Hypotheses and MLR

Lily Tomkovic via Nick Rosenberger

"Lecture"

Linear models and categorical indeendent variables

Examples:

$$\mu = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2$$

Where β_j is a partial correlation coefficient. You can end up with fun multiple level ANOVA analyses.

Take-Home Message

Two key questions:

1. What are the dependent and independent variables?
2. What distribution is appropriate for the dependent variable?

- characteristics of a linear model (slope veruss partial regression coefficient)
- construction and use of dummy variables to represeent categorical independent variables in linear models
- the diversity of relations between dependent and independent variables that can be represented by linear models.

Hypothesis Testing for Linear Models

$$sampling\ error = sample\ statistic - population\ parameter$$

Approaches to statistical inference:
1. Frequentist
2. Information Theory
3. Bayseian

Frequentist Hypothesis Testing

Test statistic measures "effect" (t, F, χ^2, etc), and is subject to sampling error. You can make conclusions about your data that are 'wrong' for a few different reasons:

- **Type I** Failing to detect significant effects
- **Type II** I am saying there's a significant effect (I reject the null hypothesis) when there's not

In frequentist statistics, at the end, you'll either: * accept H_0 - independent variable(s) **does not** have a *statistically significant effect* on the dependent variable
* reject H_0 - excluded independent variable(s) **does** have a *statistically significant effect* on the dependent variable

Information Theory Approach

You can use all of the parameters of interest (the "full model") or a "reduced" model which uses only the independent variables of interest. Which is the best fitting model?

$$Fit(full) - Fit(reduced) = contribution\ of\ omitted\ variable(s)$$

$$v_{full} - v_{reduced} = df_{test}$$

Example:

- Full Model: $\mu_i = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2$
- Reduced Model: $\mu_i = \beta_0$

Your full model may give you a good fit, but sometimes you might want to only include significant variables.

To test for multiple models, you can do a maximum likelihood parameter estimates via:

$$G = 2(LL_{full} - LL_{reduced})$$

Aikake's Information Criterion

$$AIC = -2lnL(b) + 2v$$

All the data is the same across the models in order to perform AIC analysis. b is the set of maximum-likelihood estimates of all parameters of a specific model. $L(b)$ is model likelihood. A smaller AIC is better. This compares across the full v. nested as well as the nested amongst themselves. It *is* possible at the end of an AIC that there is no "best-fitting" model identified. There are further methods to identify which models to retain and consider. See Richards (2005, Ecology 86:2805-2814) and Richards (2008, Journal of Applied Ecology 45:218-227).

Introduction to the Problem Set

A fisheries biologist is assessing the effects of mercury on contamination and performance of rainbow trout in small lakes in south-central British Columbia. Based on previous studies, she recognizes that mercury has diverse effects, which she is interested in studying. In particular, she intends to assess the following possible effects:

1. As the main organ of detoxification, the liver is involved in extracting mercury from the blood, and fish from lakes with high mercury concentrations are expected to have enlarged livers. Of course, liver mass is also expected to vary positively with fish size.

2. If the liver is not completely effective, then mercury should accumulate in other fish tissues. Such bioaccumulation may also depend on the mercury concentration in the water and a fish's size (large fish pass more water over their gills to acquire needed oxygen).

3. The physiological effects of mercury, including the cost of removing it from the blood, may reduce a fish's growth capacity. Of course, growth capacity at any age is also expected to vary with a fish's current size (positive) and the population density (negative because of intraspecific competition).

To test these expectations, the biologist samples water and fish from 35 lakes. For each lake she measures the mercury concentration in the water (waterconc: $\mu g/L$). She also collects a sample of 20 two-year-old fish from each lake that were individually tagged when they were stocked during the previous year. For each fish she measures body mass during the preceding year (fishmass: g), liver mass (liver: g), growth since the previous year (growth: g), and the tissue mercury concentration (fishconc: parts per million). She uses the averages for the fish from a given lake as the single observation for that lake. Finally, for each lake she conducts a netting survey to quantify the average fish density per lake (fish per cubic metre). The resulting data can be found in **"mercury.txt"** on the D2L site.

Table 1: Example of the data collected for mercury testing in lakes.

lake	waterconc	fishmass	liver	fishdens	fishconc	growth
1	0.017	22.6	1.34	1.3	0.98	12.0
2	0.019	13.2	0.90	1.1	0.61	1.9
3	0.010	24.5	0.90	1.4	0.84	15.2
4	0.002	33.3	1.10	1.1	0.74	31.0
5	0.008	19.6	0.82	0.1	0.68	11.5
6	0.006	28.3	0.98	1.2	0.86	17.9

Using these data conduct three sets of analyses. For each set of analyses, consider all possible first-order regression models (with intercepts). Based on your results, answer the following question:

For each analysis do the results of the simple regressions agree with those of the multiple regression? If not, explain the difference.

Work through code here

a. the effects of fish size and water mercury concentration on liver mass;

```
# Import car library for regression analysis
# install.packages("car") # Only necessary if package is not installed
library(car)  # includes vif

df <- read.table('../data/mercury.txt',sep="", header=TRUE)
```

```r
# Create linear models
liv <- lm(liver~fishmass+waterconc, data=df) # How does liver size vary with fishmass and waterconc
summary(liv)
```

```
##
## Call:
## lm(formula = liver ~ fishmass + waterconc, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22924 -0.07837 -0.01821  0.08663  0.22834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.174821   0.172591  -1.013    0.319
## fishmass     0.036006   0.004836   7.446 1.79e-08 ***
## waterconc   33.279404   6.045840   5.505 4.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1146 on 32 degrees of freedom
## Multiple R-squared:  0.634,  Adjusted R-squared:  0.6112
## F-statistic: 27.72 on 2 and 32 DF,  p-value: 1.035e-07
```

```r
# Note that the "Estimate" is relative to the variable scale, so even though fishmass has a much lower

liv.mass <- lm(liver~fishmass, data=df)
liv.wc <- lm(liver~waterconc, data=df)
summary(liv.mass)
```

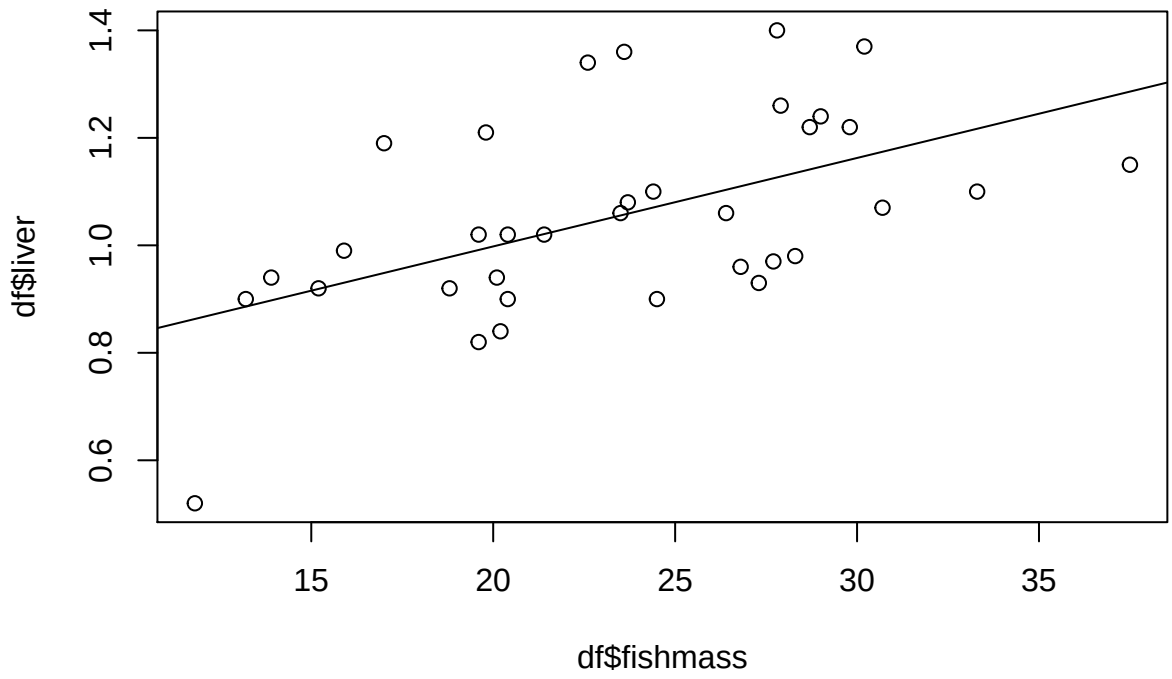
```
##
## Call:
## lm(formula = liver ~ fishmass, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34292 -0.12650  0.00444  0.06976  0.30279
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.668630   0.109131   6.127 6.64e-07 ***
## fishmass    0.016465   0.004512   3.649    9e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1574 on 33 degrees of freedom
## Multiple R-squared:  0.2875, Adjusted R-squared:  0.2659
## F-statistic: 13.32 on 1 and 33 DF,  p-value: 0.0008996
```

```r
summary(liv.wc)
```

```
##
## Call:
## lm(formula = liver ~ waterconc, data = df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53566 -0.12008 -0.03402  0.14363  0.34528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.05214    0.08351  12.600  3.7e-14 ***
## waterconc    0.23486    6.68278   0.035    0.972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1865 on 33 degrees of freedom
## Multiple R-squared:  3.743e-05,  Adjusted R-squared:  -0.03026
## F-statistic: 0.001235 on 1 and 33 DF,  p-value: 0.9722
```

```
plot(df$liver~df$fishmass)
abline(liv.mass)
```




```
# Perform AIC
AIC(liv, liv.mass, liv.wc)
```

```
##          df       AIC
## liv       4 -47.47945
## liv.mass  3 -26.16175
## liv.wc    3 -14.29808
```

```
# The more negative, the better fit!
```

b. the effects of fish size, liver mass and water mercury concentration on the mercury concentration in fish tissue; and

```
# Create linear models
mtis <- lm(fishconc~fishmass+liver+waterconc, data=df)
mtis.mass.liv <- lm(fishconc~fishmass+liver, data=df)
mtis.mass.wc <- lm(fishconc~fishmass+waterconc, data=df)
mtis.liv.wc <- lm(fishconc~liver+waterconc, data=df)
mtis.mass <- lm(fishconc~fishmass, data=df)
mtis.liv <- lm(fishconc~liver, data=df)
mtis.wc <- lm(fishconc~waterconc, data=df)

AIC(mtis, mtis.mass.liv, mtis.mass.wc, mtis.liv.wc, mtis.mass, mtis.liv, mtis.wc)
```

```
##               df       AIC
## mtis           5 -41.80896
## mtis.mass.liv  4 -33.27593
## mtis.mass.wc   4 -41.45250
## mtis.liv.wc    4 -24.48630
## mtis.mass      3 -34.16881
## mtis.liv       3 -24.44175
## mtis.wc        3 -17.92871
```

The full model is the best fitting (i.e. lowest AIC) and the next best is within 6 from AIC so we need to consider it. In this case, that's mtis and mtis.mass.wc.

```
summary(mtis)
```

```
##
## Call:
## lm(formula = fishconc ~ fishmass + liver + waterconc, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17628 -0.09553 -0.01033  0.06974  0.35566
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.199550   0.187732  -1.063  0.29601
## fishmass     0.040905   0.008559   4.779 4.04e-05 ***
## liver       -0.278112   0.189274  -1.469  0.15181
## waterconc   29.799430   9.032149   3.299  0.00244 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1227 on 31 degrees of freedom
## Multiple R-squared:  0.5689, Adjusted R-squared:  0.5272
## F-statistic: 13.64 on 3 and 31 DF,  p-value: 7.601e-06
```

```
summary(mtis.mass.wc)
```

```
##
## Call:
## lm(formula = fishconc ~ fishmass + waterconc, data = df)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.178903 -0.105788  0.002634  0.069817  0.292157
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.150930   0.188109  -0.802  0.42827
## fishmass     0.030892   0.005271   5.861 1.62e-06 ***
## waterconc   20.544033   6.589449   3.118  0.00384 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1249 on 32 degrees of freedom
## Multiple R-squared:  0.5389, Adjusted R-squared:  0.5101
## F-statistic:  18.7 on 2 and 32 DF,  p-value: 4.178e-06
```

So here, even though liver is *insignificant* in the full model, it aids in the fit of the overall model. The estimate of the coefficient in the full model for fishmass was 0.040905 and for the reduced model it is 0.030892. For waterconc it's 29.799430 in the full model and 20.544033 in the reduced. For both parameter estimates, the impact is greater or more significant if liver is included in the model.
One could conclude fish mass and water concentration are increasing the amount of mercury in the fish, and once we've accounted for the liver size, that effect is more profound. That also means that the liver size has no bearing on the amount of mercury in the fish' tissue.

c. the effects of fish size, fish density and water mercury concentration on fish growth.

```
# Create linear models
fgr <- lm(growth~fishmass+fishdens+waterconc, data=df)
fgr.mass.dens <- lm(growth~fishmass+fishdens, data=df)
fgr.dens.wc <- lm(growth~fishdens+waterconc, data=df)
fgr.mass.wc <- lm(growth~fishmass+waterconc, data=df)
fgr.mass <- lm(growth~fishmass, data=df)
fgr.dens <- lm(growth~fishdens, data=df)
fgr.wc <- lm(growth~waterconc, data=df)

AIC(fgr, fgr.mass.dens, fgr.dens.wc, fgr.mass.wc, fgr.mass, fgr.dens, fgr.wc)
```

```
##               df      AIC
## fgr            5 197.8980
## fgr.mass.dens  4 207.7371
## fgr.dens.wc    4 219.8770
## fgr.mass.wc    4 197.4371
## fgr.mass       3 207.8545
## fgr.dens       3 256.9490
## fgr.wc         3 218.1428
```

In this case the smallest AIC is fgr.mass.wc and even though the full model is within 6 AIC of it, we needn't consider the full model. Yay science!