Lab 14: Simple linear regression

STAT218

The objective of this lab is to learn to fit simple linear regression models in R and use/interpret relevant output to construct intervals and perform significance tests on the slope parameter.

Specifically, the lab covers the following:

* use of lm to fit models
* interpreting the output of summary.lm
* constructing confidence intervals using confint

We’ll use three now-familiar datsets:

* prevend: cognitive assessment scores and age from the PREVEND study
* kleiber: observations of metabolic rate and body mass among animals
* mammals: observations of brain size and body size among common species of mammal

library(tidyverse)
load('data/prevend.RData')
load('data/kleiber.RData')
load('data/mammals.RData')

### Warm-up

Consider constructing a scatterplot using formula-dataframe syntax (rather than vector inputs):

# scatterplot of rfft against age
plot(rfft ~ age, data = prevend)



The formula should specify the response variable on the left-hand side and the explanatory variable on the right-hand side, *e.g.*:

<RESPONSE> ~ <EXPLANATORY>

The variable names that appear in the formula must match exactly variables contained in the dataframe supplied as the data = ... argument.

|  |
| --- |
|  Your turn 1 |
| Construct a scatterplot of log brain size (response) against log body size (explanatory).# scatterplot of log brain size against log body size |

### Fitting SLR models

The general form of a simple linear regression model for a response $y$ in terms of an explanatory varaible $x$ is:

$$y=β\_{0}+β\_{1}x+ϵ$$

Above, the coefficients $β\_{0},β\_{1}$ are model parameters. Here you’ll see how to obtain the ‘fitted model’ in which estimates are provided for the model parameters:

$$y=\hat{β}\_{0}+\hat{β}\_{1}x$$

Note that the fitted model is written without the error term $ϵ$.

#### Computing estimates

SLR models are fitted via the function lm in R. The syntax for lm(...) – short for **l**inear **m**odel – is identical to constructing scatterplots using the formula-dataframe syntax:

# slr model of rfft (response) by age (explanatory)
fit.rfft <- lm(rfft ~ age, data = prevend)

The default output shows only the coefficient estimates. Here, the fitted model equation is:

$$RFFT=134.098052187513+1.19079380169042×age$$

The slope parameter on age captures the relationship of interest:

Each year of age is associated with an estimated decrease in mean RFFT score of 1.191.

|  |
| --- |
|  Your turn 2 |
| Fit a simple linear regression model of log brain size with log body size as the explanatory variable. Write the fitted model equation and interpret the coefficient estimate for the slope parameter in context.# slr model of log brain size by log body size |

#### Visualizing a fitted model

The model is easy to visualize using abline(...), which adds a line to an existing plot. Conveniently, the function has a method for the output of an lm call:

# reproduce the scatterplot, then add a line using abline
plot(rfft ~ age, data = prevend)
abline(reg = fit.rfft, col = 'blue', lwd = 2)



The additional arguments (color col and line width lwd) make the line easier to distinguish from the data scatter.

|  |
| --- |
|  Your turn 3 |
| Add a line showing the fitted SLR model to your scatterplot from before of log brain size against log body size. |

### Fit summary

The full set of information about a fitted model – estimates and standard errors, variance explained, significance tests, and the error variability estimate – are obtained using the summary(...) function:

# inspect model summary for estimates, inference, and measures of fit
summary(fit.rfft)

Call:
lm(formula = rfft ~ age, data = prevend)

Residuals:
 Min 1Q Median 3Q Max
-56.085 -14.690 -2.937 12.744 77.975

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.0981 6.0701 22.09 <2e-16 \*\*\*
age -1.1908 0.1007 -11.82 <2e-16 \*\*\*
---
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.52 on 206 degrees of freedom
Multiple R-squared: 0.4043, Adjusted R-squared: 0.4014
F-statistic: 139.8 on 1 and 206 DF, p-value: < 2.2e-16

Take a moment to locate the following among the output above:

* age explains 40.43% of total variability in RFFT scores
* $\hat{β}\_{1}=−1.191$
* $SE\left(\hat{β}\_{1}\right)=0.1007$
* $\hat{σ}=20.518$

Inference for simple linear regression models usually focuses on the slope parameter $β\_{1}$. The $p$-values you see in the coefficient summary table show the results of *partial significance tests*, which are *t* tests of the hypotheses:

$$\left\{\begin{matrix}H\_{0}:&β\_{j}=0\\H\_{A}:&β\_{j}\ne 0\end{matrix}\right.$$

The partial significance test on the slope parameter is typically interpreted as a test of association, since if the slope is zero then, *e.g.*, age and mean RFFT are unrelated. So the small $p$-value for the slope coefficient indicates:

The data provide very strong evidence that age is associated with RFFT score (*T* = -11.82 on 206 degrees of freedom, *p* < 0.0001).

|  |
| --- |
|  Your turn 4 |
| Inspect the model summary for your simple linear regression model of log brain and log body size and interpret the inference on the slope parameter in context. |

### Confidence intervals

Confidence intervals can be obtained for coefficient estimates using the confint(...) function; this takes a fitted lm object as input together with a confidence level.

# 95% confidence intervals
confint(object = fit.rfft, level = 0.95)

 2.5 % 97.5 %
(Intercept) 122.130647 146.0654574
age -1.389341 -0.9922471

# 99% confidence intervals
confint(object = fit.rfft, level = 0.99)

 0.5 % 99.5 %
(Intercept) 118.31647 149.8796336
age -1.45262 -0.9289675

# for slope only...
confint(object = fit.rfft, parm = 'age', level = 0.99)

 0.5 % 99.5 %
age -1.45262 -0.9289675

The interpretation is as follows:

With 99% confidence, each additional year of age is associated with an estimated decrease in mean RFFT score of between 0.929 and 1.453 points.

|  |
| --- |
|  Your turn 5 |
| Compute a 95% confidence interval for the slope parameter from your SLR model of log brain size. Interpret the interval in context. |

### Practice problem

1. [L10] The kleiber dataset contains observations of log-transformed average mass (kg) and log-transformed metabolic rate (kJ/day). Kleiber’s law refers to the relationship by which metabolism depends on body mass.
	1. Fit an SLR model of log metabolism (response) against log mass (explanatory). Write the fitted model equation and interpret the coefficient estimate for the slope parameter.
	2. Construct a scatterplot with a line overlaid to visualize the fitted model.
	3. Produce the model summary and identify the proportion of variability explained by the model.
	4. Interpret the partial significance test for the slope parameter.
	5. Construct a 99% confidence interval for the slope parameter.