Diagnostics and Remedial Measures

KNNL – Chapter 3

Diagnostics for Predictor Variables

- Problems can occur when:
  - Outliers exist among X levels
  - X levels are associated with run order when experiment is run sequentially
- Useful plots of X levels
  - Dot plot for discrete data
  - Histogram
  - Box Plot
  - Sequence Plot (X versus Run #)

Residuals

$e_i = Y_i - \hat{Y}_i$ Note this is different from $\varepsilon_i = Y_i - \left(\beta_0 + \beta_1 X_i\right)$

Properties of Residuals:

Mean: $\overline{e} = \dfrac{\sum_{i=1}^{n} e_i}{n} = 0$

Variance: $s^2 = \dfrac{\sum_{i=1}^{n}\left(e_i - \overline{e}\right)^2}{n-2} = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2} = \dfrac{SSE}{n-2} = MSE$ Correct Model $\Rightarrow$ $E\{MSE\} = \sigma^2$

Nonindependence: Residuals are not independent (based on same fitted regression).

2 Constraints: $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} X_i e_i = 0$

Not a problem if number of observations is large compared to parameters in model

Semistudentized Residuals:

$e_i^* = \dfrac{e_i - \overline{e}}{\sqrt{MSE}} = \dfrac{e_i}{\sqrt{MSE}}$ where $\sqrt{MSE}$ approximates $s\left\{e_i\right\}$

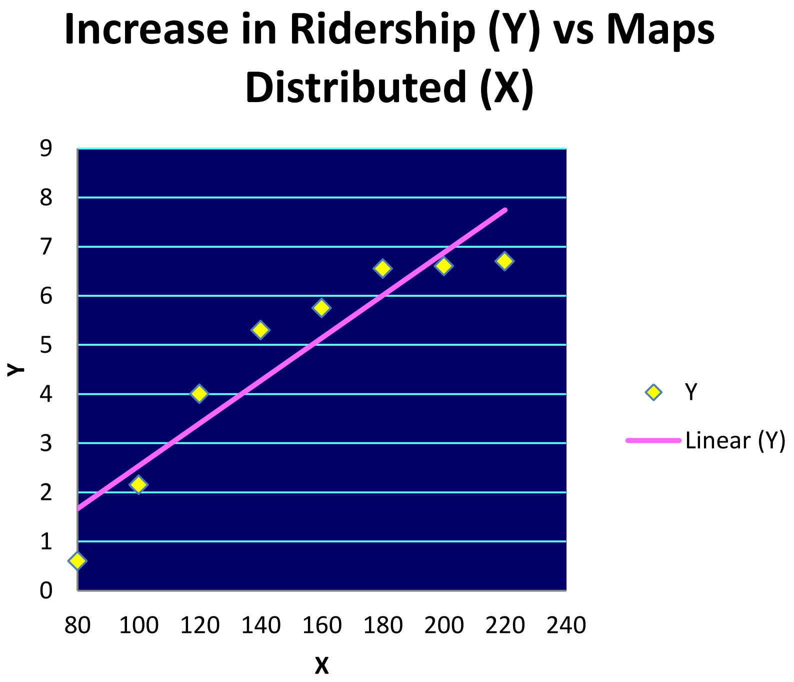
e_i^* is like a *t*-statistic and can be used to detect outliers

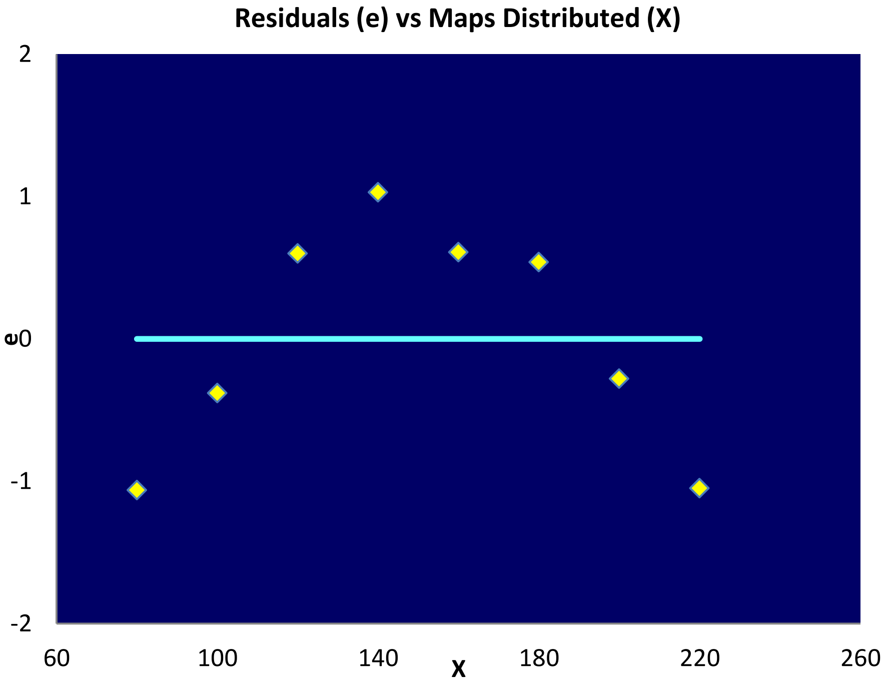
Model Departures Detected With Residuals and Plots

- Relation between Y and X is not linear
- Errors have non-constant variance
- Errors are not independent
- Existence of Outlying Observations
- Non-normal Errors
- Missing predictor variable(s)
- Common Plots
    - Residuals/Absolute Residuals versus Predictor Variable
    - Residuals/Absolute Residuals versus Predicted Values
    - Residuals versus Omitted variables
    - Residuals versus Time
    - Box Plots, Histograms, Normal Probability Plots

Detecting Nonlinearity of Regression Function

- Plot Residuals versus X
  - Random Cloud around 0 ⇒ Linear Relation
  - U-Shape or Inverted U-Shape ⇒ Nonlinear Relation
- Maps Distribution Example (Table 3.1,Figure 3.5, p.10)





Non-Constant Error Variance / Outliers / Non-Independence

- Plot Residuals versus X or Predicted Values
  - Random Cloud around 0 ⇒ Linear Relation
  - Funnel Shape ⇒ Non-constant Variance
  - Outliers fall far above (positive) or below (negative) the general cloud pattern
- Plot absolute Residuals, squared residuals, or square root of absolute residuals
  - Positive Association ⇒ Non-constant Variance
- Measurements made over time: Plot Residuals versus Time Order (Expect Random Cloud if independent)
  - Linear Trend ⇒ Process “improving” or “worsening” over time
  - Cyclical Trend ⇒ Measurements close in time are similar

Non-Normal Errors

- Box-Plot of Residuals – Can confirm symmetry and lack of outliers
- Check Proportion that lie within 1 standard deviation from 0, 2 SD, etc, where SD=sqrt(MSE)
- Normal probability plot of residual versus expected values under normality – should fall approximately on a straight line (Only works well with moderate to large samples) **qqnorm(e); qqline(e)** in R

Expected value of Residuals under Normality:

1) Rank residuals from smallest (large/negative) to highest (large/positive) Rank = k

2) Compute the percentile using $p = \dfrac{k - 0.375}{n + 0.25}$ and obtain corresponding z-value: $z(p)$

3) Multiply by $s = \sqrt{MSE}$ expected residual = $\sqrt{MSE}\left[z(p)\right]$

Omission of Important Predictors

- If data are available on other variables, plots of residuals versus X, separate for each level or sub-ranges of the new variable(s) can lead to adding new predictor variable(s)
- If, for instance, if residuals from a regression of salary versus experience was fit, we could view residuals versus experience for males and females separately, if one group tends to have positive residuals, and the other negative, then add gender to model

Test for Independence – Runs Test

- Runs Test (Presumes data are in time order)
  1) Write out the sequence of +/- signs of the residuals
  2) Count n_1 = # of positive residuals, n_2 = # of negative residuals
  3) Count u = # of “runs” of positive and negative residuals
  4) If $n_1 + n_2 \leq 20$, refer to Table of critical values (Not random if u is too small)
  5) If $n_1 + n_2 > 20$, use a large-sample (approximate) z-test:

Under Independence:

$$E\{u\} = \mu_u = \frac{2n_1 n_2}{n_1 + n_2} + 1 \qquad \sigma\{u\} = \sigma_u = \sqrt{\frac{2n_1 n_2 \left(2n_1 n_2 - n_1 - n_2\right)}{\left(n_1 + n_2\right)^2 \left(n_1 + n_2 - 1\right)}}$$

$$z_u = \frac{u - \mu_u + 0.5}{\sigma_u} \qquad \text{P-value} = P\left(Z \leq z_u\right)$$

Test For Independence - Durbin-Watson Test

$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t \quad u_t \sim NID\left(0, \sigma^2\right) \quad |\rho| < 1$

$H_0 : \rho = 0 \quad \Rightarrow$ Errors are uncorrelated over time

$H_A : \rho > 0 \quad \Rightarrow$ Positively correlated

1) Obtain Residuals from Regression

2) Compute Durbin-Watson Statistic (given below)

3) Obtain Critical Values from Table B.7, pp. 1330-1331

If $DW < d_L(1, n)$ Reject H_0

If $DW > d_U(1, n)$ Conclude H_0

Otherwise Inconclusive

Test Statistic: $DW = \dfrac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2}$

Note: This generalizes to any number of Predictors $(p-1)$

Test for Normality of Residuals

- Correlation Test
  1) Obtain correlation between observed residuals and expected values under normality (see slide 7)
  2) Compare correlation with critical value based on α-level from Table B.6, page 1329
  3) Reject the null hypothesis of normal errors if the correlation falls below the table value
- Shapiro-Wilk Test – Performed by most software packages. Related to correlation test, but more complex calculations – see NFL Point Spreads and Actual Scores Case Study for description of one version

Equal (Homogeneous) Variance - I

Brown-Forsythe Test:

$H_0:$ Equal Variance Among Errors $\sigma^2\left\{\varepsilon_i\right\}=\sigma^2\ \forall\ i$

$H_A:$ Unequal Variance Among Errors (Increasing or Decreasing in X)

1) Split Dataset into 2 groups based on levels of X (or fitted values) with sample sizes: $n_1, \quad n_2$

2) Compute the median residual in each group: $\tilde{e}_1, \quad \tilde{e}_2$

3) Compute absolute deviation from group median for each residual:

$$d_{ij}=\left|e_{ij}-\tilde{e}_j\right| \quad i=1,\ldots,n_j \quad j=1,2$$

4) Compute the mean and variance for each group of d_{ij}: $\quad \bar{d}_1, s_1^2 \quad \bar{d}_2, s_2^2$

5) Compute the pooled variance: $s^2=\dfrac{\left(n_1-1\right)s_1^2+\left(n_2-1\right)s_2^2}{n_1+n_2-2}$

Test Statistic: $t_{BF}=\dfrac{\bar{d}_1-\bar{d}_2}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \overset{H_0}{\sim} t\left(n_1+n_2-2\right)$

Reject H_0 if $\left|t_{BF}\right|\geq t\left(1-\left(\alpha/2\right);n-2\right)$

Equal (Homogeneous) Variance - II

Breusch-Pagan (aka Cook-Weisberg) Test:

$H_0:$ Equal Variance Among Errors $\sigma^2\left\{\varepsilon_i\right\} = \sigma^2 \ \forall\, i$

$H_A:$ Unequal Variance Among Errors $\sigma_i^2 = \sigma^2 h\left(\gamma_1 X_{i1} + \ldots + \gamma_p X_{ip}\right)$

1) Let $SSE = \sum_{i=1}^{n} e_i^2$ from original regression

2) Fit Regression of e_i^2 on $X_{i1},\ldots X_{ip}$ and obtain $SS\left(\text{Reg}^*\right)$

Test Statistic: $X_{BP}^2 = \dfrac{SS\left(\text{Reg}^*\right)/2}{\left(\sum_{i=1}^{n} e_i^2 \big/ n\right)^2} \overset{H_0}{\sim} \chi_p^2$

Reject H$_0$ if $X_{BP}^2 \geq \chi^2\left(1-\alpha; p\right)$ $\quad$ p = # of predictors

Linearity of Regression

F-Test for Lack-of-Fit (n_j observations at c distinct levels of "X")

$$H_0 : E\left(Y_i\right) = \beta_0 + \beta_1 X_i \quad H_A : E\left(Y_i\right) = \mu_i \neq \beta_0 + \beta_1 X_i$$

Compute fitted value Y_j and sample mean $\overline{Y}_j$ for each distinct X level

Lack-of-Fit: $SS\left(LF\right) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(\overline{Y}_j - Y_j\right)^2 \quad df_{LF} = c - 2$

Pure Error: $SS\left(PE\right) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2 \quad df_{PE} = n - c$

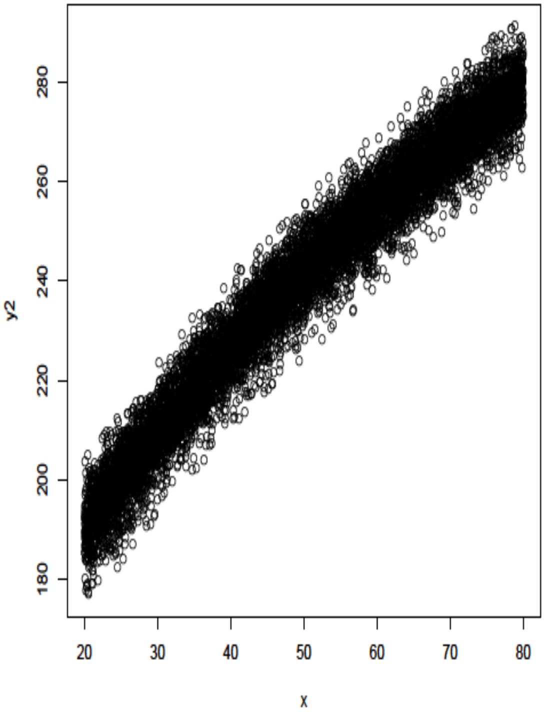
Test Statistic: $F_{LOF} = \dfrac{\left(SS(LF)/(c-2)\right)}{\left(SS(PE)/(n-c)\right)} = \dfrac{MS(LF)}{MS(PE)} \quad \overset{H_0}{\sim} \quad F_{c-2,n-c}$

Reject H$_0$ if $F_{LOF} \geq F\left(1-\alpha; c-2, n-c\right)$

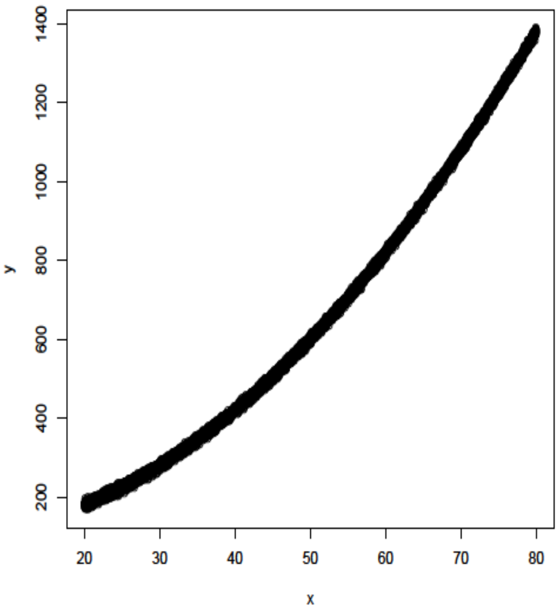
Remedial Measures

- Nonlinear Relation – Add polynomials, fit exponential regression function, or transform *Y* and/or *X*
- Non-Constant Variance – Weighted Least Squares, transform *Y* and/or *X*, or fit Generalized Linear Model
- Non-Independence of Errors – Transform *Y* or use Generalized Least Squares
- Non-Normality of Errors – Box-Cox tranformation, or fit Generalized Linear Model
- Omitted Predictors – Include important predictors in a multiple regression model
- Outlying Observations – Robust Estimation

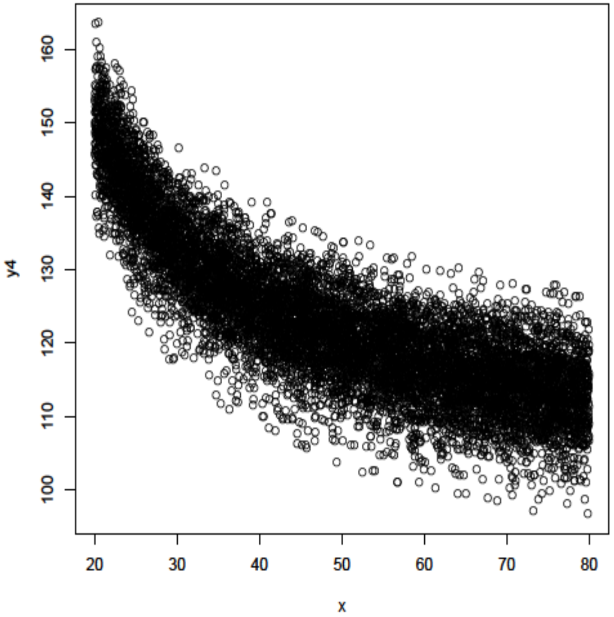
Transformations for Non-Linearity – Constant Variance



$X' = \sqrt{X}$ $X' = \ln(X)$

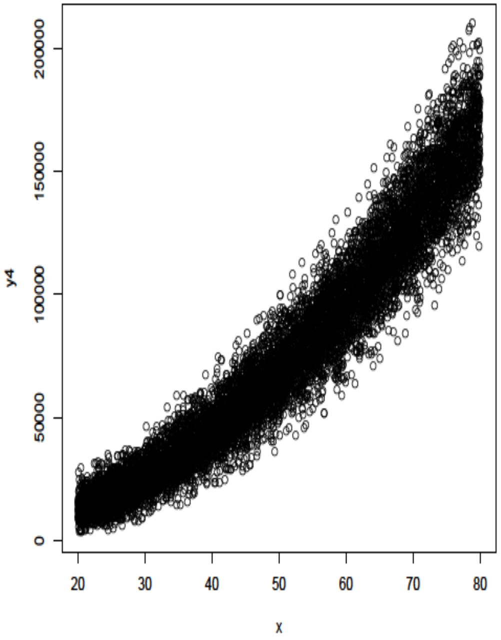


$X' = X^2$ $X' = e^X$

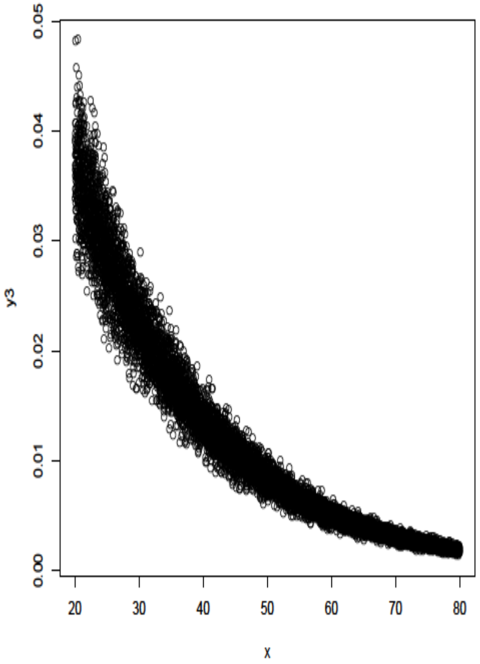


$X' = 1/X$ $X' = e^{-X}$

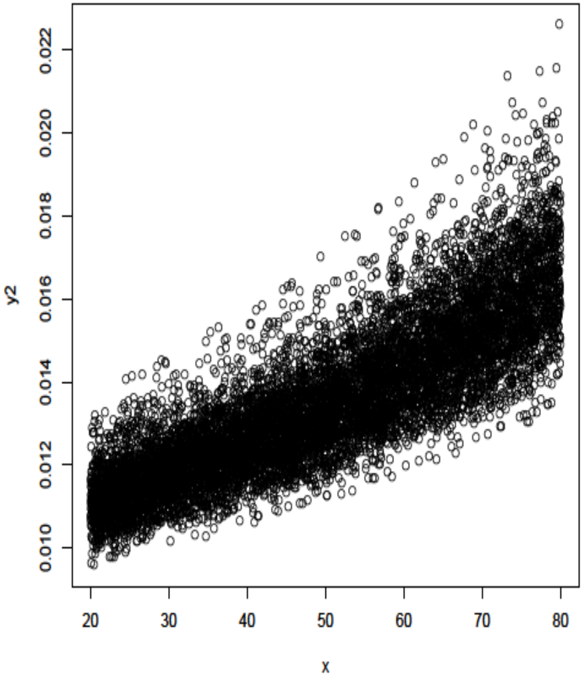
Transformations for Non-Linearity – Non-Constant Variance



Y’ = √Y



Y’ = ln(Y)



Y’ = 1/Y

Box-Cox Transformations

- Automatically selects a transformation from power family with goal of obtaining: normality, linearity, and constant variance (not always successful, but widely used)
- Goal: Fit model: $Y' = \beta_0 + \beta_1 X + \varepsilon$ for various power transformations on Y, and selecting transformation producing minimum SSE (maximum likelihood)
- Procedure: over a range of λ from, say -2 to +2, obtain W_i and regress W_i on X (assuming all $Y_i > 0$, although adding constant won't affect shape or spread of Y distribution)

$$W_i = \begin{cases} K_1\left(Y_i^{\lambda} - 1\right) & \lambda \neq 0 \\ K_2 \ln\left(Y_i\right) & \lambda = 0 \end{cases}$$

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} \qquad K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

Lowess (Smoothed) Plots

- Nonparametric method of obtaining a smooth plot of the regression relation between *Y* and *X*
- Fits regression in small neighborhoods around points along the regression line on the X axis
- Weights observations closer to the specific point higher than more distant points
- Re-weights after fitting, putting lower weights on larger residuals (in absolute value)
- Obtains fitted value for each point after “final” regression is fit
- Model is plotted along with linear fit, and confidence bands, linear fit is good if lowess lies within bands