“Simple” Linear Regression

20191004 – Sleuth 3 Chapter 7

Example

We have a data set with information about 152 flights by Endeavour Airlines that departed from JFK airport in New York to either Nashville (BNA), Cincinnati (CVG), or Minneapolis-Saint Paul (MSP) in January 2012.

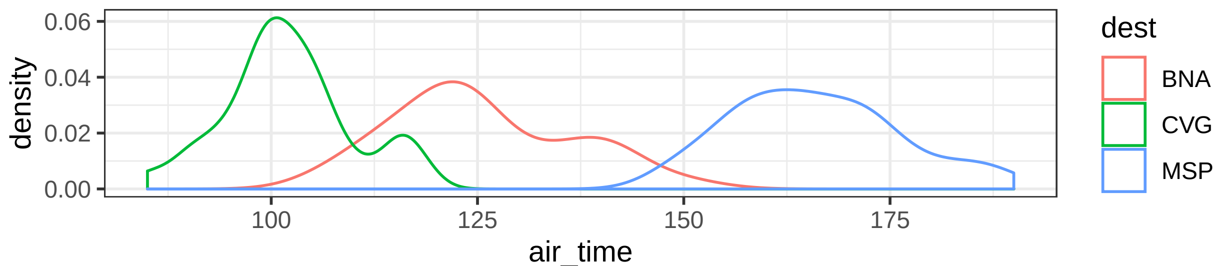
```r
head(flights, 4)
```

```
## # A tibble: 4 x 3
##   distance air_time dest
##      <dbl>    <dbl> <chr>
## 1     1029      189 MSP
## 2      765      150 BNA
## 3     1029      173 MSP
## 4      589      118 CVG
```

So Far: ANOVA Model

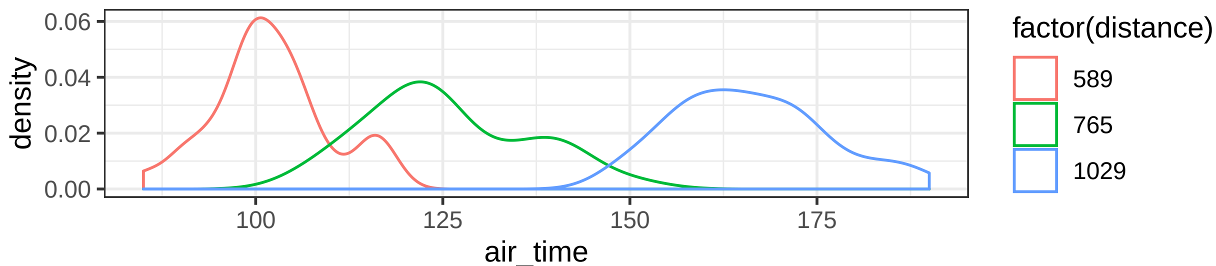
- Observations in group i follow a Normal(μ_i, σ^2) distribution
- Observations are independent of each other

```r
ggplot(data = flights, mapping = aes(x = air_time, color = dest)) +
  geom_density() +
  theme_bw()
```



Note: The picture would look exactly the same if we treated distance as a categorical variable:

```r
ggplot(data = flights, mapping = aes(x = air_time, color = factor(distance))) +
  geom_density() +
  theme_bw()
```

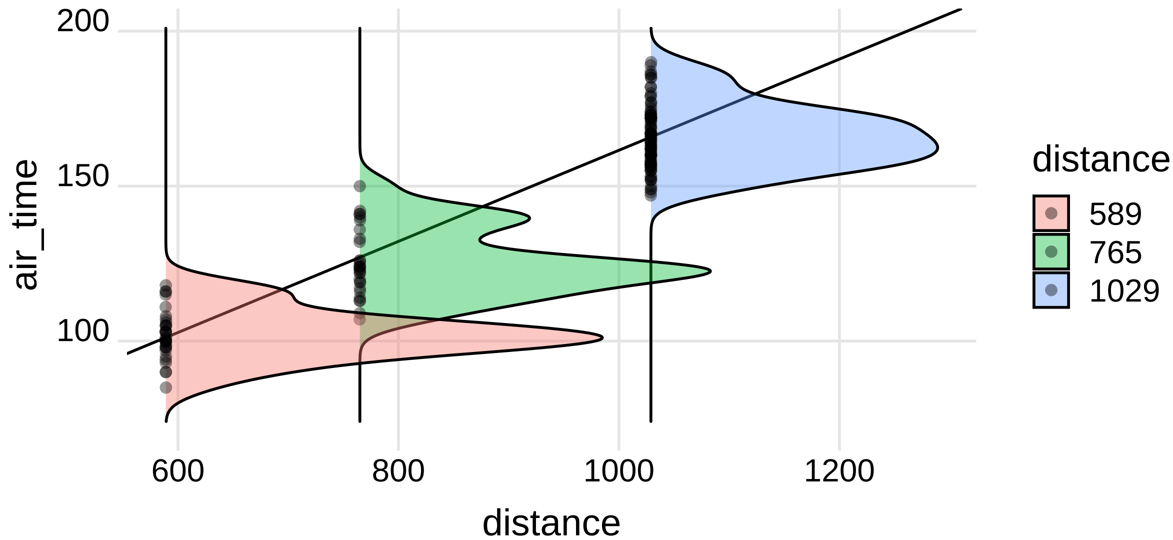


Old idea: Each group has a normal distribution with its own mean

- Categorical explanatory variable

New idea: Each group has a normal distribution with a mean that is a linear function of distance

- Quantitative (numeric) explanatory variable



The simple linear regression is exactly like the ANOVA model, with the one new restriction that the means fall along a line.

Two ways to write the model:

Focusing on the mean (book)

Values of the response variable are independent and normally distributed with mean $\mu(Y|X) = \beta_0 + \beta_1 X$

- Read as "The mean of Y for a given value of X"
- In our example, Y is air time and X is distance.

Written for a single observation, number i (my preference)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

- In our example, Y_i is the air time for flight number i and x_i is the distance for flight number i.

Parameter interpretations

- β_0 is intercept for the population: mean value of the response when $X = 0$, in the population
- β_1 is slope for the population: change in mean response when X increases by 1 unit, in the population.
- β_0 and β_1 are unknown population paramters. We estimate them with the intercept and slope of a line describing our sample.

Conditions: spells "LINE-O"

Exactly the same as conditions for ANOVA, with addition that the mean of the response is a linear function of the explanatory varible:

- **Linear** relationship between explanatory and response variables
- **Independent** observations (knowing that one observation is above its mean wouldn't give you any information about whether or not another observation is above its mean)
- **Normal** distribution
- **Equal standard deviation** of response for all values of X
  - Denote this standard deviation by σ
- **no Outliers** (not a formal part of the model, but important to check in practice)