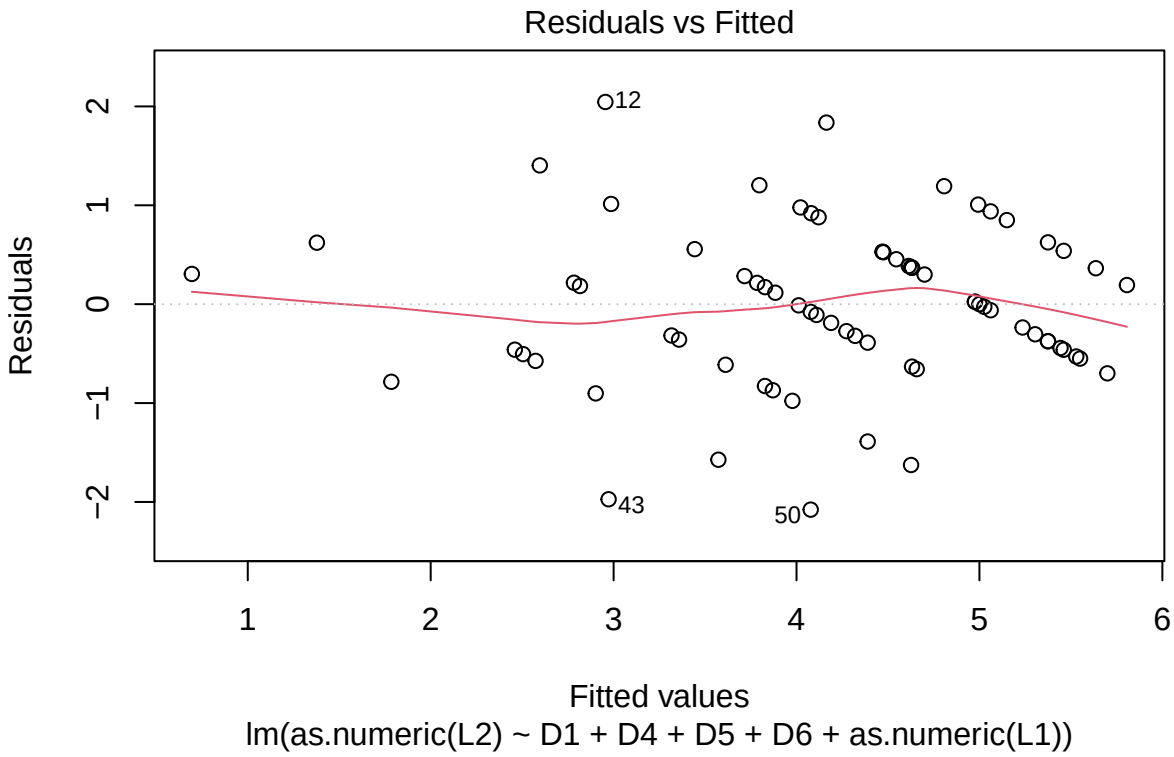
To test the effect of a predictor at the factor-level, an analysis of variance (ANOVA) is run on the model and summarized below.

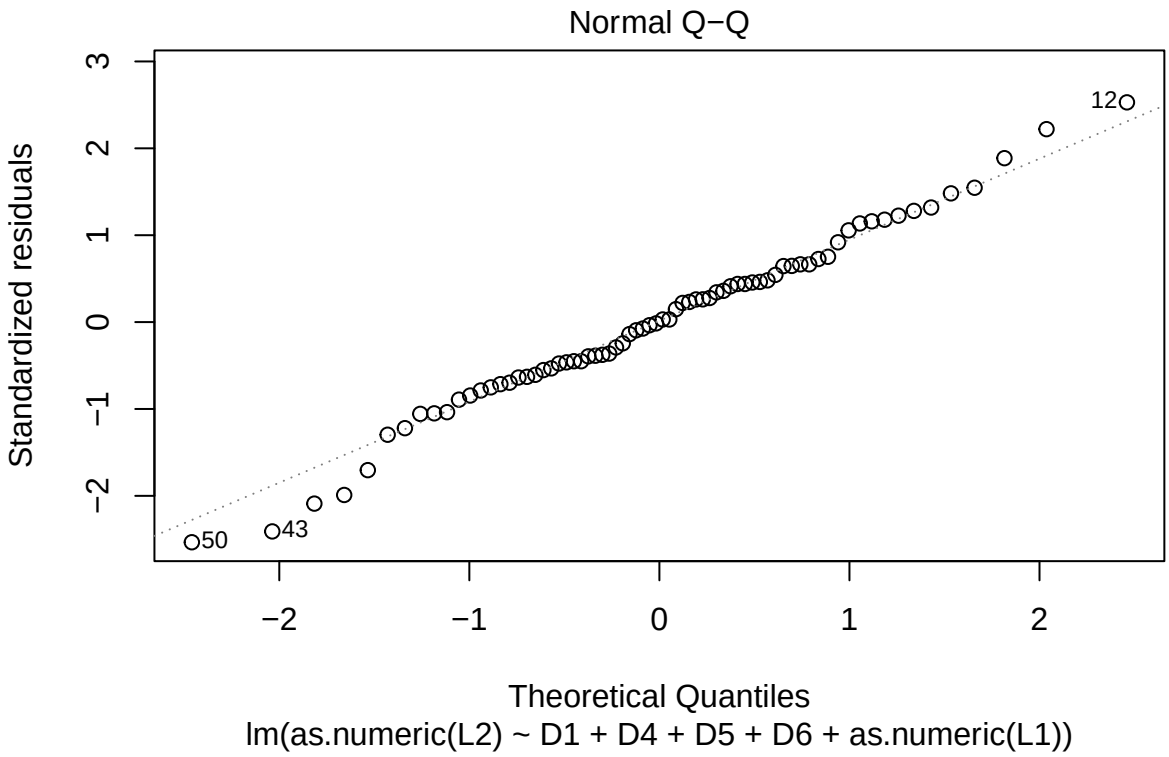
```
## Analysis of Variance Table
##
## Response: as.numeric(L2)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## D1         1  5.654  5.6543  7.2840  0.009139 **
## D4         4  9.866  2.4665  3.1773  0.019987 *
## D5         4 10.054  2.5135  3.2379  0.018341 *
## D6         2  0.497  0.2487  0.3203  0.727197
## L1         4 56.366 14.0915 18.1528 1.153e-09 ***
## Residuals 57 44.247  0.7763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

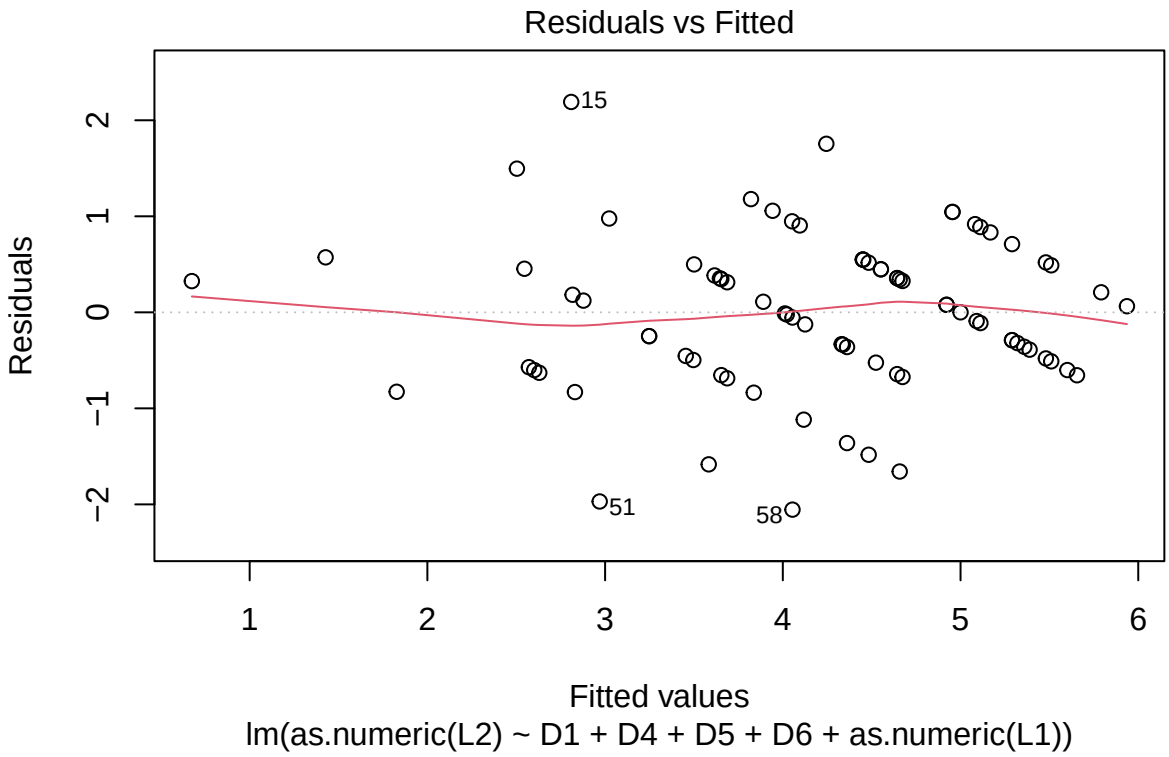
ANOVA estimates using type I SS tell us how much of the variability in $L2$ can be explained by our predictor variables *in sequential order*. This means we first measure the variability of $L2$ that is explained by $D1$, then how much of the *residual variability* can be explained by $D4$, how much of the *remaining residual variability* can be explained by $D5$, and so on, in order of their specification in the model.

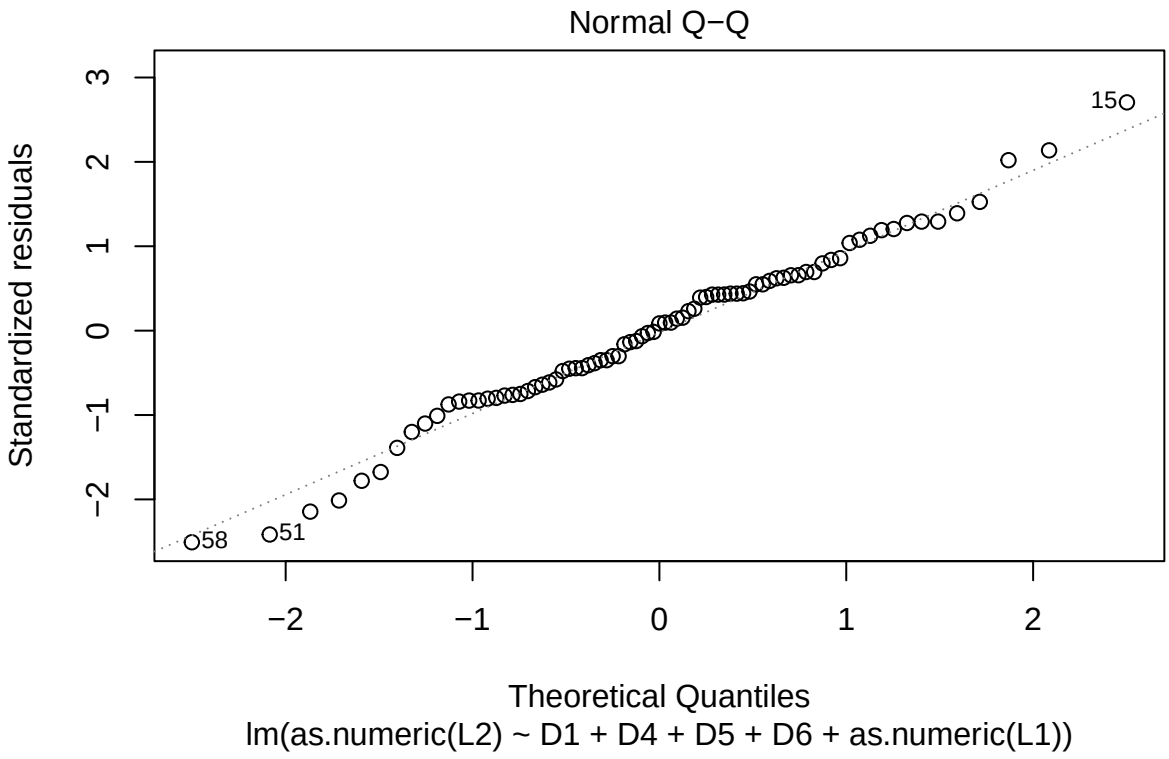
The output above indicates that the variability of $D1$ explains a significant portion of the variability in $L2$ at the 99% confidence level. Additionally, the *remaining residual variability* is signicantly explained by all the variables except $D6$.

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 $L2$, which is required to be continuous in the context of linear regression. That output is summarized below.


Residuals vs Fitted
Residuals
Fitted values
lm(as.numeric(L2) ~ D1 + D4 + D5 + D6 + as.numeric(L1))
12
43
50
2
1
0
−1
−2
1
2
3
4
5
6


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
12
50
43
−2
−1
0
1
2
Theoretical Quantiles
lm(as.numeric(L2) ~ D1 + D4 + D5 + D6 + as.numeric(L1))


Residuals vs Fitted
Residuals
Fitted values
lm(as.numeric(L2) ~ D1 + D4 + D5 + D6 + as.numeric(L1))
15
51
58
−2
−1
0
1
2
1
2
3
4
5
6




```
##
## Call:
## lm(formula = as.numeric(L2) ~ D1 + D4 + D5 + D6 + as.numeric(L1),
##     data = reg.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0776 -0.4609  0.0000  0.5250  2.0449
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.31392    0.45362   2.897  0.00526 **
## D12            -0.78625    0.41339  -1.902  0.06198 .
## D42            -0.22490    0.31143  -0.722  0.47302
## D43            -0.46784    0.27083  -1.727  0.08924 .
## D44            -1.18794    0.44126  -2.692  0.00919 **
## D45             0.10699    0.34669   0.309  0.75868
## D51            -0.08607    0.26597  -0.324  0.74735
## D52             1.02486    0.54810   1.870  0.06639 .
## D53            -0.66275    0.55340  -1.198  0.23578
## D54             0.17654    0.60159   0.293  0.77018
## D61             0.06845    0.21203   0.323  0.74793
## D62            -0.46086    0.91203  -0.505  0.61519
## as.numeric(L1)  0.82939    0.09790   8.472 7.71e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.8738 on 60 degrees of freedom
## Multiple R-squared:  0.6384, Adjusted R-squared:  0.5661
## F-statistic: 8.827 on 12 and 60 DF,  p-value: 2.114e-09
```

```
##
## Call:
## lm(formula = as.numeric(L2) ~ D1 + D4 + D5 + D6 + as.numeric(L1),
##     data = pre.reg.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05449 -0.50688  0.03181  0.49675  2.19066
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.29589    0.43678   2.967  0.00415 **
## D12            -0.79603    0.38455  -2.070  0.04225 *
## D40             0.60376    0.68107   0.886  0.37848
## D42            -0.12169    0.29452  -0.413  0.68077
## D43            -0.39838    0.25449  -1.565  0.12214
## D44            -1.14239    0.43259  -2.641  0.01025 *
## D45             0.28113    0.32543   0.864  0.39071
## D51            -0.19042    0.24668  -0.772  0.44283
## D52             0.97115    0.52358   1.855  0.06796 .
## D53            -0.66201    0.51526  -1.285  0.20322
## D54             0.09532    0.57011   0.167  0.86772
## D61             0.03018    0.19997   0.151  0.88050
## D62            -0.48014    0.90096  -0.533  0.59583
## as.numeric(L1)  0.83685    0.09412   8.892  5.3e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8663 on 68 degrees of freedom
## Multiple R-squared:  0.6345, Adjusted R-squared:  0.5646
## F-statistic:  9.08 on 13 and 68 DF,  p-value: 1.872e-10
```

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. However, we lose significance in $D12$ and $D52$.

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

```
## Analysis of Variance Table
##
## Response: as.numeric(L2)
##                Df Sum Sq Mean Sq F value    Pr(>F)
## D1              1  5.654   5.654  7.4055  0.008498 **
## D4              4  9.866   2.466  3.2303  0.018169 *
## D5              4 10.054   2.514  3.2919  0.016636 *
## D6              2  0.497   0.249  0.3257  0.723300
## as.numeric(L1)  1 54.801  54.801 71.7727 7.711e-12 ***
## Residuals      60 45.812   0.764
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: as.numeric(L2)
##                Df Sum Sq Mean Sq F value    Pr(>F)
## D1              1  6.398   6.398  8.5261  0.004745 **
## D4              5 13.220   2.644  3.5233  0.006841 **
## D5              4  9.027   2.257  3.0071  0.024046 *
## D6              2  0.605   0.303  0.4034  0.669648
## as.numeric(L1)  1 59.330  59.330 79.0600 5.299e-13 ***
## Residuals      68 51.030   0.750
## ---
## 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 $L2$ 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,

$$logit(P(L2 \leq j)) = \beta_{j0} - \eta_1 D1 - \eta_2 D4 - \eta_3 D5 - \eta_4 D6 - \eta_1 L1$$

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.

```
## # A tibble: 20 x 6
##    term  estimate std.error statistic coef.type   fake.p.value
##    <chr>    <dbl>     <dbl>     <dbl> <chr>              <dbl>
##  1 D12    -2.56       0.987   -2.60   coefficient  0.00940
##  2 D42    -0.562      0.727   -0.773  coefficient  0.440
##  3 D43    -0.810      0.639   -1.27   coefficient  0.205
##  4 D44    -3.12       1.09    -2.87   coefficient  0.00405
##  5 D45     0.816      0.810    1.01   coefficient  0.314
##  6 D51    -0.373      0.658   -0.567  coefficient  0.571
##  7 D52     3.28       1.36     2.42   coefficient  0.0157
##  8 D53    -1.49       1.21    -1.23   coefficient  0.217
##  9 D54     0.872      1.38     0.632  coefficient  0.527
## 10 D61     0.0373     0.510    0.0730 coefficient  0.942
## 11 D62    -1.37       2.11    -0.650  coefficient  0.516
## 12 L1.L    5.48       1.25     4.37   coefficient  0.0000126
## 13 L1.Q    1.32       1.07     1.23   coefficient  0.220
## 14 L1.C   -0.759      0.825   -0.920  coefficient  0.357
## 15 L1^4   -0.337      0.691   -0.488  coefficient  0.626
## 16 0|1    -5.29       1.04    -5.07   scale        0.000000402
## 17 1|2    -2.88       0.825   -3.50   scale        0.000472
## 18 2|3    -1.23       0.766   -1.60   scale        0.110
## 19 3|4     0.752      0.733    1.03   scale        0.305
## 20 4|5     4.29       0.820    5.23   scale        0.000000168
```

Ordinal logistic regression does not operate on the same normality assumptions as linear regression, rendering p-values spurious. I have nonetheless calculated p-values based on the test statistics, but these should be taken with a grain of salt. In the interpretation that follow, I focus on variables which have a relatively large test statistic (and therefore relatively small, but spurious, p-values).

The results from the ordinal logistic regression show that for respondents who identified as Program Directors, the log odds of answering 0 compared to 1, or 1 compared to 2 on $L2$ are 2.56 points *higher* compared to Residents/Fellows. That is, Program Directors are more likely to attribute less importance to the inclusion of education on climate change and mental health, compared to Residents/Fellows.

Similarly, we see that there is relatively high test-statistic on $D44$. This shows us that respondents from the SE region have a logs odds of answering 0 compared to 1, or 1 compared to 2, etc., are 3.12 points *higher* than respondents from the NE region. That is, respondents from the SE region are *more likely* to attribute less importance to the inclusion of education on climate change and mental health, compared to respondents from the NE Region.

We observe a significant effect on $D52$, which can be interpreted analogously to the variables above.

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 less importance to the inclusion of education on climate change and mental health. That is, the more worries a respondent is about climate change, the more likely they are to attribute importance to includsion of climate change curricula.

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: L2
##    LR Chisq Df Pr(>Chisq)
## D1    6.824  1  0.0089938 **
## D4   11.984  4  0.0174731 *
## D5   19.964  4  0.0005075 ***
## D6    0.455  2  0.7966594
## L1   56.804  4   1.36e-11 ***
## ---
## 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 similar results, qualitatively, to our linear regression ANOVAs. All of the variables in the model expect $D6$ explain a significant portion of the variability in $L2$ at or beyond the 95% confidence level.

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 - L2.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 $L2$ 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

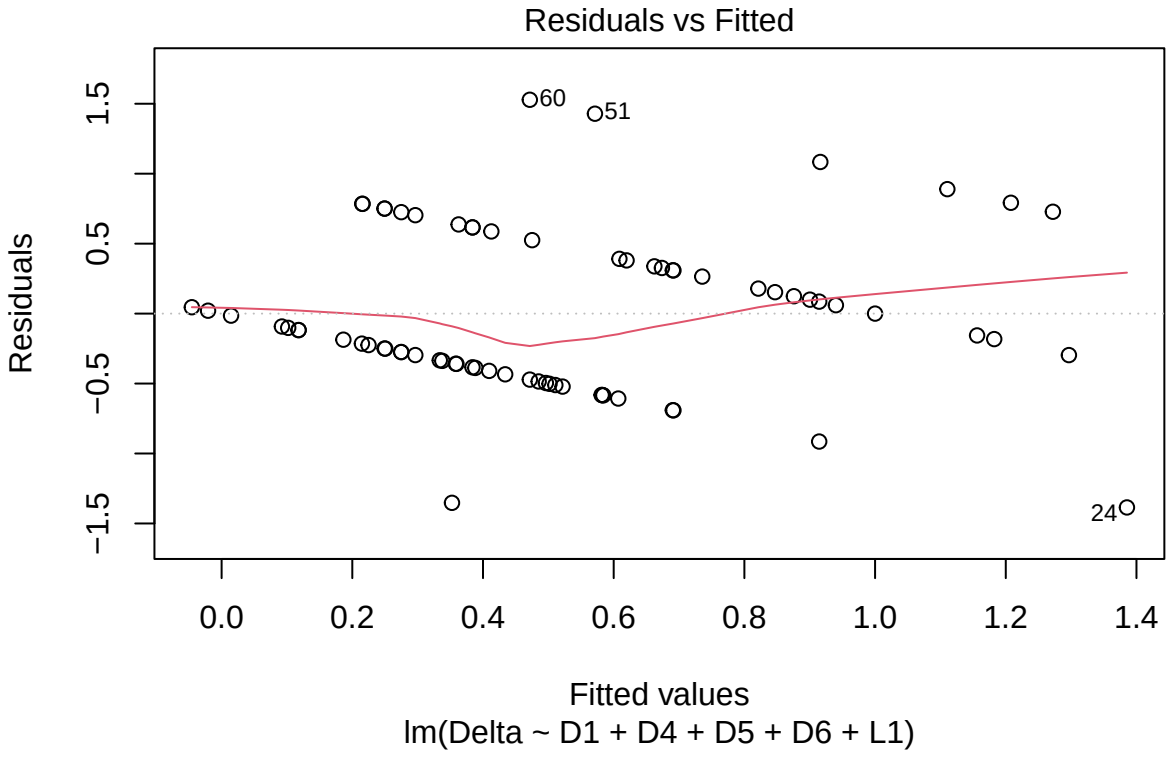
To assess the affect of the video on respondent's view on including climate change curricula, we run a multivariate linear regression where the independent variable is the change (Δ) between the response before

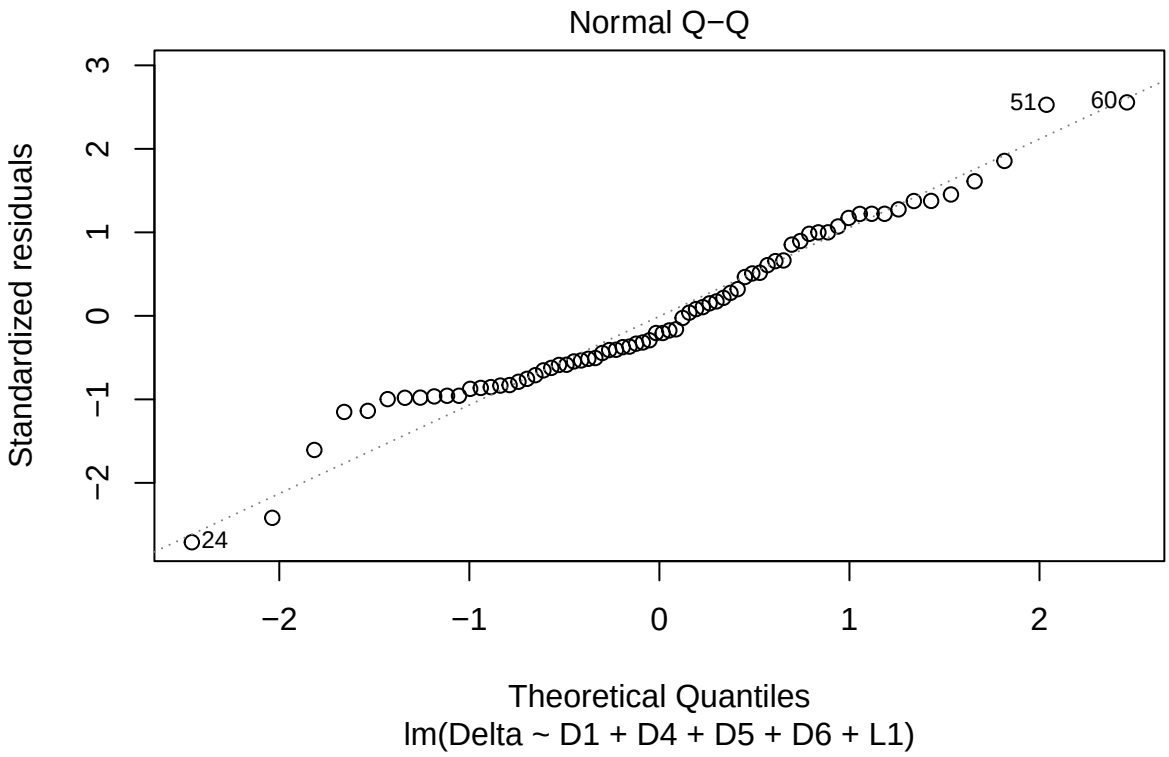
and after watching the accompanying video. We must treat this as continuous because $\Delta = L3 - L2$. Thus, our regression will be of the form,

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

where DX indicates categorical demographic variables and 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
## -1.3854 -0.3882 -0.1178  0.3803  1.5283
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.16769    0.20330   0.825   0.4129
## D12          0.58063    0.32463   1.789   0.0790 .
## D42          0.13782    0.23371   0.590   0.5577
## D43          0.33205    0.20278   1.638   0.1070
## D44          0.57191    0.32934   1.737   0.0879 .
## D45          0.26031    0.26297   0.990   0.3264
## D51          0.10898    0.20376   0.535   0.5949
## D52         -0.84112    0.43696  -1.925   0.0592 .
## D53          0.04474    0.41847   0.107   0.9152
## D54         -0.87559    0.45929  -1.906   0.0616 .
## D61         -0.02454    0.16471  -0.149   0.8821
## D62          0.72503    0.68100   1.065   0.2915
## L1.L         0.40247    0.34260   1.175   0.2450
## L1.Q        -0.36781    0.32984  -1.115   0.2695
## L1.C         0.08462    0.25406   0.333   0.7403
## L1^4         0.18890    0.22326   0.846   0.4011
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6505 on 57 degrees of freedom
## Multiple R-squared:  0.2502, Adjusted R-squared:  0.05286
## F-statistic: 1.268 on 15 and 57 DF,  p-value: 0.2525
```

Here we see none of the factors in our linear model are significant at the 95% confidence level. However, at the 10% level, $D12$, $D44$, $D52$, and $D54$ are significant.

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         1  0.1459 0.14587  0.3447 0.55942
## D4         4  2.1026 0.52565  1.2423 0.30338
## D5         4  4.0980 1.02450  2.4213 0.05868 .
## D6         2  0.5666 0.28330  0.6696 0.51592
## L1         4  1.1340 0.28350  0.6700 0.61546
## Residuals 57 24.1173 0.42311
## ---
## 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 (at 90% level) explains any variability in Δ, even after accounting for the variability explained by the previous variables ($D1$ and $D4$).

Test for Significant Difference in Means between $L2$ and $L3$

We want to use a paired test here because we are looking the same measurements from the same respondents at two different points in time. First I will perform and paired t-test.

```
##
##  Paired t-test
##
## data:  as.numeric(reg.data$L3) and as.numeric(reg.data$L2)
## t = 6.8294, df = 72, p-value = 2.277e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3783027 0.6901904
## sample estimates:
## mean of the differences
##               0.5342466
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  as.numeric(reg.data$L3) and as.numeric(reg.data$L2)
## V = 615, p-value = 1.281e-07
## alternative hypothesis: true location shift is not equal to 0
```

```
## 95 percent confidence interval:
##  0.9999677 1.0000284
## sample estimates:
## (pseudo)median
##       1.000032
```

Test for Differences in Pre and Post Survey Populations

```
## [1] "D1 : 0.836165650048555"
## [1] "D2 : 0.685946934508615"
## [1] "D3 : 0.918385345619561"
## [1] "D4 : 0.756759755718156"
## [1] "D5 : 0.831240784343665"
## [1] "D6 : 0.630971314075571"
## [1] "L1 : 0.823063008897787"
## [1] "Q1 : 0.923631572017434"
## [1] "L2 : 0.926081380836025"
```

Here we see we fail to reject the null hypothesis, that the distributions are identical, for all the variables shared by the pre and post survey populations.