PubH 7405:
REGRESSION ANALYSIS

INTRODUCTORY MULTIPLE REGRESSION

SIMPLE LINEAR REGRESSION

- The method we have learned so far is called "**SIMPLE LINEAR REGRESSION**" (& simple linear correlation).
- It's "**linear**" because we assume that the relationship is represented by a straight line.
- It's "**simple**" because it only allows us to investigate the role of "one predictor at a time" (& its effect on the response).

Most of the times, **we have more than just two measurements**; may be only one "Response" but many Predictors – at least many "**Potential Predictors**". Then you could ask **why predicting the response from just one predictor?** Or, if there are more than one predictors, then which predictor? Or which would be the most valuable one? **Can we take advantage of their availability to use them all?**

#1. Drinking in College

- Estimate how predictable heavy **drinking in the first semester of college** is on the basis of information obtained prior to college
- Identify **pre-college variables** that are important **predictors** of heavy drinking in the first semester (**many potential predictors**)
- (from a past Group Presentation)

#2. SLEEP DURATION & OBESITY

- Cohort study, n=991
- 4/1999 – 4/2004
- (past Group Presentation)
- Response is **Obesity**
- **<u>Primary predictor</u>: Sleep Duration; but information on others are available.**

Continuous: Age, Sleep Duration, Depression Score
Binary: Sex, Educational Achievement, Marital Status
Categorical: Physical Job Demand, Household Income, Alcohol Consumption, Snoring

#3. RESEARCH IN AN AMUSEMENT PARK

- They do it for business planning: designing questionnaires, selecting samples, conducting interviews, and analyzing data that provide information about visitors' attitudes, perceptions, and preferences.
- **Information** about visitors themselves, where they come from and why they came: **Many predictors**
- Results would be variety of plans, strategies, and decisions on how to draw visitors to the park & make them to spend more.

#4. In a study of Lung Health using FEV (Forced Expiratory Volume) as the Outcome or Response Variable, we could investigate the role of Age, Height, and Weight, etc… as potential predictors.

#5. In investigating "severity" of prostate cancer, potential predictors include age at diagnosis and level of serum acid phosphatase, X-ray reading, grade (pathology reading of a biopsy of the tumor obtained by needle), and stage (a rough measure of the size and location of the tumor obtained by palpation with fingers via the rectum).

Statistically, is there any problem with using one predictor at a time using Simple Linear Regression? (A similar but more simple question: suppose we want to compare several population means, what is the problem with using the two-sample t-test to compare two means at a time?) –

A few problems in addition to the obvious loss of opportunities.

REGRESSION MODEL

- With one predictor X (& the response Y), we have Model: $Y = \beta_0 + \beta_1 x + \varepsilon$ where β_0 and β_1 are two parameters called "regression coefficients", the Intercept and the Slope, respectively. The **fixed** part is the **mean response**, $E(Y)=\beta_0 + \beta_1 x$, when $X = x$. The roles of "response" and "predictor" are clearly defined.
- What would be wrong if we consider more than one predictors: **How do other factors affect the prediction of Y using one particular predictor?**

THE SLOPE

- We said that the Slope is the more important parameter because **it plays a dual role**:
- (i) It plays a role in the prediction process
- (ii) **It also tells aspects of the correlation**: the slope β_1 and the coefficient of correlation r are of the same "sign"; β_1 is positive for a positive association and negative for a negative association – and if one is zero, so is the other.

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$= r\frac{s_y}{s_x}$$

Any factor or factors that mess up or confound the relationship also affect the slope – and the prediction.

#6. Data are shown below for two groups of patients who died of acute myelogenous leukemia (AML). Patients were classified into the two groups according to the presence or absence of a morphologic characteristic of white cells. Patients termed "AG positive" were identified by the presence of Auer rods and/or significant granulature of the leukemic cells in the bone marrow at diagnosis. For the AG negative patients these factors were absent.

AG-Positive, n = 17		AG-Negative, n = 16	
White Blood count (WBC)	Survival Time (weeks)	White Blood Count (WBC)	Survival Time (weeks)
2,300	65	4,400	56
750	156	3,000	65
4,300	100	4,000	17
2,600	134	1,500	7
6,000	16	9,000	16
10,500	108	5,300	22
10,000	121	10,000	3
17,000	4	19,000	4
5,400	39	27,000	2
7,000	143	28,000	3
9,400	56	31,000	8
32,000	26	26,000	4
35,000	22	21,000	3
100,000	1	79,000	30
100,000	1	100,000	4
52,000	5	100,000	43
100,000	65		

AG-Positive, n = 17		AG-Negative, n = 16	
White Blood count (WBC)	Survival Time (weeks)	White Blood Count (WBC)	Survival Time (weeks)
2,300	65	4,400	56
750	156	3,000	65
4,300	100	4,000	17
2,600	134	1,500	7
6,000	16	9,000	16
10,500	108	5,300	22
10,000	121	10,000	3
17,000	4	19,000	4
5,400	39	27,000	2
7,000	143	28,000	3
9,400	56	31,000	8
32,000	26	26,000	4
35,000	22	21,000	3
100,000	1	79,000	30
100,000	1	100,000	4
52,000	5	100,000	43
100,000	65		

Leukemia is a cancer characterized by an over-proliferation of white blood cells; the higher the white blood count (WBC), the more severe the disease; WBC is an important predictor of Survival Time (the Response). But what about "AG", that morphologic characteristic of white cells?

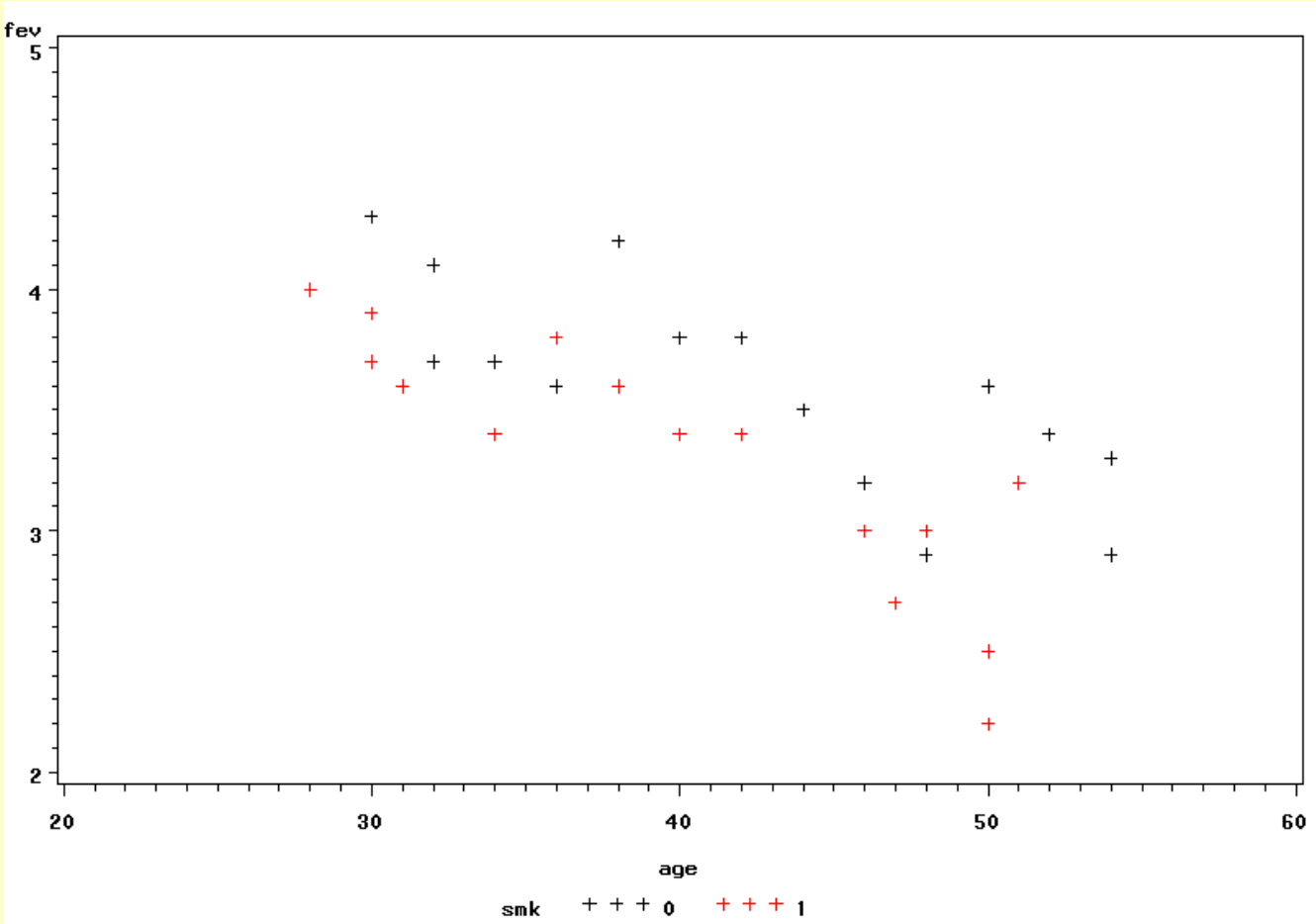
AG-Positive, n = 17		AG-Negative, n = 16	
White Blood count (WBC)	Survival Time (weeks)	White Blood Count (WBC)	Survival Time (weeks)
2,300	65	4,400	56
750	156	3,000	65
4,300	100	4,000	17
2,600	134	1,500	7
6,000	16	9,000	16
10,500	108	5,300	22
10,000	121	10,000	3
17,000	4	19,000	4
5,400	39	27,000	2
7,000	143	28,000	3
9,400	56	31,000	8
32,000	26	26,000	4
35,000	22	21,000	3
100,000	1	79,000	30
100,000	1	100,000	4
52,000	5	100,000	43
100,000	65		

It can be easily seen that, among AG-positive patients, WBC and Survival Time are negatively correlated – as noted that "the higher the white blood count (WBC), the more severe the disease". But that is not necessarily true for AG-negative patients: AG modifies the effect of WBC.

#7. Pulmonary Function

Dependent Variable: Forced Expired Volume (FEV), a measure of lung health.

Independent Variables: Age and Smoking Status



Some quick observations:

(1) Generally, FEV is linearly related to Age,

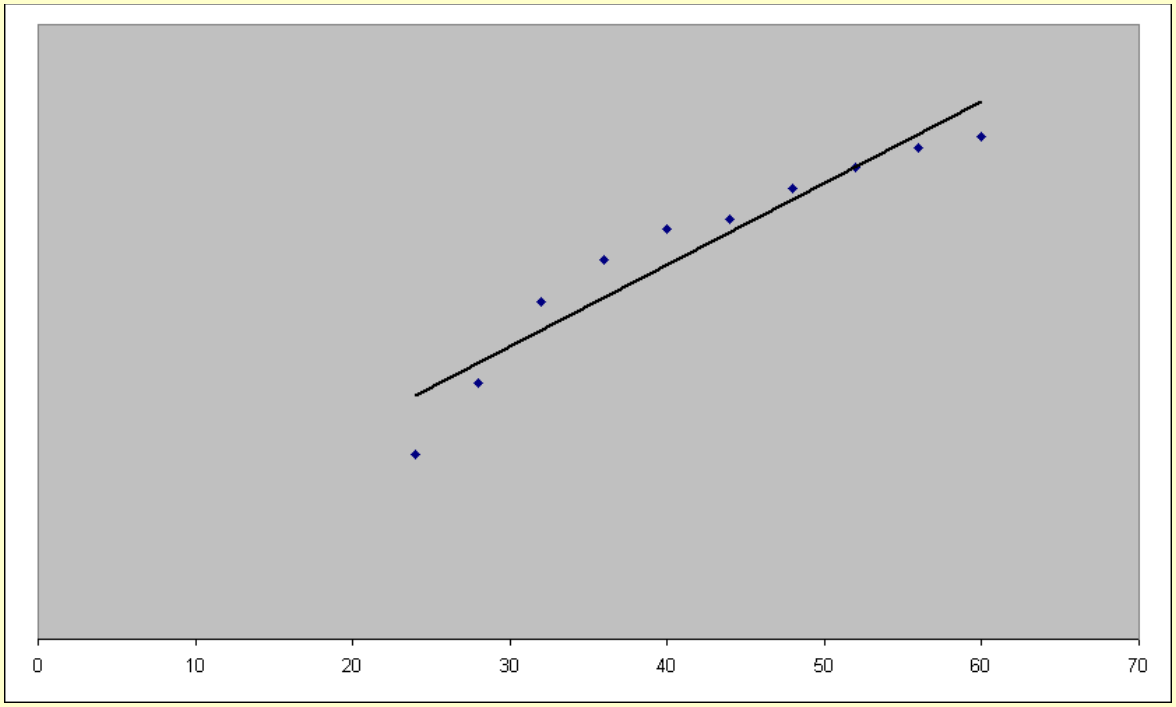
(2) Non-smokers have better lung health as compared to smokers; the Regression Line for smokers is “steeper”, i.e. large slope: Smoking could potentially modify the natural effect of Age on Lung Health.

EFFECT MODIFICATION

- **The effect of some factor X on the dependent variable Y may be influenced by the presence of other factors through "interactions"; also called "Effect Modification" in applied fields**
- **For example, (i) The presence of AG could modify the effect of WBC on Survival Time of AML patients); (ii) Smoking could modify the effect of Age on FEV (a measure of Lung Health).**
- **SLR does not allow us to study "effect modifications"; we could compare the slopes but that is not very precise and only simple problems when no other factors are considered.**

#8. The following data were collected during an experiment in which 10 laboratory animals were inoculated with a pathogen. The variables are Time after inoculation (X, in minutes) and Temperature (Y, in Celsius degrees).

	X, Time Minutes)	Y, Temperature (^{0}C)
	24	38.8
	28	39.5
	32	40.3
	36	40.7
	40	41.0
	44	41.1
	48	41.4
	52	41.6
	56	41.8
	60	41.9
Total	**420**	408.1



Strong positive correlation but the relationship may not be linear; curving up in the middle

NON-LINEAR RELATIONSHIPS

- **Not all relationships are “linear”; high blood pressure (hypertension) is not good but very low blood pressure (hypotension) is also bad;**
- **Simple Linear Regression does not allow us to study, say, quadratic relationships – or higher-order polynomials.**
- **And the case of seasonal diseases with time (around the year) as a predictor!**

We could use “data transformation” to establish “linearity”. However, sometimes it’s not possible. Sometimes, it is possible but the transformation is so complicated making it hard to interpret the results.

Age	Weight	Height	FEV1
53	161	61	2.63
40	198	72	3.95
26	210	69	3.87
34	187	68	3.74
46	131	62	2.9
44	193	72	4.91
35	135	64	3.39
45	166	69	4.19
45	180	68	4.29
30	176	66	4.49
46	188	70	3.9
50	179	68	3.24
31	210	74	4.88
37	195	67	4.01
40	190	70	4.56
32	170	71	4.41
37	218	74	4.64
54	223	72	2.55
34	176	66	3.83
40	198	69	4.59
48	179	64	1.86
37	131	63	2.87
30	110	57	2.04
47	122	65	2.75
35	145	62	2.92

#9. Lung Health

Y = FEV1,

X_1 = Age

X_2 = Weight

X_3 = Height

SIMPLE LINEAR REGRESSION RESULTS				
X = Weight				
R^2 = .5643				
	df	*SS*	*MS*	*F*
Regression	1	6.0316	6.0316	10.7446
Residual	23	12.9112	0.5614	
Total	24	18.9428		
X = Height				
R^2 = .7369				
	df	*SS*	*MS*	*F*
Regression	1	10.2890	10.289	27.3460
Residual	23	8.6538	0.3763	
Total	24	18.9428		
X = Age				
R^2 = .3452				
	df	*SS*	*MS*	*F*
Regression	1	2.2574	2.2574	3.1117
Residual	23	16.6854	0.7255	
Total	24	18.9428		

It is important to note that the sum of three R^2 exceeds 1.0; why, what does this mean?

Even without interactions, information provided by different factors may be "redundant" (overlapsed). There is a difference between "contribution" and "marginal contribution". Marginal contribution of a factor is its contribution on top of or in addition to the contributions by other factors.

MARGINAL CONTRIBUTION

In a study of Lung Health using FEV (Forced Expiratory Volume) as the Outcome or Response Variable, we could investigate the role of Age, Height, and Weight, etc… as potential predictors. Information contained in Age and Weight would be redundant (the sum of r^2 could exceed 100%); if we use Age, is the "remaining" information in Weight (not overlap with Age) still valuable?

SLR does not allow us to investigate "marginal contribution" of factors to adjust for redundancies

In a more technical topic: In “parallel-line bioassays”, for example, **Simple Linear Regression does not provide way to “fit” two regression lines – simultaneously – with the same slope** (we had to fit the lines separately then calculating the weighted average of the two slopes which may not be precise or ideal to use).

Similarly, in “slope-ratio bioassays”, Simple Linear Regression does not provide way **to “fit” two regression lines – simultaneously – with the same intercept.**

There are still many questions; most go beyond the territory of Simple Linear Regression. That is, we cannot answer them just using what we could in the context of SLR.

EXAMPLES

1. Is Age related to FEV independent of smoking status? Whether FEV & Smoking Status are related among the subjects having the same Age?
2. Is smoking status related to FEV independent of age? or whether FEV and Age are related among, say, smokers (or non-smokers)?
3. How much of the variability in FEV is explained by age and smoking combined? Would Smoking Status contribute beyond the contribution by Age?

SOME NEW ISSUES

- Is smoking related to FEV independent of age? A "**conditional**" question.
- How much of the variability in FEV is explained by age and smoking combined? Want to know the "**combined effects**"
- Would Smoking Status contribute beyond the contribution by Age? It concerns "**marginal contribution**".

SOME OTHER EXAMPLES

- How much of the variation in “test score” or “grade” can be explained by several student’s characteristics: age, gender, number of credits taken, family income?
- Is calcium intake related to Blood Pressure independent of Age?
- Is the relationship between Age and Blood Pressure the same for men and women?

A POSSIBLE SOLUTION

In order to provide more comprehensive prediction of the dependent variable Y – say the outcome of certain treatment, it is very desirable to:

(1) **Consider a large number of factors** – with available data - and

(2) **Sort out which ones are most closely related to that outcome.**

THE NEED

- We need a (**“multivariate” method** for “**risk determination**” or **“factor/risk analysis”.** That multivariate method is **Multiple (Linear) Regression**
- **Multiple linear regression** (MLR) involves **a linear combination of the explanatory or independent variables**.

THE 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 “fixed” part is the **Mean** and the “random” part is the error

THE MEAN IN A MLR MODEL

- Suppose we want to consider k independent variables simultaneously, the simple linear model can be easily generalized and expressed as:

  Mean of $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

- The β's are the (k+1) unknown parameters; β_0 is the intercept and β_i's are the slopes, one slope for each independent variables; x_i is the value of the ith independent variable (i = 1 to k) – considered as "fixed" or "designed".

A FEW THINGS WE CAN DO:

The response variable depends on more than one explanatory variable:

- See how combinations of several variables are associated with and can **jointly predict the dependent variable.**
- How much of the total variability (among the responses) can be explained (by predictors)?
- How to control for confounding effects (**interested in the effect of one variable but want to "adjust" for another variable**)
- Explore possible interactions

MODEL WITH TWO PREDICTORS

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

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

Easier to “see” with two independent variables; and what we know about models with two variables apply to models with more than two variables.

BINARY X_2

$$E(Y \mid X_1 = x_1, X_2 = 0) = \beta_0 + \beta_1 x_1$$

$$E(Y \mid X_1 = x_1, X_2 = 1) = \beta_0 + \beta_1 x_1 + \beta_2$$

$$E(Y \mid X_1 = x_1, X_2 = 1) - E(Y \mid X_1 = x_1, X_2 = 0) = \beta_2$$

Almost the same interpretation: If X_2 is binary (=0/1) representing an exposure, β_2 represents the increase in the mean of Y associated with the exposure X_2 (or a decrease if β_2 is negative) - provided that X_1 is fixed.

Example: X_1 = Age & X_2 = Smoking; β_2 represents the effect of Smoking among people of the same age – or effect of Smoking adjusted for Age.

CONTINUOUS X_2

$$E(Y \mid X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$E(Y \mid X_1 = x_1, X_2 = x_2 + 1) = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + 1)$$

$$E(Y \mid X_1 = x_1, X_2 = x_2 + 1) - E(Y \mid X_1 = x_1, X_2 = x_2) = \beta_2$$

Almost the same interpretation: If X_2 is on a continuous scale, β_2 represents the increase in the mean of Y (or a decrease if β_2 is negative) associated with one unit increase in the value of X_2, $X_2 = x_2+1$ vs. $X_2 = x_2$, **provided that X_1 is fixed**

Example: X_1 = Smoking & X_2 = Age; β_2 represents the effect on lung health due to one year of age among people with the same smoking status– or effect of Age adjusted for Smoking.

In other words, each slope parameter represents the "marginal contribution" of that independent variable – **above and beyond what contributed by other variables already in the model.**

EXAMPLE: MODELING "FEV"

Smokers

$FEV = b_0 + b_1 + b_2 age$

Non Smokers

$FEV = b_0 + b_2 age$

$FEV = b_0 + b_1(smoking) + b2\ (age)$

Smoking = 0/1 (non-smoker/smoker)

b_1 is the effect of smoking for fixed levels of age

b_2 is the effect of age adjusted for smoking status.

This "basic" model assumes the relation of age to FEV is the same for smokers and non-smokers (that is, no interaction)

MANY POSSIBILITIES

- If the Univariate Null Hypothesis is not rejected:
  - **X_1 is not related to Y regardless of X_2**
  - **But Y & X_1 maybe related after adjusting for X_2,**
- If the Univariate Null Hypothesis is rejected:
  - **X_1 is related to Y regardless of X_2**
  - **Y & X_1 maybe not related to Y after adjusting for X_2**
- **Relation of X_1 with Y could get stronger after adjusting for X_2**
- **Relation of X_1 with Y could get weaker after adjusting for X_2**

EFFECT MODIFICATIONS

- Consider the multiple regression model involving **2 independent variables** X_1 and X_2: Mean of $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- Basically, The effect of X_1 depends on the value of X_2 and vice versa. This phenomenon is called "effect modification" (or "interaction"), i.e. **one factor modifies the effect of the other.**

BINARY X_1 & CONTINUOUS X_2

$$\mathbf{E(Y \mid X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}$$

$$E(Y \mid X_1 = 0, X_2 = x_2 + 1) = \beta_0 + \beta_2 (x_2 + 1)$$

$$E(Y \mid X_1 = 0, X_2 = x_2) = \beta_0 + \beta_2 x_2$$

$$E(Y \mid X_1 = 0, X_2 = x_2 + 1) - E(Y \mid X_1 = 0, X_2 = x_2) = \beta_2$$

This (β_2) is the effect due to one unit of X_2 (say, Age) when $X_1 = 0$ (non-smokers)

BINARY X_1 & CONTINUOUS X_2

$$E(Y \mid X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

$$E(Y \mid X_1 = 1, X_2 = x_2 + 1) = \beta_0 + \beta_1 + \beta_2 (x_2 + 1) + \beta_3 (x_2 + 1)$$

$$E(Y \mid X_1 = 1, X_2 = x_2) = \beta_0 + \beta_1 + \beta_2 x_2 + \beta_3 x_2$$

$$E(Y \mid X_1 = 1, X_2 = x_2 + 1) - E(Y \mid X_1 = 1, X_2 = x_2) = \beta_2 + \beta_3$$

This (β_2+β_3) is the effect due to one unit of X_2 (say, Age) when X_1 = 1 (smokers)

(1)The effect due to one unit of X_2 (say, Age) when $X_1 = 0$ (non-smokers): β_2

(2) The effect due to one unit of X_2 (say, Age) when $X_1 = 1$ (smokers): $\beta_2 + \beta_3$

If $\beta_3 \neq 0$, Smoking modifies the effect of Age; Smoking is an "**Effect Modifier**".

Key for investigating: Focus on β_3; it measures the size/degree of effect modification.

QUADRATIC RELATIONSHIP

- Consider the multiple regression model involving **one independent variable** X: $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ where X is a continuous covariate.
- Look as 2 predictors: $X_1 = X$ and $X_2 = X^2$
- Quadratic and other "polynomial" models allow us to investigate non-linear relationships; however, polynomial models with an independent variable present in **higher powers** than the second are much less often used/seen.

POSSIBLE APPLICATIONS

- The second degree polynomial may **provide a <u>better fit</u> than a linear model;**
- **A quadratic model could be fitted for the purpose of <u>establishing the linearity</u>**; the key item to look for is whether the coefficient of the second power is zero.
- If a quadratic model fits, perhaps an useful application would be to **"optimize" the Mean of Y**: $\beta 0 + \beta 1x + \beta 2\ x2$, in order to determine the value of X at which the Mean of Y attains its maximum or minimum value (depending on the sign of β_2).

Problem #1: **The meaning of the regression coefficients here is not the same** as that given earlier because of the quadratic term:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$E(Y \mid X = x+1) = \beta_0 + \beta_1 (x+1) + \beta_2 (x+1)^2$$

$$E(Y \mid X = x+1) - E(Y \mid X = x) = \beta_1 + \beta_2 (2x+1)$$

which is not only β_1 but **a function of X = x**

Problem #2: Effect or contribution of a Predictor should be judged differently, e.g.:

In the model :

$$E(Y \mid X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$$

Effect of X_2 is tested against :

$$\mathbf{H_0 : \beta_2 = \beta_3 = 0}$$

GENERAL LINEAR MULTIPLE REGRESSION MODEL

- The linear multiple regression model can be generalized and expressed as:
  Mean of $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$
  plus a normally-distributed error term.
- However, **the X's 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.**
- The terms involved could be continuous, binary, or categorical (more than 2 categories)

The uses of products and higher power terms provide power tool and makes regression analysis even more popular but **they should be used cautiously** – especially the polynomial terms

BRIEF ON DATA ANALYSIS

- As parameters, regression coefficients are unknown.
- We estimate these unknown parameters by the very same method of least squares . We can evoke the “normal error regression model” and obtain standard errors for least squares estimates.
- Given the estimates of all regression coefficient, **we can estimate the mean response Y** (given a set of values of predictors X’s) and **individual response –** just as in the case of simple regression..

Data : $\{(x_i, y_i)\}_{i=1}^{n}$

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

Solve :

SLR: Least Squares Method

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

$$= 0$$

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

$$= 0$$

THE CASE OF 2 PREDICTORS

- By the Model, X=x, the Mean of Y is $\beta_0 + \beta_1 x_1 + \beta_2 x_2$.
- Let b_0, b_1, and b_2 are estimates of β_0, β_1 and β_2, respectively; then $(b_0 + b_1 x_1 + b_1 x_2)$ is an estimate of y. The error of that estimate is $[y - (b_0 + b_1 x_1 + b_1 x_2)]$ so that $Q=\Sigma [y - (b_0 + b_1 x_1 + b_1 x_2)]^2$ represents the "total errors" (not distinguishing an under-estimation from an over-estimation); called "the sum of squared errors"
- The **method of least squares** requires that we find "good estimates" of β_0, β_1 and β_2 the values b_0, b_1, and b_2 so as to minimize the "sum of squared errors Q".

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

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

Solve :

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

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

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

ANOVA IN REGRESSION

- The variation in Y is conventionally measured in terms of the deviations $(Y_i - \overline{Y})$'s; the total variation, denoted by SST, is the sum of squared deviations: $SST = \Sigma(Y_i - \overline{Y})^2$. For example, SST = 0 when all observations are the same; SST is the numerator of the sample variance of Y, the greater SST the greater the variation among Y-values.
- When we use the regression approach, the variation in Y is decomposed into two components:
$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$

ANOVA IN REGRESSION

- In the decomposition: $(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$
- The first term reflects the variation around the regression mean; the part than **cannot be explained** by the regression itself with the sum of squared deviations: $SSE = \Sigma(Y_i - \hat{Y}_i)^2$.
- The difference between the above two sums of squares, $SSR = SST - SSE = \Sigma(\hat{Y}_i - \overline{Y})^2$, is called the **regression sum of squares; SSR** may be considered a measure of the variation in Y associated with the regression model.

COEFFICIENT OF DETERMINATION

- The ratio, called the coefficient of 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.

ANALYSIS OF VARIANCE

- SST measures the “total variation” in the sample (of values of the dependent variable) with (n-1) degrees of freedom, n is the sample size. It is decomposed into: SST=SSE+SSR
- (1) SSE measures the variation cannot be explained by the regression with (n-k-1) degrees of freedom, and
- (2) SSR measures the variation in Y associated with the regression line with k degrees of freedom representing the k slopes.

“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.

TESTING HYPOTHESES

- Once we have fitted a multiple linear regression model and obtained estimates for the various parameters of interest, we want to **answer questions about the contributions of various factors** to the prediction of the future of patients. There are three types of tests:

(1) An overall test

(2) Test for the value of a single factor

(3) Test for contribution of a group of variables

OVERALL TEST

- The question is: " Taken collectively, does the entire set of explanatory or independent variables contribute significantly to the prediction of the Dependent Variable Y?".
- The **Null Hypothesis** for this test may stated as: "All k independent variables, considered together, do not explain the variation in the values of Y". In other words,

$$H_o : \beta_1 = \beta_2 = ... = \beta_k = 0$$

Global Test Of Significance

- **Hypotheses**

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

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

- **Test Statistic**

  F = 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).

TEST FOR SINGLE FACTOR

- The **question** is: “Does the addition of one **particular factor** of interest add significantly to the prediction of Dependent Variable over and above that achieved by other factors?”.
- The **Null Hypothesis** for this test may stated as: "Factor X_i does not have any value added to the explain the variation in Y-values when other factors are included in the model". In other words,

$$H_0 : \beta_i = 0$$

Test of Significance for each Slope

- **Hypotheses**

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

- **Test Statistic**

$$t = \frac{b_i}{s(b_i)}$$

- **Rejection Rule**

Reject H_0 for small or large t

TEST FOR A GROUP OF VARABLES

- The **question** is: “Does the addition of **a group of factors** add significantly to the prediction of Y over and above that achieved by other factors?
- The **Null Hypothesis** for this test may stated as: "Factors $\{X_i, X_{i+1}, \ldots, X_{i+m}\}$, considered together as a group, do not have any value added to the prediction of the Mean of Y that other factors are already included in the model". In other words,

$$H_0 : \beta_i = \beta_{i+1} = \ldots = \beta_{i+m} = 0$$

Application #1:

This "**multiple contribution**" Test is often used to test whether a similar group of variables, such as demographic characteristics, is important for the prediction of the mean of Y; these variables have some trait in common.

Application #2:

collection of powers and/or product terms.

It is of interest to assess powers & interaction effects collectively before considering individual interaction terms in a model. It reduces the total number of tests & helps to provide better control of overall Type I error rates which may be inflated due to multiple testing.

Example: PULMONARY FUNCTION

- Dependent Variable: Forced Expired Volume (FEV)
- Independent Variables:
  - Age of person
  - Smoking status of person
- Three Basic 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

```
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	R-Square	0.7038
Dependent Mean	3.44667		
Coeff Var	8.07136		

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

Proportion of variance explained by both variables

You can compare the results of multiple regression analysis versus results of simple regression analyses in terms of:
(1) regression coefficients, and
(2) coefficients of determination
And you could fit subgroup SLR as a way to explore possible interaction.

PROC REG; MODEL fev = age smk;
PROC REG; MODEL fev = age ; WHERE smk = 0;
PROC REG; MODEL fev = age ; WHERE smk = 1;

All subjects

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.58114	0.27653	20.18	<.0001
age	1	-0.04702	0.00634	-7.42	<.0001
smk	1	-0.40384	0.10242	-3.94	0.0005

Non-smokers

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.24764	0.38050	13.79	<.0001
age	1	-0.03911	0.00887	-4.41	0.0007

Smokers

Intercept	1	5.50002	0.36163	15.21	<.0001
age	1	-0.05508	0.00885	-6.22	<.0001

Readings & Exercises

- Readings: A thorough reading of the text's section 6.1 (pp.214-221) is highly recommended.
- Exercises: The following exercises are good for practice, all from chapter 6 of text: 6.5(b), 6.6(a,b), 6.7(a), and 6.10(a).

Due As Homework

#11.1 We have a data set on 56 people, 18 of them are non-smokers (File: Tobacco); two outcome or response variables are urinary Cotinine (a derivative of Nicotine), and urinary NNAL (a derivative of NNN, a toxin only comes from tobacco products). Data for 3 other explanatory variables are also included: Age, Gender, and type of tobacco used (non-smokers, smokers, and e-cigarette users). Some non=smokers are included because Cotinine and NNAL have been known transferable through environmental (or second-hand) smoke. Define an indicator (0/1) for Gender and two indicator variables representing type of tobacco used.

a) Let Y= Cotinine, fit the multiple linear regression model with all covariates (Age, Gender, and Type of tobacco used) and interpret the results – with emphasis on emphasis is on the difference between smokers and e-cigarette users and the difference between e-cigarette users and non-smokers (Note: this may affect your choice of baseline for type of tobacco used).

b) Repeat analyses in (a) with Y = NNAL. What do we do to decide whether to use NNAL on the log scale as most investigators do?

#11.2 We have data consisting of age and vital capacity (VC, liters) for each of 84 men working in the cadmium industry; data are stored in file “Cadmium Fumes”. They are divided into 3 groups: A1, exposed to cadmium fumes for at least 10 years; A2, exposed to fumes for less than 10 years; and B, not exposed to fumes. The three groups are represented by two indicator variables, with group B serves as baseline: X2 = 1 if man belongs to group A1 and X2 =0 otherwise; X3 = 1 if man belongs to group A2 and X3=0 otherwise.

a) A multiple regression model with Y= (100)(vital capacity) as the dependent variable and three independent variables (X1=Age, X2, and X3) was fitted; interpret the results (values of the estimated regression coefficients).

b) Plot the residuals against values of each of the two predictors; What do the plots suggest, any clear departures from the model?

c) Repeat (a) using only data from groups A1 and A2; Independent variables include X1=Age and X2 defined as (X2 = 1 if man belongs to group A1, and X2 = 0 if man belongs to group A2).