Сравнение линейных моделей

Линейные модели, осень 2014

Марина Варфоломеева
Каф. Зоологии беспозвоночных, СПбГУ

Сравнение линейных моделей

- Зачем нужно сравнивать модели?
- Принципы выбора лучшей линейной модели
- Тестирование гипотез при помощи сравнения линейных моделей
- Сравнение моделей по качеству подгонки к данным*
- Сравнение предсказательной силы линейных моделей с использованием кросс-валидации

Вы сможете

- Объяснить связь между качеством описания существующих данных и краткостью модели
- Объяснить, что такое "переобучение" модели
- Рассказать, каким образом происходит кросс-валидация моделей
- Протестировать влияние отдельных параметров линейной регрессии при помощи сравнения вложенных моделей
- Подобрать модель с оптимальной точностью подгонки к данным, оцененной по коэффициенту детерминации с поправкой или по C_p Маллоу
- Оценить предсказательную силу модели при помощи k-кратной кросс-валидации
-

Зачем нужно сравнивать модели?

Пример: птицы в лесах Австралии

От каких характеристик лесного участка зависит обилие птиц в лесах юго-западной Виктории, Австралия (Loyn, 1987)

56 лесных участков:

- ABUND - обилие птиц
- YR.ISOL - год изоляции участка
- GRAZE - пастбищная нагрузка (1-5)
- ALT - высота над уровнем моря
- L10DIST - логарифм расстояния до ближайшего леса
- L10LDIST - логарифм расстояния до ближайшего большого леса
- L10AREA - логарифм площади

```
birds <- read.csv("loyn.csv")
```

Mystic Forest - Warburton, Victoria by ¡kuba! on flickr

Нужна оптимальная модель

От каких характеристик лесного участка зависит обилие птиц в лесах юго-западной Виктории, Австралия (Loyn, 1987)

Переменных много, хотим из них выбрать **оптимальный небольшой** набор:

- При помощи разных критериев подберем несколько подходящих кандидатов
- Выберем лучшую модель с небольшим числом параметров

"Essentially, all models are wrong, but some are useful"
(Georg E. P. Box)

Принципы выбора лучшей линейной модели

Принципы выбора лучшей модели

Эти критерии конкурируют друг с другом

Хорошее описание существующих данных

Если мы включим много переменных, то лучше опишем данные

Стандартные ошибки параметров будут большие, интерпретация сложная

Большой R^2, маленький MSe

Парсимония

Минимальный набор переменных, который может объяснить существующие данные

Стандартные ошибки параметров будут низкие, интерпретация простая

Компромисс при подборе оптимальной модели:

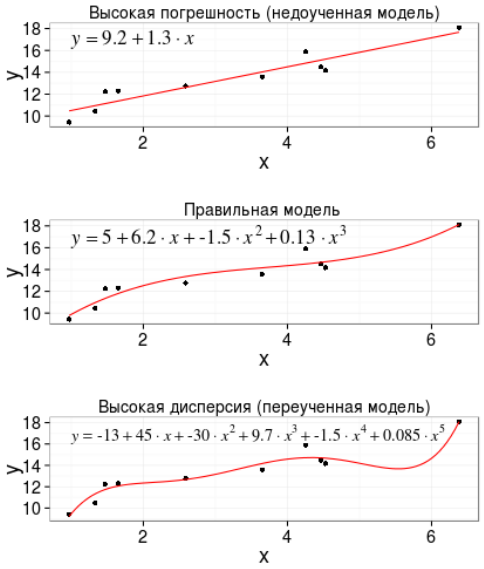
точность / смещенная оценка

Переобучение

Переобучение происходит, когда модель, из-за избыточного усложнения, описывает не только отношения между переменными, но и случайный шум

При увеличении числа предикторов в модели (при ее усложнении), она точнее опишет данные, по которым подобрана, но на новых данных точность предсказаний будет низкой из-за "переобучения" (overfitting).

Легче всего проиллюстрировать на примере полиномиальной регрессии



Критерии и методы выбора моделей зависят от задачи

Объяснение закономерностей

- Тестирование гипотез о влиянии факторов или удаление влияния одних переменных, для изучения других
- Нужны точные тесты влияния предикторов: F-тесты (о нем сейчас) или likelihood-ratio тесты

Описание закономерностей

- Описание функциональной зависимости между зависимой переменной и предикторами
- Нужна точность оценки параметров и парсимония: C_p Маллоу, "информационные" критерии (AIC, BIC, AICc, QAIC, и т.д.)

Предсказание

- Предсказание значений зависимой переменной для **новых** данных
- Нужна оценка качества модели на новых данных с использованием кросс-валидации (о ней сейчас)

Не позволяйте компьютеру думать за вас!

Дополнительные критерии для сравнения моделей:

- Диагностические признаки и качество подгонки:
  - остатки, автокорреляция, кросс-корреляция, распределение ошибок, выбросы и проч.
- Посторонние теоритические соображения:
  - разумность, целесообразность модели, простота, ценность выводов

Тестирование гипотез при помощи сравнения линейных моделей

Для тестирования гипотез о влиянии фактора можно сравнить модели с этим фактором и без него.

- Можно сравнивать для тестирования гипотез только вложенные модели (справедливо для F-критерия и для likelihood-ratio тестов)

Вложенные модели (nested models)

Две модели являются вложенными, если одну из них можно получить из другой, приравнивая некоторые коэффициенты более сложной модели к 0.

Какие из этих моделей вложены и в какие именно?

Полная модель (full model)

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

Неполные модели (reduced models)

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$$

$$y_i = \beta_0 + \beta_2 x_2 + \epsilon_i$$

Нулевая модель (null model)

$$y_i = \beta_0 + \epsilon_i$$

- Неполные модели являются вложенными по отношению к полной модели, нулевая модель - вложенная по отношению к полной и к неполным.
- Неполные модели по отношению друг к другу - **не** вложенные

Задание:

Запишите все вложенные модели для данной полной модели

(1) $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i$

- Модели:
  - (2) $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$
  - (3) $y_i = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon_i$
  - (4) $y_i = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_i$
  - (5) $y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$
  - (6) $y_i = \beta_0 + \beta_2 x_2 + \epsilon_i$
  - (7) $y_i = \beta_0 + \beta_3 x_3 + \epsilon_i$
  - (8) $y_i = \beta_0 + \epsilon_i$

- Вложенность:
  - (2)-(4) - вложены в (1)
  - (5)-(7) - вложены в (1), при этом
    - (5) вложена в (1), (2), (3);
    - (6) вложена в (1), (2), (4);
    - (7) вложена в (1), (3), (4)
  - (8) - нулевая модель - вложена во все

Сравнение линейных моделей при помощи F-критерия

Полная модель

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \ldots + \beta_p x_{ip} + \epsilon_i$$

$$df_{reduced,full} = p,\ df_{error,full} = n - p - 1$$

Уменьшенная модель

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

$$df_{reduced,reduced} = k,\ df_{error,reduced} = n - k - 1$$

F-критерий для сравнения моделей

Есть ли выигрыш от включения фактора в модель?

$$F = \frac{(SS_{error,reduced} - SS_{error,full})/(df_{reduced,full} - df_{reduced,reduced})}{(SS_{error,full})/df_{error,full}}$$

Задание:

- Запишите формулу модели, которая описывает, как зависит обилие птиц в лесах Австралии (ABUND) от переменных:
  - YR.ISOL - год изоляции участка
  - GRAZE - пастбищная нагрузка (1-5)
  - ALT - высота над уровнем моря
  - L10DIST - логарифм расстояния до ближайшего леса
  - L10LDIST - логарифм расстояния до ближайшего большого леса
  - L10AREA - логарифм площади

```
frm_full <-
```

- Подберите модель, используя эту формулу
- Какие переменные можно протестировать на предмет возможности исключения из модели?

Решение

L10DIST, L10LDIST, YR.ISOL не влияют

```
frm_full <- ABUND ~ L10AREA + L10DIST + YR.ISOL + L10LDIST + GRAZE
lm_full <- lm(frm_full, birds)
summary(lm_full)
```

```
#
# Call:
# lm(formula = frm_full, data = birds)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -16.368  -3.521   0.527   2.637  14.794
#
# Coefficients:
#             Estimate Std. Error t value  Pr(>|t|)
# (Intercept) -111.7820    89.7814   -1.25     0.219
# L10AREA        7.7557     1.4175    5.47 0.0000014 ***
# L10DIST       -1.2567     2.6321   -0.48     0.635
# YR.ISOL        0.0694     0.0447    1.55     0.127
# L10LDIST      -1.1065     2.0399   -0.54     0.590
# GRAZE         -1.8843     0.8881   -2.12     0.039 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 6.36 on 50 degrees of freedom
```

Сравнение линейных моделей при помощи (частного) F-критерия

функция `anova(модель_1, модель_2)` в R

Модели обязательно должны быть вложенными!

Протестируем, нужны ли переменные L10LDIST, L10DIST, YR.ISOL

Переменные, удаление которых **не ухудшает** модель, можно будет удалить и получить минимальную осмысленную модель (не термин:)

Тестируем L10LDIST

```
frm_ldist <- ABUND ~ L10AREA + L10DIST + YR.ISOL + GRAZE
lm_ldist <- lm(frm_ldist, birds)
anova(lm_ldist, lm_full)
```

```
# Analysis of Variance Table
#
# Model 1: ABUND ~ L10AREA + L10DIST + YR.ISOL + GRAZE
# Model 2: ABUND ~ L10AREA + L10DIST + YR.ISOL + L10LDIST + GRAZE
#   Res.Df  RSS Df Sum of Sq    F Pr(>F)
# 1     51 2036
# 2     50 2024  1      11.9 0.29   0.59
```

- L10LDIST не улучшает модель - выбрасываем

Тестируем L10DIST, при условии, что L10LDIST уже нет в модели

```
frm_dist <- ABUND ~ L10AREA + YR.ISOL + GRAZE
lm_dist <- lm(frm_dist, birds)
anova(lm_dist, lm_ldist)
```

```
# Analysis of Variance Table
#
# Model 1: ABUND ~ L10AREA + YR.ISOL + GRAZE
# Model 2: ABUND ~ L10AREA + L10DIST + YR.ISOL + GRAZE
#   Res.Df  RSS Df Sum of Sq    F Pr(>F)
# 1     52 2071
# 2     51 2036  1      35.4 0.89   0.35
```

- L10DIST не улучшает модель - выбрасываем

Тестируем YR.ISOL, при условии, что L10DIST и L10LDIST нет в модели

```
frm_yrisol <- ABUND ~ L10AREA + GRAZE
lm_yrisol <- lm(frm_yrisol, birds)
anova(lm_yrisol, lm_dist)
```

```
# Analysis of Variance Table
#
# Model 1: ABUND ~ L10AREA + GRAZE
# Model 2: ABUND ~ L10AREA + YR.ISOL + GRAZE
#   Res.Df  RSS Df Sum of Sq    F Pr(>F)
# 1     53 2201
# 2     52 2071  1       130 3.26  0.077 .
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- L10LDIST не улучшает модель - выбрасываем

А вот GRAZE выкинуть не получится

```
frm_graze <- ABUND ~ L10AREA
lm_graze <- lm(frm_graze, birds)
anova(lm_graze, lm_dist)
```

```
# Analysis of Variance Table
#
# Model 1: ABUND ~ L10AREA
# Model 2: ABUND ~ L10AREA + YR.ISOL + GRAZE
#   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
# 1     54 2867
# 2     52 2071  2       796 9.99 0.00021 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- GRAZE улучшает модель - нужно оставить

Минимальная модель

```
frm_yrisol
```

```
# ABUND ~ L10AREA + GRAZE
```

Сравнение моделей по качеству подгонки к данным*

в этой лекции - без информационных критериев

Коэффициент детерминации

Обычный коэффициент детерминации оценивает долю объясненной изменчивости

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

$R^2_{adjusted}$

Доля объясненной изменчивости с поправкой на число предикторов

$$R^2_{adjusted} = 1 - (1 - R^2)\,\frac{n-1}{n-k} \leq R^2$$

n - число наблюдений,
k - количество параметров в модели

У хорошей модели будет большой $R^2_{adjusted}$

C_p Мэллоу (Mallow's C_p)

Оценивает "общую ошибку предсказания" с использованием *p*-параметров

$$C_p = \frac{SS_{error,p-predictors}}{MS_{error,full}} - (n - 2p)$$

C_p Мэллоу связан с F-критерием

$$C_p = p + (F_p - 1)(m + 1 - p)$$

m - общее число возможных параметров
p - число параметров в уменьшенной модели

У хорошей модели $C_p \approx p$

- Если нет ошибки предсказания, то $F_p \approx 1$ и $C_p \approx p$
- Если есть ошибка предсказания, то $F_p > 1$ и $C_p > p$

Найдем лучшую из всех моделей по коэффициенту детерминации

```
library(leaps)
crit_ar2 <- leaps(x = birds[, c(3, 6:10)], y = birds$ABUND,
                  names = names(birds[, c(3, 6:10)]),
                  method = "adjr2")
# crit_ar2$size # число предикторов
# crit_ar2$which # предикторы в модели
# crit_ar2$adjr2 # R^2 adj. для модели

# Номер строки лучшей модели (модели с макс. adjr2)
best_ar2 <- which.max(crit_ar2$adjr2)
# Какие переменные входят в модель?
crit_ar2$which[best_ar2, ]
```

```
#  YR.ISOL    GRAZE      ALT  L10DIST L10LDIST  L10AREA
#     TRUE     TRUE     TRUE    FALSE    FALSE     TRUE
```

```
# Записываем формулу лучшей модели по adjr2
frm_ar2 <- ABUND ~ YR.ISOL + GRAZE + ALT + L10AREA
```

В нашем случае переменных немного, можем перебрать все модели кандидаты. Если переменных много, можно использовать пошаговые процедуры (опасно) или тестировать несколько осмысленных кандидатов

Задание:

Выберите лучшую модель по значению C_p Маллоу

Нужно изменить параметр `method` в функции `leaps()`

У лучшей модели будет минимальным модуль разницы между ее числом параметров и C_p

Решение

```
crit_cp <- leaps(x = birds[, c(3, 6:10)], y = birds$ABUND, names = names(birds[, c(3, 6:10)]), method = "
# Ищем лучшую модель
# полную модель нужно исключить
n_mod <- length(crit_cp$size)
best_cp <- which.min(abs(crit_cp$size[-n_mod] - crit_cp$Cp[-n_mod]))
# Какие переменные входят в модель?
crit_cp$which[best_cp, ]
```

```
#  YR.ISOL    GRAZE      ALT  L10DIST L10LDIST  L10AREA
#    FALSE     TRUE    FALSE    FALSE     TRUE     TRUE
```

```
frm_cp <- ABUND ~ GRAZE + L10DIST + L10AREA
```

Какую из моделей выбрать, если мы хотим предсказывать с их помощью

Теперь у нас есть три многообещающих модели кандидата

```
frm_yrisol
```

```
# ABUND ~ L10AREA + GRAZE
```

```
frm_ar2
```

```
# ABUND ~ YR.ISOL + GRAZE + ALT + L10AREA
```

```
frm_cp
```

```
# ABUND ~ GRAZE + L10DIST + L10AREA
```

Оценим их предсказательную силу

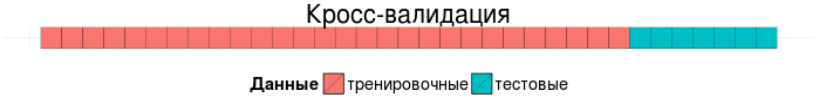
Сравнение предсказательной силы линейных моделей с использованием кросс-валидации

Кросс-валидация

Если оценивать качество модели по тем же данным, по которым она была подобрана, оценки будут завышенными

Кросс-валидация решает эту проблему

Делим данные **случайным образом** на **тренировочное и тестовое подмножества**, обычно в пропорции 60:40, 70:30 или 80:20



Тренировочные данные

Используются для подбора модели (для обучения)

Чтобы модель была хорошей, тренировочных данных **должно быть много**

Тестовые данные

Используются для оценки качества модели

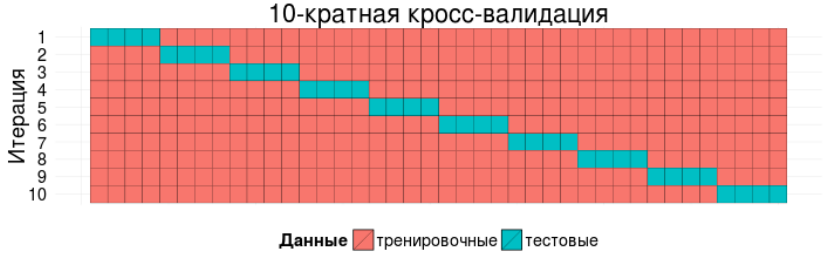
Чтобы надежно оценить качество модели, тестовых данных **тоже должно быть много**

K-кратная кросс-валидация (k-fold cross-validation)

Делим данные **случайным образом** на k частей

$k-1$ часть используется для обучения, на k-й части тестируется модель

Процедура повторяется k раз



k-кратная кросс-валидация лучше обычной, особенно, если данных не много

RMSE - стандартная ошибка предсказания

$$RMSE = \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{n}}$$

Это параметр, который определяет ширину доверительных интервалов предсказаний

Очень чувствительна к выборосам (альтернатива - MAE - средний модуль ошибок)

Можно сравнивать между моделями, только если они в одинаковых единицах (исходные данные моделей (не)преобразованы одинаково, зависимая переменная в одних и тех же единицах)

Нет жестких границ для RMSE "хорошей" модели, это относительная величина.

Бывает, что критерии противоречат друг другу, тогда решаем с учетом других соображений, например, простоты и интерпретируемости. Лучше меньше параметров.

Этапы сравнения моделей с использованием кросс-валидации

- Делим данные на тренировочное и тестовое подмножества
- Для каждой из моделей-кандидатов повторяем следующие шаги
  - Подбираем на тренировочном подмножестве модель-кандидат
  - Используя тестовые данные, предсказываем ожидаемые значение y используя модель-кандидат
  - Рассчитываем RMSE для модели-кандидата (стандартное отклонение остатков)

$$RMSE = \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{n}}$$

- Сравниваем RMSE всех моделей кандидатов. Модель, у которой минимальное значение RMSE - лучшая

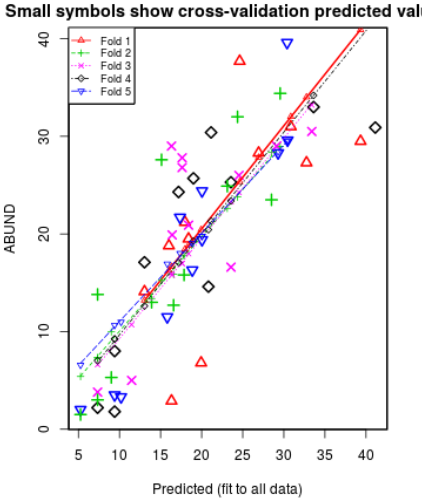
Кросс-валидация для линейных моделей

CVlm(df = исходные_данные, form.lm = формула, m = кратность)

```
library(DAAG)
val_yrisol <- CVlm(df = birds, form.lm = frm_yrisol, m = 5)
```

```
# Analysis of Variance Table
#
# Response: ABUND
#           Df Sum Sq Mean Sq F value  Pr(>F)
# L10AREA    1   3471    3471    83.6 1.8e-12 ***
# GRAZE      1    666     666    16.0 0.00019 ***
# Residuals 53   2201      42
# ---
# Signif. codes:
# 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#
# fold 1
# Observations in test set: 11
#                  6    13    18     25    28    31
# Predicted   13.042 17.97  16.3 18.404 16.01  19.9
# cvpred      13.211 18.45  16.7 18.829 16.25  20.4
# ABUND       14.100 21.20   2.9 19.500 18.80   6.8
# CV residual  0.889  2.75 -13.8  0.671  2.55 -13.6
#                 45    47    50    52    55
```



Задание:

Посчитайте RMSE для модели val_yrisol

$$RMSE = \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_i$ - это предсказанные во время кросс-валидации значения - `val_yrisol$cvpred`

y_i - это реальные наблюдаемые значения зависимой переменной - `val_yrisol$ABUND`

Решение

```
# RMSE вручную
sqrt(mean((val_yrisol$cvpred - val_yrisol$ABUND)^2))
```

```
# [1] 6.5
```

Можно создать пользовательскую функцию для рассчета RMSE

```
rmse <- function(cv_obj, y_name){
  sqrt(mean((cv_obj$cvpred - cv_obj[, y_name])^2))
}

# теперь можно пользоваться функцией
rmse(val_yrisol, "ABUND")
```

```
# [1] 6.5
```

Задание

- Сделайте 5-кратную кросс-валидацию оставшихся двух моделей и полной модели

```
frm_cp
frm_ar2
frm_full
```

- Посчитайте их RMSE

Какая из моделей-кандидатов дает более качественные предсказания?

Решение

Кросс-валидация

```
val_cp <- CVlm(df = birds, form.lm = frm_cp, m = 5)
val_ar2 <- CVlm(df = birds, form.lm = frm_ar2, m = 5)
val_full <- CVlm(df = birds, form.lm = frm_full, m = 5)
```

Считаем RMSE

```
rmse(val_cp, "ABUND")
```

```
# [1] 6.5
```

```
rmse(val_ar2, "ABUND")
```

```
# [1] 6.84
```

```
rmse(val_full, "ABUND")
```

```
# [1] 6.93
```

Какие модели дают более качественные предсказания?

```
rmse(val_yrisol, "ABUND"); rmse(val_cp, "ABUND")
```

```
# [1] 6.5
```

```
# [1] 6.5
```

```
rmse(val_ar2, "ABUND"); rmse(val_full, "ABUND")
```

```
# [1] 6.84
```

```
# [1] 6.93
```

Судя по значениям RMSE, это модели

```
frm_yrisol
```

```
# ABUND ~ L10AREA + GRAZE
```

```
frm_cp
```

```
# ABUND ~ GRAZE + L10DIST + L10AREA
```

Takehome messages

- Модели, которые качественно описывают существующие данные включают много параметров, но предсказания с их помощью менее точны из-за переобучения
- Для выбора оптимальной модели используются разные критерии в зависимости от задачи
  - Сравнивая вложенные модели можно отбраковать переменные, включение которых в модель не улучшает ее
  - Оптимальный набор переменных для более качественного описания **существующих данных** можно подобрать сравнивая модели по $R^2_{adjusted}$ и C_p Маллоу
  - Оценить предсказательную силу модели на **новых данных** можно при помощи кросс-валидации сравнив ошибки предсказаний

Дополнительные ресурсы

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. Springer.

- 2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability
- 2.2.2 The Bias-Variance Trade-Off
- 3.2.2 Some Important Questions

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer.

- 1.1 Prediction Versus Interpretation
- 1.2 Key Ingredients of Predictive Models
- 4 Over-Fitting and Model Tuning
- 5 Measuring Performance in Regression Models

Quinn, G.G.P., Keough, M.J., 2002. Experimental design and data analysis for biologists. Cambridge University Press.

- 6.1.15 Finding the “best” regression model
- 6.1.16 Hierarchical partitioning