**Statistics Final Project**

For the midterm you provided one variable from either the states or countries data set. You then chose another variable that one of your classmates had provided and engaged in some analysis. The main question for the midterm was to explore how (and to what extent) YOUR variable (or a variable derived from it) could be used to predict the variable of your classmate. In this new part of your overall project we want you do three MORE things.

1. Extending your two-variable regression analysis to consider

a) Whether there is strong evidence that there is a non-zero correlation between your two variables, and

b) What a predicted response variable would be for a given value of the explanatory variable. (I want to see both a prediction interval and a confidence interval as well as an explanation for the difference IN CONTEXT)

1. Split YOUR variable into at least two samples and compare them using both a hypothesis test and a confidence interval for the difference in means for that one variable between the two samples. In splitting up the observations you can use a category provided such as region or come up with a new meaningful way to split the states or countries.

If you wish to redo your regression analysis from the midterm, please choose a DIFFERENT response variable than you originally chose for the midterm.

Note the rubric on the next page is the same as for the midterm except for the two new added categories of confidence intervals and hypothesis testing.

That means, there is a zero chance that

Final Project Rubric

Name: Total Grade:

Data Gathering

a) Discusses any biases in the data set

b) Discusses ways to improve the data gathering methods

c) Discusses how the data gathering methods might affect the analysis of and conclusions based on the data set

Describing Data

a) Displays data and chooses summary statistics appropriately

b) Describes data displays appropriately

c) Interprets data displays and summary statistics appropriately

Linear Regression

a) Checks appropriate conditions for linear regression

b) Comments on the strength of linear associations appropriately

c) Interprets and uses linear regression models appropriately

d) Deals with outliers and influential points appropriately (where applicable)

e) Works with subsets of data appropriately (where applicable)

f) Straightens relationships appropriately (where applicable)

Confidence Intervals

1. Checks appropriate sampling distribution conditions.
2. Creates a confidence interval for a statistic appropriately.
3. Interprets the confidence interval appropriately (including margin of error).
4. Analyzes confidence interval appropriately (determine sample size required or confidence level used (where applicable)

Hypothesis Testing

1. Checks appropriate sampling distribution conditions
2. Creates appropriate null and alternative hypotheses
3. Employs appropriate tests for the given hypotheses (may use confidence intervals)
4. Interprets the results of the tests appropriately
5. Discusses possible errors appropriately

Data Analysis Habits

a) Creates and follows his/her own questions when investigating a data set

b) Explores multiple variables (or subsets of variables) for relationships

c) Considers evidence for and against a claim (avoids cherry picking)

d) Considers appropriate units for data (for example thinking about whether a variable should be per capita)

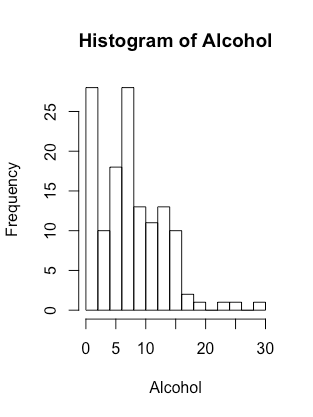
Communication Habits

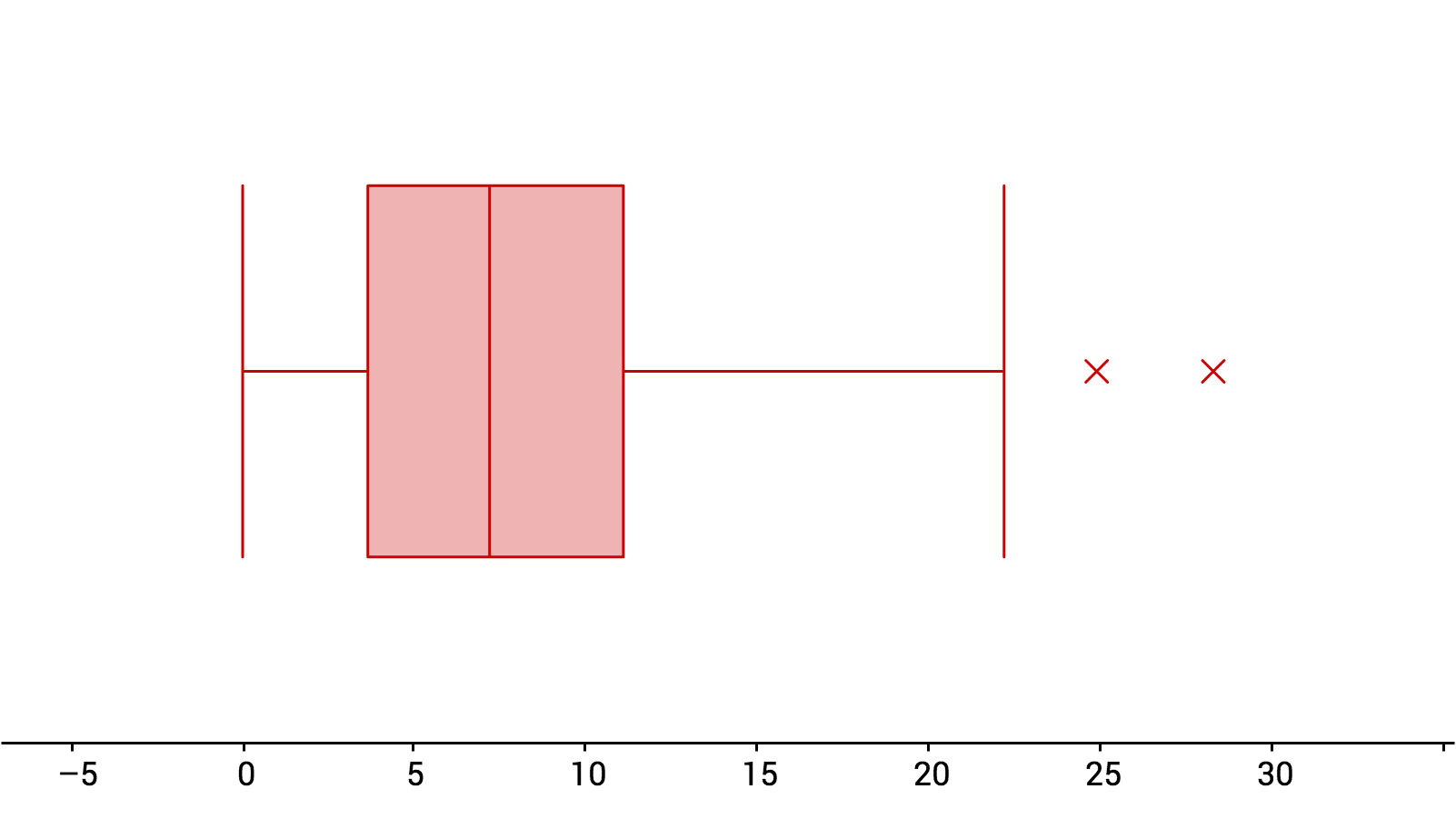
a) The expression of ideas is engaging and makes it enjoyable to read

b) The integration of visual aids and discussion is well balanced

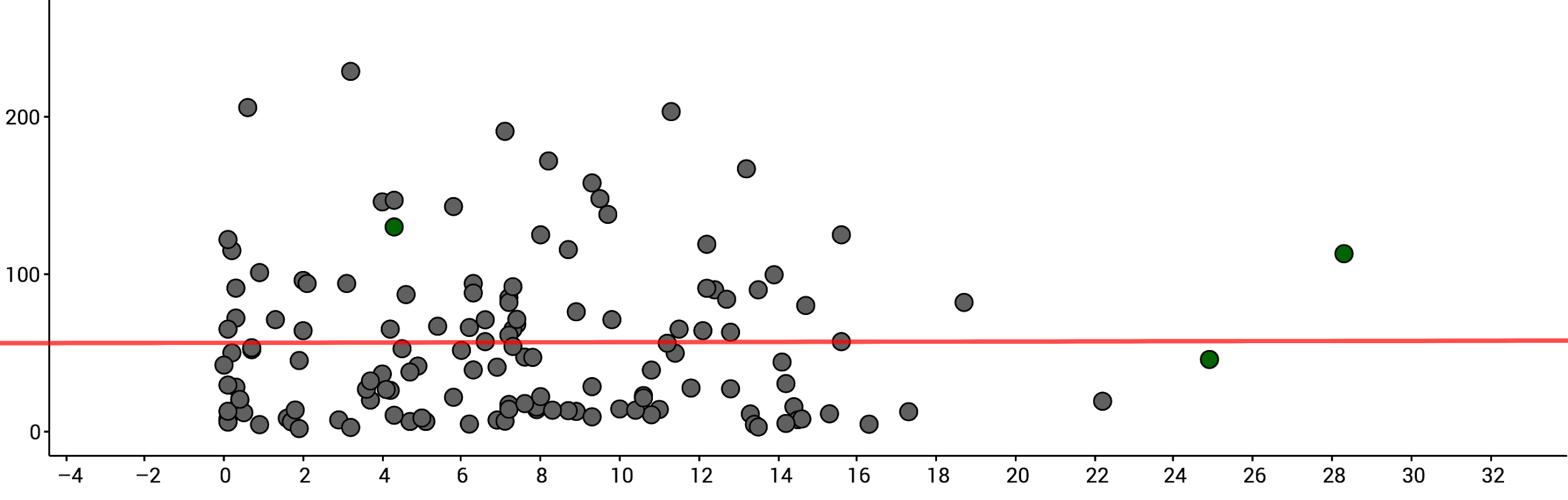
c) The expression of ideas is succinct yet thorough

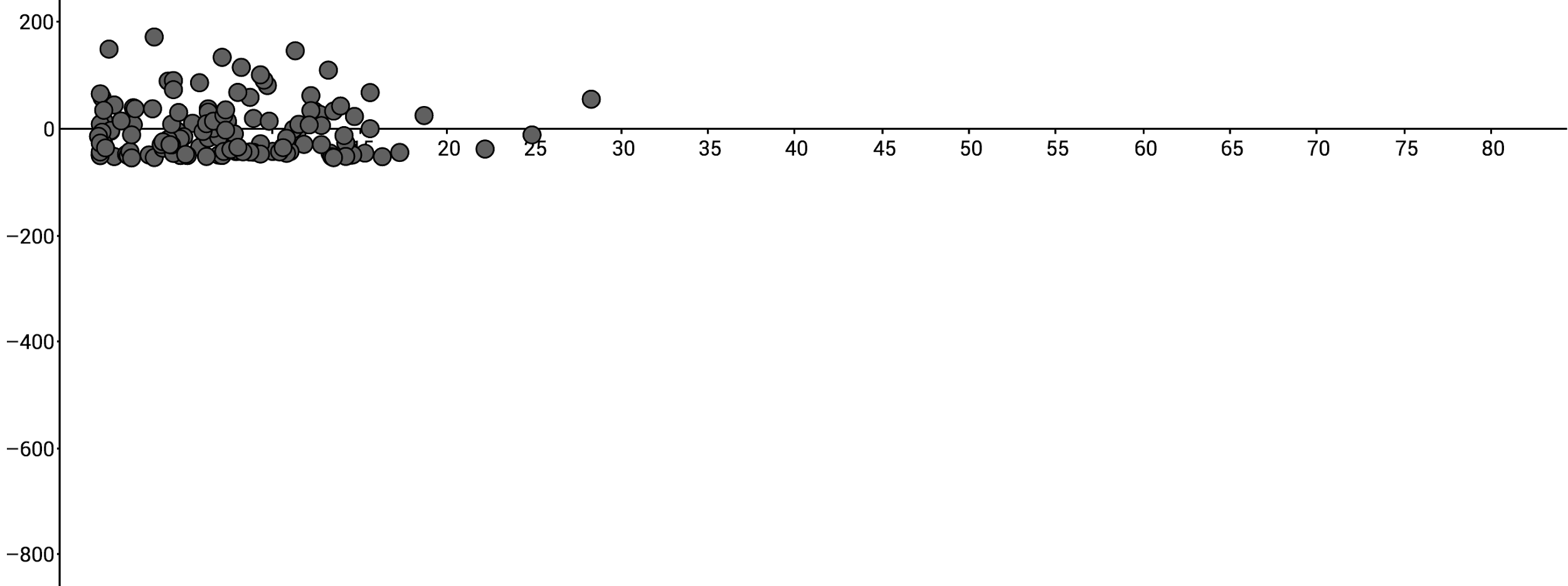
d) The organization of the written work helps to communicate ideas

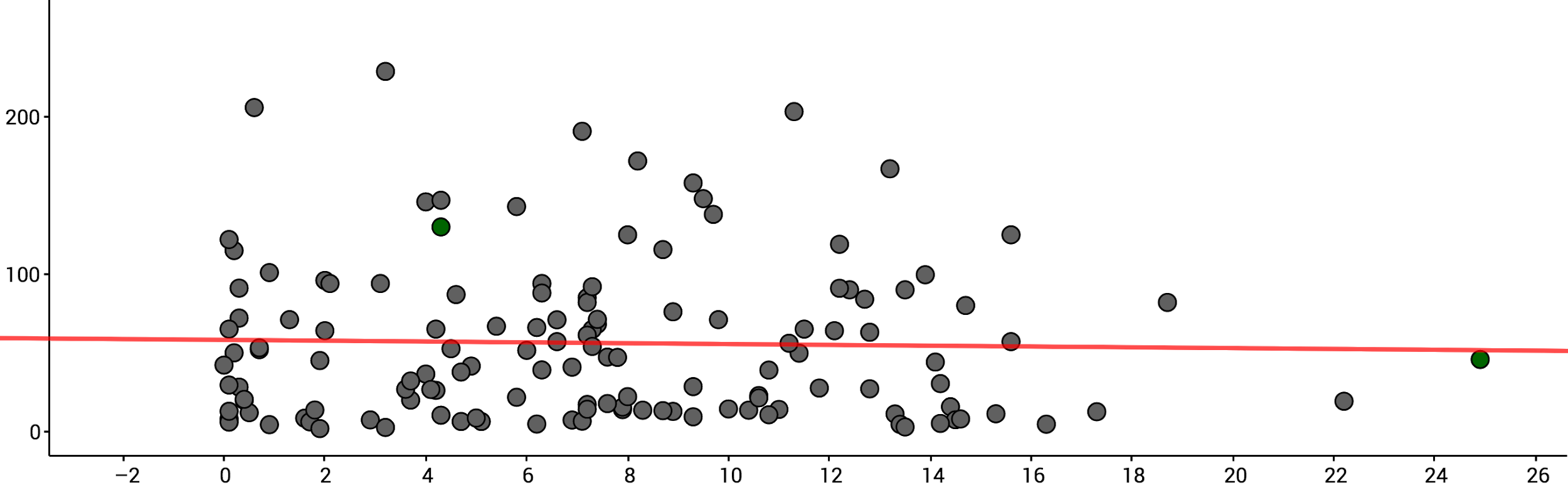
**Data Gathering**  
 15-19 years old total alcohol per capita consumption Data on alcohol consumption among young people stem mainly from surveillance systems focusing on different health behaviors, including alcohol consumption. Surveys are often conducted in schools because they provide a means to sample a large youth population in an affordable manner. Data is also weighted by population size in each country.   
 Many high-income countries participate in school surveys, data on alcohol consumption among young people from low-income countries are scarce. Surveys are often not precise and also often often under report the use of alcohol. Furthermore, developed countries are more willing to report data. Underdeveloped data as are often not as available.   
Data gathering methods for the alcohol consumption of youth are likely not able to be improved in any other method than a survey. Perhaps researchers could calculate the difference between survey data and real data. This difference can then be created into a formula that is then applied to survey data to create more accurate results.   
These methods might affect the analysis of and conclusions based on the data set by underreporting data. Outliers that are usually regarded may be paid even more attention to because of the unreliability of these methods.   
  
**Describing Data**  
Below shows graphical displays of the data including a histogram with 20 bins and a boxplot.   
The data is skewed towards the right and is bimodal with the data being most centered at 0 and 7. The shape may be a result of the legal age of drinking in many countries being 18 and older whereas our data is collected from those 15-19 years old. The median is the more appropriate measure of center because the data has two extreme outliers. Furthermore, because of the data gathering method and the subject of youth and alcohol, the median may be a more accurate measure of center. The interquartile range is also thus a more accurate measure of spread as it employs the median as it’s center.   
  
Mean: 7.4963  
Median: 7.2  
 Standard Deviation: 5.3829, 5.4028  
 Interquartile range: 7.45  


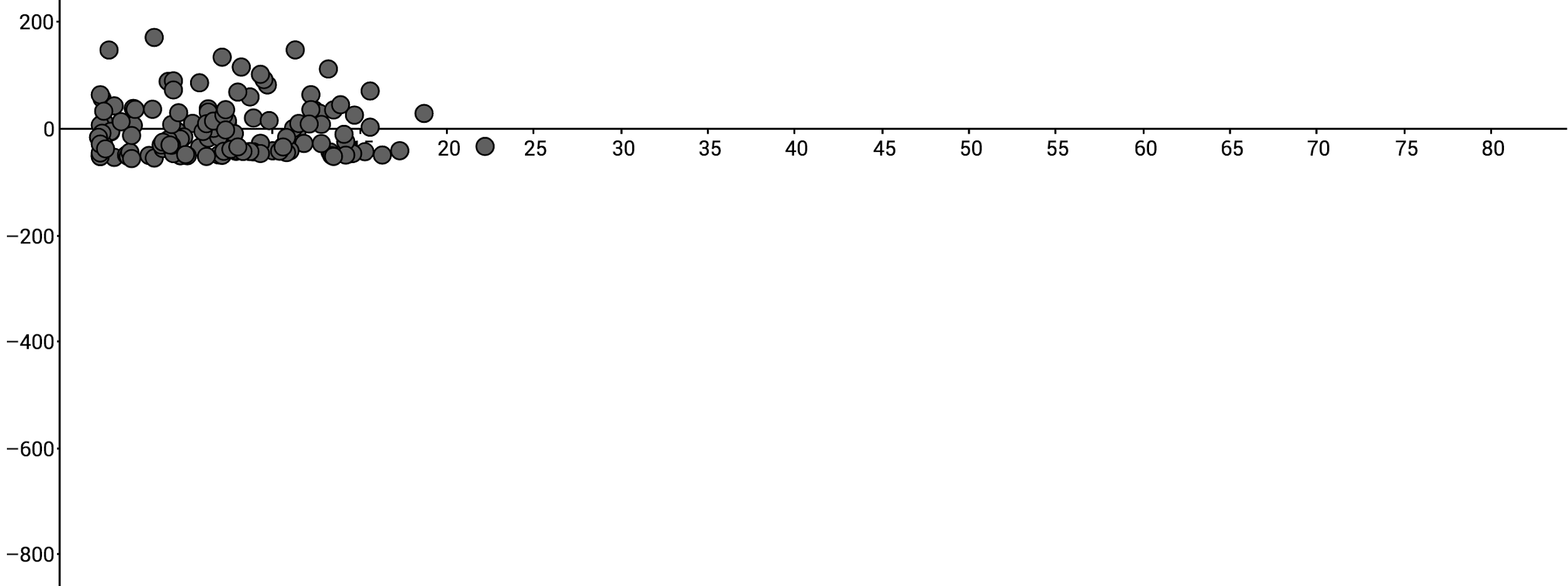
  
**Linear Regression**  
Below shows the scatter plot of “15-19 years old total alcohol consumption in litres of pure alcohol in 2010” as the independent and “Adolescent birth rate (per 1000 women aged 15-19 years)” as the dependent variable, and the linear regression model right below it. Further down shows the scatterplot and regression model with the outliers removed.   
There is not a strong linear association at all in the original data with an r-squared value of 0. In the data with the removed outliers, the r-squared value is 0.0001, which also shows a very weak linear association. This is interesting, because one would assume a strong correlation with teenage and drinking, but this is not the case.  
The equation of the line of the line of best fit is y=0.04x+55.96. Using this model, a nation in which adolescents from age 15-19 that drank 0.1 liters of alcohol would have 55.9692 births per 1000 women aged 15-19 that year. The actual amount is 102 births, which is twice the calculated number. It is thus shown from linear regression model that this model is not very accurate.

In fact, the data with removed outliers show an incredibly weak negative association. The two extreme outliers are a result of the outliers from the alcohol consumption data set.









**Confidence Intervals**

A T-estimate for difference of means was taken for the average teen alcohol per capita consumption and the average teen birth rate in each country under a 95% confidence interval.

There is uncertainty in our model itself since the coefficients we found are estimates of the “true” parameters. Thus there is also uncertainty in what is the “true” mean. There is also uncertainty because the mean itself is just an estimate. We expect some variation around the mean of teen alcohol per capita consumption of 10 liters. To distinguish the these ideas we call the narrow interval of just where our mean is likely to be the confidence interval for the mean predicted value at x=10 liters, and the wider interval the prediction interval.

Confidence interval: 7.767539 +/- 0.8909296

Prediction interval: 8.991140 +/- 10.32451

Standard Error: 0.009062

The confidence interval is fairly small whereas the prediction interval is fairly large in terms of range. This indicates that that the margin of error is fairly low. This means that average teen alcohol per capita consumption is not a statistically significant predictor of teen pregnancy rate per capita at the 95% confidence level.

> summary(No...Sheet1)

Call:

lm(formula = Alcohol ~ Pregnancy)

Residuals:

Min 1Q Median 3Q Max

-8.9503 -3.6819 0.0735 3.1218 19.7354

Coefficients:

Estimate Std. Error t value

(Intercept) 9.237106 0.675768 13.669

Pregnancy -0.031135 0.009062 -3.436

> plot(Alcohol,Pregnancy)

> mathfit=lm(Alcohol~Pregnancy)

> mathfit

Call:

lm(formula = Alcohol ~ Pregnancy)

Coefficients:

(Intercept) Pregnancy

9.23711 -0.03113

> abline(mathfit,col="red")

> newdata=data.frame(Alcohol=c(10))

> predict(mathfit,newdata,interval="confidence")

fit lwr upr

7.767539 6.8766094 8.6584686

> predict(mathfit,newdata,interval="prediction")

fit lwr upr

8.991140 -1.33337 19.31565

**Hypothesis Testing**

A two sample t-test of difference of means was taken to compare Asian countries’ alcohol consumption to European countries’. The null-hypothesis here is that there is no “true” difference between the variables, and the alternative hypothesis is that there is a difference between the variables.

east=subset(stats...Sheet1,Region=="Asia")  
> attach(east)

> west=subset(stats...Sheet1,Region=="Europe")  
> attach(west)

> t.test(east$Alcohol,west$Alcohol)  
  
 Welch Two Sample t-test  
  
data: east$Alcohol and west$Alcohol  
t = -7.5109, df = 62.882, p-value =  
2.586e-10  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -11.024848 -6.391045  
sample estimates:  
mean of x mean of y   
 4.551429 13.259375

The t-test estimated that the mean alcohol consumption per capita of Asian countries was 4.551429 liters, whereas the mean alcohol consumption per capita of European countries was 13.259375 liters. This parameter of the t-test was 62.882 degrees of freedom. Our 95% confidence interval for the value of the the difference of the true mean is between -11.024848 and -6.391045. The p-value being below 0.05 means that we reject the null hypothesis and the data shows that there is a statistically significant difference between the mean alcohol consumption of Asian countries and European countries.

However, a type 1 error could have been made in which the true null hypothesis could have been mistakenly rejected. Given that there are more first world countries in Europe than in Asia, data collection could be possibly more underreported in Asian countries than European countries. 37 European countries were sampled while 41 Asian countries were sampled, which may not be a high enough sample size.