Categorical Variables in Regression

University of California, San Diego
Instructor: Armin Schwartzman

Dataset

- We use again the `04cars` dataset. (For the purposes of introducing this material, we use `04cars.version2` instead.) We now focus on the following variables:
  - `mpg` : Highway MPG
  - `type` : Vehicle Type
    (1 = SportsCar, 2 = SportUtility, 3 = Wagon, 4 = Minivan, 5 = Pickup)
  - `drive` : Drive Train
    (0 = Front-Wheel, 1 = Rear-Wheel, 2 = All-Wheel)
  - `cyl` : Number of Cylinders
  - `hp` : Horsepower
  - `wt` : Weight
- Our goal is still to explain `mpg` as a function of the other variables.

Dataset

- The variables are of different types:
  - `hp` and `wt` are numerical
  - `type` and `drive` are categorical
  - `cyl` is discrete ordinal
- Categorical variables are often called factors and the different values they take levels. For example, `drive` is a factor with 3 levels.

Discrete variables: numerical or categorical?

- Say we want to regress `mpg` on `cyl` only.
- The variable `cyl` is discrete and ordinal.
- We have (at least) two choices:
  1. Consider `cyl` numerical and proceed as usual.
  2. Consider `cyl` categorical and introduce indicator variables.

Discrete variables as numerical variables

- When a discrete variable is ordinal, then it can be considered numerical:

$$\mathbb{E}(\texttt{mpg}|\texttt{cyl}) = \beta_0 + \beta_1 \texttt{cyl}$$

- This implicitly constrains the difference in gas consumption to be the same (on average) between cars with 8 cylinders and cars with 6 cylinders, and cars with 6 cylinders and cars with 4 cylinders.

Discrete variables as categorical variables

- We introduce indicator (also called dummy) variables:
  - `cyl6` $= 1$ if `cyl` $=$ 6 and 0 otherwise
  - `cyl8` $= 1$ if `cyl` $=$ 8 and 0 otherwise

A categorical variable with ℓ levels requires $\ell - 1$ indicator variables.

- Effectively, a categorical variable divides the sample into subgroups according to the different categories. For example, `cyl` partitions the dataset according to the number of cylinders.

- The regression is then `mpg` on `cyl6` and `cyl8`:

$$\mathbb{E}(\texttt{mpg}|\texttt{cyl}) = \beta_0 + \beta_1 \texttt{cyl6} + \beta_2 \texttt{cyl8}$$

  Put differently, we fit one mean per group (according to `cyl`).

  For example, the mean mpg for cars with `cyl` $=$ 4 is β_0, while the mean mpg for cars with `cyl` $=$ 6 is $\beta_0 + \beta_1$.

- This is the same model as (but expressed differently)

$$\mathbb{E}(\texttt{mpg}|\texttt{cyl}) = \gamma_1 \texttt{cyl4} + \gamma_2 \texttt{cyl6} + \gamma_3 \texttt{cyl8}$$

- The main question is whether there is a difference in group means.
- The situation may be visualized using side-by-side boxplots.
- The question may be formalized as a hypothesis testing problem for $\beta_1 = \beta_2 = 0$. The test of choice is the ANOVA F-test.

Two categorical predictors without interactions

- Suppose we want to regress `mpg` on `cyl` and `drive` only.
- Consider the model *without* interactions:

$$\mathbb{E}(\texttt{mpg}|\texttt{cyl},\texttt{drive}) = \beta_0 + \beta_1\texttt{drive1} + \beta_2\texttt{drive2} + \beta_3\texttt{cyl6} + \beta_4\texttt{cyl8}$$

- It implicitly assumes that (expected) `mpg` increases with `drive` in the same way regardless of `cyl`. In that case, we say that there are no interaction between factors `drive` and `cyl` (in the model).

Two-way ANOVA table

- Sequential model comparison using F-tests:
  - 1st row: $\boxed{1}$ versus $\boxed{\texttt{drive}}$
  - 2nd row: $\boxed{\texttt{drive}}$ versus $\boxed{\texttt{drive} + \texttt{cyl}}$
- Note that R uses the SSE from the full (last) model in both tests.
- Also note that the order matters since `drive` and `cyl` are not orthogonal (as predictor vectors).

Two categorical predictors with interactions

- Consider the corresponding model *with* interactions:

$$\begin{aligned}\mathbb{E}(\texttt{mpg}|\texttt{cyl},\texttt{drive}) = \beta_0 + \beta_1\texttt{drive1} + \beta_2\texttt{drive2} + \beta_3\texttt{cyl6} + \beta_4\texttt{cyl8} \\ + \beta_5\texttt{drive1} * \texttt{cyl6} + \beta_6\texttt{drive1} * \texttt{cyl8} \\ + \beta_7\texttt{drive2} * \texttt{cyl6} + \beta_8\texttt{drive2} * \texttt{cyl8}\end{aligned}$$

- In this model accounts for a different line per group defined by `cyl`. The only thing that ties the models together is in the error: when the model is homoscedastic, the variance of the residual error is the same across group.
- The presence of interactions may be visualize using an interaction plot.

Two-way ANOVA table

- Sequential model comparison using F-tests:
  - 1st row: $\boxed{1}$ versus $\boxed{\texttt{drive}}$
  - 2nd row: $\boxed{\texttt{drive}}$ versus $\boxed{\texttt{drive} + \texttt{cyl}}$
  - 3rd row: $\boxed{\texttt{drive} + \texttt{cyl}}$ versus $\boxed{\texttt{drive} + \texttt{cyl} + \texttt{drive} * \texttt{cyl}}$

(Same comments as before.)

Numerical and categorical predictors w/ interactions

- Interactions may be defined between all kinds of variables.

- Suppose we want to regress `mpg` on `wt` and `drive` allowing for possibly different increases with `wt` within each engine type `drive`:

$$\begin{aligned}\mathbb{E}(\texttt{mpg}|\texttt{cyl},\texttt{drive}) = \beta_0 &+ \beta_1\texttt{wt} \\ &+ \beta_2\texttt{drive1} + \beta_3\texttt{drive2} \\ &+ \beta_4\texttt{wt}*\texttt{drive1} + \beta_5\texttt{wt}*\texttt{drive2}\end{aligned}$$

ANCOVA (Analysis of Covariance) table

- Sequential model comparison using F-tests:

  - 1st row: $\boxed{1}$ versus $\boxed{\texttt{wt}}$
  - 2nd row: $\boxed{\texttt{wt}}$ versus $\boxed{\texttt{wt}+\texttt{drive}}$
  - 3rd row: $\boxed{\texttt{wt}+\texttt{drive}}$ versus $\boxed{\texttt{wt}+\texttt{drive}+\texttt{wt}*\texttt{drive}}$

  (Same comments as before.)

- Since `wt` is numerical (taking a comparatively large number of different values), the last model is *not* the most complex model we can fit with `wt` and `drive`.