Simple Linear Regression Summary

Population Model

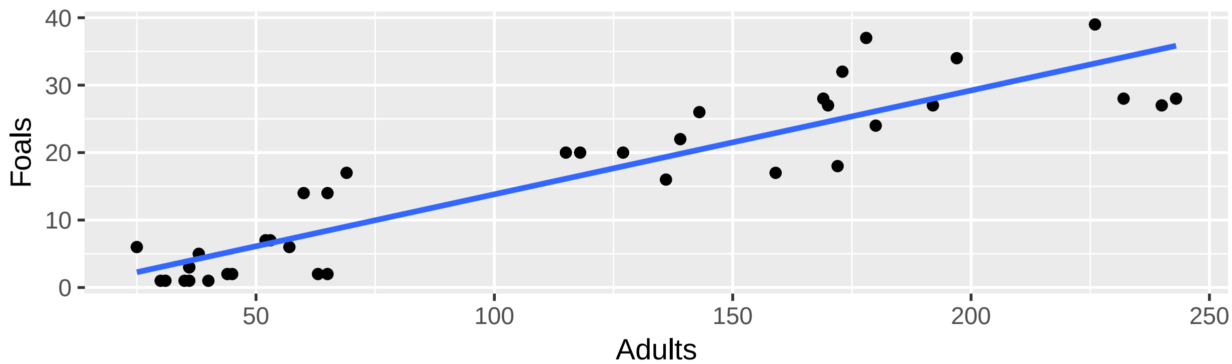
- The relationship between explanatory and response variable in the population is described by a line with intercept β_0 and slope β_1, with normally distributed "errors" around the line

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\varepsilon_i \sim \text{Normal}(0, \sigma)$$

- Y_i is response variable value for observational unit number i (number of foals in i'th herd)
- X_i is explanatory variable value for observational unit number i (number of adults in i'th herd)
- β_0 and β_1 are **population parameters** we want to estimate

Plot line based on sample data

```r
ggplot(data = horses, mapping = aes(x = Adults, y = Foals)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Fit linear regression model, print summary (`Foals` is response, `Adults` is explanatory)

```r
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
```

b_0, estimated intercept

b_1, estimated slope

Standard Error for b_1; this is an estimate of the variability in values of b_1 we will obtain from different samples

t statistic for a test of whether $\beta_1 = 0$

p value for a test of whether $\beta_1 = 0$

Residual standard deviation

Degrees of freedom: n - 2

R^2

Conditions for inference

Representative sample; No **O**utliers; **L**inear relationship; **I**ndependent observations; **N**ormally distributed residuals; **E**qual variance of residuals

Equation of estimated line based on sample

Predicted Foals = -1.578 + 0.154 × Adults

Interpretation of Estimated Intercept

The model predicts that if a herd contains 0 Adults, there will be -1.578 Foals born.

Interpretation of Estimated Slope

The model predicts that for each additional Adult in a herd of horses, an additional 0.154 Foals will be born.

Prediction for a Herd with 50 Adults

Predicted Foals = -1.578 + 0.154 × 50 = 6.122.

The model predicts that a herd with 50 adults will have 6.122 foals.

Find and interpret a 95% confidence interval for β_1 (procedure similar for β_0)

Confidence Interval for β_1: $b_1 \pm t^* SE(b_1)$, where:

- b_1 is estimate of slope based on this sample (from the R summary output)
- t^* is the critical value from a t distribution with $n-2$ degrees of freedom (from `qt`)
- $SE(b_1)$ is the standard error of b_1 (from the R summary output)

```
confint(lm_fit, level = 0.95)
```

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

```
0.154 - qt(0.975, df = 36) * 0.0114
```

```
## [1] 0.1309
```

```
0.154 + qt(0.975, df = 36) * 0.0114
```

```
## [1] 0.1771
```

We are 95% confident that the slope of a line describing the relationship between the number of adults in a herd of horses and the number of foals born to that herd, in the population of all herds of horses, is between 0.13 and 0.18.

Conduct a hypothesis test with null hypothesis $\beta_1 = 0$

Test statistic: $t = \frac{b_1 - \beta_1^{null}}{SE(b_1)} \sim t_{n-2}$

```
(0.154 - 0)/0.0114
```

```
## [1] 13.51
```

```
2 * pt(-13.5, df = 36)
```

```
## [1] 1.175e-15
```

Note that the third column of the "Coefficients:" table on the previous page also has the test statistic for this test, and the fourth column has the p-value. The notation $1.175e{-}15$ means $1.175 * 10^{-15} = 0.000000000000001175$. Since the p-value is very small, we reject the null hypothesis. The data provide strong evidence that there is an association between the number of adults in a herd and the number of foals born to that herd.

Use the residual standard deviation to describe how good the model's predictions are.

About 95% of predictions from this model are within plus or minus 9.88 foals of the actual number of foals produced by a herd. (9.88 is two times the residual standard deviation from the R summary output.)

Use the R^2 value to describe how useful the model is (not that important, included for completeness)

The R^2 value for this regression is 0.835. This is close to 1, indicating that the points fall fairly close to the line. (Recall that R^2 is the square of the correlation between the explanatory and response variables.) This linear model accounts for about 83.5% of the variation in the response variable.