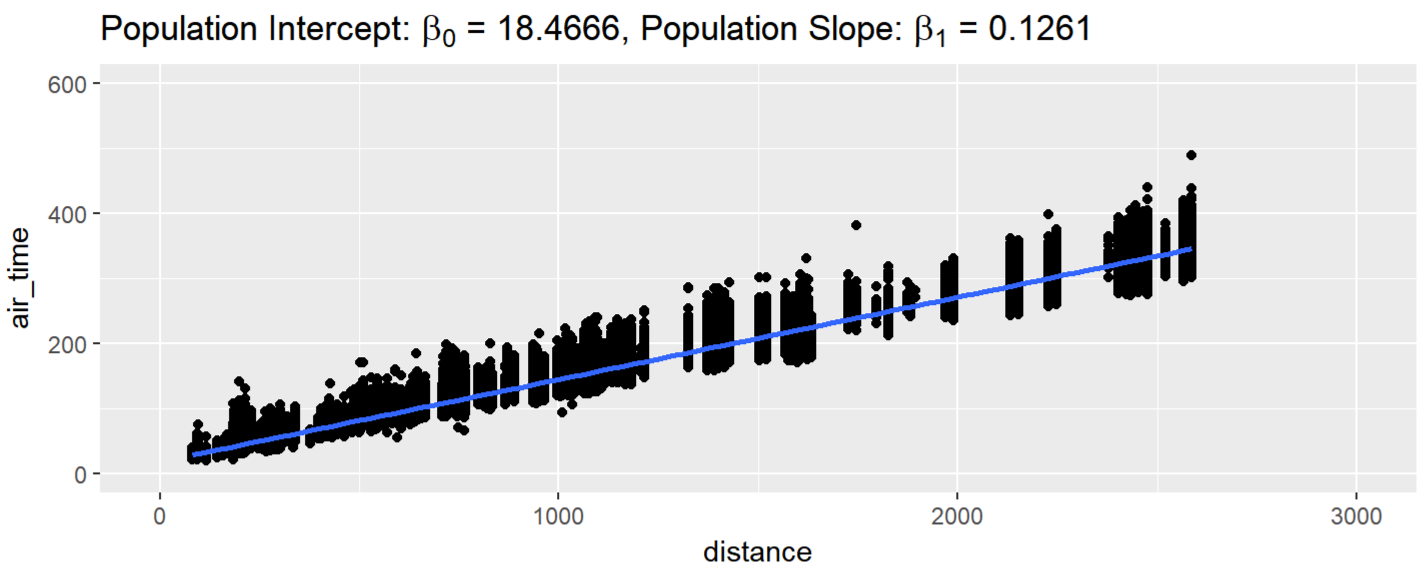
Linear Model Theory (ish)

Evan L. Ray

Population: Every Flight that Departed from NYC in 2013



Population Model:

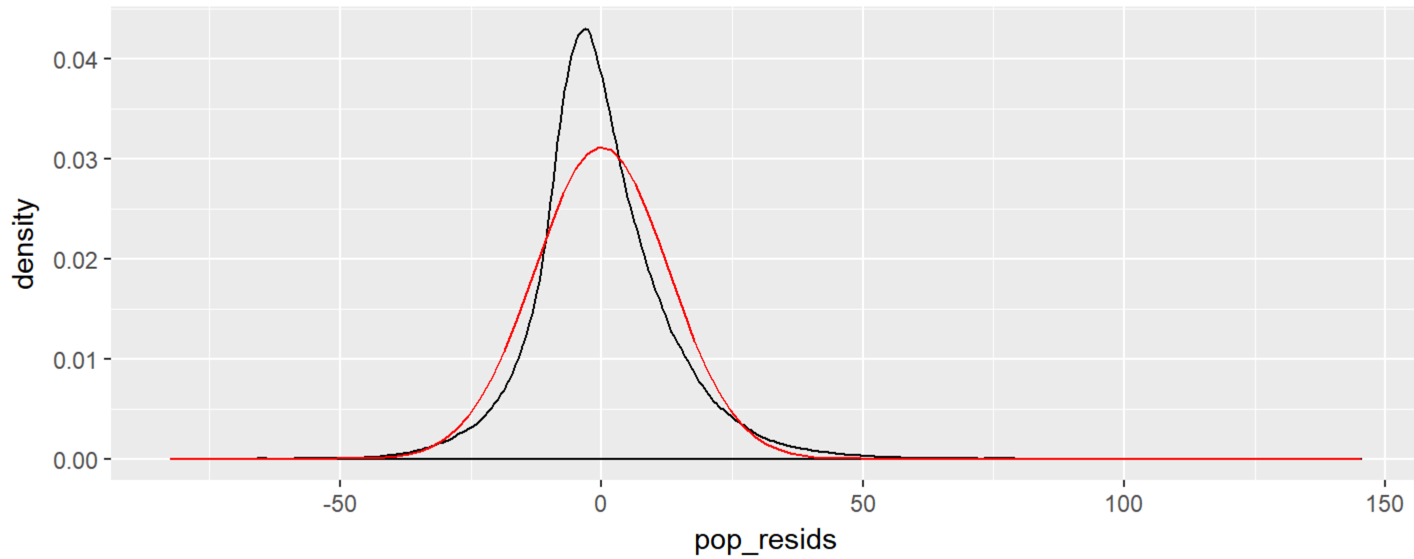
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$\varepsilon_i \sim \text{Normal}(0, \sigma)$

Residuals Distribution in Population:

Not exactly normal, but close enough.

Black: Actual distribution of residuals; Red: The closest normal approximation

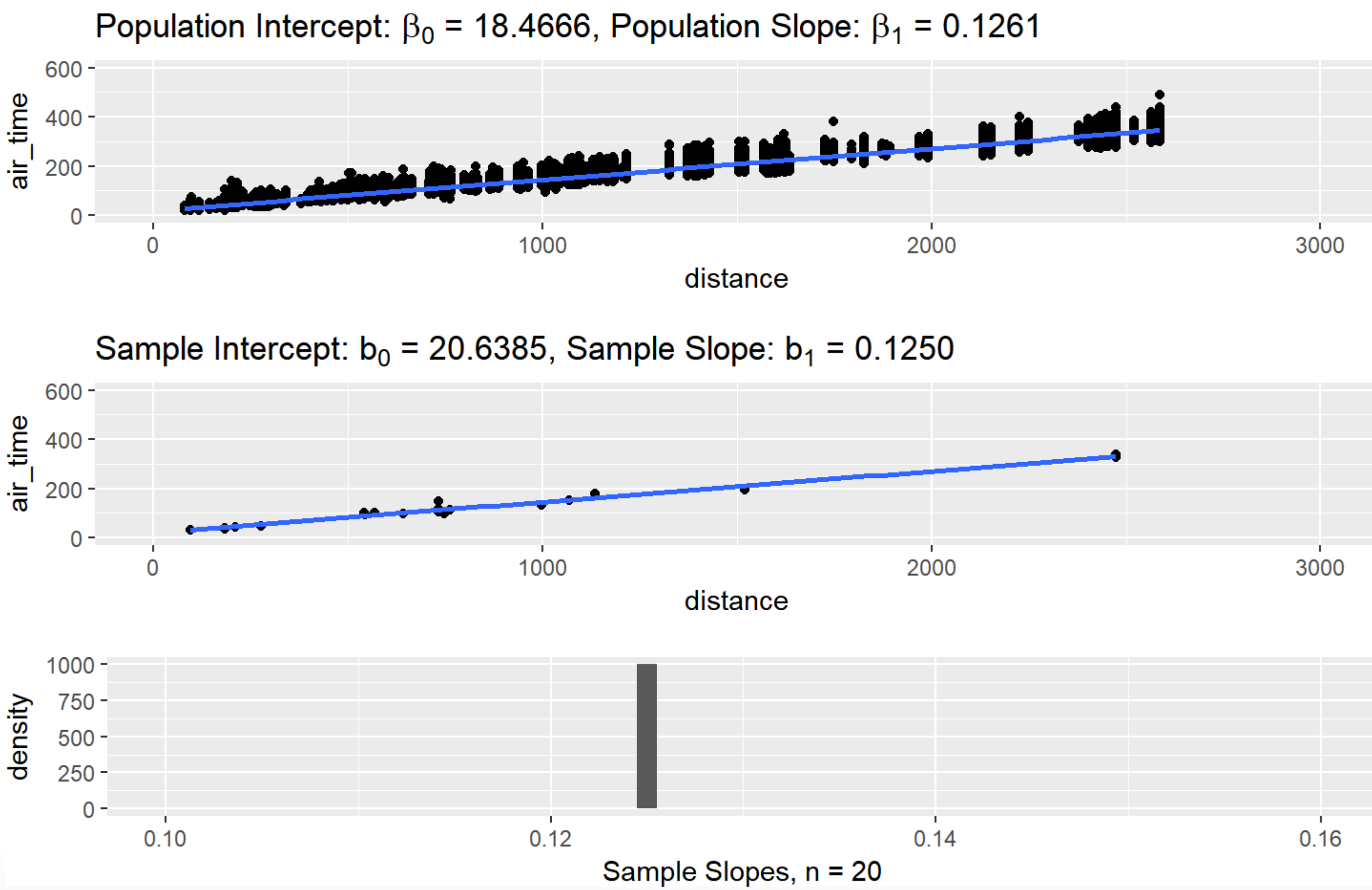


Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples

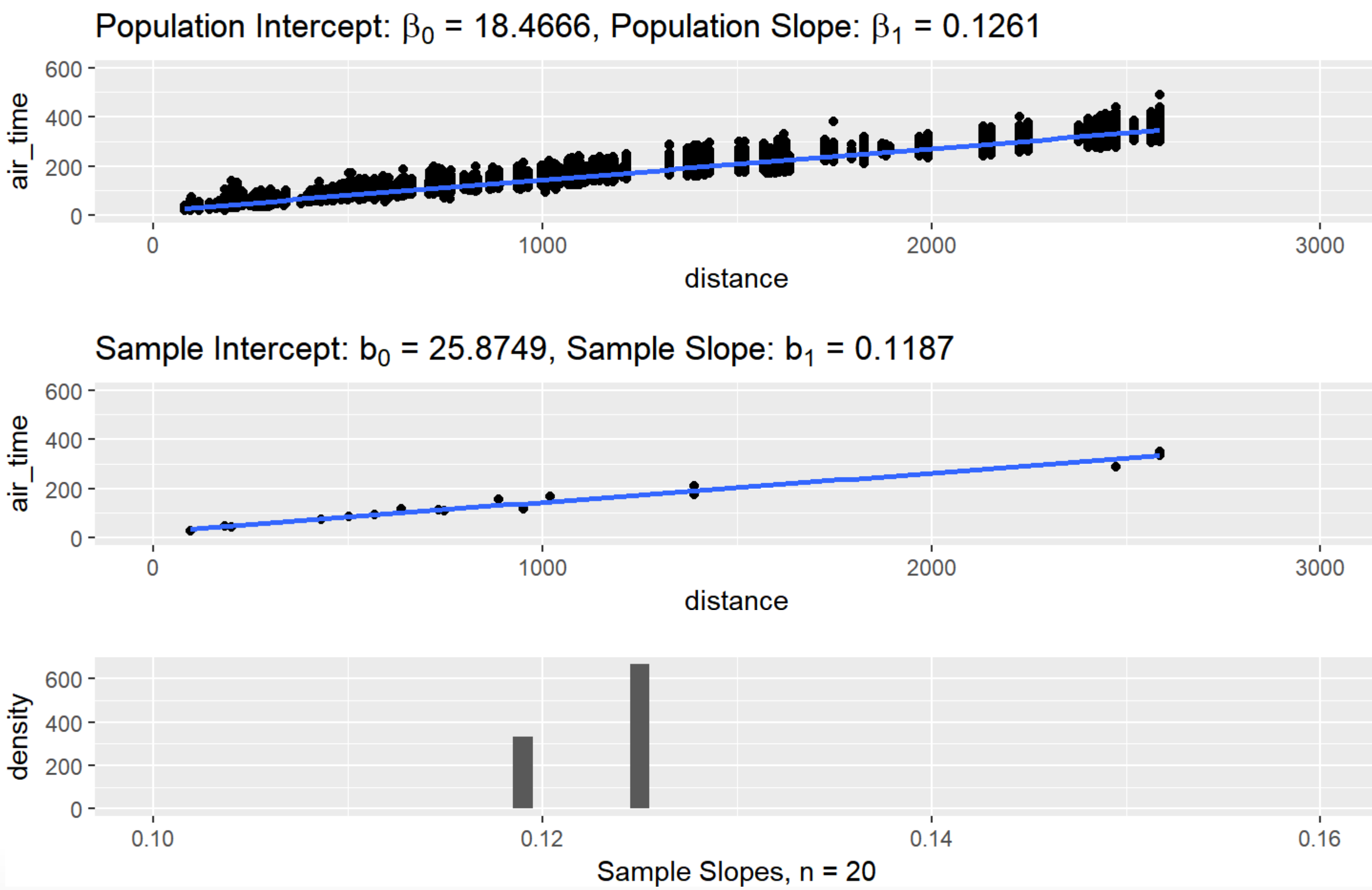
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



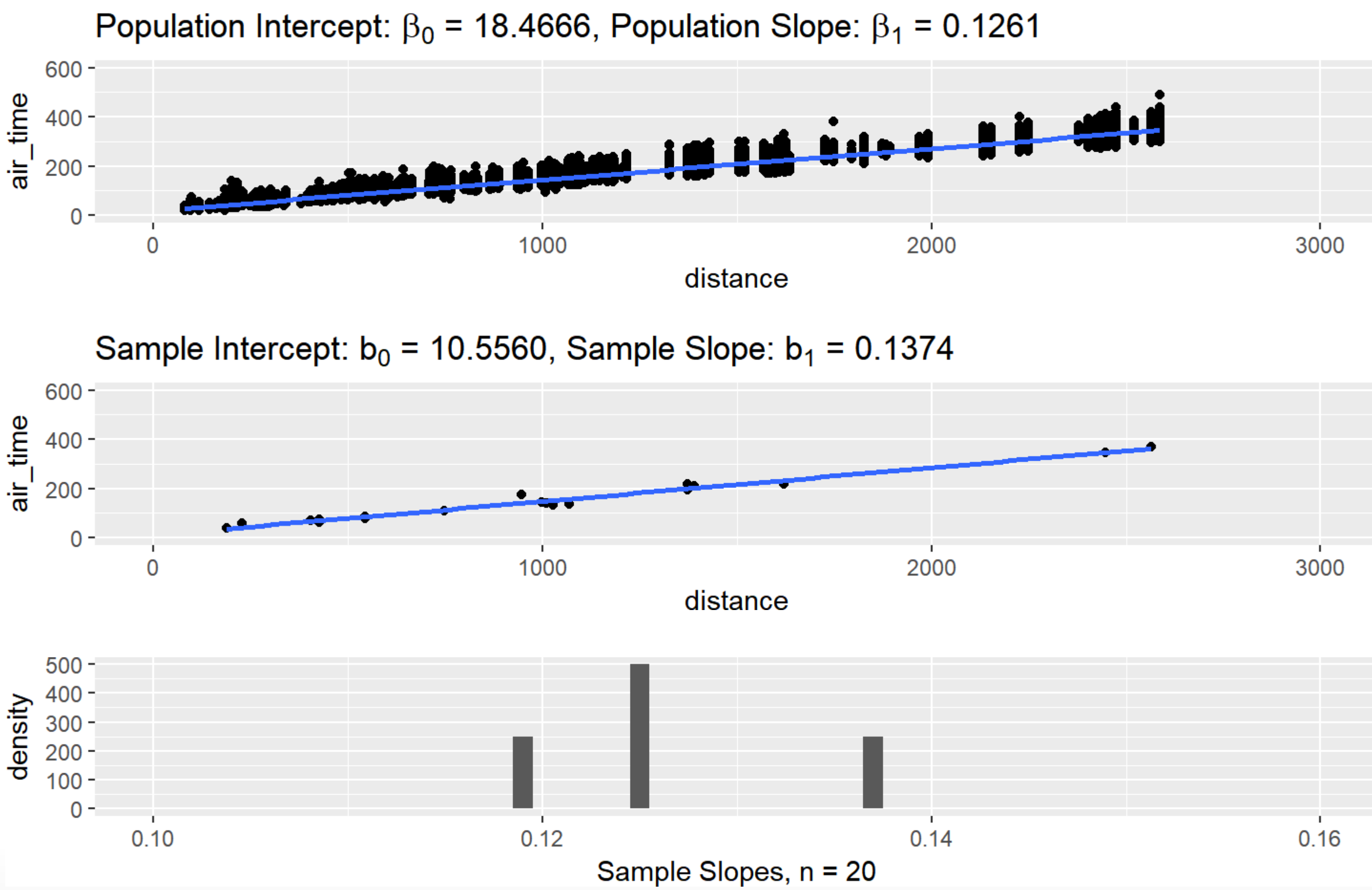
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



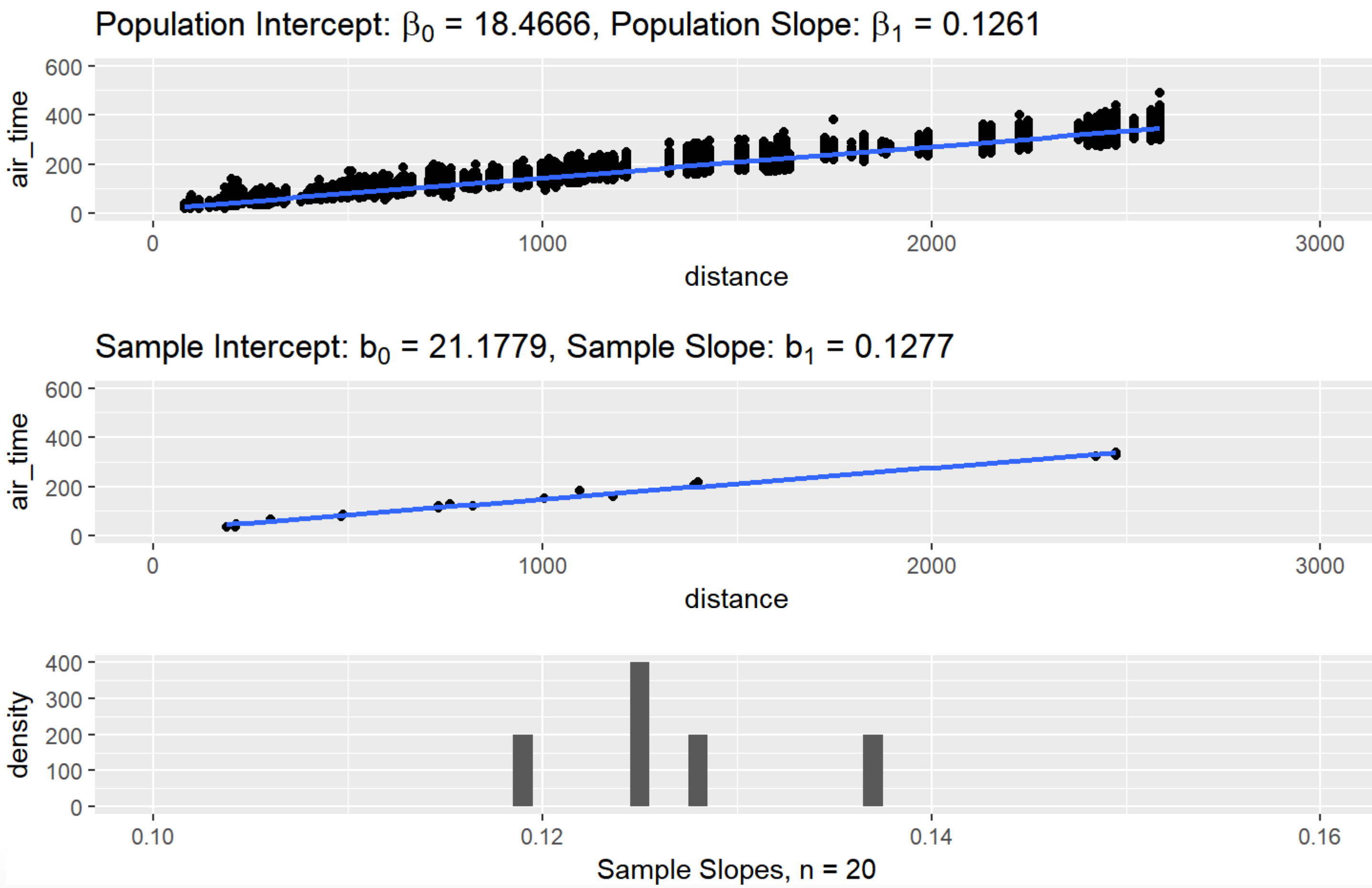
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



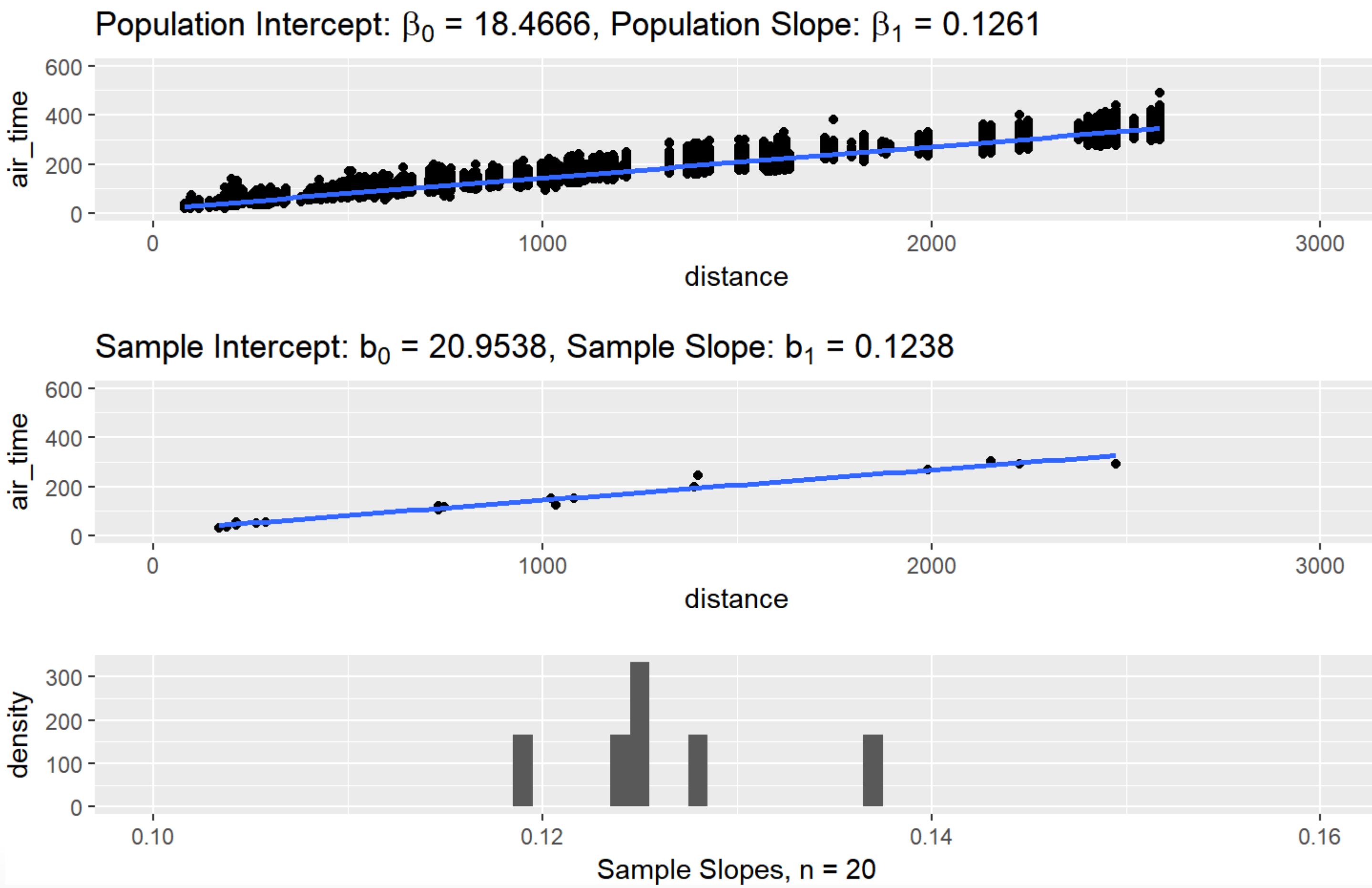
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



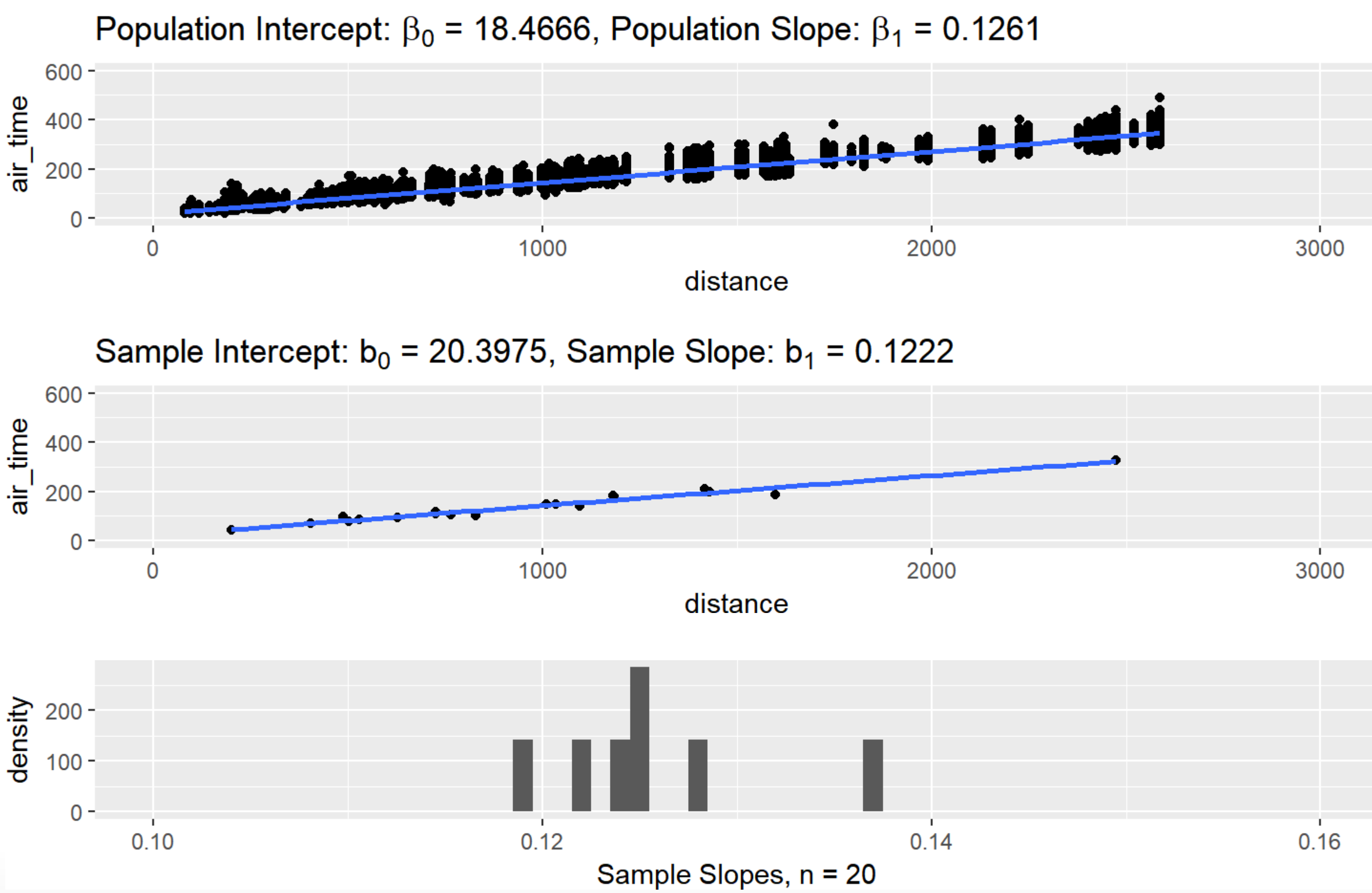
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



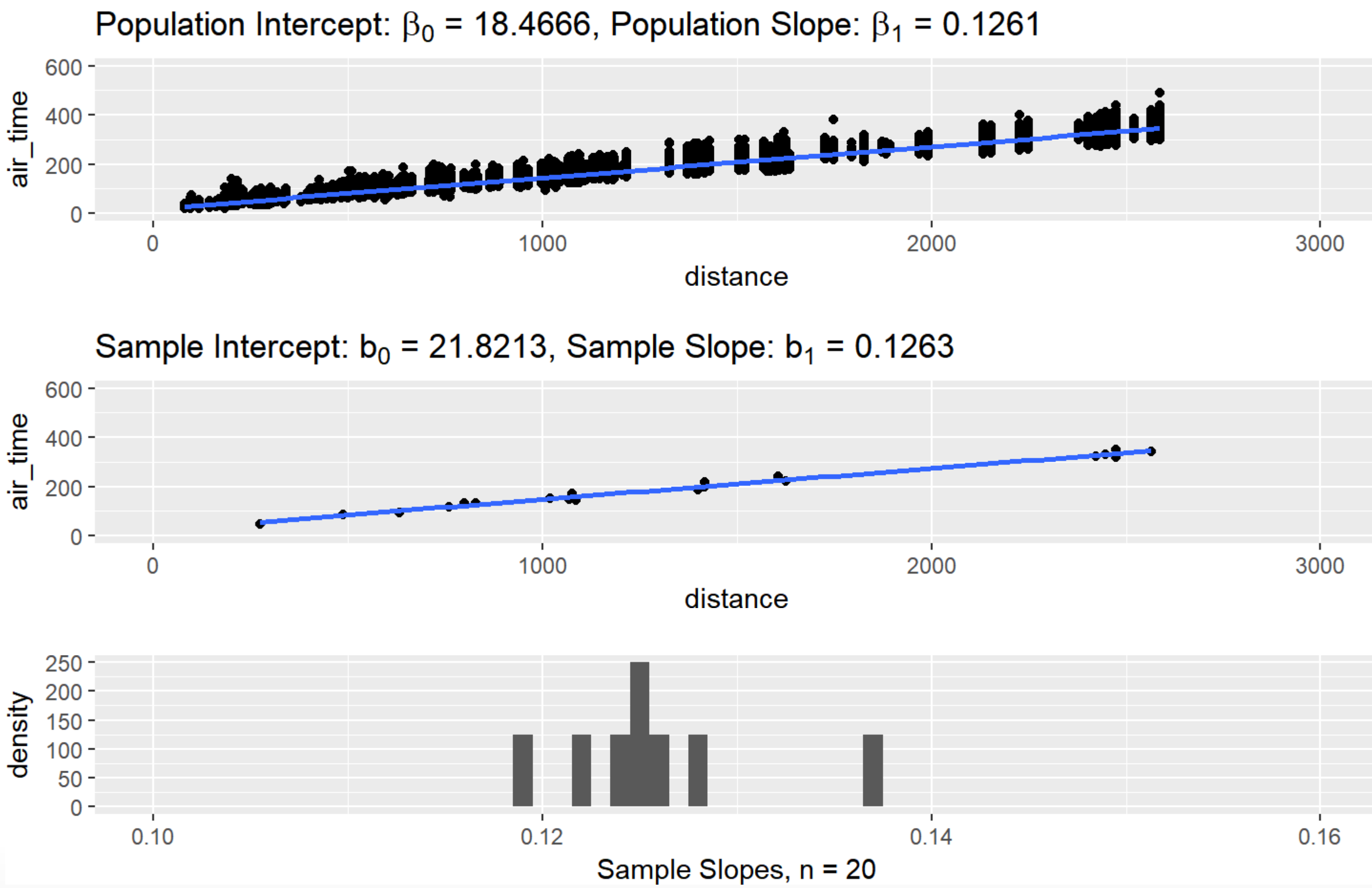
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



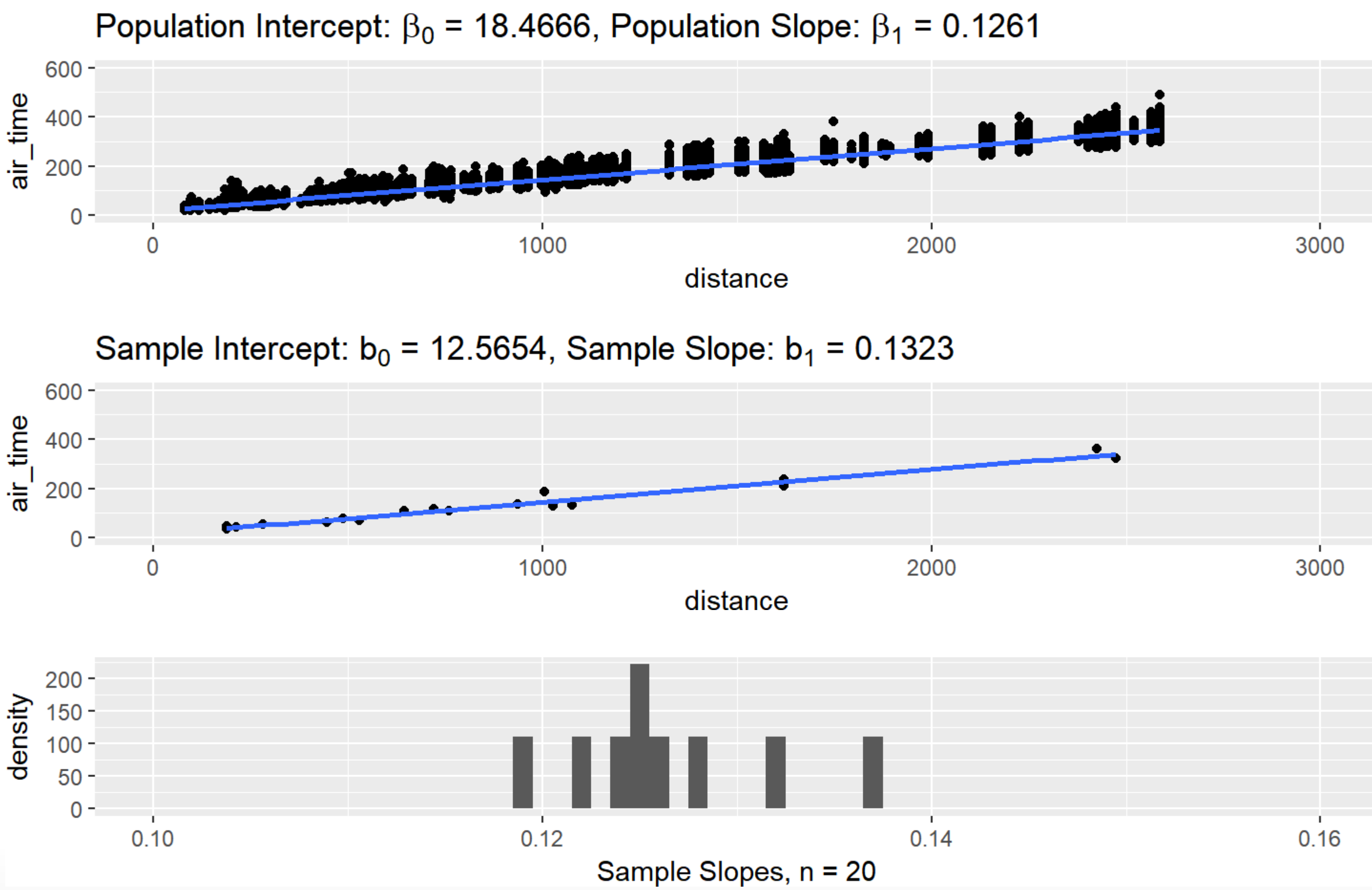
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



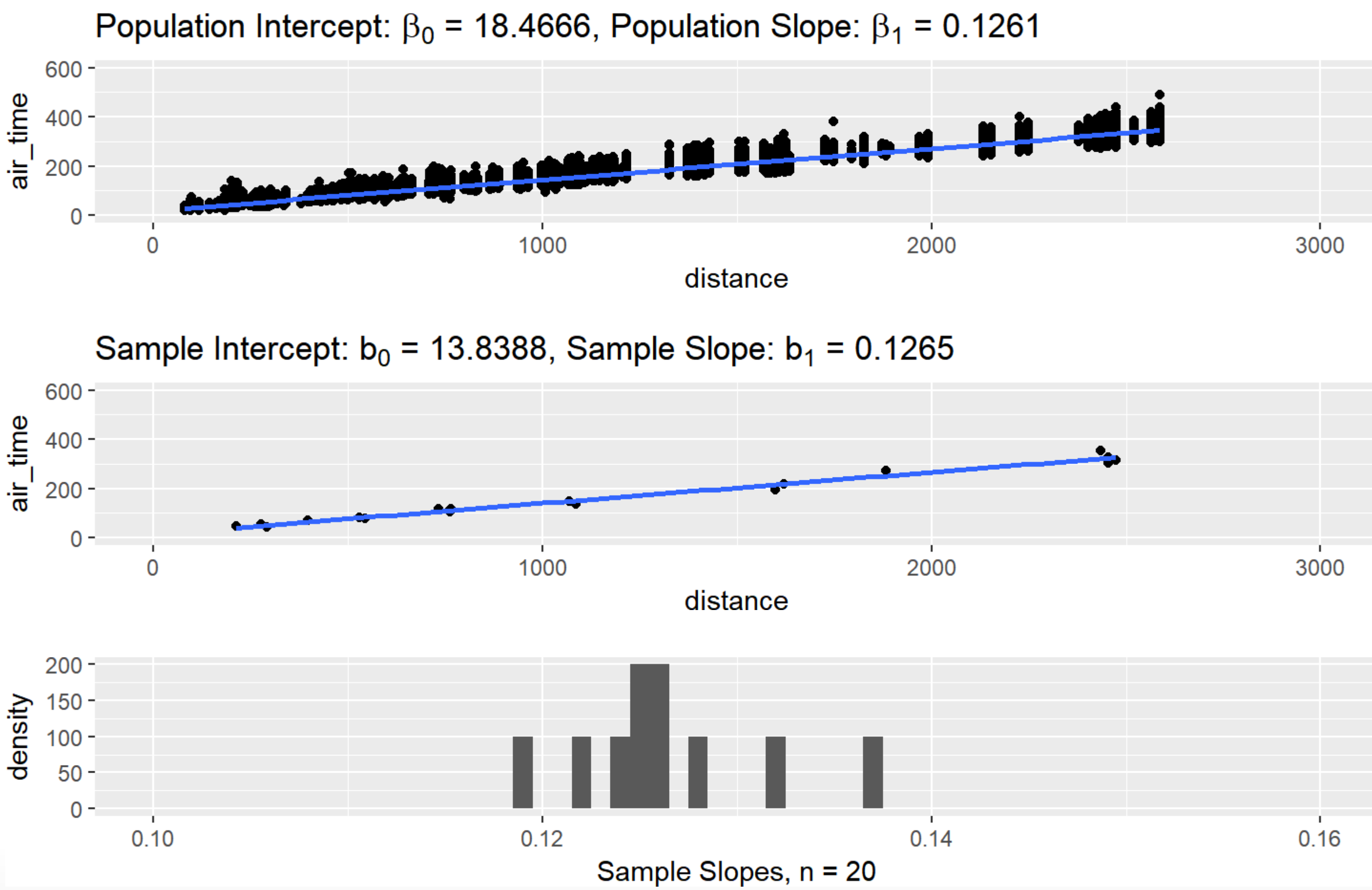
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



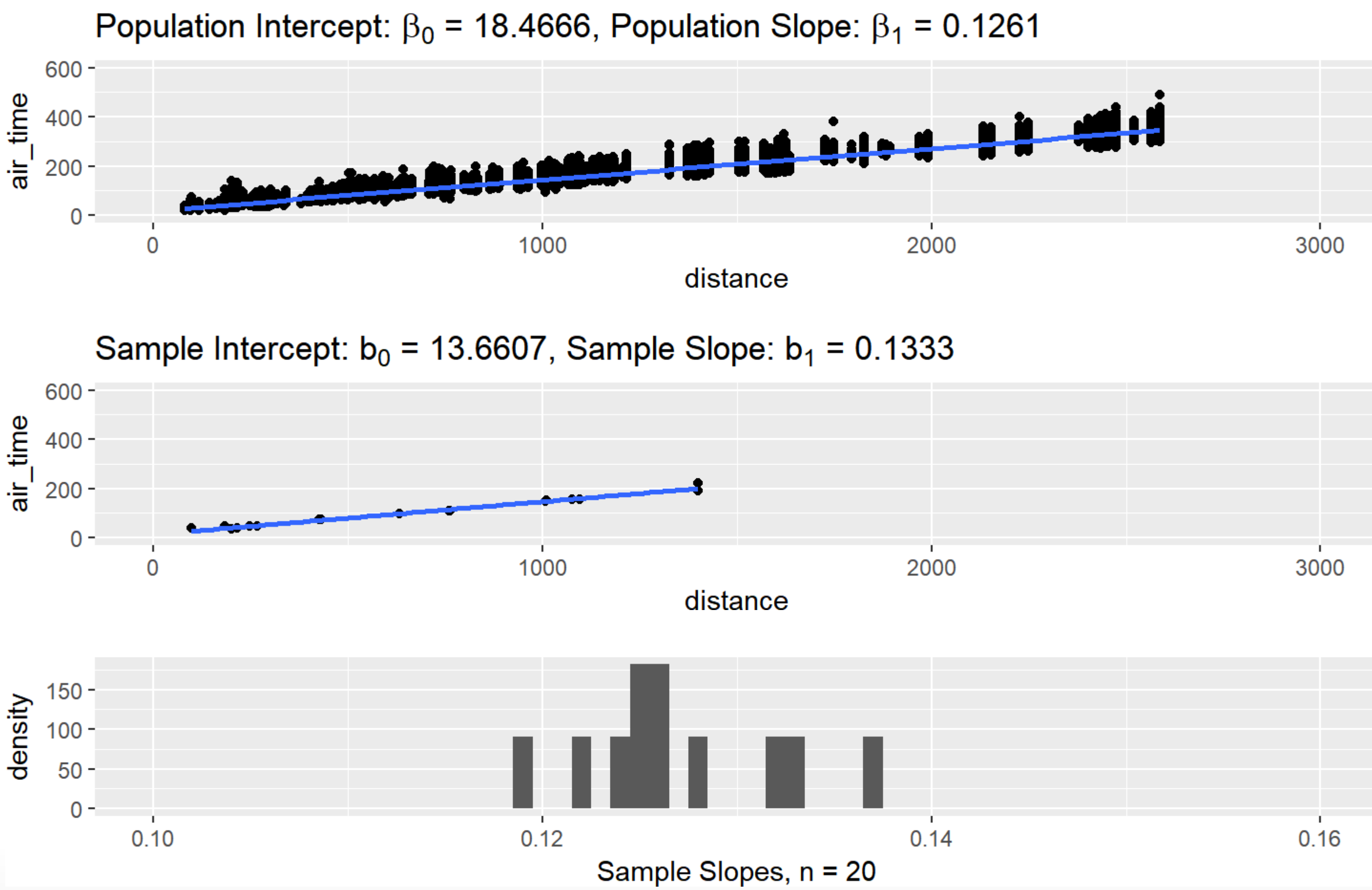
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



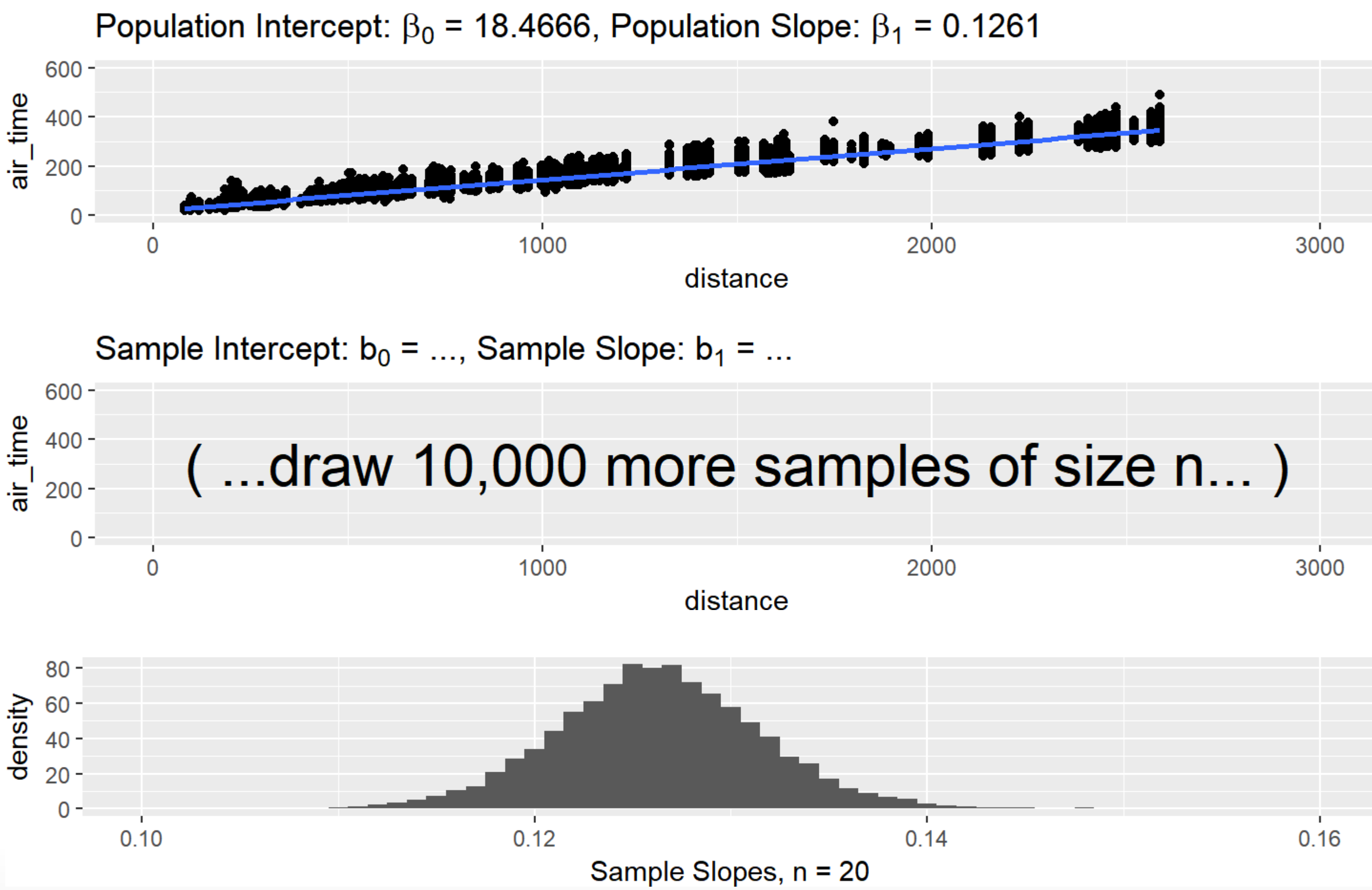
Sampling Distribution of b_1

The distribution of slope estimates b_1, across all different samples



Sampling Distribution of b_1

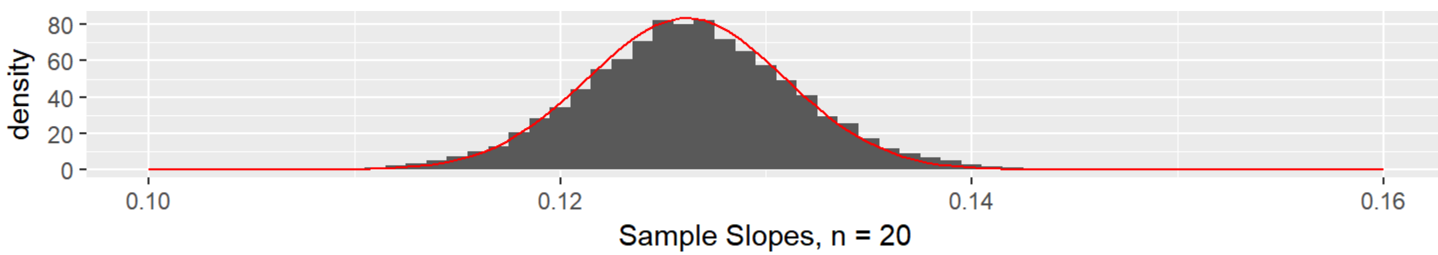
The distribution of slope estimates b_1, across all different samples



Sampling Distribution of b_1

- If all of the conditions for inference are satisfied (R. O'LINE) then

$b_1 \sim \text{Normal}\left(\beta_1, SD(b_1)\right)$, where $SD(b_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$



- (This is also still approximately true if most of the assumptions are mostly satisfied.)
- Recall: Probabilities involving the normal distributions only depend on how many standard deviations away from the mean we are:

$\frac{b_1 - \beta_1}{SD(b_1)}$

- **Problem**: This is not useful in practice, because we do not know σ (actual standard deviation of residuals in the population), so can't find $SD(b_1)$

What can we do?

- Estimate $SD(b_1)$. An estimate of a standard deviation is called a standard error.

$$SE(b_1) = \left[\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n-2) \sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{1/2}$$

$$\frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}$$

How to use this for hypothesis tests?

Null hypothesis: $\beta_1 = 0$

Alternative hypothesis: $\beta_1 \neq 0$

- **p-value:** Probability of getting a test statistic at least as extreme as what we got based on this sample, assuming the null hypothesis is true.
- **test statistic:** $t = \frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}$
- If null hypothesis is true, $t = \frac{b_1 - \beta_1}{SE(b_1)} = \frac{b_1 - 0}{SE(b_1)}$
  - "How many estimated standard deviations away from the hypothesized slope was our sample slope?"
- (Calculation of p-value hand-drawn on board)

How to use this for Conf. Intervals?

- For a 95% CI, find the value t^* with $P(-t^* \leq \frac{b_1-\beta_1}{SE(b_1)} \leq t^*) = 0.95$
- This means that for 95% of samples, $-t^* \leq \frac{b_1-\beta_1}{SE(b_1)} \leq t^*$
- …so for 95% of samples, $-t^* SE(b_1) \leq b_1 - \beta_1 \leq t^* SE(b_1)$
- …so for 95% of samples, $-b_1 - t^* SE(b_1) \leq -\beta_1 \leq -b_1 + t^* SE(b_1)$
- …so for 95% of samples, $b_1 - t^* SE(b_1) \leq \beta_1 \leq b_1 + t^* SE(b_1)$
- Confidence interval: $[b_1 - t^* SE(b_1), b_1 + t^* SE(b_1)]$
- In R, t^* = `qt(0.975, df = n - 2)` for a 95% CI.