Linear Regression Inference

Statistics on a linear model

Download the section 123.Rmd handout to `STAT240/lecture/sect13-regression-inference`.

Download the files `lake-monona-winters-2024.csv`, `riley.txt` and `lions.csv` to `STAT240/data`.

We've seen how to estimate a linear model, but we have not done any statistics.

$$\hat{y}_i \;=\; \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Is there an actual linear relationship between x and y? There is when the slope β_1 is nonzero.

We will extend our point estimate for slope into an interval estimate for $\hat{\beta}_1$.

$$\hat{\beta}_1 \pm \frac{\alpha}{2} \text{ Critical value } \times \text{ Standard error of } \hat{\beta}_1$$

This is a $1 - \alpha$ confidence interval for the slope.

What is the estimation error of $\hat{\beta}_1$?

$$\hat{se}(\hat{\beta}_1) = \frac{S}{\sqrt{(n-1)s_X^2}}$$

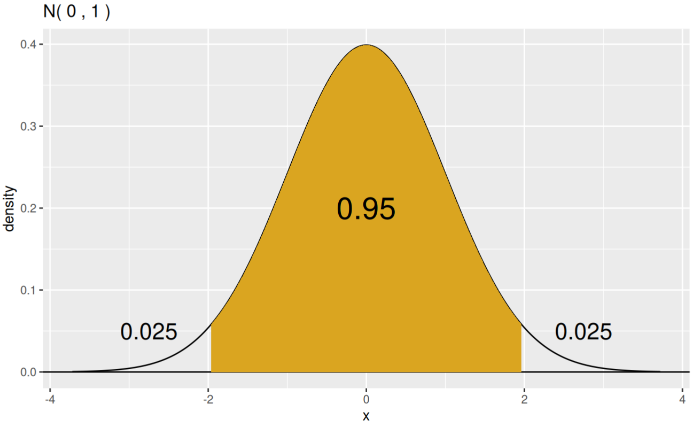
- Numerator: estimate for σ
- Denominator: variability of X

How do we guarantee $1 - \alpha$ coverage? Use a quantile on the sampling distribution.

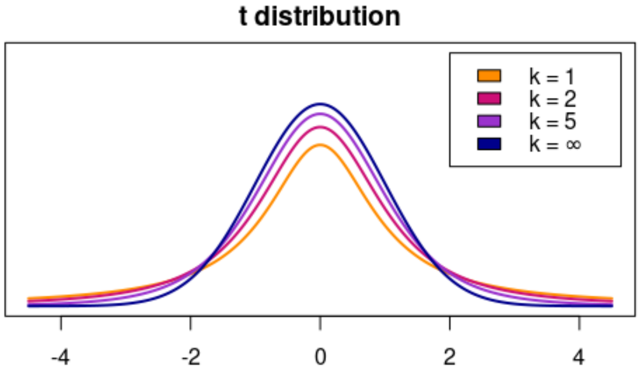
The sampling distribution for $\hat{\beta}_1$ is related to the **Student's T distribution**.

The T is similar to $N(0, 1)$.

For 95% confidence:



What does the T look like?



The T has heavier tails than $N(0,1)$, controlled by degrees of freedom.

In simple linear regression, df $= n - 2$.

Find critical values (quantiles) with `qt`.

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \times \text{ Standard error of } \hat{\beta}_1$$

Find these values with the `lm` summary.

A 95% CI for slope in the height model is:

$$(0.244, 0.256)$$

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \times \text{Standard error of } \hat{\beta}_1$$

Build and interpret a 98% CI for the slope of the Lake Monona linear model.

- Use `qt` to find the critical value

Are we confident that year and duration are related?

Formally, the hypothesis testing procedure is as follows:

- Write **hypotheses** about parameter
- Calculate **test statistic**
- Identify **null distribution**
- Calculate **p-value** on the null

The test statistic is the evidence against the null hypothesis in our data. Usually looks like this:

$$\frac{\text{Esimated value} \; - \; \text{Value under null}}{\text{Estimation error}}$$

If the null is true, the test statistic is close to 0.

The p-value is the probability of seeing our data or something more extreme, under the null.

The calculation depends on what we're trying to detect.

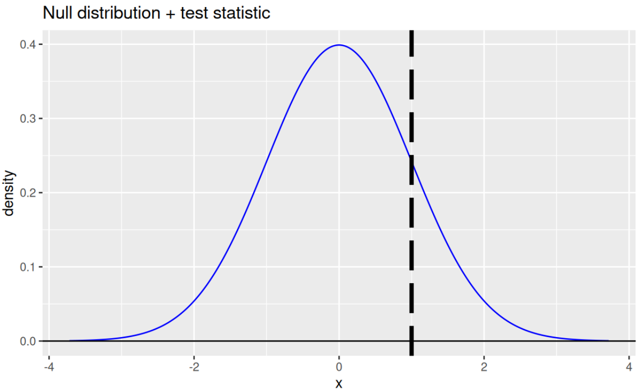
If we want to test whether x and y have a linear relationship, we need to test whether β_1 is zero or nonzero.

Do x and y have a linear relationship?

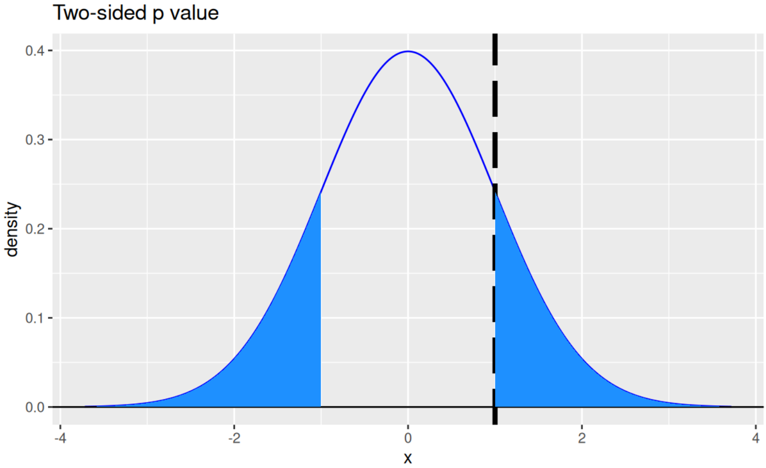
$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0$$

This is a **two-sided** test.

- p-value: outcomes more extreme than test statistic on both sides



Suppose we have a positive test statistic.

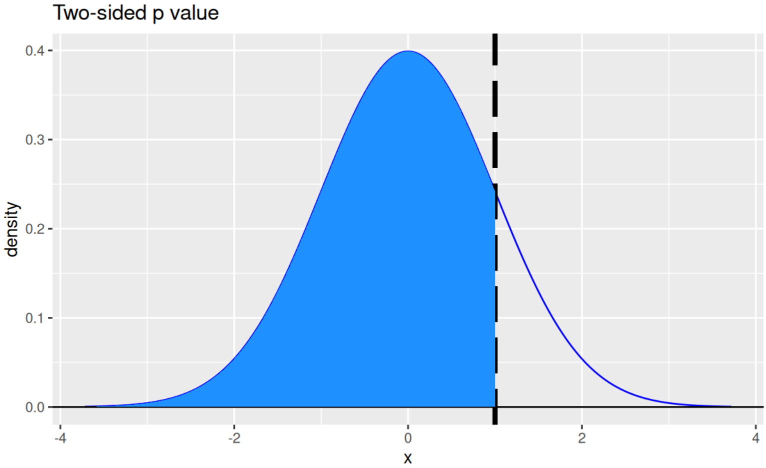

Two-sided p value
density
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
x

Do x and y have a negative relationship?

$$H_0 : \beta_1 \geq 0 \quad \text{versus} \quad H_A : \beta_1 < 0$$

This is a **one-sided** (negative) test.

- p-value: outcomes less than test statistic

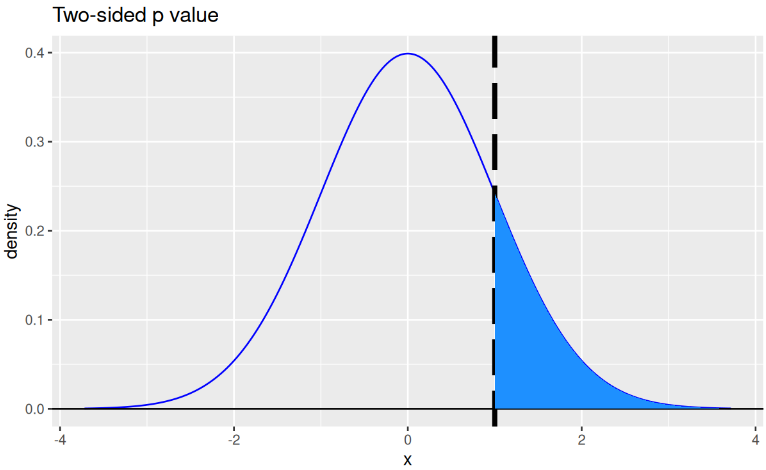

Two-sided p value
0.4
0.3
0.2
0.1
0.0
density
-4
-2
0
2
4
x

Do x and y have a positive relationship?

$$H_0 : \beta_1 \leq 0 \quad \text{versus} \quad H_A : \beta_1 > 0$$

This is a **one-sided** (positive) test.

- p-value: outcomes greater than test statistic


Two-sided p value
density
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
x

Let's test whether the Monona slope is negative, with $\alpha = 0.05$.

For a slope test, our test statistic is

$$T = \frac{\hat{\beta}_1 - \beta_{null}}{\hat{se}(\hat{\beta}_1)}$$

If H_0 is true and $\beta_1 \geq 0$, then T follows a T distribution with $n - 2$ degrees of freedom.

We have

$$t_{obs} = \frac{-0.223 - 0}{0.02667} = -8.36$$

Estimated slope below 0 $\Rightarrow$ negative t_{obs}.

Is this value consistent with a t_{n-2} distribution?

What if we were doing a two-sided test instead?

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0$$

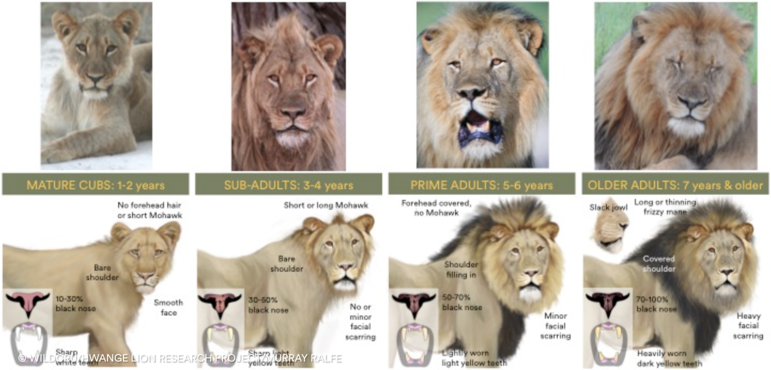
We would have the same test statistic, but a different p-value.

This is also given in the `lm` output.

We've seen how to predict a value with a linear model. Let's turn that into an interval.

Predicted value $\pm$ Critical value $\times$ Prediction error

In the lion ages data, we want to relate a lion's age to the % of its nose that is black.



Six traits can be used to accurately estimate a lion's age: nose darkness, mane growth (in males), facial scarring, teeth color and wear, and jowl slackness. Due to variance between individuals, age should be estimated based on multiple characteristics.

We predict a 5-year-old lion to have a 36% black nose. Formally,

$$\left(\hat{y}|x^* = 5\right) \;=\; 0.36$$

$$(\text{Fitted value given } x^*) \;=\; \hat{\beta}_0 + \hat{\beta}_1 x^*$$

The uncertainty in this point estimate depends on what exactly we are predicting.

- Predicting the position of the line itself. This is the *average* nose % for all 5-year-old lions.
- Predicting the nose % for a *single* 5 year old lion.

The first type of prediction, the position of the line, is $E(\hat{y} \mid x^*)$.

Let's investigate this with simulation.

- Generate n random points from $\beta_0 + \beta_1 x + \epsilon$
- Calculate $\hat{\beta}_1$ and $\hat{\beta}_0$ and plot the line

The estimated standard error of the position of the line is

$$\hat{se}(E(\hat{y} \mid x^*)) = S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

The critical value for our CI is the same as before. It is a $\alpha/2$ critical value from the T with $n-2$ degrees of freedom.

$$\hat{y}|x^* \pm t_{\alpha/2,n-2} \times S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

This is what `geom_smooth` is doing!

Use `predict` to calculate the CI for us.

Set `interval = "confidence"`. We can also plot this against the data.

The uncertainty in this point estimate depends on what exactly we are predicting.

- Predicting the position of the line itself. This is the *average* nose % for all 5-year-old lions.
- Predicting the nose % for a *single* 5 year old lion.

The second type of prediction is $\hat{y} \mid x^*$.

Again, the point estimate is just found by plugging x^* into the model.

This type of prediction has more error than predicting the position of the line.

The estimated standard error of a new prediction is

$$\hat{se}(\hat{y} \mid x^*) \ = \ S\sqrt{1 + \frac{1}{n} \ + \ \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

We have an extra $+1$ term for applying our model to a new data point.

This gives us a **prediction** interval.

$$\hat{y}|x^* \;\pm\; t_{\alpha/2,n-2} \times S\sqrt{1 + \frac{1}{n} \;+\; \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Again, use `predict`.

Confidence interval: position of the line
Prediction interval: new y value

- PIs are wider than CIs
- Both intervals are wider when we are further from $\bar{x}$

The **coefficient of determination** R^2 is

$$R^2 = \frac{\text{Total variability of y} - \text{Model error}}{\text{Total variability of y}}$$

R^2 is the fraction of the total variability in y explained by x (via the regression line).

We can find R^2 in the summary output of an `lm` object in R.

If we only have two variables (x and y), then R^2 is equal to the square of the correlation coefficient.

$$R^2 = r^2$$

(This does not necessarily hold for more complex models).

R^2 is a useful measure of how well x explains y, but it does not help us in evaluating assumptions.