Multiple regression and R-squared

Michael Lopez, Skidmore College

Preamble

Improve with the following commands in live coding

1. group_by()
2. summarize()
3. mutate()

Sample analysis from class

```
library(Lahman)
library(tidyverse)

HallOfFame %>% head()

### Track counts of categorical variables
HallOfFame %>% count(category)

### Random sample of rows
HallOfFame %>% sample_n(6)

HallOfFame %>% count(inducted)

### Question 1: Total number of people inducted in each year
HallOfFame %>% count(yearID, inducted)  ## counts of year/induction status that exist

summary_elect_year <- HallOfFame %>%
  group_by(yearID) %>%
  summarise(number_inducted = sum(inducted == "Y"))

### Two equals signs: asking a question if inducted = "Y"

ggplot(data = summary_elect_year, aes(number_inducted)) +
  geom_histogram(binwidth = 1)

## What happened with the outlier?
summary_elect_year %>% arrange(-number_inducted)

ggplot(data = summary_elect_year, aes(x = yearID, y = number_inducted)) +
  geom_point() +
  geom_line() ## connect points with a line
```

Overview

In this lab, we'll learn model comparison tools using multivariate regression. Next, we'll apply our tools to derive predictions of pitcher performance.

First, recall that we have to load the requisite libraries that we'll need (and we may have to install them, too. As always, once a package is downloaded, you do not need to run the `install.packages()` code again.

```
library(Lahman)
library(tidyverse)
```

We're going to start by using the `Teams` data.

```
data(Teams)
head(Teams)
tail(Teams)
```

```
Teams.1 <- filter(Teams, yearID >= 1970)
head(Teams.1)
```

Comparing multiple regression models.

There's an old saying in statistics, attributed to George Box: *all models are wrong, some are useful.* In practice, we never know if our regression model is correctly specified; e.g, that it is really the case the y, x_1, $\ldots$ and x_{p-1} are linearly related. All we can do is hope…and try a few analytical tools.

Let's try to come up with a few models of `RA`: runs against. First, we start with a recap of multiple regression.

```
library(broom)
options(digits = 3)
fit.1 <- lm(RA ~ HRA + BBA + SOA + HA + attendance, data = Teams.1)
tidy(fit.1)
```

1. Write the estimated model above.
2. Using the model in question (1), interpret the coefficient on `HRA`.
3. Using the model in (1), interpret the coefficient on `attendance`. Then come up with a better way to interpret the coefficient on `attendance`.
4. Remove `HRA` from the model and re-fit. Do your other coefficients change? Can you explain the difference?

After fitting a linear regression model, it is appropriate to check assumptions. First, we check the appropriateness of the normal distribution for residuals.

```
qqnorm(fit.1$resid)
qqline(fit.1$residuals)
```

Next, we compare the residuals to the fitted values, checking for the assumptions of independence among the residuals, as well as the constant variance assumption.

```
Teams.1 <- Teams.1 %>%
  mutate(fitted.value = predict(fit.1),
         residuals = RA - fitted.value)
```

In the code above, `fitted.value` stores the predictions for each row of the data set, generated from `fit.1`. In more technical terms, these are the $\hat{y}$'s.

Here's a plot of the

```
ggplot(data = Teams.1, aes(x = fitted.value, y = residuals)) +
         geom_point() +
  geom_smooth()
```

5. What do the residual plots suggest about the assumptions of our linear regression model? What about the model makes it possibly a poor fit?

Speaking of residuals, lets take a deeper look at individual predictions.

The 1970 Atlanta Braves allowed 185 home runs, 478 walks, struck out 960 batters, and gave up 1451 hits. Their attendance was 1078848. As it turns out, the Braves are the first row of our data set.

```
Teams.1 %>%
  slice(1)
```

6. The `predict` command calculates the fitted number of runs using a model such as ours (our model is `fit.1`. Using the code above, how many runs did our model predict that Braves to have allowed? What is the residual for the number of runs allowed by the Braves? Did our model overestimate or underestimate Atlanta's performance?

7. Take a look at the residuals for all teams in the 2019 season (`yearID` variable). Which team had the highest residual in 2019? Anything noticeable about all of the residuals?

R-squared

There are lots of ways to measure the success of a regression model. The most common metric is `R-squared`, which you'll recall is the fraction of variability in the outcome which is explained by the regression model. Larger R-squared's are, in principal, better.

You can access the R-squared via the `summary` command:

```
summary(fit.1)
```

Using our model, we would interpret the R-squared as follows:

90.09% of the variability in the number of runs allowed by a team can be explained by the linear model with HRA, BBA, SOA, HA, and attendance.

While popular, the traditional R-squared is also flawed. Let's see how. In the following code, we'll create two new variables in the `Teams.1` data set, `rand1` and `rand2`, which are random normal variables.

```
set.seed(0)
Teams.1 <- mutate(Teams.1,
                  rand1 = rnorm(nrow(Teams.1)),
                  rand2 = rnorm(nrow(Teams.1)))
head(Teams.1)
```

Let's see what happens when we include `rand1` and `rand2` to our regression fit.

```
fit.2 <- lm(RA ~ HRA + BBA + SOA + HA + attendance + rand1 + rand2, data = Teams.1)
summary(fit.2)
```

Even when we added random noise to the model, R-squared went up!

That's not a good thing, at least when it comes to making model comparisons. In fact, its a property of R-squared that, no matter what variable you add to a given model, the R-squared cannot go down. As a result, R-squared is not useful for model comparisons, but more to gain a sense of how much of a drop in the variability in the outcome can be explained by the model's fit.

In place of R-squared, R also shows a formula for an adjusted R-squared, which penalizes models for adding unneeded parameters. However, this metric also has weaknesses.

8. When *would* be an appropriate time to compare R-squared's from two different models?

9. What other approaches to picking a model may be more appropriate than R-squared?