

# Least squares fitting and inference for linear models

Kerby Shedden

Department of Statistics, University of Michigan

September 25, 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, exogenous 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):

$$\begin{aligned}y_i &\in \mathcal{R} \\x_i &= (x_{i1}, \dots, x_{ip})' \in \mathcal{R}^p \\i &= 1, \dots, n.\end{aligned}$$

where  $n$  is the sample size.

The data point  $(y_i, x_i)$  reflects the  $i^{\text{th}}$  “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  $\beta$  for fixed  $x$ , and it is linear in  $x$  for fixed  $\beta$  (technically, it is “bilinear”).

# Least Squares Fitting

To estimate  $f$ , we need to estimate the  $\beta_j$ . One approach to doing this is to minimize the following function of  $\beta$ :

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

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  $\alpha$  and  $\beta$ :

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

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

is the “working residual” (requires working values for  $\alpha$  and  $\beta$ ).

# Simple linear regression

Setting

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

and solving for  $\alpha$  and  $\beta$  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  $\alpha$  and  $\beta$  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  $\alpha$  and  $\beta$ .

At the least squares solution,

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

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:

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

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

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

## Simple linear regression

Note that

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

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

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

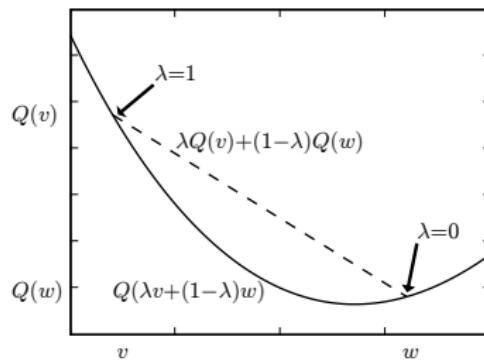
where  $\widehat{\text{cov}}$  and  $\widehat{\text{var}}$  are the usual unbiased estimates of variance and covariance.

# Convex functions

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**.



## 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

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

## Convexity of quadratic functions

If  $b \in \text{col}(A)$ , 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

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

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

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

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

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  $v_\ell$  to get the  $\ell^{\text{th}}$  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  $\ell^{\text{th}}$  component of  $2v'A'$  is the inner product of  $v$  with the  $\ell^{\text{th}}$  row (or column) of  $A$ . The derivative of this inner product with respect to a second index  $\ell'$  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  $\widehat{\text{var}}[x]$  (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  $x. = \sum_i x_i$ .

## Uniqueness of the simple linear regression least squares fit

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

$$|2n| > 0$$

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

# 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

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

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

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

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

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

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 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**:

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

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

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

where  $\beta_0$  is the intercept and  $\beta_k$  is the slope corresponding to the  $k^{\text{th}}$  covariate. For a given working covariate vector  $\beta$ , 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  $\beta$  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

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

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

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

# 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  $\nabla f$  is evaluated at  $g(z)$ .

## The least squares gradient function

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

$$\mathbf{X} = \mathbf{Q}\mathbf{R}$$

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:

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

# 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**

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

## Matrix products

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

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

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**)

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

then

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

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  $V$ :

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

# 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

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

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  $\theta$  and  $V$ . Using calculus, we differentiate with respect to  $\gamma$  and solve for the stationary point:

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

The solution is

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

so

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

# 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  $\eta \in \text{col}(\mathbf{X})$ .

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

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  $\text{col}(\mathbf{X})$  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  $\text{col}(\mathbf{X})$  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  $\text{col}(\mathbf{X})$ .

## 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  $i^{\text{th}}$  row of  $\mathbf{X}$  and  $y_i$  is the  $i^{\text{th}}$  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  $S = \text{col}(X)$ . The sum of residuals can be written

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

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

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

so

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

## Orthogonal matrices

An **orthogonal matrix**  $\mathbf{X}$  satisfies  $\mathbf{X}'\mathbf{X} = \mathbf{I}$ . 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  $\mathbf{XX}' = \mathbf{I}$ .

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

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

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  $X = (X_1, \dots, X_p)$  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  $i^{\text{th}}$  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

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

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

# 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 ( $E[Y|X = x]$ ,  $\text{var}[Y|X = x]$ ) 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  $\epsilon$  is a random variable with expected value zero.

If we specify the distribution of  $\epsilon$ , 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  $\theta$  is a finite dimensional parameter vector.

Examples:

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

Models 1 and 2 are both “linear models” because they are linear in  $\theta$ .  
The Gompertz curve is a non-linear model because it is not linear in  $\theta$ .

## 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  $\text{var}[\epsilon|x] = \sigma^2$  exists and is constant across cases
- iii the  $\epsilon$  random variables are uncorrelated across cases (given  $x$ ).

We will not assume here that  $\epsilon$  follows a particular distribution, e.g. a Gaussian distribution.

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

If we treat  $X$  as 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:

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

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

then  $E\epsilon = 0$  and  $\text{cor}(X, \epsilon) = 0$ , but  $E[\epsilon|X]$  is not identically zero. When  $\epsilon$  and  $X$  are jointly Gaussian,  $\text{cor}(\epsilon, X) = 0$  implies that  $\epsilon$  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 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 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 as 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

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

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

Note that we are using the useful fact that

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

## 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:

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

since  $\text{cov}(\epsilon_i, \bar{\epsilon}|X) = \sigma^2/n$  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

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

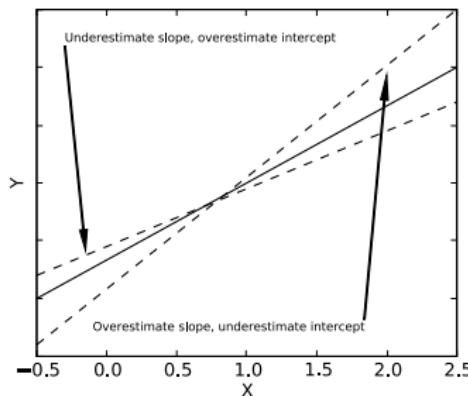
$$\begin{aligned}\text{var}[\hat{\alpha}|X] &= \text{var}[\bar{y} - \hat{\beta}\bar{x}|X] \\ &= \text{var}[\alpha + \beta\bar{x} + \bar{\epsilon} - \hat{\beta}\bar{x}|X] \\ &= \text{var}[\bar{\epsilon}|X] + \bar{x}^2 \text{var}[\hat{\beta}|X] - 2\bar{x} \text{cov}[\bar{\epsilon}, \hat{\beta}|X] \\ &=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

# Sampling covariance of parameter estimates

The **sampling covariance** between the slope and intercept is

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

When  $\bar{x} > 0$ , it's easy to see what the expression for  $\text{cov}(\hat{\alpha}, \hat{\beta}|X)$  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  $\bar{x}$ .
- ▶  $\text{var}[\hat{\alpha}]$  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  $Y_i = \alpha + \beta x_i + \epsilon_i$  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”  $\epsilon_i$ . 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  $\beta$  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  $i^{\text{th}}$  residual is

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

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.
- ▶  $\text{var}[R_i] < \text{var}[\epsilon_i]$  – the residuals are less variable than the errors.

## Some properties of residuals

Since  $E[R_i] = 0$  it follows that  $\text{var}[R_i] = E[R_i^2]$ . 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 $\sigma^2$

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  $\sigma^2$ . 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

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

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

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

Let  $u_i$  be the  $i^{\text{th}}$  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

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

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

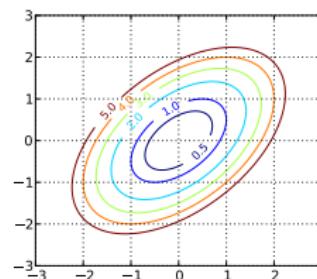
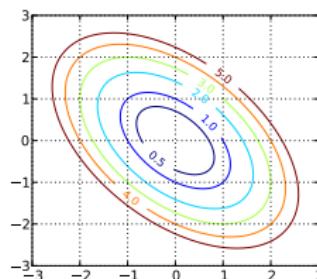
$$C = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

so that

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

Left side: level curves of  $g(x) = x' C x$

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”

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

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

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

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:

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

where  $P$  is the projection matrix onto  $\text{col}(\mathbf{X})$ . 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  $\text{RSS} = Y'(I - P)Y$  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 circulent property states that

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

## The expected value of the RSS

By linearity we have

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

and

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

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  $E[\text{RSS}] = \sigma^2\text{tr}(I - P)$ .

## 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

**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  $\lambda, v$  is an eigenvalue/eigenvector pair:

$$Pv = \lambda v.$$

If  $P$  is the projection onto a subspace  $S$ , this implies that  $\lambda v$  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  $\lambda v$ , 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  $E[\text{RSS}] = \sigma^2 \text{tr}(I - P)$ . Since  $I - P$  is the projection onto  $\text{col}(\mathbf{X})^\perp$ ,  $I - P$  has rank  $n - \text{rank}(\mathbf{X}) = n - p - 1$ . Thus

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

so

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

is an unbiased estimate of  $\sigma^2$ .

## 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.

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

# Distribution of the RSS

The RSS can be written

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

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

## Distribution of the RSS

If the distribution of  $\epsilon$  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  $\text{col}(I - P)$  to  $\mathcal{S}$ ). Then

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

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

# Optimality

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

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

for all  $\beta$ . 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  $\beta^* = M\mathbf{Y}$  be any linear unbiased estimator of  $\beta$ , when

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

this implies that for any fixed vector  $\theta$ ,

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

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  $\beta^*$  of  $\beta$ ,  $\beta^* - \hat{\beta}$  and  $\hat{\beta}$  are uncorrelated. It follows that

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

To prove the theorem note that

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

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

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

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

# Optimality (proof of GMT)

Therefore

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

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

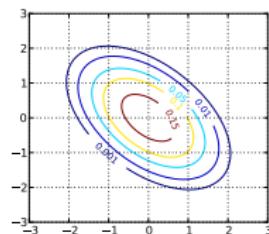
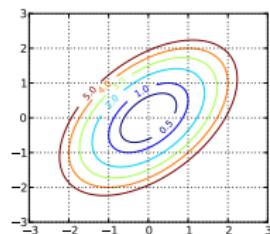
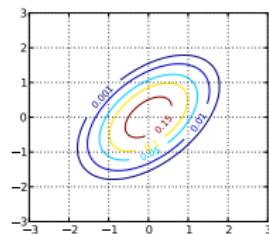
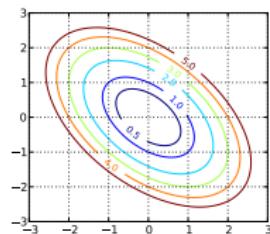
$$\text{cov}(\beta^* - \hat{\beta} | \mathbf{X}) = \sigma^2(M - B)(M - B)'.$$

# 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  $\phi : \mathcal{R}^+ \rightarrow \mathcal{R}^+$  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  $g(x) = x' C x$

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



## Regression inference with Gaussian errors

The random vector  $X = (X_1, \dots, X_p)'$  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:

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

# Regression inference with Gaussian errors

If we transform

$$Z = \mu + AX,$$

we get a random variable satisfying

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

## 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

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

where

$$S_{xx} = \sum_i (x_i - \mu)(x_i - \mu)' / n.$$

# The Cholesky decomposition

If  $\Sigma$  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  $\theta$  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  $\theta$  to a square invertible matrix

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

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

The matrix  $\theta^*$  can be chosen such that

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

by the Gram-Schmidt procedure. Let

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

## Properties of the multivariate normal distribution:

Therefore

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

and

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

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 $\chi^2$ 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  $\chi_1^2$  distribution. The  $\chi_p^2$  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.

## The moments of the $\chi^2$ distribution

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

$$EF = 1 \quad \text{var}[F] = 2.$$

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

## The $\chi^2$ 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 $\chi^2$ distribution and the RSS

This gives us

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

which follows a  $\chi_k^2$  distribution.

It follows that

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

# The $\chi^2$ distribution and the RSS

Thus when the errors are Gaussian, we have

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

and

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

## The $t$ distribution

Suppose  $Z$  is standard normal,  $V \sim \chi_p^2$ , 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

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

## 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

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

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

Since  $\hat{\sigma}^2$  is a function of the residuals, it follows that if  $\epsilon$  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

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

If the  $\epsilon$  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

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

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  $q^{\text{th}}$  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  $\alpha$ .

## 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

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

## 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  $\alpha$ .

## 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  $\alpha$  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  $\text{var}(\hat{\beta}_k | \mathbf{X}) = \sigma^2 V_k$ , 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  $|\hat{\beta}_k| / \sqrt{\hat{\sigma}^2 V_k} > F_t^{-1}(1 - \alpha/2, n - p - 1)$ , where  $F_t(\cdot, d)$  is the student t CDF with  $d$  degrees of freedom, and  $\alpha$  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

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

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,  $P_2P_1 = P_1$  and  $P_1P_2 = P_1$ . Therefore

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

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

Since

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

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

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

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

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

and

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

## 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

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

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

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

Since  $\|\hat{y}^{(2)} - \hat{y}^{(1)}\|^2$  will tend to be large when  $E[y] \notin \text{col}(\mathbf{X}_1)$ , 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

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  $\theta$  is a fixed vector, we can cover the value  $\theta' \beta$  with probability  $\alpha$  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  $\mathcal{T} \subset \mathbb{R}^{p+1}$  of vectors  $\theta$ , 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

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

## 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  $q_\alpha$ .

## The Scheffé approach to simultaneous confidence intervals

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

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

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

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  $\epsilon$  onto the line spanned by  $M_\theta$  (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,

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

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

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

## The Scheffé approach to simultaneous confidence intervals

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

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

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  $\alpha$  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

$$\hat{\sigma} \sqrt{q Q_F V_\theta}$$

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  $\chi^2$  is a  $\chi^2$  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

$$E[Y|X_1, \dots, X_q] = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p.$$

It is possible to accommodate 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 regression uses the mean structure

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

Note that this is still a linear model, as it is linear in the coefficients  $\beta$ . 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 are unbounded and it is often desirable to model  $E[Y|X]$  as a bounded function.

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

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

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

Since the mean function is linear in the unknown parameters  $\beta_j$ , 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  $\alpha$  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 combination of the  $\hat{\gamma}_j$ , so if we simultaneously cover all linear combinations of the  $\gamma_j$ , we will have our confidence band. Thus the Scheffé procedure can be applied with  $\mathcal{T} = \mathcal{R}^q$ .

# Decomposing Variance

Kerby Shedden

Department of Statistics, University of Michigan

October 9, 2019

## Law of total variation

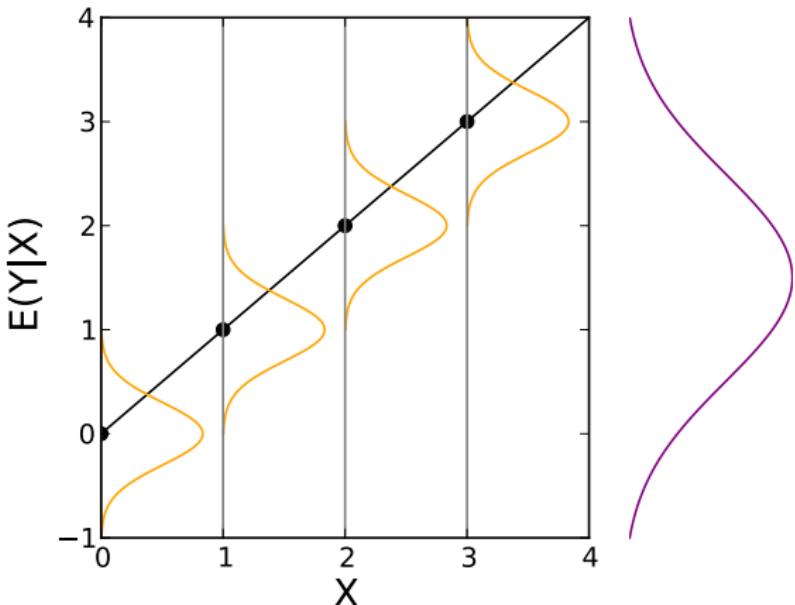
For any regression model involving a response  $Y \in \mathcal{R}$  and a covariate vector  $X \in \mathcal{R}^P$ , we can decompose the marginal variance of  $Y$  as follows:

$$\text{var}(Y) = \text{var}_X E(Y|X) + E_X \text{var}(Y|X).$$

- ▶ If the population is **homoscedastic**,  $\text{var}(Y|X)$  does not depend on  $X$ , so we can simply write  $\text{var}(Y|X) = \sigma^2$ , and we get  
$$\text{var}(Y) = \text{var}_X E(Y|X) + \sigma^2.$$
- ▶ If the population is **heteroscedastic**,  $\text{var}(Y|X)$  is a function  $\sigma^2(X)$  with expected value  $\sigma^2 = E_X \sigma^2(X)$ , and again we get  
$$\text{var}(Y) = \text{var}_X E(Y|X) + \sigma^2.$$

If we write  $Y = f(X) + \epsilon$  with  $E(\epsilon|X) = 0$ , then  $E(Y|X) = f(X)$ , and  $\text{var}_X E(Y|X)$  summarizes the variation of  $f(X)$  over the marginal distribution of  $X$ .

## Law of total variation



**Orange curves:** conditional distributions of  $Y$  given  $X$

**Purple curve:** marginal distribution of  $Y$

**Black dots:** conditional means of  $Y$  given  $X$

## Pearson correlation

The population Pearson correlation coefficient of two jointly distributed scalar-valued random variables  $X$  and  $Y$  is

$$\rho_{XY} \equiv \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Given data  $y = (y_1, \dots, y_n)'$  and  $x = (x_1, \dots, x_n)'$ , the Pearson correlation coefficient is estimated by

$$\hat{\rho}_{xy} = \frac{\widehat{\text{cov}}(x, y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}} = \frac{(x - \bar{x})'(y - \bar{y})}{\|x - \bar{x}\| \cdot \|y - \bar{y}\|}.$$

When we write  $y - \bar{y}$  here, this means  $y - \bar{y} \cdot \mathbf{1}$ , where  $\mathbf{1}$  is a vector of 1's, and  $\bar{y}$  is a scalar.

## Pearson correlation

By the Cauchy-Schwartz inequality,

$$\begin{aligned}-1 &\leq \rho_{xy} \leq 1 \\ -1 &\leq \hat{\rho}_{xy} \leq 1.\end{aligned}$$

The sample correlation coefficient is slightly biased, but the bias is so small that it is usually ignored.

# Pearson correlation and simple linear regression slopes

For the simple linear regression model

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

if we view  $X$  as a random variable that is uncorrelated with  $\epsilon$ , then

$$\text{cov}(X, Y) = \beta \sigma_X^2$$

and the correlation is

$$\rho_{XY} \equiv \text{cor}(X, Y) = \frac{\beta}{\sqrt{\beta^2 + \sigma^2/\sigma_X^2}}.$$

The sample correlation coefficient for data  $x = (x_1, \dots, x_n)$  and  $y = (y_1, \dots, y_n)$  is related to the least squares slope estimate:

$$\hat{\beta} = \frac{\widehat{\text{cov}}(x, y)}{\hat{\sigma}_x^2} = \hat{\rho}_{xy} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}.$$

## Orthogonality between fitted values and residuals

Recall that the fitted values are

$$\hat{y} = x\hat{\beta} = Py$$

where  $y \in \mathcal{R}^n$  is the vector of observed responses, and  $P \in \mathcal{R}^{n \times n}$  is the projection matrix onto  $\text{col}(\mathbf{X})$ .

The residuals are

$$r = y - \hat{y} = (I - P)y \in \mathcal{R}^n.$$

Since  $P(I - P) = \mathbf{0}_{n \times n}$  it follows that  $\hat{y}'r = 0$ .

since  $\bar{r} = 0$ , it is equivalent to state that the sample correlation between  $r$  and  $\hat{y}$  is zero, i.e.

$$\widehat{\text{cor}}(r, \hat{y}) = 0.$$

## Coefficient of determination

A descriptive summary of the explanatory power of  $x$  for  $y$  is given by the coefficient of determination, also known as the proportion of explained variance, or multiple  $R^2$ . This is the quantity

$$R^2 \equiv 1 - \frac{\|y - \hat{y}\|^2}{\|y - \bar{y}\|^2} = \frac{\|\hat{y} - \bar{y}\|^2}{\|y - \bar{y}\|^2} = \frac{\widehat{\text{var}}(\hat{y})}{\widehat{\text{var}}(y)}.$$

The equivalence between the two expressions follows from the identity

$$\begin{aligned}\|y - \bar{y}\|^2 &= \|y - \hat{y} + \hat{y} - \bar{y}\|^2 \\&= \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}\|^2 + 2(y - \hat{y})'(\hat{y} - \bar{y}) \\&= \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}\|^2,\end{aligned}$$

It should be clear that  $R^2 = 0$  iff  $\hat{y} = \bar{y}$  and  $R^2 = 1$  iff  $\hat{y} = y$ .

# Coefficient of determination

The coefficient of determination is equal to

$$\widehat{\text{cor}}(\hat{y}, y)^2.$$

To see this, note that

$$\begin{aligned}\widehat{\text{cor}}(\hat{y}, y) &= \frac{(\hat{y} - \bar{y})'(y - \bar{y})}{\|\hat{y} - \bar{y}\| \cdot \|y - \bar{y}\|} \\ &= \frac{(\hat{y} - \bar{y})'(y - \hat{y} + \hat{y} - \bar{y})}{\|\hat{y} - \bar{y}\| \cdot \|y - \bar{y}\|} \\ &= \frac{(\hat{y} - \bar{y})'(y - \hat{y}) + (\hat{y} - \bar{y})'(\hat{y} - \bar{y})}{\|\hat{y} - \bar{y}\| \cdot \|y - \bar{y}\|} \\ &= \frac{\|\hat{y} - \bar{y}\|}{\|y - \bar{y}\|}.\end{aligned}$$

# Coefficient of determination in simple linear regression

In general,

$$R^2 = \widehat{\text{cor}}(y, \hat{y})^2 = \frac{\widehat{\text{cov}}(y, \hat{y})^2}{\widehat{\text{var}}(y) \cdot \widehat{\text{var}}(\hat{y})}.$$

In the case of simple linear regression,

$$\begin{aligned}\widehat{\text{cov}}(y, \hat{y}) &= \widehat{\text{cov}}(y, \hat{\alpha} + \hat{\beta}x) \\ &= \hat{\beta} \widehat{\text{cov}}(y, x),\end{aligned}$$

and

$$\begin{aligned}\widehat{\text{var}}(\hat{y}) &= \widehat{\text{var}}(\hat{\alpha} + \hat{\beta}x) \\ &= \hat{\beta}^2 \widehat{\text{var}}(x)\end{aligned}$$

Thus for simple linear regression,  $R^2 = \widehat{\text{cor}}(y, x)^2 = \widehat{\text{cor}}(y, \hat{y})^2$ .

## Relationship to the F statistic

The F-statistic for the null hypothesis

$$\beta_1 = \dots = \beta_p = 0$$

is

$$\frac{\|\hat{y} - \bar{y}\|^2}{\|y - \hat{y}\|^2} \cdot \frac{n - p - 1}{p} = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p},$$

which is an increasing function of  $R^2$ .

## Adjusted $R^2$

The sample  $R^2$  is an estimate of the population  $R^2$ :

$$1 - \frac{E_X \text{var}(Y|X)}{\text{var}(Y)}.$$

Since it is a ratio, the plug-in estimate  $R^2$  is biased, although the bias is not large unless the sample size is small or the number of covariates is large. The adjusted  $R^2$  is an approximately unbiased estimate of the population  $R^2$ :

$$1 - (1 - R^2) \frac{n - 1}{n - p - 1}.$$

The adjusted  $R^2$  is always less than the unadjusted  $R^2$ . The adjusted  $R^2$  is always less than or equal to one, but can be negative.

## The unique variation in one covariate

How much “information” about  $y$  is present in a covariate  $x_k$ ? This question is not straightforward when the covariates are non-orthogonal, since several covariates may contain overlapping information about  $y$ .

Let  $x_k^\perp \in \mathcal{R}^n$  be the residual of the  $k^{\text{th}}$  covariates,  $x_k \in \mathcal{R}^n$ , after regressing it against all other covariates (including the intercept). If  $P_{-k}$  is the projection onto  $\text{span}(\{x_j, j \neq k\})$ , then

$$x_k^\perp = (I - P_{-k})x_k.$$

We could use  $\widehat{\text{var}}(x_k^\perp)/\widehat{\text{var}}(x_k)$  to assess how much of the variation in  $x_k$  is “unique” in that it is not also captured by other predictors.

But this measure doesn’t involve  $y$ , so it can’t tell us whether the unique variation in  $x_k$  is useful in the regression analysis.

## The unique regression information in one covariate

To learn how  $x_k$  contributes “uniquely” to the regression, we can consider how introducing  $x_k$  to a working regression model affects the  $R^2$ .

Let  $\hat{y}_{-k} = P_{-k}y$  be the fitted values in the model omitting covariate  $k$ .

Let  $R^2$  denote the multiple  $R^2$  for the full model, and let  $R_{-k}^2$  be the multiple  $R^2$  for the regression omitting covariate  $x_k$ . The value of

$$R^2 - R_{-k}^2$$

is a way to quantify how much unique information about  $y$  in  $x_k$  is not captured by the other covariates. This is called the **semi-partial  $R^2$** .

## Identity involving norms of fitted values and residuals

Before we continue, we will need a simple identity that is often useful.

In general, if  $a$  and  $b$  are orthogonal, then  $\|a + b\|^2 = \|a\|^2 + \|b\|^2$ .

If  $a$  and  $b - a$  are orthogonal, then

$$\|b\|^2 = \|b - a + a\|^2 = \|b - a\|^2 + \|a\|^2.$$

Thus in this setting we have  $\|b\|^2 - \|a\|^2 = \|b - a\|^2$ .

Applying this fact to regression, we know that the fitted values and residuals are orthogonal. Thus for the regression omitting variable  $k$ ,  $\hat{y}_{-k}$  and  $y - \hat{y}_{-k}$  are orthogonal, so  $\|y - \hat{y}_{-k}\|^2 = \|y\|^2 - \|\hat{y}_{-k}\|^2$ .

By the same argument,  $\|y - \hat{y}\|^2 = \|y\|^2 - \|\hat{y}\|^2$ .

## Improvement in $R^2$ due to one covariate

Now we can obtain a simple, direct expression for the semi-partial  $R^2$ .

Since  $x_k^\perp$  is orthogonal to the other covariates,

$$\hat{y} = \hat{y}_{-k} + \frac{\langle y, x_k^\perp \rangle}{\langle x_k^\perp, x_k^\perp \rangle} x_k^\perp,$$

and

$$\|\hat{y}\|^2 = \|\hat{y}_{-k}\|^2 + \langle y, x_k^\perp \rangle^2 / \|x_k^\perp\|^2.$$

## Improvement in $R^2$ due to one covariate

Thus we have

$$\begin{aligned} R^2 &= 1 - \frac{\|y - \hat{y}\|^2}{\|y - \bar{y}\|^2} \\ &= 1 - \frac{\|y\|^2 - \|\hat{y}\|^2}{\|y - \bar{y}\|^2} \\ &= 1 - \frac{\|y\|^2 - \|\hat{y}_{-k}\|^2 - \langle y, x_k^\perp \rangle^2 / \|x_k^\perp\|^2}{\|y - \bar{y}\|^2} \\ &= 1 - \frac{\|y - \hat{y}_{-k}\|^2}{\|y - \bar{y}\|^2} + \frac{\langle y, x_k^\perp \rangle^2 / \|x_k^\perp\|^2}{\|y - \bar{y}\|^2} \\ &= R_{-k}^2 + \frac{\langle y, x_k^\perp \rangle^2 / \|x_k^\perp\|^2}{\|y - \bar{y}\|^2}. \end{aligned}$$

## Semi-partial $R^2$

Thus the semi-partial  $R^2$  is

$$R^2 - R_{-k}^2 = \frac{\langle y, x_k^\perp \rangle^2 / \|x_k^\perp\|^2}{\|y - \bar{y}\|^2} = \frac{\langle y, x_k^\perp / \|x_k^\perp\| \rangle^2}{\|y - \bar{y}\|^2}.$$

Since  $x_k^\perp / \|x_k^\perp\|$  is centered and has length 1, it follows that

$$R^2 - R_{-k}^2 = \widehat{\text{cor}}(y, x_k^\perp)^2.$$

Thus the semi-partial  $R^2$  for covariate  $k$  has two interpretations:

- ▶ It is the improvement in  $R^2$  resulting from including covariate  $k$  in a working regression model that already contains the other covariates.
- ▶ It is the  $R^2$  for a simple linear regression of  $y$  on  $x_k^\perp = (I - P_{-k})x_k$ .

## Partial $R^2$

The **partial  $R^2$**  is

$$\frac{R^2 - R_{-k}^2}{1 - R_{-k}^2} = \frac{\langle y, x_k^\perp \rangle^2 / \|x_k^\perp\|^2}{\|y - \hat{y}_{-k}\|^2}.$$

The partial  $R^2$  for covariate  $k$  is the fraction of the maximum possible improvement in  $R^2$  that is contributed by covariate  $k$ .

Let  $\hat{y}_{-k}$  be the fitted values for regressing  $y$  on all covariates except  $x_k$ .

Since  $\hat{y}'_{-k} x_k^\perp = 0$ ,

$$\frac{\langle y, x_k^\perp \rangle^2}{\|y - \hat{y}_{-k}\|^2 \cdot \|x_k^\perp\|^2} = \frac{\langle y - \hat{y}_{-k}, x_k^\perp \rangle^2}{\|y - \hat{y}_{-k}\|^2 \cdot \|x_k^\perp\|^2}$$

The expression on the left is the usual  $R^2$  that would be obtained when regressing  $y - \hat{y}_{-k}$  on  $x_k^\perp$ . Thus the partial  $R^2$  is the same as the usual  $R^2$  for  $(I - P_{-k})y$  regressed on  $(I - P_{-k})x_k$ .

## Decomposition of projection matrices

Suppose  $P \in \mathcal{R}^{n \times n}$  is a rank- $d$  projection matrix, and  $U$  is a  $n \times d$  orthogonal matrix whose columns span  $\text{col}(P)$ . If we partition  $U$  by columns

$$U = \begin{pmatrix} & & \cdots & \\ | & U_1 & U_2 & \cdots & U_d \\ & | & | & \cdots & | \end{pmatrix},$$

then  $P = UU'$ , so we can write

$$P = \sum_{j=1}^d U_j U'_j.$$

Note that this representation is not unique, since there are different orthogonal bases for  $\text{col}(P)$ .

Each summand  $U_j U'_j \in \mathcal{R}^{n \times n}$  is a rank-1 projection matrix onto  $\langle U_j \rangle$ .

## Decomposition of $R^2$

**Question:** In a multiple regression model, how much of the variance in  $y$  is explained by a particular covariate?

**Orthogonal case:** If the design matrix  $\mathbf{X}$  is orthogonal ( $\mathbf{X}'\mathbf{X} = I$ ), the projection  $P$  onto  $\text{col}(\mathbf{X})$  can be decomposed as

$$P = \sum_{j=0}^p P_j = \frac{\mathbf{1}\mathbf{1}'}{n} + \sum_{j=1}^p x_j x_j',$$

where  $x_j$  is the  $j^{\text{th}}$  column of the design matrix (assuming here that the first column of  $\mathbf{X}$  is an intercept).

## Decomposition of $R^2$ (orthogonal case)

The  $n \times n$  rank-1 matrix

$$P_j = x_j x_j'$$

is the projection onto  $\text{span}(x_j)$  (and  $P_0$  is the projection onto the span of the vector of 1's). Furthermore, by orthogonality,  $P_j P_k = 0$  unless  $j = k$ . Since

$$\hat{y} - \bar{y} = \sum_{j=1}^p P_j y,$$

by orthogonality

$$\|\hat{y} - \bar{y}\|^2 = \sum_{j=1}^p \|P_j y\|^2.$$

Here we are using the fact that if  $U_1, \dots, U_m$  are orthogonal, then

$$\|U_1 + \cdots + U_m\|^2 = \|U_1\|^2 + \cdots + \|U_m\|^2.$$

## Decomposition of $R^2$ (orthogonal case)

The  $R^2$  for simple linear regression of  $y$  on  $x_j$  is

$$R_j^2 \equiv \|\hat{y} - \bar{y}\|^2 / \|y - \bar{y}\|^2 = \|P_j y\|^2 / \|y - \bar{y}\|^2,$$

so we see that for orthogonal design matrices,

$$R^2 = \sum_{j=1}^p R_j^2.$$

That is, the overall coefficient of determination is the sum of univariate coefficients of determination for all the explanatory variables.

## Decomposition of $R^2$

**Non-orthogonal case:** If  $\mathbf{X}$  is not orthogonal, the overall  $R^2$  will not be the sum of single covariate  $R^2$ 's.

If we let  $R_j^2$  be as above (the  $R^2$  values for regressing  $Y$  on each  $X_j$ ), then there are two different situations:  $\sum_j R_j^2 > R^2$ , and  $\sum_j R_j^2 < R^2$ .

## Decomposition of $R^2$

Case 1:  $\sum R_j^2 > R^2$

It's not surprising that  $\sum_j R_j^2$  can be bigger than  $R^2$ . For example, suppose that the population model is

$$Y = X_1 + \epsilon$$

is the data generating model, and  $X_2$  is highly correlated with  $X_1$  (but is not part of the data generating model).

For the regression of  $Y$  on both  $X_1$  and  $X_2$ , the multiple  $R^2$  will be  $1 - \sigma^2/\text{var}(Y)$  (since  $E(Y|X_1, X_2) = E(Y|X_1) = X_1$ ).

The  $R^2$  values for  $Y$  regressed on either  $X_1$  or  $X_2$  separately will also be approximately  $1 - \sigma^2/\text{var}(Y)$ .

Thus  $R_1^2 + R_2^2 \approx 2R^2$ .

## Decomposition of $R^2$

Case 2:  $\sum_j R_j^2 < R^2$

This is more surprising, and is sometimes called **enhancement**.

As an example, suppose the data generating model is

$$Y = Z + \epsilon,$$

but we don't observe  $Z$  (for simplicity assume  $EZ = 0$ ). Instead, we observe a value  $X_1$  that satisfies

$$X_1 = Z + X_2,$$

where  $X_2$  has mean 0 and is independent of  $Z$  and  $\epsilon$ .

Since  $X_2$  is independent of  $Z$  and  $\epsilon$ , it is also independent of  $Y$ , thus  $R_2^2 \approx 0$  for large  $n$ .

## Decomposition of $R^2$ (enhancement example)

The multiple  $R^2$  of  $Y$  on  $X_1$  and  $X_2$  is approximately  $\sigma_Z^2 / (\sigma_Z^2 + \sigma^2)$  for large  $n$ , since the fitted values will converge to  $\hat{Y} = X_1 - X_2 = Z$ .

To calculate  $R_1^2$ , first note that for the regression of  $y$  on  $x_1$ , where  $y, x_1 \in \mathbb{R}^n$  are data vectors

$$\hat{\beta} = \frac{\widehat{\text{cov}}(y, x_1)}{\widehat{\text{var}}(x_1)} \rightarrow \frac{\sigma_Z^2}{\sigma_Z^2 + \sigma_{X_2}^2}$$

and

$$\hat{\alpha} \rightarrow 0.$$

## Decomposition of $R^2$ (enhancement example)

Therefore for large  $n$ ,

$$\begin{aligned} n^{-1}\|y - \hat{y}\|^2 &\approx n^{-1}\|z + \epsilon - \sigma_Z^2 X_1 / (\sigma_Z^2 + \sigma_{X_2}^2)\|^2 \\ &= n^{-1}\|\sigma_{X_2}^2 z / (\sigma_Z^2 + \sigma_{X_2}^2) + \epsilon - \sigma_Z^2 x_2 / (\sigma_Z^2 + \sigma_{X_2}^2)\|^2 \\ &= \sigma_{X_2}^4 \sigma_Z^2 / (\sigma_Z^2 + \sigma_{X_2}^2)^2 + \sigma^2 + \sigma_Z^4 \sigma_{X_2}^2 / (\sigma_Z^2 + \sigma_{X_2}^2)^2 \\ &= \sigma_{X_2}^2 \sigma_Z^2 / (\sigma_Z^2 + \sigma_{X_2}^2) + \sigma^2. \end{aligned}$$

Therefore

$$\begin{aligned} R_1^2 &= 1 - \frac{n^{-1}\|y - \hat{y}\|^2}{n^{-1}\|y - \bar{y}\|^2} \\ &\approx 1 - \frac{\sigma_{X_2}^2 \sigma_Z^2 / (\sigma_Z^2 + \sigma_{X_2}^2) + \sigma^2}{\sigma_Z^2 + \sigma^2} \\ &= \frac{\sigma_Z^2}{(\sigma_Z^2 + \sigma^2)(1 + \sigma_{X_2}^2 / \sigma_Z^2)} \end{aligned}$$

## Decomposition of $R^2$ (enhancement example)

Thus

$$R_1^2/R^2 \approx 1/(1 + \sigma_{X_2}^2/\sigma_Z^2),$$

which is strictly less than one if  $\sigma_{X_2}^2 > 0$ .

Since  $R_2^2 = 0$ , it follows that  $R^2 > R_1^2 + R_2^2$ .

The reason for this is that while  $X_2$  contains no directly useful information about  $Y$  (hence  $R_2^2 = 0$ ), it can remove the “measurement error” in  $X_1$ , making  $X_1$  a better predictor of  $Z$ .

## Decomposition of $R^2$ (enhancement example)

We can now calculate the limiting partial  $R^2$  for adding  $X_2$  to a model that already contains  $X_1$ :

$$\frac{\sigma_{X_2}^2}{\sigma_{X_2}^2 + \sigma^2(1 + \sigma_{X_2}^2/\sigma_Z^2)}.$$

## Partial $R^2$ example 2

Suppose the design matrix satisfies

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

and the data generating model is

$$Y = X_1 + X_2 + \epsilon$$

with  $\text{var } \epsilon = \sigma^2$ .

## Partial $R^2$ example 2

We will calculate the partial  $R^2$  for  $X_1$ , using the fact that the partial  $R^2$  is the regular  $R^2$  for regressing

$$(I - P_{-1})y$$

on

$$(I - P_{-1})x_1$$

where  $y, x_1, x_2 \in \mathcal{R}^n$  are data vectors distributed like  $Y$ ,  $X_1$ , and  $X_2$ , and  $P_{-1}$  is the projection onto  $\text{span}(\{1, x_2\})$ .

Since this is a simple linear regression, the partial  $R^2$  can be expressed

$$\widehat{\text{cor}}((I - P_{-1})y, (I - P_{-1})x_1)^2.$$

## Partial $R^2$ example 2

We will calculate the partial  $R^2$  in a setting where all conditional means are linear. This would hold if the data are jointly Gaussian (but this is not a necessary condition for conditional means to be linear).

The numerator of the partial  $R^2$  is the square of

$$\begin{aligned}\widehat{\text{cov}}((I - P_{-1})y, (I - P_{-1})x_1) &= y'(I - P_{-1})x_1/n \\ &= (x_1 + x_2 + \epsilon)'(x_1 - rx_2)/n \\ &\rightarrow 1 - r^2.\end{aligned}$$

## Partial $R^2$ example 2

The denominator contains two factors. The first is

$$\begin{aligned}\|(I - P_{-1})x_1\|^2/n &= x_1'(I - P_{-1})x_1/n \\ &= x_1'(x_1 - rx_2)/n \\ &\rightarrow 1 - r^2.\end{aligned}$$

## Partial $R^2$ example 2

The other factor in the denominator is  $y'(I - P_{-1})y/n$ :

$$\begin{aligned} y'(I - P_{-1})y/n &= (x_1 + x_2)'(I - P_{-1})(x_1 + x_2)/n + \epsilon'(I - P_{-1})\epsilon/n + \\ &\quad 2\epsilon'(I - P_{-1})(x_1 + x_2)/n \\ &\approx (x_1 + x_2)'(x_1 - rx_2)/n + \sigma^2 \\ &\rightarrow 1 - r^2 + \sigma^2. \end{aligned}$$

Thus we get that the partial  $R^2$  is approximately equal to

$$\frac{1 - r^2}{1 - r^2 + \sigma^2}.$$

If  $r = 1$  then the result is zero ( $X_1$  has no unique explanatory power), and if  $r = 0$ , the result is  $1/(1 + \sigma^2)$ , indicating that after controlling for  $X_2$ , around  $1/(1 + \sigma^2)$  fraction of the remaining variance is explained by  $X_1$  (the rest is due to  $\epsilon$ ).

# Summary

Each of the three  $R^2$  values can be expressed either in terms of variance ratios, or as a squared correlation coefficient:

	Multiple $R^2$	Semi-partial $R^2$	Partial $R^2$
VR	$\ \hat{Y} - \bar{Y}\ ^2 / \ Y - \bar{Y}\ ^2$	$R^2 - R_{-k}^2$	$(R^2 - R_{-k}^2) / (1 - R_{-k}^2)$
Correlation	$\widehat{\text{cor}}(\hat{Y}, Y)^2$	$\widehat{\text{cor}}(Y, X_k^\perp)^2$	$\widehat{\text{cor}}((I - P_{-k})Y, X_k^\perp)^2$

# Specification Errors, Measurement Errors, Confounding

Kerby Shedden

Department of Statistics, University of Michigan

October 21, 2019

## An unobserved covariate

Suppose we have a data generating model of the form

$$Y = \alpha + \beta X + \gamma Z + \epsilon.$$

The usual conditions  $E[\epsilon|X = x, Z = z] = 0$  and  $\text{var}[\epsilon|X = x, Z = z] = \sigma^2$  hold.

The covariate  $X$  is observed, but  $Z$  is not observable.

If we regress  $y$  on  $x$ , the model we are fitting differs from the data generating model. What are the implications of this?

Does the fitted regression model  $\hat{y} = \hat{\alpha} + \hat{\beta}x$  estimate  $E[Y|X = x]$ , and does the MSE  $\hat{\sigma}^2$  estimate  $\text{var}[Y|X = x]$ ?

## An unobserved independent covariate

The simplest case is where  $X$  and  $Z$  are independent (and for simplicity  $E[Z] = 0$ ). The slope estimate  $\hat{\beta}$  has the form

$$\begin{aligned}\hat{\beta} &= \sum_i y_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 \\ &= \sum_i (\alpha + \beta x_i + \gamma z_i + \epsilon_i)(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 \\ &= \beta + \gamma \sum_i z_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 + \sum_i \epsilon_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2\end{aligned}$$

By the double expectation theorem,

$$E[\epsilon|X = x] = E_{Z|X} E[\epsilon|X = x, Z = z] = 0$$

and since  $Z$  and  $X$  are independent

$$E\left[\sum_i z_i(x_i - \bar{x})|x\right] = \sum_i (x_i - \bar{x}) E[z_i|x] = E[Z] \times \sum_i (x_i - \bar{x}) = 0.$$

## An unobserved independent covariate

Therefore  $\hat{\beta}$  remains unbiased if there is an unmeasured covariate  $Z$  that is independent of  $X$ .

What about  $\hat{\sigma}^2$ ? What does it estimate in this case?

## An unobserved independent covariate

The residuals are

$$(I - P)y = (I - P)(\gamma z + \epsilon)$$

So the residual sum of squares is

$$y'(I - P)y = \gamma^2 z'(I - P)z + \epsilon'(I - P)\epsilon + 2\gamma z'(I - P)\epsilon.$$

The expected value is therefore

$$\begin{aligned} E[y'(I - P)y|x] &= \gamma^2 \text{var}[Z]\text{rank}(I - P) + \sigma^2 \text{rank}(I - P) \\ &= (\gamma^2 \text{var}[Z] + \sigma^2)(n - 2). \end{aligned}$$

Hence the MSE of  $\hat{\sigma}^2$  has expected value  $\gamma^2 \text{var}(Z) + \sigma^2$ .

## An unobserved independent covariate

Are our inferences correct?

We can set  $\tilde{\epsilon} = \gamma z + \epsilon$  as being the error term of the model. Since

$$E[\tilde{\epsilon}|X=x] = 0 \quad \text{cov}[\tilde{\epsilon}|X=x] = (\gamma^2 \text{var}[Z] + \sigma^2)I \propto I,$$

all the results about estimation of  $\beta$  in a correctly-specified model hold in this setting.

In general, we may wish to view any unobserved covariate as simply being another source of error, like  $\epsilon$ . But we will see next that this cannot be done if  $Z$  is dependent with  $X$ .

# Confounding

As above, continue to take the data generating model to be

$$y = \alpha + \beta x + \gamma z + \epsilon,$$

but now suppose that  $X$  and  $Z$  are correlated.

As before,  $Z$  is not observed so our analysis will be based on  $Y$  and  $X$ .

A variable such as  $Z$  that is associated with both the dependent and independent variables in a regression model is called a **confounder**.

# Confounding

Suppose  $X$  and  $Z$  are standardized, and  $\text{cor}[X, Z] = r$ . Further suppose that  $E[Z|X] = rX$ .

Due to the linearity of  $E[Y|X, Z]$ :

- ▶ If  $X$  increases by one unit and  $Z$  remains fixed, the expected response increases by  $\beta$  units.
- ▶ If  $Z$  increases by one unit and  $X$  remains fixed, the expected response increases by  $\gamma$  units.

However, if we select a pair of cases with  $X$  values differing by one unit at random (without controlling  $Z$ ), their  $Z$  values will differ on average by  $r$  units. Therefore these expected responses for these two cases differ by  $\beta + r\gamma$  units.

## Known confounders

Suppose we are mainly interested in the relationship between a particular variable  $X$  and an outcome  $Y$ . A **measured confounder** is a variable  $Z$  that can be measured and included in a regression model along with  $X$ . A measured confounder generally does not pose a problem for estimating the effect of  $X$ , unless it is highly collinear with  $X$ .

For example, suppose we are studying the health effects of second-hand smoke exposure. We measure the health outcome ( $Y$ ) directly. Subjects who smoke are of course at risk for many of the same bad outcomes that may be associated with second-hand smoke exposure. Thus, it would be very important to determine which subjects smoke, and include that information as a covariate (a measured confounder) in a regression model used to assess the effects of second-hand smoke exposure.

## Unmeasured confounders

An **unmeasured confounder** is a variable that we know about, and for which we may have some knowledge of its statistical properties, but is not measured in a particular data set.

For example, we may know that certain occupations (like working in certain types of factories) may produce risks similar to the risks of exposure to second-hand smoke. If occupation data is not collected in a particular study, this is an unmeasured confounder.

Since we do not have data for unmeasured confounders, their omission may produce bias in the estimated effects for variables of interest. If we have some understanding of how a certain unmeasured confounder operates, we may be able to use a **sensitivity analysis** to get a rough idea of how much bias is present.

## Unknown confounders

An **unknown confounder** is a variable that affects the outcome of interest, but is unknown to us.

For example, there may be genetic factors that produce similar health effects as second-hand smoke, but we have no knowledge of which specific genes are involved.

## Randomization and confounding

Unknown confounders and unmeasured confounders place major limits on our ability to interpret regression models causally or mechanistically.

**Randomization:** One way to substantially reduce the risk of confounding is to randomly assign the values of  $X$ . In this case, there can be no systematic association between  $X$  and  $Z$ , and in large enough samples the actual (sample-level) association between  $X$  and  $Z$  will be very low, so very little confounding is possible.

In small samples, randomization cannot guarantee approximate orthogonality against unmeasured confounders.

# Confounding

For simplicity, suppose that  $Z$  has mean 0 and variance 1, and we use least squares to fit a working model

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

We can work out the limiting value of the slope estimate as follows.

$$\begin{aligned}\hat{\beta} &= \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{\sum_i (\alpha + \beta x_i + \gamma z_i + \epsilon_i)(x_i - \bar{x})/n}{\sum_i (x_i - \bar{x})^2/n} \\ &\rightarrow \beta + \gamma r.\end{aligned}$$

Note that if either  $\gamma = 0$  ( $Z$  is independent of  $Y$  given  $X$ ) or if  $r = 0$  ( $Z$  is uncorrelated with  $X$ ), then  $\beta$  is estimated correctly.

# Confounding

Since

$$\hat{\beta} \rightarrow \beta + \gamma r,$$

and it is easy to show that  $\hat{\alpha} \rightarrow \alpha$ , the fitted model is approximately

$$\hat{y} \approx \alpha + \beta x + \gamma rx = \alpha + (\beta + \gamma r)x.$$

How does the fitted model relate to the **marginal model**  $E[Y|X]$ ? It is easy to get

$$E[Y|X = x] = \alpha + \beta x + \gamma E[Z|X = x],$$

so the fitted regression model agrees with  $E[Y|X = x]$  as long as

$$E[Z|X = x] = rx.$$

# Confounding

Turning now to the variance structure of the fitted model, the limiting value of  $\hat{\sigma}^2$  is

$$\begin{aligned}\hat{\sigma}^2 &= \sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 / (n - 2) \\ &\approx \sum_i (\gamma z_i + \epsilon_i - \gamma rx_i)^2 / n \\ &\rightarrow \sigma^2 + \gamma^2(1 - r^2).\end{aligned}$$

Ideally this should estimate the variance function of the **marginal model**  $\text{var}[Y|X]$ .

# Confounding

By the law of total variation,

$$\begin{aligned}\text{var}[Y|X = x] &= E_{Z|X} \text{var}[Y|X = x, Z] + \text{var}_{Z|X=x} E[Y|X = x, Z] \\ &= \sigma^2 + \text{var}_{Z|X=x}(\alpha + \beta x + \gamma Z) \\ &= \sigma^2 + \gamma^2 \text{var}[Z|X = x].\end{aligned}$$

So for  $\hat{\sigma}^2$  to estimate  $\text{var}[Y|X = x]$  we need

$$\text{var}[Z|X = x] = 1 - r^2.$$

## The Gaussian case

Suppose

$$Y = \begin{pmatrix} A \\ B \end{pmatrix}$$

is a Gaussian random vector, where  $Y \in \mathcal{R}^n$ ,  $A \in \mathcal{R}^q$ , and  $B \in \mathcal{R}^{n-q}$ .

Let  $\mu = EY$  and  $\Sigma = \text{cov}[Y]$ . We can partition  $\mu$  and  $\Sigma$  as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where  $\mu_1 \in \mathcal{R}^q$ ,  $\mu_2 \in \mathcal{R}^{n-q}$ ,  $\Sigma_{11} \in \mathcal{R}^{q \times q}$ ,  $\Sigma_{12} \in \mathcal{R}^{q \times n-q}$ ,  
 $\Sigma_{22} \in \mathcal{R}^{n-q \times n-q}$ , and  $\Sigma_{21} = \Sigma'_{12}$ .

## The Gaussian case

It is a fact that  $A|B$  is Gaussian with mean

$$E[A|B = b] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(b - \mu_2)$$

and covariance matrix

$$\text{cov}[A|B = b] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

## The Gaussian Case

Now we apply these results to our model, taking  $X$  and  $Z$  to be jointly Gaussian.

The mean vector and covariance matrix are

$$E \begin{bmatrix} Z \\ X \end{bmatrix} = 0 \quad \text{cov} \begin{bmatrix} Z \\ X \end{bmatrix} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

so we get

$$E[Z|X = x] = rx \quad \text{cov}[Z|X = x] = 1 - r^2.$$

These are exactly the conditions stated earlier that guarantee the fitted mean model converges to the marginal regression function  $E[Y|X]$ , and the fitted variance model converges to the marginal regression variance  $\text{var}[Y|X]$ .

# Consequences of confounding

How does the presence of unmeasured confounders affect our ability to interpret regression models?

## Population average covariate effect

Suppose we specify a value  $X_*$  in the covariate space and randomly select two subjects  $i$  and  $j$  having  $X$  values  $X_i = X_* + 1$  and  $X_j = X_*$ . The inter-individual difference is

$$y_i - y_j = \beta + \gamma(z_i - z_j) + \epsilon_i - \epsilon_j,$$

which has a mean value (**marginal effect**) of

$$E[Y_i - Y_j | X_i = x_* + 1, X_j = x_*] = \beta + \gamma(E[Z|X = x_* + 1] - E[Z|X = x_*]),$$

which agrees with what would be obtained by least squares analysis as long as  $E[Z|X = x] = rx$ .

## Population average covariate effect

The variance of  $y_i - y_j$  is

$$2\sigma^2 + 2\gamma^2 \text{var}[Z|X],$$

which also agrees with the results of least squares analysis as long as  $\text{var}[Z|X = x] = 1 - r^2$ .

## Individual treatment effect

Now suppose we match two subjects  $i$  and  $j$  having  $X$  values differing by one unit, and who also have the same values of  $Z$ .

This is what one expects to see as the pre-treatment and post-treatment measurements following a treatment that changes an individual's  $X$  value by one unit, if the treatment does not affect  $Z$  (the within-subject treatment effect).

## Individual treatment effect

The mean difference (individual treatment effect) is

$$E[Y_i - Y_j | X_i = x_* + 1, X_j = x_*, Z_i = z_j] = \beta$$

and the variance is

$$\text{var}[Y_i - Y_j | X_i = x_* + 1, X_j = x_*, Z_i = z_j] = 2\sigma^2.$$

These do not in general agree with the estimates obtained by using least squares to analyze the observable data for  $X$  and  $Y$ . Depending on the sign of  $\gamma\theta$ , we may either overstate or underestimate the individual treatment effect  $\beta$ , and the population variance of the treatment effect will always be overstated.

## Measurement error for linear models

Suppose the data generating model is

$$Y = Z\beta + \epsilon,$$

with the usual linear model assumptions, but we do not observe  $Z$ .

Rather, we observe

$$X = Z + \tau,$$

where  $\tau$  is a random vector of covariate measurement errors with  $E[\tau] = 0$ . Assuming  $X_1 = 1$  is the intercept, it is natural to set the first column of  $\tau$  equal to zero.

This is called an **errors in variables model**, or a **measurement error model**.

## Measurement error for linear models

When covariates are measured with error, least squares point estimates may be biased and inferences may be incorrect.

Intuitively it seems that slope estimates should be “attenuated” (biased toward zero). The reasoning is that as the measurement error grows very large, the observed covariate  $X$  becomes equivalent to noise, so the slope estimate should go to zero.

## Measurement error for linear models

Let  $\mathbf{X}$  and  $\mathbf{Z}$  now represent the  $n \times p + 1$  observed and ideal design matrices. The least squares estimate of the model coefficients is

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

We will make the simplifying assumption that the covariate measurement error is uncorrelated with the covariate levels, so

$$\mathbf{Z}'\tau/n \rightarrow 0,$$

and that the covariate measurement error  $\tau$  and observation error  $\epsilon$  are uncorrelated, so

$$\tau'\epsilon/n \rightarrow 0.$$

# Measurement error for linear models

Under these circumstances,

$$\hat{\beta} \rightarrow (\mathbf{Z}'\mathbf{Z}/n + \tau'\tau/n)^{-1}\mathbf{Z}'Y/n.$$

Let  $M_z$  be the limiting value of  $\mathbf{Z}'\mathbf{Z}/n$ , and let  $M_\tau$  be the limiting value of  $\tau'\tau/n$ . Thus the limit of  $\hat{\beta}$  is

$$\begin{aligned}(M_z + M_\tau)^{-1}Z'Y/n &= (I + M_z^{-1}M_\tau)^{-1}M_z^{-1}\mathbf{Z}'Y/n \\ &\rightarrow (I + M_z^{-1}M_\tau)^{-1}\beta \\ &\equiv \beta_0.\end{aligned}$$

and hence the limiting bias is

$$\beta_0 - \beta = ((I + M_z^{-1}M_\tau)^{-1} - I)\beta.$$

## Measurement error for linear models

What can we say about the bias?

Note that the matrix  $M_z^{-1}M_\tau$  has non-negative eigenvalues, since it shares its eigenvalues with the positive semi-definite matrix

$$M_z^{-1/2}M_\tau M_z^{-T/2}.$$

It follows that all eigenvalues of  $I + M_z^{-1}M_\tau$  are greater than or equal to 1, so all eigenvalues of  $(I + M_z^{-1}M_\tau)^{-1}$  are less than or equal to 1.

This means that  $(I + M_z^{-1}M_\tau)^{-1}$  is a contraction, so  $\|\beta_0\| \leq \|\beta\|$ .

Therefore the sum of squares of fitted slopes is smaller on average than the sum of squares of actual slopes, due to measurement error.

# Types of measurement error

The “classical” measurement error model

$$X = Z + \tau,$$

where  $Z$  is the true value and  $X$  is the observed value, is the one most commonly considered.

Alternatively, in the case of an experiment it may make more sense to use the **Berkson error model**:

$$Z = X + \tau.$$

For example, suppose we aim to study a chemical reaction when a given concentration  $X$  of substrate is present. However, due to our inability to completely control the process, the actual concentration of substrate  $Z$  differs randomly from  $X$ , by an unknown amount  $\tau$ .

## Types of measurement error

You cannot simply rearrange  $Z = X + \tau$  to  $X = Z - \tau$  and claim that the two situations are equivalent.

In the first case,  $\tau$  is independent of  $Z$  but dependent with  $X$ . In the second case,  $\tau$  is independent of  $X$  but dependent with  $Z$ .

## SIMEX

SIMEX (simulation-extrapolation) is a relatively straightforward way to adjust for the effects of measurement error – if the variances and covariances among the measurement errors can be considered known.

Régress  $Y$  on  $X + \lambda E$ , where  $E$  is simulated noise having the same variance as the assumed measurement error. Denote the coefficient vector of this fit as  $\hat{\beta}_\lambda$ .

Repeat this for several values of  $\lambda \geq 0$ , leading to a set of  $\hat{\beta}_\lambda$  vectors.

Ideally,  $\hat{\beta}_{-1}$  would approximate the coefficient estimates under no measurement error.

By fitting a line or smooth curve to the  $\hat{\beta}_\lambda$  values (separately for each component of  $\beta$ ), it becomes possible to extrapolate back to  $\hat{\beta}_{-1}$ .

# Regression diagnostics

Kerby Shedden

Department of Statistics, University of Michigan

November 20, 2019

## Motivation

When working with a linear model with design matrix  $\mathbf{X}$ , the conventional linear model is based on the following conditions:

$$E[Y|\mathbf{X}] \in \text{col}(\mathbf{X}) \quad \text{and} \quad \text{var}[Y|\mathbf{X}] = \sigma^2 I.$$

Least squares point estimates depend on the first condition approximately holding. Least squares inferences depend on both of the above conditions approximately holding.

Inferences for small sample sizes may also depend on the distribution of  $Y - E[Y|\mathbf{X}]$  being approximately multivariate Gaussian, but for moderate or large sample sizes this condition is not critical.

Regression diagnostics for linear models are approaches for assessing how well a particular data set fits these two conditions.

# Residuals

Linear models can be expressed in two equivalent ways:

- ▶ Focus only on moments:

$$E[Y|\mathbf{X}] \in \text{col}(\mathbf{X}) \quad \text{and} \quad \text{var}[Y|\mathbf{X}] = \sigma^2 I.$$

- ▶ Use a "generative model", in this case an additive error model of the form  $Y = \mathbf{X}\beta + \epsilon$ , where  $\epsilon$  is random with  $E[\epsilon|\mathbf{X}] = 0$ , and  $\text{cov}[\epsilon|\mathbf{X}] \propto I$ .

Since the residuals can be viewed as predictions of the errors, it turns out that regression model diagnostics can often be developed using the residuals.

Recall that the residuals can be expressed

$$R \equiv (I - P)y$$

where  $P$  is the projection onto  $\text{col}(\mathbf{X})$  and  $y \in caIR^n$  is the vector of observed responses.

## Residuals

The residuals have two key mathematical properties regardless of the correctness of the model specification:

- ▶ The residuals sum to zero, since  $(I - P)\mathbf{1} = 0$  and hence  $\mathbf{1}'r = \mathbf{1}'(I - P)y = 0$ .
- ▶ The residuals and fitted values are orthogonal (they have zero sample covariance):

$$\begin{aligned}\widehat{\text{cov}}(r, \hat{y} | \mathbf{X}) &\propto (r - \bar{r})' \hat{Y} \\ &= r' \hat{y} \\ &= y'(I - P)Py \\ &= 0.\end{aligned}$$

These properties hold as long as an intercept is included in the model (so  $P \cdot \mathbf{1} = \mathbf{1}$ , where  $\mathbf{1} \in \mathcal{R}^n$  is a vector of 1's).

# Residuals

If the basic linear model conditions hold, these two properties have population counterparts:

- ▶ The expected value of each residual is zero:

$$\begin{aligned} E[R|\mathbf{X}] &= (I - P)E[Y|\mathbf{X}] \\ &= 0 \in \mathcal{R}^n. \end{aligned}$$

- ▶ The population covariance between any residual and any fitted value is zero:

$$\begin{aligned} \text{cov}(r, \hat{y}|\mathbf{X}) &= E[r\hat{y}'] \\ &= (I - P)\text{cov}(Y|\mathbf{X})P \\ &= \sigma^2(I - P)P \\ &= 0 \in \mathcal{R}^{n \times n}. \end{aligned}$$

# Residuals

If the model is correctly specified, there is a simple formula for the variances and covariances of the residuals:

$$\begin{aligned}\text{cov}(R|\mathbf{X}) &= (I - P)E[YY'](I - P) \\ &= (I - P)(\mathbf{X}\beta\beta'\mathbf{X}' + \sigma^2 I)(I - P) \\ &= \sigma^2(I - P).\end{aligned}$$

If the model is correctly specified, the **standardized residuals**

$$\frac{y_i - \hat{y}_i}{\hat{\sigma}}$$

and the **Studentized residuals**

$$\frac{y_i - \hat{y}_i}{\hat{\sigma}(1 - P_{ii})^{1/2}}$$

approximately have mean zero and variance one.

# Residuals

The standardized residuals are crudely standardized using a **plug-in** logic.  
The standardized “errors” are

$$\frac{y_i - E[Y|X = x_i]}{\sigma},$$

which have mean zero and unit variance (exactly).

If we plug-in  $\hat{y}_i$  for  $E[Y|X = x_i]$  and  $\hat{\sigma}$  for  $\sigma$ , then we get the standardized residual.

However we know that the actual variance of  $y_i - \hat{y}_i$  is  $\sigma^2(I - P)_{ii}$ , so it is more precise to scale by  $\sigma\sqrt{(I - P)_{ii}}$  rather than scaling by  $\sigma$ .

Since we plug in the estimate  $\hat{\sigma}$  for the population parameter  $\sigma$ , even the Studentized residual does not have variance exactly equal to 1.

## External standardization of residuals

Let  $\hat{\sigma}_{-i}^2$  be the estimate of  $\sigma^2$  obtained by fitting a regression model omitting the  $i^{\text{th}}$  case. It turns out that we can calculate this value without actually refitting the model:

$$\hat{\sigma}_{-i}^2 = \frac{(n - p - 1)\hat{\sigma}^2 - r_i^2 / (1 - P_{ii})}{n - p - 2}$$

where  $r_i$  is the residual for the model fit to all data.

The “externally standardized” residuals are

$$\frac{y_i - \hat{y}_i}{\hat{\sigma}_{-i}},$$

The “externally Studentized” residuals are

$$\frac{y_i - \hat{y}_i}{\hat{\sigma}_{-i}(1 - P_{ii})^{1/2}}.$$

## Outliers and masking

In some settings, residuals can be used to identify “outliers”. However, in a small data set, a large outlier will increase the value of  $\hat{\sigma}$ , and hence may **mask** itself.

Externally Studentized residuals solve the problem of a single large outlier masking itself. But masking may still occur if multiple large outliers are present.

## Outliers and masking

If multiple large outliers may be present we may use alternate estimates of the scale parameter  $\sigma$ :

- ▶ **Interquartile range (IQR)**: this is the difference between the 75<sup>th</sup> percentile and the 25<sup>th</sup> percentile of the distribution or data. The IQR of the standard normal distribution is 1.35, so  $\text{IQR}/1.35$  can be used to estimate  $\sigma$ .
- ▶ **Median Absolute Deviation (MAD)**: this is the median value of the absolute deviations from the median of the distribution or data, i.e.  $\text{median}(|Z - \text{median}(Z)|)$ . The MAD of the standard normal distribution is 0.65, so  $\text{MAD}/0.65$  can be used to estimate  $\sigma$ .

These alternative estimates of  $\sigma$  can be used in place of the usual  $\hat{\sigma}$  for standardizing or Studentizing residuals.

## Leverage

**Leverage** is a measure of how strongly the data for case  $i$  determine the fitted value  $\hat{y}_i$ .

Since  $\hat{y} = Py$ , where  $P$  is the projection matrix onto  $\text{col}(\mathbf{X})$ , then

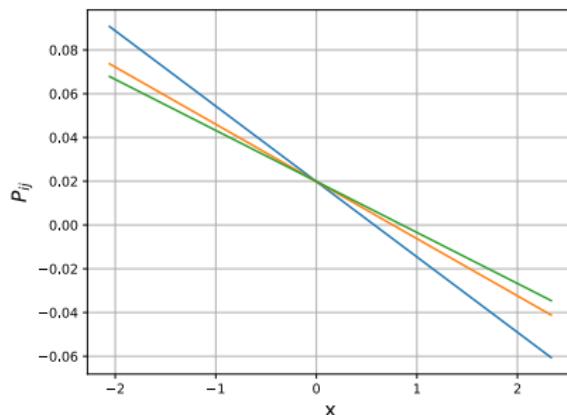
$$\hat{y}_i = \sum_j P_{ij} y_j.$$

It is natural to define the leverage for case  $i$  as  $P_{ii}$ ,

This is related to the fact that the variance of the  $i^{\text{th}}$  residual is  $\sigma^2(1 - P_{ii})$ . Since the residuals have mean zero, when  $P_{ii}$  is close to 1, the residual will likely be close to zero. This means that fitted line will usually pass close to  $(x_i, y_i)$  if it is a high leverage point.

## Leverage

These are the coefficients  $P_{ij}$  plotted against  $x_j$  (for three different values of  $i$ ), in a simple linear regression:

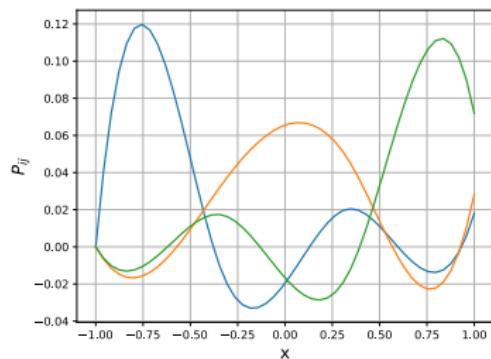


The pattern is linear due to the following identity (which only holds for simple linear regression):

$$\hat{y}_j = \sum_i \frac{S + n(x_i - \bar{x})(x_j - \bar{x})}{nS} y_i$$

# Leverage

If we use basis functions, the coefficients in each row of  $P$  are much more “local”.



For example, the blue curve above shows  $P_{i1}, \dots, P_{in}$  where  $x_i = -0.8$ . Note that the  $P_{ij}$  values are largest where  $x \approx -0.8$ .

## Leverage

What is a big leverage? The average leverage is  $\text{trace}(P)/n = (p+1)/n$ . If the leverage for a particular case is two or more times greater than the average leverage, it may be considered to have high leverage.

In simple linear regression, we showed earlier that

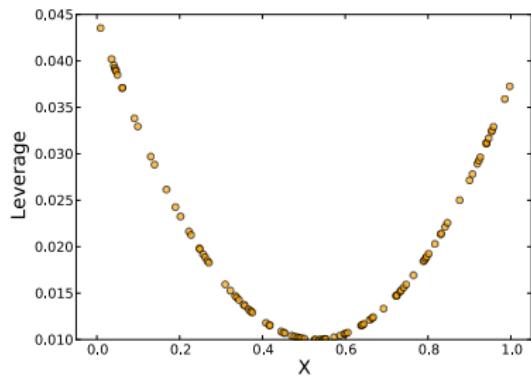
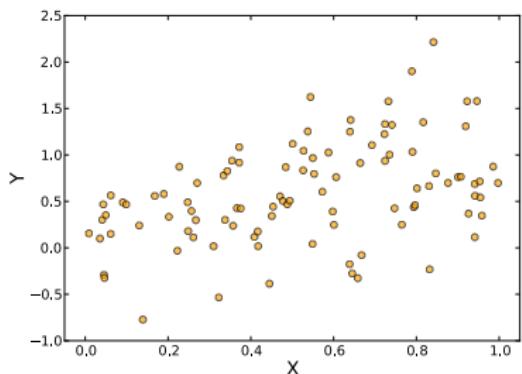
$$\text{var}(y_i - \hat{\alpha} - \hat{\beta}x_i) = (n-1)\sigma^2/n - \sigma^2(x_i - \bar{x})^2 / \sum_j (x_j - \bar{x})^2.$$

This implies that when  $p = 1$ ,

$$P_{ii} = 1/n + (x_i - \bar{x})^2 / \sum_j (x_j - \bar{x})^2.$$

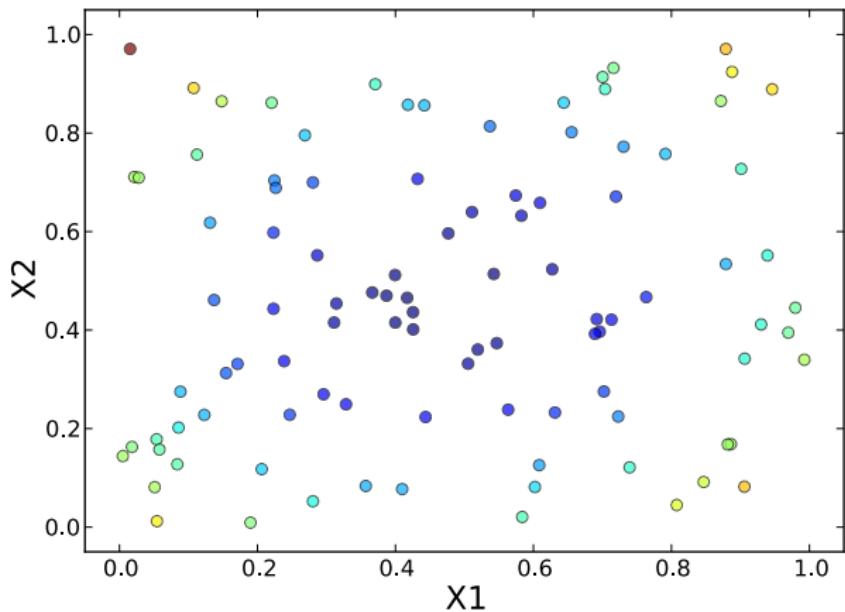
# Leverage

Leverage values in a simple linear regression:



# Leverage

Leverage values in a linear regression with two independent variables:



## Leverage

In general,

$$P_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' = \mathbf{x}_i'(\mathbf{X}'\mathbf{X}/n)^{-1}\mathbf{x}_i'/n$$

where  $\mathbf{x}_i$  is the  $i^{\text{th}}$  row of  $\mathbf{X}$  (including the intercept).

Let  $\tilde{\mathbf{x}}_i$  be row  $i$  of  $\mathbf{X}$  without the intercept, let  $\tilde{\mu}$  be the sample mean of the  $\tilde{\mathbf{x}}_i$ , and let  $\Sigma_X$  be the sample covariance matrix of the  $\tilde{\mathbf{x}}_i$  (scaled by  $n$  rather than  $n - 1$ ). It is a fact that

$$\mathbf{x}_i'(\mathbf{X}'\mathbf{X}/n)^{-1}\mathbf{x}_i' = (\tilde{\mathbf{x}}_i - \tilde{\mu})\Sigma_X^{-1}(\tilde{\mathbf{x}}_i - \tilde{\mu})' + 1$$

and therefore

$$P_{ii} = ((\tilde{\mathbf{x}}_i - \tilde{\mu}_X)\Sigma_X^{-1}(\tilde{\mathbf{x}}_i - \tilde{\mu}_X)' + 1) / n.$$

Note that this implies that  $P_{ii} \geq 1/n$ .

# Leverage

The expression

$$(\tilde{\mathbf{x}}_i - \tilde{\mu}_X) \Sigma_X^{-1} (\tilde{\mathbf{x}}_i - \tilde{\mu}_X)'$$

is the **Mahalanobis distance** between  $\tilde{\mathbf{x}}_i$  and  $\tilde{\mu}_X$ . Thus there is a direct relationship between the Mahalanobis distance of a point relative to the center of the covariate set, and its leverage.

# Influence

**Influence** measures the degree to which deletion of a case changes the fitted model.

We will see that this is different from leverage – a high leverage point has the potential to be influential, but is not always influential.

The **deleted slope** for case  $i$  is the fitted slope vector that is obtained upon deleting case  $i$ . The following identity allows the deleted slopes to be calculated efficiently

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{r_i}{1 - P_{ii}} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}'_i,$$

where  $r_i$  is the  $i^{\text{th}}$  residual, and  $\mathbf{x}_i$  is row  $i$  of the design matrix.

# Influence

The vector of all deleted fitted values  $\hat{y}_{(i)}$  are

$$\hat{y}_{(i)} = \mathbf{X}\hat{\beta}_{(i)} = \hat{y} - \frac{r_i}{1 - P_{ii}} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Influence can be measured by **Cook's distance**:

$$\begin{aligned} D_i &\equiv \frac{1}{(p+1)\hat{\sigma}^2} (\hat{y} - \hat{y}_{(i)})' (\hat{y} - \hat{y}_{(i)}) \\ &= \frac{r_i^2}{(1 - P_{ii})^2(p+1)\hat{\sigma}^2} \mathbf{x}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i' \\ &= \frac{P_{ii} r_i^{s2}}{(1 - P_{ii})(p+1)}, \end{aligned}$$

where  $r_i$  is the residual and  $r_i^s$  is the studentized residual.

## Influence

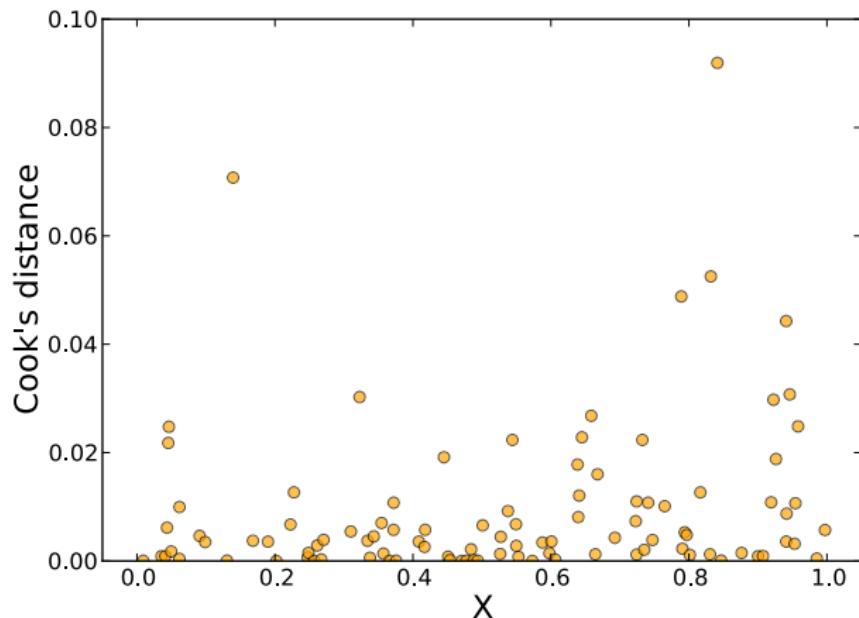
Cook's distance approximately captures the average squared change in fitted values due to deleting case  $i$ , in error variance units.

Cook's distance is large only if both the leverage  $P_{ii}$  is high, and the studentized residual for the  $i^{\text{th}}$  case is large.

As a general rule,  $D_i$  values from  $1/2$  to  $1$  are high, and values greater than  $1$  are considered to be "very high".

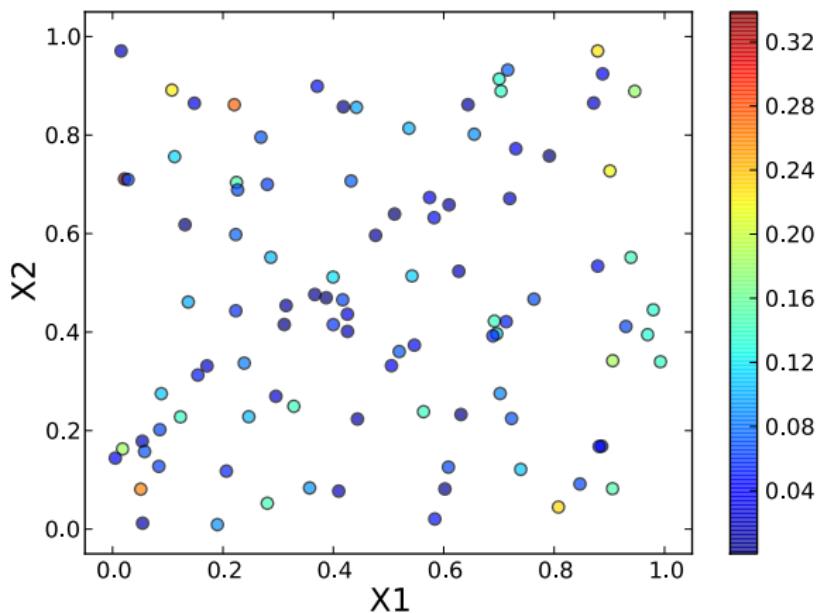
# Influence

Cook's distances in a simple linear regression:



# Influence

Cook's distances in a linear regression with two variables:



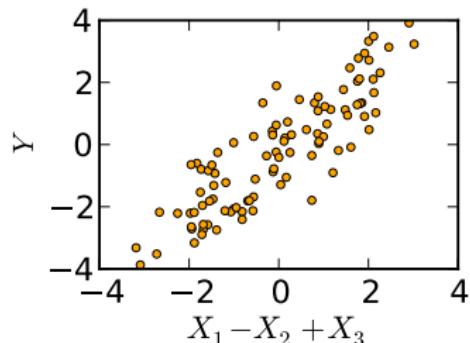
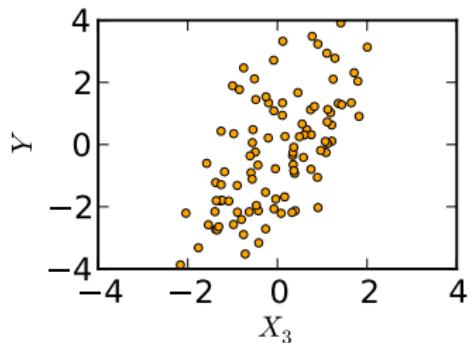
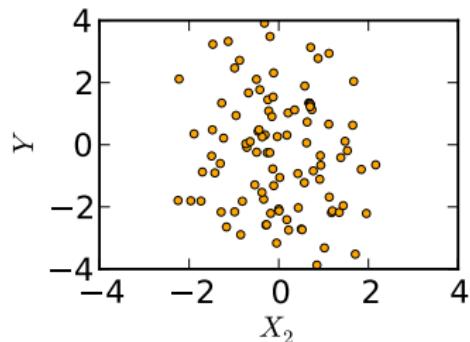
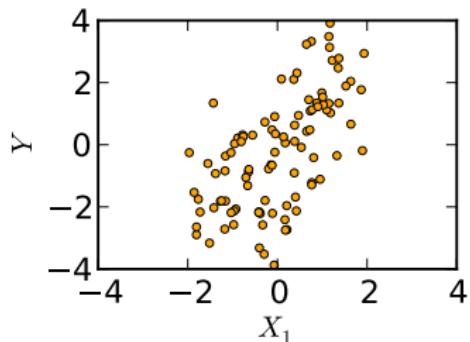
# Regression graphics

Quite a few graphical techniques have been proposed to aid in visualizing regression relationships. We will discuss the following plots:

1. Scatterplots of  $y$  against individual covariates  $x_j$
2. Scatterplots of two covariates  $x_j, x_k$  against each other
3. Residuals versus fitted values plot
4. Added variable plots
5. Partial residual plots
6. Residual quantile plots

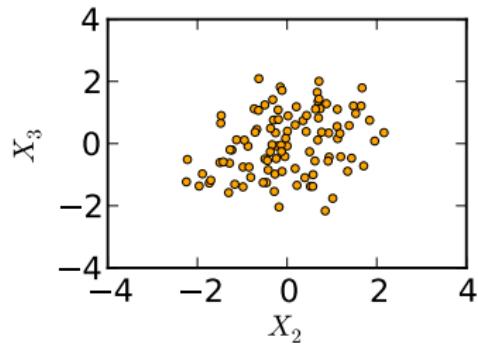
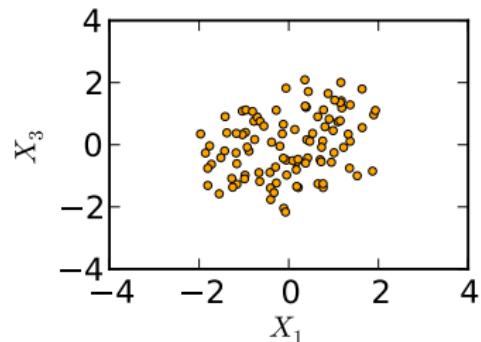
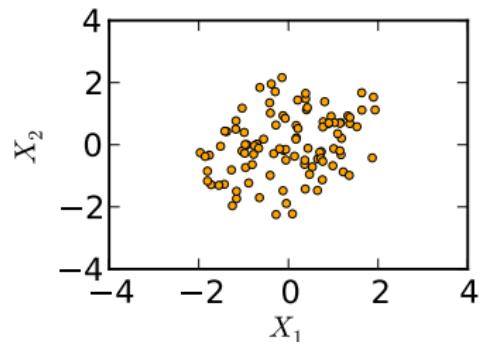
# Scatterplots of $Y$ against individual $X$ variables

$$E[Y|X] = X_1 - X_2 + X_3, \text{ var}[Y|X] = 1, \text{ var}(X_j) = 1, \text{ cor}(X_j, X_k) = 0.3$$



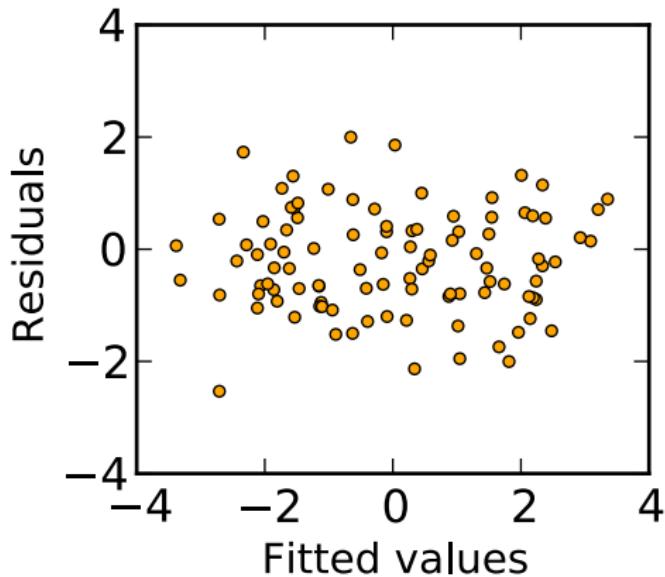
# Scatterplots of $X$ variables against each other

$$E[Y|X] = X_1 - X_2 + X_3, \text{ var}[Y|X] = 1, \text{ var}(X_j) = 1, \text{ cor}(X_j, X_k) = 0.3$$



## Residuals against fitted values plot

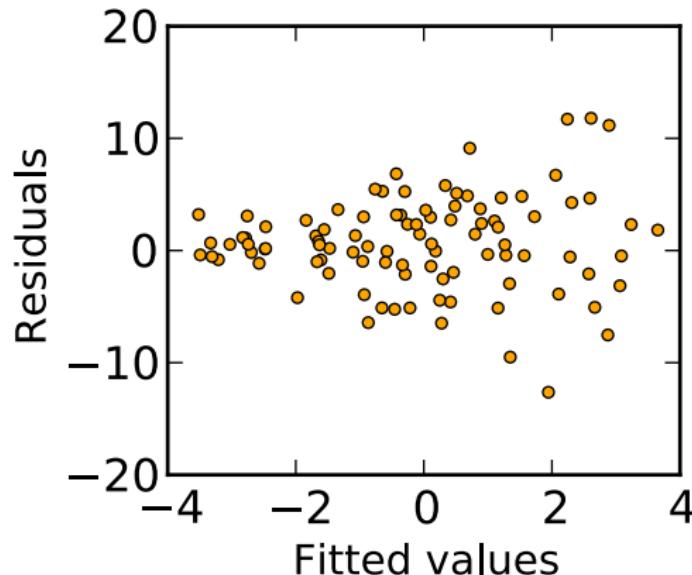
$$E[Y|X] = X_1 - X_2 + X_3, \text{ var}[Y|X] = 1, \text{ var}(X_j) = 1, \text{ cor}(X_j, X_k) = 0.3$$



# Residuals against fitted values plots

Heteroscedastic errors:

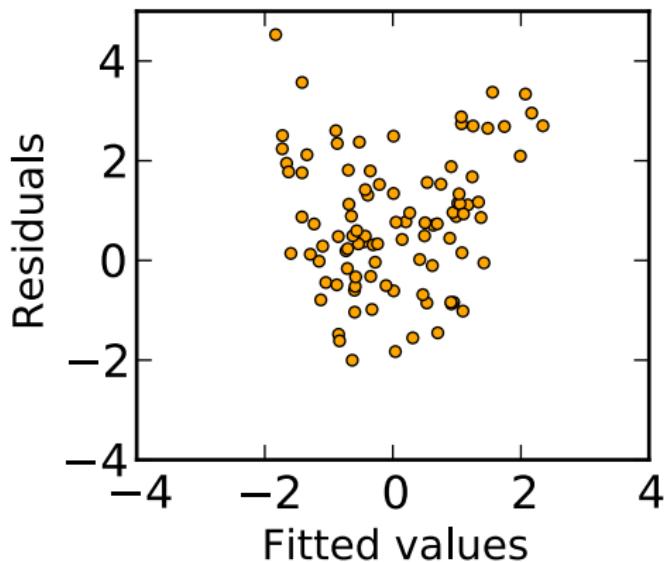
$$E[Y|X] = X_1 + X_3, \quad \text{var}[Y|X] = 4 + X_1 + X_3, \quad \text{var}(X_j) = 1,$$
$$\text{cor}(X_j, X_k) = 0.3$$



# Residuals against fitted values plots

Nonlinear mean structure:

$$E[Y|X] = X_1^2, \text{ var}[Y|X] = 1, \text{ var}(X_j) = 1, \text{ cor}(X_j, X_k) = 0.3$$



## Added variable plots

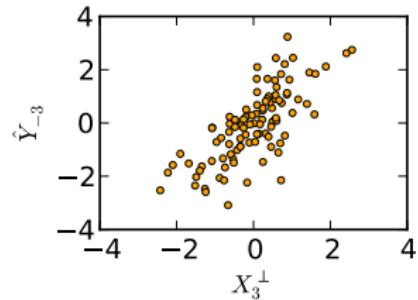
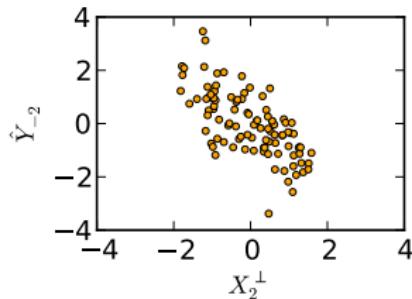
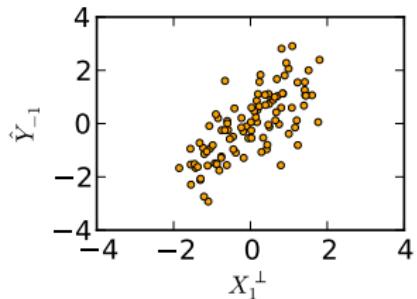
Suppose  $P_{-j}$  is the projection onto the span of all covariates except  $X_j$ , and define  $\hat{Y}_{-j} = P_{-j} Y$ ,  $X_j^* = P_{-j} X_j$ . The **added variable plot** is a scatterplot of  $Y - \hat{Y}_{-j}$  against  $X - X_j^*$ .

The squared correlation coefficient of the points in the added variable plot is the partial  $R^2$  for variable  $j$ .

Added variable plots are also called **partial regression plots**.

## Added variable plots

$$E[Y|X] = X_1 - X_2 + X_3, \text{ var}[Y|X] = 1, \text{ var}(X_j) = 1, \text{ cor}(X_j, X_k) = 0.3$$



## Partial residual plot

Suppose we fit the model

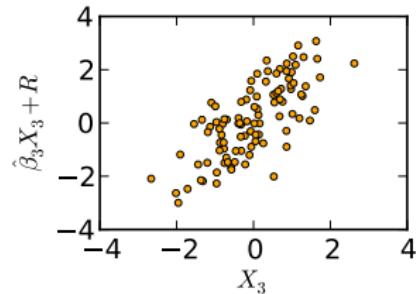
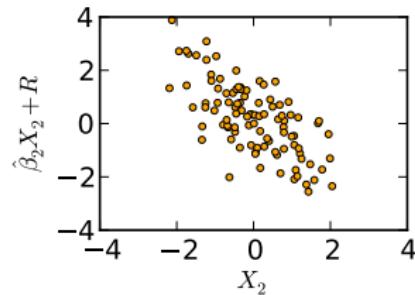
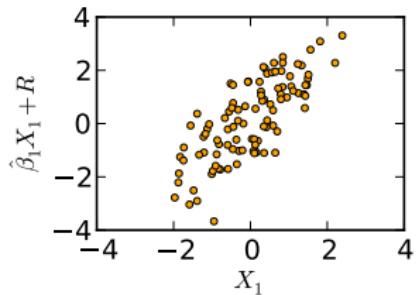
$$\hat{Y}_i = \hat{\beta}' X_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}.$$

The **partial residual plot** for covariate  $j$  is a plot of  $\hat{\beta}_j X_{ij} + R_i$  against  $X_{ij}$ , where  $R_i$  is the residual.

The partial residual plot attempts to show how covariate  $j$  is related to  $Y$ , if we control for the effects of all other covariates.

## Partial residual plot

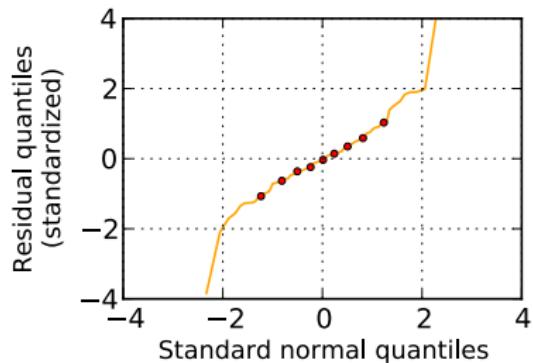
$$E[Y|X] = X_1 - X_2 + X_3, \text{ var}[Y|X] = 1, \text{ var}(X_j) = 1, \text{ cor}(X_j, X_k) = 0.3$$



# Residual quantile plots

$$E[Y|X] = X_1 - X_2 + X_3, \text{ var}[Y|X] = 1, \text{ var}(X_j) = 1, \text{ cor}(X_j, X_k) = 0.3$$

$t_4$  distributed errors



## Transformations

As noted above, the linear model imposes two main constraints on the population that is under study. Specifically, the conditional mean function should be linear, and the conditional variance function should be constant.

If it appears that  $E[Y|X = x]$  is not linear in  $x$ , or that  $\text{Var}[Y|X = x]$  is not constant in  $x$ , it may be possible to continuously transform either  $y$  or  $x$  so that the linear model becomes more consistent with the data.

## Variance stabilizing transformations

Many populations encountered in practice exhibit a **mean/variance relationship**, where  $E[Y_i]$  and  $\text{var}[Y_i]$  are related.

Suppose that

$$\text{var}[Y_i] = g(E[Y_i])\sigma^2,$$

and let  $f(\cdot)$  be a transform to be applied to the  $y_i$ . The goal is to find a transform such that the variances of the transformed responses are constant. Using a Taylor expansion,

$$f(Y_i) \approx f(E[Y_i]) + f'(E[Y_i])(Y_i - E[Y_i]).$$

# Variance stabilizing transformations

Therefore

$$\text{var}[f(Y_i)] \approx f'(E[Y_i])^2 \cdot \text{var}[Y_i] = f'(E[Y_i])^2 \cdot g(E[Y_i])\sigma^2.$$

The goal is to find  $f$  such that  $f' = 1/\sqrt{g}$ .

**Example:** Suppose  $g(z) = z^\lambda$ . This includes the “Poisson regression” case  $\lambda = 1$ , where the variance is proportional to the mean, and the case  $\lambda = 2$  where the standard deviation is proportional to the mean.

When  $\lambda = 1$ ,  $f$  solves  $f'(z) = 1/\sqrt{z}$ , so  $f$  is the square root function.

When  $\lambda = 2$ ,  $f$  solves  $f'(z) = 1/z$ , so  $f$  is the logarithm function.

## Log/log regression

Suppose we fit a simple linear regression of the form

$$E[\log(Y) | \log(X)] = \alpha + \beta \log(X).$$

$$E[\log(Y) | X = x + 1] - E[\log(Y) | X = x] = \beta$$

Using the crude approximation  $\log E[Y|X] \approx E[\log(Y)|X]$ , we conclude  $E[Y|X]$  is approximately scaled by a factor of  $e^\beta$  when  $X$  is scaled by a factor of  $e$ .

Thus in a log/log model, we may say that a  $f\%$  change in  $X$  is approximately associated with a  $f^\beta\%$  change in the expected response.

## Maximum likelihood estimation of a data transformation

The Box-Cox family of transforms is

$$y \longmapsto \frac{y^\lambda - 1}{\lambda},$$

which makes sense only when all  $Y_i$  are positive.

The Box-Cox family includes the identity ( $\lambda = 1$ ), all power transformations such as the square root ( $\lambda = 1/2$ ) and reciprocal ( $\lambda = -1$ ), and the logarithm in the limiting case  $\lambda \rightarrow 0$ .

## Maximum likelihood estimation of a data transformation

Suppose we assume that for some value of  $\lambda$ , the transformed data follow a linear model with Gaussian errors. We can then set out to estimate  $\lambda$ .

The joint log-likelihood of the transformed data is

$$-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (y_i^{(\lambda)} - x_i' \beta)^2.$$

Next we transform this back to a likelihood in terms of  $y_i = g_\lambda^{-1}(y_i^{(\lambda)})$ .  
This joint log-likelihood is

$$-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (g_\lambda(y_i) - x_i' \beta)^2 + \sum_i \log J_i$$

where the Jacobian is

$$\log J_i = \log g'_\lambda(y_i) = (\lambda - 1) \log y_i.$$

## Maximum likelihood estimation of a data transformation

The joint log likelihood for the  $y_i$  is

$$-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (g_\lambda(y_i) - x_i' \beta)^2 + (\lambda - 1) \sum_i \log y_i.$$

This likelihood is maximized with respect to  $\lambda$ ,  $\beta$ , and  $\sigma^2$  to identify the MLE.

## Maximum likelihood estimation of a data transformation

To do the maximization, let  $y^{(\lambda)} \equiv g_\lambda(y)$  denote the transformed observed responses, and let  $\hat{y}^{(\lambda)}$  denote the fitted values from regressing  $Y^{(\lambda)}$  on  $X$ . Since  $\sigma^2$  does not appear in the Jacobian,

$$\hat{\sigma}_\lambda^2 \equiv n^{-1} \|y^{(\lambda)} - \hat{y}^{(\lambda)}\|^2$$

will be the maximizing value of  $\sigma^2$ . Therefore the MLE of  $\beta$  and  $\lambda$  will maximize

$$-\frac{n}{2} \log \hat{\sigma}_\lambda^2 + (\lambda - 1) \sum_i \log y_i.$$

## collinearity Diagnostics

collinearity inflates the sampling variances of covariate effect estimates.

To understand the effect of collinearity on  $\text{var}[\hat{\beta}_j | \mathbf{X}]$ , reorder the columns and partition the design matrix  $\mathbf{X}$  as

$$\mathbf{X} = ( X_j \mid \mathbf{X}_0 ) = ( X_j - X_j^\perp + X_j^\perp \mid \mathbf{X}_0 )$$

where  $\mathbf{X}_0$  is the  $n \times p$  matrix consisting of all columns in  $\mathbf{X}$  except  $X_j$ , and  $\mathbf{X}_j^\perp$  is the projection of  $\mathbf{X}_j$  onto  $\text{col}(\mathbf{X}_0)^\perp$ . Therefore

$$H \equiv \mathbf{X}'\mathbf{X} = \left( \begin{array}{c|c} X_j' X_j & (X_j - X_j^\perp)' \mathbf{X}_0 \\ \mathbf{X}_0' (X_j - X_j^\perp) & \mathbf{X}_0' \mathbf{X}_0 \end{array} \right).$$

$\text{var} \hat{\beta}_j = \sigma^2 H_{11}^{-1}$ , so we want a simple expression for  $H_{11}^{-1}$ .

## collinearity Diagnostics

A symmetric block matrix can be inverted using:

$$\begin{pmatrix} A & B \\ B' & C \end{pmatrix}^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}BC^{-1} \\ -C^{-1}B'S^{-1} & C^{-1} + C^{-1}B'S^{-1}BC^{-1} \end{pmatrix},$$

where

$$S = A - BC^{-1}B'.$$

Therefore

$$H_{1,1}^{-1} = \frac{1}{\|X_j\|^2 - (X_j - X_j^\perp)'P_0(X_j - X_j^\perp)},$$

where  $P_0 = \mathbf{X}_0(\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{X}_0$  is the projection matrix onto  $\text{col}(\mathbf{X}_0)$ .

## collinearity Diagnostics

Since  $X_j - X_j^\perp \in \text{col}(\mathbf{X}_0)$ , we can write

$$H_{1,1}^{-1} = \frac{1}{\|X_j\|^2 - \|X_j - X_j^\perp\|^2},$$

and since  $X_j^{\perp'}(X_j - X_j^\perp) = 0$ , it follows that

$$\|X_j\|^2 = \|X_j - X_j^\perp + X_j^\perp\|^2 = \|X_j - X_j^\perp\|^2 + \|X_j^\perp\|^2,$$

so

$$H_{1,1}^{-1} = \frac{1}{\|X_j^\perp\|^2}.$$

This makes sense, since smaller values of  $\|X_j^\perp\|^2$  correspond to greater collinearity.

## collinearity Diagnostics

Let  $R_{jx}^2$  be the coefficient of determination (multiple  $R^2$ ) for the regression of  $X_j$  on the other covariates.

$$R_{jx}^2 = 1 - \frac{\|X_j - (X_j - X_j^\perp)\|^2}{\|X_j - \bar{X}_j\|^2} = 1 - \frac{\|X_j^\perp\|^2}{\|X_j - \bar{X}_j\|^2}.$$

Combining the two equations yields

$$H_{11}^{-1} = \frac{1}{\|X_j - \bar{X}_j\|^2} \cdot \frac{1}{1 - R_{jx}^2}.$$

## collinearity Diagnostics

The two factors in the expression

$$H_{11}^{-1} = \frac{1}{\|X_j - \bar{X}_j\|^2} \cdot \frac{1}{1 - R_{jx}^2}.$$

reflect two different sources of variance of  $\hat{\beta}_j$ :

- ▶  $1/\|X_j - \bar{X}_j\|^2 = 1/((n-1)\widehat{\text{var}}(X_j))$  reflects the scaling of  $X_j$
- ▶ The **variance inflation factor** (VIF)  $1/(1 - R_{jx}^2)$  is scale-free. It is always greater than or equal to 1, and is equal to 1 only if  $X_j$  is orthogonal to the other covariates. Large values of the VIF indicate that parameter estimation is strongly affected by collinearity.

# Prediction

Kerby Shedden

Department of Statistics, University of Michigan

November 9, 2018

## Prediction analysis

In a prediction-oriented analysis, we are interested in fitting a model to capture the mean relationship between independent variables  $x$  and a dependent variable  $y$ , then using the fitted model to make predictions on an independent data set.

The model has the form  $f_\theta$ , where for each  $\theta$ , we have a function from  $\mathcal{R}^P$  to  $\mathcal{R}$ . Thus  $\{f_\theta\}$  is a family of functions indexed by a parameter  $\theta$ .

We use the data to obtain an estimate  $\hat{\theta}$  of  $\theta$ , which in turn leads us to an estimate  $\hat{f}_\theta$  of the regression function.

It is helpful to think in terms of **training data**  $\{(y_i, x_i)\}$  that are used to fit the model, so  $\hat{\theta} = \hat{\theta}(\{(y_i, x_i)\})$ , and **testing data**  $\{(y_i^*, x_i^*)\}$  on which predictions are made.

## Quantifying prediction error

Prediction analysis focuses on prediction errors, for example through the **mean squared prediction error** (MSPE):

$$E|Y^* - f_{\hat{\theta}}(\mathbf{X}^*)|^2,$$

and its sample analogue

$$\sum_{i=1}^{n^*} \|y_i^* - f_{\hat{\theta}}(\mathbf{x}_i^*)\|^2 / n^*,$$

where  $n^*$  is the size of the testing set.

Prediction analysis does not usually focus on properties of the parameter estimates themselves, e.g. bias  $E\hat{\theta} - \theta$ , or parameter MSE  $E(\hat{\theta} - \theta)^2$ .

## MSPE for OLS analysis

The mean squared prediction error for OLS regression is easy to derive. The testing data follow  $y^* = \mathbf{X}^* \beta + \epsilon^*$ . Let  $\hat{y}^* = \mathbf{X}^* \hat{\beta}$  denote the predicted values in the test set. Then

$$\begin{aligned} E\|y^* - \hat{y}^*\|^2 &= E\|\mathbf{X}^* \beta + \epsilon^* - \mathbf{X}^* \hat{\beta}\|^2 \\ &= E\|\mathbf{X}^*(\beta - \hat{\beta})\|^2 + E\|\epsilon^*\|^2 \\ &= E[(\hat{\beta} - \beta)'(\mathbf{X}^{*\prime} \mathbf{X}^*)(\hat{\beta} - \beta)] + n^* \sigma^2 \\ &= \text{tr}(\mathbf{X}^{*\prime} \mathbf{X}^* \cdot E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']) + n^* \sigma^2 \\ &= \text{tr}(\mathbf{X}^{*\prime} \mathbf{X}^* \cdot \Sigma_{\hat{\beta}}) + n^* \sigma^2, \end{aligned}$$

where  $\Sigma_{\hat{\beta}}$  is the covariance matrix of  $\hat{\beta}$  from the training process. Note the requirement for  $\hat{Y}^*$  and  $Y^*$  to be independent (given  $\mathbf{X}$  and  $\mathbf{X}^*$ ).

## MSPE for OLS analysis

The MSPE for OLS is

$$\text{tr} \left( (\mathbf{X}^{*\prime} \mathbf{X}^* / n^*) \cdot \Sigma_{\hat{\beta}} \right) + \sigma^2.$$

If  $\mathbf{X}$  is the training set design matrix, then  $\Sigma_{\hat{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ , so if  $\mathbf{X} = \mathbf{X}^*$ , then

$$E\|y^* - \hat{y}\|^2 = \sigma^2(p + 1 + n^*),$$

and the MSPE in this case is

$$\sigma^2(p + 1)/n^* + \sigma^2 = \sigma^2(p + 1)/n + \sigma^2.$$

## MSPE for OLS analysis

More generally, suppose  $\mathbf{X}'\mathbf{X}/n = \mathbf{X}^{*\prime}\mathbf{X}^*/n^*$ . Then

$$\Sigma_{\hat{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 n^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}/n.$$

Thus the MSPE is

$$\text{tr} \left( (\mathbf{X}^{*\prime}\mathbf{X}^*/n^*) \cdot \Sigma_{\hat{\beta}} \right) + \sigma^2 = \sigma^2(p+1)/n + \sigma^2.$$

## MSPE in practice

The MSPE discussed here is the primary population quantity of interest for prediction. It is not straightforward to estimate however.

The task of **model selection**, discussed later, can be viewed as aiming to identify the model with the lowest MSPE (among a set of candidate models under consideration).

Note that the candidate models we fit to data may not be correctly-specified, so the usual estimate of  $\hat{\sigma}^2$  may be biased.

## PRESS residuals

One way to estimate the MSPE with few theoretical conditions is using cross validation. We will briefly introduce this idea here, then return to it and cover it in more detail when we talk about model selection.

If case  $i$  is deleted and a prediction of  $y_i$  is made from the remaining data, we can compare the observed and predicted values to get the **prediction residual**:

$$r_{(i)} \equiv y_i - \hat{y}_{(i)i}.$$

where  $\hat{y}_{(i)i}$  is the prediction of  $y_i$  based on a data set in which case  $i$  was removed.

## PRESS residuals

A simple formula for the prediction residual in OLS is given by

$$\begin{aligned} r_{(i)} &= y_i - \mathbf{x}_i \hat{\beta}_{(i)} \\ &= y_i - \mathbf{x}_i (\hat{\beta} - r_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i / (1 - P_{ii})) \\ &= r_i / (1 - P_{ii}). \end{aligned}$$

where  $\mathbf{X}$  is the design matrix,  $\mathbf{x}_i$  is row  $i$  of the design matrix, and  $P$  is the projection matrix (for the full sample).

The sum of squares of the prediction residuals is called **PRESS** (prediction residual error sum of squares). It is equivalent to using leave-one-out cross validation to estimate the “generalization error rate”.

## Bias and variance in prediction

If we are using a function  $f_{\hat{\theta}}$  to predict  $y$  from  $x$ , we can view the prediction error as arising from contributions of **bias** and **variance**.

The bias is

$$b(x) \equiv E[f_{\hat{\theta}}(x)|x] - E[y|x].$$

The variance is  $v(x) \equiv \text{var}[f_{\hat{\theta}}(x)|x]$ .

The MSPE is

$$E[(y - f_{\hat{\theta}}(x))^2] = b(x)^2 + v(x).$$

## Bias and variance in prediction

While having zero bias is an important consideration in some statistical analyses, arguably the overall accuracy, as measured by MSPE, should be the dominant consideration.

The MSPE results from a combination of squared bias and variance. If we want to minimize the MSPE we should consider using a biased estimator, if by doing so we attain better MSPE (due to it having much smaller variance).

The relationship between bias and variance discussed here is often referred to as the **bias/variance tradeoff**.

## Ridge regression

Ridge regression uses the minimizer of a penalized squared error loss function to estimate the regression coefficients:

$$\hat{\beta} \equiv \operatorname{argmin}_{\beta} \|y - \mathbf{X}\beta\|^2 + \lambda \beta' D \beta.$$

Typically  $D$  is a diagonal matrix with 0 in the 1,1 position and ones on the rest of the diagonal. In this case,

$$\beta' D \beta = \sum_{j \geq 1} \beta_j^2.$$

This makes most sense when the covariates have been standardized, so it is reasonable to penalize the  $\beta_j$  equally.

## Ridge regression

Ridge regression is a compromise between fitting the data as well as possible (by making  $\|y - \mathbf{X}\beta\|^2$  small), while not allowing any one fitted coefficient to get very large (which causes  $\beta'D\beta$  to get large).

## Ridge regression and collinearity

Suppose  $x_1$  and  $x_2$  are standardized vectors with a substantial positive correlation (i.e.  $x'_1 x_2$  is large), and the population slopes are  $\beta_1$  and  $\beta_2$ , i.e.  $E[y|x_1, x_2] = \beta_1 x_1 + \beta_2 x_2$ .

Fits of the form

$$(\beta_1 + \gamma)x_1 + (\beta_2 - \gamma)x_2 = E[y|x_1, x_2] + \gamma(x_1 - x_2)$$

have similar MSE values as  $\gamma$  varies, since  $x_1 - x_2$  is small when  $x_1$  and  $x_2$  are strongly positively associated.

In other words, OLS can't easily distinguish among these fits.

For example, if  $x_1 \approx x_2$ , then  $3x_1 + 3x_2$ ,  $4x_1 + 2x_2$ ,  $5x_1 + x_2$ , etc. all produce similar fitted values.

## Ridge regression and collinearity

For large  $\lambda$ , ridge regression favors the fits that minimize

$$(\beta_1 + \gamma)^2 + (\beta_2 - \gamma)^2.$$

This expression is minimized at  $\gamma = (\beta_2 - \beta_1)/2$ , giving the fit

$$(\beta_1 + \beta_2)x_1/2 + (\beta_1 + \beta_2)x_2/2.$$

⇒ Ridge regression favors coefficient estimates for which strongly positively correlated covariates have similar estimated effects.

## Calculation of ridge regression estimates

For a given value  $\lambda > 0$ , ridge regression is no more difficult computationally than ordinary least squares, since

$$\frac{\partial}{\partial \beta} \|y - \mathbf{X}\beta\|^2 + \lambda \beta' D \beta = -2\mathbf{X}'y + 2\mathbf{X}'\mathbf{X}\beta + 2\lambda D\beta,$$

so the ridge estimate  $\hat{\beta}$  solves the system of linear equations

$$(\mathbf{X}'\mathbf{X} + \lambda D)\beta = \mathbf{X}'y.$$

This equation can have a unique solution even when  $\mathbf{X}'\mathbf{X}$  is singular. Thus one application of ridging is to produce regression estimates for singular design matrices.

## Ridge regression bias and variance

Ridge regression estimates are biased, but may be less variable than OLS estimates. If  $\mathbf{X}'\mathbf{X}$  is non-singular, the ridge estimator can be written

$$\begin{aligned}\hat{\beta}_\lambda &= (\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'y \\ &= (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \\ &= (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon) \\ &= (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}\beta + (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon.\end{aligned}$$

Thus the bias is

$$E[\hat{\beta}_\lambda | \mathbf{X}] - \beta = ((I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1} - I)\beta$$

## Ridge regression bias and variance

The variance of the ridge regression estimates is

$$\text{var} \hat{\beta}_\lambda = \sigma^2 (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1} (\mathbf{X}'\mathbf{X})^{-1} (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-T}.$$

## Ridge regression bias and variance

Next we will show that  $\text{var}[\hat{\beta}] \geq \text{var}[\hat{\beta}_\lambda]$ , in the sense that

$$\text{var}[\hat{\beta}] - \text{var}[\hat{\beta}_\lambda]$$

is a non-negative definite matrix.

First let  $M = \lambda(\mathbf{X}'\mathbf{X})^{-1}D$ , and note that

## Ridge regression bias and variance

$$\begin{aligned} v'(\text{var}\hat{\beta} - \text{var}\hat{\beta}_\lambda)v &\propto v' ((\mathbf{X}'\mathbf{X})^{-1} - (I + M)^{-1}(\mathbf{X}'\mathbf{X})^{-1}(I + M)^{-T}) v \\ &= u' ((I + M)(\mathbf{X}'\mathbf{X})^{-1}(I + M)' - (\mathbf{X}'\mathbf{X})^{-1}) u \\ &= u' (M(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}M' + M(\mathbf{X}'\mathbf{X})^{-1}M') u \\ &= u' (2\lambda(\mathbf{X}'\mathbf{X})^{-1}D(\mathbf{X}'\mathbf{X})^{-1} + \\ &\quad \lambda^2(\mathbf{X}'\mathbf{X})^{-1}D(\mathbf{X}'\mathbf{X})^{-1}D(\mathbf{X}'\mathbf{X})^{-1}) u \end{aligned}$$

where  $u = (I + M)^{-T}v$ .

We can conclude that for any fixed vector  $\theta$ ,

$$\text{var}(\theta'\hat{\beta}_\lambda) \leq \text{var}(\theta'\hat{\beta}).$$

## Ridge regression effective degrees of freedom

Like OLS, the fitted values under ridge regression are linear functions of the observed values

$$\hat{Y}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'Y$$

In OLS regression, the degrees of freedom is the number of free parameters in the model, which is equal to the trace of the projection matrix  $P$  that satisfies  $\hat{Y} = PY$ .

Fitted values in ridge regression are not a projection of  $Y$ , but the matrix

$$\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'.$$

plays an analogous role to  $P$ .

## Ridge regression effective degrees of freedom

The **effective degrees of freedom** for ridge regression is defined as

$$\text{EDF}_\lambda = \text{tr} [\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'] .$$

The trace can be easily computed using the identity

$$\text{trace} (\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}') = \text{trace} ((\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'\mathbf{X}) .$$

## Ridge regression effective degrees of freedom

$\text{EDF}_\lambda$  is monotonically decreasing in  $\lambda$ . To see this we will use the following fact about matrix derivatives

$$\partial \text{tr}(A^{-1}B) / \partial A = -A^{-T}B'A^{-T}.$$

By the chain rule, letting  $A = \mathbf{X}'\mathbf{X} + \lambda D$ , we have

$$\begin{aligned}\partial \text{tr}(A^{-1}\mathbf{X}'\mathbf{X}) / \partial \lambda &= \sum_{ij} \frac{\partial \text{tr}(A^{-1}\mathbf{X}'\mathbf{X})}{\partial A_{ij}} \cdot \frac{\partial A_{ij}}{\partial \lambda} \\ &= -\sum_{ij} [A^{-T}(\mathbf{X}'\mathbf{X})A^{-T}]_{ij} \cdot D_{ij} \\ &= -\sum_i [A^{-T}(\mathbf{X}'\mathbf{X})A^{-T}]_{ii} \cdot D_{ii} \\ &\leq 0.\end{aligned}$$

## Ridge regression effective degrees of freedom

$\text{EDF}_\lambda$  equals  $\text{rank}(\mathbf{X})$  when  $\lambda = 0$ . To see what happens as  $\lambda \rightarrow \infty$ , we can apply the Sherman-Morrison-Woodbury identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Let  $G = \mathbf{X}'\mathbf{X}$ , and write  $D = FF'$ , where  $F$  has independent columns (usually  $F$  will be  $p+1 \times p$  as we do not penalize the intercept).

## Ridge regression effective degrees of freedom

Applying the SMW identity and letting  $\lambda \rightarrow \infty$  we get

$$\begin{aligned}\text{tr}[(G + \lambda D)^{-1} G] &= \text{tr}[(G^{-1} - G^{-1}F(I/\lambda + F'G^{-1}F)^{-1}F'G^{-1}) G] \\ &= \text{tr}[I_{p+1} - G^{-1}F(I/\lambda + F'G^{-1}F)^{-1}F'] \\ &\rightarrow \text{tr}I_{p+1} - \text{tr}[G^{-1}F(F'G^{-1}F)^{-1}F'] \\ &\rightarrow \text{tr}I_{p+1} - \text{tr}[(F'G^{-1}F)^{-1}F'G^{-1}F] \\ &= p + 1 - \text{rank}(F).\end{aligned}$$

Therefore in the usual case where  $F$  has rank  $p$ ,  $\text{EDF}_\lambda$  converges to 1 as  $\lambda$  grows large, reflecting the fact that all coefficients other than the intercept are forced to zero.

## Ridge regression and the SVD

Suppose we are fitting a ridge regression with  $D = I$ , and we factor  $\mathbf{X} = USV'$  using the singular value decomposition (SVD), so that  $U$  and  $V$  are orthogonal matrices, and  $S$  is a diagonal matrix with non-negative diagonal elements.

The fitted coefficients are

$$\begin{aligned}\hat{\beta}_\lambda &= (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'Y \\ &= (VS^2V' + \lambda VV')^{-1}VSU'Y \\ &= V(S^2 + \lambda I)^{-1}SU'Y\end{aligned}$$

Note that for OLS ( $\lambda = 0$ ), we get  $\hat{\beta} = VS^{-1}U'Y$ . The effect of ridging is to replace  $S^{-1}$  in this expression with  $(S^2 + \lambda I)^{-1}S$ , which are uniformly smaller values when  $\lambda > 0$ .

## Ridge regression tuning parameter

There are various ways to set the ridge parameter  $\lambda$ .

Cross-validation can be used to estimate the MSPE for any particular value of  $\lambda$ . Then this estimated MSPE could be minimized by checking its value at a finite set of  $\lambda$  values.

Generalized cross validation, which minimizes the following over  $\lambda$ , is a simpler, and more commonly used approach.

$$\text{GCV}(\lambda) = \frac{\|Y - \hat{Y}_\lambda\|^2}{(n - \text{EDF}_\lambda)^2}.$$

# Model selection

Kerby Shedden

Department of Statistics, University of Michigan

November 14, 2018

## Background

Suppose we observe data  $y$  and are considering a family of models  $f_\theta$  that may approximately describe how  $y$  was generated.

If we are mainly interested in the individual model parameters, we will focus on how close  $\hat{\theta}_j$  is to  $\theta_j$  (e.g. in terms of its bias, variance, MSE).

Alternatively, our focus may be on the probability distribution  $f$  that is the data-generating model for  $y$ . In this case, we are more interested in whether  $f_{\hat{\theta}}$  approximates  $f$ , rather than in whether  $\hat{\theta}_j$  approximates  $\theta_j$ .

The term model **model selection** is used to describe statistical estimation in a context when the focus is more on the fitted model than on the individual parameters.

## Model complexity and parsimony

The discrepancy between  $f_\theta$  and  $f_{\hat{\theta}}$  is strongly influenced by how complex of a model we decide to fit.

Suppose we have  $p = 30$  covariates and  $n = 50$  observations. We could consider the following two alternatives:

1. We could fit a model using all of the covariates. In this case,  $\hat{\theta}$  is unbiased for  $\theta$  (in a linear model fit using OLS). But  $\hat{\theta}$  has very high variance.
2. We could fit a model using only the five strongest effects. In this case,  $\hat{\theta}$  will be biased for  $\theta$ , but it will have lower variance (compared to the estimate including all covariates).

If our goal is for  $f_{\hat{\theta}}$  and  $f_\theta$  to be close, either approach 1 or approach 2 could perform better, depending on the circumstances.

## Assessing model fit

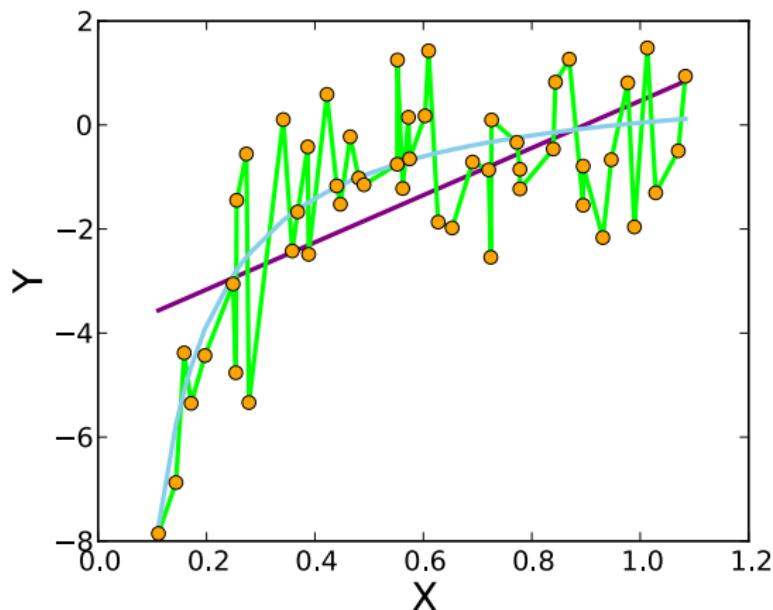
A more complex model will usually fit the data better than a more parsimonious (simpler) model. This is called **overfitting**.

Due to overfitting, we cannot simply compare models in terms of how well they fit the data (e.g. in terms of the residual sum of squares, or the height of the likelihood). The more complex model will always appear to be better if we do this.

To overcome this, most model selection procedures balance a measure of a model's fit with a measure of its complexity. To select the more complex model, it must fit the data substantially better than the simpler model.

## Fit and parsimony

**Example:** The purple, green, and blue curves below are estimates of  $E(Y|X)$ . The green line fits the data better but is more complex. Which estimate is closest to the truth?



## F-tests

Suppose we are comparing two nested families of models  $\mathcal{F}_1 \subset \mathcal{F}_2$ , both of which are linear subspaces. An F-test can be used to select between  $\mathcal{F}_1$  and  $\mathcal{F}_2$ .

## Mallows' $C_p$

Suppose we postulate the model

$$y = \mathbf{X}\beta + \epsilon$$

but in fact  $E[y] \notin \text{col}(\mathbf{X})$ . We'll continue to assume that the homoscedastic variance structure  $\text{cov}(\epsilon|\mathbf{X}) = \sigma^2 I$  holds. Denote this model as  $M$ .

Denote the error in estimating  $E[y]$  under model  $M$  as

$$D_M = \hat{y}_M - E[y],$$

where  $\hat{y}_M$  is the projection of  $y$  onto  $\text{col}(\mathbf{X})$ .

## Mallows' $C_p$

Write

$$E[y] = \theta_X + \theta_X^\perp,$$

where  $\theta_X \in \text{col}(X)$  and  $\theta_X^\perp \in \text{col}(X)^\perp$ . Since  $y = \theta_X + \theta_X^\perp + \epsilon$ , it follows that  $\hat{y} = \theta_X + \epsilon_X$ , where  $\epsilon_X$  is the projection of  $\epsilon$  onto  $\text{col}(\mathbf{X})$ .

Therefore

$$\begin{aligned} ED_M D_M' &= E(\hat{y}_M - Ey)(\hat{y}_M - Ey)' \\ &= E(\epsilon_X - \theta_X^\perp)(\epsilon_X - \theta_X^\perp)' \\ &= \theta_X^\perp \theta_X^{\perp'} + \sigma^2 P_X \end{aligned}$$

where  $P_X$  is the projection matrix onto  $\text{col}(\mathbf{X})$ .

## Mallows' $C_p$

Taking the trace of both sides, yields

$$E\|D_M\|^2 = \|\theta_X^\perp\|^2 + (p+1)\sigma^2,$$

where  $p+1$  is the rank of  $P_X$ .

Mallows'  $C_p$  aims to estimate

$$C_p^* = E\|D_M\|^2/\sigma^2 = \|\theta_X^\perp\|^2/\sigma^2 + p + 1$$

The model that minimizes  $C_p^*$  is the closest to the true model in this particular sense.

## Mallows' $C_p$

We need an estimate of  $C_p^*$ .

To begin, we can derive the expected value of

$$\hat{\sigma}^2 = \|y - \hat{y}_M\|^2 / (n - p - 1)$$

in the case where  $E[y]$  is not necessarily in  $\text{col}(\mathbf{X})$ :

$$\begin{aligned} E\hat{\sigma}^2 &= E[y'(I - P)y / (n - p - 1)] \\ &= E[(\theta_X + \theta_X^\perp + \epsilon)'(I - P)(\theta_X + \theta_X^\perp + \epsilon) / (n - p - 1)] \\ &= E[\text{tr}(I - P)(\theta_X^\perp + \epsilon)(\theta_X^\perp + \epsilon)' / (n - p - 1)] \\ &= \|\theta_X^\perp\|^2 / (n - p - 1) + \sigma^2. \end{aligned}$$

## Mallows' $C_p$

Now suppose we have an unbiased estimate of  $\sigma^2$ . This could come from a regression against a much larger design matrix that is thought to contain  $E[y]$ . Call this estimate  $\hat{\sigma}^2$ . Then

$$(n - p - 1)E(\hat{\sigma}^2 - \sigma^{*2}) = \|\theta_X^\perp\|^2.$$

Therefore we can estimate  $C_p^*$  using

$$C_p = (n - p - 1)(\hat{\sigma}^2 - \sigma^{*2})/\sigma^{*2} + p + 1.$$

The model  $M$  with the smallest value of  $C_p$  is selected.

Suppose we are selecting from a family of linear models with design matrices  $X_M$ , for  $M \in \mathcal{M}$ .

For each  $X_M$ , the model parameters (slopes and error variance) can be estimated using least squares (and method of moments for the error variance) as a vector  $\hat{\theta}_M$ . This allows us to construct a **predictive density**:

$$p(y; X_M, \hat{\theta}_M).$$

The **Kullback-Leibler divergence** (“KL-divergence”) between the predictive density and the actual density  $p(y)$  is

$$E_y \log \left( \frac{p(y)}{p(y; X_M, \hat{\theta}_M)} \right) = \int \log \left( \frac{p(y)}{p(y; X_m, \hat{\theta}_M)} \right) p(y) dy \geq 0.$$

Here we are considering  $\hat{\theta}_M$  to be fixed.

Small values of the KL divergence indicate that the predictive density is close to the actual density.

# AIC

Akaike's Information Criterion (AIC) aims to estimate the KL divergence between a candidate model and the data-generating model  $p(Y)$  unbiasedly. We can then select the candidate model that has the smallest estimated KL-divergence relative to  $p(y)$ .

The KL-divergence can be written

$$E_y \log(p(y)) - E_y \log(p(y; \hat{\theta}_M, X_M)).$$

We can ignore the first term since it doesn't depend on  $M$ . Thus it will be equivalent to select the model that maximizes the **predictive log likelihood**:

$$E_y \log(p(y; \hat{\theta}_M, X_M)) = \int \log(p(y; X_M, \hat{\theta}_M)) p(y) dY.$$

The predictive log likelihood is the expected value of

$$\log p(y^*; X_M, \hat{\theta}_M(y))$$

taken over the joint distribution of  $y$  and  $y^*$ , which are independent copies of the data.

The parameter estimates  $\hat{\theta}_M = \hat{\theta}_M(y)$  are determined from  $y$ , which you can think of as a “training set”, and the log-likelihood is evaluated at  $y^*$ , using  $\hat{\theta}_M(y)$  to set the parameters.

Since we don't have both  $y$  and  $y^*$ , it is natural to use the plug-in estimator of the predictive log likelihood:

$$\log p(y; X_M, \hat{\theta}_M(y))$$

But this is biased upward, due to overfitting.

Surprisingly, this upward bias can be shown to be approximately equal to the dimension of  $M$ , which is  $p + 1$  for regression ( $p + 2$  if you count  $\sigma^2$ ).

# AIC

Thus we may take

$$\log(p(y_{\text{train}}; X_M, \hat{\theta}_M)) - p - 1$$

as a model selection statistic to be maximized (commonly this is multiplied by -2, in which case it is to be minimized). This quantity is the AIC.

# AIC

To apply the AIC to linear models, we assume the error values  $\epsilon$  are multivariate normal, so the log-likelihood becomes

$$-\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|y - \mathbf{X}\beta\|^2.$$

If we work with the profile likelihood over  $\beta$ , we get  $-n \log(\hat{\sigma}^2)/2$  (plus a constant). Therefore maximizing the AIC is equivalent to maximizing

$$\underbrace{-n \log(\hat{\sigma}^2)}_{\text{fit}} - \underbrace{2(p+1)}_{\text{complexity}}$$

## AIC and likelihood ratios

The AIC does not require the models being compared to be nested, but let's consider this special case.

Let  $L_1$  and  $L_0$  be the maximized log likelihoods for two nested models (so  $L_1 \geq L_0$ ).

We know that  $2(L_1 - L_0)$  approximately follows a  $\chi_q^2$  distribution, where  $q$  is the difference between the number of parameters of the two models.

If  $q = 1$ , we select the larger model if  $L_1 - L_0 \geq 1.92$  (the usual likelihood ratio test at the 0.05 type I error rate).

If the additional parameters are not needed, then  $E[L_1 - L_0] = 0.5$  (so 0.5 is the lowest possible threshold for  $L_1 - L_0$  that could ever be considered).

Under AIC, we select the larger model if  $L_1 - L_0 > 1$  (less strict than the likelihood ratio test).

## Bayesian Information Criterion (BIC)

A different criterion that we will not derive here is the “Bayesian information criterion” (BIC)

$$\underbrace{-n \log(\hat{\sigma}^2)}_{\text{fit}} - \underbrace{(p+1) \log(n)}_{\text{complexity}}$$

The complexity penalty in BIC,  $\log(n)(p+1)$ , will always be larger than the corresponding AIC penalty, which is  $2(p+1)$ . Thus the BIC will always favor simpler models than the AIC.

## Model selection based on prediction

Many approaches to model selection attempt to identify the model that predicts best on independent data.

If independent “training” and “test” sets are available, for each model  $M$  the parameters of  $M$  can be fit using the training data, yielding  $\hat{\theta}_M$ . Predictions can then be made on the test set

$$\hat{Y}_{M,\text{test}} = X_{M,\text{test}} \hat{\theta}_M$$

and the quality of prediction can be assessed, for example, using the **Mean Squared Prediction Error** (MSPE):

$$\|Y_{\text{test}} - \hat{Y}_{M,\text{test}}\|^2 / n.$$

## Cross-validation

Separate training and test sets are usually not available. Cross validation is a direct method for obtaining unbiased estimates of the prediction mean squared error when only training data are available.

In  $k$ -fold cross validation, the data are partitioned into  $k$  disjoint subsets ("folds"), denoted  $S_1 \cup \dots \cup S_k = \{1, \dots, n\}$ .

Let  $\hat{\beta}_j$  be the fitted coefficients omitting the  $j^{\text{th}}$  of these subsets, and let

$$\text{CV}_k = n^{-1} \sum_{j=1}^k \sum_{i \in S_j} (Y_i - X'_i \hat{\beta}_j)^2$$

This is an approximately unbiased (but potentially very imprecise) estimate of the MSPE on a test set.

The special case of **leave one out cross validation** (LOOCV) is when  $k = n$ .

## Cross-validation

For OLS regression,  $\text{CV}_n$  (also known as “**prediction residual error sum of squares**”, or PRESS), can be computed rapidly:

$$\text{CV}_n = n^{-1} \sum_i R_i^2 / (1 - P_{ii})^2.$$

The **generalized cross-validation** (GCV) criterion replaces  $P_{ii}$  with the average diagonal element of  $P$ , which is  $\text{trace}(P)/n$ :

$$\text{GCV}_n = n^{-1} \sum_i R_i^2 / (1 - \text{tr}(P)/n)^2 = n^{-1} \frac{\|Y - \hat{Y}\|^2}{(1 - \text{tr}(P)/n)^2}.$$

# Regression analysis with dependent data

Kerby Shedden

Department of Statistics, University of Michigan

November 25, 2019

## Clustered data

Clustered data are sampled from a population that can be viewed as the union of a number of related subpopulations.

Write the data as

$$y_{ij} \in \mathcal{R}, \mathbf{x}_{ij} \in \mathcal{R}^p,$$

$$i = 1, \dots, m \quad j = 1, \dots, n_i$$

where  $i$  indexes the subpopulation and  $j$  indexes the individuals in the sample belonging to the  $i^{\text{th}}$  subpopulation.

There are  $m$  clusters (subpopulations), and there are  $n_i$  observations in cluster  $i$ . If the  $n_i$  are all the same, we have balanced clustering.

## Clustered data

It may happen that units from the same subpopulation are more alike than units from different subpopulations, i.e. for  $i \neq i'$ ,  $j \neq j'$ ,

$$\text{cor}(y_{ij}, y_{ij'}) > \text{cor}(y_{ij}, y_{i'j'}).$$

Part of the within-cluster similarity may be explained by the covariates, i.e. units from the same subpopulation may have similar  $\mathbf{x}_{ij}$  values, which leads to similar  $y$  values. In this case,

$$\text{cor}(y_{ij}, y_{ij'}) > \text{cor}(y_{ij}, y_{ij'} | \mathbf{x}_{ij}, \mathbf{x}_{ij'}).$$

Even after accounting for measured covariates, units in a cluster may still resemble each other more than units in different clusters:

$$\text{cor}(y_{ij}, y_{ij'} | \mathbf{x}_{ij}, \mathbf{x}_{ij'}) > \text{cor}(y_{ij}, y_{i'j'} | \mathbf{x}_{ij}, \mathbf{x}_{i'j'}).$$

## Clustered data

A simple working model to account for this dependence is

$$\hat{y}_{ij} = \hat{\theta}_i + \hat{\beta}' \mathbf{x}_{ij}.$$

The idea here is that  $\beta' \mathbf{x}_{ij}$  explains the variation in  $y$  that is related to the measured covariates, and the  $\theta_i$  explain variation in  $y$  that is related to the clustering.

This working model would be correct if there were  $q$  omitted variables  $\mathbf{z}_{ij\ell}$ ,  $\ell = 1, \dots, q$ , that were constant for all units in the same cluster (i.e.  $\mathbf{z}_{ij\ell}$  depends on  $\ell$  and  $i$ , but not on  $j$ ).

In that case,  $\hat{\theta}_i$  would stand in for the value of  $\sum_{\ell} \hat{\psi}_{\ell} \mathbf{z}_{ij\ell}$  that we would have obtained if the  $\mathbf{z}_{ij\ell}$  were observed.

## Clustered data

As an alternate notation, we can vectorize the data to express  $y \in \mathcal{R}^n$  and  $\mathbf{X} \in \mathcal{R}^{n \times p+1}$ , then write

$$\hat{y} = \sum_i \hat{\theta}_i I_i + \mathbf{X} \hat{\beta},$$

where  $I_i \in \mathcal{R}^n$  is the indicator of which subjects belong to cluster  $i$ .

If the observed covariates in  $\mathbf{X}$  are related to the clustering (i.e. if the columns of  $\mathbf{X}$  and the  $I_i$  are not orthogonal), then OLS apportions the overlapping variance between  $\hat{\beta}$  and  $\hat{\theta}$ .

## Test score example

Suppose  $y_{ij}$  is a reading test score for the  $j^{\text{th}}$  student in the  $i^{\text{th}}$  classroom, out of a large number of classrooms that are considered. Suppose  $\mathbf{x}_{ij}$  is the income of a student's family.

We might postulate as a population model

$$y_{ij} = \theta_i + \beta \mathbf{x}_{ij} + \epsilon_{ij},$$

which can be fit as

$$\hat{y}_{ij} = \hat{\theta}_i + \hat{\beta} \mathbf{x}_{ij}.$$

## Test score example

Ideally we would want the parameter estimates to reflect sources of variation as follows:

- ▶ “Direct effects” of parent income such as access to books, life experiences, good health care, etc. should go entirely to  $\hat{\beta}$ .
- ▶ Attributes of classrooms that are not related to parent income, for example, the effect of an exceptionally good or bad teacher, should go entirely to the  $\hat{\theta}_i$ .
- ▶ Attributes of classrooms that are correlated with parent income, such as teacher salary, training, and resources, will be apportioned by OLS between  $\hat{\theta}_i$  and  $\hat{\beta}$ .
- ▶ Unique events affecting particular individuals, such as the severe illness of the student or a family member, should go entirely to  $\epsilon$ .

## Other examples of clustered data

- ▶ Treatment outcomes for patients treated in various hospitals.
- ▶ Crime rates in police precincts distributed over a number of large cities (the precincts are the units and the cities are the clusters).
- ▶ Prices of stocks belonging to various business sectors.
- ▶ Surveys in which the data are collected following a cluster sampling approach.

## What if we ignore the $\theta_i$ ?

The  $\theta_i$  are usually not of primary interest, but we should be concerned that by failing to take account of the clustering, we may incorrectly assess the relationship between  $y$  and  $\mathbf{X}$ .

If the  $\theta_i$  are nonzero, but we fail to include them in the model, the working model is misspecified.

Let  $\mathbf{X}$  be the design matrix without intercept, and let  $Q$  be the matrix of cluster indicators (which includes the intercept):

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \\ \vdots \end{pmatrix} \quad Q = \begin{pmatrix} 1 & 0 & \cdots \\ 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ 0 & 1 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_{111} & \mathbf{x}_{112} & \cdots \\ \mathbf{x}_{121} & \mathbf{x}_{122} & \cdots \\ \mathbf{x}_{211} & \mathbf{x}_{212} & \cdots \\ \mathbf{x}_{221} & \mathbf{x}_{222} & \cdots \\ \mathbf{x}_{231} & \mathbf{x}_{232} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

## What if we ignore the $\theta_i$ ?

Let  $\tilde{\mathbf{X}} = [1_n | \mathbf{X}]$ . The estimate that results from regressing  $y$  on  $\tilde{\mathbf{X}}$  is

$$\begin{aligned} E\hat{\beta}^* &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'E[y] \\ &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{X}\beta + Q\theta). \end{aligned}$$

where  $\hat{\beta}^* = (\hat{\beta}_0, \hat{\beta}')'$ .

For the bias of  $\hat{\beta}$  for  $\beta$  to be zero, we need

$$E[\hat{\beta}^*] - \tilde{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{X}\beta + Q\theta - \tilde{\mathbf{X}}\tilde{\beta}) = 0,$$

where  $\beta_0 = E[\hat{\beta}_0]$  and  $\tilde{\beta} = (\beta_0, \beta')'$ . Thus we need

$$0 = \tilde{\mathbf{X}}'(\mathbf{X}\beta + Q\theta - \tilde{\mathbf{X}}\tilde{\beta}) = \tilde{\mathbf{X}}'(Q\theta - \beta_0).$$

## What if we ignore the $\theta_i$ ?

Let  $S = Q\theta$  be the vector of cluster effects. Since the first column of  $\tilde{\mathbf{X}}$  consists of 1's, we have that

$$\bar{S} = \beta_0.$$

For any other covariate  $\mathbf{X}_j$ , we have that

$$\mathbf{X}'_j S = \beta_0 \mathbf{X}'_j \mathbf{1}_n,$$

which implies that  $\mathbf{X}_j$  and  $S$  have zero sample covariance.

This is unlikely in many studies, where people tend to cluster (in schools, hospitals, etc.) with other people having similar covariate levels.

## Fixed effects analysis

In a **fixed effects** approach, we model the  $\theta_i$  as regression parameters, by including additional columns in the design matrix whose covariate levels are the cluster indicators, i.e. we regress  $y$  on  $[\mathbf{X} \ Q]$ .

As the sample size grows, in most applications the cluster sizes  $n_i$  will remain bounded (e.g. a primary school classroom might have up to 30 students). Thus the number of clusters must grow, so the dimension of the parameter vector  $\theta$  grows.

This puts us in a setting where the model dimension ( $p$ ) and sample size ( $n$ ) are growing together. This is not typical in standard regression modeling, and leads to the “Neyman Scott problem”. Contemporary methods for “high dimensional” regression provide one way to work around this challenge.

## Cluster effect examples

What does the inclusion of fixed effects do to the parameter estimates  $\hat{\beta}$ , which are often of primary interest?

The following slides show scatterplots of an outcome  $y$  against a scalar covariate  $x$ , in a setting where there are three clusters (indicated by color).

Above each plot are the coefficient estimates, Z-scores, and  $r^2$  values for fits in which the cluster effects are either included (top line) or excluded (second line). These estimates are obtained from the following two working models:

$$\hat{y} = \hat{\alpha} + \hat{\beta}\mathbf{x} + \sum \hat{\theta}_i I_i$$

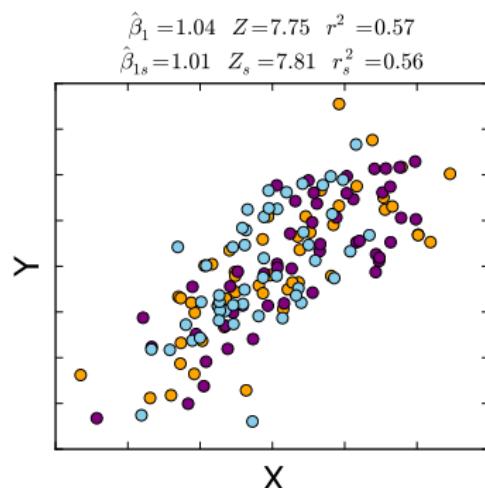
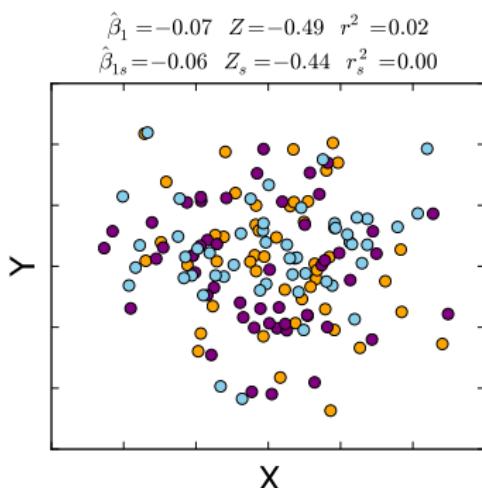
$$\hat{y} = \hat{\alpha}_s + \hat{\beta}_s \mathbf{x}.$$

The Z-scores are  $\hat{\beta}/\text{SD}(\hat{\beta})$  and the  $r^2$  values are  $\text{cor}(\hat{y}, y)^2$ .

## Cluster effect examples

Left:  $y$  is independent of both clusters and  $x$ ;  $x$  and clusters are independent;  $\hat{\beta} \approx 0$  with and without cluster terms in model.

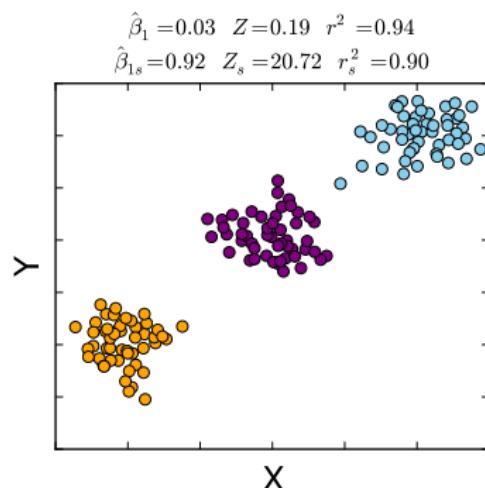
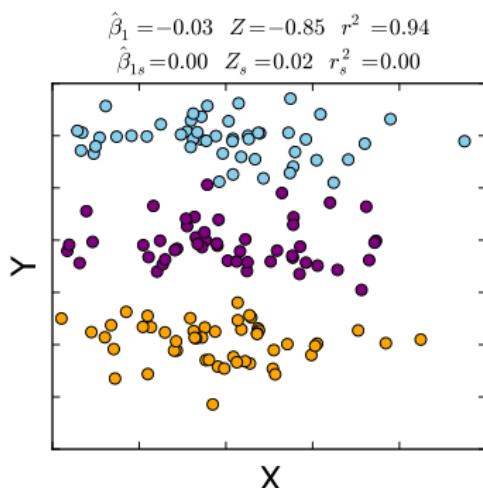
Right:  $y$  relates to  $x$ , but not to the clusters;  $x$  and clusters are independent;  $\hat{\beta}$ ,  $Z$ , and  $r^2$  are similar with and without cluster effects in model.



## Cluster effect examples

Left:  $y$  relates to clusters, but not to  $x$ ;  $x$  and clusters are independent;  $\hat{\beta} \approx 0$  with and without cluster effects in model.

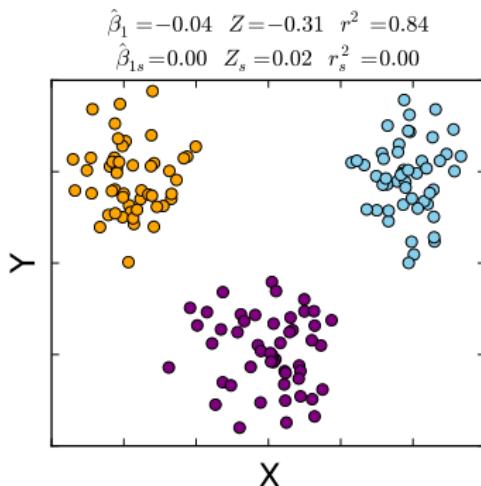
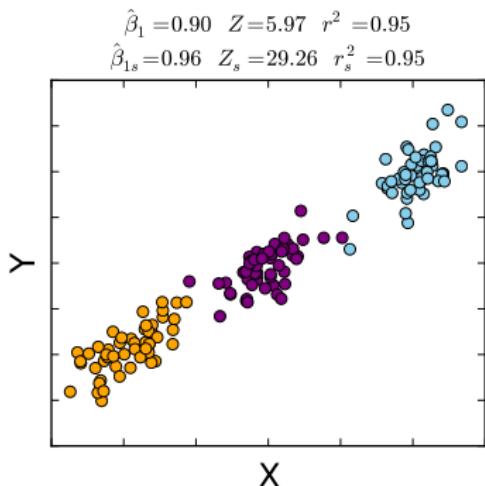
Right:  $y$  relates to clusters but not to  $x$ ;  $x$  and clusters are dependent; when cluster effects are not modeled their effect is picked up by  $x$ .



## Cluster effect examples

Left:  $y$  relates to both clusters and  $x$ ;  $x$  and clusters are dependent.

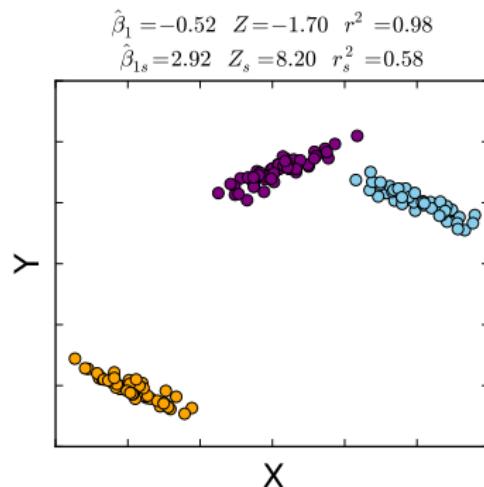
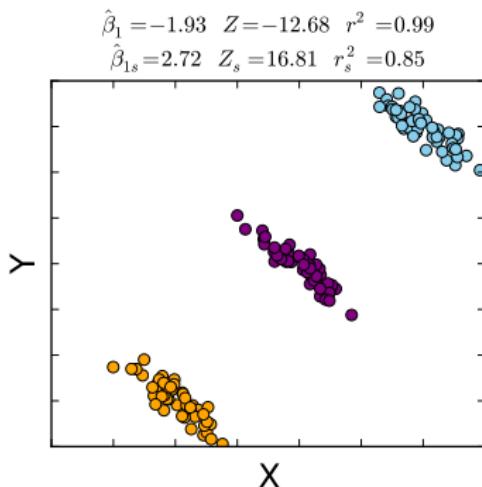
Right:  $y$  relates to clusters but not to  $x$ ;  $x$  and clusters are dependent but the net cluster effect is not linearly related to  $x$ .



## Cluster effect examples

Left:  $y$  relates to clusters and to  $x$ ;  $x$  and clusters are dependent; the  $x$  effect has the opposite sign as the net cluster effect.

Right:  $y$  relates to clusters and to  $x$ ;  $x$  and clusters are dependent; the signs and magnitudes of the  $x$  effects are cluster-specific.



## Implications of fixed effects analysis for observational data

A **stable confounder** is a confounding factor that is approximately constant within clusters. A stable confounder will become part of the net cluster effect.

If a stable confounder is correlated with an observed covariate  $X$ , then this will create non-orthogonality between the cluster effects and the effects of the observed covariates.

## Implications of fixed effects analysis for experimental data

Experiments are often carried out on “batches” of objects (specimens, parts, etc.) in which uncontrolled factors cause elements of the same batch to be more similar than elements of different batches.

If treatments are assigned randomly within each batch, there are no stable confounders (in general there are no confounders in experiments). Therefore the overall OLS estimate of  $\beta$  is unbiased as long as the standard linear model assumptions hold.

## Implications of fixed effects analysis for experimental data

Suppose the design is balanced (e.g. exactly half of each batch is treated and half is untreated). This is an orthogonal design, so the estimate based on the working model

$$\hat{y}_{ij} = \hat{\beta} \mathbf{x}_{ij}$$

and the estimate based on the working model

$$\hat{y}_{ij} = \hat{\theta}_i + \hat{\beta}^* \mathbf{x}_{ij}$$

are identical ( $\hat{\beta} = \hat{\beta}^*$ ). Thus they have the same variance. But the estimated variance of  $\hat{\beta}$  will be greater than the estimated variance of  $\hat{\beta}^*$  (since the corresponding estimate of  $\sigma^2$  is greater), so it will have lower power and wider confidence intervals.

## Random cluster effects

As we have seen, the cluster effects  $\theta_i$  can be treated as unobserved constants, and estimated as we would estimate any other regression coefficient.

An alternative way to handle cluster effects is to view the  $\theta_i$  as unobserved (latent) random variables.

In doing this, we now have two random variables in the model:  $\theta_i$  and  $\epsilon_{ij}$ , which are taken to be independent of each other.

If the  $\theta_i$  are independent and identically distributed, we can combine them with the error terms to get a single random error term per observation:

$$\epsilon_{ij}^c = \theta_i + \epsilon_{ij}.$$

## Random cluster effects

Let  $y_i \equiv (y_{i1}, \dots, y_{in_i})'$  denote the vector of responses in the  $i^{\text{th}}$  cluster, let  $\mathbf{x}_i \in \mathcal{R}^{n_i \times p}$  denote the matrix of predictor variables for the  $i^{\text{th}}$  cluster, and let  $\epsilon_i^c \equiv (\epsilon_{i1}^c, \dots, \epsilon_{in_i}^c)'$  denote the vector of random “errors” for the  $i^{\text{th}}$  cluster.

Thus we have the model

$$y_i = \mathbf{x}_i \beta + \epsilon_i^c,$$

for clusters  $i = 1, \dots, m$ . The  $\epsilon_i$  are taken to be uncorrelated between clusters, i.e.

$$\text{cov}(\epsilon_i^c, \epsilon_{i'}^c) = 0_{n_i \times n_{i'}}$$

for  $i \neq i'$ .

# Random cluster effects

The structure of the covariance matrix

$$S_i \equiv \text{cov}(\epsilon_i^c | \mathbf{x}_i)$$

is

$$S_i = \begin{pmatrix} \sigma^2 + \sigma_\theta^2 & \sigma_\theta^2 & \dots & \dots & \sigma_\theta^2 \\ \sigma_\theta^2 & \sigma^2 + \sigma_\theta^2 & \sigma_\theta^2 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_\theta^2 & \sigma_\theta^2 & \dots & \dots & \sigma^2 + \sigma_\theta^2 \end{pmatrix} = \sigma^2 I + \sigma_\theta^2 \mathbf{1}\mathbf{1}^T,$$

where  $\text{var}(\epsilon_{ij}) = \sigma^2$  and  $\text{var}(\theta_i) = \sigma_\theta^2$ .

## Generalized Least Squares

Suppose we have a linear model with mean structure  $E[y|\mathbf{X}] = \mathbf{X}\beta$  for  $y \in \mathcal{R}^n$ ,  $\mathbf{X} \in \mathcal{R}^{n \times p+1}$ , and  $\beta \in \mathcal{R}^{p+1}$ , and variance structure  $\text{Cov}[y|\mathbf{X}] \propto \Sigma$ , where  $\Sigma$  is a given  $n \times n$  matrix.

We can write the model in generative form as  $y = \mathbf{X}\beta + \epsilon$ , where  $\epsilon \in \mathcal{R}^n$  with  $E[\epsilon|\mathbf{X}] = 0$ ,  $\text{Cov}[\epsilon|\mathbf{X}] = \Sigma$ .

Factor the covariance matrix as  $\Sigma = GG'$ , and consider the transformed model

$$G^{-1}y = G^{-1}\mathbf{X}\beta + G^{-1}\epsilon.$$

Then letting  $\eta \equiv G^{-1}\epsilon$ , it follows that  $\text{Cov}(\eta) = I_{n \times n}$ , and note that the population slope vector  $\beta$  of the transformed model is identical to the population slope vector of the original model.

## Generalized Least Squares

The GLS estimator of  $\beta$  is defined to be the OLS estimator of  $\beta$  for the **decorrelated response**  $G^{-1}y$  and the **decorrelated predictors**  $G^{-1}\mathbf{X}$ .

The GLS estimate of the regression slope can be expressed in terms of the original design matrix  $\mathbf{X}$  and response vector  $y$ :

$$\begin{aligned}\hat{\beta}_{\text{GLS}} &= ((G^{-1}\mathbf{X})'G^{-1}\mathbf{X})^{-1}(G^{-1}\mathbf{X})'G^{-1}y \\ &= (\mathbf{X}'G^{-T}G^{-1}\mathbf{X})^{-1}\mathbf{X}'G^{-T}G^{-1}y \\ &= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}y.\end{aligned}$$

## Generalized Least Squares

We can use GLS to analyze clustered data with random cluster effects.

Let  $y_i^* \equiv S_i^{-1/2} y_i$ ,  $\epsilon_i^* \equiv S_i^{-1/2} \epsilon_i^c$ , and  $\mathbf{x}_i^* = S_i^{-1/2} \mathbf{x}_i$ .

Let  $y^*$ ,  $\mathbf{X}^*$ , and  $\epsilon^*$  denote the result of stacking  $y_i^*$ ,  $\mathbf{x}_i^*$ , and  $\epsilon_i^*$  over  $i$ , respectively.

## Generalized Least Squares

Since  $\text{cov}(\epsilon_i^*) \propto I$ , the OLS estimate of  $\beta$  for the model

$$y^* = \mathbf{X}^* \beta + \epsilon^*$$

is the best estimate of  $\beta$  that is linear in  $y^*$  (by the Gauss-Markov theorem).

Since the set of linear estimates based on  $y^*$  is the same as the set of linear estimates based on  $y$ , it follows that the GLS estimate of  $\beta$  based on  $y$  is the best unbiased estimate of  $\beta$  that is linear in  $y$ .

## Generalized Least Squares

The covariance matrix of  $\hat{\beta}$  in GLS is

$$\text{Cov}[\hat{\beta} | \mathbf{X}] = (\mathbf{X}^{*T} \mathbf{X}^*)^{-1} = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}$$

Note that there is no  $\sigma^2$  term on the left because we have transformed the error distribution to have covariance matrix  $I$ .

To apply GLS, it is only necessary to know  $\Sigma$  up to a multiplicative constant. The same estimated slopes  $\hat{\beta}$  are obtained if we decorrelate with  $\Sigma$  or with  $k\Sigma$  for  $k > 0$ . However we need to estimate  $\sigma^2$  and use

$$\widehat{\text{Cov}}[\hat{\beta} | \mathbf{X}] = \hat{\sigma}^2 (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}$$

for inference when  $\Sigma$  is only correct up to a scalar multiple.

## Generalized Least Squares

Since the sampling covariance matrix for GLS is proportional to  $(\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}$ , the parameter estimates are uncorrelated with each other if and only if  $\mathbf{X}^T \Sigma^{-1} \mathbf{X}$  is diagonal – that is, if the columns of  $\mathbf{X}$  are mutually orthogonal with respect to the Mahalanobis metric defined by  $\Sigma$ .

This is related to the fact that  $\hat{y}$  in GLS is the projection of  $y$  onto  $\text{col}(\mathbf{X})$  in the Mahalanobis metric defined by  $\Sigma$ . This generalizes the fact that  $\hat{y}$  in OLS is the projection of  $y$  onto  $\text{col}(\mathbf{X})$  in the Euclidean metric.

## Generalized least squares with a “working” covariance

What if we perform GLS using a possibly miss-specified “working covariance”  $S$ ?

For example, what happens if we use OLS when the actual covariance matrix of the errors is  $\Sigma \neq I$ ?

Since

$$E[\epsilon^* | \mathbf{X}^*] = E[\epsilon^* | \mathbf{X}] = 0,$$

the estimate  $\hat{\beta}$  remains unbiased. However it has two problems: it may not be the BLUE (i.e. it may not have the least variance among unbiased estimates), and the usual linear model inference procedures will be wrong.

## Generalized least squares with “working” covariance

The sampling covariance when the error structure is mis-specified is given by the “sandwich expression:”

$$\text{cov}[\hat{\beta}] = (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \Sigma_{\epsilon^*} \mathbf{X}^* (\mathbf{X}^{*'} \mathbf{X}^*)^{-1}$$

where

$$\Sigma_{\epsilon^*} = \text{cov}(\epsilon^* | \mathbf{X}^*).$$

This result covers two special situations: (i) we use OLS, so  $\mathbf{X}^* = \mathbf{X}$  and  $\Sigma_{\epsilon^*} = \Sigma_{\epsilon}$  is the covariance matrix of the errors, and (ii) we use a “working covariance” to define the decorrelating matrix  $G$ , and this working covariance is not equal to  $\Sigma^*$ .

## Generalized least squares with “working” covariance

Another way to write the covariance of  $\hat{\beta}$  is as follows:

$$\text{cov}[\hat{\beta}] = (\mathbf{X}' \Sigma_w^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_w^{-1} \Sigma \Sigma_w^{-1} \mathbf{X} (\mathbf{X}' \Sigma_w^{-1} \mathbf{X})^{-1}$$

where  $\Sigma_w$  is the “working” (possibly incorrectly specified) covariance matrix for  $\epsilon$ .

From this expression, it is clear that if  $\Sigma_w = \Sigma$  (the working model is correct), then

$$\text{cov}[\hat{\beta}] = (\mathbf{X}' \Sigma_w^{-1} \mathbf{X})^{-1} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1}.$$

## Iterated GLS

In practice, we will generally not know  $\Sigma$ , so it must be estimated from the data.  $\Sigma$  is the covariance matrix of the errors, so we can estimate  $\Sigma$  from the residuals.

Typically a low-dimensional parameteric model  $\Sigma = \Sigma(\alpha)$  is used.

Given a model for the the unexplained variation (i.e. for  $\Sigma$ ), then the fitting algorithm alternates between estimating  $\alpha$  and estimating the regression slopes  $\beta$ .

## Iterated GLS

For example, suppose our working covariance has the form

$$S_i(j, j) = \nu \quad S_i(j, k) = r\nu \quad (j \neq k).$$

This is an exchangeable model with “intraclass correlation coefficient” (ICC)  $r$  between two observations in the same cluster.

There are several closely-related ICC estimates. One approach is based on the standardized residuals  $R_{ij}^s$ :

$$\sum_i \sum_{j < j'} R_{ij}^s R_{ij'}^s / (\sum_i n_i(n_i - 1)/2 - p - 1)$$

where  $n_i$  is the size of the  $i^{\text{th}}$  cluster.

## Iterated GLS

Another common model for the error covariance is the first order autoregressive (AR-1) model:

$$\Sigma_{ij} = \alpha^{|i-j|},$$

where  $|\alpha| < 1$  is a parameter.

There are several possible estimates of  $\alpha$  borrowing ideas from time series analysis.

## Generalized Least Squares and stable confounders

For GLS to give unbiased estimates of  $\beta$ , we must have

$$E[\epsilon_{ij}^* | \mathbf{X}] = E[\theta_i + \epsilon_{ij} | \mathbf{X}] = 0.$$

Since  $E[\epsilon_{ij} | \mathbf{X}] = 0$ , this is equivalent to requiring that  $E[\theta_i | \mathbf{X}] = 0$ .

Thus if the covariates have distinct within-cluster mean values, and the within-cluster mean values of the covariates are correlated with the  $\theta_i$ , then the GLS estimate of  $\beta$  will be biased.

# Likelihood inference for the random intercepts model

## The random intercepts model

$$y_{ij} = \theta_i + \beta' \mathbf{x}_{ij} + \epsilon_{ij},$$

can be analyzed using a likelihood approach, using:

- ▶ A random intercepts density:

$$\phi(\theta | \mathbf{X})$$

- ▶ A density for the data given the random intercept:

$$f(y | \mathbf{x}, \theta)$$

## Multilevel models and conditional independence

The models are hierarchical (multilevel), and encode conditional independence relationships.

At the base level of the hierarchy, the random effect  $\theta_i$  are independent and identically distributed, and in particular are unrelated to  $\mathbf{X}$ :

$$p(\theta_1, \dots, \theta_m | \mathbf{X}) = p(\theta_1, \dots, \theta_m) = \prod_{i=1}^m \phi(\theta_i)$$

Conditionally on the random effects, the observed data  $\{y_{ij}\}$  are independent, and follow distributions that depend on the (unobserved) random effects and the observed covariates:

$$p(\{y_{ij}\} | \{\theta_i\}, \mathbf{X}) = \prod_{ij} f(y_{ij} | \theta_i, \mathbf{X})$$

## Likelihood inference for the random intercepts model

Since the random effects are not observed, we must estimate the parameters in terms of the **marginal model** for the observed data.

This is a model that depends on the **structural parameters**, which are  $(\beta, \sigma_\theta^2, \sigma^2)$  in this case.

The marginal density is obtained by integrating out the random effects

$$p(y_{i1}, \dots, y_{in_i} | \mathbf{X}) = \int f(y_{i1}, \dots, y_{in_i} | \mathbf{X}, \theta_i) \phi(\theta_i) d\theta_i.$$

## Likelihood inference for the random intercepts model

For linear multilevel models, the marginal distribution  $p(y|\mathbf{X})$  can be calculated explicitly. It is Gaussian, and therefore is characterized by its moments:

$$E[y_{ij}|\mathbf{x}_{ij}] = \beta' \mathbf{x}_{ij}$$

$$\text{Var}[y_{ij}|\mathbf{x}_{ij}] = \sigma_\theta^2 + \sigma^2,$$

$$\text{Cov}[y_{i_1j_1}, y_{i_2j_2} | \mathbf{x}_{i_1j_1}, \mathbf{x}_{i_2j_2}] = 0 \quad (\text{if } i_1 \neq i_2).$$

$$\text{Cov}[y_{ij_1}, y_{ij_2} | \mathbf{x}_{ij_1}, \mathbf{x}_{ij_2}] = \sigma_\theta^2 \quad (\text{if } j_1 \neq j_2).$$

## Marginal form of the generative model

Suppose

$$\epsilon|\mathbf{X} \sim N(0, \sigma^2) \quad \theta|\mathbf{X} \sim N(0, \sigma_\theta^2).$$

In this case  $y|\mathbf{X}$  is Gaussian, with mean and variance as given above. Thus the random intercept model can be equivalently written in marginal form as

$$y_{ij} = \beta' \mathbf{x}_{ij} + \epsilon_{ij}^*$$

where  $\epsilon_{ij}^* = \theta_i + \epsilon_{ij}$ .

It follows that  $E[\epsilon_{ij}^*|\mathbf{X}] = 0$ ,  $\text{Var}[\epsilon_{ij}^*|\mathbf{X}] = \sigma_\theta^2 + \sigma^2$ , and the  $\epsilon_{ij}^*$  values have correlation coefficient  $\sigma_\theta^2 / (\sigma^2 + \sigma_\theta^2)$  within clusters.

## Likelihood computation

Maximum likelihood estimates for the model can be calculated using a gradient-based optimization procedure applied to the marginal log-density, or using the EM algorithm.

Asymptotic standard errors can be obtained from the inverse of the Fisher information matrix. Likelihood ratio tests, AIC values, and other likelihood-based inference tools can be used.

For linear multilevel models, the parameter estimates from iterated GLS and the MLE for the random intercepts model will be similar but not identical. Both are consistent, and in general will be asymptotically equivalent.

## Predicting the random intercepts

Since the model is fit by optimizing the marginal log-likelihood, we obtain an estimate of  $\sigma_\theta^2 \equiv \text{Var}[\theta_i]$ , but we don't automatically learn anything about the individual  $\theta_i$ .

If there is an interest in the individual  $\theta_i$  values, we can predict them using the **best linear unbiased predictor** (BLUP).

The population version of the BLUP is:

$$E_{\beta, \sigma^2, \sigma_\theta^2}[\theta_i | y_i, \mathbf{X}] = \text{Cov}[\theta_i, y_i | \mathbf{X}] \cdot \text{Cov}[y_i | \mathbf{X}]^{-1} \cdot (y_i - E[y_i | \mathbf{X}])$$

Since this depends on things we don't know, in practice we use the sample version of the BLUP (sometimes called the eBLUP):

$$E_{\hat{\beta}, \hat{\sigma}^2, \hat{\sigma}_\theta^2}[\theta_i | y_i, \mathbf{X}] = \widehat{\text{Cov}}[\theta_i, y_i | \mathbf{X}] \cdot \widehat{\text{Cov}}[y_i | \mathbf{X}]^{-1} \cdot (y_i - \hat{E}[y_i | \mathbf{X}])$$

## Predicting the random intercepts

The BLUP is truly a linear function of  $y$ , is unbiased, and is “best” in the sense of minimizing the expected squared prediction error.

However the eBLUP is none of these things (it is not linear or unbiased, and it is unclear if it is “best”).

Note also that due to the hierarchical structure of the model, to predict  $\theta_i$  we only need to condition on  $y_i$  (the other  $y_{i'}$ , for  $i' \neq i$ , contain no information). Thus the BLUP for  $\theta_i$  only depends on the data through  $y_i$ . But in the eBLUP, all the data are used indirectly, through the estimates of the structural parameters.

## Predicting the random intercepts

The estimated second moments needed to calculate the BLUP are:

$$\widehat{\text{Cov}}[\theta_i, y_i | \mathbf{X}] = \hat{\sigma}_\theta^2 \cdot \mathbf{1}_{n_i}$$

and

$$\widehat{\text{Cov}}[y_i | \mathbf{X}] = \hat{\sigma}^2 I + \hat{\sigma}_\theta^2 \mathbf{1} \mathbf{1}^T.$$

where  $n_i$  is the size of the  $i^{\text{th}}$  group.

## Predicting the random intercepts

For a given set of parameter values, the BLUP for the random intercepts model is a linear function of the data, with the following form:

$$E_{\hat{\beta}, \hat{\sigma}^2, \hat{\sigma}_\theta^2}[\theta_i | y, \mathbf{X}] = n_i \sigma_\theta^2 / (\hat{\sigma}^2 + n_i \hat{\sigma}_\theta^2) \cdot \mathbf{1}^T (y_i - \hat{E}[y_i | \mathbf{X}]) / n_i,$$

## Predicting the random intercepts

In the BLUP for  $\theta_i$ , this term is the mean of the residuals:

$$\mathbf{1}^T(y_i - \hat{E}[y_i|\mathbf{X}])/n_i$$

and it is shrunk by this factor:

$$n_i\sigma_\theta^2/(\hat{\sigma}^2 + n_i\hat{\sigma}_\theta^2)$$

This shrinkage allows us to interpret the random intercepts model as a “partially pooled” model that is intermediate between the fixed effects model and the model that completely ignores the clusters.

## Predicting the random intercepts

In the fixed effects model, the parameter estimates  $\theta_i$  are unbiased, but they are “overdispersed”, meaning that the sample variance of the  $\hat{\theta}_i$  will generally be greater than  $\sigma_\theta^2$ .

In the random intercepts model, the variance parameter  $\sigma_\theta^2$  is estimated with low bias, and the BLUP's of the  $\theta_i$  are shrunk toward zero.

## Random slopes

Suppose we are interested in a measured covariate  $z$  whose effect may vary by cluster. We might start with the model

$$y_i = \mathbf{x}_i\beta + \gamma z_i + \epsilon,$$

For example, suppose that  $z_i \in \{0, 1\}^{n_i}$  is a vector of treatment indicators (1 for treated subjects, 0 for untreated subjects). Then  $\gamma$  is the average change associated with being treated (the population treatment effect).

In many cases, it is reasonable to consider the possibility that different clusters may have different treatment effects – that is, different clusters have different  $\gamma$  values. In this case we can let  $\gamma_i$  be the treatment response for cluster  $i$ .

## Random slopes

Suppose we model the random slopes  $\gamma_i$  as being Gaussian (given  $X$ ) with variance  $\sigma_\gamma^2$ . The marginal model is

$$E[y_i | \mathbf{x}_i, z_i] = \mathbf{x}'_i \beta$$

and

$$\text{Cov}[y_i | \mathbf{x}_i, z_i] = \sigma_\gamma^2 z_i z'_i + \sigma^2 I.$$

Again we can form a BLUP  $E_{\hat{\beta}, \hat{\sigma}_\gamma^2, \hat{\sigma}^2}[\gamma_i | y, \mathbf{X}, z]$ , and the BLUP turns out to be a shrunken version of what would be obtained in a fixed effects model, where we regress  $y$  on  $\mathbf{x}$  and  $z$  within every cluster.

## Linear mixed effects models

The random intercept and random slope model are special cases of the **linear mixed effects model**:

$$y_i = \mathbf{X}_i\beta + Z_i\gamma_i + \epsilon_i$$

Here,  $\mathbf{X}_i$  is a  $n_i \times p$  design matrix for cluster  $i$ ,  $Z$  is a  $n_i \times q$  random effects design matrix for cluster  $i$ ,  $\gamma_i \in \mathcal{R}^q$  is a random vector with mean 0 and covariance matrix  $\Psi$ , and  $\epsilon_i \in \mathcal{R}^{n_i}$  is a random vector with mean 0 and covariance matrix  $\sigma^2 I$ .

# Generalized Linear Models

Kerby Shedden

Department of Statistics, University of Michigan

December 9, 2019

## Motivation for nonlinear models

The key properties of a linear model are that

$$E[y|\mathbf{X}] = \mathbf{X}\beta \quad \text{and} \quad \text{var}[y|\mathbf{X}] \propto I.$$

In some cases where these conditions are not met, we can transform  $y$  so that the properties of a linear model are well-satisfied.

However it is often difficult to find a transformation that simultaneously linearizes the mean and gives constant variance.

Also, if  $y$  lies in a restricted domain (e.g.  $y_i \in \{0, 1\}$ ), parameterizing  $E[y|\mathbf{X}]$  as a linear function of  $\mathbf{X}$  violates the domain restriction.

**Generalized linear models** (GLM's) are a class of nonlinear regression models that can be used in certain cases where linear models do not fit well.

## Logistic regression

Logistic regression is a specific type of GLM. We will develop logistic regression from first principles before discussing GLM's in general.

Logistic regression is used for binary outcome data, where  $y_i = 0$  or  $y_i = 1$ . It is defined by the probability mass function

$$P(y_i = 1 | \mathbf{x}_i = \mathbf{x}) = \frac{\exp(\beta' \mathbf{x})}{1 + \exp(\beta' \mathbf{x})} = \frac{1}{1 + \exp(-\beta' \mathbf{x})},$$

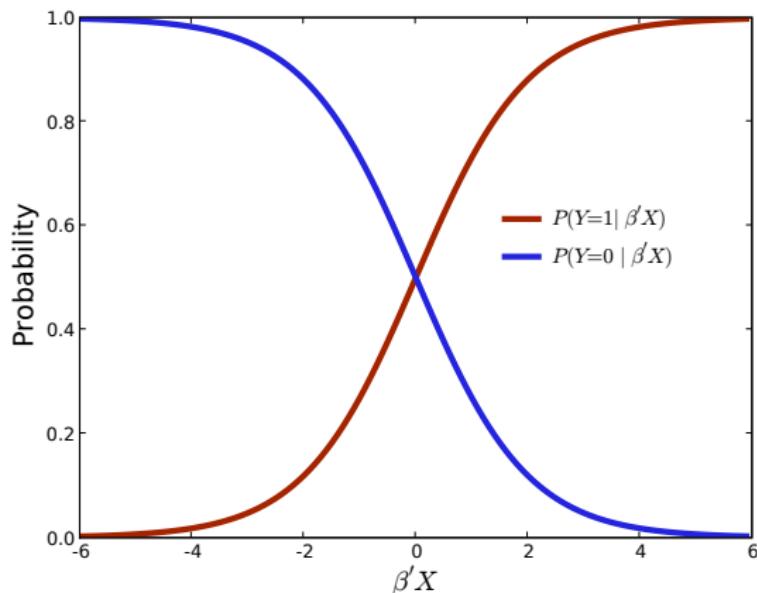
which implies that

$$P(y_i = 0 | \mathbf{x}_i = \mathbf{x}) = 1 - P(y_i = 1 | \mathbf{x}_i = \mathbf{x}) = \frac{1}{1 + \exp(\beta' \mathbf{x})},$$

where in most cases,  $\mathbf{x}(0) = 1$  so  $\beta_0$  is the intercept.

# Logistic regression

This plot shows  $P(y = 1|\mathbf{x})$  and  $P(y = 0|\mathbf{x})$ , plotted as functions of  $\beta' \mathbf{x}$ :



# Logistic regression

## The logit function

$$\text{logit}(x) = \log(x/(1-x))$$

maps the unit interval onto the real line. The inverse logit function, or expit function

$$\text{expit}(x) = \text{logit}^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}$$

maps the real line onto the unit interval.

In logistic regression, the logit function is used to map the linear predictor  $\beta'x$  to a probability.

## Logistic regression

The linear predictor in logistic regression is the **conditional log odds**:

$$\log \left[ \frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})} \right] = \beta' \mathbf{x}.$$

Thus one way to interpret a logistic regression model is that a one unit increase in  $\mathbf{x}(j)$  results in a change of  $\beta_j$  in the conditional log odds.

Or, a one unit increase in  $\mathbf{x}(j)$  results in a multiplicative change of  $\exp(\beta_j)$  in the conditional odds.

## Latent variable model for logistic regression

It may make sense to view the binary outcome  $y$  as being a dichotomization of a latent continuous outcome  $y_c$ ,

$$y = \mathcal{I}(y_c \geq 0).$$

Suppose  $y_c|X$  follows a logistic distribution, with CDF

$$F(y_c|\mathbf{x}) = \frac{\exp(y_c - \beta' \mathbf{x})}{1 + \exp(y_c - \beta' \mathbf{x})}.$$

In this case,  $y|\mathbf{x}$  follows the logistic regression model:

$$P(y = 1|\mathbf{x}) = P(y_c \geq 0|\mathbf{x}) = 1 - \frac{\exp(0 - \beta' \mathbf{x})}{1 + \exp(0 - \beta' \mathbf{x})} = \frac{\exp(\beta' \mathbf{x})}{1 + \exp(\beta' \mathbf{x})}.$$

## Mean/variance relationship for logistic regression

Since the mean and variance of a Bernoulli trial are linked, the mean structure

$$E[y|\mathbf{x}] = P(y = 1|\mathbf{x}) = \frac{\exp(\beta' \mathbf{x})}{1 + \exp(\beta' \mathbf{x})}$$

also determines the variances

$$\text{var}[y|\mathbf{x}] = P(y = 1|\mathbf{x}) \cdot P(y = 0|\mathbf{x}) = \frac{1}{2 + \exp(\beta' \mathbf{x}) + \exp(-\beta' \mathbf{x})}.$$

Since the variance depends on  $\mathbf{x}$ , logistic regression models are always heteroscedastic.

# Logistic regression and case-control studies

Suppose we sample people based on their disease status  $D$  ( $D = 1$  is a **case**,  $D = 0$  is a **control**).

We are interested in a binary marker  $M \in \{0, 1\}$  that may predict a person's disease status.

The **prospective log odds**

$$\log \left[ \frac{P(D = 1|M = m)}{P(D = 0|M = m)} \right]$$

measures how informative the marker is for the disease.

# Logistic regression and case-control studies

Suppose we model  $P(M = m|D = d)$  using logistic regression, so

$$P(M = 1|D = d) = \frac{\exp(\alpha + \beta d)}{1 + \exp(\alpha + \beta d)}$$

$$P(M = 0|D = d) = \frac{1}{1 + \exp(\alpha + \beta d)}.$$

More generally,

$$P(M = m|D = d) = \frac{\exp(m(\alpha + \beta d))}{1 + \exp(\alpha + \beta d)}.$$

# Logistic regression and case-control studies

Since

$$\log \frac{P(M = 1|D = d)}{P(M = 0|D = d)} = \alpha + \beta d$$

we see that  $\beta$  is the coefficient of  $d$  in the retrospective log odds.

## Logistic regression and case-control studies

The prospective log odds can be written

$$\begin{aligned}\log \frac{P(D = 1|M = m)}{P(D = 0|M = m)} &= \log \frac{P(M = m|D = 1)P(D = 1)/P(M = m)}{P(M = m|D = 0)P(D = 0)/P(M = m)} \\ &= \log \frac{P(M = m|D = 1)P(D = 1)}{P(M = m|D = 0)P(D = 0)}\end{aligned}$$

## Logistic regression and case-control studies

Continuing from the previous slide, we have

$$\log \frac{P(M = m|D = 1)P(D = 1)}{P(M = m|D = 0)P(D = 0)} = \\ \log \left[ \frac{\exp(m \cdot (\alpha + \beta))/(1 + \exp(\alpha + \beta))}{\exp(m \cdot \alpha)/(1 + \exp(\alpha))} \cdot \frac{P(D = 1)}{P(D = 0)} \right],$$

which equals

$$\beta m + \log \left[ \frac{1 + \exp(\alpha)}{1 + \exp(\alpha + \beta)} \cdot \frac{P(D = 1)}{P(D = 0)} \right].$$

Thus  $\beta$  is both the coefficient of  $d$  in the retrospective log odds, and it is the coefficient of  $m$  in the prospective log odds. This is sometimes called **case/control convertibility**.

## Estimation and inference for logistic regression

Assuming independent cases, the log-likelihood for logistic regression is

$$\begin{aligned} L(\beta|y, \mathbf{X}) &= \log \prod_i \frac{\exp(y_i \cdot \beta' \mathbf{x}_i)}{1 + \exp(\beta' \mathbf{x}_i)} \\ &= \sum_{i:y_i=1} \beta' \mathbf{x}_i - \sum_i \log(1 + \exp(\beta' \mathbf{x}_i)). \end{aligned}$$

This likelihood is for the conditional distribution of  $y$  given  $\mathbf{X}$ .

As in linear regression, we do not model the marginal distribution of  $\mathbf{x}$  (a row of  $\mathbf{X}$ ).

## Estimation and inference for logistic regression

Logistic regression models are usually fit using maximum likelihood estimation.

This means that the parametric likelihood above is maximized as a function of  $\beta$ .

The gradient of the log-likelihood function (the **score function**) is

$$\begin{aligned} G(\beta|y, \mathbf{X}) &= \frac{\partial}{\partial \beta} L(\beta|y, \mathbf{X}) \\ &= \sum_{i:y_i=1} \mathbf{x}_i - \sum_i \frac{\exp(\beta' \mathbf{x}_i)}{1 + \exp(\beta' \mathbf{x}_i)} \mathbf{x}_i \\ &= \sum_i \left( y_i - \frac{\exp(\beta' \mathbf{x}_i)}{1 + \exp(\beta' \mathbf{x}_i)} \right) \mathbf{x}_i. \end{aligned}$$

## Estimation and inference for logistic regression

The Hessian of the log-likelihood is

$$H(\beta|y, \mathbf{X}) = \frac{\partial^2}{\partial \beta \partial \beta'} L(\beta|y, \mathbf{X}) = - \sum_i \frac{\exp(\beta' \mathbf{x}_i)}{(1 + \exp(\beta' \mathbf{x}_i))^2} \mathbf{x}_i \mathbf{x}_i'.$$

The Hessian is strictly negative definite as long as the design matrix has independent columns. Therefore  $L(\beta|y, \mathbf{X})$  is a concave function of  $\beta$ , so has a unique maximizer, and hence the MLE is unique.

## Estimation and inference for logistic regression

From the general theory of the MLE, the Fisher information

$$I(\beta) = -(E[H(\beta|y, \mathbf{X})|\mathbf{X}])^{-1}$$

is the asymptotic sampling covariance matrix of the MLE  $\hat{\beta}$ . Since  $H(\beta|y, \mathbf{X})$  does not depend on  $y$ ,  $I(\beta) = -H(\beta|y, \mathbf{X})^{-1}$ .

Since  $\hat{\beta}$  is an MLE for a regular problem, it is consistent, asymptotically unbiased, and asymptotically normal if the model is correctly specified.

## Poisson regression

The Poisson distribution is a single-parameter family of distributions on the sample space  $\{0, 1, 2, \dots\}$ .

A key property of the Poisson distribution is that the mean is equal to the variance.

The Poisson distribution is usually parameterized in terms of a parameter  $\lambda$  that is equal to the common mean and variance.

In regression, we don't want just a single distribution. Instead we want a family of distributions indexed by the covariate vector  $\mathbf{x}$ .

To create a regression methodology based on the Poisson distribution, we can formulate a regression model in which  $y|\mathbf{x}$  is Poisson, with mean and variance equal to  $\lambda(x) = \exp(\beta' \mathbf{x})$ .

## Poisson regression

Since the mean function in a Poisson distribution has an exponential form, the covariates are related multiplicatively to the mean.

If we contrast the mean value for two different covariate vectors,  $\mathbf{x}^{(1)}$  and  $\mathbf{x}^{(2)}$ , such that  $x_j^{(1)} - x_j^{(2)} = 1$ , and  $x_k^{(1)} = x_k^{(2)}$  for  $k \neq j$ , then the means at these two points are related through:

$$\lambda(\mathbf{x}^{(1)}) = \exp(\beta_j) \lambda(\mathbf{x}^{(2)}).$$

## Poisson regression

Setting the mean to be  $\lambda_i = \exp(\beta' \mathbf{x}_i)$ , the PMF for one observation in a Poisson regression model is

$$\exp^{-\lambda_i} \lambda_i^{y_i} / y_i!$$

The corresponding contribution to the log likelihood is

$$-\lambda_i + y_i \log(\lambda_i) - \log(y_i!) = -\exp(\beta' \mathbf{x}_i) + y_i \cdot \beta' \mathbf{x}_i - \log(y_i!),$$

and the contribution to the score function is

$$-\mathbf{x}_i \exp(\beta' \mathbf{x}_i) + y_i \cdot \mathbf{x}_i = (y_i - \exp(\beta' \mathbf{x}_i)) \mathbf{x}_i.$$

## Score equations

The MLE is a stationary point of the score function. Thus, for logistic regression, the following equation is satisfied at the MLE:

$$\sum_i \left( y_i - \frac{\exp(\beta' \mathbf{x}_i)}{1 + \exp(\beta' \mathbf{x}_i)} \right) \mathbf{x}_i = 0.$$

For Poisson regression, this equation is satisfied at the MLE:

$$\sum_i (y_i - \exp(\beta' \mathbf{x}_i)) \mathbf{x}_i$$

We also know that for OLS (viewed here as a Gaussian regression model), this equation is satisfied at the MLE

$$\sum_i (y_i - \beta' \mathbf{x}_i) \mathbf{x}_i = 0.$$

## Score equations

Writing  $\mu_i = E[y_i | \mathbf{x}_i]$ , we see that for all three types of regression models, the following equation is satisfied.

$$\sum_i (y_i - \mu_i) x_i = 0.$$

This shows that the residuals are orthogonal to each covariate in all of these models, and that achieving this orthogonality characterizes the MLE.

This turns out to be a useful generic framework for regression, as many different mean functions  $\mu(\beta)$ , and variance functions  $v(\mu)$  or  $v(\beta)$  can be substituted into this equation, and the solution of the equation can be used to estimate  $\beta$ .

## Relationship between the mean and variance

We have seen three parametric regression models, each of which expresses the mean in terms of the **linear predictor**. The **family** is the distributional family used to form the log-likelihood and score functions.

For each of these models, the variance can also be related to the mean.

Family	Mean ( $\mu$ )
Gaussian	$\beta'x$
Binomial	$1/(1 + \exp(-\beta'x))$
Poisson	$\exp(\beta'x)$

Note that in each case,  $d\mu/d\beta$  is proportional to  $v \cdot x$ , where  $v$  is the variance.

## Estimating equations

For the three examples we are focusing on here, the MLE can be defined as the solution to the **estimating equations**:

$$\sum_i \partial\mu_i/\partial\beta \cdot (y_i - \mu_i(\beta))/v_i(\beta) = 0$$

Note that this is a system of  $p = \dim(x)$  equations in  $p$  unknowns. It should be solvable unless there is some degeneracy in the equations.

In the “canonical setting”,  $(d\mu_i/d\beta)/v_i(\beta) = x_i$ , so these equations are generally equivalent to the orthogonality between residuals and covariates. But we will see below that the form given here extends to some “non-canonical” settings and hence is somewhat more general.

## General development of GLM's

A GLM is based on the following conditions:

- ▶ The  $y_i$  are conditionally independent given  $\mathbf{X}$ .
- ▶ The probability mass function or density can be written

$$\log p(y_i|\theta_i, \phi, \mathbf{x}_i) = w_i(y_i\theta_i - \gamma(\theta_i))/\phi + \tau(y_i, \phi/w_i),$$

where  $w_i$  is a known weight,  $\theta_i = g(\beta' \mathbf{x}_i)$  for an unknown vector of regression slopes  $\beta$ ,  $g(\cdot)$  and  $\gamma(\cdot)$  are smooth functions,  $\phi$  is the “scale parameter” (which may be either known or unknown), and  $\tau(\cdot)$  is a known function.

## General development of GLM's

The log-likelihood function is

$$L(\beta, \phi | y, \mathbf{X}) = \sum_i w_i (y_i \theta_i - \gamma(\theta_i)) / \phi + \tau(y_i, \phi / w_i).$$

The score function with respect to  $\theta_i$  is

$$w_i (y_i - \gamma'(\theta_i)) / \phi.$$

## General development of GLM's

Next we need a fundamental fact about score functions.

Let  $f_\theta(y)$  be a density in  $y$  with parameter  $\theta$ . The score function is

$$\frac{\partial}{\partial \theta} \log f_\theta(y) = f_\theta(y)^{-1} \frac{\partial}{\partial \theta} f_\theta(y).$$

The expected value of the score function is

$$\begin{aligned} E \frac{\partial}{\partial \theta} \log f_\theta(y) &= \int f_\theta(y)^{-1} \left( \frac{\partial}{\partial \theta} f_\theta(y) \right) f_\theta(y) dy \\ &= \frac{\partial}{\partial \theta} \int f_\theta(y) dy \\ &= 0. \end{aligned}$$

Thus the score function has expected value 0 when  $\theta$  is at its true value.

## General development of GLM's

Since the expected value of the score function is zero, we can conclude that

$$E[w_i(y_i - \gamma'(\theta_i))/\phi | \mathbf{X}] = 0,$$

so

$$E[y_i | \mathbf{X}] = \gamma'(\theta_i) = \gamma'(g(\beta' \mathbf{x}_i)).$$

Note that this relationship does not depend on  $\phi$  or  $\tau$ .

## General development of GLM's

Using a similar approach, we can relate the variance to  $w_i$ ,  $\phi$ , and  $\gamma'$ . By direct calculation,

$$\partial^2 L(\theta_i | y_i, \mathbf{x}_i, \phi) / \partial \theta_i^2 = -w_i \gamma''(\theta_i) / \phi.$$

Returning to the general density  $f_\theta(y)$ , we can write the Hessian as

$$\frac{\partial}{\partial \theta \theta'} \log f_\theta(y) = f_\theta(y)^{-2} \left( f_\theta(y) \frac{\partial^2}{\partial \theta \theta'} f_\theta(y) - \frac{\partial f_\theta(y)}{\partial \theta} \cdot \frac{\partial f_\theta(y)}{\partial \theta'} \right).$$

## General development of GLM's

The expected value of the Hessian is

$$\begin{aligned} E \frac{\partial}{\partial \theta \theta'} \log f_\theta(y) &= \int \frac{\partial}{\partial \theta \theta'} \log f_\theta(y) \cdot f_\theta(y) dy \\ &= \frac{\partial}{\partial \theta \theta'} \int f_\theta(y) dy - \int \left( \frac{\partial f_\theta(y)/\partial \theta}{f_\theta(y)} \cdot \frac{\partial f_\theta(y)/\partial \theta'}{f_\theta(y)} \right) f_\theta(y) dy \\ &= -\text{cov} \left( \frac{\partial}{\partial \theta} \log f_\theta(y) | \mathbf{X} \right). \end{aligned}$$

Therefore

$$w_i \gamma''(\theta_i)/\phi = \text{var}(w_i(y_i - \gamma'(\theta_i))/\phi | \mathbf{X})$$

$$\text{so } \text{var}[y_i | \mathbf{X}] = \phi \gamma''(\theta_i)/w_i.$$

## Examples of GLM's

**Gaussian linear model:** The density of  $y|\mathbf{X}$  can be written

$$\begin{aligned}\log p(y_i|\mathbf{x}_i) &= -\log(2\pi\sigma^2)/2 - \frac{1}{2\sigma^2}(y_i - \beta'\mathbf{x}_i)^2 \\ &= -\log(2\pi\sigma^2)/2 - y_i^2/2\sigma^2 + (y_i\beta'\mathbf{x}_i - (\beta'\mathbf{x}_i)^2/2)/\sigma^2.\end{aligned}$$

This can be put into GLM form by setting  $g(x) = x$ ,  $\gamma(x) = x^2/2$ ,  $w_i = 1$ ,  $\phi = \sigma^2$ , and  $\tau(y_i, \phi) = -\log(2\pi\phi)/2 - y_i^2/2\phi$ .

## Examples of GLM's

**Logistic regression:** The mass function of  $Y|X$  can be written

$$\begin{aligned}\log p(y_i|\mathbf{x}_i) &= y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \\ &= y_i \log(p_i/(1 - p_i)) + \log(1 - p_i),\end{aligned}$$

where

$$p_i = \text{logit}^{-1}(\beta' \mathbf{x}_i) = \frac{\exp(\beta' \mathbf{x}_i)}{1 + \exp(\beta' \mathbf{x}_i)}.$$

Since  $\log(p_i/(1 - p_i)) = \beta' \mathbf{x}$ , this can be put into GLM form by setting  $g(x) = x$ ,  $\gamma(x) = -\log(1 - \text{logit}^{-1}(x)) = \log(1 + \exp(x))$ ,  $\tau(y_i, \phi) \equiv 0$ ,  $w = 1$ , and  $\phi = 1$ .

## Examples of GLM's

**Poisson regression:** In Poisson regression, the distribution of  $Y|X$  follows a Poisson distribution, with the mean response related to the covariates via

$$\log E[y|\mathbf{x}] = \beta' \mathbf{x}.$$

It follows that  $\log \text{var}[y|\mathbf{x}] = \beta' \mathbf{x}$  as well. The mass function can be written

$$\log p(y_i|\mathbf{x}_i) = y_i \beta' \mathbf{x}_i - \exp(\beta' \mathbf{x}_i) - \log(y_i!),$$

so in GLM form,  $g(x) = x$ ,  $\gamma(x) = \exp(x)$ ,  $w = 1$ ,  
 $\tau(y_i) = -\log(y_i!)$ , and  $\phi = 1$ .

## Examples of GLM's

**Negative binomial regression:** In negative binomial regression, the probability mass function for the dependent variable  $Y$  is

$$P(y_i = y | \mathbf{x}) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^y.$$

The mean of this distribution is  $\mu_i$  and the variance is  $\mu_i + \alpha\mu_i^2$ . If  $\alpha = 0$  we get the same mean/variance relationship as the Poisson model. As  $\alpha$  increases, we get increasingly more overdispersion.

## Examples of GLM's

### Negative binomial regression (continued):

The log-likelihood (dropping terms that do not involve  $\mu$ ) is

$$\log P(y_i = y | \mathbf{x}_i) = y \log\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right) - \alpha^{-1} \log(1 + \alpha\mu_i)$$

Suppose we model the mean as  $\mu_i = \exp(\beta' \mathbf{x}_i)$ . Then in the standard GLM notation, we have

$$\theta_i = \log\left(\frac{\alpha \exp(\beta' X_i)}{1 + \alpha \exp(\beta' \mathbf{x}_i)}\right),$$

so  $g(x) = \log(\alpha) + x - \log(1 + \alpha \exp(x))$ , and  
 $\gamma(x) = -\alpha^{-1} \log(1 - \exp(x))$ .

## Link functions

In a GLM, the link function maps the mean to the linear predictor  $\eta_i = \mathbf{x}'_i \beta$ . Since

$$E[y_i | \mathbf{x}_i] = \gamma'(g(\eta)),$$

it follows that the link function is the inverse of  $\gamma' \circ g$ .

For example, in the case of logistic regression,

$$\gamma'(g(\eta)) = \exp(\eta) / (1 + \exp(\eta)),$$

which is the expit function. The inverse of this function is the logit function  $\log(p/(1 - p))$ , so the logit function is the link in this case.

## Link functions

When  $g(x) = x$ , the resulting link function is called the **canonical link function**.

In the examples above, linear regression, logistic regression, and Poisson regression all used the canonical link function, but negative binomial regression did not.

The canonical link function for negative binomial regression is  $1/x$ , but this does not respect the domain and is harder to interpret than the usual log link.

Another setting where non-canonical links arise is the use of the log link function for logistic regression. In this case, the coefficients  $\beta$  are related to the log relative risk rather than to the log odds.

## Estimating equations and quasi-likelihood

As noted above, the regression parameters in a GLM can be estimated by solving these estimating equations:

$$\sum_i \partial\mu_i/\partial\beta \cdot (y_i - \mu_i(\beta))/v_i(\beta) = 0$$

Note that we only need to correctly specify  $v_i(\beta)$  up to a constant. For example, in the Gaussian case, we can set  $v_i(\beta) = 1$ .

## Estimating equations and quasi-likelihood

This opens up the possibility of specifying a large class of regression models through their first two moments, represented by the functions  $\mu(\beta)$  and  $v(\beta)$ .

For example, we can get quasi-Poisson regression by specifying  $\mu_i(\beta) = \exp(x'_i\beta)$  and  $v_i(\beta) = \mu_i(\beta)$ . This formulation of quasi-Poisson regression never refers to the Poisson distribution directly, it only depends on moments.

It can be shown that solving the quasi-likelihood equations generally gives consistent estimates of  $\beta$ , as long as the data are sample from a population in which the specified mean and variance functions are correct.

## Estimating equations and quasi-likelihood

Wedderburn introduced a “quasi-likelihood” function that can be used when working with estimating equations. It has the form

$$Q(y; \mu, v) = \int_0^\mu \frac{y - u}{v(u)} du$$

Since

$$\partial Q / \partial \beta = \partial \mu / \partial \beta \cdot \partial Q / \partial \mu,$$

and  $\partial Q / \partial \mu = (y - \mu) / v(\mu)$  by the fundamental theorem of calculus, we see that  $\partial Q / \partial \beta$  gives the estimating equations discussed above.

## Estimating equations and quasi-likelihood

In some cases the quasi-likelihood is an actual likelihood, but even if it is not, we can use it in place of a likelihood for many purposes.

For example, we can define a “Quasi Information Criterion” QIC, analogous to AIC for model selection as

$$\sum_i Q(y_i, \hat{\mu}_i, v) - p,$$

where  $p = \dim(\beta)$ .

## Scale parameters and quasi-likelihood

Since the quasi-likelihood estimating equations are homogeneous, we can estimate the mean structure  $\mu = \exp(\beta' \mathbf{x})$  in a setting where the specified variance is off by a multiplicative constant. For example, these estimating equations can be used to consistently estimate  $\beta$  in a **quasi-Poisson** model where  $\text{Var}[y_i | \mathbf{x}_i] = \phi E[y_i | \mathbf{x}_i]$ .

This is a quasi-likelihood estimator, because there is no single "quasi-Poisson distribution". There are many distributions that have this variance structure, but the solution to these estimating equations is not the MLE for a specific distribution.

This can be viewed as a way to construct a consistent estimator that can be used for any distribution where the conditional variance has this structure.

## Scale parameters and quasi-likelihood

In a quasi-likelihood analysis, the scale parameter is usually estimated in a separate step, after the regression parameters ( $\beta$ ) are estimated by solving the estimating equations.

There are several related ways to estimate the scale parameter. A common approach is to use

$$\hat{\phi} = \frac{\sum_i (y_i - \hat{\mu}_i)^2 / \hat{v}_i}{n - p}.$$

## Overdispersion

Under the Poisson model,  $\text{var}[y|\mathbf{x}] = E[y|\mathbf{x}]$ . A Poisson model results from using the Poisson GLM with the scale parameter  $\phi$  fixed at 1.

The **quasi-Poisson** model is the Poisson model with a scale parameter that may be any non-negative value. Under the quasi-Poisson model,  $\text{var}[y|\mathbf{x}] \propto E[y|\mathbf{x}]$ .

The negative binomial GLM allows the variance to be non-proportional to the mean.

Any situation in which  $\text{var}[y|\mathbf{x}] > E[y|\mathbf{x}]$  is called **overdispersion**. Overdispersion is often seen in practice.

One mechanism that may give rise to overdispersion is **heterogeneity**. Suppose we have a hierarchical model in which  $\lambda$  follows a  $\Gamma$  distribution, and  $y|\lambda$  is Poisson with mean parameter  $\lambda$ . Then marginally,  $y$  is negative binomial.

## Shape and other auxiliary parameters

We have seen that the scale parameter can be estimated independently of the regression parameters ( $\beta$ ). Some GLM's (or quasi-GLM's) contain additional parameters that cannot be estimated independently of  $\beta$ .

One example of this is the **shape parameter**  $\alpha$  in the negative binomial GLM. The shape parameter can be estimated by maximum likelihood, together with  $\beta$  (using a profile likelihood technique).

Gamma and beta GLM's also have auxiliary parameters that are estimated in this way.

## Model comparison for GLM's

If  $\phi$  is held fixed across models, then twice the log-likelihood ratio between two nested models  $\hat{\theta}^{(1)}$  and  $\hat{\theta}^{(2)}$  is

$$L \equiv 2 \sum_i (y_i \hat{\theta}_i^{(1)} - \gamma(\hat{\theta}_i^{(1)})) / \phi - 2 \sum_i (y_i \hat{\theta}_i^{(2)} - \gamma(\hat{\theta}_i^{(2)})) / \phi,$$

where  $\hat{\theta}^{(2)}$  is nested within  $\hat{\theta}^{(1)}$ , so  $L \geq 0$ . This is called the **scaled deviance**.

The statistic  $D = \phi L$ , which does not depend explicitly on  $\phi$ , is called the **deviance**.

## Model comparison for GLM's

Suppose that  $\hat{\theta}^{(1)}$  is the saturated model, in which  $\theta_i = Y_i$ . If the GLM is Gaussian and  $g(x) \equiv x$ , as discussed above, the deviance is

$$\begin{aligned} D &= 2 \sum_i (y_i^2 - y_i^2/2) - 2 \sum_i (y_i \hat{\theta}_i^{(2)} - \hat{\theta}_i^{(2)2}/2) \\ &= \sum_i y_i^2 - 2Y_i \hat{\theta}_i^{(2)} + \hat{\theta}_i^{(2)2} \\ &= \sum_i (y_i - \hat{\theta}_i^{(2)})^2. \end{aligned}$$

## Model comparison for GLM's

Thus in the Gaussian case, the deviance is the residual sum of squares for the smaller model ( $\hat{\theta}^{(2)}$ ).

In the Gaussian case,  $D/\phi = L \sim \chi_{n-p-1}^2$ .

When  $\phi$  is unknown, we can turn this around to produce an estimate of the scale parameter

$$\hat{\phi} = \frac{D}{n - p - 1}.$$

This is an unbiased estimate in the Gaussian case, but is useful for any GLM.

## Model comparison for GLM's

Now suppose we want to compare two nested generalized linear models with deviances  $D_1 < D_2$ . Let  $p_1 > p_2$  be the number of covariates in each model. The likelihood ratio test statistic is

$$L_2 - L_1 = \frac{D_2 - D_1}{\phi}$$

which asymptotically has a  $\chi^2_{p_1 - p_2}$  distribution.

If  $\phi$  is unknown, we can estimate it as described above (using the larger of the two models).

The “plug-in” likelihood ratio statistic  $(D_2 - D_1)/\hat{\phi}$  is still asymptotically  $\chi^2_{p_1 - p_2}$ , as long as  $\hat{\phi}$  is consistent.

The finite sample distribution may be better approximated using

$$\frac{D_2 - D_1}{\hat{\phi}(p_1 - p_2)} \approx F_{p_1 - p_2, n - p_1},$$

## Model comparison for GLM's

We can compare any two fitted GLM's using model selection statistics like AIC or BIC.

AIC favors models having small values of  $L_{\text{opt}} - \text{df}$ , where  $L_{\text{opt}}$  is the maximized log-likelihood, and df is the degrees of freedom.  
Equivalently, the AIC can be expressed

$$-D/2\hat{\phi} - p - 1.$$

The same  $\hat{\phi}$  value should be used for all models being compared (i.e. by using the one from the largest model).