

# DSC 330 - Business Statistics

## "Real-world Case" Project

### Task

As an employee of a car rental company in a large U.S. city, you've been asked to provide a rigorous statistical analysis of some data on recent car models to support ongoing decisions about which make/models of car should be added to the company's pool. In particular, the company hopes to gain competitive advantage by focusing on high fuel-efficiency cars; consumer research has indicated that this will allow the company to legitimately promote itself as "environmentally conscious" and hopefully gain appreciable market share.

*In your group*, build a regression model to determine which factors involved in the construction of a car affect miles-per-gallon in city driving for the **CARSPROJECT** data file. Prepare a professional document that presents a detailed discussion of your analysis. You should use SPSS to analyze the data by following these guidelines:

### Guidelines

- The response (dependent) variable is "Y" = city miles per gallon.
- There are four potential quantitative predictor (independent) variables ("X1" = number of cylinders, "X2" = horsepower in hundreds, "X3" = torque in thousands, and "X4" = weight in thousands of pounds) and one potential qualitative predictor variable (with levels "manual" and "automatic") coded with a dummy indicator variable ("D" = 0 for manual transmission, 1 for automatic). Do *not* use "PRICE" (in thousands of dollars) as a predictor in any models, or "GROUP" (which labels the cars from 1 to 4 depending on how expensive they are).
- Look at the data in a scatterplot matrix; do you notice anything that might have an impact on subsequent regression analyses?
- As a first model, try using all five predictors ("D", "X1", "X2", "X3", and "X4") just as they are (untransformed). Construct some residual plots to see whether there is *clear* violation of any assumptions for this first model.
- If you decide that this first model has some clear problems, then you'll need to try a different approach. One thing to try is transforming some of the predictors, e.g. compute reciprocal transformations of "X2," "X3," and "X4" (call the transformations "X2r," "X3r," and "X4r") and fit a second model with these five predictors: "D," "X1," "X2r," "X3r," and "X4r." Does this second model improve on the first model? How do you know?
- The second model has each of the quantitative predictors affecting the response variable similarly for manual and automatic cars (to see this write out two equations for the model, one for manuals and the other for automatics). Consider generalizing the second model so that each of the quantitative predictors affects the response variable *differently* for manuals and automatics (you will need to compute interactions between "D" and each of "X1," "X2r," "X3r," and "X4r" before fitting this third model).
- Does this third model have some irrelevant terms included, making it hard to interpret? Do a nested model F-test to see whether "X2r" and the "DX2r" interaction term are needed.
- If you decided "X2r" and the "DX2r" interaction term are not needed, delete them to produce a fourth model. Does this fourth model still have some irrelevant terms included, making it hard to interpret? Do a nested model F-test to see whether the "DX1" and "DX3r" interaction terms are needed.
- If you decided the "DX1" and "DX3r" interaction terms are not needed, delete them to produce a fifth model. Construct some residual plots to see whether there is *clear* violation of any assumptions for this model. Are you happy to use this as a final model or can you think of

- anything else to try to improve things?
- Are there any outliers? (If so, you might want to consider removing the cases in question and re-fitting the model without them - just make sure you explain what you've done in your report.)
  - Does your final model provide a good compromise between explanatory power and parsimony (i.e. does it adequately explain the relationships between the response and predictors without being overly complicated and including irrelevant terms)?
  - How does mileage depend on number of cylinders? For example, does mileage decrease or increase as number of cylinders increases (holding other predictors fixed)? Does the relationship depend on whether the car is manual or automatic?
  - Similarly, how does mileage depend on torque? (Be careful with your explanation here since we've transformed "X3" in the final model.)
  - Similarly, how does mileage depend on weight? (Be especially careful with your explanation here since we've transformed "X4" in the final model, *and* you should have a "DX4r" interaction term in your final model.) You might like to draw a "line plot" to show this particular effect graphically.
  - Does an automatic car tend to have higher or lower mileage than a manual car (holding other predictors fixed)? Or does the answer to this question depend on the value of one or more of the other predictors? You might like to use the "line plot" from the previous question to show this graphically.
  - Car rental customers generally take into account additional factors other than mileage when deciding which make/model of car they would like to rent, e.g. comfort, safety, drive-ability, price, etc. (note that some of these other factors will be related to the predictors in the dataset you've just analyzed). Thus, if your company wants to attract customers based on the make/models it carries, it would be advantageous to use more than just fuel-efficiency figures when deciding what those make/models should be. In particular, it is more meaningful to know which cars have the best mileage when you *take into account* other factors, rather than just which cars have the best mileage when you *ignore* other factors. Extend this argument to explain how your regression analysis results can be used to support decisions about which make/models of car should be added to your company's pool.
  - Two new cars are coming onto the market:
    - Wallace Jalopy ( $Y=19, X1=4, X3=1.5, X4=3.0, D=1$ )
    - Gromit Speedster ( $Y=18, X1=6, X3=3.5, X4=4.0, D=0$ )
 Would you recommend that your company buy either of these cars? Why?
  - Use your final model results to recommend two or three specific make/models of car for each of the four price categories labeled by the "GROUP" variable (which labels the cars from 1 to 4 depending on how expensive they are).

## Grading

The maximum number of points available for the project is broken down as follows:

- Technique: 60 points
- Clarity: 30 points
- Focus: 30 points
- Presentation: 30 points

**Technique:** Have you used appropriate statistical techniques and checked assumptions underlying inferences?

**Clarity:** Have you explained clearly and concisely what you have done and why, and what your conclusions are?

**Focus:** Have you met the goals of the project fully and without superfluous analyses or discussion?

**Overall presentation:** Is the report presented neatly and in a professional manner?

A suggested **layout** for the report (to be typed using a computer, not handwritten) is as follows:

- *Introduction:* Briefly state the goals of your analysis and the data used.
- *Analysis:* Describe statistical techniques applied and assumptions underlying any inferences.
- *Results:* Summarize the results of the analysis, including any relevant plots.
- *Conclusions:* Describe how the results accomplish the goals of the analysis. You should include the line graphs discussed on pages 188-194 of the text-book in this section.

**Length:** Your report must be *no more than 8 pages* (using a font size no smaller than 10-point). *Anything beyond 8 pages will be **ignored** in grading.*

If you *don't do* the project, the mid-term portion of your overall score for the class remains as a score out of 300. If you *do* the project, your mid-term portion is converted to a score out of 150, and you can get up to 150 points for the project. (Unless this would give you a score less than your original mid-term score out of 300, in which case you would retain your original mid-term score and score nothing for the project.) Note that the homework, "Statistics in Action" sessions, mini-quizzes, and final exam count in everyone's final grade.

So, you can choose not to do this project if you are content to have your grade based solely on homework, class participation, quizzes and exams. If there is a lack of consensus within a group such that some group members want to do the project but others do not, the instructor may reorganize some groups so that everyone wishing to do the project is able to do so in a group of three or four. **If you would like to do the project, let the instructor know your intentions as soon as possible (in class, during office hours, by e-mail or phone).** Project groups will be finalized *in class*, when you will also have a chance to ask questions about the project.

## Due date

**Hard copy** (i.e., printed, not electronic) to be handed in by 10am on Monday of exam week.

## Extra hints

- The Final Exam Preparation (to be handed out in class) provides an example of the type of *layout* (not content) of the report expected for the project. In particular, the project just asks you to build a *regression* model and does not cover material from the "Foundations" chapter, whereas the final exam preparation covers everything from the quarter.
- Make sure you address all the questions in the guidelines in your write-up.
- The second Case Study in chapter 6 is based on a similar dataset to the one considered here, but **it is not the same.**