Model Diagnostics

Lorenz Wolf

January 2022

1 Introduction

During my MSc we used to have an assessed project every Monday. With a total of 9h to complete a project you can imagine that speed matters. Usually I would write some concise notes of my prep the day before with practical tips and references to useful sources. This is one of those on model diagnostics for normal linear models (NLM) - this is not supposed to be a complete guide but rather a 'checklist' with some key ideas and practical notes.

In the following let Y be the response, p the number of covariates, n the number of observations, and X the design matrix.

Note: I wrote this markdown very quickly, so notation is not very clear for now - might update it later

2 Coefficient of determination R^2

- A measure of the goodness of fit for a NLM
- Larger models with more parameters will have a smaller residual sum of squares (RSS)

For models with an intercept we have:

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

which means $0 \leq R^2 \leq 1$

- The R^2 can be interpreted as the proportion of variance in the data that is explained by the model

3 Adjusted R^2

$$\bar{R}^2 = 1 - \frac{RSS}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}\frac{n-1}{n-p}$$

- takes into account the number of model parameters
- can be negative
- tries to balance out the effect of decreasing RSS when increasing the number of parameters

4 Model Checking

- F-Test and T-test results (p-values) can not be trusted if model is not satisfying assumptions
- QQ-plot to check normality
- Y against x_i to check linearity of mean
- residual plots to check homoscedasticity
- partial regression (added variable plots) to investigate the effect of a particular predictor on the dependent variable while holding all other predictor variables constant - use **avPlots(lm)** from the car package
- partial residual plots for better detection of non-linearity, use **prplot(lm, variable)** from the faraway package

5 Outliers

Potential causes of outliers can be: errors in data recording, data is a mixture of different populations, a bad model. For 2 or only 1 variable outliers can easily be detected for example via: boxplots, scatterplots, residuals.

In general it is always sensible to consider several measures to identify potential outliers.

5.1 Residuals

- $e = Y - \hat{Y} = (I_n - P)Y$, where $P = X(X^T X)^{-1} X^T$
- $cov(e) = \sigma^2(I_n - P)$ and $E(e) = 0$ and it follows $e \sim N(0, \sigma^2(I_n - P))$
- **Standardized residuals** $\left(\frac{e_i}{\sqrt{\sigma^2(I_n-P)_{ii}}} \sim N(0,1) \text{ for } i = 1, ..., n\right)$ more effective for detecting outliers, since account for residuals having slightly different variances
- However, σ^2 unknown hence need estimate $\hat{\sigma}^2$
- **Studentized residuals** $\left(r_i := \frac{e_i}{\sqrt{\hat{\sigma}^2(I_n-P)_{ii}}} \text{ for } i = 1, ..., n\right)$ loose normality but should be approximately normal
- Plots of r_i against any other variable should not reveal any trends or patterns

5.2 Heteroscedastic errors

Non constant variance, can check this by a regression of for example residuals on fitted values ($\hat{Y}$).
Fix: transformation of variables, weighted least squares (downweight observations with high variance)

5.3 Leverages

$h_{ii} = 1 - P_{ii}$ is leverage of observation i.

- high leverage implies small variance
- look at observations with leverage $h_{ii} > \frac{2r}{n}$ where $r = rank(X)$ (this compares the leverage to the average of the leverages $\frac{r}{n}$

5.4 Deleted residuals

Deleted residuals are obtained by fitting the model without the ith observation.

- Fit model excluding ith observation
- predict ith observation, denote expectation by $\hat{Y}_{(i)}$
- the deleted residual is $d_i = Y_i - \hat{Y}_{(i)}$
- can be obtained without fitting a new model by $d_i = \frac{e_i}{1-h_{ii}}$
- estimated variance of d_i is $\frac{\hat{\sigma}^2_{(i)}}{1-h_{ii}}$, where $\hat{\sigma}^2_{(i)} = \frac{RSS}{n-1}$ (RSS with ith case omitted)

Studentized deleted residuals allow hypothesis testing.

$$t_i = \frac{d_i}{\sqrt{\hat{\sigma}^2_{(i)}/(1-h_{ii})}} = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}}, \quad t_i \sim t_{n-p-1}(\text{assuming correct model})$$

How to test whether an observation is an outlier? Let $T \sim t_{n-p-1}$ and $t_{\alpha/2}$ s.t. $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$. Then $\mathcal{I} = (-t_{\alpha/2}, t_{\alpha/2})$ is a $1 - \alpha$ confidence interval for t_i, thus if $t_i \notin \mathcal{I}$ have evidence to reject H_0: $t_i \sim t_{n-p-1}$.

5.5 Cook's distance

The Cook's distance

$$c_i = r_i^2 \frac{h_{ii}}{(1-h_{ii})r}, \quad r = rank(X)$$

combines the residual and the leverage in one measure and is crucial for outlier detection.

5.6 Some useful commands

- influence(model) where model is an lm object in R. This gives the leave one out change in parameters and $\hat{\sigma}^2_{(i)}$
- identify(.) for interactive point selection in scatter plot (Faraway p.77)

6 Multicollinearity

Multicollinearity is when $X^T X$ is close to singular. This leads to:

- imprecise least squares estimates $\hat{\beta}$
- inflated standard errors, which means that t-tests performed on the paramter estimates may fail to show significance
- the model will be sensitive to measurement errors (small change in y leads to large change in β)

How to detect multicollinearity:

- check the correlation matrix of the predictors for high correlations between them
- consider the variance inflation factors $\frac{1}{1-R_j^2}$, where R_j^2 is the R^2 value obtained by regressing predictor x_j on the other predictors. Note that $var(\hat{\beta}_j) = \sigma^2 \frac{1}{1-R_j^2} \frac{1}{\sum_i (X_i j - \bar{x}_j)^2}$
- to check the singularity of X check the condition number $\sqrt{\frac{\lambda_1}{\lambda_p}}$, use $eigen(X^T X)$ to obtain the eigen decomposition

7 Some useful pointers

- In R var.test(.) to test whether variances of two groups are equal, test statistic ratio of variances (Faraway p.61)
- Transformation or weighted least squares (Faraway p.62), e.g. $h(Y) = log(Y)$ or $h(Y) = \sqrt{Y}$
- long-tailed: Cauchy (Faraway p.65)
- Shapiro Wilk test for normality (Faraway p.66)

The page numbers refer to 'Linear Models with R' by J. Faraway, pdf available here:
http://www.utstat.toronto.edu/ brunner/books/LinearModelsWithR.pdf