Imperial College London

MSc Machine Learning & Data Science

Applicable Mathematics

Ethics of Data Science – Part II

Measuring feature effects in classical models: linear regression

Dr. Chris Anagnostopoulos, Hon. Senior Lecturer

Linear Regression

```
Call:
lm(formula = Girth ~ Height + Volume, data = trees[train_i, ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.9257 -0.5487 -0.2153  0.5752  1.0926

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.61355    2.29077   6.379 4.05e-06 ***
Height      -0.08630    0.03225  -2.676    0.015 *
Volume       0.18305    0.01125  16.269 1.31e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6731 on 19 degrees of freedom
Multiple R-squared:  0.9455,    Adjusted R-squared:  0.9398
F-statistic: 164.9 on 2 and 19 DF,  p-value: 9.862e-13
```

$$\text{Girth} = \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Volume}$$

Linear Regression

```
Call:
lm(formula = Girth ~ Height + Volume, data = trees[train_i, ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.9257 -0.5487 -0.2153  0.5752  1.0926

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.61355    2.29077   6.379 4.05e-06 ***
Height      -0.08630    0.03225  -2.676    0.015 *
Volume       0.18305    0.01125  16.269 1.31e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6731 on 19 degrees of freedom
Multiple R-squared:  0.9455,    Adjusted R-squared:  0.9398
F-statistic: 164.9 on 2 and 19 DF,  p-value: 9.862e-13
```

$$\text{Girth} = \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Volume}$$

Linear Regression

```
Call:
lm(formula = Girth ~ Height + Volume, data = trees[train_i, ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.9257 -0.5487 -0.2153  0.5752  1.0926

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.61355    2.29077   6.379 4.05e-06 ***
Height      -0.08630    0.03225  -2.676    0.015 *
Volume       0.18305    0.01125  16.269 1.31e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6731 on 19 degrees of freedom
Multiple R-squared:  0.9455,    Adjusted R-squared:  0.9398
F-statistic: 164.9 on 2 and 19 DF,  p-value: 9.862e-13
```

$$\text{Girth} = \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Volume}$$

Linear Regression

```
Call:
lm(formula = Girth ~ Height, data = trees[train_i, ])

Residuals:
    Min      1Q  Median      3Q     Max
-3.6000 -2.6459 -0.0891  2.0705  3.9916

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.96983    7.72636  -0.255   0.8014
Height       0.21354    0.09968   2.142   0.0447 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.535 on 20 degrees of freedom
Multiple R-squared:  0.1866,    Adjusted R-squared:  0.146
F-statistic: 4.589 on 1 and 20 DF,  p-value: 0.04467
```

$$\text{Girth} = \beta_0 + \beta_1 \text{Height}$$

Linear Regression

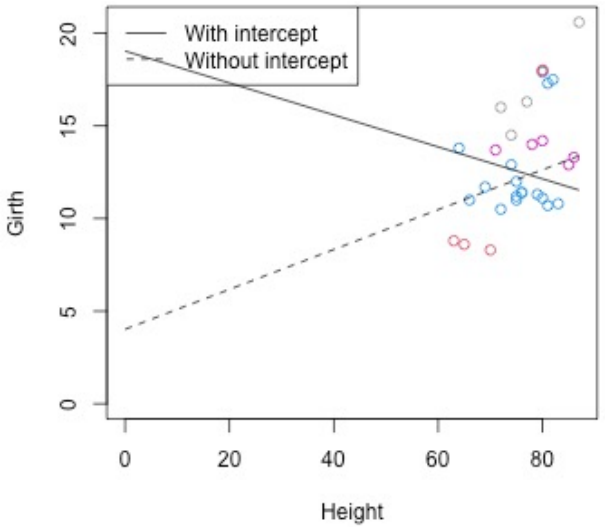
$$\text{Girth} = \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Volume}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.61355    2.29077   6.379 4.05e-06 ***
Height      -0.08630    0.03225  -2.676    0.015 *
Volume       0.18305    0.01125  16.269 1.31e-12 ***
---
```

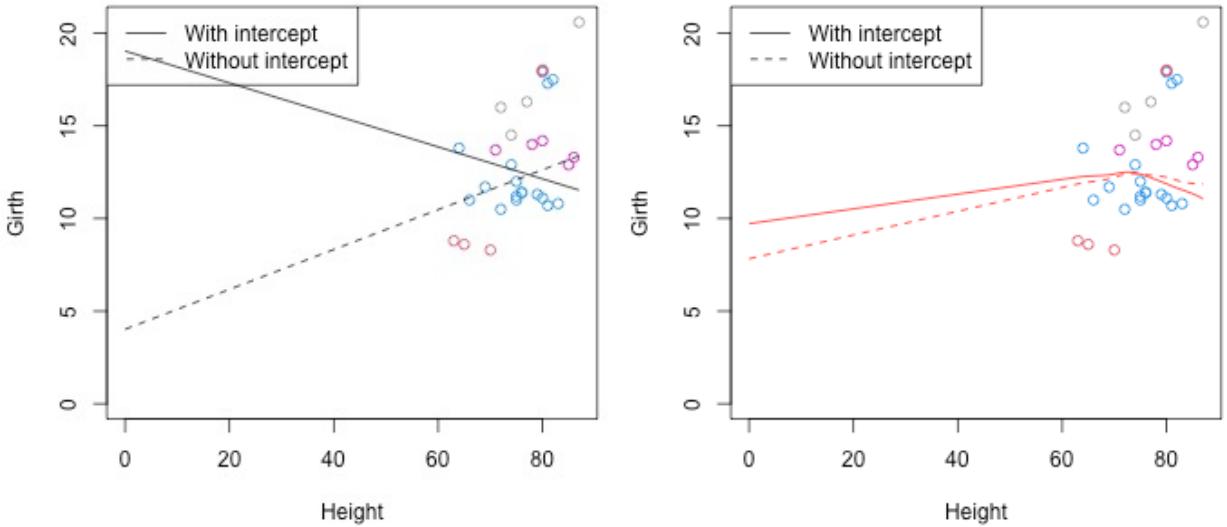
$$\text{Girth} = \beta_0 + \beta_1 \text{Height}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.96983    7.72636  -0.255   0.8014
Height       0.21354    0.09968   2.142   0.0447 *
```

Linear Regression



Linear Regression



Summary: revisiting Linear Regression

- Importance but also direction of effects depends on everything else that is in the model and can change if variables are added/dropped.
- Interpreting effects is also affected by model mis-specification: a non-linear relationship fitted with a linear model can be positive for some parts of the space and negative for others.

Summary: revisiting Linear Regression

- Importance but also direction of effects depends on everything else that is in the model and can change if variables are added/dropped.
- Interpreting effects is also affected by model mis-specification: a non-linear relationship fitted with a linear model can be positive for some parts of the space and negative for others.

- Reasons for optimism:
  - Non-linearity helps.
  - Causal reasoning helps.
  - Sensitivity analysis helps.

- Manage expectations, be humble.