Contents

Model Diagnostics: An Overview

- Basic diagnostics, review
- Model adequacy for a predictor variable – added-variable plots
- Outlying Y observation and studentized/deleted residuals
- Outlying X observation and hat matrix/leverage values
- Influential cases
- Multicollinearity diagnostics and variance inflation factor

Model Assumptions

- Recall multiple linear regression model, for $i = 1, \ldots, n$

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim \text{iid } N(0, \sigma^2).$$

  - Relationship between Y and $\boldsymbol{X}$: $\mathbb{E}(Y_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}$.
  - Homogeneous variance: $\text{Var}(Y_i) = \text{Var}(\epsilon_i) \equiv \sigma^2$.
  - Independence: $\text{Cov}(\epsilon_i, \epsilon_j) = \text{Cov}(Y_i, Y_j) = 0, \quad i \neq j$.
  - Normal distribution: $Y_i \sim N(\beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}, \sigma^2)$.

Basic Diagnostics

- Exploratory Data Analysis
  - Same as before: scatterplots, boxplots, histograms, summaries
  - New: scatterplot matrices, split boxplots, brush/spin, "coplots"
- Linearity, Homoscedasticity, Normality
  - Same as before: (externally studentized) residuals vs. each X, against $\widehat{Y}$, and against time (also note: ACF plot), QQplot.
  - Tests: e.g., F test for lack of fit, Breusch-Pagan, etc. (see Chapter 6.8 KNNL)
- Outliers, Influence, and Correlated Predictors
  - Major focus of this set of notes

Residuals—Review

- Recall that the residuals $\boldsymbol{e} = (e_1, \ldots, e_n)^T = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, where $\boldsymbol{H}$ is the hat/projection matrix.
- The mean of the residuals is

  $\boldsymbol{e}\mathbf{1}^T =$

- The variance-covariance matrix of the residuals is

  $\text{Var}\{\boldsymbol{e}\} =$

  and is estimated by

  $s^2\{\boldsymbol{e}\} =$

Residuals—Review

- Denote $\boldsymbol{H} = [h_{ij}]_{i,j=1}^{n}$.
- Then we have variance of e_i

$$\text{Var}\{e_i\} = \sigma^2(1 - h_{ii}),$$

estimated by

$$s^2\{e_i\} = MSE(1 - h_{ii})$$

- The covariance of e_i and e_j $(i \neq j)$ is

$$\text{Cov}\{e_i, e_j\} = \sigma^2(0 - h_{ij}) = -\sigma^2 h_{ij}$$

estimated by

$$s^2\{e_i\} = -MSE \times h_{ij}.$$

Studentized Residuals—Review

- The variance of e_i is not constant and the covariance of e_i, e_j is not zero.
  - Observations with a large residual relatively to its standard deviation may be outlying.
- To compare n residuals, standardize so that the residuals are on the same scale.
  - Studentized residuals (a.k.a. internally studentized) are defined as

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}.$$

  - If the model is appropriate, the studentized residuals $\{r_i\}$ have constant variance, while the ordinary residuals $\{e_i\}$ do not.

Deleted Residuals—Review

- Influence: the ith point can pull the line response surface strongly toward it if it is highly influential. This masks the point's influence.
  - Strategy: define the residual for the ith point as the prediction error for that point using the model fit to the data omitting that point.
- Deleted residuals are defined as

$$d_i = Y_i - \hat{Y}_{i(-i)}.$$

- It can be shown that

$$d_i = Y_i - \hat{Y}_{i(-i)} = \frac{e_i}{1 - h_{ii}} = \frac{Y_i - \hat{Y}_i}{1 - h_{ii}}.$$

Deleted Residuals—Review

- Let $\boldsymbol{X}_i = (1, X_{i1}, \ldots, X_{i,p-1})$ (a row vector).
- Let $\boldsymbol{X}_{-i}$ and MSE_{-i} denote the design matrix and the MSE with the ith row (observation) deleted.
- Recall that $s^2\{\text{pred}\} = MSE(1 + \boldsymbol{X}_h(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_h^T)$, one can show

Studentized Deleted Residuals—Review

- The studentized deleted residuals (a.k.a. externally studentized) are defined as, for $i = 1, \ldots, n$,

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{\frac{e_i}{1-h_{ii}}}{\sqrt{\frac{MSE_{-i}}{1-h_{ii}}}} = \frac{e_i}{\sqrt{MSE_{-i}(1-h_{ii})}}.$$

- Note that $(n-p)MSE = (n-p-1)MSE_{-i} + \frac{e_i^2}{1-h_{ii}}$, so

and there is no need to fit n separate regressions.

Outlying Y Observation

- Outlying observations are well separated from the remainder of the data.
- Consider three types of outlying observations:
  1. Outlying not in X but in $Y|X$: Usually not influential.
  2. Outlying in X but not $Y|X$: Usually not influential.
  3. Outlying in X and $Y|X$: Can be very influential.
- Goal: Identify outlying and influential observations.
  - The task is relatively straightforward for 1-2 predictor variables but becomes more challenging for more than 2 predictor variables.
  - "Hidden Extrapolation".
- Basic idea: Outlying observations may involve large residuals and often have large impact on the model fit.

Identifying Outlying Y Observations

- Basic idea: the ith observation is outlying in Y if t_i is large.
- Under H_0 : observation i is not outlying in Y

$$t_i = \frac{d_i}{s\{d_i\}} \sim t_{n-p-1}.$$

- The decision rule is

- Need Bonferroni adjustment, why?
  - n multiple comparisons.
  - For most n and p, $t_{1-\frac{\alpha}{2n};n-p-1}$ at the $\alpha = 5\%$ level is greater than 3. In practice, $|t_i| > 3$ then observation i is a possible outlier.

Hat Matrix and Leverages

- Basic idea: use the hat matrix to identify outliers in X.
- Recall that $\boldsymbol{H} = [h_{ij}]_{i,j=1}^{n}$ and $h_{ii} = \boldsymbol{X}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i^T$.
  - The diagonal elements h_{ii} are called leverages.
  - Properties of leverages h_{ii}:
    1. $0 \leq h_{ii} \leq 1$ (can you show this?)
    2. $\sum_{i=1}^{n} h_{ii} = p \Rightarrow \bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$ (show it).
    3. h_{ii} is a measure of the distance between X values of the ith observation and the means of the X values for all n observations (show: $h_{ii} = 1/n + (x_{1i} - \bar{x}_1)^T(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}(x_{1i} - \bar{x}_1)$, where $\boldsymbol{X}_c$ is the centerred design matrix X.)

Identifying Outlying X Observation

- Effects of hat values: if the ith data point is outlying in X with a high leverage h_{ii}, it can influence the fitted response $\hat{Y}_i$.
  - A higher leverage h_{ii} results in more weight of Y_i in determining $\hat{Y}_i$ (as $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$).
  - A higher leverage h_{ii} results in a smaller $s\{e_i\}$, as $\hat{Y}_i$ is closer to Y_i.
  - Connections to nonparametric smoothing.
- What is a bad hat value?
  1. If $h_{ii} > 2p/n$, then observation i is considered to be outlying in X.
  2. Moderate leverage if $h_{ii} \in [0.2, 0.5)$ and high leverage if $h_{ii} \in [0.5, 1]$.
  3. Draw a histogram, stem-and-leaf, or other plot of h_{ii}. Outlying observations tend to be large and there tends to be a gap between the outlying group and other leverage values.

Hidden Extrapolation

- $\boldsymbol{H}$ can be used to identify hidden extrapolation for large p.
  - It is possible for a point $\boldsymbol{X}_{new}$ to have each $X_{new,i}$ $(i = 1, \ldots, p)$ within the corresponding marginal range of $\boldsymbol{X}$, but for the p-dim point $\boldsymbol{X}_{new}$ to lie outside the support region of the empirical joint distribution of $\boldsymbol{X}$.
  - Can be very difficult to detect, especially if no 2-way scatterplot or 3-way brush/spin illustrates it.
- Consider

$$h_{\text{new,new}} = \boldsymbol{X}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_{\text{new}}^T.$$

If $h_{\text{new,new}} \leq max_i h_{ii}$, then it is fine to make predictions at $\boldsymbol{X}_{\text{new}}$.

Identifying Influential Observations

- An observation is influential if its deletion leads to major changes in the fitted regression.
  - Not all outlying observations are influential.
  - Main idea: Leave-one-out approach like the deleted residuals.
- Consider 3 measures:
  1. DFFITS
  2. Cook's distance
  3. DFBETAS
- No diagnostics identify all possible problems. For example, leave-one-out methods do not address multiple influential observations.
- More complicated methods are possible: bootstrap, highd-dimensional situations.

DFFITS

- DFFITS measures the effect of the ith case on fitted value of Y_i

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{-i}}{\sqrt{MSE_{-i} h_{ii}}}$$

and we can show

$$DFFITS_i = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

where t_i is the ith studentized deleted residual.

- For small to medium data sets, $|DFFITS_i| > 1$ implies that the ith observation may be influential.
- For large data sets, $|DFFITS_i| > 2\sqrt{p/n}$ implies that the ith observation may be influential.

Cook's Distance

- Cook's distance measures the influence of the ith observation on all n fitted values.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(-i)})^2}{p \cdot MSE_{-i}}.$$

and show

$$D_i = \left(\frac{r_i^2}{p}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

where r_i is the studendized residual.

Cook's Distance

- Cook's D is large when $|r_i|$ is large and h_{ii} is large
- $D_i < F_{p,n-p,0.2}$ (the 20th percentile) is no concern
- $D_i > F_{p,n-p,0.5}$ indicate substantial influence
- What about between?
  - Crude rule of thumb: If $D_i > 1$, investigate the ith observation as possibly influential.
  - If $p \to \infty$, what happens?

DFBETAS

- DFBETAS measures the influence of the ith observation on a single coefficient β_k.

$$DFBETAS_{k(i)} = (\hat{\beta}_k - \hat{\beta}_{k(-i)})/\sqrt{MSE_{-i}c_{kk}}$$

  where $c_{kk} = [(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{kk}$
  - Recall that $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.
- Larger $|DFBETAS_{k(i)}|$ indicates larger impact of observation i on $\hat{\beta}_k$.
  - For small to medium data sets, if $|DFBETAS_{k(i)}| > 1$, then the ith observation may be influential.
  - For large data sets, if $|DFBETAS_{k(i)}| > 2/\sqrt{n}$, then the ith observation may be influential.
  - The sign of $DFBETAS_{k(i)}$ tells whether inclusion of observation i leads to an increase (+) or decrease (-) in $\hat{\beta}_k$.

Multicollinearity

- When the predictor variables are correlated among themselves, multicollinearity among them is said to exist.
- Consider two extreme cases.
  - Uncorrelated predictor variables.
  - Predictor variables are perfectly correlated.

Linearly Independent Predictor Variables

- Consider $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
  - Suppose $X_1 \perp X_2$, i.e. $\text{Corr}(\hat{X}_1, X_2) = 0$.
  - We can show

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_{i1} - \bar{X}_1)}{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2}, \quad \hat{\beta}_2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_{i2} - \bar{X}_2)}{\sum_{i=1}^n (X_{i2} - \bar{X}_2)^2}.$$

  - The LS estimate of β_1 is not affected by X_2 and vice versa.
- Also, the order in which the predictor variables are put in the model is inconsequential.
- Interpretation of regression coefficients is clear: β_1 is the expected change in Y for one unit increase in X_1 with X_2 held constant.

Predictor Variables are Linearly Dependent

- Again, suppose $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
- But $X_2 = 2X_1 + 1$.
- Suppose $\beta_0 = 3, \beta_1 = 2, \beta_2 = 5$.
- Then all the following models give the same fit for Y:
  - $Y = 3 + 2X_1 + 5X_2 + \epsilon$.
  - $Y = 8 + 12X_1 + \epsilon$.
  - $Y = 2 + 6X_2 + \epsilon$.

- What is still fine.
  - Prediction of Y is fine within the model/data scope, but unreliable outside the model/data scope.
- What is not.
  - The $\widehat{\beta}$'s are not unique because $\boldsymbol{X}$ is reduced rank (why?) and $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible.
  - Interpretation of "the effect of the jth predictor holding all other variables constant" is difficult.
  - A regression coefficient may no longer reflect the effect of its corresponding predictor variable.
- Even worse: multicollinearity does not violate any model assumptions!

Concerns with Multicollinearity

- Multicollinearity could be between 3 or more variables, rather than just a correlated pair. That would be harder to detect.
- Effects of multicollinearity on the inference of regression coefficients:
  - Large changes in the fitted $\hat{\beta}_k$ when another X is added or deleted
  - Small changes in the data lead to very large changes in $\hat{\boldsymbol{\beta}}$
  - Large $s\{\hat{\beta}_k\}$. Makes the $\hat{\beta}_k$ seem non-significant even though the predictors are jointly significant and R^2 is large.
  - More difficult to interpret $\hat{\beta}_k$ as the effect of X_k on Y because the other X's cannot be held constant.
  - Estimated coefficients may have wrong sign or implausible magnitudes.

Some Diagnostics for Multicollinearity

- Multicollinearity is harmless for estimation of mean response and prediction of new observation at $\boldsymbol{X}_h$.
  - Assuming no extrapolation!
- Diagnosing multicollinearity
  - Large changes in $\hat{\beta}$'s when a predictor (or an observation) is added or deleted.
  - Important predictors are not statistically significant (large p-values) in individual tests.
  - Wide confidence intervals for β's corresponding to important predictor variables.
  - The sign of $\hat{\beta}$ is counter-intuitive.
  - Predictors are highly correlated.

Variance Inflation Factor (VIF)

- Variance inflation factor (VIF) for $\hat{\beta}_k$:

$$\text{VIF}_k = \frac{1}{1 - R_k^2}, \quad k = 1, \ldots, p-1$$

  where R_k^2 is the R^2 for a regression of X_k on the other predictor variables.
  - VIF measures the increase in the standard error of β_k due to the presence of other variables.
  - If $\max_k \text{VIF}_k > 10$, multicollinearity may have a large impact on the inference.
  - If $\sum_{j=1}^{p-1} \text{VIT}_j > p - 1$, there may be serious multicollinearity problems (for large p).

Variance Inflation Factor (VIF)

- As R_k^2 is the coefficient of multiple determination R^2 of the model

$$X_{ik} = \beta_0 + \sum_{j \neq k}^{p-1} X_{ij} + \epsilon.$$

  - $\sigma^2\{\hat{\beta}_k\} \approx \sigma^2 \text{VIF}_k = \frac{\sigma^2}{1-R_k^2}$.
    1. When R_k^2 decreases, $\sigma^2\{\hat{\beta}_k\}$ decreases.
    2. When R_k^2 increases, $\sigma^2\{\hat{\beta}_k\}$ increases.

- In fact, $\text{VIF}_k = (n-1)\left((\boldsymbol{X}_c\boldsymbol{X}_c)^{-1}\right)_{kk}$ where $\boldsymbol{X}_c$ is the scaled design matrix. (Can you show this?)

Some Remedial Measures for Multicollinearity

- Classical method.
  - Drop one or more predictor variables from the model (selection, frontier of statistics).
  - For polynomial or interaction regression models, use centered predictor variables $X_{ik} - \bar{X}_k$ to reduce multicollinearity (Gram-Schmidt transformation, why?)
- Modern method.
  - Create new predictor variables: principal component regression, PLRS, dimension-reductions.
  - Use shrinkage regression such as ridge, LASSO, SCAD, group LASSO, adaptive LASSO, $\cdots$

$$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

  - Although $\hat{\boldsymbol{\beta}}_R$ has a smaller variance, it is a biased estimator of $\boldsymbol{\beta}$.
  - Going into very frontier of Statistical Machine Learning and High-dimensional Inference.