Indicator Variables for Categorical Variables

Example: On StatCrunch, we have data for 61 roller coasters. We'd like to predict the "***Duration***" of a ride based on any number of *x*-variables.

Check conditions and run all possible regressions. Check for leverage points and large Studentized residuals. Come up with a final, decent model.

- No x's vs y = Duration are curved so OK to fit all models
- Fit all models, best by R^2_{adj} has x_1 = Drop
  x_2 = Length
  x_3 = Year Opened
  but Drop, year opened not significant.
- Drop, length ✗, ✓ , Height, Length ✗, ✓ , Speed, Length ✗, ✓
- So only x term that is significant is x = length.
- Check residuals v. Predicted → No pattern ✓ Equal spread. ✓
- Check boxplot of residuals → No outliers ✓
- Check boxplot of Cook's, 7 potential leverage points.
- Large Cook's is $\frac{4}{n} = \frac{4}{61} = 0.066$

  3 Coaster's exceed it

  Milenium Force, Canyon Blaster

  Nitro

- Histogram with these 3 coasters highlighted (for Length) shows nothing that unusual.
- Scatterplot with these 3 highlighted (y = Duration, x = Length) Doesn't show much either.
- Set indicator variables for these 3, add to model, run mult. linear model

with
- x_1 = Length (sig)
- x_2 = Milenium Force (not sig)
- x_3 = Nitro (sig)
- x_4 = Canyon Blaster (sig)

at $\alpha = 0.05$ level

Remove Milenium Force, rerun model:

$$\widehat{\text{Duration}} : 63.22 + 0.0208(\text{Length (feet)}) + 64.60(\text{Nitro}) - 53.61(\text{Canyon Blaster})$$

Nitro is exceptionally long
Canyon is exceptionally short
> Compared to rest when using the length to predict Duration.

Now, we have a hunch that roller coasters with inversion might have a different durations that coasters without it.

Make scatterplots to make sure both relationships are still linear (with and without inversion).

Scatterplot → Group by inversion, put on same plot.
Slopes look similar, but coasters with inversion seem to be shifted up (same slope, different intercept)

Run separate linear regressions for the coasters with and without inversion. Compare the slopes.

Simple Linear
y = Duration
x = Length

With: Slope on length: 0.0299
Without: Slope on length: 0.0265

Slopes similar

Since the slopes are similar, we can add an indicator variable to our model, using both "***Length***" and "***Inversion***" to predict the "***Duration***" of a roller coaster ride:

Need to code "Inversion" as
yes = 1
no = 0

DATA | Compute Exp | If Else

- When doing prediction, we plug in a value of 0 for coasters without inversion, and a value of 1 for coasters with inversion. The result is an added amount to the "***Duration***" of a ride.

- Predict the duration of a ride for a 5000 foot long roller coaster. Do it for both a coaster with and without inversion.

New Model: y = Duration
X_1 = Length (Sig)
X_2 = Nitro Indicator (Sig)
X_3 = Canyon Indicator (Sig)
X_4 = Inversion Indicator (Sig)

Save new predicted, residuals, Cook's.
(computer glitch on one row?)

Equal Spread, no pattern: Res. v. Predicted

Boxplot of Residuals: 2 Coasters
now potential
larger outliers
(should investigate---)

Boxplot of Cook's: 1 possible new
influence point
(should investigate).

"Finalized" Model:

$$\widehat{\text{Duration}} = 26.83 + 0.027(\text{Length}) + 68.65(\text{Nitro}) - 64.12(\text{Canyon}) + 32.37(\text{Inversion})$$

We'd expect coasters with inversion to last ~ 32 seconds longer.

5000 foot coaster prediction:

With Inversion: $\widehat{\text{Duration}} = 26.83 + 0.027(5000) + 68.65(0) - 64.12(0) + 32.37(1)$

$= 194.2$ seconds

Without: $\widehat{\text{Duration}} = 26.83 + 0.027(5000) + 0 + 0 + 0 =$

$= 161.83$ seconds.

Model Has $R^2_{ADJ} = 75.28\%$

RMSE = 20.88 seconds

(Actual durations vary about the predicted line by 20.88 seconds)