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

Evan Ray

November 29, 2017

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?

```
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()
## )
```

```
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
```

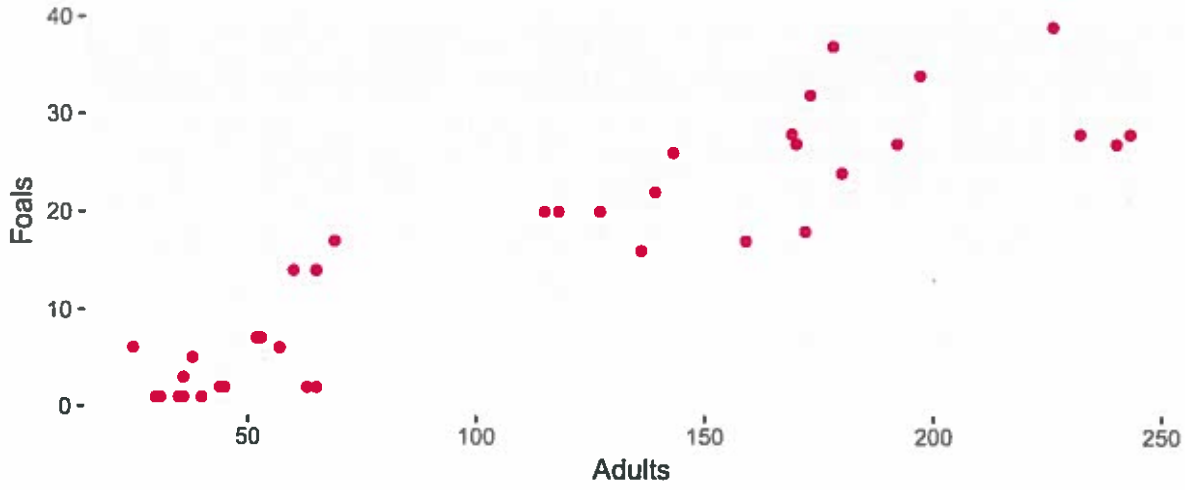
```
nrow(horses)
```

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

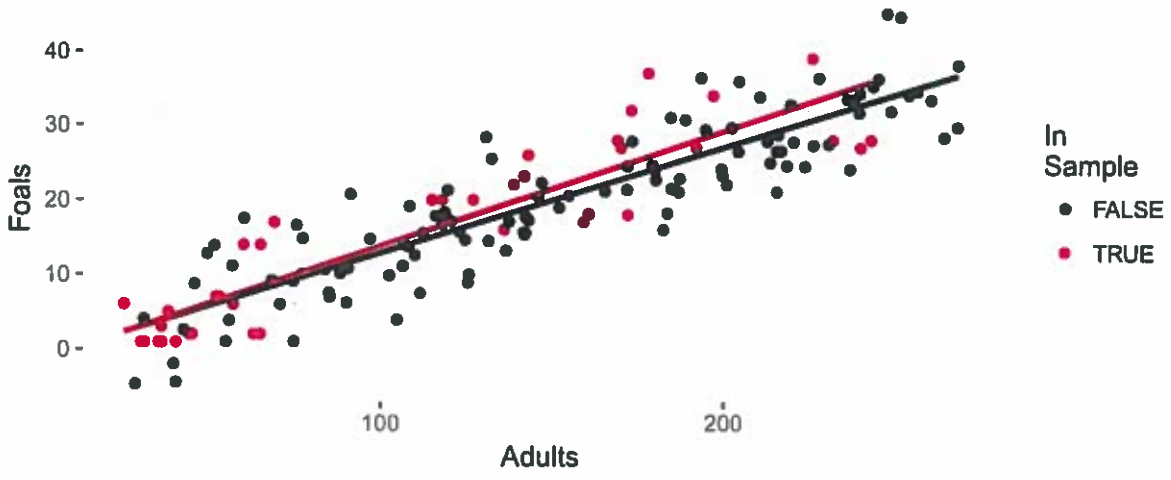
Questions to Start With:

- What is the observational unit? A herd of horses
- What are the variable data types (**categorical** or **quantitative**)?
  - **Foals:**
  - **Adults:** both quantitative
- Which of these variables is the **explanatory** variable and which is the **response**?
  - **Explanatory:** Adults
  - **Response:** Foals

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



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



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

We'll add a new condition to our list for linear regression:

- **Independence**
  - Randomization/no connection between different observational units

To remember this, think of a helpful leprechaun named Patrick O'LINE:

- (No) **Outliers**
- **Linear** Relationship
- **Independent** Observations
- **Normal** Distribution of Residuals
- **Equal** Variance of Residuals

- (No) **Outliers**

  Ok - there are no outliers in the scatter plot at the top of page 2.

- **Linear** Relationship (Straight Enough)

  The relationship between the # of Adults & the # of foals is approximately linear.

- **Independent** Observations (Randomization)

  We can't really assess this with the information we are given. We would need to assume that there is no connection between the herds in our sample. For example the herds should not contain horses that are directly related to each other if that might affect their birth rates.

- **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)

  Can't check until after we fit the model

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

  The amount of spread in the points around the linear trend is consistent across the range of values for the number of adults in the herd.

All assumptions are generally met, although we can't be sure about the assumption of independence.

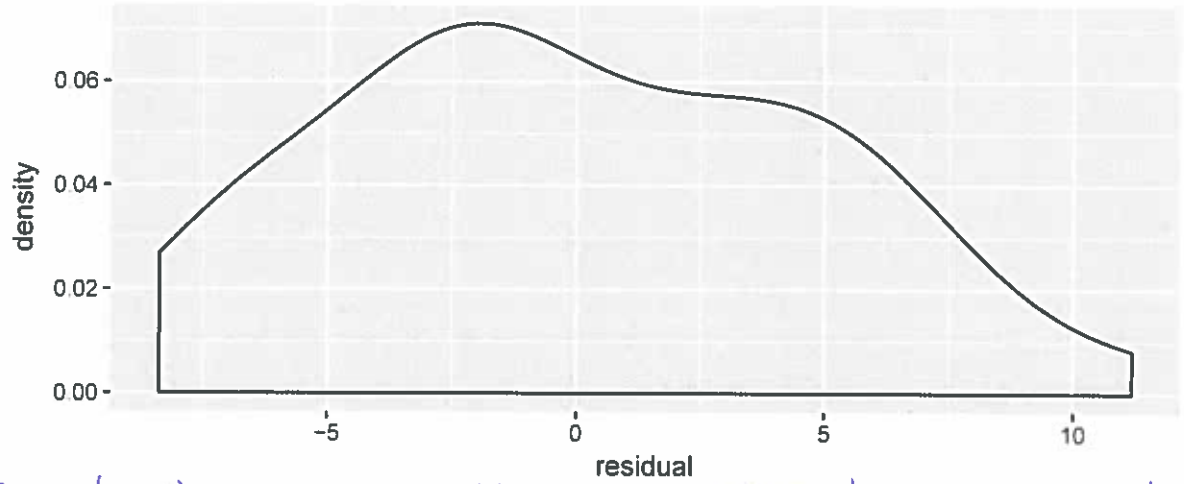
(b) Fit the linear model

```
# 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)
```



This distribution is unimodal, but it is skewed a little to the right. However, it is not too badly skewed, and I think that using a linear model should be OK even if the assumptions are not perfectly satisfied - as long as they are reasonably close.

(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.

Intercept: If the number of adults in a herd were 0, the linear model predicts that the herd would have −1.58 foals.

Slope: For each additional adult in the herd, the model's predicted number of foals increases by 0.154.

In general, the model predicts that larger herds have more foals.

(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.

$H_0: \beta_0 = 0$
$H_A: \beta_0 \neq 0$

From the R output, the p-value for this test is 0.3. Since $0.3 > \alpha = 0.05$ we fail to reject the null hypothesis. The data do not offer enough evidence to conclude that when there are 0 adults in a herd, the number of foals born to that herd will be anything other than 0.

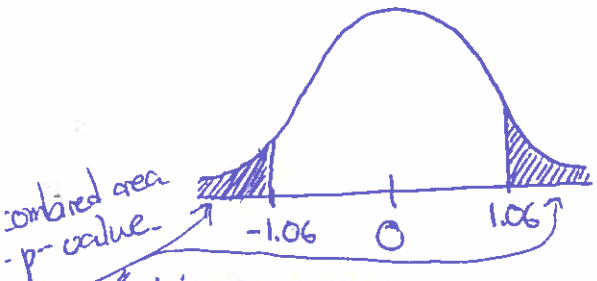
(f) Draw a picture of a relevant t distribution for the hypothesis test in part (e) and shade in the region corresponding to the p-value. How would you calculate the p-value for part (e) using the pt function in R and the given estimate and standard error?

We know that $\frac{b_0 - \beta_0}{SE(b_0)} \sim t_{n-2}$.

Assuming the null hypothesis is true, $\beta_0 = 0$, so $\frac{b_0}{SE(b_0)} \sim t_{n-2}$.

In this case, $\frac{b_0}{SE(b_0)} = \frac{-1.5784}{1.4916} = -1.06$ ← This is our test statistic.

make sure you know where to find all of these numbers in the R output on p. 4!



The p-value is the probability of getting a test statistic "at least as extreme" as the one we actually observed assuming H_0 is true. Since the H_A has the form $\beta_0 \neq 0$, "at least as extreme" means at least as far from 0 in either direction. So the p-value is the shaded area above. We can calculate the area on the left (less than −1.06) using the command $pt(-1.06, df = 36)$, which gives the output 0.148.

t_{n-2}

The p-value is twice this, so p-value $= 2 * 0.148 = 0.296$

(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. $H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$.

From the R output on page 4, the p-value for this test is $1.2e\text{-}15 = 1.2 * 10^{-15} = 0.000000000000012$.

This p-value is less than typical significance levels such as 0.05, 0.01, or 0.001. The data offer enough evidence to reject the null hypothesis and conclude that the slope of a line describing the relationship between the number of adults & # of foals in the population of all herds is not equal to 0, at a significance level of $\alpha = 0.001$.

(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
```

We are 99% confident that the slope (β_1) of a line describing the relationship between the number of adults in a herd and the number of foals born to that herd, in the population of all horse herds, is between 0.1229 and 0.185.

If we took many different samples and calculated a different 99% confidence interval for β_1 based on the data in each sample, about 99% of those confidence intervals would contain the true population slope.

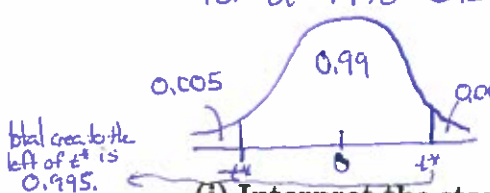
(i) How would you calculate the confidence interval for part (f) using the qt function in R and the given estimate and standard error?

The Confidence interval has the form $b_1 \pm t^* \cdot SE(b_1)$

From the R output on page 4, $b_1 = 0.154$ and $SE(b_1) = 0.0114$.

For a 99% C.I., t^* will be the 99.5th percentile of a t_{n-2} distribution: We can find this in R with: qt(0.995, df=36), which gives the output 2.72.

$0.154 - 2.72 * 0.0114 = 0.123$ and $0.154 + 2.72 * 0.0114 = 0.185$, which agree with the results in part (h) above.



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

If we knew the true standard deviation of b_1, $SD(b_1)$, we would say that for 95% of samples, the sample slope b_1 will be within $\pm 2 \cdot SD(b_1)$ of the population slope β_1. We don't know $SD(b_1)$, but $SE(b_1)$ is our best guess of it. See next page for a picture.

If assumptions (O'LINE) are satisfied,

$b_1 \sim \text{Normal}(\beta_1, SD(b_1))$

This is a sampling distribution for b_1: it describes the distribution of values of b_1 that you would get across all possible samples of a certain size.

95% of samples will have a regression line slope, b_1, that is between $\beta_1 - 2SD(b_1)$ and $\beta_1 + 2SD(b_1)$.

Unfortunately, we don't actually know $SD(b_1)$ - but we can estimate it by $SE(b_1)$.