Population Model:

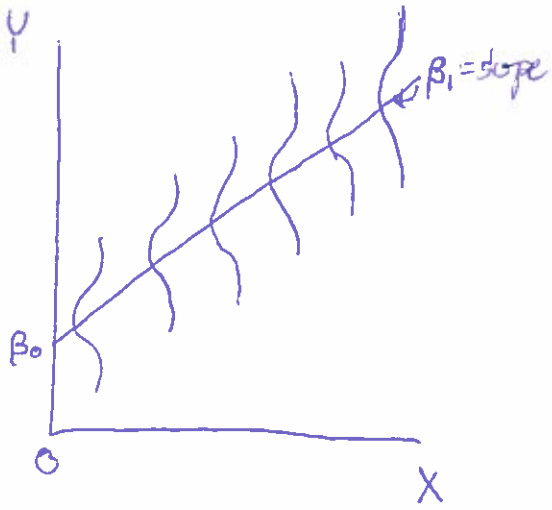
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

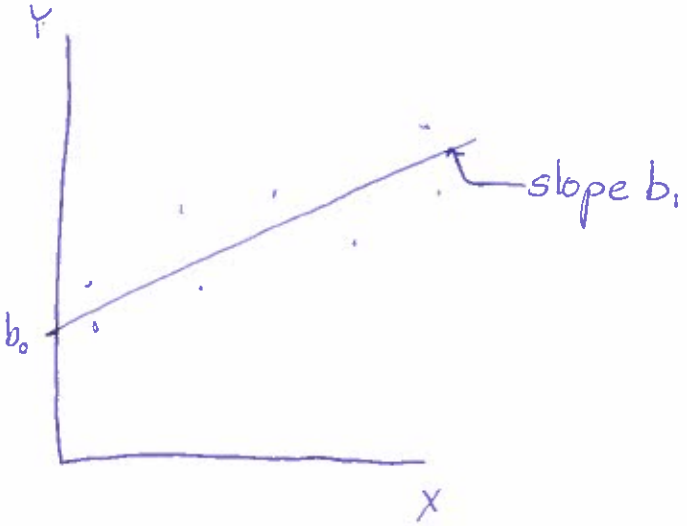
$\varepsilon_i \sim \text{Normal}(0, \sigma_e)$

Sample Regression Line

Predicted $\hat{y}_i = b_0 + b_1 x_i$

Residual: $e_i = y_i - \hat{y}_i$





We use the data in the sample to try to say some thing about the relationship between X and Y in the population.
Each sample from the population has a different regression line.

Sampling distributions:

$b_0 \sim \text{Normal}(\beta_0, SD(b_0)) \longrightarrow \frac{b_0 - \beta_0}{SE(b_0)} \sim t_{n-2}$

$b_1 \sim \text{Normal}(\beta_1, SD(b_1)) \longrightarrow \frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}$

↑ slope from the sample (sample statistic)

↑ slope from the population (population parameter)

Assumptions:

- Outliers (no outliers)
- Linear relationship ("straight enough")
- Independent
- Normal distribution of residuals (at least approximately)
- Equal variance of residuals ("does the plot thicken?")

Confidence Interval format:

$$b_1 \pm t^* \cdot SE(b_1)$$

t^* is the critical value for a t distribution with $n-2$ degrees of freedom:

e.g.: qt(0.975, df = 36)

if $n = 38$ is the sample size and we want a 95% confidence interval.

(b) Fit the linear model

```
# format is: lm(response_variable ~ explanatory_variable, data = data_frame)
lm_fit <- lm(Foals ~ Adults, data = horses)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Foals ~ Adults, data = horses)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.374 -3.312 -0.965  3.686 11.172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5784     1.4916   -1.06      0.3
## Adults        0.1540     0.0114   13.49 1.2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.94 on 36 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.83
## F-statistic:  182 on 1 and 36 DF,  p-value: 1.19e-15
```

This is the pvalue for a test of $H_0: \beta_0 = 0$ vs. $H_A: \beta_0 \neq 0$.

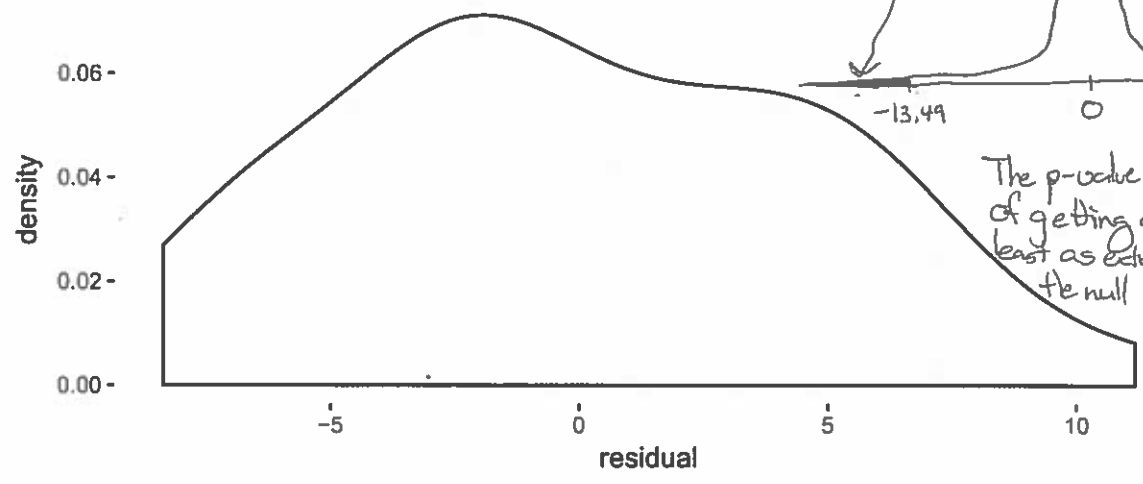
This is the p-value for a test of $H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$

The test statistic for this test is (use the value from the null hypothesis)

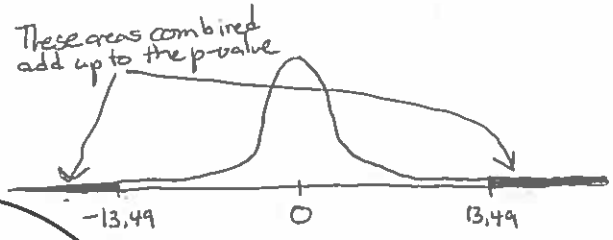
$$\frac{b_1 - \beta_1}{SE(b_1)} = \frac{0.154 - 0}{0.0114} = 13.49$$

(c) Check that the residuals follow a nearly normal distribution

```
horses <- mutate(horses,
  residual = residuals(lm_fit),
  predicted = predict(lm_fit))
ggplot() +
  geom_density(mapping = aes(x = residual), data = horses)
```



If the null hypothesis is true this statistic follows a t_{n-2} distribution:



The p-value is the probability of getting a test statistic at least as extreme as this, assuming the null hypothesis is true.

You can use the pt function in R to calculate these probabilities.