STAT 401A - Statistical Methods for Research Workers

Multiple regression inference

Jarad Niemi (Dr. J)

Iowa State University

last updated: November 10, 2014

Multiple regression model

The multiple regression model is

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}, \sigma^2)$$

Scientific questions/hypotheses can typically be written in one of the following forms:

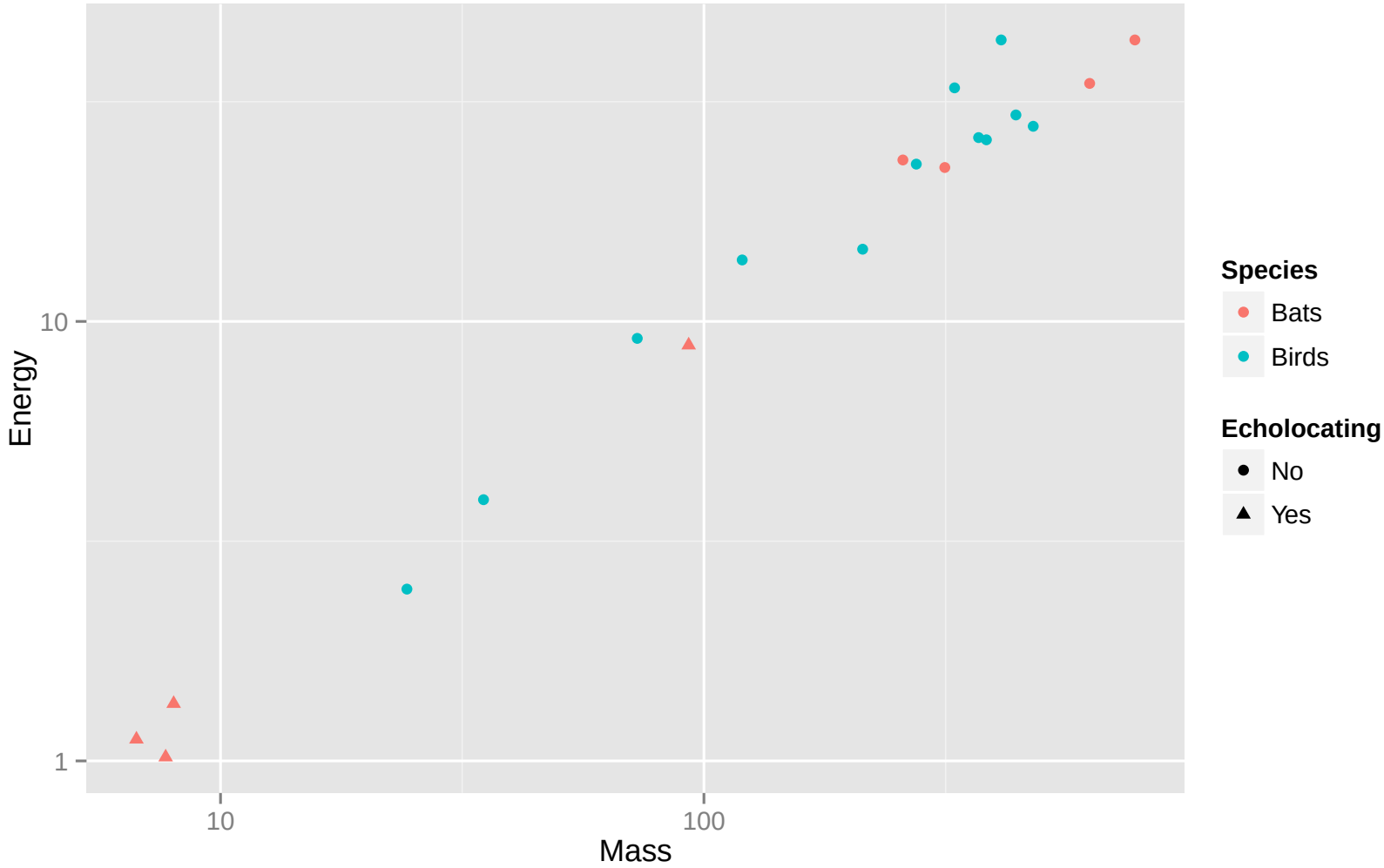
	Estimate	Null Hypothesis
Single coefficient	β_j	$\beta_j = 0$
Linear combination	$\gamma = C_0\beta_0 + C_1\beta_1 + \cdots C_p\beta_p$	$\gamma = 0$
F-test		a set of β_j's are zero
Prediction	$\mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$	

Echolocation energy cost example

Questions:

1. Do body mass or species type have any effect on energy expenditure?
2. After accounting for species type, what is the effect of body mass?
3. After accounting for body mass, is there any difference in energy expenditure amongst the species types?
4. After accounting for body mass, what are the pairwise differences in energy expenditure amongst the species types?
5. What would we expect the energy expenditure to be for an echolocating bat with body mass of 50 grams?

Echolocation energy cost example



Echolocation energy cost example

Consider the model

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2) \qquad \mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3}$$

where, for observation i, we have

- Y_i is log energy expenditure (W)
- $X_{i,1}$ is log body mass (g)
- $X_{i,2}$ is 1 if non-echolocating bats and 0 otherwise
- $X_{i,3}$ is 1 if non-echolocating birds and 0 otherwise

1. F-test: $\beta_1 = \beta_2 = \beta_3 = 0$
2. Coefficient: β_1
3. F-test: $\beta_2 = \beta_3 = 0$
4. Coefficient: β_2, β_3 and Contrast: $\beta_2 - \beta_3$
5. Prediction:

Single coefficient

Hypothesis test:

$$H_0 : \beta_j = 0 \text{ v } H_1 : \beta_j \neq 0$$

calculate the t-statistic and a (two-sided) pvalue

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \qquad p = 2P(t_{n-p} < -|t|).$$

$100(1-\alpha)\%$ two-sided confidence interval:

$$\hat{\beta}_j \pm t_{n-p}(1-\alpha/2)SE(\hat{\beta}_j)$$

Linear combination

Let

$$C_0\beta_0 + C_1\beta_1 + \cdots C_p\beta_p.$$

Hypothesis test:

$$H_0 : \gamma = 0 \text{ v } H_1 : \gamma \neq 0$$

calculate the t-statistic and a (two-sided) pvalue

$$t = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \qquad p = 2P(t_{n-p} < -|t|).$$

$100(1-\alpha)\%$ two-sided confidence interval:

$$\hat{\gamma} \pm t_{n-p}(1-\alpha/2)SE(\hat{\gamma})$$

Testing Composite hypotheses

Comparing two models

- H_0 : (reduced)
- H_1 : (full)

Do the following

1. Calculate extra sum of squares.
2. Calculate extra degrees of freedom
3. Calculate

$$\text{F-statistic} = \frac{\text{Extra sum of squares} \ / \ \text{Extra degrees of freedom}}{\hat{\sigma}^2_{full}}$$

4. Compare this to an F-distribution with
   - numerator degrees of freedom = extra degrees of freedom
   - denominator degrees of freedom = degrees of freedom in estimating $\hat{\sigma}^2_{full}$

What do we say about Y when $X_1 = x_1, \ldots, X_p = x_p$?

We can estimate

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Calculation of the standard error is not simple, but it is straight-forward. We'll just refer to it as the standard error of the mean, $SE(\hat{\mu}\{Y|X\})$.

Just like before, we also have a standard error for a prediction:

$$SE(Pred\{Y|X\})^2 = \hat{\sigma}^2 + SE(\hat{\mu}\{Y|X\})^2.$$

SAS Code and Output

```
DATA case1002;
  INFILE 'case1002.csv' DSD FIRSTOBS=2;
  LENGTH Type $ 30;
  INPUT Mass Type $ Energy;

DATA case1002new;
  INPUT Mass Type & $30.;
  DATALINES;
  50 echolocating bats
;

DATA case1002;
  SET case1002 case1002new;
  lMass   = log(Mass) ;
  lEnergy = log(Energy);
  RUN;

PROC PRINT DATA=case1002; RUN;

PROC GLM DATA=case1002 PLOTS=all;
  CLASS Type(REF='echolocating bats');
  MODEL lEnergy = lMass Type / SOLUTION CLPARM;
  LSMEANS Type / PDIFF CL;
  ESTIMATE 'neBird - neBat' Type 0 -1 1;
  OUTPUT OUT=case1002reg PREDICTED=predicted LCL=lcl UCL=ucl LCLM=lclm UCLM=uclm;

PROC PRINT DATA=case1002reg;
  WHERE Energy=.;
  RUN;
```

SAS Code and Output - ANOVA

The F-test from the ANOVA table tests the null hypothesis $\beta_1 = \cdots = \beta_p = 0$.

The GLM Procedure

Dependent Variable: lEnergy

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	29.42148268	9.80716089	283.59	<.0001
Error	16	0.55331753	0.03458235		
Corrected Total	19	29.97480021			

R-Square	Coeff Var	Root MSE	lEnergy Mean
0.981541	7.491872	0.185963	2.482201

SAS Code and Output - Parameter Estimates

The parameter estimates table provides tests and confidence intervals for individual β_j's.

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-1.497696514 B	0.14986901	-9.99	<.0001
lMass		0.814957494	0.04454143	18.30	<.0001
Type	non-echolocating bats	-0.078663681 B	0.20267926	-0.39	0.7030
Type	non-echolocating birds	0.023598237 B	0.15760050	0.15	0.8828
Type	echolocating bats	0.000000000 B	.	.	.

Parameter		95% Confidence Limits	
Intercept		-1.815404627	-1.179988400
lMass		0.720533885	0.909381102
Type	non-echolocating bats	-0.508324522	0.350997161
Type	non-echolocating birds	-0.310499899	0.357696373
Type	echolocating bats	.	.

SAS Code and Output - LSMEANS

The LSMEANS statement performs pairwise differences.

The GLM Procedure
Least Squares Means

Least Squares Means for effect Type
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: lEnergy

i/j	1	2	3
1		0.3837	0.7030
2	0.3837		0.8828
3	0.7030	0.8828	

Least Squares Means for Effect Type

i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.102262	-0.344318	0.139794
1	3	-0.078664	-0.508325	0.350997
2	3	0.023598	-0.310500	0.357696

SAS Code and Output - ESTIMATE statement

The ESTIMATE statement can be used for specific comparisons.

The GLM Procedure

Dependent Variable: lEnergy

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
neBird - neBat	-0.02359824	0.15760050	-0.15	0.8828	-0.35769637	0.31049990

SAS Code and Output - Type I and III SS

Type I and III SS tables perform specific F-tests:

- Type I performs sequential tests
- Type III performs conditional tests

In this case, the line for lMass in the

	H_0	H_1
Type I SS	β_0	$\beta_0 + \beta_1 X_1$
Type III SS	$\beta_0 + \beta_2 X_2 + \beta_3 X_3$	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Source	DF	Type I SS	Mean Square	F Value	Pr > F
lMass	1	29.39190909	29.39190909	849.91	<.0001
Type	2	0.02957359	0.01478680	0.43	0.6593

Source	DF	Type III SS	Mean Square	F Value	Pr > F
lMass	1	11.57700181	11.57700181	334.77	<.0001
Type	2	0.02957359	0.01478680	0.43	0.6593

SAS Code and Output - OUTPUT statement

PRINTing the data set constructed in the OUTPUT statement provides the predictions and uncertainties

Obs	Type	Mass	Energy	lMass	Energy	predicted	lcl	ucl	lclm	uclm
21	echolocating bats	50	.	3.91202	.	1.69044	1.23358	2.14729	1.45956	1.92132

Now exponentiate since we used log(Energy).

R Code and Output - ANOVA Table

For F-tests in R, fit both models and then use anova to compare them.

```
m0 = lm(log(Energy)~1, case1002)
m1 = lm(log(Energy)~log(Mass)+Type, case1002)
anova(m0,m1)


Analysis of Variance Table

Model 1: log(Energy) ~ 1
Model 2: log(Energy) ~ log(Mass) + Type
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     19 29.9748
2     16  0.5533  3    29.422 283.59 4.464e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R Code and Output - Parameter estimates

```
summary(m1)
```

```
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)               -1.49770    0.14987  -9.993 2.77e-08 ***
log(Mass)                  0.81496    0.04454  18.297 3.76e-12 ***
Typenon-echolocating bats -0.07866    0.20268  -0.388    0.703
Typenon-echolocating birds 0.02360    0.15760   0.150    0.883
---
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
```

```
confint(m1)
```

```
                              2.5 %     97.5 %
(Intercept)               -1.8154046 -1.1799884
log(Mass)                  0.7205339  0.9093811
Typenon-echolocating bats -0.5083245  0.3509972
```

R Code and Output - LSMEANS

Compared to the SAS output, these pvalues are adjusted.

```
library(lsmeans)
lsmeans(m1, 'Type', contr='pairwise')


$lsmeans
 Type                    lsmean         SE df lower.CL upper.CL
 echolocating bats      3.042364 0.16031730 16 2.702507 3.382222
 non-echolocating bats  2.963701 0.09593823 16 2.760321 3.167081
 non-echolocating birds 3.065963 0.05580097 16 2.947670 3.184255

Confidence level used: 0.95

$contrasts
 contrast                                        estimate         SE df t.ratio p.value
 echolocating bats - non-echolocating bats       0.07866368 0.2026793 16   0.388  0.9207
 echolocating bats - non-echolocating birds     -0.02359824 0.1576005 16  -0.150  0.9877
 non-echolocating bats - non-echolocating birds -0.10226192 0.1141826 16  -0.896  0.6507

P value adjustment: tukey method for a family of 3 means
```

R Code and Output - F-tests

Type III SS F-tests, i.e. drop 1 term

```
drop1(m1, test='F')


Single term deletions

Model:
log(Energy) ~ log(Mass) + Type
          Df Sum of Sq     RSS     AIC  F value    Pr(>F)
<none>                  0.5533 -63.751
log(Mass)  1   11.5770 12.1303  -4.000 334.7662 3.758e-12 ***
Type       2    0.0296  0.5829 -66.710   0.4276    0.6593
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R Code and Output - F-tests

or you could fit the models and compare them using anova

```
anova(lm(log(Energy)~Type, case1002),      m1)


Analysis of Variance Table

Model 1: log(Energy) ~ Type
Model 2: log(Energy) ~ log(Mass) + Type
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     17 12.1303
2     16  0.5533  1    11.577 334.77 3.758e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


anova(lm(log(Energy)~log(Mass), case1002), m1)


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
```

R Code and Output - Predictions

```
new = data.frame(Mass=50, Type='echolocating bats')
exp(predict(m1, new, interval='confidence'))
```

```
       fit      lwr      upr
1 5.421844 4.304047 6.829942
```

```
exp(predict(m1, new, interval='prediction'))
```

```
       fit      lwr      upr
1 5.421844 3.433494 8.561654
```