

PubH 7405: REGRESSION ANALYSIS



MILR: INFERENCE, Part I

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 factor or factors** to the prediction of the dependent variable Y. 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 factors**

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, we can write:

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

Logically, we could look for some individual effect – even without the overall test – if it’s the “primary aim” of the research project. For example, this factor represents “**treatment assignment**” in a **clinical trial** (say, =1 if treated & 0 if placebo)

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 in the model?**”.
- The **Null Hypothesis** for this test may be stated as: “Factor X_i does not have any value added to the explain the variation in Y-values over and above that achieved by other factors “. In other words, we can write:

$$H_0 : \beta_i = 0$$

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 $\{X_{i+1}, X_{i+2}, \dots, X_{i+m}\}$, considered together as a group, do not have any value added to the prediction of the Mean of Y over and above that achieved by other factors ". In other words,

$$H_0 : \beta_{i+1} = \beta_{i+2} = \dots = \beta_{i+m} = 0$$

TEST FOR A GROUP OF VARIABLES

- 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**.
- Other groupings: **Psychological and Social** aspects of health in **QofL research**

Another application: **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**.

This lecture is aimed to provide the details for these tests of significance. Since the primary focus is the **regression approach**, we'll go through the “**decomposition of the sums of squares**” – the basic approach of ANOVA (Analysis of “**Variance**”).

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**:
$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Three Basic SUMS OF SQUARES

- 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 (fitted value); the part that 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, $SSR = SST - SSE = \Sigma(\hat{Y}_i - \bar{Y})^2$, is called the **regression sum of squares**; SSR measures of the variation in Y associated with the regression model.
- Three basic sums of squares are: $SST = SSR + SSE$ (this result can be easily proved – and we did).

“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.

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, we can write:

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

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

Since:

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

If H_0 is true, the **average/mean response** has nothing to do with the predictors.

GLOBAL TEST OF SIGNIFICANCE

- Hypotheses

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

H_A : One or some 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).

In using “Multiple Regression”, the emphasis is in “**Marginal Contribution**”. For example, if we use the following two-factor” model:

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

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

The results would tell us:

(a) The effects of X_1 on Y after adjusted for X_2

(b) The effects of X_2 on Y after adjusted for X_1

It is possible that, individually, each of the two factors are highly correlated to the response; however, considered together in a multiple regression model, neither one is significant. In other words, **the “contribution” of each is significant but the “marginal contributions” are not significant – because they are redundant.**

SEQUENTIAL PROCESS

- SSR measures of the variation in Y associated with the regression model with k variables; Regression helps to “reduce” SST by the amount SSR to what left unexplained in SSE.
- Instead of a model with k independent variables, **let consider the sequential process to include one or some variables at a time.**
- By doing this we can focus on the “marginal contribution” of the variable or variables to be added – as **further reduction** in SSE.

“EXTRA” SUMS OF SQUARES

- An extra sum of squares measures the (marginal) reduction in the error sum of squares **when one or several independent variables are added to the regression model**, given that the other variables are already in the model.
- Equivalently, an extra sum of squares measures the (marginal) increase in the regression sum of squares when one or several independent variables are added to the regression model, given that the other variables are already in the model.

EXAMPLE

| 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 |

$$n = 25$$

$$Y = \text{FEV1},$$

a measure of Lung Health

$$X_1 = \text{Age}$$

$$X_2 = \text{Weight}$$

$$X_3 = \text{Height}$$

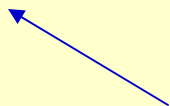
REGRESSION of FEV ON AGE

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| <i>Regression Statistics</i> | | | | |
| Multiple R | 0.3452 | | | |
| R Square | 0.1192 | | | |
| Observations | 25 | | | |
| <i>ANOVA</i> | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 1 | 2.257 | 2.257 | 3.112 |
| Residual | 23 | 16.685 | 0.725 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | 5.2531 | 0.9211 | 5.703 | <.001 |
| Age | -0.0401 | 0.0227 | -1.764 | 0.091 |

Note:

$$SSR(X_1) = 2.257$$

$$SSE(X_1) = 16.685$$



We add in (X_1) to identify the **model** used: X_1 only.

Regression of FEV ON AGE &

WEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| Regression Statistics | | | | |
| Multiple R | 0.6693 | | | |
| R Square | 0.4480 | | | |
| Observations | 25 | | | |
| ANOVA | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 2 | 8.486 | 4.243 | 8.927 |
| Residual | 22 | 10.457 | 0.475 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | 2.4241 | 1.0801 | 2.244 | 0.03512 |
| Age | -0.0418 | 0.0184 | -2.272 | 0.0332 |
| Weight | 0.0166 | 0.0046 | 3.620 | 0.0015 |

$$SSR(X_1, X_2) = 8.486$$

$$SSE(X_1, X_2) = 10.457$$

We add in (X_1, X_2) to identify the model used: X_1 & X_2



THE CHANGES

The first model, X_1 only:

$$SSR(X_1) = 2.257$$

$$SSE(X_1) = 16.685$$

The second model, X_1 and X_2 :

$$SSR(X_1, X_2) = 8.486$$

$$SSE(X_1, X_2) = 10.457$$

From the **first** to the **second** model: **SSR increases** and **SSE decreases**, by the **same** amount (because **SSR + SSE = SST**; **SST is constant** - unchanged regardless of the regression model used).

MARGINAL CONTRIBUTION

- The changes from the model containing only X_1 to the model containing both X_1 and X_2 represent the “marginal contribution” by X_2 .
- The marginal contribution represent the part (of SSE) that is further explained by X_2 (in addition to what already explained by X_1).
- **$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$** is the “**extra sum of squares due to the addition of X_2 to the model that already includes X_1 .**

The ANOVA Table was constructed to show the decomposition of SST into $SSR(X_1, X_2)$ and $SSE(X_1, X_2)$; **$SSR(X_1, X_2)$** represents the **combined contribution by X_1 and X_2** .

In the sequential process, we can further decompose **$SSR(X_1, X_2)$** to show the contribution of X_1 (which enters first), **$SSR(X_1)$** , and the marginal contribution of X_2 , **$SSR(X_2|X_1)$** ; this involves addition of one single factor, so this $SSR(X_2|X_1)$ is associated with one degree of freedom.

(Amended) ANOVA TABLE

- The breakdowns of the total sum of squares and its associated degree of freedom could be displayed as:
- (amended) ANOVA Table:

| Source of Variation | SS | df | MS |
|-------------------------|---------------|-----|---------------|
| Regression | $SSR(X1, X2)$ | 2 | $MSR(X1, X2)$ |
| Due to X1 | $SSR(X1)$ | 1 | $MSR(X1)$ |
| Addition of X2: $X2 X1$ | $SSR(X2 X1)$ | 1 | $MSR(X2 X1)$ |
| Error | $SSE(X1, X2)$ | n-3 | $MSE(X1, X2)$ |
| Total | SST | n-1 | |

Regression of FEV ON AGE, WEIGHT, & HEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| Regression Statistics | | | | |
| Multiple R | 0.821 | | | |
| R Square | 0.674 | | | |
| Observations | 25 | | | |
| ANOVA | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 3 | 12.767 | 4.256 | 14.471 |
| Residual | 21 | 6.176 | 0.294 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | -6.0588 | 2.3801 | -2.546 | 0.0188 |
| Age | -0.0404 | 0.0145 | -2.79111 | 0.0109 |
| Weight | -0.0046 | 0.0066 | -0.698 | 0.4928 |
| Height | 0.1802 | 0.0472 | 3.815 | 0.0010 |

$$X_3 = \text{Height}$$

$$\begin{aligned} \text{SSR}(X_1, X_2, X_3) &= 12.767 \\ \text{SSE}(X_1, X_2, X_3) &= 6.176 \end{aligned}$$

← We add in (X_1, X_2, X_3) to identify the model used: X_1 , X_2 , and X_3 .

THE CHANGES

The second model, X_1 and X_2 :

$$\text{SSR}(X_1, X_2) = 8.486$$

$$\text{SSE}(X_1, X_2) = 10.457$$

The third model, X_1 , X_2 and X_3 :

$$\text{SSR}(X_1, X_2, X_3) = 12.767$$

$$\text{SSE}(X_1, X_2, X_3) = 6.176$$

From the second to the third model: **SSR increases** and **SSE decreases**, by the **same** amount (because $\text{SSR} + \text{SSE} = \text{SST}$ regardless of the model used).

MARGINAL CONTRIBUTION

- The changes from the model containing X_1 and X_2 to the model containing X_1 , X_2 , and X_3 represent the “**marginal contribution**” by X_3 .
- The marginal contribution represent the part (of SSE) that is further explained by X_3 (in addition to what already explained by X_1 and X_2).
- $SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$ is the “extra sum of squares due to the **addition of X_3** to the model that already includes X_1 and X_2 .”

ANOVA TABLE

- The breakdowns of the total sum of squares and its associated degree of freedom could be displayed as:
- (amended) ANOVA Table:

| Source of Variation | SS | df | MS |
|------------------------------------|----------------------|-----|----------------------|
| Regression | $SSR(X_1, X_2, X_3)$ | 3 | $MSR(X_1, X_2, X_3)$ |
| Due to X_1 | $SSR(X_1)$ | 1 | $MSR(X_1)$ |
| Addition of X_2 : $X_2 X_1$ | $SSR(X_2 X_1)$ | 1 | $MSR(X_2 X_1)$ |
| Addition of X_3 : $X_3 X_1, X_2$ | $SSR(X_3 X_1, X_2)$ | 1 | $MSR(X_3 X_1, X_2)$ |
| Error | $SSE(X_1, X_2, X_3)$ | n-4 | $MSE(X_1, X_2, X_3)$ |
| Total | SST | n-1 | |

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 **over and above that achieved by other factors**". In other words,

$$H_0 : \beta_i = 0$$

TEST FOR SINGLE FACTOR #1

- The Null Hypothesis is $H_0 : \beta_i = 0$
- Regardless of the number of variables in the model, one simple approach is using
$$t = \frac{b_i}{SE(b_i)}$$
- Refer it to the percentiles of the **t-distribution** with df is the error degree of freedom, where b_i is the corresponding estimated regression coefficient and $SE(b_i)$ is the standard error of β_i , both of which are provided by any computer package – in **one “run”**.

TEST FOR SINGLE FACTOR #2a

- Suppose we have a multiple regression model with 2 independent variables (the **Full Model**):

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

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

& Suppose we are interested in the Null Hypothesis:

$$H_0 : \beta_2 = 0$$

- We can compare the Full Model to the **Reduced Model**:

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

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

TEST FOR SINGLE FACTOR #1a

- Test Statistic

$$F^* = \text{MSR}(X_2|X_1)/\text{MSE}(X_1, X_2)$$

- Rejection Rule

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

where F_α is based on an F distribution with **one (1) degree of freedom (numerator) and (n-3) degrees of freedom (denominator)**.

Regression of FEV ON AGE & WEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| <i>Regression Statistics</i> | | | | |
| Multiple R | 0.6693 | | | |
| R Square | 0.4480 | | | |
| Observations | 25 | | | |
| <i>ANOVA</i> | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 2 | 8.486 | 4.243 | 8.927 |
| Age | 1 | 2.257 | 2.257 | |
| Addition of Weight | 1 | 6.229 | 6.229 | 13.104 |
| Residual | 22 | 10.457 | 0.475 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | 2.4241 | 1.081 | 2.244 | 0.0352 |
| Age | -0.0418 | 0.0184 | -2.272 | 0.0332 |
| Weight | 0.0166 | 0.0046 | 3.620 | 0.0015 |

$$\begin{aligned} \text{MSR}(W|A) &= 6.229, \\ \text{MSE}(A,W) &= .475; \end{aligned}$$

$$F^* = 13.104, \text{ df} = (1,22); p = .0015$$

TEST FOR SINGLE FACTOR #2b

- Suppose we have a multiple regression model with 3 independent variables (the Full Model):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

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

& Suppose we are interested in the Null Hypothesis:

$$H_0 : \beta_3 = 0$$

- We can compare the Full Model to the Reduced Model:

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

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

TEST FOR SINGLE FACTOR #2b

- Test Statistic

$$F^* = \text{MSR}(X_3|X_1, X_2) / \text{MSE}(X_1, X_2, X_3)$$

- Rejection Rule

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

where F_α is based on an F distribution with **one degree of freedom (numerator) and (n-4) degrees of freedom (denominator)**.

Regression of FEV ON AGE, WEIGHT, & HEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| <i>Regression Statistics</i> | | | | |
| Multiple R | 0.8210 | | | |
| R Square | 0.6740 | | | |
| Observations | 25 | | | |
| <i>ANOVA</i> | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 3 | 12.767 | 4.256 | 14.471 |
| Age, Weight | 2 | 8.486 | | |
| Addition of Height | 1 | 4.281 | 4.281 | 14.557 |
| Residual | 21 | 6.176 | 0.294 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | -6.0588 | 2.3800 | -2.546 | 0.0188 |
| Age | -0.0404 | 0.0145 | -2.791 | 0.0109 |
| Weight | -0.0046 | 0.0066 | -0.699 | 0.4928 |
| Height | 0.1802 | 0.0472 | 3.815 | 0.0010 |

$$\begin{aligned} \text{MSR}(H|A, W) &= 4.281, \\ \text{MSE}(A, W, H) &= .294; \end{aligned}$$

$$F^* = 14.557, \text{ df} = (1, 21); p = .0010$$

We can perform an F test or a t-test.

Actually, the two tests are equivalent; numerical value of F^* is equal to the square of the calculated value of “t”: $F^* = t^2$. The t-test needs one computer run and the F test requires two computer runs – one with all variables in and one with the variable under investigated set aside.

Regression of FEV ON AGE & WEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| Regression Statistics | | | | |
| Multiple R | 0.6693 | | | |
| R Square | 0.4480 | | | |
| Observations | 25 | | | |
| ANOVA | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 2 | 8.486 | 4.243 | 8.927 |
| Age | 1 | 2.257 | 2.257 | |
| Addition of Weight | 1 | 6.229 | 6.229 | 13.104 |
| Residual | 22 | 10.457 | 0.475 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | 2.4241 | 1.081 | 2.244 | 0.0352 |
| Age | -0.0418 | 0.0184 | -2.272 | 0.0332 |
| Weight | 0.0166 | 0.0046 | 3.620 | 0.0015 |

$$\text{MSR}(W|A) = 6.229, \quad \text{MSE}(A,W) = .475;$$

$$F^* = 13.104, \text{ df} = (1,22); \text{ p} = .0015$$

$$13.104 = (3.62)**2$$

Regression of FEV ON AGE, WEIGHT, & HEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| Regression Statistics | | | | |
| Multiple R | 0.8210 | | | |
| R Square | 0.6740 | | | |
| Observations | 25 | | | |
| ANOVA | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 3 | 12.767 | 4.256 | 14.471 |
| Age, Weight | 2 | 8.486 | | |
| Addition of Height | 1 | 4.281 | 4.281 | 14.557 |
| Residual | 21 | 6.176 | 0.294 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | -6.0588 | 2.3800 | -2.546 | 0.0188 |
| Age | -0.0404 | 0.0145 | -2.791 | 0.0109 |
| Weight | -0.0046 | 0.0066 | -0.699 | 0.4928 |
| Height | 0.1802 | 0.0472 | 3.815 | 0.0010 |

HEIGHT

MSR(H|A,W) = 4.281, MSE(A,W,H) = .294;
 $F^* = 14.557$, $df = (1,21)$; $p = .0010$

$$14.557 = (3.815)**2$$

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 $\{X_{i+1}, X_{i+2}, \dots, X_{i+m}\}$, 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,

$$H_0 : \beta_{i+1} = \beta_{i+2} = \dots = \beta_{i+m} = 0$$

The “ANOVA Approach”, through the use of “extra sums of squares”, **was not really useful for testing concerning single factors** because we could more easily use the point estimate and standard error of the regression coefficient to form a t-test. In fact, we do not have to do anything; all results are obtained “automatically” from any computer package. However, the story is very **different for tests concerning a group of several variable: NO EASY WAY OUT; In the sequential process, we can add in more than one independent variables at a time.**

Example:

Back to the data set on Lung Health ($n = 25$):

$Y = \text{FEV1}$,

$X_1 = \text{Age}$

$X_2 = \text{Weight}$

$X_3 = \text{Height}$

Let consider the **additional combined value of Weight and Height** in the prediction of FEV1; the following models are fitted:

Model A: only Age

Model B: All three factors

| MODEL A | | | | |
|----------------|--------------|----------------|------------|---------|
| | df | SS | MS | F |
| Regression | 1 | 2.257 | 2.257 | 3.112 |
| Residual | 23 | 16.686 | 0.725 | |
| Total | 24 | 18.943 | | |
| | | | | |
| | Coefficients | Standard Error | t Stat | P-value |
| Intercept | 5.2531 | 0.9211 | 5.703 | <.001 |
| Age | -0.041 | 0.0227 | -1.7639997 | 0.091 |
| | | | | |
| MODEL B | | | | |
| | df | SS | MS | F |
| Regression | 3 | 12.767 | 4.256 | 14.471 |
| Residual | 21 | 6.176 | 0.294 | |
| Total | 24 | 18.943 | | |
| | | | | |
| | Coefficients | Standard Error | t Stat | P-value |
| Intercept | -6.0588 | 2.3801 | -2.546 | 0.0188 |
| Age | -0.0404 | 0.0145 | -2.791 | 0.0109 |
| Weight | -0.0046 | 0.0066 | -0.698 | 0.4928 |
| Height | 0.18023 | 0.0472 | 3.815 | 0.0010 |

THE CHANGES

In **Model A**, with X_1 only:

$$\text{SSR}(X_1) = 2.257$$

$$\text{SSE}(X_1) = 16.685$$

In **Model B**, with X_1 , X_2 , and X_3 :

$$\text{SSR}(X_1, X_2, X_3) = 12.767$$

$$\text{SSE}(X_1, X_2, X_3) = 6.176$$

From Model A to Model B: **SSR increases** and **SSE decreases**, by the **same amount** (because $\text{SSR} + \text{SSE} = \text{SST}$ regardless of the model used).

MARGINAL CONTRIBUTION

- The changes from the model containing only X_1 to the model containing all three variables represent the “marginal contribution” by the **addition of X_2 & X_3** .
- The marginal contribution represent the part (of SSE) that is further explained by X_2 and X_3 (in addition to what already explained by X_1).
- **$SSR(X_2, X_3 | X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$** is the “extra sum of squares due to the addition of X_2 and X_3 to the model that already includes X_1 .”

The ANOVA Table was constructed to show the decomposition of SST into $SSR(X_1, X_2, X_3)$ and $SSE(X_1, X_2, X_3)$; $SSR(X_1, X_2, X_3)$ represents the contribution by X_1 , X_2 , and X_3 .

In the sequential process, we can further decompose **$SSR(X_1, X_2, X_3)$** to show the contribution of X_1 (which enters first), **$SSR(X_1)$** , and the marginal contribution of X_2 & X_3 , **$SSR(X_2, X_3 | X_1)$** ; this involves addition of two variables, so this extra sum of square is associated with 2 degrees of freedom.

ANOVA TABLE

- The breakdowns of the total sum of squares and its associated degree of freedom could be displayed as:
- (amended) ANOVA Table:

| Source of Variation | SS | df | MS |
|-----------------------|----------------------|-----|----------------------|
| Regression | $SSR(X_1, X_2, X_3)$ | 3 | $MSR(X_1, X_2, X_3)$ |
| Due to X1 | $SSR(X_1)$ | 1 | $MSR(X_1)$ |
| Addition of X2, X3 X1 | $SSR(X_2, X_3 X_1)$ | 2 | $MSR(X_2, X_3 X_1)$ |
| Error | $SSE(X_1, X_2, X_3)$ | n-4 | $MSE(X_1, X_2, X_3)$ |
| Total | SST | n-1 | |

Regression of FEV ON AGE, WEIGHT, & HEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| <i>Regression Statistics</i> | | | | |
| Multiple R | 0.8210 | | | |
| R Square | 0.6740 | | | |
| Observations | 25 | | | |
| <i>ANOVA</i> | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 3 | 12.767 | 4.256 | 14.471 |
| Age | 1 | 2.257 | | |
| Addition of Weight & Height | 2 | 10.510 | 5.255 | 17.868 |
| Residual | 21 | 6.176 | 0.294 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | -6.0588 | 2.3801 | -2.546 | 0.0188 |
| Age | -0.0404 | 0.0145 | -2.791 | 0.0109 |
| Weight | -0.0046 | 0.0066 | -0.698 | 0.4928 |
| Height | 0.1802 | 0.0472 | 3.815 | 0.0010 |

$$\begin{aligned} \text{MSR(H,W|A)} &= 5.255 \\ \text{MSE(A,H,W)} &= .294 \end{aligned}$$

$$F^* = 17.868, df = (2,21); p = .00003$$

TEST FOR A GROUP OF VARIABLES

- 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.
- Another application: 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:

Back to the data set on Lung Health ($n = 25$):

$Y = \text{FEV1}$,

$X_1 = \text{Age}$

$X_2 = \text{Weight}$

$X_3 = \text{Height}$

Let consider the combined value of Weight and Height in the prediction of FEV1; the following models are fitted:

Model A: only 3 original variables (Age, Weight, Height)

Model B: All three original variables plus 3 squared terms (A^{**2} , W^{**2} , H^{**2}), and 3 product terms ($A*W$, $A*H$, $W*H$) – for a total 9 independent variables.

RESULTS FOR TWO MODELS

| MODEL A (3 variables) | | | | | |
|------------------------------|-----------|---------------|-----------|----------|-----------------------|
| <i>Regression Statistics</i> | | | | | |
| Multiple R | 0.8210 | | | | |
| R Square | 0.6740 | | | | |
| Observations | 25 | | | | |
| <i>ANOVA</i> | | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| Regression | 3 | 12.767 | 4.256 | 14.471 | <.001 |
| Residual | 21 | 6.176 | 0.294 | | |
| Total | 24 | 18.943 | | | |
| MODEL B (9 variables) | | | | | |
| <i>Regression Statistics</i> | | | | | |
| Multiple R | 0.9097 | | | | |
| R Square | 0.8276 | | | | |
| Observations | 25 | | | | |
| <i>ANOVA</i> | | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| Regression | 9 | 15.678 | 1.742 | 8.003 | <.001 |
| Residual | 15 | 3.265 | 0.218 | | |
| Total | 24 | 18.943 | | | |

| (Amended) ANOVA | | | | | |
|----------------------------------|----------|--------------|--------------|--------------|----------------|
| | df | SS | MS | F | Significance F |
| Regression | 9 | 15.678 | 1.742 | 8.003 | <.001 |
| Age, Weight, & Height | 3 | 12.767 | 4.256 | 14.47 | <.001 |
| Add Powers & Products | 6 | 2.911 | 0.485 | 2.229 | 0.0975 |
| Residual | 15 | 3.265 | 0.218 | | |
| Total | 24 | 18.943 | | | |

Together, the effects of the quadratic terms and product terms are marginal ($p = .0975$).

In the decomposition of the sums of squares; the “extra sums of squares” are not only useful for testing for the marginal contribution of individual variable or group of variables; they can also be used to “**measure**” these contributions. Unlike the statistical tests of significant, **the measures are not affected by the sample size.** Let start with the coefficient of multiple determination as a measure of the **proportionate reduction in the variation of Y** achieved by the regression model.

COEFFICIENT OF DETERMINATION

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

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

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

COEFFICIENT OF PARTIAL DETERMINATION #a

- Suppose we have a multiple regression model with 2 independent variables (the Full Model) and suppose we are interested in the marginal contribution of X_2 :

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

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

- The **coefficient of partial determination**, between Y and X_2 measures the marginal reduction in the variation of Y associated with the addition of X_2 , when X_1 is already in the model:

$$R_{Y2|1}^2 = \frac{SSR(X_2 | X_1)}{SSE(X_1)}$$

Regression of FEV ON AGE & WEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| Regression Statistics | | | | |
| Multiple R | 0.3452 | | | |
| R Square | 0.1192 | | | |
| Observations | 25 | | | |
| ANOVA | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 1 | 2.257 | 2.257 | 3.112 |
| Residual | 23 | 16.685 | 0.725 | |
| Total | 24 | 18.943 | | |
| Coefficients | | | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | 5.2531 | 0.9211 | 5.703 | <.001 |
| Age | -0.0401 | 0.0227 | -1.764 | 0.091 |

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| Regression Statistics | | | | |
| Multiple R | 0.6693 | | | |
| R Square | 0.4480 | | | |
| Observations | 25 | | | |
| ANOVA | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 2 | 8.486 | 4.243 | 8.927 |
| Age | 1 | 2.257 | 2.257 | |
| Addition of Weight | 1 | 6.229 | 6.229 | 13.104 |
| Residual | 22 | 10.457 | 0.475 | |
| Total | 24 | 18.943 | | |
| Coefficients | | | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | 2.4241 | 1.081 | 2.244 | 0.0352 |
| Age 1 | -0.0418 | 0.0184 | -2.272 | 0.0332 |
| Weight | 0.0166 | 0.0046 | 3.620 | 0.0015 |

$$SSE(A) = 16.686$$

$$SSR(W | A) = 6.229$$

$$R^2_{FEV \& W | A} = \frac{6.229}{16.686} = .3733, \text{ or } 37.33\%$$

COEFFICIENT OF PARTIAL DETERMINATION #b

- Suppose we have a multiple regression model with 3 independent variables (the Full Model) and suppose we are interested in the marginal contribution of X_3 :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

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

- The coefficient of partial determination, between Y and X_3 measures the marginal reduction in the variation of Y associated with the **addition of X_3 , when X_1 and X_2 are already in the model:**

$$R_{Y3|12}^2 = \frac{SSR(X_3 | X_1, X_2)}{SSE(X_1, X_2)}$$

Regression of FEV ON AGE, WEIGHT, & HEIGHT

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| Regression Statistics | | | | |
| Multiple R | 0.6693 | | | |
| R Square | 0.4480 | | | |
| Observations | 25 | | | |
| ANOVA | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 2 | 8.486 | 4.243 | 8.927 |
| Residual | 22 | 10.457 | 0.475 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | 2.4241 | 1.0801 | 2.244 | 0.03512 |
| Age | -0.0418 | 0.0184 | -2.272 | 0.0332 |
| Weight | 0.0166 | 0.0046 | 3.620 | 0.0015 |

| SUMMARY OUTPUT | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|
| Regression Statistics | | | | |
| Multiple R | 0.8210 | | | |
| R Square | 0.6740 | | | |
| Observations | 25 | | | |
| ANOVA | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> |
| Regression | 3 | 12.767 | 4.256 | 14.471 |
| Age, Weight | 2 | 8.486 | | |
| Addition of Height | 1 | 4.281 | 4.281 | 14.557 |
| Residual | 21 | 6.176 | 0.294 | |
| Total | 24 | 18.943 | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept | -6.0588 | 2.3800 | -2.546 | 0.0188 |
| Age | -0.0404 | 0.0145 | -2.791 | 0.0109 |
| Weight | -0.0046 | 0.0066 | -0.699 | 0.4928 |
| Height | 0.1802 | 0.0472 | 3.815 | 0.0010 |

$$SSE(A, W) = 10.457$$

$$SSR(H | A, W) = 4.281$$

$$R^2_{FEV \& H | A, W} = \frac{4.281}{10.157} = .4215, \text{ or } 42.15\%$$

Similarly, we can define:

$$R_{Y_{2,3}|1}^2 = \frac{SSR(X_2, X_3 | X_1)}{SSE(X_1)}$$

COEFFICIENTS OF PARTIAL CORRELATION

The square root of a coefficient of partial determination is called a **coefficient of partial correlation**. It is given the same sign as that of the corresponding regression coefficient in the fitted regression model. Coefficients of partial correlation are frequently used in practice, although they do not have a clear meaning as coefficients of partial determination nor the (single) coefficient of correlation.

AN IMPORTANT RESULT

Let both the response variable Y and the predictor under investigation (say, X_1) be both regressed against the other predictor variables already in the regression model and the residuals are obtained for each. These two sets of residuals reflect the part of each (Y and X_1) that is not linearly associated with the other predictor variables.

The (simple) correlation between the above two sets of residuals is equal to the Coefficient of Partial Correlation between Y and X_1 .

Readings & Exercises

- Readings: A thorough reading of the text's sections 7.1-7.4 (pp.256-271) is recommended.
- Exercises: The following exercises are good for practice, all from chapter 7 of text: 7.3-7.7, 7.11.

Due As Homework

18.1 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 $X_3 = \text{Gender}$:

a) Obtain the ANOVA table that decomposes the Sum of Squares Regression (SSR) into Sum of Squares associated with X_1 , extra Sum of Squares associated with X_2 given X_1 , and extra Sum of Squares associated with X_3 given X_1 and X_2 .

b) Test whether X_3 can be dropped from the regression model given that X_1 and X_2 are retained

c) Test whether X_2 and X_3 can be both dropped from the regression model given that X_1 is retained

d) Is it always the case that $SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2)$?

18.2 Answer the 4 questions of Exercise 18.1 using dataset “Infants” with $Y = \text{Birth Weight}$, $X_1 = \text{Gestational Weeks}$, $X_2 = \text{Mother's Age}$, and $X_3 = \text{Toxemia}$ (toxemia is a pregnancy condition resulting from metabolic disorder).

Only #18.2 is required