MT5753 - Week 2 Practical

2017

1 Objectives for this practical:

In this practical we are going to investigate the nature of any impacts using the `EIA` data by:

- fitting a multiple linear model
- carrying out stepwise model selection and all-possible subsets selection
- interpreting model results
- assessing the predictive power of the model
- identifying and assessing model assumptions
- improve your model using a generalised least squares framework

注意这个和课件的区别

需要很久，至少几小时

---

2 Investigating impacts using the EIA data

We are going to use multiple linear regression to model the EIA data. Specifically, we will use the variables provided (including `impact`) to model animal density (`density`) in the `EIA`. Note: if you did not save your data from practical one you will need to re-make the `density` column.

Use **both** the `step` function and the `dredge` function (in the `MuMIn` library) to carry out model selection for your data and be sure you assess model assumptions and understand if they are valid in this case.

---

3 Assessment

- This practical is submitted via a Moodle quiz.
- Marks allocated for each question are indicated inside square brackets. There are a total of 30 marks available.
- **This practical is to be submitted by 12.00pm (noon) Monday 6th November 2017**.
- This practical counts for 10% of your course grade.

4 Analysis of EIA data

Start by loading the EIA data into your workspace. If you did not save your data from practical one, the you will need to re-make the density column. Remember not to attach several EIA datasets to your workspace. If you are unsure, use the `search()` function to see what is in your workspace.

---

1. State whether `density`, `hour`, `day`, `month`, `impact` and `gridcode` are discrete ordinal, discrete nominal or continuous. [2]

Explore the relationship between each predictor and `density`. Do you think that linearity is reasonable for these variables? Try using the `qplot` function (from the `ggplot2` library) to make some plots.

As an aside, here is how you might plot the geo-spatial data using ggplot2:

```
# first group the data by gridcodes and find the mean density for each cell
require(dplyr)
newdata<-group_by(EIA, GridCode)%>%
  summarise(x.pos=first(x.pos), y.pos=first(y.pos), area=first(area), density=mean(density))
# pick a nice colour scheme
col<-colorRampPalette(rev(rgb(c(231,117,27),c(41,112,158),c(138,179,119),max=255)))(100)
# plot the data
p<-ggplot(newdata)
p<-p + geom_tile(aes(x=x.pos, y=y.pos, fill=density, height=1000, width=1000)) +
  scale_fill_gradientn(colours=col, space="Lab", na.value="grey50", guide="colourbar")
p + theme_bw() + coord_equal()
```

---

2. Which of the following statements relating to factor variables is **FALSE**? [1]

a) We don't need to assume linearity between factor variables and the response

b) A different coefficient is estimated for each (non-baseline) factor level

c) We can predict between factor levels

d) More coefficients are estimated when we use factors

4.1 Fitting Multiple Covariate Linear Models

In order to assess which covariates may play a role in determining the density of animals, begin by fitting the two multiple covariate linear models below.

```
# month as continuous
fit.full<- lm(density ~ tidestate + observationhour + DayOfMonth +
                MonthOfYear + impact + Year + x.pos + y.pos, data=EIA)

# month as a factor
fit.full.fac<- lm(density ~ tidestate + observationhour + DayOfMonth +
                  as.factor(MonthOfYear) + impact + Year +
                  x.pos + y.pos, data=EIA)
```

Note: `gridcode` is omitted as it may be correlated with `x.pos` and `y.pos`. `Year` is not tried as a factor as it provides similar information as `impact` and so will likely be correlated too.

---

3. How many coefficients does changing month to a factor add to the model? [1]

4.2 Confidence Intervals

4. Calculate a 95% confidence interval for the `Year` coefficient using the `fit.full` model. Give your answers to two decimal places. [1]

4.3 Model performance

5. What is the adjusted R^2 value for the `fit.full` model? Give your answer to four decimal places. [1]

---

6. True or False? The adjusted R^2 for the `fit.full` model shows that the model covariates explain approximately 0.01% of the variance in the response variable. [1]

Use the `VIF` function on the `fit.full.fac` model, to assess collinearity and use AIC to choose the best model.

Use the `stepVIF` function from the `pedometrics` `R` library on the `fit.full.fac` model. Use the default threshold value for this function.

7. Which of the following about collinearity is FALSE? [1]

a) A confidence interval for the Year coefficient is over 48 times wider than it would be for a model with no collinear variables.

b) The GVIF is equivalent to the VIF but adjusted for multiple covariates.

c) It is appropriate to assess models with and without the collinear variables and use AIC score to choose the best model. In this instance the preferred model is one with impact removed.

d) The 'stepVIF' function identifies two collinear variables and uses R-squared to determine which variable to drop. In this case Year is dropped from the model.

e) Pairwise comparisons between covariates, such as scatter plots and covariance values may be used to assess collinearity prior to modelling.

4.4 Model Selection

In this section you will use hypothesis tests and information criteria to perform backwards, stepwise and all possible subsets selection.

Remove the impact covariate from the `fit.full.fac` model. Use this new model (`fit.fullfac.noimp`) to continue.

Use F-tests to decide which variables appear to be statistically significant in the model at this stage.

---

8. True or False? The null hypothesis for the F-test is that a model with a particular covariate included is no better than a model with that covariate removed. In this case, all variables except `DayOfMonth` appear to have significant relationships with density. [1]

Perform a stepwise automated selection using AIC on the model fitted to date (`fit.fullfac.noimp`). Use the `step` function with `direction = 'both'`

Now do all possible subsets selection using the `dredge` function and the default, AICc.

Note that the `dredge` function is in the `MuMIn` library and that you will need to run the code `options(na.action='na.fail')` before using the `dredge` function.

```
require(MuMIn)
options(na.action='na.fail')
dredge(fit.fullfac.noimp)
```

---

9. Which of the following about model selection is FALSE? [1]
   a) The best stepwise model is the same as that from all possible subsets
   b) The model with 'DayOfMonth' included is within 2 AIC points of the model with it not included (28% weight compared with ~70% weight).
   c) The best all possible subsets model contains the intercept, observation hour, x.pos, y.pos and year.
   d) Owing to the large sample size, it makes no difference to covariate selection whether we do all possible subsets selection using AIC or AICc.

---

10. True or False? Using BIC for stepwise selection does not change the covariates selected in the final model. [1]

Continue using the best model from AIC stepwise selection. Update this model to include interaction terms for both x.pos and y.pos with year.

Using the model with interaction terms, do F-tests to assess covariate significance and compare this with the output from all possible subsets (default settings) and forwards and backwards stepwise selection (AIC, direction = both) on the same model.

---

11. Which of the following statements about model selection is FALSE? [1]

a) Hypothesis testing (F-test) suggests that the model without the year:y.pos interaction is preferred to the full model.

b) Dredge, using AICc, suggests that there is little to choose between a model a) without either interaction term, b) with both interaction terms and c) with only the x.pos interaction term retained.

c) Forwards and backwards stepwise selection using AIC chooses the full model with the year:x.pos interaction term retained.

d) Forwards and backwards stepwise and all possible subsets selection using BIC return the same model as backwards selection using hypothesis testing.

Continue using the model returned from the stepwise AIC function:

---

12. Using an appropriate plot to investigate the similarity of the observed and fitted values from the stepwise AIC model, which one of the following statements is FALSE? [1]

a) The range of the data is not equivalent to the predictions.

b) There is little agreement between observed and fitted values.

c) The large values are over predicted.

d) For a perfect model fit, the points on an observed vs fitted plot should lie on the line 'abline(0,1)'.

---

13. Which combination of model covariates gives rise to the lowest animal density? [1]

a) year = 2012, tidestate='EBB', observation hour = 20, month = 9, low x and y positions.

b) year = 2009, tidestate='FLOOD', observation hour = 20, month = 1, low x and y positions.

c) year = 2010, tidestate='EBB', observation hour = 4, month = 12, high x and y-position.

d) year = 2009, tidestate='EBB', observation hour = 4, month = 1, high x and y-position.

e) year = 2012, tidestate='FLOOD', observation hour = 4, month = 9, low x and y positions.

---

14. Which one of the following would you use to assess predictive power of this model? [1]
    a) Median Residual
    b) *p*-value
    c) Adjusted R-Squared
    d) F-statistic
    e) Multiple R-Squared
    f) Residual Standard Error

---

15. What is the estimate of the error variance for this model? Give your answer to two decimal places. [1]

4.5 Linear Model Assumptions

16. Which of the following statements about assumption tests is FALSE? [1]
    a) The null hypothesis for 'shapiro.test' is that the errors are normally distributed
    b) The null hypothesis for 'ncvTest' is that there is non-constant error variance
    c) The null hypothesis for 'durbinWatsonTest' is that the errors are independent

---

17. Which of the following, about the validity of assumptions for the AIC-based stepwise selection model with interaction terms is TRUE? [1]
    a) Independence is probably not reasonable given the sampling design and the result of the independence assumption test. There is evidence of non-constant error variance (very small *p*-value) and non-normality (right skewed histogram)
    b) Independence is probably reasonable given the sampling design and the result of the independence assumption test. There is evidence of non-constant error variance (very small *p*-value) and non-normality (right skewed histogram)
    c) Independence is probably not reasonable given the sampling design and the result of the independence assumption test. There is evidence of constant error variance (very small *p*-value) and non-normality (right skewed histogram)

---

18. Fill in the *blanks* with the words 'critical' or 'not critical' in relation to the effect these assumptions have on our inference from a model. Assume the answers relate to a data set of a similar size to the `EIA` data. [2]
    a) Linearity for continuous variables is a *blank* assumption
    b) Independence of errors is a *blank* assumption
    c) Constant error variance is a *blank* assumption
    d) Normality of erorrs is a *blank* assumption

4.6 Generalised Least Squares Modelling

For this section we will use the square root of density as the response and investigate using the `impact` variable instead of year.

Fit two GLS models with an exponential mean-variance relationship and one with a power based mean-variance relationship.

Use the following code to help you:

```
require(nlme)
EIA$sqrtdensity<-sqrt(density)
fit.gls<-gls(sqrtdensity ~ tidestate + observationhour + impact + x.pos + y.pos +
               MonthOfYear + impact:x.pos + impact:y.pos, data = EIA, method='ML',
             weights=????)
```

---

19. True or False? The exponential model is a better representation of the error variance than a constant. [1]

Using the exponential model, make a plot of the observed mean-variance relationship and add the relationship estimated from the model.

- Bin the fitted values
- calculate the variance of the residuals in each bin.
- Plot the mean fitted value in each bin against the variance of residiuals in each bin.
- Add a line to show the estimated mean-variance relationship.

Use the following code to help you:

```
plot(fitted(fit.gls), residuals(fit.gls, type='response'))

cut.fit<-cut(fitted(fit.gls), breaks=quantile(fitted(fit.gls), probs=seq(0,1,length=20)))
means1<- tapply(fitted(fit.gls), cut.fit, mean)
```

---

20. Save your mean-variance plot and upload to Moodle. Use sensible axis labels and give your plot a title. Your file should be one of jpeg, png or pdf. [2]

---

21. Which of the following about the mean-variance relationship is FALSE? [1]

a) The exponential model slightly underestimates the variance for the smaller predicted root-density values

b) The exponential model severely overestimates the variance for the highest predicted root-density values

c) The residual variance suggests that a variance term that is allowed to both increase and decrease might be preferred.

4.7 Dealing with correlated errors

We have updated the mean-variance relationship but still find some correlation in model residuals. Use an acf plot to visualise this:

```
par(mfrow=c(1,2))
acf(residuals(fit.gls, type='response'))
acf(residuals(fit.gls, type='normalized'))
```

These acf plots should look identical as we have not dealt with any correlation yet.

Update your GLS model to include an AR(1) correlation matrix with gridcode/day as a blocking structure. Note that you will need to use the new dataset (created below) and that the fitting of GLS models may take several minutes. Use the following code to help you:

```
EIA$block<-paste(Year, MonthOfYear, DayOfMonth, GridCode, sep='')
require(dplyr)
EIA2<-arrange(EIA, block, Year, MonthOfYear, DayOfMonth, GridCode)
```

Having fitted an AR(1) model, also try an AR(2).

---

22. Which of the following about GLS models is FALSE? [1]
    a) The AR(2) model is the best model since the AIC score is the lowest
    b) We cannot use the AIC to choose between models with different covariates unless the models are fitted using maximum likelihood
    c) We can use AIC to choose between models with the same covariates but differing correlation structures if the models are fitted using REML.
    d) The normalized residual acf plot for the AR(2) confirms the AIC result, that the AR(2) model fits better the correlation in the residuals than the AR(1) model.
    e) The AR(2) model reduces the correlation at most lags to near zero and so has dealt with the issues regarding correlation in model residuals. We can expect that the standard errors for the estimated coefficients are of an appropriate size and we can trust any model selection results that use hypothesis testing.

Use hypothesis testing (F-tests) for backwards model selection and answer the following question.

---

23. After backwards selection using hypothesis testing, select all the variables that remain in your model. [1]
    a) tidestate
    b) observation hour
    c) month of the year
    d) x-position
    e) y-position
    f) impact
    g) x-position:impact
    h) y-position:impact

---

24. Make a prediction from your best model for both before and after impact using the relevant covariate values given below. Give your answers in density and to 2 decimal places. [1]

Use the following code to help you calculate your predictions as we will use the standard errors in the next question. The standard errors are only available from the `MuMIn` version of the `predict.gls` function and not from the `nlme` package.

```
myprediction<-MuMIn:::predict.gls(finalmodel, newdata = newdat, se.fit=TRUE)
```

a) tidestate = SLACK

b) observation hour = 10am

c) month of the year = 6

d) x-position = 1500

e) y-position = 1000

f) impact = 0 and 1

---

25. Calculate a 95% confidence interval for the **before impact** prediction from the previous question. Please give each of your answers as a density and to two decimal places. [1]

---

26. Calculate a 95% confidence interval for the **after impact** prediction. Please give each of your answers as a density and to two decimal places. [1]

---

27. Which of the following about summarising the models you have fitted is FALSE? [1]

a) The best linear model underestimated the variance and found significant covariates, owing to unaccounted for residual correlation and a poor model of the mean-variance relationship.

b) The addition of a more appropriate mean-variance relationship (exponential model) using a GLS model found that month of the year was no longer significant.

c) The inclusion of a model for the correlation in residuals removed several covariates, which would otherwise lead a researcher to come to the wrong conclusions.

d) It is unlikely that there are any further improvements to be made to this model now that all the assumptions are met.
