Chapter 1 Part 3

04/20/2021

New Model

Based on residual diagnostics, we should test a new model, in which a quadratic term is added to the linear term in Model 2.

$$Y_i = \beta_0 + \beta_1 \text{Yearnew}_i + \beta_2 \text{Yearnew}_i^2 + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2).$$

- What could this model suggest about how the *rate* of increase of horse speeds is changing over time?
- Lets fit this model in R!

Comparing Modles

There is evidence that the quadratic model improves upon the linear model - What is evidence of this?

Comparisons

- Visually the quadratic looks like a better fit
- The proportion of year-to-year variability in winning speeds explained by the model, has increased from 51.3% to 64.1% (R^2)
- The pattern in the Residuals vs. Fitted plot of Figure has disappeared
  - Any other concerning patterns?
- Normality is a little sketchier in the left tail
- The large negative coefficient for β_2 suggests that the rate of increase is indeed slowing in more recent years.

Linear Regression with a Binary Predictor

- We also may want to include track condition as an explanatory variable.
- We could start by using `fast` as the lone predictor:
  - Do winning speeds differ for fast and non-fast conditions?
  - `fast` is considered an **indicator variable**—it takes on only the values 0 and 1,
  - 1 indicates presence of a certain attribute (like fast racing conditions). - Since `fast` is numeric, we can use simple linear regression techniques to fit a model

Binary Predictor Model

$$Y_i = \beta_0 + \beta_1 \text{Fast}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2). \tag{1}$$

- Here, it's easy to see the meaning of our slope and intercept by writing out separate equations for the two conditions:
- Good or slow conditions (`fast` $= 0$)

$$Y_i = \beta_0 + \epsilon_i$$

- Fast conditions (`fast` $= 1$)

$$Y_i = (\beta_0 + \beta_1) + \epsilon_i$$

Binary Predictor Model

- β_0 is the expected winning speed under good or slow conditions,
- β_1 is the difference between expected winning speeds under fast conditions vs. non-fast conditions.
- Lets fit this in R

```
model3 <- lm(speed ~ fast, data = derby.df)
```

Results

- The estimated winning speed under non-fast conditions is 52.0 ft/s
- mean winning speeds under fast conditions are estimated to be 1.6 ft/s higher.
- Is this any different than a two-sample t test?

Our Model vs T test

- They are the same, testing β_1's significance is the same as a two sample t test
- We do require equal variances

Multiple Linear Regression with Two Predictors

- Now we want to include *Yearnew* and *Fast*

$$Y_i = \beta_0 + \beta_1 \mathrm{Yearnew}_i + \beta_2 \mathrm{Fast}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2). \quad (2)$$

and linear least squares regression (LLSR) provides the following parameter estimates:

- In R

Compare

- How does this compare to our past models?
- How can we make it possibly better?

Inference in Multiple Linear Regression: Normal Theory

- We have been using linear regression for descriptive purposes, which is an important task.
- Next we want to conduct statistical inference as well
  - determining if effects are statistically significant
  - quantifying uncertainty in effect size estimates with confidence intervals
  - quantifying uncertainty in model predictions with prediction intervals.
- Under LINE assumptions, all of these inferential tasks can be completed with the help of the t-distribution and estimated standard errors.
- Lets get some CI's in R!

Examples of Inferential Statements Based on our Latest Model

- We can be 95% confident that average winning speeds under fast conditions are between 0.93 and 1.53 ft/s higher than under non-fast conditions, after accounting for the effect of year.
- Fast conditions lead to significantly faster winning speeds than non-fast conditions ($t = 8.14$ on 119 df, $p < .001$), holding year constant.
- Based on our model, we can be 95% confident that the winning speed in 2017 under fast conditions will be between 53.4 and 56.3 ft/s. Note that Always Dreaming's actual winning speed barely fit within this interval—the 2017 winning speed was a borderline outlier on the slow side.

Inference in Multiple Linear Regression: Bootstrapping

- Remember that you must check LINE assumptions using the same residual plots to ensure that the inferential statements in the previous section are valid.
- In cases when model assumptions are shaky, one alternative approach to statistical inference is **bootstrapping**;
  - bootstrapping is a robust approach to statistical inference that we will use frequently throughout this book because of its power and flexibility.
  - In bootstrapping, we use only the data we've collected and computing power to estimate the uncertainty surrounding our parameter estimates.
    - Our primary assumption is that our original sample represents the larger population, and then we can learn about uncertainty in our parameter estimates through repeated samples (with replacement) from our original sample.

CI's Using Bootstrapping

To generate a bootstrap CI for out latest model, we can follow these steps:

- take a (bootstrap) sample of 122 years of Derby data with replacement, so that some years will get sampled several times and others not at all.
  - This is **case resampling**, so that all information from a given year (winning speed, track condition, number of starters) remains together.
- fit model 4 to the bootstrap sample, saving $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.
- repeat the two steps above a large number of times (say 1000).

CI's Using Bootstrapping

- the 1000 bootstrap estimates for each parameter can be plotted to show the **bootstrap distribution** .
- a 95% confidence interval for each parameter can be found by taking the middle 95% of each bootstrap distribution—i.e., by picking off the 2.5 and 97.5 percentiles. This is called the **percentile method**.
- Lets do this!

Comparing Theoretical to Bootstrap

- In this case, we see that 95% bootstrap confidence intervals for β_0, β_1, and β_2 are very similar to the normal-theory confidence intervals we found earlier.
- The normal-theory confidence interval for the effect of fast tracks is 0.93 to 1.53 ft/s, while the analogous bootstrap confidence interval is 0.91 to 1.57 ft/s.
- There are many variations on this bootstrap procedure.
  - You could sample residuals rather than cases
  - You could conduct a parametric bootstrap in which error terms are randomly chosen from a normal distribution
- Researchers have devised other ways of calculating confidence intervals besides the percentile method, including normality, studentized, and bias-corrected and accelerated methods (@Hesterberg2015; @Efron1993; @Davison1997).
- We will focus on case resampling and percentile confidence intervals for now for their understandability and wide applicability.

Multiple Linear Regression with an Interaction Term

- One limitation of model 4, however, is that we assumed that the effect of track condition has been the same for 122 years, or conversely that the yearly improvements in winning speeds are identical for all track conditions.
- To expand our modeling capabilities to allow the effect of one predictor to change depending on levels of a second predictor, we need to consider **interaction terms**.
- If we create a new variable by taking the product of `yearnew` and `fast` (i.e., the **interaction** between `yearnew` and `fast`), adding that variable into our model will have the desired effect.

Consider Model 5

$$Y_i = \beta_0 + \beta_1 \text{Yearnew}_i + \beta_2 \text{Fast}_i + \beta_3 \text{Yearnew}_i \times \text{Fast}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2)$$

where LLSR provides the following parameter estimates: - Lets fit this in R!

Our New Model

$$\hat{Y}_i = 50.53 + 0.031\text{Yearnew}_i + 1.83\text{Fast}_i - 0.011\text{Yearnew}_i \times \text{Fast}_i. \tag{3}$$

- Interpretations of model coefficients are most easily seen by writing out separate equations for fast and non-fast track conditions:

$$\text{Fast} = 0:$$
$$\hat{Y}_i = 50.53 + 0.031\text{Yearnew}_i$$
$$\text{Fast} = 1:$$
$$\hat{Y}_i = (50.53 + 1.83) + (0.031 - 0.011)\text{Yearnew}_i$$

Interpretations

Interpretations for estimated model coefficients:

- $\hat{\beta}_0 = 50.53$. The expected winning speed in 1896 under non-fast conditions was 50.53 ft/s.
- $\hat{\beta}_1 = 0.031$. The expected yearly increase in winning speeds under non-fast conditions is 0.031 ft/s.
- $\hat{\beta}_2 = 1.83$. The winning speed in 1896 was expected to be 1.83 ft/s faster under fast conditions compared to non-fast conditions.
- $\hat{\beta}_3 = -0.011$. The expected yearly increase in winning speeds under fast conditions is 0.020 ft/s, compared to 0.031 ft/s under non-fast conditions, a difference of 0.011 ft/s.

Summarizing Interactions

- Interaction allows us to model the relationships we noticed earlier where both the intercept and slope describing the relationships between `speed` and `year` differ depending on whether track conditions were fast or not.
- Note that we interpret the coefficient for the interaction term by comparing slopes under fast and non-fast conditions; this produces a much more understandable interpretation for a reader than attempting to interpret the -0.011 directly.

Building a Multiple Linear Regression Model

- We now begin iterating toward a "final model" for these data, on which we will base conclusions.
- Typical features of a "final multiple linear regression model" include:
  - explanatory variables allow one to address primary research questions
  - explanatory variables control for important covariates
  - potential interactions have been investigated
  - variables are centered where interpretations can be enhanced
  - unnecessary terms have been removed
  - LINE assumptions and the presence of influential points have both been checked using residual plots
  - the model tells a "persuasive story parsimoniously"

The Process

- Although the process of reporting and writing up research results often demands the selection of a sensible final model, it's important to realize that a statisticians typically will examine and consider an entire taxonomy of models when formulating conclusions, and b different statisticians sometimes select different models as their "final model" for the same set of data. Choice of a "final model" depends on many factors, such as
  - primary research questions,
  - purpose of modeling,
  - tradeoff between parsimony and quality of fitted model,
  - underlying assumptions, etc.

More on Modeling

- Modeling decisions should never be automated or made completely on the basis of statistical tests
- Subject area knowledge should always play a role in the modeling process.
- You should be able to defend any final model you select, but you should not feel pressured to find the one and only "correct model",
- Most good models will lead to similar conclusions.

Model Comparisons

Several tests and measures of model performance can be used when comparing different models for model building:

- R^2. Measures the variability in the response variable explained by the model. One problem is that R^2 always increases with extra predictors, even if the predictors add very little information.
- adjusted R^2. Adds a penalty for model complexity to R^2 so that any increase in performance must outweigh the cost of additional complexity. We should ideally favor any model with higher adjusted R^2, regardless of size, but the penalty for model complexity (additional terms) is fairly ad-hoc.

Model Comparisons

- AIC (Akaike Information Criterion). Again attempts to balance model performance with model complexity, with smaller AIC levels being preferable, regardless of model size.
- The BIC (Bayesian Information Criterion) is similar to the AIC, but with a greater penalty for additional model terms.

Model Comparisons

- extra sum of squares F test.
  - This is a generalization of the t-test for individual model coefficients which can be used to perform significance tests on **nested models**,
  - One model is a reduced version of the other.
- For example, we could test whether our final model (below) really needs to adjust for track condition, which is comprised of indicators for both fast condition and good condition (leaving slow condition as the reference level).
- Our null hypothesis is then $\beta_3 = \beta_4 = 0$. We have statistically significant evidence (F $= 57.2$ on 2 and 116 df, p $< .001$) that track condition is associated with winning speeds, after accounting for quadratic time trends and number of starters.

Potential Final Model

One potential final model for predicting winning speeds of Kentucky Derby races is:

$$\begin{aligned} Y_i = \beta_0 + \beta_1 \text{Yearnew}_i + \beta_2 \text{Yearnew}_i^2 + \beta_3 \text{Fast}_i \\ + \beta_4 \text{Good}_i + \beta_5 \text{Starters}_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2) \end{aligned} \tag{4}$$

and LLSR provides the following parameter estimates: - Lets fit this model in R!

Describing this Model

- Accounts for the slowing annual increases in winning speed with a negative quadratic term
- Adjusts for baseline differences stemming from track conditions
- Suggests that, for a fixed year and track condition, a larger field is associated with slower winning times
  - unlike the positive relationship we saw between speed and number of starters in our exploratory analyses.
- The model explains 82.7% of the year-to-year variability in winning speeds, - Residual plots show no serious issues with LINE assumptions.
- We tested interaction terms for different effects of time or number of starters based on track condition, but we found no significant evidence of interactions.