Adding Categorical Variables

Multiple linear regression – Stat 230

In this class, you will extend the multiple linear regression model to include categorical variables. You will consider how to do this, how to interpret the results, and how to determine whether the addition of a categorical variable discernibly improves the model (i.e., explains discernibly more variability in the response).

In this activity we will revisit the `Cars` data set described in Sections 3.1 and the first page of Section 3.3 (page 70).

```
cars <- read.csv("https://aloy.rbind.io/kuiper_data/cars.csv")
```

Adding a categorical variable

Task 1. (Activity 17) Make boxplots or individual value plots of `log(Price)` versus the categorical variables `Make`, `Model`, `Trim`, and `Type`.

Task 2. (Adaptation of Activity 18) Create five indicator variables that can be used to represent the six possible values of `Make` (Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn). To answer this, just describe *how* you would define them.

Task 3. (Adaptation of Activity 19) Build (i.e., fit) a multiple regression model using `log(Price)` as the response, and `Mileage` and the indicator variables for `Make`. To do this, you can simply add `Make` into the `lm()` statements we saw last class. R will automatically convert a categorical explanatory variable into a set of indicator variables. Report the fitted regression equation and the R^2 value. Which level of `Make` is the baseline (i.e., is represented by 0's across all other indicator variables).

Task 4. The inclusion of `Make` allows each `Make` to have a different y-intercept but a common slope. Write the fitted regression equation for each `Make` (you will have six equations in total).

Task 5. Now, let's plot the fitted "parallel slopes" model that you just wrote out. To do this, we can use the `plotModel()` function in the {mosaic} package. Use this plot to verify your work in the previous task.

```
# Replace the ___ with the name of your fitted model
# Be sure to load mosaic in the setup chunk!
plotModel(___)
```

Warning

If you transform the response variable you will get an error. To avoid this, you can create a new column in the data set for the transformed response and then refit your regression model. In this example, we do the following:

```
cars$logprice <- log(cars$Price)
cars_lm2 <- lm(logprice ~ Mileage + Make, data = cars)
plotModel(cars_lm2)
```

Testing a categorical variable

Is there evidence that `Make` is associated with price after accounting for mileage? To answer this we need to use an extra sums of squares F-test.

Task 6. Write down the hypotheses for the sums of squares F-test that can be used to determine whether `Make` is associated with price after accounting for mileage. Remember that this F-test is comparing two nested models, so start by thinking about what the full model is and what the reduced model is.

Task 7. You have already fit the full model above. Fit the reduced model and use the `anova()` command to run the extra sums of squares F-test. State your conclusion to this F-test.

Note

To run an extra sums of squares F-test, you can fit the `full` and `reduced` models and then pass them into `anova()` in the order you see below.

```
anova(reduced, full)
```

One thing to note is that R reports the SSE (which is labels `RSS`) for each model, not the SSR. Why? It turns out that the extra sums of squares F-test has two equivalent formulas: one in terms of the increase in explained variability (SSR), and the other in terms of the decreases in unexplained variability (SSE).

$$F = \frac{(\text{SSR}_{full} - \text{SSR}_{reduced})/(\text{df}_{full} - \text{df}_{reduced})}{\text{MSE}_{full}} = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(\text{df}_{reduced} - \text{df}_{full})}{\text{MSE}_{full}}$$

R chooses the the version subtracting SSE in the numerator.

Task 8. If you prefer to work with SSR you can run `anova()` on only the full model, but you need to be careful about the order you add the variables into your multiple linear regression model (to make your life easier). Refit your model (if needed) to add `Mileage` as the first explanatory variable and `Make` as the second explanatory variable. Then run `anova()` and only pass in your full model. You should see an ANOVA table with rows `Mileage`, `Make`, and `Residuals`, in that order. Verify that the F-statistic for `Make` is the same as the F-statistic calculated in the previous task.

Adding additional categorical variables.

Optional task. If you get done with the above tasks before the end of class, revisit the exploratory plots you created in Task 1 and expand your multiple linear regression model to include the other potentially useful categorical variables. Think about how they can be interpreted and determine whether they discernibly improve the model.