**Exploring Okcupid Personal Statements:**

**Through Age, Gender, Sexual Orientation and Relationship Status**

**Background:**

Online dating has become increasingly prevalent in recent years. 20% of current committed relationship began online and 40% of Americans use online dating. The goal of this experiment was to examine the connection between different types of people and the amount they wrote on their online dating profile.

**Description of the experiment:**

The San Francisco OKCupid Users data set was collected in June 2012 with the use of a python script that pulled data from public profiles on [www.okcupid.com](http://www.okcupid.com). It has 59946 entries, which includes people within a 25-mile radius of San Francisco, who were online in the last year (June 2011), with at least one profile picture. The original dataset includes 32 variables but for this analysis I will only be utilizing five of them: Age, Gender, Sexual Orientation, Relationship Status and Essay (Personal Statements).

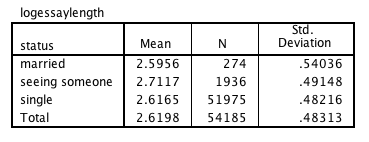
The gender and orientation variables need not require any cleaning however the age variable had to be converted from an integer to a factor. I did this by creating 4 groups based on the values of the quartiles. The first group consists of people age (since all users are 18 or older) 18-26, the second group is ages 26-32, the third is 32-37 and the fourth is all people over age 37. I also removed two entries whose reported age was over 100 since these were obviously an error. The status variable originally had four (married, seeing someone, single and unknown) levels however I removed the cases of unknown (10 entries) since I did not think they added to the analysis and caused issues with group size. For my response I took the essay0 or “My self summary” variable and converted the text into a numeric factor based on the number of characters in the essay. Finally I removed all the cases where the respondent did not write anything in their personally essay since this analysis is focused on what factors cause someone to write more not wither or not they wrote anything. I also removed any incomplete entries. After all of this the data set includes 54185 entries so 5761 entries in total were removed during the data cleaning process.

**The data and the experimental design:**

Since I have 4 factors of interest and no blocking I originally wanted to use a four-way Anova as my model. However 2 of the groups in the four-way had only 1 entry. In order to combat this I then decided to do a series of four three-way Anovas which would give me all the same interactions besides the 4 way interaction (age\*gender\*orientation\*status). However two of these three-way anovas still included groups with less then 10 entries so I decided to conduct two three-way anovas (age\*gender\*orientation and age\*gender\*status) since in both of these models there are at least 10 entries in each group (see tables 5 and 6).

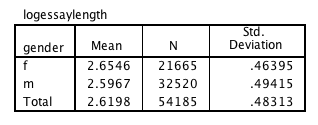
For all four factors there appear to be differences in the means. For the status variable (see Table 1) the mean for the married is the lowest at 394.09) (10^2.5956=394.09) characters and the seeing someone mean is the highest with an average of 514.87 (10^2.7117=514.87) characters written. The mean for single people falls a bit above married people at 413.52 (10^2.6165=413.52) characters. There is only a difference of about 20 characters between married and single people however there is a difference of about a 120 between married people and people seeing someone which seems like a significant difference. There is also about a 100 character difference between single people and people seeing someone which also indicates there is likely a significant difference between these means.

Table Comparison of means for married, seeing someone and single groups



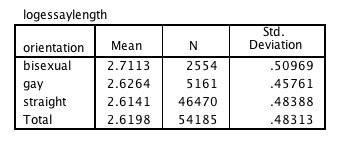
The means for the two groups in the gender factor also have an apparent difference. The average number of characters women write is 451.43 (10^2.6546=451.43) and for men it is 395.09 (10^2.5967=395.09) characters. This amounts to a difference of about 56 characters, which is about the length of one sentence so it seems like this is a notable difference.

Table Comparison of means for f (female) and m (male) groups



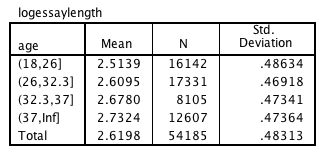
The average number of characters for the sexual orientation factor is 514.39 (10^2.7113) characters for bisexual people, 423.05 (10^2.6264) characters for gay people and 411.24 (10^ 2.6141) characters for straight people (see Table 3). The difference in means seems to lie mainly between gay and straight people and bisexual people since only 12 characters, one or two words, separate straight and gay people.

Table Comparison of means for bisexual, gay and straight groups



For the age factor there does appear to be difference between all four of the means (see table 4) and that the number of characters used increased with age. The 18-26 group used an average of 326.51 (10^2.5139) characters. The 26-32 group used 406.91 (10^2.6095) characters on average. The 32-37 group used 476.43 (10^2.6780) characters on average and the 37 and older group used 540.01 (10^2.6780) characters on average.

Table Comparison for means between 18-26 age group, 26-32 age group, 32-37 age group and 37 and above age group



The essay length variable required a transformation since the plots for all four factors indicated that there was a difference in spread across the groups. These plots also show a large number of extreme values due to this spread problem. (see fig 1-4). After trying square root, reciprocal, reciprocal of the square root and log transformations (see fig 5-8) it was clear that the log transformation yielded the greatest improvement. Even after the transformation the plots still indicate a difference in spread across groups (fig 8-11) however this is somewhat expected in a data set of this size so I don’t think it is a concern at this time. Additionally there are still a number of values that SPSS identifies as extreme however with such a large data set this is not unexpected since there is more room for variation.

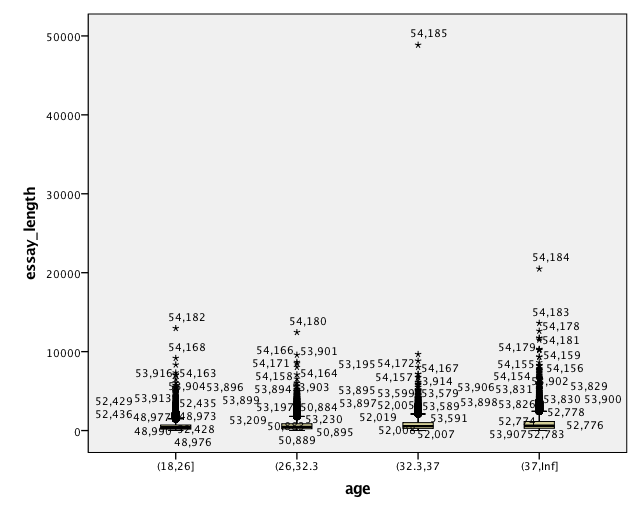
****

Figure Boxplot of age vs. essay length prior to transformation

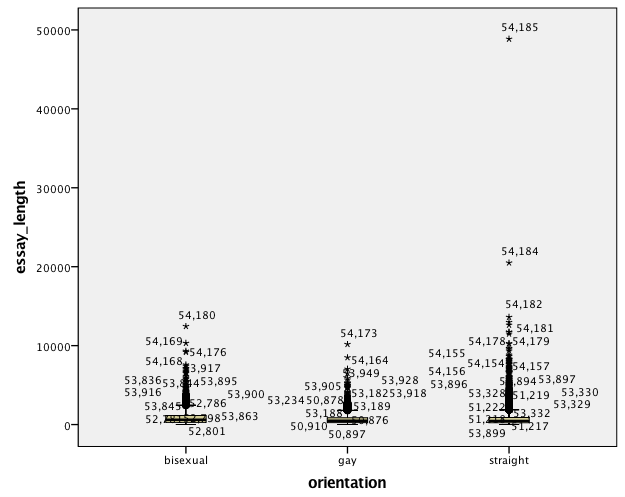
****

Figure Boxplot of orientation vs. essay length prior to transformation

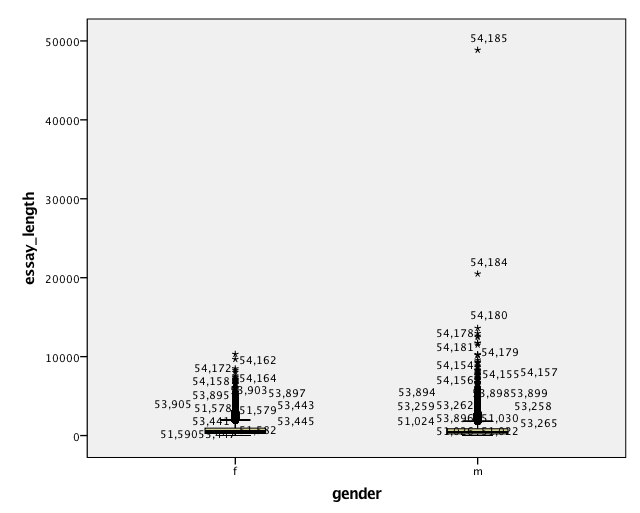
****

Figure Boxplot of gender vs. essay length prior to transformation

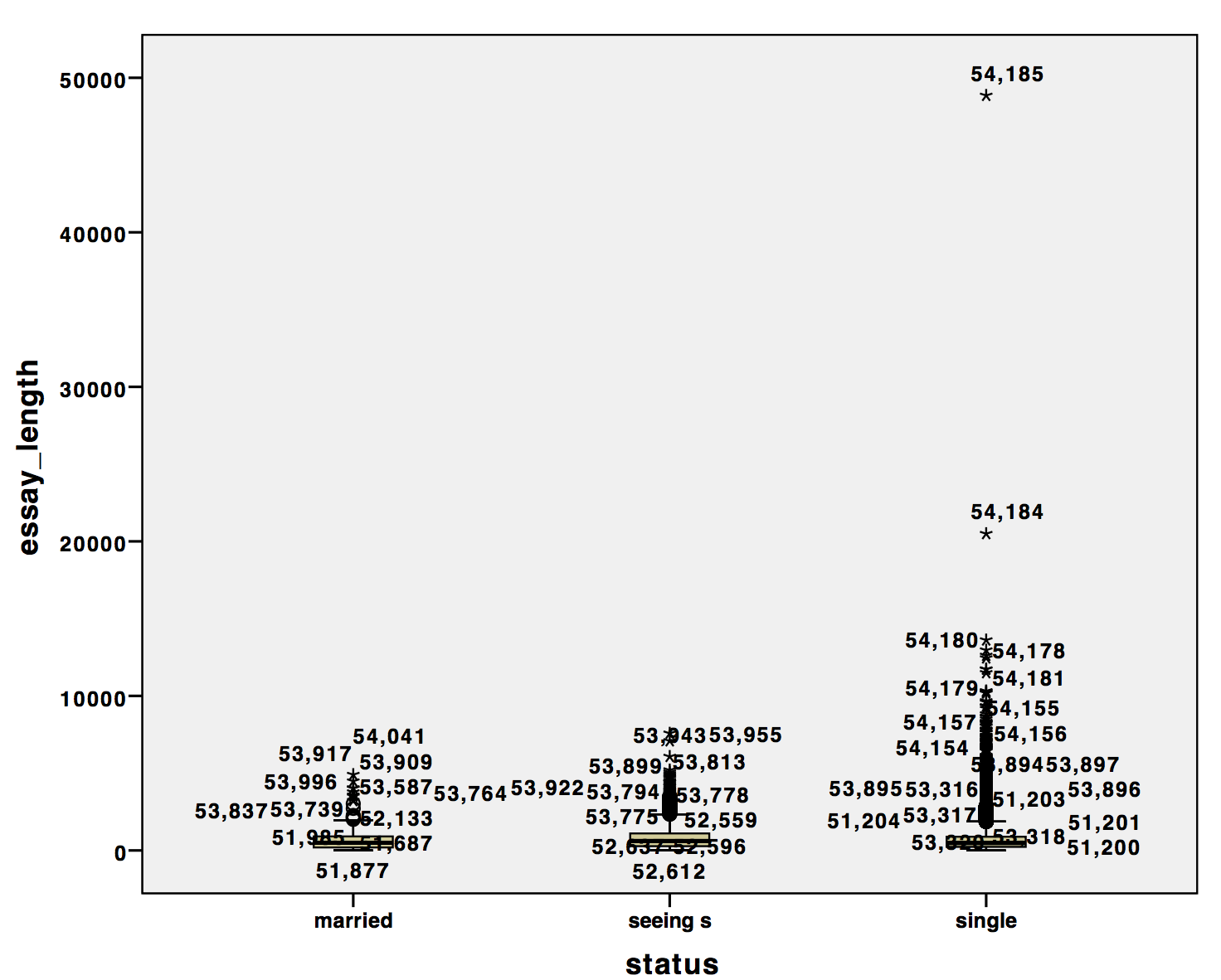
****

Figure Boxplot of status vs. essay length prior to transformation

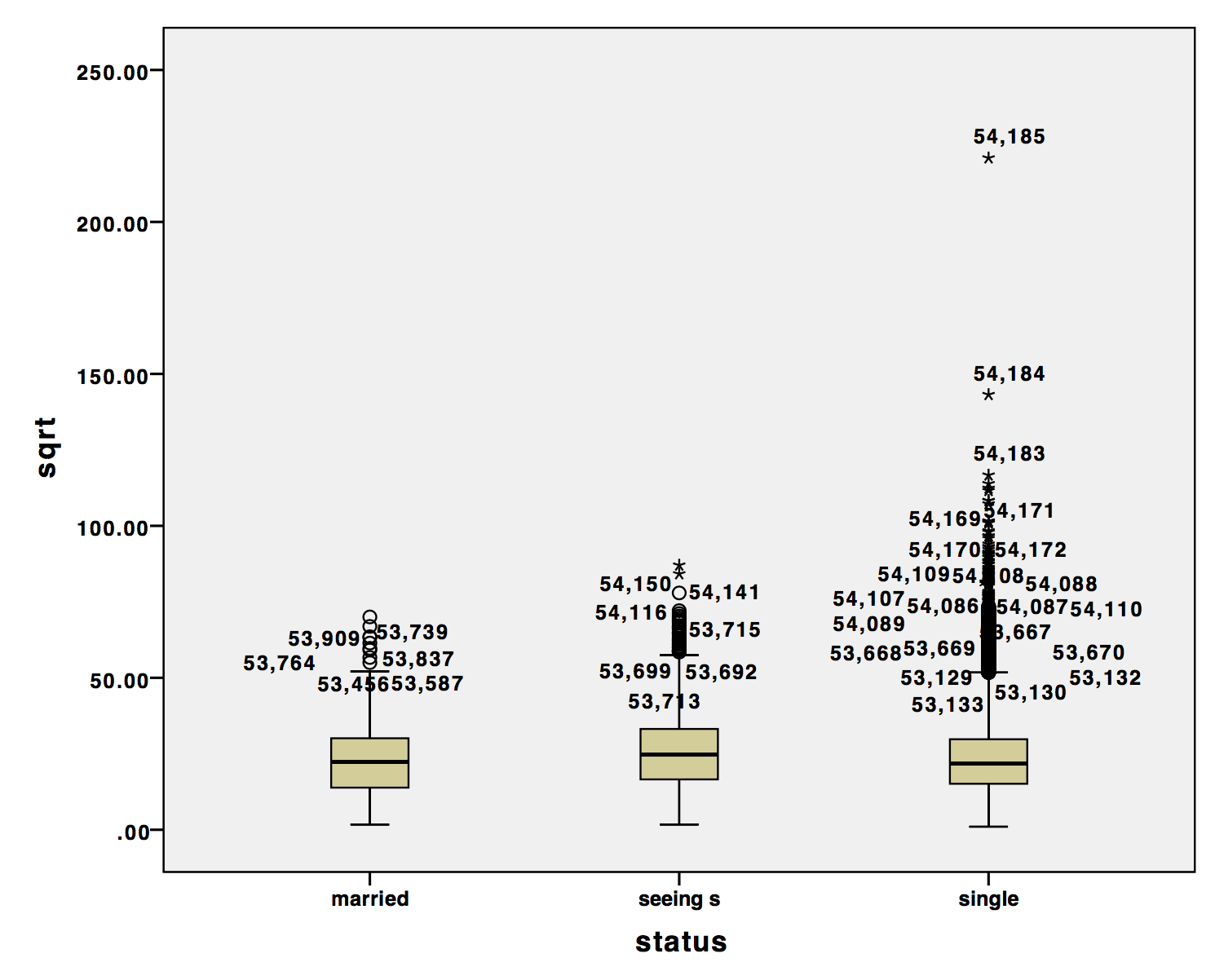
****

Figure Square Root transformation of essay length vs status

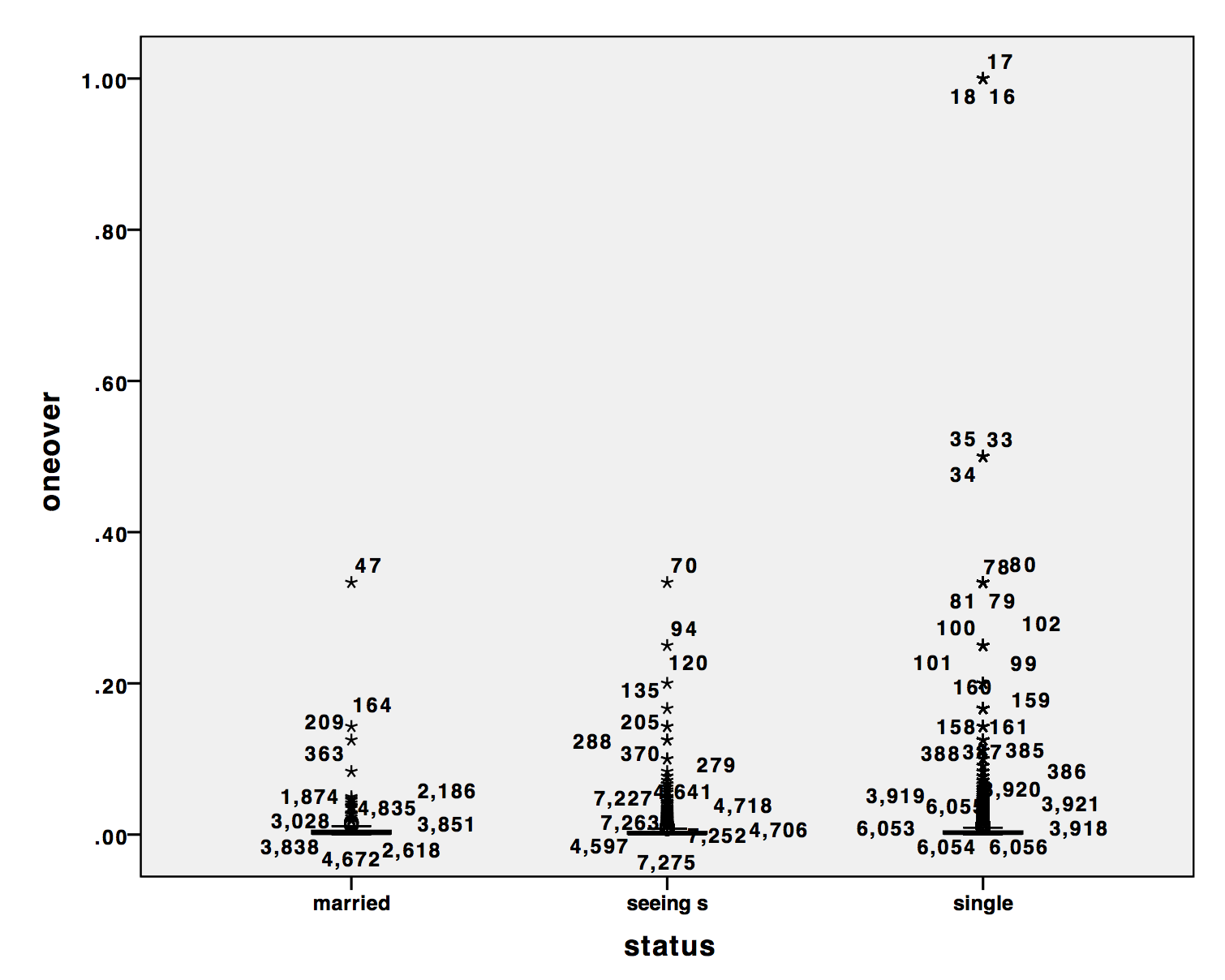
****

Figure Reciprocal transformation of essay length vs. status

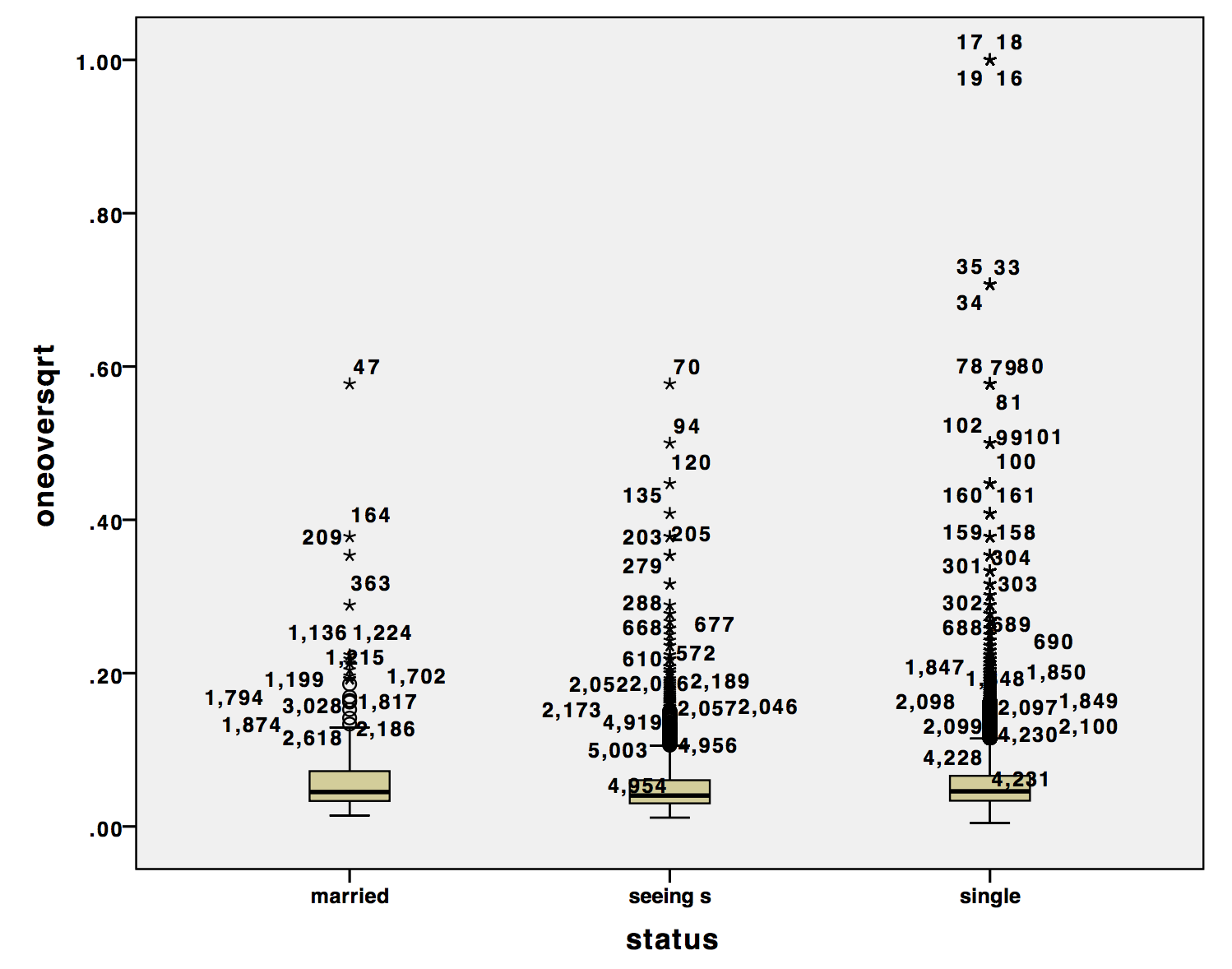
****

Figure Reciprocal of the square root transformation of essay length vs. status

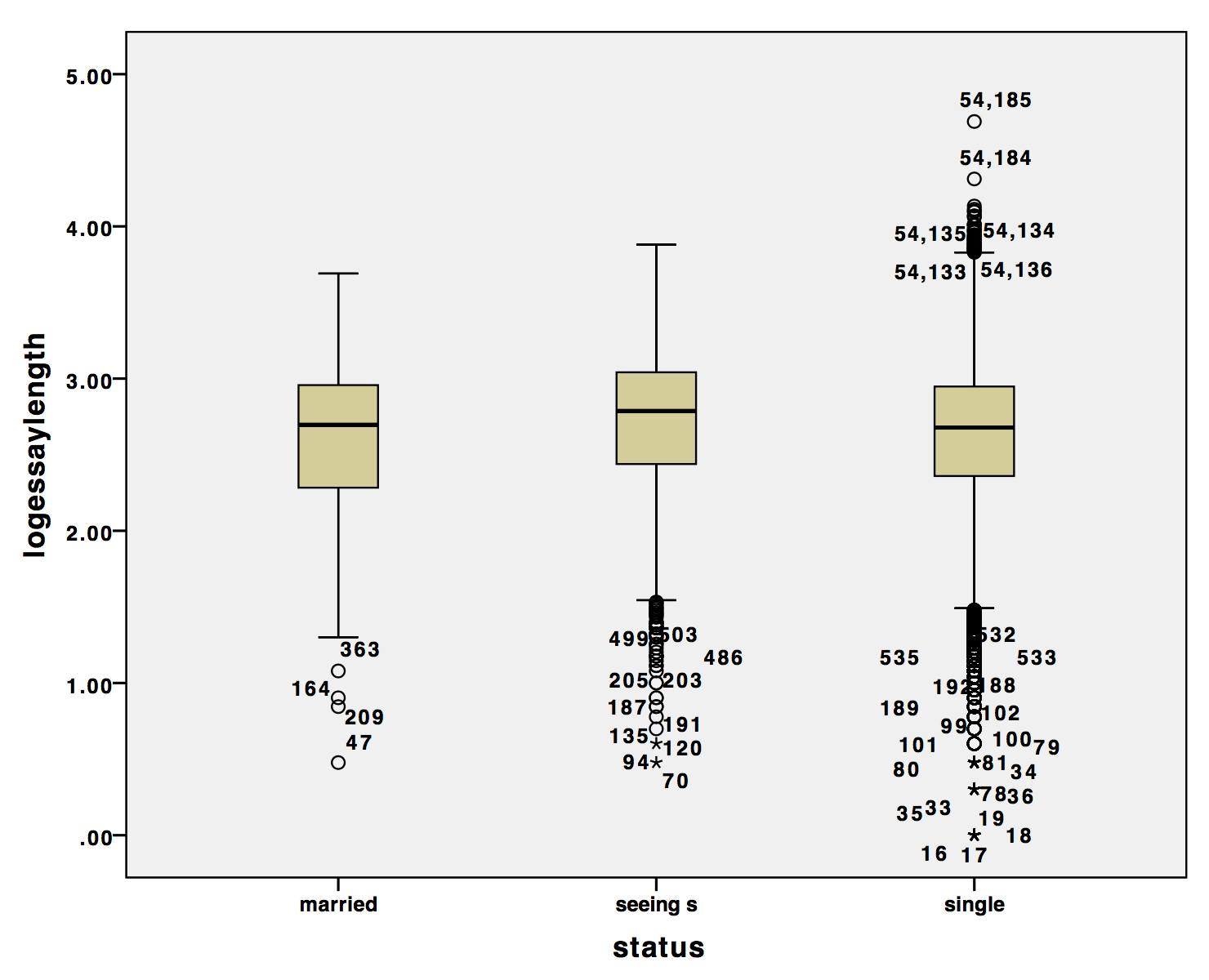
****

Figure Log (base 10) transformation of essay length vs. status

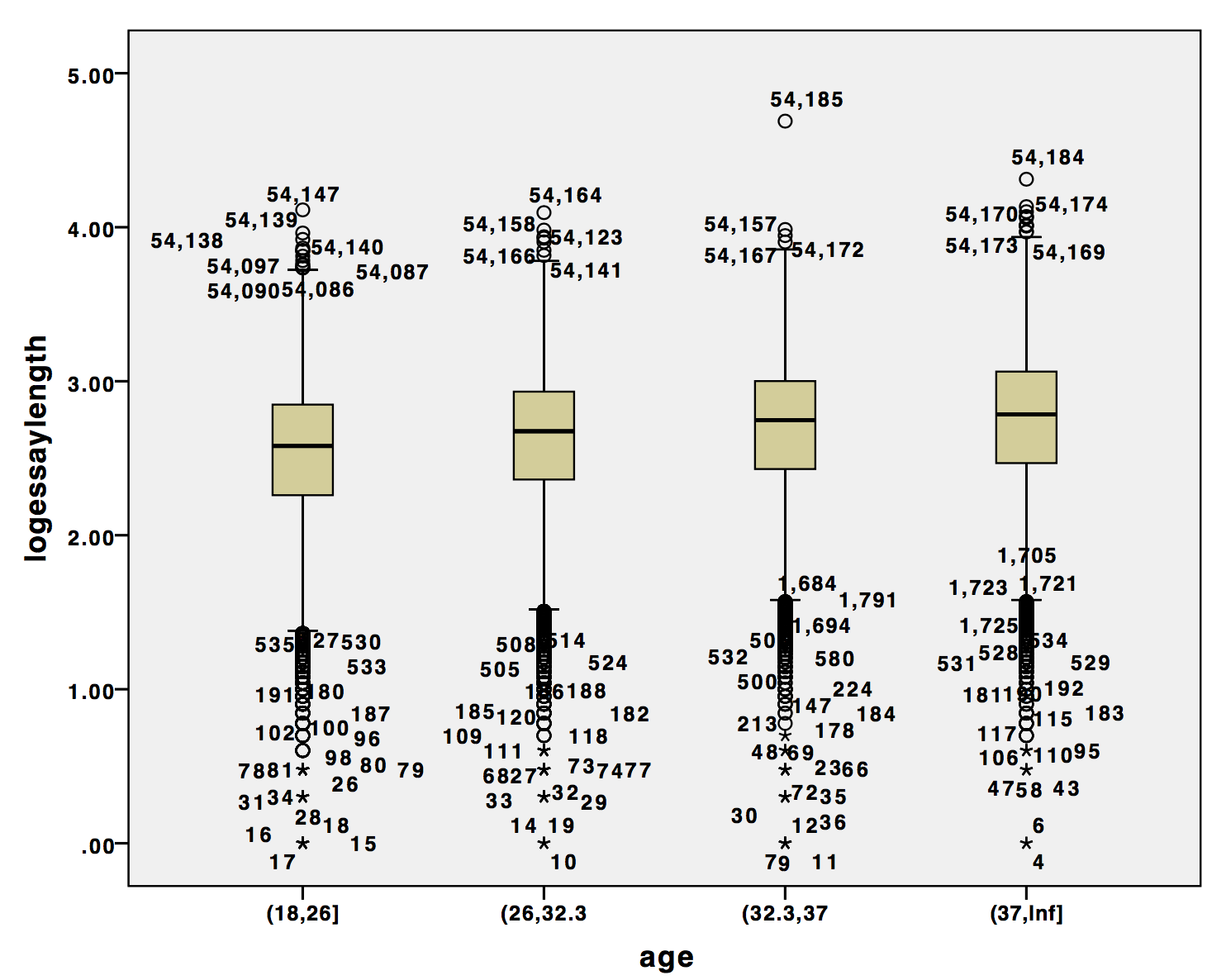
****

Figure Log essay length vs age

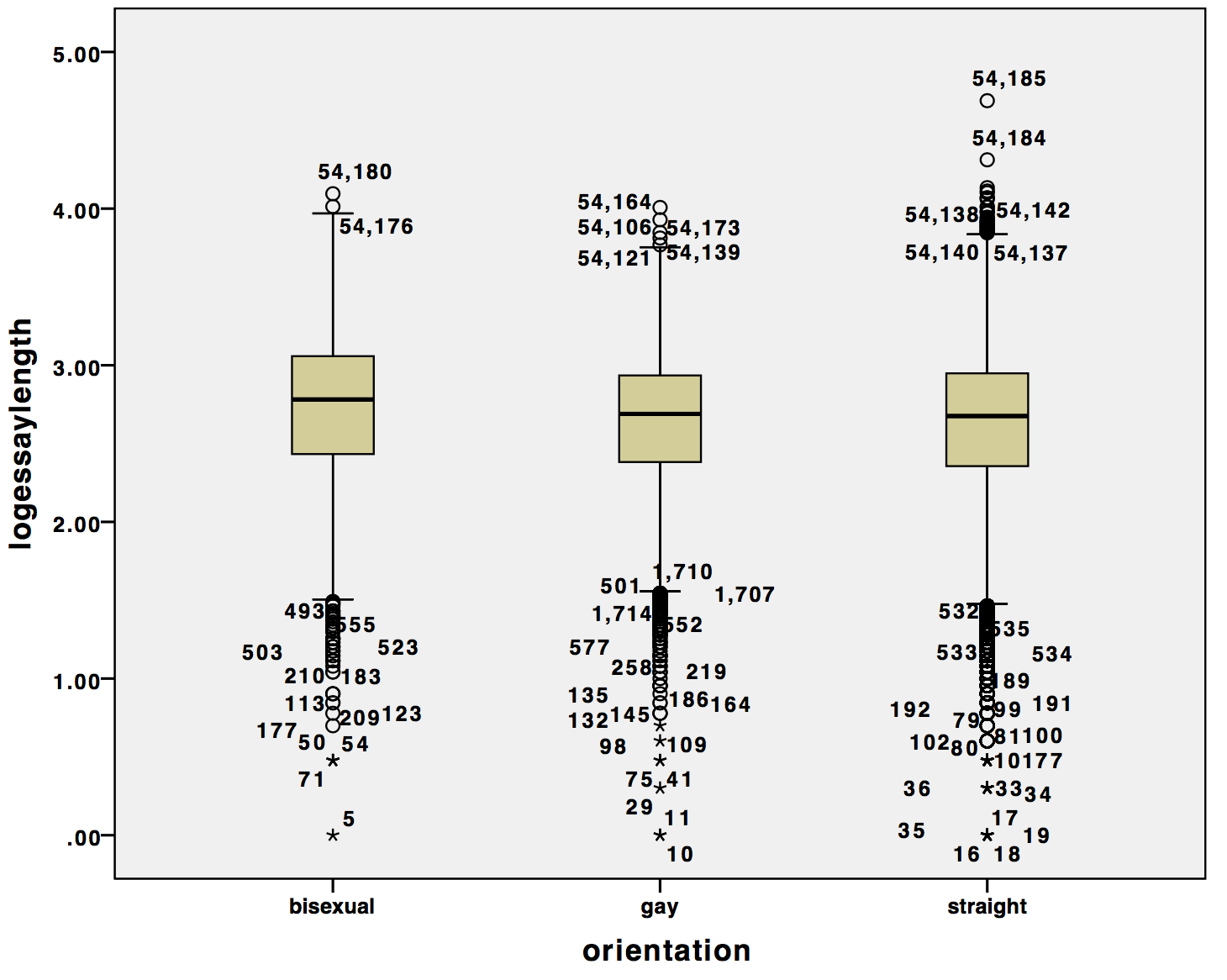
****

Figure Log essay length vs orientation

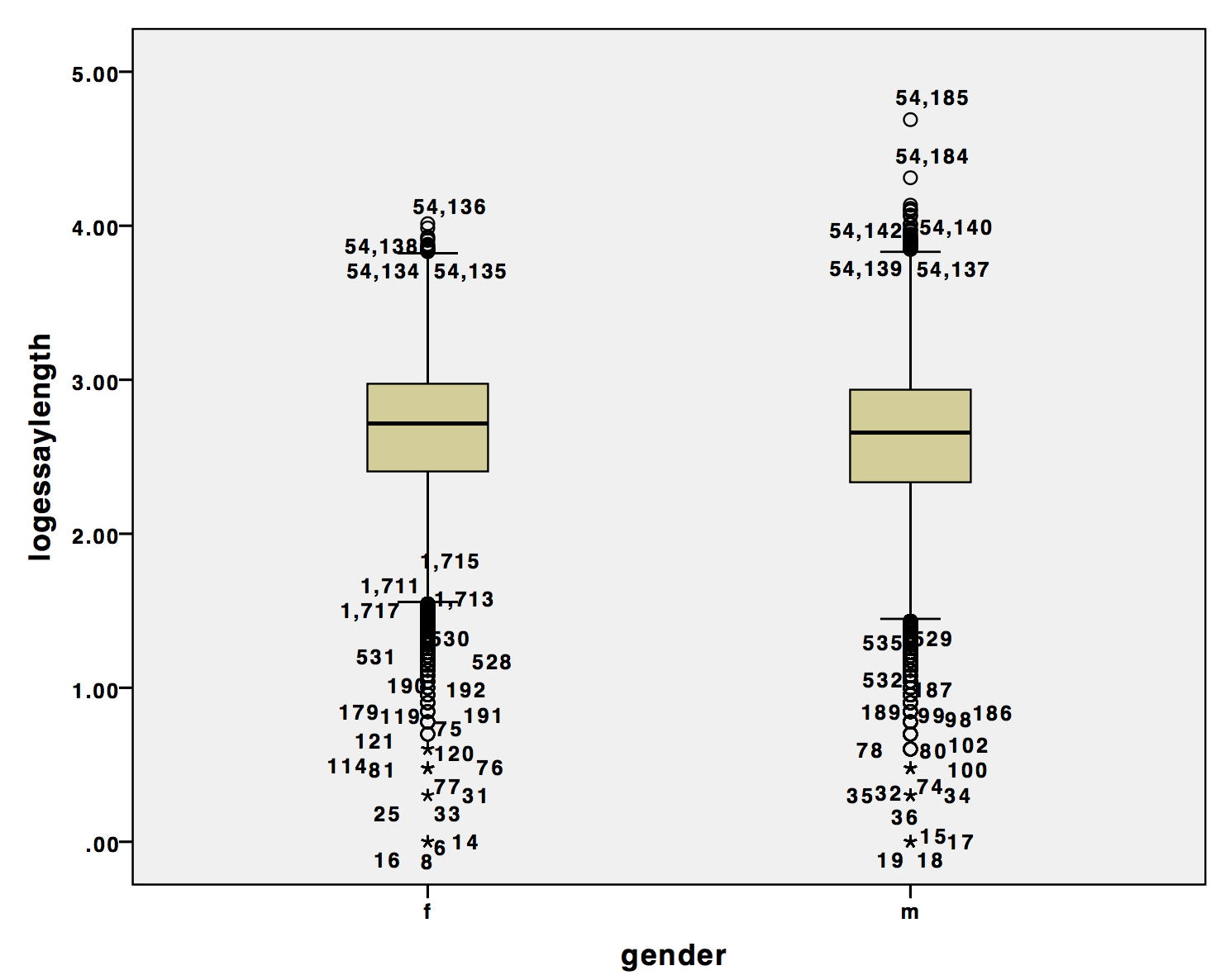
****

Figure Log essay length vs. gender

The SDmax/SDmin for the age\*gender\*orientation anova is .54146/.39867=1.358(see Table 5) and the SDmax/SDmin for the age\*gender\*status anova is .65322/.41409=1.577 (see table 6). These are both reasonably small and indicate that the assumption of homogeneity of variances is met.

Table Descriptive statistics for all groups in Age\*Gender\*Orientation Anova (values referenced highlighted)

|  |
| --- |
| **Report** |
| logessaylength |
| age | gender | orientation | Mean | N | Std. Deviation |
| (18,26] | f | bisexual | 2.6242 | 848 | .49253 |
| gay | 2.5889 | 536 | .44753 |
| straight | 2.5291 | 4980 | .47111 |
| Total | 2.5468 | 6364 | .47328 |
| m | bisexual | 2.6062 | 262 | .52441 |
| gay | 2.5763 | 1127 | .46779 |
| straight | 2.4776 | 8389 | .49433 |
| Total | 2.4924 | 9778 | .49349 |
| Total | bisexual | 2.6199 | 1110 | .50006 |
| gay | 2.5804 | 1663 | .46126 |
| straight | 2.4968 | 13369 | .48643 |
| Total | 2.5139 | 16142 | .48634 |
| (26,32.3] | f | bisexual | 2.7122 | 519 | .53038 |
| gay | 2.6604 | 456 | .44439 |
| straight | 2.6313 | 5593 | .44807 |
| Total | 2.6397 | 6568 | .45534 |
| m | bisexual | 2.7226 | 208 | .49158 |
| gay | 2.6129 | 1108 | .46298 |
| straight | 2.5856 | 9447 | .47729 |
| Total | 2.5910 | 10763 | .47651 |
| Total | bisexual | 2.7152 | 727 | .51926 |
| gay | 2.6267 | 1564 | .45801 |
| straight | 2.6026 | 15040 | .46715 |
| Total | 2.6095 | 17331 | .46918 |
| (32.3,37] | f | bisexual | 2.7945 | 230 | .47174 |
| gay | 2.7065 | 169 | .39867 |
| straight | 2.7163 | 2902 | .44577 |
| Total | 2.7212 | 3301 | .44568 |
| m | bisexual | 2.7450 | 99 | .54146 |
| gay | 2.6294 | 522 | .45142 |
| straight | 2.6483 | 4183 | .49248 |
| Total | 2.6483 | 4804 | .48939 |
| Total | bisexual | 2.7796 | 329 | .49344 |
| gay | 2.6483 | 691 | .44007 |
| straight | 2.6762 | 7085 | .47505 |
| Total | 2.6780 | 8105 | .47341 |
| (37,Inf] | f | bisexual | 2.9317 | 248 | .45481 |
| gay | 2.7240 | 315 | .43426 |
| straight | 2.7516 | 4869 | .44274 |
| Total | 2.7582 | 5432 | .44440 |
| m | bisexual | 2.8650 | 140 | .48750 |
| gay | 2.6591 | 928 | .46260 |
| straight | 2.7176 | 6107 | .49756 |
| Total | 2.7129 | 7175 | .49377 |
| Total | bisexual | 2.9076 | 388 | .46735 |
| gay | 2.6756 | 1243 | .45629 |
| straight | 2.7327 | 10976 | .47430 |
| Total | 2.7324 | 12607 | .47364 |
| Total | f | bisexual | 2.7115 | 1845 | .50664 |
| gay | 2.6533 | 1476 | .44128 |
| straight | 2.6489 | 18344 | .46087 |
| Total | 2.6546 | 21665 | .46395 |
| m | bisexual | 2.7108 | 709 | .51790 |
| gay | 2.6157 | 3685 | .46361 |
| straight | 2.5914 | 28126 | .49701 |
| Total | 2.5967 | 32520 | .49415 |
| Total | bisexual | 2.7113 | 2554 | .50969 |
| gay | 2.6264 | 5161 | .45761 |
| straight | 2.6141 | 46470 | .48388 |
| Total | 2.6198 | 54185 | .48313 |

Table Descriptive statistics for all groups in Age\*Gender\*Status Anova (values referenced highlighted)

|  |
| --- |
| **Report** |
| logessaylength |
| age | gender | status | Mean | N | Std. Deviation |
| (18,26] | f | married | 2.5442 | 20 | .65322 |
| seeing someone | 2.6366 | 425 | .50831 |
| single | 2.5404 | 5919 | .46940 |
| Total | 2.5468 | 6364 | .47328 |
| m | married | 2.6703 | 11 | .49400 |
| seeing someone | 2.6072 | 289 | .52185 |
| single | 2.4887 | 9478 | .49220 |
| Total | 2.4924 | 9778 | .49349 |
| Total | married | 2.5890 | 31 | .59612 |
| seeing someone | 2.6247 | 714 | .51367 |
| single | 2.5086 | 15397 | .48420 |
| Total | 2.5139 | 16142 | .48634 |
| (26,32.3] | f | married | 2.6264 | 52 | .60158 |
| seeing someone | 2.7494 | 349 | .48449 |
| single | 2.6336 | 6167 | .45151 |
| Total | 2.6397 | 6568 | .45534 |
| m | married | 2.4683 | 36 | .46787 |
| seeing someone | 2.7343 | 405 | .42294 |
| single | 2.5858 | 10322 | .47763 |
| Total | 2.5910 | 10763 | .47651 |
| Total | married | 2.5617 | 88 | .55346 |
| seeing someone | 2.7413 | 754 | .45223 |
| single | 2.6037 | 16489 | .46859 |
| Total | 2.6095 | 17331 | .46918 |
| (32.3,37] | f | married | 2.8108 | 26 | .41409 |
| seeing someone | 2.7781 | 104 | .44085 |
| single | 2.7186 | 3171 | .44601 |
| Total | 2.7212 | 3301 | .44568 |
| m | married | 2.5015 | 28 | .51812 |
| seeing someone | 2.7494 | 156 | .49002 |
| single | 2.6457 | 4620 | .48881 |
| Total | 2.6483 | 4804 | .48939 |
| Total | married | 2.6504 | 54 | .49190 |
| seeing someone | 2.7609 | 260 | .47030 |
| single | 2.6754 | 7791 | .47319 |
| Total | 2.6780 | 8105 | .47341 |
| (37,Inf] | f | married | 2.6585 | 30 | .48694 |
| seeing someone | 2.9166 | 68 | .46883 |
| single | 2.7567 | 5334 | .44350 |
| Total | 2.7582 | 5432 | .44440 |
| m | married | 2.5723 | 71 | .56418 |
| seeing someone | 2.8050 | 140 | .55293 |
| single | 2.7125 | 6964 | .49145 |
| Total | 2.7129 | 7175 | .49377 |
| Total | married | 2.5979 | 101 | .54142 |
| seeing someone | 2.8415 | 208 | .52839 |
| single | 2.7317 | 12298 | .47174 |
| Total | 2.7324 | 12607 | .47364 |
| Total | f | married | 2.6585 | 128 | .55163 |
| seeing someone | 2.7139 | 946 | .49568 |
| single | 2.6518 | 20591 | .46169 |
| Total | 2.6546 | 21665 | .46395 |
| m | married | 2.5405 | 146 | .52600 |
| seeing someone | 2.7096 | 990 | .48768 |
| single | 2.5934 | 31384 | .49378 |
| Total | 2.5967 | 32520 | .49415 |
| Total | married | 2.5956 | 274 | .54036 |
| seeing someone | 2.7117 | 1936 | .49148 |
| single | 2.6165 | 51975 | .48216 |
| Total | 2.6198 | 54185 | .48313 |

**Analysis of data:**

**Age\*Gender\*Orientation:**

The formal analysis of variance of the first model (age\*gender\*orientation) showed that all three factors, age, gender and orientation were statistically significant at the .05 level. The interaction of age and orientation was also significant at this level (see table 7). From our informal analysis we expected significance across all three factors. Additionally post-hoc tests show that there is a significant difference between all four age groups, as we predicted. Additionally for the orientation factor these was a significant difference between gay people and bisexual as well as straight people and bisexual people but there was not a difference between gay and straight people (see tables 8 and 9). This result is consistent with the trends noted in the informal analysis. The gender factor is also significant indicating there is a significant difference between the amount men and women write on their online dating profiles. Additionally from these three factors we know that women on average write more characters then men, bisexual people write more then straight and gay people on average and that older people tend to write more as well. However from the interaction of age and orientation we can see that for the two older groups the amount that straight people increased enough to surpass what older gay people wrote. In other words older straight people write significantly more then older gay people (see fig 12 and 13).

Table Anova table for Age\*Gender\*Orientation

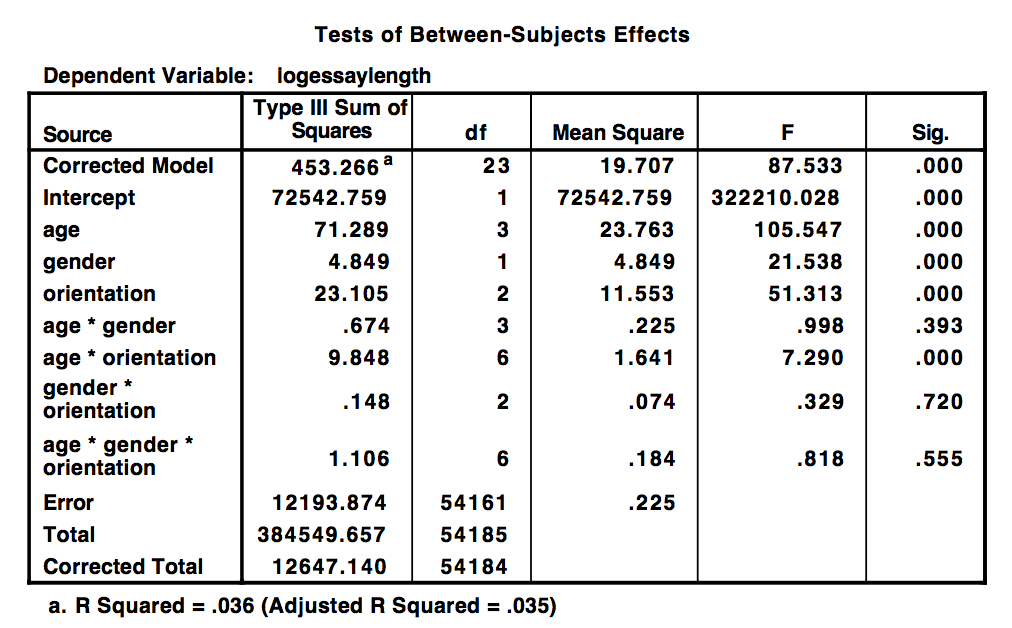


Table Post Hoc tests for Age

|  |
| --- |
| **Multiple Comparisons** |
| Dependent Variable:   logessaylength |
|  | (I) age | (J) age | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval |
|  | Lower Bound | Upper Bound |
| Scheffe | (18,26] | (26,32.3] | -.0956\* | .00519 | .000 | -.1101 | -.0811 |
| (32.3,37] | -.1641\* | .00646 | .000 | -.1822 | -.1461 |
| (37,Inf] | -.2186\* | .00564 | .000 | -.2343 | -.2028 |
| (26,32.3] | (18,26] | .0956\* | .00519 | .000 | .0811 | .1101 |
| (32.3,37] | -.0685\* | .00639 | .000 | -.0864 | -.0507 |
| (37,Inf] | -.1230\* | .00555 | .000 | -.1385 | -.1074 |
| (32.3,37] | (18,26] | .1641\* | .00646 | .000 | .1461 | .1822 |
| (26,32.3] | .0685\* | .00639 | .000 | .0507 | .0864 |
| (37,Inf] | -.0544\* | .00676 | .000 | -.0733 | -.0356 |
| (37,Inf] | (18,26] | .2186\* | .00564 | .000 | .2028 | .2343 |
| (26,32.3] | .1230\* | .00555 | .000 | .1074 | .1385 |
| (32.3,37] | .0544\* | .00676 | .000 | .0356 | .0733 |
| Bonferroni | (18,26] | (26,32.3] | -.0956\* | .00519 | .000 | -.1093 | -.0819 |
| (32.3,37] | -.1641\* | .00646 | .000 | -.1812 | -.1471 |
| (37,Inf] | -.2186\* | .00564 | .000 | -.2334 | -.2037 |
| (26,32.3] | (18,26] | .0956\* | .00519 | .000 | .0819 | .1093 |
| (32.3,37] | -.0685\* | .00639 | .000 | -.0854 | -.0517 |
| (37,Inf] | -.1230\* | .00555 | .000 | -.1376 | -.1083 |
| (32.3,37] | (18,26] | .1641\* | .00646 | .000 | .1471 | .1812 |
| (26,32.3] | .0685\* | .00639 | .000 | .0517 | .0854 |
| (37,Inf] | -.0544\* | .00676 | .000 | -.0723 | -.0366 |
| (37,Inf] | (18,26] | .2186\* | .00564 | .000 | .2037 | .2334 |
| (26,32.3] | .1230\* | .00555 | .000 | .1083 | .1376 |
| (32.3,37] | .0544\* | .00676 | .000 | .0366 | .0723 |
| Based on observed means.   The error term is Mean Square(Error) = .225. |  |  |  |  |  |  |  |
| \*. The mean difference is significant at the 0 |  |  |  |  |  |  |  |

Table Post Hoc tests for Orientation

|  |
| --- |
| **Multiple Comparisons** |
| Dependent Variable:   logessaylength |
|  | (I) orientation | (J) orientation | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval |
|  | Lower Bound | Upper Bound |
| Scheffe | bisexual | gay | .0849\* | .01148 | .000 | .0568 | .1130 |
| straight | .0972\* | .00964 | .000 | .0736 | .1208 |
| gay | bisexual | -.0849\* | .01148 | .000 | -.1130 | -.0568 |
| straight | .0124 | .00696 | .207 | -.0047 | .0294 |
| straight | bisexual | -.0972\* | .00964 | .000 | -.1208 | -.0736 |
| gay | -.0124 | .00696 | .207 | -.0294 | .0047 |
| Bonferroni | bisexual | gay | .0849\* | .01148 | .000 | .0574 | .1124 |
| straight | .0972\* | .00964 | .000 | .0742 | .1203 |
| gay | bisexual | -.0849\* | .01148 | .000 | -.1124 | -.0574 |
| straight | .0124 | .00696 | .228 | -.0043 | .0290 |
| straight | bisexual | -.0972\* | .00964 | .000 | -.1203 | -.0742 |
| gay | -.0124 | .00696 | .228 | -.0290 | .0043 |
| Based on observed means.   The error term is Mean Square(Error) = .225. |  |  |  |  |  |  |  |
| \*. The mean difference is significant at the 0 |  |  |  |  |  |  |  |

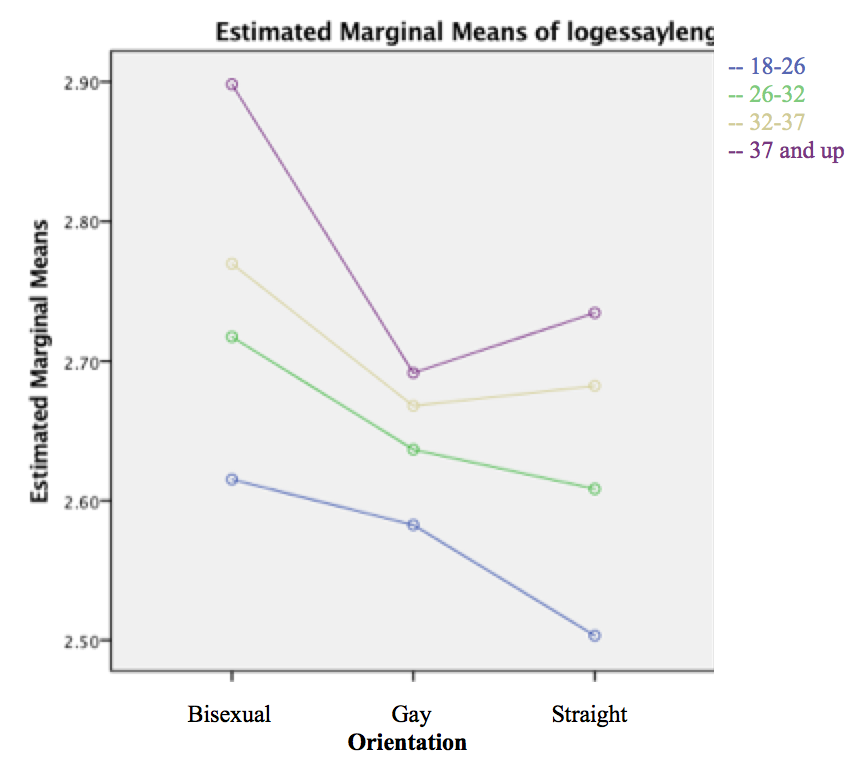


Figure Interaction graph for orientation and age

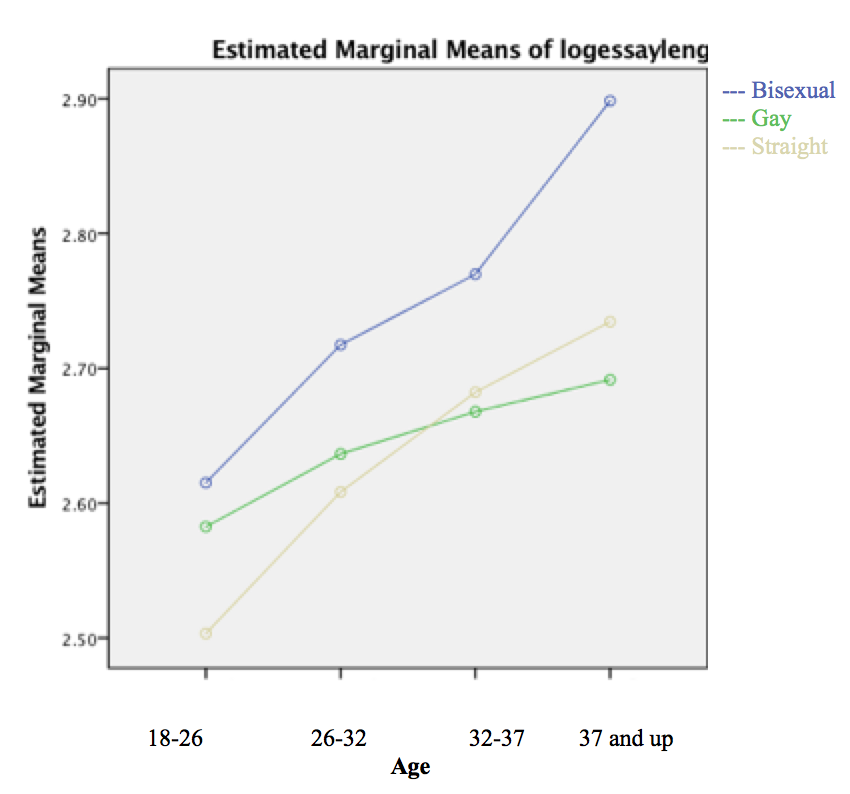


Figure Interaction graph for age and orientation

As you can see from the histogram of the residuals (see fig 14) the residuals follow a normal curve. It is slightly right skewed and there are a number of values that are beyond 3 standard deviations away from 0 however with a dataset this large it is expected that there would be some values that could fall beyond this point. If we expect 99.73% of the values to fall within three standard deviations we can expect around 146 values to fall outside (54,185-.9973\*54185=146.2). We also expected some level of skew-ness since the transformation did not full eliminate the spread problem since in the data. The P-P plot (see fig 15) also shows a slight lack of normality however given the size of the data set and the central limit theorem, which states that the larger the n the closer to normality, I am confident that the data does meet the assumption of normality. Finally we can see from the residuals vs. fit plot that the residuals are fairly evenly spread across predictions, this also supports that the assumption of homogeneity of variance is met (see fig 16).

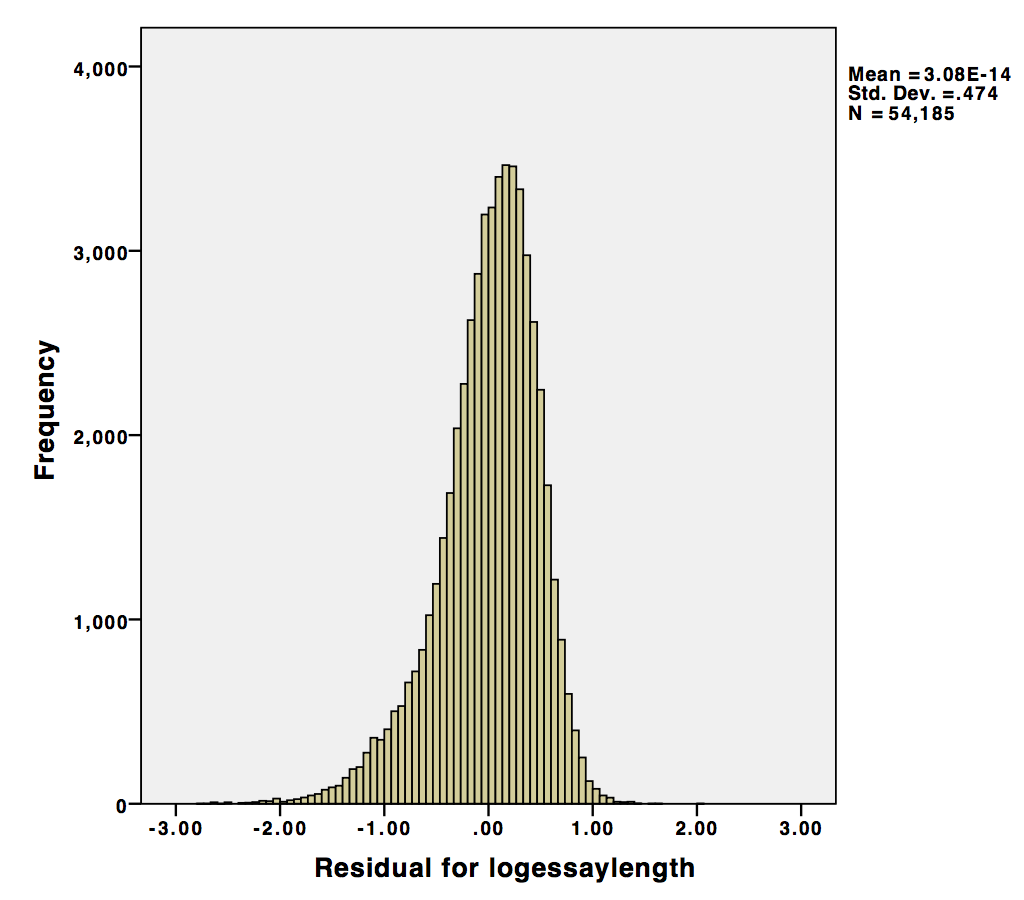


Figure Histogram of the residuals for age\*gender\*orientation Anvova

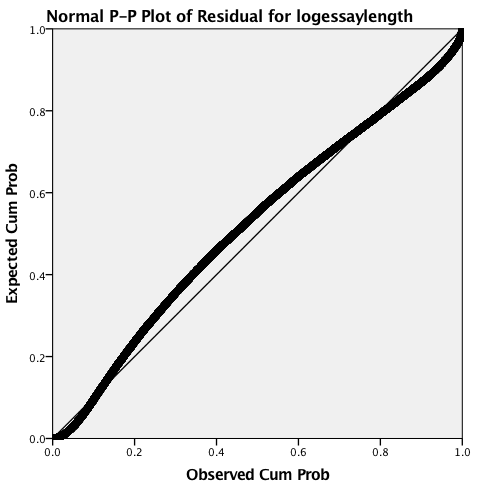


Figure P-P plot for age\*gender\*orientation

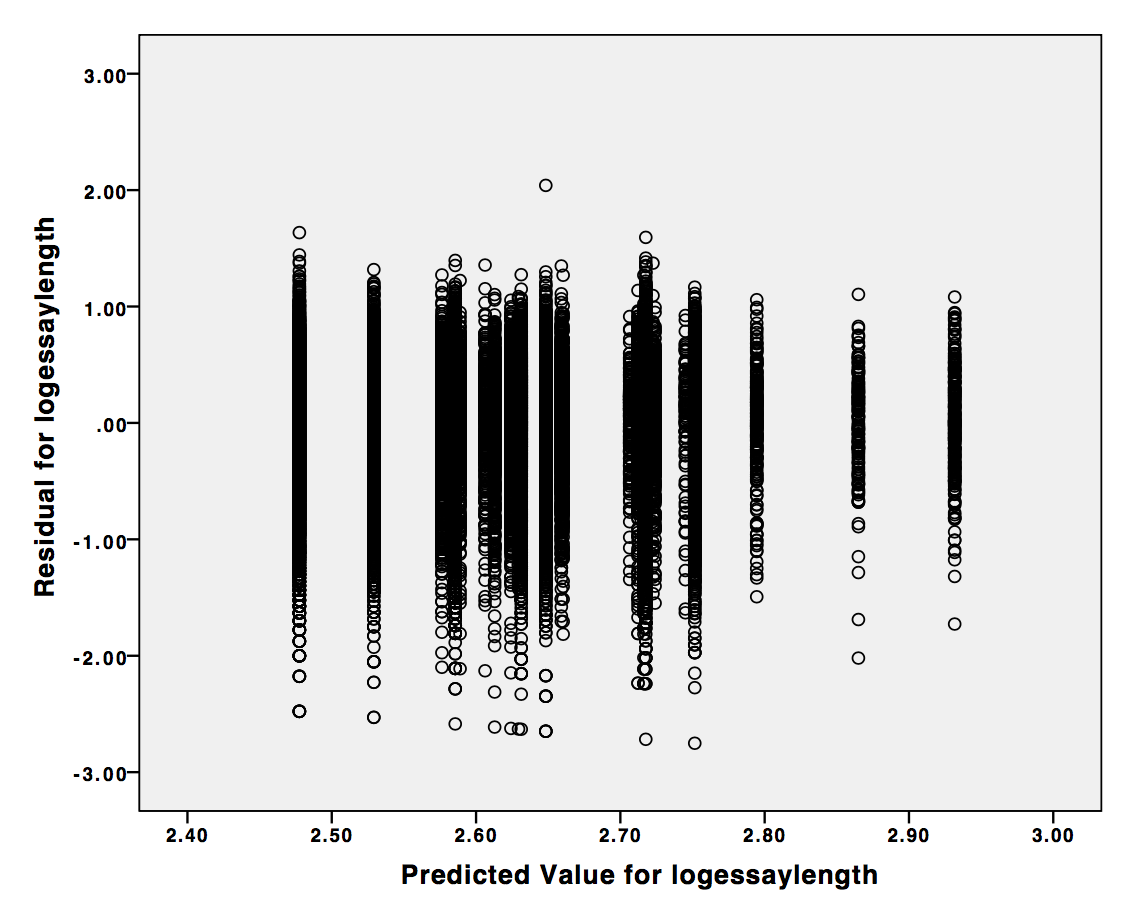


Figure Predicted vs Residuals for age\*gender\*orientation

**Age\*Gender\*Status:**

From the second three-way Anova we once again see that age and gender are significant factors at the .05 level. Additionally relationship status is also significant at this level and none of the interactions between status and, age and gender were significant (see table 10). From the post hoc tests we know that there is not a significant difference between single and married people but there is a difference between people who are seeing someone and people who are married and people who are seeing someone and people who are single (see table 11). Once again the residuals are not perfectly normally however given the size of the data set they are not overly unexpected or unusual (see fig 17 and 18). The residual vs. fit plot does not indicate that homogeneity of variance has been violated since the values of the residuals are fairly evenly distributed across the predicted values (see fig 19).

Table Anova table for age\*gender\*status

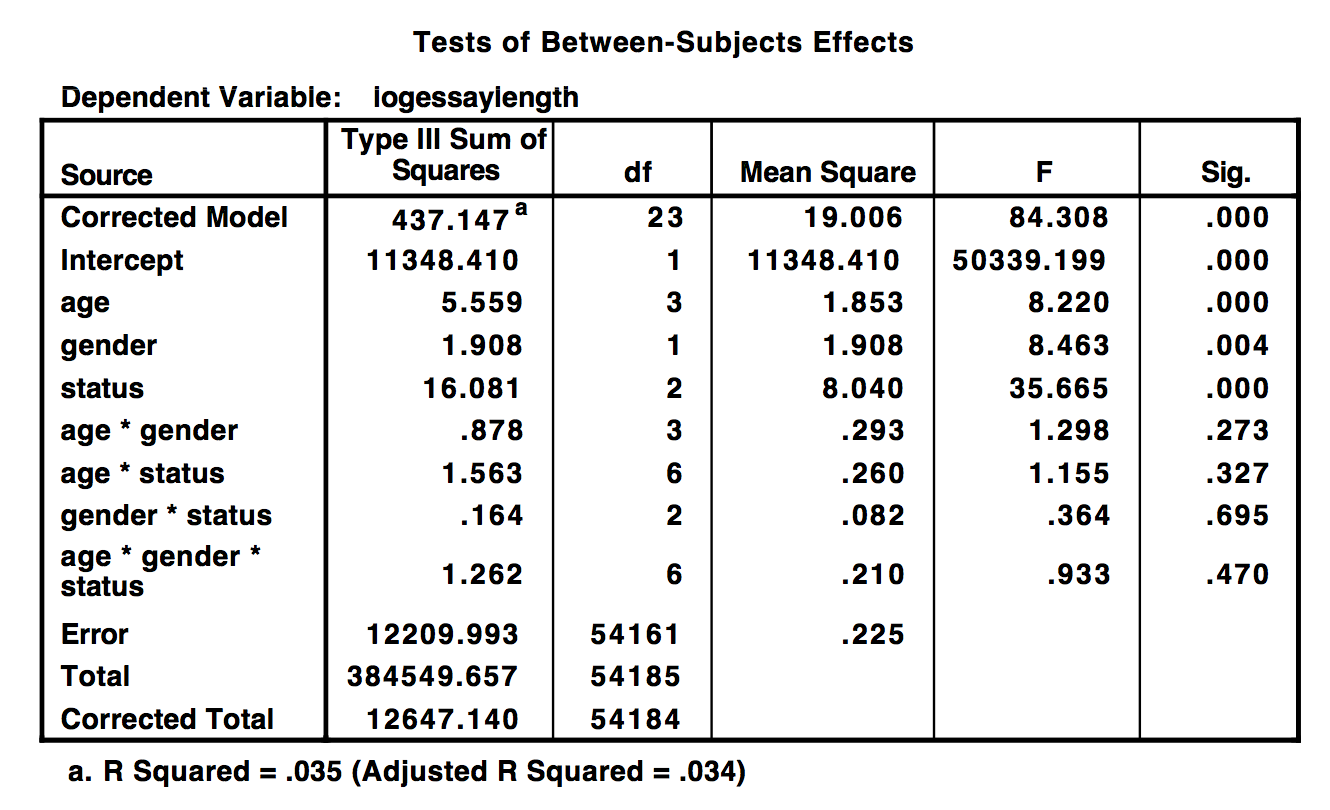
****

Table Post Hoc tests for status

|  |
| --- |
| **Multiple Comparisons** |
| Dependent Variable:   logessaylength |
|  | (I) status | (J) status | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval |
|  | Lower Bound | Upper Bound |
| Scheffe | married | seeing someone | -.1161\* | .03065 | .001 | -.1911 | -.0411 |
| single | -.0209 | .02876 | .767 | -.0913 | .0495 |
| seeing someone | married | .1161\* | .03065 | .001 | .0411 | .1911 |
| single | .0951\* | .01099 | .000 | .0682 | .1221 |
| single | married | .0209 | .02876 | .767 | -.0495 | .0913 |
| seeing someone | -.0951\* | .01099 | .000 | -.1221 | -.0682 |
| Bonferroni | married | seeing someone | -.1161\* | .03065 | .000 | -.1894 | -.0427 |
| single | -.0209 | .02876 | 1.000 | -.0898 | .0479 |
| seeing someone | married | .1161\* | .03065 | .000 | .0427 | .1894 |
| single | .0951\* | .01099 | .000 | .0688 | .1215 |
| single | married | .0209 | .02876 | 1.000 | -.0479 | .0898 |
| seeing someone | -.0951\* | .01099 | .000 | -.1215 | -.0688 |
| Based on observed means.   The error term is Mean Square(Error) = .225. |  |  |  |  |  |  |  |
| \*. The mean difference is significant at the 0 |  |  |  |  |  |  |  |

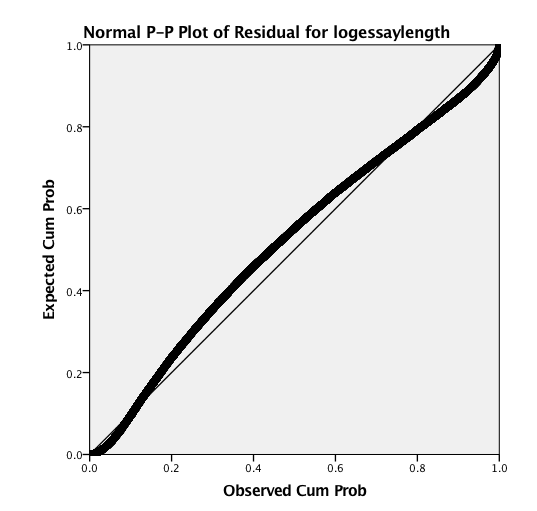
****

Figure P-P plot for age\*gender\*status

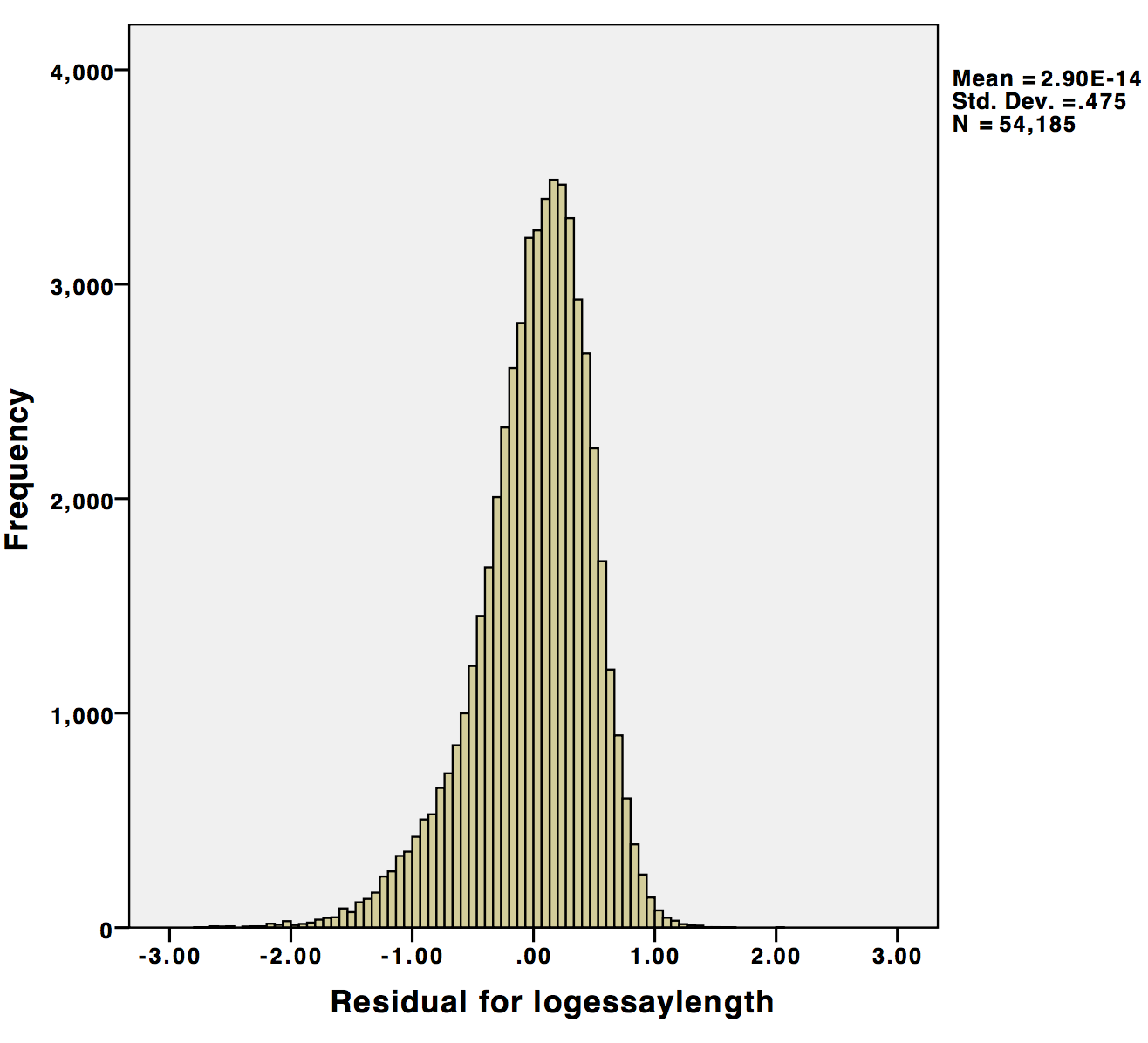
****

Figure Histogram of the residuals for age\*gender\*status

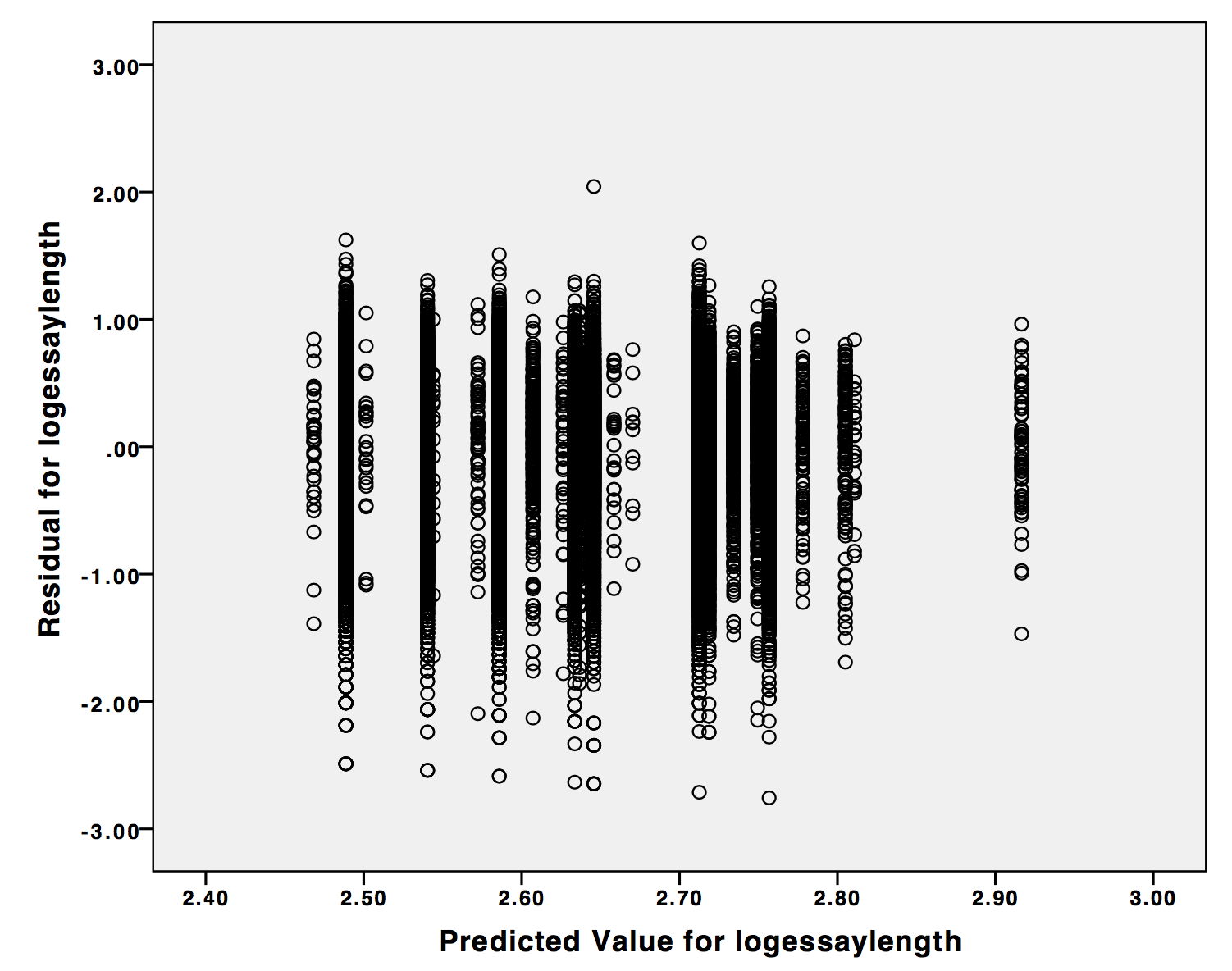
****

Figure Predicted vs. Residual for age\*gender\*status

**Discussion:**

From this analysis it is clear there is a connection between age, gender, orientation, status, age\*orientation, and the amount of characters someone writes in their personal statement on their OKcupid dating profile. There are many reasons these connections may exist perhaps people whose status is ‘seeing someone’ feel the need to explain there situation more then someone who is single or married and that is why they write more. Equally likely however is a confounding factor such as the personality type of people who mark their status as seeing someone. The same rational could be used for those who identify as bisexual, perhaps they feel the need to explain their situation more then straight or gay people. Or perhaps people who are bisexual are more open about their personally life and that is why they write more. Because this data set is observation and therefore more susceptible to confounding factors, such as the examples above we cannot draw any firm conclusions about what the connections between the factors and the response mean. Additionally the data it is also self reported and therefore less reliable. In one survey 53% of people admitted to lying on their online dating profile. The factor most likely to be affected by dishonesty is age since people, particularly women, are more most likely to lie about their age. This dataset also only included people within 25 miles of San Francisco and therefore cannot be generalized beyond the scope of the San Francisco area. All that said it does show interesting trends in the behavior of the different types of people using online dating. In the future it would be really interesting to create an experiment where people fill out a fake dating profile so that we could gather more concrete data on how different factors affect how much and what people write on their profiles.

**References:**

* R statistical package
* SPSS statistical package
* <https://github.com/rudeboybert/JSE_OkCupid/blob/master/okcupid_codebook.txt>
* <http://www.eharmony.com/online-dating-statistics/>
* <https://strainindex.wordpress.com/2008/07/28/the-average-sentence-length/>
* textbook