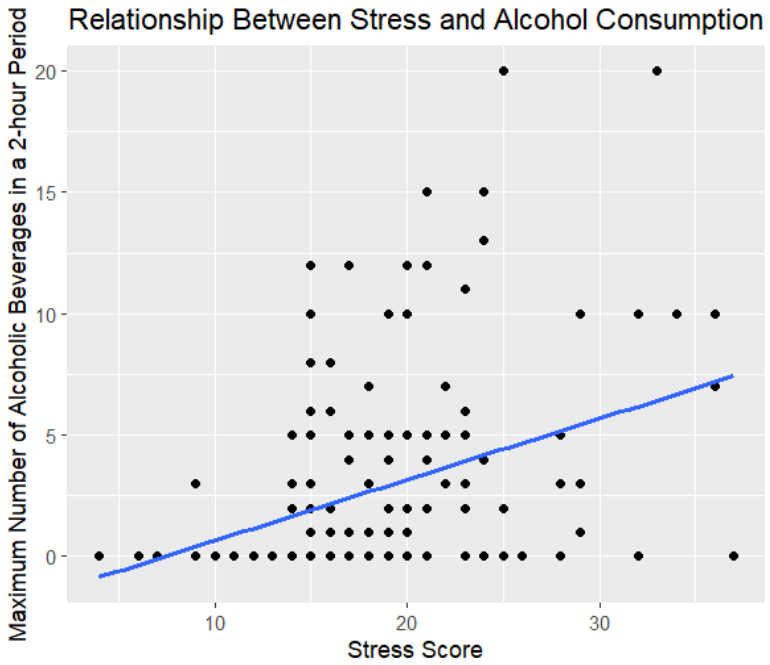
Data Analysis - Correlation and Linear Regression

Student: Kevin Ivanov

The Assignment

1) Construct a scatterplot with a linear regression line. Use ggplot2.



2) Calculate the correlation coefficient. Interpret the correlation coefficient in terms of the strength and direction of the linear relation.

[1] 0.3263685

The correlation between the Stress Score of students and their self-reported Maximum Alcohol consumption is r = 0.326, which indicates a very low to moderate positive linear relationship.

3) Conduct the linear regression analysis.

```
Call:
lm(formula = Stress.Score ~ Maximum.Alcohol)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0504  -3.1653  -0.1115   2.6763  18.9496

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     18.05037    0.52871  34.141  < 2e-16 ***
Maximum.Alcohol  0.42444    0.09971   4.257 3.61e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.421 on 152 degrees of freedom
Multiple R-squared:  0.1065,    Adjusted R-squared:  0.1006
F-statistic: 18.12 on 1 and 152 DF,  p-value: 3.615e-05
```

4) Reading from the output, what is the equation of the linear regression line for this data? State the answer as an equation.

Interpretation Output for the Regression Line

The linear regression model: Stress Score = 18.0504 + 0.4244 (Maximum Alcohol)
=18.0504+0.4244x

5) What is the predicted number of maximum alcoholic beverages consumed for a student who has a stress score of 24? Round answers to the nearest unit. Use R for the computation. (Use R as a calculator)

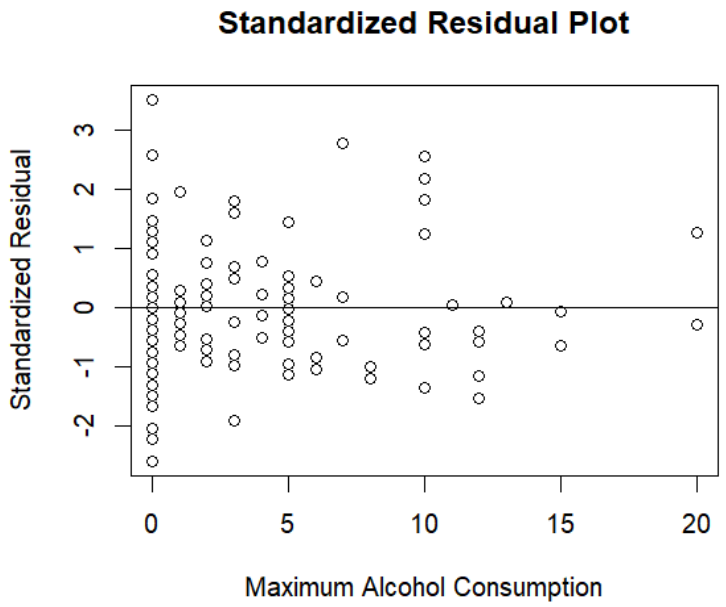
Answer: 28

```
> intercept <- 18.05037
> slope <- 0.42444
>
> stress_score <- 24
>
> predicted_alcohol <- intercept + slope * stress_score
>
> rounded_predicted_alcohol <- round(predicted_alcohol)
> rounded_predicted_alcohol
[1] 28
```

6) What percent of the variation of the response variable is explained by a linear relationship between the given explanatory and response variables?

R^2=0.1065 = Approximately 10.65% of the variation in Stress Score consumption can be explained by the linear relationship with the Maximum Alcohol

7) Construct the standardize residual plot for this analysis.



8) Approximately how many residuals are more than 2 standard deviations away from the linear regression line?

About 8

Code

7.8.2 Scatterplot with Linear Regression Line

```
Scatterplot <- ggplot(LabData, aes(x = Stress.Score, y = Maximum.Alcohol)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +  labs(
    x = "Stress Score",
    y = "Max Alcoholic Drinks in 2 Hours",
    title = "Relationship Between Stress and Alcohol Consumption"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
```

10.1 Correlation

```
cor(Stress.Score, Maximum.Alcohol, use = "complete.obs")
```

[1] 0.3263685.

10.2 LINEAR REGRESSION AND COEFFICIENT OF DETERMINATION

```
> summary(lm(Stress.Score ~Maximum.Alcohol))
```

10.3 STANDARDIZED RESIDUAL PLOT

```
Std.Resid <- rstandard(lm(Stress.Score ~ Maximum.Alcohol))
> plot(Maximum.Alcohol, Std.Resid, main = "Standardized Residual Plot",
+      xlab = "Maximum Alcohol Consumption", ylab = "Standardized Residual") +
+     abline(0, 0)
```