Multiple Regression:

The example we use: predict the number of accidents (Y) from the number of licensed vehicles (X_1) and the number of police (X_2)

comm	vehicles (X_1)	police (X_2)	accidents (Y)
	(1000s)		(100s)
1	4	20	1
2	10	6	4
3	15	2	5
4	12	8	4
5	8	9	3
6	16	8	4
7	5	12	2
8	7	15	1
9	9	10	4
10	10	10	2

Plotting our data in a three-dimensional space where the x, y, and z axes correspond to X_1, X_2, and Y, respectively,

the least-squares task is to look for a best-fitting plane of the form

$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$.

The sum of the squared lengths of the vertical projections to the plane is to be minimized.

Some computational formulas:

SSTO $= \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2 =$

$\boldsymbol{Y}'\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}'\mathbf{1}\mathbf{1}'\boldsymbol{Y} =$ (where $\mathbf{1}$ is a column vector of one's)

$\boldsymbol{Y}'[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}]\boldsymbol{Y}$ (where $\boldsymbol{J}$ is a matrix of one's)

$$\text{SSE} = \sum e_i^2 = \sum (Y_i - \widehat{Y}_i)^2 =$$

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y} =$$

$$\boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y} = \boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}$$

$$\text{SSR} = \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y} - \tfrac{1}{n}\boldsymbol{Y}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{Y} =$$

$$\boldsymbol{Y}'[\boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \tfrac{1}{n}\boldsymbol{J}]\boldsymbol{Y} = \boldsymbol{Y}'[\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}]\boldsymbol{Y}$$

A “quadratic form” has the form (pun intended):

$$\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y} = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}Y_iY_j$$

where $\boldsymbol{A}$ is a symmetric matrix.

Note that $\boldsymbol{Y}'[\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}]\boldsymbol{Y}$, $\boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}$, and

$\boldsymbol{Y}'[\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}]\boldsymbol{Y}$ are all quadratic forms, implying that all of the various sums of squares can be put into this framework.

Thus, distributional results obtained for quadratic forms generally can be applied directly to our sums of squares.

The covariance between two linear combination:

Suppose I have n random variables, $X_1, \ldots, X_n$, and two linear combinations:

$a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{in}X_n$ and

$a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jn}X_n$

The covariance between these two linear combinations is

$a_{i1}a_{j1}V(X_1) + \cdots + a_{in}a_{jn}V(X_n) +$

$2a_{i1}a_{j2}Cov(X_1, X_2) + \cdots + 2a_{i1}a_{jn}Cov(X_1, X_n) +$

$2a_{i2}a_{j3}Cov(X_2, X_3) + \cdots + 2a_{i2}a_{jn}Cov(X_2, X_n)$

$+ \cdots + 2a_{i(n-1)}a_{jn}Cov(X_{n-1}, X_n)$

Let $\boldsymbol{a}_{\boldsymbol{i}}' = [a_{i1}\ a_{i2}\ \cdots\ a_{in}]$,

$\boldsymbol{a}_{\boldsymbol{j}}' = [a_{j1}\ a_{j2}\ \cdots\ a_{jn}]$, and

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

$$\sigma^2(\boldsymbol{X}) = \begin{bmatrix} V(X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & & \vdots \\ Cov(X_n, X_1) & \cdots & V(X_n) \end{bmatrix}$$

$Cov(\boldsymbol{a}_{\boldsymbol{i}}'\boldsymbol{X}, \boldsymbol{a}_{\boldsymbol{j}}'\boldsymbol{X}) =$

$\boldsymbol{a}_{\boldsymbol{i}}'[\sigma^2(\boldsymbol{X})]\boldsymbol{a}_{\boldsymbol{j}}$

$$E(\boldsymbol{a}_{\boldsymbol{i}}^{'}\boldsymbol{X}) = \boldsymbol{a}_{\boldsymbol{i}}^{'}E(\boldsymbol{X}) = a_{i1}E(X_1)+\cdots+a_{in}E(X_n)$$

$$E(\boldsymbol{a}_{\boldsymbol{j}}^{'}\boldsymbol{X}) = \boldsymbol{a}_{\boldsymbol{j}}^{'}E(\boldsymbol{X}) = a_{j1}E(X_1)+\cdots+a_{jn}E(X_n)$$

Let

$$\boldsymbol{A}_{r\times n} = \begin{bmatrix} \boldsymbol{a}_{\boldsymbol{1}}^{'} \\ \boldsymbol{a}_{\boldsymbol{2}}^{'} \\ \vdots \\ \boldsymbol{a}_{\boldsymbol{r}}^{'} \end{bmatrix}$$

$$\boldsymbol{A}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{a}_{\boldsymbol{1}}^{'}\boldsymbol{X} \\ \boldsymbol{a}_{\boldsymbol{2}}^{'}\boldsymbol{X} \\ \vdots \\ \boldsymbol{a}_{\boldsymbol{r}}^{'}\boldsymbol{X} \end{bmatrix}$$

1) $E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X})$

2) $\sigma^2(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\sigma^2(\boldsymbol{X})\boldsymbol{A}^{'}$

Inferences in Regression – Matrix Re-expression:

$$\boldsymbol{b} = \left[\begin{array}{c} b_0 \\ b_1 \end{array} \right] =$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{A}_{\boldsymbol{2}\times\boldsymbol{n}}\boldsymbol{Y}_{\boldsymbol{n}\times\boldsymbol{1}}$$

$$\sigma^2\{\boldsymbol{b}\} = \boldsymbol{A}\sigma^2\{\boldsymbol{Y}\}\boldsymbol{A}' = \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}' =$$

$$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2 \left[\begin{array}{cc} \frac{\sum X^2}{n\sum(X_i - \bar{X})^2} & \frac{-\bar{X}}{\sum(X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\sum(X_i - \bar{X})^2} & \frac{1}{\sum(X_i - \bar{X})^2} \end{array} \right]$$

Notation:

$$s^2\{\boldsymbol{b}\} = MSE(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

One application:

Let $\widehat{Y_h} = b_0 + b_1 X_h =$

$$[1 \;\; X_h] \left[\begin{array}{c} b_0 \\ b_1 \end{array} \right] =$$

$$Var(\widehat{Y_h}) = [1 \;\; X_h] \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \left[\begin{array}{c} 1 \\ X_h \end{array} \right] =$$

$\sigma^2 [\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}]$

Multiple Regression Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and all ϵ_i are independent.

Again, have n observations –

$$\boldsymbol{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}; \boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}; \boldsymbol{\epsilon}_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix};$$

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} 1 & X_{11} & \cdots X_{1(p-1)} \\ \vdots & & \vdots \\ 1 & X_{n1} & \cdots X_{n(p-1)} \end{bmatrix};$$

$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where

$E(\boldsymbol{\epsilon}) = \mathbf{0}$

$\sigma^2\{\boldsymbol{\epsilon}\} = \sigma^2\boldsymbol{I}$

I just need to translate my results for the case of one dependent and one independent variables:

Least-squares estimates:

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

Unbiasedness: $E(\boldsymbol{b}) = \boldsymbol{\beta}$

Residuals: $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$

Fitted values: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b}$

SSTO = $\sum(Y_i - \bar{Y})^2$

SSE = $\sum e_i^2 = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}$

SSR is by subtraction

$\sigma^2\{\boldsymbol{b}\} = \sigma^2(X'X)^{-1}$

$s^2\{\boldsymbol{b}\} = MSE(X'X)^{-1}$

Source	df	SS	MS	F
Regression	$p-1$	SSR	MSR	MSR/MSE
Error	$n-p$	SSE	MSE	$\sim F_{p-1,n-p}$
Total	$n-1$	SSTO		

I could also do Error = SSPE + SSLF, but I need sets of identical observations on the variables and these are usually hard to get.

The test is for $H_o : \beta_1 = \cdots = \beta_{p-1} = 0$ simultaneously.

In MSE there are $n-p$ degrees of freedom; you lose p because you estimate p parameters, $\beta_0, \ldots, \beta_{p-1}$

Coefficient of multiple determination: $R^2 = 1 - \frac{SSE}{SSTO}$;

the (positive) square root is the coefficient of multiple correlation.

This is the correlation between Y_i and $\widehat{Y}_i$ over all i from 1 to n

Among all linear combinations this correlation is the maximum possible.

Adjusted R^2:

Even though we define $R^2 = 1 - \frac{SSE}{SSTO}$,

the expectation is not exactly zero when $H_o : \beta_1 = \cdots = \beta_{p-1} = 0$ is true.

i.e., $E(\frac{SSE}{SSTO}) = \frac{n-p}{n-1}$, which implies $E(R^2) = 1 - \frac{n-p}{n-1} = \frac{p-1}{n-1}$

So, the suggested correction called adjusted R^2:

$$R^2_{adjusted} = 1 - (\frac{n-1}{n-p})(\frac{SSE}{SSTO}),$$

which has expectation of zero when H_o is true.

This is an attempt to get closer to an unbiased estimate in the correlational model we will talk about later.

Called Wherry's "shrinkage" formula inappropriately ("shrinkage" refers to cross-validation, which is not what this is).

Note:

$R_a^2 = \frac{R^2 - E(R^2)}{1 - E(R^2)} =$

$1 - (\frac{n-1}{n-p})(\frac{SSE}{SSTO})$

In simple linear regression, we had confidence intervals for each of the regression coefficients, for the mean value of Y at a particular level of X, and a prediction interval for a new observation.

We have all of these in the multiple regression context as well.

a) For β_k:

$$(b_k - \beta_k)/s(b_k) \sim t_{n-p}$$

where $s(b_k)$ is the square root of the appropriate diagonal entry in $s^2\{\boldsymbol{b}\} = MSE(\boldsymbol{X}'\boldsymbol{X})^{-1}$

b) For $\widehat{Y}_h = b_0 + b_1 X_{h1} + \cdots + b_{p-1} X_{h(p-1)}$,

$Var(\widehat{Y}_h) = \boldsymbol{X}_{\boldsymbol{h}}'[\boldsymbol{\sigma}^{\boldsymbol{2}}(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}]\boldsymbol{X}_{\boldsymbol{h}}$, where

$$\boldsymbol{X_h} = \begin{bmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{h(p-1)} \end{bmatrix}$$

Thus,

$$\frac{\widehat{Y}_h - (b_0 + b_1 X_{h1} + \cdots + b_{p-1} X_{h(p-1)})}{\sqrt{MSE(\boldsymbol{X_h'}[(\boldsymbol{X'X})^{-1}]\boldsymbol{X_h})}} \sim t_{n-p}$$

c) For a new observation, Y_h:

$$\frac{(Y_h - \widehat{Y}_h)}{\sqrt{MSE(1 + \boldsymbol{X_h'}[(\boldsymbol{X'X})^{-1}]\boldsymbol{X_h})}} \sim t_{n-p}$$

Regression coefficients:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i$$

The betas are sometime called "partial regression coefficients". Other things being held constant (ceteris paribus), one unit change in X_{ik} results in a β_k change in Y_i

This is not, however, a means of arguing for the importance of a particular variable in predicting Y (because all of the independent variables are intercorrelated among themselves).

Note:

$$\frac{Y_i}{\sigma_Y} = \frac{\beta_0}{\sigma_Y} + \beta_1 \frac{\sigma_{X_1}}{\sigma_Y} (\frac{X_{i1}}{\sigma_{X_1}}) + \cdots +$$

$$\beta_{p-1} \frac{\sigma_{X_{p-1}}}{\sigma_Y} (\frac{X_{i(p-1)}}{\sigma_{X_{p-1}}}) + \frac{\epsilon_i}{\sigma_Y}$$

$\beta_k \frac{\sigma_{X_k}}{\sigma_Y}$ are “standardized regression coefficients” (because all the variables now have unit variance.

$b_k \frac{S_{X_k}}{S_Y}$ are called (confusingly) “beta” coefficients.

This does get rid of the scale problem but not the problem of interpreting importance in the framework of a correlated system of independent variables –

this lingers on like a curse for anyone using multiple regression in a data analytic context.

Inference in Multiple Regression:

The “Extra Sum of Squares Principle” (Model Comparisons)

This is a very general strategy for testing hypotheses about regression coefficients.

In the general model:

$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

we know how to test:

a) $H_o : \beta_1 = \cdots = \beta_{p-1} = 0$, with

$\frac{MSR}{MSE} \sim F_{p-1,n-p}$

b) $H_o : \beta_k = 0$, with

$[\frac{b_k}{s(b_k)}]^2 \sim [t_{n-p}]^2 \sim F_{1,n-p}$

Thus we know how to test “one” or “all”

Now I want a strategy for

$H_o : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$, i.e., we set $p - q$ of the betas equal to zero.

In general, because we can label and order the variables as we wish, this is a way of testing any $p - q$ coefficients against zero.

If $q = p - 1$, we get $H_o : \beta_{p-1} = 0$

If $q = 1$, we get $H_o : \beta_1 = \cdots = \beta_{p-1} = 0$

Now, suppose we fit the Full Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i$$

and get: $SSTO = SSE(F) + SSR(F)$, where "F" stands for "Full"

Now, consider the reduced model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)} + \epsilon_i$$

Fitting this, we get

$SSTO = SSE(R) + SSR(R)$, where "R" stands for "Reduced"

Consider the "Extra Sum of Squares" defined by $SSE(R) - SSE(F)$. First, what is the sign of the extra sum of squares? (does SSE(R) need to be bigger than SSE(F)?)

If $H_o : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$ is true, what do you expect $SSE(R) - SSE(F)$ to reflect? (answer: just error)

Degrees of Freedom:

$SSE(F) \to n - p$;

$SSE(R) \to n - q$

Thus, $SSE(R) - SSE(F) \to p - q$

So, I could form

$\frac{SSE(R)-SSE(F)}{p-q}$

which estimates σ^2 unbiasedly under $H_o : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$

If H_o is not true, it will be expected to be larger than σ^2

I need a denominator that always estimates σ^2 unbiasedly

Assuming the Full Model is the “true” one

$\frac{SSE(F)}{n-p}$ is it –

Thus, under $H_o : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$,

$\frac{[SSE(R)-SSE(F)]/(p-q)}{SSE(F)/(n-p)} \sim F_{p-q,n-p}$

If $q = p - 1$, we get $H_o : \beta_{p-1} = 0$ and $F_{1,n-p}$

If $q = 1$, we get $H_o : \beta_1 = \cdots = \beta_{p-1} = 0$ and $F_{p-1,n-p}$

We call the term, $SSE(R)-SSE(F)$, the “extra sum of squares”; it is the amount added to SSR if the additional variables need to reach the Full model are added to the Reduced model

Denoted as

$SSR(X_q, X_{q+1}, \ldots, X_{p-1}|X_1, \ldots, X_{q-1})$

We could build this up sequentially:

$SSR(X_1, \ldots, X_{p-1}) = SSR(X_1)+SSR(X_2|X_1)+ \cdots + SSR(X_{p-1}|X_1, \ldots, X_{p-2})$

The procedure of fitting Full and Reduced models is very general. Suppose, the null hypothesis is

$H_o : \beta_i = \beta_j$

Then $SSE(F)$ has $n-p$ degrees of freedom; $SSE(R)$ has $n-(p-1)$ degrees of freedom;

Thus, $\frac{[SSE(R)-SSE(F)]/1}{SSE(F)/(n-p)} \sim F_{1,n-p}$

Nonlinear models fit by multiple linear regression:

a) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$,

where $X_{i2} = X_{i1}^2$ gives a curve when plotting the fitted values in the X_1 by Y space (this is polynomial regression).

b) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$

includes the "interaction" of X_1 and X_2;

rewriting:

$Y_i = \beta_0 + \beta_1 X_{i1} + (\beta_2 + \beta_3 X_{i1}) X_{i2} + \epsilon_i$

Thus, the regression coefficient on X_2 changes depending on the level of X_1

(or we could rephrase as the regression coefficient on X_1 changing depending on the level of X_2)

We refer to the model as “nonadditive” when an interaction product term is included.

c) $Y_i = \beta_0 X_i^{\beta_1} \rightarrow$

$\log(Y_i) = \log(\beta_0) + \beta_1 \log(X_i) \rightarrow$

$Y_i^{'} = \beta_0^{'} + \beta_1 X_i^{'}$