Least squares fitting and inference for linear models

Kerby Shedden

Department of Statistics, University of Michigan

December 9, 2019

Goals and terminology

The goal is to learn about an unknown function f that relates variables $y \in \mathcal{R}$ and $x \in \mathcal{R}^p$ through a relationship $y \approx f(x)$. The variables y and x have distinct roles:

- Independent variables (predictors, regressors, covariates, exogeneous variables): $x \in \mathcal{R}^p$, a vector.
- Dependent variable (response, outcome, endogeneous variables): $y \in \mathcal{R}$, a scalar.

Note that the terms "independent" and "dependent" do not imply statistical independence or linear algebraic independence of these variables. Instead, these terms suggest that the value of x can be manipulated, or exhibits variation "on its own", and we observe the consequent changes in y.

Goals and terminology

The analysis is empirical (based on a sample of data):

$$y_i \in \mathcal{R}$$
$$x_i = (x_{i1}, \ldots, x_{ip})' \in \mathcal{R}^p$$
$$i = 1, \ldots, n.$$

where n is the sample size.

The data point (y_i, x_i) reflects the i^{th} "analysis unit" (also called "case" or "observation").

Why do we want to do this?

- *Prediction:* Estimate f to predict the typical value of y at a given x point using $\hat{f}(x)$ (x is not necessarily one of the x_i points in the data).
- *Learning about structure:* Inductive learning about the relationship between x and y, such as understanding which predictors or combinations of predictors are associated with particular changes in y.

"Learning about structure" often has the goal of better understanding the physical (biological, social, etc.) mechanisms underlying the relationship between x and y.

We will see that inferences about mechanisms based on regression analysis can be misleading, particularly when based on observational data.

Examples:

- An empirical model for the weather conditions 48 hours from now could be based on current and historical weather conditions. Such a model could have a lot of practical value, but it 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 any risk due to lead exposure, and to estimate the overall effects of lead exposure in a large population. 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

$$f(x) \equiv E[Y|X = x],$$

where Y and X are random variables.

Less commonly, the regression function is defined as a conditional quantile, such as the median

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

or even some other quantile $f(x) \equiv Q_p(Y|X = x)$. This is called quantile regression.

The conditional expectation function

The conditional expectation $E[Y|X]$ 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 occur when $X = x$. If there are densities we can write:

$$E[Y|X = x] = \int y \cdot f(y|X = x)dy.$$

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.

In regression analysis, 1 is much more commonly used than 2.

Least Squares Fitting

In a linear model, the independent variable x is postulated to be related to the dependent variable y via a linear relationship

$$y_i \approx \sum_{j=1}^{p} \beta_j x_{ij} = \beta' x_i.$$

This is a "linear model" in two senses: it is linear in β for fixed x, and it is linear in x for fixed β (technically, it is "bilinear").

Least Squares Fitting

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

$$L(\beta) = \sum_i (y_i - \sum_j \beta_j x_{ij})^2 = \sum_i (y_i - \beta' x_i)^2$$

This is called least squares estimation.

Simple linear regression

A special case of the linear model is simple linear regression, when there is $p = 1$ covariate and an intercept (a covariate whose value is always 1).

$$L(\alpha, \beta) = \sum_i (y_i - \alpha - \beta x_i)^2$$

We can differentiate with respect to α and β:

$$\begin{array}{rcccl} \partial L/\partial \alpha & = & -2\sum_i (y_i - \alpha - \beta x_i) & = & -2\sum r_i \\ \partial L/\partial \beta & = & -2\sum_i (y_i - \alpha - \beta x_i)x_i & = & -2\sum_i r_i x_i \end{array}$$

$$r_i = y_i - \alpha - \beta x_i$$

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

Simple linear regression

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 (^) distinguishes the least-squares optimal values of α and β from an arbitrary pair of parameter values.

Simple linear regression

We will call $\hat{\alpha}$ and $\hat{\beta}$ the "least squares estimates" of the model parameters α and β.

At the least squares solution,

$$\begin{array}{rcccl} \partial L/\partial \alpha & = & 2\sum r_i & = & 0 \\ \partial L/\partial \beta & = & -2\sum_i r_i x_i & = & 0 \end{array}$$

we have the following basic properties for the least squares estimates:

- The residuals $r_i = y_i - \alpha - \beta x_i$ sum to zero
- The residuals are orthogonal to the independent variable x.

Recall that two vectors $v, w \in \mathcal{R}^d$ are orthogonal if $\sum_j v_j w_j = 0$.

Two important identities

Centered and uncentered sums of squares can be related as follows:

$$\sum x_i^2/n - \bar{x}^2 = \sum_i (x_i - \bar{x})^2/n.$$

Centered and uncentered cross-products can be related as follows:

$$\sum y_i x_i/n - \bar{y}\bar{x} = \sum y_i(x_i - \bar{x})/n = \sum_i (y_i - \bar{y})(x_i - \bar{x})/n.$$

Simple linear regression

Note that

$$\sum_i (x_i - \bar{x})^2/n \qquad \text{and} \qquad \sum_i (y_i - \bar{y})(x_i - \bar{x})/n.$$

are essentially the sample variance of $x_1, \ldots, x_n$, 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

$$\hat{\beta} = \frac{\widehat{\mathrm{cov}}(y, x)}{\widehat{\mathrm{var}}(x)}$$

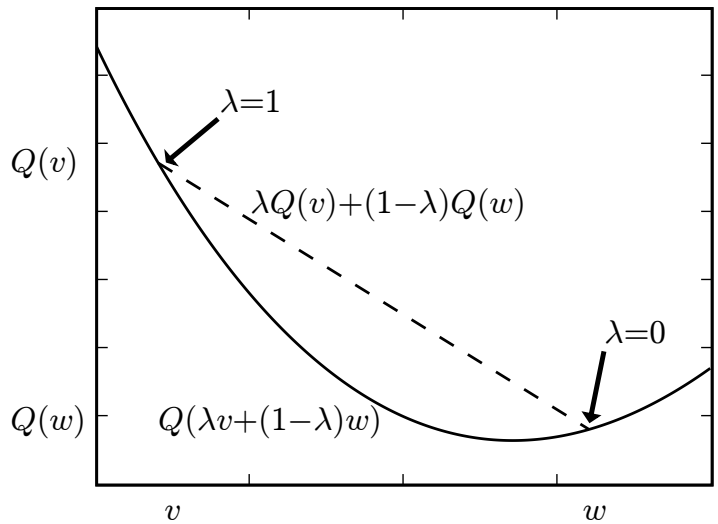
where $\widehat{\mathrm{cov}}$ and $\widehat{\mathrm{var}}$ are the usual unbiased estimates of variance and covariance.

Convex functions

A map $Q : \mathcal{R}^d \to \mathcal{R}$ 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.



Convex functions

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

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 of quadratic functions

A general quadratic function in d dimensions can be written

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

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

Note that

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

Convexity of quadratic functions

If $b \in \mathrm{col}(A)$, we can complete the square to eliminate the linear term, giving us

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

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

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

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.

Convexity of quadratic functions

Two key definitions:

- A square matrix A is positive definite if $v'Av > 0$ for all vectors $v \neq 0$.
- A square 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.

Note that without loss of generality, A is symmetric, since otherwise $\tilde{A} \equiv (A + A')/2$ gives the same quadratic form as A.

Convexity of quadratic functions

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

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

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

Convexity of quadratic functions

The gradient, or Jacobian of a scalar-valued function $y = f(x)$ ($y \in \mathcal{R}$, $x \in \mathcal{R}^d$) is

$$(\partial f/\partial x_1, \ldots, \partial f/\partial x_d),$$

which is viewed as a row vector by convention.

Convexity of quadratic functions

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

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

and differentiate with respect to v_ℓ to get the ℓ^{th} component of the Jacobian:

$$2v_\ell A_{\ell\ell} + \sum_{j \neq \ell} v_j A_{\ell j} + \sum_{i \neq \ell} v_i A_{i\ell} = 2(Av)_\ell.$$

Convexity of quadratic functions

The Hessian of a scalar-valued function $y = f(x)$ ($y \in \mathcal{R}$, $x \in \mathcal{R}^d$) is the matrix

$$H_{ij} = \partial^2 f / \partial x_j \partial x_i.$$

If A is symmetric, the Hessian of the quadratic form Q is $2A$.

To see this, note that the ℓ^{th} component of $2v'A'$ is the inner product of v with the ℓ^{th} row (or column) of A. The derivative of this inner product with respect to a second index ℓ' is $2A_{\ell\ell'}$.

It follows that a quadratic function is strictly convex iff its Hessian is positive definite (more generally, any continuous, twice differentiable function is convex on $\mathcal{R}^d$ if and only if its Hessian matrix is everywhere positive definite).

Uniqueness of the simple linear regression least squares fit

The least squares solution for simple linear regression, $\hat{\alpha}$, $\hat{\beta}$, is unique as long as $\widehat{\mathrm{var}}[x]$ (the sample variance of the covariate) is positive.

To see this, note that the Hessian (second derivative matrix) of $L(\alpha, \beta)$ is

$$H = \begin{pmatrix} \partial^2 L/\partial\alpha^2 & \partial^2 L/\partial\alpha\partial\beta \\ \partial^2 L/\partial\alpha\partial\beta & \partial^2 L/\partial\beta^2 \end{pmatrix} = \begin{pmatrix} 2n & 2x_\cdot \\ 2x_\cdot & 2\sum x_i^2 \end{pmatrix}$$

where $x_\cdot = \sum_i x_i$.

Uniqueness of the simple linear regression least squares fit

If $\widehat{\mathrm{var}}(x) > 0$ then this is a positive definite matrix since all the principal submatrices have positive determinants:

$$|2n| > 0$$

$$\begin{vmatrix} 2n & 2x_. \\ 2x_. & 2\sum x_i^2 \end{vmatrix} = 4n\sum x_i^2 - 4(x_.)^2$$
$$= 4n(n-1)\widehat{\mathrm{var}}(x)$$
$$> 0.$$

The fitted line and the data center

The fitted line

$$y = \hat{\alpha} + \hat{\beta}x$$

passes through the center of mass of the data $(\bar{x}, \bar{y})$.

This can be seen by substituting $x = \bar{x}$ into the equation of the fitted line, which yields $\bar{y}$.

Regression slopes and the Pearson correlation

The sample Pearson correlation coefficient between x and y is

$$\hat{\rho} = \frac{\widehat{\mathrm{cov}}(y, x)}{\widehat{\mathrm{SD}}(y)\widehat{\mathrm{SD}}(x)}$$

The relationship between $\hat{\beta}$ and $\hat{\rho}$ is

$$\hat{\beta} = \hat{\rho} \cdot \frac{\widehat{\mathrm{SD}}(y)}{\widehat{\mathrm{SD}}(x)}.$$

Thus the fitted slope has the same sign as the Pearson correlation coefficient between y and x.

Reversing x and y

If x and y are reversed in a simple linear regression, the slope is

$$\hat{\beta}_* = \frac{\widehat{\mathrm{cov}}(y,x)}{\widehat{\mathrm{var}}(y)} = \widehat{\mathrm{cor}}(y,x)\frac{\widehat{\mathrm{SD}}(x)}{\widehat{\mathrm{SD}}(y)}.$$

If the data fall exactly on a line, then $\mathrm{cor}(y,x) = 1$, 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 property does not hold.

Norms

The Euclidean norm on vectors (the most commonly used norm) is:

$$\|v\| = \sqrt{\sum_i v_i^2} = \sqrt{v'v}.$$

Here are some useful identities, for vectors $v, w \in \mathcal{R}^p$:

$$\|v + w\|^2 = \|v\|^2 + \|w\|^2 + 2v'w$$

$$\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2v'w$$

Fitting multiple linear 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.

Fitting multiple linear regression models

The linear model coefficients are written as a vector

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

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.

Fitting multiple linear regression models

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

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

Estimating β by minimizing $L(\beta)$ is called ordinary least squares (OLS).

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

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

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

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

The multivariate chain rule

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 $g(z)$.

The least squares gradient function

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 $f(x) = \sum_j x_j^2$.

Normal equations

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.

Solving the normal equations

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

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

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

$$\mathbf{R}\beta = \mathbf{Q}'Y,$$

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

Matrix products

In multiple regression we encounter the matrix product $\mathbf{X}'\mathbf{X}$. Let's review some ways to think about matrix products.

If $\mathbf{A} \in \mathcal{R}^{n \times m}$ and $\mathbf{B} \in \mathcal{R}^{n \times p}$ are matrices, we can form the product $\mathbf{A}'\mathbf{B} \in \mathcal{R}^{m \times p}$.

Suppose we partition $\mathbf{A}$ and $\mathbf{B}$ by rows:

$$\mathbf{A} = \begin{pmatrix} - & a_1 & - \\ - & a_2 & - \\ & \cdots & \\ & \cdots & \\ - & a_n & - \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} - & b_1 & - \\ - & b_2 & - \\ & \cdots & \\ & \cdots & \\ - & b_n & - \end{pmatrix}$$

Matrix products

Then

$$\mathbf{A}'\mathbf{B} = a_1' b_1 + a_2' b_2 + \cdots + a_n' b_n,$$

where each $a_j' b_j$ term is an outer product

$$a_j' b_j = \begin{pmatrix} a_{j1} b_{j1} & a_{j1} b_{j2} & \cdots \\ a_{j2} b_{j1} & a_{j2} b_{j2} & \cdots \\ \cdots & \cdots & \end{pmatrix} \in \mathcal{R}^{m \times p}.$$

Matrix products

Now if we partition $\mathbf{A}$ and $\mathbf{B}$ by columns

$$\mathbf{A} = \begin{pmatrix} | & | & | \\ a_1 & a_2 & a_3 \\ | & | & | \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} | & | & | \\ b_1 & b_2 & b_3 \\ | & | & | \end{pmatrix},$$

then $\mathbf{A}'\mathbf{B}$ is a matrix of inner products

$$\mathbf{A}'\mathbf{B} = \begin{pmatrix} a_1' b_1 & a_1' b_2 & \cdots \\ a_2' b_1 & a_2' b_2 & \cdots \\ \cdots & \cdots & \cdots \\ & & \end{pmatrix}.$$

Matrix products

Thus we can view the product $\mathbf{X}'\mathbf{X}$ involving the design matrix in two different ways. If we partition $\mathbf{X}$ by rows (the cases)

$$\mathbf{X} = \begin{pmatrix} - & x_1 & - \\ - & x_2 & - \\ & \cdots & \\ & \cdots & \\ - & x_n & - \end{pmatrix}$$

then

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} x_i' x_i$$

is the sum of the outer product matrices of the cases.

Matrix products

If we partition $\mathbf{X}$ by the columns (the variables)

$$\mathbf{X} = \begin{pmatrix} | & | & | \\ x_1 & x_2 & x_3 \\ | & | & | \end{pmatrix},$$

then $\mathbf{X}'\mathbf{X}$ is a matrix whose entries are the pairwise inner products of the variables ($[\mathbf{X}'\mathbf{X}]_{ij} = x_i' x_j$).

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.

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:

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

Projections

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

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

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 $P_S(V)$.

Projections

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 $V_{S^\perp} = V - P_V(S)$.

Projections

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

This follows from property 2 – if there exists a V for which $V_1 \neq V_2 \in S$ both minimize the distance from V to S, then $V = V_1 + U_1$ and $V = V_2 + U_2$ ($U_j = V - V_j$) would be distinct decompositions of V as a sum of a vector in S and a vector in $S^{\perp}$, contradicting property 2.

Projections

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.

Projections

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 $\theta_A + \theta_B = P_S(A + B)$.

Projections

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

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

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 $S^{\perp}$.

Projections

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 $Q = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so for an arbitrary vector V,

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

and note that the first summand is in S while the second summand (by direct calculation) is perpendicular to the first summand, hence is in $S^{\perp}$.

Projections

Property 7 (continued):

To show that the second summand is in $S^{\perp}$, take $s \in S$ 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.

Projections

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:

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

The solution is

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

so

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

Projections

Property 8: The composition of projections onto S and $S^{\perp}$ (in either order) is identically zero.

$P_S \circ P_{S^{\perp}} = P_{S^{\perp}} \circ P_S \equiv 0$. 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

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

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

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

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

Row-wise and column-wise geometry of least squares

The rows of the design matrix $\mathbf{X}$ are vectors in $\mathcal{R}^{p+1}$, the columns of the design matrix are vectors in $\mathcal{R}^n$.

Both of these spaces are usually too big to explicitly draw, but we can think visually about both the row-wise and column-wise geometries of the least squares problem.

The n-dimensional space containing Y and $\mathrm{col}(\mathbf{X})$ is called the variable space.

The $p+2$-dimensional space containing (x_i, y_i) is called the case space.

Variable space geometry of least squares

Thinking column-wise, we are working in $\mathcal{R}^n$. The vector Y containing all values of the dependent variable is a vector in $\mathcal{R}^n$, and $\mathrm{col}(\mathbf{X})$ is a $p+1$ dimensional subspace of $\mathcal{R}^n$.

The least squares problem can be seen to have the goal of producing a vector of values that are in $\mathcal{R}^n$, and that are as close as possible to Y among all such vectors. We will usually write this vector as $\hat{Y}$. It is obtained by projecting Y onto $\mathrm{col}(\mathbf{X})$.

Case space geometry of least squares

Thinking row-wise, we are working in $\mathcal{R}^{p+1}$, or R^{p+2}. The data for one case can be written (x_i, y_i), where x_i is the i^{th} row of $\mathbf{X}$ and y_i is the i^{th} element of Y.

$(x_i, y_i) \in \mathcal{R}^{p+2}$ is the "cloud of data points", where each point includes both the independent and dependent variables in a single vector.

Alternatively, we can think of $x_i \in \mathcal{R}^{p+1}$ as being the domain of the regression function $E[Y|X = x]$, which forms a surface above this domain.

Properties of the least squares fit

- 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

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

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 $\bar{Y} = \mathbf{1}'Y/n$.

Properties of the least squares fit

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

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

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

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

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

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

so

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

Orthogonal matrices

An orthogonal matrix $\mathbf{X}$ satisfies $\mathbf{X}'\mathbf{X} = I$. This is equivalent to stating that the columns of $\mathbf{X}$ are mutually orthonormal.

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

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

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

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

Orthogonal matrices

- If $\mathbf{X}$ is orthogonal, the slopes obtained by using multiple regression of Y on $X = (X_1, \ldots, X_p)$ 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

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

Comparing multiple regression and simple regression slopes

- 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

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

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

Comparing multiple regression and simple regression slopes

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 $r_{1y}, r_{2y}, r_{12} \geq 0$, then if $r_{12} > r_{1y}/r_{2y}$, then $\hat{\beta}_1$ has opposite signs in single and multiple regression. Note that if $r_{1y} > r_{2y}$ it is impossible for $r_{12} > r_{1y}/r_{2y}$. Thus, the effect direction for the covariate that is more strongly marginally correlated with Y cannot be reversed.

This is an example of "Simpson's paradox".

Comparing multiple regression and simple regression slopes

Numerical example: if $r_{1y} = 0.1$, $r_{2y} = 0.6$, $r_{12} = 0.6$, then X_1 is (marginally) positively associated with Y, but for fixed values of X_2, the association between Y and X_1 is negative.

Formulation of regression models in terms of probability distributions

Up to this point, we have primarily expressed regression models in terms of the mean structure, e.g.

$$E[Y|X = x] = \beta' x.$$

Regression models are commonly discussed in terms of moment structures ($E[Y|X = x]$, $\mathrm{var}[Y|X = x]$) or quantile structures ($Q_p[Y|X = x]$), rather than as fully-specified probability distributions.

Formulation of regression models in terms of probability distributions

If we want to specify the model more completely, we can think in terms of a random "error term" that describes how the observed value y deviates from the ideal value $f(x)$, where (x, y) is generated according to the regression model.

A very general regression model is

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

where ϵ is a random variable with expected value zero.

If we specify the distribution of ϵ, then we have fully specified the distribution of $Y|X$.

Formulation of regression models in terms of probability distributions

A more restrictive "additive error" model is:

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

Under this model,

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

Without loss of generality, $E[\epsilon|X] = 0$, so $E[Y|X] = f(X)$.

Regression model formulations and parameterizations

A parametric regression model is:

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

where θ is a finite dimensional parameter vector.

Examples:

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

Models 1 and 2 are both "linear models" because they are linear in θ.
The Gompertz curve is a non-linear model because it is not linear in θ.

Basic inference for simple linear regression

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

Specifically, we will assume derive properties of $\hat{\beta}$ that hold for the generating model $y = x'\beta + \epsilon$, where:

i $E[\epsilon|x] = 0$

ii $\mathrm{var}[\epsilon|x] = \sigma^2$ exists and is constant across cases

iii the ϵ random variables are uncorrelated across cases (given x).

We will not assume here that ϵ follows a particular distribution, e.g. a Gaussian distribution.

The relationship between $E[\epsilon|X]$ and $\mathrm{cov}(\epsilon, X)$

If we treat X as a random variable, the condition that

$$E[\epsilon|X] = 0$$

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

$$\begin{aligned}
\mathrm{cov}(X, \epsilon) &= E[X\epsilon] - E[X] \cdot E[\epsilon] \\
&= E[X\epsilon] \\
&= E_X[E[\epsilon X|X]] \\
&= E_X[XE[\epsilon|X]] \\
&= E_X[X \cdot 0] \\
&= 0.
\end{aligned}$$

The relationship between $E[\epsilon|X]$ and $\text{cov}(\epsilon, X)$

The converse is not true. If $\text{cor}(X, \epsilon) = 0$ and $E[\epsilon] = 0$ it may not be the case that $E[\epsilon|X] = 0$.

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 $E[\epsilon|X] = E\epsilon = 0$.

Data sampling

To be able to interpret the results of a regression analysis, we need to know how the data were sampled. In particular, it 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 two important situations:

- Designed experiment: We are studying the effect of temperature on reaction yield in a chemical synthesis. The temperature X is controlled and set by the experimenter. In this case, Y is randomly sampled conditionally on X, but X is not randomly sampled, so it doesn't make sense to consider X to be a random variable.
- Observational study: We are interested in the relationship between cholesterol level X and blood pressure Y. We sample people at random from a well defined population (e.g. residents of Michigan) and measure their blood pressure and cholesterol levels. In this case, X and Y are sampled from their joint distribution, and both can be viewed a random variables.

Data sampling

There is another design that we may encounter:

- 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 residents of Michigan. We wish to select a subset of 500 individuals to contact for acquiring cholesterol measures, and it is decided that studying the 250 people with greatest blood pressure together with the 250 people with lowest blood pressure will be most informative. In this case 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 are sampled also affects our interpretation of the results.

Basic inference for simple linear regression via OLS

Above we showed that the slope and intercept estimates are

$$\hat{\beta} = \frac{\widehat{\mathrm{cov}}(Y, X)}{\widehat{\mathrm{var}}[X]} = \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2},$$

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

Note that we are using the useful fact that

$$\sum_i (y_i - \bar{y})(x_i - \bar{x}) = \sum_i y_i(x_i - \bar{x}) = \sum_i (y_i - \bar{y})x_i.$$

Sampling means of parameter estimates

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 $E[\hat{\beta}|X] = \beta$. Thus $\hat{\beta}$ is unbiased (a parameter estimate is unbiased if its sampling mean is the same as the population value of the parameter).

The intercept is also unbiased:

$$\begin{aligned}E[\hat{\alpha}|X] &= E(\bar{y} - \hat{\beta}\bar{x}|X) \\ &= \alpha + \beta\bar{x} + E[\bar{\epsilon}|X] - E[\hat{\beta}\bar{x}|X] \\ &= \alpha\end{aligned}$$

Sampling variances of parameter estimates

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

First we will need the following result:

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

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

Sampling variances of parameter estimates

To derive the sampling variances, start with the identity:

$$\hat{\beta} = \beta + \sum_i \epsilon_i (x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2.$$

The sampling variances are

$$\begin{aligned} \mathrm{var}[\hat{\beta}|X] &= \sigma^2 / \sum_i (x_i - \bar{x})^2 \\ &= \sigma^2 / \left((n-1) \widehat{\mathrm{var}}[x] \right). \end{aligned}$$

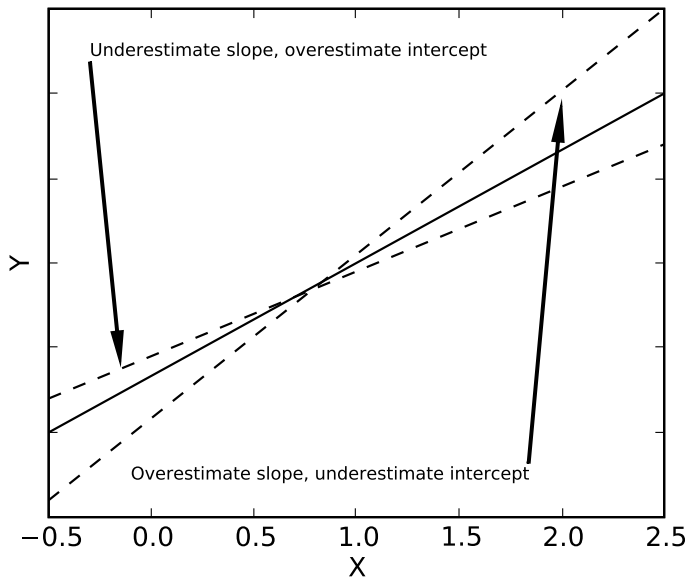
$$\begin{aligned} \mathrm{var}[\hat{\alpha}|X] &= \mathrm{var}[\bar{y} - \hat{\beta}\bar{x}|X] \\ &= \mathrm{var}[\alpha + \beta\bar{x} + \bar{\epsilon} - \hat{\beta}\bar{x}|X] \\ &= \mathrm{var}[\bar{\epsilon}|X] + \bar{x}^2 \mathrm{var}[\hat{\beta}|X] - 2\bar{x}\mathrm{cov}[\bar{\epsilon}, \hat{\beta}|X] \\ &= \sigma^2/n + \bar{x}^2 \sigma^2 / ((n-1)\widehat{\mathrm{var}}[x]). \end{aligned}$$

Sampling covariance of parameter estimates

The sampling covariance between the slope and intercept is

$$\begin{aligned}
\mathrm{cov}[\hat{\alpha}, \hat{\beta}|X] &= \mathrm{cov}[\bar{y} - \hat{\beta}\bar{x}, \hat{\beta}|X] \\
&= \mathrm{cov}[\bar{y}, \hat{\beta}|X] - \bar{x}\mathrm{var}[\hat{\beta}|X] \\
&= \mathrm{cov}[\bar{\epsilon}, \hat{\beta}|X] - \bar{x}\mathrm{var}[\hat{\beta}|X] \\
&= -\sigma^2\bar{x}/((n-1)\widehat{\mathrm{var}}[X]).
\end{aligned}$$

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



Basic inference for simple linear regression via OLS

Some observations:

- All variances scale with sample size like $1/n$.
- $\hat{\beta}$ does not depend on $\bar{x}$.
- $\mathrm{var}[\hat{\alpha}]$ 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 $Y_i = \alpha + \beta x_i + \epsilon_i$ and therefore $\bar{Y} = \alpha + \beta \bar{x} + \bar{\epsilon}$, we can subtract to get

$$Y_i - \bar{Y} = \beta(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon}$$

it follows that

$$R_i = (\beta - \hat{\beta})(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon}.$$

Since $E\hat{\beta} = \beta$ and $E[\epsilon_i - \bar{\epsilon}] = 0$, it follows that $ER_i = 0$. Note that this is a distinct fact from the identity $\sum_i R_i = 0$.

Some properties of residuals

It is important to distinguish the residual R_i from the "observation errors" ϵ_i. The identity $R_i = (\beta - \hat{\beta})(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon}$ shows us that the centered observation error $\epsilon_i - \bar{\epsilon}$ is one part of the residual. The other term

$$(\beta - \hat{\beta})(x_i - \bar{x})$$

reflects the fact that the residuals are also influenced by how well we recover the true slope β through our estimate $\hat{\beta}$.

Some properties of residuals

To illustrate, consider two possibilities:

- We overestimate the slope ($\beta - \hat{\beta} < 0$). The residuals to the right of the right of the mean (i.e. when $X > \bar{x}$) are shifted down (toward $-\infty$), and the residuals to the left of the mean are shifted up.
- We underestimate the slope ($\beta - \hat{\beta} > 0$). The residuals to the right of the right of the mean are shifted up, and the residuals to the left of the mean are shifted down.

Some properties of residuals

We can use the identity $R_i = (\beta - \hat{\beta})(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon}$ to derive the variance of each residual. We have:

$$\text{var}[(\beta - \hat{\beta})(x_i - \bar{x})] = \sigma^2 (x_i - \bar{x})^2 / \sum_j (x_j - \bar{x})^2$$

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

$$\begin{aligned} \text{cov}\left((\beta - \hat{\beta})(x_i - \bar{x}), \epsilon_i - \bar{\epsilon}|X\right) &= -(x_i - \bar{x})\text{cov}(\hat{\beta}, \epsilon_i|X) \\ &= -\sigma^2 (x_i - \bar{x})^2 / \sum_j (x_j - \bar{x})^2, \end{aligned}$$

Some properties of residuals

Thus the variance of the i^{th} residual is

$$\begin{aligned}
\text{var}[R_i|X] &= \text{var}\left((\beta - \hat{\beta})(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon}|X\right) \\
&= \text{var}\left((\beta - \hat{\beta})(x_i - \bar{x})|X\right) + \text{var}[\epsilon_i - \bar{\epsilon}|X] + \\
&\qquad 2\text{cov}\left((\beta - \hat{\beta})(x_i - \bar{x}), \epsilon_i - \bar{\epsilon}|X\right) \\
&= (n-1)\sigma^2/n - \sigma^2(x_i - \bar{x})^2/\sum_j (x_j - \bar{x})^2.
\end{aligned}$$

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

$$(x_i - \bar{x})^2/\sum_j (x_j - \bar{x})^2 \leq (n-1)/n.$$

Some properties of residuals

- The residuals are not iid – they are correlated with each other, and they have different variances.
- $\mathrm{var}[R_i] < \mathrm{var}[\epsilon_i]$ – the residuals are less variable than the errors.

Some properties of residuals

Since $E[R_i] = 0$ it follows that $\mathrm{var}[R_i] = E[R_i^2]$. 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 the error variance σ^2

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. That is

$$E\left(\sum_i R_i^2/(n-2)\right) = \sigma^2.$$

Basic inference for multiple linear regression via OLS

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

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

Letting

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

we see that $E[\eta|\mathbf{X}] = 0$ and

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

- $\mathrm{var}[\epsilon|\mathbf{X}] = \sigma^2 I$ since (i) the ϵ_i are uncorrelated and (ii) the ϵ_i have constant variance given X.

Variance of $\hat{\beta}$ in multiple regression OLS

Let u_i be the i^{th} row of the design matrix $\mathbf{X}$. Then

$$\mathbf{X}'\mathbf{X} = \sum_i u_i' u_i$$

where $u_i' u_i$ is an outer-product (a $p+1 \times p+1$ matrix).

If we have a limiting behavior

$$n^{-1} \sum_i u_i' u_i \to Q,$$

for a fixed $p+1 \times p+1$ matrix Q, then $\mathbf{X}'\mathbf{X} \approx nQ$, so

$$\text{cov}(\hat{\beta}|\mathbf{X}) \approx \sigma^2 Q^{-1}/n.$$

Thus we see the usual influence of sample size on the standard errors of the regression coefficients.

Level sets of quadratic forms

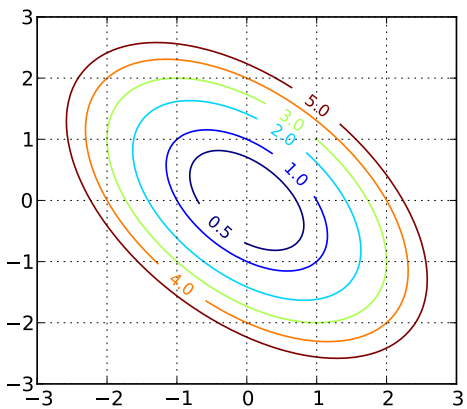
Suppose

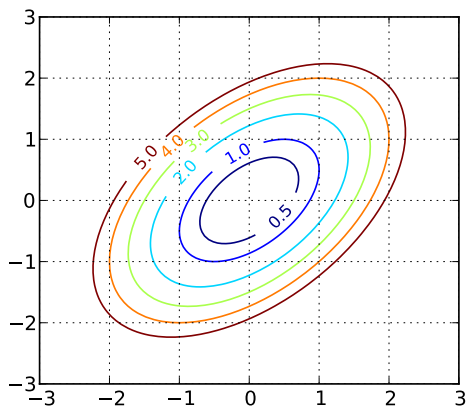
$$C = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

so that

$$C^{-1} = \frac{2}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

Left side: level curves of $g(x) = x'Cx$
Right side: level curves of $h(x) = x'C^{-1}x$





Eigen-decompositions and quadratic forms

The dominant eigenvector of C maximizes the "Rayleigh quotient"

$$g(x) = x'Cx/x'x,$$

thus it points in the direction of greatest change of g.

If

$$C = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

the dominant eigenvector points in the $(1, 1)$ direction. The dominant eigenvector of C^{-1} points in the $(-1, 1)$ direction.

If the eigendecomposition of C is $\sum_j \lambda_j v_j v_j'$ then the eigendecomposition of C^{-1} is $\sum_j \lambda_j^{-1} v_j v_j'$ – thus directions in which the level curves of C are most spread out are the directions in which the level curves of C^{-1} are most compressed.

Variance of $\hat{\beta}$ in multiple regression OLS

If $p = 2$ and

$$X'X/n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & r \\ 0 & r & 1 \end{pmatrix},$$

then

$$\mathrm{cov}(\hat{\beta}) = \sigma^2 n^{-1} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/(1-r^2) & -r/(1-r^2) \\ 0 & -r/(1-r^2) & 1/(1-r^2) \end{pmatrix}.$$

So if $p = 2$ and X_1 and X_2 are positively colinear (meaning that $(X_1 - \bar{x}_1)'(X_2 - \bar{x}_2) > 0$), the corresponding slope estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are negatively correlated.

The residual sum of squares

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

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

where P is the projection matrix onto $\mathrm{col}(\mathbf{X})$. 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 $\mathrm{RSS} = Y'(I - P)Y$ is a quadratic form in Y, and we can write

$$Y'(I - P)Y = \mathrm{tr}\left(Y'(I - P)Y\right) = \mathrm{tr}\left((I - P)YY'\right),$$

where the second equality uses the circulant property of the trace.

For three factors, the circulent property states that

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

The expected value of the RSS

By linearity we have

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

and

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

Since $P\mathbf{X} = \mathbf{X}$ and hence $(I-P)\mathbf{X} = 0$,

$$(I-P)E[YY'] = \sigma^2(I-P).$$

Therefore the expected value of the RSS is $E[\mathrm{RSS}] = \sigma^2\mathrm{tr}(I-P)$.

Four more properties of projection matrices

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

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

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

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

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

$$E[\mathrm{RSS}] = \sigma^2 (n - p - 1),$$

so

$$\mathrm{RSS}/(n - p - 1)$$

is an unbiased estimate of σ^2.

Covariance matrix of residuals

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

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

Distribution of the RSS

The RSS can be written

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

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

Distribution of the RSS

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 $\mathbf{X}$ 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 $\mathrm{col}(I-P)$ to $\mathcal{S}$). Then

$$\begin{aligned} \mathrm{tr}[(I-P)\epsilon\epsilon'] &\stackrel{d}{=} \mathrm{tr}[(I-P)Q\epsilon(Q\epsilon)'] \\ &= \mathrm{tr}[Q'(I-P)Q\epsilon\epsilon'] \end{aligned}$$

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

Optimality

For a given design matrix $\mathbf{X}$, 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

$$E[MY|\mathbf{X}] = \beta$$

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.

Optimality (BLUE)

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

$$\mathrm{cov}(\hat{\beta}|\mathbf{X}) \leq \mathrm{cov}(\beta^*|\mathbf{X}),$$

this implies that for any fixed vector θ,

$$\mathrm{var}[\theta'\hat{\beta}|\mathbf{X}] \leq \mathrm{var}[\theta'\beta^*|\mathbf{X}].$$

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

Optimality (proof of GMT)

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

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

To prove the theorem note that

$$E[\beta^*|\mathbf{X}] = M \cdot E[Y|\mathbf{X}] = M\mathbf{X}\beta = \beta$$

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

$$E[\hat{\beta}|\mathbf{X}] = B\mathbf{X}\beta = \beta$$

for all β, so $(M - B)\mathbf{X} \equiv 0$.

Optimality (proof of GMT)

Therefore

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

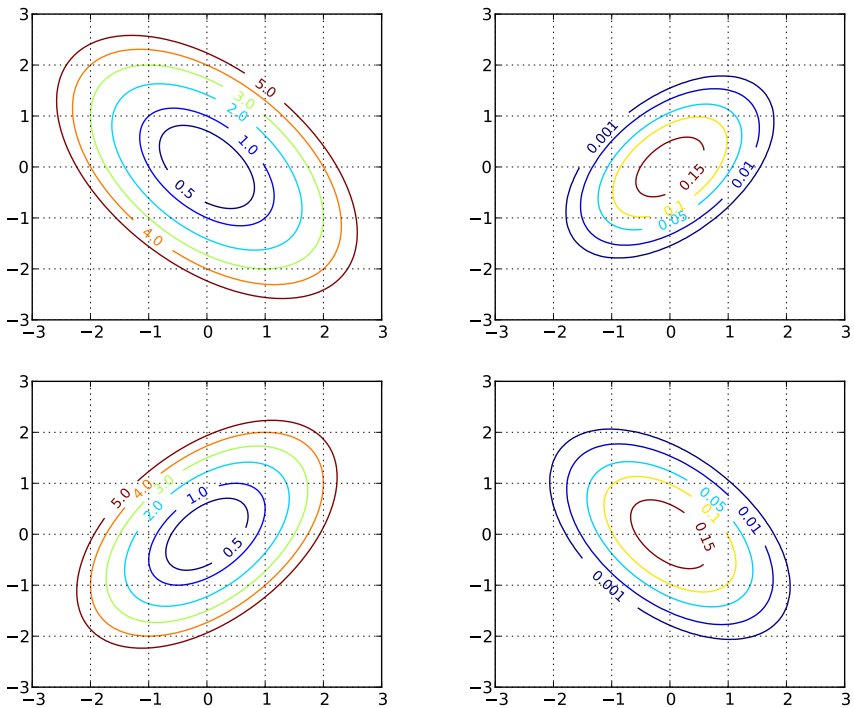
Note that we have an explicit expression for the gap between $\mathrm{cov}(\hat{\beta})$ and $\mathrm{cov}(\beta^*)$:

$$\mathrm{cov}(\beta^* - \hat{\beta}|\mathbf{X}) = \sigma^2(M-B)(M-B)'.$$

Multivariate density families

If C is a covariance matrix, many families of multivariate densities have the form $c \cdot \phi((x-\mu)'C^{-1}(x-\mu))$, where $c > 0$ and $\phi : \mathcal{R}^+ \to \mathcal{R}^+$ is a function, typically decreasing with a mode at the origin (e.g. for the multivariate normal density, $\phi(u) = \exp(-u/2)$ and for the multivariate t-distribution with d degrees of freedom, $\phi(u) = (1 + u/d)^{-(d+1)/2}$) .

Left side: level sets of $g(x) = x'Cx$
Right side: level sets of a density $c \cdot \phi((x-\mu)'C^{-1}(x-\mu))$



Regression inference with Gaussian errors

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

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

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

Regression inference with Gaussian errors

If we transform

$$Z = \mu + AX,$$

we get a random variable satisfying

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

Regression inference with Gaussian errors

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.

Regression inference with Gaussian errors

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

$$-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_i (x_i-\mu)'\Sigma^{-1}(x_i-\mu) = -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\mathrm{tr}\left(S_{xx}\Sigma^{-1}\right)$$

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, let $Z \sim N(0, I_p)$, and write

$$X = \mu + AZ$$

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

Properties of the multivariate normal distribution:

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

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

by the Gram-Schmidt procedure. Let

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

Properties of the multivariate normal distribution:

Therefore

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

and

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

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.

Properties of the multivariate normal distribution:

- 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^2_1 distribution. The χ^2_p distribution is defined to be the distribution of

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

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

The moments of the χ^2 distribution

By direct calculation, if $F \sim \chi^2_1$,

$$EF = 1 \qquad\qquad \mathrm{var}[F] = 2.$$

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

The density of the χ^2 distribution

The χ^2_p 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^2_{p-1} density is

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

The χ^2_p density is the density of the sum of a χ^2_{p-1} random variable and a χ^2_1 random variable, which can be written

The density of the χ^2 distribution

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

where $u = s/x$.

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

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

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, i.e.

$$QPQ' = \begin{pmatrix} \mathbf{I}_k & 0 \\ 0 & 0 \end{pmatrix},$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix.

The χ^2 distribution and the RSS

This gives us

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

which follows a χ^2_k distribution.

It follows that

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

The χ^2 distribution and the RSS

Thus when the errors are Gaussian, we have

$$\mathrm{var}[\hat{\sigma}^2] = \frac{2\sigma^4}{n-p-1}.$$

and

$$\mathrm{SD}[\hat{\sigma}^2] = \sigma^2 \left(\frac{2}{n-p-1}\right)^{1/2}.$$

The t distribution

Suppose Z is standard normal, $V \sim \chi^2_p$, 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 t distribution

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

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

The t distribution

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

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

Independence of estimated mean and variance parameters

To review, the linear model residuals are

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

and the estimated coefficients are

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

Independence of estimated mean and variance parameters

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

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

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

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

so that

$$\mathrm{var}[\hat{\beta}_k|\mathbf{X}] = \sigma^2 V_k.$$

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

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

Confidence intervals for a regression coefficient

Therefore

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

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

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

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

Confidence intervals for a regression coefficient

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

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

which has coverage probability α.

Confidence intervals 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

$$\mathrm{var}[\hat{\beta}' x^*] = \sigma^2 x^{*\prime}(\mathbf{X}'\mathbf{X})^{-1} x^* \equiv \sigma^2 V_{x^*}.$$

Confidence intervals for the expected response

As above we have that

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

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

Therefore

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

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:

$$E[Y^* - \hat{\beta}' x^* | \mathbf{X}] = \beta' x^* - \beta' x^* = 0$$

and

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

Prediction intervals

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

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

Wald tests

Suppose we want to carry out a formal hypothesis test for the null hypothesis $\beta_k = 0$, for some specified index k.

Since $\mathrm{var}(\hat{\beta}_k|\mathbf{X}) = \sigma^2 V_k$, the "Z-score"

$$\hat{\beta}_k / \sqrt{\sigma^2 V_k}$$

follows a standard normal distribution under the null hypothesis. The "Z-test" or "asymptotic Wald test" rejects the null hypothesis if $|\hat{\beta}_k| / \sqrt{\sigma^2 V_k} > F^{-1}(1 - \alpha/2)$, where F is the standard normal cumulative distribution function (CDF).

More generally, we can also test hypotheses of the form $\beta_k = c$ using the Z-score

$$(\hat{\beta}_k - c) / \sqrt{\sigma^2 V_k}$$

Wald tests

If the errors are normally distributed, then

$$\hat{\beta}_k / \sqrt{\hat{\sigma}^2 V_k}$$

follows a t-distribution with $n - p - 1$ degrees of freedom, and the Wald test rejects the null hypothesis if $|\hat{\beta}_k| / \sqrt{\sigma^2 V_k} > F_t^{-1}(1 - \alpha/2, n - p - 1)$, where $F_t(\cdot, d)$ is the student t CDF with d degrees of freedom, and α is the type-I error probability (e.g. $\alpha = 0.05$).

Wald tests for contrasts

More generally, we can consider the contrast $\theta'\beta$ defined by a fixed vector $\theta \in \mathcal{R}^{p+1}$. The population value of the contrast can be estimated by the plug-in estimate $\theta'\hat{\beta}$.

We can test the null hypothesis $\theta'\beta = 0$ with the Z-score

$$\theta'\hat{\beta}/\sqrt{\hat{\sigma}^2\theta' V\theta}.$$

This approximately follows a standard normal distribution, and more "exactly" (under specified assumptions) it follows a student t-distribution with $n - p - 1$ degrees of freedom (all under the null hypothesis).

F-tests

Suppose we have two nested design matrices $\mathbf{X}_1 \in \mathcal{R}^{n \times p_1}$ and $\mathbf{X}_2 \in \mathcal{R}^{n \times p_2}$, such that

$$\mathrm{col}(\mathbf{X}_1) \subset \mathrm{col}(\mathbf{X}_2).$$

We may wish to compare the model defined by $\mathbf{X}_1$ to the model defined by $\mathbf{X}_2$. To do this, we need a test statistic that discriminates between the two models.

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.

F-tests

Due to the nesting, $P_2 P_1 = P_1$ and $P_1 P_2 = P_1$. Therefore

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

so $P_2 - P_1$ is a projection that projects onto $\mathrm{col}(\mathbf{X}_2) - \mathrm{col}(\mathbf{X}_1)$, the complement of $\mathrm{col}(\mathbf{X}_1)$ in $\mathrm{col}(\mathbf{X}_2)$.

Since

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

it follows that if $E[y] \in \mathrm{col}(\mathbf{X}_2)$ then

$$\mathrm{Cov}(y - \hat{y}^{(2)}, \hat{y}^{(2)} - \hat{y}^{(1)}) = E[(I - P_2) YY' (P_1 - P_2)] = 0.$$

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

F-tests

Since $P_2\mathbf{X}_1 = P_1\mathbf{X}_1 = \mathbf{X}_1$, we have

$$(I - P_2)\mathbf{X}_1 = (P_2 - P_1)\mathbf{X}_1 = 0.$$

Now suppose we take as the null hypothesis that $E[y] \in \mathrm{col}(\mathbf{X}_1)$, so we can write $y = \theta + \epsilon$, where $\theta \in \mathrm{col}(\mathbf{X}_1)$. Therefore under the null hypothesis

$$\|y - \hat{y}^{(2)}\|^2 = \mathrm{tr}[(I - P_2)yy'] = \mathrm{tr}[(I - P_2)\epsilon\epsilon']$$

and

$$\|\hat{y}^{(2)} - \hat{y}^{(1)}\|^2 = \mathrm{tr}[(P_2 - P_1)yy'] = \mathrm{tr}[(P_2 - P_1)\epsilon\epsilon'].$$

F-tests

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

$$\|y - \hat{y}^{(2)}\|^2/\sigma^2 = \|(I - P_2)y\|^2/\sigma^2 \sim \chi^2_{n-p_2}$$

and under the null hypothesis

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

Therefore

$$\frac{\|\hat{y}^{(2)} - \hat{y}^{(1)}\|^2/(p_2 - p_1)}{\|y - \hat{y}^{(2)}\|^2/(n - p_2)}.$$

Since $\|\hat{y}^{(2)} - \hat{y}^{(1)}\|^2$ will tend to be large when $E[y] \notin \mathrm{col}(\mathbf{X}_1)$, i.e. when the null hypothesis is false, this quantity can be used as a test-statistic. It is called the F-test statistic.

The F distribution

The F-test statistic is the ratio between two rescaled independent χ^2 draws.

If $U \sim \chi^2_p$ and $V \sim \chi^2_q$, 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 F distribution

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

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

The F distribution

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

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

The F distribution

Now if we let

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

then

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

The F distribution

Therefore, the F-test statistic follows an $F_{p_2-p_1,n-p_2}$ distribution

$$\frac{\|\hat{y}^{(2)} - \hat{y}^{(1)}\|^2/(p_2 - p_1)}{\|y - \hat{y}^{(2)}\|^2/(n - p_2)} \sim F_{p_2-p_1,n-p_2}.$$

Simultaneous confidence intervals

If θ is a fixed vector, we can cover the value $\theta'\beta$ with probability α by pivoting on the t_{n-p-1}-distributed pivotal quantity

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

where

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

Now suppose we have a set $\mathcal{T} \subset \mathcal{R}^{p+1}$ 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 set of simultaneous confidence intervals for $\{\theta'\beta; \theta \in \mathcal{T}\}$.

The Bonferroni approach to simultaneous confidence intervals

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 the union bound implies that

$$\begin{aligned} P(I_1 \text{ and } I_2 \cdots \text{and } I_k) &= 1 - P(I_1' \text{ or } I_2' \cdots \text{or } I_k') \\ &\geq 1 - \sum_j P(I_j'). \end{aligned}$$

The Bonferroni approach to simultaneous confidence intervals

As long as

$$1 - \alpha \geq \sum_j P(I'_j),$$

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 to simultaneous confidence intervals

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

Begin with the pivotal quantity

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

and postulate that a symmetric interval can be found so that

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

Equivalently, we can write

$$P\left(\sup_{\theta\in\mathcal{T}} \frac{(\theta'\hat{\beta} - \theta'\beta)^2}{\hat{\sigma}^2\theta'(\mathbf{X}'\mathbf{X})^{-1}\theta} \leq Q_\alpha^2\right) = \alpha.$$

The Scheffé approach approach to simultaneous confidence intervals

Since

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

we have

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

where

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

The Scheffé approach to simultaneous confidence intervals

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

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

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

The Scheffé approach to simultaneous confidence intervals

Moreover,

$$\|P\epsilon\|^2/\sigma^2 \sim \chi^2_q$$

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

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

Thus

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

The Scheffé approach to simultaneous confidence intervals

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

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

so

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

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

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

The Scheffé approach to simultaneous confidence intervals

The "multiplier" for the Scheffé simultaneous confidence interval is

$$\hat{\sigma}\sqrt{qQ_F V_\theta}$$

where the F distribution has $q, n-p-1$ degrees of freedom. For large n, we can approximate this with

$$\hat{\sigma}\sqrt{Q_{\chi^2} V_\theta}$$

where χ^2 is a χ^2 distribution with q degrees of freedom.

Instead of the usual factor of 2, we have $\sqrt{Q_{\chi^2}}$. Note that this equals 2 when $q = 1$, and grows fairly slowly with q, i.e. it is 3.3 when $q = 5$ and 4.3 when $q = 10$.

Polynomial regression

The conventional linear model has the mean specification

$$E[Y|X = x] = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p.$$

It is possible to accomodate nonlinear relationships while still working with linear models.

Polynomial regression is a traditional approach to doing this. If there is only one predictor variable, polynomial reqression uses the mean structure

$$E[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p.$$

Note that this is still a linear model, as it is linear in the coefficients β. Multiple regression techniques (e.g. OLS) can be used for estimation and inference.

Functional linear regression

A drawback of polynomial regression is that the polynomials can be highly colinear (e.g. $\text{cor}(U, U^2) \approx 0.97$ if U is uniform on $0, 1$). Also, polynomials have unbounded support, and it is often desirable to model $E[Y|X = x]$ as a linear combination of functions with local (bounded) suport.

If there is a single covariate X, we can use a model of the form

$$E[Y|X = x] = \beta_1\phi_1(x) + \cdots + \beta_p\phi_p(x)$$

where the $\phi_j(\cdot)$ are basis functions that are chosen based on what we think $E[Y|X = x]$ might look like. Splines, sinusoids, wavelets, and radial basis functions are possible choices.

Since the mean function is linear in the unknown parameters β_j, this is a linear model and can be estimated using multiple linear regression (OLS) techniques.

Confidence bands

Suppose we have a functional linear model of the form

$$E[Y|X = x, t] = \beta' x + f(t),$$

and we model $f(t)$ as

$$f(t) = \sum_{j=1}^{q} \gamma_j \phi_j(t)$$

where the $\phi_j(\cdot)$ are basis functions.

A confidence band for f with coverage probability α is an expression of the form

$$\hat{f}(t) \pm M(t)$$

such that

$$P\left(\hat{f}(t) - M(t) \leq f(t) \leq \hat{f}(t) + M(t) \quad \forall t\right) = \alpha.$$

Confidence bands

We can use as a point estimate

$$\hat{f}(t) \equiv \sum_{j=1}^{q} \hat{\gamma}_j \phi_j(t)$$

For each fixed t, $\sum_{j=1}^{q} \gamma_j \phi_j(t)$ is a linear combniation of the $\hat{\gamma}_j$, so if we simultaneously cover all linear combinations of the γ_j, we will have our confidence band. Thus the Scheffé procedure can be applied with $\mathcal{T} = \mathcal{R}^q$.