Statistics 203: Introduction to Regression and Analysis of Variance

Multiple Linear Regression: Diagnostics

Jonathan Taylor

Today

- Splines + other bases.
- Diagnostics

Spline models

- Splines are piecewise polynomials functions, i.e. on an interval between "knots" (t_i, t_{i+1}) the spline $f(x)$ is polynomial but the coefficients change within each interval.
- Example: cubic spline with knows at $t_1 < t_2 < \cdots < t_h$

$$f(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{h} \beta_i (x - t_i)_+^3$$

where

$$(x - t_i)_+ = \begin{cases} x - t_i & \text{if } x - t_i \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Here is an example.
- Conditioning problem again: B-splines are used to keep the model subspace the same but have the design less ill-conditioned.
- Other bases one might use: Fourier: $\sin$ and $\cos$ waves; Wavelet: space/time localized basis for functions.

What are the assumptions?

- What is the full model for a given design matrix X ?

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{i,p-1} + \varepsilon_i$$

- Errors $\varepsilon \sim N(0, \sigma^2 I)$.
- What can go wrong?
  - Regression function can be wrong – missing predictors, nonlinear.
  - Assumptions about the errors can be wrong.
  - Outliers & influential observations: both in predictors and observations.

Problems in the regression function

- True regression function may have higher-order non-linear terms i.e. X_1^2 or even *interactions* $X_1 \cdot X_2$.
- How to fix? Difficult in general – we will look at two plots "added variable" plots and "partial residual" plots.

Partial residual plot

- For $1 \leq j \leq p-1$ let

$$e_{ij}^{*} = e_i + \widehat{\beta}_j X_{ij}.$$

- Can help to determine if variance depends on $\boldsymbol{X}_j$ and outliers.
- If there is a non-linear trend, it is evidence that linear is not sufficient.

Added-variable plot

- For $1 \leq j \leq p-1$ let $H_{(j)}$ be the Hat matrix with this predictor deleted. Plot

$$(I - H_{(j)})Y \, \text{vs.} (I - H_{(j)})X_j.$$

- Plot should be linear and slope should be β_j. Why?

$$\begin{aligned} Y &= X_{(j)}\beta_{(j)} + \beta_j X_j + \varepsilon \\ (I - H_{(j)})Y &= (I - H_{(j)})X_{(j)}\beta_{(j)} + \beta_j (I - H_{(j)})X_j + (I - H_{(j)})\varepsilon \\ (I - H_{(j)})Y &= \beta_j (I - H_{(j)})X_j + (I - H_{(j)})\varepsilon \end{aligned}$$

- Also can be helpful for detecting outliers.
- If there is a non-linear trend, it is evidence that linear is not sufficient.

Problems with the errors

- Errors may not be normally distributed. We will look at QQplot for a graphical check. May not effect inference in large samples.
- Variance may not be constant. Transformations can sometimes help correct this. Non-constant variance affects our estimates of $SE(\widehat{\beta})$ which can change t and F statistics substantially!
- Graphical checks of non-constant variance: added variable plots, partial residual plots, fitted vs. residual plots.
- Errors may not be independent. This can seriously affect our estimates of $SE(\widehat{\beta})$.

Outliers & Influence

- Some residuals may be much larger than others which can affect the overall fit of the model. This may be evidence of an outlier: a point where the model has very poor fit. This can be caused by many factors and such points should not be automatically deleted from the dataset.
- Even if an observation does not have a large residual, it can exert a strong *influence* on the regression function.
- General stragegy to measure influence: for each observation, drop it from the model and measure "how much does the model change"?

Dropping an observation

- A $\cdot_{(i)}$ indicates i-th observation was not used in fitting the model.
- For example: $\widehat{Y}_{j(i)}$ is the regression function evaluated at the j-th observations predictors BUT the coefficients $(\widehat{\beta}_{0,(i)}, \dots, \widehat{\beta}_{p-1,(i)})$ were fit after deleting i-th row of data.
- Basic idea: if $\widehat{Y}_{j(i)}$ is very different than $\widehat{Y}_j$ (using all the data) then i is an influential point for determining $\widehat{Y}_j$.

Different residuals

- Ordinary residuals: $e_i = Y_i - \widehat{Y}_i$
- Standardized residuals: $r_i = e_i/s(e_i) = e_i/\widehat{\sigma}\sqrt{1-H_{ii}}$, H is the "hat" matrix. (`rstandard`)
- Studentized residuals: $t_i = e_i/\widehat{\sigma_{(i)}}\sqrt{1-H_{ii}} \sim t_{n-p-1}$. (`rstudent`)

Crude outlier detection test

- If the studentized residuals are large: observation may be an outlier.
- Problem: if n is large, if we "threshold" at $t_{1-\alpha/2,n-p-1}$ we will get many outliers by chance even if model is correct.
- Solution: Bonferroni correction, threshold at $t_{1-\alpha/(2*n),n-p-1}$.

Bonferroni correction for multiple comparisons

- If we are doing many t (or other) tests, say $m > 1$ we can control overall false positive rate at α by testing each one at level α/m.
- Proof:

$$
\begin{aligned}
P\left(\text{at least one false positive}\right) \\
= P\left(\cup_{i=1}^m |T_i| \geq t_{1-\alpha/(2*m),n-p-2}\right) \\
\leq \sum_{i=1}^m P\left(|T_i| \geq t_{1-\alpha/(2*m),n-p-2}\right) \\
= \sum_{i=1}^m \frac{\alpha}{m} = \alpha.
\end{aligned}
$$

$DFFITS$

- $$DFFITS_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{\widehat{\sigma}_{(i)}\sqrt{H_{ii}}}$$
- This quantity measures how much the regression function changes at the i-th observation when the i-th variable is deleted.
- For small/medium datasets: value of 1 or greater is considered "suspicious". For large dataset: value of $2\sqrt{p/n}$.

Cook's distance

- $$D_i = \frac{\sum_{j=1}^n (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p\,\widehat{\sigma}^2}$$
- This quantity measures how much the entire regression function changes when the i-th variable is deleted.
- Should be comparable to $F_{p,n-p}$: if the "p-value" of D_i is 50 percent or more, then the i-th point is likely influential: investigate this point further.

$DFBETAS$

- $$DFBETAS_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\widehat{\sigma}^2_{(i)}(X^TX)^{-1}_{jj}}}.$$
- This quantity measures how much the coefficients change when the i-th variable is deleted.
- For small/medium datasets: value of 1 or greater is "suspicious". For large dataset: value of $2/\sqrt{n}$.
- Here is an example.