Regression Diagnostics

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Linear Regression Assumptions

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I)$$

Assumptions

- **L**inear relationship
- **I**ndependent observations
- **N**ormally distributed residuals
- **E**qual variance across X's
- Plus need to check for influential points and outliers: one or a few observations should not dominate the model fit

Independence Assumption

The responses are independent of one another. In simplest terms independence means that knowing one response tells us nothing about another one. It largely depends on how the data are collected. Key designs where independence is unlikely include:

Independence Assumption

The responses are independent of one another. In simplest terms independence means that knowing one response tells us nothing about another one. It largely depends on how the data are collected. Key designs where independence is unlikely include:

- **Times series studies**: observations close together in time tend to be more similar than observations far apart in time

Independence Assumption

The responses are independent of one another. In simplest terms independence means that knowing one response tells us nothing about another one. It largely depends on how the data are collected. Key designs where independence is unlikely include:

- **Times series studies**: observations close together in time tend to be more similar than observations far apart in time
- **Repeated observations** on each of many persons followed in a cohort study: Two observations for the same person are likely to be more similar than two from different persons.

Independence Assumption

The responses are independent of one another. In simplest terms independence means that knowing one response tells us nothing about another one. It largely depends on how the data are collected. Key designs where independence is unlikely include:

- **Times series studies**: observations close together in time tend to be more similar than observations far apart in time
- **Repeated observations** on each of many persons followed in a cohort study: Two observations for the same person are likely to be more similar than two from different persons.
- **Family studies**: responses from multiple members of the same family tend to be correlated because family members share genes and environments in ways that are not easily measured by our predictor variables.

Independence Assumption

The responses are independent of one another. In simplest terms independence means that knowing one response tells us nothing about another one. It largely depends on how the data are collected. Key designs where independence is unlikely include:

- **Times series studies**: observations close together in time tend to be more similar than observations far apart in time
- **Repeated observations** on each of many persons followed in a cohort study: Two observations for the same person are likely to be more similar than two from different persons.
- **Family studies**: responses from multiple members of the same family tend to be correlated because family members share genes and environments in ways that are not easily measured by our predictor variables.
- **Clustered designs**: when subjects are samples in clusters: families, neighborhoods, schools, etc. responses from the same cluster tend to be correlated.

When Ys are not independent...

$$
\begin{aligned}
Y &= X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \Sigma) \quad \text{(True)} &(1)\\
Y &= X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I) &(2)
\end{aligned}
$$

- Are the least square estimates unbiased?

When Ys are not independent...

$$
\begin{aligned}
Y &= X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \Sigma) \quad \text{(True)} && (1) \\
Y &= X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I) && (2)
\end{aligned}
$$

- Are the least square estimates unbiased? $E(\hat{\beta}) = ?$
- std error($\hat{\beta}$) are incorrect, sometimes grossly incorrect.
- Hence confidnece intervals, t-tests and F-tests obtained from a least squares procedure will be wrong.
- Need to use "robust" or "bootstrap" standard errors estimates.
- Or consider other models that account for correlations. See Analysis of Longitudinal Data by Diggle, Heagerty, Liang and Zeger.

Robust Standard Error Estimates in R

```
>setwd("/Users/yen-yiho/Desktop/STAT704/Data")
>data<-read.csv("data.csv", header=T, stringsAsFactor=F)
>reg <- lm(weight ~ lag_calories+lag_cycling+
            I(lag_calories*lag_cycling),
          data=data)
>summary(reg)
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  75.0209516  0.0908520 825.749  < 2e-16 ***
lag_calories                  0.0011774  0.0003311   3.556 0.000449 ***
lag_cycling                   0.3464949  0.3059401   1.133 0.258476
I(lag_calories * lag_cycling) -0.0014470  0.0004103  -3.527 0.000500 ***

>summary(reg,robust = T)
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  75.0209516  0.0988184 759.180  < 2e-16 ***
lag_calories                  0.0011774  0.0002454   4.799 2.74e-06 ***
lag_cycling                   0.3464949  0.2898921   1.195    0.233
I(lag_calories * lag_cycling) -0.0014470  0.0003241  -4.464 1.21e-05 ***
```

Linear Regression Assumptions

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I)$$

Assumptions

- **L**inear relationship
- **I**ndependent observations
- **N**ormally distributed residuals
- **E**qual variance across X's
- Plus need to check for influential points and outliers: one or a few observations should not dominate the model fit

When sample size is small ...

- When sample size $\to \infty$, even if Y is not normal

$$\hat{\beta} = (X'X)^{-1}X'Y = WY \to N(\beta, W\text{Var}(Y)W') \quad (CLT)$$

- The normality of $\hat{\beta}$ depends on
  - the total sample size
  - the departure of Y from Gaussian
  - the design matrix
- If you worry about Non-Gaussianity in small sample size, what would you do?

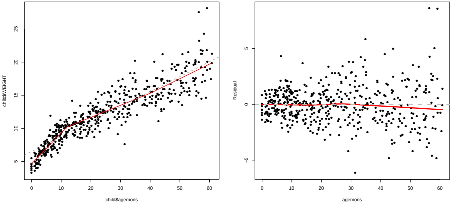
When sample size is small ...

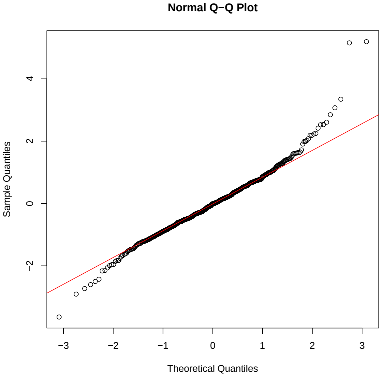
- When sample size $\to \infty$, even if Y is not normal

$$\hat{\beta} = (X'X)^{-1}X'Y = WY \to N(\beta, W\mathrm{Var}(Y)W') \quad (CLT)$$

- The normality of $\hat{\beta}$ depends on
  - the total sample size
  - the departure of Y from Gaussian
  - the design matrix
- If you worry about Non-Gaussianity in small sample size, what would you do?
  bootstrap standard error

Check Normality: Q-Qplot of standardized residuals





Q-Q plot

```
>age12<-ifelse(child$agemons>12, child$agemons-12, 0)
>age30<-ifelse(child$agemons>30, child$agemons-30,0)
>fit1sp<-lm(WEIGHT ~ agemons + age12 + age30, data=child)
>std.resid<-rstandard(fit1sp)
>qqnorm(std.resid)
>qqline(std.resid, col=2)
```