Stat 344 – HW 14

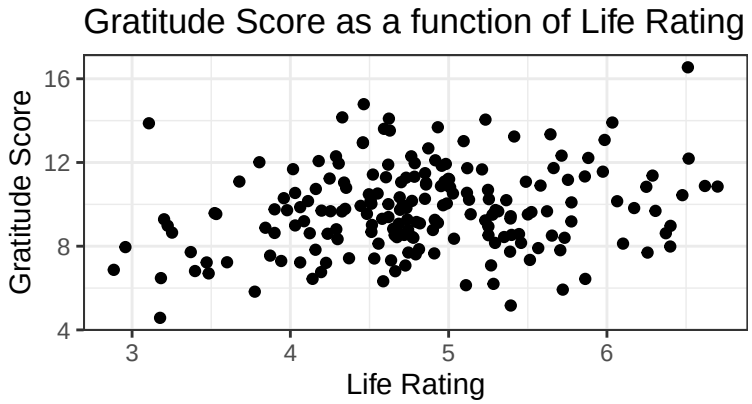
Trey Tipton

April 13, 2022

A. Exploration

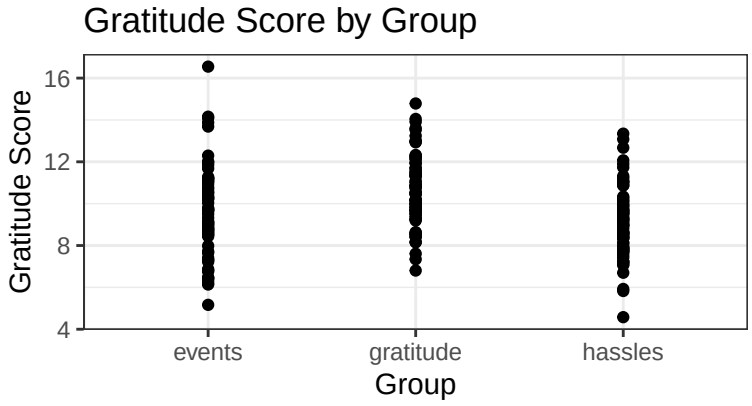
```
grateful <- read.csv('https://sldr.netlify.com/data/gratitude-experiment.csv')

gm <- lm(gratitude_score ~ life_rating + group + illness_score, data = grateful)
```

```
gf_point(gratitude_score ~ life_rating, data = grateful) %>%
  gf_labs(title = "Gratitude Score as a function of Life Rating",
          x = "Life Rating", y = "Gratitude Score")
```



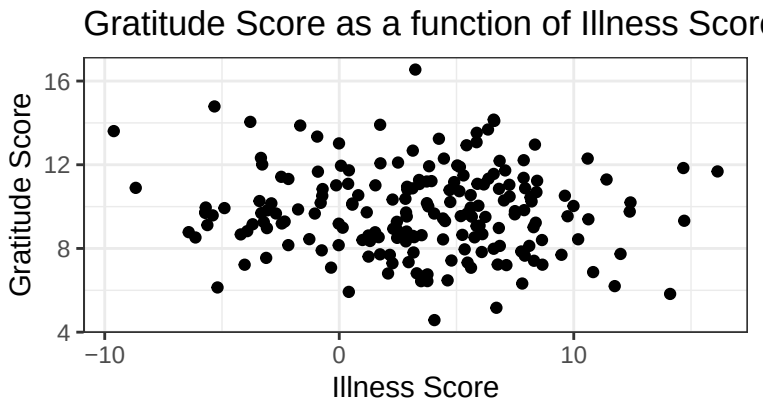
The gratitude score by life rating score appears to be randomly scattered and there does not seem to be a relationship between the two.

```
gf_point(gratitude_score ~ group, data = grateful) %>%
  gf_labs(title = "Gratitude Score by Group",
          x = "Group", y = "Gratitude Score")
```



The mean of gratitude scores by group appear to go from highest to lowest from gratitude, events, and hassles respectively. The events group also appears to have the largest variance.

```
gf_point(gratitude_score ~ illness_score, data = grateful) %>%
  gf_labs(title = "Gratitude Score as a function of Illness Score",
          x = "Illness Score", y = "Gratitude Score")
```



Similar to the life rating, illness score does not appear to have an effect on the gratitude score as they seem to be randomly scattered.

Models and R-squared

```
m1 <- lm(gratitude_score ~ 1, data = grateful)
summary(m1)
```

```
##
## Call:
## lm(formula = gratitude_score ~ 1, data = grateful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1784 -1.3126 -0.0853  1.3087  6.7927
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.7556     0.1442   67.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.018 on 195 degrees of freedom
```

```
m2 <- lm(gratitude_score ~ life_rating, data = grateful)
summary(m2)
```

```
##
## Call:
## lm(formula = gratitude_score ~ life_rating, data = grateful)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.891 -1.466 -0.098  1.219  5.899
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.2008     0.8828   8.157 4.22e-14 ***
## life_rating   0.5295     0.1806   2.932  0.00377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.98 on 194 degrees of freedom
## Multiple R-squared:  0.04243,    Adjusted R-squared:  0.03749
## F-statistic: 8.596 on 1 and 194 DF,  p-value: 0.003775
```

```
m3 <- lm(gratitude_score ~ group, data = grateful)
summary(m3)
```

```
##
## Call:
## lm(formula = gratitude_score ~ group, data = grateful)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.527 -1.208 -0.128  1.160  7.016
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.5317     0.2383  40.006  < 2e-16 ***
## groupgratitude   1.1096     0.3369   3.293  0.00118 **
## grouphassles    -0.4280     0.3357  -1.275  0.20382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.921 on 193 degrees of freedom
## Multiple R-squared:  0.1036, Adjusted R-squared:  0.0943
## F-statistic: 11.15 on 2 and 193 DF,  p-value: 2.612e-05
```

```
m4 <- lm(gratitude_score ~ illness_score, data = grateful)
summary(m4)
```

```
##
## Call:
## lm(formula = gratitude_score ~ illness_score, data = grateful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1600 -1.2910 -0.1164  1.3388  6.7890
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.84731    0.17698  55.641   <2e-16 ***
## illness_score -0.02714    0.03034  -0.895    0.372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.019 on 194 degrees of freedom
## Multiple R-squared:  0.00411,    Adjusted R-squared:  -0.001024
```

```
## F-statistic: 0.8006 on 1 and 194 DF,  p-value: 0.372
```

```
m5 <- lm(gratitude_score ~ life_rating + group, data = grateful)
summary(m5)
```

```
##
## Call:
## lm(formula = gratitude_score ~ life_rating + group, data = grateful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7003 -1.3238 -0.1666  1.0438  6.1955
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.5188     0.8501   8.845 5.93e-16 ***
## life_rating      0.4352     0.1766   2.464  0.01461 *
## groupgratitude   0.9397     0.3397   2.767  0.00622 **
## grouphassles    -0.5188     0.3334  -1.556  0.12131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.896 on 192 degrees of freedom
## Multiple R-squared:  0.1311, Adjusted R-squared:  0.1175
## F-statistic: 9.654 on 3 and 192 DF,  p-value: 5.764e-06
```

```
m6 <- lm(gratitude_score ~ life_rating + illness_score, data = grateful)
summary(m6)
```

```
##
## Call:
## lm(formula = gratitude_score ~ life_rating + illness_score, data = grateful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8005 -1.4093 -0.0906  1.1929  5.8957
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.29224    0.88881   8.204 3.21e-14 ***
## life_rating    0.52962    0.18069   2.931  0.00378 **
## illness_score -0.02720    0.02976  -0.914  0.36187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.981 on 193 degrees of freedom
## Multiple R-squared:  0.04655,    Adjusted R-squared:  0.03667
## F-statistic: 4.712 on 2 and 193 DF,  p-value: 0.01005
```

```
m7 <- lm(gratitude_score ~ illness_score + group, data = grateful)
summary(m7)
```

```
##
## Call:
## lm(formula = gratitude_score ~ illness_score + group, data = grateful)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5185 -1.1751 -0.1266  1.1498  7.0025
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.577493   0.276048  34.695  < 2e-16 ***
## illness_score  -0.009785   0.029590  -0.331  0.74125
## groupgratitude  1.085615   0.345423   3.143  0.00194 **
## grouphassles   -0.442148   0.339157  -1.304  0.19391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.925 on 192 degrees of freedom
## Multiple R-squared:  0.1041, Adjusted R-squared:  0.0901
## F-statistic: 7.437 on 3 and 192 DF,  p-value: 9.752e-05
```

```
## gm is m8, the full model
summary(gm)
```

```
##
## Call:
## lm(formula = gratitude_score ~ life_rating + group + illness_score,
##     data = grateful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6766 -1.2877 -0.2173  1.0723  6.1700
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.56331    0.85764   8.819  7.2e-16 ***
## life_rating     0.43882    0.17717   2.477   0.0141 *
## groupgratitude  0.90631    0.34850   2.601   0.0100 *
## grouphassles   -0.53846    0.33696  -1.598   0.1117
## illness_score  -0.01307    0.02923  -0.447   0.6553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.9 on 191 degrees of freedom
## Multiple R-squared:  0.132,  Adjusted R-squared:  0.1138
## F-statistic:  7.26 on 4 and 191 DF,  p-value: 1.834e-05
```

Models 1 through 8 results for r^2, adjusted r^2.

gratitude_score ~ 1

0, 0

gratitude_score ~ life_rating

0.04243, 0.03749

gratitude_score ~ group

0.1036, 0.0943

gratitude_score ~ illness_score

0.00411, -0.001024

gratitude_score ~ life_rating + group

0.1311, 0.1175

gratitude_score ~ life_rating + illness_score

0.04655, 0.03667

gratitude_score ~ illness_score + group

0.1041, 0.0901

gratitude_score ~ illness_score + group + illness_score

0.132, 0.1138

The highest r^2 value is the full model of all three predictors with an r^2 of 0.132. However, the highest adjusted r^2 is model 5, which includes only life rating and group as predictors for gratitude score with an adjusted r^2 of 0.1175.

If presenting results, I would schoose to report the adjusted r^2 since we are comparing models with a different number of predictors. The adjusted r^2 value takes into account how many predictors the model has, k, while the multiple r^2 value does not and gets better if there is simply more predictors. For this reason, I would report the model 5 to be a better r^2 than the full model, because it is a higher r^2 value when taking into account the number of predictors.

Model Selection

Approach 1: Backwards Step-wise Selection

```
car::Anova(gm)
```

```
## Anova Table (Type II tests)
##
## Response: gratitude_score
##               Sum Sq  Df F value    Pr(>F)
## life_rating    22.15   1  6.1345 0.0141260 *
## group          67.86   2  9.3984 0.0001279 ***
## illness_score   0.72   1  0.1999 0.6553434
## Residuals     689.59 191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the original model, the largest large p-value is generated by illness_score, so we should remove it. Remove illness_score

```
car::Anova(m5)
```

```
## Anova Table (Type II tests)
##
## Response: gratitude_score
##             Sum Sq  Df F value    Pr(>F)
## life_rating  21.83   1  6.0722   0.01461 *
## group        70.42   2  9.7933 8.913e-05 ***
## Residuals   690.32 192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The remaining p-values are small, so each predictor now plays a significant role in the model. The best model includes life_rating and group as predictors for gratitude_score, which lines up with what our adjusted r^2 value told us earlier.

Approach 2: Forward Selection

```
car::Anova(m1)
```

```
## Warning in Anova.lm(m1): the model contains only an intercept: Type III test
## substituted
```

```
## Anova Table (Type III tests)
##
## Response: gratitude_score
##              Sum Sq  Df F value    Pr(>F)
## (Intercept) 18653.5   1  4578.6 < 2.2e-16 ***
## Residuals     794.4 195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(m2)
```

```
## Anova Table (Type II tests)
##
## Response: gratitude_score
##             Sum Sq  Df F value   Pr(>F)
## life_rating  33.71   1  8.5956 0.003775 **
## Residuals   760.74 194
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(m3)
```

```
## Anova Table (Type II tests)
##
## Response: gratitude_score
##           Sum Sq  Df F value    Pr(>F)
## group      82.30   2  11.152 2.612e-05 ***
## Residuals 712.15 193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(m4)
```

```
## Anova Table (Type II tests)
##
## Response: gratitude_score
##               Sum Sq  Df F value Pr(>F)
## illness_score   3.26   1  0.8006  0.372
## Residuals     791.18 194
```

The smallest small p-value is from m3, the model with only group as a predictor. So we add group as a predictor and try the models with group + other predictors individually.

```
car::Anova(m5)
```

```
## Anova Table (Type II tests)
##
## Response: gratitude_score
##             Sum Sq  Df F value    Pr(>F)
```

```
## life_rating  21.83   1  6.0722   0.01461 *
## group        70.42   2  9.7933 8.913e-05 ***
## Residuals   690.32 192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(m7)
```

```
## Anova Table (Type II tests)
##
## Response: gratitude_score
##               Sum Sq  Df F value    Pr(>F)
## illness_score   0.41   1  0.1093    0.7412
## group          79.44   2 10.7144 3.878e-05 ***
## Residuals     711.74 192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The next smallest small p-value comes from m5, the model with group and life_rating as predictors. So we should now have group and life_rating as predictors and try the models with those and the other predictors individually. There is only one more predictor to try, which is the original model.

```
car::Anova(gm)
```

```
## Anova Table (Type II tests)
##
## Response: gratitude_score
##               Sum Sq  Df F value    Pr(>F)
## life_rating    22.15   1  6.1345 0.0141260 *
## group          67.86   2  9.3984 0.0001279 ***
## illness_score   0.72   1  0.1999 0.6553434
## Residuals     689.59 191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since, illness_score does not have a small p-value, we do not add it to the model. Therefore, the best model is m5 where group and life_rating are predictors of gratitude_score.

Both methods yield the same answer, neither with the full model, meaning that there is reason to remove illness score as a predictor.

I do believe that m5 is the best model. This is the model with life_rating and group are predictors for gratitude_score and illness_score is removed. This appears to be the best model because all three methods to check the goodness of fit for the model yielded m5 as the model, including the adjusted r^2 value, backwards step-wise selection, and forward selection. The only method that provided a different answer was the multiple r^2 value, which makes sense because it does not account for k number of predictors and gets better for models with more predictors. Therefore, m5 overwhelmingly outperforms the other models and is the overall best model.

Overall, I can conclude that life_rating and group are the best predictors of the gratitude score out of the predictors I chose. However, according to the scatterplots I showed and the relatively low adjusted r^2 value from not only the full model but all of them, the relationship does not appear to be overwhelming and I would not say that there is a good linear relationship between gratitude_score and the predictors I chose. The best predictor appears to be group according to the scatter plot and anova p-values, while the worst predictor was illness_score.