BIOSTAT 651

Notes #2: Linear regression review

- Lecture Topics:
  - Review of linear regression
  - Weighted least squares

Linear regression

- response: Y_i
- covariate: $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \cdots, x_{ip})$
- i: index of subject
- n: total number of subjects in the data

Model

- Linearly relate predictors to the mean response (assume X is deterministic)
- For i-th subject,

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \\
&= \mathbf{x}_i{}^T \boldsymbol{\beta} + \epsilon_i,
\end{aligned}
$$

where $\epsilon_i \sim N(0, \sigma^2)$.

- Matrix form:
  - set $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$
  - design matrix:

$$\mathbf{X} \;=\; \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

- Model:

$$\mathbf{Y} \;=\; \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \;\sim\; MVN_n(\mathbf{0}, \sigma^2 I)$$

i.e.

$$\mathbf{Y} \;\sim\; MVN_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I).$$

Model

- Assumptions:
  - Systematic component: predictor effect through linear regression on the mean (*linearity assumption*)

$$E[Y_i|\mathbf{x}_i] \;=\; \mu_i \;=\; \mathbf{x}_i^T\boldsymbol{\beta}$$

  - Random component: at each level of the predictor, variation in the response is characterized as

$$N(0, \sigma^2)$$

  - Independence (between subjects)

Interpretation

- In simple linear regression:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i,$$

  - β_1: change in mean response per unit increase of x_1.

$$\beta_1 = E[Y_i | x_{i1} = a + 1] - E[Y_i | x_{i1} = a]$$

- Multiple linear regression: need to adjust for other covariates (or holding them constant)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

  - β_1: change in mean response per unit increase of x_1, adjusting for x_2 (holding x_2 constant)

$$\begin{aligned} \beta_1 &= E[Y_i | x_{i1} = a + 1, x_{i2} = c] \\ &\quad - E[Y_i | x_{i1} = a, x_{i2} = c] \end{aligned}$$

Parameter Estimation

- Least Squares Estimation (LSE): minimize the sum of squared errors

$$\min_{\boldsymbol{\beta}} \ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Estimator:

$$\widehat{\boldsymbol{\beta}} \ = \ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

where $\mathbf{X}$ is of full rank.

$$\widehat{\boldsymbol{\beta}} \; = \; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

- $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

$$E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

- Variance of $\widehat{\boldsymbol{\beta}}$:

$$V(\widehat{\boldsymbol{\beta}}) \; = \; \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

- σ^2 estimator:

$$\widehat{\sigma}^2 \; = \; \frac{1}{n-p-1}SSE \; = \; \frac{1}{n-p-1}\mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

- Residual:

$$\mathbf{e} \; = \; \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}.$$

Analysis of Variance

- ANOVA
  - SST : total variation of $\mathbf{Y}$ around mean

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \mathbf{Y}^T(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{Y}$$

  - SSR : variation of $\mathbf{Y}$ explained by regression

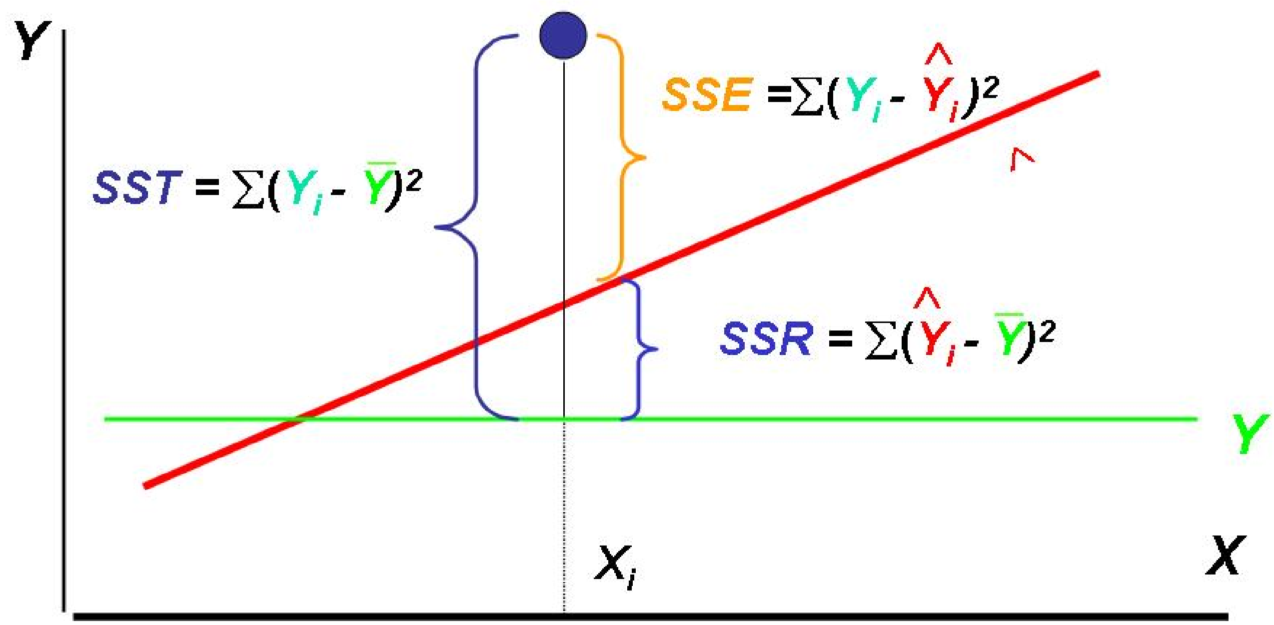
$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \mathbf{Y}^T(\mathbf{H} - \mathbf{1}\mathbf{1}^T/n)\mathbf{Y}$$

  - SSE : variation of $\mathbf{Y}$ unexplained by regression

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

  - $SST \;=\; SSR + SSE$

- Sum of squares



[http://www.trizsigma.com/regression.html]

- ANOVA table

Source	SS	DF	MS	F
Regression	SSR	p	SSR/p	
Error	SSE	n-p-1	SSE/(n-p-1)	
Total	SST	n-1	SST/(n-1)	

- R^2: explained sum of squares over total sum of square

$$R^2 = SSR/SST$$

Example: Child obesity data

- Example: A study on childhood obesity examined the relationship between a child's weight, age and exposure to pre-natal smoke.
  - Response: weight (kg)
  - Predictors: age, pre-natal smoke
- Linear regression model

$$
\begin{aligned}
Y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\
&= \beta_0 + \beta_1 A_i + \beta_2 S_i + \beta_3 A_i \times S_i + \epsilon_i
\end{aligned}
$$

where

- A_i: Age in years
- S_i: =1 if exposed; =0 if unexposed
- $A_i \times S_i$: interaction term

Example: Child obesity data

```
DATA Weights;
INPUT id wt age smoke obesity;
wt_kg = wt / 1000;
A_S = age * smoke;
datalines;
1 22509.41 7 0 1
2 33452.27 7 1  0
3 13380.91 3 0  1
4 24947.45 8 1  1
5 15875.65 4 1 1
. . .

proc reg;
model wt_kg = age smoke A_S;
run;
```

Linear regression for continuous responses

The REG Procedure
Model: MODEL1
Dependent Variable: wt_kg

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	973.37997	324.45999	6.74	0.0016
Error	26	1250.94969	48.11345		
Corrected Total	29	2224.32967			

Root MSE	6.93639	**R-Square**	0.4376
Dependent Mean	23.22381	**Adj R-Sq**	0.3727
Coeff Var	29.86756		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	4.95865	6.65202	0.75	0.4627
age	1	2.87548	1.05916	2.71	0.0116
smoke	1	0.52160	8.58019	0.06	0.9520
A_S	1	0.20133	1.37335	0.15	0.8846

Hypothesis Test

- Test whether βs or linear combinations of βs have specific values:
  - $H_0 : \beta_1 = 0$
  - $H_0 : \beta_1 = \beta_2 = 0$
  - $H_0 : \beta_1 - \beta_2 = 1$
- Can be written as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b}$$

  (most often $\mathbf{b} = \mathbf{0}$), where $\mathbf{C}$ is of rank r.

- Test statistics:

$$\frac{(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{b})^T \{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{b})/r}{\widehat{\sigma}^2} \sim F_{r,n-p-1}.$$

- For $\mathbf{b} = \mathbf{0}$:

  i. If $\mathbf{C} = (0, \ldots, 0, 1, 0, \ldots)$, a single vector with $c_j = 1$ and $c_k = 0$ for all $k \neq j$. Then it is the same as the t-test for $\beta_j = 0$,

  $$t = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)} \sim t_{n-p-1},$$

  and $t^2 = F$ where $F \sim F_{1,n-p-1}$.

  ii. If $\mathbf{C} = \mathrm{diag}(0, 1, 1, \cdots, 1)$. Then it is an overall F test (i.e. $\beta_1 = \cdots = \beta_p = 0$).

- Hypothesis test using full and reduced model:

$$\begin{aligned} F &= \frac{SSR(\text{full}) - SSR(\text{reduced})/\Delta df}{SSE(\text{full})/(n-p-1)} \\ &\sim F_{\Delta df, n-p-1} \end{aligned}$$

or

$$\begin{aligned} F &= \frac{SSE(\text{reduced}) - SSE(\text{full})/\Delta df}{SSE(\text{full})/(n-p-1)} \\ &\sim F_{\Delta df, n-p-1} \end{aligned}$$

  - Exactly same result as previous!

Example: Child obesity data

- Test for the main and interaction effect of Smoke.
  - H0: $\beta_2 = \beta_3 = 0$
- Use the contrast matrix

$$C =$$

- Use full and reduced models
  - Full Model:
  - Reduced Model:

Example: Child obesity data

- Use the contrast matrix:

```
proc reg;
model wt_kg = age smoke A_S;
age: test smoke=0, A_S=0;
run;
```

- Fit the full and the reduced models:

```
proc reg;
model wt_kg = age smoke A_S;
run;
proc reg;
model wt_kg = age ;
run;
```

The REG Procedure
Model: MODEL1

Test age Results for Dependent Variable wt_kg				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	9.75141	0.20	0.8178
Denominator	26	48.11345		

The REG Procedure
Model: MODEL1
Dependent Variable: wt_kg

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	973.37997	324.45999	6.74	0.0016
Error	26	1250.94969	48.11345		
Corrected Total	29	2224.32967			

Root MSE	6.93639	**R-Square**	0.4376
Dependent Mean	23.22381	**Adj R-Sq**	0.3727
Coeff Var	29.86756		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	4.95865	6.65202	0.75	0.4627
age	1	2.87548	1.05916	2.71	0.0116
smoke	1	0.52160	8.58019	0.06	0.9520
A_S	1	0.20133	1.37335	0.15	0.8846

The REG Procedure
Model: MODEL1
Dependent Variable: wt_kg

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	953.87715	953.87715	21.02	<.0001
Error	28	1270.45252	45.37330		
Corrected Total	29	2224.32967			

Root MSE	6.73597	**R-Square**	0.4288
Dependent Mean	23.22381	**Adj R-Sq**	0.4084
Coeff Var	29.00459		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	5.41375	4.07439	1.33	0.1947
age	1	3.00170	0.65467	4.59	<.0001

Example: Child obesity data

- Use the contrast matrix:
  - Test statistics:
  - Null distribution:
- Use the full and the reduced model:
  - Test statistics:
  - Null distribution:

Diagnostics: Violations of Assumptions

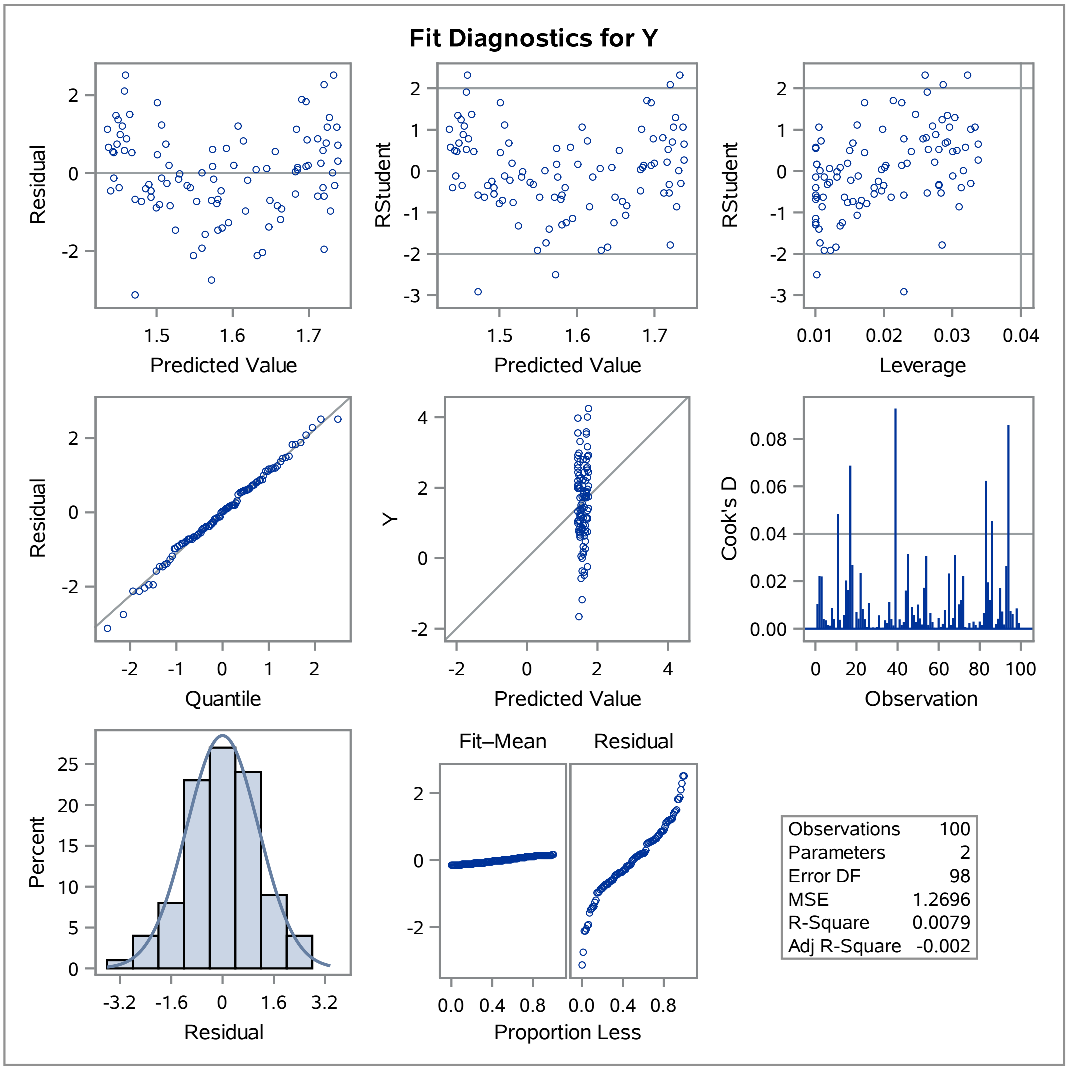
- Assumptions:
  - Linearity: $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$
  - Normality
  - Equal variance (homoscedasticity)
  - Independence

Diagnostics: Violations of Assumptions

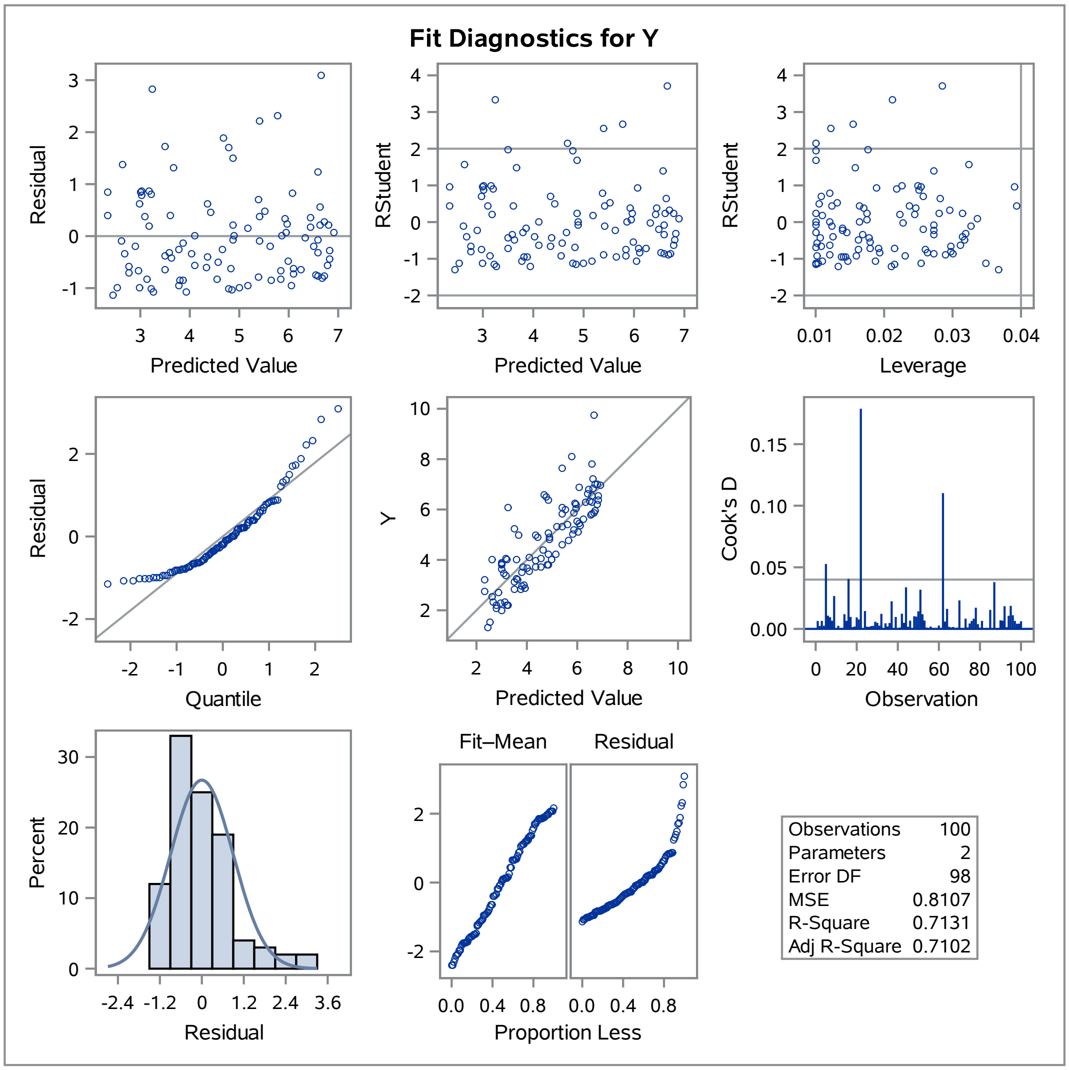
- Linearity:
  - Check: Partial regression plot, residual plot.
  - Remedy: Transformation, Add another regressor (ex. add x^2)
- Normality:
  - Check: Normal quantile plot, Statistical tests for normality (ex. Shapiro-Wilk test)
  - Remedy: Transformation, GLM
- Equal variance:
  - Check: Residual plot
  - Remedy: Transformation (ex. log), Weighted Least Square, GLM
- Independence:
  - Check: Done by intuition (e.g., repeatedly measured..), Residual plots
  - Remedy: Longitudinal, Time series

- Violation of which assumption?

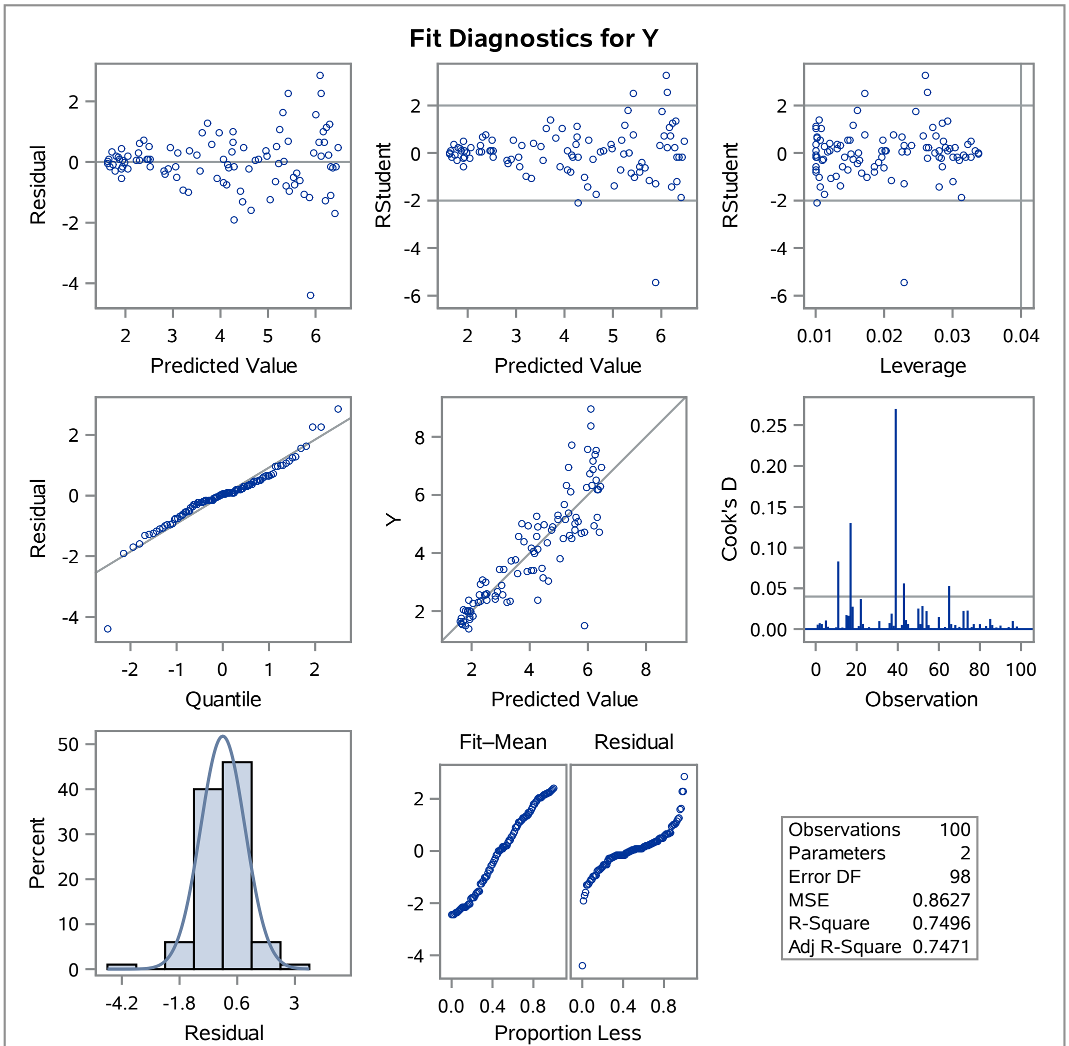
The REG Procedure
Model: MODEL1
Dependent Variable: Y



The REG Procedure
Model: MODEL1
Dependent Variable: Y



The REG Procedure
Model: MODEL1
Dependent Variable: Y



Observations	100
Parameters	2
Error DF	98
MSE	0.8627
R-Square	0.7496
Adj R-Square	0.7471

Weighted Least Squares

- Suppose that each observation has difference variance (heteroscedasticity):

$$\sigma_1 \neq \sigma_2 \neq \cdots \neq \sigma_n$$

$$V = diag[\sigma_1^2, \cdots, \sigma_n^2]$$

  - Grouped (Aggregate) data: Y_i is an average of n_i observations:

$$Y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} E_{ij}, \quad Var(E_{ij}) = \sigma^2,$$

    and then

$$Var(Y_i) = \sigma_i^2 = \sigma^2 / n_i$$

  - Count data: variance increases as the mean increases:

$$\sigma_i^2 \approx \mu_i \sigma^2$$

- Original model:

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

- Transformed model:

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

  where $\tilde{Y} = V^{-1/2}Y$, $\tilde{X} = V^{-1/2}X$ and $\tilde{\boldsymbol{\epsilon}} = V^{-1/2}\boldsymbol{\epsilon}$

  – Satisfy the equal variance assumption

$$\begin{aligned} Var(\tilde{\boldsymbol{\epsilon}}) &= V^{-1/2}Var(\boldsymbol{\epsilon})V^{-1/2} \\ &= V^{-1/2}VV^{-1/2} = I \end{aligned}$$

- Weighted Least Squares (WLS) estimator

$$\begin{aligned} \beta_{wls} &= (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} \\ &= (X^TV^{-1}X)^{-1}X^TV^{-1}Y \qquad (1) \end{aligned}$$

Example: Apple shots

- Researchers recorded the average number of stem shots in apple trees in each day. Varying numbers of trees n_i are observed in each day.
- Model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2/n_i)$$

  - Response: $Y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} E_{ij}$
  - E_{ij} : number of stem shoots from the jth tree on the ith day of the growing season.
  - x_i: number of days since dormancy.

Example: Apple shots

```
data apple;
input day ni Y;
cards;
0 5 10.2
3 5 10.4
7 5 10.6
13 6 12.5
18 5 12.0
24 4 15.0
25 6 15.17
32 5 17.0
38 7 18.71
. . .
```

Example: Apple shots

```
proc reg data=apple;
model Y = day;
weight ni;
run;
```

- SAS will construct a weight matrix equals to

$$V^{-1} = \begin{pmatrix} w_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n \end{pmatrix}$$

where $w_i = n_i$ in this example.

Example: Apple shots

- Use IML to estimate β

```
proc iml;
 use apple;
 read all var {Y} into Y;
 read all var {day} into X_1;
 read all var {ni} into W;

 n=nrow(Y);
 one_n=j(n,1,1);
 X=one_n||X_1;
 V_inv=DIAG(W);

 beta=inv(t(X)*V_inv*X)*t(X)*V_inv * Y;

 print beta;

quit;
```

PROC REG

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	22
Number of Observations Used	22

Weight: ni

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	6164.27627	6164.27627	1657.24	<.0001
Error	20	74.39209	3.71960		
Corrected Total	21	6238.66835			

Root MSE	1.92863	**R-Square**	0.9881
Dependent Mean	21.42212	**Adj R-Sq**	0.9875
Coeff Var	9.00297		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	9.97375	0.31427	31.74	<.0001
day	1	0.21733	0.00534	40.71	<.0001

IML

beta
9.9737537
0.2173303

Weighted Least Squares

- Important technique in linear regression to address for heteroscedasticity.
- GLM model: WLS is used to estimate parameters
  - Iteratively Reweighted Least Squares (IRWLS)