

Linear Regression Inference

Statistics on a linear model

Download the section 16 .Rmd handout to
STAT240/lectures/16-regression-inference.

Material in this section is covered by Chapter 14 on
the notes website.

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

We have not done any statistics on our model yet.

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

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

Let's build an interval estimate for $\hat{\beta}_1$.

point estimate \pm critical value \times standard error

Here, the point estimate is $\hat{\beta}_1$.

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

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

- Numerator: spread of points around line
- Denominator: variability of X

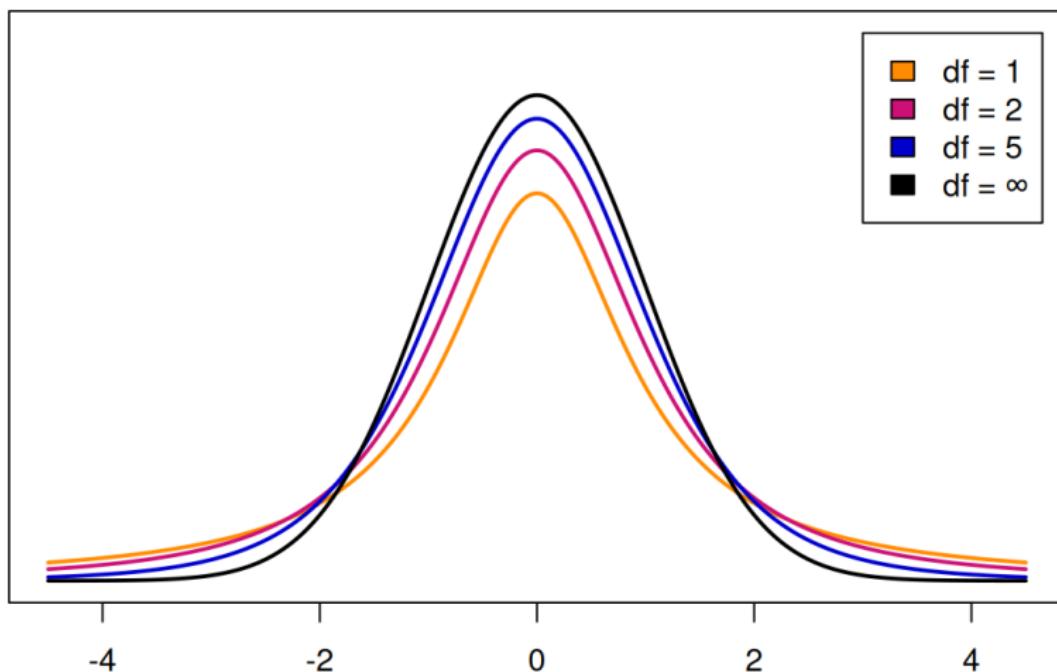
To guarantee $1 - \alpha$ coverage, we use a quantile from the T distribution.

Why? Need to estimate σ with S .

- “Residual standard error”

In simple linear regression, $\text{df} = n - 2$.

T distributions



$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \times \frac{s}{\sqrt{(n-1)s_X^2}}$$

Find these values with the `lm` summary.

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

$$(0.244, 0.256)$$

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \times \frac{S}{\sqrt{(n-1)s_X^2}}$$

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

- Try to calculate SE by hand

Are we confident that year and duration are related?

We can also perform a T test for the slope. The test statistic looks like:

$$\frac{\text{Estimated slope} - \text{Slope from } H_0}{\text{Estimation error}}$$

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

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

This is a one-sided test of

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

For a slope test, our test statistic is

$$T = \frac{\hat{\beta}_1 - \beta_{null}}{\frac{s}{\sqrt{(n-1)s_X^2}}}$$

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

We have

$$t_{obs} = \frac{-0.224 - 0}{0.02636} = -8.5$$

Is this value consistent with a t_{n-2} distribution?
No - very small p-value in “less than” direction

What if we just want to know if x and y are related?

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

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.

$\text{predicted value} \pm \text{critical value} \times \text{prediction error}$

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



MATURE CUBS: 1-2 years



SUB-ADULTS: 3-4 years



PRIME ADULTS: 5-6 years



OLDER ADULTS: 7 years & older



© WILDLIFE WANGE LION RESEARCH PROJECT, RAY RALFE

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,

$$(\hat{y} \mid x^* = 5) = 0.36$$

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

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 $E(\hat{y} | x^*)$.

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

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

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.

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

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 $\hat{y} | x^*$.

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

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

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

This gives us a **prediction** interval.

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

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}