JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

Lecture 9

Model Checking and Key Extensions

Review of where we left off

1. We have established the multiple linear regressio model:

$$Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}, \epsilon_{n\times 1} \sim MVN(0_{n\times 1}, \sigma^2 I_{n\times n})$$

2. We know that:

$$\hat{\beta} \text{ satisfies } X'(Y - X\beta) = 0 \text{ and minimizes } \sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

3. We have defined:
   - $\hat{Y} = X\hat{\beta} = HY,$ where $H = X(X'X)^{-1}X'$
   - $\hat{R} = Y - \hat{Y} = Y - X\hat{\beta} = (I - H)Y$

4. Then we showed that:
   - $\hat{\beta} \sim MVN(\beta, \sigma^2(X'X)^{-1})$
   - $\hat{Y} \sim MVN(X\beta, \sigma^2 H)$
   - $\hat{R} \sim MVN(0, \sigma^2(I - H))$

Review of where we left off

Target	Estimate $\sim$ Sampling Distn	95% CI for target	Test statistic for H0: Target = 0
β_j	$\hat{\beta}_j \sim N(\beta_j, [\sigma^2(X'X)^{-1})]_{jj})$	$\hat{\beta}_j \pm t \times \hat{se}(\hat{\beta}_j)$	$\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$
$A\beta$	$A\hat{\beta} \sim N(A\beta, \sigma^2 A(X'X)^{-1}A')$	$A\hat{\beta} \pm t \times \hat{se}(A\hat{\beta})$	$\frac{A\hat{\beta}_j}{\hat{se}(A\hat{\beta}_j)}$
$g(\beta_j)$	$g(\hat{\beta}_j) \sim N(g(\beta_j), [g'(\beta_j)]^2[\sigma^2(X'X)^{-1}]_{jj})$	$g(\hat{\beta}_j) \pm t \times \hat{se}(g(\hat{\beta}_j))$	$\frac{g(\hat{\beta}_j)}{\hat{se}(g(\hat{\beta}_j))}$
$g(\beta)$	$g(\hat{\beta}) \sim N(g(\beta), g'(\beta)'[\sigma^2(X'X)^{-1}]g'(\beta))$	$g(\hat{\beta}) \pm t \times \hat{se}(g(\hat{\beta}))$	$\frac{g(\hat{\beta})}{\hat{se}(g(\hat{\beta}))}$
$\mu_i = E(Y_i \mid X_i)$	$\hat{Y}_i \sim N(\mu_i, \sigma^2[H]_{ii})$	$\hat{Y}_i \pm t \times \hat{se}(\hat{Y}_i)$	$\frac{\hat{Y}_i}{\hat{se}(\hat{Y}_i)}$
$\mu(x_0) = E(Y \mid x_0)$	$x_0'\hat{\beta} \sim N(x_0'\beta, \hat{\sigma}^2 x_0'(X'X)^{-1}x_0)$	$x_0'\hat{\beta} \pm t \times \hat{se}(x_0'\hat{\beta})$	$\frac{x_0'\hat{\beta}}{\hat{se}(x_0'\hat{\beta}}$

Key Assumptions by Order of Importance

- $E(Y|X) = X\beta$
- Residuals are independent
- Variance of residuals is constant
- Residuals are normally distributed
- There are not a small number of highly influencial observations

Omitted Variable Bias

Omitted Variable Bias

Simulation exercise

- Within your breakout group, design a simulation study that would numerically demonstrate the result we just derived.
  - You go work for 15 minutes and then we will review together

Correct Functional Form for Continuous X

- To explore the assumption that E(Y|X) = Xβ, you can make the following plots:
  1. Plot $\hat{R}$ vs. $X_j, j = 1, \ldots, p$.
     - Recall that the residuals are independent of X if the model is correctly specified
  2. Plot $\hat{R}$ vs. $\hat{Y}$.
     - The residuals and predicted values are independent if the model is correctly specified

- Never plot $\hat{R}$ vs. Y because these are correlated!

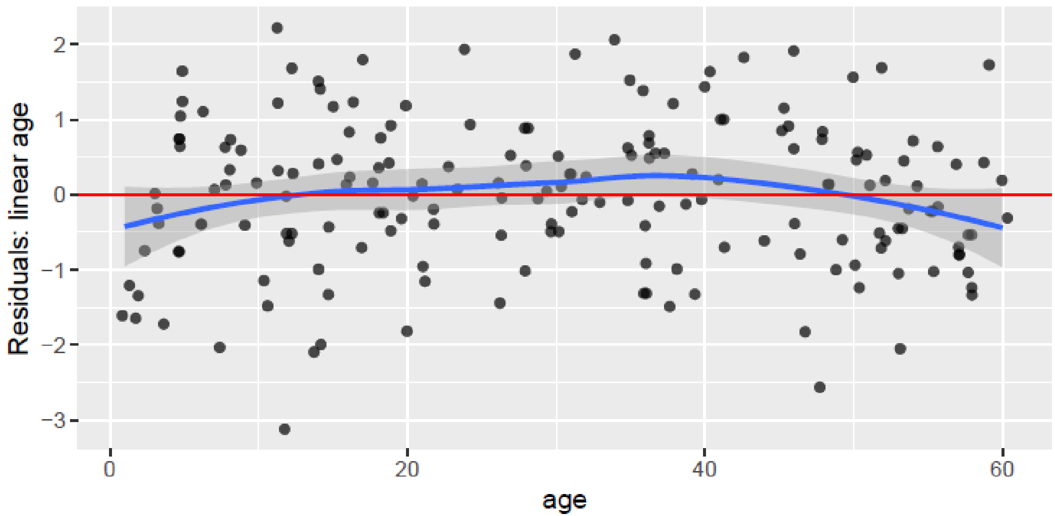
- Based on the figures from 1. and 2., you could modify the model to increase/decrease the complexity of the functional form of the variables.

Example: Nepali Anthropometry Data

```r
reg0<-lm(data=d.cc, arm~age)
d.cc$residuals = residuals(reg0)
```

```r
ggplot(d.cc,aes(x=age, y=residuals)) +
    geom_jitter(alpha = 0.7) +
    geom_smooth() +          geom_hline(yintercept=0,color="red") +
    labs(y="Residuals: linear age")
```

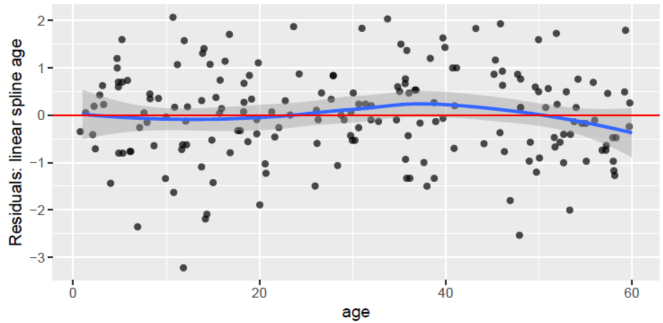
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

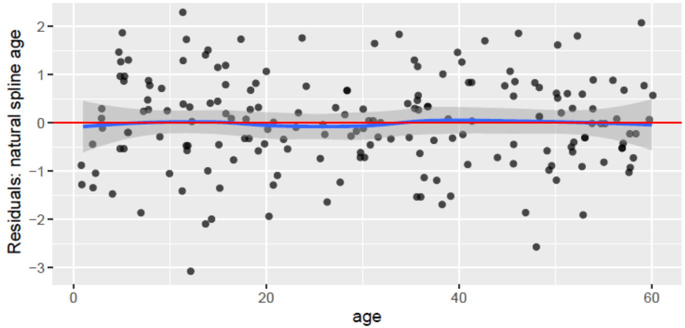


Example: Nepali Anthropometry Data

Update the model to include a smooth function of age via linear splines or natural splines.

```
reg1 = lm(arm~age + agesp6 + agesp12,data=d.cc)
reg2 = lm(arm~ns(age,3),data=d.cc)
d.cc$residuals1 = residuals(reg1)
d.cc$residuals2 = residuals(reg2)
```





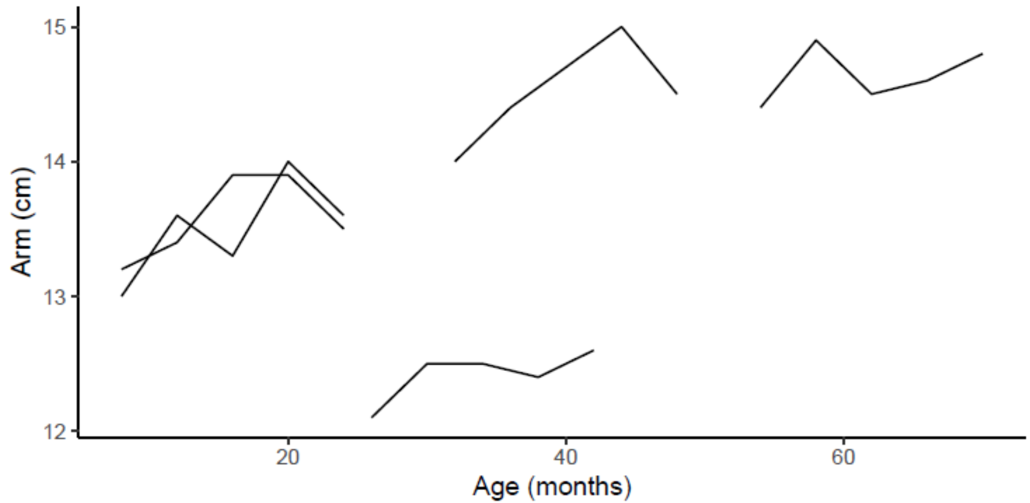
Independence Assumption

- Driven by the design of the study
- Longitudinal design
- Clustered design
- Why do we care?

Example: Nepali Anthropometry Data

- Design: i = 1, …, m = 200 children each measured at baseline (j = 1) and then every 4 months for 4 follow-up visits (j = 2, 3, 4, 5).

```
ggplot(d5,aes(x=age,y=arm,group = factor(id))) +
    geom_line() +
    labs(x='Age (months)', y ='Arm (cm)') +
    theme_classic()
```



Checking the Independence Assumption

- Don't need to, we know the independence assumption is violated based on knowledge of the design

- We can explore covariance/correlation in the observed data

- Example: Consider the Nepali Anthropometry data where we have data for i = 1, …, m = 200 children each measured at baseline (j = 1) and then every 4 months for 4 follow-up visits (j = 2, 3, 4, 5).
  - Step 1: Regress *Y* on *X* assuming independence and estimate β and *R*
  - Step 2: Plot $\hat{R}_{ij}$ vs. $\hat{R}_{ik}$ for all j,k
  - Compute $Cov\left(\hat{R}_{ij}, \hat{R}_{ik}\right) = \sqrt{Var(\hat{R}_{ij})} \times \sqrt{Var(\hat{R}_{ik})} \times Corr(\hat{R}_{ij}, \hat{R}_{ik})$
  - Or standardize the residuals and plot $Corr(\hat{R}_{sij}, \hat{R}_{sik})$

How do we rethink the model?

What if we apply least squares to correlated data?

What if we apply least squares to correlated data?

Solution: Weighted least squares

Solution: Weighted least squares

Next time....

- More model checking....
  - Robust variance estimation
  - Non-constant variance
  - Non-normal residuals
  - Influence and leverage statistics