STAT6038 Week 9 Lecture Notes

Rui Qiu

2017-05-03

1 Wednesday's Lecture

1.1 Some data playaround

page 47 Example 5 body fat data

`pairs()` draws pairwise plots to check linearity.

Another thing you can do (play around with data) is to use `cor(bodyfat)` command to check correlation (which is basically the same thing).

The standardized residuals vs fitted values plots suggest some potential high leverage points on the left hand side.

The qq-plot(standardized residuals) looks fine. (3 most extreme points are 13, 14, 19 in this plot)

But the cook's distance plot suggests 1, 3, 14 have high leverages. But not very relatively high, with the highest not even over 0.5.

So the model is fine(?).

1.2 Standardized Residuals – two types

1. PRESS/standardized/internally studentised
    - r_i's in the book
    - `rstandard()` in R
    - used in the default `plot(model)` version of the Normal qq plot, but they could also be used in an improved version of the **main** residuals vs fitted values plot.
2. studentized/externally studentized
    - t_i's in the book
    - `rstudent()` in R
    - they can be assessed against a Student's t distribution with $(n-p)-1$ degrees of freedom

Body Fat Example - 14 observations, t_{15} 95% cuts off point ± 2.13

A test of $H_0 : \Delta_{\text{obs14}} = 0$, where Δ is the mean shift deviation due to observation 14.

Observed $t_{14} = +2.016 < 2.13$. Not reject the null hypothesis.

As $p = 0.062 = 2 \times 0.031 \not< \alpha = 0.05$, do not reject H_0.

2 Thursday's Lecture

2.1 Model Refinement: Added Variable Plots

The aim of multiple regression (MR) modelling is to use the X or explanatory variables/predictors to explain what is happening with the Y or response variable.

In analysis of variance terms we want to maximize the variance explained by the model (involving the X variables) and minimize the unexplained variance (in Y). In fact, we would like (and we assume) that this unexplained variance will be just random stochastic variability, i.e.

$$\epsilon \sim N(0, \sigma^2)$$

"the errors are **independent** and identically (**normally**) distributed with **constant variance**, σ^2. (*Homogeneity of Variance, a.k.a. Homoscedasticity*)

However, the residuals (observed error) from a multiple regression model often show non-random structure. (i.e. they are **not** independent) or it might be random, but simply too variable (the unexplained variance is large and the explained variance is relatively small and the overall F-test is not significant).

If we have other X variables (in the data but not in the model), we can try adding them to the model and see if they help explain the unwanted structure and/or some of the unexplained variations.

Added variable plots (AVP) are a way of assessing if, but also how, we might include another X variable in the model.

Note, if we don't have any other candidate X variables in the data (there are no other observed X variables) then the structure and/or excess variability is "unobserved heterogeneity".

AVP for a model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

and an additional variables X_{k+1}.

PLOTPLOTPLOTPLOTPLOTPLOTPLOT

Note: "SLR MODEL $\equiv$ model with a 0 intercept, k slope equal to a partial regression coefficient for β_{k+1} in an expanded model."

Here the vertical axis represents what is left in Y that is not explained by $X_1, \ldots, X_k$.

And the horizontal axis is what is in X_{k+1} that is not explained by $X_1, \ldots, \ldots, X_k$.

We can check if the relationship on the AVP is linear by adding a simple linear regression (SLR) line.

Or we could propose a different way to include X_{k+1} in the model, i.e. some transformation or higher order term if there is apparent relationship which is non-linear!

3 Friday's Lecture

Standardized residuals

- internally standardized residuals
- externally studentized residuals

Influence Measures:

1. DFFITS `dffits()`
2. DFBETAS `dfbetas()`
3. Cook's distance `cooks.distance()`
4. Covariance Ratio (COVRATIO) `covratio()`

Cook's distance can be seen as measuring the overall effect of the ith observation on parameters, whereas the DFBETAS measure the effects on each of the individual parameters.

...

Cook's distance is also a function of both the standard residual value and the leverage values, which is why values of Cook's D can be drawn as a series of curves on the default plot `plot(model, which=5)` in R, but the choice of 0.5 and 1 used in this plot are definitely arbitrary choices. (no real bases!)

...

COVRATIO is OUTSIDE the range $(1 - 3p/n, 1 + 3p/n)$ should be considered highly influential.

`influnece()` gives us **hat values, sigma (a vector whose i-th element contains the estimate of the residual standard deviation obtained when the i-th case is dropped from the regression.), weighted residuals**