SF2930 Regression analysis

Exercise session 4 - Ch. 6: Diagnostic for leverage and influence, Ch. 9: Multicollinearity

In class:

1. Montgomery et al., 6.10. Formally show that (the Cook's distance)

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}. \tag{1}$$

2. Assessing model diagnostic plots requires experience. Often it is difficult to decide whether a deviation is a systematic one (i.e. needing correction) or a random one (i.e. just variability in the data). Experience can be gained by performing model diagnostics on problems where it is known whether the model assumptions hold or do not hold. This allows us to identify the naturally occurring variability in the results.

   In the following we simulate one predictor `xx` and four responses `yy.a, yy.b, yy.c`, and `yy.d`.

```
> set.seed(21)
> n <- 100
> xx <- 1:n
> yy.a <- 2+1*xx+rnorm(n)
> yy.b <- 2+1*xx+rnorm(n)*(xx)
> yy.c <- 2+1*xx+rnorm(n)*(1+xx/n)
> yy.d <- cos(xx*pi/(n/2)) + rnorm(n)
```

   Fit four simple linear regression models using `xx` as the predictor.

   a) For each model, create a scatter plot with the regression line, plot the four standard residual plots and the plot containing Cook's distance (this is provided by `plot.lm()` with the argument `which = 1:5`). Decide for each model which of the standard regression assumptions are fulfilled and which ones are violated. Verify your claims with the construction of the responses.

   b) Instead of `plot.lm()` use the function `resplot()` which is available from the course webpage. The function `resplot()` uses resampling to visualize whether a model violation is present. How does the function perform for the four models?

c) Repeat generating the random numbers a few times (i.e. use different random seeds) and study the variation in the resulting plots. You can also change the number of observations and track the changes in the plots.

3. Montgomery et al., 6.1. Perform a thorough influence analysis of the solar thermal energy test data given in Table B.2. Discuss your results.

4. Montgomery et al., 9.6. Use the regressors x_2 (passing yards), x_7 (percentage of rushing plays) and x_8 (opponents yards rushing) for the National Football League data in Table B.1.

   a) Does the correlation matrix give any indication of multicollinearity?

   b) Calculate the variance inflation factor and the condition number of $\mathbf{X}'\mathbf{X}$. Is there any evidence for multicollinearity?

5. Montgomery et al., 9.10. Analyze the housing price data in Table B.4 for multicollinearity. Use the variance inflation factors and the condition number of $\mathbf{X}'\mathbf{X}$.

Recommended exercises:

Book	Theory	Implementation
Rawlings et al.:	11.3, 11.4	
Montgomery et al.:	6.11	6.2, 6.8, 9.11

References

D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 2012. ISBN 9780470542811.

J.O. Rawlings, S.G. Pantula, and D.A. Dickey. *Applied Regression Analysis: A Research Tool*. Springer Texts in Statistics. Springer New York, 2001. ISBN 9780387984544.