Stat 140: Inference for Simple Linear Regression Example - Wild Horses

Evan Ray

Wild Horses

What is the relationship between the size of a herd of horses and the number of foals (baby horses!!) that are born to that herd in a year?

```r
horses <- read_csv("https://mhc-stat140-2017.github.io/data/sdm4/Wild_Horses.csv")
```

```
## Parsed with column specification:
## cols(
##   Foals = col_integer(),
##   Adults = col_integer()
## )
```

```r
head(horses)
```

```
## # A tibble: 6 x 2
##   Foals Adults
##   <int>  <int>
## 1    28    232
## 2    18    172
## 3    16    136
## 4    20    127
## 5    20    118
## 6    20    115
```

```r
nrow(horses)
```

```
## [1] 38
```

Questions to Start With:

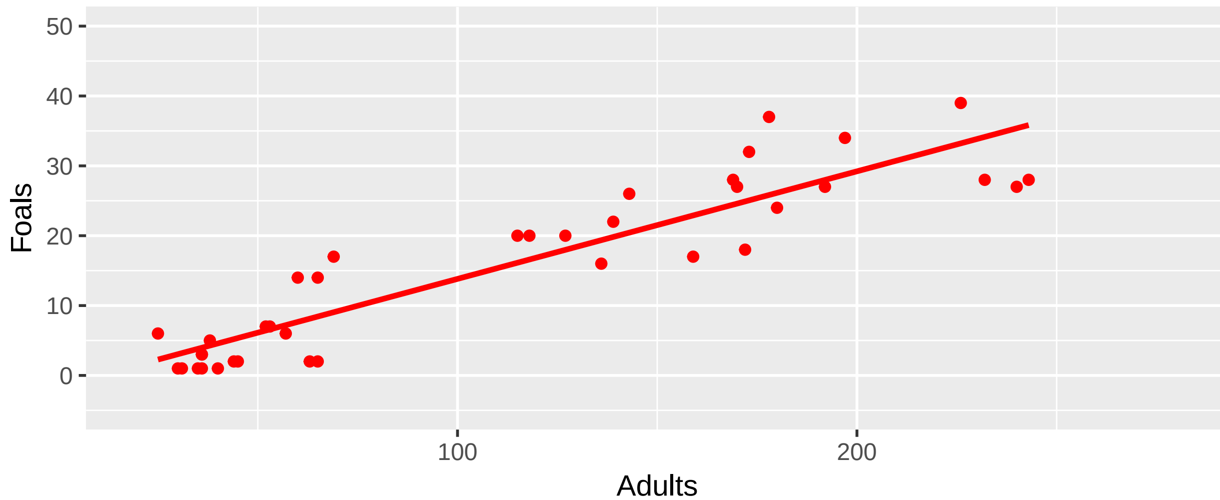
- What is the observational unit?

- What are the variable data types (**categorical** or **quantitative**)?
  - **Foals**:
  - **Adults**:

- Which of these variables is the **explanatory** variable and which is the **response**?
  - **Explanatory**:
  - **Response**:

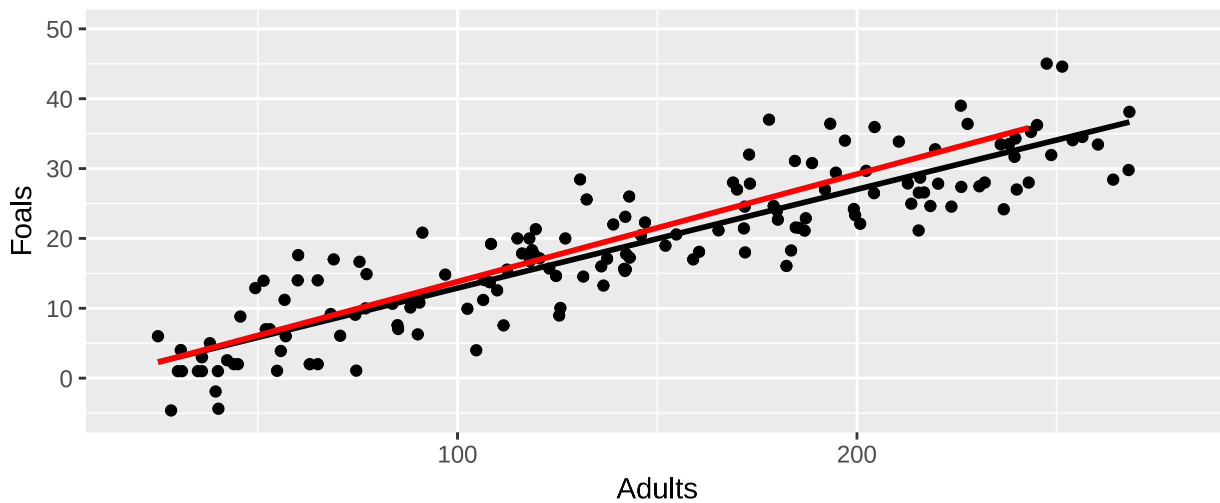
Previously: Fit linear regression to describe the relationship between number of adults and number of foals in the *sample.*

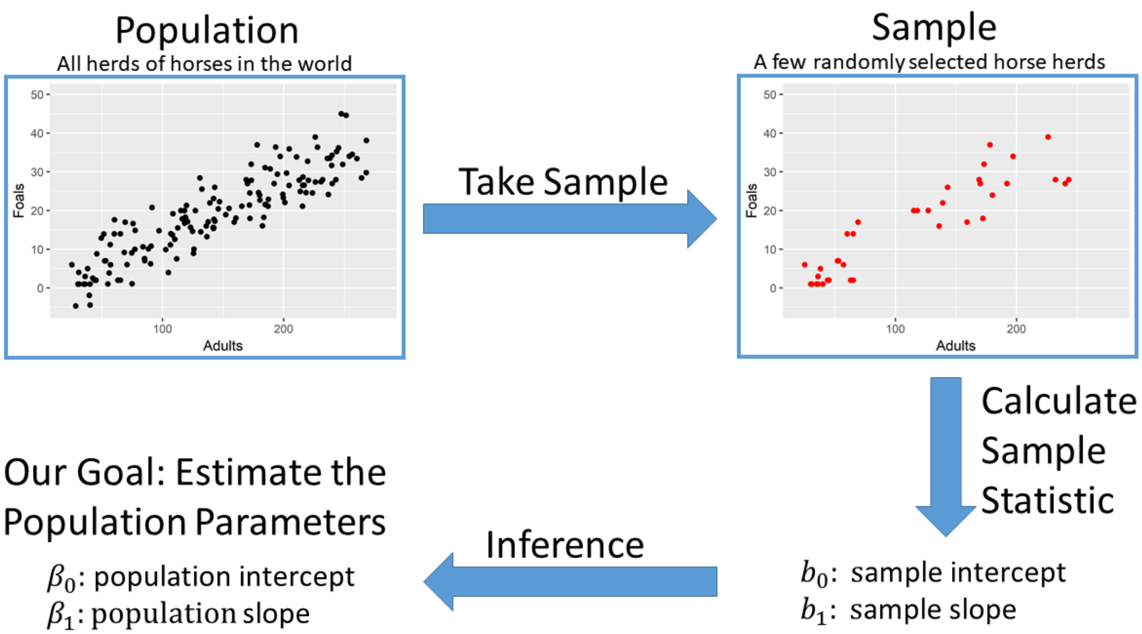
- Estimated Line: $\hat{y}_i = b_0 + b_1 x_i$
- b_0 and b_1 are **sample statistics**: they describe the data in our sample



Today: Use data from this sample to learn about the relationship between number of adults and number of foals in the *population*

- Population Model:
  - $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
  - $\varepsilon_i \sim \text{Normal}(0, \sigma)$
- β_0 and β_1 are **population parameters**: they describe the population




Population
All herds of horses in the world
Foals
Adults
Take Sample
Sample
A few randomly selected horse herds
Foals
Adults
Calculate
Sample
Statistic
b_0: sample intercept
b_1: sample slope
Inference
Our Goal: Estimate the
Population Parameters
β_0: population intercept
β_1: population slope

(a) Are the assumptions for the linear regression model met?

- **Representative** Sample

- (No) **Outliers**

- **Linear** Relationahip (Straight Enough)

- **Independent** Observations (Randomization)

- **Normal** Distribution of Residuals (Can't check this yet – need to look at a histogram or density plot of the residuals after fitting the model)

- **Equal** Variance of Residuals (Does the Plot Thicken?)

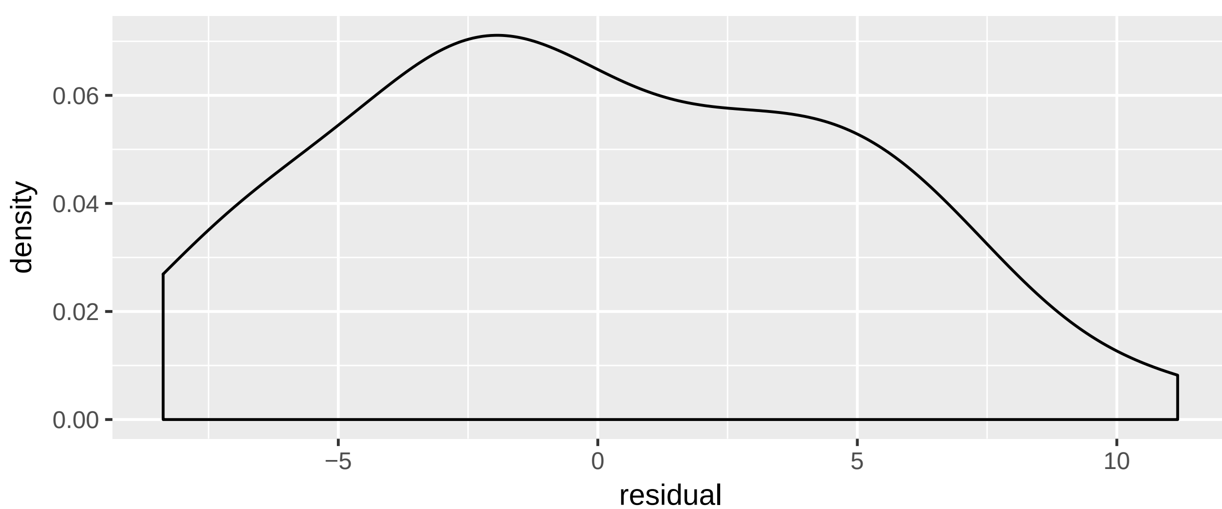
(b) Fit the linear model

```
# 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
```

(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)
```



(d) Explain in context what the regression says about the relationship between the number of adult horses in a herd and the number of foals born to that herd. Interpret both the intercept and the slope in context.

(e) Conduct a hypothesis test of the claim that when there are 0 adults in a herd, there will be 0 foals born to that herd.

(f) Draw a picture of a relevant t distribution for the hypothesis test in part (e) and shade in the region corresponding to the p-value. How would you calculate the p-value for part (e) using the pt function in R and the given estimate and standard error?

(g) Conduct a hypothesis test of the claim that there is no relationship between the number of adults in a herd and the number of foals who are born to that herd.

(h) Obtain a 99% confidence interval for the population intercept, β_0, and for the population slope, β_1. Interpret the confidence interval for β_1 in context.

```
## Note that unlike every other confidence interval function we've looked at,
## we set the confidence level with an argument called level, not conf.level
confint(lm_fit, level = 0.99)
```

```
##               0.5 % 99.5 %
## (Intercept) -5.6347  2.478
## Adults       0.1229  0.185
```

(i) How would you calculate the confidence interval for part (f) using the `qt` function in R and the given estimate and standard error?

```
qt(0.995, df = 38 - 2)
```

```
## [1] 2.719
```

(j) Interpret the standard error for the slope using the "95" part of the 68-95-99.7 rule.