# Regression II
# Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-10-13 14:32

# Outline

- ANOVA
- Matrix formulation
- Two-sample t-test
- Diagnostics
- Measurement error

# Analysis of Variance

- Recall under $H_0 : \beta = 0$

$$t = \frac{\hat{\beta}}{\sqrt{s^2_{y.x} / \sum_i (X_i - \bar{X})^2}} \sim t_{N-2}$$

- Equivalently

$$t = \frac{[XY]/[X^2]}{\sqrt{s^2_{y.x}/[X^2]}} \sim t_{N-2}$$

- In general, if $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$. Thus

$$t^2 = \frac{[XY]^2/[X^2]}{s^2_{y.x}} \sim F_{1,N-2}$$

# Analysis of Variance

- Note

$$\begin{aligned} SSR &= \sum(\hat{Y}_i - \overline{Y})^2 = \sum(\hat{\alpha} + \hat{\beta}X_i - \overline{Y})^2 \\ &= \sum(\overline{Y} - \hat{\beta}\overline{X} + \hat{\beta}X_i - \overline{Y})^2 \\ &= \sum \hat{\beta}^2(X_i - \overline{X})^2 \\ &= \frac{[XY]^2}{[X^2]^2}\sum(X_i - \overline{X})^2 = \frac{[XY]^2}{[X^2]} \end{aligned}$$

- Thus

$$t^2 = \frac{SSR}{MSE} = \frac{SSR}{SSE/(N-2)}$$

# Analysis of Variance

- If $\beta = 0$,

$$\frac{SSR}{\sigma^2} \sim \chi^2_1 \perp \frac{SSE}{\sigma^2} \sim \chi^2_{N-2}$$

(*Cochran's theorem*: cf Neter et al p.76, 1996)

- Thus

$$t^2 = \frac{SSR/1}{SSE/(N-2)} \sim F_{1,N-2}$$

# Analysis of Variance

- For $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$, can use $F$ with

$$C_\alpha = \{F : F > F_{1-\alpha;1,N-2}\}$$

- For two sided alternative $F$ and $t$ tests equivalent
- For one sided alternative, use $t$

# Analysis of Variance

- ANOVA table:

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression | 1 | SSR | SSR | MSR/MSE |
| Residual | $N-2$ | SSE | SSE/$(N-2)$ | |
| Total | $N-1$ | SST | | |

# Matrix Formulation

- Let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

- Linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

# Matrix Formulation

- Equations (1) and (2) from previous slides

$$-\bar{Y} + \alpha + \beta\bar{X} = 0$$

$$-\sum_i X_i Y_i + \alpha \sum_i X_i + \beta \sum_i X_i^2 = 0$$

- Equivalent to

$$\boldsymbol{X'X\beta = X'Y}$$

# Matrix Formulation

- Therefore

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- Can also show

$$SST = \boldsymbol{Y}'\boldsymbol{Y} - \frac{1}{N}\boldsymbol{Y}'\boldsymbol{J}\boldsymbol{Y}$$

$$SSR = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - \frac{1}{N}\boldsymbol{Y}'\boldsymbol{J}\boldsymbol{Y}$$

$$SSE = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$$

where $\boldsymbol{J}$ is an $n \times n$ matrix of 1's

# Linear Regression and 2 Sample t-test

- Define

$$X = \begin{cases} 1 \text{ if calcium} \\ \\ 0 \text{ if placebo} \end{cases}$$

- $X$ is called an *indicator* or *dummy* variable
- Model

$$Y = \alpha + \beta X + \epsilon$$

# Linear Regression and 2 Sample t-test

- Suppose we have 2 groups of observations:

  $Y_{1i}$ for $i = 1, \ldots, n_1$ and $Y_{2i}$ for $i = 2, \ldots, n_2$

- Recall test statistic

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{1/n_1 + 1/n_2}}$$

  where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{N - 2}$$

# Linear Regression and 2 Sample t-test

- Let

$$N = n_1 + n_2$$

$$(Y_1, \ldots, Y_{n_1}) = (Y_{11}, \ldots, Y_{1n_1})$$

$$(Y_{n_1+1}, \ldots, Y_N) = (Y_{21}, \ldots, Y_{2n_2})$$

$$X_i = \begin{cases} 1 \text{ if group 1} \\ \\ 0 \text{ if group 2} \end{cases}$$

# Linear Regression and 2 Sample t-test

- Consider the regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i; i = 1, 2, 3, \ldots, N$$

- Note

$$[X^2] = \sum_i (X_i - \bar{X})^2 = \sum X_i^2 - N\bar{X}^2$$

$$= n_1 - N\left(\frac{n_1}{N}\right)^2 = n_1\left(1 - \frac{n_1}{N}\right) = \frac{n_1 n_2}{N}$$

# Linear Regression and 2 Sample t-test

- Recall

$$\hat{\beta} = \sum c_i Y_i$$

  where $c_i = (X_i - \bar{X})/[X^2]$

- Thus

$$\hat{\beta} = \frac{(1-\bar{X})\sum_{i=1}^{n_1} Y_i}{[X^2]} + \frac{(-\bar{X})\sum_{i=n_1+1}^{N} Y_i}{[X^2]}$$

$$= \bar{Y}_1 - \bar{Y}_2$$

- Can show

$$s^2_{Y \cdot X} = s^2_p$$

# Linear Regression and 2 Sample t-test

- Therefore:

$$t = \frac{\hat{\beta}}{\sqrt{s^2_{Y \cdot X} / \sum_i (X_i - \bar{X})^2}}$$

$$= \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{N/(n_1 n_2)}}$$

# Linear Regression and 2 Sample t-test

- Example: Body fat in Native American children
- Percent body fact (PBF) measured by bioelectric impedance and skinfolds
- Two tribes: Apache (mountains) and O'Odham (desert)
- Question: Is the mean PBF the same in Apache and O'Odham children?
- Samples: O'Odham ($n = 63$); Apache ($n = 35$)

# Linear Regression and 2 Sample t-test

- Two sample t-test results:

```
Two-sample t test with equal variances            1: Number of obs =       35
                                                  2: Number of obs =       63

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Apache  |  32.77771    1.163163   28.1798   0.0000       30.41388    35.14154
 O'Odham |  37.92245     1.09047   34.7762   0.0000       35.74263    40.10227
---------+--------------------------------------------------------------------
    diff | -5.144741    1.701678  -3.02333   0.0032      -8.522545   -1.766937
------------------------------------------------------------------------------
Degrees of freedom: 96
```

# Linear Regression and 2 Sample t-test

- Model

$$Y = \alpha + \beta X + \epsilon$$

where

$$Y = PBF$$

and

$$X = \begin{cases} 1 & \text{If Apache} \\ 0 & \text{If O'Odham} \end{cases}$$

# Linear Regression and 2 Sample t-test

```
regress pbf siten

  Source |       SS       df       MS              Number of obs =      98
---------+------------------------------           F(  1,    96) =    9.14
   Model |  595.538076     1  595.538076            Prob > F      =  0.0032
Residual |  6254.73035    96  65.1534412            R-squared     =  0.0869
---------+------------------------------           Adj R-squared =  0.0774
   Total |  6850.26843    97   70.621324            Root MSE      =  8.0718

------------------------------------------------------------------------------
     pbf |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   siten |   5.144741   1.701678      3.023   0.003       1.766937    8.522545
   _cons |   27.63297   2.912094      9.489   0.000       21.85251    33.41343
------------------------------------------------------------------------------
```

# Diagnostics

- Assumptions for linear regression
  1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
  2. $X$'s are fixed constants
  3. $\epsilon_i$ iid $\sim N(0, \sigma^2)$

# Diagnostics

- Assumptions: Linear model and homogeneity of variance
- *Residual plot*: Scatterplot of

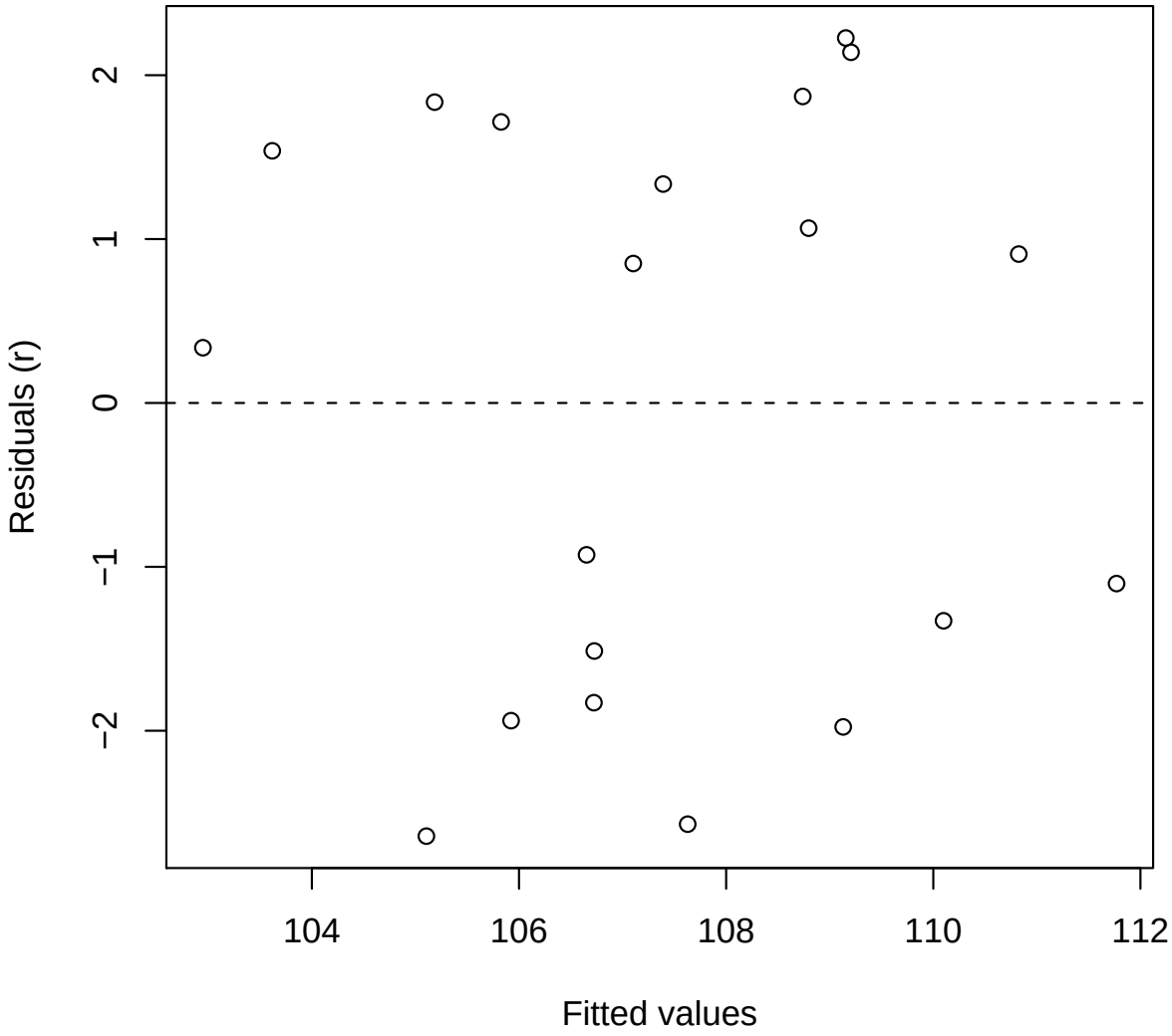
$$(\hat{Y}_i, r_i) = (\hat{Y}_i, Y_i - \hat{Y}_i)$$

- If we see lack of homogeneity of variance or linearity, consider transformations; See Table 10.28 (page 399) of text

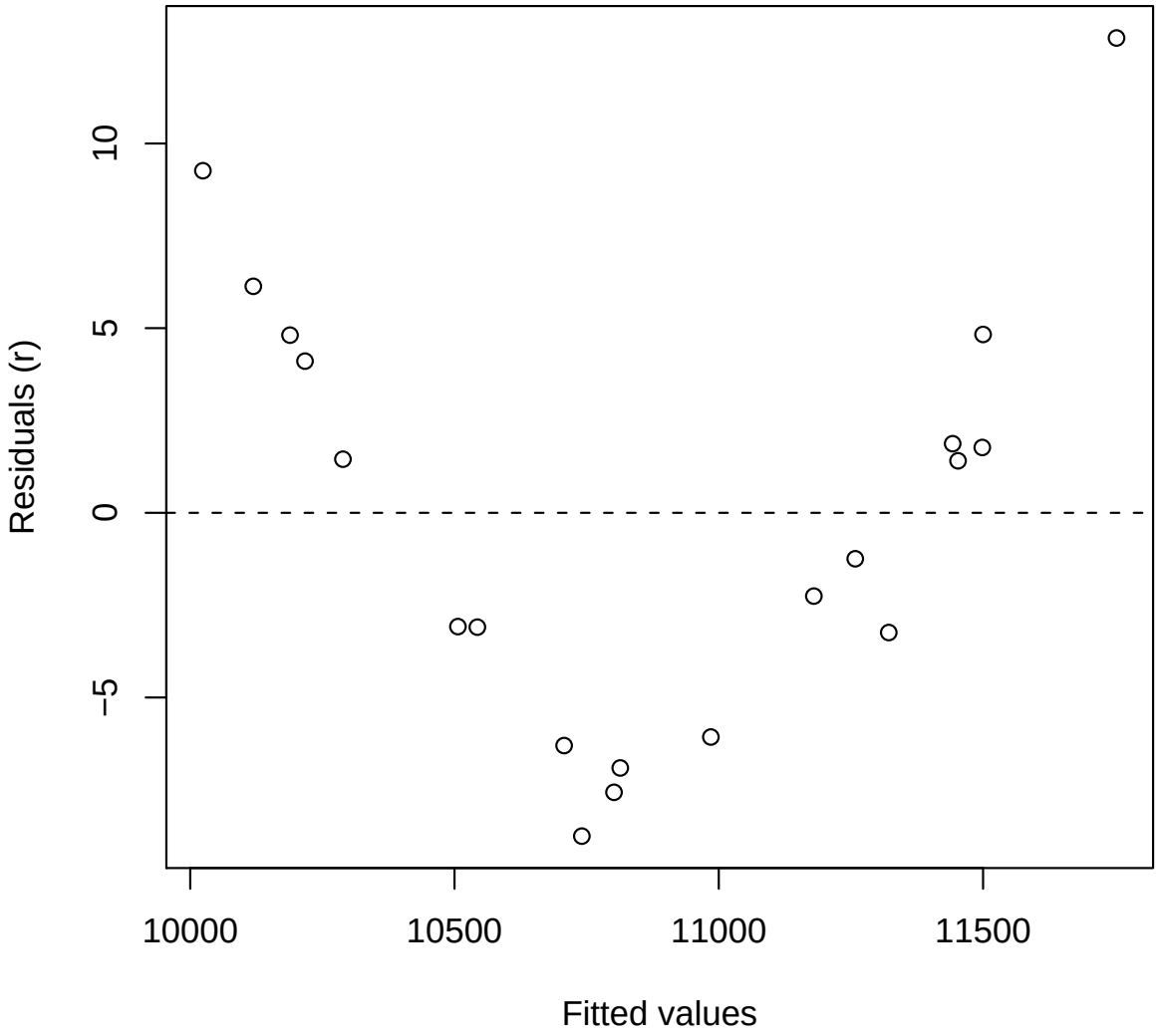
# Diagnostics

- The following three slides are prototypical residual plots indicating
  1. linear regression model is appropriate
  2. assumption of linearity questionable
  3. assumption of constant variance questionable

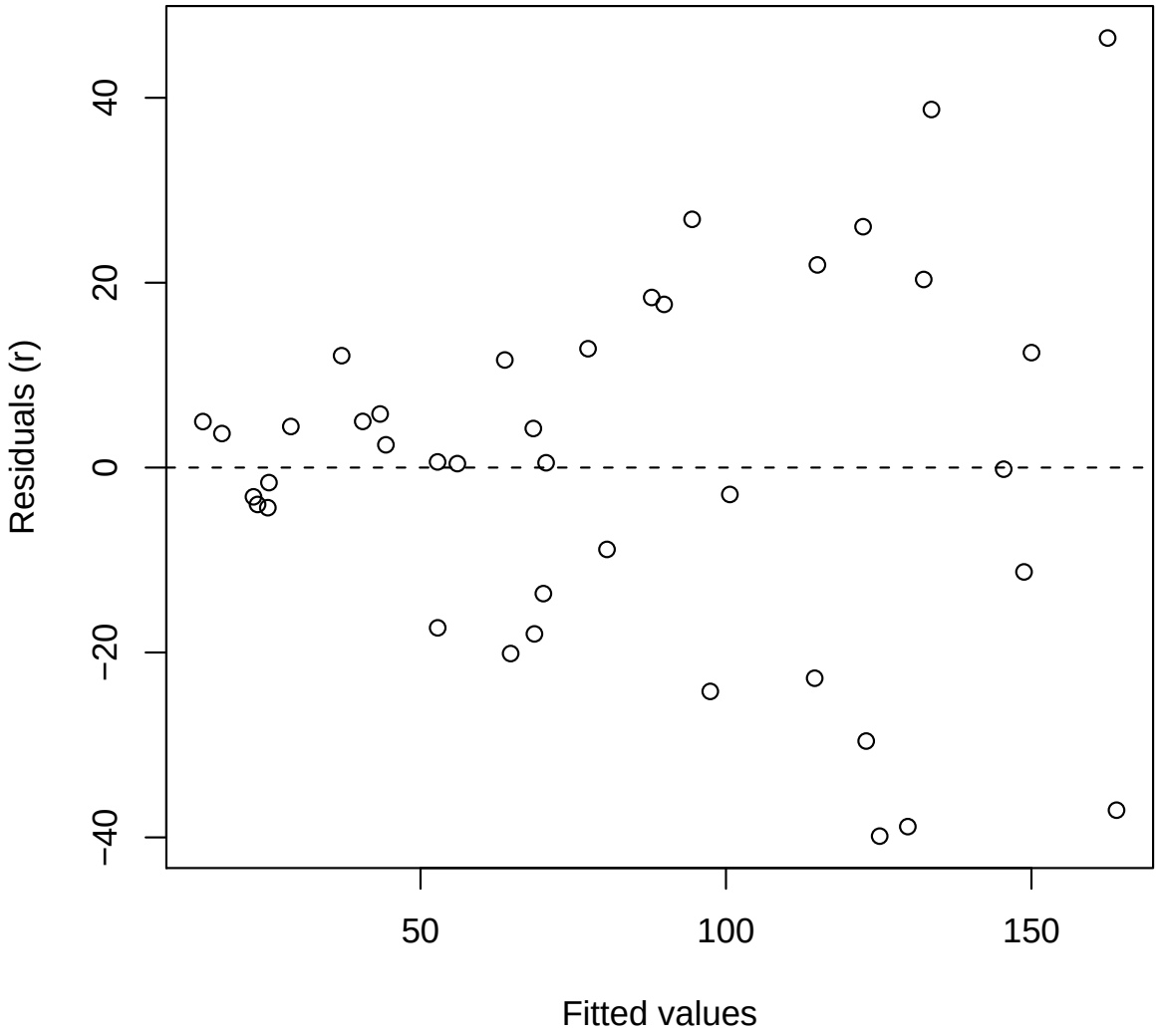
# Regression: Residuals



# Regression: Residuals



# Regression: Residuals



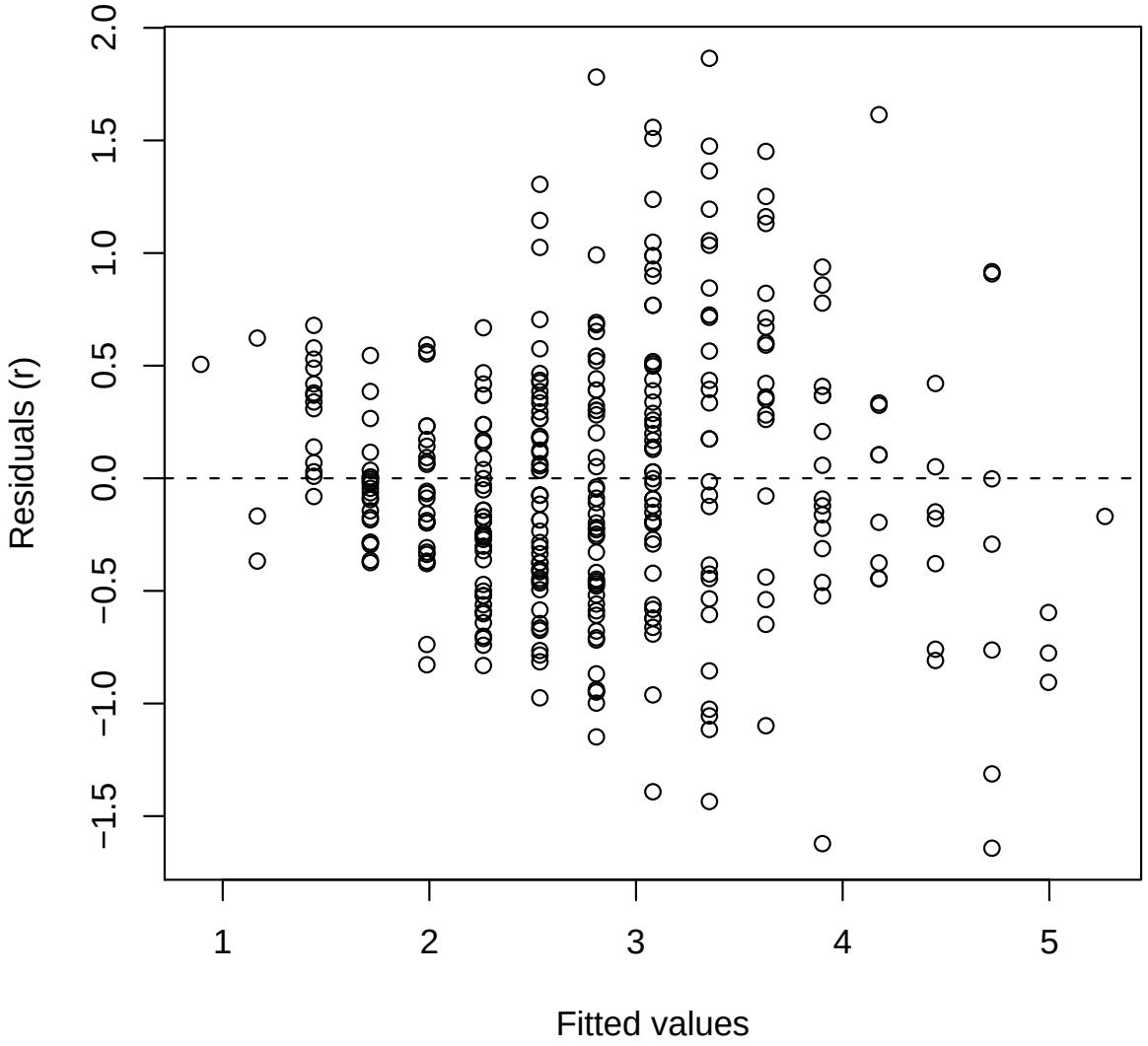
# Regression: Example

- FEV and age if sex=1

```
      Source |       SS       df       MS              Number of obs =     336
-------------+------------------------------           F(  1,   334) =  641.57
       Model |  221.896 397     1  221.896397           Prob > F      =  0.0000
    Residual |  115.518401    334  .345863477           R-squared     =  0.6576
-------------+------------------------------           Adj R-squared =  0.6566
       Total |  337.414798    335  1.00720835           Root MSE      =   .5881

------------------------------------------------------------------------------
         FEV |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2734776   .0107969    25.33   0.000     .2522391    .2947161
       _cons |   .0736006   .1127891     0.65   0.514    -.1482659    .2954671
------------------------------------------------------------------------------
```

# Regression: Example



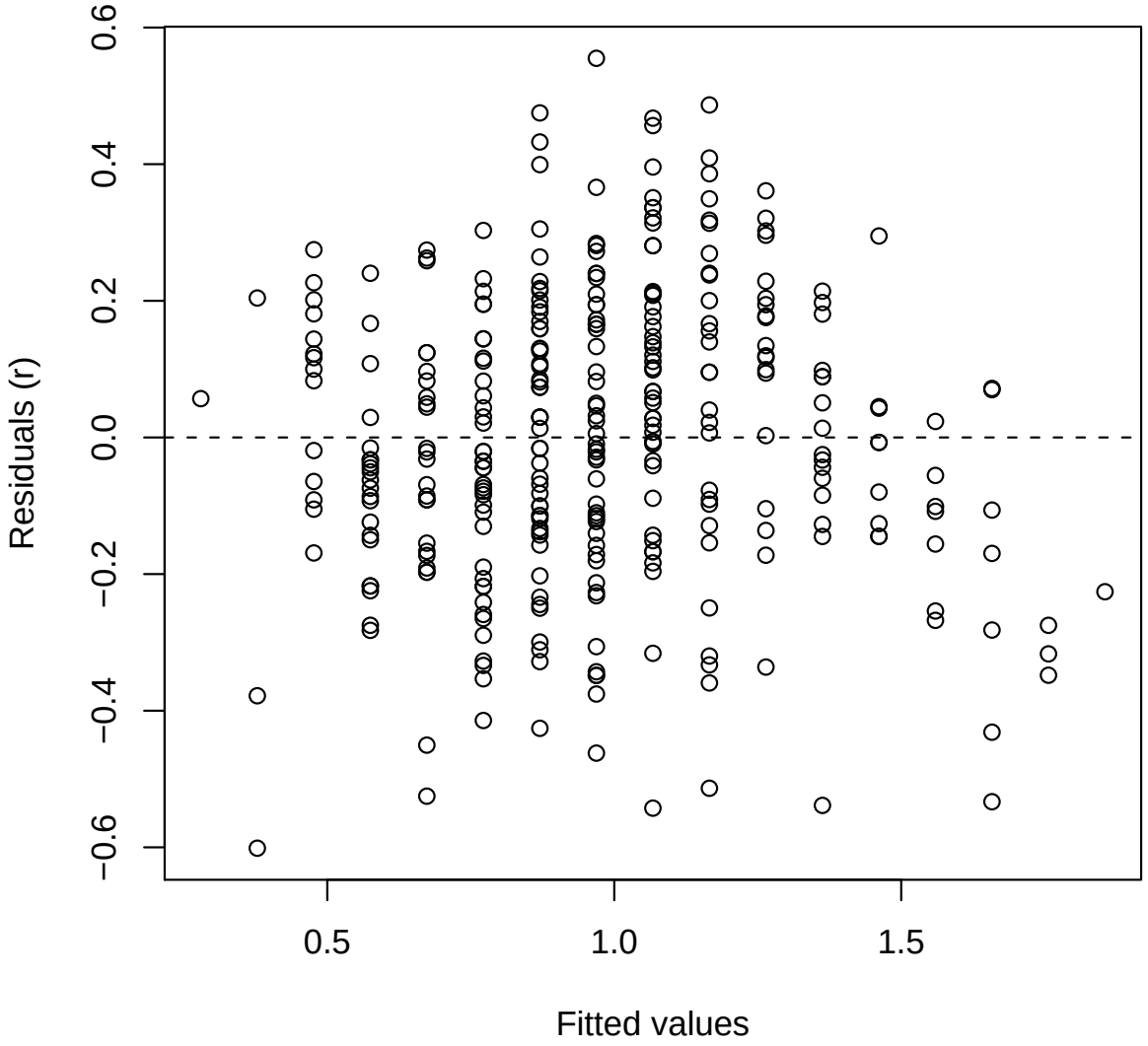
# Regression: Example

- Regress log(FEV) age if sex =1

```
      Source |       SS       df       MS              Number of obs =     336
-------------+------------------------------           F(  1,   334) =  651.53
       Model |  28.7636241     1  28.7636241           Prob > F      =  0.0000
    Residual |  14.7454315   334  .044147998           R-squared     =  0.6611
-------------+------------------------------           Adj R-squared =  0.6601
       Total |  43.5090556   335  .129877778           Root MSE      =  .21011

------------------------------------------------------------------------------
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .098462   .0038575    25.53   0.000      .090874      .10605
       _cons |  -.0156867   .0402968    -0.39   0.697    -.0949541    .0635808
------------------------------------------------------------------------------
```

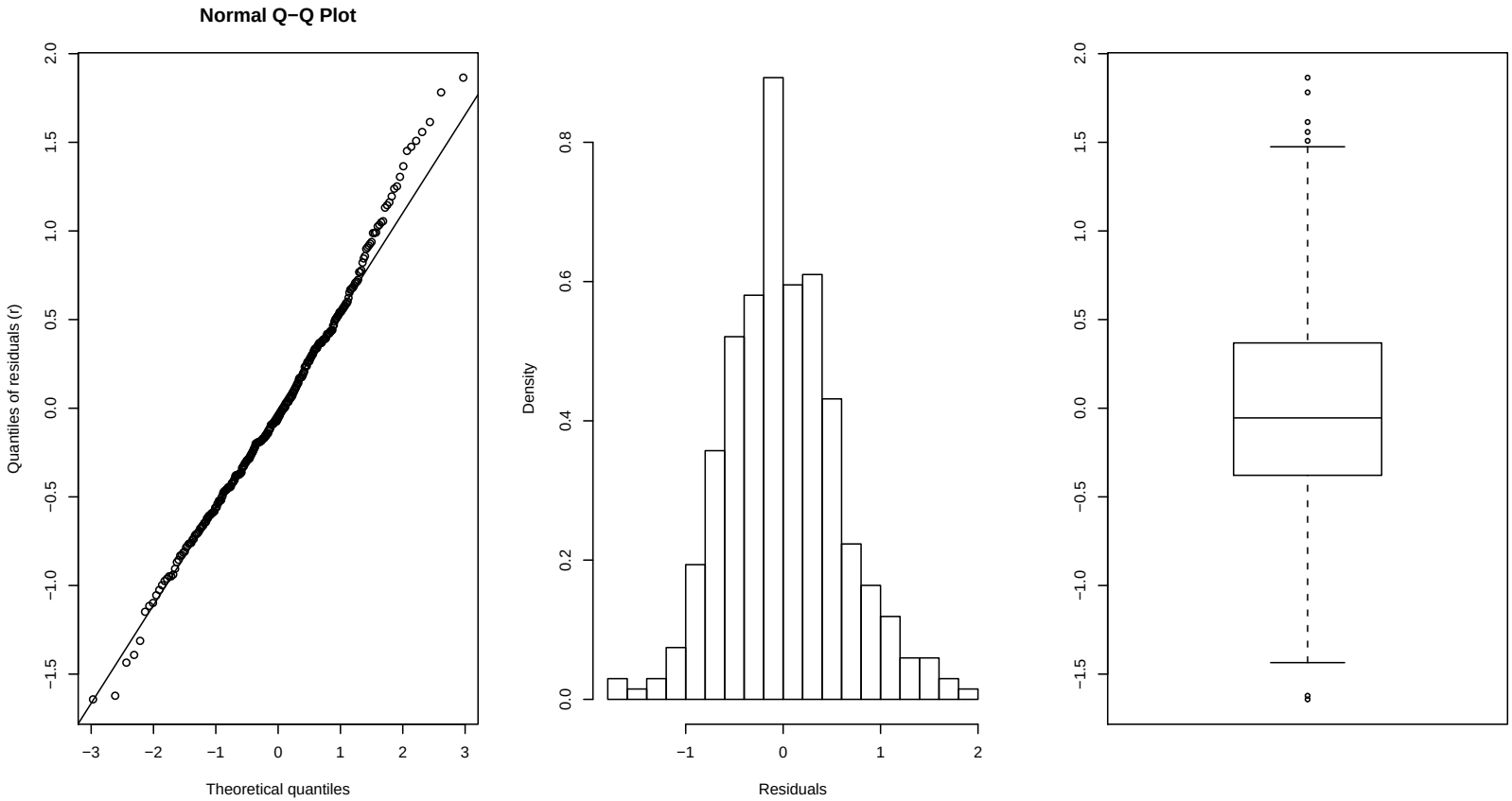
# Regression: Example



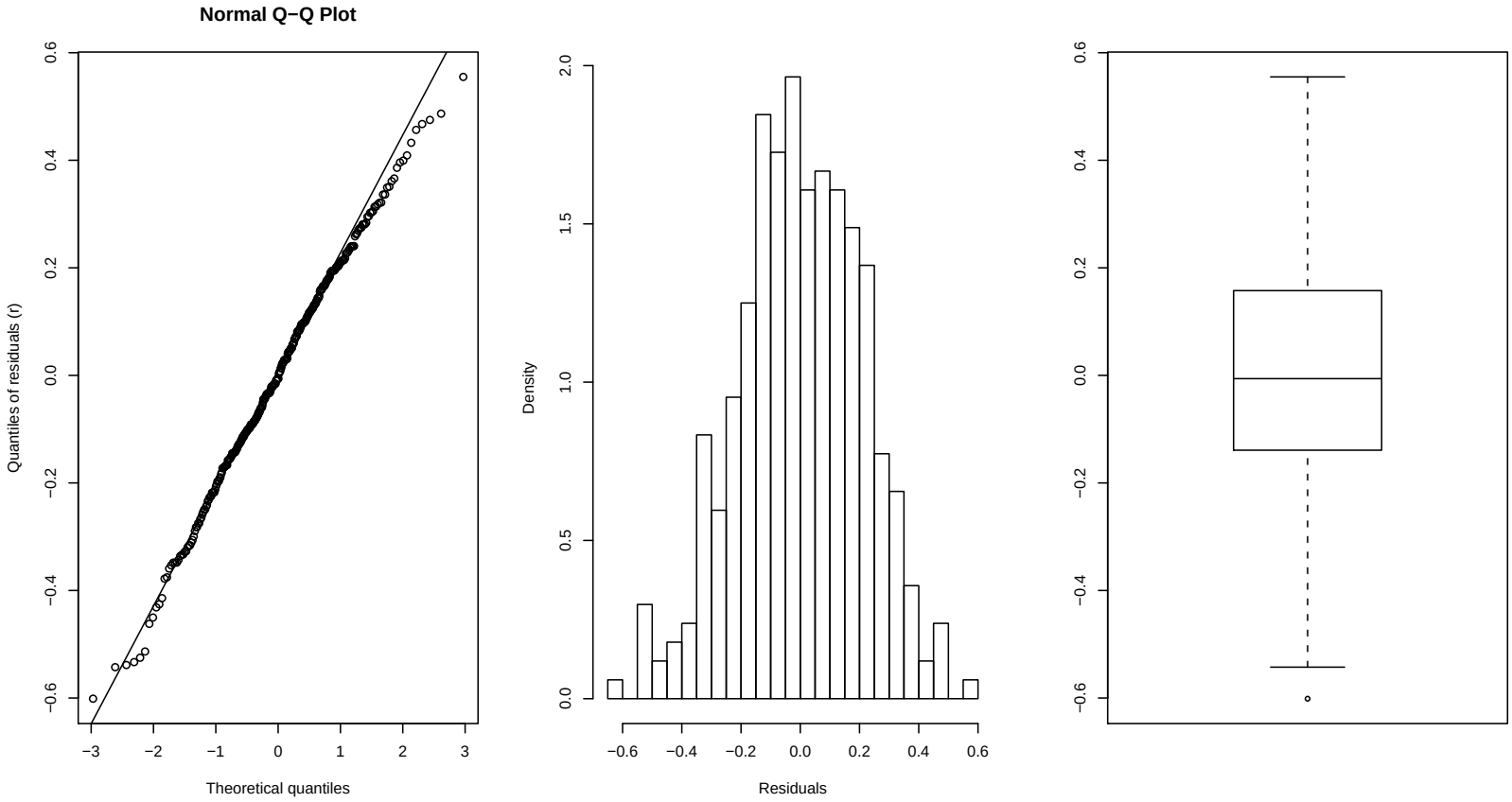
# Normality Diagnostics

- Assumption: $\epsilon_i$'s are normally distributed
- This assumption is not as important if $N$ is large (CLT)
- Inference robust to small departures from normality
- Violations of other assumptions can suggest non-normality
- Tests of normality on residuals; beware lack of power
- qq-plot, histogram, boxplot of residuals

# Normality Diagnostics: FEV

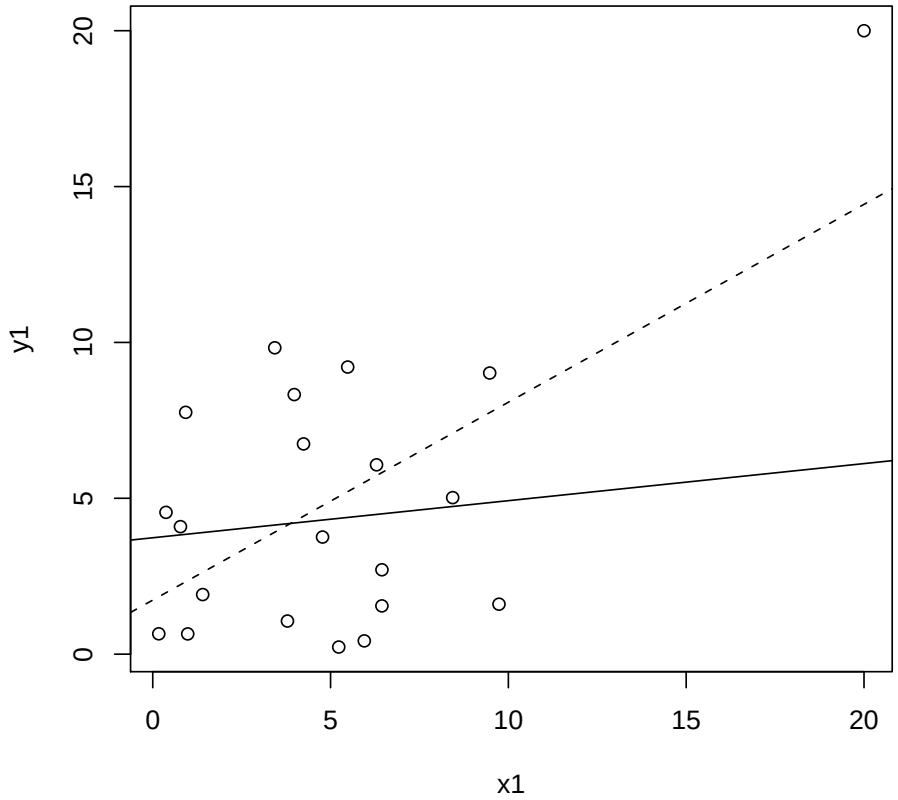


# Normality Diagnostics: log(FEV)



# Regression: Diagnostics

- Beware influential observations; always check scatterplot



# Remedial Measures

- Transformations, e.g., $\log(Y) = \alpha + \beta X$
- Multiple regression, e.g., $Y = \alpha + \beta_1 X + \beta_2 X^2$
- Nonparametric procedures, e.g., Kendall's tau
- More sophisticated models allowing for
  - dependenies/clusters (e.g., GEE)
  - heterogeneity of variance (e.g., weight least squares)

# Regression: $X$ random

- Assumption: $X$'s are known
- Suppose $X$ and $Y$ are both RVs

$$Y = \alpha + \beta_{Y.X} X + \epsilon$$

$$X \perp \epsilon; V(X) = \delta^2$$

- Results on estimation, testing, and prediction still hold (Neter et al 1996 p 85; Section 2.9.2 of Abraham and Ledolter 2006)

# Regression: $X$ random

- Now

$$\beta_{Y \cdot X} = \frac{Cov(Y, X)}{V(X)}$$

- Proof: recall

$$Cov(a + bW, U) = bCov(W, U)$$

and

$$Cov(W, U + V) = Cov(W, U) + Cov(W, V)$$

# Regression: $X$ random

- Thus

$$\begin{aligned} Cov(Y, X) &= Cov(\alpha + \beta_{Y.X} X + \epsilon, X) \\ &= \beta_{Y.X} Cov(X, X) + Cov(\epsilon, X) \\ &= \beta_{Y.X} V(X) \end{aligned}$$

# Measurement Error

- Instead of observing $X$, we observe

$$W = X + U$$

where $U$ is a RV with

$$E(U) = 0, V(U) = \tau^2$$

$$U \perp X, U \perp Y$$

- Then

$$Cov(W, Y) = Cov(X + U, Y)$$

$$= Cov(X, Y) + Cov(U, Y) = Cov(X, Y)$$

# Measurement Error

- By independence

$$V(W) = V(X) + V(U) = \delta^2 + \tau^2$$

- Thus

$$\begin{aligned}
\beta_{Y \cdot W} &= \frac{Cov(Y,W)}{V(W)} \\
&= \frac{Cov(Y,X)}{\delta^2+\tau^2} \\
&= \frac{\delta^2}{\delta^2+\tau^2}\frac{Cov(Y,X)}{\delta^2} \\
&= \frac{\delta^2}{\delta^2+\tau^2}\beta_{Y \cdot X}
\end{aligned}$$

# Measurement Error

- Since

$$0 \leq \frac{\delta^2}{\delta^2 + \tau^2} \leq 1,$$

  it follows

$$|\beta_{Y \cdot W}| \leq |\beta_{Y \cdot X}|$$

- Attenuation towards the null

# Measurement Error

- Thus if $X$ is not determined precisely, we underestimate the strength of association between $X$ and $Y$
- Reliability coefficient of $X$:

$$R_{el} = \frac{\delta^2}{\delta^2 + \tau^2}$$

- If $R_{el}$ is known,

$$\tilde{\beta} = R_{el}^{-1} \hat{\beta}_{Y \cdot W}$$

  is an unbiased estimator of $\beta_{Y \cdot X}$

# Measurement Error

- Since

$$V(\tilde{\beta}) = R_{el}^{-2} V(\hat{\beta}_{Y \cdot W})$$

the $t$-statistic for testing $H_0 : \beta_{Y \cdot X} = 0$ is

$$t_{Y \cdot X} = \frac{\tilde{\beta}}{\sqrt{V(\tilde{\beta})}} = \frac{R_{el}^{-1} \hat{\beta}_{Y \cdot W}}{\sqrt{R_{el}^{-2} V(\hat{\beta}_{Y \cdot W})}} = t_{Y \cdot W}$$

# Measurement Error

- Suppose there are $k$ measures of $W$ made on each person in the study
- It can be shown that

$$V(\bar{W}_k) = \delta^2 + \frac{\tau^2}{k}$$

- Therefore

$$\beta_{Y \cdot \bar{W}_k} = \frac{\delta^2}{\delta^2 + \tau^2/k} \beta_{Y \cdot X} \rightarrow \beta_{Y \cdot X} \text{ as } k \rightarrow \infty$$

# Measurement Error

- For example, suppose $W$ is a physiological variable such as BP or cholesterol
- If we get 2 or more measures of $W$, the bias will be reduced
- For cholesterol, $R_{el} \approx 0.8$ and $\delta^2 + \tau^2 \approx 1600$
- Therefore

$$\tau^2 = .2(1600) = 320$$

- If $k = 2$, $1280/(1280 + 320/2) = 0.89$

  If $k = 3$, $1280/(1280 + 320/3) = 0.92$

# Measurement Error

- Measurement error likely present in most situation, however usually ignored because:
  - Practically negligible (e.g., precise instrumentation)
  - Interest in inference/prediction based on observable random variables
- Random measurement error in $Y$ can be absorbed into $\epsilon$