Model Selection

R Club

Jacinta Kong

16/2/2022

Choosing the best model

The best model depends on your question/intention

- Find the best descriptor of your data? Regardless of predictors
- Test effect of a predictor on response? Or overall model variation?
- Cases when P values are inappropriate?

How many predictors?

> *Underfitting is bad but adding more variables comes at a cost: overfitting.*

Difference between dredging for significant predictors and making pre-determined comparisons for hypothesis testing.
What is your hypothesis?

Consider:

- Confounding variables
- Covariates
- Simpson's paradox

Simpson's Paradox

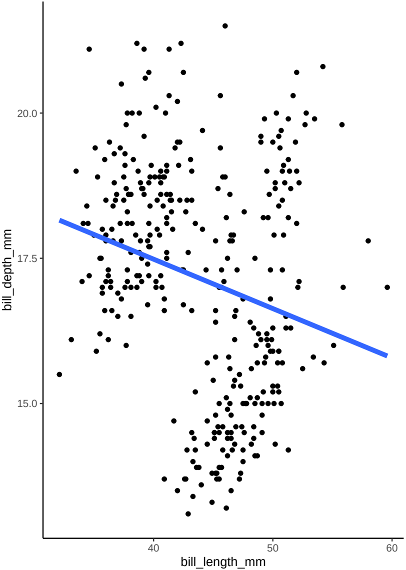

Figure 1: Reversal of correlation

Philosophical basis of model selection

***Parsimony**: Explain the most variation in Y using the fewest terms (variables) possible*

- Trade-off precision, generality, realism
- Not more parameters than observations
- Cannot explain *all* variation

Full vs reduced models

Full:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Additive:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Reduced:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Dropping variables

One sample t-test on **additional effect of β coefficient** on Y

- H0: effect $= 0$
- H1: effect $\neq 0$

If coefficient does not significantly explain more variation, drop it.
Check `summary`.

Non-significant interaction

```
Call:
lm(formula = Height ~ Wr.Hnd * Sex, data = MASS::survey)

Residuals:
    Min      1Q  Median      3Q     Max
-17.285  -5.037   0.978   4.274  19.807

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    147.4497     9.1625  16.093   <2e-16 ***
Wr.Hnd           1.0385     0.5203   1.996   0.0473 *
SexMale         -7.1567    12.2915  -0.582   0.5610
Wr.Hnd:SexMale   0.9020     0.6627   1.361   0.1750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.973 on 203 degrees of freedom
  (30 observations deleted due to missingness)
Multiple R-squared:  0.5107,    Adjusted R-squared:  0.5035
F-statistic: 70.63 on 3 and 203 DF,  p-value: < 2.2e-16
```

ANOVA with non-significant interaction

```
Analysis of Variance Table

Response: Height
            Df Sum Sq Mean Sq  F value    Pr(>F)
Wr.Hnd       1 7298.7  7298.7 150.1286 < 2.2e-16 ***
Sex          1 2912.4  2912.4  59.9052 4.604e-13 ***
Wr.Hnd:Sex   1   90.1    90.1   1.8526     0.175
Residuals  203 9869.1    48.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Drop interaction: t-test changes

```
Call:
lm(formula = Height ~ Wr.Hnd + Sex, data = MASS::survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.7479  -4.1830   0.7749   4.6665  21.9253 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 137.6870     5.7131  24.100  < 2e-16 ***
Wr.Hnd        1.5944     0.3229   4.937 1.64e-06 ***
SexMale       9.4898     1.2287   7.724 5.00e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.987 on 204 degrees of freedom
  (30 observations deleted due to missingness)
Multiple R-squared:  0.5062,	Adjusted R-squared:  0.5014 
F-statistic: 104.6 on 2 and 204 DF,  p-value: < 2.2e-16
```

Reduced model ANOVA

```
Analysis of Variance Table

Response: Height
           Df Sum Sq Mean Sq F value    Pr(>F)
Wr.Hnd      1 7298.7  7298.7 149.504 < 2.2e-16 ***
Sex         1 2912.4  2912.4  59.656 4.998e-13 ***
Residuals 204 9959.2    48.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactive model does not explain more varation than more parsimonious additive model

- SS of predictors does not change
- Variation added to residuals

Additive model is more parsimonious

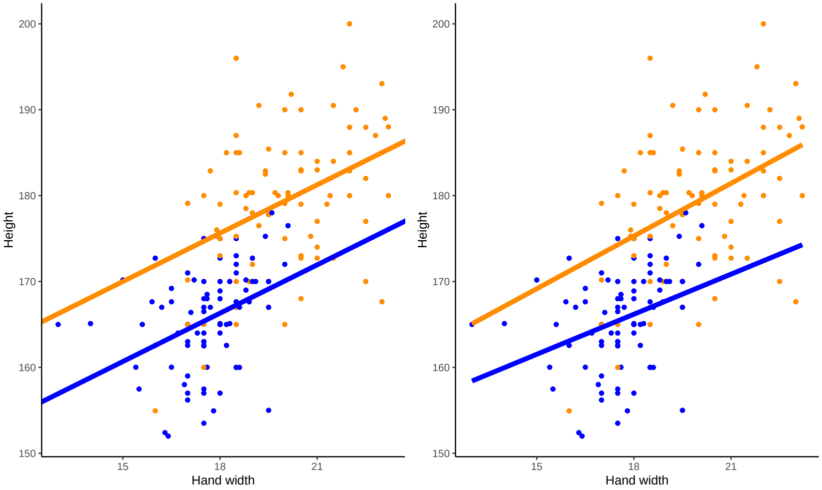

Figure 2: Additive (left) and interactive (right) model for hand span and height for male (orange) and female (blue) students

Analysis of Deviance

- More formal test of dropping variables
  - Likelihood ratio test (Goodness of Fit)
  - e.g. with and without predictor
- Compare test statistics between two models

```
two_pred <- lm(Y ~ X1 + X2, data) # Full model
one_pred <- lm(Y ~ X1, data) # Reduced model
anova(two_pred, one_pred, test = "Chisq")
```

H0: No difference between models - pick reduced model
H1: Additional term explains significant variation - pick full model

MASS::survey Analysis of Deviance

```r
mod_int <- lm(Height ~ Wr.Hnd * Sex, survey)
mod_add <- lm(Height ~ Wr.Hnd + Sex, survey)
anova(mod_int, mod_add, test = "Chisq")
```

```
Analysis of Variance Table

Model 1: Height ~ Wr.Hnd * Sex
Model 2: Height ~ Wr.Hnd + Sex
  Res.Df    RSS Df Sum of Sq Pr(>Chi)
1    164 7266.6
2    165 7287.3 -1   -20.634    0.495
```

Same conclusion as before: Choose additive model.
Can do the same thing for `Height ~ Wr.Hnd` vs `Height ~ Wr.Hnd + Sex`.

The problem with model selection with P values

- Dropping variables solely based on P values is error prone in more complex models
  - e.g. Mixed effects model where estimating P values is uncertain
- Alternative to use information theoretic approach
  - AIC/BIC
  - No P values, not hypothesis testing, no null model, models are not "rejected"
  - Allows model averaging - take weighted average of estimated parameters from set of models

AIC/BIC

Akaike's Information Criterion: More formal quantification of model prediction error based on log-likelihood method of parameter estimation.

- Penalise for number of terms - parsimony
  - Smaller AIC = better fit

Bayesian Information Criterion: Similar to AIC but stronger penalty for parameters

Information theoretic approach for inferences

- Compare set of models fitted to same dataset - rank equal candidate models
- Ideally models represent **alternative hypotheses**
- **AIC weights** = relative likelihood model is best model in set
  - Proportion: 0 to 1
  - Higher value = better model

Practical considerations

- Cannot compare too many models at once - spurious choice
- Influenced by small sample sizes (use second order AIC: AICc)
- "Best model" $\neq$ true model
  - Depends on sample from true population

Additive model is parsimonious

```r
mod_one <- lm(Height ~ Wr.Hnd, survey)
model_names <- c("mod_int", "mod_add", "mod_one")
mod_AIC <- c(AIC(mod_int), # full model
             AIC(mod_add), # additive model
             AIC(mod_one)) # reduced model
mid_weights <- MuMIn::Weights(mod_AIC)
results1 <- as.data.frame(cbind(model_names, round(mod_AIC ,3),
                                round(mid_weights, 3)))
names(results1) <- c("Model", "AIC", "wi")
knitr::kable(results1)
```

Model	AIC	wi
mod_int	1119.634	0.318
mod_add	1118.11	0.682
mod_one	1161.436	0

Another example, same result

- Does AICc by default
- Sorted best to worst

Additive model is $0.682/0.318 = 2.1$ times more likely to be the best model.

```r
MuMIn::model.sel(mod_int, mod_add, mod_one, rank = AIC)[,c(-5)]
```

```
Model selection table
        (Int) Sex Wr.Hnd Sex:Wr.Hnd df   logLik    AIC delta weight
mod_add 132.5   +  1.876             4 -555.055 1118.1  0.00  0.682
mod_int 138.2   +  1.555          +  5 -554.817 1119.6  1.52  0.318
mod_one 109.0      3.378             3 -577.718 1161.4 43.33  0.000
Models ranked by AIC(x)
```

Data dredging

Only use for exploratory analyses

- Automatic model selection from a full model based on AIC
- No *a priori* models/hypotheses
- Fixed effects only

Two ways for stepwise selection:

1. Forwards
   - Adding terms
2. Backwards
   - Removing terms (e.g. above)

`MASS::stepAIC` or `MuMIn::dredge`

MASS::stepAIC

Does `forward`, `backward` or `both`. No missing data.

```
full_model <- lm(Y ~., data)
# . fits all predictors without interactions
step_model <- stepAIC(full_model,
                      direction = "both",
                      trace = FALSE)
summary(step_model)
```

For fully crossed model: `lm(Y ~ (.)^2, data)`

MASS::survey

11 predictor variables.

$$
\begin{aligned}
\text{Height} = {} & \beta_0 + \beta_1(\text{Sex}_{\text{Male}}) + \beta_2(\text{Wr. Hnd}) + \beta_3(\text{NW. Hnd}) + \\
& \beta_4(\text{W. Hnd}_{\text{Right}}) + \beta_5(\text{Fold}_{\text{Neither}}) + \beta_6(\text{Fold}_{\text{R on L}}) + \beta_7(\text{Pulse}) + \\
& \beta_8(\text{Clap}_{\text{Neither}}) + \beta_9(\text{Clap}_{\text{Right}}) + \beta_{10}(\text{Exer}_{\text{None}}) + \beta_{11}(\text{Exer}_{\text{Some}}) + \\
& \beta_{12}(\text{Smoke}_{\text{Never}}) + \beta_{13}(\text{Smoke}_{\text{Occas}}) + \beta_{14}(\text{Smoke}_{\text{Regul}}) + \beta_{15}(\text{M. I}_{\text{Metric}}) \\
& \beta_{16}(\text{Age}) + \epsilon
\end{aligned} \tag{1}
$$

Stepwise MASS::survey

```
Call:
lm(formula = Height ~ Sex + Wr.Hnd + NW.Hnd + Clap + Exer, data = surve

Residuals:
     Min       1Q   Median       3Q      Max 
-18.8384  -3.8184   0.8951   3.8444  17.6725 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 137.8541     6.0041  22.960  < 2e-16 ***
SexMale       9.5747     1.2922   7.410  6.9e-12 ***
Wr.Hnd        3.3089     0.9865   3.354 0.000994 ***
NW.Hnd       -1.5229     0.9744  -1.563 0.120050    
ClapNeither  -1.8885     1.6529  -1.143 0.254926    
ClapRight    -2.9877     1.3711  -2.179 0.030788 *  
ExerNone     -5.2955     1.8543  -2.856 0.004863 ** 
ExerSome     -2.7728     1.0676  -2.597 0.010272 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.378 on 160 degrees of freedom
Multiple R-squared:  0.6071,     Adjusted R-squared:  0.5899
```

Comparing models after stepwise selection

$$\begin{aligned}\text{Height} = \beta_0 + \beta_1(\text{Sex}_{\text{Male}}) + \beta_2(\text{Wr. Hnd}) + \beta_3(\text{NW. Hnd}) + \\ \beta_4(\text{W. Hnd}_{\text{Right}}) + \beta_5(\text{Fold}_{\text{Neither}}) + \beta_6(\text{Fold}_{\text{R on L}}) + \beta_7(\text{Pulse}) + \\ \beta_8(\text{Clap}_{\text{Neither}}) + \beta_9(\text{Clap}_{\text{Right}}) + \beta_{10}(\text{Exer}_{\text{None}}) + \beta_{11}(\text{Exer}_{\text{Some}}) + \\ \beta_{12}(\text{Smoke}_{\text{Never}}) + \beta_{13}(\text{Smoke}_{\text{Occas}}) + \beta_{14}(\text{Smoke}_{\text{Regul}}) + \beta_{15}(\text{M. I}_{\text{Metric}}) \\ \beta_{16}(\text{Age}) + \epsilon\end{aligned} \tag{2}$$

$$\begin{aligned}\text{Height} = \beta_0 + \beta_1(\text{Sex}_{\text{Male}}) + \beta_2(\text{Wr. Hnd}) + \beta_3(\text{NW. Hnd}) + \\ \beta_4(\text{Clap}_{\text{Neither}}) + \beta_5(\text{Clap}_{\text{Right}}) + \beta_6(\text{Exer}_{\text{None}}) + \beta_7(\text{Exer}_{\text{Some}}) + \\ \epsilon\end{aligned} \tag{3}$$

MuMIn::dredge

- Backwards
- Need to change how R handles missing values
- Shows *all* possible combinations

```
options(na.action = "na.fail")
# change missing values behaviour
dd <- MuMIn::dredge(full_model)
summary(MuMIn::get.models(dd, 1)[[1]]) # get best model
```

Dredging MASS::survey

```
Call:
lm(formula = Height ~ Age + Exer + Sex + Wr.Hnd + 1, data = survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.3220  -3.5480   0.8529   3.7239  17.8312 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 137.31124    6.03757  22.743  < 2e-16 ***
Age          -0.13581    0.08202  -1.656  0.09970 .  
ExerNone     -4.85879    1.85983  -2.612  0.00983 ** 
ExerSome     -3.09869    1.05542  -2.936  0.00381 ** 
SexMale       9.06311    1.27571   7.104 3.63e-11 ***
Wr.Hnd        1.86605    0.33129   5.633 7.67e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.419 on 162 degrees of freedom
Multiple R-squared:  0.597, Adjusted R-squared:  0.5846 
F-statistic:    48 on 5 and 162 DF,  p-value: < 2.2e-16
```

Comparing models after dredging

$$\begin{aligned}\text{Height} = {} & \beta_0 + \beta_1(\text{Sex}_{\text{Male}}) + \beta_2(\text{Wr. Hnd}) + \beta_3(\text{NW. Hnd}) + \\ & \beta_4(\text{W. Hnd}_{\text{Right}}) + \beta_5(\text{Fold}_{\text{Neither}}) + \beta_6(\text{Fold}_{\text{R on L}}) + \beta_7(\text{Pulse}) + \\ & \beta_8(\text{Clap}_{\text{Neither}}) + \beta_9(\text{Clap}_{\text{Right}}) + \beta_{10}(\text{Exer}_{\text{None}}) + \beta_{11}(\text{Exer}_{\text{Some}}) + \\ & \beta_{12}(\text{Smoke}_{\text{Never}}) + \beta_{13}(\text{Smoke}_{\text{Occas}}) + \beta_{14}(\text{Smoke}_{\text{Regul}}) + \beta_{15}(\text{M. I}_{\text{Metric}}) \\ & \beta_{16}(\text{Age}) + \epsilon\end{aligned} \tag{4}$$

$$\begin{aligned}\text{Height} = {} & \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Exer}_{\text{None}}) + \beta_3(\text{Exer}_{\text{Some}}) + \\ & \beta_4(\text{Sex}_{\text{Male}}) + \beta_5(\text{Wr. Hnd}) + \epsilon\end{aligned} \tag{5}$$

Summary: Caution

Automatic model selection (dredging) is risky from a modelling philosophy perspective

- Not hypothesis driven
  - Ensure model is sensible and meaningful
  - Discarding biologically relevant variables?
- Is the process justified?
  - Not an unbiased process - P-value fishing?
  - E.g. exploratory analyses

Chance of spurious “best” model - Think properly about data!

Other approaches

- Ridge or lasso regression - weighted regressions
- Principle Component Analysis (PCA)
- Multivariate multiple regression ($\geq$ 2 response variables)