Simple Linear Regression

David Gerard

2018-12-07

Objectives

- Intuitively understand simple linear regression.
- Ch 7 in the book.

Case Study

- The theory of Big Bang suggests a formal relationship between the distance between any two celestial objects (Y) and the recession velocity (X) between them (how fast they are moving apart) given the (unknown) age of the universe (T):

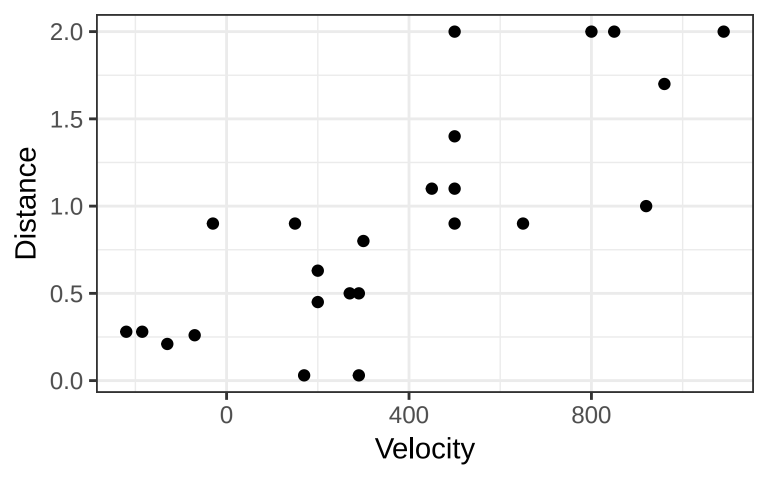
$$Y = TX$$

- Distance vs velocity measurements of multiple nebulae

```
library(Sleuth3)
data("case0701")
```

Scatterplot

```r
library(ggplot2)
qplot(Velocity, Distance, data = case0701, geom = "point")
```



Questions of Interest

- The formula describes a line with zero intercept. Is the intercept zero?
- What is the age of the universe (estimate T)?

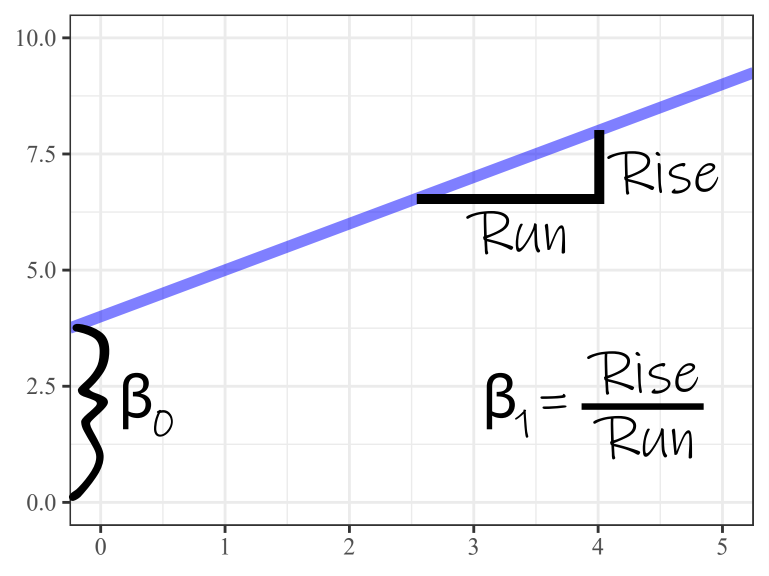
Review: Lines

- Every line may be represented by a formula of the form

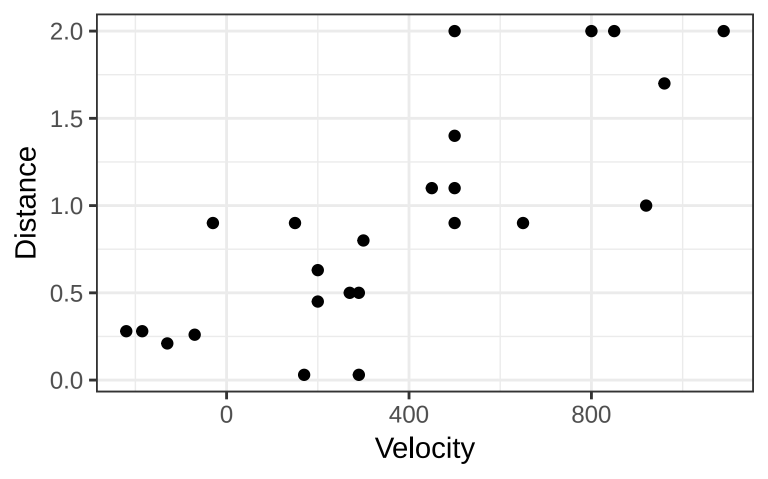
$$Y = \beta_0 + \beta_1 X$$

- Y = response variable on y-axis
- X = explanatory variable on the x-axis
- β_1 = slope (rise over run)
  - How much larger is Y when X is increased by 1.
- β_0 = y-intercept (the value of the line at $X = 0$)

Review Lines



A line doesn't exactly fit



A line plus noise

- The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

A line plus noise

- The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i: distance from earth of nebula i

A line plus noise

- The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i

A line plus noise

- The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i
- β_0: The intercept of the mean line ("regression line")
    - Mean when $X_i = 0$

A line plus noise

- The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i
- β_0: The intercept of the mean line ("regression line")
  - Mean when $X_i = 0$
- β_1: Slope of the regression line.
  - Difference in mean distance between two nebula when they differ by only 1 velocity unit.

A line plus noise

- The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i
- β_0: The intercept of the mean line ("regression line")
  - Mean when $X_i = 0$
- β_1: Slope of the regression line.
  - Difference in mean distance between two nebula when they differ by only 1 velocity unit.
- $\beta_0 + \beta_1 X_i$: the mean distance at velocity X_i

A line plus noise

- The linear regression model

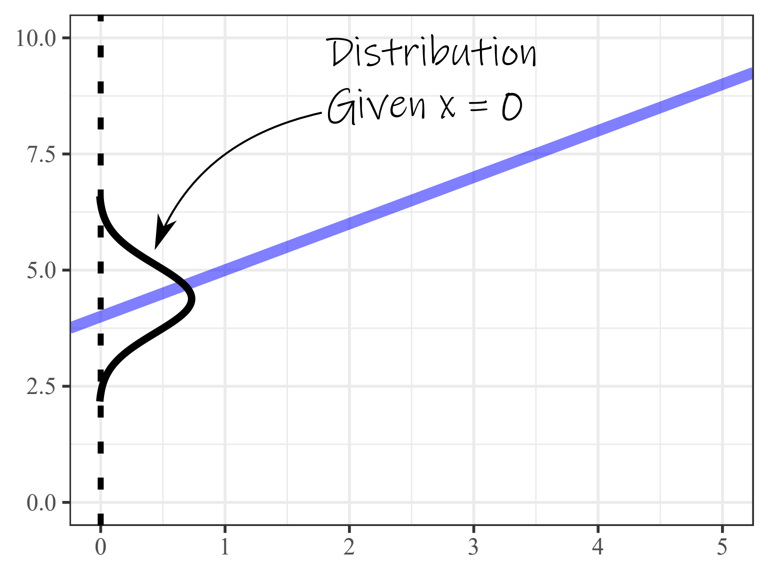
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i: distance from earth of nebula i
- X_i: recession velocity of nebula i
- β_0: The intercept of the mean line ("regression line")
    - Mean when $X_i = 0$
- β_1: Slope of the regression line.
    - Difference in mean distance between two nebula when they differ by only 1 velocity unit.
- $\beta_0 + \beta_1 X_i$: the mean distance at velocity X_i
- ϵ_i: Individual noise with mean 0 and variance σ^2. Ideally normally distributed.

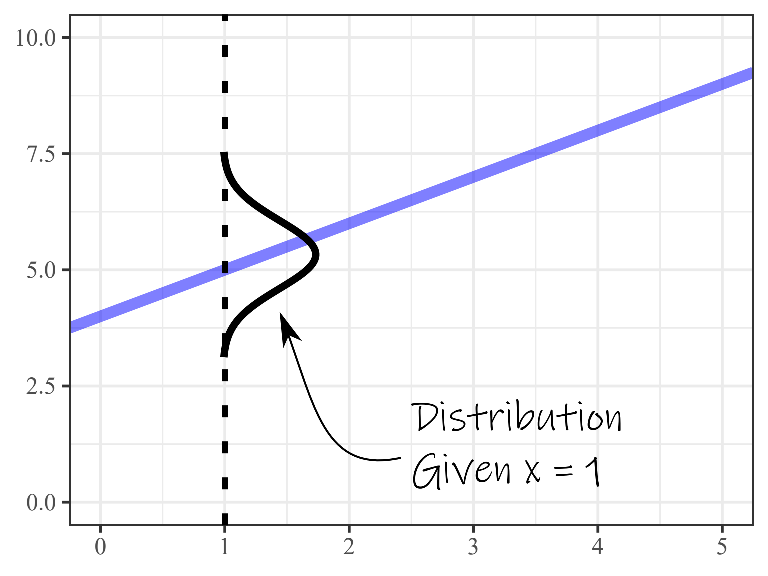
Some intuition

- The distribution of Y is *conditional* on the value of X.
- The distribution of Y is assumed to have the **same variance**, σ^2 for **all possible values of** X.
- This last one is a considerable assumption.

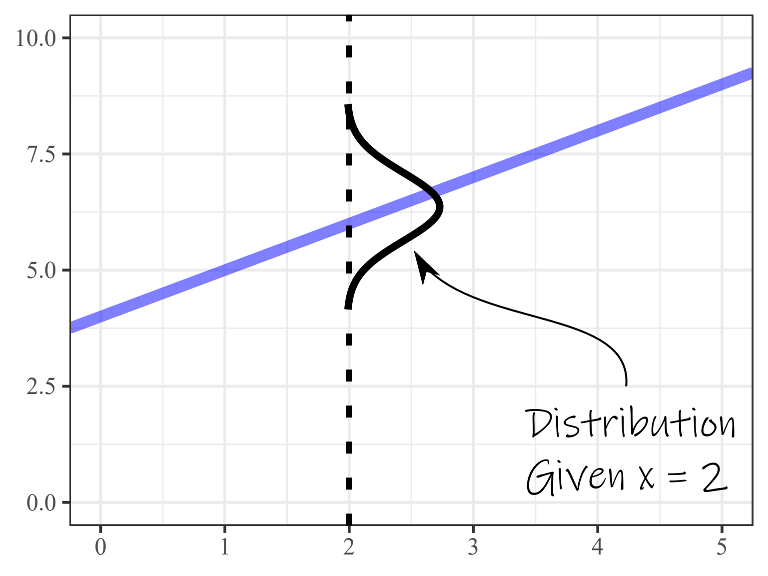
Conditional Distributions



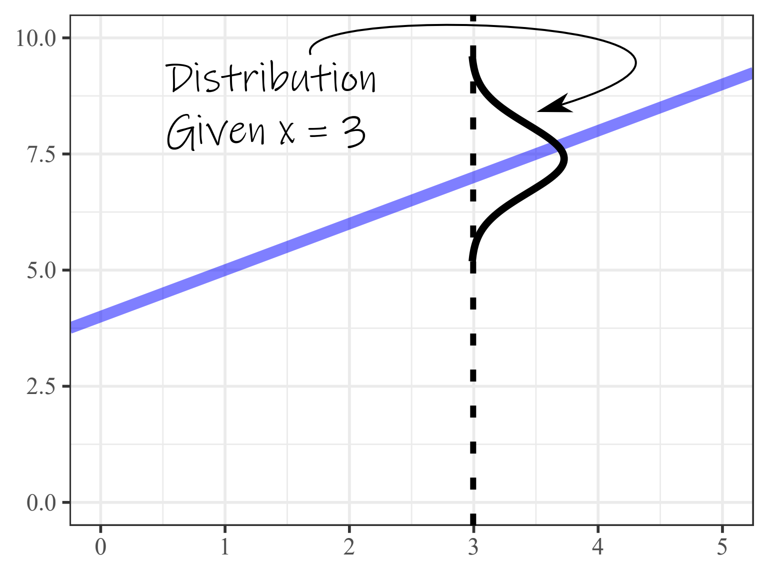
Conditional Distributions



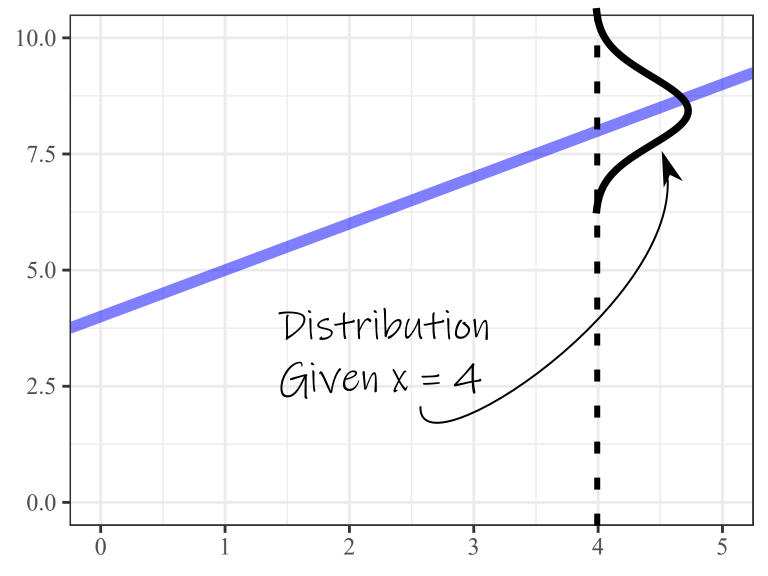
Conditional Distributions



Conditional Distributions



Conditional Distributions



How do we estimate β_0 and β_1?

- β_0 and β_1 are **parameters**
- We want to estimate them from our **sample**

How do we estimate β_0 and β_1?

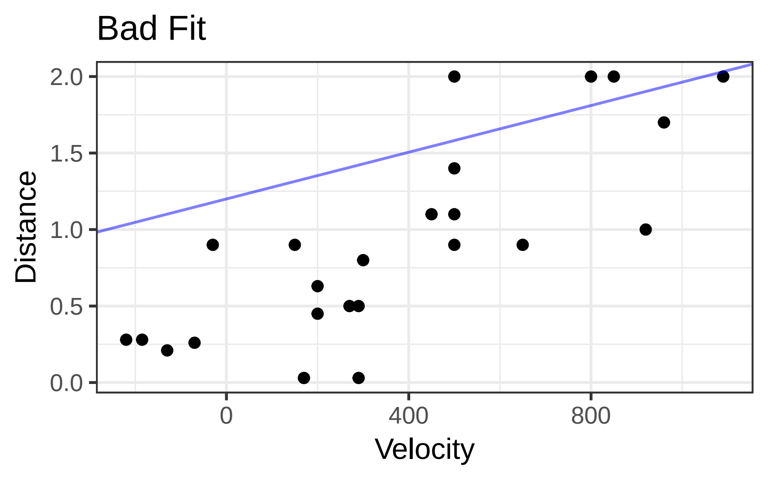
- β_0 and β_1 are **parameters**
- We want to estimate them from our **sample**
- Idea: Draw a line through the cloud of points and calculate the slope and intercept of that line?
- Problem: Subjective

How do we estimate β_0 and β_1?

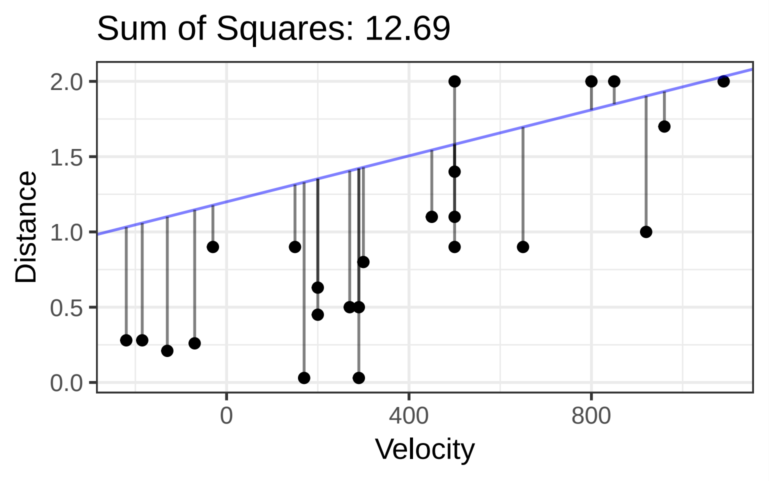
- β_0 and β_1 are **parameters**
- We want to estimate them from our **sample**
- Idea: Draw a line through the cloud of points and calculate the slope and intercept of that line?
- Problem: Subjective
- Another idea: Minimize residuals (sum of squared residuals).

Ordinary Least Squares

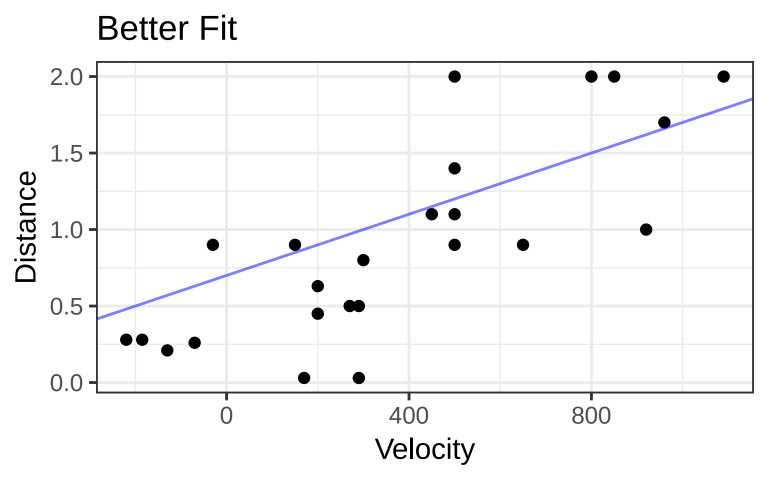
- Residuals: $\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$
- Sum of squared residuals: $\hat{\epsilon}_1^2 + \hat{\epsilon}_2^2 + \cdots + \hat{\epsilon}_n^2$
- Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that have small sum of squared residuals.
- The obtained estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, are called the **ordinary least squares** (OLS) estimates.


Bad Fit
2.0
1.5
1.0
0.5
0.0
Distance
0
400
800
Velocity

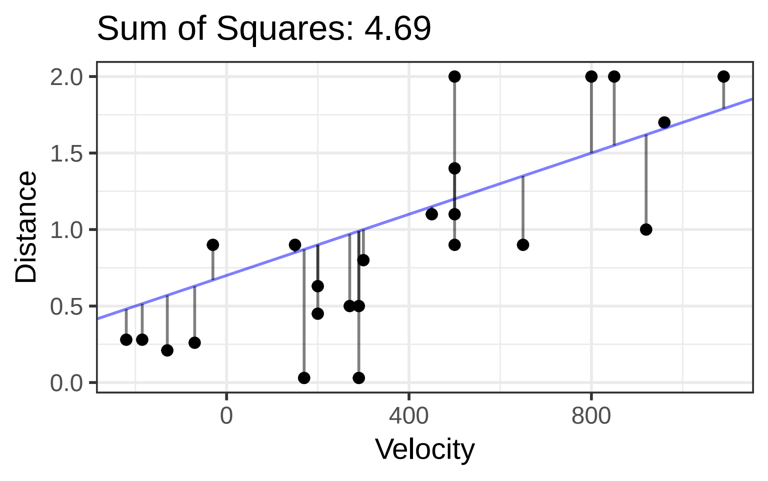
Bad Fit



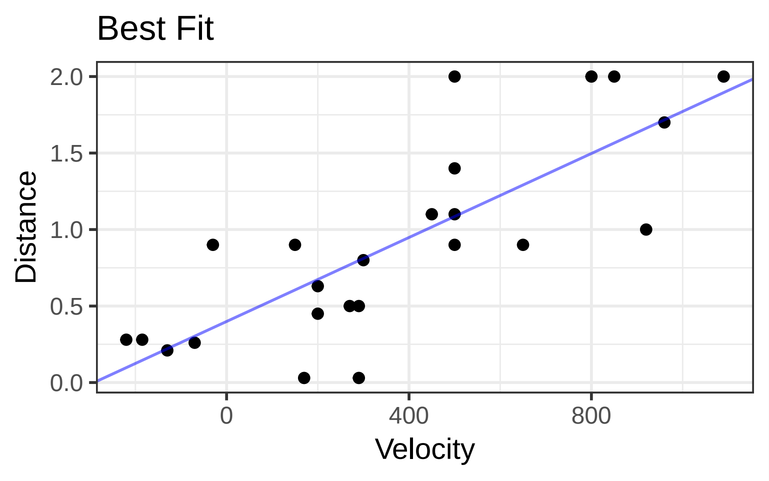
Better Fit



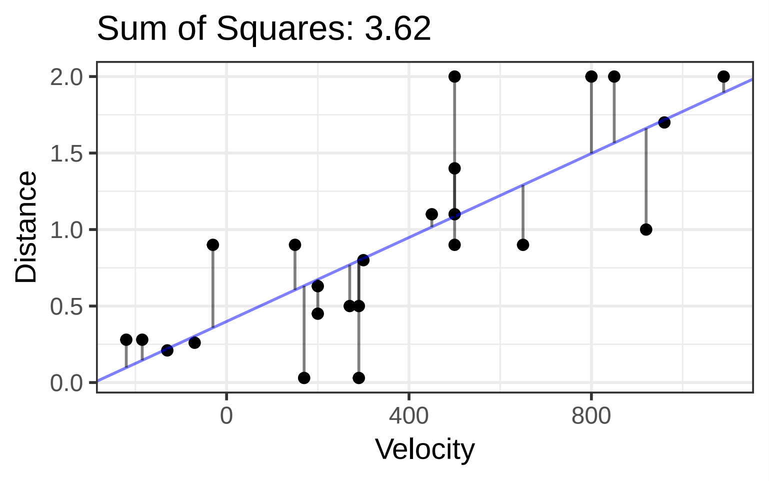
Better Fit



Best Fit (OLS Fit)



Best Fit (OLS Fit)



Closed Form Solutions

- You can use calculus to prove that the OLS fits are
- $\hat{\beta}_1 = \frac{s_y}{s_x}\rho$
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$

where

- $s_y =$ sample standard deviation of the Y_i's
- $s_x =$ sample standard deviation of the X_i's
- $\rho =$ sample correlation between the X_i's and Y_i's.

Estimate of σ^2

- Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, we can estimate the variance σ^2 using the residuals.
- $\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 - \hat{\beta}_1 X_i)$

Estimate of σ^2

- Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, we can estimate the variance σ^2 using the residuals.
- $\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 - \hat{\beta}_1 X_i)$
- $\hat{\sigma}^2 = (\hat{\epsilon}_1^2 + \hat{\epsilon}_2^2 + \cdots + \hat{\epsilon}_n^2)/\nu$
- $\hat{\sigma}^2 =$ Sum of squared residuals divided by the degrees of freedom.

Estimate of σ^2

- Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, we can estimate the variance σ^2 using the residuals.
- $\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 - \hat{\beta}_1 X_i)$
- $\hat{\sigma}^2 = (\hat{\epsilon}_1^2 + \hat{\epsilon}_2^2 + \cdots + \hat{\epsilon}_n^2)/\nu$
- $\hat{\sigma}^2 =$ Sum of squared residuals divided by the degrees of freedom.
- $\nu =$ degrees of freedom $= n - \#$parameters $= n - 2$

In R

- Use the `lm()` function (for **L**inear **M**odel)
- Always save this output.
- `coef()` returns the estimates of the regression "coefficients" (β_0 and β_1).

```
lmout <- lm(Distance ~ Velocity, data = case0701)
coef(lmout)
```

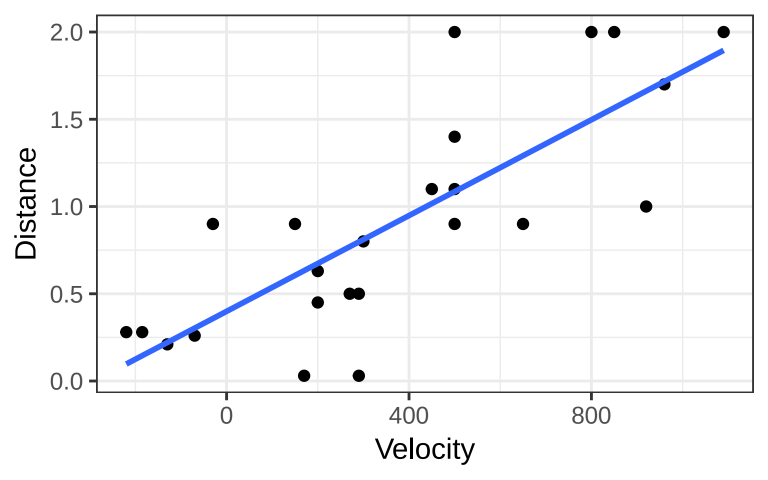
```
## (Intercept)    Velocity
##    0.399170    0.001372
```

- `sigma()` returns the estimate of the standard deviation.

```
## [1] 0.4056
```

Plot regression line

```r
qplot(Velocity, Distance, data = case0701,
      geom = "point") +
  geom_smooth(method = "lm", se = FALSE)
```



Sampling Distribution

- $\hat{\beta}_0$ and $\hat{\beta}_1$ both have *sampling distributions*.

Sampling Distribution

- $\hat{\beta}_0$ and $\hat{\beta}_1$ both have *sampling distributions*.
- Collect a new sample where **the new sample points have the same values of X_i**.

Sampling Distribution

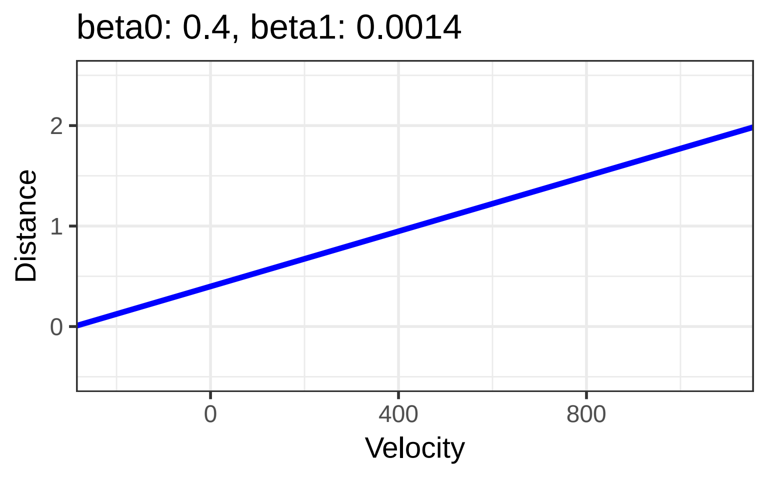
- $\hat{\beta}_0$ and $\hat{\beta}_1$ both have *sampling distributions*.
- Collect a new sample where **the new sample points have the same values of X_i**.
- Recalculate the least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$.

Sampling Distribution

- $\hat{\beta}_0$ and $\hat{\beta}_1$ both have *sampling distributions*.
- Collect a new sample where **the new sample points have the same values of** X_i.
- Recalculate the least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Repeat

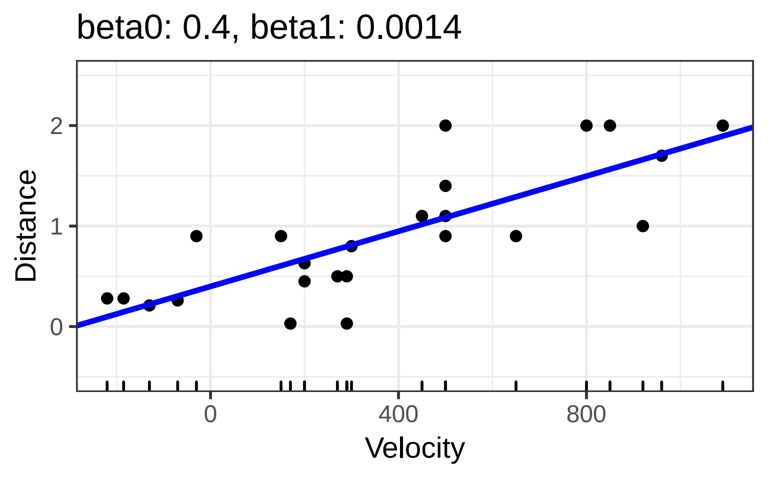
Sampling Distribution

- Ground Truth

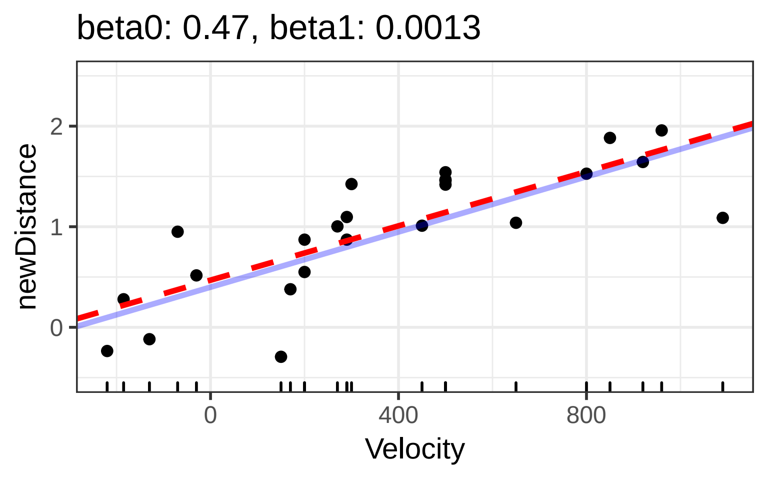


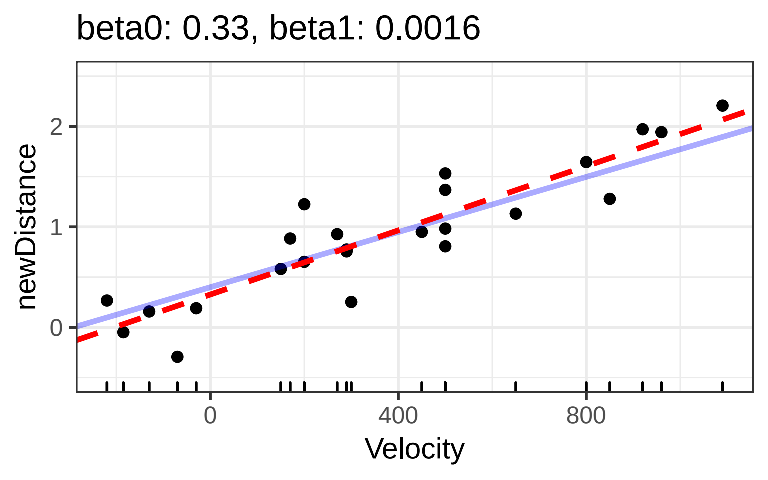
Sampling Distribution

- Our Observed Data

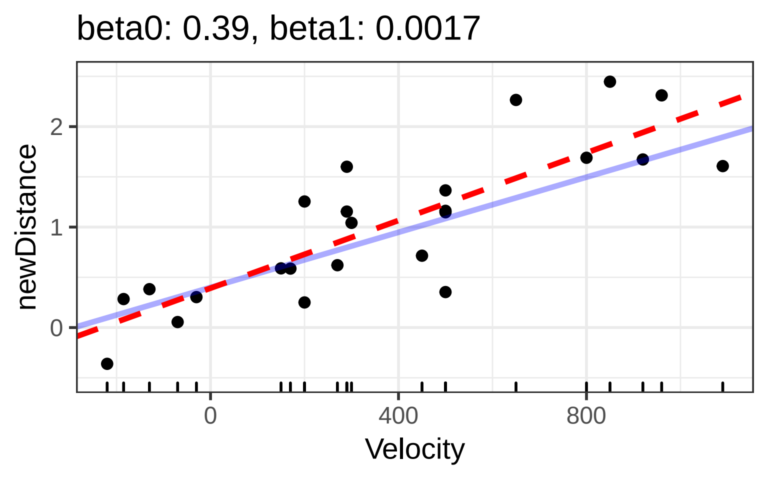


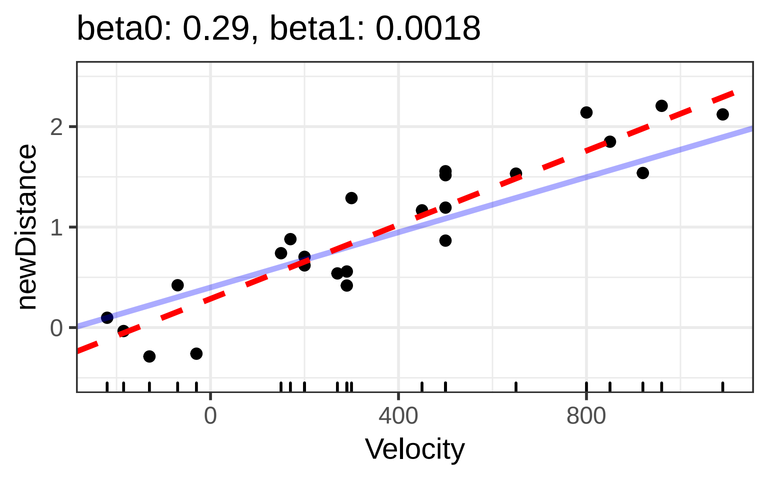
Sampling Distribution i



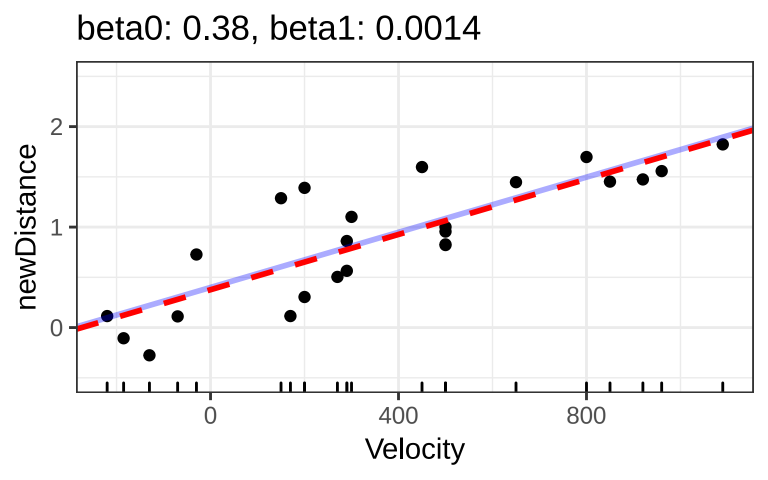

beta0: 0.33, beta1: 0.0016
newDistance
2
1
0
0
400
800
Velocity

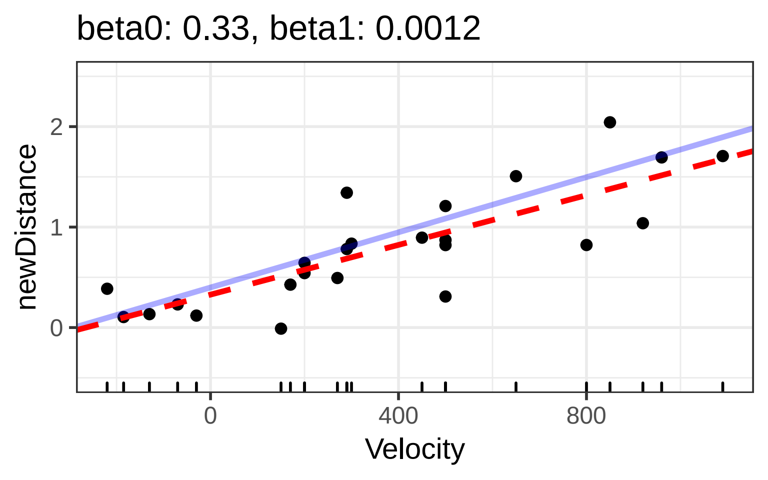
Sampling Distribution iii



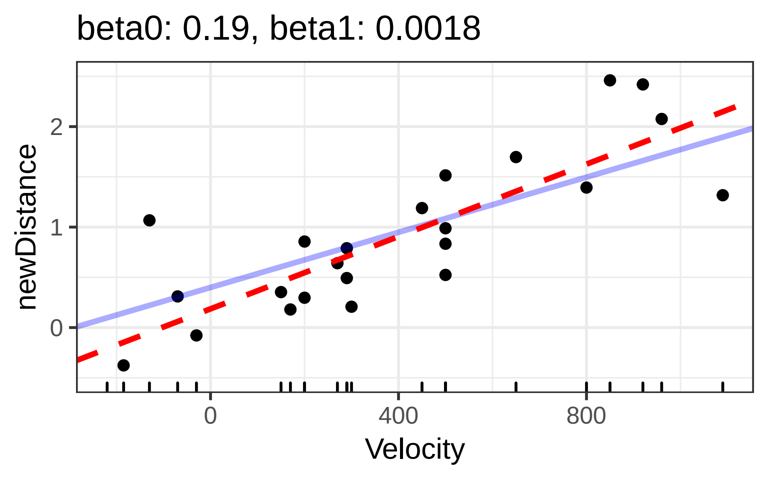

beta0: 0.29, beta1: 0.0018
newDistance
2
1
0
0
400
800
Velocity

Sampling Distribution v

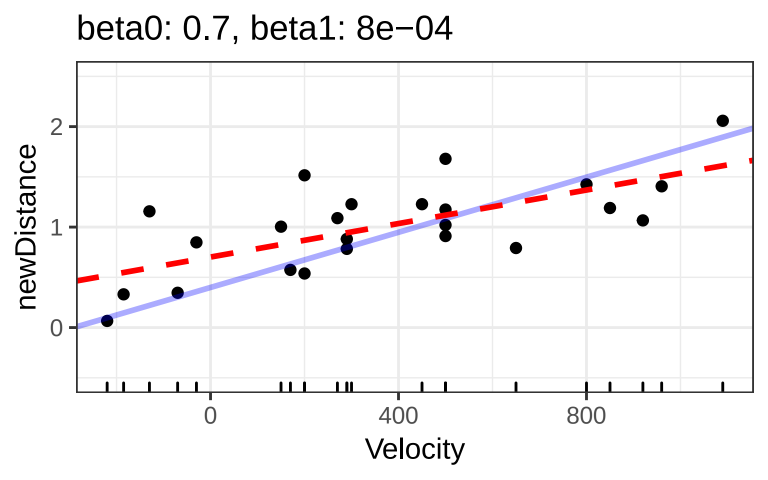



beta0: 0.33, beta1: 0.0012
newDistance
2
1
0
0
400
800
Velocity

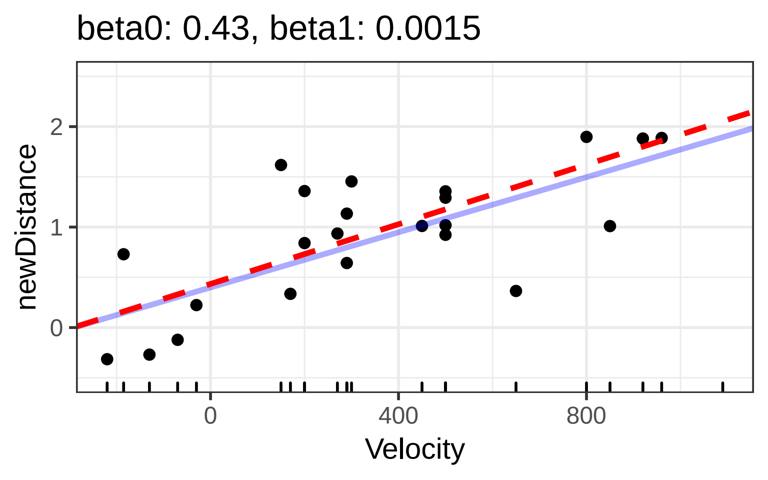
Sampling Distribution vii



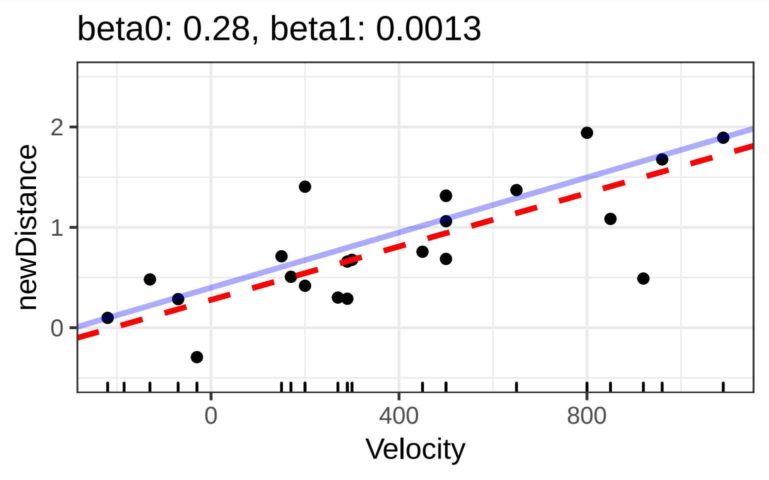
Sampling Distribution viii



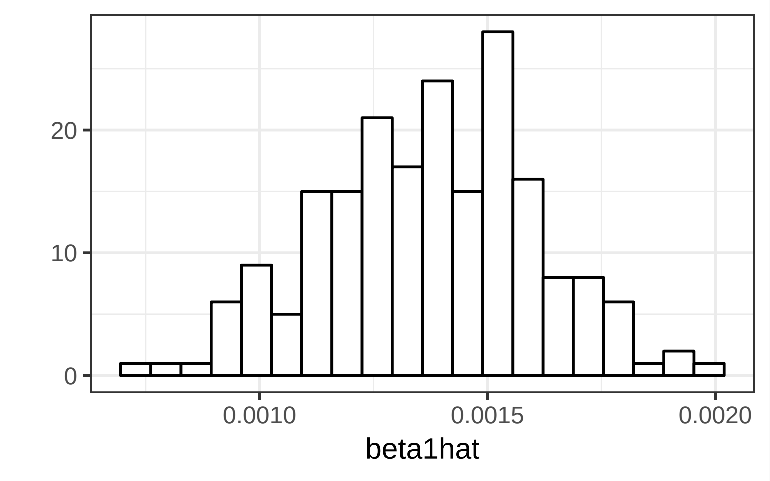
Sampling Distribution ix



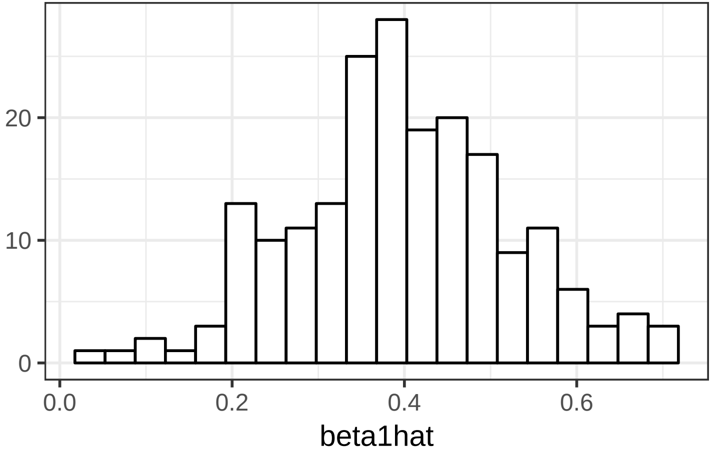
Sampling Distribution x



Sampling Distribution of $\hat{\beta}_1$



Sampling Distribution of $\hat{\beta}_0$



Theoretical Sampling Distributions

- A variant of the central limit theorem can be used to show that for large n
- $\hat{\beta}_1 \sim N(\beta_1, SD(\hat{\beta}_1))$
- $SD(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{(n-1)s_X^2}}$
- $\hat{\beta}_0 \sim N(\beta_0, SD(\hat{\beta}_0))$
- $SD(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}$
- The standard deviation formulas are complex (and not too important for you), but a computer can calculate them easily.

t-ratios

- So we have
- $\frac{\hat{\beta}_1 - \beta_1}{SD(\hat{\beta}_1)} \sim N(0,1)$
- $\frac{\hat{\beta}_0 - \beta_0}{SD(\hat{\beta}_0)} \sim N(0,1)$

t-ratios

- So we have
- $\frac{\hat{\beta}_1 - \beta_1}{SD(\hat{\beta}_1)} \sim N(0,1)$
- $\frac{\hat{\beta}_0 - \beta_0}{SD(\hat{\beta}_0)} \sim N(0,1)$
- $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$
- $\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t_{n-2}$

t-ratios

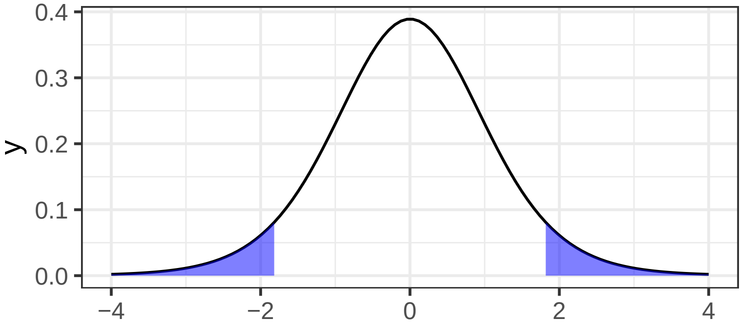
- So we have
- $\frac{\hat{\beta}_1 - \beta_1}{SD(\hat{\beta}_1)} \sim N(0,1)$
- $\frac{\hat{\beta}_0 - \beta_0}{SD(\hat{\beta}_0)} \sim N(0,1)$
- $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$
- $\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t_{n-2}$
- $SE(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}$
- $SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}$

Use t-ratios for testing hypotheses

- Under $H_0 : \beta_1 = 0$, we have

$$\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

- Compare observed t-statistic to theoretical t_{n-2} distribution and calculate p-values



Use t-ratios for confidence intervals

- The following is satisfied in 95% of repeated samples (again, where the covariate levels do not change):

$$t_{n-2}(0.025) \leq \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1))} \leq t_{n-2}(0.975)$$

- Solve for β_1 to get a 95% confidence interval

$$\hat{\beta}_1 \pm t_{n-2}(0.975) SE(\hat{\beta}_1)$$

Obtaining these in R

```r
lmout <- lm(Distance ~ Velocity, data = case0701)
summary(lmout)
```

```
##
## Call:
## lm(formula = Distance ~ Velocity, data = case0701)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7672 -0.2352 -0.0108  0.2108  0.9146
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.399170   0.118666    3.36   0.0028
## Velocity    0.001372   0.000228    6.02  4.6e-06
##
## Residual standard error: 0.406 on 22 degrees of freedom
## Multiple R-squared:  0.623,  Adjusted R-squared:  0.605
## F-statistic: 36.3 on 1 and 22 DF,  p-value: 4.61e-06
```

Obtaining these in R

```r
confint(lmout)
```

```
##                   2.5 %   97.5 %
## (Intercept) 0.1530719 0.645269
## Velocity    0.0008999 0.001845
```

Interpretation of Coefficient Estimates

Randomized Experiments

- A one unit increase in X results in a β_1 unit increase in Y.
- E.g. Every hour after slaughter decreases the pH in the postmortem muscle of a steer carcus by 0.21 pH units ($p < 0.001$, 95% CI -0.25 to -0.16).
- The words and phrases “decreases”, “increases”, “results in” are causal.

Interpretation of Coefficient Estimates

Observational Study

- Populations that differ only by one unit of X tend to differ by β_1 units Y.
- E.g. Nebulae that have a receding velocity 1 km/sec faster tend to be 0.0014 megaparsecs further from Earth ($p < 0.001$, 95% CI of 0.00090 0.0018).
- The words "differ" and "difference" are less causal.

Back to Big Bang

Case Study

- The theory of Big Bang suggests a formal relationship between the distance between any two celestial objects (Y) and the recession velocity (X) between them (how fast they are moving apart) given the (unknown) age of the universe (T):

$$Y = TX$$

Questions of Interest

- The formula describes a line with zero intercept. Is the intercept zero?
- What is the age of the universe (estimate T)?

Test if β_0 is 0

```r
sumout <- summary(lmout)
coef(sumout)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.399170  0.1186662   3.364 2.803e-03
## Velocity    0.001372  0.0002278   6.024 4.608e-06
```

- We reject H_0 and conclude that the intercept is not 0.

Estimate Age of Universe

- If the big-bang theory were correct, $\beta_0 = 0$, so we would fit assuming $\beta_0 = 0$ to estimate β_1 (the age of the universe)

```
lm_noint <- lm(Distance ~ Velocity - 1, data = case0701)
cbind(coef(lm_noint), confint(lm_noint))
```

```
##                      2.5 %   97.5 %
## Velocity 0.001921 0.001526 0.002317
```

- Estimated age is 0.001921 megaparsec-second per km, with a 95% confidence interval of 0.001526 to 0.002317 megaparsec-second per km.
- Possible to convert these units to years.