SOME FAVORITE CAR ACTIVITIES

Robin Lock
St. Lawrence University
rlock@stlawu.edu

We describe a series of class activities and student project topics that all revolve around data about cars that can be used at various points of a first or second course in statistics. In several of the projects, students work with their own datasets where they have some choice in selecting the population to sample from, via a convenient web form. However, we maintain a consistent set of variables and format to facilitate grading. Topics include correlation, regression (simple and multiple), ANOVA for means, and ANCOVA. We discuss ideas for organizing these activities and using RMarkdown documents to facilitate assessment.

INTRODUCTION

One challenge with using real data when teaching statistics is to find contexts that students find familiar and interesting. Applications in areas such as sports, medicine, or business might have great appeal to some students, whereas other students might have little interest or be unfamiliar with issues and jargon of the area. One topic that seems to be more universally accepted is cars. Students might own their own car, have a family car, or dream of owning a car after graduation. They are generally familiar with terms such as city miles per gallon rating or mileage on a car and have some intuition about how such variables might be related. We discuss several activities and project topics that build on this interest in cars.

Another challenge is the tension between giving students ownership of individualized data while minimizing the burden of grading projects and ensuring that the data illustrate the desired concepts. For many of these car-based projects, students can choose their own car model and geographic area to sample, but the variables measured will be consistent across projects. We develop RMarkdown documents to produce individualized keys to assist with the grading process.

We start with an in-class activity for an introductory statistics course that introduces the concept of correlation. The remaining projects and activities come at various points in a second statistics course, covering topics such as simple linear regression, multiple regression, analysis of variance for means, and analysis of covariance. Sample handouts, datasets, web links, and RMarkdown grading documents for adapting these activities and projects to other teaching situations can be found at http://myslu.stlawu.edu/~rlock/icots11.

CAR CORRELATIONS

This activity guides students in an introductory class to discover properties of sample correlations. The dataset contains information on a sample of 110 different new car models (Consumer Reports, 2020). Variables include weight, height, city miles per gallon, fuel capacity, time to go a quarter mile, and time to accelerate from 0 to 60 mph. Students work in pairs or triples to first brainstorm (before seeing any data) the direction and strength of association they might expect to see between six pairs of variables by putting the letter for each pair somewhere along the scale in Figure 1.

(a) Weight vs. CityMPG (b) Weight vs. FuelCap (c) Weight vs. QtrMile
(d) Acc060 vs. QtrMile (e) Acc060 vs. Height (f) CityMPG vs. QtrMile

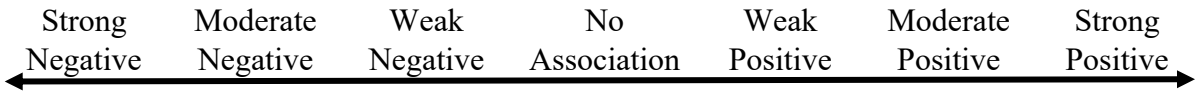

Figure 1. Scale for recording strength and direction of associations

After making initial guesses, students get access to the data and work together (using Minitab in a computer classroom) to make scatterplots for each of the relationships and then revise their guesses on a second version of Figure 1. We then introduce the idea of correlation as a numerical measure for

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

such relationships. Students use the software to find correlations for the six variable pairs, and then are asked to discuss what they think might be the properties of correlation.

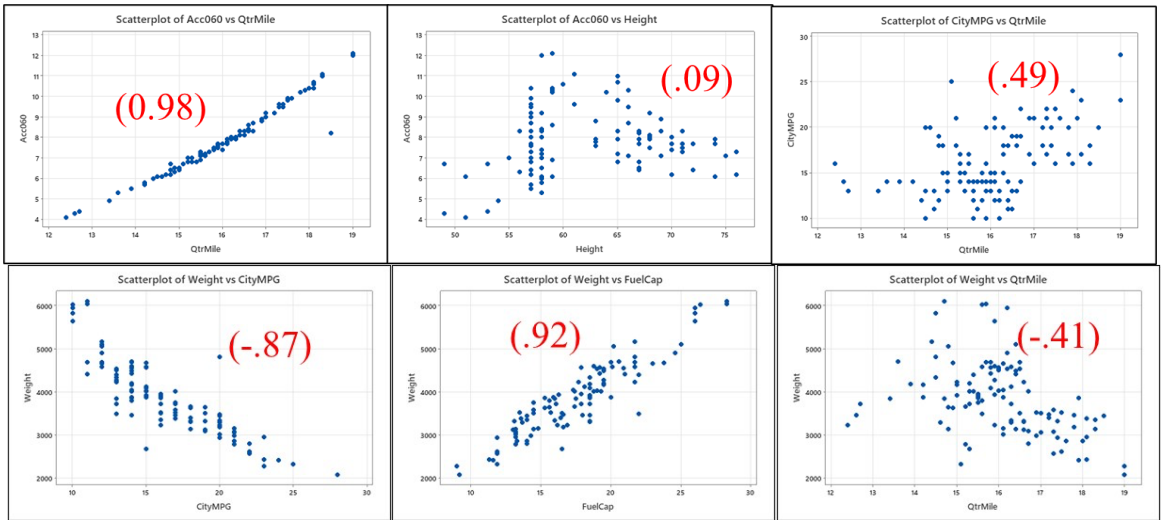

Figure 2. Car correlations

Students readily surmise that correlations will be values between -1 and $+1$, with the sign of the correlation indicating the direction of association, and closer to ± 1 showing stronger associations. Putting a scale from -1 to 0 to $+1$ on what they have labeled in Figure 1 helps students make connections between their intuitions about relationships, what they see in the scatterplots, and the numerical summary provided by the correlation coefficient.

INDIVIDUALIZED CAR DATA

The next series of project activities go throughout an Applied Regression (Stat2) course. We recognize the value of each student working with their own datasets but want to maintain a predictable structure for what the data are likely to show. To accomplish this, we use a common theme of modelling prices for used cars. We use an online source (https://autotrader.com) for students to obtain data. They choose a car model and location (by zip code) that they want to sample. To facilitate the sampling process, we use a web app (thanks to Dr. Choong-Soo Lee) as shown in Figure 3 that automatically scrapes the needed data.

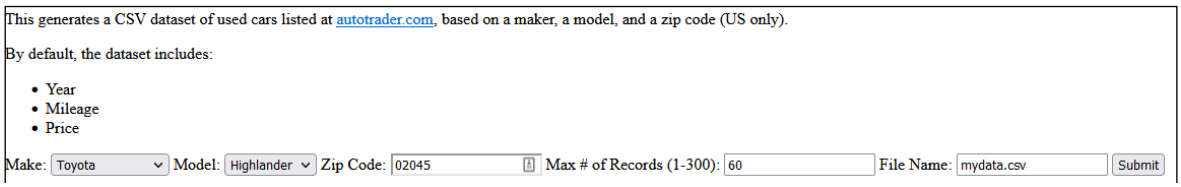

Figure 3. Collecting car data at *http://myslu.stlawu.edu/~clee/dataset/autotrader/*

The app selects a sample of used cars of their model for sale in the vicinity of the chosen zip code and records the *year*, *mileage* (in thousands of miles), and *price* (in $1,000's) to save in a CSV file. We want each student to have a sample of 50 prices for their car models but suggest that they originally request a few more because they might find some unusable cases (such as a new car or a missing value for price or mileage) that they can then trim down to 50 cases in the final dataset. Depending on the car model and zip code, there may not be 50 cars for sale in that region (students see the sample size before accepting it). In that case they can choose another model or zip code. (The 02045 code in Figure 2 is from near Boston, Massachusetts, so it generally has a lot of cars for most models.) We also ask students to add an *age* variable to the CSV dataset, which they find by subtracting the *year* from the current year. Students submit their datasets electronically (before the first project is due) so that we can verify that

they are using the proper format and so that we have copies of their data for checking work and later projects. Now they are ready for the first project.

SIMPLE LINEAR REGRESSION PROJECT

The first project using their car data involves regression with a single predictor, *age* of the car, for the *price*. Not surprisingly, these variables almost always show a negative association. Although we introduce linear regression in the introductory statistics course, the Stat2 course starts with a review and adds more computational details (such as formulas for coefficient estimates), which we ask students to work out in this project. Students do project work in an RMarkdown document using RStudio to assist with those calculations. Figure 4 shows a typical plot for a student who sampled prices for a Subaru Forester near zip code 94566 (Pleasanton, California).

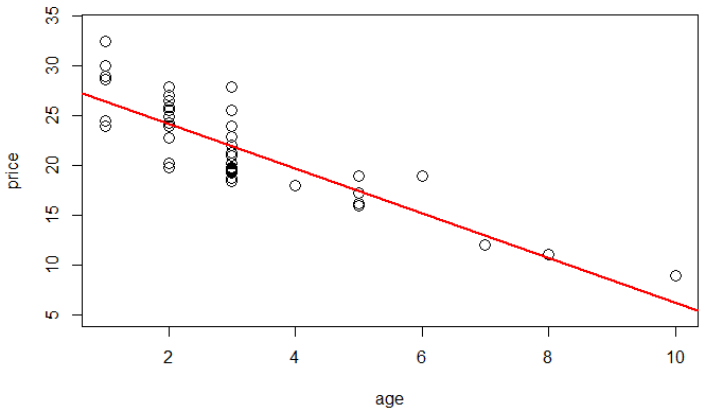

Figure 4. Price vs. Age for Subaru Forester

Tasks for this project include:

- Estimate regression coefficients from basic sums.
- Use plots to check regression conditions.
- Use studentized residuals, leverage, etc. to identify any unusual points.
- Do inference for the slope and correlation.
- Connect r^2 to the ANOVA table.
- Find a 95% prediction interval for the price of a three-year-old car.
- Answer: Is there an age at which this car model is expected to be free?

But Isn't Grading Projects Hard When Everyone Gets Different Answers?

Here is one of the advantages of using a common data theme. We can relatively easily produce an RMarkdown file that generates a "key" with calculations, plots, and other relevant output based on any individual student's dataset. If we save each student's submitted data in a CSV file with their name, we can readily run the document to print (or store) the key to use when grading and later attach to the graded project to return to the student. Some examples of such RMarkdown files for the projects described here can found at http://myslu.stlawu.edu/~rlock/icots11.

Most students get similar results throughout the project. They see a negative association that is usually quite statistically significant. Samples tend to be skewed towards younger cars (as seen in Figure 3). Some students see this as a problem with checking regression conditions (for example, when viewing a residuals vs. fits plot), but it is actually just a natural part of the sampling process because more "younger" used cars tend to be advertised for sale. When asked about the "free car" phenomenon, most students can find a point when the predicted value drops to zero and then tend to offer one of two reactions for what this says about the model. Some simply say it just shows the model is flawed (because cars shouldn't be free or negatively priced), others point out the dangers of extrapolation—especially when the "free" age is beyond any ages in their sample. Although often subtle, there is often a bit of curvature to this relationship; prices tend to decrease more rapidly in the early ages and slow down for older cars. The next project can help address this issue.

MULTIPLE REGRESSION PROJECT

A bit later in the Stat2 course, we return to the car data to give students some practice with multiple regression. An obvious extension is to include the *mileage* variable that appears in the original dataset. This can cause some interesting results when included with *age* in a multiple regression to predict *price* because *age* and *mileage* are often strongly correlated themselves. For example, when using the two-predictor model for the *Forester.csv* data, we find that *age* is still clearly significant ($\hat{\beta}_1 = -1.71, p = 1.8\times 10^{-5}$), whereas *mileage* is not quite significant at a 5% level ($\hat{\beta}_2 = -0.044, p = 0.073$)—even though, by itself, *mileage* is a strong predictor ($r = -0.776, p = 3.5\times 10^{-11}$).

We can also address the curvature issue by trying a quadratic model based on *age* (Figure 5a). Because prices tend to depreciate more rapidly in the early years, this often produces a concave up shape that eventually hits a low point and begins to turn upward. This may eliminate the "free car phenomenon but now raises a different concern about why prices might rise. Again, students generally use one of two possible explanations, a general flaw in the quadratic model or, more creatively, a "classic" effect that sees prices rising for vintage older cars (although it is probably hard to justify this for a Subaru Forester!). A more reasonable model might use a log transformation on the prices, as seen in Figure 5(b), that avoids both the zero-price issue and the antique car effect.

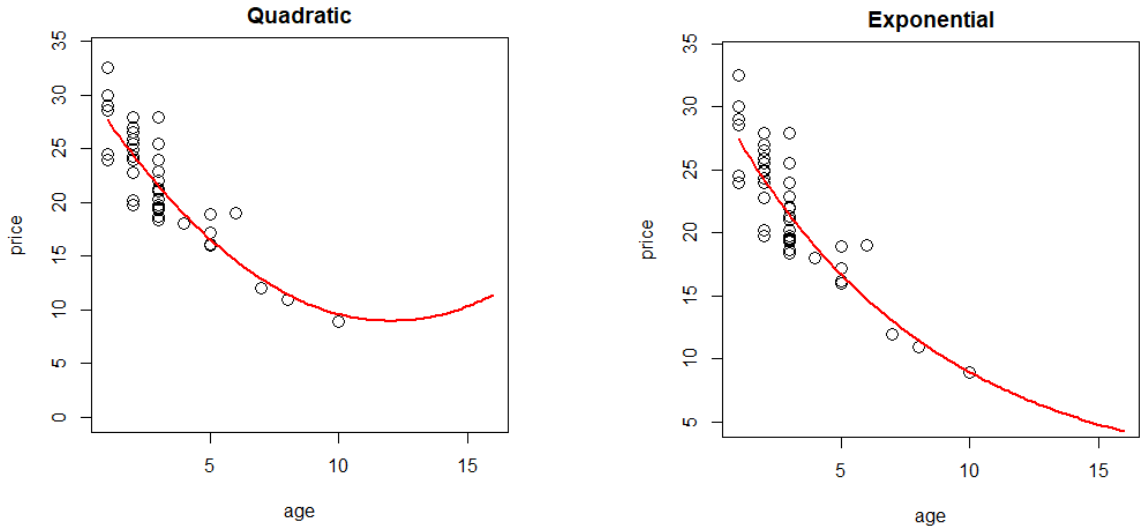

Figure 5. (a) Quadratic fit (b) Exponential fit

ANOVA TO COMPARE MEANS

We use the car data again with a project to let students work with a one-way ANOVA model. Here we have each student work with four different car models, their original car (Car #1) and the data from three other students. We create a table such as in Table 1 to assign the groups and create a folder with the datasets from Project #1. Although some car models are re-used, no two students get the same set of four car models. Students merge their data with data for the three students in their row and add the car model names as a variable.

Table 1. First few lines of a table to assign car models for an ANOVA project

Car #1	Car #2	Car #3	Car #4
Beckman (Camry)	Rossman (CRV)	Pearl (Jetta)	Garfield (Outback)
Chance (Elantra)	Pearl (Jetta)	Peck (Mazda6)	Franklin (Passat)
delMas (Durango)	Wild (Explorer)	Garfield (Outback)	Rossman (CRV)
Garfield (Outback)	Franklin (Passat)	Witmer (Rav4)	Hartlaub (Mustang)
Moore (Mustang)	Chance (Elantra)	Cannon (Forester)	Witmer (Rav4)

Part of the project has students investigate pairwise differences (using Fisher's Least Significant Difference [LSD] and Tukey's Honestly Significant Difference [HSD]) and also check the conditions for a one-way ANOVA. We would generally like there to be some significant differences (but not all four groups) and for the conditions to not be egregiously violated. Thus, to help assign the groups, we produce a set of boxplots showing prices for all the car models as displayed in Figure 6.

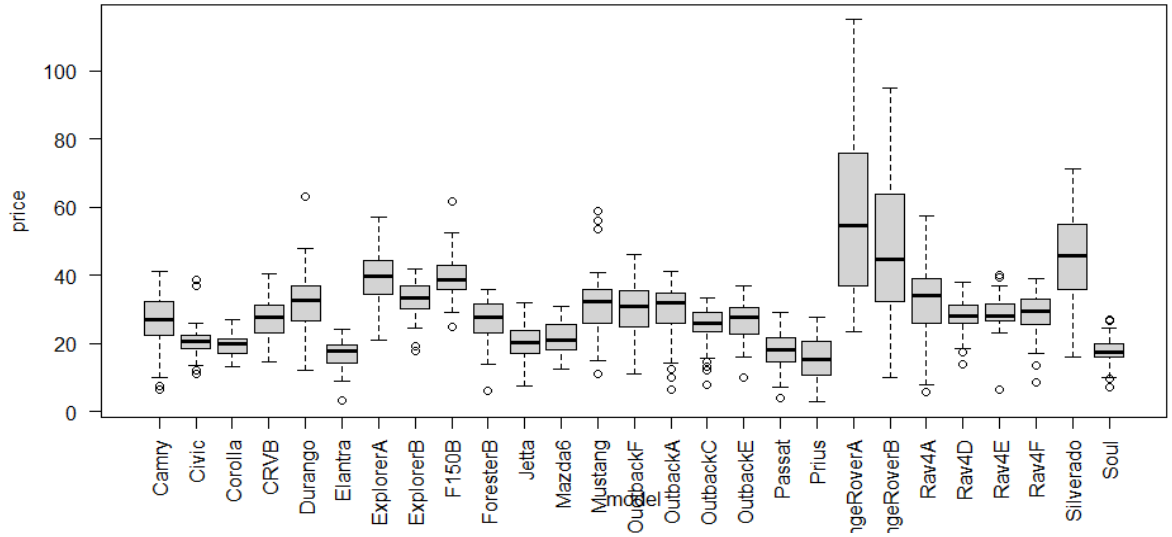

Figure 6. Boxplots of prices for 27 student car models

Some typical student results are shown in Figure 7. The mean price for Jetta (19.64 thousand) is clearly less than the other three models. Tukey's Honestly Significant Difference (HSD =3.61) shows no significant difference in mean price among the other three models, but Fisher's Least Significant Difference (LSD =2.41) indicates the mean price for Outback (29.44 thousand) may indicate a discernable difference from the Camry (26.92 thousand) and is close to showing a difference from the CRV (27.21 thousand).

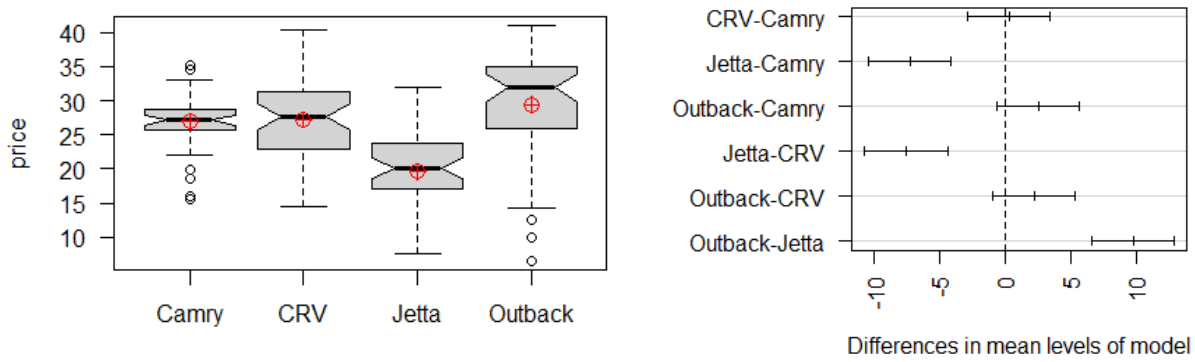

Figure 7. Boxplots of prices and Tukey HSD intervals for pairwise differences of 4 car models

ANAYSIS OF COVARIANCE

The car-themed data come back one more time as a class activity to illustrate ideas of analysis of covariance. Here the students all work with the same four car models (see Table 2).

Table 2. Means and standard deviations of prices for four other car models

Car	Audi	BMW	Mercedes	Yukon
Mean price	30.985	32.091	35.352	36.826
Std. dev.	122.15	11.25	18.59	17.84

When they run a one-way ANOVA for the prices, they find no convincing evidence for a discernable difference in means among the four car models ($F = 1.59$, $p = 0.192$). But they know from earlier projects that prices depend on age, so they can add *age* as a covariate to the model to produce the output shown in Figure 8.

```
               Estimate  Std. Error t value Pr(>|t|)
(Intercept)     53.1624      1.3928  38.170  < 2e-16 ***
age             -3.6717      0.1483 -24.760  < 2e-16 ***
ModelBMW        -2.7134      1.5162  -1.790  0.07507 .
ModelMercedes    0.4010      1.5168   0.264  0.79178
ModelYukon       4.5923      1.5092   3.043  0.00267 **
---
Residual standard error: 7.542 on 195 degrees of freedom
Multiple R-squared:  0.7644, Adjusted R-squared:  0.7596
F-statistic: 158.2 on 4 and 195 DF,  p-value: < 2.2e-16

Response: price
           Df Sum Sq Mean Sq  F value    Pr(>F)
age         1  34627   34627 608.8049 < 2.2e-16 ***
model       3   1361     454   7.9758 4.831e-05 ***
Residuals 195  11091      57
```

Figure 8. ANCOVA output using *age* and car *model* as predictors for price.

After accounting for age, students see that now the model factor explains a significant proportion of the remaining variability in prices. We ask them to brainstorm why this might happen and (with some prodding) they eventually come up with a couple of reasonable possibilities. Obviously *age* explains a lot of the variability in *prices*. This makes the residual standard error in the ANOVA model (7.54) about half of what it was using the car model alone (15.31). Thus, a difference in means that might not be considered significant with the ANOVA alone has a better chance of being convincing when that extra variability has been accounted for and removed. From the coefficients of the indicators in the model, we see that the adjustment for Yukon is a significant effect from the base case (Audi) and the BMW is mildly surprising in the other direction. Students find this curious because the mean Audi price in the sample is actually less than the BMW. What could account for this? Further investigation shows that the average age of the cars in the Audi sample (6.04) is more than a year more than the BMW (5.00) or the Mercedes (4.96), whereas the mean Yukon age (5.70) is almost as large as the Audi. So, the Yukons have the highest mean price, but also tend to be older—so at the same age they would tend to be even more expensive than the BMWs or Mercedes. Seeing this in such a practical setting helps students internalize what we mean by "comparing prices for car models after accounting for age."

CONCLUSION

Each of the activities and projects we describe here use car-based data with variables that are familiar to most students. Using the web form, students can tailor their data by choosing the car model to sample and zip code to sample from. Relationships between variables such as the *age*, *mileage*, and *price* of the cars for that model are relatively predictable, so we can anticipate what students are likely to find and can easily generate individualized "keys" to help check their work. Threading the same data situation throughout the course helps students see how different statistical techniques can be applied to similar questions for the same data.

REFERENCES

Consumer Reports. (2020). *Cars.* https://www.consumerreports.org/cars/