

Least Squares Fitting and Inference

Kerby Shedden
August, 2008

Definitions and Motivation

Independent variables (predictors, regressors, covariates):

$$X = (X_1, \dots, X_p)'$$

Dependent variable (response, outcome): Y

The goal is to learn about an unknown function f , where

$$Y \approx f(X)$$

based on a finite collection of data

$$X_i = (X_{i1}, \dots, X_{ip})' \\ i = 1, \dots, n.$$

where n is the sample size.

Why do we want to do this?

- *Prediction*: Evaluate $f(X)$ to predict the “typical” value of Y at a given X point (not necessarily one of the X_i points in the data).
- *Model inference*: Inductive learning about the relationship between X and Y , such as better understanding which predictors or combinations of predictors are associated with particular changes in Y .

Model inference often has the goal of better understanding the physical (biological, social, etc.) mechanism underlying the relationship between X and Y (but keep in mind that with observational data, causal inferences are difficult or impossible to make).

Examples:

- An empirical model for the weather 48 hours from now based on current satellite data, historical patterns, etc. would primarily be of interest for prediction, rather than inference. Such a model would have a lot of practical value (if it were actually predictive), but would not necessarily provide a lot of insight into the atmospheric processes that underly changes in the weather.
- A study of the relationship between childhood lead exposure and subsequent behavioral problems would primarily be of interest for inference, rather than prediction. Such a model could be used to assess whether there is a risk due to lead exposure, and to estimate for a large population how many children are adversely affected. The effect of lead exposure on an individual child is probably too small in relation to numerous other risk factors for such a model to be of predictive value at the individual level.

Statistical interpretations of the regression function

The most common way of putting curve fitting into a statistical framework is to define f as the conditional expectation

$$f(X) \equiv E(Y|X).$$

Much less commonly (but occasionally usefully) the regression function is defined as a conditional quantile, such as the median

$$f(X) \equiv \text{median}(Y|X)$$

or even some other quantile $f(X) \equiv Q_p(Y|X)$.

The conditional expectation function

The conditional expectation function $E(Y|X)$ can be viewed in two ways:

1. As a deterministic function of X , essentially what we would get if we sampled a large number of X, Y pairs from their joint distribution, and took the average of the Y values that are coupled with a specific X value. If there are densities we can write:

$$E(Y|X) = \int Y f(Y|X) dY.$$

2. As a scalar random variable. A realization is obtained by sampling X from its marginal distribution, then plugging this value into the deterministic function described in 1 above.

Least Squares Fitting

In a linear model, one postulates that the independent variables are related to Y via a linear relationship

$$Y_i \approx \sum_{j=1}^p \beta_j X_{ij} = \beta' X_i.$$

In this case, to estimate f we need to estimate the β_j . One approach to doing this is to minimize the following function of β :

$$L(\beta) = \sum_i (Y_i - \sum_j \beta_j X_{ij})^2 = \sum_i (Y_i - \beta' X_i)^2$$

A special case is “simple linear regression,” when there is $p = 1$ covariate.

$$L(\alpha, \beta) = \sum_i (Y_i - \alpha - \beta X_i)^2$$

Here we can differentiate with respect to α and β :

$$\begin{aligned} \partial L / \partial \alpha &= -2 \sum_i (Y_i - \alpha - \beta X_i) &= -2 \sum_i R_i \\ \partial L / \partial \beta &= -2 \sum_i (Y_i - \alpha - \beta X_i) X_i &= -2 \sum_i R_i X_i \end{aligned}$$

$$R_i = Y_i - \alpha - \beta X_i$$

is the “working residual” (requires working values for α and β).

Setting

$$\partial L/\partial\alpha = \partial L/\partial\beta = 0$$

and solving for α and β yields

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{\sum Y_i X_i/n - \bar{Y}\bar{X}}{\sum X_i^2/n - \bar{X}^2}$$

where $\bar{Y} = \sum Y_i/n$ and $\bar{X} = \sum X_i/n$ are the sample mean values (averages).

The “hat” notation ($\hat{}$) distinguishes the least-squares optimal values of α and β from an arbitrary pair of values.

Two important identities:

$$\sum X_i^2/n - \bar{X}^2 = \sum_i (X_i - \bar{X})^2/n$$

$$\sum Y_i X_i/n - \bar{Y} \bar{X} = \sum Y_i (X_i - \bar{X})/n = \sum_i (Y_i - \bar{Y})(X_i - \bar{X})/n.$$

Note that

$$\sum_i (X_i - \bar{X})^2/n \quad \text{and} \quad \sum_i (Y_i - \bar{Y})(X_i - \bar{X})/n.$$

are essentially the sample variance of X_1, \dots, X_n , and the sample covariance of the (X_i, Y_i) pairs. Since $\hat{\beta}$ is their ratio, we can replace n in the denominator with $n - 1$ so that

$$\hat{\beta} = \frac{\widehat{\text{cov}}(Y, X)}{\widehat{\text{var}}(X)}$$

where $\widehat{\text{cov}}$ and $\widehat{\text{var}}$ are the usual unbiased estimates of variance and covariance.

Review of norms

The Euclidean norm on vectors (the main norm we will use in this class) is:

$$\|V\| = \sqrt{\sum_i V_i^2}.$$

Some useful identities:

$$\|V + W\|^2 = \|V\|^2 + \|W\|^2 + 2V'W$$

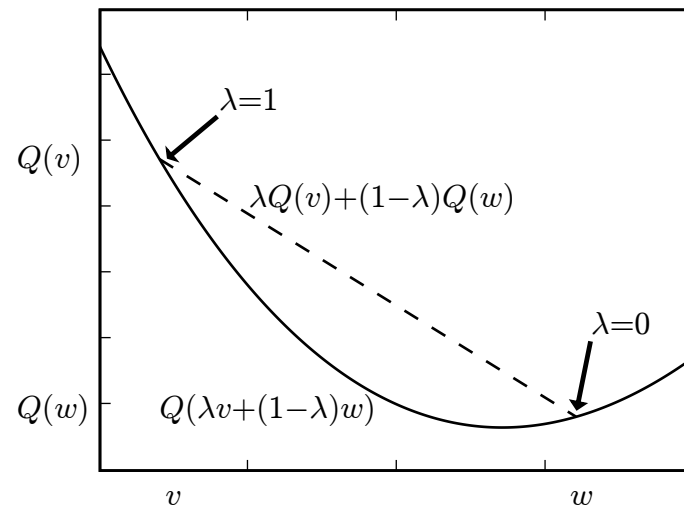
$$\|V - W\|^2 = \|V\|^2 + \|W\|^2 - 2V'W$$

Review of convexity

A map $Q : \mathcal{R}^d \rightarrow \mathcal{R}$ is convex if for any $v, w \in \mathcal{R}^d$,

$$Q(\lambda v + (1 - \lambda)w) \leq \lambda Q(v) + (1 - \lambda)Q(w),$$

for $0 \leq \lambda \leq 1$. If the inequality is strict for $0 < \lambda < 1$ and all $v \neq w$, then Q is strictly convex.



A key property of strictly convex functions is that they have a unique global minimizer. That is, there exists $v \in \mathcal{R}^d$ such that $Q(v) < Q(w)$ for all $w \in \mathcal{R}^d \setminus v$.

The proof is simple. Suppose there exists $v \neq w$ such that

$$Q(v) = Q(w) = \inf_{u \in \mathcal{R}^d} Q(u).$$

If Q is strictly convex and $\lambda = 1/2$, then

$$Q(v/2 + w/2) < (Q(v) + Q(w))/2 = \inf_{u \in \mathcal{R}^d} Q(u),$$

Thus $z = (v + w)/2$ has the property that $Q(z) < \inf_{u \in \mathcal{R}^d} Q(u)$, a contradiction.

Convexity for quadratic functions

A general quadratic function can be written

$$Q(v) = v'Av + b'v + c$$

where A is a $d \times d$ matrix, b is a vector in \mathcal{R}^d , and c is a scalar.

Note that

$$v'Av = \sum_{i,j} v_i v_j A_{ij} \quad b'v = \sum_j b_j v_j.$$

Convexity for quadratic functions

If $b \in \text{col}(A)$, we can complete the square to get

$$Q(v) = (v - f)'A(v - f) + s.$$

where f is any vector satisfying $Af = -b/2$, and $s = c - f' Af$.

If A is invertible, we can take $f = -A^{-1}b/2$.

Convexity for quadratic functions

Since the property of being convex is invariant to translations in both the domain and range, without loss of generality we can assume $f = 0$ and $s = 0$ for purposes of analyzing the convexity of Q .

We will also need the following definition: a square matrix A is positive definite if $v'Av > 0$ for all vectors $v \neq 0$. The matrix is positive semidefinite if $v'Av \geq 0$ for all v .

We will now show that the quadratic function $Q(v) = v'Av$ is strictly convex if and only if A is positive definite (without loss of generality, A is symmetric, since otherwise $(A + A')/2$ gives the same quadratic form).

$$\begin{aligned} Q(\lambda v + (1 - \lambda)w) &= (\lambda v + (1 - \lambda)w)'A(\lambda v + (1 - \lambda)w) \\ &= \lambda^2 v'Av + (1 - \lambda)^2 w'Aw + 2\lambda(1 - \lambda)v'Aw. \end{aligned}$$

$$\begin{aligned} \lambda Q(v) + (1 - \lambda)Q(w) - Q(\lambda v + (1 - \lambda)w) &= \lambda(1 - \lambda)(v'Av + w'Aw - 2w'Av) \\ &= \lambda(1 - \lambda)(v - w)'A(v - w) \\ &\geq 0 \end{aligned}$$

This is a strict inequality for all $v \neq w$, $0 < \lambda < 1$ if and only if A is positive definite.

When A is symmetric, the Hessian of $Q(v) = v'Av$ is $2A$. To see this, write

$$v'Av = \sum_i v_i^2 A_{ii} + 2 \sum_{i < j} v_i v_j A_{ij}$$

and differentiate with respect to v_i , then v_j .

It follows that any quadratic function is strictly convex iff its Hessian is positive definite.

Properties of the simple least squares fit

- The least squares solution $\hat{\alpha}$, $\hat{\beta}$ is unique as long as $\widehat{\text{var}}(X)$ (the sample variance of the covariate) is positive. To see this, note that the Hessian (second derivative matrix) of $L(\alpha, \beta)$ is

$$H = \begin{pmatrix} \partial^2 L / \partial \alpha^2 & \partial^2 L / \partial \alpha \partial \beta \\ \partial^2 L / \partial \alpha \partial \beta & \partial^2 L / \partial \beta^2 \end{pmatrix} = \begin{pmatrix} 2n & 2X. \\ 2X. & 2 \sum X_i^2 \end{pmatrix}$$

where $X. = \sum_i X_i$.

If $\text{var}(X) > 0$ then this is a positive definite matrix since all the principle submatrices have non-negative determinants:

$$|2n| \geq 0$$

$$\begin{aligned} \begin{vmatrix} 2n & 2X. \\ 2X. & 2 \sum X_i^2 \end{vmatrix} &= 4n \sum X_i^2 - 4(X.)^2 \\ &= 4n(n-1)\widehat{\text{var}}(X) \end{aligned}$$

- The “fitted line” is

$$Y = \hat{\alpha} + \hat{\beta}X.$$

Substituting $X = \bar{X}$ yields \bar{Y} . Thus the fitted line always passes through the “center” of the data (\bar{X}, \bar{Y}) .

- The fitted slope can be written

$$\hat{\beta} = \widehat{\text{cor}}(Y, X) \frac{\widehat{\text{SD}}(Y)}{\widehat{\text{SD}}(X)},$$

so the fitted slope has the same sign as the Pearson correlation coefficient between Y and X .

- The residuals

$$r_i = Y_i - \hat{Y}_i.$$

are the vertical differences between the observed responses Y_i and the fitted values

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i.$$

By direct calculation,

$$\begin{aligned}\sum_i r_i &= Y. - n\hat{\alpha} - \hat{\beta}X. \\ &= Y. - (Y. - \hat{\beta}X.) - \hat{\beta}X. \\ &= 0.\end{aligned}$$

- If X and Y are reversed, the slope is

$$\hat{\beta}_* = \frac{\widehat{\text{cov}}(Y, X)}{\widehat{\text{var}}(Y)} = \widehat{\text{cor}}(Y, X) \frac{\widehat{\text{SD}}(X)}{\widehat{\text{SD}}(Y)}.$$

If the data fall exactly on a line, then $\text{cor}(Y, X) = 1$, so $\hat{\beta}_* = 1/\hat{\beta}$, which is consistent with algebraically rearranging

$$Y = \alpha + \beta X$$

to

$$X = -\alpha/\beta + Y/\beta.$$

But if the data do not fit a line exactly, this consistency does not hold.

Fitting multiple regression models

For multiple regression ($p > 1$), the covariate data define the “design matrix:”

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ & & & \cdots & \\ & & & \cdots & \\ & & & \cdots & \\ & & & \cdots & \\ & & & \cdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

Note that in some situations the first column of 1's (the intercept) will not be included.

The linear model coefficients are written as a vector

$$\beta = (\beta_0, \beta_1, \dots, \beta_p)'$$

where β_0 is the intercept and β_k is the slope corresponding to the k^{th} covariate. For a given working covariate vector β , the vector of fitted values is given by the matrix-vector product

$$\hat{Y} = \mathbf{X}\beta,$$

which is an n -dimensional vector.

The vector of residuals $Y - \mathbf{X}\beta$ is also an n -dimensional vector.

The goal for least-squares estimation is to minimize the sum of squared differences between the fitted and observed values.

$$L(\beta) = \sum_i (Y_i - \hat{Y}_i)^2 = \|Y - \mathbf{X}\beta\|^2.$$

The multivariate chain rule

Suppose $g(\cdot)$ is a map from \mathcal{R}^m to \mathcal{R}^n and $f(\cdot)$ is a scalar-valued function on \mathcal{R}^n . If $h = f \circ g$, i.e. $h(z) = f(g(z))$. Let $f_j(x) = \partial f(x)/\partial x_j$, let

$$\nabla f(x) = (f_1(x), \dots, f_n(x))'$$

denote the gradient of f , and let J denote the Jacobian of g

$$J_{ij}(z) = \partial g_i(z)/\partial z_j.$$

Then

$$\begin{aligned} \partial h(z)/\partial z_j &= \sum_i f_i(g(z)) \partial g_i(z)/\partial z_j \\ &= [J(z)' \nabla f(g(z))]_j \end{aligned}$$

Thus we can write the gradient of h as a matrix-vector product between the (transposed) Jacobian of g and the gradient of f :

$$\nabla h = J' \nabla f$$

where J is evaluated at z and ∇f is evaluated at $g(z)$.

For the least squares problem, the gradient of $L(\beta)$ with respect to β is

$$\partial L / \partial \beta = -2\mathbf{X}'(Y - \mathbf{X}\beta).$$

This can be seen by differentiating

$$L(\beta) = \sum_i (Y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j)^2$$

element-wise, or by differentiating

$$\|Y - \mathbf{X}\beta\|^2$$

using the multivariate chain rule, letting $g(\beta) = Y - \mathbf{X}\beta$ and $f(x) = \sum_j x_j^2$.

Setting $\partial L/\partial\beta = 0$ yields the “normal equations:”

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'Y$$

Thus calculating the least squares estimate of β reduces to solving a system of $p + 1$ linear equations. Algebraically we can write

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y,$$

which is often useful for deriving analytical results. However this expression should not be used to numerically calculate the coefficients.

The most standard numerical approach is to calculate the QR decomposition of

$$\mathbf{X} = \mathbf{QR}$$

where Q is a $n \times p + 1$ orthogonal matrix (i.e. $Q'Q = I$) and R is a $p + 1 \times p + 1$ upper triangular matrix.

The QR decomposition can be calculated rapidly, and highly precisely. Once it is obtained, the normal equations become

$$R\beta = Q'Y,$$

which is an easily solved $p + 1 \times p + 1$ triangular system.

Mathematical properties of the multiple regression fit

- The multiple least square solution is unique as long as the columns of \mathbf{X} are linearly independent. Here is the proof:
 1. The Hessian of $L(\beta)$ is $2\mathbf{X}'\mathbf{X}$.
 2. For $v \neq 0$, $v'(\mathbf{X}'\mathbf{X})v = (\mathbf{X}v)'\mathbf{X}v = \|\mathbf{X}v\|^2 > 0$, since the columns of \mathbf{X} are linearly independent. Therefore the Hessian of L is positive definite.
 3. Since $L(\beta)$ is quadratic with a positive definite Hessian matrix, it is convex and hence has a unique global minimizer.

Review of projections

Suppose S is a subspace of \mathcal{R}^d , and V is a vector in \mathcal{R}^d . The projection operator P_S maps V to the vector in S that is closest to V :

$$P_S(V) = \operatorname{argmin}_{\eta \in S} \|V - \eta\|^2.$$

Review of projections (continued)

Property 1: $(V - P_S(V))'s = 0$ for all $s \in S$. To see this, let $s \in S$. Without loss of generality $\|s\| = 1$ and $(V - P_S(V))'s \leq 0$. Let $\lambda \geq 0$, and write

$$\|V - P_S V + \lambda s\|^2 = \|V - P_S V\|^2 + \lambda^2 + 2\lambda(V - P_S(V))'s.$$

If $(V - P_S(V))'s \neq 0$, then for sufficiently small $\lambda > 0$, $\lambda^2 + 2\lambda(V - P_S(V))'s < 0$. This means that $P_S(V) - \lambda s$ is closer to V than $P_S(V)$, contradicting the definition of $P_S(V)$.

Review of projections (continued)

Property 2: Given a subspace S of \mathcal{R}^d , any vector $V \in \mathcal{R}^d$ can be written uniquely in the form $V = V_S + V_{S^\perp}$, where $V_S \in S$ and $s'V_{S^\perp} = 0$ for all $s \in S$. To prove uniqueness, suppose

$$V = V_S + V_{S^\perp} = \tilde{V}_S + \tilde{V}_{S^\perp}.$$

Then

$$\begin{aligned} 0 &= \|V_S - \tilde{V}_S + V_{S^\perp} - \tilde{V}_{S^\perp}\|^2 \\ &= \|V_S - \tilde{V}_S\|^2 + \|V_{S^\perp} - \tilde{V}_{S^\perp}\|^2 + 2(V_S - \tilde{V}_S)'(V_{S^\perp} - \tilde{V}_{S^\perp}) \\ &= \|V_S - \tilde{V}_S\|^2 + \|V_{S^\perp} - \tilde{V}_{S^\perp}\|^2. \end{aligned}$$

which is only possible if $V_S = \tilde{V}_S$ and $V_{S^\perp} = \tilde{V}_{S^\perp}$. Existence follows from Property 1, with $V_S = P_V(S)$ and $V_{S^\perp} = V - P_V(S)$.

Review of projections (continued)

Property 3: The projection $P_S(V)$ is unique.

This follows from properties 1 and 2.

Review of projections (continued)

Property 4: $P_S(P_S(V)) = P_S(V)$. The proof of this is simple, since $P_S(V) \in S$, and any element of S has zero distance to itself.

A matrix or linear map with this property is called “idempotent.”

Review of projections (continued)

Property 5: P_S is a linear operator.

Let A, B be vectors with $\theta_A = P_S(A)$ and $\theta_B = P_S(B)$. Then we can write

$$A + B = \theta_A + \theta_B + (A + B - \theta_A - \theta_B),$$

where $\theta_A + \theta_B \in S$, and

$$s'(A + B - \theta_A - \theta_B) = s'(A - \theta_A) + s'(B - \theta_B) = 0$$

for all $s \in S$. By Property 2 above, this representation is unique, so $\theta_A + \theta_B = P_S(A + B)$.

Property 6

Suppose P_S is the projection operator onto a subspace S . Then $I - P_S$, where I is the identity matrix, is the projection operator onto the subspace

$$S^\perp \equiv \{u \in \mathcal{R}^d \mid u's = 0 \text{ for all } s \in S\}.$$

To prove this, write

$$V = (I - P_S)V + P_S V,$$

and note that $((I - P_S)V)'s = 0$ for all $s \in S$, so $(I - P_S)V \in S^\perp$, and $u'P_S V = 0$ for all $u \in S^\perp$. By property 2 this decomposition is unique, and therefore $I - P_S$ is the projection operator onto S^\perp .

Review of projections (continued)

Property 7: Since $P_S(V)$ is linear, it can be represented in the form $P_S(V) = P_S \cdot V$ for a suitable square matrix P_S . Suppose S is spanned by the columns of a non-singular matrix \mathbf{X} . Then

$$P_S = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

To prove this, let $Q = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$$\begin{aligned} V &= QV + (V - QV) \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V + (I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')V \end{aligned}$$

and note that the first summand is in S .

To show that the second summand is in S^\perp , take $s \in S^\perp$ and write $s = \mathbf{X}b$. Then

$$\begin{aligned} s'(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')V &= b'\mathbf{X}'(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')V \\ &= 0. \end{aligned}$$

Therefore this is the unique decomposition from Property 2, above, so P_S must be the projection.

Property 7 (continued)

An alternate approach to property 7 is constructive. Let $\theta = \mathbf{X}\gamma$, and suppose we wish to minimize the distance between θ and V . Using calculus, we differentiate with respect to γ and solve for the stationary point:

$$\begin{aligned}\partial\|V - \mathbf{X}\gamma\|^2/\partial\gamma &= \partial(V'V - 2V'\mathbf{X}\gamma + \gamma'\mathbf{X}'\mathbf{X}\gamma)/\partial\gamma \\ &= -2\mathbf{X}'V + 2\mathbf{X}'\mathbf{X}\gamma \\ &= 0.\end{aligned}$$

The solution is

$$\gamma = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V$$

so

$$\theta = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V.$$

Review of projections (continued)

Property 8

$P_S \circ P_{S^\perp} = P_{S^\perp} \circ P_S \equiv 0$. This can be shown by direct calculation using the representations of P_S and P_{S^\perp} given above.

Least squares and projections

The least squares problem of minimizing

$$\|Y - X\beta\|^2$$

is equivalent to minimizing $\|Y - \eta\|^2$ over $\eta \in \text{col}(X)$.

Therefore the minimizing value $\hat{\eta}$ is the projection of Y onto $\text{col}(X)$.

If the columns of X are linearly independent, there is a unique vector $\hat{\beta}$ such that $X\hat{\beta} = \hat{\eta}$.

Properties of the least squares solution

- The fitted regression surface passes through the mean point $(\bar{\mathbf{X}}, \bar{Y})$. To see this note that the fitted surface at $\bar{\mathbf{X}}$ (the vector of column-wise means) is

$$\begin{aligned}\bar{\mathbf{X}}'\hat{\beta} &= (\mathbf{1}'\mathbf{X}/n)\hat{\beta} \\ &= \mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}/n,\end{aligned}$$

where $\mathbf{1}$ is a column vector of 1's. Since $\mathbf{1} \in \text{col}(\mathbf{X})$, it follows that

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{1} = \mathbf{1},$$

which gives the result, since $\bar{Y} = \mathbf{1}'\mathbf{Y}/n$.

Properties of the least squares solution

- The multiple regression residuals sum to zero. The multiple regression residuals are

$$\begin{aligned} R &\equiv Y - \hat{Y} \\ &= Y - P_S Y \\ &= (I - P_S)Y \\ &= P_{S^\perp} Y, \end{aligned}$$

where $S = \text{col}(X)$. The sum of residuals can be written

$$\mathbf{1}'(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')Y,$$

where $\mathbf{1}$ is a vector of 1's. If \mathbf{X} includes an intercept, $P_S \mathbf{1} = \mathbf{1}$, so

$$\mathbf{1}'I = \mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{1}',$$

so

$$\mathbf{1}'(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = 0.$$

Orthogonal matrices

An orthogonal matrix X satisfies $X'X = I$.

If X is square, then $X' = X^{-1}$ and also $XX' = I$.

If X is orthogonal then the projection onto $\text{col}(X)$ simplifies to

$$X(X'X)^{-1}X' = XX'$$

If X is orthogonal and the first column of X is constant, it follows that the remaining columns of X are centered and have sample variance $1/(n - 1)$.

- If \mathbf{X} is orthogonal, the slopes obtained by using multiple regression of Y on $X = (X_1, \dots, X_p)$ are the same as the slopes obtained by carrying out p simple linear regressions of Y on each covariate separately.

To see this, note that the multiple regression slope estimate for the i^{th} covariate is

$$\hat{\beta}_{m,i} = \mathbf{X}'_{:i} Y$$

where $\mathbf{X}_{:i}$ is column i of \mathbf{X} . Since $\mathbf{X}'\mathbf{X} = I$ it follows that each covariate has zero sample mean, and sample variance equal to $1/(n - 1)$. Thus the simple linear regression slope for covariate i is

$$\hat{\beta}_i = \widehat{\text{cov}}(\mathbf{X}_{:i}, Y) / \widehat{\text{var}}(\mathbf{X}_{:i}) = \mathbf{X}'_{:i} Y = \hat{\beta}_{m,i}.$$

- The signs of the multiple regression slopes need not agree with the signs of the corresponding simple regression slopes. For example, suppose there are two covariates, both with mean zero and variance 1, and for simplicity assume that Y has mean zero and variance 1. Let r_{12} be the correlation between the two covariates, and let r_{1y} and r_{2y} be the correlations between each covariate and the response. It follows that

$$\mathbf{X}'\mathbf{X}/(n-1) = \begin{pmatrix} n/(n-1) & 0 & 0 \\ 0 & 1 & r_{12} \\ 0 & r_{12} & 1 \end{pmatrix}$$

$$\mathbf{X}'Y/(n-1) = \begin{pmatrix} 0 \\ r_{1y} \\ r_{2y} \end{pmatrix}.$$

So we can write

$$\hat{\beta} = (\mathbf{X}'\mathbf{X}/(n-1))^{-1}(\mathbf{X}'Y/(n-1)) = \frac{1}{1-r_{12}^2} \begin{pmatrix} 0 \\ r_{1y} - r_{12}r_{2y} \\ r_{2y} - r_{12}r_{1y} \end{pmatrix}.$$

Thus, for example, if $r_{2y} = cr_{1y}$ for $c > 1$, and $r_{1y}, r_{2y}, r_{12} \geq 0$, then if $r_{12} > 1/c$, $\hat{\beta}_1$ has opposite signs in single and multiple regression. Note that the sign of the covariate that is more strongly correlated with Y cannot be reversed.

Regression model formulations and parameterizations

A very general regression model is

$$Y = f(X, \epsilon)$$

where ϵ is unobserved “error” with expected value zero.

A more restrictive “additive error” model is:

$$Y = f(X) + \epsilon.$$

Under this model,

$$\begin{aligned} E(Y|X) &= E(f(X)|X) + E(\epsilon|X) \\ &= f(X) + E(\epsilon|X). \end{aligned}$$

If we assume that $E(\epsilon|X) = 0$ (which would follow if (i) $E\epsilon = 0$ and (ii) ϵ and X are independent random variables), then $E(Y|X) = f(X)$.

A parametric regression model is:

$$Y = f(X; \theta) + \epsilon,$$

where θ is finite dimensional and the mapping $X \rightarrow f(X; \theta)$ is determined by θ .

Examples:

1. The linear response surface model $f(X; \theta) = \theta'X$
2. The quadratic response surface model $f(X; \theta) = \theta_1 + \theta_2X + \theta_3X^2$
3. The Gompertz curve $f(X; \theta) = \theta_1 \exp(\theta_2 \exp(\theta_3X))$ $\theta_2, \theta_3 \leq 0$.

Models 1 and 2 are both “linear models” because they are linear in θ (although 2 is not linear in X). The Gompertz curve is a “non-linear” model because it is not linear in θ .

Basic Inference for Least Squares Fits

This section deals with statistical properties of least square fits that can be derived under minimal modeling assumptions.

Specifically, we will assume that $Y = X\beta + \epsilon$, where:

- i** $E(\epsilon|X) = 0$
- ii** $\text{var}(\epsilon|X)$ exists and is constant
- iii** the ϵ random variables are uncorrelated across cases. We will not assume here that ϵ follows a particular parametric distribution, e.g. a Gaussian distribution.

Note about the meaning of $E(\epsilon|X)$ and $\text{var}(\epsilon|X)$.

When X is not a random variable, $E(\epsilon|X)$ means “the expected value of ϵ for a given value of X ,” and similarly $\text{var}(\epsilon|X)$ means “the variance of ϵ for a given value of X .”

When X is a random variable, the condition that

$$E(\epsilon|X) = 0$$

for all X implies that $\text{cor}(X, \epsilon) = 0$. This follows from the double expectation theorem:

$$\begin{aligned}
\text{cov}(X, \epsilon) &= EX\epsilon - EX \cdot E\epsilon \\
&= EX\epsilon \\
&= E_X E(\epsilon X | X) \\
&= E_X X E(\epsilon | X) \\
&= E_X X \cdot 0 \\
&= 0.
\end{aligned}$$

Note that $\text{cor}(X, \epsilon) = 0$ and $E\epsilon = 0$ do not imply that $E(\epsilon|X) = 0$ in general. For example, if $X \in \{-1, 0, 1\}$ and $\epsilon \in \{-1, 1\}$, with joint distribution

	-1	1
-1	1/12	3/12
0	4/12	0
1	1/12	3/12

then $E\epsilon = 0$ and $\text{cor}(X, \epsilon) = 0$, but $E(\epsilon|X)$ is not identically zero. When ϵ and X are jointly Gaussian, $\text{cor}(\epsilon, X) = 0$ implies that ϵ and X are independent, which in turn implies that $E(\epsilon|X) = E\epsilon = 0$.

To be able to interpret the results of a regression analysis, it is very important to consider how the data were sampled. In particular, it is important to consider whether X and/or Y and/or the pair X, Y should be considered as random draws from a population. Here are some typical situations:

- **Designed experiment:** We are studying the effect of temperature on reaction yield in a chemical synthesis. The temperature (X) is controlled by the experimenter. In this case it doesn't make sense to consider the X value to be a random variable.
- **Observational study:** We are interested in the relationship between cholesterol level (X), and blood pressure (Y). Suppose we sample people at random from a well defined population (e.g. all registered nurses in the state of Michigan) and measure their blood pressure and cholesterol levels. In this case X and Y are both random variables.

- Case/control study: Again suppose we are interested in the relationship between cholesterol level (X), and blood pressure (Y). But now suppose we have an exhaustive list of blood pressure measurements for all nurses in the state of Michigan. We wish to select a practical number of individuals (say 500) to contact for acquiring cholesterol measures, and it is decided that studying the 250 nurses with greatest blood pressure together with the 250 nurses with least blood pressure will be most informative. In this case both X is randomly sampled conditionally on Y , but Y is not randomly sampled.

Summary: Regression models are formulated in terms of the conditional distribution of Y given X . The statistical properties of $\hat{\beta}$ are easiest to calculate and interpret as being conditional on X . The way that the data were sampled affects our interpretation of the results.

Basic inference for simple least squares

Above we showed that the slope and intercept estimates are

$$\hat{\beta} = \frac{\widehat{\text{cov}}(Y, X)}{\widehat{\text{var}}(X)} = \frac{\sum_i Y_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2},$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

Note that we are using the useful fact that

$$\sum_i (Y_i - \bar{Y})(X_i - \bar{X}) = \sum_i Y_i (X_i - \bar{X}) = \sum_i (Y_i - \bar{Y})X_i.$$

First we will calculate the *sampling means* of $\hat{\alpha}$ and $\hat{\beta}$. A useful identity is that

$$\begin{aligned}\hat{\beta} &= \sum_i (\alpha + \beta X_i + \epsilon_i)(X_i - \bar{X}) / \sum_i (X_i - \bar{X})^2 \\ &= \beta + \sum_i \epsilon_i (X_i - \bar{X}) / \sum_i (X_i - \bar{X})^2.\end{aligned}$$

From this identity it is clear that $E\hat{\beta} = \beta$. Thus $\hat{\beta}$ is unbiased (a parameter estimate is unbiased if its sampling mean is the same as the population value of the parameter).

Continuing with the intercept:

$$\begin{aligned}E\hat{\alpha} &= E(\bar{Y} - \hat{\beta}\bar{X}) \\ &= \alpha + \beta\bar{X} + E\bar{\epsilon} - E\hat{\beta}\bar{X} \\ &= \alpha\end{aligned}$$

Thus $\hat{\alpha}$ is also unbiased.

Next we will calculate the *sampling variances and covariance* of $\hat{\alpha}$ and $\hat{\beta}$. These capture the variability of the parameter estimates over replicated studies or experiments from the same population.

First we will need the following result:

$$\begin{aligned}\text{cov}(\hat{\beta}, \bar{\epsilon}) &= \sum_i \text{cov}(\epsilon_i, \bar{\epsilon})(X_i - \bar{X}) / \sum_i (X_i - \bar{X})^2 \\ &= 0,\end{aligned}$$

since $\text{cov}(\epsilon_i, \bar{\epsilon}) = \sigma^2/n$ does not depend on i .

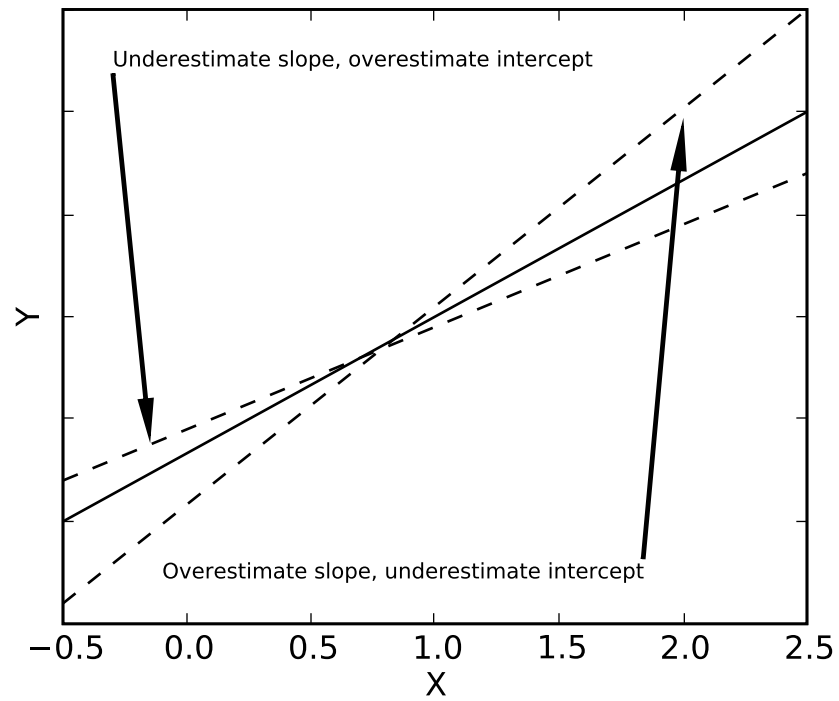
To derive the sampling variances and covariance, start with the “useful identity” above.

$$\begin{aligned}\text{var}(\hat{\beta}) &= \sigma^2 / \sum_i (X_i - \bar{X})^2 \\ &= \sigma^2 / ((n - 1)\text{var}(X)).\end{aligned}$$

$$\begin{aligned}\text{var}(\hat{\alpha}) &= \text{var}(\bar{Y} - \hat{\beta}\bar{X}) \\ &= \text{var}(\alpha + \beta\bar{X} + \bar{\epsilon} - \hat{\beta}\bar{X}) \\ &= \text{var}(\bar{\epsilon}) + \bar{X}^2\text{var}(\hat{\beta}) - 2\bar{X}\text{cov}(\bar{\epsilon}, \hat{\beta}) \\ &= \sigma^2/n + \bar{X}^2\sigma^2/((n - 1)\text{var}(X)).\end{aligned}$$

$$\begin{aligned}\text{cov}(\hat{\alpha}, \hat{\beta}) &= \text{cov}(\bar{Y}, \hat{\beta}) - \bar{X}\text{var}(\hat{\beta}) \\ &= -\sigma^2\bar{X}/((n - 1)\text{var}(X)).\end{aligned}$$

When $\bar{X} > 0$, it's easy to see what the expression for $\text{cov}(\hat{\alpha}, \hat{\beta})$ is telling us:



Some observations:

- All variances scale with sample size like $1/n$.
- $\hat{\beta}$ does not depend on \bar{X} .
- $\text{var}(\hat{\alpha})$ is minimized if $\bar{X} = 0$.
- $\hat{\alpha}$ and $\hat{\beta}$ are uncorrelated if $\bar{X} = 0$.

Some properties of residuals

Start with the following useful expression:

$$\begin{aligned} R_i &\equiv Y_i - \hat{\alpha} - \hat{\beta}X_i \\ &= Y_i - \bar{Y} - \hat{\beta}(X_i - \bar{X}). \end{aligned}$$

Since

$$Y_i - \bar{Y} = \beta(X_i - \bar{X}) + \epsilon_i - \bar{\epsilon}$$

it follows that

$$\begin{aligned} R_i &= (\beta - \hat{\beta})(X_i - \bar{X}) + \epsilon_i - \bar{\epsilon} \\ &\sim \text{estimation error} + \text{observation error.} \end{aligned}$$

- The “estimation error” and “observation error” have mean zero, therefore $ER_i = 0$. Note that this is a distinct property from the fact that $\sum_i R_i = 0$, which requires no statistical assumptions about the model.

The “estimation error” and “observation error” are negatively correlated:

$$\begin{aligned}\text{cov}((\beta - \hat{\beta})(X_i - \bar{X}), \epsilon_i - \bar{\epsilon}) &= -(X_i - \bar{X})\text{cov}(\hat{\beta}, \epsilon_i) \\ &= -\sigma^2(X_i - \bar{X})^2 / \sum_j (X_j - \bar{X})^2,\end{aligned}$$

thus the “estimation error” and “observation error” partially cancel each other out, particularly for the cases with relatively larger values of $(X_i - \bar{X})^2$.

Moreover, $\epsilon_i - \bar{\epsilon}$ tends to be smaller than ϵ_i , since

$$\begin{aligned}\text{var}(\epsilon_i - \bar{\epsilon}) &= \text{var}(\epsilon_i) + \text{var}(\bar{\epsilon}) - 2\text{cov}(\epsilon_i, \bar{\epsilon}) \\ &= \sigma^2 + \sigma^2/n - 2\sigma^2/n \\ &= \sigma^2(n - 1)/n.\end{aligned}$$

These are symptoms of *overfitting* – the residuals R_i tend to be smaller than the actual errors ϵ_i .

Overfitting is also reflected in the variance of the residuals:

$$\begin{aligned}\text{var}(R_i) &= \text{var}((\beta - \hat{\beta})(X_i - \bar{X}) + \epsilon_i - \bar{\epsilon}) \\ &= (X_i - \bar{X})^2 \text{var}(\hat{\beta}) + \text{var}(\epsilon_i - \bar{\epsilon}) - 2\sigma^2(X_i - \bar{X})^2 / \sum_j (X_j - \bar{X})^2 \\ &= (n-1)\sigma^2/n - \sigma^2(X_i - \bar{X})^2 / \sum_j (X_j - \bar{X})^2.\end{aligned}$$

Note the following identity, which ensures that this expression is non-negative:

$$(X_i - \bar{X})^2 / \sum_j (X_j - \bar{X})^2 \leq (n-1)/n.$$

Since $ER_i = 0$ it follows that $\text{var}(R_i) = ER_i^2$. Therefore the expected value of $\sum_i R_i^2$ is

$$(n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2$$

since the $(X_i - \bar{X})^2 / \sum_j (X_j - \bar{X})^2$ sum to one.

Estimating $\sigma^2 = \text{var}(Y|X) = \text{var}(\epsilon)$

Since

$$E \sum_i R_i^2 = (n - 2)\sigma^2$$

it follows that

$$\sum_i R_i^2 / (n - 2)$$

is an unbiased estimate of σ^2 .

Basic inference for multiple least squares

We have the following useful identity for the multiple least squares fit:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + \epsilon) \\ &= \beta + (X'X)^{-1}X'\epsilon.\end{aligned}$$

Letting

$$\eta = (X'X)^{-1}X'\epsilon$$

we see that $E(\eta|X) = 0$ and

$$\begin{aligned}\text{var}(\eta|X) = \text{var}(\hat{\beta}|X) &= (X'X)^{-1}X'\text{var}(\epsilon|X)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\end{aligned}$$

- $\text{var}(\epsilon|X) = \sigma^2I$ by the assumptions that (i) the ϵ_i are uncorrelated and (ii) the ϵ_i have constant variance given X .

The residual sum of squares

The residual sum of squares (RSS) is the squared norm of the residual vector:

$$\begin{aligned}\text{RSS} &= \|Y - \hat{Y}\|^2 \\ &= \|Y - PY\|^2 \\ &= \|(I - P)Y\|^2 \\ &= Y'(I - P)(I - P)Y \\ &= Y'(I - P)Y,\end{aligned}$$

where P is the projection matrix onto $\text{col}(X)$. The last equivalence follows from the fact that $I - P$ is a projection and hence is idempotent.

The expected value of the RSS

The expression $RSS = Y'(I - P)Y$ is a quadratic form in Y , and we can write

$$Y'(I - P)Y = \text{tr}(Y'(I - P)Y) = \text{tr}((I - P)YY'),$$

where the second equality uses the circulant property of the trace. For three factors, the circulant property means that

$$\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA).$$

By linearity we have

$$E\text{tr}((I - P)YY') = \text{tr}((I - P) \cdot EYY'),$$

and

$$\begin{aligned} EYY' &= EX\beta\beta'X' + EX\beta\epsilon' + E\epsilon\beta'X' + E\epsilon\epsilon' \\ &= X\beta\beta'X' + E\epsilon\epsilon' \\ &= X\beta\beta'X' + \sigma^2I. \end{aligned}$$

Since $PX = X$ and hence $(I - P)X = 0$,

$$(I - P)EYY' = \sigma^2(I - P).$$

Therefore the expected value of the RSS is

$$ERSS = \sigma^2\text{tr}(I - P).$$

Four more properties of projection matrices

Property 9: A projection matrix P is symmetric. One way to show this is to let V_1, \dots, V_q be an orthonormal basis for S , where P is the projection onto S . Then complete the V_j with V_{q+1}, \dots, V_d to get a basis. By direct calculation,

$$(P - \sum_{j=1}^q V_j V_j') V_k = 0$$

for all k , hence $P = \sum_{j=1}^q V_j V_j'$ which is symmetric.

Four more properties of projection matrices (continued)

Property 10: A projection matrix is positive semidefinite. Let V be an arbitrary vector and write $V = V_1 + V_2$, where $V_1 \in S$ and $V_2 \in S^\perp$. Then

$$(V_1 + V_2)'P(V_1 + V_2) = V_1'V_1 \geq 0.$$

Four more properties of projection matrices (continued)

Property 11: The eigenvalues of a projection matrix P must be zero or one.

Suppose λ, v is an eigenvalue/eigenvector pair:

$$Pv = \lambda v.$$

If P is the projection onto a subspace S , this implies that λv is the closest element of S to v . But if $\lambda v \in S$ then $v \in S$, and is strictly closer to v than λv , unless $\lambda = 1$ or $v = 0$. Therefore only 0 and 1 can be eigenvalues of P .

Four more properties of projection matrices (continued)

Property 12: The trace of a projection matrix is its rank.

The rank of a matrix is the number of nonzero eigenvalues. The trace of a matrix is the sum of all eigenvalues. Since the nonzero eigenvalues of a projection matrix are all 1, the rank and the trace must be identical.

The expected value of the RSS (continued)

We know that $ERSS = \sigma^2 \text{tr}(I - P)$. Since $I - P$ is the projection onto $\text{col}(X)^\perp$, $I - P$ has rank $n - \text{rank}(X) = n - p - 1$. Thus

$$ERSS = \sigma^2(n - p - 1),$$

so

$$RSS/(n - p - 1)$$

is an unbiased estimate of σ^2 .

Covariance matrix of residuals

Since $E\hat{\beta} = \beta$, it follows that $E\hat{Y} = X\beta = EY$. Therefore we can derive the following simple expression for the covariance matrix of the residuals.

$$\begin{aligned}\text{cov}(Y - \hat{Y}) &= E(Y - \hat{Y})(Y - \hat{Y})' \\ &= (I - P)EYY'(I - P) \\ &= (I - P)(X\beta\beta'X' + \sigma^2I)(I - P) \\ &= \sigma^2(I - P)\end{aligned}$$

Variance and distribution of the RSS

The RSS can be written

$$\begin{aligned}\text{RSS} &= \text{tr}((I - P)YY') \\ &= \text{tr}((I - P)\epsilon\epsilon')\end{aligned}$$

Therefore, the distribution of the RSS does not depend on β . It also depends on X only through $\text{col}(X)$.

If the distribution of ϵ is invariant under orthogonal transforms, i.e.

$$\epsilon \stackrel{d}{=} Q\epsilon$$

when Q is a square orthogonal matrix, then we can make the stronger statement that the distribution of the RSS only depends on X through its rank.

To see this, construct a square orthogonal matrix Q so that $Q'(I - P)Q$ is the projection onto a fixed subspace \mathcal{S} of dimension $n - p - 1$ (so Q' maps $\text{col}(I - P)$ to \mathcal{S}). Then

$$\begin{aligned}\text{tr}((I - P)\epsilon\epsilon') &= \text{tr}((I - P)Q\epsilon(Q\epsilon)') \\ &= \text{tr}(Q'(I - P)Q\epsilon\epsilon')\end{aligned}$$

Note that since Q is square we have $QQ' = Q'Q = I$.

Optimality

For a given design matrix X , there are many linear estimators that are unbiased for β . That is, there are many matrices $M \in \mathcal{R}^{p+1 \times n}$ such that

$$EMY = \beta$$

for all β . The Gauss-Markov theorem states that among these, the least squares estimate is “best,” in the sense that its covariance matrix is “smallest.”

Here we are using the definition that a matrix A is “smaller” than a matrix B if

$$B - A$$

is positive definite.

Letting $\beta^* = MY$ be any linear unbiased estimator of β , when

$$\text{cov}\hat{\beta} \leq \text{cov}\beta^*,$$

this implies that for any fixed vector θ ,

$$\text{var}(\theta'\hat{\beta}) \leq \text{var}(\theta'\beta^*).$$

The Gauss-Markov theorem implies that the least squares estimate $\hat{\beta}$ is the “BLUE” (best linear unbiased estimator) for the least squares model.

The idea of the proof is to show that for any linear unbiased estimate β^* of β , $\beta^* - \hat{\beta}$ and $\hat{\beta}$ are uncorrelated. It follows that

$$\begin{aligned}\text{cov}(\beta^*) &= \text{cov}(\beta^* - \hat{\beta} + \hat{\beta}) \\ &= \text{cov}(\beta^* - \hat{\beta}) + \text{cov}(\hat{\beta}) \\ &\geq \text{cov}(\hat{\beta}).\end{aligned}$$

To prove the theorem note that

$$E\beta^* = M(EY) = MX\beta = \beta$$

for all β , and let $B = (X'X)^{-1}X'$, so that

$$E\hat{\beta} = BX\beta = \beta$$

for all β , so

$$(M - B)X \equiv 0.$$

Therefore

$$\begin{aligned}\text{cov}(\beta^* - \hat{\beta}, \hat{\beta}) &= E(M - B)Y(BY - \beta)' \\ &= (M - B)(X\beta\beta'X' + \sigma^2I)B' - (M - B)X\beta\beta' \\ &= \sigma^2(M - B)B' \\ &= 0.\end{aligned}$$

Note that we have an explicit expression for the gap between $\text{cov}(\hat{\beta})$ and $\text{cov}(\beta^*)$:

$$\text{cov}(\beta^* - \hat{\beta}) = \sigma^2(M - B)(M - B)'.$$

Regression inference with Gaussian errors

The multivariate normal distribution

The random vector $X = (X_1, \dots, X_p)'$ has a p -dimensional standard multivariate normal distribution if its components are independent and standard normal.

The density of X is the product of p standard normal densities:

$$p(X) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2} \sum_j X_j^2\right) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2} X'X\right).$$

If we transform

$$Z = \mu + AX,$$

we get a random variable satisfying

$$\begin{aligned} EZ &= \mu \\ \text{cov}(Z) &= A\text{cov}(X)A' = AA' \equiv \Sigma. \end{aligned}$$

The density of Z can be obtained using the change of variables formula:

$$\begin{aligned} p(Z) &= (2\pi)^{-p/2} |A^{-1}| \exp\left(-\frac{1}{2}(Z - \mu)' A^{-T} A^{-1} (Z - \mu)\right) \\ &= (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(Z - \mu)' \Sigma^{-1} (Z - \mu)\right) \end{aligned}$$

This distribution is denoted $N(\mu, \Sigma)$. The log-density is

$$-\frac{1}{2} \log |\Sigma| - \frac{1}{2} (Z - \mu)' \Sigma^{-1} (Z - \mu),$$

with the constant term dropped.

The joint log-density for an *iid* sample of size n is

$$-\frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_i (X_i - \mu)' \Sigma^{-1} (X_i - \mu) = -\frac{n}{2} \log |\Sigma| - \frac{n}{2} \text{tr} (S_{xx} \Sigma^{-1})$$

where

$$S_{xx} = \sum (X_i - \mu)(X_i - \mu)' / n.$$

The Cholesky decomposition

If Σ is a non-singular covariance matrix, there is a lower triangular matrix A with positive diagonal elements such that

$$AA' = \Sigma$$

This matrix can be denoted $\Sigma^{1/2}$, and is called the “Cholesky square root.”

Properties of the multivariate normal distribution:

- A linear function of a multivariate normal random vector is also multivariate normal. Specifically, if

$$X \sim N(\mu, \Sigma)$$

is p -variate normal, and θ is a $q \times p$ matrix with $q \leq p$, then $Y \sim \theta X$ has a

$$N(\theta\mu, \theta\Sigma\theta')$$

distribution.

To prove this fact, write

$$X = \mu + AZ$$

where $AA' = \Sigma$ is the Cholesky decomposition.

Next, extend θ to a square invertible matrix

$$\tilde{\theta} = \begin{pmatrix} \theta \\ \theta^* \end{pmatrix}.$$

where $\theta^* \in \mathcal{R}^{p-q \times p}$.

The matrix θ^* can be chosen such that

$$\theta \Sigma \theta^{*'} = 0,$$

by the Gram-Schmidt procedure. Let

$$\tilde{Y} = \tilde{\theta} X = \begin{pmatrix} Y \\ Y^* \end{pmatrix} = \begin{pmatrix} \theta \mu + \theta AZ \\ \theta^* \mu + \theta^* AZ \end{pmatrix}.$$

Therefore

$$\text{cov}(\tilde{Y}) = \begin{pmatrix} \theta \Sigma \theta' & 0 \\ 0 & \theta^* \Sigma \theta^{*'} \end{pmatrix},$$

and

$$\text{cov}(\tilde{Y})^{-1} = \begin{pmatrix} (\theta \Sigma \theta')^{-1} & 0 \\ 0 & (\theta^* \Sigma \theta^{*'})^{-1} \end{pmatrix},$$

Using the change of variables formula, and the structure of the multivariate normal density, it follows that

$$p(\tilde{Y}) = p(Y)p(Y^*).$$

This implies that Y and Y^* are independent, and by inspecting the form of their densities, both are seen to be multivariate normal.

- A consequence of the above argument is that in general, uncorrelated components of a multivariate normal vector are independent.
- If X is a standard multivariate normal vector, and Q is a square orthogonal matrix, then QX is also standard multivariate normal. This follows directly from the fact that $QQ' = I$.

The χ^2 distribution

If z is a standard normal random variable, the density of z^2 can be calculated directly as

$$p(z) = z^{-1/2} \exp(-z/2) / \sqrt{2\pi}.$$

This is the χ_1^2 distribution. The χ_p^2 distribution is defined to be the distribution of

$$\sum_{j=1}^p z_j^2$$

where z_1, \dots, z_p are iid standard normal random variables.

By direct calculation, if $F \sim \chi_1^2$,

$$EF = 1 \quad \text{var}(F) = 2.$$

Therefore the mean of the χ_p^2 distribution is p and the variance is $2p$.

The χ_p^2 density is

$$p(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} \exp(-x/2).$$

To prove this by induction, assume that the χ_{p-1}^2 density is

$$\frac{1}{\Gamma((p-1)/2)2^{(p-1)/2}} x^{(p-1)/2-1} \exp(-x/2).$$

The χ_p^2 density is the density of the sum of a χ_{p-1}^2 random variable and a χ_1^2 random variable, which can be written

$$\begin{aligned}
& \frac{1}{\Gamma((p-1)/2)\Gamma(1/2)2^{p/2}} \int_0^x s^{(p-1)/2-1} \exp(-s/2)(x-s)^{-1/2} \exp(-(x-s)/2) ds = \\
& \frac{1}{\Gamma((p-1)/2)\Gamma(1/2)2^{p/2}} \exp(-x/2) \int_0^x s^{(p-1)/2-1} (x-s)^{-1/2} ds = \\
& \frac{1}{\Gamma((p-1)/2)\Gamma(1/2)2^{p/2}} \exp(-x/2) x^{p/2-1} \int_0^1 u^{p/2-3/2} (1-u)^{-1/2} du.
\end{aligned}$$

where $u = s/x$.

The density for χ_p^2 can now be obtained by applying the identities

$$\begin{aligned}
\Gamma(1/2) &= \sqrt{\pi} \\
\int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du &= \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta).
\end{aligned}$$

The χ^2 distribution and the RSS

Let P be a projection matrix and Z be a *iid* vector of standard normal values. For any square orthogonal matrix Q ,

$$Z'PZ = (QZ)'QPQ'(QZ).$$

Since QZ is equal in distribution to Z , $Z'PZ$ is equal in distribution to

$$Z'QPQ'Z.$$

If the rank of P is k , we can choose Q so that QPQ' is the projection onto the first k canonical basis vectors.

This gives us

$$Z'PZ \stackrel{d}{=} Z'QPQ'Z = \sum_{j=1}^k Z_j^2$$

which follows a χ_k^2 distribution.

It follows that

$$\begin{aligned} \frac{n-p-1}{\sigma^2} \hat{\sigma}^2 &= Y'(I-P)Y/\sigma^2 \\ &= (\epsilon/\sigma)'(I-P)(\epsilon/\sigma) \\ &\sim \chi_{n-p-1}^2. \end{aligned}$$

Thus when the errors are Gaussian, we have

$$\text{var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n - p - 1}.$$

The t distribution

Suppose Z is standard normal, $V \sim \chi_p^2$, and V is independent of Z . Then

$$T = \sqrt{p}Z/\sqrt{V}$$

has a “ t distribution with p degrees of freedom.”

Note that by the law of large numbers, V/p converges almost surely to 1. Therefore T converges in distribution to a standard normal distribution.

To derive the t density apply the change of variables formula. Let

$$\begin{pmatrix} U \\ W \end{pmatrix} \equiv \begin{pmatrix} Z/\sqrt{V} \\ V \end{pmatrix}$$

The Jacobian is

$$J = \begin{vmatrix} 1/\sqrt{V} & -Z/2V^{3/2} \\ 0 & 1 \end{vmatrix} = V^{-1/2} = W^{-1/2}.$$

Since the joint density of Z and V is

$$p(Z, V) \propto \exp(-Z^2/2)V^{p/2-1} \exp(-V/2),$$

it follows that the joint density of U and W is

$$\begin{aligned} p(U, W) &\propto \exp(-U^2W/2)W^{p/2-1/2} \exp(-W/2) \\ &= \exp(-W(U^2 + 1)/2)W^{p/2-1/2}. \end{aligned}$$

Therefore

$$\begin{aligned} p(U) &= \int p(U, W) dW \\ &\propto \int \exp(-W(U^2 + 1)/2) W^{p/2-1/2} dW \\ &= (U^2 + 1)^{-p/2-1/2} \int \exp(-Y/2) Y^{p/2-1/2} dY \\ &\propto (U^2 + 1)^{-p/2-1/2}. \end{aligned}$$

where

$$Y = W(U^2 + 1).$$

Finally, write $T = \sqrt{p}U$ to get that

$$p(T) \propto (T^2/p + 1)^{-(p+1)/2}.$$

Confidence intervals

The linear model residuals are

$$R \equiv Y - \hat{Y} = (I - P)Y$$

and the estimated coefficients are

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Recalling that $(I - P)X = 0$, and $E(Y - \hat{Y}) = 0$, it follows that

$$\begin{aligned}\text{cov}(Y - \hat{Y}, \hat{\beta}) &= E(Y - \hat{Y})\hat{\beta}' \\ &= E(I - P)YY'X(X'X)^{-1} \\ &= (I - P)(X\beta\beta'X' + \sigma^2I)X(X'X)^{-1} \\ &= 0.\end{aligned}$$

Therefore every estimated coefficient is uncorrelated with every residual. If ϵ is Gaussian, they are also independent.

Since $\hat{\sigma}^2$ is a function of the residuals, it follows that if ϵ is Gaussian, then $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

Confidence interval for a regression coefficient

Let

$$V_k = [(X'X)^{-1}]_{kk}$$

so that

$$\text{var}\hat{\beta}_k = \sigma^2 V_k.$$

If the ϵ are multivariate Gaussian $N(0, \sigma^2 I)$, then

$$(\hat{\beta}_k - \beta_k) / \sigma \sqrt{V_k} \sim N(0, 1).$$

Therefore

$$(n - p - 1)\hat{\sigma}^2/\sigma^2 \sim \chi_{n-p-1}^2$$

and using the fact that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent, it follows that

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 V_k}}$$

has a t-distribution with $n - p - 1$ degrees of freedom.

Therefore if $Q_T(q, k)$ is the q^{th} quantile of the t-distribution with k degrees of freedom, then for $0 \leq \alpha \leq 1$, and $q_\alpha = Q_T(1 - (1 - \alpha)/2, n - p - 1)$,

$$P\left(-q_\alpha \leq (\hat{\beta}_k - \beta_k) / \sqrt{\hat{\sigma}^2 V_k} \leq q_\alpha\right) = \alpha.$$

Rearranging terms we get the confidence interval

$$P\left(\hat{\beta}_k - \sqrt{\hat{\sigma}^2 V_k} q_\alpha \leq \beta_k \leq \hat{\beta}_k + \sqrt{\hat{\sigma}^2 V_k} q_\alpha\right) = \alpha$$

which has coverage probability α .

Confidence interval for the expected response

Let x^* be any point in \mathcal{R}^{p+1} . The expected response at $X = x^*$ is

$$E(Y|X = x^*) = \beta'x^*.$$

A point estimate for this value is

$$\hat{\beta}'x^*,$$

which is unbiased since $\hat{\beta}$ is unbiased, and has variance

$$\text{var}(\hat{\beta}'x^*) = \sigma^2 x^{*'}(X'X)^{-1}x^* \equiv \sigma^2 V_{x^*}.$$

As above we have that

$$\frac{\widehat{\beta}'x^* - \beta'x^*}{\sqrt{\widehat{\sigma}^2 V_{x^*}}}$$

has a t -distribution with $n - p - 1$ degrees of freedom.

Therefore

$$P(\widehat{\beta}'x^* - q_\alpha \sqrt{\widehat{\sigma}^2 V_{x^*}} \leq \beta'x^* \leq \widehat{\beta}'x^* + q_\alpha \sqrt{\widehat{\sigma}^2 V_{x^*}}) = \alpha$$

defines a CI for $E(Y|X = x^*)$ with coverage probability α .

Prediction intervals

Suppose Y^* is a new observation at $X = x^*$, independent of the data used to estimate $\hat{\beta}$ and $\hat{\sigma}^2$. If the errors are Gaussian, then $Y^* - \hat{\beta}'x^*$ is Gaussian, with the following mean and variance:

$$E(Y^* - \hat{\beta}'x^*) = \beta'x^* - \beta'x^* = 0$$

and

$$\text{var}(Y^* - \hat{\beta}'x^*) = \sigma^2(1 + V_{x^*}),$$

It follows that

$$\frac{Y^* - \hat{\beta}'x^*}{\sqrt{\hat{\sigma}^2(1 + V_{x^*})}}$$

has a t -distribution with $n - p - 1$ degrees of freedom. Therefore a prediction interval at x^* with coverage probability α is defined by

$$P\left(\hat{\beta}'x^* - q_\alpha\sqrt{\hat{\sigma}^2(1 + V_{x^*})} \leq Y^* \leq \hat{\beta}'x^* + q_\alpha\sqrt{\hat{\sigma}^2(1 + V_{x^*})}\right) = \alpha.$$

The F distribution

If $U \sim \chi_p^2$ and $V \sim \chi_q^2$, then

$$\frac{U/p}{V/q}$$

has an “F-distribution with p, q degrees of freedom,” denoted $F_{p,q}$.

To derive the kernel of the density, let

$$\begin{pmatrix} X \\ Y \end{pmatrix} \equiv \begin{pmatrix} U/V \\ V \end{pmatrix}.$$

The Jacobian of the map is

$$\begin{vmatrix} 1/V & -U/V^2 \\ 0 & 1 \end{vmatrix} = 1/V.$$

The joint density of U and V is

$$p(U, V) \propto U^{p/2-1} \exp(-U/2) V^{q/2-1} \exp(-V/2).$$

The joint density of X and Y is

$$p(X, Y) \propto X^{p/2-1} Y^{(p+q)/2-1} \exp(-Y(X+1)/2).$$

Now let

$$Z = Y(X + 1),$$

so

$$p(X, Z) \propto X^{p/2-1} Z^{(p+q)/2-1} (X + 1)^{-(p+q)/2} \exp(-Z/2)$$

and hence

$$p(X) \propto X^{p/2-1} / (X + 1)^{(p+q)/2}.$$

Now if we let

$$F = \frac{U/p}{V/q} = \frac{q}{p}X$$

then

$$p(F) \propto F^{p/2-1} / (pF/q + 1)^{(p+q)/2}.$$

F-tests

Suppose we have two nested design matrices $X_1 \in \mathcal{R}^{n \times p_1}$ and $X_2 \in \mathcal{R}^{n \times p_2}$, such that

$$\text{col}(X_1) \subset \text{col}(X_2).$$

Let P_1 and P_2 be the corresponding projections, and let

$$\begin{aligned}\hat{Y}^{(1)} &= P_1 Y \\ \hat{Y}^{(2)} &= P_2 Y\end{aligned}$$

be the fitted values.

Due to the nesting, $P_2P_1 = P_1$ and $P_1P_2 = P_1$. Therefore

$$(P_2 - P_1)^2 = P_2 - P_1,$$

so this is a projection. It projects onto $\text{col}(X_2) - \text{col}(X_1)$, the complement of $\text{col}(X_1)$ in $\text{col}(X_2)$.

Since

$$(I - P_2)(P_2 - P_1) = 0,$$

it follows that

$$\text{cov}(Y - \hat{Y}^{(2)}, \hat{Y}^{(2)} - \hat{Y}^{(1)}) = 0.$$

If the linear model errors are Gaussian, $Y - \hat{Y}^{(2)}$ and $\hat{Y}^{(2)} - \hat{Y}^{(1)}$ are independent.

Since $P_2X_1 = P_1X_1 = X_1$, we have

$$(I - P_2)X_1 = (P_2 - P_1)X_1 = 0.$$

Now suppose we take as the null hypothesis that $EY \in \text{col}(X_1)$, so we can write $Y = \theta + \epsilon$, where $\theta \in \text{col}(X_1)$. Therefore under the null hypothesis

$$\|Y - \hat{Y}^{(2)}\|^2 = \text{tr}(I - P_2)YY' = \text{tr}(I - P_2)\epsilon\epsilon'$$

and

$$\|\hat{Y}^{(2)} - \hat{Y}^{(1)}\|^2 = \text{tr}(P_2 - P_1)YY' = \text{tr}(P_2 - P_1)\epsilon\epsilon'.$$

Since $I - P_2$ and $P_2 - P_1$ are projections onto subspaces of dimension $n - p_2$ and $p_2 - p_1$, respectively, it follows that

$$\|Y - \hat{Y}^{(2)}\|^2/\sigma^2 = \|(I - P_2)Y\|^2/\sigma^2 \sim \chi_{n-p_2}^2$$

and under the null hypothesis

$$\|\hat{Y}^{(2)} - \hat{Y}^{(1)}\|^2/\sigma^2 = \|(P_2 - P_1)Y\|^2/\sigma^2 \sim \chi_{p_2-p_1}^2.$$

Therefore

$$\frac{\|\hat{Y}_2 - \hat{Y}_1\|^2/(p_2 - p_1)}{\|Y - \hat{Y}_2\|^2/(n - p_2)} \sim F_{p_2-p_1, n-p_2}.$$

Since $\hat{Y}_2 - \hat{Y}_1$ will tend to be large when $EY \notin \text{col}(X_1)$, i.e. when the null hypothesis is false, this quantity can be used as a test-statistic with p-values determined by the $F_{p_2-p_1, n-p_2}$ null distribution.

Simultaneous confidence intervals

If θ is a fixed vector, we can cover the value

$$\theta'\beta$$

with probability $1 - \alpha$ by pivoting on the t_{n-p-1} -distributed quantity

$$\frac{\theta'\hat{\beta} - \theta'\beta}{\sqrt{\hat{\sigma}^2 V_\theta}},$$

where

$$V_\theta = \theta'(X'X)^{-1}\theta.$$

Now suppose we have a set \mathcal{T} of vectors θ , and we want to construct a set of confidence intervals such that

$$P(\text{all } \theta'\beta \text{ covered, } \theta \in \mathcal{T}) = \alpha.$$

We call this a “simultaneous set of confidence intervals for $\{\theta'\beta; \theta \in \mathcal{T}\}$.”

The Bonferroni approach

The Bonferroni approach can be applied when \mathcal{T} is a finite set, $|\mathcal{T}| = k$. Let

$$I_j = \mathcal{I}(\text{CI } j \text{ covers } \theta'_j \beta)$$

and

$$I'_j = \mathcal{I}(\text{CI } j \text{ does not cover } \theta'_j \beta).$$

Then

$$\begin{aligned} P(I_1 \text{ and } I_2 \cdots \text{ and } I_k) &= 1 - P(I'_1 \text{ or } I'_2 \cdots \text{ or } I'_k) \\ &\geq 1 - \sum_i P(I'_i). \end{aligned}$$

Therefore as long as

$$1 - \alpha \geq \sum_i P(I'_i),$$

the intervals cover simultaneously. One way to achieve this is if each interval individually has probability

$$\alpha' \equiv 1 - (1 - \alpha)/k$$

of covering its corresponding true value. To do this, use the same approach as used to construct single confidence intervals, but with a much larger value of q_α .

The Scheffé approach

The Scheffé approach can be applied if \mathcal{T} is a linear subspace of \mathcal{R}^{p+1} .

Begin with the pivotal quantity

$$\frac{\theta' \hat{\beta} - \theta' \beta}{\sqrt{\hat{\sigma}^2 \theta' (X' X)^{-1} \theta}},$$

and postulate that a symmetric interval can be found so that

$$P \left(-M_\alpha \leq \frac{\theta' \hat{\beta} - \theta' \beta}{\sqrt{\hat{\sigma}^2 \theta' (X' X)^{-1} \theta}} \leq M_\alpha \text{ for all } \theta \in \mathcal{T} \right) = \alpha.$$

Equivalently, we can write

$$P \left(\sup_{\theta \in \mathcal{I}} \frac{(\theta' \hat{\beta} - \theta' \beta)^2}{\hat{\sigma}^2 \theta' (X' X)^{-1} \theta} \leq M_\alpha^2 \right) = \alpha.$$

Since

$$\hat{\beta} - \beta = (X' X)^{-1} X' \epsilon,$$

we have

$$\begin{aligned} \frac{(\theta' \hat{\beta} - \theta' \beta)^2}{\hat{\sigma}^2 \theta' (X' X)^{-1} \theta} &= \frac{\theta' (X' X)^{-1} X' \epsilon \epsilon' X (X' X)^{-1} \theta}{\hat{\sigma}^2 \theta' (X' X)^{-1} \theta} \\ &= \frac{M'_\theta \epsilon \epsilon' M_\theta}{\hat{\sigma}^2 M'_\theta M_\theta}, \end{aligned}$$

where

$$M_\theta = X(X'X)^{-1}\theta.$$

Note that

$$\frac{M_\theta' \epsilon \epsilon' M_\theta}{\hat{\sigma}^2 M_\theta' M_\theta} = \langle \epsilon, M_\theta / \|M_\theta\| \rangle^2 / \hat{\sigma}^2,$$

i.e. it is the squared length of the projection of ϵ onto the line spanned by M_θ (divided by $\hat{\sigma}^2$).

The quantity $\langle \epsilon, M_\theta / \|M_\theta\| \rangle^2$ is maximized at $\|P\epsilon\|^2$, where P is the projection onto the linear space

$$\{X(X'X)^{-1}\theta \mid \theta \in \mathcal{T}\} = \{M_\theta\}.$$

Therefore

$$\sup_{\theta \in \mathcal{T}} \langle \epsilon, M_\theta / \|M_\theta\| \rangle^2 / \hat{\sigma}^2 = \|P\epsilon\|^2 / \hat{\sigma}^2,$$

and since $\{M_\theta\} \subset \text{col}(X)$, it follows that $P\epsilon$ and $\hat{\sigma}^2$ are independent.

Moreover,

$$\|P\epsilon\|^2/\sigma^2 \sim \chi_q^2$$

where $q = \dim(\mathcal{T})$, and as we know,

$$\frac{n-p-1}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-p-1}^2.$$

Thus

$$\frac{\|P\epsilon\|^2/q}{\hat{\sigma}^2} \sim F_{q, n-p-1}.$$

Let Q_F be the α quantile of the $F_{q,n-p-1}$ distribution. Then

$$P \left(\frac{|\theta' \hat{\beta} - \theta' \beta|}{\sqrt{\theta' (X' X)^{-1} \theta}} \leq \hat{\sigma} \sqrt{q Q_F} \text{ for all } \theta \right) = \alpha,$$

so

$$P \left(\theta' \hat{\beta} - \hat{\sigma} \sqrt{q Q_F V_\theta} \leq \theta' \beta \leq \theta' \hat{\beta} + \hat{\sigma} \sqrt{q Q_F V_\theta} \text{ for all } \theta \right) = \alpha$$

defines a level α simultaneous confidence set for $\{\theta' \beta \mid \theta \in \mathcal{T}\}$, where

$$V_\theta = \theta' (X' X)^{-1} \theta.$$