Linear Regression part 2

Frank Edwards

Linear regression with multiple predictors

We can extend the linear regression model:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Linear regression with multiple predictors

We can extend the linear regression model:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

to include more than one predictor. We rewrite the equation as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots \beta_p x_p + \varepsilon$$

Linear regression with multiple predictors

We can extend the linear regression model:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

to include more than one predictor. We rewrite the equation as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots \beta_p x_p + \varepsilon$$

In matrix notation

To be more compact:

$$Y = \beta X + \varepsilon$$

Where Y is the vector of predictors, β is the vector of coefficients (including the intercept), X is the matrix of all predictors, and ε is the error term.

Estimating a multiple linear regression in R

```
library(gapminder)
gm_model<-lm(lifeExp ~log(gdpPercap) + continent,
             data = gapminder)
```

Interpreting a regression model with multiple coefficients

```
broom::tidy(gm_model)
```

```
## # A tibble: 6 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           2.32     1.36       1.70 8.85e-  2
## 2 log(gdpPercap)        6.42     0.183     35.0  1.50e-202
## 3 continentAmericas     7.01     0.554     12.7  3.99e- 35
## 4 continentAsia         5.91     0.477     12.4  7.30e- 34
## 5 continentEurope       9.58     0.604     15.9  6.70e- 53
## 6 continentOceania      9.21     1.54       6.00 2.42e-  9
```

Interpreting a regression model with multiple coefficients

```
broom::tidy(gm_model)
```

```
## # A tibble: 6 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           2.32     1.36       1.70 8.85e-  2
## 2 log(gdpPercap)        6.42     0.183     35.0  1.50e-202
## 3 continentAmericas     7.01     0.554     12.7  3.99e- 35
## 4 continentAsia         5.91     0.477     12.4  7.30e- 34
## 5 continentEurope       9.58     0.604     15.9  6.70e- 53
## 6 continentOceania      9.21     1.54       6.00 2.42e-  9
```

Recall that $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

What is our expected value for life expectancy:

- if GDP per capita = 1000 and continent = Europe

Interpreting a regression model with multiple coefficients

```
broom::tidy(gm_model)
```

```
## # A tibble: 6 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)          2.32      1.36       1.70 8.85e-  2
## 2 log(gdpPercap)       6.42      0.183     35.0  1.50e-202
## 3 continentAmericas    7.01      0.554     12.7  3.99e- 35
## 4 continentAsia        5.91      0.477     12.4  7.30e- 34
## 5 continentEurope      9.58      0.604     15.9  6.70e- 53
## 6 continentOceania     9.21      1.54       6.00 2.42e-  9
```

Recall that $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

What is our expected value for life expectancy:

- if GDP per capita = 1000 and continent = Europe
- if GDP per capita = 2000 and continent = Asia

Interpreting a regression model with multiple coefficients

```
broom::tidy(gm_model)
```

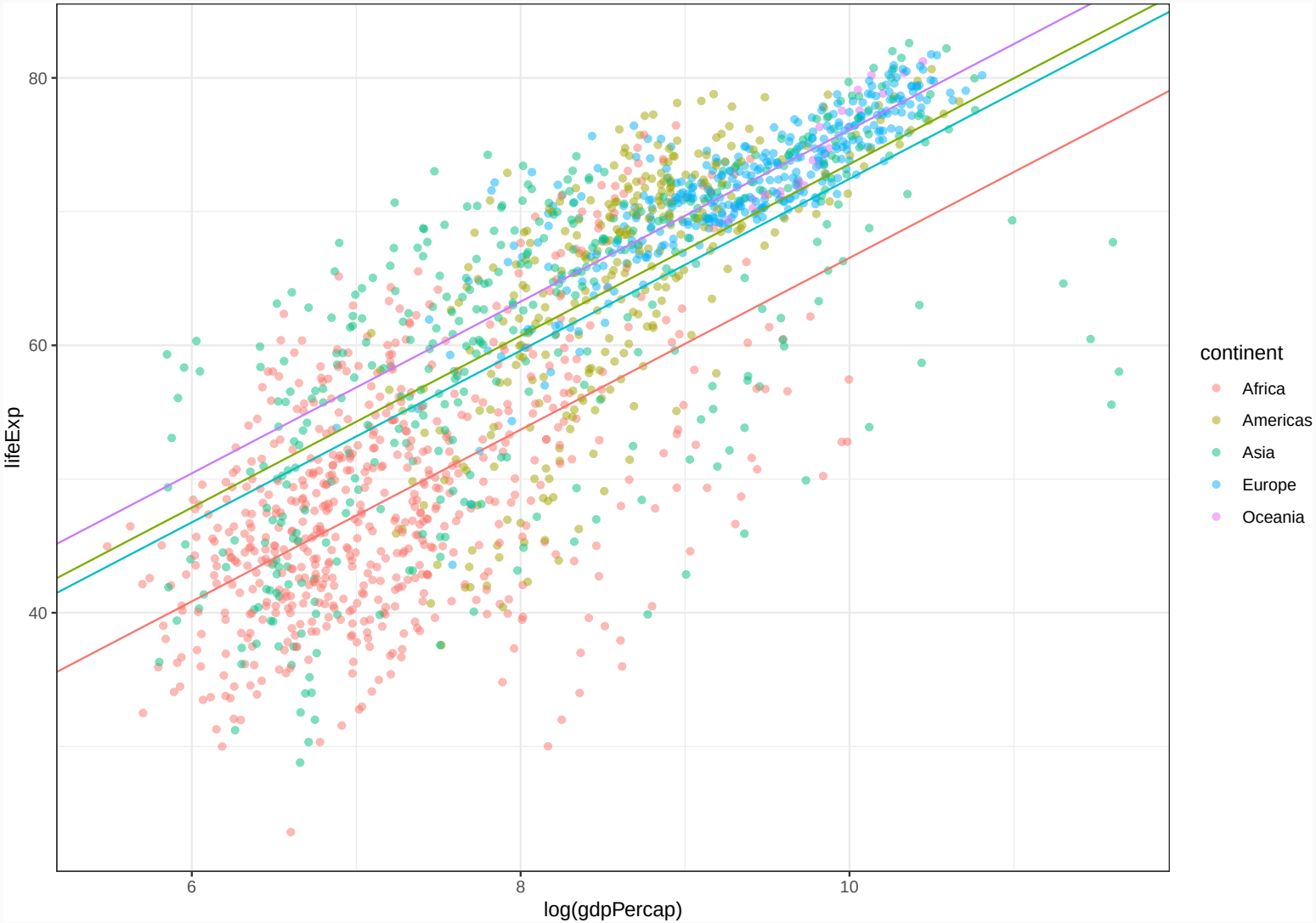
```
## # A tibble: 6 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)          2.32      1.36       1.70 8.85e-  2
## 2 log(gdpPercap)       6.42      0.183     35.0  1.50e-202
## 3 continentAmericas    7.01      0.554     12.7  3.99e- 35
## 4 continentAsia        5.91      0.477     12.4  7.30e- 34
## 5 continentEurope      9.58      0.604     15.9  6.70e- 53
## 6 continentOceania     9.21      1.54       6.00 2.42e-  9
```

Recall that $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

What is our expected value for life expectancy:

- if GDP per capita = 1000 and continent = Europe
- if GDP per capita = 2000 and continent = Asia
- if GDP per capita = 3000 and continent = Africa

Visualizing the model



Let’s try this with different data

```
### load in the 'mtcars' data
data(mtcars)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3
```

Variables in the mtcars data

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs Engine (0 = V-shaped, 1 = straight)
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears
- carb Number of carburetors

Let’s build a model for fuel efficiency with this data (theory!)

Our outcome of interest is `mpg`. What measured features of these cars do we think might be related to fuel efficiency?

Start simple

$$\mathrm{E}[\mathrm{mpg}_i] = \beta_0 + \beta_1 \mathrm{hp}_i$$

Start simple

$$\mathrm{E}[\mathrm{mpg}_i] = \beta_0 + \beta_1 \mathrm{hp}_i$$

$$\mathrm{mpg}_i = \beta_0 + \beta_1 \mathrm{hp}_i + \varepsilon$$

How would we estimate this model in R?

Now let's add complexity

$$\mathrm{E}[\mathrm{mpg}_i] = \beta_0 + \beta_1 \mathrm{hp}_i + \beta_2 \mathrm{wt}_i$$

Now let's add complexity

$$\mathrm{E}[\mathrm{mpg}_i] = \beta_0 + \beta_1 \mathrm{hp}_i + \beta_2 \mathrm{wt}_i$$

Build up and interpret an MPG model

Interactions

The prior model allowed each continent to have its own intercept fpr life expectancy. What if we thought the relationship of between GDP and life expectancy varied by continent?

We can add *interaction terms* to our model to model processes where we believe the relationship between y and x_1 is a function of x_2.

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

Estimating interactions in R

```
gm_model2<-lm(lifeExp ~ log(gdpPercap) * continent,
              data = gapminder)
```

Interpreting an interaction model

```
broom::tidy(gm_model2)
```

```
## # A tibble: 10 x 5
##    term                                estimate std.error statistic
##    <chr>                                  <dbl>     <dbl>     <dbl>
##  1 (Intercept)                             7.60      2.37      3.21
##  2 log(gdpPercap)                          5.69      0.325    17.5
##  3 continentAmericas                     -26.7       5.63     -4.74
##  4 continentAsia                           1.98      3.23      0.614
##  5 continentEurope                         5.37      5.26      1.02
##  6 continentOceania                      -39.8      42.6      -0.934
##  7 log(gdpPercap):continentAmericas        4.03      0.674     5.97
##  8 log(gdpPercap):continentAsia            0.561     0.421     1.33
```

Visualizing interactions