BIOS 755: Model Diagnostics

Alexander McLain

February 20, 2025

First, let's review model diagnostics for standard regression models

Two model diagnostic topics that are of interest:

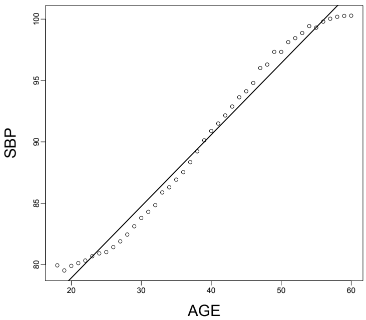
1. Examine whether the assumptions of linear regression models satisfied
   - Linear, Independent, Normality, and Equal variance (LINE)
   - Most of the assumptions of linear regression can be checked by plotting the estimated residuals
   $$e_i = Y_i - \hat{Y}_i$$
2. Identify points that potentially influence the fitted regression line (outliers and influential points)

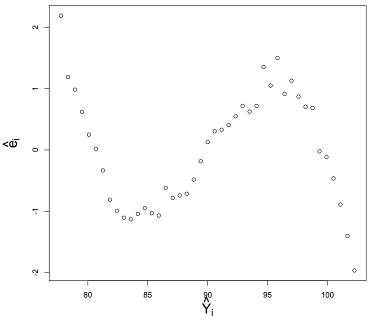
Line

- Assumption: there is a linear relationship between Y and the X's. I.e.,
  - The expected value of the residual is zero for all combinations of X's
  - $E(e_i) = 0$.
- How to check
  - Plot Y_i vs. X_i.
  - Plot e_i vs. $\hat{Y}_i$

Line

Example Violation:





Corrective actions:

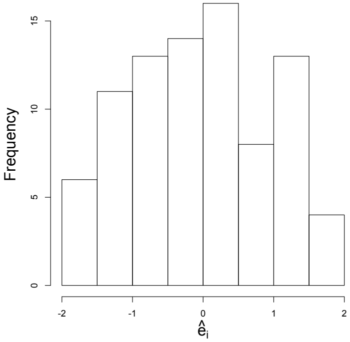
- Fit a different regression model.
- Add quadratic of cubic components.
- Transformation of Y and/or X

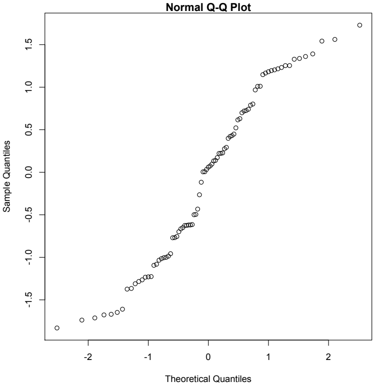
Normality

- Assumption: The residuals are normally distributed.
- How to check
  - Histogram of the e_i's.
  - Normal probability plot of e_i.
  - Tests of normality on e_i.
  - Outlier tests.

Normality

Example Violation:





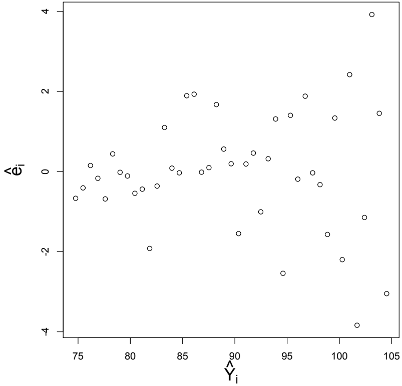
Corrective actions:

- Examine outliers to determine if they are contaminated
- Transformation of Y and/or X

Equal Variance

- **Assumption:** the variance of the residuals is equal for all X's.
- How to check
  - Plot Y_i vs. X_i.
  - Plot e_i vs. $\hat{Y}_i$

  Example Violation $\Longrightarrow$



- **Corrective action:**
  - Transformation of Y.
  - Patterned covariance model

If an assumption is violated, what is the effect?

- Line:
  - Predictive estimates ($\hat{Y}_i$) are biased estimates of $E(Y_i)$.
  - **Predictive intervals** of Y_i are not valid.
  - *Confidence intervals and hypothesis tests* of β, $\hat{Y}_i$ are not valid.
- Independent:
  - **Predictive intervals** of Y_i are not valid.
  - *Confidence intervals and hypothesis tests* of β, $\hat{Y}_i$ are not valid.
- Normality:
  - **Predictive intervals** of Y_i are not valid.
  - *Confidence intervals and hypothesis tests* of β, $\hat{Y}_i$ are not valid for small samples (i.e., $n < 30$).
- Equal variance:
  - **Predictive intervals** of Y_i are not valid.
  - *Confidence intervals and hypothesis tests* of β, $\hat{Y}_i$ are not valid.

Are these a problem? Depends on your question.

Linear Mixed Model

- The linear mixed model can be expressed as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$$

where

- $\boldsymbol{X}_i$ – $n_i \times p$ matrix of covariates
- $\boldsymbol{Z}_i$ – $n_i \times q$ matrix of covariates (usually the columns of $\boldsymbol{Z}_i$ are a subset of the columns of $\boldsymbol{X}_i$ and $q < k$).
- β – $k \times 1$ vector of fixed effects.
- $\boldsymbol{b}_i$ – $q \times 1$ vector of random effects, $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$,
- $\boldsymbol{e}_i$ – $n_i \times 1$ vector of errors and $\boldsymbol{e}_i \sim N(0, \boldsymbol{R}_i)$.

Types of residuals

Four types of residuals

- **Unconditional residuals**

$$r_i^u = Y_i - \boldsymbol{X}_i\hat{\beta}$$

These are not as good for model diagnostics.

- **Conditional residuals**

$$r_i^c = Y_i - (\boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)$$

where $\hat{\boldsymbol{b}}_i$ are the EBLUP random effects.

- **Scaled residuals**

$$r_i^s = \hat{L}_i^{-1} r_i^u$$

where $\hat{\Sigma}_i = \hat{L}_i\hat{L}_i'$ and $\hat{\Sigma}_i = \boldsymbol{Z}_i\hat{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$.

Types of residuals

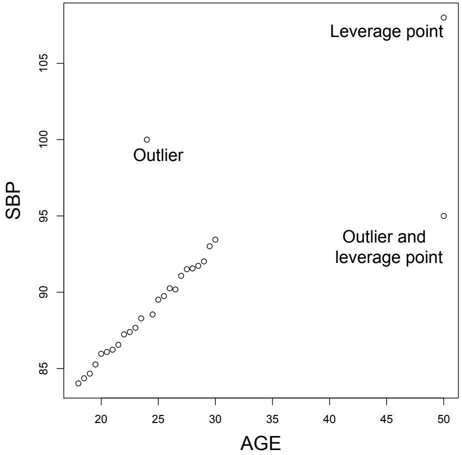
- **Studentized conditional residuals**
  - Each conditional residual is divided by its estimated standard deviation

$$r_i^{st} = \frac{r_i^c}{\sqrt{Var(r_i^c)}}$$

  - By scaling the residuals, we can remove the effect of unequal variance (i.e., by time point) in our model.
  - These are appropriate to use for linear mixed models.
- For patterned covariance models, scaled residuals are good.
- We'll use residuals similar to how they are used for linear regression (see above).

Introduction

We will discuss the detection of **outliers** and **leverage points**.



What's the difference?

- **Leverage point:** unusual predictor combination, i.e. $(X_{i1}, \ldots, X_{ip})$.
- **Outlier:** Unusual Y_i values.

How to identify influence?

- The basic procedure for quantifying influence is simple:
  1. Fit the model to the data and obtain estimates of all parameters.
  2. Remove one or more data points from the analysis and compute updated estimates of model parameters.
  3. Based on full- and reduced-data estimates, contrast quantities of interest to determine how the absence of the observations changes the analysis.
- We will look at the influence an observation has on the **overall** model fit, as well as on **mean and variance** parameter estimates.

Overall influence

- The overall influence of an individual on the model fit can be measured by the change in the likelihood when the individual is not included.
- The is measured by the likelihood and restricted likelihood distances

$$
\begin{aligned}
LD_i &= 2\{l(\hat{\theta}) - l(\hat{\theta}_{-i})\} \\
RLD_i &= 2\{l_R(\hat{\theta}) - l_R(\hat{\theta}_{-i})\}
\end{aligned}
$$

- The likelihood distance is a global summary measure, expressing the joint influence of the individual on the parameter estimates of θ

Change in Parameter Estimates

- Change in the parameter estimates can be estimated in a similar manner.
- Cook's D and the multivariate DFFITS statistic are two measures of how much the parameter estimates change when an individual is removed from the analysis.
- While we won't go over the specific formulas, for both statistics were concerned about large values, indicating that the change in the parameter estimate is large relative to the variability of the estimate.

Change in Parameter Estimates Variance

- Another way to measure the influence of an individual is to look at how the variance of parameter estimates changes.
- Data points that have a small Cook's D, for example, can still greatly affect hypothesis tests and confidence intervals.
- COVTRACE and COVRATIO are two quantities in SAS that measure the impact on the covariance estimates.
- The benchmarks of "no influence" are zero for the covariance trace and one for the covariance ratio.

Detecting points with high leverage

How do we detect X_i values with large leverage?

- Calculation of leverage of X_i is done by h_{ii} is the diagonal elements of the so-called hat matrix H ($H = X'(X'X)^{-1}X$)
- The greater the h_{ii}, the greater the leverage
- The both high and low $\boldsymbol{X}_i$ compared to the sample mean $\bar{\boldsymbol{X}}$ gives large leverage.