Stats 401 Lab 13

Ed Wu

12/7/2018

Outline

- Interpreting Coefficients
- Collinearity
- Interactions

Interpretation: Simple Linear Regression

- Let's first consider the case with 1 predictor
- Suppose we've fit a linear model

$$y_i = b_0 + b_1 x_{i1} + e_i$$

  where, $i = 1, \ldots, n$
- What is the interpretation of the sample coefficient b_0?

Interpretation: Simple Linear Regression

- Recall: we interpret b_1 as the expected change in our response variable for every 1 unit increase in our explanatory variable

Example

We consider pitching data for 2015, which we have obtained using the R package `Lahman`:

```
matrix(names(pitchers),nrow = 3,byrow = T)
```

```
##      [,1]         [,2]     [,3]   [,4]
## [1,] "playerID"   "teamID" "W"    "L"
## [2,] "G"          "GS"     "BB"   "SO"
## [3,] "ERA"        "Kp9"    "IPpG" "starter"
```

For this lab, we will be interested in the last 3 variables: strikeouts per 9 innings (`Kp9`), innings pitched per game(`IPpG`), and a dummy variable for starter (`starter`), which is 1 if the pitcher is a starter and 0 if the pitcher is a reliever.

Example

We fit a model to predict strikeouts per 9 innings (Kp9) using innings pitched per game(IPpG)

```
lm_simple = lm(Kp9 ~ IPpG, data = pitchers)
round(summary(lm_simple)$coefficients,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3912     0.1880 44.6283    0e+00
## IPpG         -0.1937     0.0483 -4.0104    1e-04
```

How can we interpret the coefficient for IPpG?

Interpretation: Multiple Regression

- Now suppose we include a second predictor, x_{i2}, and instead fit the model

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$$

  for $i = 1, \ldots, n$
- Is our interpretation of b_1 the same?

Interpretation: Multiple Regression

- The interpretation is **different**!
- When we include additional predictors, the value of b_1 will change
- We can think of b_1 as the expected change in our response variable for every 1 unit increase in x_1 for a fixed value of x_2

Example (Continued)

Suppose we now predict strikeouts per 9 innings using innings pitched per game and an indicator of whether the pitcher is a starting pitcher

```
lm_mult = lm(Kp9 ~ IPpG + starter, data = pitchers)
round(summary(lm_mult)$coefficients,4)
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.6043     0.2395 31.7452        0
## IPpG               0.9099     0.2210  4.1169        0
## starterStarter    -5.4789     1.0720 -5.1109        0
```

Collinearity

- Collinearity (or Multicollinearity) occurs when explanatory variables are linearly related to each other
- If we can write one variable as a linear combination of the other variables, then we have collinearity
- For example, suppose we have the vectors of predictors $\mathbf{w}$, $\mathbf{x}$, and $\mathbf{z}$, and that $\mathbf{w} = 2\mathbf{x} - 3\mathbf{z}$. Then the predictors in this case are collinear

Why is this a problem?

- When we have collinear variables, it becomes impossible to distinguish between the effects of the variables
  - Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are collinear, and that they both have an effect on $\mathbf{y}$
  - When we try to fit the model $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$, it is impossible to tell which value to assign to b_1 and which to assign to b_2
- This is exactly the problem we saw when we discussed over-specified models with factor variables

Example

Recall our baseball pitchers example, where we fit a model predicting strikeouts per 9 innings using innings pitched per game:

```
##                Estimate Std. Error
## (Intercept)        8.39       0.19
## IPpG              -0.19       0.05
```

Suppose in our data set, we also have a variable "outs per game" and that we accidentally include it in our model. Since outs are equal to innings times 3, outs per game and innings per game are collinear.

Example

- Our sample model is $y_i = 8.39 - 0.19x_{i1} + e_i$ where y_i is strikeouts per 9 innings and x_{i1} is innings pitched per game for pitcher i.
- These coefficients are the unique least squares solution and minimize the sum of square residuals
- Let x_{i2} be outs per game. Because the predictors are collinear ($\mathbf{x}_2 = 3\mathbf{x}_1$), the model $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$ does not have a unique solution!
- There are infinitely many solutions, each of which is just as "correct" as the others – we can't tell which value to assign to b_1 and which to b_2

Lab Activity (Part 1)

We are interested in studying the weights of grapefruit by type:

```
##   weight type circumference
## 1    8.3  red          11.0
## 2    7.0  red          10.0
## 3    7.5 pink          10.6
## 4    9.0  red          10.2
## 5    6.0 pink           7.0
```

Let $\mathbf{x}_1$ be a dummy variable for red grapefruits, $\mathbf{x}_2$ be a dummy variable for pink grapefruits, and $\mathbf{x}_3 = (11, 10, 10.6, 10.2, 7)$ be circumference. Let $\mathbf{y} = (8.3, 7, 7.5, 9, 6)$ be the weights of the grapefruits. For which model(s) would the explanatory variables be collinear?

1. $y_i = b_0 + b_2 x_{i2} + b_3 x_{i3} + e_i$
2. $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + e_i$
3. $y_i = b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + e_i$
4. $y_i = b_0 + b_1 x_{i1} + b_3 x_{i3} + b_4 x_{i3}^2 + e_i$

Approximate Collinearity

- It is also an issue when variables are close to collinear
- When predictors are approximately collinear, it becomes difficult to disentangle the associations of the variables with the outcome variable

Another Way to Think of Collinearity

- Exact collinearity causes the variance of $\hat{\beta}_i$ to be infinite (because there are infinitely many least squares solutions)
- When variables are close to collinear, the variance is not infinite, but very high

Why do we care?

- The sample coefficients are the estimates for the true coefficients
- If the estimates are unreliable, it is difficult to make good inferences
- For example, with a high variance, we would have large confidence intervals and it would be hard to tell if the observed coefficient is statistically significant

Collinearity in R

- When we attempt to fit a model with exact collinearity, R will recognize that the variables are collinear and drop one of the variables

```
pitchers$OpG = 3*pitchers$IPpG
lm_mult = lm(Kp9 ~ IPpG + starter + OpG, data = pitchers)
coef(lm_mult)
```

```
##     (Intercept)            IPpG starterStarter
##       7.6043474       0.9098721      -5.4789241
```

- If the variables are only approximately collinear, then R will still fit the model with all variables – we need to be aware that our coefficient estimates could be unreliable

Example

Almost collinear variables often show up in data sets (especially when there are many predictors). Let's examine the data set from homework 11:

```
senic = read.table("https://ionides.github.io/401f18/hw/hw1
matrix(names(senic),ncol = 3)
```

```
##      [,1]             [,2]         [,3]
## [1,] "Hospital"       "Culture"    "Region"
## [2,] "Length.of.stay" "X.ray"      "Patients"
## [3,] "Age"            "Beds"       "Nurses"
## [4,] "Infection.risk" "Med.school" "Facilities"
```

We can see there are some variables that appear to measure similar things. Specifically, patients, beds, and nurses are all measuring the number of patients in some way.

Example

Using the `cor()` function, we can obtain the pairwise correlations for a matrix or data frame. We can use this to verify our hypothesis above.

```
cor(senic[,c(7,10,11)])
```

```
##                Beds  Patients    Nurses
## Beds      1.0000000 0.9809977 0.9155042
## Patients 0.9809977 1.0000000 0.9078970
## Nurses   0.9155042 0.9078970 1.0000000
```

We can see that these 3 variables are highly correlated. We will examine how this approximate collinearity affects the model fit in the lab ticket.

Note: it is often a good idea to examine the full correlation matrix to see if there are potential collinearity issues (i.e., using ‘cor(senic)‘ will give all the pairwise correlations in the data set).

Interaction Terms

- So far, we've examined the effects of 1 predictor at a time on the outcome
- For example, we could answer the question: is the linear association between strike outs per 9 innings and innings pitched per game different from 0?
- What if we wanted to examine the effects of 2 predictors simultaneously?
- For example, we might want to know if the effect of innings pitched per game on strikeouts per 9 innings is different for starting and relief pitchers

Interaction Terms

- To accomplish this we use interaction terms: interaction terms are the product of predictors
- Suppose we have the following model:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i1} \times x_{i2} + e_i$$

  where for each i, y_i is the strikeouts per 9 innings, x_{i1} is the innings pitched per game, x_{i2} is a dummy variable for a starting pitcher
- Recall that by including the dummy variable for pitcher type, we have two different intercepts
- By including the interaction term $x_{i1} \times x_{i2}$, we have two different slopes
- How should we interpret b_3?

Interaction Terms

- The example above included the interaction between a continuous and categorical variable
- Interactions can also be done between 2 continuous variables, as well as 2 categorical variables

Lab Ticket

1. Read in the SENIC data using the following command:

```
file = "https://ionides.github.io/401f18/hw/hw11/senic.txt"
senic = read.table(file,header = T)
```

2. Fit two models in R. In both cases, use `Infection.risk` as outcome

- Predictors: `Beds` and `Patients`
- Predictor: `Beds`

3. How do coefficients for `Beds` differ between the two models? What about the standard errors, t statistics, and p-values?