"Simple" Linear Regression (Sleuth 3 Chapter 7)

Big points for today

- Simple linear regression is exactly like the ANOVA model, but the group means are on a line.
- First R commands
- Interpretation of slope and intercept
- Hypothesis tests and confidence intervals about slope and intercept

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.

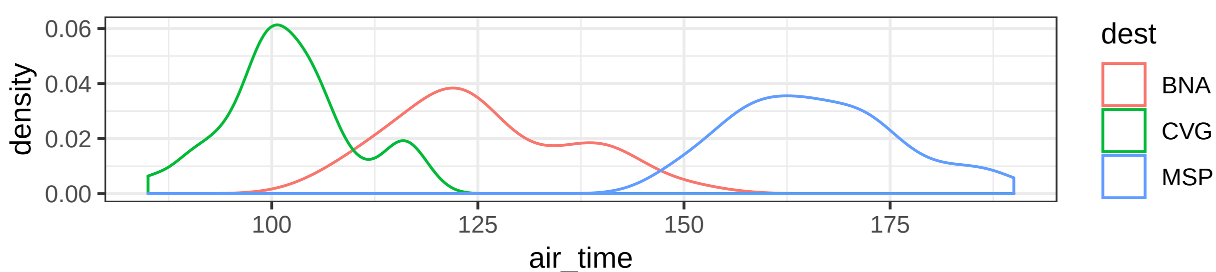
```
head(flights)
```

```
## # A tibble: 6 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
## 5      589      115 CVG
## 6     1029      153 MSP
```

So Far: ANOVA Model

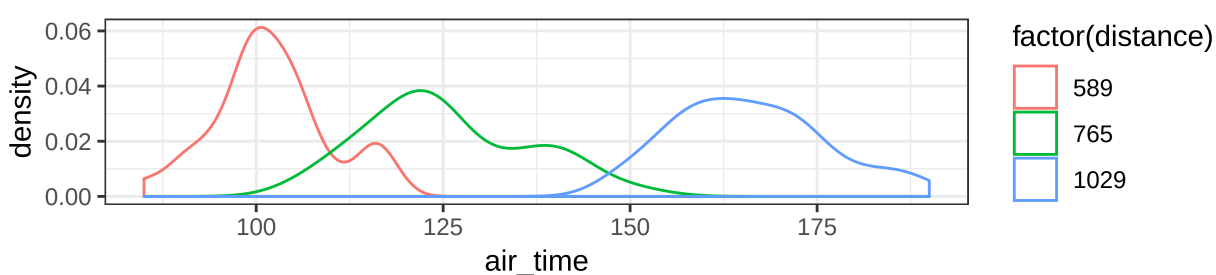
- Observations in group i follow a Normal(μ_i, σ^2) distribution
- Observations are independent of each other

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



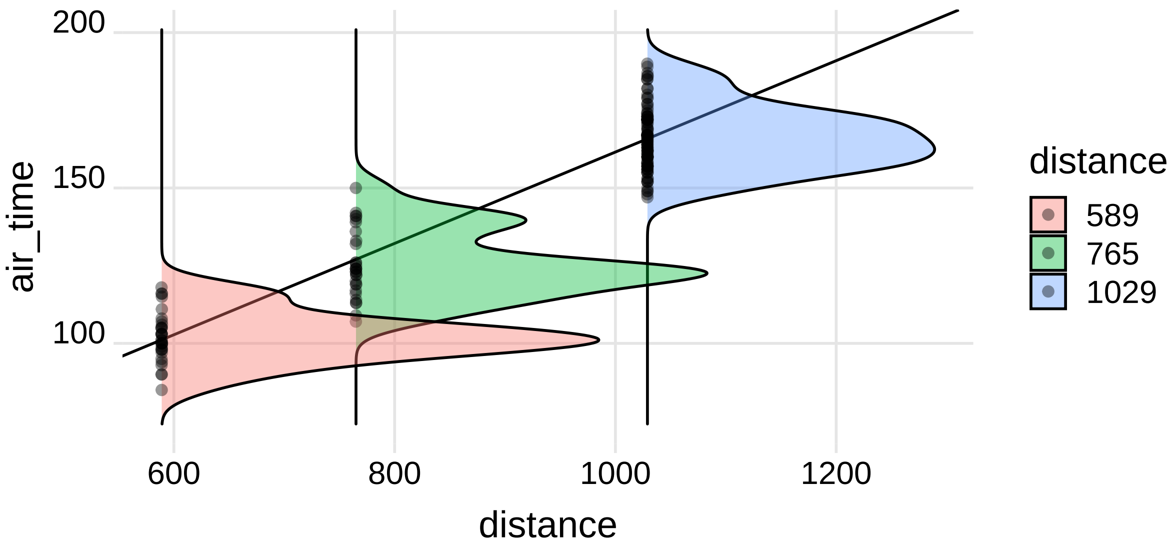
Note: The picture would look exactly the same if we treated distance as a categorical variable:

```
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

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



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

Conditions: spells "LINE-O"

- **Linear** relationship between explanatory and response variables: $\mu(Y|X) = \beta_0 + \beta_1 X$
  - Read as "The mean of Y for a given value of X"
  - β_0 is intercept: mean response when $X = 0$
  - β_1 is slope: change in mean response when X increases by 1 unit.
  - β_0 and β_1 are **parameters** describing the relationship between X and Y **in the population**
- **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)

R Code

```
model_fit <- lm(air_time ~ distance, data = flights)
summary(model_fit)
```

```
##
## Call:
## lm(formula = air_time ~ distance, data = flights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.022  -7.054  -1.086   6.170  24.170
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.567729   3.955477   3.683 0.000321 ***
## distance     0.146999   0.004372  33.624  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.881 on 150 degrees of freedom
## Multiple R-squared:  0.8829, Adjusted R-squared:  0.8821
## F-statistic:  1131 on 1 and 150 DF,  p-value: < 2.2e-16
```

1. What is the estimated intercept and its interpretation?

2. Conduct a hypothesis test of the claim that when a flight travels 0 miles, its air time is 0 minutes.

3. What is the estimated slope and its interpretation?

4. Conduct a hypothesis test of the claim that a flight's air time is unrelated to the distance travelled.

5. Conduct a hypothesis test of the claim that these planes are flying at an average speed that's the same as the typical cruising speed of commercial passenger aircraft.

According to Wikipedia, the typical cruising speed of commercial passenger aircraft is about 560 miles per hour (https://en.wikipedia.org/wiki/Cruise_(aeronautics)). After some unit changes, this works out to about 0.107 minutes per mile.

```r
# calculate t statistic
(0.147 - 0.107) / 0.0044
```

```
## [1] 9.090909
```

```r
# calculate 2-sided p-value
pt(-9.09, df = 152 - 2) + pt(9.09, df = 152 - 2, lower.tail = FALSE)
```

```
## [1] 5.49638e-16
```

6. Find and interpret a 95% confidence interval for the slope of the line

```r
# automatic calculations
confint(model_fit)
```

```
##                 2.5 %     97.5 %
## (Intercept) 6.7520812 22.3833778
## distance    0.1383611  0.1556377
```

```r
# manual calculations from the formula: get the multiplier for an individual 95% CI
qt(0.975, df = 152 - 2)
```

```
## [1] 1.975905
```

```r
# calculate lower and upper endpoints of confidence interval
0.147 - 1.976 * 0.00437
```

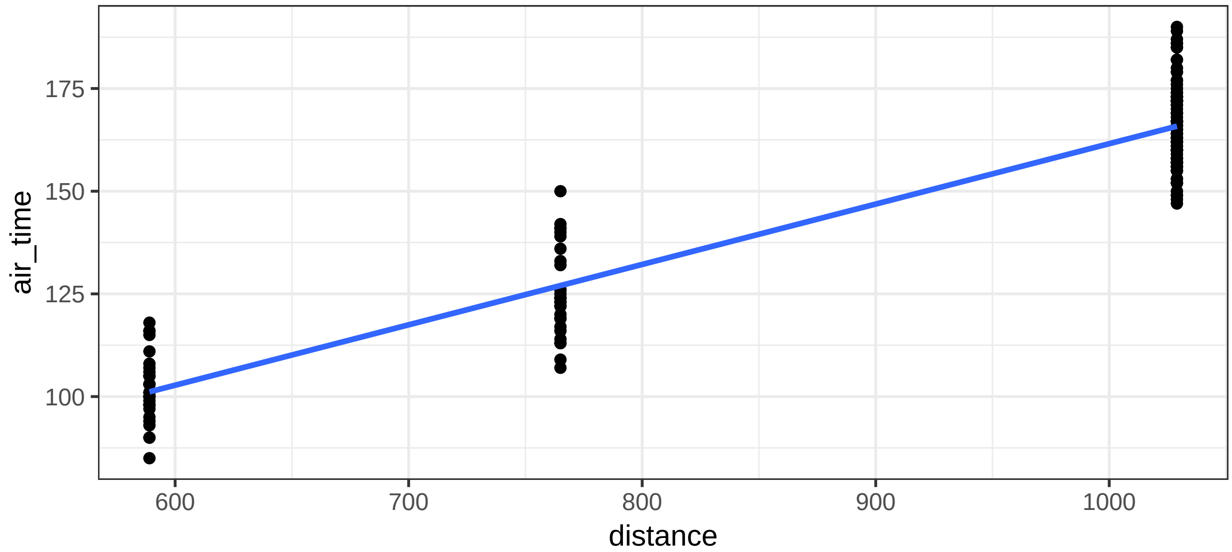
```
## [1] 0.1383649
```

```r
0.147 + 1.976 * 0.00437
```

```
## [1] 0.1556351
```

R Code to make scatterplot with estimated line overlaid

```
ggplot(data = flights, mapping = aes(x = distance, y = air_time)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```




```
ggplot(data = flights, mapping = aes(x = distance, y = air_time)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()
```