

PubH 7405: REGRESSION ANALYSIS



MULTIPLE REGRESSION ANALYSIS

THE DRAWBACKS OF SLR

- There are effect modifiers; SLR does not allow us to study **effect modifications**
- Even without interactions, information provided by different factors may be **redundant**. There are confounders; SLR does not allow us to **investigate marginal contribution** - contribution of a factor adjusted for other factors.
- SLR does not allow us to study or investigate **non-linear relationships**.

NORMAL ERROR REGRESSION MODEL

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

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

The terms involved could be continuous, binary, or categorical (several categories); they do not need to represent k different predictors; some term could be the product of two predictors, some term could be the quadratic power of another predictor.

In particular, Multiple Regression allows us to get into two new areas that were not possible with Simple Linear Regression:

- (i) Interaction or Effect Modification, and
- (ii) Non-linear Relationship.

The primary reasons for emphasizing matrices are: (1) Simple Linear Regression and Multiple Linear Regression look the same in matrix terms; we do not have to prove some of the results again; and (2) It lead to the same computational tools/software and it allows more theoretical works.

OBSERVATIONS & ERRORS

$$\mathbf{Y}_{nx1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \boldsymbol{\varepsilon}_{nx1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

X: the “Design Matrix” (a matrix of constants X 's)

$$\mathbf{X}_{nx(k+1)} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdot & x_{k1} \\ 1 & x_{12} & x_{22} & \cdot & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdot & x_{kn} \end{bmatrix}$$

First subscript: Variable;
Second subscript: Subject

The dimension of “**Design Matrix**” X is changed to handle more predictors: one column for each predictor (the number of rows is still the sample size. The first column (filled with “1”) is still “optional”; not included when doing “**Regression through the origin**” (i.e. no intercept).

β : Regression Coefficient
(a column vector of parameters)

$$\mathbf{\beta}_{(k+1) \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

MLR MODEL IN MATRIX TERMS

$$\mathbf{Y}_{n-by-1} = \mathbf{X}_{n-by-(k+1)} \boldsymbol{\beta}_{(k+1)-by-1} + \boldsymbol{\varepsilon}_{n-by-1}$$

$$\mathbf{E}(\mathbf{Y})_{n-by-1} = \mathbf{X}\boldsymbol{\beta}$$

$$\sigma^2(\mathbf{Y})_{n-by-n} = \sigma^2 \mathbf{I}$$

OPERATIONS ON BASIC DATA MATRICES

$$\mathbf{Y}'\mathbf{Y} = [y_1 \quad y_2 \quad \cdots \quad y_n] \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = [\sum y_i^2]$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & \cdot & 1 \\ x_{11} & x_{12} & x_{13} & \cdot & x_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \cdot & x_{kn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \vdots \\ \sum x_{ki} y_i \end{bmatrix}$$

$$\begin{aligned}
 \mathbf{X}'\mathbf{X} &= \begin{bmatrix} \mathbf{1} & \mathbf{1} & \cdots & \mathbf{1} \\ \mathbf{x}_{11} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_{k1} & \mathbf{x}_{k2} & \cdots & \mathbf{x}_{kn} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{x}_{11} & \cdots & \mathbf{x}_{k1} \\ \mathbf{1} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{k2} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{1} & \mathbf{x}_{1n} & \cdots & \mathbf{x}_{kn} \end{bmatrix} \\
 &= \begin{bmatrix} \mathbf{1} & \sum \mathbf{x}_{1i} & \cdots & \sum \mathbf{x}_{ki} \\ \sum \mathbf{x}_{1i} & \sum \mathbf{x}_{1i}^2 & \cdots & \sum \mathbf{x}_{1i} \mathbf{x}_{ki} \\ \vdots & \vdots & \cdots & \vdots \\ \sum \mathbf{x}_{ki} & \sum \mathbf{x}_{1i} \mathbf{x}_{ki} & \cdots & \sum \mathbf{x}_{1i}^2 \end{bmatrix}
 \end{aligned}$$

$\mathbf{X}'\mathbf{X}$ is a square matrix filled with sums of squares and sums of products; we can form $\mathbf{X}\mathbf{X}'$ but it is a different n -by- n matrix which we do not need.

LEAST SQUARE METHOD

$$\text{Data : } \{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)\}_{i=1}^n$$

$$Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \dots - \beta_k x_{ki})^2$$

We solve equations :

$$\frac{\delta Q}{\delta \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \dots - \beta_k x_{ki}) = 0$$

$$\frac{\delta Q}{\delta \beta_1} = -2 \sum_{i=1}^n x_{1i} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \dots - \beta_k x_{ki}) = 0$$

...

$$\frac{\delta Q}{\delta \beta_k} = -2 \sum_{i=1}^n x_{ki} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \dots - \beta_k x_{ki}) = 0$$

Data : The case of 2 variables $\{(x_{1i}, x_{2i}, y_i)\}_{i=1}^n$

$$Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

Solve:

$$\frac{\delta Q}{\delta \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) = 0$$

$$\frac{\delta Q}{\delta \beta_1} = -2 \sum_{i=1}^n x_{1i} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) = 0$$

$$\frac{\delta Q}{\delta \beta_2} = -2 \sum_{i=1}^n x_{2i} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) = 0$$

NORMAL EQUATIONS

$$\frac{\delta Q}{\delta \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \dots - \beta_k x_{ki}) = \sum_{i=1}^n e_i = 0$$

$$\frac{\delta Q}{\delta \beta_1} = -2 \sum_{i=1}^n x_{1i} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \dots - \beta_k x_{ki}) = \sum_{i=1}^n x_{1i} e_i = 0$$

...

$$\frac{\delta Q}{\delta \beta_k} = -2 \sum_{i=1}^n x_{ki} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \dots - \beta_k x_{ki}) = \sum_{i=1}^n x_{ki} e_i = 0$$

SSE MATRIX NOTATION

Sum of squared errors :

$$Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} \dots - \beta_k x_{ki})^2$$

In matrix notation :

$$\begin{aligned} Q &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Normal Equations :

$$\begin{aligned} Q &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

$$\frac{\delta}{\delta\boldsymbol{\beta}} (Q) = \begin{bmatrix} \frac{\delta Q}{\delta\beta_0} \\ \frac{\delta Q}{\delta\beta_1} \\ \dots \\ \frac{\delta Q}{\delta\beta_k} \end{bmatrix}$$

$$= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

$$= 0 \Leftrightarrow \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

LEAST SQUARE ESTIMATES

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

Note the error in equation (6.25) of the text

Recall :

$$\mathbf{E}(AY) = A\mathbf{E}(Y)$$

$$\mathbf{b} = [(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}']\mathbf{Y}$$
$$= \mathbf{A}\mathbf{Y}$$

$$\mathbf{E}(\mathbf{b}) = \mathbf{A}\mathbf{E}(\mathbf{Y})$$
$$= [(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'][\mathbf{X}\boldsymbol{\beta}]$$
$$= [(\mathbf{X}'\mathbf{X})^{-1}][\mathbf{X}'\mathbf{X}]\boldsymbol{\beta}$$
$$= \boldsymbol{\beta}$$

**Estimates of intercept and
all slopes are unbiased**

Recall :

$$\sigma^2(\mathbf{AY}) = \mathbf{A}\sigma^2(\mathbf{Y})\mathbf{A}'$$

THE HAT MATRIX

$$\hat{Y} = Xb$$

$$= X[(X'X)^{-1} X'Y]$$

$$= [X(X'X)^{-1} X']Y$$

$$= HY$$

$H = X(X'X)^{-1} X'$ is called the "Hat Matrix"

IDEMPOTENCY

the "Hat Matrix":

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is "idempotent":

$$\mathbf{H}\mathbf{H} = \mathbf{H}$$

RESIDUALS

Model :

$$\mathbf{Y}_{nx1} = \mathbf{X}_{nx2} \boldsymbol{\beta}_{2x1} + \boldsymbol{\varepsilon}_{nx1}$$

Fitted Value :

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

Residuals :

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}\mathbf{b} \\ &= \mathbf{Y} - \mathbf{H}\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \end{aligned}$$

Like the Hat Matrix \mathbf{H} , $(\mathbf{I}-\mathbf{H})$ is symmetric & idempotent

Recall :

$$\sigma^2(\mathbf{AY}) = \mathbf{A}\sigma^2(\mathbf{Y})\mathbf{A}'$$

VARIANCE OF RESIDUALS

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\sigma^2(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2(\mathbf{Y})(\mathbf{I} - \mathbf{H})'$$

$$= (\mathbf{I} - \mathbf{H})(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{H})'$$

$$= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})'$$

$$= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})$$

$$= \sigma^2(\mathbf{I} - \mathbf{H})$$

$$\hat{=} \text{MSE}(\mathbf{I} - \mathbf{H})$$

VARIANCE OF REGRESSION COEFFICIENTS

$$\mathbf{b} = [(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}']\mathbf{Y}$$
$$= \mathbf{A}\mathbf{Y}$$

$$\sigma^2(\mathbf{b}) = \mathbf{A}\sigma^2(\mathbf{Y})\mathbf{A}'$$
$$= \mathbf{A}\sigma^2\mathbf{I}\mathbf{A}'$$
$$= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$
$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{s}^2(\mathbf{b}) = \text{MSE}(\mathbf{X}'\mathbf{X})^{-1}$$

The matrix notation and derivations may be deceiving because **they hide enormous computational complexities**. To find the inverse of a 10×10 matrix, for example $X'X$ with $k = 9$, requires tremendous amount of computation. However, the actual computations will be done by computer; hence it does not matter to us whether $X'X$ represents a 2×2 or a 6×6 matrix.

THE MEAN RESPONSE

Let $\mathbf{X} = \mathbf{x}_h$ denote the level of X for which we wish to estimate the mean response, i.e. $E(Y|\mathbf{X}=\mathbf{x}_h)$. The only thing new is that \mathbf{X} and \mathbf{x}_h are vector; $\mathbf{x}_h = (x_{1h}, x_{2h}, \dots, x_{kh})$. The point estimate of the response is:

$$\begin{aligned} E(Y | \mathbf{X} = \mathbf{x}_h) &= \hat{Y}_h \\ &= b_0 + b_1 x_{1h} + b_2 x_{2h} + \dots + b_k x_{kh} \end{aligned}$$

In matrix terms:

$$\hat{Y} = \mathbf{X}'_{\mathbf{h}} \mathbf{b}$$

$$\begin{aligned} \sigma^2(\hat{Y}) &= \mathbf{X}'_{\mathbf{h}} \boldsymbol{\sigma}^2(\mathbf{b}) \mathbf{X}_{\mathbf{h}} \\ &= \sigma^2 \mathbf{X}'_{\mathbf{h}} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_{\mathbf{h}} \end{aligned}$$

$$s^2(\hat{Y}) = \text{MSE}(\mathbf{X}'_{\mathbf{h}} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_{\mathbf{h}})$$

PRICTION OF NEW OBSERVATION

Let $X = x_h$ denote the level of X under investigation, at which the mean response is $E(Y|X=x_h)$. Let $Y_{h(new)}$ be the value of the new individual response of interest. The point estimate is still the same $E(Y|X=x_h)$:

$$\begin{aligned}\hat{Y}_{h(new)} &= b_0 + b_1 x_{1h} + b_2 x_{2h} + \dots + b_k x_{kh} \\ &= \mathbf{X}'_h \mathbf{b}\end{aligned}$$

$$\mathbf{Var}(\hat{Y}_{h(new)}) = \sigma^2 \{1 + \mathbf{X}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_h\}$$

$$\mathbf{s}^2(\hat{Y}_{h(new)}) = MSE\{1 + \mathbf{X}'_h (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_h\}$$

ANALYSIS OF VARIANCE

In matrix terms, we can write – with **J** being the square $n \times n$ matrix with all elements equal to 1:

$$SST = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

$$SSE = \mathbf{e}'\mathbf{e}$$

$$= (\mathbf{Y} - \mathbf{X}\mathbf{b})(\mathbf{Y} - \mathbf{X}\mathbf{b})'$$

$$= \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$$

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

All three ANOVA sums of squares (SST, SSR, and SSE) are quadratic forms (H is the hat matrix):

$$SST = \mathbf{Y}' \left[\mathbf{I} - \left(\frac{1}{n} \right) \mathbf{J} \right] \mathbf{Y}$$

$$SSE = \mathbf{Y}' (\mathbf{I} - \mathbf{H}) \mathbf{Y}$$

$$SSR = \mathbf{Y}' \left[\mathbf{H} - \left(\frac{1}{n} \right) \mathbf{J} \right] \mathbf{Y}$$

“ANOVA” TABLE

- The breakdowns of the total sum of squares and its associated degree of freedom are displayed in the form of an “analysis of variance table” (ANOVA table) for regression analysis as follows:

Source of Variation	SS	df	MS	F Statistic	p-value
Regression	SSR	k	MSR	MSR/MSE	
Error	SSE	n-k-1	MSE		
Total	SST	n-1			

- MSE, the “error mean square”, serves as an estimate of the constant variance σ^2 as stipulated by the regression model.

Global Test Of Significance

- **Hypotheses**

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_a : One and/or both of the parameters not equal to zero.

- **Test Statistic**

$$F = \text{MSR/MSE (From ANOVA table)}$$

- **Rejection Rule**

Reject H_0 if $F > F_\alpha$

where F_α is based on an F distribution with k d.f. (numerator) and $(n - k - 1)$ d.f. (denominator).

COEFFICIENT OF DETERMINATION

- The ratio, called the **coefficient of multiple determination**, defined as:

$$R^2 = \frac{SSR}{SST}$$

- representing the portion of total variation in y-values attributable to difference in values of independent variables or covariates.

1) It can be shown that the **coefficient of multiple determination R^2** can be expressed as the square of the coefficient of correlation between the response Y and its fitted value.

2) To distinguish between the coefficients of determination for simple and multiple regression, the former is sometime referred to as the **coefficient of simple determination.**

COEFFICIENT OF CORRELATION

- 1) The coefficient of multiple correlation R is defined as the positive square root of the coefficient of multiple determination R^2
- 2) The coefficient of multiple correlation R , except for the case of $k=1$, is less useful; the only exception is its use as **the coefficient of correlation between the response Y and its fitted value.**

SCATTER DIAGRAMS & CORRELATION MATRIX

- Diagnostics play an important role in the development and evaluation of MLR models.
- We need a **scatter diagram for Y against each of the X's**; Y versus X_1 & Y versus X_2 for $k=2$
- That would be nicely supplemented by a “**correlation matrix**” showing coefficients of simple correlation between y and each X as well as correlation among predictor variables. This matrix is symmetric and its main diagonal contains 1's (correlation between a variable and itself; the dimension is $(k+1) \times (k+1)$). **PROC CORR gives us this matrix.**

RESIDUAL PLOTS

- A plot of residuals plotted against each of the **predictors** provides information about **adequacy of the regression model with respect to that predictor variable**,
- Just like in SLR, a **plot of residuals plotted against the fitted values** is useful for assessing **the appropriateness of the model in general**, the constancy of the variance, and information about possible outliers,
- A plot of residuals plotted against time or some sequence marker would tell about possible correlation between error terms, etc...

AN EXAMPLE WITH 2 PREDICTORS

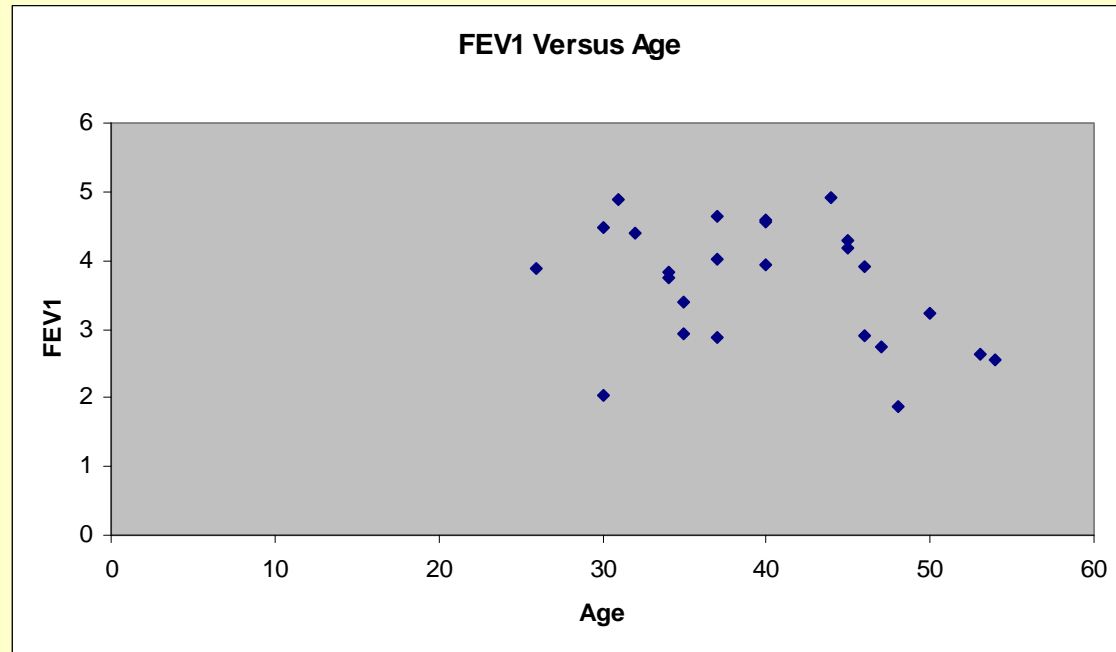
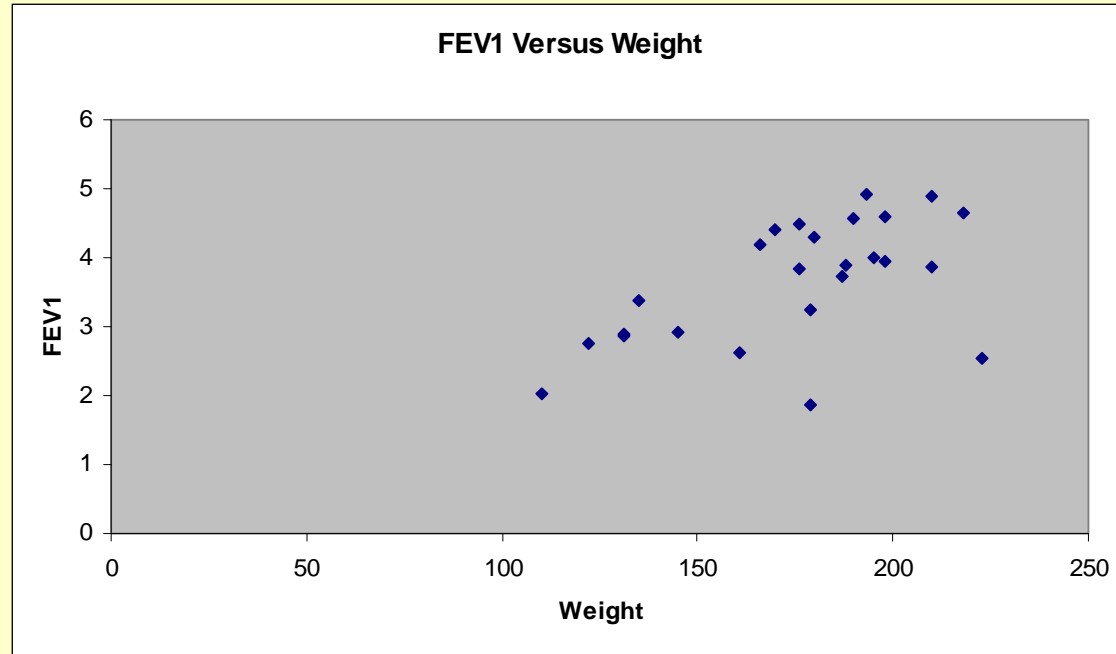
Age	Weight	FEV1
53	161	2.63
40	198	3.95
26	210	3.87
34	187	3.74
46	131	2.9
44	193	4.91
35	135	3.39
45	166	4.19
45	180	4.29
30	176	4.49
46	188	3.9
50	179	3.24
31	210	4.88
37	195	4.01
40	190	4.56
32	170	4.41
37	218	4.64
54	223	2.55
34	176	3.83
40	198	4.59
48	179	1.86
37	131	2.87
30	110	2.04
47	122	2.75
35	145	2.92

FEV1 is popular measure of lung health

	Age	Weight	FEV1
Age	1.0000	0.0259	-0.3452
Weight		1.0000	0.5643
FEV1			1.0000

Preliminary Results:

FEV1 seems to correlate negatively to Age ($r=-.3452$) but positively to Weight ($r=.5643$).



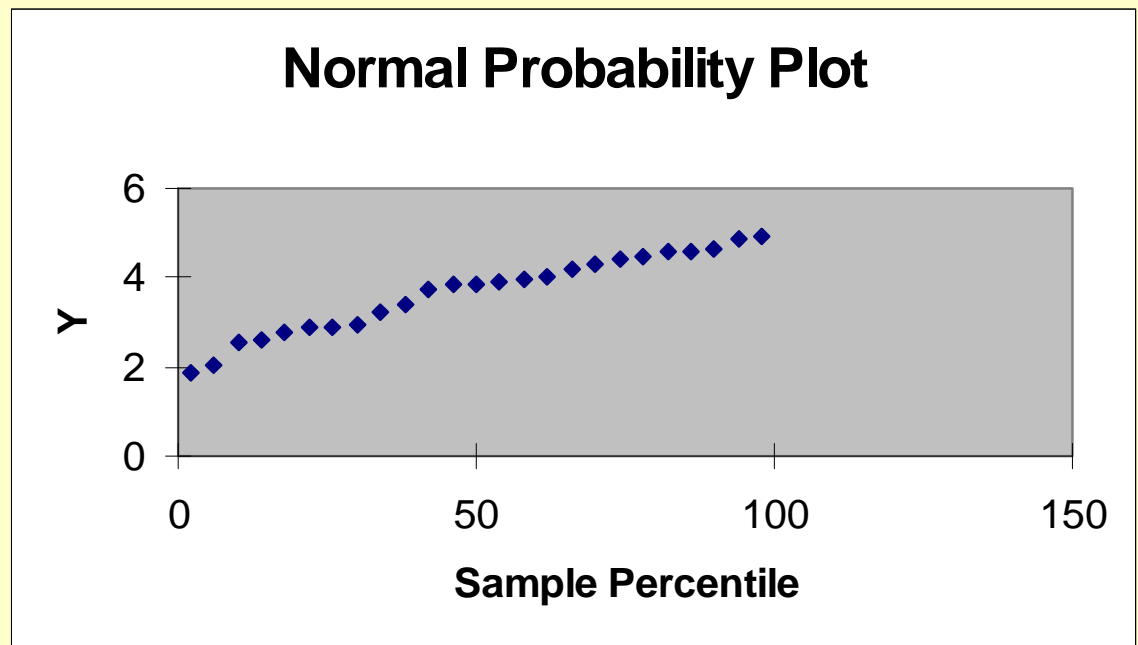
SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.669311
R Square	0.447977
Adjusted R Square	0.397793
Standard Error	0.689428
Observations	25

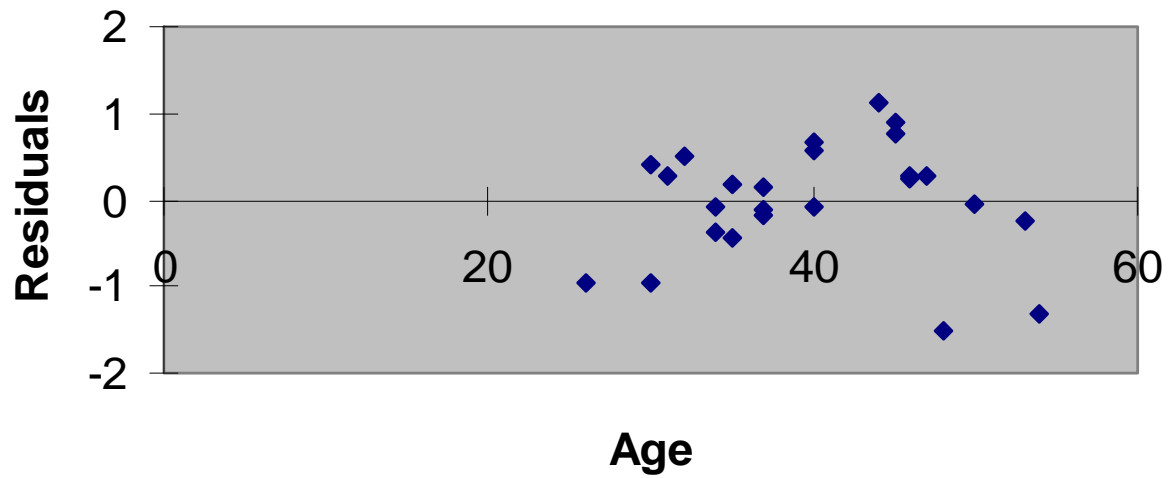
We'll talk about "adjusted R square" later

ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	8.4859284	4.242964	8.926707	0.001450537	
Residual	22	10.456848	0.475311			
Total	24	18.942776				
	<i>Coefficients</i>	<i>St Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	2.424082643	1.080088	2.244338	0.035198	0.184117314	4.664047972
X1 = Age	-0.04180436	0.0183968	-2.27238	0.033187	-0.07995689	-0.003651829
X2 = Weight	0.016574028	0.0045785	3.619964	0.001517	0.007078786	0.02606927

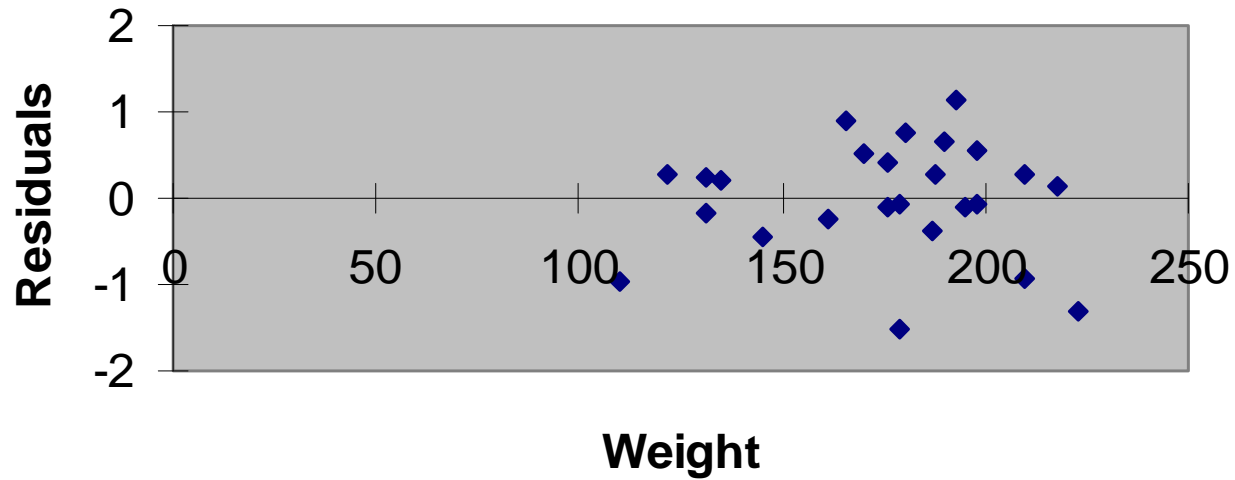
PROBABILITY OUTPUT	
Percentile	Y
2	1.86
6	2.04
10	2.55
14	2.63
18	2.75
22	2.87
26	2.9
30	2.92
34	3.24
38	3.39
42	3.74
46	3.83
50	3.87
54	3.9
58	3.95
62	4.01
66	4.19
70	4.29
74	4.41
78	4.49
82	4.56
86	4.59
90	4.64
94	4.88
98	4.91



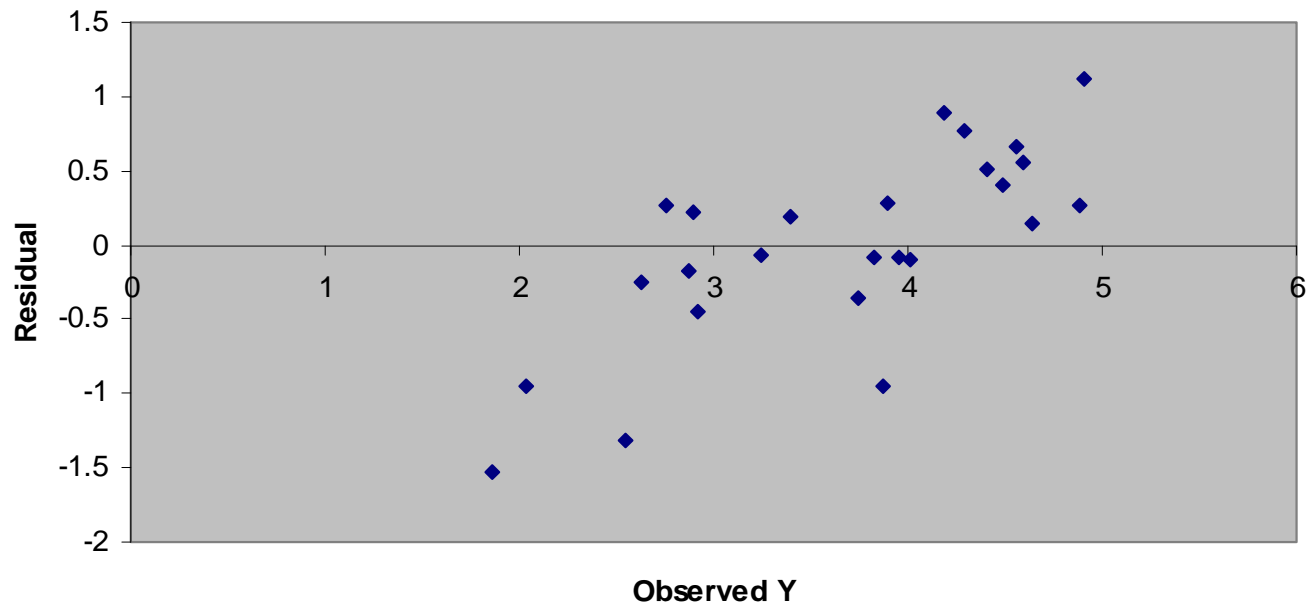
X1 = Age Residual Plot



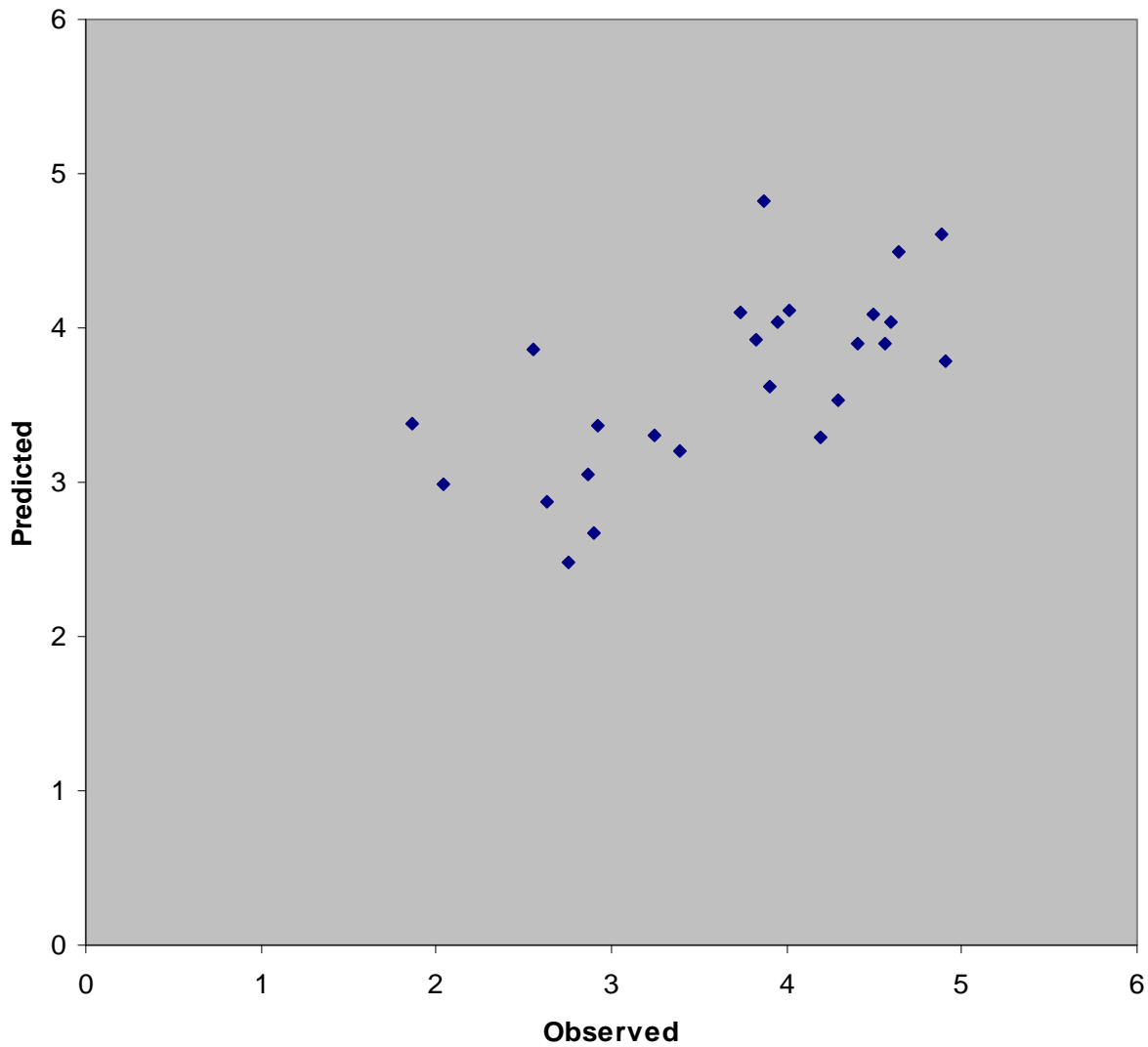
X2 = Weight Residual Plot



Residual Versus Observed

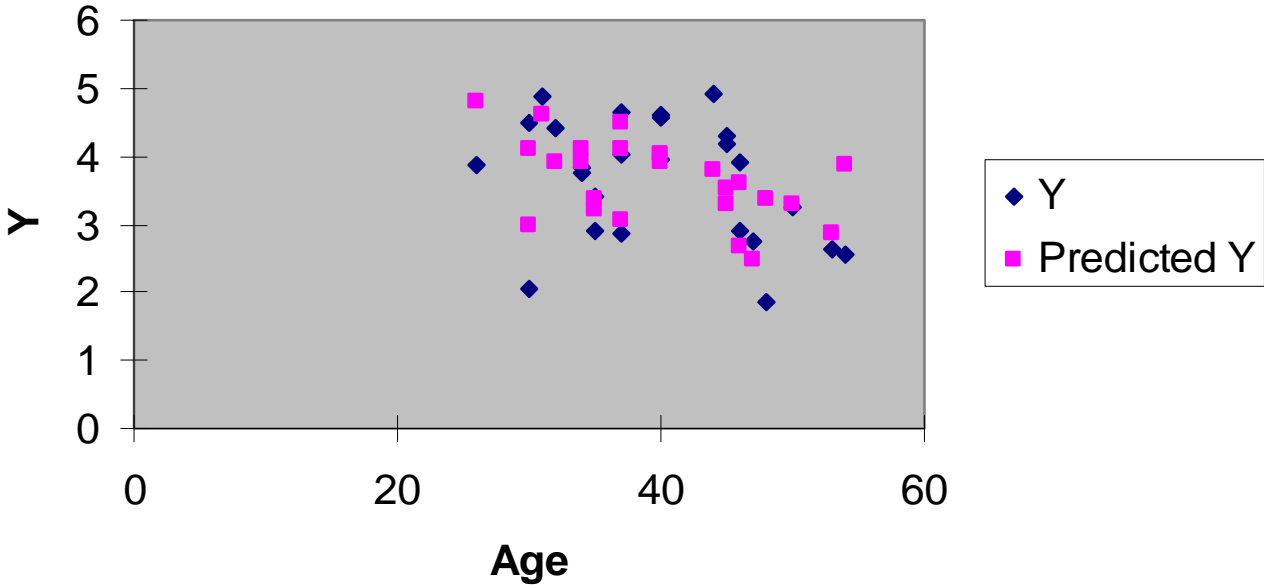


Predicted Versus Observed

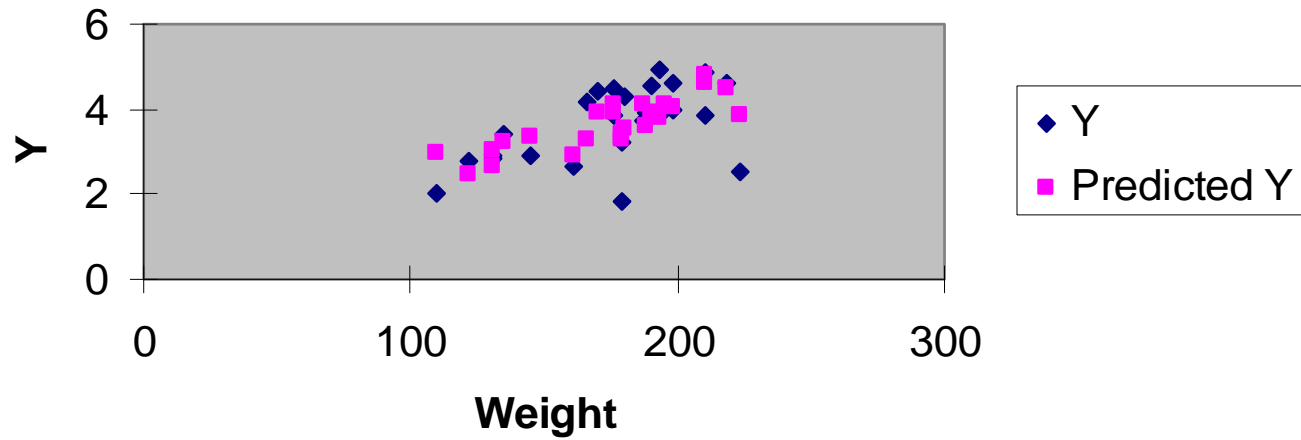


$R = .6693$

X1 = Age Line Fit Plot



X2 = Weight Line Fit Plot



Example: PULMONARY FUNCTION

- **Dependent Variable:** Forced Expired Volume (FEV)
- **Independent Variables:**
 - Age of person
 - Smoking status of person
- **Questions:**
 - Is age related to FEV independent of smoking status
 - Is smoking status related to FEV independent of age
 - How much of the variability in FEV is explained by age and smoking combined

Model for FEV Example

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

X_1 = Smoking status (1=smoker, 0=nonsmoker)

X_2 = Age

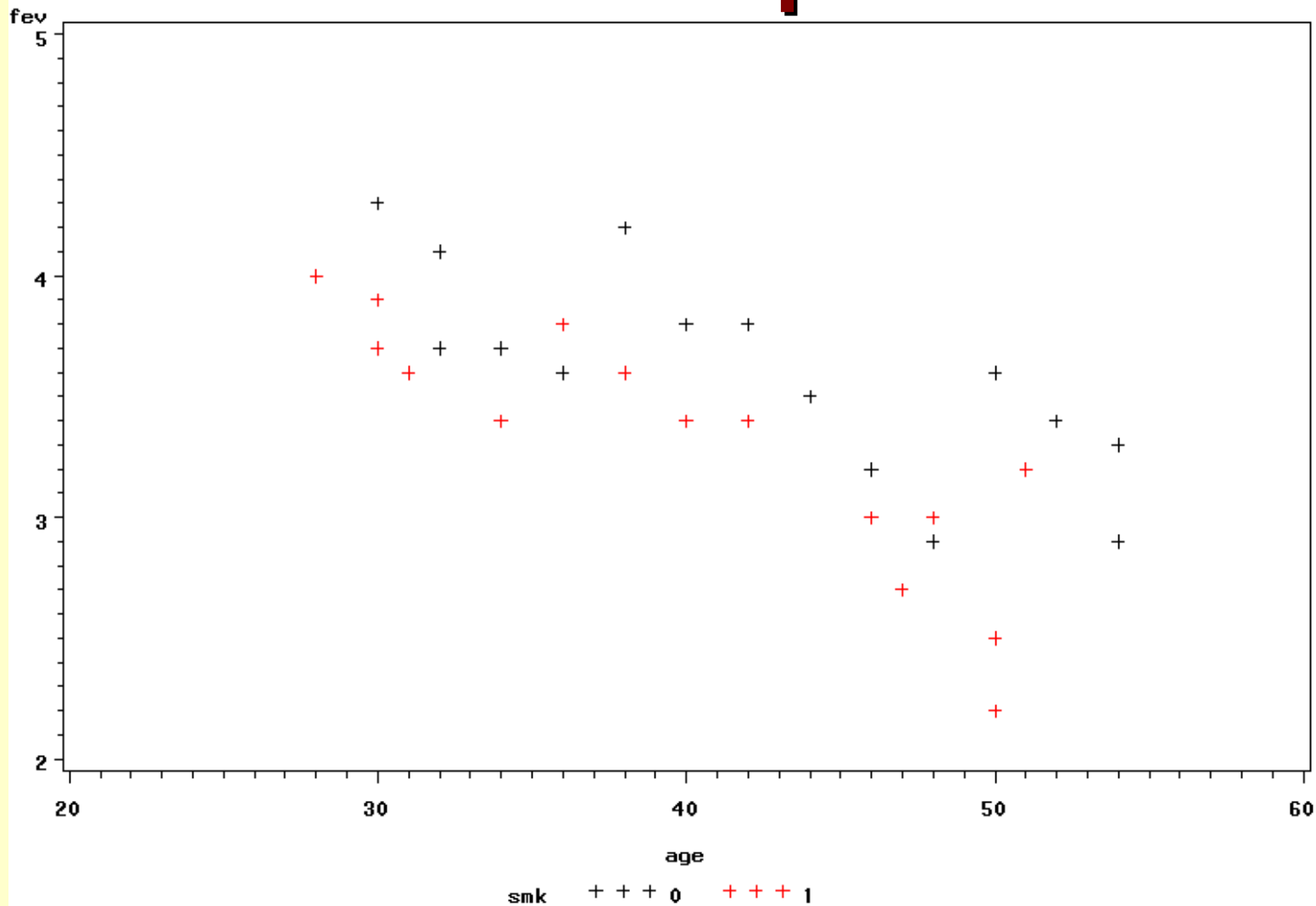
Smokers

$$E(\text{FEV}) = \beta_0 + \beta_1 + \beta_2 * \text{Age}$$

Non Smokers

$$E(\text{FEV}) = \beta_0 + \beta_2 * \text{Age}$$

Pulmonary Function Example Scatterplot



Pulmonary Function Example

SAS Output with Proc Corr

PROC CORR DATA = FEV;

Pearson Correlation Coefficients, n = 30

	age	smk	fev
age	1.00000	-0.12788	-0.73024
smk	0.5007	1.00000	<.0001
fev	-0.12788	-0.31620	1.00000
	0.5007	0.0887	
	<.0001	0.0887	

```

PROC REG;
  MODEL fev = age smk ;
RUN;

```

Dependent Variable: FEV

Analysis of Variance

Source	DF		Sum of Squares	Mean Square	F Value	Pr > F
Model	2	SSR	4.96510	2.48255	32.08	<.0001
Error	27	SSE	2.08957	0.07739		
Corrected Total	29	SST	7.05467			

Root MSE 0.27819
 Dependent Mean 3.44667
 Coeff Var 8.07136

R-Square 0.7038

Tests Ho: $\beta_1 = \beta_2 = 0$

Proportion of variance explained by both variables

Readings & Exercises

- Readings: A thorough reading of the text's sections 6.2-6.9 (pp.222-247) is recommended.
- Exercises: The following exercises are good for practice, all from chapter 6 of text: 6.5(b-d), 6.7, 6.10(a-d), and 6.15(a-f).

Due As Homework

- 16.1** Refer to dataset “Cigarettes”, let $Y = \log(\text{NNAL})$ and consider a model with three independent variables, $X_1 = \text{CPD}$, $X_2 = \text{Age}$, and $X_3 = \text{Gender}$:
- Fit the model and interpret the results, especially the values of the estimated regression coefficients.
 - Plot the residuals against predicted values; What do the plots suggest, any clear departures from the model?
 - Conduct the Brown-Forsythe test for constancy of error variance.
 - Add a term to model (a) to test if X_3 modify the effect of X_1
 - Add a quadratic term to model (a) to test if effect of X_1 is linear.
- 16.2** Answer the 5 questions of Exercise 16.1 using dataset “Infants” with $Y = \text{Birth Weight}$, $X_1 = \text{Gestational Weeks}$ and $X_2 = \text{Mother's Age}$, and $X_3 = \text{Toxemia}$ (toxemia is a pregnancy condition resulting from metabolic disorder).

Only #16.2 is required