Inference for Linear Regression

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?

```
## # 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
```

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

Questions to Start With:

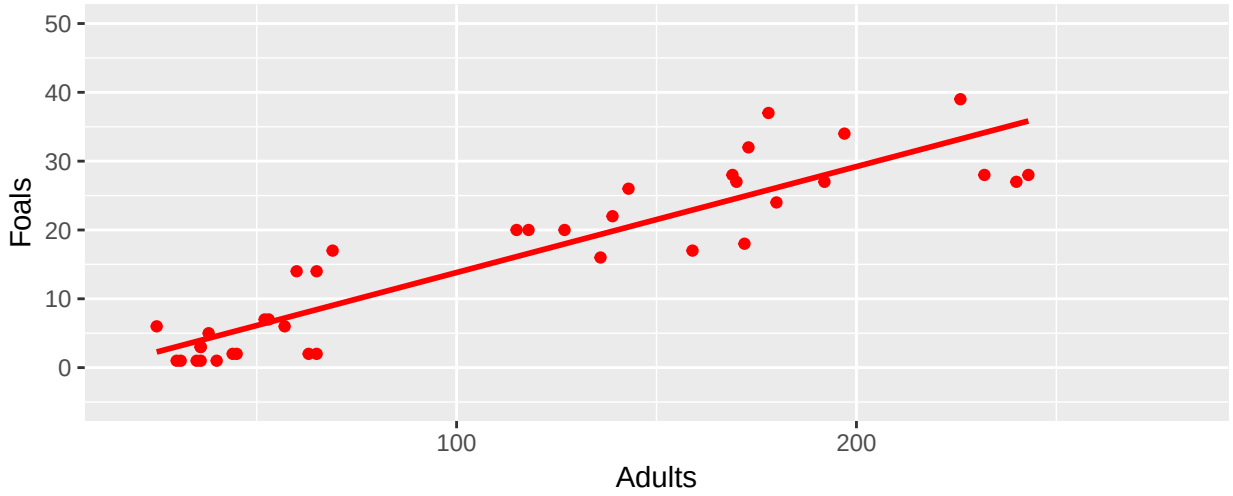
- What is the observational unit?

- 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**:

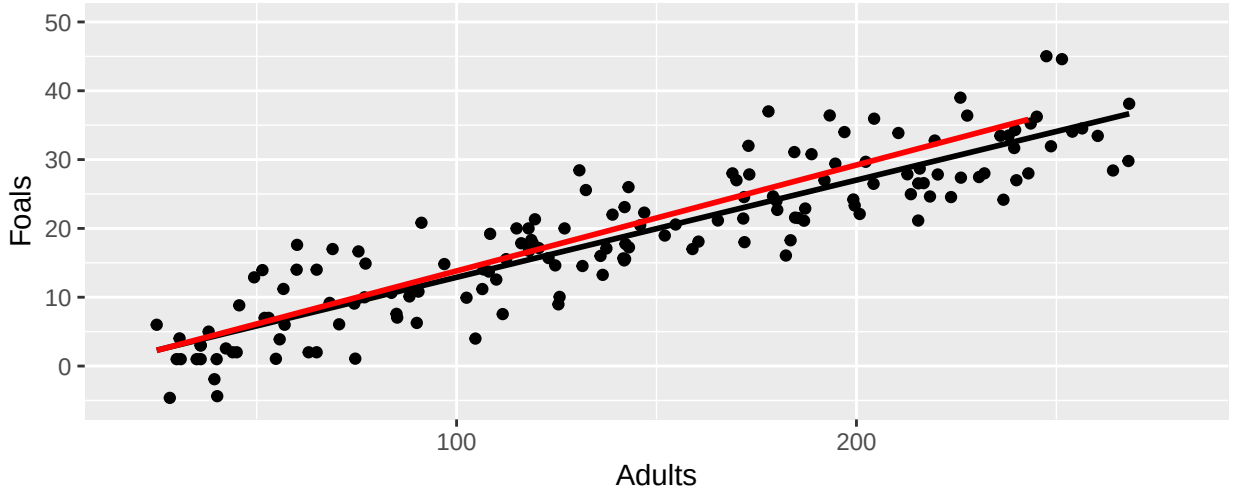
Previously: Fit linear regression to describe the relationship between number of adults and number of foals in the *sample*.

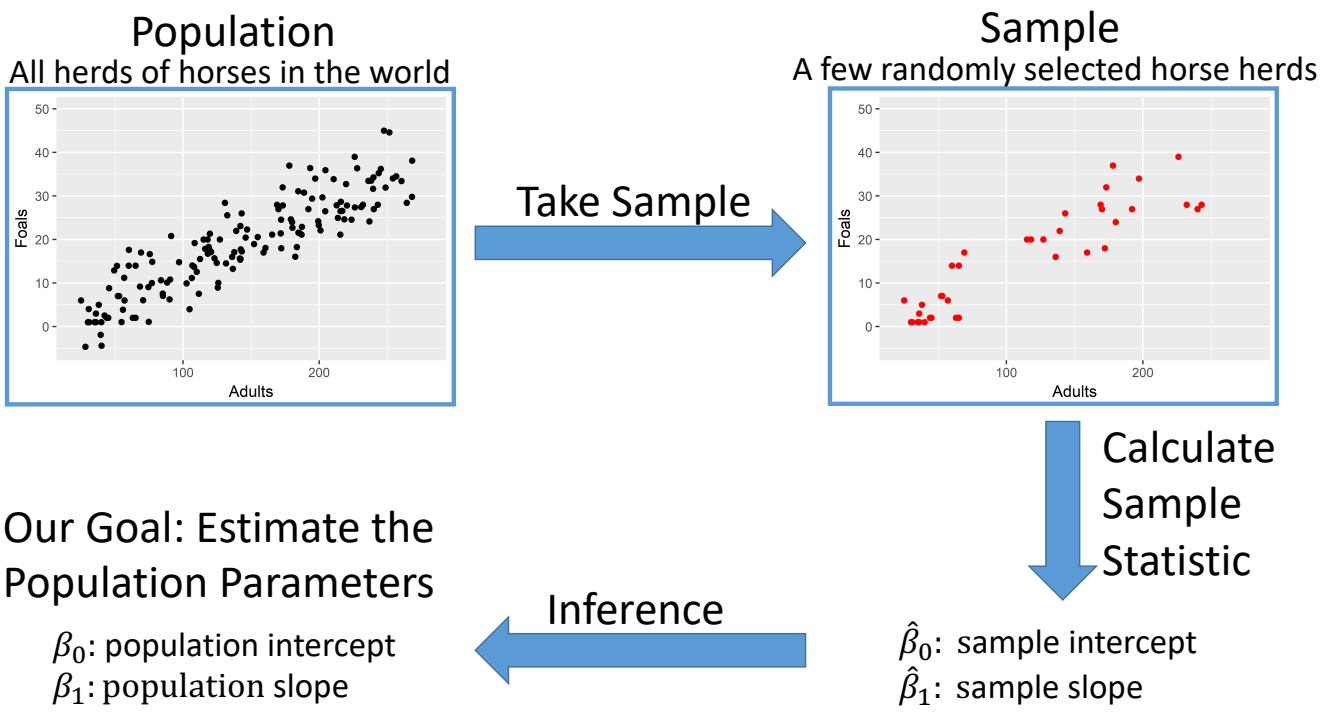
- Estimated Line: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- b_0 and b_1 are **sample statistics**: they describe the data in our sample



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





Sampling Distribution of $\hat{\beta}_1$ (similar for $\hat{\beta}_0$)

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

$$\hat{\beta}_1 \sim \text{Normal}\left(\beta_1, SD(\hat{\beta}_1)\right), \text{ where } SD(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- This is alsoapproximately true if most of the conditions are mostly satisfied and the sample size is "large".
- Recall: Probabilities involving the normal distributions only depend on how many standard deviations away from the mean we are:

$$\frac{\hat{\beta}_1 - \beta_1}{SD(\hat{\beta}_1)}$$

  - Example: For about 95% of samples, the estimate $\hat{\beta}_1$ is within $\pm 2SD(\hat{\beta}_1)$ of the true value of β_1 in the population.
- **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 $SD(\hat{\beta}_1)$

What can we do?

- Estimate $SD(\hat{\beta}_1)$. An estimate of a standard deviation is called a standard error.

$SE(\hat{\beta}_1) = \left[\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^{1/2}$

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

How to use this for hypothesis tests?

Null hypothesis: $\beta_1 = \beta_1^{null}$

Alternative hypothesis: $\beta_1 \neq \beta_1^{null}$

- **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:** $t = \frac{\hat{\beta}_1 - \beta_1^{null}}{SE(\hat{\beta}_1)}$
  - "How many estimated standard deviations away from the hypothesized slope was our sample slope?"
- If null hypothesis is true, $t \sim t_{n-2}$
- (Calculation of p-value hand-drawn on board)

How to use this for Conf. Intervals?

- For a 95% CI, find the value t^* with $P(-t^* \leq \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \leq t^*) = 0.95$
- This means that for 95% of samples, $-t^* \leq \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \leq t^*$
- ...so for 95% of samples, $-t^* SE(\hat{\beta}_1) \leq \hat{\beta}_1 - \beta_1 \leq t^* SE(\hat{\beta}_1)$
- ...so for 95% of samples, $-b_1 - t^* SE(\hat{\beta}_1) \leq -\beta_1 \leq -\hat{\beta}_1 + t^* SE(b_1)$
- ...so for 95% of samples, $b_1 - t^* SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t^* SE(b_1)$
- Confidence interval: $[\hat{\beta}_1 - t^* SE(\hat{\beta}_1), \hat{\beta}_1 + t^* SE(\hat{\beta}_1)]$
- In R, t^* = `qt(0.975, df = n - 2)` for a 95% CI.

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

- **Representative** Sample

We don't know how our sample was selected, so this is difficult to assess. To proceed, we need to assume that there are no forms of bias in the sampling design that could be related to the birth rates of foals in our herds, and/or restrict the conclusions of our analysis to the "population" of horses that our sample is representative of.

- (No) **Outliers**

The scatter plot of the sample data (top of page 2) shows no outliers.

- **Linear** Relationahip (Straight Enough)

The scatter plot of the sample data (top of page 2) shows an approximately linear association between the number of adults in a herd of horses and the number of foals born to that herd.

- **Independent** Observations (Randomization)

We don't know how our sample was selected, so this is difficult to assess. To proceed, we need to assume that there is no connection between the *residuals* for different herds of horses. For example, this condition might be violated if some of the herds in our sample were from the same region, and could have been affected by the same natural events that year.

- **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?)

The scatter plot of the sample data (top of page 2) shows approximately equal variability of the points around the trend across all values of the explanatory variable.

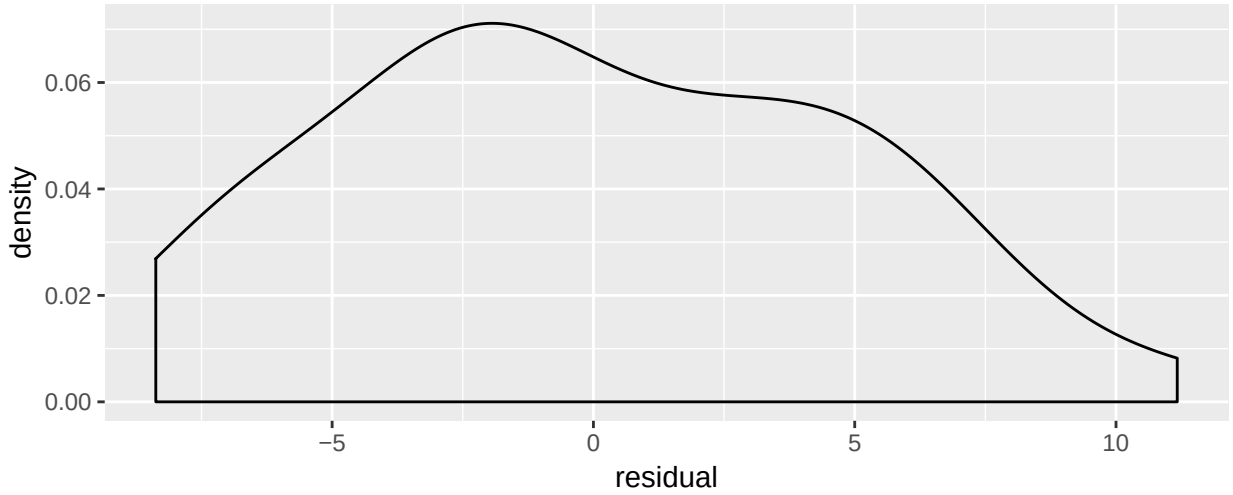
(b) Fit the linear model

```
# format: 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 (and/or the sample size is large enough to apply the Central Limit Theorem)

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



The density plot of the residuals shows that their distribution is unimodal, skewed slightly to the right, and there are no outliers. This distribution is probably close enough to normal to give reasonable results?

(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.

The estimated intercept is $\hat{\beta}_0 = -1.5784$. The model predicts that a herd of horses with 0 adults in it would have -1.5784 babies.

The estimated slope is $\hat{\beta}_1 = 0.154$. The model predicts that for each additional adult in a herd, an additional 0.154 foals will be born to that herd.

(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) How would you calculate the p-value for part (e) using the pt function in R and the given estimate and standard error? 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.

```
(-1.578 - 0)/1.4916
```

```
## [1] -1.058
```

```
2 * pt(-1.06, df = 38 - 2)
```

```
## [1] 0.2962
```

(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 95% confidence interval for the population intercept, β_0, and for the population slope, β_1. Interpret the confidence interval for β_1 in context.

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 = 0.95)
```

```
##                2.5 % 97.5 %
## (Intercept) -4.6035 1.4468
## Adults       0.1308 0.1771
```

(i) How would you calculate the confidence interval for β_1 using the `qt` function in R and the given estimate and standard error?

```
qt(0.975, df = 38 - 2)
```

```
## [1] 2.028
```

```
0.1540 - 2.028 * 0.011
```

```
## [1] 0.1317
```

```
0.1540 + 2.028 * 0.011
```

```
## [1] 0.1763
```

(j) Interpret the standard error for the slope using the "2 standard deviations" rule.