Simple Linear Regression + Inference for Slope

Marine Ecology

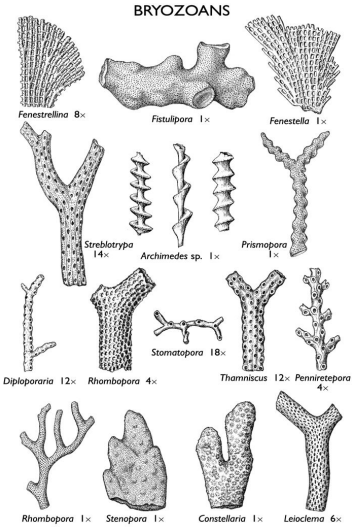
From an evolutionary perspective, an individual organism's success is judged by its ability to reproduce, and the most successful organisms are those which make the biggest contribution to their species' gene pool.

In many species, one factor associated with reproductive success is size, but there are competing explanations for this association.

In Pettersen, White, and Marshall (2015) the authors collected data on two species of marine bryozoan (simple, aquatic invertebrate animals).

The mass and metabolic rate of these individuals were recorded, and regression was used to model the relationship between mass and metabolic rate.

From "Regression, Transformations, and Mixed-Effects with Marine Bryozoans" by Cieran Evans



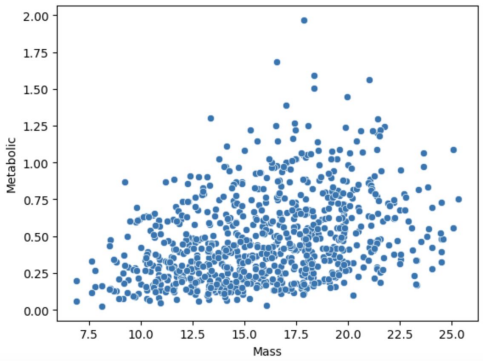
https://isgs.illinois.edu/outreach/geology-resources/bryozoans

CHOOSE a simple linear regression (SLR) model

Describe the pattern:

- Shape/form
- Direction
- Strength

Form of a model:

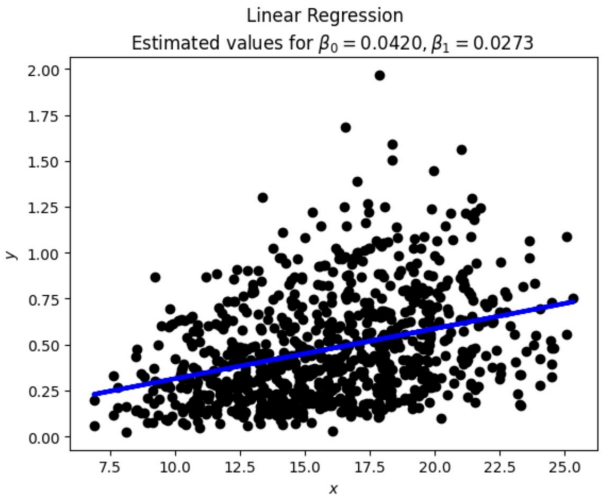


There are three parameters to estimate for a SLR model:

FIT a simple linear model

Least Squares Regression

We will use Python to calculate the estimated values.



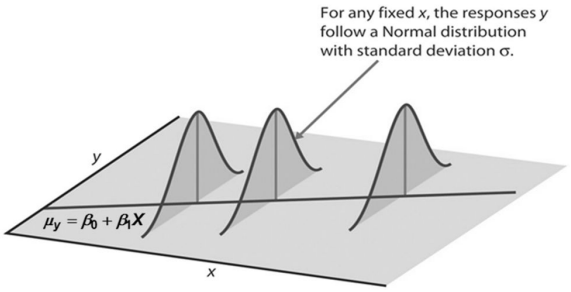
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0420	0.043	0.979	0.328	-0.042	0.126
Mass	0.0273	0.003	10.394	0.000	0.022	0.032

The Simple Linear Regression Model

For a quantitative response variable Y and a single quantitative explanatory variable X, the simple linear regression model is

where ϵ follows a Normal distribution, that is, ϵ~N(0, σ_ϵ), and the errors are independent from one another.

- μ_Y: population mean of the response variable
- We assume in a SLR model that μ_Y is _____ to the value of the explanatory variable.



FIT: Estimating the standard deviation of the error

Standard Error of Regression

For a simple linear regression model, the estimated standard deviation of the error term, based on the least squares regression line, is

$$\hat{\sigma}_\epsilon = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

Use Python for this:

	df	sum_sq	mean_sq	F	PR(>F)
Mass	1.0	8.410391	8.410391	108.043047	7.363579e-24
Residual	821.0	63.909077	0.077843	NaN	NaN

More about the regression standard error

Calculation:

Degrees of freedom =

The regression standard error depends on the residuals.

It represents the “typical variation” in an observation from the mean response.

ASSESS how well the model describes the data

Conditions required for simple linear regression:

Linearity:

Normality:

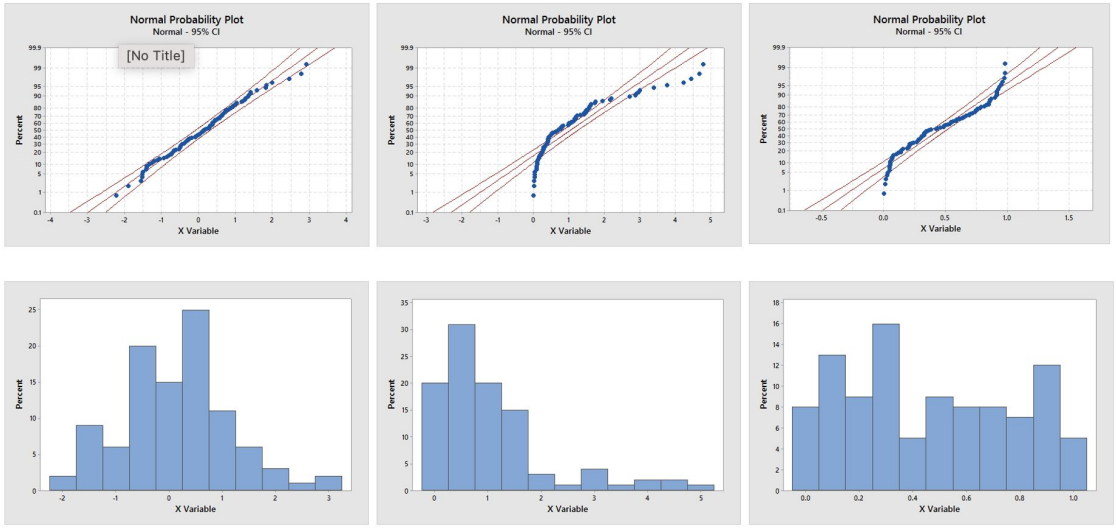
Zero Mean:

Constant Variance:

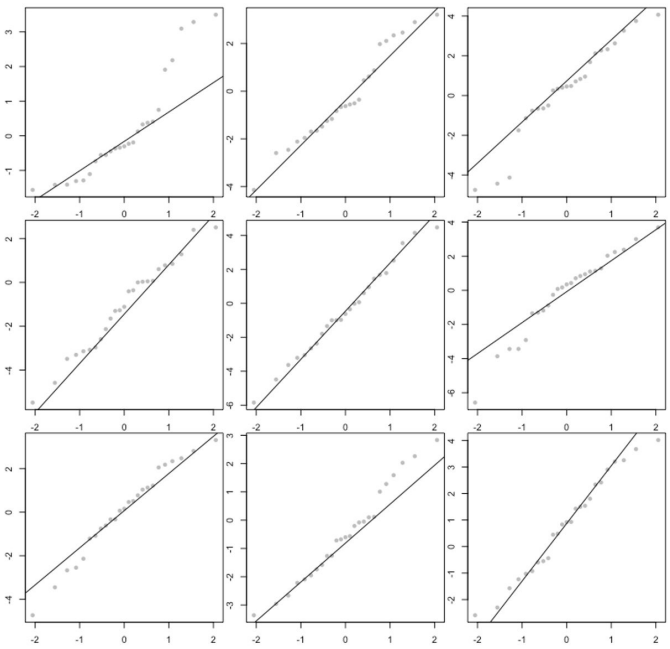
Independence:

Random:

ASSESS normality of residuals

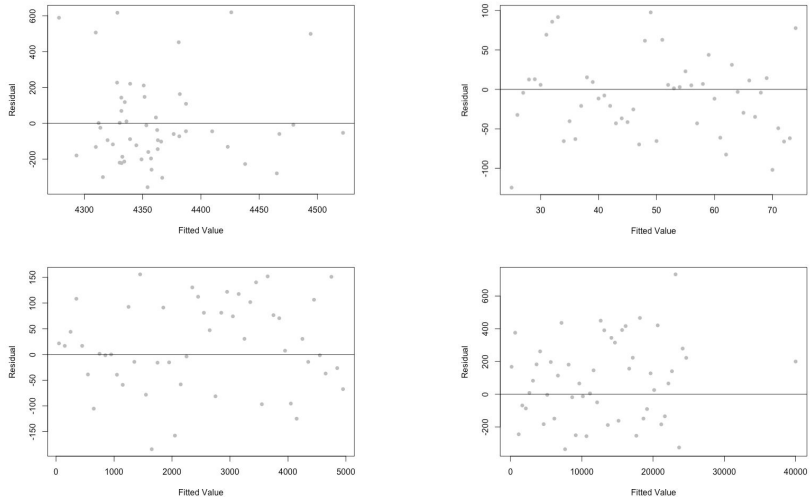


ASSESS normality

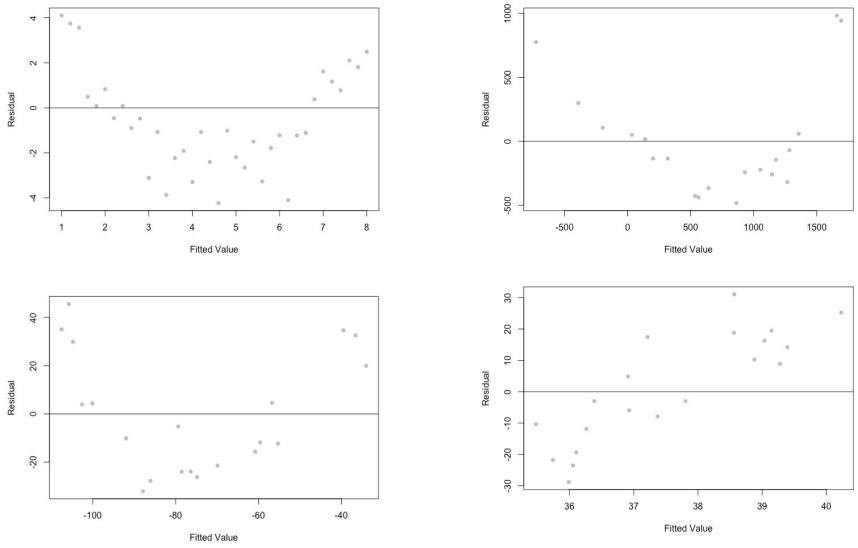


ASSESS linearity and constant variance

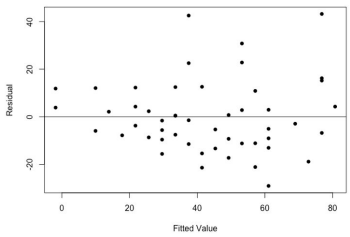
Reasonable Residual Plots (residuals on y-axis, fitted value on x-axis) with no serious problems.

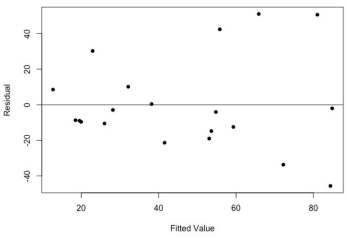


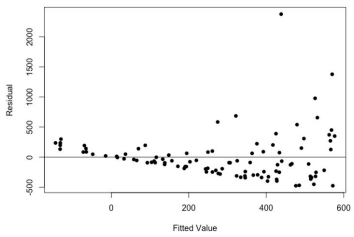
ASSESS: lack of linearity



ASSESS: non-constant variance







USE: Interpret Estimates

Recall our equation:

Slope:

Y-Intercept:

Residuals (R^2 and Regression Standard Error):

```
=========================================
  R-squared:                        0.116
  Adj. R-squared:                   0.115
  F-statistic:                      108.0
  Prob (F-statistic):           7.36e-24
  Log-Likelihood:                -116.20
  AIC:                              236.4
  BIC:                              245.8
-----------------------------------------
```

Descriptive → Inference
Is the slope significantly different from zero?

Types of Inference

Significance Tests:

Confidence Intervals:

Sampling Distribution of Slope Coefficient

The **sampling distribution** of a statistic is the distribution of…

For least squares estimates from a SLR model,

Test statistic:

Sampling Distribution of Slope Coefficient

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_a: \beta_1 \neq 0$$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0420	0.043	0.979	0.328	-0.042	0.126
Mass	0.0273	0.003	10.394	0.000	0.022	0.032

Strength of Evidence

p-value:

"Large" values: a high probability suggests that our result is

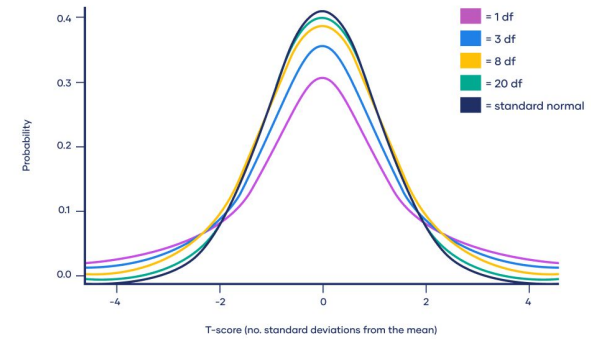
"Small" values: a small value indicates that if the true slope really was zero,

Thus, the assumption that the slope is zero

Sampling Distribution of Slope Coefficient

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0420	0.043	0.979	0.328	-0.042	0.126
Mass	0.0273	0.003	10.394	0.000	0.022	0.032

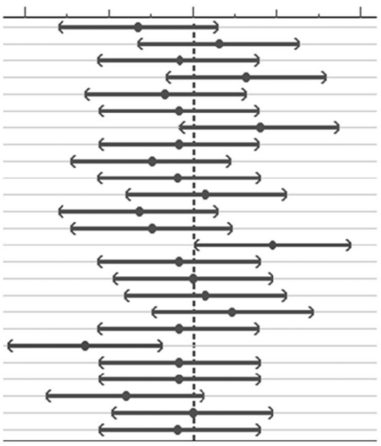
No. Observations:	823
Df Residuals:	821
Df Model:	1



Scribbr https://istats.shinyapps.io/tdist/

Confidence Intervals

- Finds a “plausible” range of values for the slope
- Useful when we want to estimate the _____ of the effect.
- “We are 95% confident that the true slope is in this interval.”
-



Calculate a 95% Confidence Interval

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0420	0.043	0.979	0.328	-0.042	0.126
Mass	0.0273	0.003	10.394	0.000	0.022	0.032

- General format of a (1-alpha)*100% confidence interval:
- Formula for a (1-alpha)*100% confidence interval for the slope:
- Calculation:

Calculate and interpret a 95% Confidence Interval

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0420	0.043	0.979	0.328	-0.042	0.126
Mass	0.0273	0.003	10.394	0.000	0.022	0.032

```
from scipy.stats import t
t.ppf(1-0.05/2, 821) ## get critical value
```

```
1.9628576665180677
```