Assignment: Homework 5 Report

Name: Horace Fung

Instructor: Professor Simonoff

Course: Regression and Multivariate Data Analysis

Modelling the number of days dogs stay at a shelter before adoption

This report examines the relationship between the number of months a dog stays at the Austin Animal Center before adoption and their breed and age. Understanding factors that affect the length of stay is particularly important to the Austin Animal Center because it is a no-kill shelter— its resources are spread thin when certain dogs stay for too long while new dogs arrive. Moreover, the center can promote dogs at risk of staying for too long to potential adopters.

The data contains all adoptions for 2019 and is from the City of Austin Data Portal and the Austin Animal Center. The total dataset has 1,636 observations but 492 observations (30% of the data) were randomly held out for validation. The original dataset contained over 100 breeds; based on classification data from the American Kennel Club, these breeds were grouped into 7 groups— Hounding, Toy, Working, Terrier, Herding, Sporting and Non-sporting. There are three caveats with the dataset. First, some dogs were adopted and returned several times. In this analysis, the failed adoptions were ignored, and the time spent at the center is simply from the first arrival date to the most reception adoption date. Second, only successful adoptions were included in the dataset, which means information on dogs that are struggling to attract any adoptions at all may not be reflected in the model. Lastly, the breed categorical variable only contains six groups because none of the observations belong to the non-sporting breed.

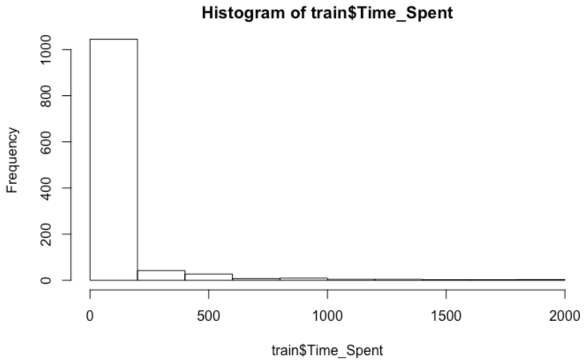
1. Definition of Variables:

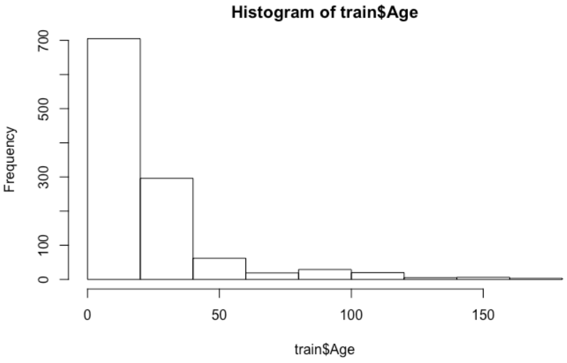
The initial data contains two variables—a numerical variable Age and a categorical variable Group.

1. *Response Variable Time_Spent:* The number of days between a dog's arrival date to the center and its most recent adoption date.
2. *Age:* A numerical variable for age of the dog in months.
3. *Group*: A categorical variable for which group the dog belongs to (Hounding, Toy, Working, Terrier, Herding and Sporting). In the effect coding, Sporting was left out (indicated by -1 for the effect coding).

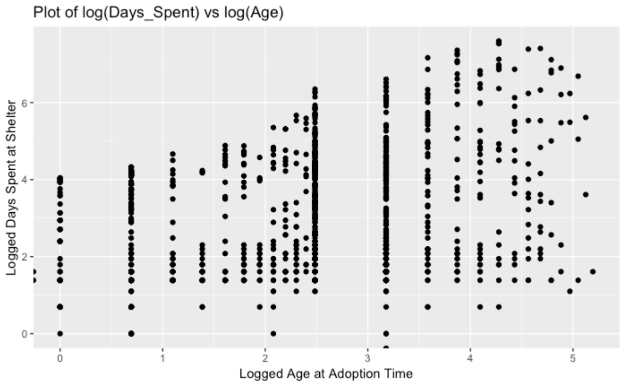
.

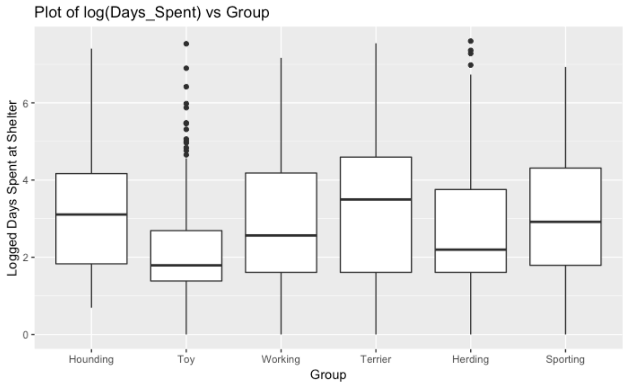
2. Initial Plots and log-Transformations:

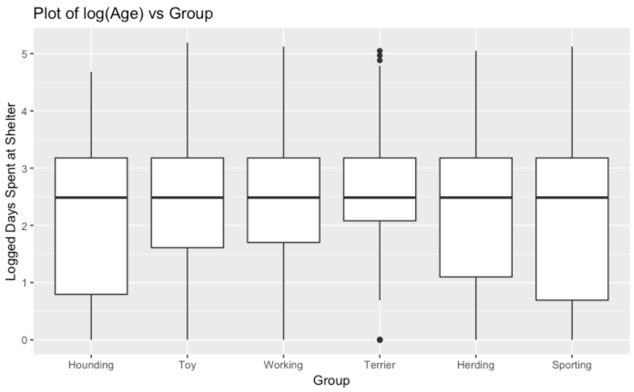




Histograms for both Time_Spent and Age reveal severe long right tails. The logs of both variables were taken with base 10. The log-log relationship between Time_Spent and Age means a 1% change in Age is associated with β% change in Time_Spent. This seems to make intuitive sense; risks such as health complications increase rapidly with a dog's age and may have a multiplicative effect on how long they stay at the shelter since potential adopters may be less willing to choose older dogs or suitable adopters are harder to find. The scatterplots and boxplots of the variables are shown below:







There seems to be a slightly positive relationship between log(Time_Spent) and log(Age) but the relationship is weak. In addition, there is still non-constant variance despite the log-transformations. For the Groups variable, it does seem like log(Time_Spent) varies between some of the groups and could be informative. Furthermore, there is clear non-constant variance across the groups, especially the Toy breeds, which suggests weighted least squares may be appropriate. Finally, as shown below, the number of observations in each group is sufficently balanced:

```
Hounding       Toy  Working  Terrier  Herding Sporting
     102       207      123      265      250      198
```

3. Constant Shift Model and Unusual Observations:

The first type of model is a simpler constant shift model. As the name suggests, the differences in log(Time_Spent) for different dog Groups is explained by a shift in the line, but the slope that represents the relationship between log(Time_Spent) and log(Age) is unchanged. The regression outputs for a constant shift model is shown below:

```
Anova Table (Type III tests)

Response: log_Time_Spent
            Sum Sq   Df F value    Pr(>F)
(Intercept) 165.28    1 425.060 < 2.2e-16 ***
log_Age      61.58    1 158.369 < 2.2e-16 ***
Group        37.91    5  19.497 < 2.2e-16 ***
Residuals   442.51 1138

lm(formula = log_Time_Spent ~ log_Age + Group, data = train)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.83788    0.04064  20.617  < 2e-16 ***
log_Age      0.43819    0.03482  12.584  < 2e-16 ***
Hounding     0.15306    0.05415   2.826 0.004790 **
```

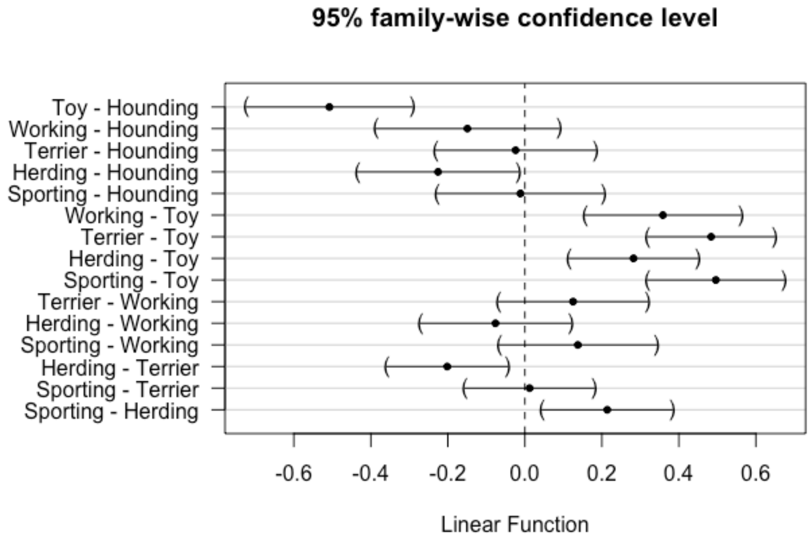
```
Toy         -0.35473    0.04053  -8.751  < 2e-16 ***
Working      0.00372    0.04992   0.075 0.940617
Terrier      0.12907    0.03709   3.480 0.000521 ***
Herding     -0.07254    0.03771  -1.924 0.054658 .

Residual standard error: 0.6236 on 1138 degrees of freedom
Multiple R-squared:  0.1792,    Adjusted R-squared:  0.1749
F-statistic:  41.4 on 6 and 1138 DF,  p-value: < 2.2e-16
```

```
Tables of means
Grand mean

1.279174

 Group
    Hounding      Toy Working Terrier Herding Sporting
       1.443   0.9369   1.294   1.421   1.218    1.431
rep  102.000 207.0000 123.000 265.000 250.000  198.000

Ordinary Unadjusted Means
Group
 Hounding       Toy   Working   Terrier   Herding  Sporting
1.4542173 0.9485287 1.3060770 1.4327883 1.2293414 1.4424048
```

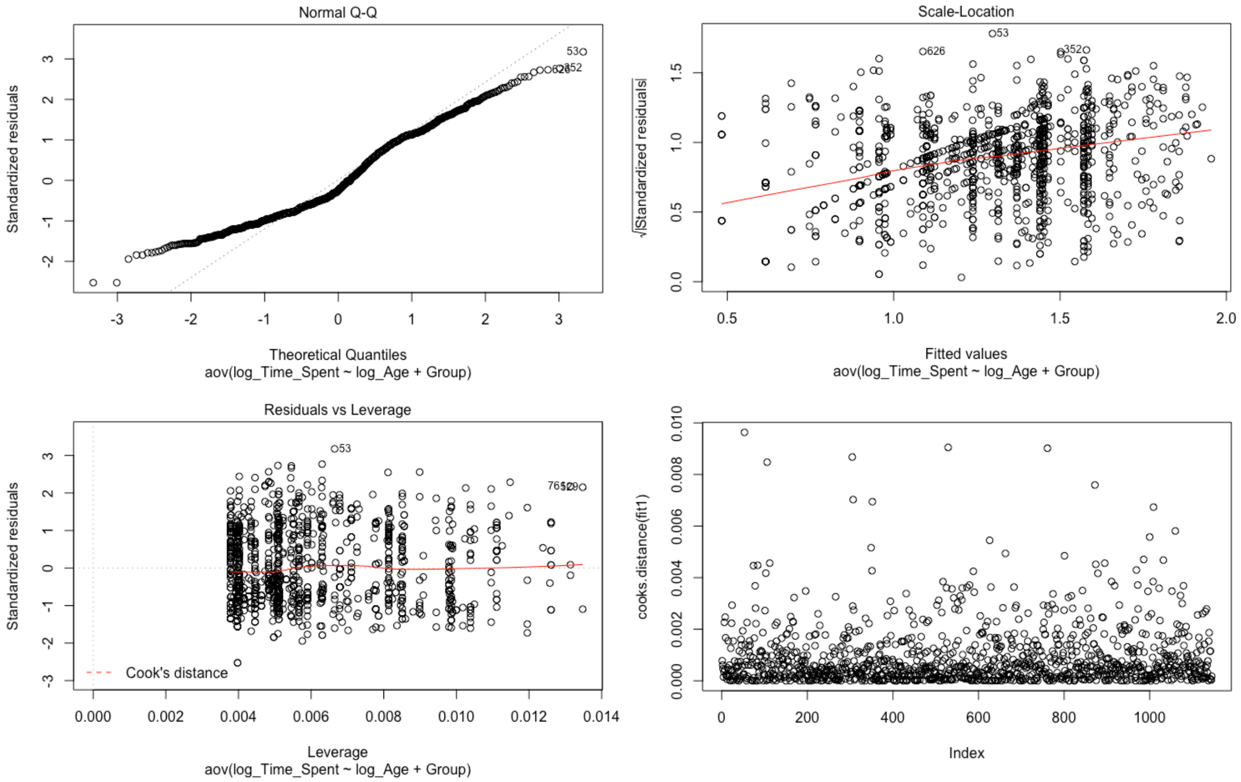
In the ANOVA table, it is clear that both log(Age) and Group are significant predictors of log(Time_Spent). The coefficient for log(Age) is 0.43819, which means a 1% in increase in a dog’s age in months is associated with a 0.43819% increase in the number of days the dog spends at the shelter, given the Group is fixed. The table of means show the estimated expected log(Time_Spent) for each Group given a log(Age) of 1.279174 (~19 months). Toy breeds spend the shortest time at the shelter on average while Hounding, Terrier and Sporting breeds spend longer periods on average. Working and Herding breeds are in between, but tilts closer to longer expected log(Time_Spent). The ordinary unadjusted means tell a similar story. The adjusted R^2 is 17.49%, which is quite low.

A Tukey comparison can be performed for the constant shift model to check which groups are significantly different. The output and confidence intervals plot are displayed below:

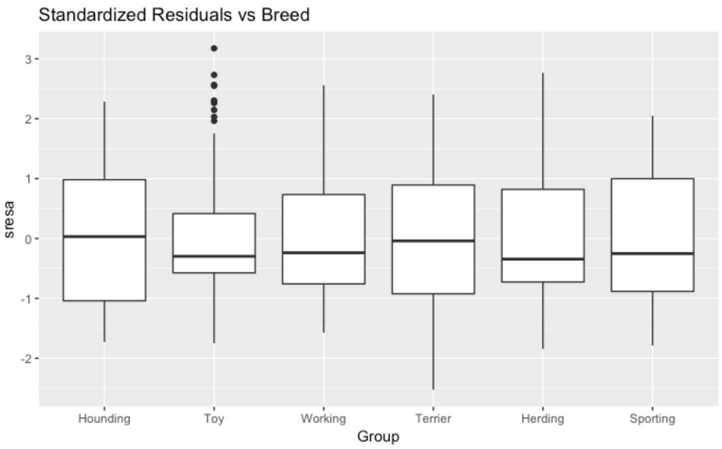
```
Multiple Comparisons of Means: Tukey Contrasts

Linear Hypotheses:        Estimate   Std. Error t value Pr(>|t|)
Toy - Hounding == 0       -0.50778    0.07561  -6.716  < 0.001 ***
Working - Hounding == 0  -0.14934    0.08356  -1.787  0.46898
Terrier - Hounding == 0  -0.02399    0.07293  -0.329  0.99948
Herding - Hounding == 0  -0.22560    0.07329  -3.078  0.02536 *
Sporting - Hounding == 0 -0.01164    0.07600  -0.153  0.99999
Working - Toy == 0         0.35845    0.07103   5.047  < 0.001 ***
Terrier - Toy == 0         0.48379    0.05785   8.362  < 0.001 ***
Herding - Toy == 0         0.28219    0.05869   4.808  < 0.001 ***
Sporting - Toy == 0        0.49614    0.06223   7.973  < 0.001 ***
Terrier - Working == 0     0.12535    0.06812   1.840  0.43468
Herding - Working == 0    -0.07626    0.06869  -1.110  0.87481
Sporting - Working == 0    0.13770    0.07167   1.921  0.38418
Herding - Terrier == 0    -0.20160    0.05516  -3.655  0.00353 **
Sporting - Terrier == 0    0.01235    0.05895   0.210  0.99994
Sporting - Herding == 0    0.21396    0.05936   3.604  0.00424 **
```



There are several pairings that significantly different. The pairs with very significant p-values and confidence intervals far from zero are Toy-Hounding, Working-Toy, Terrier-Toy, Herding-Toy and Sporting-Toy. While there are other pairs that are also statistically different such as Herding-Terrier, the Tukey comparisons test strongly suggests that Toy breeds are quite different from the rest in terms of how long they stay at the shelter. The diagnostic plots of the constant shift model are shown below:



There is non-normality in the normality plot, which suggests a linear model may not fully capture the relationship between these predictors and log(Time_Spent). In the residuals vs fitted values plot, there are clear outliers where standardized residuals exceed |2.5|. Furthermore, between fitted values 1.0 and 1.5, the residuals seem to cluster around e = 0.56 and variance seem lower, suggesting non-constant variance. Finally, all hat values are below 2.5*(7+1)/1144 = 0.017, which means there are no clear leverage points. Before handling the unusual observations, the group related non-constant variance needs to be accounted for with a weighted least squares regression. Below are the residuals vs Group plot and Levene's test output:

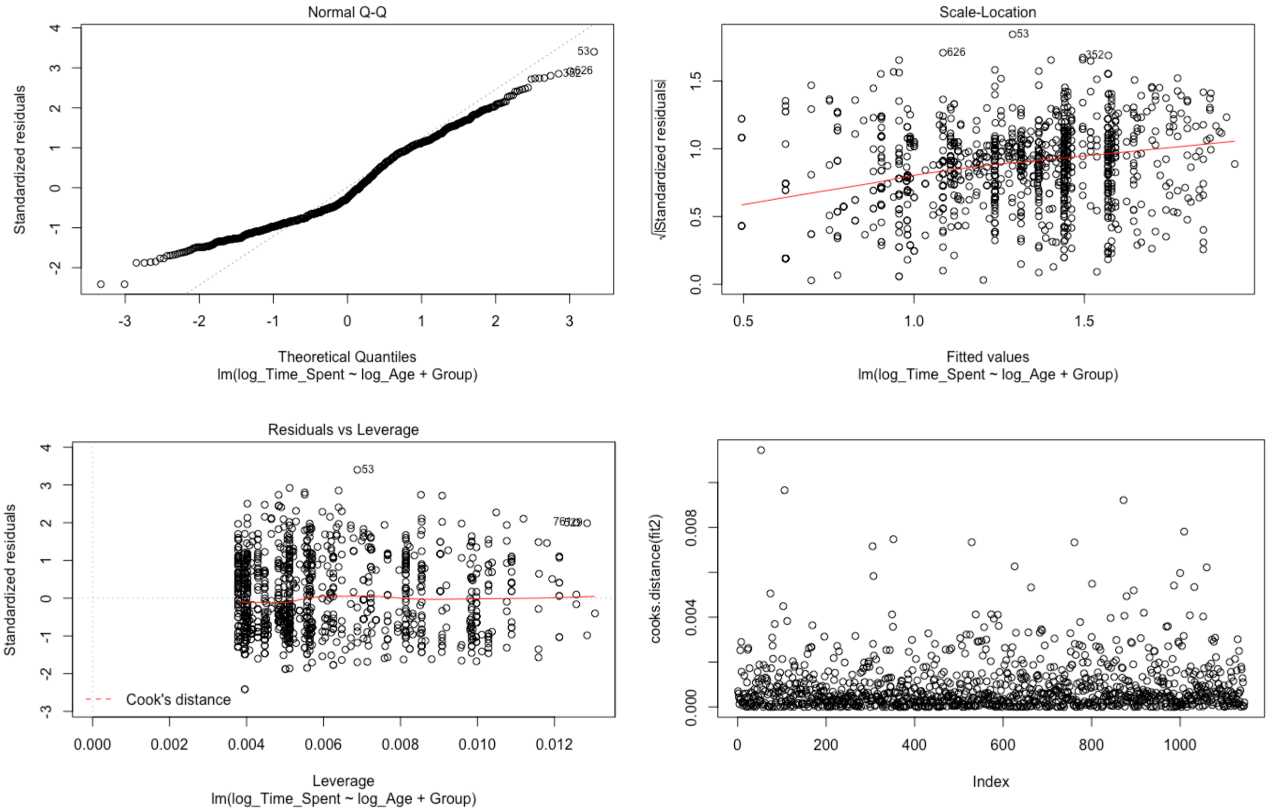


```
Analysis of Variance Table

Response: absres
            Df  Sum Sq Mean Sq F value    Pr(>F)
Group        5   6.883 1.37666  5.1967 0.0001009 ***
Residuals 1139 301.731 0.26491

lm(formula = log_Time_Spent ~ log_Age + Group, data = train,
    weights = wt)
```

The Levene’s test is very significant and strongly suggests that Group explains some of the non-constant variance. A weighted least squares regression was performed and the output and diagnostics are shown below:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.848263   0.040571  20.908  < 2e-16 ***
log_Age      0.428041   0.034630  12.360  < 2e-16 ***
Hounding     0.152303   0.058585   2.600  0.00945 **
Toy         -0.353998   0.038712  -9.144  < 2e-16 ***
Working      0.003812   0.047690   0.080  0.93630
Terrier      0.130125   0.038372   3.391  0.00072 ***
Herding     -0.072783   0.037156  -1.959  0.05037 .
```

```
Residual standard error: 0.6219 on 1138 degrees of freedom
Multiple R-squared:  0.1784,    Adjusted R-squared:  0.174
F-statistic: 41.18 on 6 and 1138 DF,  p-value: < 2.2e-16
```



Some non-constant variance remains, but it seems weighted least squares has controlled for most of it. A Levene's test for the weighted model outputs p-value = 0.191 which is no longer significant. The final step for the constant shift model is to remove unusual observations, particularly outliers with extremely high standardized residuals. In total, 19 observations were removed which accounts for 1.66% of the training data. These observations are posted in Appendix 1 and all of them were removed due to high standardized residuals. Unfortunately, there is really no information as to why these specific dogs had unusual time spent at the shelter without visiting or asking the Austin Animal Center. The final WLS constant shift model output and diagnostics are shown below:

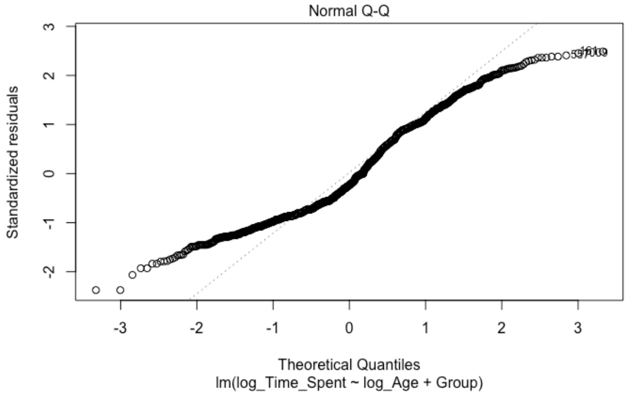
Anova Table (Type III tests)

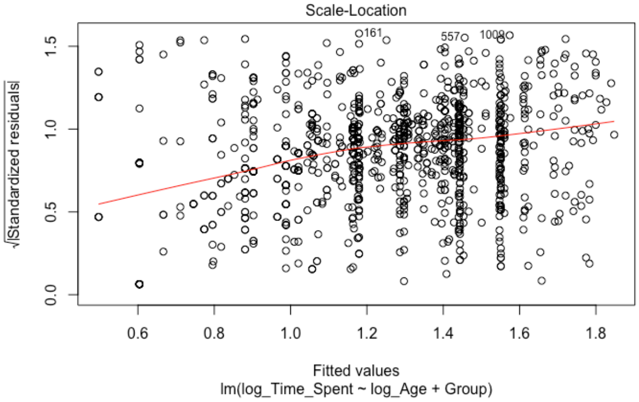
```
Response: log_Time_Spent
            Sum Sq   Df F value    Pr(>F)
(Intercept) 195.32    1 555.679 < 2.2e-16 ***
log_Age      42.42    1 120.691 < 2.2e-16 ***
Group        59.28    5  33.732 < 2.2e-16 ***
Residuals   393.32 1119
```

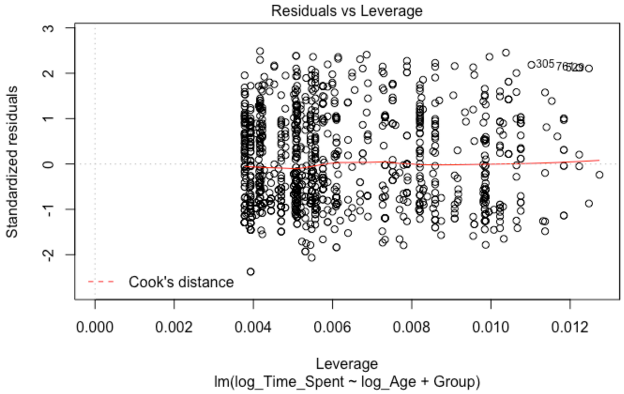
```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.89860    0.03812  23.573  < 2e-16 ***
log_Age      0.35565    0.03237  10.986  < 2e-16 ***
Hounding     0.17058    0.05829   2.927  0.00350 **
Toy         -0.40165    0.03395 -11.831  < 2e-16 ***
Working      0.01477    0.04667   0.317  0.75167
Terrier      0.16134    0.03820   4.224  2.6e-05 ***
Herding     -0.10300    0.03447  -2.988  0.00286 **

Residual standard error: 0.5929 on 1119 degrees of freedom
Multiple R-squared:  0.2005,    Adjusted R-squared:  0.1962
F-statistic: 46.78 on 6 and 1119 DF,  p-value: < 2.2e-16
```

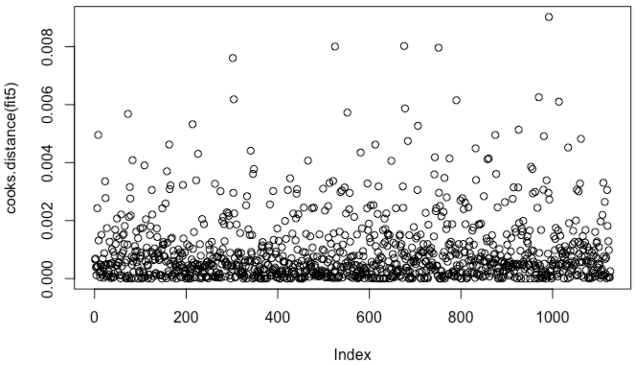
Tables of means

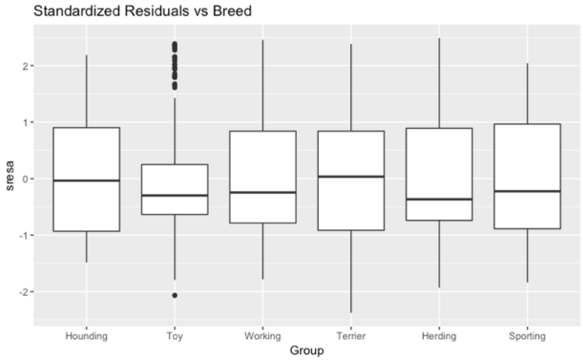
```
Grand mean
1.254277
      Hounding   Toy Working Terrier Herding Sporting
       1.437   0.8603   1.279   1.422   1.163    1.425
rep  102.000 197.0000 122.000 265.000 242.000  198.000

Ordinary Unadjusted Means
Hounding      Toy  Working  Terrier  Herding Sporting
1.449822 0.872923 1.291319 1.434511 1.175193 1.437598
```

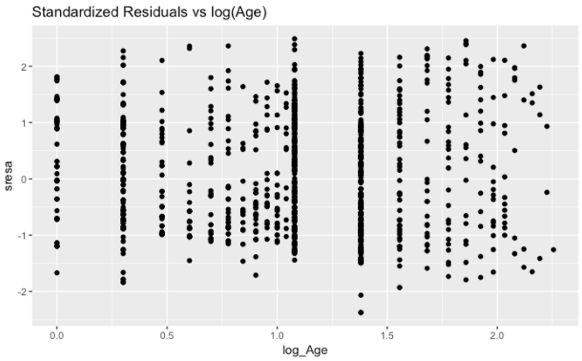












The unusual observations were not influential; the conclusions of the model are largely the same as before with both log(Age) and Group statistically significant. There are still some high standardized residuals but they are not influential or clear outliers. In addition, the residuals have some non-normality and may require a more complicated model or additional predictors to achieve a better fit. The p-value on Levene's test is 0.0717. While that is generally a level which is not significant, it is still concerning that the p-value has dropped from 0.191. However, this remains the best constant shift model.

4. Full Model and Unusual Observations:

An alternative to the constant shift model is a full model with interaction terms. The analysis begins with fitting the original training data before unusual observations were removed:

Anova Table (Type III tests)

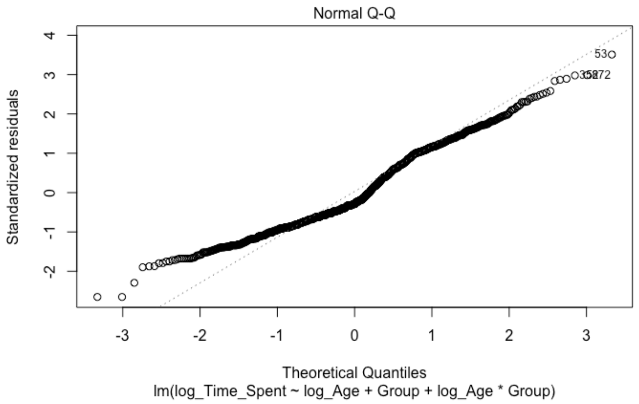
```
Response: log_Time_Spent
              Sum Sq   Df  F value    Pr(>F)
(Intercept)   140.84    1 369.0928 < 2.2e-16 ***
log_Age        57.30    1 150.1658 < 2.2e-16 ***
Group           3.36    5   1.7636    0.1175
log_Age:Group  10.17    5   5.3310 7.526e-05 ***
Residuals     432.34 1133
```

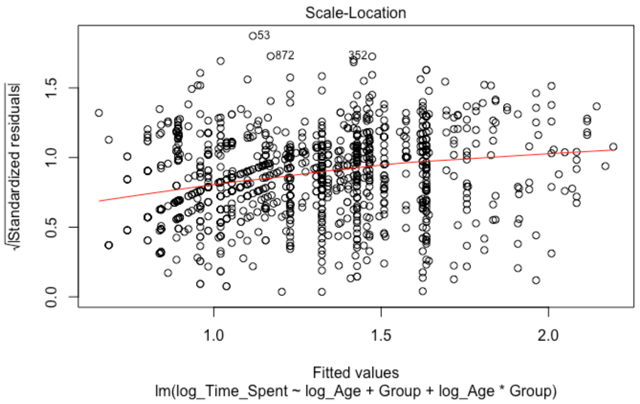
The partial F-test shows the interaction term log(Age)*Group is very significant and the full model performance could be better than the constant shift model. The regression output, diagnostics and Levene’s test output are displayed below:

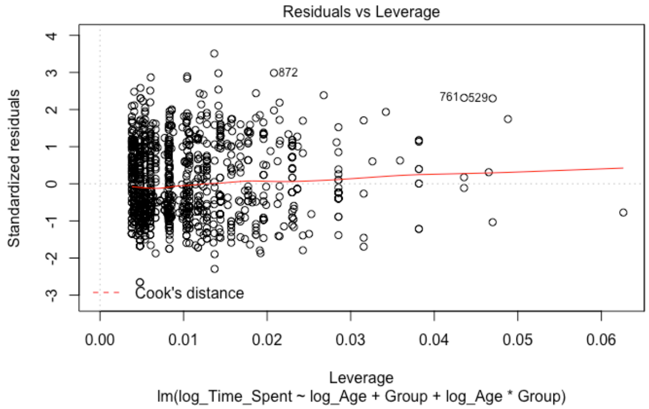
Coefficients:

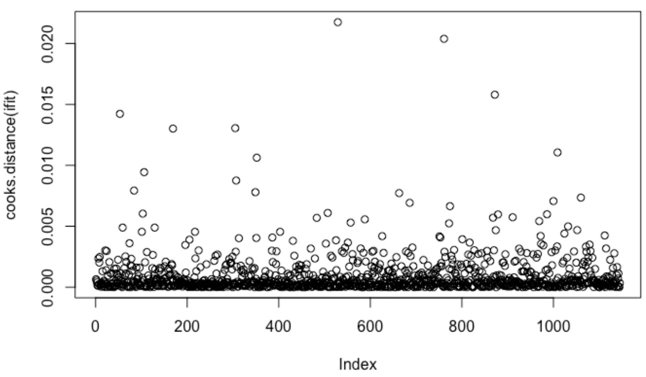
```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.81756    0.04256  19.212  < 2e-16 ***
log_Age           0.45651    0.03725  12.254  < 2e-16 ***
Hounding          0.22033    0.10732   2.053  0.04030 *
Toy              -0.07602    0.08751  -0.869  0.38518
Working          -0.16099    0.11928  -1.350  0.17740
Terrier          -0.13143    0.08841  -1.487  0.13738
Herding           0.07095    0.07949   0.892  0.37234
log_Age:Hounding -0.06785    0.09703  -0.699  0.48452
log_Age:Toy      -0.25449    0.07243  -3.513  0.00046 ***
log_Age:Working   0.16107    0.10543   1.528  0.12686
log_Age:Terrier   0.23094    0.07323   3.154  0.00165 **
log_Age:Herding  -0.14171    0.06994  -2.026  0.04299 *
```

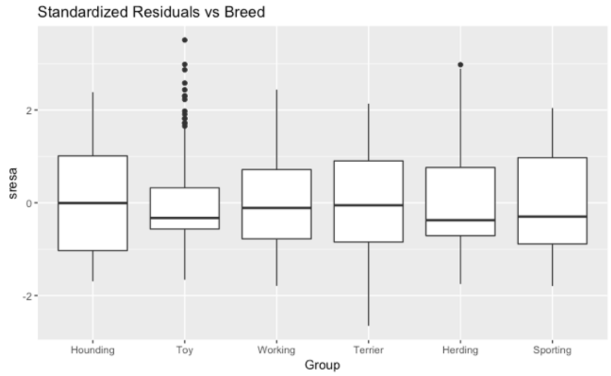
```
Residual standard error: 0.6177 on 1133 degrees of freedom
Multiple R-squared:  0.1981,    Adjusted R-squared:  0.1903
F-statistic: 25.44 on 11 and 1133 DF,  p-value: < 2.2e-16
```









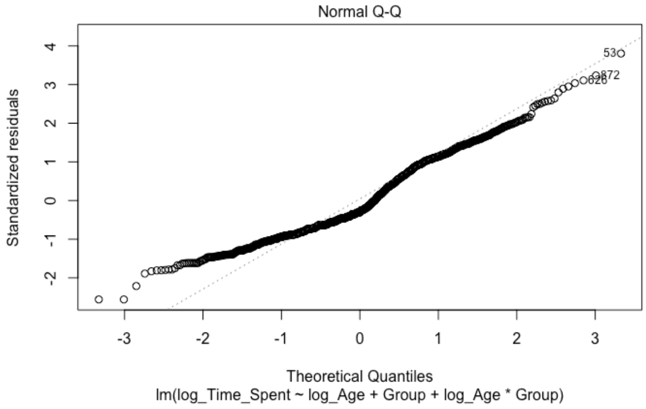


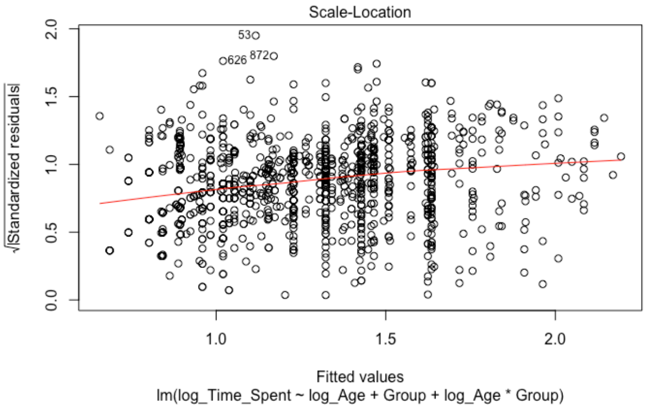
Analysis of Variance Table

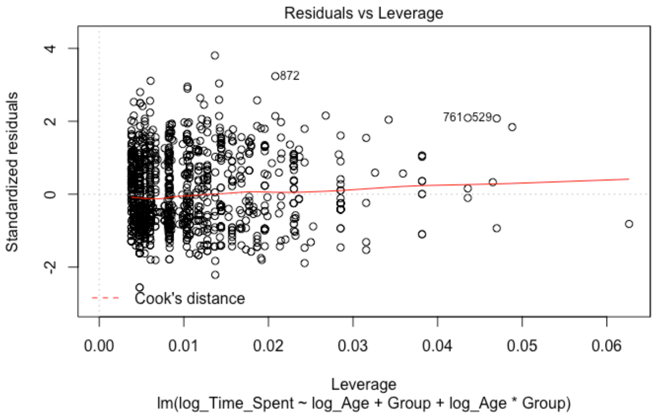
```
Response: absres

             Df  Sum Sq Mean Sq F value    Pr(>F)
Group         5   6.869 1.37374  5.0005 0.0001547 ***
Residuals 1139 312.904 0.27472
```

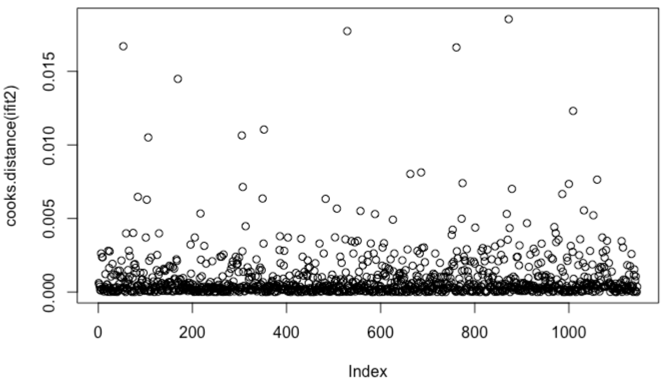
Just like the constant shift case, the Levene's test is statistically significant and the residuals vs group plot shows non-constant variance— a weighted least squares regression is appropriate here. In addition, there are very large outliers and leverage points that need to be examined and possibly removed. Another note is that adjusted R^2 compared to the constant shift model (without weights) is only about 2% higher; the extra complexity offers little improvement to model performance. The WLS full model regression output and diagnostics are shown below:

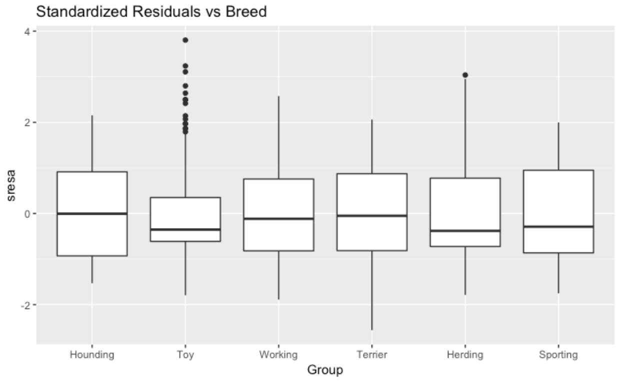
```
Coefficients:      Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.81756    0.04285  19.082  < 2e-16 ***
log_Age            0.45651    0.03760  12.140  < 2e-16 ***
Hounding           0.22033    0.11721   1.880 0.060378 .
Toy               -0.07602    0.08255  -0.921 0.357298
Working           -0.16099    0.11398  -1.412 0.158082
Terrier           -0.13143    0.09105  -1.444 0.149150
Herding            0.07095    0.07857   0.903 0.366724
log_Age:Hounding -0.06785    0.10609  -0.640 0.522596
log_Age:Toy       -0.25449    0.06856  -3.712 0.000216 ***
log_Age:Working    0.16107    0.10076   1.599 0.110203
log_Age:Terrier    0.23094    0.07541   3.063 0.002246 **
log_Age:Herding   -0.14171    0.06917  -2.049 0.040729 *

Residual standard error: 0.6159 on 1133 degrees of freedom
Multiple R-squared:  0.1987,    Adjusted R-squared:  0.1909
F-statistic: 25.54 on 11 and 1133 DF,  p-value: < 2.2e-16
```











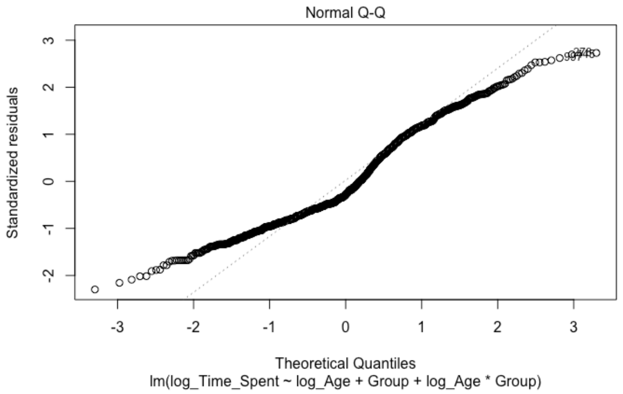
```
Analysis of Variance Table

Response: absres
            Df Sum Sq Mean Sq F value Pr(>F)
Group        5   1.59 0.31739  1.1336 0.3406
Residuals 1139 318.90 0.27998
```

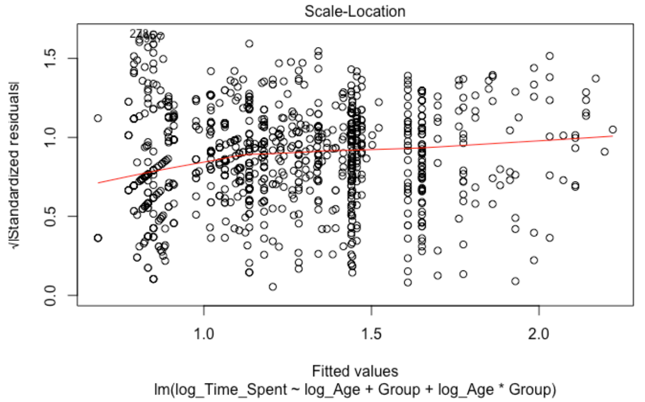
The Levene's test is no longer significant after performing weighted least squares. Handling unusual observations for this model and dataset was quite difficult— when influential points were removed, new outliers and leverage points appear in the next fit. In fact, following this cycle of removing influential outliers and leverage points will result in removing all observations that belong to Hounding breeds, which is problematic since the validation set and out-of-sample instances will contain dogs from this group. A possible explanation is the swamping effect, where good observations are mistakenly flagged as unusual and removed. A solution is to apply more advanced anomaly detection techniques rather than the current procedure of removing observations outside-in, but that is beyond the scope of this report. Instead, unusual observations were removed until the regression output barely changed, and the interpretation of the model is stable. In addition, Cook's distance plots were greatly used to assess how influential the unusual observations were. While Cook's distance may not capture all influential observations, it can be used to further support the belief that the model is relatively unaffected by these unusual observations. In total, 128 observations were removed (11% of the training data) and are listed in Appendix 1. The final full model regression output and relevant plots are displayed below:

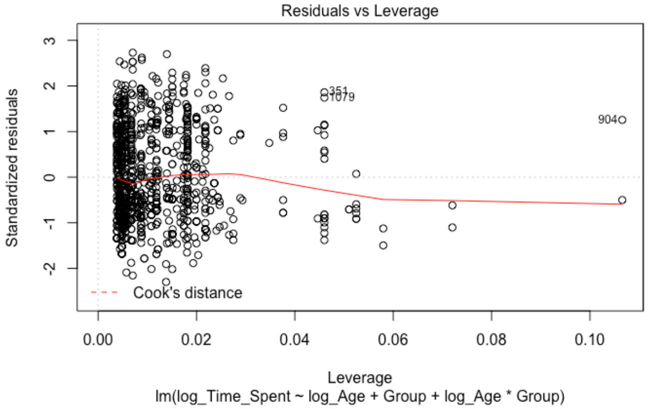
Anova Table (Type III tests)

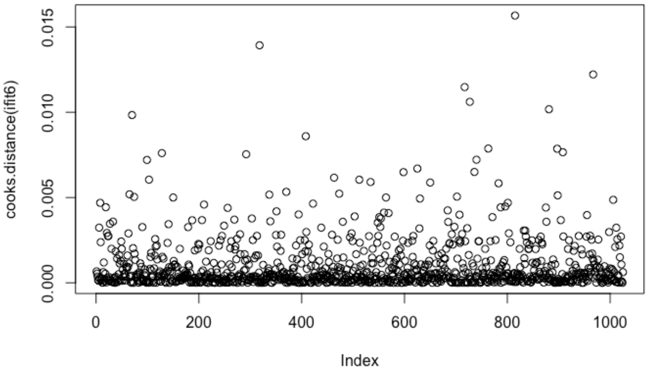
```
Response: log_Time_Spent
              Sum Sq   Df F value    Pr(>F)
(Intercept)    32.40    1 103.324 < 2.2e-16 ***
log_Age         3.74    1  11.926 0.0005765 ***
Group           4.49    5   2.863 0.0141788 *
log_Age:Group  20.09    5  12.814 4.225e-12 ***
Residuals     317.65 1013
```

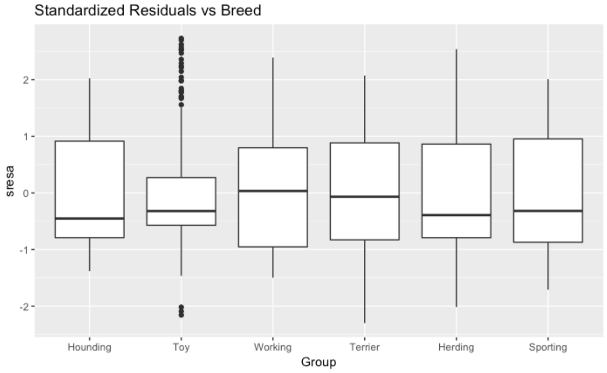
```
Coefficients:     Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.957459   0.093791  10.208  < 2e-16 ***
log_Age           0.290945   0.081008   3.592 0.000344 ***
Hounding         -0.194698   0.299052  -0.651 0.515160
Toy              -0.165592   0.107966  -1.534 0.125401
Working           0.674225   0.348844   1.933 0.053545 .
Terrier          -0.273609   0.122365  -2.236 0.025567 *
Herding           0.006899   0.109540   0.063 0.949796
log_Age:Hounding  0.201524   0.276094   0.730 0.465611
log_Age:Toy      -0.230592   0.092412  -2.495 0.012743 *
log_Age:Working  -0.469995   0.284662  -1.651 0.099034 .
log_Age:Terrier   0.409590   0.103241   3.967 7.78e-05 ***
log_Age:Herding  -0.125558   0.095758  -1.311 0.190084

Residual standard error: 0.5662 on 1020 degrees of freedom
Multiple R-squared:  0.2458,    Adjusted R-squared:  0.2376
F-statistic: 30.22 on 11 and 1020 DF,  p-value: < 2.2e-16
```

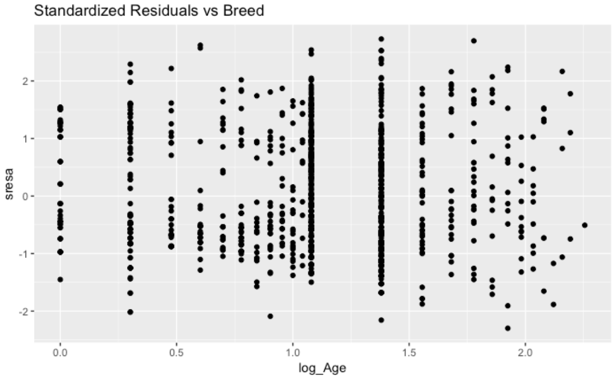




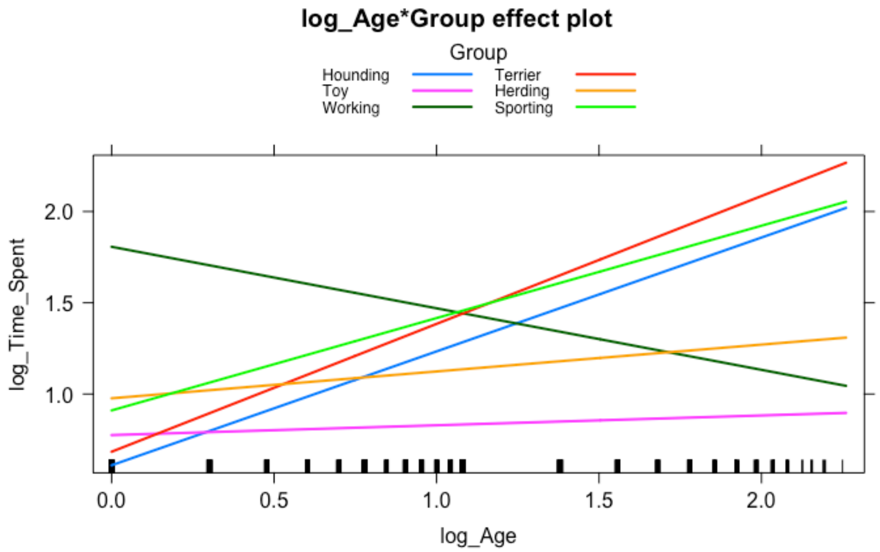








```
Analysis of Variance Table

Response: absres
                Df  Sum Sq Mean Sq F value  Pr(>F)
itrain5$Group    5   2.715 0.54305  2.0039 0.07566 .
Residuals     1019 276.151 0.27100
```



The partial F-test for Log(Age)*Group is still significant which means the interaction term can be kept in the model. There are clear outliers and leverage points but they are likely not influential. Cook's distances are also far below 1. Similar to the constant shift model, the normality plot shows clear non-normality in the residuals. Levene's test is not significant, however, the p-value is quite low and the standardized residuals vs group boxplot suggests non-constant variance may still be an issue. Finally, the effect plot suggests three subgroups for log(Time_Spent) vs log(Age) relationship. Working breeds have a negative slope and appear to be very different from the rest. Herding and Toy breeds have relatively flat positive slope. Terrier, Sporting and Hounding breeds all have relatively steep positive slopes.

5. Model Selection:

In this analysis, there are three possible models— a simple regression with log(Age), a constant shift model, and a full model with interaction. Simple regression can be easily ruled out as the best model; partial F-test showed that Group is highly significant and the simple regression's adjusted R^2 = 10.81%, much lower than the constant shift model. The unweighted constant shift model and full model performs similarly in terms of adjusted R^2; constant shift adjusted R^2 = 17.49% and full model adjusted R^2 = 19.03%. A few predictions by the final WLS versions of both models are displayed below:

Constant Shift Model Predictions (in days)

	fit	lwr	upr	Time Spent (in days)
1	7.99	0.59	108.74	4
2	62.52	4.59	850.76	21
3	14.17	1.04	192.39	1
4	9.23	0.68	125.51	105
5	18.80	1.10	320.18	4
6	18.39	1.10	306.99	46
7	7.99	0.59	108.74	28
8	27.78	1.63	472.75	250
9	41.06	2.41	699.24	239
10	13.47	0.81	224.28	1442

Full Model Predictions (in days)

	fit	lwr	upr	Time Spent (in days)
1	10.49	0.89	124.23	4
2	91.67	7.52	1117.41	21
3	13.30	1.13	156.30	1
4	11.14	0.94	131.44	105
5	12.75	0.87	186.72	4
6	7.78	0.54	111.99	46
7	10.49	0.89	124.23	28
8	27.53	1.89	400.56	250
9	59.44	4.07	868.53	239
10	7.43	0.52	106.09	1442

Both models are poor predictors of how long a dog stays at the Austin Animal Center. The prediction intervals are far too wide and the fitted values are also very different from the actual values. Furthermore, some of the actual values lie outside of the 95% prediction interval. Overall, given that both models perform poorly, the better model is the constant shift model. Firstly, it is simpler and will likely generalize better for new data. Secondly, the analysis failed to properly handle the unusual observations in relations to the full model, therefore, it is possible that current full model is still affected by influential points.

6. Interpreting the best model:

$y = \mu + \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 E_3 + \alpha_4 E_4 + \alpha_5 E_5 + \beta_1 x_1$

$y = 0.90 + 0.17E_1 - 0.40E_2 + 0.01E_3 + 0.16E_4 - 0.10E_4 + 0.36x_1$

E_1: Effect coding where 1 indicates Hounding breed
E_2: Effect coding where 1 indicates Toy breed
E_3: Effect coding where 1 indicates Working breed
E_4: Effect coding where 1 indicates Terrier breed
E_5: Effect coding where 1 indicates Herding breed
x_1: log(Age) of a dog with Age in months

The intercept μ is the overall level and is very significant. For predictions, however, it has very little practice use since every dog belongs to one of the Groups, which means there will never be an instance where all predictors are 0 and $y = \mu$. The coefficients of the effect codings can be interpreted by undoing the logs:

$$\log(y) = \alpha_i E_i$$

$$\log(y) = \log(z_i) E_i$$

$$y = z_i^{E_i}$$

Therefore, coefficient $\alpha_1 = 0.17$ means that when a dog belongs to the Hounding group, it is associated with an increase of $10^{0.17} = 1.04$ in Time_Spent at the shelter in months, given all else is held fixed. The covariate coefficient $\beta_1 = 0.36$ is a log-log relationship and means that a 1% change in Age is associated with a 0.36% change in Time_Spent at the shelter in months, given all else is held fixed.

6. Conclusion:

The best model performs does not predict the time a dog spends at the Austin Animal Center accurately or precisely, therefore, it has extremely limited practical use. However, the model does demonstrate that Age and Group are significant predictors of Time_Spent, and perhaps could be included in more complicated models. The model also provides some practical insights; the coefficients for Hounding, Working, Terrier and Sporting are all positive, which means that any dog in these Group may spend longer times at the shelter compared to the overall level and may require more promoting to potential adopters. Overall, the model could be improved on with additional predictors or more sophisticated ways of handling unusual observations in the full model case.

Data:

1) "Austin Animal Center Outcomes." *Austin Animal Center*, https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Outcomes/9t4d-g238.

2) "Austin Animal Center Intakes." *Austin Animal Center*, https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Intakes/wter-evkm.

3) "List of Breeds by Group." *American Kennel Club*, https://www.akc.org/public-education/resources/general-tips-information/dog-breeds-sorted-groups/.