EGM101 – Skills Toolbox

Week 6, Part 3: (Linear) Regression

Refresher: Lines

- General equation:

$$y = \beta + \alpha x$$

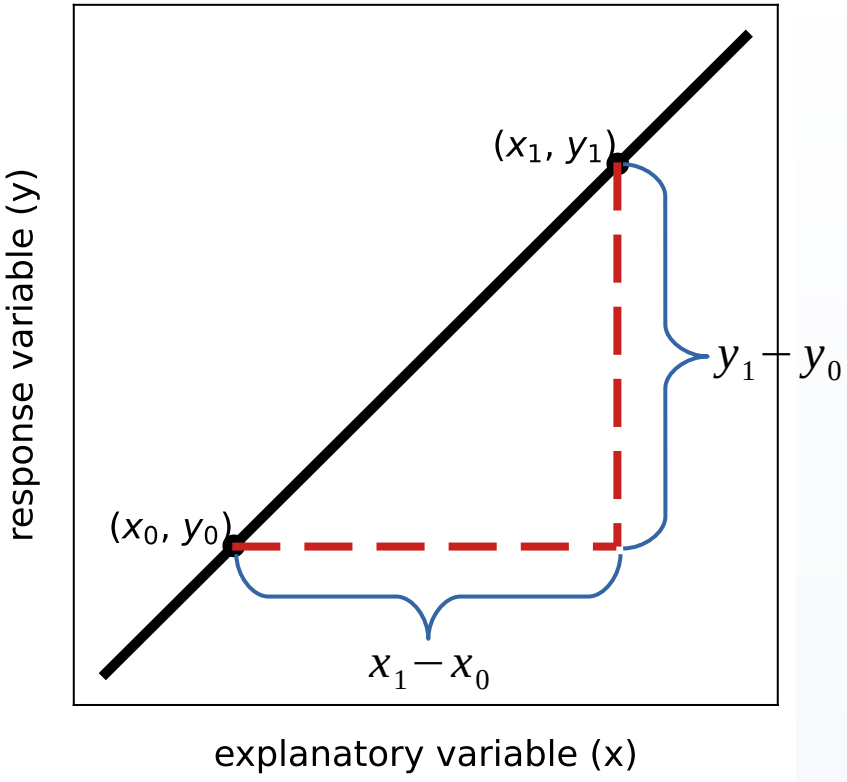
- Slope, α:

$$\alpha = \frac{y_1 - y_0}{x_1 - x_0}$$

  - Describes relationship between change in x, change in y

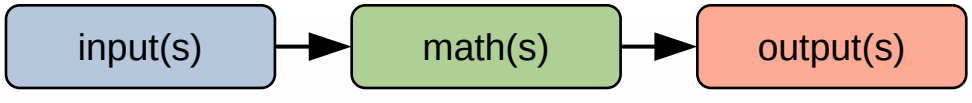
- Intercept, β:

$$\beta = y - \alpha x$$



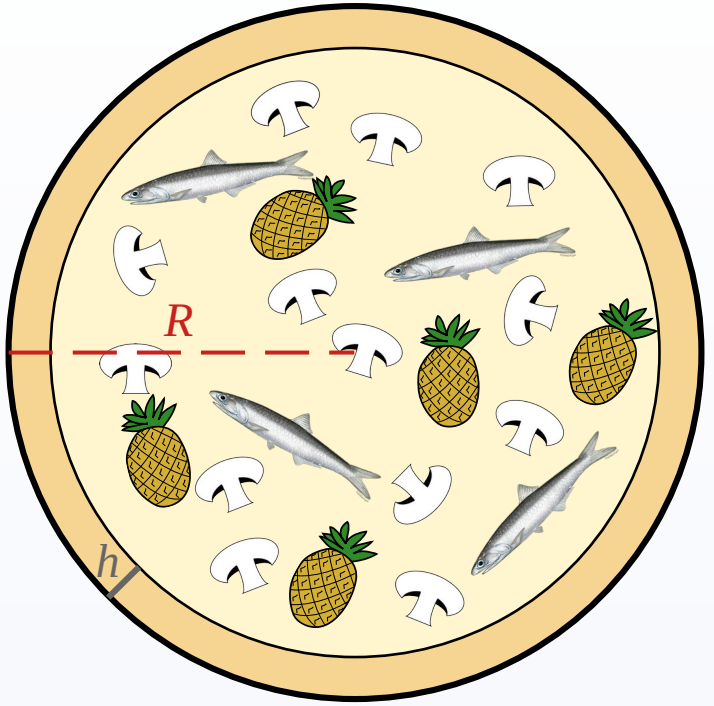
Mathematical models

- In science, seek to:
  - Understand (explain)
  - Make credible predictions
- One tool: mathematical models
- Example: surface area (A) of pizza
  - Simple!
  - But: what about crust?
  - But: crusts aren't even thickness, pizza isn't perfectly round, …
- Ultimately: "all models are wrong, but some are useful" (G. E. P. Box)



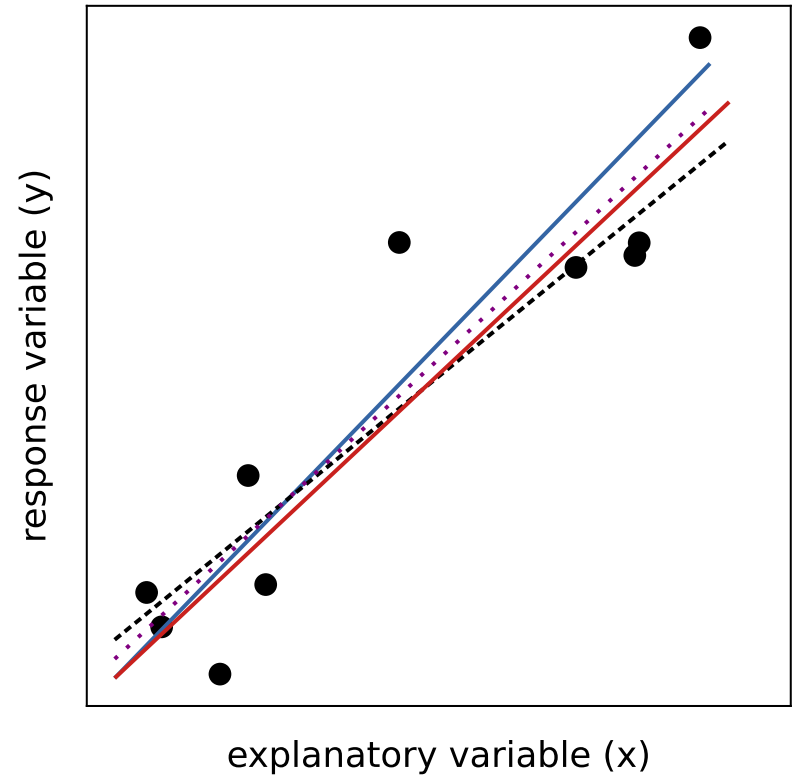
$A = \pi R^2$

$A = \pi (R - h)^2$



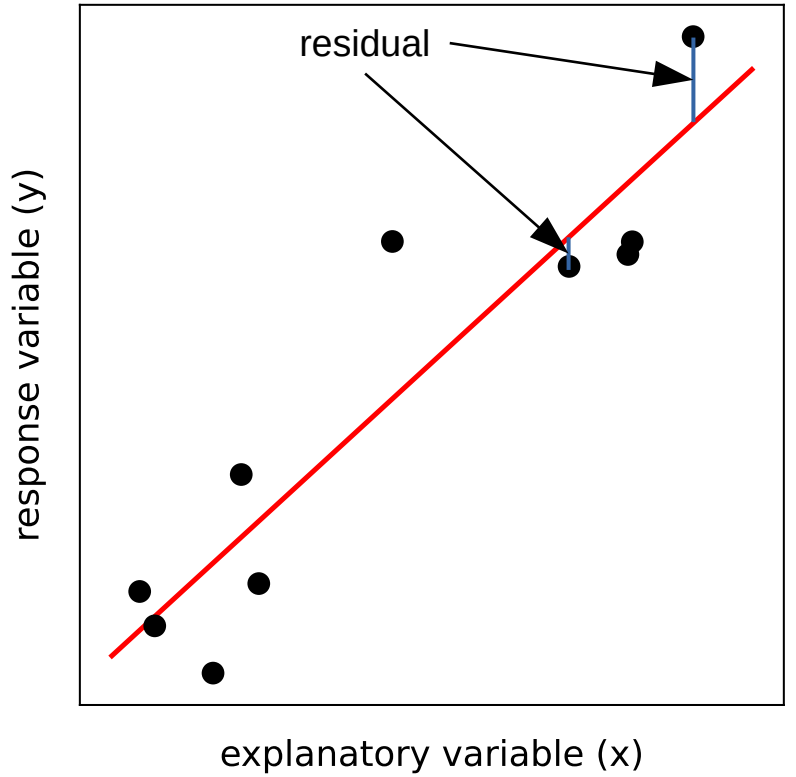
Linear Relationships

- "Linear relationship" means we can find a line that "fits" the data:
  - Explains relationship
  - Predicts unknown values
  - In other words: a model!
- Problem: there are many such lines
  - Question: how do we find the "best" line?



Finding a Linear Model

- First: what do we mean by “best”?
- Remember:
  - Model is a *prediction*
  - Difference between prediction, observation: residual (error)
- Very often: “best” model minimizes the total prediction error



Simple Linear Regression

- Regression: process of finding “best” fitting model
  - Linear regression: using a linear model
  - Simple linear regression: using a single independent (explanatory) variable
- General linear model*:

$$\hat{y} = \beta + \alpha x + \varepsilon$$

  - $\hat{y}$: predicted value of y (response/dependent variable)
  - ε: residuals (error) between predicted y, observed y
- Goal: find α, β that give the smallest ε

*for two variables

(Ordinary) Least-squares Regression

- One solution: minimize sum of $(y - \hat{y})^2$ (*least-squares*)
- Potential issues:
  - Assumes that there is a linear relationship
  - Works best when ε are normally distributed
  - Sensitive to extreme values (cf. standard deviation)
  - Explanatory variables cannot be the same value (must have non-zero variance)

$$\hat{y} = \beta + \alpha x + \varepsilon$$

$$\alpha = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = r \frac{s_y}{s_x}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\beta = \bar{y} - \alpha \bar{x}$$

Properties of Least-squares Regression

- Slope:
  - Describes relationship of changes in explanatory, response variables
  - Depends on covariance of x, y:
    - Negative covariance → negative slope
    - Positive covariance → positive slope

$$\alpha = \frac{\mathrm{cov}(x, y)}{var(x)} = r\frac{s_y}{s_x}$$

- With non-zero β:
  - Line always passes through point ($\bar{x}$, $\bar{y}$)
  - Sum (and therefore mean) of residuals is zero
  - Residuals are not correlated with x

Summary

- We use mathematical models to help explain, predict the “real world”
- Goal with linear regression is to find a model that:
  - Explains the variation in y based on variation in x
  - Can “predict” values
  - Minimizes the difference between predicted, observed values
- Least-squares is most common, but not the only way

Additional Resources

- Illowsky and Dean, Chapters 12.1 – 12.3
- Caswell, Chapters 9.1 – 9.4
- Weiss, Chapters 14.1 – 14.3
- 7 Classical Assumptions of OLS Linear Regression [Jim Frost]
- Intro to residuals and least squares regression [Khan Academy]
- Regression line example [Khan Academy]