Multiple Linear Regression - Examples

Susana Faria

Universidade do Minho

- The Kentucky Derby is a 1.25-mile horse race held annually at the Churchill Downs race track in Louisville, Kentucky.
- Our data set derbyplus.csv contains, for the years 1896-2017:
  - year: the year of the race,
  - winner: the winning horse,
  - condition: the condition of the track,
  - speed: the average speed (in feet per second) of the winner,
  - starters: the number of starters (field size, or horses who raced).
- We would like to use least squares linear regression techniques to model the speed of the winning horse as a function of track condition, field size, and trends over time.

We must create new variables:

- fast: taking the value 1 for races run on fast tracks, and 0 for races run under other conditions,
- good: another indicator variable, taking the value 1 for races run under good conditions, and 0 for races run under other conditions,
- yearnew: a centered variable, where we measure the number of years since 1896,
- fastfactor replaces fast = 0 with the description "not fast", and fast = 1 with the description "fast". Changing a numeric categorical variable to descriptive phrases can make plot legends more meaningful.

```
1
2 derby <- read.csv("derbyplus.csv")
3
4 derby$fast <- ifelse(derby$condition=="fast",1,0)
5 derby$good <- ifelse(derby$condition=="good",1,0)
6 derby$yearnew <- derby$year-1896
7 derby$fastfactor <- ifelse(derby$fast==1,"fast","not fast")
```

With any statistical analysis, our first task is to explore the data, examining distributions of individual responses and predictors using graphical and numerical summaries, and beginning to discover relationships between variables.

This should always be done before any model fitting!

- We will examine the response variable and each potential covariate individually;
- Continuous variables can be summarized using histograms and statistics indicating center and spread;
- Categorical variables can be summarized with tables and possibly bar charts.

```
1
2 summary(derby)
3
4 library(ggplot2)
5
6 speed_hist <- ggplot(data = derby, aes(x = speed)) +
7   geom_histogram(binwidth = 0.5, fill = "white",
8                  color = "black") +
9   xlab("Winning speed (ft/s)") + ylab("Frequency") + labs(title="(a)")
10 speed_hist
11
12 starters_hist <- ggplot(data = derby, aes(x = starters)) +
13   geom_histogram(binwidth = 3, fill = "white",
14                  color = "black") +
15   xlab("Number of starters") + ylab("Frequency") + labs(title="(b)")
16 starters_hist
```

We see that the response variable, winning speed, follows a distribution with a slight left skew, with a large number of horses winning with speeds between 53-55 feet per second and that the number of starters is mainly distributed between 5 and 20, with the largest number of races having between 15 and 20 starters.

The next step in an initial exploratory analysis is the examination of numerical and graphical summaries of relationships between exploratory variables and responses.

- The relationship between two continuous variables is depicted with scatterplots;
- Relationships between categorical variables like track condition and continuous variables can be illustrated with side-by-side boxplots.

```
1
2 gg <- ggpairs(data = derby,
3                columns = c("condition", "year", "starters", "speed"))
4 gg[4,1] <- gg[4,1] + geom_histogram( binwidth = 0.75)
5 gg[2,1] <- gg[2,1] + geom_histogram( binwidth = 20)
6 gg[3,1] <- gg[3,1] + geom_histogram( binwidth = 3)
7 gg
```

- We see that higher winning speeds are associated with more recent years, while the relationship between winning speed and number of starters is less clear cut;
- We also see a somewhat strong correlation between year and number of starters;
- We should be aware of highly correlated explanatory variables whose contributions might overlap too much;
- We see evidence of higher speeds on fast tracks and also a tendency for recent years to have more fast conditions;
- Finally, notice that the diagonal illustrates the distribution of individual variables, using density curves for continuous variables and a bar chart for categorical variables.

We will begin by modeling the winning speed as a function of time. Let Y_i be the speed of the winning horse in year i:

$$Y_i = \beta_0 + \beta_1 \, Year_i + \epsilon_i; \;\; \epsilon_i \sim N(0, \sigma^2)$$

$$model1 <- lm(speed \sim year, data = derby)$$

- According to our simple linear regression model, winning horses of the Kentucky Derby have an estimated winning speed of 2.05 ft/s in Year 0 and the winning speed improves by an estimated 0.026 ft/s every year.
- With an R^2 of 0.513, the regression model explains a moderate amount (51.3%) of the year-to-year variability in winning speeds, and the trend toward a linear rate of improvement each year is statistically significant at the 0.05 level.

In Model 1 that the intercept has little meaning in context, since it estimates a winning speed in Year 0, when 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.

$$model2 <- lm(speed \sim yearnew, data = derby)$$

Note that the only thing that changes from

Model 1 to Model 2 is the estimated intercept.

We should also attempt 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:

$$plot(model2)$$

- 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;
- 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;
- 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.

- 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;

We see:

- 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.

$$model3 <- lm(speed \sim yearnew + I(yearnew^2), data = derby)$$

- There is evidence that the quadratic model improves upon the linear model;
- R^2, the proportion of year-to-year variability in winning speeds explained by the model, has increased to 64.1%;
- The pattern in the Residuals vs. Fitted plot has disappeared, although normality is a little sketchier in the left tail;
- The larger mass of points with fitted values near 54 appears to have slightly lower variability;
- The significantly negative coefficient for β_2 suggests that the rate of increase is indeed slowing in more recent years.

If we pretend to known: Do winning speeds differ for fast and non-fast conditions?

$$Y_i = \beta_0 + \beta_1 \, Fast_i + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, \sigma^2)$$

- Good or slow conditions (fast = 0)

$$Y_i = \beta_0 + \epsilon_i$$

- Fast conditions (fast = 1)

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

$$model4 < -lm(speed \sim fast, data = derby)$$

- β_0: the expected winning speed under good or slow conditions;
- β_1:the difference between expected winning speeds under fast conditions vs. non-fast conditions.

If we simply wanted to compare mean winning speeds under fast and non-fast conditions, why didn't we just run a two-sample t-test? The answer is: we did! The t-test to β_1 is equivalent to an independent-samples t-test under equal variances.

A model with terms for both year and track condition will estimate the difference between winning speeds under fast and non-fast conditions for a fixed year:

$$Y_i = \beta_0 + \beta_1\, yearnew_i + \beta_2\, Fast_i + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, \sigma^2)$$

- Our new model estimates that winning speeds are, on average, 1.23 ft/s faster under fast conditions after accounting for time trends, which is down from an estimated 1.63 ft/s without accounting for time;
- It appears our original model may have overestimated the effect of fast conditions by conflating it with improvements over time;
- Through our new model, we also estimate that winning speeds increase by 0.023 ft/s per year, after accounting for track condition;
- This yearly effect is also smaller than the 0.026 ft/s per year we estimated in Model 1, without adjusting for track condition;
- Based on the R^2 value, Model 5 explains 68.7% of the year-to-year variability in winning speeds, a noticeable increase over using either explanatory variable alone.

- One limitation of Model 5, however, is that we must assume 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;
- Amazingly, 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.

$$Y_i = \beta_0 + \beta_1\ yearnew_i + \beta_2\ Fast_i + \beta_3\ \ yearnew_i \times Fast_i + \epsilon_i \quad \text{with} \ \ \epsilon_i \sim N(0, \sigma^2)$$

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

Fast = 0 :

$$\hat{Y}_i = 50.53 + 0.031 yearnew_i$$

Fast = 1 :

$$\hat{Y}_i = (50.53 + 1.83) + (0.031 - 0.011) yearnew_i$$

leading to the following 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.