

STAT 462

Applied Regression Analysis

Lesson 5: Multiple Linear Regression (MLR) Model & Evaluation

Overview of this Lesson

In this lesson, we make our first (and last?!) major jump in the course. We move from the simple linear regression model with one predictor to the multiple linear regression model with two or more predictors. That is, we use the adjective "simple" to denote that our model has only predictor, and we use the adjective "multiple" to indicate that our model has at least two predictors.

In the multiple regression setting, because of the potentially large number of predictors, it is more efficient to use matrices to define the regression model and the subsequent analyses. This lesson considers some of the more important multiple regression formulas in matrix form. If you're unsure about any of this, it may be a good time to take a look at this [Matrix Algebra Review](https://onlinecourses.science.psu.edu/statprogram/matrix_review) (https://onlinecourses.science.psu.edu/statprogram/matrix_review) .

The good news is that everything you learned about the simple linear regression model extends — with at most minor modification — to the multiple linear regression model. Think about it — you don't have to forget all of that good stuff you learned! In particular:

- The models have similar "LINE" assumptions. The only real difference is that whereas in simple linear regression we think of the distribution of errors at a fixed value of the single predictor, with multiple linear regression we have to think of the distribution of errors at a fixed set of values for all the predictors. All of the model checking procedures we learned earlier are useful in the multiple linear regression framework, although the process becomes more involved since we now have multiple predictors. We'll explore this issue further in Lesson 6.
- The use and interpretation of r^2 (which we'll denote R^2 in the context of multiple linear regression) remains the same. However, with multiple linear regression we can also make use of an "adjusted" R^2 value, which is useful for model building purposes. We'll explore this measure further in Lesson 11.
- With a minor generalization of the degrees of freedom, we use t -tests and t -intervals for the regression slope coefficients to assess whether a predictor is significantly linearly related to the response, after controlling for the effects of all the other predictors in the model.
- With a minor generalization of the degrees of freedom, we use confidence intervals for estimating the mean response and prediction intervals for predicting an individual response. We'll explore these further in Lesson 6.

For the simple linear regression model, there is only one slope parameter about which one can perform hypothesis tests. For the multiple linear regression model, there are three different hypothesis tests for slopes that one could conduct. They are:

- a hypothesis test for testing that *one* slope parameter is 0
- a hypothesis test for testing that *all* of the slope parameters are 0
- a hypothesis test for testing that a *subset* — more than one, but not all — of the slope parameters are 0

In this lesson, we also learn how to perform each of the above three hypothesis tests.

Key Learning Goals for this Lesson:

- Be able to interpret the coefficients of a multiple regression model.
- Understand what the scope of the model is in the multiple regression model.
- Understand the calculation and interpretation of R^2 in a multiple regression setting.
- Understand the calculation and use of adjusted R^2 in a multiple regression setting.
- Translate research questions involving slope parameters into the appropriate hypotheses for testing.
- Know how to calculate a confidence interval for a single slope parameter in the multiple regression setting.
- Understand the general idea behind the general linear F-test.
- Understand the decomposition of a regression sum of squares into a sum of sequential sums of squares.
- Calculate a sequential sums of squares using either of the two definitions.
- Know how to obtain a two (or more)-degree-of-freedom sequential sum of squares.
- Perform a general hypothesis test using the general linear F-test and relevant statistical software output.
- Know how to specify the null and alternative hypotheses and be able to draw a conclusion given appropriate software output for the overall F -test for $H_0: \beta_1 = \dots = \beta_k = 0$.
- Know how to specify the null and alternative hypotheses and be able to draw a conclusion given appropriate software output for the general linear F -test for any subset of the slope parameters.
- Know how to specify the null and alternative hypotheses and be able to draw a conclusion given appropriate software output for the t -test or general linear F -test for $H_0: \beta_p = 0$.
- Understand that the t -test for a slope parameter tests the *marginal* significance of the predictor after adjusting for the other predictors in the model (as can be justified by the equivalence of the t -test and the corresponding general linear F -test for one slope).
- Calculate and understand partial R^2 .

- 5.1 - Example on IQ and Physical Characteristics (/stat462/node/129)
- 5.2 - Example on Underground Air Quality (/stat462/node/130)
- 5.3 - The Multiple Linear Regression Model (/stat462/node/131)
- 5.4 - A Matrix Formulation of the Multiple Regression Model (/stat462/node/132)
- 5.5 - Three Types of MLR Parameter Tests (/stat462/node/134)
- 5.6 - The General Linear F-Test (/stat462/node/135)
- 5.7 - MLR Parameter Tests (/stat462/node/137)
- 5.8 - Partial R-squared (/stat462/node/138)
- 5.9 - Further MLR Examples (/stat462/node/133)

STAT 462

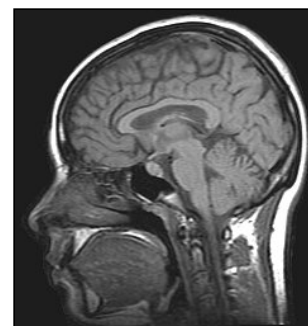
Applied Regression Analysis

5.1 - Example on IQ and Physical Characteristics

Let's jump in and take a look at some "real-life" examples in which a multiple linear regression model is used. Make sure you notice, in each case, that the model has more than one predictor. You might also try to pay attention to the similarities and differences among the examples and their resulting models. Most of all, don't worry about mastering all of the details now. In the upcoming lessons, we will re-visit similar examples in greater detail. For now, my hope is that these examples leave you with an appreciation of the richness of multiple regression.

Are a person's brain size and body size predictive of his or her intelligence?

Interested in answering the above research question, some researchers (Willerman, *et al*, 1991) collected the following data (iqsize.txt
(/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/iqsize.txt)) on a sample of $n = 38$ college students:



- Response (y): Performance IQ scores (**PIQ**) from the revised Wechsler Adult Intelligence Scale. This variable served as the investigator's measure of the individual's intelligence.
- Potential predictor (x_1): Brain size based on the count obtained from **MRI** scans (given as count/10,000).
- Potential predictor (x_2): **Height** in inches.
- Potential predictor (x_3): **Weight** in pounds.

As always, the first thing we should want to do when presented with a set of data is to plot it. And, of course, plotting the data is a little more challenging in the multiple regression setting, as there is one scatter plot for each pair of variables. Not only do we have to consider the relationship between the response and each of the predictors, we also have to consider how the predictors are related among each other.

A common way of investigating the relationships among all of the variables is by way of a "**scatter plot matrix**." Basically, a scatter plot matrix contains a scatter plot of each pair of variables arranged in an orderly array. Here's what one version of a scatter plot matrix looks like for our brain and body size example:



Click to enable Adobe Flash Player

For each scatter plot in the matrix, the variable on the y -axis appears at the left end of the plot's row and the variable on the x -axis appears at the bottom of the plot's column. Try to identify the variables on the y -axis and x -axis in each of the six scatter plots appearing in the matrix. You can check your understanding by **rolling your mouse over each scatter plot** appearing in the above matrix.

Incidentally, in case you are wondering, the tick marks on each of the axes are located at 25% and 75% of the data range from the minimum. That is:

- the first tick = $((\text{maximum} - \text{minimum}) * 0.25) + \text{minimum}$
- the second tick = $((\text{maximum} - \text{minimum}) * 0.75) + \text{minimum}$

Sometimes software packages use a different scheme in labeling the scatter plot matrix. For each plot in the following scatter plot matrix, the variable on the y -axis appears at the right end of the plot's row and the variable on the x -axis appears at the top of the plot's column. Again, you can **roll your mouse over each scatter plot** appearing in the matrix to make sure you understand this different labeling scheme:



Click to enable Adobe Flash Player

Now, what does a scatter plot matrix tell us? Of course, one use of the plots is simple data checking. Are there any egregiously erroneous data errors? The scatter plots also illustrate the "**marginal relationships**" between each pair of variables *without regard to the other variables*. For example, it appears that brain size is the best single predictor of PIQ, but none of the relationships are particularly strong. In multiple linear regression, the challenge is to see how the response y relates to all three predictors simultaneously.

Loading [MathJax]/extensions/MathZoom.js

We always start a regression analysis by formulating a model for our data. One possible multiple linear regression **model with three quantitative predictors** for our brain and body size example is:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

where:

- y_i is the intelligence (**PIQ**) of student i
- x_{i1} is the brain size (**MRI**) of student i
- x_{i2} is the height (**Height**) of student i
- x_{i3} is the weight (**Weight**) of student i

and the **independent** error terms ϵ_i follow a **normal** distribution with mean 0 and **equal variance** σ^2 .

A couple of things to note about this model:

- Because we have more than one predictor (x) variable, we use slightly modified notation. The x -variables (e.g., x_{i1} , x_{i2} , and x_{i3}) are now subscripted with a 1, 2, and 3 as a way of keeping track of the three different quantitative variables. We also subscript the slope parameters with the corresponding numbers (e.g., β_1 , β_2 and β_3).
- The "LINE" conditions must still hold for the multiple linear regression model. The linear part comes from the formulated regression function — it is, what we say, "**linear in the parameters**." This simply means that each beta coefficient multiplies a predictor variable or a transformation of one or more predictor variables. We'll see in Lesson 7 that this means that, for example, the model, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, is a multiple *linear* regression model even though it represents a curved relationship between y and x .

Of course, our interest in performing a regression analysis is almost always to answer some sort of research question. Can you think of some research questions that the researchers might want to answer here? How about the following set of questions? What procedure would you use to answer each research question? (Do the procedures that appear in parentheses seem reasonable?)

- Which, if any, predictors — brain size, height, or weight — explain some of the variation in intelligence scores? (Conduct hypothesis tests for individually testing whether each slope parameter could be 0.)
- What is the effect of brain size on PIQ, after taking into account height and weight? (Calculate and interpret a confidence interval for the brain size slope parameter.)
- What is the PIQ of an individual with a given brain size, height, and weight? (Calculate and interpret a prediction interval for the response.)

Let's take a look at statistical software output for the multiple regression model we formulated above:

Regression Analysis: PIQ versus Brain, Height, Weight

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	5572.7	1857.58	4.74	0.007
Brain	1	5239.2	5239.23	13.37	0.001
Height	1	1934.7	1934.71	4.94	0.033
Weight	1	0.0	0.00	0.00	0.998
Error	34	13321.8	391.82		
Total	37	18894.6			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
19.7944	29.49%	23.27%	12.76%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	111.4	63.0	1.77	0.086	
Brain	2.060	0.563	3.66	0.001	1.58
Height	-2.73	1.23	-2.22	0.033	2.28
Weight	0.001	0.197	0.00	0.998	2.02

Regression Equation

$$\text{PIQ} = 111.4 + 2.060 \text{ Brain} - 2.73 \text{ Height} + 0.001 \text{ Weight}$$

My hope is that you immediately observe that much of the output looks the same as before! The only substantial differences are:

- More predictors appear in the estimated regression equation and therefore also in the column labeled "Term" in the table of estimates.
- There is an additional row for each predictor term in the Analysis of Variance Table. The label "Adj SS" indicates that these represent the increase in regression sums of squares for each term relative to a model that contains all of the other terms in the model (so-called Adjusted or Type III sums of squares). It is usually possible to instead use Sequential or Type I sums of squares, which represent the increase in regression sums of squares when a term is added to a model that contains only the terms listed before it in the ANOVA table.

We'll learn more about these differences later, but let's focus now on what you already know. The output tells us that:

- The R^2 value is 29.49%. This tells us that 29.49% of the variation in intelligence, as quantified by PIQ, is reduced by taking into account brain size, height and weight.
- The Adjusted R^2 value — denoted "**R-sq(adj)**" — is 23.27%. When considering different multiple linear regression models for PIQ, we could use this value to help compare the models.
- The P -values for the t -tests appearing in the table of estimates suggest that the slope parameters for Brain ($P = 0.001$) and Height ($P = 0.033$) are significantly different from 0, while the slope parameter for Weight ($P = 0.998$) is not.
- The P -value for the analysis of variance F -test ($P = 0.007$) suggests that the model containing Brain, Height and Weight is more useful in predicting intelligence than not taking into account the three predictors. (Note that the model with the three predictors is the *best* model!)

Loading [MathJax]/extensions/MathZoom.js

So, we already have a pretty good start on this multiple linear regression stuff. Let's take a look at another example.

◀ Lesson 5: Multiple Linear Regression (MLR) Model & Evaluation (/stat462/node/83)	up (/stat462/node/83)	5.2 - Example on Underground Air Quality ▶ (/stat462/node/130)
---	--	--

STAT 462

Applied Regression Analysis

5.2 - Example on Underground Air Quality

What are the breathing habits of baby birds that live in underground burrows?

Some mammals burrow into the ground to live. Scientists have found that the quality of the air in these burrows is not as good as the air above ground. In fact, some mammals change the way that they breathe in order to accommodate living in the poor air quality conditions underground.

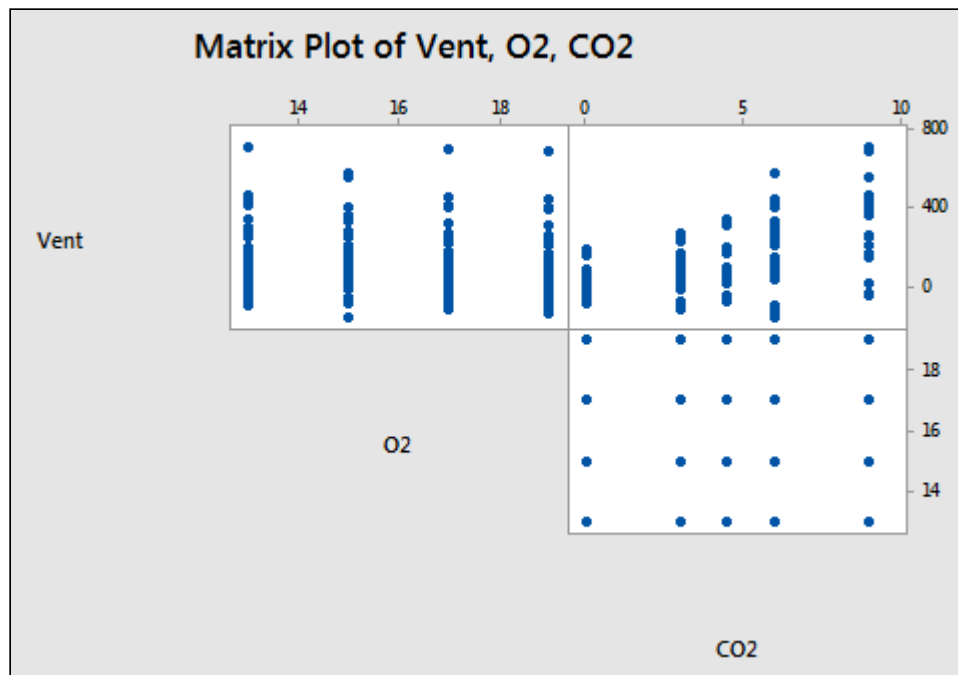
Some researchers (Colby, *et al*, 1987) wanted to find out if nestling bank swallows, which live in underground burrows, also alter how they breathe. The researchers conducted a randomized experiment on $n = 120$ nestling bank swallows. In an underground burrow, they varied the percentage of oxygen at four different levels (13%, 15%, 17%, and 19%) and the percentage of carbon dioxide at five different levels (0%, 3%, 4.5%, 6%, and 9%). Under each of the resulting $5 \times 4 = 20$ experimental conditions, the researchers observed the total volume of air breathed per minute for each of 6 nestling bank swallows. In this way, they obtained the following data (babybirds.txt

(/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/babybirds.txt)) on the $n = 120$ nestling bank swallows:

- Response (y): percentage increase in "minute ventilation," (**Vent**), *i.e.*, total volume of air breathed per minute.
- Potential predictor (x_1): percentage of oxygen (**O2**) in the air the baby birds breathe.
- Potential predictor (x_2): percentage of carbon dioxide (**CO2**) in the air the baby birds breathe.

Here's a scatter plot matrix of the resulting data obtained by the researchers:

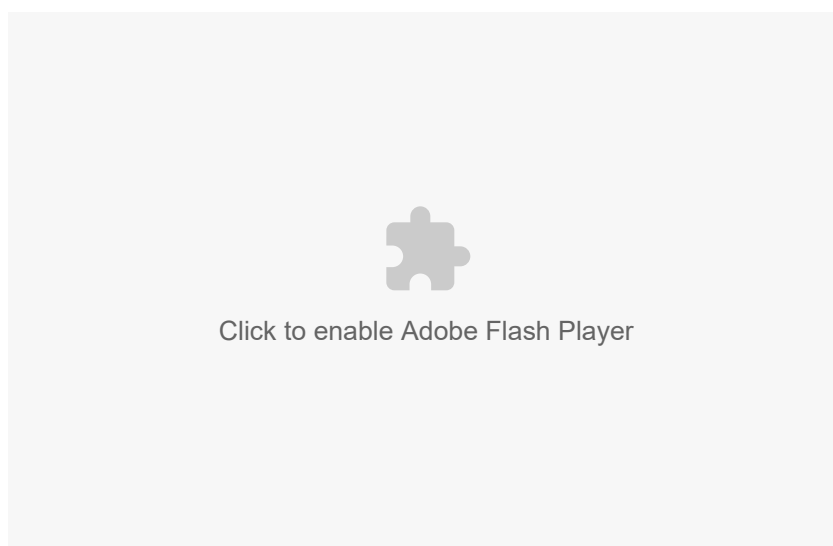




What does this particular scatter plot matrix tell us? Do you buy into the following statements?

- There doesn't appear to be a substantial relationship between minute ventilation (**Vent**) and percentage of oxygen (**O2**).
- The relationship between minute ventilation (**Vent**) and percentage of carbon dioxide (**CO2**) appears to be curved and with increasing error variance.
- The plot between percentage of oxygen (**O2**) and percentage of carbon dioxide (**CO2**) is the classical appearance of a scatter plot for the experimental conditions. The plot suggests that there is no correlation at all between the two variables. You should be able to observe from the plot the 4 levels of **O2** and the 5 levels of **CO2** that make up the $5 \times 4 = 20$ experimental conditions.

When we have one response variable and only two predictor variables, we have another sometimes useful plot at our disposal, namely a "**three-dimensional scatter plot**:"



If we added the estimated regression equation to the plot, what one word do you think describes what it would look like? Click the "**Draw Plane**" button in the above animation to draw the plot of the estimated regression equation for this data. Does it make sense that it looks like a "**plane**?" Incidentally, it is still important to remember that the plane depicted in the plot is just an estimate of the actual plane in the population that we are trying to study.

Here is a reasonable "**first-order**" model with **two quantitative predictors** that we could consider when trying to summarize the trend in the data:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) + \epsilon_i$$

where:

- y_i is percentage of minute ventilation of nestling bank swallow i
- x_{i1} is percentage of oxygen exposed to nestling bank swallow i
- x_{i2} is percentage of carbon dioxide exposed to nestling bank swallow i

and the **independent** error terms ϵ_i follow a **normal** distribution with mean 0 and **equal variance** σ^2 .

The adjective "**first-order**" is used to characterize a model in which the highest power on all of the predictor terms is one. In this case, the power on x_{i1} , although typically not shown, is one. And, the power on x_{i2} is also one, although not shown. Therefore, the model we formulated can be classified as a "first-order model." An example of a second-order model would be $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$.

Do you have your research questions ready? How about the following set of questions? (Do the procedures that appear in parentheses seem appropriate in answering the research question?)

- Is oxygen related to minute ventilation, after taking into account carbon dioxide? (Conduct a hypothesis test for testing whether the O2 slope parameter is 0.)
- Is carbon dioxide related to minute ventilation, after taking into account oxygen? (Conduct a hypothesis test for testing whether the CO2 slope parameter is 0.)
- What is the mean minute ventilation of all nestling bank swallows whose breathing air is comprised of 15% oxygen and 5% carbon dioxide? (Calculate and interpret a confidence interval for the mean response.)

Here's statistical software output for the multiple regression model we formulated above:

Regression Analysis: Vent versus O2, CO2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1061819	530909	21.44	0.000
O2	1	17045	17045	0.69	0.408
CO2	1	1044773	1044773	42.19	0.000
Error	117	2897566	24766		
Lack-of-Fit	17	91284	5370	0.19	1.000
Pure Error	100	2806283	28063		
Total	119	3959385			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
157.371	26.82%	25.57%	22.78%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	86	106	0.81	0.419	
O2	-5.33	6.42	-0.83	0.408	1.00
CO2	31.10	4.79	6.50	0.000	1.00

Regression Equation

$$\text{Vent} = 86 - 5.33 \text{ O2} + 31.10 \text{ CO2}$$

What do we learn from the output?

- Only 26.82% of the variation in minute ventilation is reduced by taking into account the percentages of oxygen and carbon dioxide.
- The P -values for the t -tests appearing in the table of estimates suggest that the slope parameter for carbon dioxide level ($P < 0.001$) is significantly different from 0, while the slope parameter for oxygen level ($P = 0.408$) is not. Does this conclusion appear consistent with the above scatter plot matrix and the three-dimensional plot? Yes!
- The P -value for the analysis of variance F -test ($P < 0.001$) suggests that the model containing oxygen and carbon dioxide levels is more useful in predicting minute ventilation than not taking into account the two predictors. (Again, the F -test does not tell us that the model with the two predictors is the *best* model! For one thing, we have performed no model checking yet!)

◀ 5.1 - Example on IQ and Physical Characteristics (/stat462/node/129)

up (/stat462/node/83)

5.3 - The Multiple Linear Regression Model › (/stat462/node/131)

STAT 462

Applied Regression Analysis

5.3 - The Multiple Linear Regression Model

Notation for the Population Model

- A population model for a multiple linear regression model that relates a y -variable to k x -variables is written as

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_k x_{i,k} + \epsilon_i.$$

- Here we're using " k " for the number of predictor variables, which means we have $k+1$ regression parameters (the β coefficients). Some textbooks use " p " for the number of regression parameters and $p-1$ for the number of predictor variables.
- We assume that the ϵ_i have a normal distribution with mean 0 and constant variance σ^2 . These are the same assumptions that we used in simple regression with one x -variable.
- The subscript i refers to the i^{th} individual or unit in the population. In the notation for the x -variables, the subscript following i simply denotes which x -variable it is.
- The word "linear" in "multiple linear regression" refers to the fact that the model is *linear in the parameters*, $\beta_0, \beta_1, \dots, \beta_k$. This simply means that each parameter multiplies an x -variable, while the regression function is a sum of these "parameter times x -variable" terms. Each x -variable can be a predictor variable or a transformation of predictor variables (such as the square of a predictor variable or two predictor variables multiplied together). Allowing non-linear transformation of predictor variables like this enables the multiple linear regression model to represent non-linear relationships between the response variable and the predictor variables. We'll explore predictor transformations further in Lesson 7. Note that even β_0 represents a "parameter times x -variable" term if you think of the x -variable that is multiplied by β_0 as being the constant function "1."

Estimates of the Model Parameters

- The estimates of the β coefficients are the values that minimize the sum of squared errors for the sample. The exact formula for this is given in the next section on matrix notation.
- The letter b is used to represent a sample estimate of a β coefficient. Thus b_0 is the sample estimate of β_0 , b_1 is the sample estimate of β_1 , and so on.
- $MSE = \frac{SSE}{n-(k+1)}$ estimates σ^2 , the variance of the errors. In the formula, n = sample size, $k+1$ = number of β coefficients in the model (including the intercept) and SSE = sum of squared errors. Notice that simple linear regression has $k=1$ predictor variable, so $k+1 = 2$. Thus, we get the formula for MSE that we introduced in that context of one predictor.
- $S = \sqrt{MSE}$ estimates σ and is known as the *regression standard error* or the *residual standard error*.
- In the case of two predictors, the estimated regression equation yields a plane (as opposed to a line in the simple linear regression setting). For more than two predictors, the estimated regression equation yields a hyperplane.

Interpretation of the Model Parameters

- Each β coefficient represents the change in the mean response, $E(y)$, per unit increase in the associated predictor variable when all the other predictors are held constant.
- For example, β_1 represents the change in the mean response, $E(y)$, per unit increase in x_1 when x_2, x_3, \dots, x_k are held constant.
- The intercept term, β_0 , represents the mean response, $E(y)$, when all the predictors x_1, x_2, \dots, x_k , are all zero (which may or may not have any practical meaning).

Fitted Values and Residuals

- A **fitted (or predicted) value** is calculated as $\hat{y}_i = b_0 + b_1x_{i,1} + b_2x_{i,2} + \dots + b_kx_{i,k}$, where the b values come from statistical software and the x -values are specified by us.
- A **residual (error)** term is calculated as $e_i = y_i - \hat{y}_i$, the difference between an actual and a predicted value of y .
- A **plot of residuals versus fitted values** ideally should resemble a horizontal random band. Departures from this form indicates difficulties with the model and/or data.
- Other residual analyses can be done exactly as we did in simple regression. For instance, we might wish to examine a normal probability plot (NPP) of the residuals. Additional plots to consider are plots of residuals versus each x -variable separately. This might help us identify sources of curvature or nonconstant variance. Plots of residuals versus potential x -variables not included in the model might help us identify x -variables that should be included. We'll explore this further in Lesson 6.

ANOVA Table

Source	df	SS	MS	F
Regression	k	SSR	$MSR = SSR / k$	MSR / MSE
Error	$n - (k+1)$	SSE	$MSE = SSE / (n - (k+1))$	
Total	$n - 1$	SSTO		

Coefficient of Determination, R-squared, and Adjusted R-squared

- As in simple linear regression, $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$, and represents the proportion of variation in y (about its mean) "explained" by the multiple linear regression model with predictors, x_1, x_2, \dots
- If we start with a simple linear regression model with one predictor variable, x_1 , then add a second predictor variable, x_2 , SSE will decrease (or stay the same) while $SSTO$ remains constant, and so R^2 will increase (or stay the same). In other words, R^2 always increases (or stays the same) as more predictors are added to a multiple linear regression model, *even if the predictors added are unrelated to the response variable*. Thus, by itself, R^2 cannot be used to help us identify which predictors should be included in a model and which should be excluded.
- An alternative measure, adjusted R^2 , does not necessarily increase as more predictors are added, and can be used to help us identify which predictors should be included in a model and which should be excluded. Adjusted $R^2 = 1 - \left(\frac{n-1}{n-(k+1)} \right) (1 - R^2)$, and, while it has no practical interpretation, is useful for such model building purposes. Simply stated, when comparing two models used to predict the same response variable, we generally prefer the model with the higher value of adjusted R^2 – see Lesson 11 for more details.

Significance Testing of Each Variable

Within a multiple regression model, we may want to know whether a particular x -variable is making a useful contribution to the model. That is, given the presence of the other x -variables in the model, does a particular x -variable help us predict or explain the y -variable? For instance, suppose that we have three x -variables in the model. The general structure of the model could be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

As an example, to determine whether variable x_1 is a useful predictor variable in this model, we could test

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_A : \beta_1 &\neq 0. \end{aligned}$$

If the null hypothesis above were the case, then a change in the value of x_1 would not change y , so y and x_1 are not linearly related. Also, we would still be left with variables x_2 and x_3 being present in the model. When we cannot reject the null hypothesis above, we should say that we do not need variable x_1 in the model given that variables x_2 and x_3 will remain in the model. In general, the interpretation of a slope in multiple regression can be tricky. Correlations among the predictors can change the slope values dramatically from what they would be in separate simple regressions.

To carry out the test, statistical software will report p -values for all coefficients in the model. Each p -value will be based on a t -statistic calculated as

$$t^* = (\text{sample coefficient} - \text{hypothesized value}) / \text{standard error of coefficient}.$$

For our example above, the t -statistic is:

$$t^* = \frac{b_1 - 0}{\text{se}(b_1)} = \frac{b_1}{\text{se}(b_1)}.$$

Note that the hypothesized value is usually just 0, so this portion of the formula is often omitted.

Multiple linear regression, in contrast to simple linear regression, involves multiple predictors and so testing each variable can quickly become complicated. For example, suppose we apply two separate tests for two predictors, say x_1 and x_2 , and both tests have high p -values. One test suggests x_1 is not needed in a model with all the other predictors included, while the other test suggests x_2 is not needed in a model with all the other predictors included. But, this doesn't necessarily mean that *both* x_1 and x_2 are not needed in a model with all the other predictors included. It may well turn out that we would do better to omit either x_1 or x_2 from the model, but not both. How then do we determine what to do? We'll explore this issue further later in this lesson.

◀ 5.2 - Example on Underground Air Quality
(/stat462/node/130)

up
(/stat462/node/83)

5.4 - A Matrix Formulation of the Multiple
Regression Model ▶ (/stat462/node/132)

STAT 462

Applied Regression Analysis

5.4 - A Matrix Formulation of the Multiple Regression Model

Note: This portion of the lesson is most important for those students who will continue studying statistics after taking Stat 462. We will only rarely use the material within the remainder of this course.

A matrix formulation of the multiple regression model

In the multiple regression setting, because of the potentially large number of predictors, it is more efficient to use matrices to define the regression model and the subsequent analyses. Here, we review basic matrix algebra, as well as learn some of the more important multiple regression formulas in matrix form.

As always, let's start with the simple case first. Consider the following simple linear regression function:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i = 1, \dots, n$$

If we actually let $i = 1, \dots, n$, we see that we obtain n equations:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

Well, that's a pretty inefficient way of writing it all out! As you can see, there is a pattern that emerges. By taking advantage of this pattern, we can instead formulate the above simple linear regression function in matrix notation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$Y = X\beta + \epsilon$$

That is, instead of writing out the n equations, using matrix notation, our simple linear regression function reduces to a short and simple statement:

$$Y = X\beta + \epsilon$$

Now, what does this statement mean? Well, here's the answer:

- X is an $n \times 2$ **matrix**.
- Y is an $n \times 1$ **column vector**, β is a 2×1 column vector, and ϵ is an $n \times 1$ column vector.
- The matrix X and vector β are multiplied together using the techniques of **matrix multiplication**.
- And, the vector $X\beta$ is added to the vector ϵ using the techniques of **matrix addition**.

Now, that might not mean anything to you, if you've never studied matrix algebra — or if you have and you forgot it all! So, let's start with a quick and basic review.

Definition of a matrix

An $r \times c$ **matrix** is a rectangular array of symbols or numbers arranged in r rows and c columns. A matrix is almost always denoted by a single capital letter in boldface type.

Here are three examples of simple matrices. The matrix A is a 2×2 **square matrix** containing numbers:

$$A = \begin{bmatrix} 1 & 2 \\ 6 & 3 \end{bmatrix}$$

The matrix B is a 5×3 matrix containing numbers:

$$B = \begin{bmatrix} 1 & 80 & 3.4 \\ 1 & 92 & 3.1 \\ 1 & 65 & 2.5 \\ 1 & 71 & 2.8 \\ 1 & 40 & 1.9 \end{bmatrix}$$

And, the matrix X is a 6×3 matrix containing a column of 1's and two columns of various x variables:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \\ 1 & x_{51} & x_{52} \\ 1 & x_{61} & x_{62} \end{bmatrix}$$

Definition of a vector and a scalar

A **column vector** is an $r \times 1$ matrix, that is, a matrix with only one column. A vector is almost often denoted by a single lowercase letter in boldface type. The following vector q is a 3×1 column vector containing numbers:

$$q = \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix}$$

A **row vector** is an $1 \times c$ matrix, that is, a matrix with only one row. The vector h is a 1×4 row vector containing numbers:

$$h = [21 \quad 46 \quad 32 \quad 90]$$

A 1×1 "matrix" is called a **scalar**, but it's just an ordinary number, such as 29 or σ^2 .

Matrix multiplication

Recall that $X\beta$ that appears in the regression function:

$$Y = X\beta + \epsilon$$

is an example of matrix multiplication. Now, there are some restrictions — you can't just multiply any two old matrices together. **Two matrices can be multiplied together only if** the number of columns of the first matrix equals the number of rows of the second matrix. Then, when you multiply the two matrices:

- the number of rows of the resulting matrix equals the number of rows of the first matrix, and
- the number of columns of the resulting matrix equals the number of columns of the second matrix.

For example, if A is a 2×3 matrix and B is a 3×5 matrix, then the matrix multiplication AB is possible. The resulting matrix $C = AB$ has 2 rows and 5 columns. That is, C is a 2×5 matrix. Note that the matrix multiplication BA is not possible.

For another example, if X is an $n \times (k+1)$ matrix and β is a $(k+1) \times 1$ column vector, then the matrix multiplication $X\beta$ is possible. The resulting matrix $X\beta$ has n rows and 1 column. That is, $X\beta$ is an $n \times 1$ column vector.

Okay, now that we know when we can multiply two matrices together, how do we do it? Here's the basic rule for multiplying A by B to get $C = AB$:

The entry in the i^{th} row and j^{th} column of C is the **inner product** — that is, element-by-element products added together — of the i^{th} row of A with the j^{th} column of B .

For example:

$$C = AB = \begin{bmatrix} 1 & 9 & 7 \\ 8 & 1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 2 & 1 & 5 \\ 5 & 4 & 7 & 3 \\ 6 & 9 & 6 & 8 \end{bmatrix} = \begin{bmatrix} 90 & 101 & 106 & 88 \\ 41 & 38 & 27 & 59 \end{bmatrix}$$

That is, the entry in the **first row** and **first column** of C , denoted c_{11} , is obtained by:

$$c_{11} = 1(3) + 9(5) + 7(6) = 90$$

And, the entry in the **first row** and **second column** of C , denoted c_{12} , is obtained by:

$$c_{12} = 1(2) + 9(4) + 7(9) = 101$$

And, the entry in the **second row** and **third column** of C , denoted c_{23} , is obtained by:

$$c_{23} = 8(1) + 1(7) + 2(6) = 27$$

You might convince yourself that the remaining five elements of C have been obtained correctly.

Matrix addition

Recall that $X\beta + \epsilon$ that appears in the regression function:

$$Y = X\beta + \epsilon$$

is an example of matrix addition. Again, there are some restrictions — you can't just add any two old matrices together. **Two matrices can be added together only if** they have the same number of rows and columns. Then, to add two matrices, simply add the corresponding elements of the two matrices. That is:

- Add the entry in the first row, first column of the first matrix with the entry in the first row, first column of the second matrix.
- Add the entry in the first row, second column of the first matrix with the entry in the first row, second column of the second matrix.
- And, so on.

For example:

$$C = A + B = \begin{bmatrix} 2 & 4 & -1 \\ 1 & 8 & 7 \\ 3 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 7 & 5 & 2 \\ 9 & -3 & 1 \\ 2 & 1 & 8 \end{bmatrix} = \begin{bmatrix} 9 & 9 & 1 \\ 10 & 5 & 8 \\ 5 & 6 & 14 \end{bmatrix}$$

That is, the entry in the **first row** and **first column** of C , denoted c_{11} , is obtained by:

$$c_{11} = 2 + 7 = 9$$

And, the entry in the **first row** and **second column** of C , denoted c_{12} , is obtained by:

$$c_{12} = 4 + 5 = 9$$

You might convince yourself that the remaining seven elements of C have been obtained correctly.

Least squares estimates in matrix notation

Here's the punchline: the $(k+1) \times 1$ vector containing the estimates of the $(k+1)$ parameters of the regression function can be shown to equal:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X'X)^{-1}X'Y$$

where:

- $(X'X)^{-1}$ is the **inverse** of the $X'X$ matrix, and
- X' is the **transpose** of the X matrix.

As before, that might not mean anything to you, if you've never studied matrix algebra — or if you have and you forgot it all! So, let's go off and review inverses and transposes of matrices.

Definition of the transpose of a matrix

The **transpose** of a matrix A is a matrix, denoted A' or A^T , whose rows are the columns of A and whose columns are the rows of A — all in the same order. For example, the transpose of the 3×2 matrix A :

$$A = \begin{bmatrix} 1 & 5 \\ 4 & 8 \\ 7 & 9 \end{bmatrix}$$

is the 2×3 matrix A' :

$$A' = A^T = \begin{bmatrix} 1 & 4 & 7 \\ 5 & 8 & 9 \end{bmatrix}$$

And, since the X matrix in the simple linear regression setting is:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

the $X'X$ matrix in the simple linear regression setting must be:

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

Definition of the identity matrix

The square $n \times n$ identity matrix, denoted I_n , is a matrix with 1's on the diagonal and 0's elsewhere. For example, the 2×2 identity matrix is:

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The identity matrix plays the same role as the number 1 in ordinary arithmetic:

$$\begin{bmatrix} 9 & 7 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 9 & 7 \\ 4 & 6 \end{bmatrix}$$

That is, when you multiply a matrix by the identity, you get the same matrix back.

Definition of the inverse of a matrix

The **inverse** A^{-1} of a square (!) matrix A is the unique matrix such that:

$$A^{-1}A = I = AA^{-1}$$

That is, the inverse of A is the matrix A^{-1} that you have to multiply A by in order to obtain the identity matrix I . Note that I am not just trying to be cute by including (!) in that first sentence. The inverse only exists for square matrices!

Now, finding inverses is a really messy venture. The good news is that we'll always let computers find the inverses for us. In fact, we won't even know that statistical software is finding inverses behind the scenes!

An example

Ugh! All of these definitions! Let's take a look at an example just to convince ourselves that, yes, indeed the least squares estimates are obtained by the following matrix formula:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = (X'X)^{-1}X'Y$$

Let's consider the data in soapsuds.txt

(/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/soapsuds.txt) , in which the height of suds ($y = suds$) in a standard dishpan was recorded for various amounts of soap ($x = soap$, in grams) (Draper and Smith, 1998, p. 108). Using statistical software to fit the simple linear regression model to these data, we obtain:



Regression Equation

$$suds = -2.68 + 9.500 \text{ soap}$$

Let's see if we can obtain the same answer using the above matrix formula. We previously showed that:

$$X'X = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

We can easily calculate some parts of this formula:

x_i	y_i	$x_i \times y_i$	x_i^2
soap	suds	so*su	soap ²
4.0	33	132.0	16.00
4.5	42	189.0	20.25
5.0	45	225.0	25.00
5.5	51	280.5	30.25
6.0	53	318.0	36.00
6.5	61	396.5	42.25
7.0	62	434.0	49.00
---	---	-----	-----
38.5	347	1975.0	218.75

That is, the 2×2 matrix $X'X$ is:

$$X'X = \begin{bmatrix} 7 & 38.5 \\ 38.5 & 218.75 \end{bmatrix}$$

And, the 2×1 column vector $X'Y$ is:

$$X'Y = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = \begin{bmatrix} 347 \\ 1975 \end{bmatrix}$$

So, we've determined $X'X$ and $X'Y$. Now, all we need to do is to find the inverse $(X'X)^{-1}$. As mentioned before, it is very messy to determine inverses by hand. Letting computer software do the dirty work for us, it can be shown that the inverse of $X'X$ is:

$$(X'X)^{-1} = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix}$$

And so, putting all of our work together, we obtain the least squares estimates:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix} \begin{bmatrix} 347 \\ 1975 \end{bmatrix} = \begin{bmatrix} -2.67 \\ 9.51 \end{bmatrix}$$

That is, the estimated intercept is $b_0 = -2.67$ and the estimated slope is $b_1 = 9.51$. Aha! Our estimates are the same as those reported above (within rounding error)!

Further Matrix Results for Multiple Linear Regression

Matrix notation applies to other regression topics, including fitted values, residuals, sums of squares, and inferences about regression parameters. One important matrix that appears in many formulas is the so-called "hat matrix," $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, since it puts the hat on \mathbf{Y} !

Linear Dependence

There is just one more really critical topic that we should address here, and that is linear dependence. We say that the columns of the matrix \mathbf{A} :

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 & 1 \\ 2 & 1 & 8 & 6 \\ 3 & 6 & 12 & 3 \end{bmatrix}$$

are **linearly dependent**, since (at least) one of the columns can be written as a linear combination of another, namely the third column is $4 \times$ the first column. If none of the columns can be written as a linear combination of the other columns, then we say the columns are **linearly independent**.

Unfortunately, linear dependence is not always obvious. For example, the columns in the following matrix \mathbf{A} :

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 1 \\ 2 & 3 & 1 \\ 3 & 2 & 1 \end{bmatrix}$$

are linearly dependent, because the first column plus the second column equals $5 \times$ the third column.

Now, why should we care about linear dependence? Because the inverse of a square matrix exists only if the columns are linearly independent. Since the vector of regression estimates \mathbf{b} depends on $(\mathbf{X}'\mathbf{X})^{-1}$, the parameter estimates b_0 , b_1 , and so on cannot be uniquely determined if some of the columns of \mathbf{X} are linearly dependent! That is, if the columns of your \mathbf{X} matrix — that is, two or more of your predictor variables — are linearly dependent (or nearly so), you will run into trouble when trying to estimate the regression equation.

For example, suppose for some strange reason we multiplied the predictor variable *soap* by 2 in the dataset *soapsuds.txt*. That is, we'd have two predictor variables, say *soap1* (which is the original *soap*) and *soap2* (which is $2 \times$ the original *soap*):

soap1	soap2	suds
4.0	8	33
4.5	9	42
5.0	10	45
5.5	11	51
6.0	12	53
6.5	13	61
7.0	14	62

If we tried to regress $y = \text{suds}$ on $x_1 = \text{soap1}$ and $x_2 = \text{soap2}$, we see that statistical software spits out trouble:

```
* soap2 is highly correlated with other X variables
* soap2 has been removed from the equation

The regression equation is suds = - 2.68 + 9.50 soap1
```

In short, the first moral of the story is "don't collect your data in such a way that the predictor variables are perfectly correlated." And, the second moral of the story is "if your software package reports an error message concerning high correlation among your predictor variables, then think about linear dependence and how to get rid of it."

◀ 5.3 - The Multiple Linear Regression Model
(/stat462/node/131)

up
(/stat462/node/83)

5.5 - Three Types of MLR Parameter Tests ▶
(/stat462/node/134)

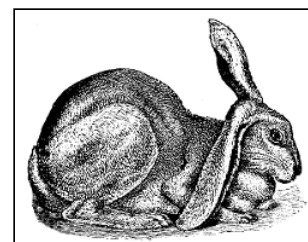
STAT 462

Applied Regression Analysis

5.5 - Three Types of MLR Parameter Tests

Let's investigate an example that highlights the differences between the three hypotheses that we learn how to test in the remainder of this lesson.

An Example: Heart attacks in rabbits. When heart muscle is deprived of oxygen, the tissue dies and leads to a heart attack ("myocardial infarction"). Apparently, cooling the heart reduces the size of the heart attack. It is not known, however, whether cooling is only effective if it takes place *before* the blood flow to the heart becomes restricted. Some researchers (Hale, *et al*, 1997) hypothesized that cooling the heart would be effective in reducing the size of the heart attack even if it takes place *after* the blood flow becomes restricted.



To investigate their hypothesis, the researchers conducted an experiment on 32 anesthetized rabbits that were subjected to a heart attack. The researchers established three experimental groups:

- Rabbits whose hearts were cooled to 6° C within 5 minutes of the blocked artery ("**early cooling**")
- Rabbits whose hearts were cooled to 6° C within 25 minutes of the blocked artery ("**late cooling**")
- Rabbits whose hearts were not cooled at all ("**no cooling**")

At the end of the experiment, the researchers measured the size of the **infarcted** (*i.e.*, damaged) **area** (in grams) in each of the 32 rabbits. But, as you can imagine, there is great variability in the size of hearts. The size of a rabbit's infarcted area may be large only because it has a larger heart. Therefore, in order to adjust for differences in heart sizes, the researchers also measured the size of the **region at risk** for infarction (in grams) in each of the 32 rabbits.

With their measurements in hand (coolhearts.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/coolhearts.txt)), the researchers' primary research question was:

Does the mean size of the infarcted area differ among the three treatment groups — no cooling, early cooling, and late cooling — when controlling for the size of the region at risk for infarction?

A regression model that the researchers might use in answering their research question is:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

where:

- y_i is the size of the infarcted area (in grams) of rabbit i
- x_{i1} is the size of the region at risk (in grams) of rabbit i
- x_{i2} is 1 if early cooling or rabbit i , 0 if not

Loading [MathJax]/extensions/MathZoom.js

- $x_{i3} = 1$ if late cooling of rabbit i , 0 if not

and the independent error terms ϵ_i follow a normal distribution with mean 0 and equal variance σ^2 .

The predictors x_2 and x_3 are "indicator variables" that translate the categorical information on the experimental group to which a rabbit belongs into a usable form. We'll learn more about such variables in Lesson 8, but for now observe that for "early cooling" rabbits $x_2 = 1$ and $x_3 = 0$, for "late cooling" rabbits $x_2 = 0$ and $x_3 = 1$, and for "no cooling" rabbits $x_2 = 0$ and $x_3 = 0$. The model can therefore be simplified for each of the three experimental groups:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2) + \epsilon_i \text{ for "early cooling" rabbits}$$

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_3) + \epsilon_i \text{ for "late cooling" rabbits}$$

$$y_i = (\beta_0 + \beta_1 x_{i1}) + \epsilon_i \text{ for "no cooling" rabbits}$$

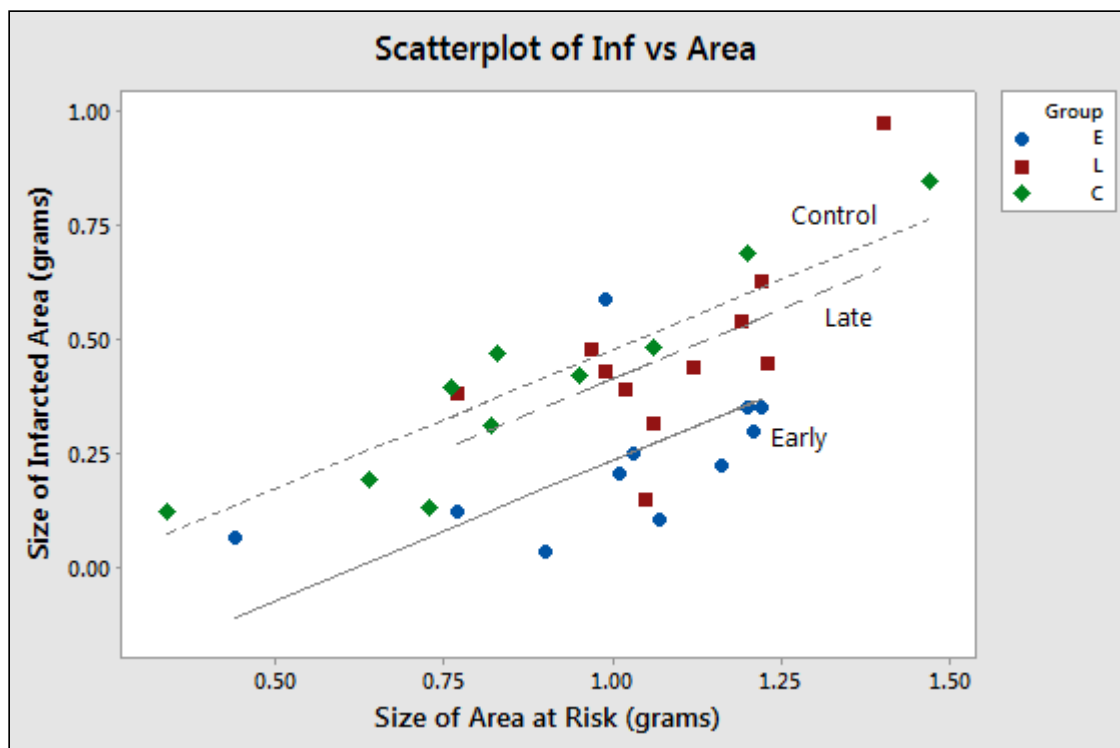
Thus, β_2 represents the difference in mean size of the infarcted area — controlling for the size of the region at risk — between "early cooling" and "no cooling" rabbits. Similarly, β_3 represents the difference in mean size of the infarcted area — controlling for the size of the region at risk — between "late cooling" and "no cooling" rabbits.

Fitting the above model to the researchers' data, statistical software reports:

Regression Equation

$$\text{Inf} = -0.135 + 0.613 \text{ Area} - 0.2435 X_2 - 0.0657 X_3$$

A plot of the data adorned with the estimated regression equation looks like:



The plot suggests that, as we'd expect, as the size of the area at risk increases, the size of the infarcted area also tends to increase. The plot also suggests that for *this sample* of 32 rabbits with a given size of area at risk, 1.0 gram infarcted area differs for the three experimental groups. But, the researchers aren't just interested in this sample. They want to be able to answer their research question for *the whole population* of rabbits.

How could the researchers use the above regression model to answer their research question? Note that the estimated slope coefficients b_2 and b_3 are -0.2435 and -0.0657, respectively. If the estimated coefficients b_2 and b_3 were instead both 0, then the average size of the infarcted area would be the same for the three groups of rabbits *in this sample*. It can be shown that the mean size of the infarcted area would be the same *for the whole population* of rabbits — controlling for the size of the region at risk — if the two slopes β_2 and β_3 simultaneously equal 0. That is, the researchers's question reduces to testing the hypothesis $H_0 : \beta_2 = \beta_3 = 0$.

I'm hoping this example clearly illustrates the need for being able to "translate" a research question into a statistical procedure. Often, the procedure involves four steps, namely:

- formulating a multiple regression model
- determining how the model helps answer the research question
- checking the model
- and performing a hypothesis test (or calculating a confidence interval)

We next learn how to perform three different hypothesis tests for slope parameters in order to answer various research questions. Let's take a look at the different research questions — and the hypotheses we need to test in order to answer the questions — for our heart attacks in rabbits example.

A research question

Consider the research question: "Is a regression model containing at least one predictor useful in predicting the size of the infarct?" Are you convinced that testing the following hypotheses helps answer the research question?

- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- $H_A : \text{At least one } \beta_i \neq 0 \text{ (for } i = 1, 2, 3)$

In this case, the researchers are interested in testing that *all* three slope parameters are zero. We'll soon see that the null hypothesis is tested using the analysis of variance F -test.

Another research question

Consider the research question: "Is the size of the infarct significantly (linearly) related to the area of the region at risk?" Are you convinced that testing the following hypotheses helps answer the research question?

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

In this case, the researchers are interested in testing that *just one* of the three slope parameters is zero. You already know how to do this, don't you? Wouldn't this just involve performing a t -test for β_1 ? We'll soon learn how to think about the t -test for a single slope parameter in the multiple regression framework.

A final research question

Consider the researcher's primary research question: "Is the size of the infarct area significantly (linearly) related to the type of treatment after controlling for the size of the region at risk for infarction?" Are you convinced that testing the following hypotheses helps answer the research question?

- $H_0 : \beta_2 = \beta_3 = 0$
- $H_A : \text{At least one } \beta_i \neq 0 \text{ (for } i = 2, 3)$

Loading [MathJax]/extensions/MathZoom.js

In this case, the researchers are interested in testing whether a *subset* (more than one, but not all) of the slope parameters are simultaneously zero. We will learn a general linear F -test for testing such a hypothesis.

◀ 5.4 - A Matrix Formulation of the Multiple Regression Model (/stat462/node/132)	up (/stat462/node/83)	5.6 - The General Linear F-Test ▶ (/stat462/node/135)
---	---------------------------------------	---

STAT 462

Applied Regression Analysis

5.6 - The General Linear F-Test

The "**general linear F-test**" involves three basic steps, namely:

1. Define a larger **full model**. (By "larger," we mean one with more parameters.)
2. Define a smaller **reduced model**. (By "smaller," we mean one with fewer parameters.)
3. Use an **F-statistic** to decide whether or not to reject the smaller reduced model in favor of the larger full model.

As you can see by the wording of the third step, the null hypothesis always pertains to the reduced model, while the alternative hypothesis always pertains to the full model.

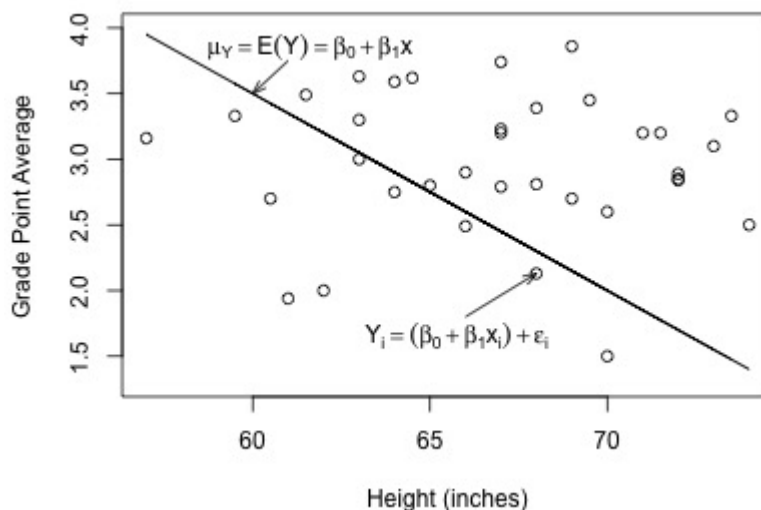
The easiest way to learn about the general linear F-test is to first go back to what we know, namely the simple linear regression model. Once we understand the general linear F-test for the simple case, we then see that it can be easily extended to the multiple case. We take that approach here.

The full model

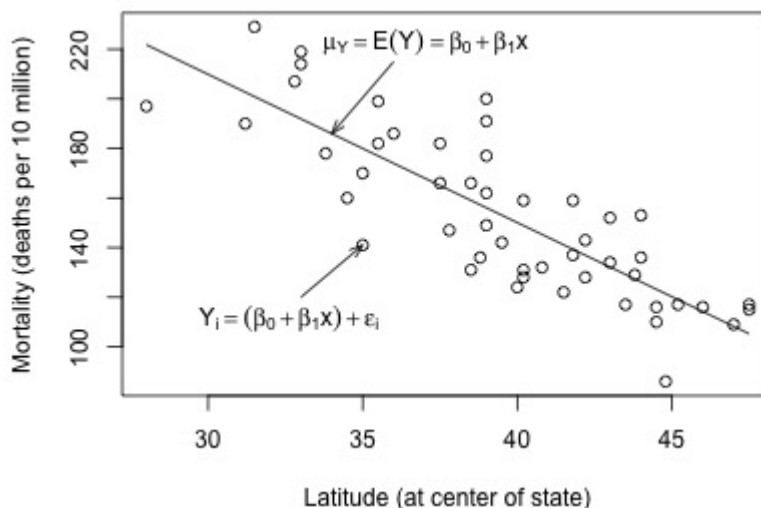
The "**full model**", which is also sometimes referred to as the "**unrestricted model**," is the model thought to be most appropriate for the data. For simple linear regression, the full model is:

$$y_i = (\beta_0 + \beta_1 x_{i1}) + \epsilon_i$$

Here's a plot of a hypothesized full model for a set of data that we worked with previously in this course (student heights and grade point averages):



And, here's another plot of a hypothesized full model that we previously encountered (state latitudes and skin cancer mortalities):



In each plot, the solid line represents what the *hypothesized* population regression line might look like for the full model. The question we have to answer in each case is "does the full model describe the data well?" Here, we might think that the full model does well in summarizing the trend in the second plot but not the first.

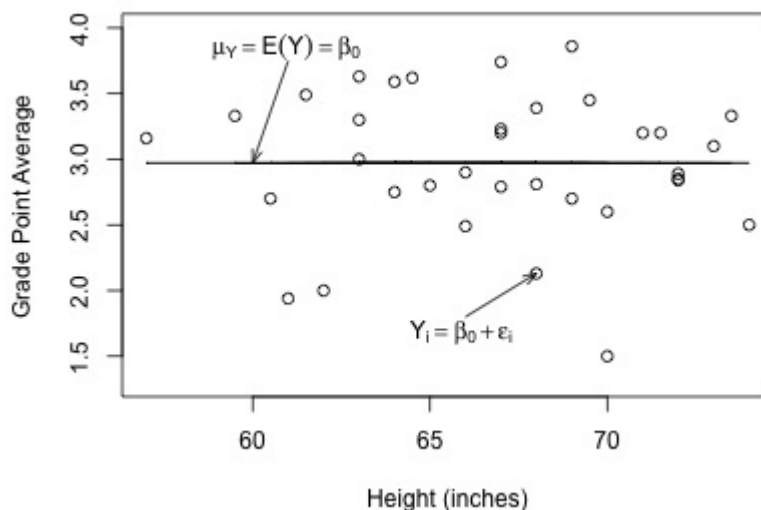
The reduced model

The "**reduced model**," which is sometimes also referred to as the "**restricted model**," is the model described by the null hypothesis H_0 . For simple linear regression, a common null hypothesis is $H_0 : \beta_1 = 0$. In this case, the reduced model is obtained by "zeroing-out" the slope β_1 that appears in the full model. That is, the reduced model is:

$$y_i = \beta_0 + \epsilon_i$$

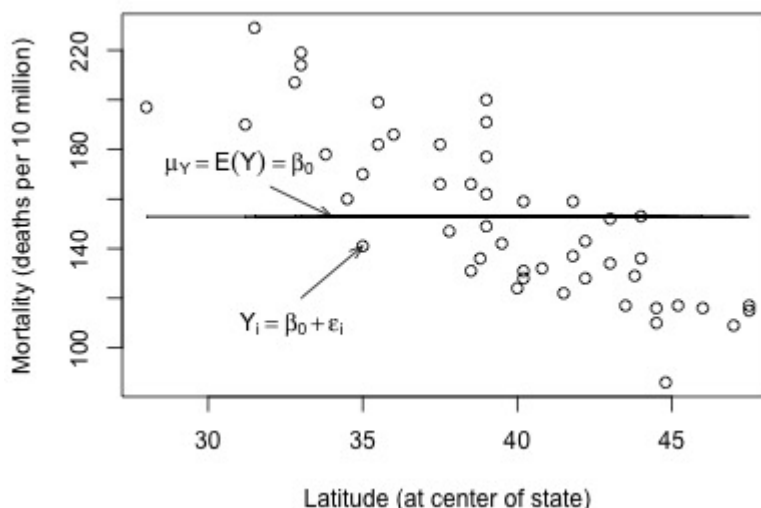
This reduced model suggests that each response y_i is a function only of some overall mean, β_0 , and some error ϵ_i .

Let's take another look at the plot of student grade point average against height, but this time with a line representing what the hypothesized population regression line might look like for the reduced model:



Not bad — there (fortunately?!) doesn't appear to be a relationship between height and grade point average. And, it appears as if the reduced model might be appropriate in describing the lack of a relationship between heights and

grade point averages. How does the reduced model do for the skin cancer mortality example?



It doesn't appear as if the reduced model would do a very good job of summarizing the trend in the population.

The test

How do we decide if the reduced model or the full model does a better job of describing the trend in the data when it can't be determined by simply looking at a plot? What we need to do is to quantify how much error remains after fitting each of the two models to our data. That is, we take the general linear F-test approach:

- **"Fit the full model"** to the data.
 - Obtain the least squares estimates of β_0 and β_1 .
 - Determine the error sum of squares, which we denote " $SSE(F)$."
- **"Fit the reduced model"** to the data.
 - Obtain the least squares estimate of β_0 .
 - Determine the error sum of squares, which we denote " $SSE(R)$."

Recall that, in general, the error sum of squares is obtained by summing the squared distances between the observed and fitted (estimated) responses:

$$\sum (\text{observed} - \text{fitted})^2$$

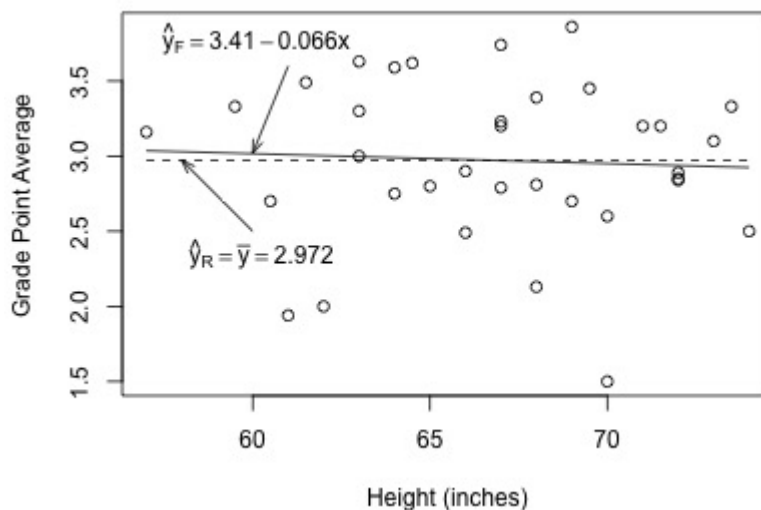
Therefore, since y_i is the observed response and \hat{y}_i is the fitted response for the **full model**:

$$SSE(F) = \sum (y_i - \hat{y}_i)^2$$

And, since y_i is the observed response and \bar{y} is the fitted response for the **reduced model**:

$$SSE(R) = \sum (y_i - \bar{y})^2$$

Let's get a better feel for the general linear F-test approach by applying it to two different two datasets. First, let's look at the heightgpa data ([/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/heightgpa.txt](http://stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/heightgpa.txt)). The following plot of grade point averages against heights contains two estimated regression lines — the solid line is the estimated line for the full model, and the dashed line is the estimated line for the reduced model:



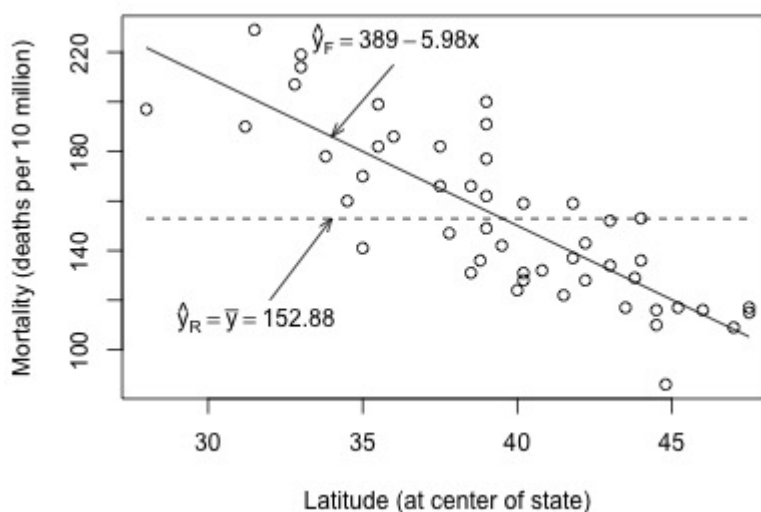
As you can see, the estimated lines are almost identical. Calculating the error sum of squares for each model, we obtain:

$$SSE(F) = \sum (y_i - \hat{y}_i)^2 = 9.7055$$

$$SSE(R) = \sum (y_i - \bar{y})^2 = 9.7331$$

The two quantities are almost identical. Adding height to the reduced model to obtain the full model reduces the amount of error by only 0.0276 (from 9.7331 to 9.7055). That is, adding height to the model does very little in reducing the variability in grade point averages. In this case, there appears to be no advantage in using the larger full model over the simpler reduced model.

Look what happens when we fit the full and reduced models to the skin cancer mortality and latitude dataset (/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/skincancer.txt) :



Here, there is quite a big difference in the estimated equation for the reduced model (solid line) and the estimated equation for the full model (dashed line). The error sums of squares quantify the substantial difference in the two estimated equations:

$$SSE(F) = \sum (y_i - \hat{y}_i)^2 = 17173$$

$$SSE(R) = \sum (y_i - \bar{y})^2 = 53637$$

Adding latitude to the reduced model to obtain the full model reduces the amount of error by 36464 (from 53637 to 17173). That is, adding latitude to the model substantially reduces the variability in skin cancer mortality. In this case, there appears to be a big advantage in using the larger full model over the simpler reduced model.

Where are we going with this general linear F-test approach? In short:

- The general linear F-test involves a comparison between $SSE(R)$ and $SSE(F)$.
- $SSE(R)$ can never be smaller than $SSE(F)$. It is always larger than (or possibly the same as) $SSE(F)$.
 - If $SSE(F)$ is close to $SSE(R)$, then the variation around the estimated full model regression function is almost as large as the variation around the estimated reduced model regression function. If that's the case, it makes sense to use the simpler reduced model.
 - On the other hand, if $SSE(F)$ and $SSE(R)$ differ greatly, then the additional parameter(s) in the full model substantially reduce the variation around the estimated regression function. In this case, it makes sense to go with the larger full model.

How different does $SSE(R)$ have to be from $SSE(F)$ in order to justify using the larger full model? The general linear F -statistic:

$$F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F} \right) \div \left(\frac{SSE(F)}{df_F} \right)$$

helps answer this question. The F -statistic intuitively makes sense — it is a function of $SSE(R) - SSE(F)$, the difference in the error between the two models. The degrees of freedom — denoted df_R and df_F — are those associated with the reduced and full model error sum of squares, respectively.

We use the general linear F -statistic to decide whether or not:

- to reject the null hypothesis H_0 : the reduced model,
- in favor of the alternative hypothesis H_A : the full model.

In general, we reject H_0 if F^* is large — or equivalently if its associated P -value is small.

The test applied to the simple linear regression model

For simple linear regression, it turns out that the general linear F -test is just the same ANOVA F -test that we learned before. As noted earlier for the simple linear regression case, the full model is:

$$y_i = (\beta_0 + \beta_1 x_{i1}) + \epsilon_i$$

and the reduced model is:

$$y_i = \beta_0 + \epsilon_i$$

Therefore, the appropriate null and alternative hypotheses are specified either as:

- $H_0: y_i = \beta_0 + \epsilon_i$
- $H_A: y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

or as:

- $H_0: \beta_1 = 0$
- $H_A: \beta_1 \neq 0$



The degrees of freedom associated with the error sum of squares for the reduced model is $n-1$, and:

$$SSE(R) = \sum (y_i - \bar{y})^2 = SSTO$$

The degrees of freedom associated with the error sum of squares for the full model is $n-2$, and:

$$SSE(F) = \sum (y_i - \hat{y}_i)^2 = SSE$$

Now, we can see how the general linear F -statistic just reduces algebraically to the ANOVA F -test that we know:

$F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F} \right) \div \left(\frac{SSE(F)}{df_F} \right)$	
	
$df_R = n - 1$	$SSE(R) = SSTO$
$df_F = n - 2$	$SSE(F) = SSE$
	
$F^* = \left(\frac{SSTO - SSE}{(n-1) - (n-2)} \right) \div \left(\frac{SSE}{(n-2)} \right) = \frac{MSR}{MSE}$	

That is, the general linear F -statistic reduces to the ANOVA F -statistic:

$$F^* = \frac{MSR}{MSE}$$

For the student height and grade point average example:

$$F^* = \frac{MSR}{MSE} = \frac{0.0276/1}{9.7055/33} = \frac{0.0276}{0.2941} = 0.094$$

For the skin cancer mortality example:

$$F^* = \frac{MSR}{MSE} = \frac{36464/1}{17173/47} = \frac{36464}{365.4} = 99.8$$

The P -value is calculated as usual. The P -value answers the question: "what is the probability that we'd get an F^* statistic as large as we did, if the null hypothesis were true?" The P -value is determined by comparing F^* to an F distribution with 1 numerator degree of freedom and $n-2$ denominator degrees of freedom. For the student height and grade point average example, the P -value is 0.761 (so we fail to reject H_0 and we favor the reduced model), while for the skin cancer mortality example, the P -value is 0.000 (so we reject H_0 and we favor the full model).

An example

Does alcoholism have an effect on muscle strength? Some researchers (Urbano-Marquez, *et al*, 1989) who were interested in answering this question collected the following data (alcoholarm.txt (<https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/alcoholarm.txt>)) on a sample of 50 alcoholic men:

- x = the total lifetime dose of alcohol (kg per kg of body weight) consumed
- y = the strength of the deltoid muscle in the man's non-dominant arm

The full model is the model that would summarize a linear relationship between alcohol consumption and arm strength. The reduced model, on the other hand, is the model that claims there is no relationship between alcohol consumption and arm strength.

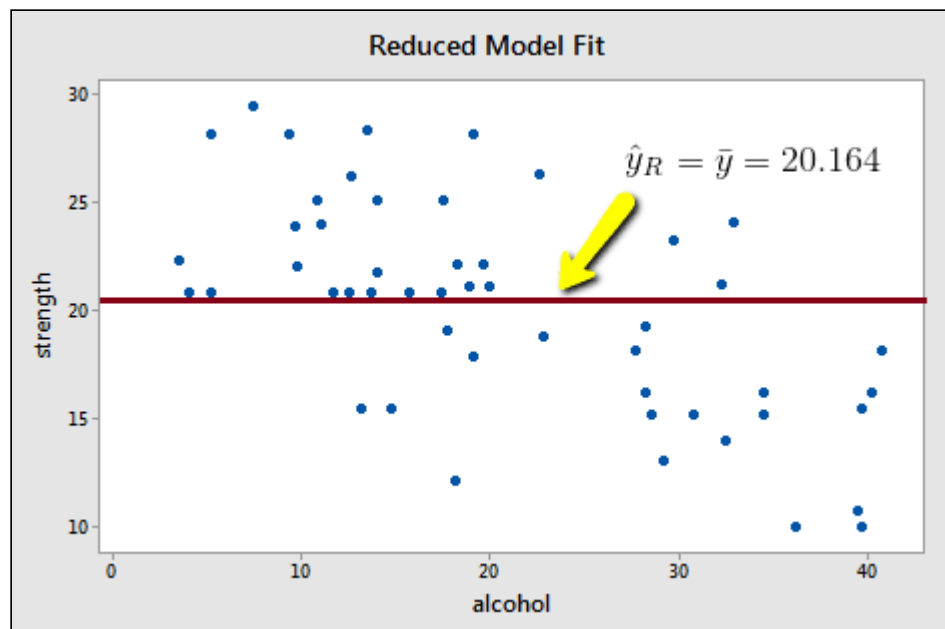
Therefore, the appropriate null and alternative hypotheses are specified either as:

- $H_0: y_i = \beta_0 + \varepsilon_i$
- $H_A: y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

or as:

- $H_0: \beta_1 = 0$
- $H_A: \beta_1 \neq 0$

Upon fitting the reduced model to the data, we obtain:

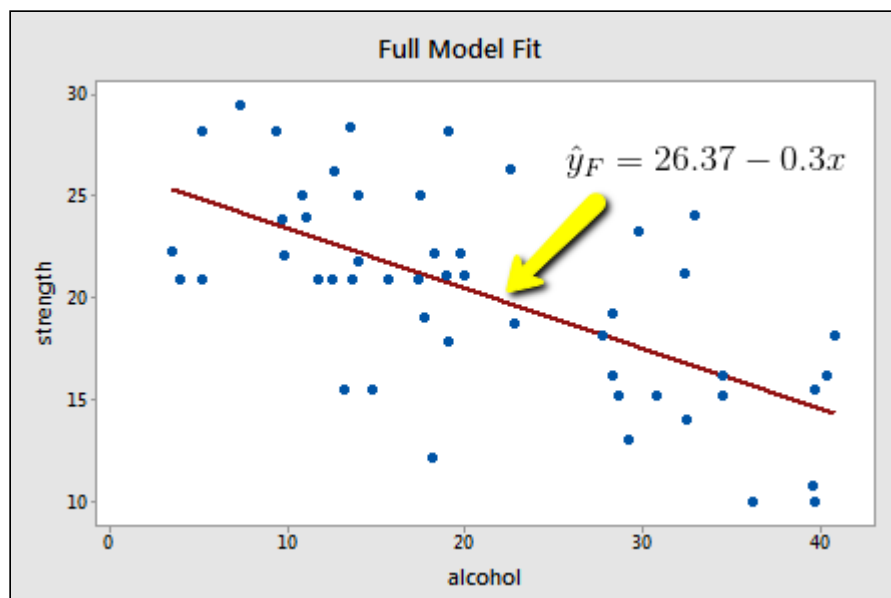


and:

$$SSE(R) = \sum (y_i - \bar{y})^2 = 1224.32$$

Note that the reduced model does not appear to summarize the trend in the data very well.

Upon fitting the full model to the data, we obtain:



and:

$$SSE(F) = \sum (y_i - \hat{y}_i)^2 = 720.27$$

The full model appears to describe the trend in the data better than the reduced model.

The good news is that in the simple linear regression case, we don't have to bother with calculating the general linear F -statistic. Statistical software does it for us in the ANOVA table:

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	504.04	504.040	33.5899	0.000
Error	48	720.27	15.006		
Total	49	1224.32			

Annotations:
 - A box labeled $SSE(R)=SSTO$ points to the Total SS (1224.32).
 - A box labeled $SSE(F)=SSE$ points to the Error SS (720.27).

As you can see, the output reports both $SSE(F)$ — the amount of error associated with the full model — and $SSE(R)$ — the amount of error associated with the reduced model. The F -statistic is:

$$F^* = \frac{MSR}{MSE} = \frac{504.04/1}{720.27/48} = \frac{504.04}{15.006} = 33.59$$

and its associated P -value is < 0.001 (so we reject H_0 and we favor the full model). We can conclude that there is a statistically significant linear association between lifetime alcohol consumption and arm strength.

STAT 462

Applied Regression Analysis

5.7 - MLR Parameter Tests

Earlier in this lesson, we translated three different research questions pertaining to the heart attacks in rabbits study (coolhearts.txt (/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/coolhearts.txt)) into three sets of hypotheses we can test using the general linear F -statistic. The research questions and their corresponding hypotheses are:

1. Is the regression model containing at least one predictor useful in predicting the size of the infarct?

- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- $H_A : \text{At least one } \beta_j \neq 0 \text{ (for } j = 1, 2, 3)$

2. Is the size of the infarct significantly (linearly) related to the area of the region at risk?

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

3. (Primary research question) Is the size of the infarct area significantly (linearly) related to the type of treatment upon controlling for the size of the region at risk for infarction?

- $H_0 : \beta_2 = \beta_3 = 0$
- $H_A : \text{At least one } \beta_j \neq 0 \text{ (for } j = 2, 3)$

Let's test each of the hypotheses now using the general linear F -statistic:

$$F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F} \right) \div \left(\frac{SSE(F)}{df_F} \right)$$

To calculate the F -statistic for each test, we first determine the error sum of squares for the reduced and full models — $SSE(R)$ and $SSE(F)$, respectively. The number of error degrees of freedom associated with the reduced and full models — df_R and df_F , respectively — is the number of observations, n , minus the number of parameters, $k+1$, in the model. That is, in general, the number of error degrees of freedom is $n - (k+1)$. We use statistical software to determine the P -value for each test.

Testing all slope parameters equal 0

To answer the research question: "Is the regression model containing at least one predictor useful in predicting the size of the infarct?," we test the hypotheses:

Loading [MathJax]/extensions/MathZoom.js

- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- $H_A : \text{At least one } \beta_j \neq 0 \text{ (for } j = 1, 2, 3)$

The full model. The full model is the largest possible model — that is, the model containing all of the possible predictors. In this case, the full model is:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

The error sum of squares for the full model, $SSE(F)$, is just the usual error sum of squares, SSE , that appears in the analysis of variance table. Because there are $k+1 = 3+1 = 4$ parameters in the full model, the number of error degrees of freedom associated with the full model is $df_F = n - 4$.

The reduced model. The reduced model is the model that the null hypothesis describes. Because the null hypothesis sets each of the slope parameters in the full model equal to 0, the reduced model is:

$$y_i = \beta_0 + \epsilon_i$$

The reduced model basically suggests that none of the variation in the response y is explained by any of the predictors. Therefore, the error sum of squares for the reduced model, $SSE(R)$, is just the total sum of squares, $SSTO$, that appears in the analysis of variance table. Because there is only one parameter in the reduced model, the number of error degrees of freedom associated with the reduced model is $df_R = n - 1$.

The test. Upon plugging in the above quantities, the general linear F -statistic:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

becomes the usual "**overall F -test**":

$$F^* = \frac{SSR}{3} \div \frac{SSE}{n - 4} = \frac{MSR}{MSE}.$$

That is, to test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, we just use the overall F -test and P -value reported in the analysis of variance table:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	0.95927	0.31976	16.43	0.000
Area	1	0.63742	0.63742	32.75	0.000
X2	1	0.29733	0.29733	15.28	0.001
X3	1	0.01981	0.01981	1.02	0.322
Error	28	0.54491	0.01946		
Total	31	1.50418			

Regression Equation

$$\text{Inf} = -0.135 + 0.613 \text{ Area} - 0.2435 \text{ X2} - 0.0657 \text{ X3}$$

$$F^* = \frac{0.95927}{3} \div \frac{0.54491}{28} = \frac{0.31976}{0.01946} = 16.43.$$

There is sufficient evidence ($F = 16.43$, $P < 0.001$) to conclude that at least one of the slope parameters is not equal to 0.

In general, to test that *all* of the slope parameters in a multiple linear regression model are 0, we use the overall F -test reported in the analysis of variance table.

Testing one slope parameter is 0

Now let's answer the second research question: "Is the size of the infarct significantly (linearly) related to the area of the region at risk?" To do so, we test the hypotheses:

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

The full model. Again, the full model is the model containing all of the possible predictors:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

The error sum of squares for the full model, $SSE(F)$, is just the usual error sum of squares, SSE . Alternatively, because the three predictors in the model are x_1 , x_2 , and x_3 , we can denote the error sum of squares as $SSE(x_1, x_2, x_3)$. Again, because there are 4 parameters in the model, the number of error degrees of freedom associated with the full model is $df_F = n - 4$.

The reduced model. Because the null hypothesis sets the first slope parameter, β_1 , equal to 0, the reduced model is:

$$y_i = (\beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

Because the two predictors in the model are x_2 and x_3 , we denote the error sum of squares as $SSE(x_2, x_3)$. Because there are 3 parameters in the model, the number of error degrees of freedom associated with the reduced model is $df_R = n - 3$.

The test. The general linear statistic:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

simplifies to:

$$F^* = \frac{SSR(x_1|x_2, x_3)}{1} \div \frac{SSE(x_1, x_2, x_3)}{n - 4} = \frac{MSR(x_1|x_2, x_3)}{MSE(x_1, x_2, x_3)}$$

Getting the numbers from the following output:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	0.95927	0.31976	16.43	0.000
X2	1	0.29733	0.29733	15.28	0.001
X3	1	0.01981	0.01981	1.02	0.322
Area	1	0.63742	0.63742	32.75	0.000
Error	28	0.54491	0.01946		
Total	31	1.50418			

Regression Equation

$$\text{Inf} = -0.135 - 0.2435 \text{ X2} - 0.0657 \text{ X3} + 0.613 \text{ Area}$$

Loading [MathJax]/extensions/MathZoom.js

we determine that value of the F -statistic is:

$$F^* = \frac{SSR(x_1|x_2, x_3)}{1} \div MSE = \frac{0.63742}{0.01946} = 32.7554.$$

The P -value is the probability — if the null hypothesis were true — that we would get an F -statistic larger than 32.7554. Comparing our F -statistic to an F -distribution with 1 numerator degree of freedom and 28 denominator degrees of freedom, the probability is close to 1 that we would observe an F -statistic *smaller* than 32.7554:

F distribution with 1 DF in numerator and 28 DF in denominator

```

      x  P( X ≤ x )
32.7554  1.00000

```

Therefore, the probability that we would get an F -statistic *larger* than 32.7554 is close to 0. That is, the P -value is < 0.001 . There is sufficient evidence ($F = 32.8$, $P < 0.001$) to conclude that the size of the infarct is significantly related to the size of the area at risk.

But wait a second! Have you been wondering why we couldn't just use the slope's t -statistic to test that the slope parameter, β_1 , is 0? We can! Notice that the P -value ($P < 0.001$) for the t -test ($t^* = 5.72$):

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.135	0.104	-1.29	0.206	
X2	-0.2435	0.0623	-3.91	0.001	1.44
X3	-0.0657	0.0651	-1.01	0.322	1.57
Area	0.613	0.107	5.72	0.000	1.14

Regression Equation

Inf = -0.135 - 0.2435 X2 - 0.0657 X3 + 0.613 Area

is the same as the P -value we obtained for the F -test. This will be always be the case when we test that *only one* slope parameter is 0. That's because of the well-known relationship between a t -statistic and an F -statistic that has one numerator degree of freedom:

$$t_{(n-(k+1))}^2 = F_{(1, n-(k+1))}$$

For our example, the square of the t -statistic, 5.72, equals our F -statistic (within rounding error). That is:

$$t^{*2} = 5.72^2 = 32.72 = F^*$$

So what have we learned in all of this discussion about the equivalence of the F -test when testing only one slope parameter and the t -test? In short:

- We can use either the F -test or the t -test to test that *only one* slope parameter is 0. Because the t -test results can be read directly from the software output, it makes sense that it would be the test that we'll use most often.
- But, we have to be careful with our interpretations! The equivalence of the t -test to the F -test when testing only one slope parameter has taught us something new about the t -test. The t -test is a test for the *marginal* significance of the x_1 predictor *after* the other predictors x_2 and x_3 have been taken into account. It does *not* test for the significance of the relationship between the response y and the predictor x_1 alone.

Testing a subset of slope parameters is 0

Finally, let's answer the third — and primary — research question: "Is the size of the infarct area significantly (linearly) related to the type of treatment upon controlling for the size of the region at risk for infarction?" To do so, we test the hypotheses:

- $H_0 : \beta_2 = \beta_3 = 0$
- $H_A : \text{At least one } \beta_j \neq 0 \text{ (for } j = 2, 3)$

The full model. Again, the full model is the model containing all of the possible predictors:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

The error sum of squares for the full model, $SSE(F)$, is just the usual error sum of squares, $SSE = 0.54491$ from the output above. Alternatively, because the three predictors in the model are x_1 , x_2 , and x_3 , we can denote the error sum of squares as $SSE(x_1, x_2, x_3)$. Again, because there are 4 parameters in the model, the number of error degrees of freedom associated with the full model is $df_F = n - 4 = 32 - 4 = 28$.

The reduced model. Because the null hypothesis sets the second and third slope parameters, β_2 and β_3 , equal to 0, the reduced model is:

$$y_i = (\beta_0 + \beta_1 x_{i1}) + \epsilon_i$$

The ANOVA table for the reduced model is:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.6249	0.62492	21.32	0.000
Area	1	0.6249	0.62492	21.32	0.000
Error	30	0.8793	0.02931		
Total	31	1.5042			

Because the only predictor in the model is x_1 , we denote the error sum of squares as $SSE(x_1) = 0.8793$. Because there are 2 parameters in the model, the number of error degrees of freedom associated with the reduced model is $df_R = n - 2 = 32 - 2 = 30$.

The test. The general linear statistic is:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} = \frac{0.8793 - 0.54491}{30 - 28} \div \frac{0.54491}{28} = \frac{0.33439}{2} \div 0.01946 = 8.59.$$

The P -value is the probability — if the null hypothesis were true — that we would observe an F -statistic more extreme than 8.59. The following output:

F distribution with 2 DF in numerator and 28 DF in denominator

x	P(X ≤ x)
8.59	0.998767

tells us that the probability of observing such an F -statistic that is *smaller* than 8.59 is 0.9988. Therefore, the

an F -statistic that is *larger* than 8.59 is $1 - 0.9988 = 0.0012$. The P -value is very

small. There is sufficient evidence ($F = 8.59$, $P = 0.0012$) to conclude that the type of cooling is significantly related to the extent of damage that occurs — after taking into account the size of the region at risk.

Summary of MLR Testing

For the simple linear regression model, there is only one slope parameter about which one can perform hypothesis tests. For the multiple linear regression model, there are three different hypothesis tests for slopes that one could conduct. They are:

- Hypothesis test for testing that **all** of the slope parameters are 0.
- Hypothesis test for testing that a **subset** — more than one, but not all — of the slope parameters are 0.
- Hypothesis test for testing that **one** slope parameter is 0.

We have learned how to perform each of the above three hypothesis tests.

The F -statistic and associated p -value in the ANOVA table are used for testing whether **all** of the slope parameters are 0. In most applications this p -value will be small enough to reject the null hypothesis and conclude that at least one predictor is useful in the model. For example, for the rabbit heart attacks study, the F -statistic is $(0.95927/3) / (0.54491/(32-4)) = 16.43$ with p -value 0.000.

To test whether a **subset** — more than one, but not all — of the slope parameters are 0, use the **general linear F-test** formula by fitting the full model to find $SSE(F)$ and fitting the reduced model to find $SSE(R)$. Then the numerator of the F -statistic is $(SSE(R) - SSE(F)) / (df_R - df_F)$. The denominator of the F -statistic is the mean squared error in the ANOVA table. For example, for the rabbit heart attacks study, the **general linear F-statistic** is $[(0.8793 - 0.54491) / (30 - 28)] / (0.54491 / 28) = 8.59$ with p -value 0.0012.

To test whether **one** slope parameter is 0, we can use an F -test as just described. Alternatively, we can use a t -test, which will have an identical p -value since in this case the square of the t -statistic is equal to the F -statistic. For example, for the rabbit heart attacks study, the F -statistic for testing the slope parameter for the *Area* predictor is $(0.63742/1) / (0.54491/(32-4)) = 32.75$ with p -value 0.000. Alternatively, the t -statistic for testing the slope parameter for the *Area* predictor is $0.613 / 0.107 = 5.72$ with p -value 0.000, and $5.72^2 = 32.72$.

Incidentally, you may be wondering why we can't just do a series of individual t -tests to test whether a subset of the slope parameters are 0. For example, for the rabbit heart attacks study, we could have done the following:

- Fit the model of $y = \text{InfSize}$ on $x_1 = \text{Area}$ and x_2 and x_3 and use an individual t -test for x_3 .
- If the test results indicate that we can drop x_3 then fit the model of $y = \text{InfSize}$ on $x_1 = \text{Area}$ and x_2 and use an individual t -test for x_2 .

The problem with this approach is we're using two individual t -tests instead of one F -test, which means our chance of drawing an incorrect conclusion in our testing procedure is higher. Every time we do a hypothesis test, we can draw an incorrect conclusion by:

- rejecting a true null hypothesis, i.e., make a type 1 error by concluding the tested predictor(s) should be retained in the model, when in truth it/they should be dropped; or
- failing to reject a false null hypothesis, i.e., make a type 2 error by concluding the tested predictor(s) should be dropped from the model, when in truth it/they should be retained.

Thus, in general, the fewer tests we perform the better. In this case, this means that wherever possible using one F -test in place of multiple individual t -tests is preferable.

[◀ 5.6 - The General Linear F-Test \(/stat462/node/135\)](#)

[up \(/stat462/node/83\)](#)

[5.8 - Partial R-squared ▶ \(/stat462/node/138\)](#)

STAT 462

Applied Regression Analysis

5.9 - Further MLR Examples

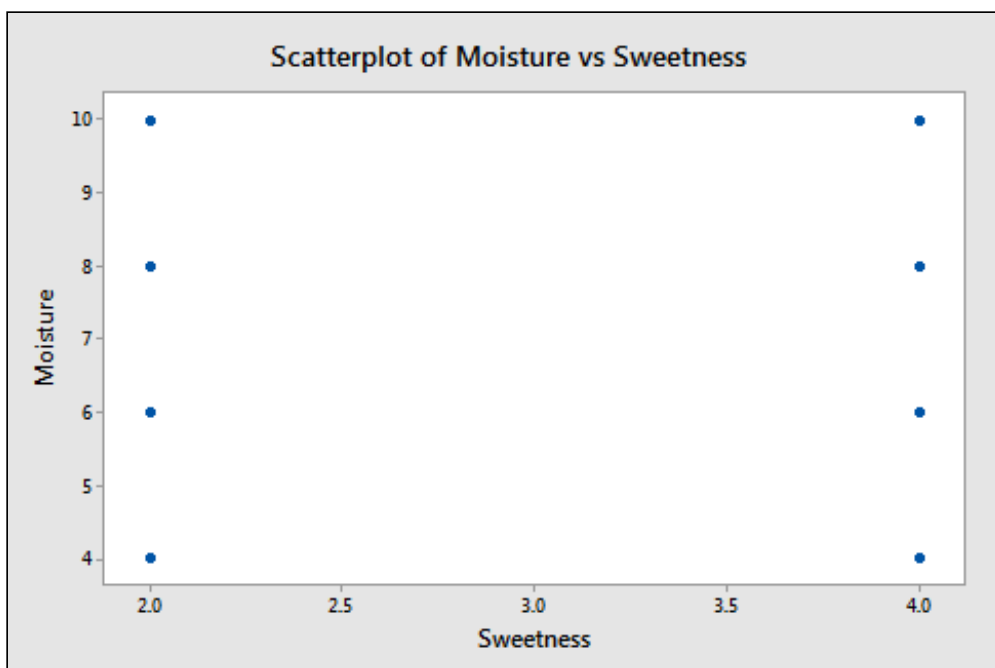
Example 1: Pastry Sweetness Data

A designed experiment is done to assess how moisture content and sweetness of a pastry product affect a taster's rating of the product (pastry.txt (/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/pastry.txt)). In a designed experiment, the eight possible combinations of four moisture levels and two sweetness levels are studied. Two pastries are prepared and rated for each of the eight combinations, so the total sample size is $n = 16$. The y -variable is the rating of the pastry. The two x -variables are moisture and sweetness. The values (and sample sizes) of the x -variables were designed so that the x -variables were not correlated.

**Correlation: Moisture, Sweetness**

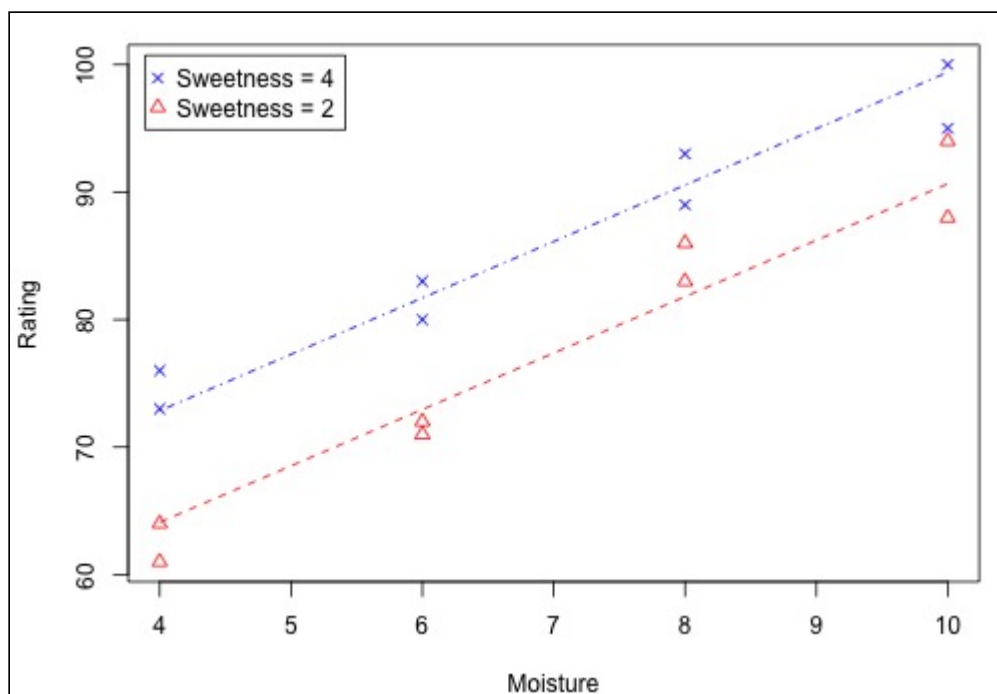
Pearson correlation of Moisture and Sweetness = 0.000
P-Value = 1.000

A plot of moisture versus sweetness (the two x -variables) is as follows:



Notice that the points are on a rectangular grid so the correlation between the two variables is 0. (Please Note: we are not able to see that actually there are 2 observations at each location of the grid!)

The following figure shows how the two x-variables affect the pastry rating.



There is a linear relationship between rating and moisture and there is also a sweetness difference. The results given in the following output are for three different regressions - separate simple regressions for each x-variable and a multiple regression that incorporates both x-variables.

Regression Analysis: Rating versus Moisture						Regression Analysis: Rating versus Sweetness					
Analysis of Variance						Analysis of Variance					
Source	DF	Adj SS	Adj MS	F-Value	P-Value	Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1566.45	1566.45	54.75	0.000	Regression	1	306.3	306.3	2.58	0.130
Moisture	1	1566.45	1566.45	54.75	0.000	Sweetness	1	306.3	306.3	2.58	0.130
Error	14	400.55	28.61			Error	14	1660.8	118.6		
Lack-of-Fit	2	15.05	7.52	0.23	0.795	Total	15	1967.0			
Pure Error	12	385.50	32.13								
Total	15	1967.00									
Model Summary						Model Summary					
S	R-sq	R-sq(adj)	R-sq(pred)			S	R-sq	R-sq(adj)	R-sq(pred)		
5.34890	79.64%	78.18%	72.71%			10.8915	15.57%	9.54%	0.00%		
Coefficients						Coefficients					
Term	Coef	SE Coef	T-Value	P-Value	VIF	Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	50.77	4.39	11.55	0.000		Constant	68.63	8.61	7.97	0.000	
Moisture	4.425	0.598	7.40	0.000	1.00	Sweetness	4.38	2.72	1.61	0.130	1.00
Regression Equation						Regression Equation					
Rating = 50.77 + 4.425 Moisture						Rating = 68.63 + 4.38 Sweetness					

Regression Analysis: Rating versus Moisture, Sweetness

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1872.70	936.35	129.08	0.000
Moisture	1	1566.45	1566.45	215.95	0.000
Sweetness	1	306.25	306.25	42.22	0.000
Error	13	94.30	7.25		
Lack-of-Fit	5	37.30	7.46	1.05	0.453
Pure Error	8	57.00	7.13		
Total	15	1967.00			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.69330	95.21%	94.47%	92.46%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	37.65	3.00	12.57	0.000	
Moisture	4.425	0.301	14.70	0.000	1.00
Sweetness	4.375	0.673	6.50	0.000	1.00

Regression Equation

Rating = 37.65 + 4.425 Moisture + 4.375 Sweetness

There are three important features to notice in the results:

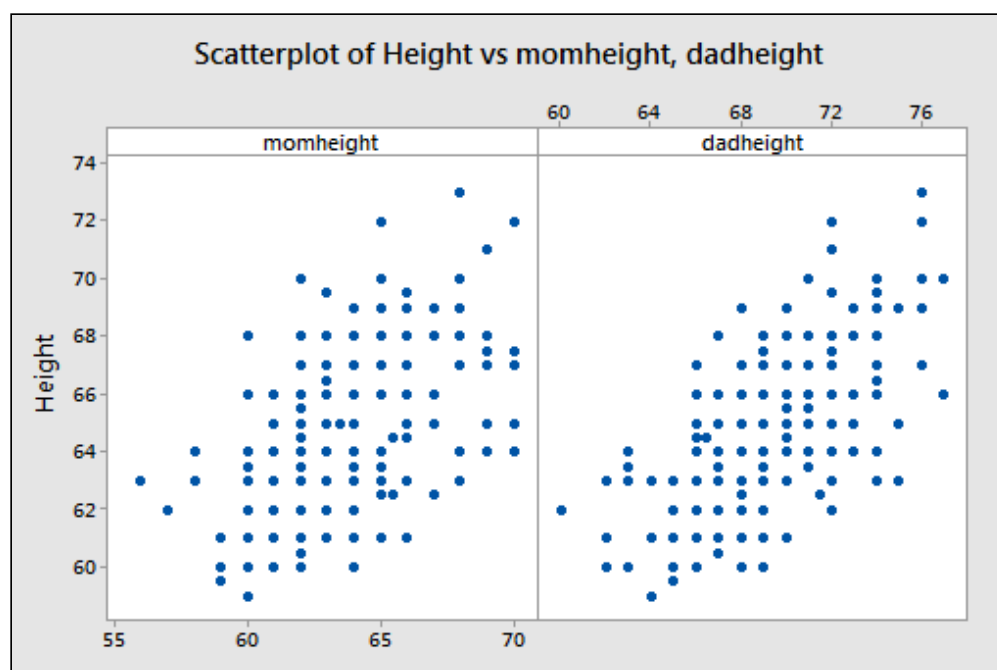
1. The sample coefficient that multiplies **Moisture** is 4.425 in both the simple and the multiple regression. The sample coefficient that multiplies **Sweetness** is 4.375 in both the simple and the multiple regression. **This result does not generally occur**; the only reason that it does in this case is that **Moisture** and **Sweetness** are not correlated, so the estimated slopes are independent of each other. For most observational studies, predictors are typically correlated and estimated slopes in a multiple linear regression model **do not match** the corresponding slope estimates in simple linear regression models.
2. The R^2 for the multiple regression, 95.21%, is the sum of the R^2 values for the simple regressions (79.64% and 15.57%). Again, **this will only happen when we have uncorrelated x-variables**.
3. The variable **Sweetness** is not statistically significant in the simple regression ($p = 0.130$), but it is in the multiple regression. This is a benefit of doing a multiple regression. By putting both variables into the equation, we have greatly reduced the standard deviation of the residuals (notice the S values). This in turn reduces the standard errors of the coefficients, leading to greater (absolute) t -values and smaller p -values.

(Data source: *Applied Regression Models*, (4th edition), Kutner, Neter, and Nachtsheim).

Example 2: Female Stat Students

The data are from $n = 214$ females in statistics classes at the University of California at Davis (stat_females.txt (/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/stat_females.txt)). The variables are y = student's self-reported height, x_1 = student's guess at her mother's height, and x_2 = student's guess at her father's height. All heights are in

inches. The scatterplots below are of each student's height versus mother's height and student's height against father's height.



Both show a moderate positive association with a straight-line pattern and no notable outliers.

Regression Analysis: Height versus momheight, dadheight

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	666.1	333.074	80.73	0.000
momheight	1	128.1	128.117	31.05	0.000
dadheight	1	278.5	278.488	67.50	0.000
Error	211	870.5	4.126		
Lack-of-Fit	101	446.3	4.419	1.15	0.242
Pure Error	110	424.2	3.857		
Total	213	1536.6			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.03115	43.35%	42.81%	41.58%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	18.55	3.69	5.02	0.000	
momheight	0.3035	0.0545	5.57	0.000	1.19
dadheight	0.3879	0.0472	8.22	0.000	1.19

Regression Equation

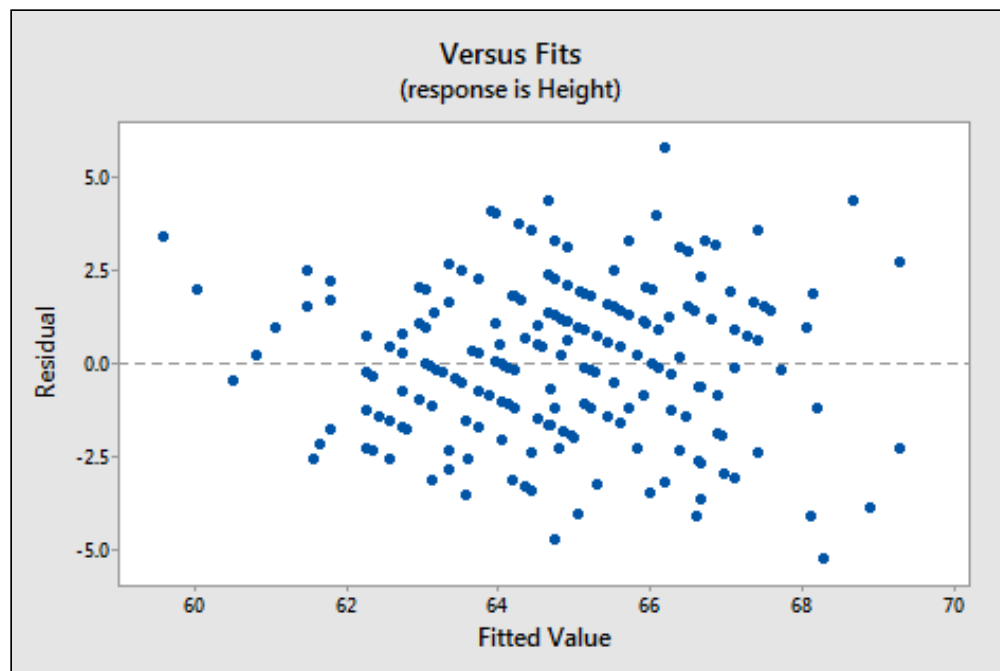
$$\text{Height} = 18.55 + 0.3035 \text{ momheight} + 0.3879 \text{ dadheight}$$

Interpretations

- The sample multiple regression equation is predicted student height = $18.55 + 0.3035 \times \text{mother's height} + 0.3879 \times \text{father's height}$. To use this equation for prediction, we substitute specified values for the two parents' heights.
- We can interpret the “slopes” in the same way that we do for a simple linear regression model but we have to add the constraint that values of other variables remain constant. For example:
 - When father's height is held constant, the average student height increases 0.3035 inches for each one-inch increase in mother's height.
 - When mother's height is held constant, the average student height increases 0.3879 inches for each one-inch increase in father's height.
- The p -values given for the two x -variables tell us that student height is significantly related to each.
- The value of $R^2 = 43.35\%$ means that the model (the two x -variables) explains 43.35% of the observed variation in student heights.
- The value $S = 2.03115$ is the estimated standard deviation of the regression errors. Roughly, it is the average absolute size of a residual.

Residual Plots

Just as in simple regression, we can use a plot of residuals versus fits to evaluate the validity of assumptions. The residual plot for these data is shown in the following figure:



It looks about as it should - a random horizontal band of points. Other residual analyses can be done exactly as we did for simple regression. For instance, we might wish to examine a normal probability plot of the residuals. Additional plots to consider are plots of residuals versus each x -variable separately. This might help us identify sources of curvature or non-constant variance.

Example 3: Hospital Data

Data from $n = 113$ hospitals in the United States are used to assess factors related to the likelihood that a hospital patients acquires an infection while hospitalized. The variables here are y = infection risk, x_1 = average length of patient stay, x_2 = average patient age, x_3 = measure of how many x-rays are given in the hospital (infectionrisk.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/infectionrisk.txt)). Statistical software output is as follows:



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.00	1.31	0.76	0.448	
Stay	0.3082	0.0594	5.19	0.000	1.23
Age	-0.0230	0.0235	-0.98	0.330	1.05
Xray	0.01966	0.00576	3.41	0.001	1.18

Regression Equation

$$\text{InfctRsk} = 1.00 + 0.3082 \text{ Stay} - 0.0230 \text{ Age} + 0.01966 \text{ Xray}$$

Interpretations for this example include:

- The p -value for testing the coefficient that multiplies **Age** is 0.330. Thus we cannot reject the null hypothesis $H_0: \beta_2 = 0$. The variable **Age** is not a useful predictor within this model that includes **Stay** and **Xrays**.
- For the variables **Stay** and **X-rays**, the p -values for testing their coefficients are at a statistically significant level so both are useful predictors of infection risk (within the context of this model!).
- We usually don't worry about the p -value for Constant. It has to do with the "intercept" of the model and seldom has any practical meaning unless it makes sense for all the x -variables to be zero simultaneously.

(Data source: *Applied Regression Models*, (4th edition), Kutner, Neter, and Nachtsheim).

Example 4: Physiological Measurements Data


For a sample of $n = 20$ individuals, we have measurements of y = body fat, x_1 = triceps skinfold thickness, x_2 = thigh circumference, and x_3 = midarm circumference (bodyfat.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/bodyfat.txt)). Statistical software results for the sample coefficients, MSE (highlighted), and $(\mathbf{X}^T \mathbf{X})^{-1}$ are given below:



Regression Analysis: Bodyfat versus Triceps, Thigh, Midarm

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	396.985	132.328	21.52	0.000
Triceps	1	12.705	12.705	2.07	0.170
Thigh	1	7.529	7.529	1.22	0.285
Midarm	1	11.546	11.546	1.88	0.190
Error	16	98.405	6.150		
Total	19	495.390			



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.47998	80.14%	76.41%	67.55%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	117.1	99.8	1.17	0.258	
Triceps	4.33	3.02	1.44	0.170	708.84
Thigh	-2.86	2.58	-1.11	0.285	564.34
Midarm	-2.19	1.60	-1.37	0.190	104.61

Regression Equation

Bodyfat = 117.1 + 4.33 Triceps - 2.86 Thigh - 2.19 Midarm

$(\mathbf{X}^T\mathbf{X})^{-1}$ - (calculated manually, see note below)

1618.87	48.8103	-41.8487	-25.7988
48.81	1.4785	-1.2648	-0.7785
-41.85	-1.2648	1.0840	0.6658
-25.80	-0.7785	0.6658	0.4139

Note: There is no real need to know how to calculate this matrix using statistical software, but in case you're curious first store the design matrix, \mathbf{X} from the regression model. Then find the transpose of \mathbf{X} and multiply the transpose of \mathbf{X} and \mathbf{X} . Finally, invert the resulting matrix.

The variance-covariance matrix of the sample coefficients is found by multiplying each element in $(\mathbf{X}^T\mathbf{X})^{-1}$ by MSE. Common notation for the resulting matrix is either $\mathbf{s}^2(\mathbf{b})$ or $\mathbf{se}^2(\mathbf{b})$. Thus, the standard errors of the coefficients given in the output above can be calculated as follows:

- $\text{Var}(b_0) = (6.15031)(1618.87) = 9956.55$, so $\text{se}(b_0) = \sqrt{9956.55} = 99.782$.
- $\text{Var}(b_1) = (6.15031)(1.4785) = 9.0932$, so $\text{se}(b_1) = \sqrt{9.0932} = 3.016$.
- $\text{Var}(b_2) = (6.15031)(1.0840) = 6.6669$, so $\text{se}(b_2) = \sqrt{6.6669} = 2.582$.
- $\text{Var}(b_3) = (6.15031)(0.4139) = 2.54561$, so $\text{se}(b_3) = \sqrt{2.54561} = 1.595$.

As an example of a covariance and correlation between two coefficients, we consider b_1 and b_2 .

- $\text{Cov}(b_1, b_2) = (6.15031)(-1.2648) = -7.7789$. The value -1.2648 is in the second row and third column of $(\mathbf{X}^T \mathbf{X})^{-1}$. (Keep in mind that the first row and first column give information about b_0 , so the second row has information about b_1 , and so on.)
- $\text{Corr}(b_1, b_2) = \text{covariance divided by product of standard errors} = -7.7789 / (3.016 \times 2.582) = -0.999$.

The extremely high correlation between these two sample coefficient estimates results from a high correlation between the Triceps and Thigh variables. The consequence is that it is difficult to separate the individual effects of these two variables.

If all x -variables are uncorrelated with each other, then all covariances between pairs of sample coefficients that multiply x -variables will equal 0. This means that the estimate of one beta is not affected by the presence of the other x -variables. Many experiments are designed to achieve this property. With observational data, however, we'll most likely not have this happen.

(Data source: *Applied Regression Models*, (4th edition), Kutner, Neter, and Nachtsheim).

Example 5: Peruvian Blood Pressure Data

This dataset consists of variables possibly relating to blood pressures of $n = 39$ Peruvians who have moved from rural high altitude areas to urban lower altitude areas (peru.txt
(<https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/peru.txt>)
) . The variables in this dataset are:



- Y = systolic blood pressure
- X_1 = age
- X_2 = years in urban area
- $X_3 = X_2 / X_1$ = fraction of life in urban area
- X_4 = weight (kg)
- X_5 = height (mm)
- X_6 = chin skinfold
- X_7 = forearm skinfold
- X_8 = calf skinfold
- X_9 = resting pulse rate

First, we run a multiple regression using all nine x -variables as predictors. The results are given below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	4358.85	484.32	6.46	0.000
Error	29	2172.58	74.92		
Total	38	6531.44			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
8.65544	66.74%	56.41%	34.45%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	146.8	49.0	3.00	0.006	
Age	-1.121	0.327	-3.43	0.002	3.21
Years	2.455	0.815	3.01	0.005	34.29
FracLife	-115.3	30.2	-3.82	0.001	24.39
Weight	1.414	0.431	3.28	0.003	4.75
Height	-0.0346	0.0369	-0.94	0.355	1.91
Chin	-0.944	0.741	-1.27	0.213	2.06
Forearm	-1.17	1.19	-0.98	0.335	3.80
Calf	-0.159	0.537	-0.30	0.770	2.41
Pulse	0.115	0.170	0.67	0.507	1.33

When looking at tests for individual variables, we see that p -values for the variables **Height**, **Chin**, **Forearm**, **Calf**, and **Pulse** are not at a statistically significant level. These individual tests are affected by correlations amongst the x -variables, so we will use the **general linear F-test** to see whether it is reasonable to declare that all five non-significant variables can be dropped from the model.

In other words, consider testing:

$$H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

$$H_A : \text{at least one of } \{\beta_5, \beta_6, \beta_7, \beta_8, \beta_9\} \neq 0$$

within the nine variable model given above. If this null is not rejected, it is reasonable to say that none of the five variables **Height**, **Chin**, **Forearm**, **Calf** and **Pulse** contribute to the prediction/explanation of systolic blood pressure.

The *full model* includes all nine variables; $SSE(\text{full}) = 2172.58$, the full error $df = 29$, and $MSE(\text{full}) = 74.92$ (we get these from the results above). The *reduced model* includes only the variables **Age**, **Years**, **fracLife**, and **Weight** (which are the remaining variables if the five possibly non-significant variables are dropped). Regression results for the reduced model are given below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	3901.7	975.43	12.61	0.000
Error	34	2629.7	77.34		
Total	38	6531.4			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
8.79456	59.74%	55.00%	44.84%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	116.8	22.0	5.32	0.000	
Age	-0.951	0.316	-3.00	0.005	2.91
Years	2.339	0.771	3.03	0.005	29.79
FracLife	-108.1	28.3	-3.81	0.001	20.83
Weight	0.832	0.275	3.02	0.005	1.88

We see that $SSE(\text{reduced}) = 2629.7$, and the reduced error $df = 34$. We also see that all four individual x -variables are statistically significant.

The calculation for the general linear F -test statistic is:

$$F = \frac{\frac{SSE(\text{reduced}) - SSE(\text{full})}{\text{error df for reduced} - \text{error df for full}}}{MSE(\text{full})} = \frac{\frac{2629.7 - 2172.58}{34 - 29}}{74.92} = 1.220$$

Thus, this test statistic comes from an $F_{5,29}$ distribution, of which the associated p -value is 0.325 (this can be found by using statistical software or looking up a table). This is not at a statistically significant level, so we do not reject the null hypothesis and we favour the reduced model. Thus it is feasible to drop the variables X_5 , X_6 , X_7 , X_8 , and X_9 from the model.

Example 6: Measurements of College Students

For $n = 55$ college students, we have measurements (Physical.txt

(<https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/Physical.txt>)) for the following five variables:

Y = height (in)

X_1 = left forearm length (cm)

X_2 = left foot length (cm)

X_3 = head circumference (cm)

X_4 = nose length (cm)

Statistical software output for the full model is given below.

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	18.50	7.83	(2.78, 34.23)	2.36	0.022	
LeftArm	0.802	0.171	(0.459, 1.145)	4.70	0.000	1.63
LeftFoot	0.997	0.162	(0.671, 1.323)	6.14	0.000	1.86
HeadCirc	0.081	0.150	(-0.220, 0.381)	0.54	0.593	1.28
nose	-0.147	0.492	(-1.136, 0.841)	-0.30	0.766	1.14

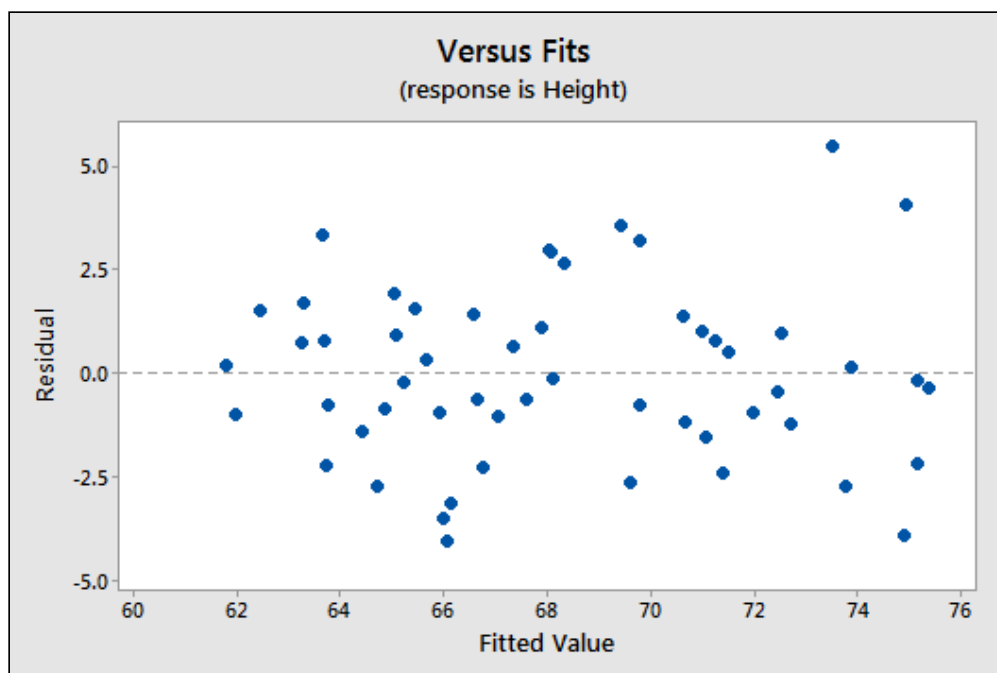
Regression Equation

$$\text{Height} = 18.50 + 0.802 \text{ LeftArm} + 0.997 \text{ LeftFoot} + 0.081 \text{ HeadCirc} - 0.147 \text{ nose}$$

The interpretations of the t -tests are as follows:

- The sample coefficients for **LeftArm** and **LeftFoot** achieve statistical significance. This indicates that they are useful as predictors of **Height**.
- The sample coefficients for **HeadCirc** and **nose** are not significant. Each t -test considers the question of whether the variable is needed, given that all other variables will remain in the model.

Below is a plot of residuals versus the fitted values and it seems suitable.



There is no obvious curvature and the variance is reasonably constant. One may note two possible outliers, but nothing serious.

The first calculation we will perform is for the general linear F -test. The results above might lead us to test

$$H_0 : \beta_3 = \beta_4 = 0$$

$$H_A : \text{at least one of } \{\beta_3, \beta_4\} \neq 0$$

in the full model. If we fail to reject the null hypothesis, we could then remove both of **HeadCirc** and **nose** as predictors.

Below is the ANOVA table for the full model.

Analysis of Variance

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	4	816.39	204.098	42.81	0.000
LeftArm	1	590.21	590.214	123.81	0.000
LeftFoot	1	224.35	224.349	47.06	0.000
HeadCirc	1	1.40	1.402	0.29	0.590
nose	1	0.43	0.427	0.09	0.766
Error	50	238.35	4.767		
Total	54	1054.75			

From this output, we see that $SSE(\text{full}) = 238.35$, with $df = 50$, and $MSE(\text{full}) = 4.77$. The reduced model includes only the two variables **LeftArm** and **LeftFoot** as predictors. The ANOVA results for the reduced model are found below.

Analysis of Variance

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	2	814.56	407.281	88.18	0.000
LeftArm	1	590.21	590.214	127.78	0.000
LeftFoot	1	224.35	224.349	48.57	0.000
Error	52	240.18	4.619		
Lack-of-Fit	44	175.14	3.980	0.49	0.937
Pure Error	8	65.04	8.130		
Total	54	1054.75			

From this output, we see that $SSE(\text{reduced}) = SSE(X_1, X_2) = 240.18$, with $df = 52$, and $MSE(\text{reduced}) = MSE(X_1, X_2) = 4.62$.

With these values obtained, we can now obtain the test statistic for testing $H_0 : \beta_3 = \beta_4 = 0$:

$$F = \frac{\frac{SSE(X_1, X_2) - SSE(\text{full})}{\text{error df for reduced} - \text{error df for full}}}{MSE(\text{full})} = \frac{\frac{240.18 - 238.35}{52 - 50}}{4.77} = 0.192$$

This value comes from an $F_{2,50}$ distribution. By using statistical software or looking up a table we find that the area to the left of $F = 0.192$ (with df of 2 and 50) is 0.174. The p -value is the area to the right of F , so $p = 1 - 0.174 = 0.826$. Thus, we do not reject the null hypothesis, we favour the reduced model, and it is reasonable to remove **HeadCirc** and **nose** from the model.

Next we consider what fraction of variation in $Y = \text{Height}$ cannot be explained by $X_2 = \text{LeftFoot}$, but can be explained by $X_1 = \text{LeftArm}$? To answer this question, we calculate the partial R^2 . The formula is:

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

The denominator, $SSE(X_2)$, measures the unexplained variation in Y when X_2 is the predictor. The ANOVA table for this regression is found in below.

Analysis of Variance

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	1	707.4	707.420	107.95	0.000
LeftFoot	1	707.4	707.420	107.95	0.000
Error	53	347.3	6.553		
Lack-of-Fit	19	113.0	5.948	0.86	0.625
Pure Error	34	234.3	6.892		
Total	54	1054.7			

These results give us $SSE(X_2) = 347.3$.

The numerator, $SSE(X_2) - SSE(X_1, X_2)$, measures the further reduction in the SSE when X_1 is added to the model. Results from the earlier output give us $SSE(X_1, X_2) = 240.18$ and now we can calculate:

$$R^2_{Y,1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{347.3 - 240.18}{347.3} = 0.308$$

Thus $X_1 = \text{LeftArm}$ explains 30.8% of the variation in $Y = \text{Height}$ that could not be explained by $X_2 = \text{LeftFoot}$.

< 5.8 - Partial R-squared (/stat462/node/138)

up
(/stat462/node/83)
