Project - Sample Solution

DDSmith

March 17, 2021

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

Data for Parts 1 and 2 comes from the `Cars93` dataset within the package `MASS`.

Table 1: Summary Statistics. Sample size is 93 observations for each variable. The mean mpg for highway driving is about 7 miles versus city driving. The variation seems to be the same as the standard deviations are comparable.

Var	SSize	Min	Q1	Med	Q3	Max	Mean	SD
MPG.Highway	93	15	18	21	25	46	22.37	5.62
MPG.City	93	20	26	28	31	50	29.09	5.33

Part 2. Histograms and boxplots of same two continuous variables.

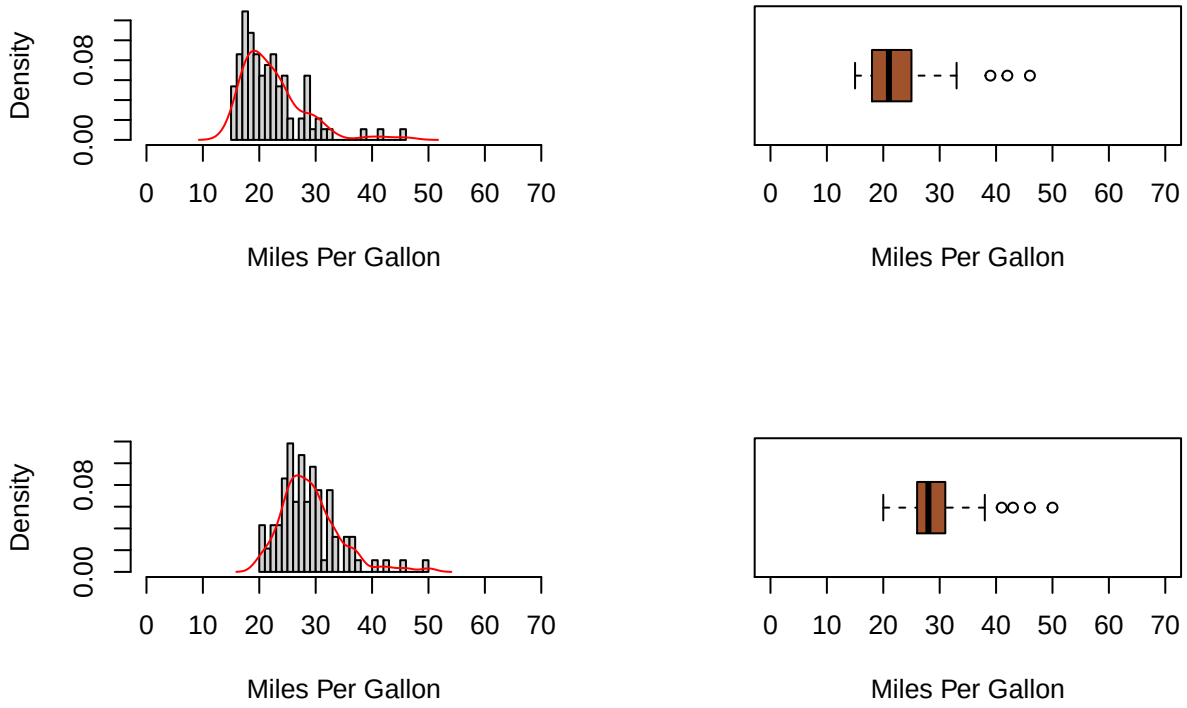

Figure 2: Univariate plots of `MPG.City` (top row) and `MPG.Highway` (bottom row). Both are unimodal and nonsymmetric. The mode for `MPG.Highway` is shifted right by about 7 mpg. Both distributions are skewed-right. The higher mpg for highway driving is also seen in the boxplots. Both have outliers in the upper tails of their distributions.

Part 3: Mosaic plot and table of zoo data.

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

```
##         legs
## predator   0   2   4   5   6   8 Sum
##    FALSE   6  16  16   0   7   0  45
##    TRUE   17  11  22   1   3   2  56
##    Sum    23  27  38   1  10   2 101
```

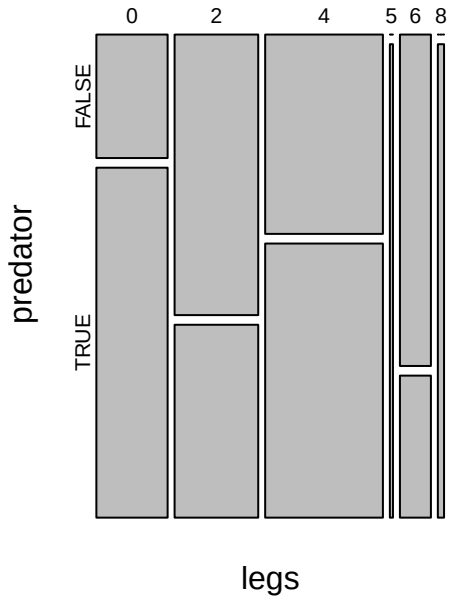

Figure 2. Data comes from the dataset `Zoo` in the `mlbench` package. The existance of whether an animal is a Predator or not is cross-classifed against the number of legs. A few interesting patterns to note: It appears that all eight-legged animals are predators (there are only two - mollusks). Most six-legged animals are not predators (insects). There are slightly more four-legged predators than non-predators.

Discussion:

- Relative Risk - Compares conditional probabilities. The conditional probability that an animal at the zoo with no legs is a predator is 17/23. The conditional probability that an animal at the zoo with no legs is not a predator is 6/23. The relative risk of an animal with no legs being a predator vs not-a-predator is $(17/23)/(6/23) = 17/6 = 2.83$. This indicates that an animal with no legs is about 2.8 times more likely to be a predator.
- Odds Ratio - Compares the Odds. There are 17 predators and 6 non-predators that have no legs. Thus, the odds of being a predator if you are an animal at the zoo with no legs is 17:6. If you are an animal at the zoo with 2 legs then the odds of being a predator are 11:16. The odds ratio of $(17/6)/(11/16) = 4.12$ indicates that the odds of an animal with no legs are about 4 times more likely of being a predator than animals with two legs.

Part 4. Scatterplot and linear regression analysis.

Data for this part originates from the `Batting` dataset found in the package `Lahman`. The dataset is rather large. I have taken a random sample of 2000 out of the 107 thousand observations.

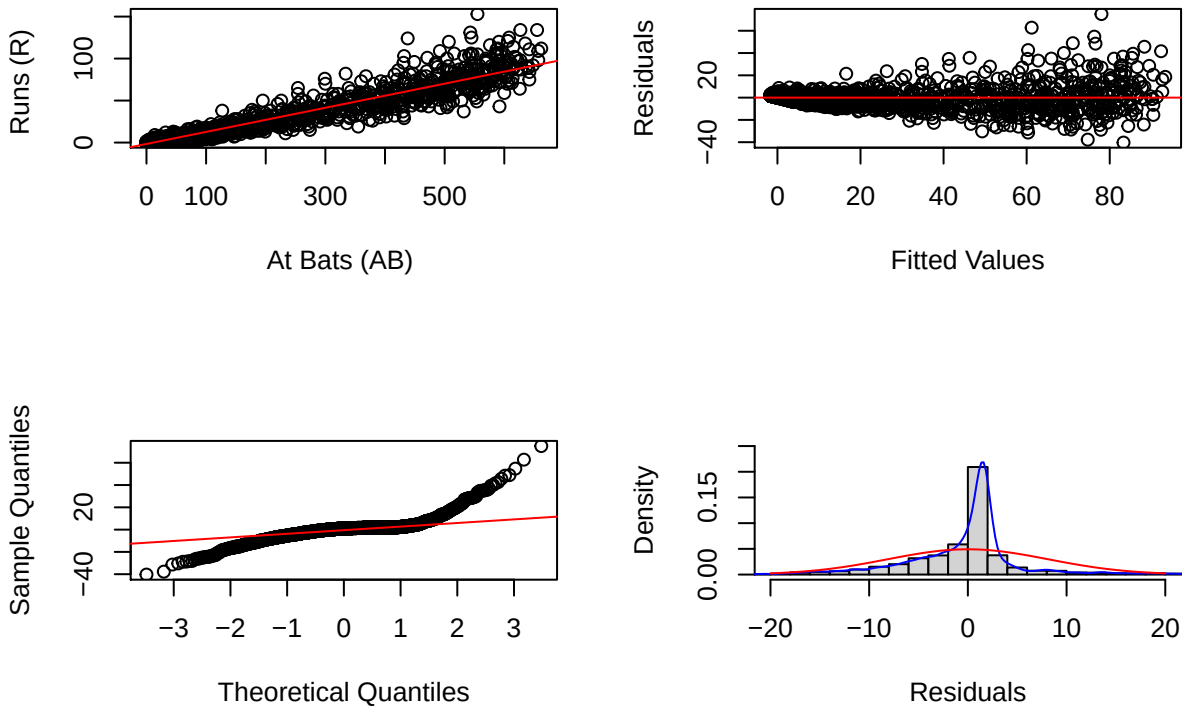

Figure 3: Plot of Runs as a function of At Bats. Linearity is plausible. It appears that Runs increases linearly as At Bats increase. Constant variance is suspect. It appears that variation in Runs increases as At Bats increases. The residuals do not appear to be Normally distributed. The residuals are more peaked than a Normal distribution and have much longer tails. This is seen in the qq-plot where the theoretical quantiles fall sharply below and above the line. This can also be seen in the histogram comparison of the Normal model (red) and density overlay (blue). It also appears that there is a slight skew left in the residuals. The linear regression estimates are $\hat{\beta}_0$ = -1.68 and $\hat{\beta}_1 = 0.14$. Mean Runs increase by .15 per At Bat. The correlation coefficient is 0.9556 with an R^2 of 91.32%. About 92% of a player's Runs can be explained by At Bats.