IDS 702

Linear Regression - 4 (Categorical predictors, outliers)

September 13, 2022
Andrea Lane, PhD

Agenda

1. Pre-class reading questions
2. Road map recap
3. Categorical predictors
4. Outliers/influential points
5. In class analysis

Learning Objectives

By the end of this class, you should be able to:

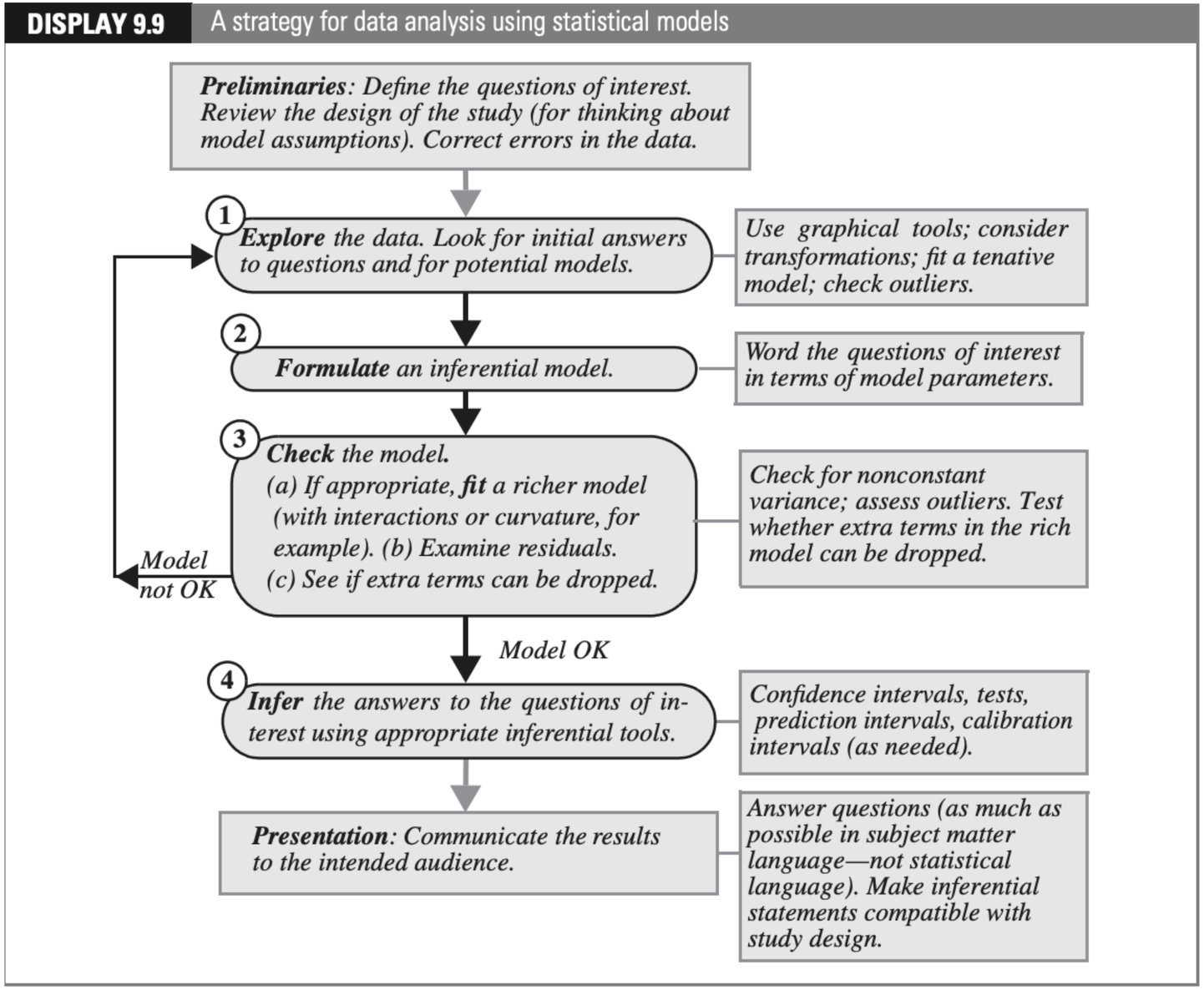
- Interpret regression output with a categorical variable
- Create a factor variable in R
- Differentiate between leverage, influence, and outliers
- Generate diagnostic plots in R

1. Pre-class reading questions

Pre-class reading questions

- What is the difference between a qualitative variable and a quantitative variable?
- What is your approach to EDA for qualitative variables? How can you visualize qualitative variables?
- What is the difference between an outlier and a high leverage point?
- At what point in the data analysis process should you address outliers?

2. Road map recap



Graphic from "The Statistical Sleuth: A Course in Methods of Data Analysis"

MLR Topics:

Last week:
- Estimation
- Inference
- Check assumptions

Today:
- Categorical predictors
- Outliers

Thursday:
- Collinearity
- Interactions

Next week:
- Prediction
- Model selection

3. Categorical Predictors

Categorical variable terms

- Levels: values of a categorical variable
- Binary variable: categorical variable with only two levels
- Factor (in R): categorical variable that stores levels and labels
- Dummy variable: numeric variable that represents a categorical variable
- Reference/baseline level: value to which other values of categorical variable are compared (important for coefficient interpretation)

Binary variable

- Example: binary variable to represent home ownership
- 2 levels: owns a home or does not own a home
- Dummy variable:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

Binary variable in R

- R creates the dummy variable automatically
- Use the factor() function to specify levels and labels
- The regression output will give a single coefficient estimate, t-value, p-value, etc.
- The reference level is based on the level (lowest numeric or first alphabetically) but can be changed with the relevel() function

Categorical variables (>2 levels)

- A single dummy variable cannot represent all values
- We also cannot have a dummy variable for every level (coefficients cannot be estimated uniquely in this case)
- We need # levels -1 dummy variables for each categorical variable
- Example: region (East, West, South) (What's the reference category?)

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West.} \end{cases}$$

Nested F test / type III test

- We may want to assess the association between a categorical variable and the outcome
- Since we have dummy variables, this requires testing if a subset of the coefficients are equal to 0
- This is called a nested F test or Type III test
- The test compares a reduced model to the full model

Let’s see it in R

- Load the Auto dataset from the ISLR2 package

`mpg` miles per gallon
`cylinders` Number of cylinders between 4 and 8
`displacement` Engine displacement (cu. inches)
`horsepower` Engine horsepower
`weight` Vehicle weight (lbs.)
`acceleration` Time to accelerate from 0 to 60 mph (sec.)
`year` Model year (modulo 100)
`origin` Origin of car (1. American, 2. European, 3. Japanese)
`name` Vehicle name

4. Outliers and influential points

Outliers

- Individual observations can have a large impact on the model (estimates, SE, R^2, RSE)
- Sometimes the points are obvious from EDA, but other times they are not
- An **outlier** is a data point whose value does not follow the general trend of the rest of the data

Leverage

- Points with extreme **predictor/covariate/feature** values are called high leverage points
- The **leverage score** measures how far away the values of the independent variables for the ith observation are from those of other observations
- The leverage score for an observation is defined as the ith diagonal element of the projection/hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}$
- $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^{n} h_{ii} = p + 1$
- High leverage points are often determined by paying attention to any observation for which $h_{ii} > 2(p+1)/n$

High leverage: what to do?

- Make sure they do not result from data entry errors
- Make sure you look at the impact of those points on the estimates: just because a point is high leverage does not mean it will have a large effect on regression!
- When a point has a large effect on the regression, we say the observation is influential
    - This depends on the value of y

Cook's distance

- Quantifies the influence of the ith observation
- $\hat{y_{j(i)}}$ is the predicted value after excluding the ith observation

$$D_i = \sum_{j=1}^{n} \frac{(\hat{y}_j - \hat{y_{j(i)}})^2}{s_e^2(p+1)}$$

Large Cook's distance: what to do?

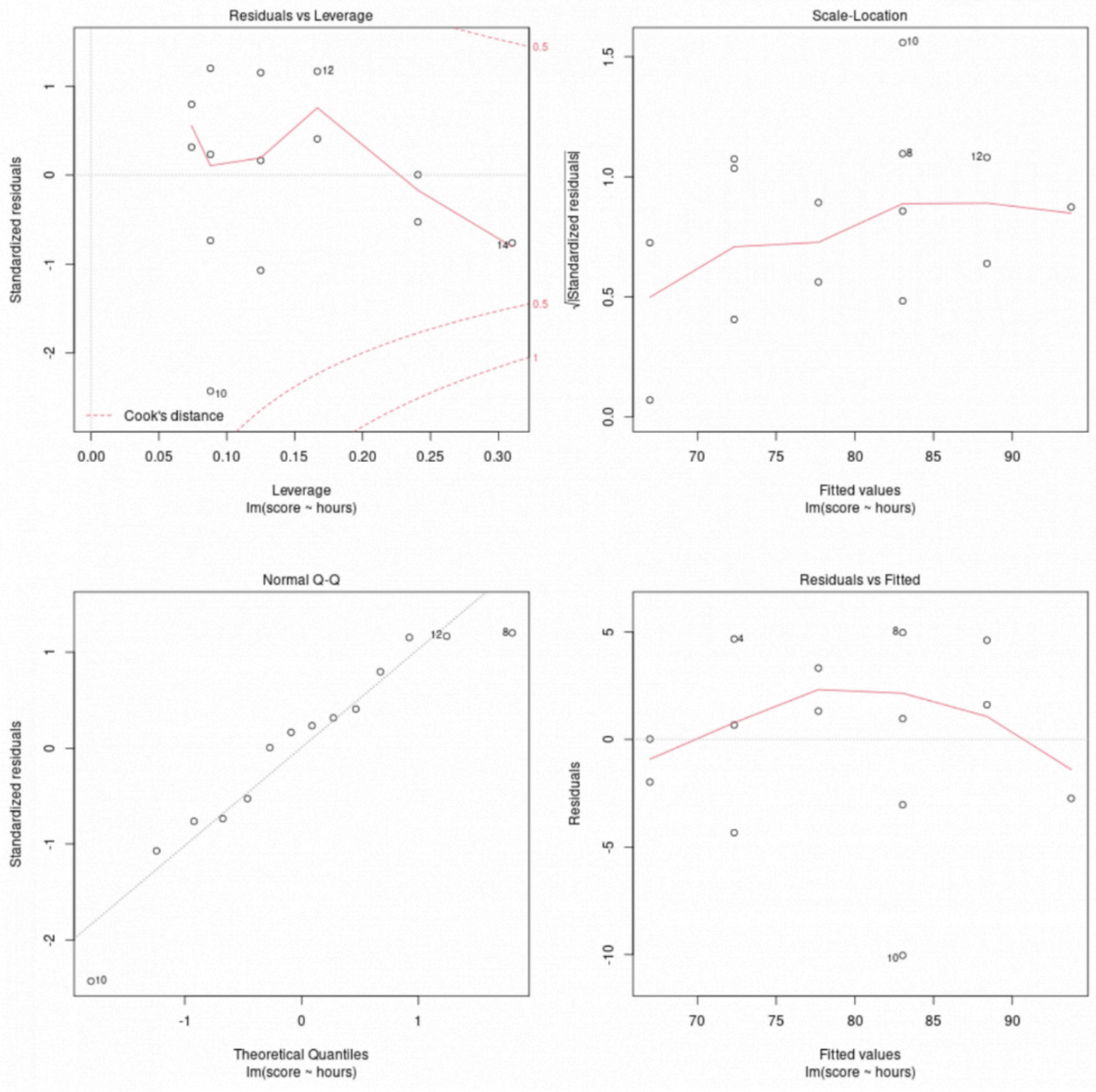
- General consensus is that $D_i > 1$ indicates an observation is an influential value, but we generally pay attention to $D_i > 0.5$
- For each observation with high Cook's distance, fit the model with and without that observation, and compare the results

Standardized/studentized residuals

- Divide residual by SE to be comparable
- Values with large standardized residuals are outliers, but not necessarily influential on the regression line

Diagnostic plots in R

plot(model)



Summary: What to do with outliers/influential observations?

- Make sure the observation does not arise from a data entry error
  - If it does, it can be changed or excluded
- May want to report results with and without influential observation(s)

5. In class analysis

Wrap-up

- Data Analysis Assignment 1 due Fri, Sept 16 11:55 PM
- All statistical reflection assignments are now posted (go live at 6:30 PM today)
  - Select from the provided list, but be clear at the top of your reflection which one you chose