ENTO / RNR 613 – Inferential Tools for Multiple Regression

T-tests are useful for making inferences about the value of *individual* regression coefficients. These regression coefficients describe the *association* between the mean response (Y) and a series of X's. They are used for both hypothesis testing and confidence interval building.

Another approach is available for making inferences about regression parameters: *partial F-tests,* also called *Extra-Sum-of-Squares F-tests*. Extra-Sum-of-Squares F-tests provide much flexibility for hypothesis testing. They can be used 1) to test the effect of a *group* of explanatory variables, and 2) to measure the contribution of one or more explanatory variables to explanation of the variation in the response variable.

Example: Some bats use echolocation to orient themselves with respect to their surroundings. To assess whether sound production is energetically costly, in-flight energy expenditure was measured in 4 non-echolocating bats, 12 non-echolocating birds, and 4 echolocating bats.

Is in-flight energy requirement different between non-echolocating bats and echolocating bats?

First, we need to choose an inferential model to answer that question. For sake of simplicity, we start with a parallel lines regression model, using as a *reference* non-echolocating bats:

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$$

(An abbreviation to indicate a categorical explanatory variable to be modeled with indicator variable is to write that variable in uppercase. To save space, the model could simply be described as: $\mu\{lenergy \mid lmass, TYPE\} = lmass + TYPE$

<<Display 10.5>>

Phrased in term of regression coefficients, our question is $\beta_3 = 0$?

Specifics of the above model:

a) The dummy variable *bird* = 1 for birds, 0 otherwise (create a column in JMP).
b) The dummy variable *ebat* = 1 for echolocating bat, 0 otherwise (another column).
c) The data were log transformed (non-linearity and non-constant variance of the responses).

We first fit the above model to check for need for *transformations*, *outliers*, etc….. (we will see later that the choice of a first model depends on sample size).

But before using this model, we must also fit a rich model to assess whether the lines are truly parallel (Lack-of-Fit test is not useful here because there are no replicates at the levels of lmass):

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat + \beta_4 (lmass \times bird) + \beta_5 (lmass \times ebat)$$

The lines are parallel if *both* $\beta_4 = 0$ *and* $\beta_5 = 0$.

T-tests cannot be used as before (e.g. linear contrasts in ANOVA) to test hypotheses involving *more than one* regression coefficient. This is because the estimates of the regression coefficients included in a model are *not statistically independent* (the estimate of a regression coefficient depends on the presence of the other coefficients in the model). This lack of independence complicates estimation of the SE required for drawing inferences on a combination of regression coefficients with a t-test procedure (calculation of the SE is based on the variance of, and covariance between, the coefficients: see Sleuth p. 288-289).

But the extra-sum-of-squares method (also called partial F-tests) is perfect for testing whether *several* coefficients are all zero.

Recall that:

Extra SS = SS_{res} from *reduced* model – SS_{res} from *full* model, (i.e. we use the "Error" SS from the ANOVA tables)
= Variation unexplained by reduced model – variation unexplained by full model
= Extra variation in the response (Y) explained by the full model

The F-statistic for the extra SS is:

$$F\text{ - }statistic = \frac{\left[\dfrac{Extra\ sum\ of\ squares}{Number\ of\ betas\ being\ tested}\right]}{Estimate\ of\ \sigma^2\ from\ full\ model}$$

Here we have:

Full model:

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat + \beta_4 (lmass \times bird) + \beta_5 (lmass \times ebat)$$

Reduced model:

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$$

ANOVA Table for the Full Model: 6 coefficients

Source	DF	**Sum of Squares**	Mean Square	F Ratio
Model	5	29.469932	5.89399	163.4404
Error	**14**	**0.504868**	**0.03606 (s^2)**	Prob>F
C Total	19	29.974800		<.0001

ANOVA Table for the Reduced Model: 4 coefficients

Source	DF	**Sum of Squares**	Mean Square	F Ratio
Model	3	29.421483	9.80716	283.5887
Error	16	**0.553318**	0.03458	Prob>F
C Total	19	29.974800		<.0001

Extra SS = 0.5533 – 0.5049 = 0.0484
Number of betas tested. = 6-4 = 2
Extra SS F-test = (0.0484/2) / 0.0361 = 0.672 with 2, 14 d.f.

Numerator d.f. = no of coefficient tested; denominator d.f. is from s^2 (Error MS) of full model

$F_{2, 14}$ = 0.672 yields P = 0.53, so there is no evidence that the association between energy expenditure and body size differs among the three types of flying vertebrates (i.e., there is no significant interaction between Body mass and flying type).

**** Extra-sum-of-squares tests are useful to select appropriate inferential models ****

Extra-sum-of-squares test for interaction term is done directly in JMP:

Fit Model (with indicator variables):

lenergy = lmass + bird + ebat + lmass*bird + lmass*ebat.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.202448	1.261334	-0.16	0.8748
lmass	0.5897821	0.206138	2.86	0.0126
bird	-1.37839	1.295241	-1.06	0.3053
ebat	-1.268068	1.28542	-0.99	0.3406
bird X lmass	0.2455883	0.213432	1.15	0.2691
ebat X lmass	0.214875	0.223623	0.96	0.3529

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
lmass	1	1	0.29520027	8.1859	0.0126
bird	1	1	0.04084066	1.1325	0.3053
ebat	1	1	0.03509494	0.9732	0.3406
bird X lmass	1	1	0.04774690	1.3240	0.2691
(ebat X lmass	1	1	0.03329584	0.9233	0.3529

Fit Model (without indicator variables):

lenergy = lmass + TYPE + lmass*TYPE

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.0846	0.439569	-2.47	0.0271
lmass	0.7432698	0.076788	9.68	<.0001
type[1]	0.1322889	0.19371	0.68	0.5058
type[2]	-0.046281	0.121225	-0.38	0.7084
type[1]*(lmass-4.8855)	-0.153488	0.141636	-1.08	0.2968
type[2]*(lmass-4.8855)	0.0921005	0.083166	1.11	0.2868

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
lmass	1	1	3.3787539	93.6929	<.0001
type	2	2	0.0169252	0.2347	0.7939
type*lmass	**2**	**2**	**0.0484495**	**0.6718**	**0.5265**

The Extra SS test comparing models with and without interaction done by hand yielded $F_{2, 14} = 0.672$ and P = 0.53, as in the Table for effect tests above.

So the parallel lines model seems reasonable:

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$$

The question of interest is: $\beta_3 = 0$?

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.57636	0.287236	-5.49	<.0001
lmass	0.8149575	0.044541	18.30	<.0001
bird	0.1022619	0.114183	0.90	0.3837
ebat	0.0786637	0.202679	0.39	0.7030

The two-sided p-value for the coefficient of *ebat* is 0.7030. This provides no evidence that β_3 is different from 0.

When a test yields a large p-value, it is *always possible* that the study was not powerful enough to detect a meaningful relationship.

Reporting a 95% CI emphasizes the fact that *power of the test may have been low* and provides a set of likely values for β_3.

The 95 % CIs for the regression coefficients on the transformed scale are: (JMP calculates this)

Term	Lower 95%	Upper 95%
Intercept	-2.185271	-0.967449
lmass	0.7205344	0.9093806
bird	-0.139793	0.3443171
ebat	**-0.350995**	**0.5083224**

The median in-flight energy expenditure is $\exp(\beta_3)$ times (i.e. 1.08 times) as great for echolocating bats as it is for non-echolocating bats *of similar body mass*. The 95% CI is obtained by taking the anti-log of the endpoints of the CI on the transformed scale: Exp (-0.351) = 0.70 to exp (0.508) = 1.66.

Note: the Dummy variable *ebat* was not log transformed, so obtaining an interpretation for its coefficient on the untransformed scale only considers the fact that Y was log transformed.

Another question: Is there variation in flight energy expenditure among the three vertebrate types?

We compare the following 2 models:

<u>Full model</u>: (parallel lines: 4 parameters)

$$\mu\{lenergy \mid lmass, TYPE\} = \beta_0 + \beta_1 lmass + \beta_2 bird + \beta_3 ebat$$

<u>Reduced model</u>: (common line: 2 parameters)

$$\mu\{lenergy \mid lmass\} = \beta_0 + \beta_1 lmass$$

Extra SS = 0.58289 – 0.55332 = 0.02957
Number of Betas tested = 4 - 2 = 2
s^2 = 0.03459 (from full model)

F-statistic = (0.2957/2) / 0.03458 = 0.428, so p-value for F = 0.428 with 2, 16 d.f. is 0.66.

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

The single-line model is therefore adequate to describe the data in this problem.

Contribution of a single variable: R^2 as a tool for building inferential models

The R-squared statistic is a valuable *descriptor* of the fit of a model.

It is calculated as: $$\mathbf{R^2} = \frac{\textit{Total SS - Residual SS}}{\textit{Total SS}} \times 100\%$$

It measures the *amount of total variation in the response variable* that is explained by the regression on the explanatory variables.

Example: Galileo measured horizontal distance covered by a bronze ball released at different heights from an inclined plane on a table. Regression can be used to describe the horizontal distance traveled, which would help figuring the type of trajectory taken by the ball.

A quadratic regression model: **Distance = 199.91 + 0.71 Height – 0.00034 Height ^{2}**

Summary of Fit

RSquare	0.990339
RSquare Adj	0.985509
Root Mean Square Error	13.6389
Mean of Response	434
Observations (or Sum Wgts)	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	76277.922	38139.0	205.0267
Error	4	744.078	186.0	Prob>F
C Total	6	77022.000		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	199.91282	16.75945	11.93	0.0003
height	0.7083225	0.074823	9.47	0.0007
height2	-0.000344	0.000067	-5.15	0.0068

Both the quadratic and linear coefficients are different from 0. From the ANOVA table, we find that the R-square is (76277.9 / 77022) x 100 % = 99.03%. The fit of the model is very good.

<<Display 10.2>>
Do we need a cubic term in the model?

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	155.77551	8.32579	18.71	0.0003
height	1.115298	0.065671	16.98	0.0004
height2	-0.001245	0.000138	-8.99	0.0029
height3	0.0000005	8.327e-8	6.58	0.0072

Summary of Fit

RSquare	0.999374
RSquare Adj	0.998747
Root Mean Square Error	4.010556
Mean of Response	434
Observations (or Sum Wgts)	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	76973.746	25657.9	1595.189
Error	3	48.254	16.1	Prob>F
C Total	6	77022.000		<.0001

The p-value for the coefficient of height-cubed provides evidence that the cubic term is different from 0. But how much more variability in the response variable does it explain?

The extra amount of variation explained in the response variable arising from addition of the cubic term is:

Extra sum of squares = SS_{res} from reduced model – SS_{res} from full model
= Unexplained by reduced model – unexplained by full
=744.078 – 48.254
= 695.824

This only represents an increase of (695.824 / 77022.00) X 100 % = 0.903 % in amount of total variation in the response variable explained by the new model.

The percentage of the total variation in the response variable *not explained* by the quadratic model was 100 – (0.990339 X 100) = 0.966 %).

So the **cubic term** explains a significant proportion of the *remaining variability from the reduced quadratic model* (0.903/0.966 = 93.5 % of the remaining variability). But the gain in term of total variation explained is small compared to what was accomplished by the quadratic model (i.e. about 1%).

When should quadratic (or higher order terms) be included in the model?

They should not be routinely included, and are useful in 4 situations:

1) When there are good reasons to suspect the response to be non-linear
2) When we search for an optimum or minimum
3) When precise predictions are needed (presumably few explanatory variables are used)
4) To produce a rich model for assessing the fit of an inferential one.

When should an Interaction term be included?

Not routinely. Inclusion is indicated:

1) When the question of interest pertains to interactions
2) When good reasons exist to suspect interactions
3) When assessing the fit of an inferential model

Occam's Razor again

R^2 can *always be made greater* by adding explanatory variables. For example fluctuation in the Dow Jones Index during nine days in June 1994 was predicted with the following seven explanatory variables:

high temperature in NY city on the previous day;
low temperature on the previous day;
an indicator variable equal to 1 if the forecast for the day was sunny and 0 otherwise;
an indicator variable equal to 1 if the New York Yankees won their baseball game on the previous day and 0 otherwise;
the number of runs the Yankee scored;
an indicator variable equal to 1 if the new York Mets won their baseball game on the previous day and 0 if not;
the number of runs the Mets scored.

The R^2 of the model was *89.6 %*. Would you use this model to invest in stocks?
I hope not! The model used 7 variables to explain variation in 9 data points. The model fitted well because there were almost as many variables as observations. That *particular* equation fits well but would be very unlikely to fit future data.

Foundation for the Occam's Razor principle: *Simple models* that adequately explain the data are more likely to have predictive power than complex models that are more likely to fit the data without reflecting any real associations between the variables. This should be kept in mind when deciding whether to keep quadratic (or higher order) terms and interactions in inferential models.

Example: Predicting Stock prices with 7 arbitrarily chosen variables (I had to try it for myself!).

Response: Stock

Summary of Fit

RSquare	0.966228
RSquare Adj	0.729821
Root Mean Square Error	15.98882
Mean of Response	58.77778
Observations (or Sum Wgts)	9

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	17.617805	41.79174	0.42	0.7460
HighT	5.1361639	4.059363	1.27	0.4258
LowT	-5.699337	4.944225	-1.15	0.4549
Sun	71.085717	23.39164	3.04	0.2024
Yank	25.81683	20.4658	1.26	0.4267
YankScore	4.44963	3.665936	1.21	0.4387
Mets	-35.30382	30.22017	-1.17	0.4507
MetsRuns	-8.082005	5.428728	-1.49	0.3765

Adjusted R-square Statistic

The adjusted R-square includes a penalty for unnecessary explanatory variables. It measures the proportion of the variation in the responses explained by the regression model, but this time the residual **mean squares** rather than the residual sums of squares are used:

$$\textbf{Adjusted } \mathbf{R^2} \textbf{ = 100 } \frac{\textit{(Total mean square)} - \textit{(Residual mean square)}}{\textit{Total mean square}} \%$$

With increased number of regression coefficients included in the model, the Residual SS always decline, so the **R-squared statistic** always increases.

However for the **Adjusted $\mathbf{R^2}$**, the number of d.f. associated with the *Residual mean square* is n - #betas [Residual mean square = Residual SS / (n - #betas)]. This tends to increase the value of the residual MS when factors included in the model do not account for much of the variation in the response variable. On the other hand, the *Total mean square* does not change when more factors are included in a model.

Thus, an increase in the number of “useless” regression coefficients in a model increases the discrepancy between the **adjusted $\mathbf{R^2}$** and the **R-squared**. The adjusted R^2 is useful for **casual assessment** of improvement of fit: factors that increase the difference between **$\mathbf{R^2}$ and Adjusted $\mathbf{R^2}$** would in general be less useful in a model.

R^2 is a better **descriptor** than adjusted R^2 of the total variation in the response variable explained by a model.

Still the Adjusted R^2 in the above model is 73.0 %. This illustrates that R-squared is a difficult statistic to use for model checking, model comparison, or inference.