MA-331 Final

Kevin Ha, Connor Haaf, Ari Birnbaum

12/14/2021

Final Project Write Up

Executive Summary

The aim of this study is to look further into the interplay between two chemicals and their relation on bone growth and bone reduction. Specifically, we look at OC and $TRAP$ then assess their impact alone and together on $VO+$ and $VO-$. As such, we run a number of statistical analysis and predictive models on these data to try to further understand their interplay.

Study

As already stated, this is a study of interactions between OC and $TRAP$ on bone growth and reduction. This is done over 31 data points of the values of OC, $TRAP$, bone growth ($VO+$) and reduction ($VO-$) for an individual. Additionally, due to some sort of skew or other imperfections in the given data, we perform these comparisons over the log of the data as well.

Software

This analysis was written and produced using tools and libraries available through R, R Studio. R Markdown was used for producing LaTeX and graphical summaries.

Methodology

In this study, we take in a set of 31 data points consisting of $VO+$, $VO-$, OC, LOC, $TRAP$, $LTRAP$, $LVO+$, and $LVO-$ (where L means log). Based on the data in this set, we assess, through regression models, the strength of association between the various data. Additionally, based on the relations among the data types, we predict what changes we should see in one as we change another.

For each such prediction, we consider the quality of the prediction using a t test and potential skewness with qq plot analysis. Using all of these tools, we build a prediction and run it both on the linearly reduced data and the logarithmically mapped data through a linear reduction.

Results and Analysis

In our study, we found that all the data was mostly normally distributed. There were slight skews in each of the graphs. Additionally, we saw good coefficients of correlation between various members of the chemical set we were observing. Regardless of this, each of the linear regressions had a positive association with various strength correlations. Additionally, we ran the same statistical modeling on the data after logging it. This resulted in about a 10% increase in

Conclusion

In conclusion, after testing for the result of changing TRAP and or OC, we were able to conclude that there was indeed a significant correlation between bone formation and resorption with these changes. This change was a positive association, and the data set could have been larger to help work out the skew that we observed. In lieu of this, we took the log of the data to help normalize it and found that the correlation with OC increased by about 10%. Overall, we conclude that there is a correlation between this formation and resorption, and this would likely be better represented with a larger sample.

11.37

a.

The plot of the residuals has a slight concave curve down when compared with the line of best fit.

b.

$TRAP_p = 0.00158 < OC_p = 0.22102$. Since $TRAP$'s p-value is significantly lower than OC's, $TRAP$ acts as a better predictor of bone formation, $VO+$.

This view is consistent with that of 11.36, since $TRAP$ had a smaller standard deviation of 6.5282397 when compared with OC's 19.6097441. Thus, $TRAP$'s volatility was lower, meaning more consistent results.

11.38

Based on the above table, the standard deviation for the estimated coefficients of the intercept, OC, and $TRAP$ are respectively $288.8035, 7.0965, \text{and } 24.4286$. While OC's is the smallest, the intercept and $TRAP$'s coefficient prove to be very volatile between the three tables.

The same volatility is apparent in the p-values of the data set as well. Specifically, note the incredibly small p-values of 11.38's $VO-$ and 11.37.a's OC.

11.39

Given that the data is skewed right, the logarithmic data set is more normalized. When we have a reduced skew, we're modeling an increased input set. Meaning, our model becomes more predictive and the correlation values are higher.

11.40

For 11.37

a.

The plot of the residuals again has a slight concave curve down when compared with the line of best fit.

b.

$TRAP_p = 0.291 > OC_p = 0.016$. Since $TRAP$'s p-value is significantly greater than OC's, OC acts as a better predictor of bone formation, $VO+$.

This view is consistent with that of 11.36, since OC had a smaller standard deviation of 6.5282397 when compared with OC's 19.6097441. Thus, $TRAP$'s volatility was lower, meaning more consistent results.

For 11.38

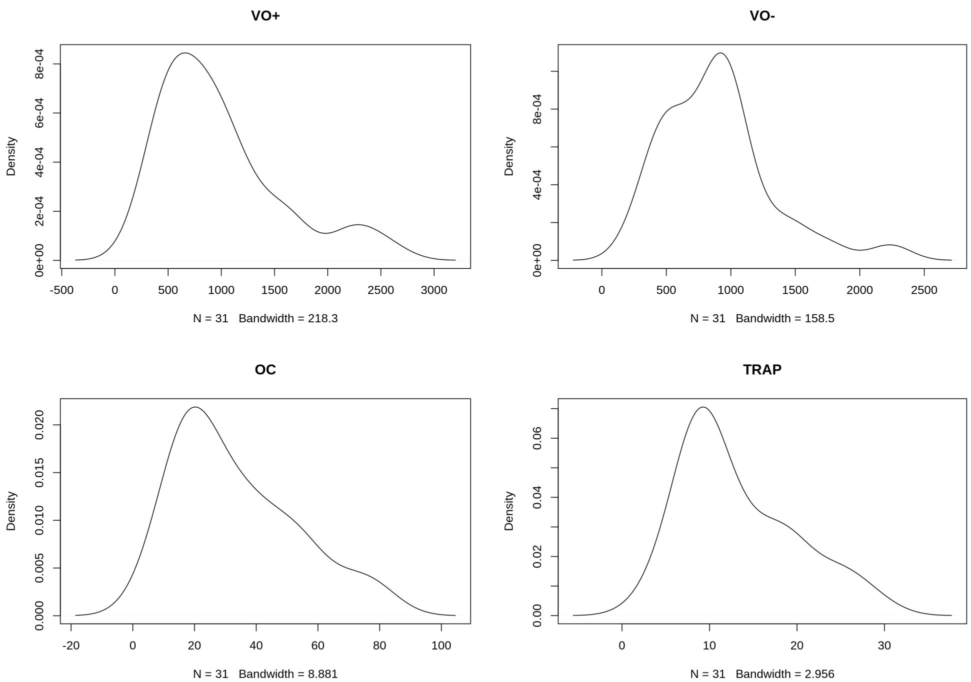
c.

Based on the above table, the estimated coefficients for the intercept, OC, and $TRAP$ are far more stable when $VO+$ and $VO-$ are revered.

The same is apparent for the p-values of the data set as well.

Problem 11.36

Problem 11.36a



Numerical summary for VO+:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   285.0   542.5   870.0   985.8  1188.5  2545.0
```

$s_{VO+} = 579.8580527$

Numerical summary for VO-:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   254.0   554.0   903.0   889.2  1023.0  2236.0
```

$s_{VO-} = 427.6161378$

Numerical summary for OC:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.10   18.60   30.20   33.42   46.05   77.90
```

$s_{OC} = 19.6097441$

Numerical summary for TRAP:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.30    8.90   10.30   13.25   18.80   28.80
```

$s_{TRAP} = 6.5282397$

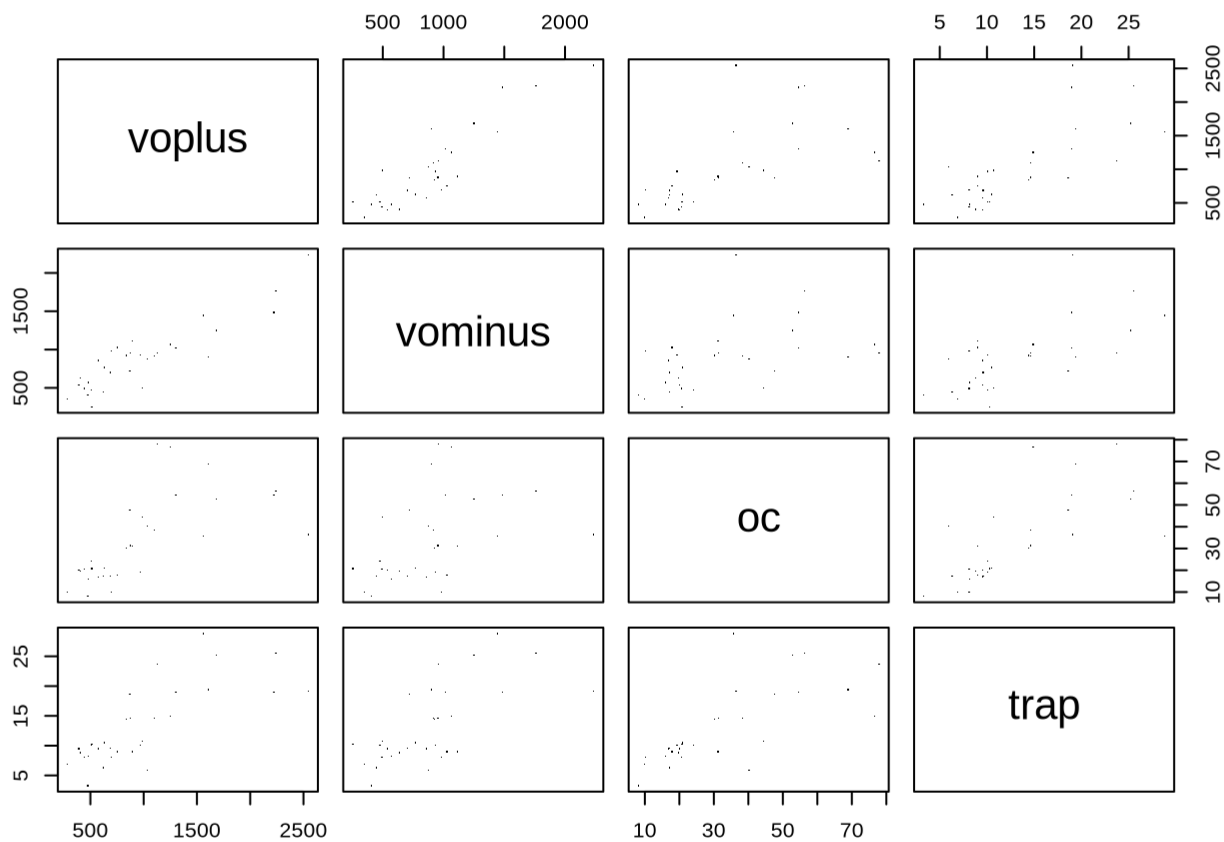
Graphical summary for all and the distribution: **Right skewed.**

Problem 11.36b

Numerical summary:

```
##            voplus   vominus        oc      trap
## voplus  1.0000000 0.8957707 0.6596140 0.7648649
## vominus 0.8957707 1.0000000 0.4547603 0.6779267
## oc      0.6596140 0.4547603 1.0000000 0.7298519
## trap    0.7648649 0.6779267 0.7298519 1.0000000
```

Graphical Summary:

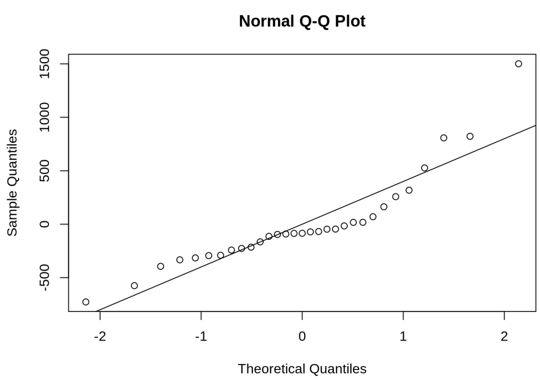


All pairs exhibit a linear and positive association with varying strength as determined by their correlation numbers.

Problem 11.37

Problem 11.37a

```
##
## Call:
## lm(formula = voplus ~ oc, data = biomark)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -727.45 -234.43  -85.08   43.66 1500.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  334.034    159.241   2.098   0.0448 *
## oc            19.505      4.127   4.726 5.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 443.3 on 29 degrees of freedom
## Multiple R-squared:  0.4351, Adjusted R-squared:  0.4156
## F-statistic: 22.34 on 1 and 29 DF,  p-value: 5.429e-05
```



$VO+ = 334 + 19.51OC$

The plot of residuals against OC is sightly curved.

Problem 11.37b

```
##
## Call:
## lm(formula = voplus ~ oc + trap, data = biomark)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -708.2 -198.6 -100.2  125.8 1224.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   57.704    156.539   0.369  0.71518
## oc             6.415      5.125   1.252  0.22102
## trap          53.874     15.393   3.500  0.00158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 376.3 on 28 degrees of freedom
## Multiple R-squared:  0.607,  Adjusted R-squared:  0.5789
## F-statistic: 21.62 on 2 and 28 DF,  p-value: 2.096e-06
```

$VO+ = 58 + 6.41OC + 53.9TRAP$

The view that TRAP is a better predictor of bone formation than OC is consistent with the results from 11.36 as the p-value for TRAP is significantly lower than the p-value for OC.

Problem 11.38

Problem 11.38a

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} + \epsilon_i$

OR

$VO+ = \beta_0 + \beta_1 OC + \beta_2 TRAP + \beta_3 VOMINUS + \epsilon_i$

Assumption: ϵ_i is assumed to be independent and $N(0, \sigma)$

Problem 11.38b

```
##
## Call:
## lm(formula = voplus ~ oc + trap + vominus, data = biomark)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -364.19 -158.57  -15.13  120.08  441.11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -243.4877    94.2183  -2.584  0.01549 *
## oc             8.2349     2.8397   2.900  0.00733 **
## trap           6.6071    10.3340   0.639  0.52797
## vominus        0.9746     0.1211   8.048  1.2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 207.8 on 27 degrees of freedom
## Multiple R-squared:  0.8844, Adjusted R-squared:  0.8715
## F-statistic: 68.84 on 3 and 27 DF,  p-value: 9.031e-13
```

Problem 11.38c

	EC	SE	T-Statistics	P-value	EC	SE	T-Statistics	P-value	EC	SE	T-Statistics	P-value
Intercept	-243.49	94.22	-2.58	0.01	57.70	156.54	0.37	0.72	334.03	159.24	2.10	0.04
OC	8.23	2.84	2.90	0.007	6.41	5.13	1.25	0.22	19.51	4.13	4.73	5.43e^-5
TRAP	6.61	10.33	0.64	0.53	53.87	15.39	3.50	0.002				
VOMINUS	0.97	0.12	8.05	1.2e^-8								

Blue = Problem 11.38
Red = Problem 11.37b
Green = Problem 11.37a

EC = Estimated Coefficient

SE = Standard Error

The coefficients and P-values differ significantly between all three tables.

Problem 11.38d

11.37a: Percent of variation is 43.51% (obtained from Multiple R-squared section in the summary) and the estimate of $\sigma = 443.3$

11.37b: Percent of variation is 60.7% (obtained from Multiple R-squared section in the summary) and the estimate of $\sigma = 376.3$

11.38b: Percent of variation is 88.44% (obtained from Multiple R-squared section in the summary) and the estimate of $\sigma = 207.8$

Problem 11.38e

11.38b suggests that TRAP coefficient is not significant, so remove it.

```
##
## Call:
## lm(formula = voplus ~ oc + vominus, data = biomark)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -350.25 -153.94  -13.22  148.19  428.09
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -234.14400   92.09009  -2.543 0.016818 *
## oc             9.40388    2.14964   4.375 0.000153 ***
## vominus        1.01857    0.09858  10.333 4.65e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 205.6 on 28 degrees of freedom
## Multiple R-squared:  0.8826, Adjusted R-squared:  0.8742
## F-statistic: 105.3 on 2 and 28 DF,  p-value: 9.418e-14
```

$VO+ = -234.14 + 9.40OC + 1.02VO-$

Better performance in terms of P-value compared to part B.

Problem 11.39

```
##         lvoplus  lvominus       loc     ltrap
## lvoplus       1 0.8396741 0.7735853 0.7549684
```

	Dependent variable:			
	lvoplus			
	(1)	(2)	(3)	(4)
loc	0.706	0.430	0.392	0.406
	p = 0.00000	p = 0.017	p = 0.003	p = 0.00004
	t = 6.574***	t = 2.560**	t = 3.398***	t = 4.925***
ltrap		0.424	0.028	
		p = 0.049	p = 0.862	
		t = 2.066**	t = 0.176	
lvominus			0.673	0.682
			p = 0.00001	p = 0.00000
			t = 5.710***	t = 6.569***
Constant	4.385	4.259	0.872	0.832
	p = 0.000	p = 0.000	p = 0.185	p = 0.169
	t = 12.036***	t = 12.147***	t = 1.361	t = 1.413
Observations	31	31	31	31

R^2	0.598	0.652	0.842	0.842
Adjusted R^2	0.585	0.627	0.825	0.831
Residual Std. Error	0.358 (df = 29)	0.339 (df = 28)	0.233 (df = 27)	0.229 (df = 28)
F Statistic	43.217*** (df = 1; 29)	26.178*** (df = 2; 28)	48.015*** (df = 3; 27)	74.588*** (df = 2; 28)
Note:				*p<0.1;* ***p<0.05;*** p<0.01

$$LVO+ = 4.38 + 0.706LOC$$

$$LVO+ = 4.26 + 0.430LOC + 0.424LTRAP$$

$$LVO+ = 0.872 + 0.392LOC + 0.028LTRAP + 0.673LVO-$$

Model without LTRAP since it is insignificant (#4 in the table above):

$$LVO+ = 0.832 + 0.406LOC + 0.682LVO-$$

Problem 11.40

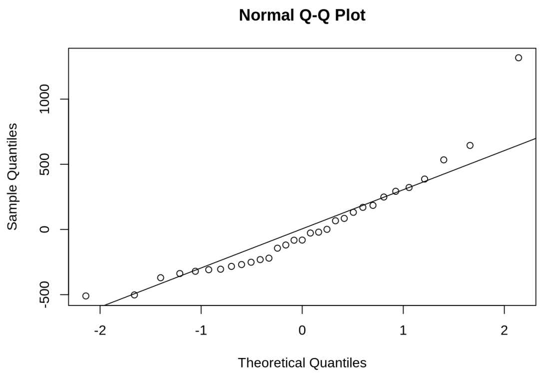
11.36:

```
##          vominus    voplus        oc      trap
## vominus       1 0.8957707 0.4547603 0.6779267
```

(Graphs remain the same and other numerical summaries remain the same)

11.37: (#1 is Part A, #2 is Part B)

	Dependent variable:	
	vominus	
	(1)	(2)
oc	9.917	-1.868
	p = 0.011	p = 0.676
	t = 2.750**	t = -0.423
trap		48.501
		p = 0.002
		t = 3.655***
Constant	557.818	309.051
	p = 0.0004	p = 0.030
	t = 4.009***	t = 2.290**
Observations	31	31
R^2	0.207	0.463
Adjusted R^2	0.179	0.425
Residual Std. Error	387.351 (df = 29)	324.353 (df = 28)
F Statistic	7.561** (df = 1; 29)	12.071*** (df = 2; 28)
Note:		*p<0.1;* ***p<0.05;*** p<0.01



The plot of residuals against OC is sightly curved again.

$VO- = 557.82 + 9.92OC$ (Part A)

$VO- = 309.05 - 1.87OC + 48.5TRAP$ (Part B)

11.38 (#1 is Part B, #2 is Part E)

	Dependent variable:	
	vominus	
	(1)	(2)
oc	-6.513	-5.254
	p = 0.016	p = 0.026
	t = -2.598**	t = -2.360**
trap	9.485	
	p = 0.291	
	t = 1.079	
voplus	0.724	0.778
	p = 0.000	p = 0.000
	t = 8.048***	t = 10.333***
Constant	267.261	298.012
	p = 0.002	p = 0.0002
	t = 3.577***	t = 4.302***
Observations	31	31
R^2	0.842	0.835
Adjusted R^2	0.824	0.823
Residual Std. Error	179.162 (df = 27)	179.688 (df = 28)
F Statistic	47.966*** (df = 3; 27)	70.949*** (df = 2; 28)
Note:		*p<0.1;* ***p<0.05;*** p<0.01

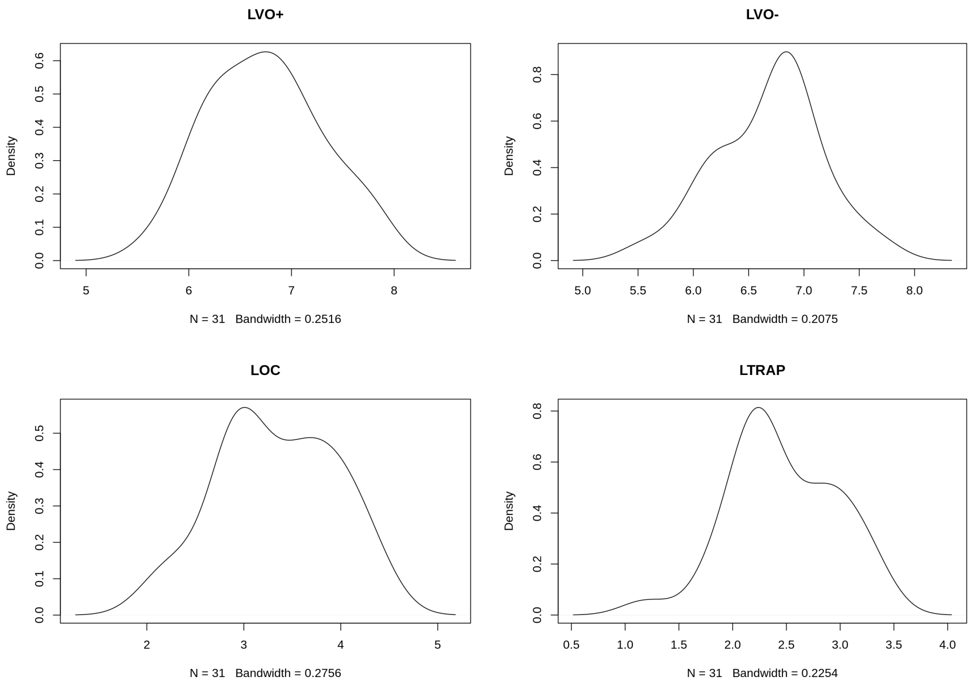
$VO- = 267.26 - 6.51OC + 9.48TRAP + 0.72VO+$ (Part B)

$VO- = 298.01 - 5.254OC + 0.78VO+$ (Part E)to me cause i dont

Problem 11.41

11.36:

```
##          lvominus   lvoplus      loc     ltrap
## lvominus        1 0.8396741 0.554607 0.6643005
```



Slight left skew when compared to the non-logarithmic versions of the graph

Numerical summary for LVO+:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.652   6.295   6.768   6.742   7.079   7.842
```

$s_{LVO+} = 0.5554948$

Numerical summary for LVO-:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.537   6.316   6.806   6.682   6.931   7.712
```

$s_{LVO-} = 0.4832368$

Numerical summary for LOC:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.092   2.922   3.408   3.338   3.829   4.355
```

$s_{LOC} = 0.6085089$

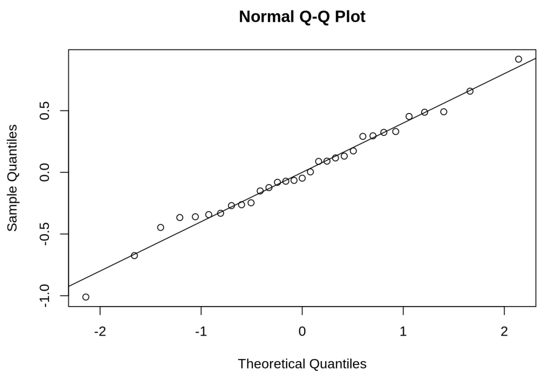
Numerical summary for LTRAP:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.194   2.186   2.332   2.467   2.933   3.360
```

$s_{LTRAP} = 0.4978053$

11.37: (#1 is Part A, #2 is Part B)

	Dependent variable:	
	lvominus	
	(1)	(2)
loc	0.440	0.057
	p = 0.002	p = 0.762
	t = 3.589***	t = 0.307
ltrap		0.590
		p = 0.015
		t = 2.611**
Constant	5.211	5.037
	p = 0.000	p = 0.000
	t = 12.524***	t = 13.063***
Observations	31	31
R^2	0.308	0.443
Adjusted R^2	0.284	0.403
Residual Std. Error	0.409 (df = 29)	0.373 (df = 28)
F Statistic	12.883*** (df = 1; 29)	11.142*** (df = 2; 28)
Note:		*p<0.1;* ***p<0.05;*** p<0.01



The plot of residuals against OC is sightly curved again.

$LVO- = 5.21 + 0.44LOC$ (Part A)

$LVO- = 5.04 - 0.06LOC + 0.59LTRAP$ (Part B)

11.38 (#1 is Part B, #2 is Part E)

	Dependent variable:	
	lvominus	
	(1)	(2)
loc	-0.293	-0.188
	p = 0.047	p = 0.141
	t = -2.083**	t = -1.519
ltrap	0.245	
	p = 0.153	
	t = 1.472	
lvoplus	0.813	0.890

	p = 0.00001	p = 0.00000
	t = 5.710***	t = 6.569***
Constant	1.573	1.311
	p = 0.025	p = 0.054
	t = 2.377**	t = 2.015*
Observations	31	31
R^2	0.748	0.728
Adjusted R^2	0.720	0.708
Residual Std. Error	0.256 (df = 27)	0.261 (df = 28)
F Statistic	26.679*** (df = 3; 27)	37.377*** (df = 2; 28)
Note:		*p<0.1;* ***p<0.05;*** p<0.01

$LVO- = 1.57 - 0.29LOC + 0.25LTRAP + 0.81LVO+$ (Part B)

$LVO- = 1.31 - 0.19LOC + 0.89LVO+$ (Part E)