Stat 412/512

Extra Sum of Squares

Jan 26 2015

Charlotte Wickham

stat512.cwick.co.nz

Inferences

On a single parameter ✓

On a linear combination of parameters

✓ (if we can redefine the baseline to get it)

On the mean response

On multiple parameters at once

Inference on the mean

We can estimate the mean response for given values of the explanatories by plugging in the values into the estimated model,

$\hat{\mu}\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\} = \hat{\beta}_0 + \hat{\beta}_1 \log \text{Mass} + \hat{\beta}_2 \text{ebat} + \hat{\beta}_3 \text{bird}$

Estimate mean for a echolocating bat that has log(Mass) = 4

Again one approach is to redefine the model.

I.e. estimate the mean log energy for a echo-locating bat with log body mass of 4:

$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\} = \beta_0 + \beta_1 (\log \text{Mass} - 4) + \beta_2 \textit{non-ebat} + \beta_3 \textit{bird}$

$\mu\{ \log \text{Mass} = 4, \text{Type} = \text{echo-b}\} = \beta_0$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.76213	0.11100	15.875	3.26e-11
I(log(Mass) - 4)	0.81496	0.04454	18.297	3.76e-12
Type2non-echolocating bats	-0.07866	0.20268	-0.388	0.703
Type2non-echolocating birds	0.02360	0.15760	0.150	0.883

```
> fit_bat <- lm(log(Energy) ~ log(Mass) + Type, data = case1002)
> summary(fit_bat)

...
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.57636    0.28724  -5.488 4.96e-05 ***
log(Mass)                     0.81496    0.04454  18.297 3.76e-12 ***
Typenon-echolocating birds  0.10226    0.11418   0.896    0.384
Typeecholocating bats       0.07866    0.20268   0.388    0.703
```

Or use predict in R.

you have to set all the explanatory variables in the model

```
> newdata <- data.frame(Mass = exp(4), Type = "echolocating bats")
> predict(fit_bats, newdata = newdata, se.fit = TRUE, interval = "confidence")
$fit
       fit      lwr      upr
1 1.762133 1.526821 1.997446

$se.fit
[1] 0.1110015

$df
[1] 16

$residual.scale
[1] 0.1859633
```

should like the dataframe we gave to lm but without our response.

```
> summary(fit_bats)
…

Residual standard error: 0.186 on 16 degrees of freedom
Multiple R-squared: 0.9815,   Adjusted R-squared: 0.9781
F-statistic: 283.6 on 3 and 16 DF,  p-value: 4.464e-14
```

A confidence interval for the mean

$$\text{estimate} \pm \text{multiplier} \times SE_{\text{estimate}}$$

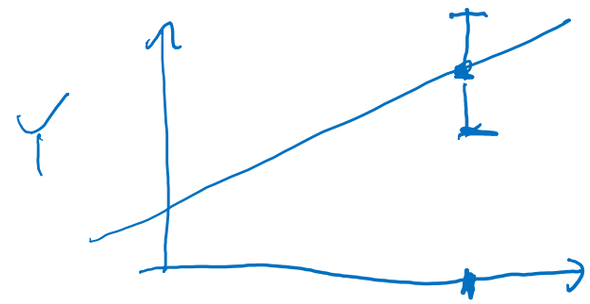
predict

→ 95% confidence interval for a single mean

$$\text{multiplier} = t_{n-p}(0.975)$$

I R

qt(0.975, 16) = 2.12

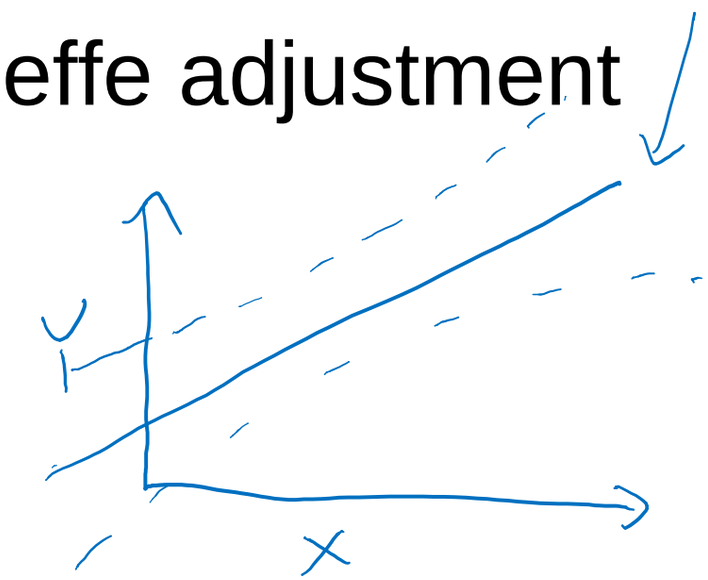


95% confidence interval the entire line

(within the range of the data)

multiplier = $\sqrt{pF_{p,n-p}(0.95)}$ Scheffe adjustment

sqrt((4)* qf(0.95, 4, 16)) = 3.47



p = number of β's

data.frame(mass

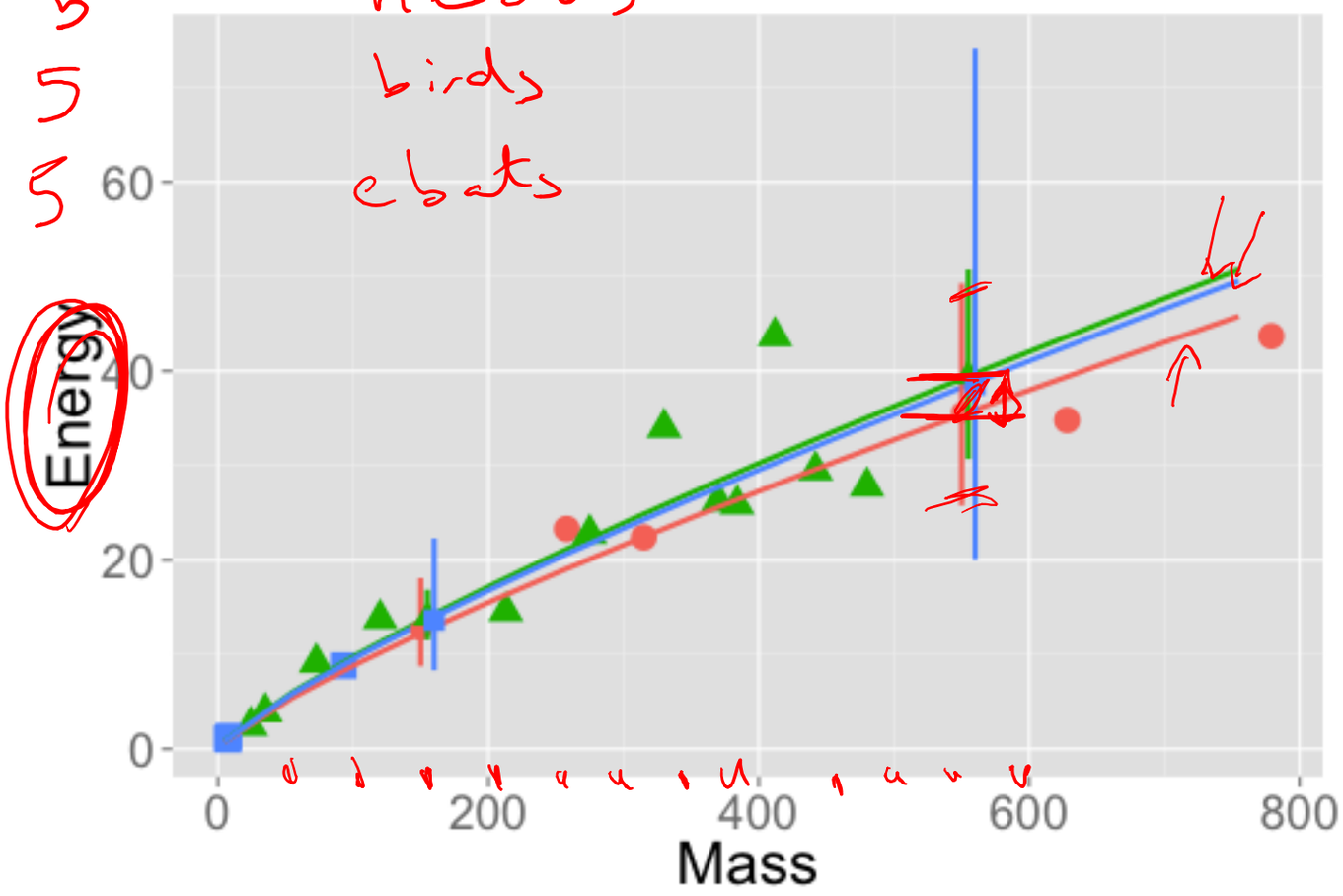
```
newdata2 <- expand.grid(Mass = seq(5, 800, 50), Type = c("non-echolocating bats",
  "non-echolocating birds", "echolocating bats"))
preds <- as.data.frame(predict(fit_bats, newdata = newdata2, se.fit = TRUE))
preds <- cbind(newdata2, preds)
```

← adding est. mean & se

seq(start, stop, increment)

newdata2

Mass	Type
5	nebats
5	birds
5	ebats



expand.grid makes a data.frame with rows that include all combination of the values in the columns specified

qplot(Mass, Energy, data = case1002, colour = Type)
+ geom_line(aes(y = fit))

Predictions

$$\text{Prediction variance} = \sigma^2 + SD\{\hat{\mu}\{Y|X\}\}^2$$

variance in estimated mean

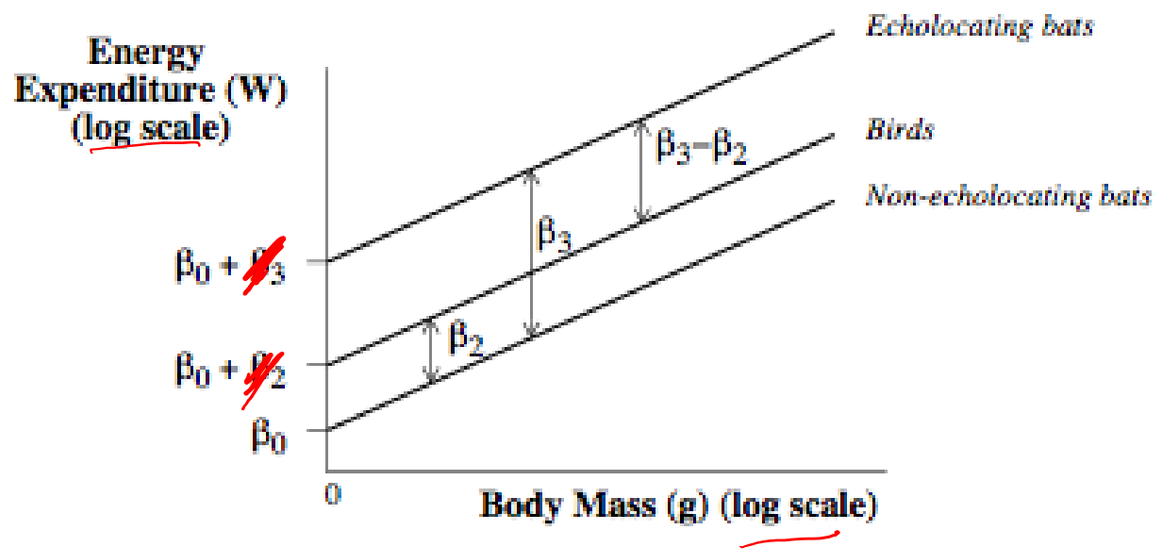
Just like in simple linear regression, the standard error on a prediction is always bigger than the standard error on the mean response at the same levels of the explanatories.

```
predict(fit_bats, newdata = newdata,
        interval = "prediction")
$fit
       fit     lwr      upr
1 1.762133 1.30302 2.221247
```

Extra Sum of Squares

Display 10.5 **p. 272**

The parallel regression lines model for the bat echolocation data



We might be interested in whether both β_2 and β_3 are zero.

If $\beta_2 = \beta_3 = 0$, then our model reduces to:

$$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\} = \beta_0 + \beta_1 \log \text{Mass}$$

equal lines

Two models

- Full model:

$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\}$

$= \beta_0 + \beta_1 \log \text{Mass} + \beta_2 \, bird + \beta_3 \, ebat$

- Reduced model:

$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\}$

$= \beta_0 + \beta_1 \log \text{Mass}$

If the reduced model is the truth:

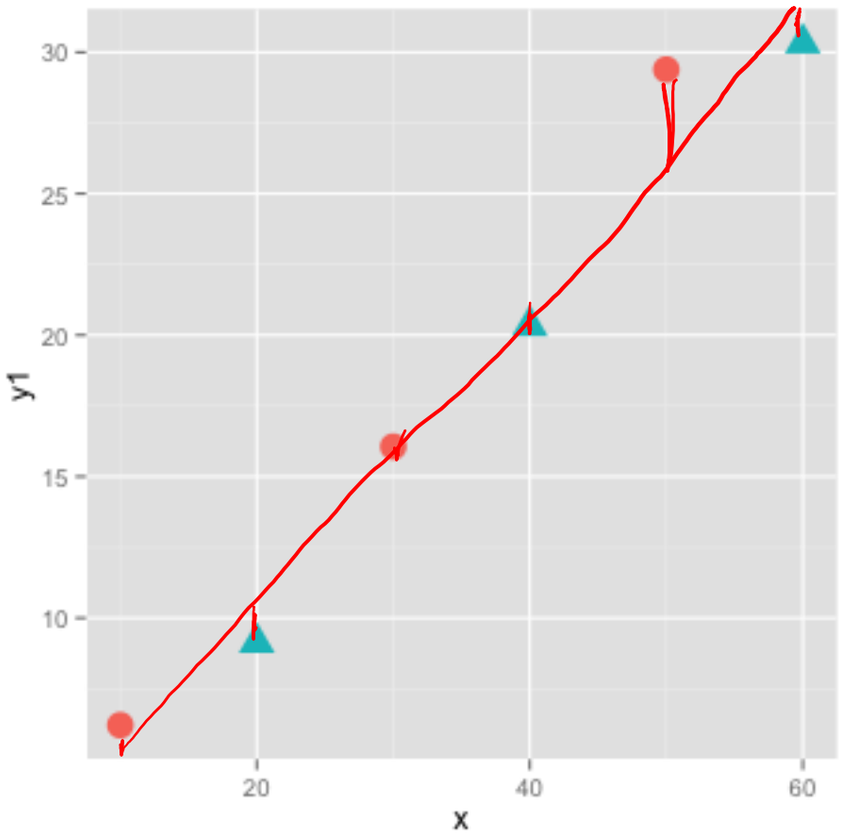
then β_2 and β_3 should be estimated close to zero

both models should fit about the same

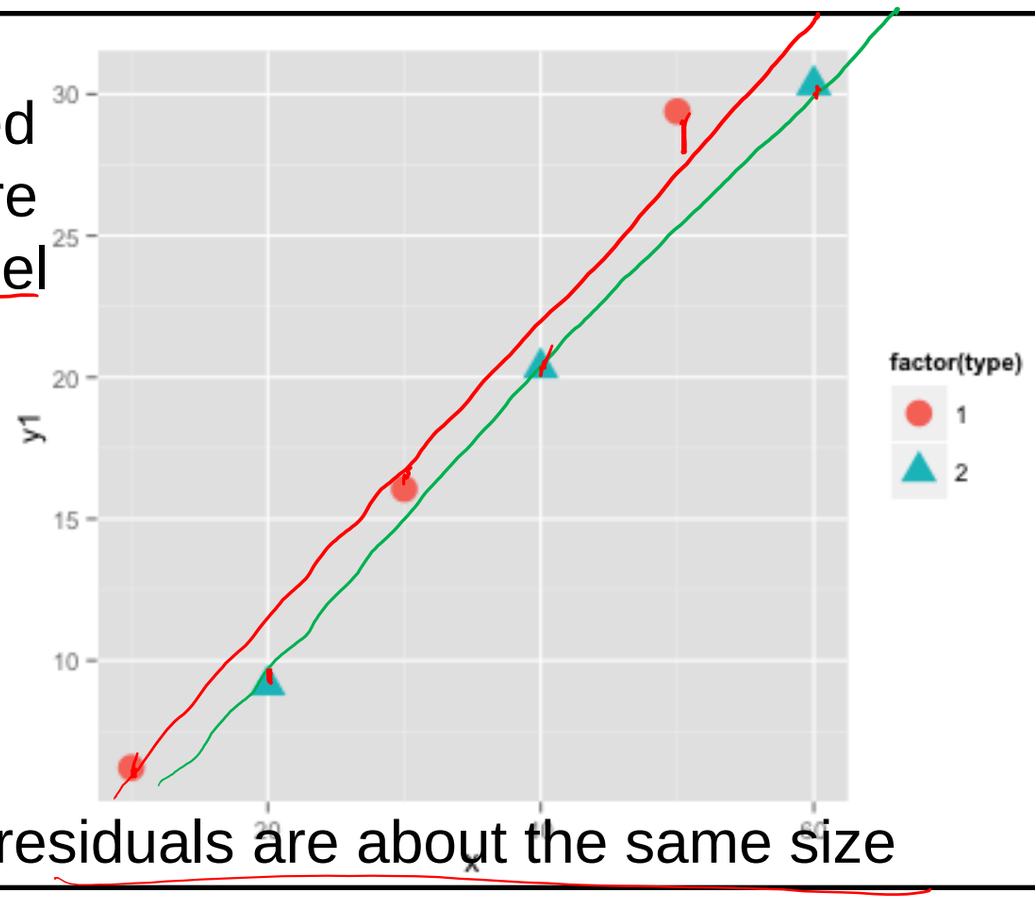
the residuals in both models should be about the same size

fit reduced model

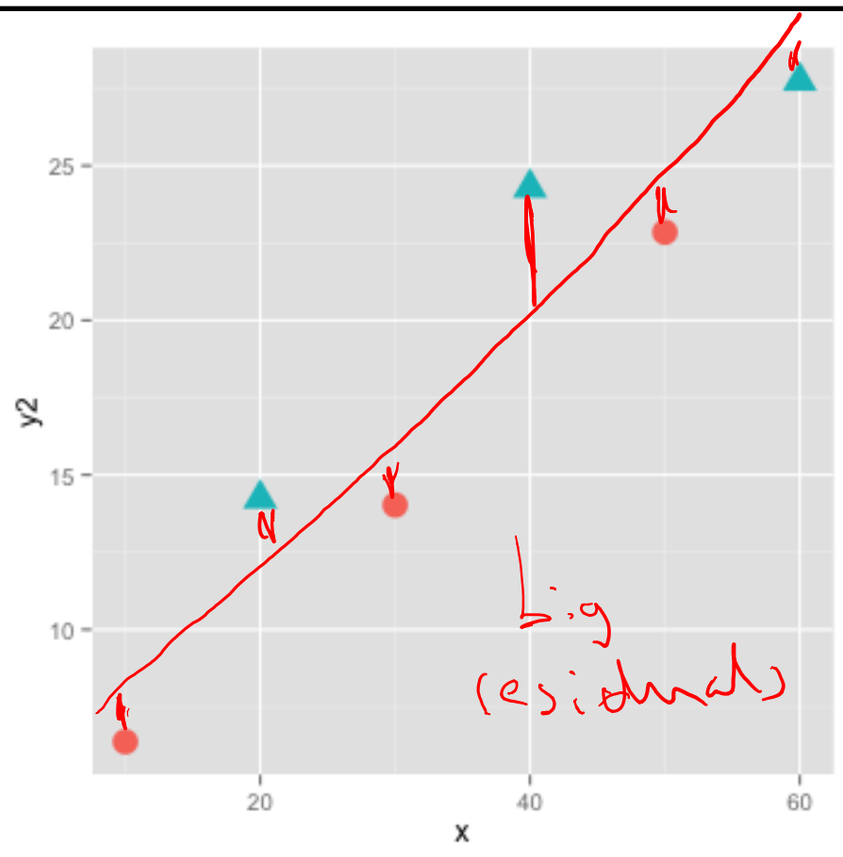
fit full model



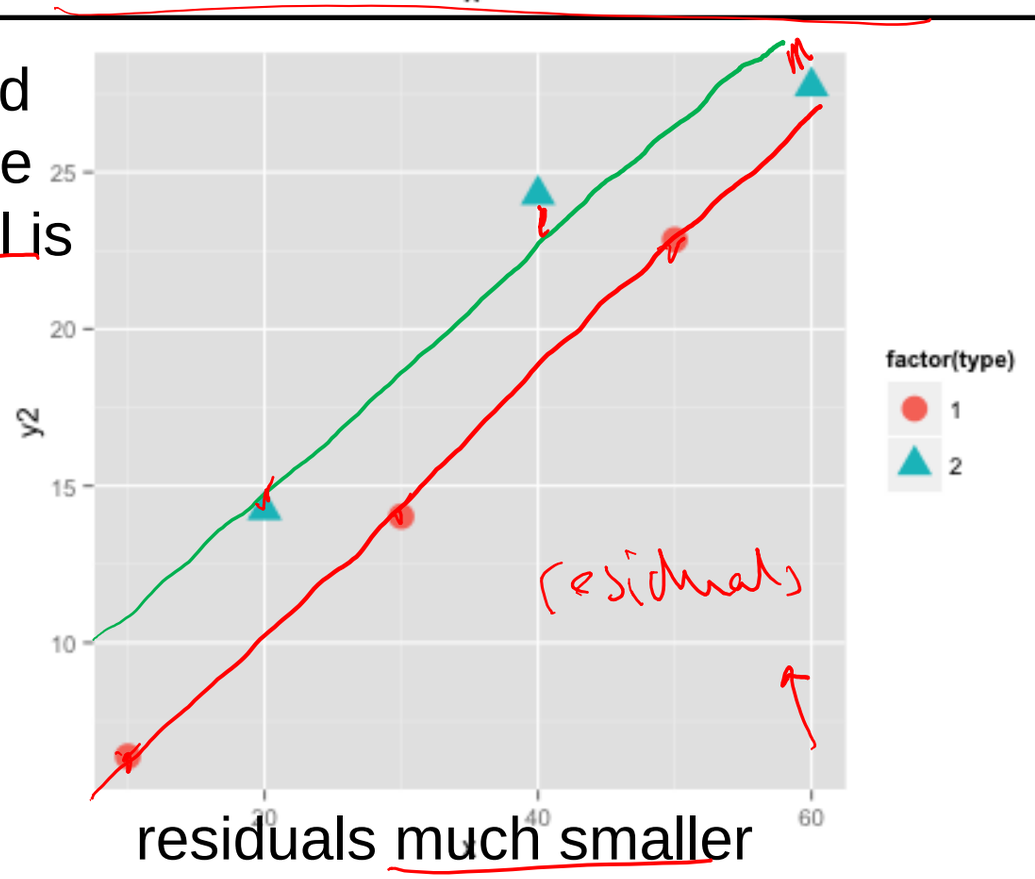
one simulated dataset where reduced model is true



residuals are about the same size



one simulated dataset where reduced model is not true



residuals much smaller

Extra SS F-test

Extra Sum of Squares =

Sum of squared residuals from reduced model -

Sum of squared residuals from full model

extra parameters

$$\text{F-stat} = \frac{\text{Extra Sum of Squares}/\text{Extra degrees of freedom}}{\hat{\sigma}^2_{\text{full}}}$$

If the reduced model is correct, the the F-statistic has an F-distribution with degrees of freedom:
number of parameters being tested and
n - number of parameters in the full model

Full model:

$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\}$

$= \beta_0 + \beta_1 \log \text{Mass} + \beta_2 \, bird + \beta_3 \, ebat$

has 4 parameters, n = 20

Reduced model:

$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\}$

$= \beta_0 + \beta_1 \log \text{Mass}$

has 2 less parameters
or has 2 parameters set to zero

$$\text{F-stat} = \frac{\text{Extra Sum of Squares}/\text{Extra degrees of freedom}}{\hat{\sigma}^2_{\text{full}}}$$

will be F with 2 and 16 degrees of freedom

Shortcut RSS = $\hat{\sigma}^2 \times$ d.f

Full model:

```
> fit_bats <- lm(log(Energy) ~ log(Mass) + Type, data = case1002)
> summary(fit_bats)

...

Residual standard error: 0.186 on 16 degrees of freedom
Multiple R-squared: 0.9815,   Adjusted R-squared: 0.9781
F-statistic: 283.6 on 3 and 16 DF,  p-value: 4.464e-14
```

$\hat{\sigma}^2$ =

RSS =

Reduced model:

```
> fit_eq <- lm(log(Energy) ~ log(Mass), data = case1002)
> summary(fit_eq)

....

Residual standard error: 0.18 on 18 degrees of freedom
Multiple R-squared: 0.9806,   Adjusted R-squared: 0.9795
F-statistic: 907.6 on 1 and 18 DF,  p-value: < 2.2e-16
```

RSS =

F-stat =

1- pf(, 2, 16) =

```
> anova(fit_eq, fit_bats)
Analysis of Variance Table

Model 1: log(Energy) ~ log(Mass)
Model 2: log(Energy) ~ log(Mass) + Type
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     18 0.58289
2     16 0.55332  2  0.029574 0.4276 0.6593
```

There is no evidence that the mean log energy differs for birds, echolocating bats and non-echolocating bats after accounting for body mass (extra sum of squares F-test, p-value = 0.66).

Another example

We relied on a parallel lines regression to answer our question of interest, we might also want to test this is reasonable.

Fit separate lines model (check assumptions look good)

Test whether interaction terms can be dropped.

Full model:

$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\}$

$= \beta_0 + \beta_1 \log \text{Mass} + \beta_2 \, bird + \beta_3 \, ebat +$

$\beta_4 \log \text{Mass} \times bird + \beta_5 \log \text{Mass} \times ebat$

6 parameters

Reduced model:

$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\}$

$= \beta_0 + \beta_1 \log \text{Mass} + \beta_2 \, bird + \beta_3 \, ebat$

4 parameters

```
> anova(fit_bats, fit_sep)
Analysis of Variance Table

Model 1: log(Energy) ~ log(Mass) + Type
Model 2: log(Energy) ~ log(Mass) + Type +
    log(Mass):Type
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     16 0.55332
2     14 0.50487  2   0.04845 0.6718 0.5265
```

There is no evidence that the relationship between mean log energy and log body mass differs for birds, echolocating bats and non-echolocating bats (extra sum of squares F-test, p-value = 0.56).

```
> summary(fit_bats)

Call:
lm(formula = log(Energy) ~ log(Mass) + Type, data = case1002)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23224 -0.12199 -0.03637  0.12574  0.34457

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.57636    0.28724  -5.488 4.96e-05 ***
log(Mass)                  0.81496    0.04454  18.297 3.76e-12 ***
Typenon-echolocating birds  0.10226    0.11418   0.896    0.384
Typeecholocating bats       0.07866    0.20268   0.388    0.703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.186 on 16 degrees of freedom
Multiple R-squared: 0.9815,  Adjusted R-squared: 0.9781
F-statistic: 283.6 on 3 and 16 DF,  p-value: 4.464e-14
```

A extra sum of squares F-test, with the reduced model:

$\mu\{ \log \text{Energy} \mid \log \text{Mass}, \text{Type}\} = \beta_0$