Lizzie Kahle

Appendix Outline: Syntax Follows Same Order

I. **Table 1: ANOVA & Post Hoc table for literacy score vs. country**

II. **Descriptive statistics tables**
1. **Table 2: B1 (level of measurement, values, codebook)**
2. **Table 3: B2 (descriptive statistics of outcome)**
3. **Table 4: B3 (descriptive statistics of predictor)**
4. **Figure 1: Frequency table (table 5) for participant responses to each of the 6 attitude questions**
    a) **C1: Figure 1: Bar graphs representing the frequencies**

III. **Table 6: Bivariate test results table**

IV. **Simple Linear and Multiple Regression**
- **B. Table 7: Correlation results table**
- **C. Figure 2: Correlation plot**
- **D. Table 8: Model 1 results**
- **E. Figure 3: Simple linear Regression plot**
- **F. Table 9: Multiple regression results table (models 1 + 2)**

Rationale for subsetting sample to only comparing U.S. and England:

```
> # install.packages("car")
>  library(car)
> Rename the variables for convenience
> names(PIRLS)[names(PIRLS) == "asrlit01"] <- "literacy_score"
> names(PIRLS)[names(PIRLS) == "idcntry"] <- "country_origin"
> # Convert country_origin to a factor with meaningful levels
> PIRLS$country_origin <- factor(PIRLS$country_origin, levels = c(170, 578, 926,
840),
+                                labels = c("Columbia", "Norway", "England",
"U.S."))
Conduct ANOVA
> anova_model <- aov(literacy_score ~ country_origin, data = PIRLS)
> summary(anova_model)
                 Df   Sum Sq Mean Sq F value Pr(>F)
country_origin    3  7848683 2616228     371 <2e-16 ***
Residuals      2875 20273332    7052
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> posthoc <- TukeyHSD(anova_model)
> print(posthoc)
  Tukey multiple comparisons of means
    95% family-wise confidence level
> Fit: aov(formula = literacy_score ~ country_origin, data = PIRLS)
```

```
$country_origin
                       diff       lwr        upr     p adj
Norway-Columbia    81.227358  69.72341  92.731302 0.0000000
England-Columbia 132.243575 120.81590 143.671246 0.0000000
U.S.-Columbia    125.760920 114.29557 137.226269 0.0000000
England-Norway    51.016216  39.71666  62.315774 0.0000000
U.S.-Norway       44.533562  33.19590  55.871222 0.0000000
U.S.-England      -6.482655 -17.74292   4.777607 0.4498208
```

Subset data:

```
# Assuming your dataset is named PIRLS
>
> # Subset the dataset to include only U.S. and England data
> subset_data_PIRLS <- subset(PIRLS, country_origin %in% c("U.S.", "England"))
>
> # Check the first few rows of the subsetted dataset
> head(subset_data_PIRLS)
```

Visualizations of attitude scale questions responses

```
# Stacked bar
plot_data <- data.frame(
Grade = factor(rep(1:4, 6)),
Frequency = c(397, 217, 684, 191, 368, 452, 399, 230, 230, 364, 196, 257, 465,
419, 171, 765, 1010, 278, 84, 85, 757, 361, 150, 193),
  Variable = rep(c("asbgrst1", "asbgrst2", "asbgrst3", "asbgrst4", "asbgrst5",
"asbgrst6"), each = 4)
)

# View the created data frame
View(plot_data)
```

Individual bar graphs

```
# Reorder the levels of the variables
subset_data_PIRLS$asbgrst1 <- factor(subset_data_PIRLS$asbgrst1, levels = c(1, 2,
3, 4))
subset_data_PIRLS$asbgrst2 <- factor(subset_data_PIRLS$asbgrst2, levels = c(1, 2,
3, 4))
subset_data_PIRLS$asbgrst3 <- factor(subset_data_PIRLS$asbgrst3, levels = c(1, 2,
3, 4))
subset_data_PIRLS$asbgrst4 <- factor(subset_data_PIRLS$asbgrst4, levels = c(1, 2,
3, 4))
subset_data_PIRLS$asbgrst5 <- factor(subset_data_PIRLS$asbgrst5, levels = c(1, 2,
3, 4))
subset_data_PIRLS$asbgrst6 <- factor(subset_data_PIRLS$asbgrst6, levels = c(1, 2,
3, 4))

# Create tables for each variable
table_asbgrst1 <- table(subset_data_PIRLS$asbgrst1)
table_asbgrst2 <- table(subset_data_PIRLS$asbgrst2)
table_asbgrst3 <- table(subset_data_PIRLS$asbgrst3)
table_asbgrst4 <- table(subset_data_PIRLS$asbgrst4)
table_asbgrst5 <- table(subset_data_PIRLS$asbgrst5)
table_asbgrst6 <- table(subset_data_PIRLS$asbgrst6)
```

```
# Plot the tables
par(mfrow = c(2, 3)) # Arrange plots in a 2x3 grid
barplot(table_asbgrst1, main = "asbgrst1")
barplot(table_asbgrst2, main = "asbgrst2")
barplot(table_asbgrst3, main = "asbgrst3")
barplot(table_asbgrst4, main = "asbgrst4")
barplot(table_asbgrst5, main = "asbgrst5")
barplot(table_asbgrst6, main = "asbgrst6")
```

Syntax

One-way ANOVA of # of books across the three levels of attitude

Bivariate statistics

```
> # Perform ANOVA with "books_at_home" as the dependent variable and
"reading_attitudes" as the independent variable
> # Convert numeric levels to categorical levels with labels
> subset_data_PIRLS$reading_attitudes <-
factor(subset_data_PIRLS$reading_attitudes,
+                                                levels = c(1, 2, 3),
+                                                labels = c("high", "medium",
"low"))
> # Check the levels of the factor variable
> levels(subset_data_PIRLS$reading_attitudes)
[1] "high"   "medium" "low"
> anova_result <- aov(books_at_home ~ reading_attitudes, data = subset_data_PIRLS)
> # View ANOVA summary
> summary(anova_result)
                    Df   Sum Sq Mean Sq F value   Pr(>F)
reading_attitudes    2   217872  108936    13.7 1.28e-06 ***
Residuals         1449 11523639    7953
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
18 observations deleted due to missingness
>
>  # Conduct post hoc test (Tukey HSD)
> posthoc <- TukeyHSD(anova_result)
> # View the results of the post hoc test
> print(posthoc)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = books_at_home ~ reading_attitudes, data = subset_data_PIRLS)

$reading_attitudes
                  diff       lwr       upr     p adj
medium-high -23.418664 -35.19785 -11.63948 0.0000101
low-high    -28.244117 -45.56740 -10.92084 0.0004003
low-medium   -4.825452 -22.12968  12.47878 0.789992
```

Table 7 # Correlation Analysis

```
correlation <- cor.test(subset_data$book2, subset_data$literacy_score, method =
"pearson")
>
> cor_results <- data.frame(corr = correlation$estimate, pvalue =
correlation$p.value,
+                           sampsize = length(subset_data$literacy_score))
>
> kable(cor_results, digits = 3, row.names = F)
```

›rr	·alue	ımpsize
cor 2281858	113248e-19	70

```
> View(cor_results)
```

Simple Linear Regression and Multiple Regression

```
> # Model 1 - Simple Linear Regression

> model1 <- lm(literacy_score ~ book2, data = subset_data)

> # Model 2 - Multiple Linear Regression

> model2 <- lm(literacy_score ~ book2 + country_origin + attitudes_read, data =
subset_data)

> # Regression Table

> stargazer(model1, model2, title="Results", header = FALSE, type = 'html')

> summary(lm_model)
```

Model 1

```
Call:> summary(lm_model)
lm(formula = literacy_score ~ book2, data = subset_data)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-296.331  -53.764    5.766   58.825  240.703

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 531.33110    3.61045  147.16   <2e-16 ***
book2         0.22578    0.02514    8.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 86.48 on 1468 degrees of freedom
Multiple R-squared:  0.05207,    Adjusted R-squared:  0.05142
F-statistic: 80.64 on 1 and 1468 DF,  p-value: < 2.2e-16

Multiple Regression
> print(model2)
> lm(formula = literacy_score ~ book2 + country_origin + attitudes_read,
    data = subset_data)

Coefficients:
       (Intercept)               book2  country_originU.S.      attitudes_read
          591.6091              0.1958             -3.1711            -32.4334
```

Model 2

```
> summary(model2)
Call:
lm(formula = literacy_score ~ book2 + country_origin + attitudes_read,
    data = subset_data)

Residuals:
     Min       1Q   Median       3Q      Max
-288.571  -49.795    5.924   58.551  229.510

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        591.60913    7.16871  82.527  < 2e-16 ***
book2                0.19579    0.02469   7.931 4.31e-15 ***
country_originU.S.  -3.17112    4.40499  -0.720    0.472
attitudes_read     -32.43337    3.22475 -10.058  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 83.75 on 1448 degrees of freedom
Multiple R-squared:  0.1157,     Adjusted R-squared:  0.1138
F-statistic: 63.13 on 3 and 1448 DF,  p-value: < 2.2e-16
```

Scatterplot Linear regression model

```
> lm_model <- lm(literacy_score ~ book2, data = subset_data)
>
> # Compute the R2
> r_squared <- summary(lm_model)$r.squared
>
```

```
> # Calculate standard error of the estimate
> se_estimate <- summary(lm_model)$sigma
>
> # Calculate predicted values from the regression model
> subset_data$predicted <- predict(lm_model)
>
> # Calculate lines 2 SD above and below the regression line
> subset_data$upper_line <- subset_data$predicted + 2 * se_estimate
> subset_data$lower_line <- subset_data$predicted - 2 * se_estimate
>
> # Scatter plot with regression line and 95% confidence interval
> library(ggplot2)
> ggplot(subset_data, aes(x = book2, y = literacy_score)) +
+     geom_point() + # Adding points
+     geom_smooth(method = "lm", se = FALSE) + # Adding regression line with 95%
CI
+     geom_line(aes(y = upper_line), color = "red", linetype = "dashed") +
+     geom_line(aes(y = lower_line), color = "red", linetype = "dashed") +
+     labs(x = "Number of Books at Home",
+          y = "Score on Literacy Section of PIRLS",
+          title = "Scatter Plot of Books at Home and 4th Grade Literacy Scores,
U.S. and England") +
+     annotate("text", x = max(subset_data$book2),
+              y = min(subset_data$literacy_score),
+              label = paste("r² =", round(r_squared, 2)), hjust = 1) +
+     theme_minimal() # Optional: Applying a minimal theme
`geom_smooth()` using formula = 'y ~ x'
```

Model 1 vs. Model 2

```
summary(anova_result)
                                  Df   Sum Sq Mean Sq F value   Pr(>F)
country_origin                     1    44186   44186   5.588   0.0182 *
reading_attitudes                  2   210777  105389  13.327 1.84e-06 ***
country_origin:reading_attitudes   2    51750   25875   3.272   0.0382 *
Residuals                       1446 11434797    7908
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
18 observations deleted due to missingness
```

Syntax for residual plots

```
> # Create a histogram of the residuals
> hist(residuals_model2,
+      main = "Histogram of Residuals in Multiple Regression",
+      xlab = "Residuals",
+      ylab = "Frequency",
+      breaks = 50  # Adjust the number of breaks as needed
+ )
>
> qqnorm(subset_data$residuals)
> plot(fitted(model2), subset_data$residuals, main = "Residuals vs Fitted",
+      xlab = "Fitted values", ylab = "Residuals")
> res.std <- rstandard(model2)
>
> plot(res.std, ylab="Standardized Residual", ylim=c(-4,4)) +
```

```
+     abline(h =c(-2,0,2), lty = 1, col = "blue") +
+     abline(h =c(-3,0,3), lty = 1, col = "red")
integer(0)
> outlierTest(model2)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
25 -3.462877         0.00055001      0.79862
```

Additional Syntax

Bivariate test: Two way anova with predictor and both controls

```
> leveneTest(books_at_home ~ country_origin * reading_attitudes, data =
subset_data_PIRLS, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
        Df F value Pr(>F)
group    5  0.3934 0.8536
      1446

anova_result <- aov(books_at_home ~ country_origin * reading_attitudes, data =
subset_data_PIRLS)
> summary(anova_result)
                                  Df   Sum Sq Mean Sq F value   Pr(>F)
country_origin                     1    44186   44186   5.588   0.0182 *
reading_attitudes                  2   210777  105389  13.327 1.84e-06 ***
country_origin:reading_attitudes   2    51750   25875   3.272   0.0382 *
Residuals                       1446 11434797    7908
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
18 observations deleted due to missingness

anova_result <- aov(books_at_home ~ country_origin * reading_attitudes, data =
subset_data_PIRLS)
> View(anova_result)
> summary(anova_result)
                                  Df   Sum Sq Mean Sq F value   Pr(>F)
country_origin                     1    44186   44186   5.588   0.0182 *
reading_attitudes                  2   210777  105389  13.327 1.84e-06 ***
country_origin:reading_attitudes   2    51750   25875   3.272   0.0382 *
Residuals                       1446 11434797    7908
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
18 observations deleted due to missingness
> Tukey_Results <- TukeyHSD(anova_result)
> Tukey_Results$reading_attitudes
```

```
                 diff       lwr       upr        p adj
medium-high -22.84119 -34.58708 -11.09531 1.635377e-05
low-high    -28.14258 -45.41689 -10.86828 4.050370e-04
low-medium   -5.30139 -22.55670  11.95392 7.512036e-01
> Tukey_Results$country_origin
                  diff       lwr       upr      p adj
U.S.-England -11.03377 -20.19011 -1.877429 0.01821937
> library(DescTools)
> scheffe_results <- ScheffeTest(anova_result)
> scheffe_results$reading_attitudes
                 diff    lwr.ci    upr.ci         pval
medium-high -22.84119 -39.52230 -6.160085 0.0009278649
low-high    -28.14258 -52.67496 -3.610202 0.0124634493
low-medium   -5.30139 -29.80679 19.204012 0.9913724829

 scheffe_results$`country_origin:reading_attitudes`
                                    diff    lwr.ci      upr.ci         pval
U.S.:high-England:high      -23.1833664 -46.87344   0.5067040 0.0598288462
England:medium-England:high -35.5878750 -59.11153 -12.0642211 0.0001267038
U.S.:medium-England:high    -33.1381926 -56.13500 -10.1413835 0.0003537629
England:low-England:high    -34.8361998 -68.75354  -0.9188576 0.0395073916
U.S.:low-England:high       -43.8104478 -78.85817  -8.7627279 0.0040300507
England:medium-U.S.:high    -12.4045086 -36.63922  11.8302068 0.7140828077
U.S.:medium-U.S.:high        -9.9548262 -33.67849  13.7688406 0.8552788341
England:low-U.S.:high       -11.6528334 -46.06715  22.7614839 0.9376276415
U.S.:low-U.S.:high          -20.6270814 -56.15597  14.9018055 0.5872841595
U.S.:medium-England:medium    2.4496824 -21.10781  26.0071701 0>.9997449187
England:low-England:medium    0.7516752 -33.54830  35.0516474 0.9999998895
U.S.:low-England:medium      -8.2225728 -43.64071  27.1955675 0.9880567628
England:low-U.S.:medium      -1.6980073 -35.63882  32.2428094 0.9999932051
U.S.:low-U.S.:medium        -10.6722552 -45.74269  24.3981825 0.9602146845
U.S.:low-England:low         -8.9742480 -52.00294  34.0544456 0.9927212519
> scheffe_results$country_origin
                  diff    lwr.ci   upr.ci      pval
U.S.-England -11.03377 -26.58625 4.518713 0.3489691
```