SAS Project 2: Correlation and Regression Analysis of GPA and Frequency of Caffeine Consumption

STAT 430

In this project, we are analyzing how often one consumes caffeine in a week(Days_Per_Wk) can affect GPA. In other words, Days_Per_Wk is the independent variable and GPA is the dependent variable.

1. Correlation Coefficient Analysis:

Correlation coefficient is a measure of the strength of the relationship between x (Days_Per_Wk) and y (GPA) values in the sample. From SAS results, the correlation coefficient (r) is -0.74177. Since the correlation coefficient is negative, this means that there is a negative correlation between the frequency of caffeine intake per week and GPA. Using the general rule of thumb, the magnitude of r = |-0.74177| = 0.74177 is between 0.5 and 0.8 which indicates a moderate correlation between GPA and frequency of caffeine intake per week. Overall, there is a negative correlation between frequency of caffeine intake per week and GPA. In other words, as caffeine consumption increases, GPA decreases (vice versa). However, we can argue that correlation coefficient of 0.74177 is high enough to indicate a strong correlation between frequency of caffeine intake per week and GPA. Thus, both of the variables (Days_Per_Wk and GPA) will be kept in the linear regression model.

Note: Correlation is not the same as causation. Thus, we cannot say that higher frequency of caffeine consumption leads to poor GPA (vice versa). Low GPA may be the result of some other unknown factors like missing classes or poor time management.

2. ANOVA Result Analysis:

We conduct a one-way ANOVA test to compare variability among sample means to variability within each sample. From ANOVA table, the p-value (Pr>F) is <0.0001 or 0.00 and the F statistic is 78.29. ANOVA is testing the hypotheses: H_0: There is no difference in GPA among group means(i.e. all group means are equal) and H_a: At least one group differs significantly from the overall mean of GPA (the dependent variable). Since p-value of 0.00 is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that at least one group differs significantly from overall mean GPA. The small p-value shows that the independent variable is significant. Additionally, the large F ratio means the variation among group means is more than you'd expect to see by chance and that the results are significant. In this scenario, we can perform post-hoc tests such as Tukey HSD test to determine exactly which groups differ from each other. We categorize the frequency of caffeine consumption in a week into 5 groups: rarely (1-2 days), sometimes (3-4 days), frequently(5-6 days), 7(everyday). From Tukey results, we look at the comparisons that are significant and indicated with ***. The difference of means between the group consuming caffeine once/twice(rarely) is different from the group consuming caffeine frequently and the group consuming caffeine everyday. The difference of means between group consuming caffeine everyday(7 days) and group consuming caffeine frequently(5 or 6 days) is also significant. There is also a difference in means of the group consuming caffeine 3 or 4 times per week(sometimes)

and the group consuming caffeine everyday. Lastly, there is a difference in the means between the group consuming caffeine sometimes and the group consuming caffeine frequently.
R-square or coefficient of determination is the proportion of variance in one variable that can be explained by variation in the other variable. From ANOVA table, we see that the R-square value is 0.5502. An interpretation of R-square is: 55.02% of the variation in frequency of caffeine consumption (independent variable) per week can be explained by variation in GPA (dependent variable). With an R-square of 0.5502, we can conclude that the model is a good model since the number is above 0.50 (or 50%) .

Note: that R-square value does not determine whether the coefficient estimates and predictions are biased (i.e we need to look at residual plots). R-square also does not indicate whether regression model is adequate. You can have low R-square value for good model or high R-square value for bad model.

3. Regression Results
Next, we build the linear regression model with PROC REG after confirming significance of our independent variable (Days_Per_Wk).
The regression equation is $\widehat{y} = 4.06491 - 0.10180\, x$ or
GPA = 4.06491 - 0.10180 * Days_Per_Wk where Days_Per Wk is the frequency of caffeine consumption in a week.
The p-value (Pr> |t|) for parameter β_0 (y-intercept) is < .0001 or 0.00.
The p-value (Pr> |t|) for parameter β_1 (slope) is <0.0001 or 0.00.
The hypotheses tested for the significance of the linear regression equation are:
H_0: $\beta_1 = 0$ (frequency of caffeine consumption has no effect on GPA; linear regression is insignificant)
H_a: $\beta_1 \neq 0$ (changes in frequency of caffeine consumption are associated with changes in GPA; linear regression is significant)
Since the p-value for β_1(0.00) is less than $\alpha = 0.05$, we reject the null hypothesis and conclude $\beta_1 \neq 0$. This means that the linear relationship between GPA and frequency of caffeine consumption is significant (i.e. linear regression is significant). Since p-value for parameter β_0 (0.00) is less than $\alpha = 0.05$, the y-intercept is significant. Thus, we keep Days_Per_Wk(independent variable) and y-intercept in the model.

4. Residual Analysis
Next, we look at the normal probability plot in the Fit Diagnostics for GPA. It is observed that the residuals are identically distributed above and below the fit line. We can draw a straight line through the middle of residuals plot which means they are identically distributed. Looking at the "residuals for GPA" plot, there are only 4 residuals that are significantly far from the fitted line

(i.e. outliers below the -0.25 line. There is constant variance amongst the residuals which indicates negative correlation. Additionally, we see that the residuals are normally distributed as the histogram in the "Fit Diagnostics" closely follows the bell-shape of the normal distribution.

5. Make sure all assumptions are met. Explain what this does to your results.

The first criteria is that the dependent variable must be continuous. The dependent variable(GPA) is continuous since it can take an infinite number of possible values on the interval from 0.0 to 4.0 scale. Since GPA is a numerical/ quantitative variable, linear regression is the correct method. The second criteria is that the data we are modeling meets the identically distributed criterion. The error terms are not independent from one another since the residuals in the "Residuals for GPA" plot does not appear to be randomly distributed (ie. they are lined up in a straight line and not scattered). Serial correlation occurs in time-series studies when errors associated with a time period carry over to future time periods. How might this affect results? Serial correlation may cause estimated variance of regression coefficients to be biased which leads to unreliable hypothesis testing. The t-statistic may appear more significant than they really are. For example if we are predicting GPA, an overestimate in one group(frequency of caffeine consumption) might lead to overestimates in succeeding frequency of caffeine consumption. Therefore, we may need to investigate the residuals using the time series method or other methods. On the other hand, the error terms are identically distributed as the Q-Q plot shows the residuals closely follows the fitted line cleanly without a lot of outliers. Though not accurate, we can also look at the histogram in Fit Diagnostics which shows that the residuals closely follows a normal distribution. The third criteria is that the error term is normally distributed with a mean of zero and a standard deviation of σ^2. Indeed, the Q-Q plot shows the residual points close to the fit line and the mean of error terms in the histogram for residuals is center at 0. Thus, the error terms follow a normal distribution, $N(0, \sigma^2)$.

5. Explain why your sample is good or not good.

I believe my sample is good since the R-square is above 0.50 or 50%. This means that the linear regression model is generally a good fit to the data (though not exactly the best since R-square is not above 0.75). The ANOVA results indicated that the independent variable is significant (i.e $Pr > |t| = 0.00 < \alpha = 0.05$ means there are differences in means of GPA for different groups based on frequency of caffeine consumption). Both of the parameters $\beta_0$and β_1 are significant with p-value (Pr>F) of $0.00 < \alpha = 0.05$. This means parameter estimates/ independent variable (GPA) is significant Additionally, the MSE (0.0424) and RMSE (0.2059) is quite low which indicates that the forecast/estimates are close to the actual values. Overall, the linear regression model seems to be a reasonable method to analyze the data.

Though we cannot simply conclude that drinking more caffeine leads to poor GPA, we know from the model that frequency of drinking caffeine has some effect on GPA.

What can be improved: The correlation coefficient might be higher (>0.8) if we sampled more people. Given the time constraints, I was not able to collect more data/ conduct another survey. Additionally, we can improve on SRS design since there is not an equal number of students from all majors. Students can voluntarily choose to participate in the study or not which may lead to biased results. In the data, there are more STEM majors than non-STEM majors which means our sample may be slightly biased.

6. Supporting documentation:

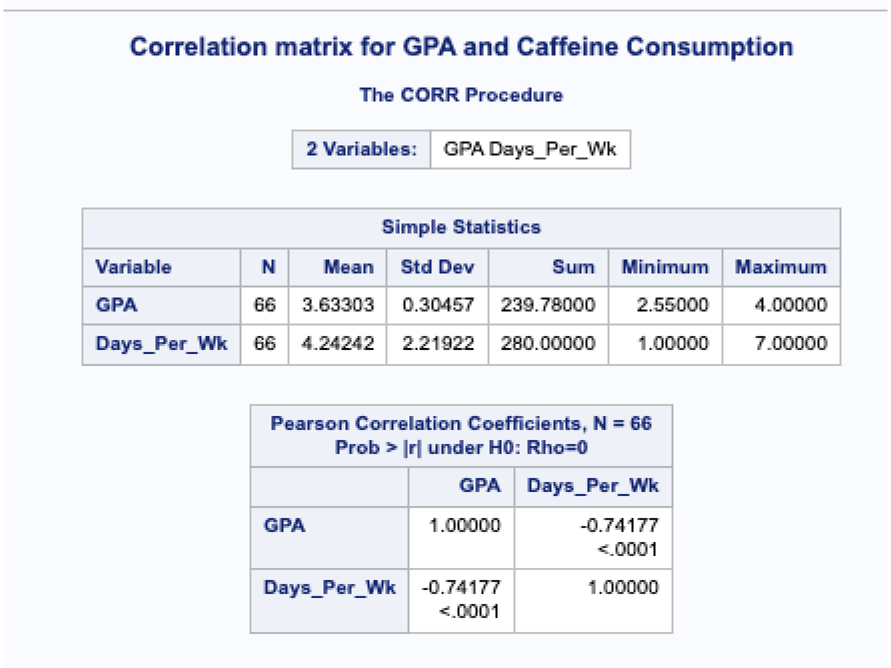
Code:

```
DATA CIMHS;
INFILE '/home/u59339458/sasuser.v94/CIMHSData.csv'
DELIMITER=',';
LENGTH Group $ 20;
INPUT MAJOR $
       GENDER $
       RACE $
       YEAR $
       RESIDENCE $
       GPA
       Days_Per_Wk
       Cups
       Type $
       Size $
       Reason $
       Diagnosis $
       Q1PHQ
       Q2PHQ
       Q3PHQ
       Q4PHQ
       Q5PHQ
       Q6PHQ
       Q7PHQ
       Q8PHQ
       Q9PHQ
       Q10GAD
       Q11GAD
       Q12GAD
       Q13GAD
       Q14GAD
       Q15GAD
       Q16GAD
       Q17PSS
       Q18PSS
       Q19PSS
       Q20PSS
       Q21PSS
       Q22PSS
       Q23PSS
       Q24PSS
       Q25PSS
       Q26PSS
       Source_stress $
       Caff_Symp $
       Caff_Grades $;
IF Size EQ 'S' THEN Caff_Lvl = 180;
ELSE IF Size EQ'M' THEN Caff_Lvl= 260;
ELSE IF Size EQ 'L' THEN Caff_Lvl= 330;
ELSE IF Size EQ 'XL' THEN Caff_Lvl= 415;
ELSE IF Size EQ 'B' THEN Caff_Lvl = 34;
/* To do ANOVA, we categorize caffeine consumption data into
groups*/
IF Days_Per_Wk EQ 1 OR Days_Per_Wk EQ 2 THEN Group= "Rarely";
ELSE IF Days_Per_Wk EQ 3 OR Days_Per_Wk EQ 4 THEN
Group="Sometimes";
ELSE IF Days_Per_Wk EQ 5 OR Days_Per_Wk EQ 6 THEN Group=
"Frequent";
ELSE IF Days_Per_WK EQ 7 THEN Group="Everyday";
/* Calculating the amount of caffeine consumed per week */
Total_Caff_Wk = Cups * Caff_Lvl * Days_Per_Wk;
RUN;

PROC CORR DATA= CIMHS;
TITLE "Correlation matrix for GPA and Caffeine Consumption";
VAR GPA Days_Per_Wk;
RUN;
PROC ANOVA DATA= CIMHS;
TITLE "Analysis of Caffeine Consumption and GPA";
CLASS Group;
MODEL GPA = Group ;
```

```
MEANS Group / TUKEY;
PROC REG DATA= CIMHS;
TITLE "Linear Regression: Effect of Caffeine Consumption on
GPA";
MODEL GPA = Days_Per_Wk;
RUN;
```

Output:



Correlation matrix for GPA and Caffeine Consumption

The CORR Procedure

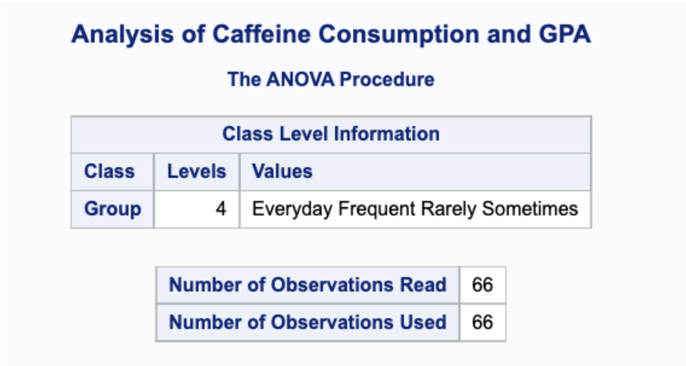
2 Variables:	GPA Days_Per_Wk

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
GPA	66	3.63303	0.30457	239.78000	2.55000	4.00000
Days_Per_Wk	66	4.24242	2.21922	280.00000	1.00000	7.00000

Pearson Correlation Coefficients, N = 66
Prob > |r| under H0: Rho=0

	GPA	Days_Per_Wk
GPA	1.00000	-0.74177 <.0001
Days_Per_Wk	-0.74177 <.0001	1.00000



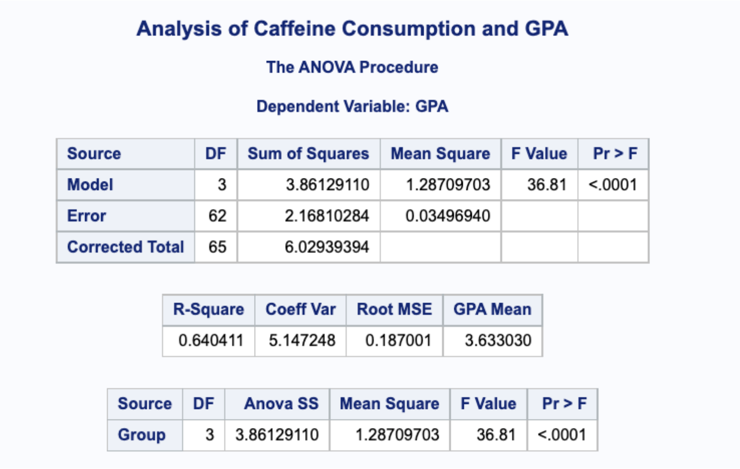
Analysis of Caffeine Consumption and GPA

The ANOVA Procedure

Class Level Information

Class	Levels	Values
Group	4	Everyday Frequent Rarely Sometimes

Number of Observations Read	66
Number of Observations Used	66



Analysis of Caffeine Consumption and GPA

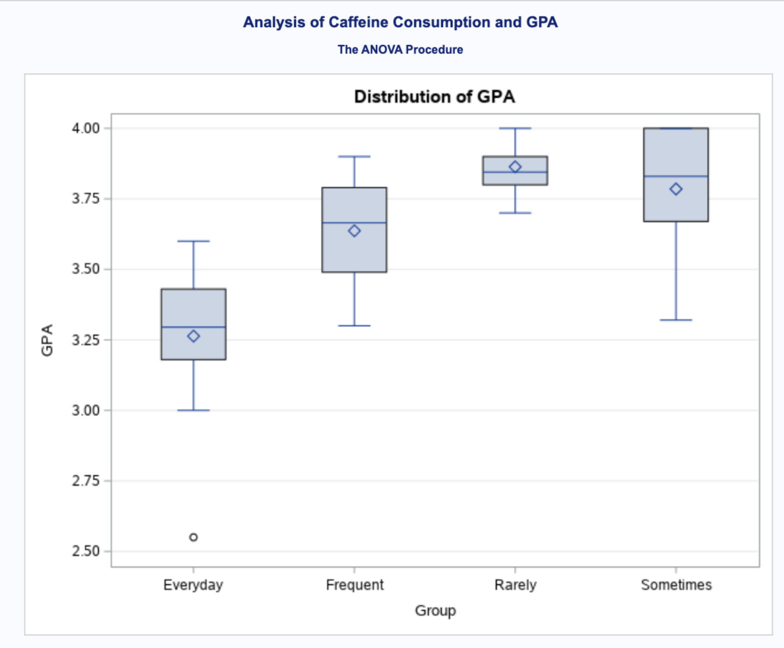
The ANOVA Procedure

Dependent Variable: GPA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3.86129110	1.28709703	36.81	<.0001
Error	62	2.16810284	0.03496940		
Corrected Total	65	6.02939394			

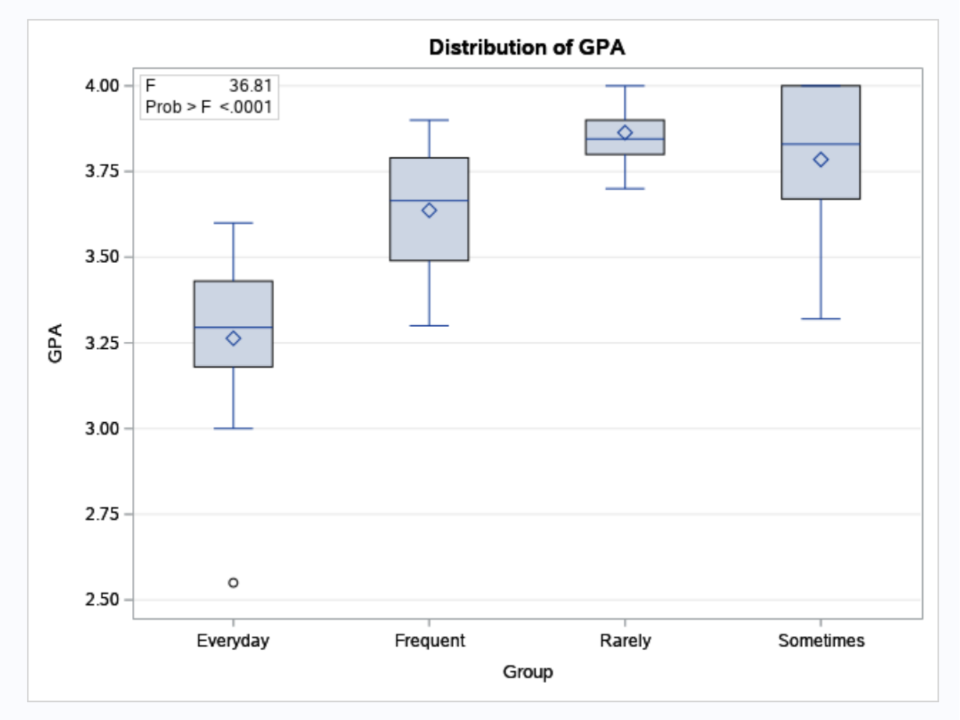
R-Square	Coeff Var	Root MSE	GPA Mean
0.640411	5.147248	0.187001	3.633030

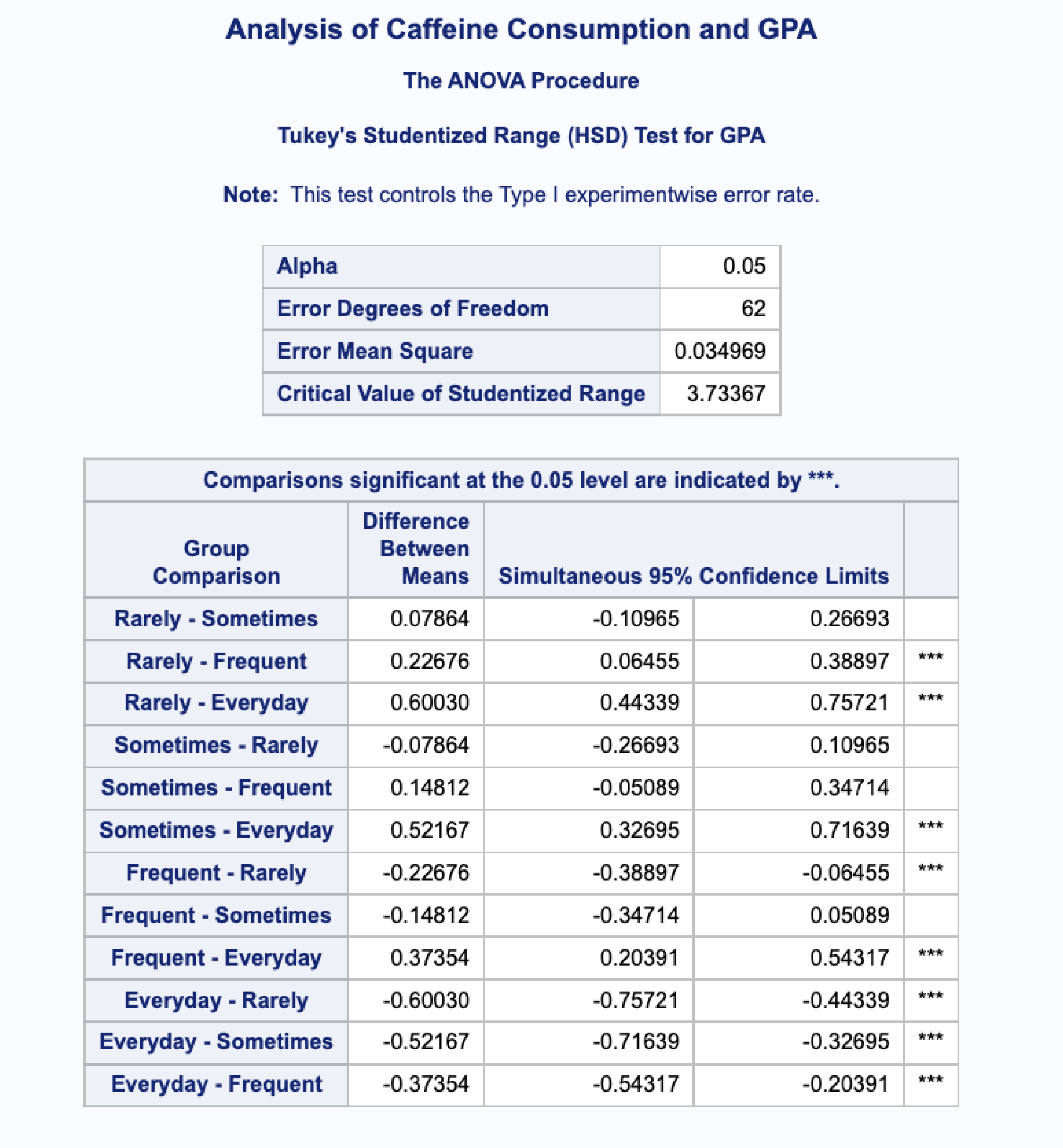
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Group	3	3.86129110	1.28709703	36.81	<.0001



Analysis of Caffeine Consumption and GPA

The ANOVA Procedure


Distribution of GPA
F 36.81
Prob > F <.0001
4.00
3.75
3.50
3.25
3.00
2.75
2.50
GPA
Everyday
Frequent
Rarely
Sometimes
Group



Analysis of Caffeine Consumption and GPA

The ANOVA Procedure

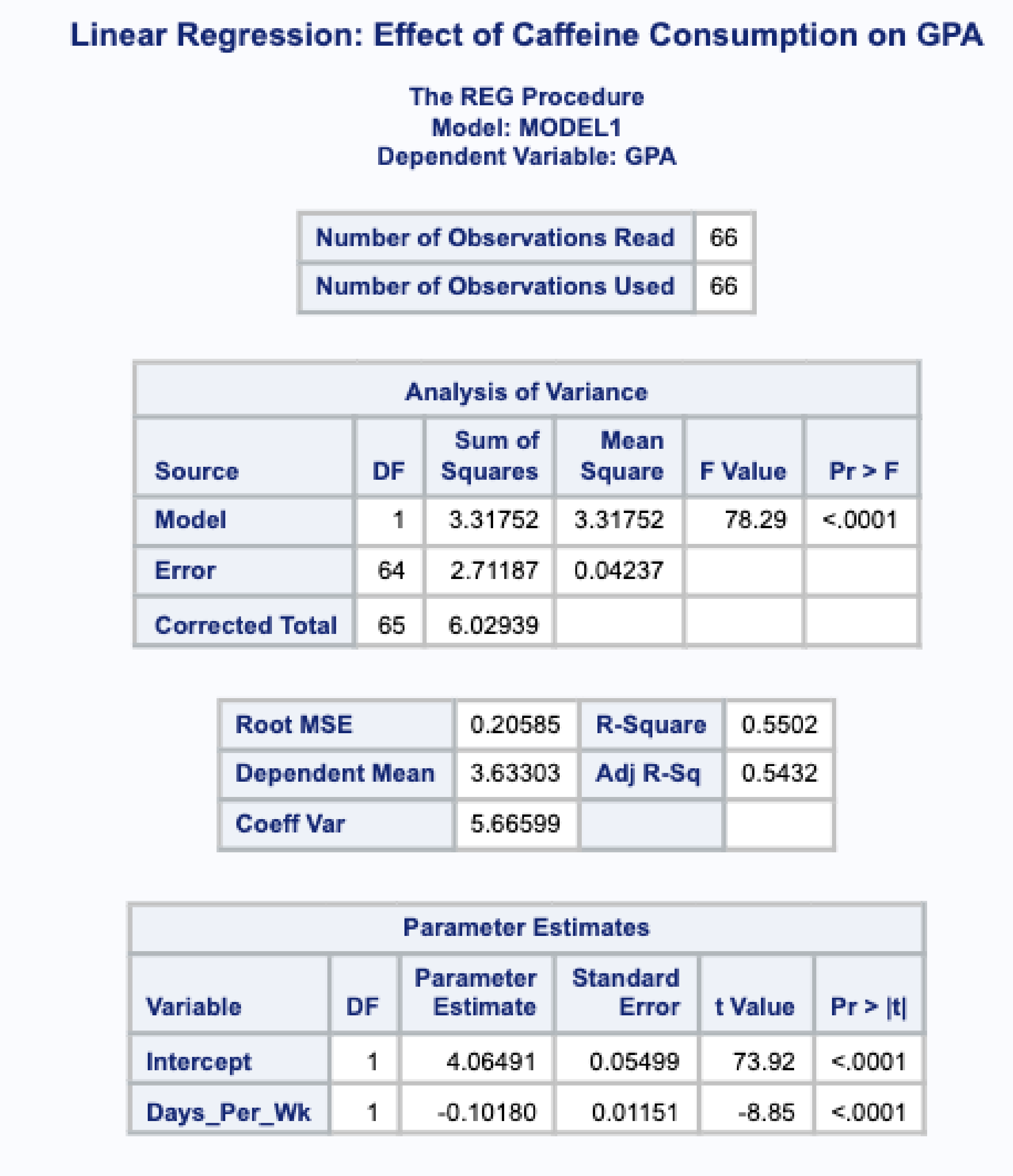
Tukey's Studentized Range (HSD) Test for GPA

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	62
Error Mean Square	0.034969
Critical Value of Studentized Range	3.73367

Comparisons significant at the 0.05 level are indicated by *.**

Group Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Rarely - Sometimes	0.07864	-0.10965	0.26693	
Rarely - Frequent	0.22676	0.06455	0.38897	***
Rarely - Everyday	0.60030	0.44339	0.75721	***
Sometimes - Rarely	-0.07864	-0.26693	0.10965	
Sometimes - Frequent	0.14812	-0.05089	0.34714	
Sometimes - Everyday	0.52167	0.32695	0.71639	***
Frequent - Rarely	-0.22676	-0.38897	-0.06455	***
Frequent - Sometimes	-0.14812	-0.34714	0.05089	
Frequent - Everyday	0.37354	0.20391	0.54317	***
Everyday - Rarely	-0.60030	-0.75721	-0.44339	***
Everyday - Sometimes	-0.52167	-0.71639	-0.32695	***
Everyday - Frequent	-0.37354	-0.54317	-0.20391	***



Linear Regression: Effect of Caffeine Consumption on GPA

The REG Procedure
Model: MODEL1
Dependent Variable: GPA

Number of Observations Read	66
Number of Observations Used	66

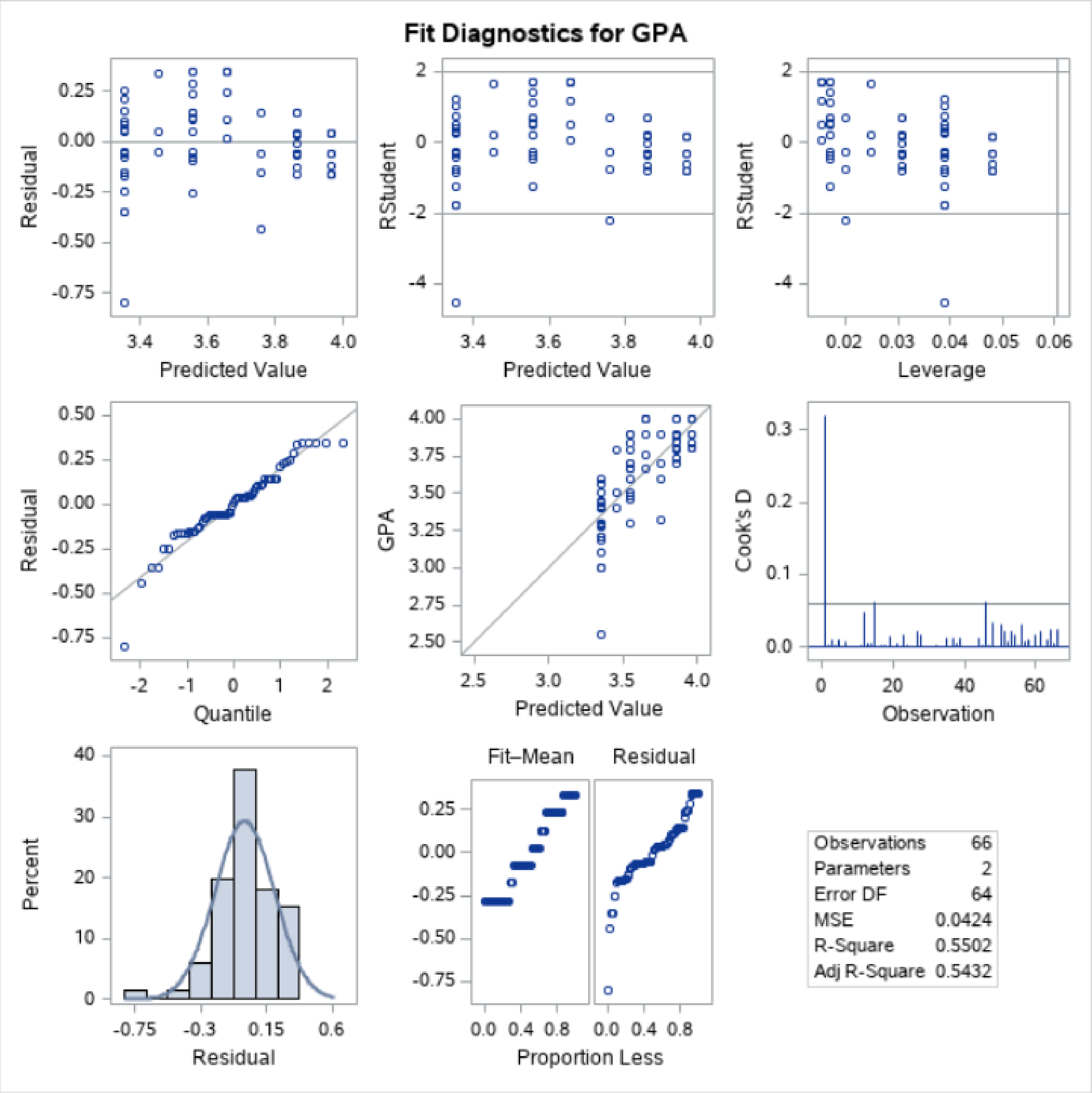
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.31752	3.31752	78.29	<.0001
Error	64	2.71187	0.04237		
Corrected Total	65	6.02939			

Root MSE	0.20585	R-Square	0.5502
Dependent Mean	3.63303	Adj R-Sq	0.5432
Coeff Var	5.66599		

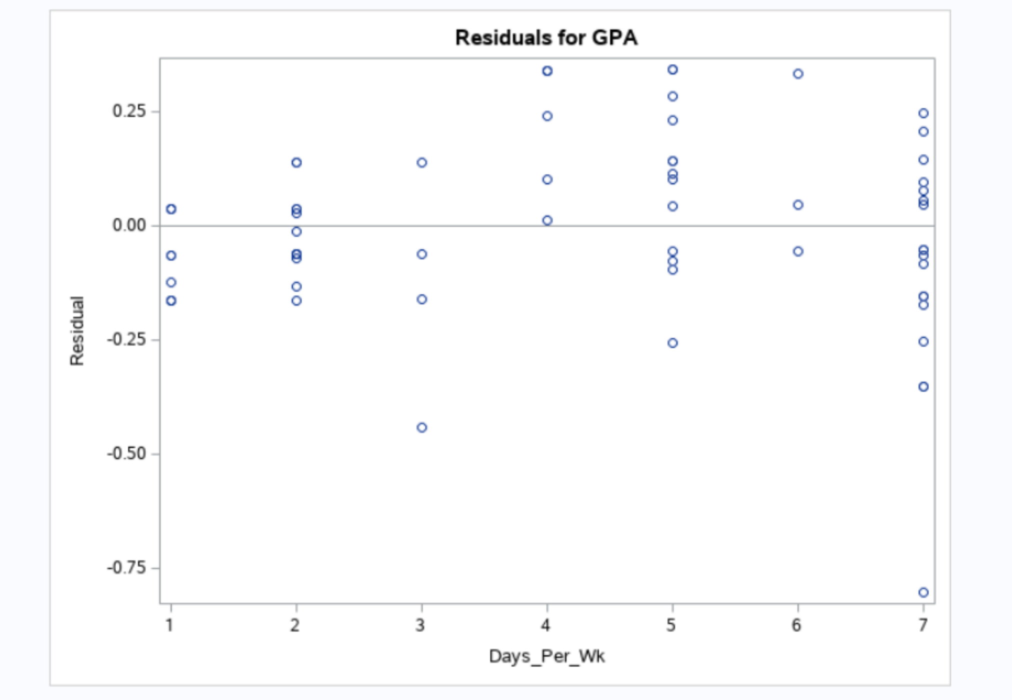
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.06491	0.05499	73.92	<.0001
Days_Per_Wk	1	-0.10180	0.01151	-8.85	<.0001

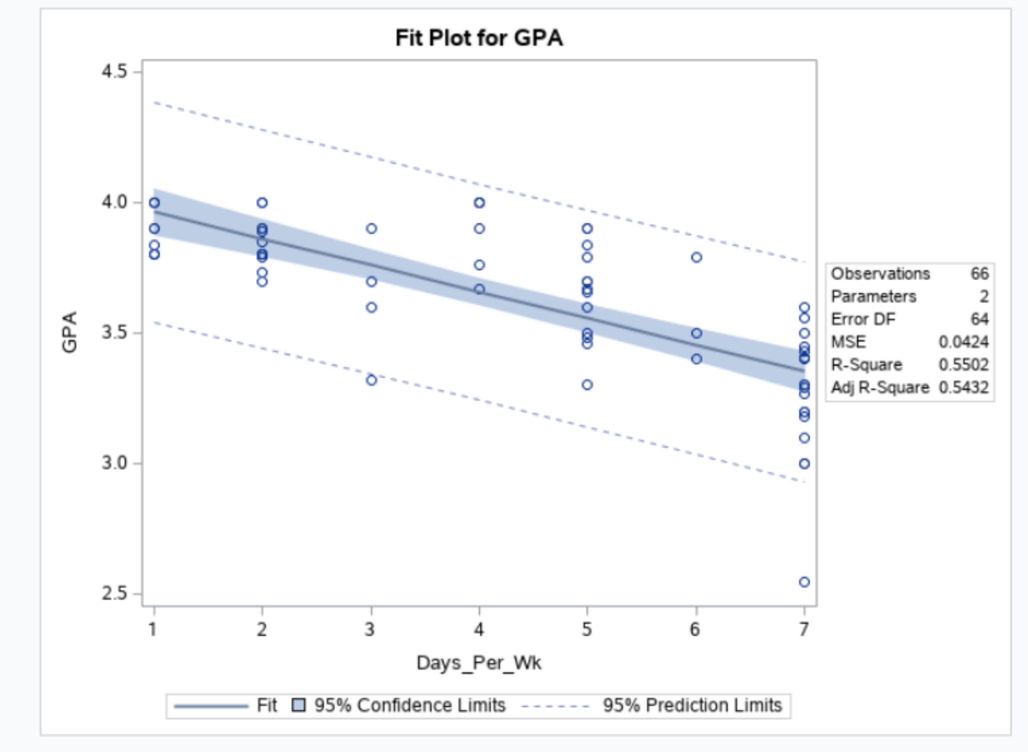
Linear Regression: Effect of Caffeine Consumption on GPA

The REG Procedure
Model: MODEL1
Dependent Variable: GPA



Observations	66
Parameters	2
Error DF	64
MSE	0.0424
R-Square	0.5502
Adj R-Square	0.5432


Residuals for GPA
Residual
0.25
0.00
-0.25
-0.50
-0.75
1
2
3
4
5
6
7
Days_Per_Wk


Fit Plot for GPA
GPA
4.5
4.0
3.5
3.0
2.5
1
2
3
4
5
6
7
Days_Per_Wk
Observations 66
Parameters 2
Error DF 64
MSE 0.0424
R-Square 0.5502
Adj R-Square 0.5432
Fit
95% Confidence Limits
95% Prediction Limits