Linear Regression with More than One Predictor

Chapter 4

Multiple regression means that the number of predictor variables exceeds one. The main difference between simple linear regression (one predictor) and multiple regression lies in the interpretation of the coefficients. In multiple regression, the coefficient of a predictor assesses the effect on *Y* of a unit increase in the predictor *when all other predictors remain the same*.[1] We say that the other predictors are *controlled for*. In simple linear regression, there is only one predictor, so there are no other predictors to control for – nothing is controlled for. So the coefficient in simple linear regression represents the *total effect* of the predictor on *Y*. Examples will make clear the distinction.

The multiple linear regression model with *k* - 1 predictor variables.

Suppose that $(x_{12}, x_{13},..., x_{1,k})$, $(x_{22}, x_{23},..., x_{2,k})$, … , $(x_{n2}, x_{n3},..., x_{n,k})$ are given constants. Let $Y_1, Y_2,..., Y_n$ be random variables such that for all $i = 1, \ldots, n$

(1) $E(Y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{i,k}$

(2) $Var(Y_i) = \sigma^2$

(3) $Y_1, Y_2,..., Y_n$ are independent

(4) Y_i is normally distributed.

If all of these assumptions hold, then the multiple linear regression model is said to be valid.

Note on notation: To keep the notation consistent with Ramanathan, the intercept term is labeled β_1, instead of α. So the first *x*-variable has coefficient β_2 and is labeled X_2, rather than X_1.

As in the simple linear regression model, the four assumptions are model **specifications**:

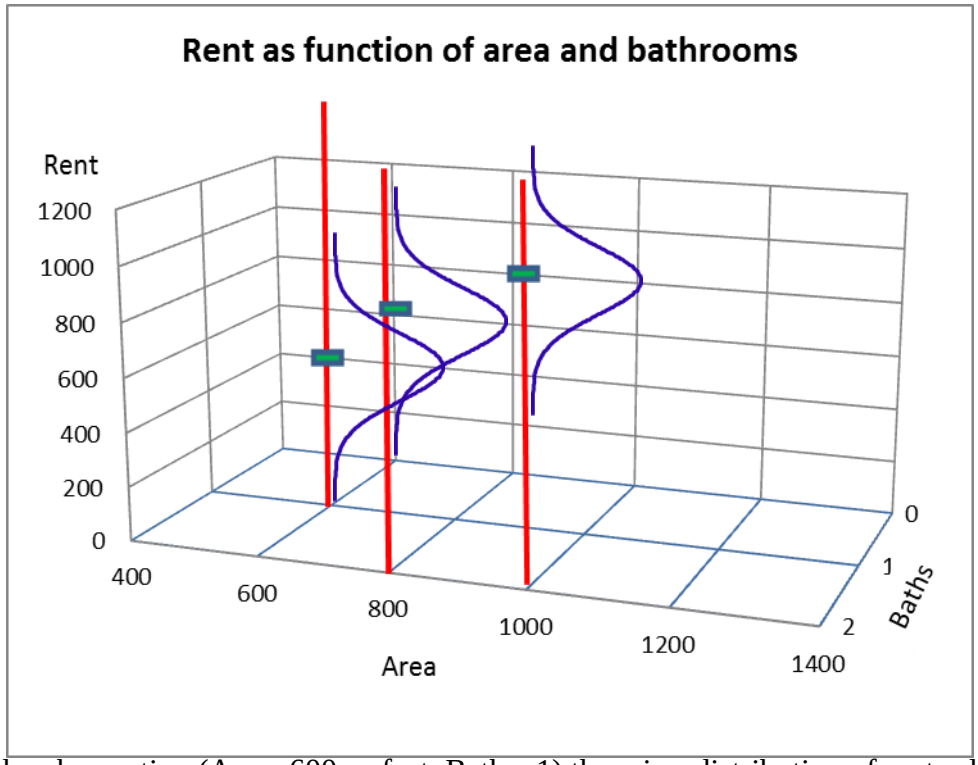
- Assumption (1) is the **linearity specification**. (**L**)
- Assumption (2) is the **homoscedasticity specification**. (**H**)
- Assumption (3) is the **independence specification**. (**I**)
- Assumption (4) is the **normal specification**. (**N**)

The multiple linear regression model is the same as the simple linear regression model, except that there are more given parameters and the **L** specification is a longer linear combination of those parameters. There are *n* units of observation. The first set of constants $x_{12}, x_{13},..., x_{1,k}$ is a vector of the values of *k* - 1 different variables for the first unit of observation. For example, if the unit of observation is an apartment, then the given constants could be the area, number of bathrooms, age, etc. of one apartment. The second set of constants $x_{22}, x_{23},..., x_{2,k}$ is the area, number of bathrooms, age, etc. for a second apartment. And so on.

In short, the model specifications say that at each vector of possible *x*-variables, there is a normal distribution of potential (*Y*) outcomes, and the mean of the normal distribution varies linearly with *x* but the variance does not vary with *x*, and the outcomes are all drawn

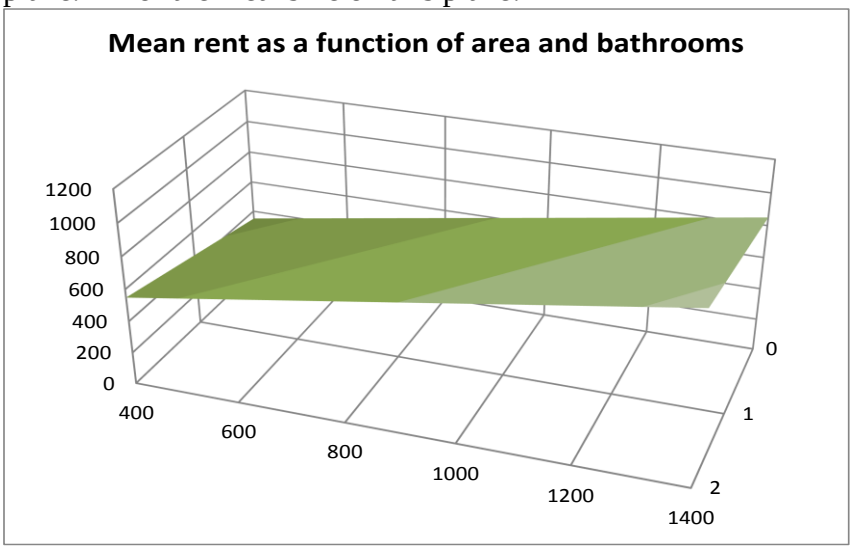
[1] The Latin phrase *ceteris paribus* is also used. It means "all other things being the same." In this context, the "all other things" that remain the same are all of the other predictors *in* the model. But potential predictors that are *not* in the model are not controlled for – they do not necessarily remain the same.

independently. The following graph illustrates matters for three observations of Y = rent, x_1 =area, x_2 = number of bathrooms of apartments:



At the observation (Area=600 sq feet, Baths=1) there is a distribution of rents along the red vertical pole. That distribution of rents is normal (blue curve) with a mean at the green tab and a certain variance. At the observation (Area=800 sq feet, Baths=2) there is another distribution of rents along the red vertical pole. That distribution of rents is normal (blue curve) with a mean at the green tab and the same variance as at the (Area=600 sq feet, Baths=1) pole. Similarly for the pole at (Area=1000, Baths=2).

When you connect the means (the green tabs) for all distributions at all poles, you get a plane. All of the means lie on this plane:



Like the simple linear regression model, the multiple linear regression model can be restated equivalently in terms of the errors:

Suppose that $(x_{12}, x_{13},..., x_{1,k})$, $(x_{22}, x_{23},..., x_{2,k})$, … , $(x_{n2}, x_{n3},..., x_{n,k})$ are given constants. Let $Y_1, Y_2,..., Y_n$ be random variables such that for all $i = 1, \ldots, n$

(1) $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + \varepsilon_i$ and $E(\varepsilon_i) = 0$

(2) $Var(\varepsilon_i) = \sigma^2$

(3) $\varepsilon_1, \varepsilon_2,..., \varepsilon_n$ are independent

(4) ε_i is normally distributed.

If all of these assumptions hold, then the multiple linear regression model is said to be valid.

In case the *x*-constants are actually realizations of random variables $X_2, X_3,..., X_k$, then we can proceed in one of two ways: First (**conditional case**), we can treat the regression as inference on the conditional mean of *Y* given $X_2, X_3,..., X_k$ and continue in the usual regression manner. In this case, our inferences are conditioned upon the given values of the *X*'s, namely $(x_{12}, x_{13},..., x_{1,k})$, $(x_{22}, x_{23},..., x_{2,k})$, … , $(x_{n2}, x_{n3},..., x_{n,k})$. Second (**unconditional case**), we can generalize our inference to cover the variability of *x*-values that were not observed. This requires adjustments in the usual regression procedures to take the sampling variability of the *x*'s into account. The usual regression estimates of the coefficients are still unbiased in this case, but the usual estimates of their variances are not. However, if the sample size is sufficiently large so that the full distribution of $X_2, X_3,..., X_k$ has been adequately sampled, then there is little difference between the conditional and unconditional inferences.

Least-squares estimation of the parameters.

I will illustrate least-squares estimation for two predictor variables ($k = 3$). It should be clear from this illustration how to extend the procedure for more than two predictors. Suppose that $y_1, y_2,..., y_n$ are a realization of the random variables $Y_1, Y_2,..., Y_n$, for which the linear regression model holds at (x_{12}, x_{13}), (x_{22}, x_{23}), … , (x_{n2}, x_{n3}). The regression model says that y_i "should be" $\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$. Thus, the principle of least squares says to estimate β_1, β_2, and β_3 by the values of β_1, β_2, and β_3 that minimize the **Error Sum of Squares** (**ESS**) = $\sum_{i=1}^{n} [y_i - (\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3})]^2$, which upon expanding is

ESS = $\sum_{i=1}^{n} [y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3}]^2 = n\beta_1^2 + 2\beta_1\beta_2 \sum x_{i2} + 2\beta_1\beta_3 \sum x_{i3} + \beta_2^2 \sum x_{i2}^2 +$

$2\beta_2\beta_3 \sum x_{i2}x_{i3} + \beta_3^2 \sum x_{i3}^2 - 2\beta_1 \sum y_i - 2\beta_2 \sum x_{i2}y_i - 2\beta_3 \sum x_{i3}y_i + \sum y_i^2$.

To minimize ESS, take the partial derivatives with respect to β_1, β_2, and β_3, and set the three expressions equal to zero. We obtain

$$\frac{\partial}{\partial \beta_1}(ESS) = 2n\beta_1 + 2\beta_2 \sum x_{i2} + 2\beta_3 \sum x_{i3} - 2\sum y_i = 0$$

$$\frac{\partial}{\partial \beta_2}(ESS) = 2\beta_1 \sum x_{i2} + 2\beta_2 \sum x_{i2}^2 + 2\beta_3 \sum x_{i2}x_{i3} - 2\sum x_{i2}y_i = 0$$

$$\frac{\partial}{\partial \beta_3}(ESS) = 2\beta_1 \sum x_{i3} + 2\beta_3 \sum x_{i2}x_{i3} + 2\beta_3 \sum x_{i3}^2 - 2\sum x_{i3}y_i = 0$$

Solve simultaneously for β_1, β_2, and β_3. It can be verified that the solutions provide a minimum. So the β_1, β_2, and β_3 that minimize the ESS are the solutions to the following simultaneous linear equations, called the **Normal Equations**:

$$n\beta_1 + \beta_2 \sum x_{i2} + \beta_3 \sum x_{i3} = \sum y_i$$

$$\beta_1 \sum x_{i2} + \beta_2 \sum x_{i2}^2 + \beta_3 \sum x_{i2}x_{i3} = \sum x_{i2}y_i$$

$$\beta_1 \sum x_{i3} + \beta_2 \sum x_{i2}x_{i3} + \beta_3 \sum x_{i3}^2 = \sum x_{i3}y_i$$

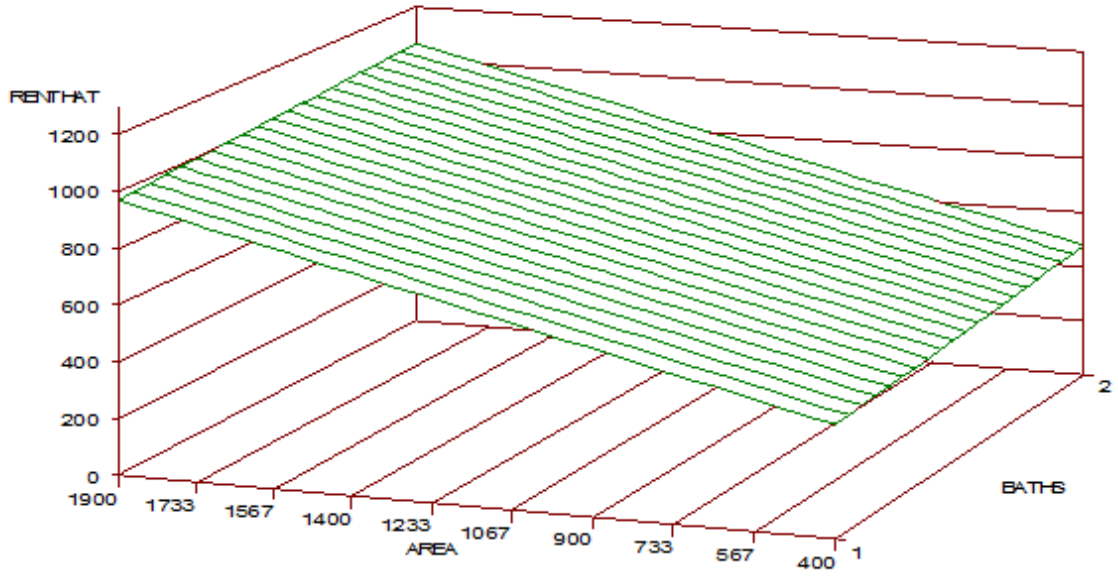
Explicit formulas for the least-squares estimators are complicated. The normal equations are important because they are easy to solve numerically. The normal equations are always linear in the parameters, and there are very good algorithms for solving systems of simultaneous linear equations efficiently.

Under the normal specification (4), the least-squares estimates are also maximum likelihood estimates (why?).

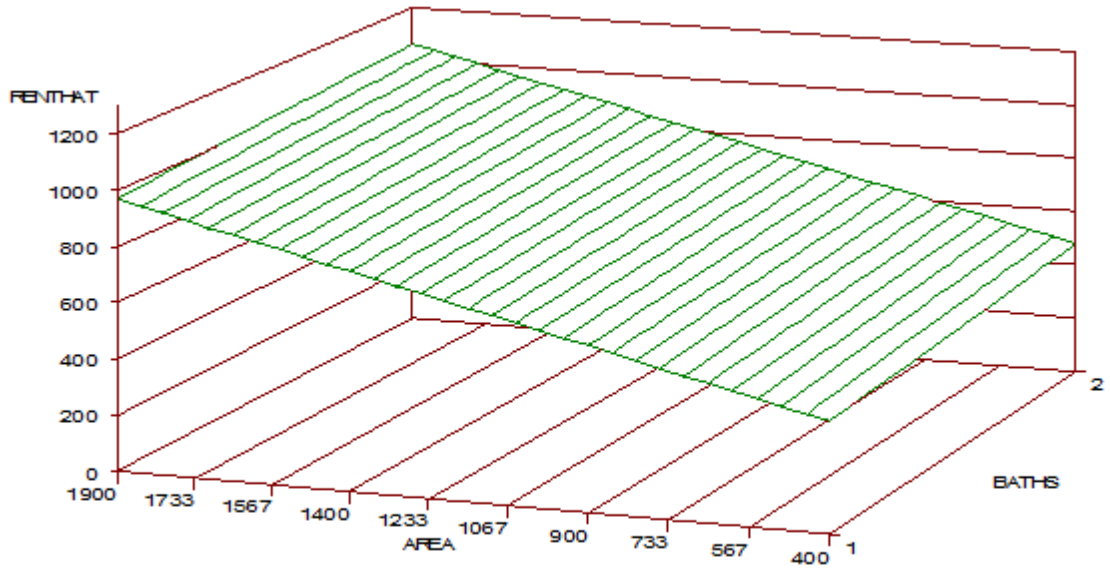
The meaning of the regression coefficients.

If the linear specification holds, then $E(Y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$. If x_2 is increased by 1 unit but no change is made in any other *x*, then *E(Y)* increases to $\beta_1 + \beta_2(x_2 + 1) + \beta_3 x_3 + \cdots + \beta_k x_k = (\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k) + \beta_2$. So the mean of *Y* increases by β_2. Thus the coefficient of each *x* should be interpreted as the change in the mean of *Y* when the *x* increases by 1, *ceteris paribus*, or *controlling for the other x's*. Notice that potential predictors that are not actually used as predictors in the model are *not* controlled for. In simple linear regression, there is only one predictor and therefore *no* predictor is controlled for.

The effect of controlling for the number of bathrooms in a multiple regression of rent (*Y*) on area (X_1) and bathrooms (X_2) is shown graphically below. In weaving, the green lines running parallel to the area axis are called *warp* lines. The slope of the warp lines is the coefficient of area in the multiple regression. For any given green line, the number of bathrooms remains fixed (ignoring for the sake of presentation that bathrooms can have only integer or half-integer values). Only the area changes. This is controlling for bathrooms. The slope is the same on each of the warp lines. This is a feature of linear regression: the slope does not depend upon the value of the controlled-for predictors.



The effect of controlling for area is shown graphically below. In weaving, the green lines running parallel to the bathroom axis are called *woof* lines. The slope of the woof lines is the coefficient of bathrooms in the multiple regression. For any given green line, the area remains fixed. Only the number of bathrooms changes (ignoring for the sake of presentation that bathrooms can have only integer or half-integer values). This is controlling for area. The slope is the same on each of the woof lines. This is a feature of linear regression: the slope does not depend upon the value of the controlled-for predictors.



An important reason for using multiple regression is to control for the effects of other predictors on *Y*, so that you can isolate the stand-alone effect of a predictor on *Y*. This is especially important the harder it is to disentangle the *Y*-effects of two predictors that are themselves closely related. An example might be the area of an apartment and the number of rooms in the apartment. The rent (*Y*) increases with increasing area and with increasing number of rooms. But it is hard to separate the effect of area on rent from the effect of rooms on rent because both area and rooms increase or decrease together. Does the rent increase because the area increases or because the number of rooms increases or both? Area and rooms are **confounded**. The effects of two predictors (on *Y*) are said to be confounded if it is hard to disentangle the effect of one predictor on *Y* from the effect of another predictor on *Y* because of a strong interrelationship between the two predictors.[2] In order to separate the effect of area from the effect of rooms, we need in our data some apartments with about the same area but different numbers of rooms and/or apartments with the same number of rooms but different areas. As long as area and rooms are not too closely entangled, multiple regression can separate the effects. Including both predictors in the regression model allows one to analytically hold the rooms constant and vary the area, or to hold the area constant and increase the rooms. Look to the coefficient of area for the separate effect of area, and look to the coefficient of rooms for the separate effect of rooms. Multiple regression can remove the confounding effect of interrelated predictors. But if the relationship between area and rooms is *extremely* close, we have the problem of (multi)collinearity, which creates further problems – to be discussed later in the course.

In simple linear regression, the least-squares slope coefficient has the formula $r\frac{s_y}{s_x}$, which we have seen has illuminating intuition-building interpretations. Do the multiple regression coefficients have similar representations? Yes. It can be shown that

$$\hat{\beta}_2 = \frac{r_{yx_2} - r_{yx_3}r_{x_2x_3}}{1 - r_{x_2x_3}^{\ 2}}\frac{s_y}{s_{x_2}} \text{ and } \hat{\beta}_3 = \frac{r_{yx_3} - r_{yx_2}r_{x_2x_3}}{1 - r_{x_2x_3}^{\ 2}}\frac{s_y}{s_{x_3}}.$$

Although these formulas do not appear initially to be more than moderately similar to the formula $r\frac{s_y}{s_x}$ for a simple regression coefficient – and at first glance, they do not appear to be at all intuitive – they can be given a very similar interpretation. Bear with me.

First, note what happens in these multiple regression formulas if $r_{x_2x_3} = 0$ (i.e., the predictors are uncorrelated with each other). Then $\hat{\beta}_2 = r_{yx_2}\frac{s_y}{s_{x_2}}$ and $\hat{\beta}_3 = r_{yx_3}\frac{s_y}{s_{x_3}}$. But these expressions are just the values that $\hat{\beta}_2$ and $\hat{\beta}_3$ have in *simple* linear regressions! Thus, when predictor variables are uncorrelated with each other, the multiple regression coefficients are the same as the simple regression coefficients. This means that controlling for the predictors makes no difference – you get the same result as if you did not control for them. This makes intuitive sense because if X_2 and X_3 are uncorrelated with each other, it should not matter whether you hold X_3 constant or not – doing so or not doing so has no effect on what X_2 does to *Y*.

[2] Often, the term *confounding* is reserved for especially severe cases of entangled effects. A strong linear relationship among the predictors is called multicollinearity, which will be discussed thoroughly later in the course.

Second, note what happens if $r_{x_2x_3} = 1$. Then the denominator of $\hat{\beta}_2$ is 0. But also the numerator is 0 because $r_{yx_2} - r_{yx_3}r_{x_2x_3} = r_{yx_2} - r_{yx_3} = 0$ (why?). So the value of $\hat{\beta}_2$ is indeterminate in this case. For the same reason, the value of $\hat{\beta}_3$ is also indeterminate. Both coefficients are also indeterminate if $r_{x_2x_3} = -1$ (why?). So if X_2 and X_3 are perfectly correlated, then a singularity occurs. The problem is that adding X_3 to the model adds no additional explanatory power. Because X_3 is perfectly correlated with X_2, all the information that X_3 would bring to the explanation of *Y* is already present in X_2. So it is expected, in some sense, that an issue would arise in this case.

Third, note that the formulas for these multiple regression coefficients can be rewritten in another way that is enlightening:

$$\hat{\beta}_2 = \frac{r_{yx_2} - r_{yx_3}r_{x_2x_3}}{1 - r_{x_2x_3}^2}\frac{s_y}{s_{x_2}} = \frac{r_{yx_2} - r_{yx_3}r_{x_2x_3}}{\sqrt{1-r_{yx_3}^2}\sqrt{1-r_{x_2x_3}^2}}\frac{s_y\sqrt{1-r_{yx_3}^2}}{s_{x_2}\sqrt{1-r_{x_2x_3}^2}} = r_{yx_2.x_3}\frac{s_{y.x_3}}{s_{x_2.x_3}}$$

$$\hat{\beta}_3 = \frac{r_{yx_3} - r_{yx_2}r_{x_2x_3}}{1 - r_{x_2x_3}^2}\frac{s_y}{s_{x_3}} = \frac{r_{yx_3} - r_{yx_2}r_{x_2x_3}}{\sqrt{1-r_{yx_2}^2}\sqrt{1-r_{x_2x_3}^2}}\frac{s_y\sqrt{1-r_{yx_2}^2}}{s_{x_3}\sqrt{1-r_{x_2x_3}^2}} = r_{yx_3.x_2}\frac{s_{y.x_2}}{s_{x_3.x_2}}$$

- a notation that parallels the formula in the simple linear regression case. In these formulas,

$$r_{yx_2.x_3} = \frac{r_{yx_2} - r_{yx_3}r_{x_2x_3}}{\sqrt{1-r_{yx_3}^2}\sqrt{1-r_{x_2x_3}^2}} \text{ and } r_{yx_3.x_2} = \frac{r_{yx_3} - r_{yx_2}r_{x_2x_3}}{\sqrt{1-r_{yx_2}^2}\sqrt{1-r_{x_2x_3}^2}}.$$

$r_{yx_2.x_3}$ is called the **partial correlation coefficient** between *y* and x_2, controlling for x_3. This measures the strength of the linear relationship between *y* and x_2 when x_3 is held constant. In case the *x*'s are realizations of random variables, the partial correlation $r_{yx_2.x_3}$ is the correlation coefficient between the *y*'s and x_2's in the conditional distribution of (Y, X_2) given X_3. $s_{y.x_3}^2 \equiv s_y^2(1-r_{yx_3}^2)$ is the (estimated) variance of the conditional distribution of *Y* given X_3, and $s_{x_2.x_3}^2 \equiv s_{x_2}^2(1-r_{x_2x_3}^2)$ is the (estimated) variance of the conditional distribution of X_2 given X_3. By the way, these variance formulas reveal that the conditional variance is smaller than the unconditional variance by a fraction ($r_{yx_3}^2$ or $r_{x_2x_3}^2$) that is equal to the proportion of variance explained by the conditioning variable X_3. This again makes intuitive sense because when you control for X_3, you "take out" the effect on *Y* that X_3 is responsible for, and you "take out" the effect on X_2 that X_3 is responsible for. Summing up, we see that the formula for the coefficient of x_2 in this multiple regression, namely $\hat{\beta}_2 = r_{yx_2.x_3}\frac{s_{y.x_3}}{s_{x_2.x_3}}$, exactly parallels the formula for the coefficient of x_2 in the simple regression of *y* on x_2, namely $\hat{\beta}_2 = r_{yx_2}\frac{s_y}{s_{x_2}}$. The multiple

regression coefficient is just the simple regression coefficient for the regression of *y* on x_2, calculated *within* the conditional distribution of (Y, X_2) given X_3.

Similarly, $r_{yx_3.x_2}$ is called the **partial correlation coefficient** between *y* and x_3, controlling for x_2; $s_{y.x_2}{}^2 \equiv s_y{}^2(1-r_{yx_2}{}^2)$ is the (estimated) variance of the conditional distribution of *Y* given X_2; and the formula for the coefficient of x_3 in this multiple regression, namely $\hat{\beta}_3 = r_{yx_3.x_2}\frac{s_{y.x_2}}{s_{x_3.x_2}}$, exactly parallels the formula for the coefficient of x_3 in the simple regression of *y* on x_3, namely $\hat{\beta}_3 = r_{yx_3}\frac{s_y}{s_{x_3}}$. Again, the multiple regression coefficient is just the simple regression coefficient for the regression of *y* on x_3, calculated *within* the conditional distribution of (Y, X_3) given X_2.

But, returning to the main story, the point is that you can get the multiple regression coefficient $\hat{\beta}_2$ from the simple regression formula $\hat{\beta}_2 = r_{yx_2}\frac{s_y}{s_{x_2}}$ by replacing each term of the formula by the corresponding "controlled for" term: r_{yx_2} by $r_{yx_2.x_3}$, s_y by $s_{y.x_3}$, and s_{x_2} by $s_{x_2.x_3}$. The result can be interpreted as *conditional* regression to the mean – i.e., regression to the mean, *for constant* x_3. Similarly for β_3.

The properties of the least-squares estimators can be derived by writing them as linear combinations of $Y_1, Y_2, ..., Y_n$, with the weights depending on the values of the *x*'s. Then apply the properties of linear combinations from Chapter 2, as was done for simple linear regression in Chapter 3. I omit the details. This gives us that the least squares estimators $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are

- unbiased
- consistent
- BLUE (minimum variance among all unbiased estimators that are linear combinations of $Y_1, Y_2, ..., Y_n$)
- normally distributed when the normal specification (4) holds
- maximum likelihood estimators when the normal specification (4) holds.

Explicit formulas for the least-squares estimators $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ as linear combinations of $Y_1, Y_2, ..., Y_n$ are complicated. Progress in analyzing the theoretical properties of least-squares estimators requires a tool to reduce the complexity. Matrix representation has proven a very effective tool in that regard.[3] But for practical purposes, the computations are best left to reliable statistical software.

[3] See my notes "Matrix Approach to Linear Regression.doc" for an introduction to the matrix approach.

Residuals

Paralleling the case of simple linear regression, the **error** that the model makes in saying that y_i "should be" $\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik}$ is $y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3} - \cdots - \beta_k x_{ik}$. This error can be estimated by the **residual** $y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik}$. The magnitude of the single residual $y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik}$ can be taken as an estimate of the "average" deviation σ at a given set of *x*'s. And $(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik})^2$ is an estimate of σ^2, even though it is based on only one data point. You can get a better estimate by using all of the residuals. A good estimate of σ^2 is given by the average

$\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik})^2}{n-k}$, and an unbiased estimator is

$\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik})^2}{n-k}$.

Note that the numerator of the variance estimate is the minimum value of the ESS. The denominator is not *n*. We divide by $n - k$ in order to get an unbiased estimator of σ^2. Unbiasedness follows from the fact that $\dfrac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik})^2}{\sigma^2}$ has a χ^2 distribution [4] with $n - k$ degrees of freedom and so has a mean of $n - k$.

The estimate $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik})^2}{n-k}$ is called the **mean-squared error** (**MSE**), and its square root $\hat{\sigma} = \sqrt{\dfrac{\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik})^2}{n-k}}$ is called the **root mean-squared error** (**RMSE**). The RMSE is, in effect, the standard deviation of the residuals because the mean of the residuals is zero. So the RMSE can be interpreted as a kind of "average" amount by which the estimates $\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \cdots + \hat{\beta}_k x_{ik}$ miss their targets y_i,

---

[4] It is tedious to show this, but it can be done with essentially just the properties given in Chapter 2. Start with the fact that $\dfrac{\sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3} - \cdots - \beta_k x_{ik})^2}{\sigma^2}$ has a χ^2 distribution with *n* degrees of freedom. (Why?) Then add and subtract the estimated regression to write $\dfrac{\sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3} - \cdots - \beta_k x_{ik})^2}{\sigma^2}$

$= \dfrac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik} + (\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 x_{i2} - \beta_2 x_{i2}) + \cdots + (\hat{\beta}_k x_{ik} - \beta_k x_{ik}))^2}{\sigma^2} =$

$\dfrac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik})^2}{\sigma^2}$ + *k* independent terms. The *k* terms are "split off", each term involving one of the parameters $\beta_1, \beta_2, \ldots, \beta_k$ and its corresponding estimate. Each of the *k* terms will be the square of a standard normal distribution and so be χ^2 with one degree of freedom. The sum of the *k* terms will be χ^2_k. So the first term on the right-hand side should be χ^2_{n-k} if it is independent of the *k* terms (it is).

in the same way that the ordinary (sample) standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ can be interpreted as a kind of "average" amount by which the estimate $\bar{x}$ misses its targets x_i. Thus, RMSE is a fundamental measure of how well the regression model fits the data.[5]

R-square

R-square for a multiple regression is similar to R-square for a simple regression. The multiple regression R-square remains the proportionate improvement in the ESS when going from using no predictors to using all of them. The improvement in ESS is $\sum(y_i - \bar{y})^2$ - $\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \cdots - \hat{\beta}_k x_{ik})^2 = \sum_{i=1}^{n}(\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \cdots + \hat{\beta}_k x_{ik} - \bar{y})^2$, which is nonnegative. So using the *x*'s improves the ESS. The improvement $\sum_{i=1}^{n}(\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \cdots + \hat{\beta}_k x_{ik} - \bar{y})^2$ is called the **regression sum of squares** (**RSS**).

The proportion of improvement is RSS/TSS = $\sum_{i=1}^{n}(\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \cdots + \hat{\beta}_k x_{ik} - \bar{y})^2 / \sum(y_i - \bar{y})^2$, which is **R-square**. As a proportion or percent, R-square necessarily lies between 0 and 1. If the regression line fits the data perfectly, then $y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \cdots + \hat{\beta}_k x_{ik}$ for all *i*, so R-square = 1 (complete improvement). If the regression line is flat, then $\hat{\beta}_i = 0$ for all $i > 1$ and $\hat{\beta}_1 = \bar{y}$ (check the formula for $\hat{\alpha}$), so R-square = 0 (no improvement).

What if you start with some predictors already in the model and add more? Since RSS ≥ 0, the ESS with predictors is never worse than the ESS without predictors. But that does not answer the question of whether the ESS could get worse if you start with some predictors already in the model and add more predictors. Intuitively, the ESS should not get worse in this case, because additional predictors bring additional information to improve the fit. The intuition is correct: Denote by ESS(*m*) the minimum possible ESS that results from using predictors $X_2, X_3, \ldots, X_m$. Denote by ESS(*k*) the minimum ESS from using predictors $X_2, X_3, \ldots, X_m, X_{m+1}, \ldots, X_k$. Since ESS(*k*) is the minimum for <u>*all*</u> choices of coefficients $\beta_1, \ldots, \beta_k$, and since one of those possible choices is to pick for $\beta_1, \ldots, \beta_m$ the values that minimize ESS(*m*) and to pick zeroes for the additional coefficients $\beta_{m+1}, \ldots, \beta_k$, then the minimum ESS(*k*) must be ≤ the ESS with that choice. But the ESS with that choice is ESS(*m*). So ESS(*k*) ≤ ESS(*m*). So adding more predictors never increases ESS and therefore never reduces RSS and therefore never reduces R-square. (This argument is a proof!)

The observation that RSS and ESS change for the better when more predictors are added can be exploited to make a test of whether the additional predictors collectively contribute to the model explanation of *Y*. We simply find a way to test whether the improvement in ESS (or RSS) is statistically significant. Intuitively, if the improvement in ESS (or RSS) is small, then the added predictors do not likely contribute much to the explanation of *Y*. If the improvement is large, then they do. But how much improvement is enough to trigger significance? Providing an answer is complicated by the dependence of ESS and RSS on the scale of measurement. Switching from feet to inches multiplies ESS and RSS by 144 without changing anything real

[5] The other fundamental measure of model fit is "R-square".

about the relationships. A unitless measure could be obtained by the change in R-square. That is, divide the change in ESS (or RSS) by the TSS. That try gets Honorable Mention. The reason it is not fully satisfactory is that our measure needs to be a statistic with a known and hopefully "easy" distribution for hypothesis testing. The numerator and denominator of the ratio (*change in ESS*) / *TSS* can be turned into chi-square random variables, so one might think the ratio could be turned into an F-statistic. But the F-statistic requires the ratio of *independent* chi-squares. It turns out that *change in ESS* and *TSS* are dependent.[6] The solution is to replace TSS in the denominator of the ratio by ESS(*k*). It turns out that *change in ESS* and ESS(*k*) are independent.

The Wald test

To develop this idea further, consider two regression models **U** and **R**. They are the same, except for the linearity specification.:

L specification for Model **U**: $E(Y_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \beta_{m+1} x_{i,m+1} + \cdots + \beta_k x_{ik}$

L specification for Model **R**: $E(Y_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_m x_{im}$

"**U**" means "**U**nrestricted" and "**R**" means "**R**estricted" because in Model **R**, the coefficients $\beta_{m+1}, \ldots, \beta_k$ are effectively restricted to be zeroes. Note that Model U has more predictors (*k*) than Model R (*m*). We start with Model **R** and add more predictors to get Model **U**. We have

TSS = ESS(**U**) + RSS(**U**)
TSS = ESS(**R**) + RSS(**R**)

Note we don't need to distinguish TSS(**U**) and TSS(**R**). The reason is that TSS(**R**) = TSS(**U**) because both are equal to $\sum (y_i - \bar{y})^2$. The TSS does not change with the number of predictors because TSS involves only *y*-data, and we have the same *y*-data however many predictors we use. Because Model **U** adds predictors, we have ESS(**U**) ≤ ESS(**R**). The change in ESS is ESS(**R**) – ESS(**U**). This change is generated by the $k - m$ added predictors $X_{m+1}, \ldots, X_k$. The change can be split into $k - m$ independent terms, each of which represents the reduction in ESS attributable to one of the added predictors. When each such term is divided by σ^2, the result is a χ_1^2 random variable, under the assumption/hypothesis that the coefficient of the corresponding *X* is zero. So their sum $\dfrac{\text{ESS(}\mathbf{R}\text{) - ESS(}\mathbf{U}\text{)}}{\sigma^2}$ is χ^2_{k-m}. Also, $\dfrac{\text{ESS(}\mathbf{U}\text{)}}{\sigma^2}$ is χ^2_{n-k}. And, as I commented previously, it can be shown that $\dfrac{\text{ESS(}\mathbf{R}\text{) - ESS(}\mathbf{U}\text{)}}{\sigma^2}$ and $\dfrac{\text{ESS(}\mathbf{U}\text{)}}{\sigma^2}$ are independent. Thus (why?) the ratio $\dfrac{\text{ESS(}\mathbf{R}\text{) - ESS(}\mathbf{U}\text{)}}{\sigma^2(k-m)} \Big/ \dfrac{\text{ESS(}\mathbf{U}\text{)}}{\sigma^2(n-k)} = \dfrac{\text{ESS(}\mathbf{R}\text{) - ESS(}\mathbf{U}\text{)}}{k-m} \Big/ \dfrac{\text{ESS(}\mathbf{U}\text{)}}{n-k}$ is $F_{k-m,n-k}$, under the hypothesis that the coefficients of all added predictors are zeroes. So we can test this hypothesis by computing the $F_{k-m,n-k}$ statistic and seeing if the computed value is likely for the $F_{k-m,n-k}$ distribution.

Formally, test $H_0 : \beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0$ vs. $H_1 :$ at least one $\beta_{m+1}, \beta_{m+2}, \ldots, \beta_k \neq 0$ by rejecting the null hypothesis if $\dfrac{\text{ESS(}\mathbf{R}\text{) - ESS(}\mathbf{U}\text{)}}{k-m} \Big/ \dfrac{\text{ESS(}\mathbf{U}\text{)}}{n-k} > F_{k-m,n-k}$ critical point. Note that this test is one-tailed. Small values of the F statistic occur if the reduction in ESS from

[6] It is intuitive that the change and TSS are positively correlated: The larger the change, the larger the TSS. In fact, it is clear that TSS cannot be less than the change.

adding the extra predictors is small, which is evidence in favor of the null hypothesis. Large values of the F statistic occur if the reduction in ESS from adding the extra predictors is large, which is evidence against the null hypothesis. This test of several coefficients is called the Wald test.

A special case of the Wald test occurs when Model R is the no-predictor model $E(Y_i) = \beta_1$. Then the null hypothesis is that *all* predictors have zero coefficients. If that were true, then none of the predictors contributes to explaining *Y*, so none of them belongs in the model. This test is performed by all regression software and the result printed on the output. It is the model F test. The idea is that it provides a quick check of whether any of the predictors might be useful in explaining *Y*.

Adjusted R-square

R-square cannot decrease when more predictors are added to the linear specification. The sampling variability of added predictors will increase the calculated R-square, even if the added predictors really have zero coefficients in the population. The sample R-square could be built up by throwing relatively worthless predictors into the model. Therefore, the sample R-square could give an inflated picture of the value of the predictors. Various adjustments to R-square have been proposed so that R-square will give a more accurate picture of the model's actual explanatory power. The most common adjustment is called **adjusted R-square**, or R-square adjusted for degrees of freedom, and is written $\overline{R}^2$. It adjusts the numerator and denominator of R-square in the following way: Since R-square = $RSS / TSS = (TSS - ESS) / TSS = 1 - ESS / TSS$ and ESS has $n - k$ degrees of freedom and TSS has $n - 1$ degrees of freedom, then adjusted R-square proposes to turn ESS and TSS into unbiased estimators of variance by dividing them by their degrees of freedom. So $\overline{R}^2 = 1 - \frac{ESS/(n-k)}{TSS/(n-1)}$. It is true that $\overline{R}^2 \leq$ R-square. But $\overline{R}^2$ can be negative.

Other properties

Many other properties of multiple regression models are similar to their analogues from simple linear regression. Among these are the distributions, confidence intervals, and hypothesis tests for individual coefficients, for the conditional mean of *Y* given the *X*'s, and for estimating an individual value of *Y* at given *X*'s. Accordingly, discussion of these items will not be given here.