STA221

Neil Montgomery

Last edited: 2017-04-03 00:07

model selection preview

Recall the Body Fat % dataset.

```
## # A tibble: 250 × 15
##   `Pct BF`   Age Weight Height  Neck Chest Abdomen    waist   Hip
##      <dbl> <int>  <dbl>  <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl>
## 1     12.3    23 154.25  67.75  36.2  93.1    85.2 33.54331  94.5
## 2      6.1    22 173.25  72.25  38.5  93.6    83.0 32.67717  98.7
## 3     25.3    22 154.00  66.25  34.0  95.8    87.9 34.60630  99.2
## 4     10.4    26 184.75  72.25  37.4 101.8    86.4 34.01575 101.2
## 5     28.7    24 184.25  71.25  34.4  97.3   100.0 39.37008 101.9
## # ... with 245 more rows, and 6 more variables: Thigh <dbl>,
## #   Knee <dbl>, Ankle <dbl>, Bicep <dbl>, Forearm <dbl>, Wrist <dbl>
```

model selection preview

We had considered these two simple regression models:

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.69314    2.76045  -5.323 2.29e-07
## Weight        0.18938    0.01533  12.357  < 2e-16
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.58078   14.15400   1.807   0.0719
## Height      -0.09316    0.20119  -0.463   0.6438
```

model selection preview

Model with both. Is this a contradiction?

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 76.78100   10.04121   7.647 4.59e-13
## Weight       0.26326    0.01536  17.136  < 2e-16
## Height      -1.48829    0.15873  -9.376  < 2e-16
##
## Residual standard error: 5.626 on 247 degrees of freedom
## Multiple R-squared:  0.5435, Adjusted R-squared:  0.5398
## F-statistic: 147.1 on 2 and 247 DF,  p-value: < 2.2e-16
```

more possibilities - indicators and interactions

indicator, or “dummy” variables

An input variable in a multiple regression model can be just about anything (with minimal technical requirements).

A special and very useful example is a variable with only two possible values: 0 and 1.

indicator, or “dummy” variables

An input variable in a multiple regression model can be just about anything (with minimal technical requirements).

A special and very useful example is a variable with only two possible values: 0 and 1.

This is called an *indicator*, or dummy variable. The 0 and 1 values have no numerical meaning. They only divide the dataset into two groups.

indicator, or “dummy” variables

An input variable in a multiple regression model can be just about anything (with minimal technical requirements).

A special and very useful example is a variable with only two possible values: 0 and 1.

This is called an *indicator*, or dummy variable. The 0 and 1 values have no numerical meaning. They only divide the dataset into two groups.

For example, question 28.2 “Pizza” has results from the assessment of $n = 29$ frozen pizza brands.

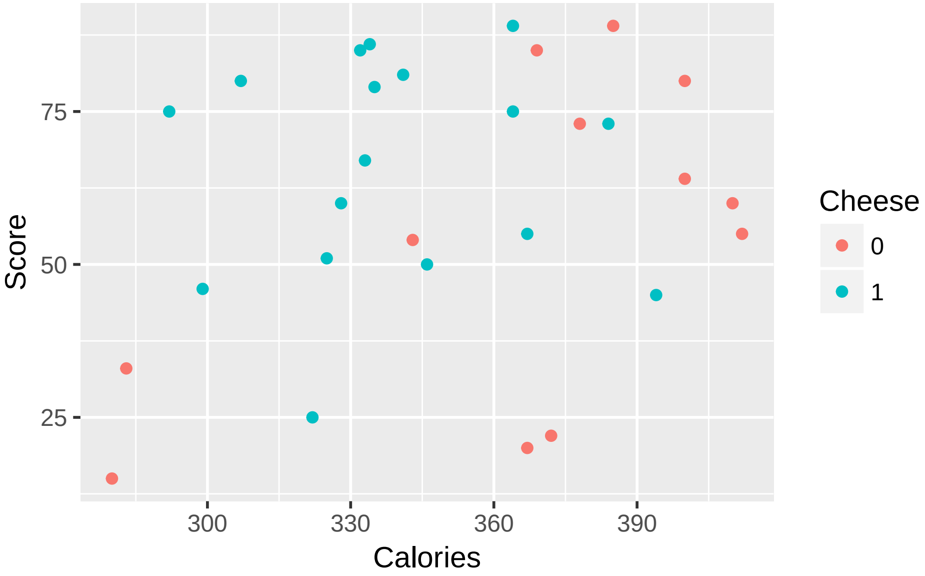
pizza

Here's a glance at the data. The last two columns are redundant.

```
## # A tibble: 29 × 7
##                       Brand Score  Cost Calories   Fat   Type Cheese
##                       <chr> <dbl> <dbl>    <dbl> <dbl> <fctr> <fctr>
## 1        Freshetta 4 Cheese    89  0.98      364    15 cheese      1
## 2 Freschetta stuffed crust    86  1.23      334    11 cheese      1
## 3                  DiGiorno    85  0.94      332    12 cheese      1
## 4              Amy's organic    81  1.92      341    14 cheese      1
## 5                   Safeway    80  0.84      307     9 cheese      1
## # ... with 24 more rows
```

They are there for "software" reasons.

Score versus Calories plotted



model with a dummy variable

What is the meaning of β_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

when x_2 is a dummy variable?

model with a dummy variable

What is the meaning of β_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

when x_2 is a dummy variable?

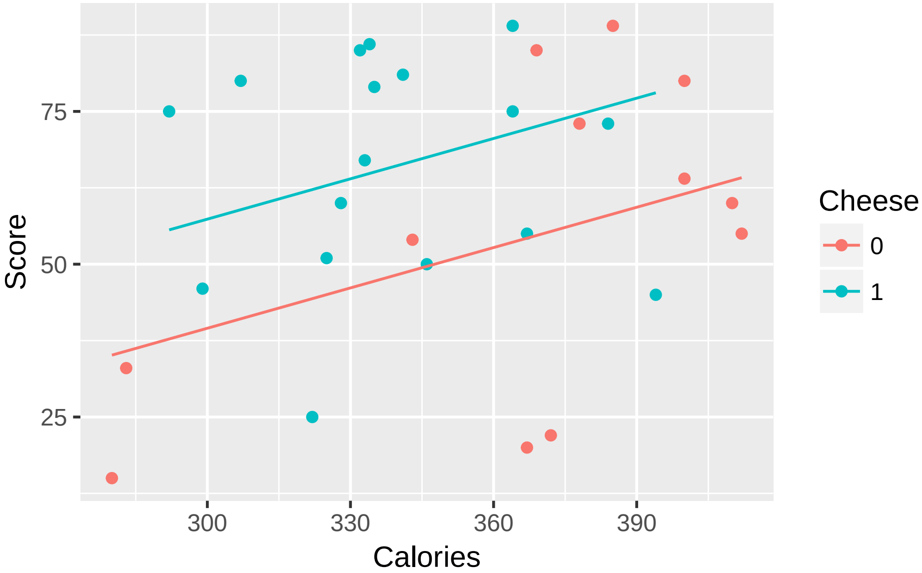
It lets you fit parallel lines with different intercepts.

pizza with Cheese dummy fitted

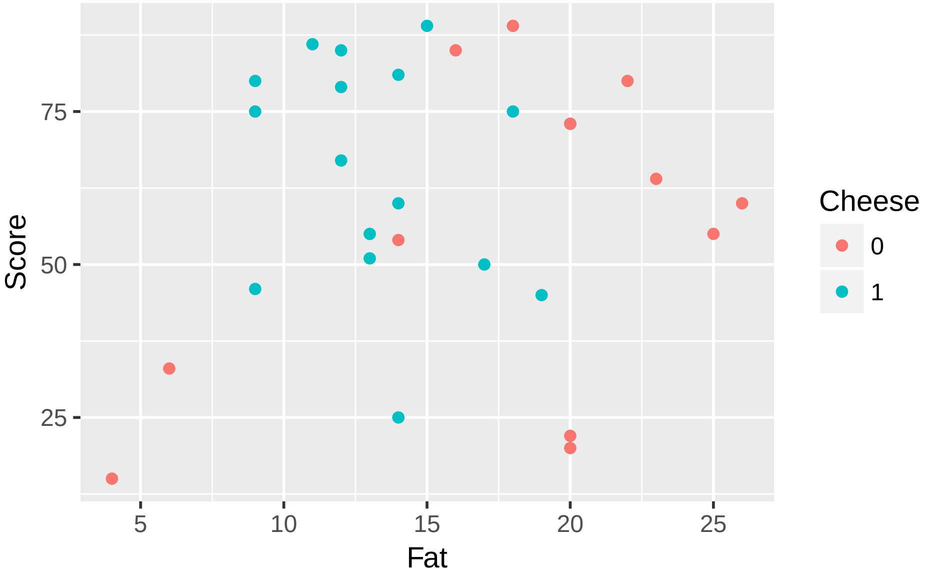
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.4510    41.2354  -0.641   0.5268
## Calories      0.2199     0.1113   1.976   0.0589
## Cheese1      17.8476     8.3603   2.135   0.0424
##
## Residual standard error: 20.65 on 26 degrees of freedom
## Multiple R-squared:  0.1929, Adjusted R-squared:  0.1308
## F-statistic: 3.107 on 2 and 26 DF,  p-value: 0.06168
```

`Cheese1` is R-speak for *this line is about the impact of ‘Cheese‘ with baseline value ‘1‘.*

pizza plotted with shifted lines (two intercepts)



Fat and Score by Cheese plotted



interaction with a dummy variable

Another use of dummy variables is to allow for different intercepts *and* slopes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

interaction with a dummy variable

Another use of dummy variables is to allow for different intercepts *and* slopes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

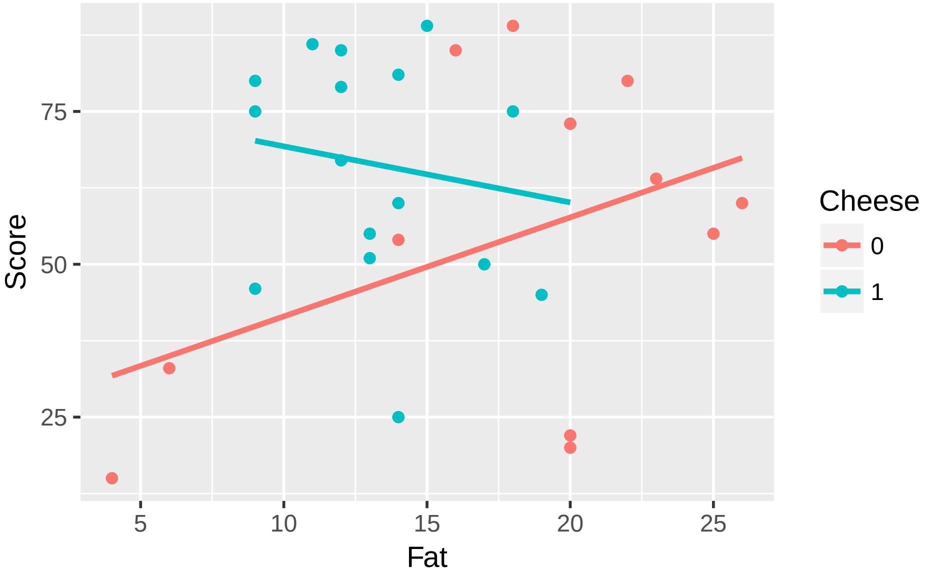
The $x_1 x_2$ term is called an *interaction* term, which allows the impact of x_1 to change as a function of x_2.

Interaction is not limited to the case of one of them being a dummy variable.

pizza with interaction

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    25.2850    17.5776   1.438   0.1627
## Fat             1.6195     0.9241   1.753   0.0919
## Cheese1        53.1752    28.1152   1.891   0.0702
## Fat:Cheese1    -2.5365     1.8217  -1.392   0.1761
##
## Residual standard error: 21.19 on 25 degrees of freedom
## Multiple R-squared:  0.1832, Adjusted R-squared:  0.08518
## F-statistic: 1.869 on 3 and 25 DF,  p-value: 0.1607
```

pizza with two slopes/intercepts



fun fact: t-test versus regression - I

```
##
##  Two Sample t-test
##
## data:  Score by Cheese
## t = -1.4441, df = 27, p-value = 0.1602
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28.64695   4.98028
## sample estimates:
## mean in group 0 mean in group 1
##        54.16667        66.00000
```

fun fact: t-test versus regression - II

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.167      6.274   8.634 3.01e-09
## Cheese1       11.833      8.194   1.444     0.16
##
## Residual standard error: 21.73 on 27 degrees of freedom
## Multiple R-squared:  0.0717, Adjusted R-squared:  0.03732
## F-statistic: 2.085 on 1 and 27 DF,  p-value: 0.1602
```

relationships among the inputs

“multicollinearity”

I stated the following fact about the b_i estimates for β_i:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k-1}$$

where c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

“multicollinearity”

I stated the following fact about the b_i estimates for β_i:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k-1}$$

where c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

It turns out that the more accurately x_i can be expressed as a linear combination of the other x_j in the model, the larger c_i gets.

"multicollinearity"

I stated the following fact about the b_i estimates for β_i:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k-1}$$

where c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

It turns out that the more accurately x_i can be expressed as a linear combination of the other x_j in the model, the larger c_i gets.

For example, when x_i and some other x_j are highly "correlated", it means they are close to linear functions of one another.

“multicollinearity”

I stated the following fact about the b_i estimates for β_i:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k-1}$$

where c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

It turns out that the more accurately x_i can be expressed as a linear combination of the other x_j in the model, the larger c_i gets.

For example, when x_i and some other x_j are highly “correlated”, it means they are close to linear functions of one another.

What happens when c_i is large?

illustration of the problem - two pairs of inputs

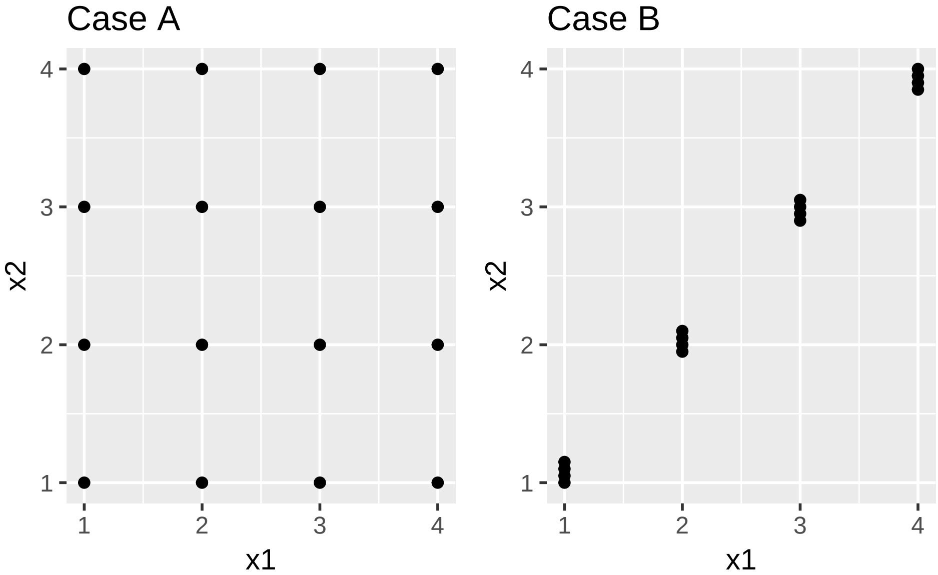


illustration of the problem

I'll generate some data from the same model in each case:

$$y = 1 + 2x_1 + 3x_2 + \varepsilon, \quad \varepsilon \sim N(0, 1)$$

Then fit the two datasets to regression models...

Case A

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5331     1.0177   1.506    0.156
## x1            1.9401     0.2744   7.069 8.43e-06
## x2            2.8854     0.2744  10.513 1.00e-07
##
## Residual standard error: 1.227 on 13 degrees of freedom
## Multiple R-squared:  0.9251, Adjusted R-squared:  0.9135
## F-statistic: 80.25 on 2 and 13 DF,  p-value: 4.843e-08
```

Case B

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5331     1.0177   1.506    0.156
## x1            4.1181     5.2218   0.789    0.444
## x2            0.7074     5.4890   0.129    0.899
##
## Residual standard error: 1.227 on 13 degrees of freedom
## Multiple R-squared:  0.9591, Adjusted R-squared:  0.9528
## F-statistic: 152.3 on 2 and 13 DF,  p-value: 9.506e-10
```

Note the small p-value for the overall F test.

Note that multicollinearity is merely a *possible* problem

Case C: same model fit to the Case B situation but with $n = 288$

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0510     0.1888   5.565 6.03e-08
## x1            2.1419     0.9690   2.210  0.02787
## x2            2.8299     1.0186   2.778  0.00583
##
## Residual standard error: 0.9663 on 285 degrees of freedom
## Multiple R-squared:  0.9693, Adjusted R-squared:  0.9691
## F-statistic:  4502 on 2 and 285 DF,  p-value: < 2.2e-16
```