BIOS 662 Fall 2018

Linear Regression, Part II

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

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

Analysis of Variance

- Recall that 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

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

- Thus

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

Analysis of Variance

- If $\beta = 0$ then

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

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

- Thus

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

Analysis of Variance

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

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

- For the two-sided alternative the F and t tests are equivalent

- For a 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

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

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

- Linear model

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

Matrix Formulation

- Equations (1) and (2) from previous set of notes:

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

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

- Equivalent to:

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

Matrix Formulation

- Therefore

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

- We can also show

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

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

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

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

Linear Regression and Two Sample t-test

- Define

$$X = \begin{cases} 1 & \text{if in group 1} \\ \\ 0 & \text{if in group 2} \end{cases}$$

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

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

Linear Regression and Two Sample t-test

- Suppose we have two groups of observations: Y_{1i} for $i = 1, \ldots, n_1$ and Y_{2i} for $i = 1, \ldots, n_2$
- Recall that the test statistic for the two sample t-test is

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

where

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

Linear Regression and Two Sample t-test

- Let

$$N = n_1 + n_2$$

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

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

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

Linear Regression and Two Sample t-test

- Consider the regression model:

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

- Note that

$$
\begin{aligned}
[X^2] &= \sum_i (X_i - \bar{X})^2 = \sum X_i^2 - N\bar{X}^2 \\
&= n_1 - N \left(\frac{n_1}{N}\right)^2 \\
&= n_1 \left(1 - \frac{n_1}{N}\right) \\
&= \frac{n_1 n_2}{N}
\end{aligned}
$$

Linear Regression and Two Sample t-test

- Recall that

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

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

- Thus

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

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

- We can show that

$$s^2_{y \cdot x} = s^2_p$$

Linear Regression and Two Sample t-test

- Therefore:

$$
\begin{aligned}
t &= \frac{\hat{\beta}}{\sqrt{s^2_{y\cdot x} / \sum_i (X_i - \bar{X})^2}} \\
&= \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{N/(n_1 n_2)}}
\end{aligned}
$$

Linear Regression and Two Sample t-test

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

Linear Regression and Two Sample t-test

- Two sample t-test:

```
proc ttest;
   var pbf;
   class tribe;

The TTEST Procedure

Variable:  pbf

tribe            N        Mean     Std Dev     Std Err

Apache          35     33.1757      6.9215      1.1700
Tohona          63     37.3615      8.0349      1.0123
Diff (1-2)             -4.1857      7.6591      1.6147

Method           Variances        DF    t Value    Pr > |t|

Pooled           Equal            96      -2.59      0.0110
Satterthwaite    Unequal      79.523      -2.71      0.0083
```

Linear Regression and Two Sample t-test

- Model

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

where

$$Y = \text{PBF}$$

and

$$X = \begin{cases} 1 & \text{if Apache} \\ \\ 0 & \text{if Tohona} \end{cases}$$

Linear Regression and Two Sample t-test

```
proc reg;
   model pbf=apache;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	394.20974	394.20974	6.72	0.0110
Error	96	5631.59441	58.66244		
Corrected Total	97	6025.80415			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	37.36147	0.96496	38.72	<.0001
apache	1	-4.18574	1.61469	-2.59	0.0110

Diagnostics

- Assumptions for linear regression
  1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
  2. Xs are fixed constants
  3. ϵ_i iid $\sim N(0, \sigma^2)$
     (*homogeneity of variance*)
- *Residual plot*: Scatterplot of

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

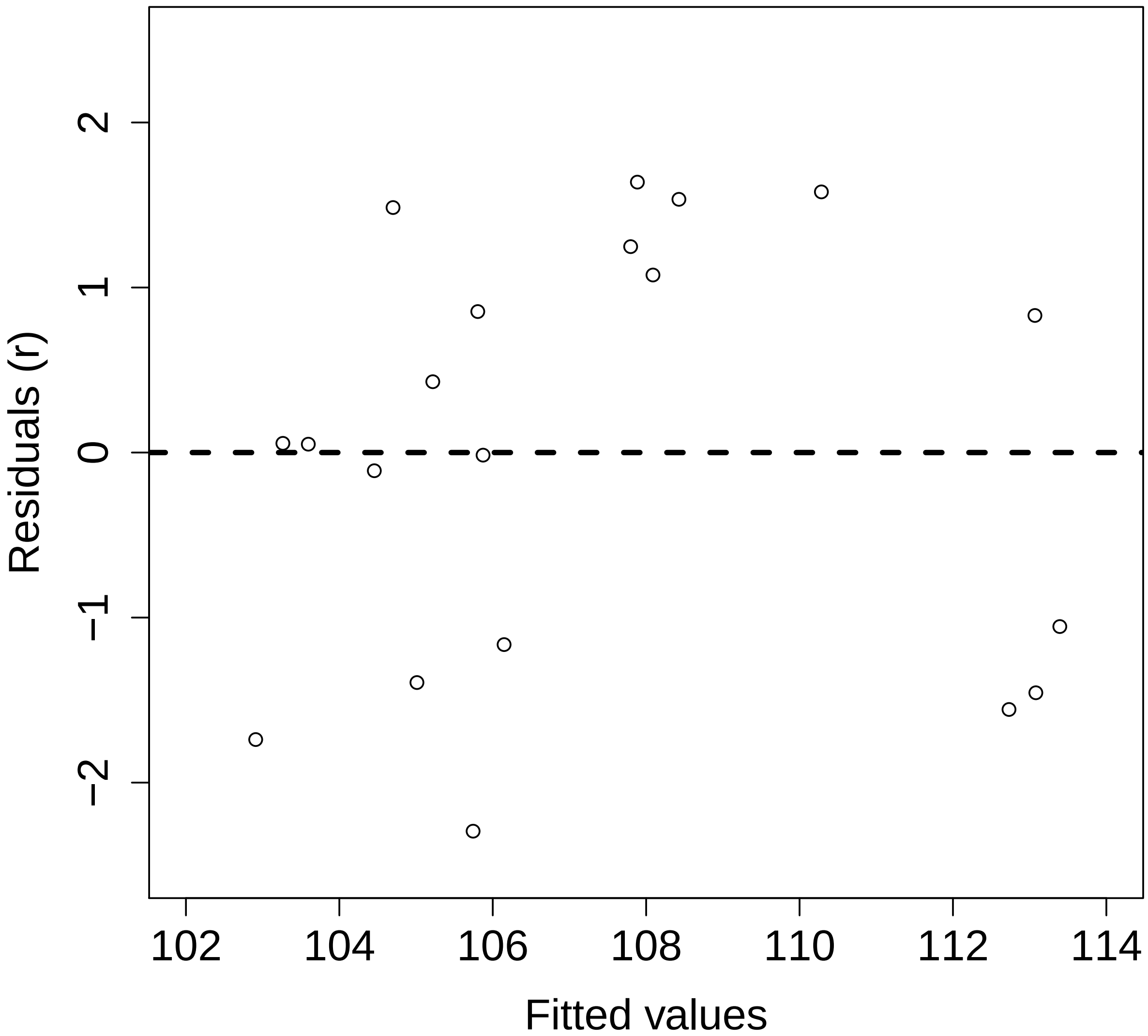
- If we see lack of homogeneity of variance or of linearity, consider transformations; see Table 10.28 (page 399) of the text

Diagnostics

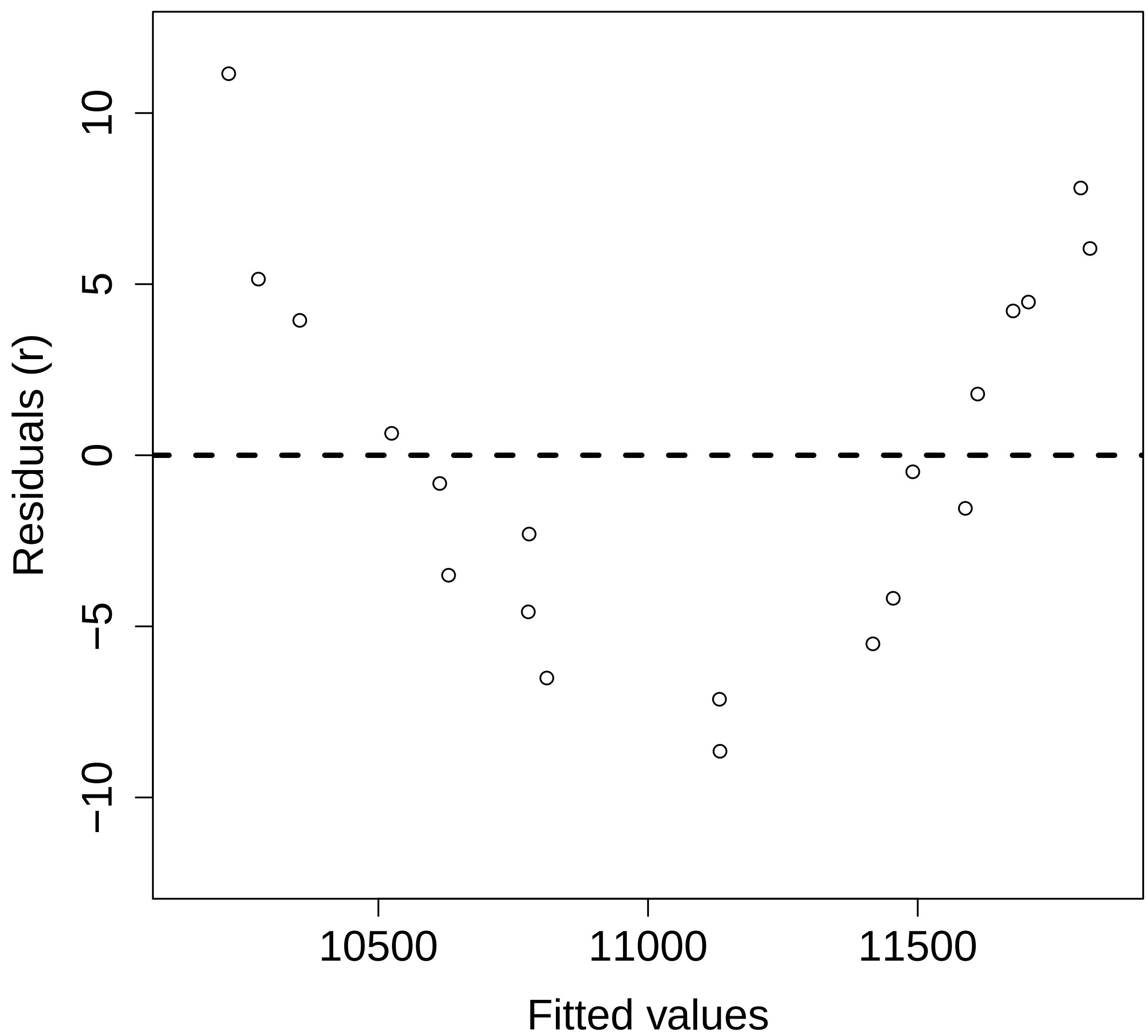
- The following three pages contain prototypical residual plots indicating successively:

  1. linear regression model is appropriate
  2. assumption of linearity questionable
  3. assumption of constant variance questionable

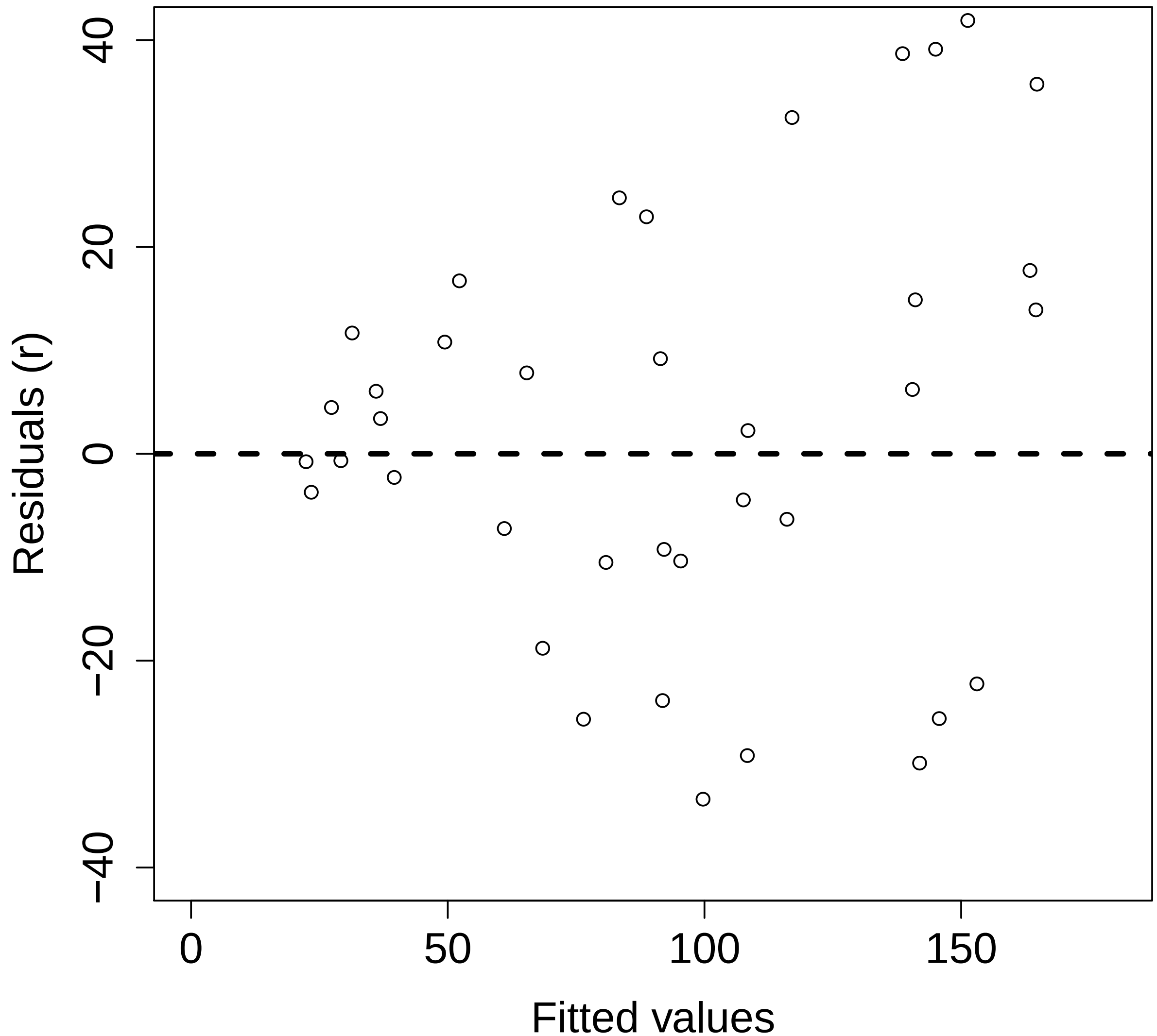
Regression: Residuals



Regression: Residuals



Regression: Residuals



Regression: Example

- FEV_1 as a function of age in male children

```
proc reg;
   model fev1=age;
```

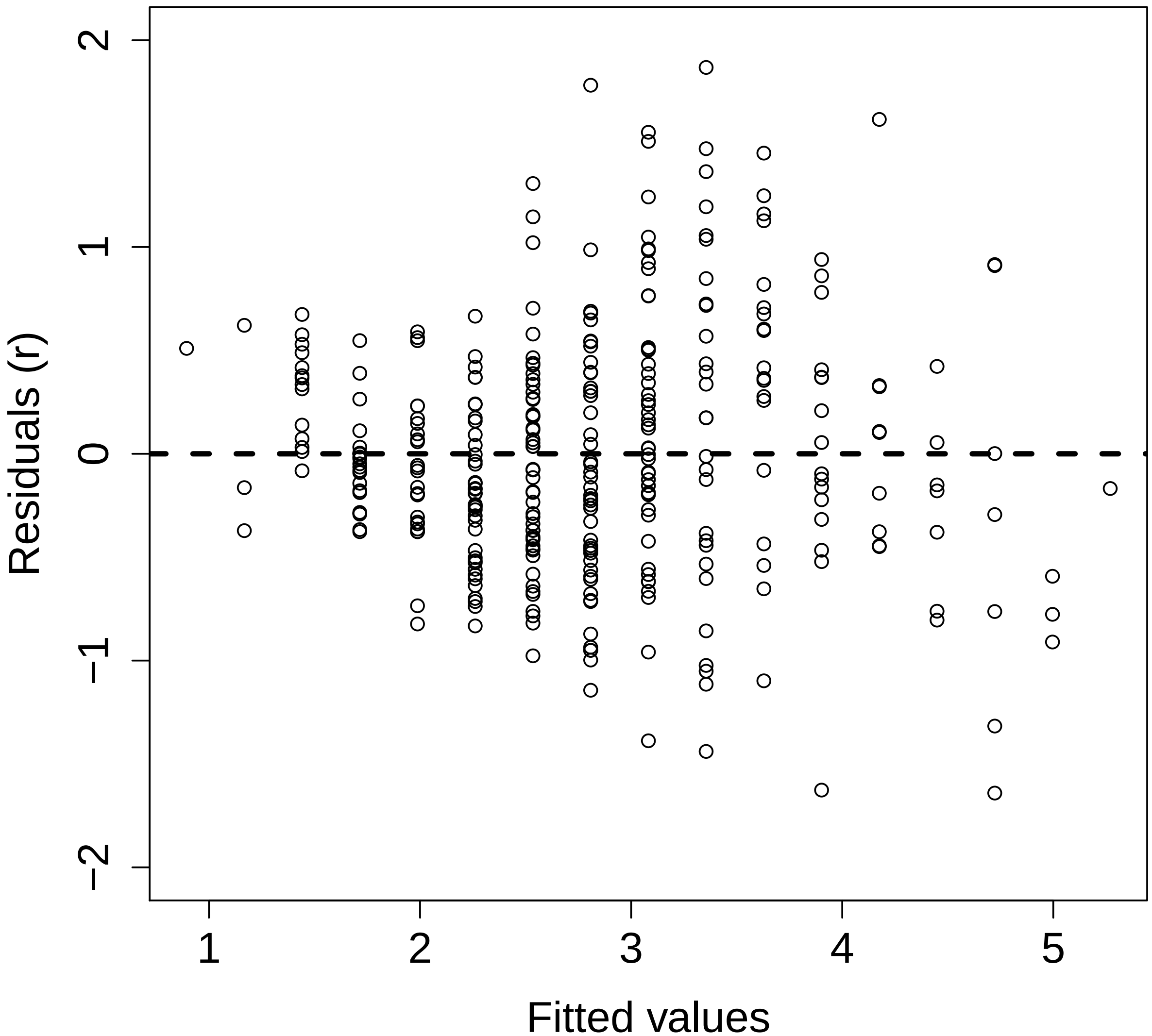
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	221.89640	221.89640	641.57	<.0001
Error	334	115.51840	0.34586		
Corrected Total	335	337.41480			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.07360	0.11279	0.65	0.5145
age	1	0.27348	0.01080	25.33	<.0001

Regression: Example cont.



Regression: Example cont.

- Regress $\log(\text{FEV}_1)$ on age for male children

```
proc reg;
   model logfev1=age;
```

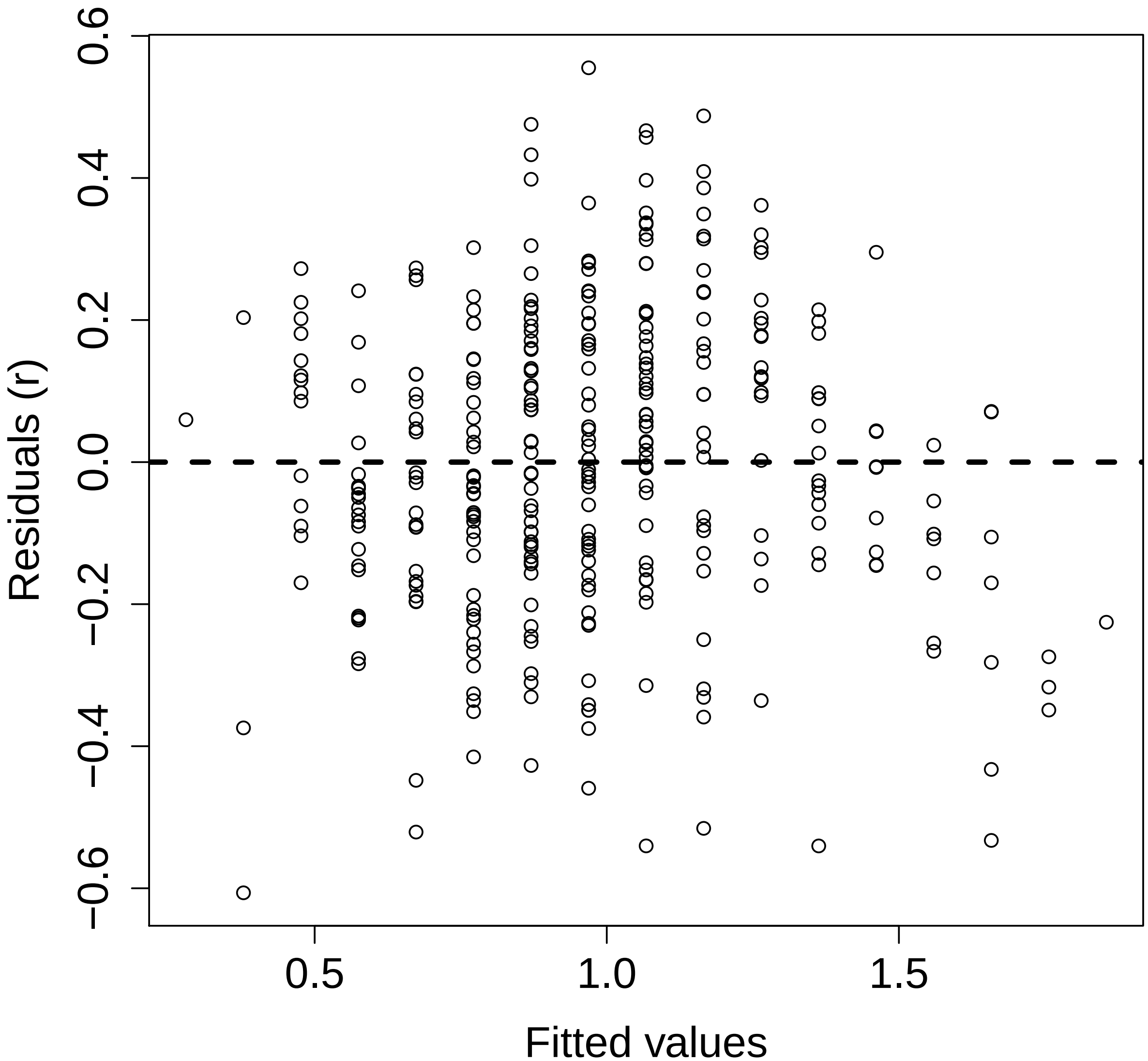
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	28.76362	28.76362	651.53	<.0001
Error	334	14.74543	0.04415		
Corrected Total	335	43.50906			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.01569	0.04030	-0.39	0.6973
age	1	0.09846	0.00386	25.53	<.0001

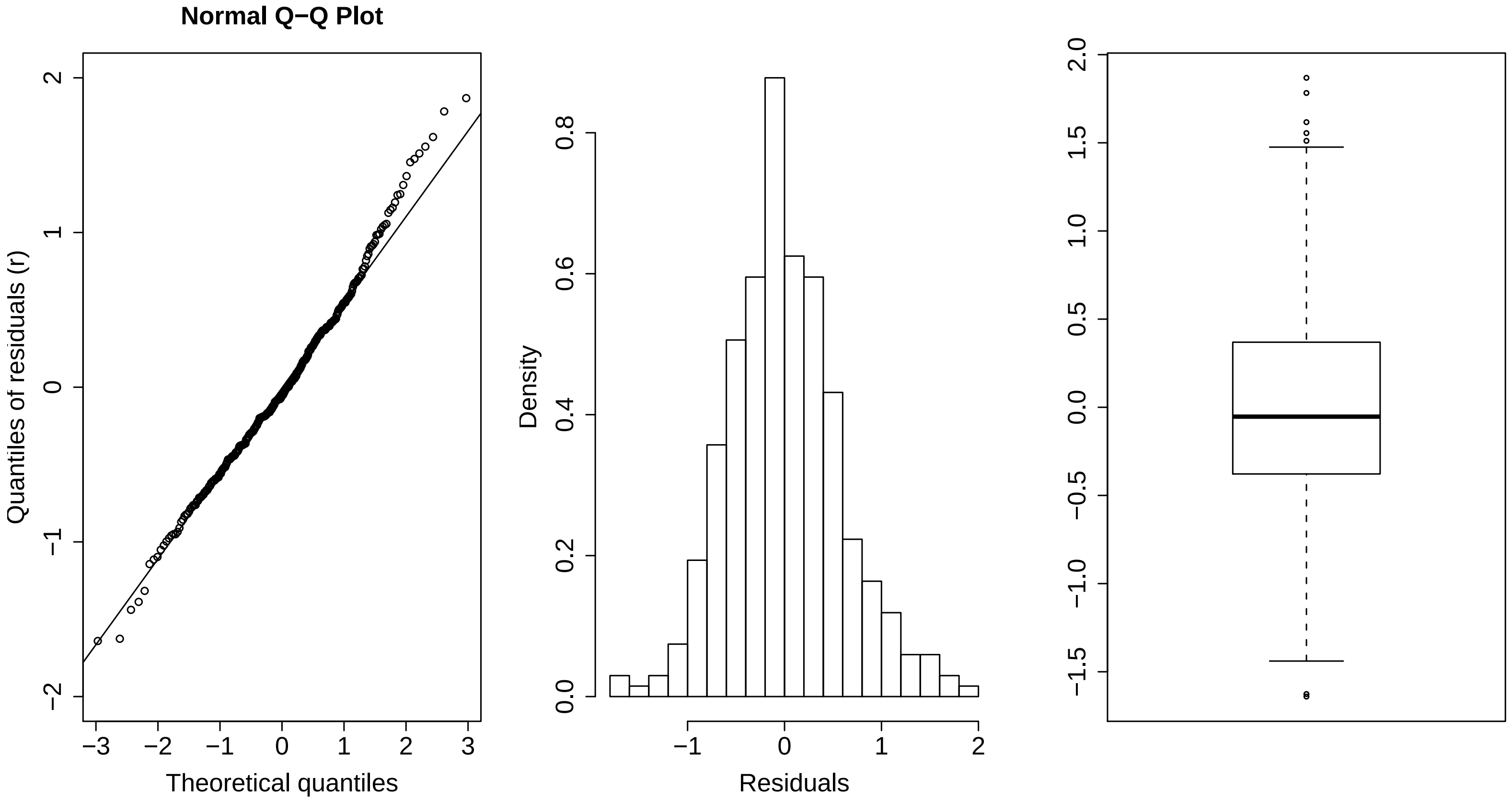
Regression: Example cont.



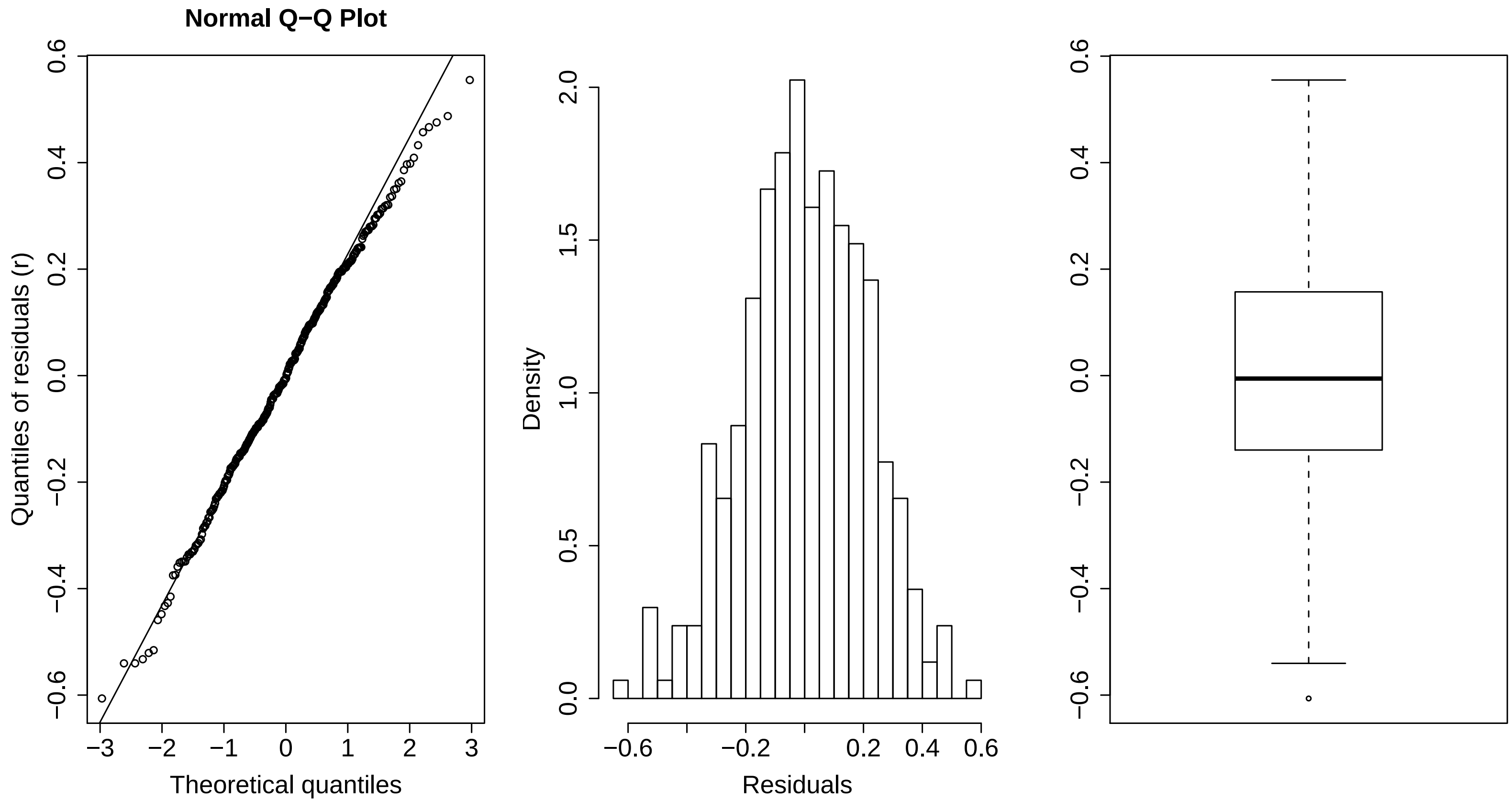
Normality Diagnostics

- Assumption: The ϵ_i 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 of residuals; beware lack of power
- qq-plot, histogram, boxplot of residuals

Normality Diagnostics: FEV_1

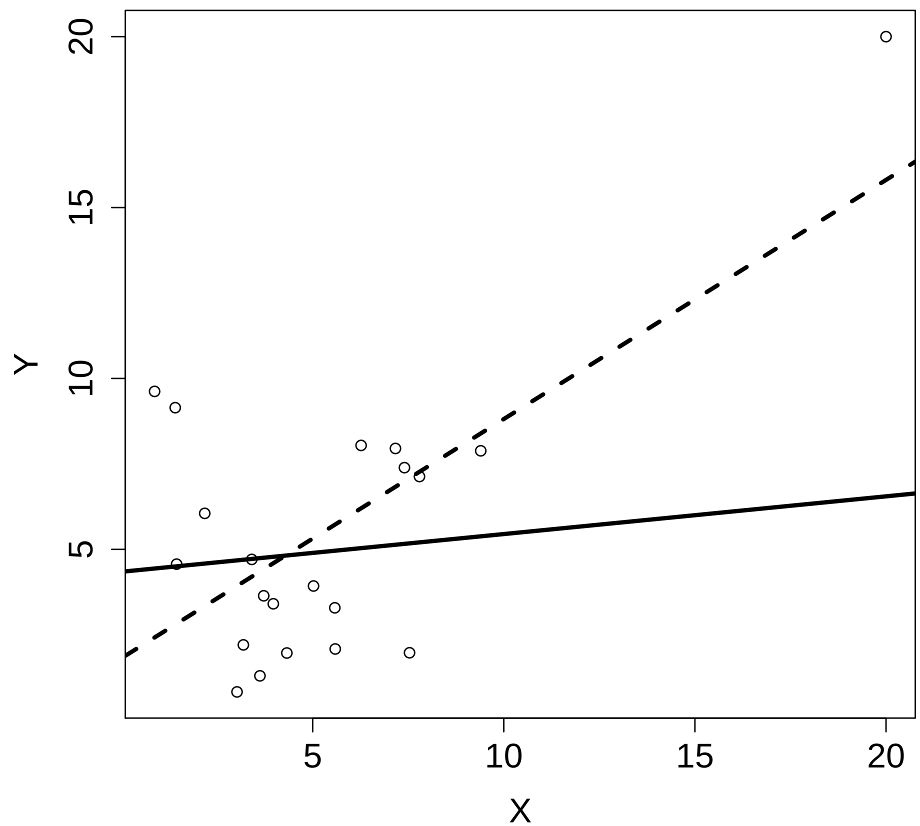


Normality Diagnostics: $\log(\text{FEV}_1)$



Regression: Diagnostics

- Beware influential observations; always check scatterplot

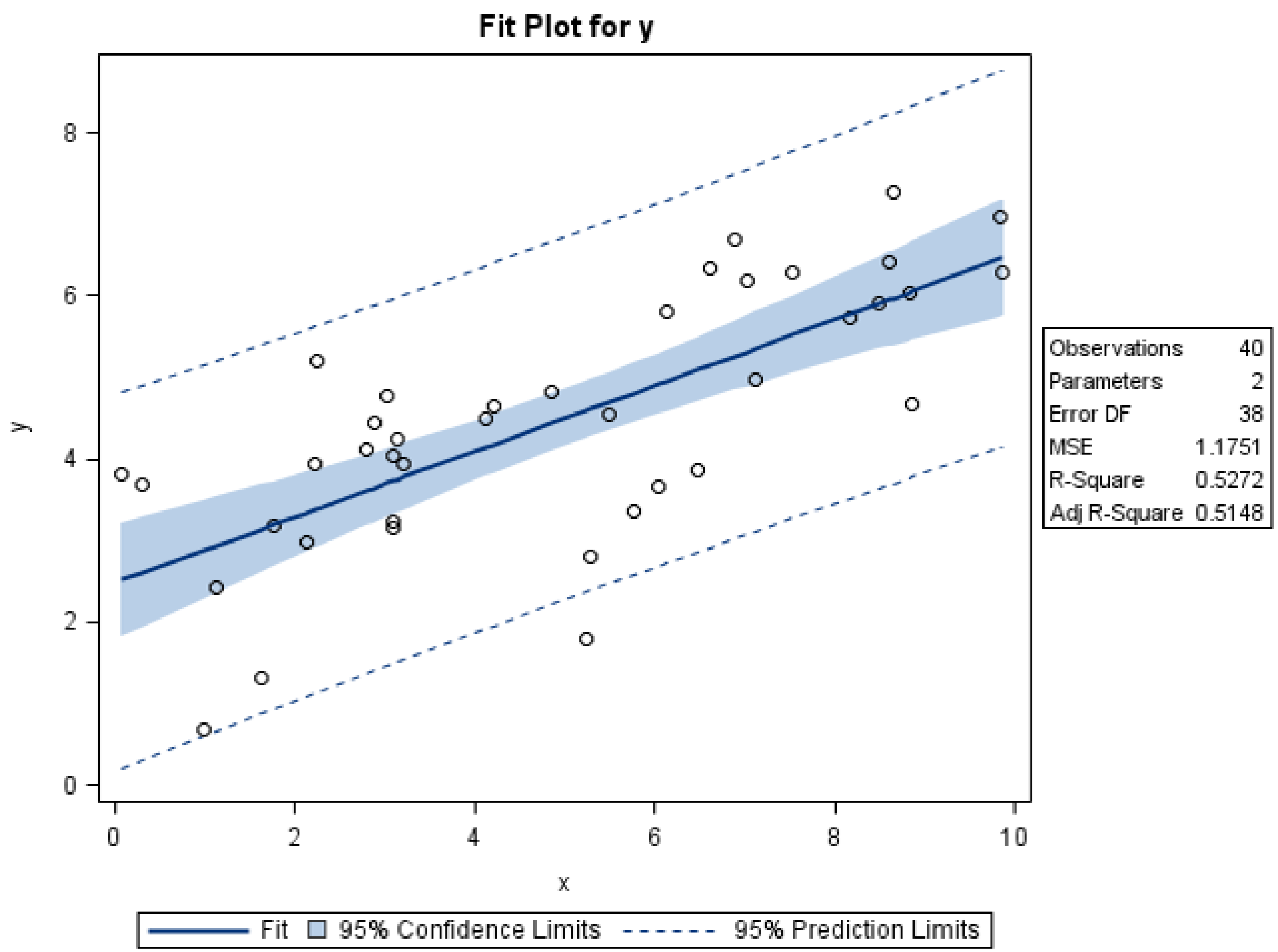


Regression: Graphical Diagnostics in SAS

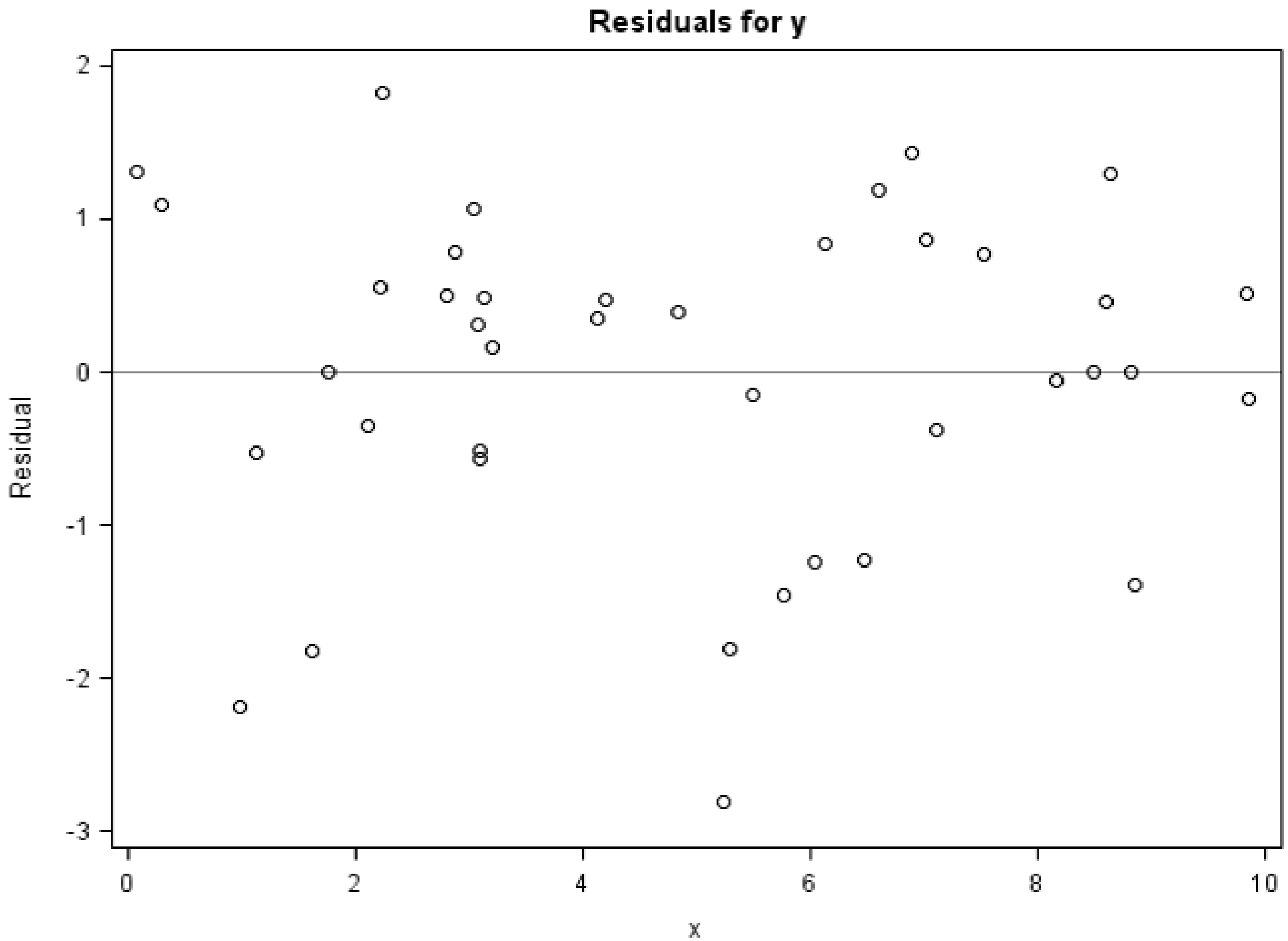
- Use ODS graphics in SAS 9.2 or later
- Default plots often sufficient, use options in plots= to specify particular plots

```
ods graphics on;
ods rtf file='diagnostics.rtf';

proc reg data=diagnostics;
   model y = x;

run; quit;
ods rtf close;
```

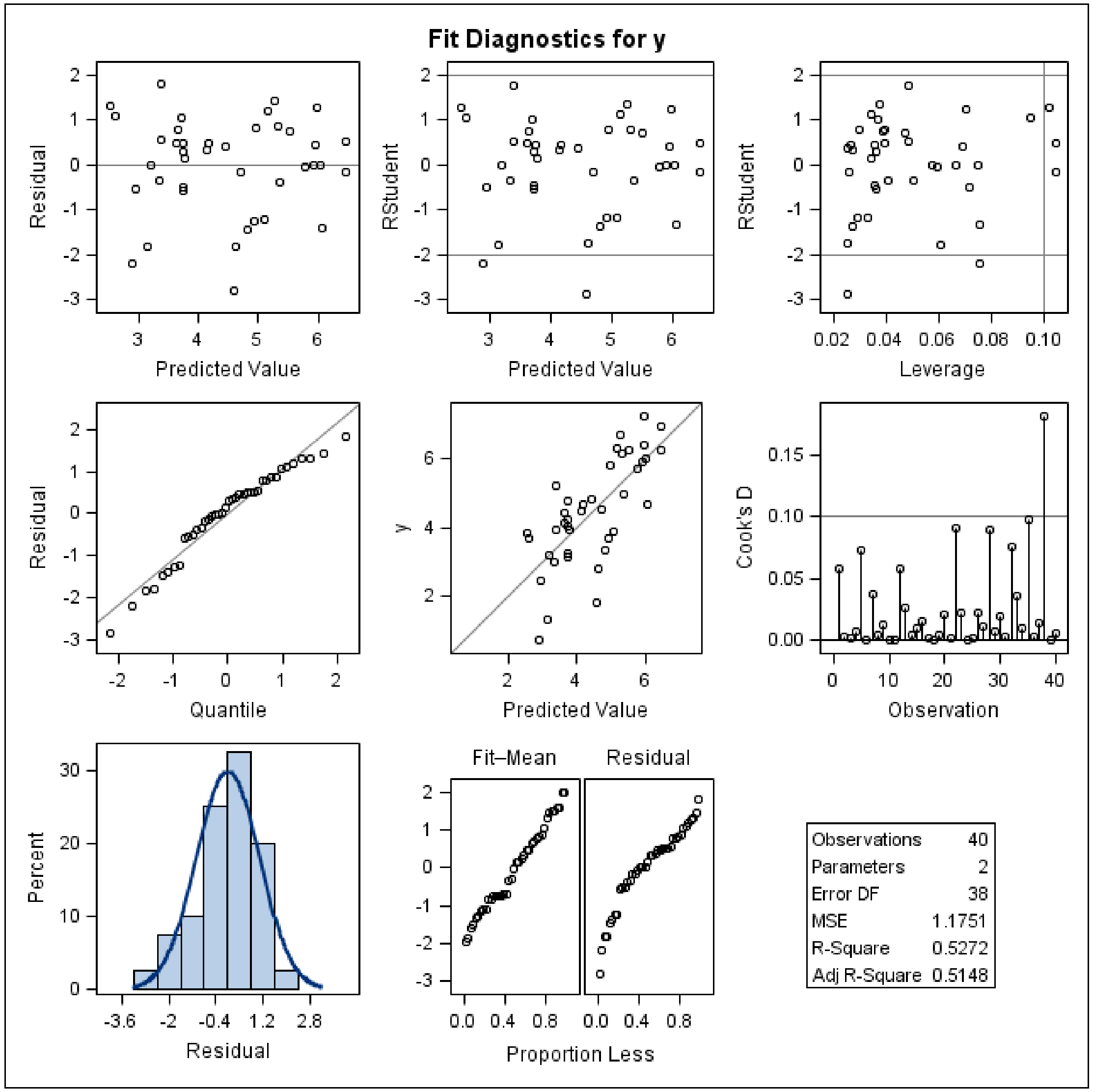
Regression: Graphical Diagnostics in SAS



Regression: Graphical Diagnostics in SAS



Regression: Graphical Diagnostics in SAS



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
  - dependencies/clusters (e.g., GEE)
  - heterogeneity of variance (e.g., weighted least squares)

Regression: X Random

- Assumption: Xs are known
- Suppose X and Y are both random variables

$$Y = \alpha + \beta_{y \cdot x} X + \epsilon$$

$$X \perp \epsilon; \;\; \mathrm{Var}(X) = \delta^2$$

- Results on estimation, testing, and prediction still hold (Neter et al., 1996 p 85; Section 2.9.2 of Abraham and Ledolter, 2006)
- The covariance between two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}$$

Regression: X Random

- Now

$$\beta_{y\cdot x} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

- Proof: We have $\text{Cov}(a + bW, U) = b\,\text{Cov}(W, U)$ and $\text{Cov}(W, U + V) = \text{Cov}(W, U) + \text{Cov}(W, V)$

- Thus

$$\begin{aligned}
\text{Cov}(Y, X) &= \text{Cov}(\alpha + \beta_{y\cdot x}X + \epsilon, X) \\
&= \beta_{y\cdot x}\text{Cov}(X, X) + \text{Cov}(\epsilon, X) \\
&= \beta_{y\cdot x}\text{Var}(X)
\end{aligned}$$

Measurement Error

- Instead of observing X, we observe

$$W = X + U$$

where U is a random variable with

$$E(U) = 0, \quad \mathrm{Var}(U) = \tau^2$$

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

- Then

$$\begin{aligned}
\mathrm{Cov}(W, Y) &= \mathrm{Cov}(X + U, Y) \\
&= \mathrm{Cov}(X, Y) + \mathrm{Cov}(U, Y) = \mathrm{Cov}(X, Y)
\end{aligned}$$

Measurement Error

- By independence

$$\text{Var}(W) = \text{Var}(X) + \text{Var}(U) = \delta^2 + \tau^2$$

- Thus

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

Measurement Error

- Because

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

it follows that

$$|\beta_{y \cdot w}| \leq |\beta_{y \cdot x}|$$

- That is, there is 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:

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

- If R_{rel} is known,

$$\tilde{\beta} = R_{\text{rel}}^{-1}\,\hat{\beta}_{y \cdot w}$$

  is an unbiased estimator of $\beta_{y \cdot x}$

Measurement Error

- Because

$$\mathrm{Var}(\tilde{\beta}) = R_{\mathrm{rel}}^{-2}\,\mathrm{Var}(\hat{\beta}_{y\cdot w})$$

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

$$t_{y\cdot x} = \frac{\tilde{\beta}}{\sqrt{\mathrm{Var}(\tilde{\beta})}} = \frac{R_{\mathrm{rel}}^{-1}\hat{\beta}_{y\cdot w}}{\sqrt{R_{\mathrm{rel}}^{-2}\,\mathrm{Var}(\hat{\beta}_{y\cdot w})}} = t_{y\cdot w}$$

Measurement Error

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

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

- Therefore

$$\beta_{y \cdot \bar{w}_k} = \frac{\delta^2}{\delta^2 + \tau^2/k} \beta_{y \cdot x} \rightarrow \beta_{y \cdot x} \quad \text{as} \quad k \rightarrow \infty$$

Measurement Error

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

$$\tau^2 = 0.2(1600) = 320$$

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

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

Measurement Error

- Measurement error is likely to be present in most situations; however, it is usually ignored because:
  - Often practically negligible (e.g., if can use precise instrumentation)
  - Interest is in inference/prediction based on observable random variables
- Random measurement error in Y is absorbed into ϵ