1 newpage

2 Data Science: Bridging Principle and Practice

2.1 Part 8: Model Selection (Bike Sharing case study)

2.2 8. Model Selection

2.3 8b. Finding The Best Model

EXERCISE:Use the widget to look for more accurate models. Try at least 10 combinations of features, and make sure to track which model seems to be performing the best. Note: it may take a few seconds for the graphs to update after you click the buttons. Record the parameters for your best model in the cell below.

NOTE: there are many possible answers to this question, because there are many possible definitions of "best" and "most accurate" models. Below is an example of the way your answer should be formatted- the exact variables and beta vector you settled on may be different.

```
In [6]: best_expl_vars = ["is work day", "weather", "temp"]

        # replace the ellipses with True if your model used an intercept term
        # otherwise, replace the ellipses with False
        best_intercept = True

        # replace the ellipses with the name of the response variable
        best_response_var = "casual riders"

        # replace the ellipses with a list of the coefficients in the beta vector
        best_beta = [ 638.25630065, -826.77232767, -168.03010361, 2048.4788933 ]
```

QUESTION: What was your approach to finding a better model? Explain which variables you tried and why, as well as what metrics showed it was the "better". Reference the scatter plots, fit lines, RMSE, etc, and record the explanatory and response variables for your best model.

Answers will vary here, depending on which variables you tried. Hopefully you touched on at least a few of the following points:

- a lower RMSE will typically indicate a better model. RMSE is usually slightly lower for the training data than for the validation data, because the model was fitted using the former but not the latter. So, lower RMSE on validation data will typically indicate a model that will perform well on new data.
- when comparing models that predicted for total riders, casual riders, and registered riders, it's important to keep in mind that each of the possible response variables have different distributions. For instance: the maximum number of casual riders is around 3500, while the maximum number of total riders is over 8000. Therefore, a model predicting the number of casual riders will usually have a lower RMSE than a model predicting total riders, simply because there are generally less casual riders, and not necessarily because it is a "better" model.
- when the predictions of models are plotted against the actual values (as in the top two plots), more accurate linear models will have points that cluster more closely around the fit lines
- many explanatory variables have non-linear relationships with the response variable. We can see this because for most of our linear models, the errors are not evenly distributed above and below the horizontal line at 0
- For some explanatory variable combinations, you may have generated strange-looking scatter plots where all the points were clustered in a few narrow bands. This occurs if you only choose a few (1-3ish) explanatory variables, and those variables are categorical. For example, "is holiday" only has

two possible values: 0 and 1. If we make a model with only "is holiday" as our explanatory variable, then our model looks like:

$$y_{predicted} = \beta_0 * is\_holiday$$

$\beta_0$is just one number, since there is just one explanatory variable. Because "is holiday" can only be 0 or 1, we can then only get two possible predictions for numbers of riders:β_0* 0 orβ_0* 1. This leads to those narrow horizontal bands of predicted values.

QUESTION:How did your model perform on the test data compared to how it performed on the training and validation data? Given what you know about how models "learn" from data, are the results you saw in line with your expectations?

Answers will vary depending on the explanatory variables chosen. Typically we would expect the model to perform better on the training data than on the test data, as about as well on the validation data as the test data. This is because the model was fitted (that is, the values of beta were determined) using the training data but not the validation or test data. Because the model has "seen" the training data before, we would expect it to make better predictions.

Ideally, the validation data acts as a measurement for how accurately the model can predict for data it *hasn't* been trained on, so the test data performance should be similar to the validation data performance.

HOWEVER, there is always a possibility that, due to random chance, the model actually performs *better* on the test data than on training data.