

PubH 7405: REGRESSION ANALYSIS



INTRODUCTORY MULTIPLE REGRESSION

SIMPLE LINEAR REGRESSION

- The method we have learned so far is called “**SIMPLE LINEAR REGRESSION**” (& simple 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?**

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.

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.

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 granulation 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?

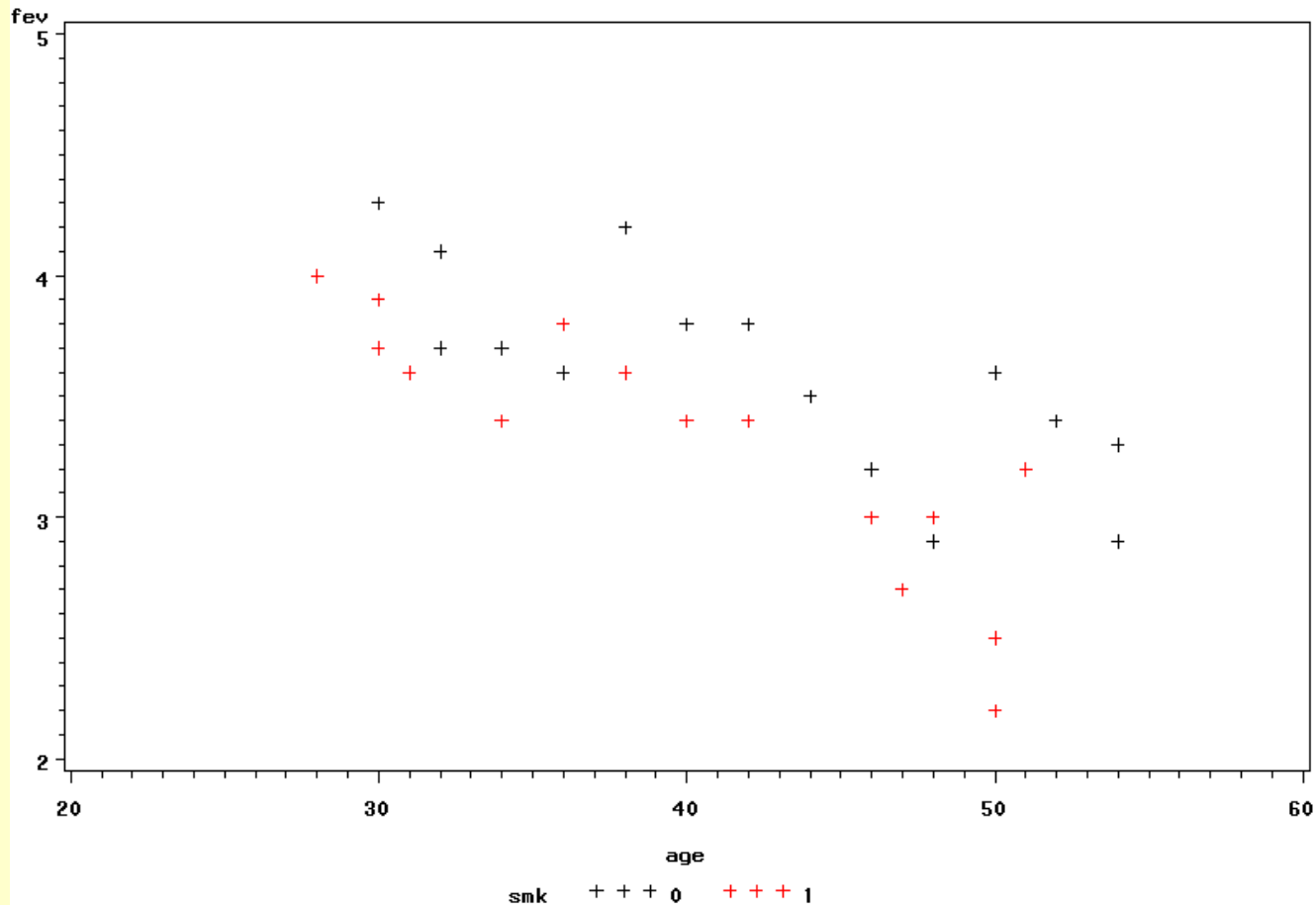
AG-Positive, n = 17		AG-Negative, n = 16	
White Blood count (WBC)	Survival Time (weeks)	White Blood Count (WBC)	Survival Time (weeks)
2,300	65	4,400	56
750	156	3,000	65
4,300	100	4,000	17
2,600	134	1,500	7
6,000	16	9,000	16
10,500	108	5,300	22
10,000	121	10,000	3
17,000	4	19,000	4
5,400	39	27,000	2
7,000	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.**

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 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.

Even without interactions, information provided by different factors may be “redundant” (overlapped). 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

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 (°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.

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 just using what we could in the context of SLR.

QUESTION #1a

Is Age related to FEV independent of smoking status? Or whether FEV and Smoking Status are related **among the subjects having the same Age?**

QUESTION #b

Is smoking status related to FEV independent of age? or whether FEV and Age are related **among**, say, smokers (or non-smokers)?

QUESTION #2

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:

$$\text{Mean of } Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- 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 i th 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

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

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

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

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: $X_1 = \text{Age}$ & $X_2 = \text{Smoking}$; β_2 represents the effect of Smoking among people of the same age – or effect of Smoking adjusted for Age.

CONTINUOUS X_2

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

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

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

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: $X_1 = \text{Smoking}$ & $X_2 = \text{Age}$; 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

$$FEV = b_0 + b_1 + b_2 \text{age}$$

Non Smokers

$$FEV = b_0 + b_2 \text{age}$$

$$FEV = b_0 + b_1(\text{smoking}) + b_2(\text{age})$$

$$\text{Smoking} = 0/1 \text{ (non-smoker/smoker)}$$

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 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
 - 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**

EXAMPLE:

EFFECT MODIFICATIONS

- Consider the multiple regression model involving two independent variables X_1 and X_2 :
Mean of $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$.
- 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. To include in the model the product of those two covariates is the way to investigate interaction.

BINARY X_1 & CONTINUOUS X_2

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

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

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

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

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

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

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

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

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

This $(\beta_2 + \beta_3)$ 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”.

EXAMPLE:

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.

The meaning of the regression coefficients here is not the same as that given earlier because of the **quadratic term**. We have, for example:

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

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

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

which is not β_1 but a function of $X = x$

POLYNOMIAL REGRESSION

- The second-order or quadratic model has been used in two different ways:
 - (i) when the true relationship is a second degree polynomial or when the true relationship is unknown but the second degree polynomial provides a better fit than a linear one, but
 - (ii) more often, a quadratic model is fitted for the purpose of establishing the linearity. The key item to look for is whether the coefficient of the second power is zero.

POLYNOMIAL REGRESSION

- If a quadratic model fits, perhaps an useful application would be to maximize/minimize the “Mean Index”,
 $\beta_1 x + \beta_2 x^2$,
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).

GENERAL LINEAR

MULTIPLE REGRESSION MODEL

- The linear multiple regression model can be generalized and expressed as:

$$\text{Mean of } Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

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)

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

$$\begin{aligned} \frac{\delta Q}{\delta \beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \frac{\delta Q}{\delta \beta_1} &= -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) \\ &= 0 \end{aligned}$$

THE CASE OF 2 PREDICTORS

- By the Model, $X=x$, the Mean of Y is $\beta_0 + \beta_1x_1 + \beta_2x_2$.
- Let b_0 , b_1 , and b_2 are estimates of β_0 , β_1 and β_2 , respectively; then $(b_0 + b_1x_1 + b_1x_2)$ is an estimate of y . The error of that estimate is $[y - (b_0 + b_1x_1 + b_1x_2)]$ so that $Q = \sum [y - (b_0 + b_1x_1 + b_1x_2)]^2$ 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 ”.

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

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

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 $(Y_i - \bar{Y})$'s; the total variation, denoted by SST , is the sum of squared deviations: $SST = \sum(Y_i - \bar{Y})^2$. 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:
$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

ANOVA IN REGRESSION

- In the decomposition: $(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$
- The first term reflects the variation around the regression mean; the part that **cannot be explained** by the regression itself with the sum of squared deviations: $SSE = \sum(Y_i - \hat{Y}_i)^2$.
- The difference between the above two sums of squares, $SSR = SST - SSE = \sum(\hat{Y}_i - \bar{Y})^2$, 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,

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

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**

$$\text{Reject } H_0 \text{ if } F > F_\alpha$$

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 VARIABLES

- 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 be stated as: "Factors $\{X_i, X_{i+1}, \dots, X_{i+m}\}$, 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,

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

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 $H_0: \beta_1 = \beta_2 = 0$

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

15.1 Refer to dataset “Cigarettes”, let $Y = \log(\text{NNAL})$ and consider a model with two independent variables, $X_1 = \text{CPD}$ and $X_2 = \text{Age}$: a) Fit the model and 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 data from the women (Gender = 2).

15.2 Answer the 3 questions of Exercise 15.1 using dataset “Infants” with $Y = \text{Birth Weight}$, $X_1 = \text{Gestational Weeks}$ and $X_2 = \text{Mother's Age}$. In part (c), use data from those case with toxemia = 1 (toxemia is a pregnancy condition resulting from metabolic disorder).

Only #15.2 is required