Chapter 1 Part 2

04/20/2021

Multiple Linear Regression Modeling

- Let us model the winning speed as a function of time
  - for example, have winning speeds increased at a constant rate since 1896?
  - Let Y_i be the speed of the winning horse in year i.

$$Y_i = \beta_0 + \beta_1 \mathrm{Year}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2).(\#eq : model1) \tag{1}$$

Understanding the Model Equation

$$Y_i = \beta_0 + \beta_1 \text{Year}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2). \tag{2}$$

- β_0 represents the true intercept—the expected winning speed during Year 0
- β_1 represents the true slope—the expected increase in winning speed from one year to the next
  - assuming the rate of increase is linear (i.e., constant with each successive year since 1896)
- The **error** (ϵ_i) terms represent the deviations of the actual winning speed in Year i (Y_i) from the expected speeds under this model ($\beta_0 + \beta_1 \text{Year}_i$)
  - the part of a horse's winning speed that is not explained by a linear trend over time.
- The variability in these deviations from the regression model is denoted by σ^2.

Fitting the Model

- The parameters in this model (β_0, β_1, and σ^2) can be estimated through *ordinary least squares methods*
- hats denote estimates of population parameters based on empirical data. - Values for $\hat{\beta}_0$ and $\hat{\beta}_1$ are selected to minimize the sum of squared residuals
  - A **residual** is simply the observed prediction error—the actual winning speed for a given year minus the winning speed predicted by the model.

Our Notation

- Predicted speed: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{Year}_i$
- Residual (estimated error): $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- Estimated variance of points around the line: $\hat{\sigma}^2 = \sum \hat{\epsilon}_i^2/(n-2)$

Lets fit our model in R!

Kentucky Derby Model

- The model:

$$\hat{Y}_i = 2.05 + 0.026\text{Year}_i$$

- The model estimates
  - $\hat{\beta}_0 = 2.05$
  - $\hat{\beta}_1 = 0.026$
  - $\hat{\sigma} = 0.90$

Interpretation

- The model estimates
  - $\hat{\beta}_0 = 2.05$
  - $\hat{\beta}_1 = 0.026$
- Winning horses of the Kentucky Derby have an estimated winning speed of 2.05 ft/s in Year 0 (more than 2000 years ago!)
- What do we call this number?
- What are some considerations before interpreting this number?
- Winning speed improves by an estimated 0.026 ft/s every year.
  - What do we call this number?
  - What are some considerations before interpreting this number?

More Interpretation

- R^2 of 0.513, the regression model explains a moderate amount (51.3%) of the year-to-year variability in winning speeds
- The trend toward a linear rate of improvement each year is statistically significant at the 0.05 level ($t(120) = 11.251$, $p < .001$).

Centering

- We didn't use any of our created/modified variables in this model
- In our first model, the intercept has little meaning in context, since it estimates a winning speed in Year 0, but the first Kentucky Derby run at the current distance (1.25 miles) was in 1896.
- One way to create more meaningful parameters is through **centering**.
- This is one reason why we create a centered variable for modeling. What other reason might we have for often centering year variables?

A New Model

$$Y_i = \beta_0 + \beta_1 \text{Yearnew}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2)$$
$$\text{and} \quad \text{Yearnew} = \text{Year} - 1896.$$

- What is the only change?
- What will change about our model coefficients?
- Once we fit our new model we can compare the two lines they define.
- Lets fit the new model in R and create a graph that allows us to compare both models.

Centerings Affect

- The only thing that changes from Model 1 to Model 2 is the estimated intercept;
- $\hat{\beta}_1$, R^2, and $\hat{\sigma}$ all remain exactly the same.
- $\hat{\beta}_0$ tells us that the estimated winning speed in 1896 is 51.59 ft/s,
- That is, estimates of the linear rate of improvement or the variability explained by the model remain the same.
- As in our comparison graph, centering year has the effect of shifting the y-axis from year 0 to year 1896, but nothing else changes.

Assumptions

- We need to verify that our LINE linear regression model assumptions fit for Model 2 if we want to make *inferential statements* (hypothesis tests or confidence intervals) about parameters or predictions.
- Most of these assumptions can be checked graphically using a set of residual plots (plot in R)
- Lets generate them in R

Using the Plots

- The upper left plot, Residuals vs. Fitted, can be used to check the Linearity assumption. Residuals should be patternless around $Y = 0$; if not, there is a pattern in the data that is currently unaccounted for.
- The upper right plot, Normal Q-Q, can be used to check the Normality assumption. Deviations from a straight line indicate that the distribution of residuals does not conform to a theoretical normal curve.
- The lower left plot, Scale-Location, can be used to check the Equal Variance assumption. Positive or negative trends across the fitted values indicate variability that is not constant.
- The lower right plot, Residuals vs. Leverage, can be used to check for influential points. Points with high leverage (having unusual values of the predictors) and/or high absolute residuals can have an undue influence on estimates of model parameters.

Well, Are our Conditions Met?

- Why or why not? What passes? Fails?

Condition Comments

- Residuals vs. Fitted plot indicates that a quadratic fit might be better than the linear fit of Model 2
- Other assumptions look reasonable
- Influential points would be denoted by high values of Cook's Distance; they would fall outside cutoff lines in the northeast or southeast section of the Residuals vs. Leverage plot. Since no cutoff lines are even noticeable, there are no potential influential points of concern.

Influential Points, Leverage, and Cooks Distance

- Video. Click HERE

Graphs!

- We will rely on graphical evidence for identifying regression model assumption violations, looking for highly obvious violations of assumptions before trying corrective actions.
- While some numerical tests have been devised for issues such as normality and influence, most of these tests are not very reliable, highly influenced by sample size and other factors.
- There is typically no residual plot, however, to evaluate the Independence assumption; evidence for lack of independence comes from knowing about the study design and methods of data collection.
- In this case, with a new field of horses each year, the assumption of independence is pretty reasonable.