Outline

1. Significance testing
   - An example with two quantitative predictors
   - ANOVA f-tests
   - Wald t-tests
   - Consequences of correlated predictors

2. Model selection
   - Sequential significance testing
     - Nested models
     - Additional Sum-of-Squares principle
     - Sequential testing
   - the adjusted R^2
   - Likelihood
   - the Akaike criterion

Pesticide example

```
> tox = read.table("toxic.txt", header=T)
> tox
    dose weight toxicity
1  0.696  0.321    0.324
2  0.729  0.354    0.367
3  0.509  0.134    0.321
4  0.559  0.184    0.375
5  0.679  0.304    0.345
6  0.583  0.208    0.341
7  0.742  0.367    0.327
8  0.781  0.406    0.256
9  0.865  0.490    0.214
10 0.723  0.223    0.501
11 0.940  0.440    0.318
12 0.903  0.403    0.317
13 0.910  0.410    0.349
14 0.684  0.184    0.402
15 0.904  0.404    0.374
16 0.887  0.387    0.340
17 0.593  0.093    0.598
18 0.640  0.140    0.444
19 0.512  0.012    0.543
```

A study was conducted to assess the toxic effect of a pesticide on a given species of insect.
`dose`: dose rate of the pesticide,
`weight`: body weight of an insect,
`tocicity`: rate of toxic action.

Candidate models

Consider 4 possible linear models for this data:

$$\begin{aligned}
y_i &= \beta_0 + e_i \\
y_i &= \beta_0 + \beta_1 \mathrm{dose}_i + e_i \\
y_i &= \beta_0 + \beta_2 \mathrm{weight}_i + e_i \\
y_i &= \beta_0 + \beta_1 \mathrm{dose}_i + \beta_2 \mathrm{weight}_i + e_i
\end{aligned}$$

Fit these models in R:

```
fit.0  = lm(toxicity ~ 1,                data=tox)
fit.d  = lm(toxicity ~ dose,             data=tox)
fit.w  = lm(toxicity ~      weight, data=tox)
fit.dw = lm(toxicity ~ dose+weight, data=tox)
fit.wd = lm(toxicity ~ weight+dose, data=tox)
```

Comparing models using anova

```
> anova(fit.0, fit.d)
Analysis of Variance Table
Model 1: toxicity ~ 1
Model 2: toxicity ~ dose
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1      18 0.1576
2      17 0.1204  1    0.0372 5.26  0.035 *

> anova(fit.w, fit.wd)
Analysis of Variance Table
Model 1: toxicity ~ weight
Model 2: toxicity ~ weight + dose
  Res.Df      RSS Df Sum of Sq      F   Pr(>F)
1      17 0.065499
2      16 0.034738  1  0.030761 14.168 0.001697 **
```

Testing $\beta_1 = 0$ (dose effect) gives a different result whether weight is included in the model or not.

Comparing models using anova

We did two different tests:

- $H_0 : [\beta_1 = 0|\beta_0]$ is testing $\beta_1 = 0$ (or not) given that only the intercept β_0 is in the model
- $H_0 : [\beta_1 = 0|\beta_0, \beta_2]$ is testing $\beta_1 = 0$ assuming that an intercept β_0 and a weight effect β_2 are in the model.

They make different assumptions, may reach different results.

The anova function, when given two (or more) different models, does an f-test by default.

Source	df	SS	MS
$\beta_2\|\beta_0$	1	SS($\beta_2\|\beta_0$)	SS($\beta_2\|\beta_0$)/1
$\beta_1\|\beta_0, \beta_2$	1	SS($\beta_1\|\beta_0, \beta_2$)	SS($\beta_1\|\beta_0, \beta_2$)/1
Error	$n-3$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	SSError/($n-3$)
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	

Fact: if H_0 is correct, $F = \mathrm{MS}(\beta_1|\beta_0, \beta_2)/\mathrm{MSError} \sim F_{1,n-3}$.

Comparing models using anova

Be very careful with anova on a single model:

```
> anova(fit.w, fit.wd)
> anova(fit.w, fit.dw)  # same output

> anova(fit.dw)
Response: toxicity
          Df   Sum Sq  Mean Sq F value    Pr(>F)
dose       1 0.037239 0.037239  17.152 0.0007669 ***
weight     1 0.085629 0.085629  39.440 1.097e-05 ***
Residuals 16 0.034738 0.002171

> anova(fit.wd)
Response: toxicity
          Df   Sum Sq  Mean Sq F value    Pr(>F)
weight     1 0.092107 0.092107  42.424 7.147e-06 ***
dose       1 0.030761 0.030761  14.168  0.001697 **
Residuals 16 0.034738 0.002171
```

Each predictor is added one by one (Type I SS).

The order matters!

Which one is appropriate to test a body weight effect?
to test a dose effect?

Comparing models using `drop1`

```
> drop1(fit.dw, test="F")
Single term deletions
Model: toxicity ~ dose + weight

       Df Sum of Sq      RSS      AIC F value     Pr(F)
<none>              0.034738 -113.783
dose    1  0.030761 0.065499 -103.733  14.168  0.001697 **
weight  1  0.085629 0.120367  -92.171  39.440 1.097e-05 ***

> drop1(fit.wd, test="F")
Single term deletions
Model: toxicity ~ weight + dose
       Df Sum of Sq      RSS      AIC F value     Pr(F)
<none>              0.034738 -113.783
weight  1  0.085629 0.120367  -92.171  39.440 1.097e-05 ***
dose    1  0.030761 0.065499 -103.733  14.168  0.001697 **
```

F-tests, to test each predictors after accounting for all others (Type III SS). The order does not matter.

Comparing models using anova

- Use anova to compare *multiple* models.
- Models are nested when one model is a particular case of the other model.
- anova can perform f-tests to compare 2 or more nested models

```
> anova(fit.0, fit.d, fit.dw)
Model 1: toxicity ~ 1
Model 2: toxicity ~ dose
Model 3: toxicity ~ dose + weight
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     18 0.157606
2     17 0.120367  1  0.037239 17.152 0.0007669 ***
3     16 0.034738  1  0.085629 39.440 1.097e-05 ***

> anova(fit.0, fit.w, fit.wd)
Model 1: toxicity ~ 1
Model 2: toxicity ~ weight
Model 3: toxicity ~ weight + dose
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     18 0.157606
2     17 0.065499  1  0.092107 42.424 7.147e-06 ***
3     16 0.034738  1  0.030761 14.168  0.001697 **
```

Parameter inference using `summary`

The `summary` function performs Wald t-tests.

```
> summary(fit.d)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.6049     0.1036   5.836 1.98e-05 ***
dose         -0.3206     0.1398  -2.293   0.0348 *  

Residual standard error: 0.08415 on 17 degrees of freedom
Multiple R-squared: 0.2363,     Adjusted R-squared: 0.1914
F-statistic: 5.259 on 1 and 17 DF,  p-value: 0.03485

> summary(fit.wd)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.22281    0.08364   2.664  0.01698 *  
weight      -1.13321    0.18044  -6.280 1.10e-05 ***
dose         0.65139    0.17305   3.764  0.00170 ** 

Residual standard error: 0.0466 on 16 degrees of freedom
Multiple R-squared: 0.7796,     Adjusted R-squared: 0.752
F-statistic:  28.3 on 2 and 16 DF,  p-value: 5.57e-06
```

Parameter inference using `summary`

The order does *not* matter for t-tests:

```
> summary(fit.wd)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22281    0.08364   2.664  0.01698 *
weight      -1.13321    0.18044  -6.280 1.10e-05 ***
dose         0.65139    0.17305   3.764  0.00170 **
...

> summary(fit.dw)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22281    0.08364   2.664  0.01698 *
dose         0.65139    0.17305   3.764  0.00170 **
weight      -1.13321    0.18044  -6.280 1.10e-05 ***

Residual standard error: 0.0466 on 16 degrees of freedom
Multiple R-squared: 0.7796,     Adjusted R-squared: 0.752
F-statistic:  28.3 on 2 and 16 DF,  p-value: 5.57e-06
```

Parameter inference

- For testing the same hypothesis, the f-test and t-test match: $(-2.293)^2 = 5.26$ and $3.764^2 = 14.168$
- But two different tests:
  - Weak evidence for a dose effect if body weight is ignored
  - Strong evidence of a dose effect after adjusting for a body weight effect.
- Results are different because dose and weight are correlated.

Consequences of correlated predictors

Also called multicollinearity.

- F-tests are order dependent
- Counter-intuitive results:

```
> summary(fit.d)
...         Estimate Std. Error t value Pr(>|t|)
dose         -0.3206     0.1398  -2.293   0.0348 *
```

Negative effect of dose, if dose alone!! As dose rate increases, the rate of toxic action decreases!? When results are against intuition, this is a warning.

Correlation between dose and body weight:

```
> plot(dose ~ weight, data=tox)
> with(tox,  cor(dose,weight))
[1] 0.8943634
> plot(toxicity ~ dose, data=tox, pch=16)
> plot(toxicity ~ dose, data=tox, pch=16, col=grey(
> plot(toxicity ~ dose, data=tox, pch=16, col=grey(
```

Can we have uncorrelated predictors?

Predictors x_1 and x_2 are uncorrelated if

$$\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0$$

- In designed experiments we can choose combination of x_{i1} and x_{i2} values so that these predictors are uncorrelated in the experiment.
- Qualitative predictors: can also be correlated
- Example: `sex` and `smoke`, in the fev data set
- Completely balanced designs (more later)

Outline

1. Significance testing
   - An example with two quantitative predictors
   - ANOVA f-tests
   - Wald t-tests
   - Consequences of correlated predictors

2. Model selection
   - Sequential significance testing
     - Nested models
     - Additional Sum-of-Squares principle
     - Sequential testing
   - the adjusted R^2
   - Likelihood
   - the Akaike criterion

Model selection

Testing parameters is the same as selecting between 2 models. In our example, we have 4 models to choose from.

1. $y_i = \beta_0 + e_i$
2. $y_i = \beta_0 + \beta_2 \text{weight}_i + e_i$
3. $y_i = \beta_0 + \beta_1 \text{dose}_i + e_i$
4. $y_i = \beta_0 + \beta_1 \text{dose}_i + \beta_2 \text{weight}_i + e_i$

- $H_0 : [\beta_2 = 0 | \beta_0]$ is a test to choose between model 1 (H_0) and model 2 (H_a).
- $H_0 : [\beta_2 = 0 | \beta_0, \beta_1]$ is a test to choose between model 3 (H_0) and model 4 (H_a).
- $H_0 : [\beta_1 = \beta_2 = 0 | \beta_0]$ is an overall test to choose between model 0 (H_0) and model 4 (H_a).

Nested models

Two models are nested if one of them is a particular case of the other one: the simpler model can be obtained by setting some coefficients of the more complex model to particular values.

Among the 4 models to explain pesticide toxicity

- which ones are nested?
- which ones are not nested?

Example: Cow data set

4 treatment with 4 levels of an additive in the cow feed:
control (0.0), low (0.1), medium (0.2) and high (0.3)
`treatment`: factor with 4 levels
`level`: numeric variable, whose values are 0, 0.1, 0.2 or 0.3.
`fat`: fat percentage in milk yield (%)
`milk`: milk yield (lbs)

Are these models nested?

1. $\text{fat}_i = \beta_0 + \beta_2 * \text{initial.weight}_i + e_i$
2. $\text{fat}_i = \beta_0 + \beta_{j(i)} + e_i$, where $j(i)$ is the treatment # for cow i
3. $\text{fat}_i = \beta_0 + \beta_1 * \text{level}_i + e_i$

Multiple R^2

R^2 is a measure of fit quality:

$$R^2 = \frac{\text{SSRegression}}{\text{SSTotal}}$$

It is the proportion of the total variation of the response variable explained by the multiple linear regression model.

Equivalently:

$$R^2 = 1 - \frac{\text{SSError}}{\text{SSTotal}}$$

- The SSError always decreases as more predictors are added to the model.
- R^2 always increases and can be artificially large.
- Cows: R^2 from model 2 is necessarily higher than R^2 from model 1. What can we say about R^2 from models 1 and 3?

Additional Sum-of-Squares principle

- ANOVA F-test, to compare two nested models: a “full” and a “reduced” model.
- we used it to test a single predictor.
- can be used to test multiple predictors at a time.

Example:
reduced: has $k = 1$ coefficient (other than intercept)

$$\mathrm{fat}_i = \beta_0 + \beta_1 * \mathrm{level}_i + e_i$$

full: has $p =$ coefficients other than intercept

$$\mathrm{fat}_i = \beta_0 + \beta_1 * \mathrm{level}_i + \beta_2 * \mathrm{initial.weight}_i + \beta_3 * \mathrm{lactation}_i + \beta_4 * \mathrm{age}_i + e_i$$

Additional Sum-of-Squares principle

- Fit “full” model: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \cdots + \beta_p x_{ip} + e_i$. Obtain $\mathrm{SSE}_{(\mathrm{full})}$ from the ANOVA:

Source	df	SS
Regression	p	$\mathrm{SSR}_{(\mathrm{full})}$
Error	$n-p-1$	$\mathrm{SSE}_{(\mathrm{full})}$
Total	$n-1$	SSTot

- Fit “reduced” model: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i$. Obtain $\mathrm{SSE}_{(\mathrm{reduced})}$ from the ANOVA:

Source	df	SS
Regression	k	$\mathrm{SSR}_{(\mathrm{reduced})}$
Error	$n-k-1$	$\mathrm{SSE}_{(\mathrm{reduced})}$
Total	$n-1$	SSTot

Example

```
> full = lm(fat ˜ level+initial.weight+lactation+age, data=cow
> reduced = lm(fat ˜ level, data=cow)
> anova(full)
> anova(reduced)
```

Source	df	SS
Regression	4	3.547
Error	45	7.952
Total	49	11.499

Source	df	SS
Regression	1	2.452
Error	48	9.047
Total	49	11.499

Additional Sum-of-Squares principle

Compute the “additional sum of squares” as

$$\mathrm{SSR}_{\text{(full)}} - \mathrm{SSR}_{\text{(reduced)}} = \mathrm{SSE}_{\text{(reduced)}} - \mathrm{SSE}_{\text{(full)}}$$

which is always ≥ 0, on $\mathrm{df} = p - k = (n - p - 1) - (n - k - 1)$

F-test

if the reduced model is true, then

$$F = \frac{(\mathrm{SSE}_{\text{(reduced)}} - \mathrm{SSE}_{\text{(full)}})/(p - k)}{(\mathrm{SSE}_{\text{(full)}})/(n - p - 1)} \sim F_{p-k,n-p-1}.$$

An f-test is used to test the reduced (H_0) versus the full (H_a) model.

Hypotheses: $e_i \sim$ normal distribution, are independent, and have homogeneous variance.

Example

Source	df	SS
Regression	4	3.547
Error	45	7.952
Total	49	11.499

Source	df	SS
Regression	1	2.452
Error	48	9.047
Total	49	11.499

So $F =$ $= 2.0651$ on df $=$ 3 and 45. Then $p = 0.12$.

```
> anova(reduced, full)
Model 1: fat ~ level
Model 2: fat ~ level + initial.weight + lactation + age
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 9.0469
2     45 7.9521  3    1.0948 2.0651 0.1182
```

Sequential testing

Often, there are *many* models we want to consider. Example: There are $2^5 = 16$ models equal or nested within each of these:

```
fat ~ initial.weight+lactation+age+treatment
fat ~ initial.weight+lactation+age+level
```

We may not analyze them all!

Various ways to do model selection:

- Many criteria: p-value from F-test, Adjusted R^2, AIC, etc.
- Different ways to search: backward elimination, forward selection, stepwise selection.

Backward elimination

1. fit the full model with all the predictors
2. find the predictor with the smallest f-value / t-value or largest associated p-value
   - if its p-value is above some threshold, go to step 3.
   - if not, keep the corresponding predictor and stop.
3. delete the predictor, re-fit the model and go to step 2.

Note: a threshold of $p > .05$ is often used, which corresponds approximately to $|t| < 2$ or $f < 4$.

There are multiple tests being done... The Bonferroni idea is rarely used, because it is overly conservative. Every term might be removed.

```
> drop1(full, test="F")
fat ˜ level + initial.weight + lactation + age
              Df Sum of Sq     RSS     AIC F value    Pr(F)
<none>                       7.952 -81.929
level          1      2.078  10.030 -72.324 11.7567 0.001308 **
initial.weight 1      0.086   8.038 -83.394  0.4845 0.489987
lactation      1      0.497   8.449 -80.898  2.8126 0.100463
age            1      0.302   8.254 -82.065  1.7091 0.197746

> newfit = update(full, . ˜ . - initial.weight)
> drop1(newfit, test="F")
fat ˜ level + lactation + age
          Df Sum of Sq     RSS     AIC F value    Pr(F)
<none>                   8.038 -83.394
level      1      2.211  10.249 -73.243 12.6541 0.000882 ***
lactation  1      0.487   8.525 -82.453  2.7869 0.101829
age        1      0.229   8.267 -83.990  1.3098 0.258357

> newfit = update(newfit, . ˜ . - age)
> drop1(newfit, test="F")
fat ˜ level + lactation
          Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                   8.267 -83.990
level      1      2.546  10.813 -72.565 14.4756 0.0004094 ***
lactation  1      0.780   9.047 -81.480  4.4365 0.0405448 *
```

Forward selection

1. fit the most simple model, using only predictors you want to force in the model, not matter what. Also prepare a list of candidate predictors.
2. find the predictor with the largest f-value / t-value or smallest associated p-value
    - if its p-value is below some threshold, go to step 3.
    - if not, stop. Do not add the predictor to the final model.
3. Add the predictor, re-fit the model and go to step 2.

Note: a threshold of $p < .05$ is often used, which corresponds approximately to $|t| > 2$ or $f > 4$.

There are multiple tests being done...

```
> basic = lm(fat ˜ 1, data=cow)
> add1(basic, test="F",
       scope = ˜initial.weight+lactation+age∗level)
fat ˜ 1
               Df Sum of Sq    RSS     AIC F value     Pr(F)
<none>                      11.499 -71.488
initial.weight  1     0.566 10.933 -72.011  2.4841 0.1215677
lactation       1     0.686 10.813 -72.565  3.0470 0.0872835
age             1     0.352 11.147 -71.043  1.5163 0.2241734
level           1     2.452  9.047 -81.480 13.0101 0.0007363

> newfit = update(basic, . ˜ . + level)
> add1(newfit, test="F",
       scope = ˜initial.weight+lactation+age∗level)
...
> newfit = update(newfit, . ˜ . + lactation)
> add1(newfit, test="F",
       scope = ˜initial.weight+lactation+age∗level)
fat ˜ level + lactation
               Df Sum of Sq    RSS     AIC F value  Pr(F)
<none>                       8.267 -83.990
initial.weight  1     0.012  8.254 -82.065  0.0694 0.7934
age             1     0.229  8.038 -83.394  1.3098 0.2584
```

Stepwise selection

- start with some model, simple or complex
- do a forward step as well as a backward step
- until no predictor should be added, and no predictor should be removed.

```
> library(MASS)

> best1 = stepAIC(full, test="F",
                  scope=~ initial.weight+lactation+age*level)
> best2 = stepAIC(basic, test="F",
                  scope=~ initial.weight+lactation+age*level)
...
Step:  AIC=-83.99
fat ~ level + lactation

                Df SumofSq     RSS     AIC F Value   Pr(F)
<none>                       8.267 -83.990
+ age            1   0.229   8.038 -83.394   1.310 0.25835
+ initial.weight 1   0.012   8.254 -82.065   0.069 0.79338
- lactation      1   0.780   9.047 -81.480   4.437 0.04054 *
- level          1   2.546  10.813 -72.565  14.476 0.00040 ***
```

Warnings

- Forward selection, backward selection, stepwise selection can all miss an optimal model. Forward selection has the potential of ’stopping short’.
- They may not agree.
- No adjustment for multiple testing... It is important to start with a model that is not too large, guided by biological sense.
- They can only compare nested models.

The adjusted R^2

Recall $R^2 = \frac{\text{SSRegression}}{\text{SSTotal}} = 1 - \frac{\text{SSError}}{\text{SSTotal}}$ always increases and can be artificially large.

Adjusted R^2

$$\text{adj}R^2 = 1 - \frac{\text{MSError}}{\text{SSTotal}/(n-1)} = 1 - \frac{n-1}{n-1-k}(1-R^2)$$

where k is the number of coefficients (other than the intercept). It is penalized version of R^2. The more complex the model, the highest the penalty.

- As k goes up, R^2 increases but $n-1-k$ decreases.
- adjusted R^2 may decrease when the added predictors do not improve the fit.
- MSError and adjusted R^2 are equivalent for choosing among models.

The adjusted R^2

Example: predict fat percentage using level and lactation.
$R^2 = 0.28$, MSError$= 0.42\%$, $n = 50$ cows and $k =$
adj$R^2 = \qquad\qquad\qquad\qquad = 0.25$

Another example:

```
> summary(lm(fat ~ treatment*age + initial.weight, data=cow))
Residual standard error: 0.4362 on 41 degrees of freedom
Multiple R-squared: 0.3215,     Adjusted R-squared: 0.1891

> summary(lm(fat ~ level + lactation, data=cow))
Residual standard error: 0.4194 on 47 degrees of freedom
Multiple R-squared: 0.2811,     Adjusted R-squared: 0.2505
```

- Are these two models nested?
- Which model would be preferred, based on adjusted R^2? based on MSError?

Likelihood

The likelihood of a particular value of a parameter is the probability of obtaining the observed data if the parameter had that value. It measures how well the data supports that particular value.

Example: tiny wasp are given the choice between two female cabbage white butterfly. One of them recently mated (so had eggs to be parasitized), the other not.

$n = 32$ wasps, $y = 23$ chose the mated female. Let
$p =$ proportion of wasps in the population that would make the good choice.

Likelihood of $p = 0.5$, as if the wasps have no clue?

Log-likelihood

Likelihood of $p = 0.5$, as if the wasps have no clue: $L(p = 0.5|Y = 23) = \mathbb{P}\{Y = 23|p = 0.5\} = 0.0065$ from Binomial formula:

$$L(p) = \binom{32}{23} p^{23}(1-p)^9$$

Most often, it is easier to work with the log of the likelihood:

$$\begin{aligned}\log L(p|Y = 23) &= \log\left(\binom{32}{23} p^{23}(1-p)^9\right) \\ &= \log\binom{32}{23} + 23\log(p) + 9\log(1-p)\end{aligned}$$

and $\log L(0.5) = \log(0.0065) = -5.031$

Maximum likelihood

The maximum likelihood estimate of a parameter is the value of the parameter for which the probability of obtaining the observed data if the highest. It's our best estimate.

- Sometimes there are analytical formulas, which coincide with other estimation methods.
- Many times we find the maximum likelihood numerically

Finding the maximum likelihood numerically

```
> dbinom(23, size=32, p=0.5)
[1] 0.00653062
> lik = function(p){ dbinom(23, size=32, p=p)}
> lik(0.5)
[1] 0.00653062
> log(lik(0.5))
[1] -5.031253
> lik(0.2)
[1] 3.158014e-10
> log(lik(0.2))
[1] -21.87591

> 23/32
[1] 0.71875
> lik(0.72)
[1] 0.1552330
> log(lik(0.72))
[1] -1.862828

> pp=seq(0.2,0.9,by=.01)
> ll=log(lik(pp))
> pp
> ll
> plot(pp,log(ll), type="l")
> abline(v=0.72)
```

Likelihood ratio test

Idea: if $p = 0.5$ is false, then the likelihood of $p = 0.5$ will be much lower than the maximum likelihood, the ratio $\frac{L(\hat{p})}{L(0.5)}$ will be large, i.e. the difference in log-likelihoods will be large: $\log L(\hat{p}) - \log L(0.5)$.

LRT to test $\alpha = \alpha_0$

- Test statistic: $X^2 = 2 * (\log L(\hat{\alpha}) - \log L(\alpha_0))$
- Null distribution: if H_0: $\alpha = \alpha_0$ is true then X^2 has a chi-square distribution approximately, with df=# of parameters in α.

Here we want to test H_0: $p = 0.5$.
$x^2 = 2 * (-1.86) - 2 * (-5.03) = 6.337$ on df$=1$ here. We get $p = 0.012$: strong evidence that $p \neq 0.5$.

Likelihood ratio test for `dose` and `weight`

LRT of H_0: $\beta_{\text{dose}} = 0$, after accounting for a weight effect:

```
> drop1(fit.dw, test="Chisq")
Single term deletions
Model: toxicity ˜ dose + weight
       Df Sum of Sq      RSS      AIC   Pr(Chi)
<none>               0.034738 -113.783
dose    1  0.030761 0.065499 -103.733  0.000518 ***
weight  1  0.085629 0.120367  -92.171 1.179e-06 ***
```

$-2 * L(\hat{\beta}_{\text{dose}} = 0) + 2 * L(\hat{\beta}_{\text{dose}}) = -103.733 + 113.783 + 2 = 12.05$ and $\mathbb{P}\{X^2_{\text{df}=1} > 12.05\} = 0.000518$.

Compare with the f-test based on SS:

```
> drop1(fit.dw, test="F")
Single term deletions
Model: toxicity ˜ dose + weight
       Df Sum of Sq      RSS      AIC F value     Pr(F)
<none>               0.034738 -113.783
dose    1  0.030761 0.065499 -103.733  14.168  0.001697 **
weight  1  0.085629 0.120367  -92.171  39.440 1.097e-05 ***
```

AIC: the Akaike criterion

- Model fit (R^2) always improves with model complexity. We would like to strike a good balance between model fit and model simplicity.
- AIC combines a measure of model fit with a measure of model complexity: The smaller, the better.

Akaike Information Criterion

For a given data set and a given model,

$$\text{AIC} = -2 \log L + 2p$$

where L is the maximum *likelihood* of the data using the model, and p is the number of parameters in the model.

- Here, $-2 \log L$ is a function of the prediction error: the smaller, the better. Measures how the model fits the data.
- $2p$ penalizes complex models: the smaller, the better.

AIC: the Akaike criterion

Strategy

Consider a number of candidate models. They need not be nested. Calculate their AIC. Choose the model(s) with the smallest AIC.

- Theoretically: AIC aims to estimate the prediction accuracy of the model for new data sets. Up to a constant.
- The absolute value of AIC is meaningless. The relative AIC values, between models, is meaningful.
- Often there are too many models, we cannot get all the AIC values. We can use stepwise selection.

Stepwise selection with AIC

Look for a model with the smallest AIC:

- start with some model, simple or complex
- do a forward step as well as a backward step based on AIC
- until no predictor should be added, and no predictor should be removed.

```
> library(MASS)
> stepAIC(basic,scope= ˜ initial.weight+lactation+age*level)
Step:  AIC=-83.99
fat ˜ level + lactation
                 Df Sum of Sq     RSS     AIC
<none>                          8.267 -83.990
+ age             1     0.229   8.038 -83.394
+ initial.weight  1     0.012   8.254 -82.065
- lactation       1     0.780   9.047 -81.480
- level           1     2.546  10.813 -72.565

> fullt = lm(fat ˜ treatment+initial.weight+lactation+age,
             data=cow)
> stepAIC(fullt,
        scope= ˜ initial.weight+lactation+age*treatment+level)
...
Step:  AIC=-80.76
fat ˜ treatment + lactation
                 Df Sum of Sq     RSS     AIC
<none>                          8.141 -80.755
+ age             1     0.256   7.885 -80.353
+ initial.weight  1     0.002   8.139 -78.766
- lactation       1     0.686   8.827 -78.710
- treatment       3     2.672  10.813 -72.565
```

BIC: the Bayesian information criterion

For standard models,

$$\text{BIC} = -2\log L + \log(n) * p$$

p is the # of parameters in the model, n is the sample size.

- Theoretically: BIC aims to approximate the posterior probability of the model. Up to a constant.
- The absolute value of BIC is meaningless. The relative BIC values, between models, is meaningful.
  The smaller, the better.
- The penalty in BIC is stronger than in AIC: AIC tends to select more complex models, BIC tends to select simpler models.
- In very simplified terms: AIC is better when the purpose is to make predictions. BIC is better when the purpose is to decide what terms truly are in the model.

BIC: the Bayesian information criterion

In R: use the option `k=log(n)` and plug-in the correct sample size n. Then remember the output is really about BIC (not AIC).

```
> stepAIC(full, scope=~ initial.weight+lactation+age*level,
                k=log(50))
...
Step:  AIC=-78.25
fat ~ level + lactation

                 Df Sum of Sq    RSS     AIC
<none>                         8.267 -78.254
- lactation       1     0.780  9.047 -77.656
+ age             1     0.229  8.038 -75.746
+ initial.weight  1     0.012  8.254 -74.417
- level           1     2.546 10.813 -68.741
```

Model selection: recap

- We can use p-values if models are nested. Or adjusted R^2 (or MSError) or information criteria like AIC or BIC.
- When there are too many candidate models, we can do a stepwise search for the best model(s).
- To describe the method, indicate both
  - the search criterion (F-test, LRT, adjusted R^2, AIC, etc.)
  - the search method (exhaustive (!), forward, backward, both)

- Use simple models. Do not start with an overly complex model: danger of data dredging and spurious relationships. Use biological knowledge to start with a sensible model.
- Sometimes there is no single “best” model. There may not be enough information in the data to tell what the truth is exactly.