Data Analysis Project

Brian Won Lee

May 18, 2023

Part 1: Table of summary statistics for two continuous variables.

Data for Parts 1 and 2 comes from the "Auto" dataset within the package "ISLR".

Table 1: Summary Statistics. Sample size is 392 observations for each variable. The difference of mean for horsepower versus acceleration is 88.96. The variation seems to be very different as the standard deviations are not comparable because each variable have different unit of measurement. Horsepower is measured in engine horsepower and accleration is the time to accelerate from 0 to 60 mph in seconds. It makes sense because as more horsepower an engine has, the faster (less time in seconds) to reach from 0 to 60 mph.

Var	SSize	Min	Q1	Med	Q3	Max	Mean	SD
horsepower	392	46	75	93.5	126	230	104.5	38.49
acceleration	392	8	13.78	15.5	17.02	24.8	15.54	2.76

Part 2. Histograms, density estimates, and boxplots of same two continuous variables.

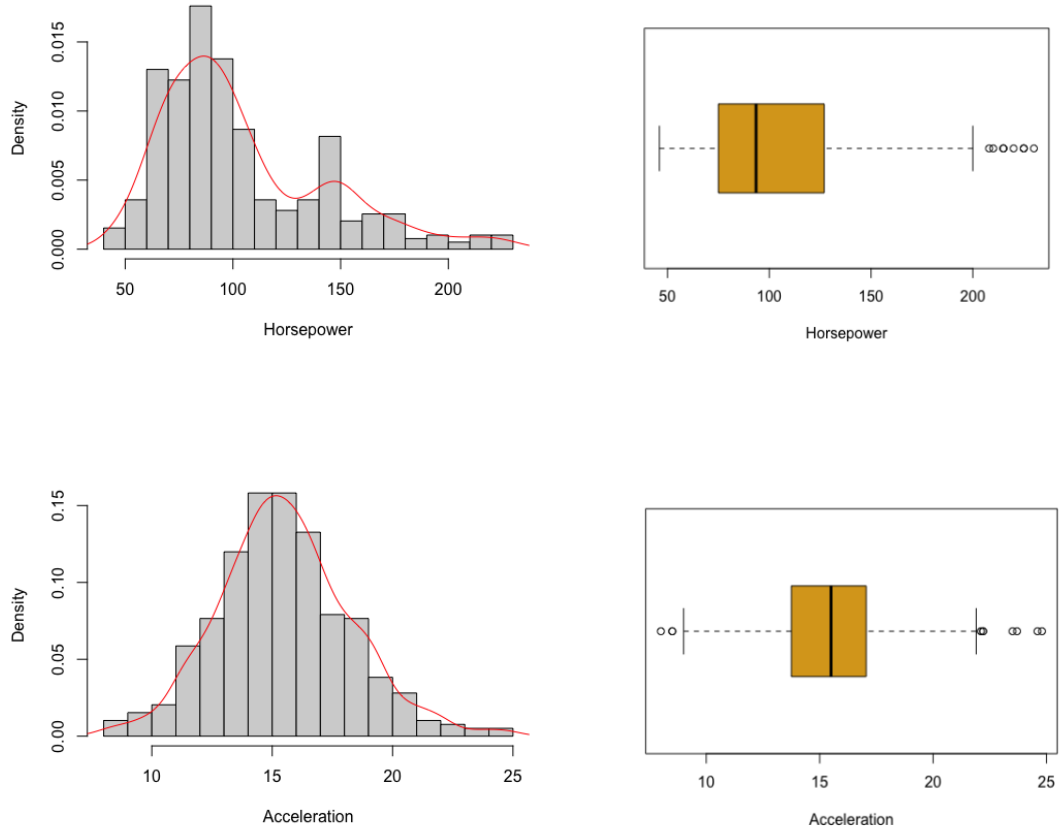


Figure 1: Univariate plots of Horsepower (top row) and Acceleration (bottom row). Horsepower histogram with density estimate is bimodal and nonsymmetric, and Acceleration histogram with density estimate is unimodal and symmetric although not a perfect symmetry. Distribution for Horsepower is skewed-right, and distribution for Acceleration has negligible skewness (almost no skewness) to the right. Boxplot for Horsepower has outliers in the upper tails of its distribution, whereas Acceleration has most of its outliers at the uppertail and little at the lowertail.

Part 3: Categorical Analysis for 2 categorical variables: Cross-tabulation and mosaic plot of HouseVotes84 data.

Data for Part 3 originates from the dataset HouseVotes84 found in the mlbench package.

		V1(handicapped-infants)		
		n	y	sum
V15(duty-free-exports)	n	146	82	228
	y	73	98	171
	sum	219	180	399

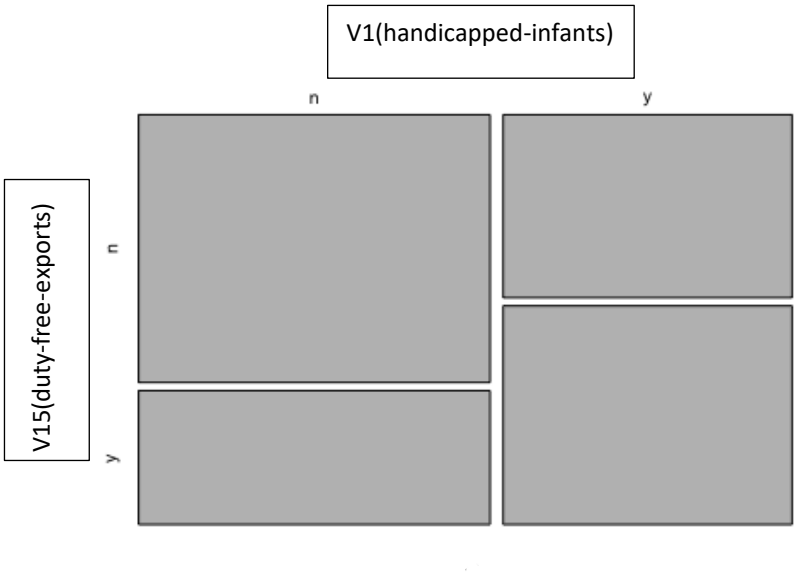


Figure 2. Data comes from the dataset HouseVotes84 in the mlbench package. The aggregate votes of Democrats and Republicans voted no(n) or yes(y) based on two different variables. Majority of the voters voted against both of these ideas.

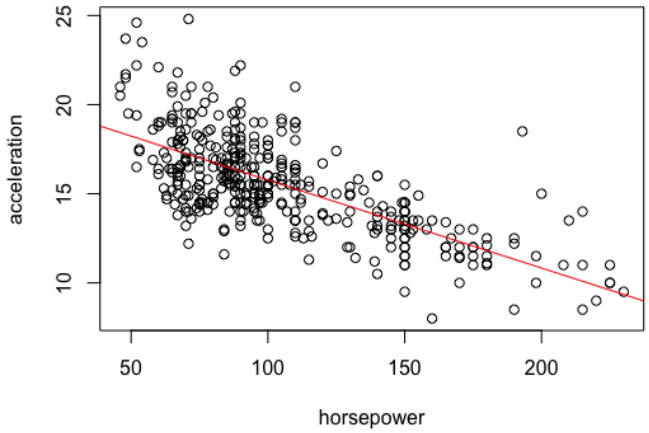
Discussion:

- Odds Ratio - Compares the Odds. Among those who voted no(n) for V15(duty-free-exports), there were 146 who voted no and 82 who voted yes(y) for V1(handicapped-infants). Thus, the odds of voting yes for V1 among those who voted no for V15 is 82:146. On the other hand, among those who voted yes for V15, there were 73 who voted no and 98 who voted yes for V1, so the odds of voting yes for V1 among those who voted yes for V15 is 98:73. The odds ratio of (98/73)/(82/146) = 2.39 indicates that the odds of voting yes for V1 among those who voted yes for V15 is approximately 2.39 times that of those who voted no for V15. In simpler terms, supporters of V15 were about 2.39 times likely to also support V1 as compared to those who did not support V15.

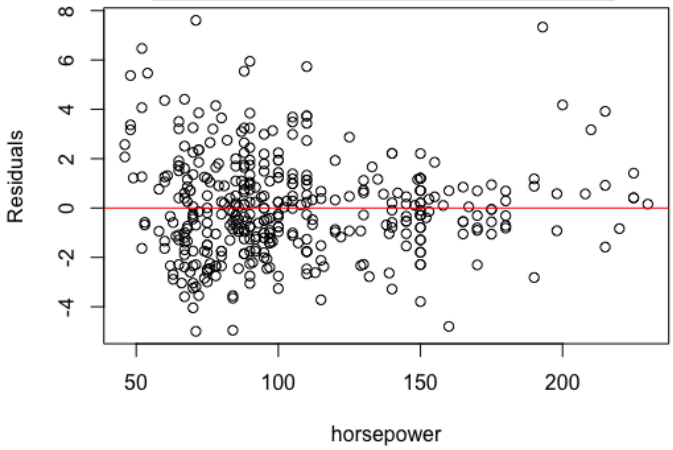
Part 4. Simple Linear Regression Analysis

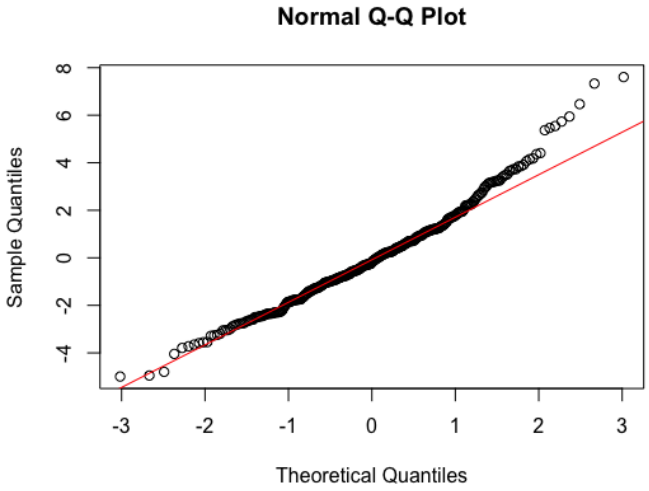
Data for this part originates from the Auto dataset found in the package ISLR. There were 392 total observations for both variables: acceleration (Y) as response and horsepower (x) as independent or explanatory variable.

Scatter Plot
(acceleration~horsepower)



Residuals vs horsepower(Fitted values)





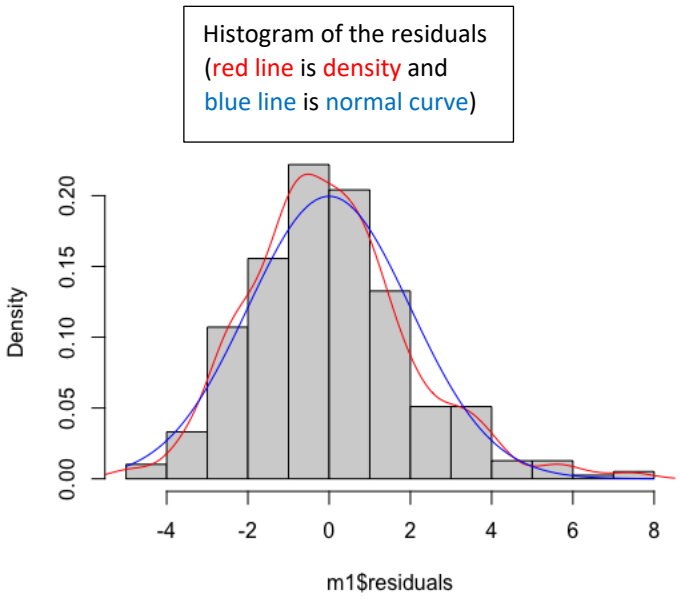


Figure 3: The first graph is a scatter plot of acceleration as the response (Y) and horsepower as the explanatory or independent variable (x). The second graph is residuals vs horsepower(fitted values). The third graph is the qqplot of the residuals. The fourth graph is the histogram of the residuals, which red line indicates density and blue line as normal curve. Linearity is plausible. It appears that horsepower increases linearly as acceleration decreases. Constant variance is questionable, because it does not look like residuals have the same amount of variation. The residuals do not appear to be normally distributed. The residuals are slightly more peaked than a normal distribution and have longer left and right tails. This can be seen in the qqplot where theoretical quantiles rise above the line. This can also be seen in the histogram comparison of the normal model (blue) and density overlay (red). It also appears that there is a slight skew right in the residuals. The linear regression estimates are $\hat{\beta}_0 = 20.70193$ and $\hat{\beta}_1 = -0.04940$. The estimated regression line is $\hat{y} = 20.70193 - 0.04940x$ (x as horsepower and $\hat{y}$ as estimated acceleration). Since $\hat{\beta}_1 = -0.04940$, the mean acceleration decreases by 0.04940 per horsepower in an engine. The correlation of determination is ($r^2 = 0.475$) so about 47.5% of the variation in acceleration is explained by horsepower. The correlation coefficient is r = $\sqrt{0.475}$ = about -0.6892. This is negative due to slope parameter and there is some degree of correlation but not as strong due to many variability.