Stat 5102 Lecture Slides: Deck 5
Linear Models

Charles J. Geyer
School of Statistics
University of Minnesota

Linear Models

We now return to frequentist statistics for the rest of the course.

The next subject is *linear models*, parts of which are variously called *regression* and *analysis of variance* (ANOVA) and *analysis of covariance* (ANCOVA), with regression being subdivided into *simple linear regression* and *multiple regression*.

Although users have a very fractured view of the subject — many think regression and ANOVA have nothing to do with each other — a unified view is much simpler and more powerful.

Linear Models (cont.)

In linear models we have data on n individuals. For each individual we observe one variable, called the *response*, which is treated as random, and also observe other variables, called *predictors* or *covariates*, which are treated as fixed.

If the predictors are actually random, then we condition on them.

Collect the response variables into a random vector $\mathbf{Y}$ of length n. In linear models we assume the components of $\mathbf{Y}$ are normally distributed and independent and have the same variance σ^2. We do not assume they are identically distributed. Their means can be different.

Linear Models (cont.)

$$E(\mathbf{Y}) = \boldsymbol{\mu} \qquad (*)$$
$$\text{var}(\mathbf{Y}) = \sigma^2\mathbf{I} \qquad (**)$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. Hence

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}) \qquad (***)$$

Recall that we are conditioning on the covariates, hence the expectation $(*)$ is actually a conditional expectation, conditioning on any covariates that are random, although we have not indicated that in the notation. Similarly, the variance in $(**)$ is a conditional variance, and the distribution in $(***)$ is a conditional distribution.

Linear Models (cont.)

One more assumption gives "linear models" its name

$$\boldsymbol{\mu} = \mathbf{M}\boldsymbol{\beta}$$

where $\mathbf{M}$ is a nonrandom matrix, which may depend on the covariates, and $\boldsymbol{\beta}$ is a vector of dimension p of unknown parameters.

The matrix $\mathbf{M}$ is called the *model matrix* or the *design matrix*. We will always use the former, since the latter doesn't make much sense except for a designed experiment.

Each row of $\mathbf{M}$ corresponds to one individual. The i-th row determines the mean for the i-th individual

$$E(Y_i) = m_{i1}\beta_1 + m_{i2}\beta_2 + \cdots + m_{ip}\beta_p$$

and m_{i1}, ..., m_{ip} depend only on the covariate information for this individual.

Linear Models (cont.)

The joint PDF of the data is

$$
\begin{aligned}
f(\mathbf{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{M}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{M}\boldsymbol{\beta})\right)
\end{aligned}
$$

Hence the log likelihood is

$$
l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{M}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{M}\boldsymbol{\beta})
$$

The Method of Least Squares

The maximum likelihood estimate for $\boldsymbol{\beta}$ maximizes the log likelihood, which is the same as minimizing the quadratic function

$$\boldsymbol{\beta} \mapsto (\mathbf{y} - \mathbf{M}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{M}\boldsymbol{\beta})$$

Hence this method of estimation is also called the "method of least squares". Historically, the method of least squares was invented about 1800 and the method of maximum likelihood was invented about 1920. So the older name still attaches to the method.

Linear Models (cont.)

Differentiating the log likelihood with respect to $\boldsymbol{\beta}$ gives

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_k} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} (y_i - \mu_i)^2$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(y_i - \mu_i) \frac{\partial \mu_i}{\partial \beta_k}$$

and since $\partial \mu_i / \partial \beta_k = m_{ik}$, this gives the matrix equation

$$\nabla_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{M}\boldsymbol{\beta})^T \mathbf{M}$$

Setting this equal to zero and multiplying both sides by $1/\sigma^2$ gives us the equations

$$(\mathbf{y} - \mathbf{M}\boldsymbol{\beta})^T \mathbf{M} = 0 \qquad \text{or} \qquad \mathbf{M}^T (\mathbf{y} - \mathbf{M}\boldsymbol{\beta}) = 0$$

to solve to obtain the MLE of $\boldsymbol{\beta}$.

Linear Models (cont.)

$$\mathbf{M}^T(\mathbf{y} - \mathbf{M}\boldsymbol{\beta}) = \mathbf{M}^T\mathbf{y} - \mathbf{M}^T\mathbf{M}\boldsymbol{\beta} = 0$$

is equivalent to

$$\mathbf{M}^T\mathbf{y} = \mathbf{M}^T\mathbf{M}\boldsymbol{\beta}$$

which is sometimes called the "normal equations" (not to be confused with the normal distribution). Their solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{y}$$

assuming the matrix $\mathbf{M}^T\mathbf{M}$ is invertible. If it is not invertible, then the MLE is not unique. More on this later.

Linear Models (cont.)

Recall that only $\mathbf{Y}$ is random. The model matrix is considered fixed. A linear function of a normal random vector is another normal random vector. Hence the MLE for β is a normal random vector with mean vector

$$\begin{aligned} E(\hat{\beta}) &= (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T E(\mathbf{Y}) \\ &= (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{M}\beta \\ &= \beta \end{aligned}$$

and variance matrix

$$\begin{aligned} \text{var}(\hat{\beta}) &= (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T \text{var}(\mathbf{Y})\mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1} \\ &= \sigma^2(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1} \\ &= \sigma^2(\mathbf{M}^T\mathbf{M})^{-1} \end{aligned}$$

Linear Models (cont.)

By invariance of maximum likelihood the MLE for $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}} = \mathrm{M}\hat{\boldsymbol{\beta}}$$

which is also a normal random vector with mean vector

$$E(\hat{\boldsymbol{\mu}}) = \mathrm{M}E(\hat{\boldsymbol{\beta}}) = \mathrm{M}\boldsymbol{\beta} = \boldsymbol{\mu}$$

and variance matrix

$$\begin{aligned} \mathrm{var}(\hat{\boldsymbol{\mu}}) &= \mathrm{M}\,\mathrm{var}(\hat{\boldsymbol{\beta}})\mathrm{M}^T \\ &= \sigma^2\mathrm{M}(\mathrm{M}^T\mathrm{M})^{-1}\mathrm{M}^T \end{aligned}$$

Regression is Projection

Let V denote the vector subspace of $\mathbb{R}^n$ consisting of all possible mean vectors

$$V = \{ \mathbf{M}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p \}$$

then the MLE for $\boldsymbol{\mu}$ solves the constrained optimization problem

$$\begin{aligned} &\text{minimize} \\ &\qquad \|\mathbf{y} - \boldsymbol{\mu}\|^2 \\ &\text{subject to} \\ &\qquad \boldsymbol{\mu} \in V \end{aligned}$$

where

$$\|\mathbf{y} - \boldsymbol{\mu}\|^2 = (\mathbf{y} - \boldsymbol{\mu})^T (\mathbf{y} - \boldsymbol{\mu})$$

is the square of the distance between $\mathbf{y}$ and $\boldsymbol{\mu}$ in n-dimensional space.

Regression is Projection (cont.)

In words, the MLE for $\boldsymbol{\mu}$ is the closest point in the set of all possible mean vectors V to the observed data $\mathbf{y}$. In mathematical terminology, $\hat{\boldsymbol{\mu}}$ is the *projection* of $\mathbf{y}$ on V.

Everything takes place in n-dimensional space, where n is the number of individuals. $\boldsymbol{\mu}$ and $\mathbf{y}$ are points in n-dimensional space, and V is a vector subspace of n-dimensional space.

The MLE of $\boldsymbol{\mu}$ is always unique. There is always a unique closest point to $\mathbf{y}$ in V.

Regression is Projection (cont.)

V is the smallest vector space containing the columns of $\mathbf{M}$, each of which is an n-dimensional vector. If the p columns of $\mathbf{M}$ are linearly independent (meaning none can be written as a linear combination of the others), then $p \leq n$ and V is a p-dimensional vector space and the map

$$\boldsymbol{\beta} \mapsto \mathbf{M}\boldsymbol{\beta}$$

is one-to-one so the linear equation

$$\hat{\boldsymbol{\mu}} = \mathbf{M}\boldsymbol{\beta}$$

has a unique solution for $\boldsymbol{\beta}$, which is the MLE for $\boldsymbol{\beta}$.

Regression is Projection (cont.)

If the p columns of $\mathbf{M}$ are not linearly independent (meaning some of them can be written as a linear combinations of the others), then V is a q-dimensional vector space, where q is the largest number of linearly independent vectors among the columns of $\mathbf{M}$. Then the map

$$\boldsymbol{\beta} \mapsto \mathbf{M}\boldsymbol{\beta}$$

is many-to-one so the linear equation

$$\hat{\boldsymbol{\mu}} = \mathbf{M}\boldsymbol{\beta}$$

has many solutions for $\boldsymbol{\beta}$, any of which is a (non-unique) MLE for $\boldsymbol{\beta}$.

Regression is Projection (cont.)

The *rank* of a matrix $\mathbf{M}$ is the largest number of linearly independent columns it has.

The rank of the model matrix $\mathbf{M}$ is the dimension q of the subspace V of all possible mean vectors.

When $q = p$ (the rank equals the column dimension), we say the model matrix is full rank.

When $q < p$ (the model matrix is not full rank), we can find a matrix $\mathbf{M}_2$ whose columns are a subset of the columns of $\mathbf{M}$ and whose rank is q.

Regression is Projection (cont.)

Then

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{M}_2^T \mathbf{M}_2)^{-1} \mathbf{M}_2^T \mathbf{y}$$

is the unique MLE for the $\boldsymbol{\beta}$ for this new problem with model matrix $\mathbf{M}_2$ and

$$\hat{\boldsymbol{\mu}}_2 = \mathbf{M}_2 (\mathbf{M}_2^T \mathbf{M}_2)^{-1} \mathbf{M}_2^T \mathbf{y}$$

is the unique MLE for $\boldsymbol{\mu}$.

Since, by construction

$$V = \{ \mathbf{M}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p \} = \{ \mathbf{M}_2 \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^q \}$$

the "regression as projection" problem is the same in both cases and $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_2$.

Regression is Projection (cont.)

Thus we have figured out how to deal with the case where the MLE for $\boldsymbol{\beta}$ is not unique.

Since every column of $\mathbf{M}_2$ is also a column of $\mathbf{M}$, $\hat{\boldsymbol{\beta}}_2$ can be thought of as the solution for the original problem subject to the constraint that $\beta_j = 0$ for all j such that the the j-th column of $\mathbf{M}$ is not a column of $\mathbf{M}_2$.

Thus we have also found a (non-unique) MLE for $\boldsymbol{\beta}$ for the original problem

$$\widehat{\beta}_j = \widehat{\beta}_{2,k}$$

when the j-th column of $\mathbf{M}$ is the k-th column of $\mathbf{M}_2$ and

$$\widehat{\beta}_j = 0$$

when the j-th column of $\mathbf{M}$ not a column of $\mathbf{M}_2$.

Regression Coefficients are Meaningless

We have seen that the MLE for $\boldsymbol{\beta}$ is not always uniquely defined but this is not a problem.

Let $\mathbf{M}_3$ be any $n \times r$ matrix such that

$$V = \{\, \mathbf{M}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p \,\} = \{\, \mathbf{M}_3\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^r \,\}$$

Since the "regression as projection" problem is the same in both cases, so is the MLE for $\boldsymbol{\mu}$. But the MLE for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_3$ seem to have no relation to each other. None of the components need be the same.

Regression Coefficients are Meaningless (cont.)

If $\mathbf{M}$ and $\mathbf{M}_3$ are both full rank, then there is a relationship between them: $\mathbf{M} = \mathbf{M}_3\mathbf{A}$ for some invertible matrix $\mathbf{A}$ and

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_3 &= (\mathbf{M}_3^T\mathbf{M}_3)^{-1}\mathbf{M}_3^T\mathbf{y} \\
\hat{\boldsymbol{\beta}} &= (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{y} \\
&= (\mathbf{A}^T\mathbf{M}_3^T\mathbf{M}_3\mathbf{A})^{-1}\mathbf{A}^T\mathbf{M}_3^T\mathbf{y} \\
&= \mathbf{A}^{-1}(\mathbf{M}_3^T\mathbf{M}_3)^{-1}(\mathbf{A}^T)^{-1}\mathbf{A}^T\mathbf{M}_3^T\mathbf{y} \\
&= \mathbf{A}^{-1}(\mathbf{M}_3^T\mathbf{M}_3)^{-1}\mathbf{M}_3^T\mathbf{y} \\
&= \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}_3
\end{aligned}$$

so there is a relationship between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_3$ but a highly non-obvious one, since we usually don't know $\mathbf{A}$ explicitly.

Regression Coefficients are Meaningless (cont.)

We need the regression coefficient vector β because we don't have an explicit representation of the subspace V of all possible mean vectors. We have to go through $\hat{\beta}$ to get to $\hat{\mu}$.

But that doesn't make $\hat{\beta}$ meaningful. The MLE $\hat{\mu}$ of the mean vector is always meaningful and interpretable. The MLE $\hat{\beta}$ of the regression coefficient vector is rarely meaningful or interpretable.

Despite this, many regression textbooks spend a large amount of time teaching how to "interpret" these meaningless quantities. This leads to many confusions on the part of the poor students so taught!

Regression Coefficients are Meaningless (cont.)

Worse, these regression textbooks do not even call $\boldsymbol{\mu}$ a parameter vector or $\hat{\boldsymbol{\mu}}$ a parameter estimate.

They write $\widehat{\mathbf{y}}$ instead of $\hat{\boldsymbol{\mu}}$ and call $\widehat{\mathbf{y}}$ the vector of *predicted values*. This makes a crazy kind of sense. If you want to predict Y_i, then the best prediction is μ_i, where "best" means minimizing expected squared prediction error (Deck 1, Slides 6–9). And your best estimate of μ_i is $\widehat{\mu}_i$, where "best" means achieving the Cramér-Rao lower bound for asymptotic variance (Deck 3, Slides 67–72).

That greatly complicates a simple situation. $\hat{\boldsymbol{\mu}}$ is a point estimate of $\boldsymbol{\mu}$, just like any other parameter estimate. $\widehat{\mathbf{y}}$ is a best prediction, where "best" involves two different bests with two different meanings.

Regression Coefficients are Meaningless (cont.)

When $\boldsymbol{\mu}$ is not called a parameter and $\hat{\boldsymbol{\mu}}$ is not called a parameter estimate, students have only one parameter vector $\boldsymbol{\beta}$ to think about. Unfortunately, they think about the meaningless one rather than the meaningful one.

The notation $\hat{\mathbf{y}}$ does not fit with the rest of mathematical statistics. Nowhere else do we put a hat on a random variable. Nowhere else do we call a parameter estimate anything other than a parameter estimate.

No wonder students taught this way are confused!

Regression is Projection (cont.)

The matrix that "puts the hat on $\mathbf{y}$" is called the *hat matrix*

$$\hat{\boldsymbol{\mu}} = \mathbf{M}\hat{\boldsymbol{\beta}} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{y}$$

so

$$\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$$

where

$$\mathbf{H} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$$

Regression is Projection (cont.)

A matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}^T$. A square matrix $\mathbf{A}$ is *idempotent* if $\mathbf{A} = \mathbf{A}^2$.

Regression is Projection (cont.)

A hat matrix is symmetric and idempotent. To check symmetry recall

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$$

and so forth. Hence

$$(\mathbf{M}^T\mathbf{M})^T = \mathbf{M}^T(\mathbf{M}^T)^T = \mathbf{M}^T\mathbf{M}$$

is symmetric. The inverse of a symmetric matrix is symmetric by Cramer's rule for construction of inverses. Hence $(\mathbf{M}^T\mathbf{M})^{-1}$ is symmetric. Hence

$$\left(\mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\right)^T = (\mathbf{M}^T)^T(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$$

is symmetric.

Regression is Projection (cont.)

The check for idempotence is straightforward

$$\mathbf{H}^2 = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T = \mathbf{H}$$

Regression is Projection (cont.)

A square matrix $\mathbf{H}$ is a *projection matrix* if $\mathbf{Hy}$ is the point in the vector subspace

$$V = \{ \mathbf{H}\boldsymbol{\eta} : \boldsymbol{\eta} \in \mathbb{R}^n \}$$

that is closest to $\mathbf{y}$.

Theorem. A square matrix $\mathbf{H}$ is a projection matrix if and only if it is symmetric and idempotent.

Proof. We already know one direction. Least squares does projection and its hat matrix is symmetric and idempotent.

Regression is Projection (cont.)

For the other direction assume $\mathbf{H}$ is symmetric and idempotent.

The matrix $\mathbf{I} - \mathbf{H}$ is also square, symmetric, and idempotent, because

$$(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - 2\mathbf{H} + \mathbf{H}^2 = \mathbf{I} - \mathbf{H}$$

Vectors $\mathbf{u}$ and $\mathbf{v}$ are *perpendicular* or *orthogonal* if $\mathbf{u}^T\mathbf{v} = 0$. $\mathbf{Hy}$ and $(\mathbf{I} - \mathbf{H})\mathbf{y}$ are perpendicular because

$$\mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{H}\mathbf{y} = 0$$

because

$$(\mathbf{I} - \mathbf{H})\mathbf{H} = \mathbf{H} - \mathbf{H}^2 = 0$$

Regression is Projection (cont.)

Let $\hat{\mu} = \mathbf{H}\mathbf{y}$ and let $\mu = \mathbf{H}\eta$ be any other point in V. Then

$$\begin{aligned}
\|\mathbf{y} - \mu\|^2 &= (\mathbf{y} - \mu)^T(\mathbf{y} - \mu) \\
&= (\mathbf{y} - \hat{\mu} + \hat{\mu} - \mu)^T(\mathbf{y} - \hat{\mu} + \hat{\mu} - \mu) \\
&= (\mathbf{y} - \hat{\mu})^T(\mathbf{y} - \hat{\mu}) + 2(\mathbf{y} - \hat{\mu})^T(\hat{\mu} - \mu) \\
&\quad + (\hat{\mu} - \mu)^T(\hat{\mu} - \mu) \\
&= \|\mathbf{y} - \hat{\mu}\|^2 + \|\hat{\mu} - \mu\|^2 + 2(\mathbf{y} - \hat{\mu})^T(\hat{\mu} - \mu) \\
&= \|\mathbf{y} - \hat{\mu}\|^2 + \|\hat{\mu} - \mu\|^2 + 2(\mathbf{y} - \mathbf{H}\mathbf{y})^T(\mathbf{H}\mathbf{y} - \mathbf{H}\eta) \\
&= \|\mathbf{y} - \hat{\mu}\|^2 + \|\hat{\mu} - \mu\|^2 + 2\mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{H}(\mathbf{y} - \eta) \\
&= \|\mathbf{y} - \hat{\mu}\|^2 + \|\hat{\mu} - \mu\|^2
\end{aligned}$$

because $(\mathbf{I} - \mathbf{H})\mathbf{H} = 0$. And that proves the theorem.

Regression is Projection (cont.)

Theorem. Suppose the assumptions for linear models. Then $\mathbf{Y} - \hat{\boldsymbol{\mu}}$ is independent of $\hat{\boldsymbol{\mu}}$ and

$$\begin{aligned} \hat{\boldsymbol{\mu}} &\sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{H}) \\ \mathbf{Y} - \hat{\boldsymbol{\mu}} &\sim \mathcal{N}\Big(0, \sigma^2(\mathbf{I} - \mathbf{H})\Big) \\ \frac{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2}{\sigma^2} &\sim \text{chi}^2(n - q) \end{aligned}$$

where q is the rank of $\mathbf{H}$.

This is a generalization of the theorem about sampling distributions for normal populations (Deck 1, Slide 58–77).

Regression is Projection (cont.)

A linear function of a normal random vector is a normal random vector. Hence

$$\begin{pmatrix} \mathbf{Y} - \hat{\boldsymbol{\mu}} \\ \hat{\boldsymbol{\mu}} \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \mathbf{H} \\ \mathbf{H} \end{pmatrix} \mathbf{Y}$$

is a normal random vector. Uncorrelated implies independent for jointly normal random vectors. Hence independence follows from

$$\begin{aligned} \operatorname{cov}(\mathbf{Y} - \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}) &= \operatorname{cov}\Big((\mathbf{I} - \mathbf{H})\mathbf{Y}, \mathbf{H}\mathbf{Y}\Big) \\ &= E\{(\mathbf{I} - \mathbf{H})(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{H}\} \\ &= \sigma^2 (\mathbf{I} - \mathbf{H})\mathbf{H} \end{aligned}$$

being zero, and this follows from $(\mathbf{I} - \mathbf{H})\mathbf{H} = 0$.

Regression is Projection (cont.)

We already saw that $\hat{\boldsymbol{\mu}}$ is a normal random vector.

$$
\begin{aligned}
E(\hat{\boldsymbol{\mu}}) &= E(\mathbf{HY}) \\
&= \mathbf{H}E(\mathbf{Y}) \\
&= \mathbf{H}\boldsymbol{\mu} \\
&= \boldsymbol{\mu} \\
\text{var}(\hat{\boldsymbol{\mu}}) &= \text{var}(\mathbf{HY}) \\
&= \mathbf{H}\,\text{var}(\mathbf{Y})\mathbf{H} \\
&= \sigma^2\mathbf{H}^2 \\
&= \sigma^2\mathbf{H}
\end{aligned}
$$

Regression is Projection (cont.)

Also $\mathbf{Y} - \hat{\boldsymbol{\mu}}$ is a normal random vector.

$$\begin{aligned}
E(\mathbf{Y} - \hat{\boldsymbol{\mu}}) &= E\{(\mathbf{I} - \mathbf{H})\mathbf{Y}\} \\
&= (\mathbf{I} - \mathbf{H})E(\mathbf{Y}) \\
&= (\mathbf{I} - \mathbf{H})\boldsymbol{\mu} \\
&= 0 \\
\text{var}(\mathbf{Y} - \hat{\boldsymbol{\mu}}) &= \text{var}\{(\mathbf{I} - \mathbf{H})\mathbf{Y}\} \\
&= (\mathbf{I} - \mathbf{H})\,\text{var}(\mathbf{Y})(\mathbf{I} - \mathbf{H}) \\
&= \sigma^2(\mathbf{I} - \mathbf{H})^2 \\
&= \sigma^2(\mathbf{I} - \mathbf{H})
\end{aligned}$$

Regression is Projection (cont.)

Consider the spectral decomposition (5101, Deck 5, Slides 103–110)

$$\mathbf{I} - \mathbf{H} = \mathbf{O}\mathbf{D}\mathbf{O}^T$$

where $\mathbf{O}$ is orthogonal and $\mathbf{D}$ is diagonal. Then

$$(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - \mathbf{H}$$

means

$$\mathbf{O}\mathbf{D}^2\mathbf{O}^T = \mathbf{O}\mathbf{D}\mathbf{O}^T$$

hence

$$\mathbf{D}^2 = \mathbf{D}$$

hence every diagonal element of $\mathbf{D}$ is its own square, hence either zero or one.

Regression is Projection (cont.)

Hence

$$\mathbf{O}\mathbf{D}\mathbf{O}^T = \mathbf{O}_1\mathbf{O}_1^T$$

where $\mathbf{O}_1$ is the matrix whose columns are the columns of $\mathbf{O}$ corresponding to diagonal elements of $\mathbf{D}$ that are equal to one.

The columns of $\mathbf{O}_1$ are an orthogonal basis for the vector subspace on which $\mathbf{I} - \mathbf{H}$ projects

$$V^{\perp} = \{\, (\mathbf{I} - \mathbf{H})\boldsymbol{\eta} : \boldsymbol{\eta} \in \mathbb{R}^n \,\}$$

Regression is Projection (cont.)

Let k be the column dimension of $\mathbf{O}_1$, and let $\mathbf{Z}$ be a standard normal random vector of dimension k. Then $\mathbf{O}_1\mathbf{Z}$ is a normal random vector having mean vector zero and variance matrix

$$\text{var}(\mathbf{O}_1\mathbf{Z}) = \mathbf{O}_1 \, \text{var}(\mathbf{Z}) \mathbf{O}_1^T = \mathbf{O}_1 \mathbf{O}_1^T = \mathbf{I} - \mathbf{H}$$

Hence $\mathbf{O}_1\mathbf{Z}$ has the same distribution as $(\mathbf{Y} - \hat{\boldsymbol{\mu}})/\sigma$. Since

$$\|\mathbf{O}_1\mathbf{Z}\|^2 = \mathbf{Z}^T \mathbf{O}_1^T \mathbf{O}_1 \mathbf{Z} = \mathbf{Z}^T \mathbf{Z} = \|\mathbf{Z}\|^2$$

and $\|\mathbf{Z}\|^2$ has the chi$^2(k)$ distribution, that proves all of the theorem except for $k = n - q$.

Regression is Projection (cont.)

Let $\mathbf{O}_2$ be the matrix whose columns are the columns of $\mathbf{O}$ corresponding to diagonal elements of $\mathbf{I}-\mathbf{D}$ that are equal to one and to diagonal elements of $\mathbf{D}$ that are equal to zero. Since

$$\mathbf{O}(\mathbf{I}-\mathbf{D})\mathbf{O}^T = \mathbf{I}-\mathbf{O}\mathbf{D}\mathbf{O}^T = \mathbf{I}-(\mathbf{I}-\mathbf{H}) = \mathbf{H}$$

and since the diagonal elements of $\mathbf{I}-\mathbf{D}$ are one when the diagonal elements of $\mathbf{D}$ are zero and vice versa, it follows that the columns of $\mathbf{O}_2$ are an orthogonal basis for V.

Since $\mathbf{O}_1$ has k columns and $\mathbf{O}_2$ has $n-k$ columns, V has dimension $n-k=q$. That finishes the proof of the theorem.

Linear Models (cont.)

With the theorem, we can now finish parameter estimation in linear models. As is usual when dealing with normal sampling distributions, we do not use the MLE for σ^2 but rather the unbiased estimator

$$\widehat{\sigma}^2 = \frac{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2}{n - q}$$

where q is the rank of the hat matrix and the model matrix (the MLE is the same except for dividing by n rather than $n - q$).

That this estimator is unbiased follows from the theorem and the expectation of the chi-square distribution.

Linear Models (cont.)

Let

$$\mathbf{W} = (\mathbf{M}^T\mathbf{M})^{-1}$$

then

$$\text{var}(\widehat{\beta}_i) = \sigma^2 w_{ii}$$

where w_{ij} denotes components of $\mathbf{W}$. Hence

$$\text{se}(\widehat{\beta}_i) = \widehat{\sigma}\sqrt{w_{ii}}$$

estimates the standard deviation of $\widehat{\beta}_i$.

Linear Models (cont.)

Moreover,

$$\frac{\widehat{\beta}_i - \beta_i}{\mathsf{se}(\widehat{\beta}_i)} = \frac{\dfrac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{w_{ii}}}}{\sqrt{\dfrac{\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2}{(n-q)\sigma^2}}} \sim t(n-q)$$

is an exact pivotal quantity, which can be used to make exact confidence intervals or hypothesis tests about β_i.

If $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the $t(n-q)$ distribution, then

$$\widehat{\beta}_i \pm t_{\alpha/2}\, \mathsf{se}(\widehat{\beta}_i)$$

is an exact $100(1-\alpha)\%$ confidence interval for the true unknown parameter β_i.

Linear Models (cont.)

If

$$t = \frac{\widehat{\beta}_i - \beta_i^*}{\mathsf{se}(\widehat{\beta}_i)}$$

and T is a $t(n-q)$ random variable, then $\mathsf{Pr}(T < t)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \beta_i \geq \beta_i^*$$
$$H_1 \colon \beta_i < \beta_i^*$$

$\mathsf{Pr}(T > t)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \beta_i \leq \beta_i^*$$
$$H_1 \colon \beta_i > \beta_i^*$$

Linear Models (cont.)

And $2\Pr(T > |t|)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \beta_i = \beta_i^*$$
$$H_1 \colon \beta_i \neq \beta_i^*$$

Linear Models (cont.)

We now repeat the last four slides changing the parameter from β_i to μ_i.

As before, define the hat matrix (slide 24)

$$\mathbf{H} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$$

then

$$\text{var}(\widehat{\mu}_i) = \sigma^2 h_{ii}$$

where h_{ij} denotes components of $\mathbf{H}$. Hence

$$\text{se}(\widehat{\mu}_i) = \widehat{\sigma}\sqrt{h_{ii}}$$

estimates the standard deviation of $\widehat{\mu}_i$.

Linear Models (cont.)

Moreover,

$$\frac{\widehat{\mu}_i - \mu_i}{\mathsf{se}(\widehat{\mu}_i)} \sim t(n-q)$$

is an exact pivotal quantity, which can be used to make exact confidence intervals or hypothesis tests about μ_i.

If $t_{\alpha/2}$ is the $1-\alpha/2$ quantile of the $t(n-q)$ distribution, then

$$\widehat{\mu}_i \pm t_{\alpha/2}\,\mathsf{se}(\widehat{\mu}_i)$$

is an exact $100(1-\alpha)\%$ confidence interval for the true unknown parameter μ_i.

Linear Models (cont.)

If

$$t = \frac{\widehat{\mu}_i - \mu_i^*}{\mathsf{se}(\widehat{\mu}_i)}$$

and T is a $t(n-q)$ random variable, then $\Pr(T < t)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \mu_i \geq \mu_i^*$$
$$H_1 \colon \mu_i < \mu_i^*$$

$\Pr(T > t)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \mu_i \leq \mu_i^*$$
$$H_1 \colon \mu_i > \mu_i^*$$

Linear Models (cont.)

And $2\Pr(T > |t|)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \mu_i = \mu_i^*$$
$$H_1 \colon \mu_i \neq \mu_i^*$$

Confidence Intervals for New Data

Suppose that, instead of a confidence interval for the mean for one of the individuals in our data, we want confidence intervals for some new individuals, whose covariate data would form a new model matrix $\mathbf{M}_{\text{new}}$. Then the vector of mean values for these new individuals is $\boldsymbol{\mu}_{\text{new}} = \mathbf{M}_{\text{new}}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the true unknown parameter vector.

Confidence Intervals for New Data (cont.)

The MLE for $\boldsymbol{\mu}_{\text{new}}$ is by invariance of maximum likelihood

$$\hat{\boldsymbol{\mu}}_{\text{new}} = \mathbf{M}_{\text{new}}\hat{\boldsymbol{\beta}}$$

and this estimator is normal with mean and variance

$$\begin{aligned}
E(\hat{\boldsymbol{\mu}}_{\text{new}}) &= \mathbf{M}_{\text{new}} E(\hat{\boldsymbol{\beta}}) \\
&= \mathbf{M}_{\text{new}}\boldsymbol{\beta} \\
&= \boldsymbol{\mu}_{\text{new}} \\
\text{var}(\hat{\boldsymbol{\mu}}_{\text{new}}) &= \mathbf{M}_{\text{new}}\, \text{var}(\hat{\boldsymbol{\beta}})\mathbf{M}_{\text{new}}^T \\
&= \sigma^2 \mathbf{M}_{\text{new}}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}_{\text{new}}^T
\end{aligned}$$

Confidence Intervals for New Data (cont.)

We repeated the four slides 40–43 changing the parameter from β_i to μ_i to get the four slides 44–47. We now repeat them again changing μ_i to $\mu_{\text{new},i}$.

Define the matrix

$$\mathbf{H}_{\text{new}} = \mathbf{M}_{\text{new}}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}_{\text{new}}^T$$

then

$$\text{var}(\widehat{\mu}_{\text{new},i}) = \sigma^2 h_{\text{new},ii}$$

where $h_{\text{new},ij}$ denotes components of $\mathbf{H}_{\text{new}}$. Hence

$$\text{se}(\widehat{\mu}_{\text{new},i}) = \widehat{\sigma}\sqrt{h_{\text{new},ii}}$$

estimates the standard deviation of $\widehat{\mu}_{\text{new},i}$.

Confidence Intervals for New Data (cont.)

Moreover,

$$\frac{\widehat{\mu}_{\text{new},i} - \mu_{\text{new},i}}{\text{se}(\widehat{\mu}_{\text{new},i})} \sim t(n-q)$$

is an exact pivotal quantity, which can be used to make exact confidence intervals or hypothesis tests about $\mu_{\text{new},i}$.

If $t_{\alpha/2}$ is the $1-\alpha/2$ quantile of the $t(n-q)$ distribution, then

$$\widehat{\mu}_{\text{new},i} \pm t_{\alpha/2}\, \text{se}(\widehat{\mu}_{\text{new},i})$$

is an exact $100(1-\alpha)\%$ confidence interval for the true unknown parameter $\mu_{\text{new},i}$.

Hypothesis Tests for New Data

If

$$t = \frac{\widehat{\mu}_{\text{new},i} - \mu^*_{\text{new},i}}{\text{se}(\widehat{\mu}_{\text{new},i})}$$

and T is a $t(n-q)$ random variable, then $\Pr(T < t)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \mu_{\text{new},i} \geq \mu^*_{\text{new},i}$$
$$H_1 \colon \mu_{\text{new},i} < \mu^*_{\text{new},i}$$

$\Pr(T > t)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \mu_{\text{new},i} \leq \mu^*_{\text{new},i}$$
$$H_1 \colon \mu_{\text{new},i} > \mu^*_{\text{new},i}$$

Hypothesis Tests for New Data (cont.)

And $2\Pr(T > |t|)$ is an exact P-value for the test with hypotheses

$$H_0 \colon \mu_{\text{new},i} = \mu^*_{\text{new},i}$$
$$H_1 \colon \mu_{\text{new},i} \neq \mu^*_{\text{new},i}$$

Prediction Intervals for New Data

Suppose that, instead of estimating the vector $\boldsymbol{\mu}_{\text{new}}$ of mean values of the response for new individuals, we wish to predict the new data $\mathbf{Y}_{\text{new}}$ for these individuals.

Since

$$\mathbf{Y}_{\text{new}} - \boldsymbol{\mu}_{\text{new}} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$
$$\hat{\boldsymbol{\mu}}_{\text{new}} - \boldsymbol{\mu}_{\text{new}} \sim \mathcal{N}(0, \sigma^2\mathbf{H}_{\text{new}})$$

and these random vectors are independent (because $\mathbf{Y}_{\text{new}}$ is independent of the original data used to estimate $\hat{\boldsymbol{\mu}}_{\text{new}}$)

$$\mathbf{Y}_{\text{new}} - \hat{\boldsymbol{\mu}}_{\text{new}} \sim \mathcal{N}\Big(0, \sigma^2(\mathbf{I} + \mathbf{H}_{\text{new}})\Big)$$

Prediction Intervals for New Data (cont.)

We repeated the four slides 40–43 changing the parameter from β_i to μ_i to get the four slides 44–47. We repeated them again changing μ_i to $\mu_{\text{new},i}$ to get the four slides 50–53. Now we partially repeat them again changing confidence intervals for $\mu_{\text{new},i}$ to prediction intervals for $Y_{\text{new},i}$.

Since

$$\text{var}(Y_{\text{new},i} - \widehat{\mu}_{\text{new},i}) = \sigma^2(1 + h_{\text{new},ii})$$

it follows that

$$\sqrt{\widehat{\sigma}^2 + \text{se}(\widehat{\mu}_{\text{new},i})^2} = \widehat{\sigma}\sqrt{1 + h_{\text{new},ii}}$$

estimates the standard deviation of $Y_{\text{new},i} - \widehat{\mu}_{\text{new},i}$.

Prediction Intervals for New Data (cont.)

Moreover,

$$\frac{Y_{\text{new},i} - \widehat{\mu}_{\text{new},i}}{\sqrt{\widehat{\sigma}^2 + \text{se}(\widehat{\mu}_{\text{new},i})^2}} \sim t(n-q)$$

is an exact pivotal quantity, which can be used to make exact prediction intervals for $Y_{\text{new},i}$.

If $t_{\alpha/2}$ is the $1-\alpha/2$ quantile of the $t(n-q)$ distribution, then

$$\widehat{\mu}_{\text{new},i} \pm t_{\alpha/2}\sqrt{\widehat{\sigma}^2 + \text{se}(\widehat{\mu}_{\text{new},i})^2}$$

is an exact $100(1-\alpha)\%$ prediction interval for new data $Y_{\text{new},i}$.

Linear Models (cont.)

We have seen that confidence intervals and hypothesis tests for β_i, for μ_i, for $\mu_{\text{new},i}$, and prediction intervals for $Y_{\text{new},i}$ are very similar.

Here's something different.

Nested Models

Suppose we have two linear models with model matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ both having the same row dimension n but different column dimensions p_1 and p_2. The corresponding hat matrices

$$\mathbf{H}_i = \mathbf{M}_i(\mathbf{M}_i^T\mathbf{M}_i)^{-1}\mathbf{M}_i^T$$

are both $n \times n$. The corresponding spaces of mean vectors are

$$V_i = \{\, \mathbf{M}_i\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^{p_i} \,\}$$

are both vector subspaces of $\mathbb{R}^n$.

We say the model 1 is *nested within* model 2 if $V_1 \subset V_2$.

Nested Models (cont.)

This technical definition of nesting is, in general, hard to verify. So we usually use a simpler notion. Model 1 is nested within model 2 if all of the columns of $\mathbf{M}_1$ are columns of $\mathbf{M}_2$ or if all of the terms in the R formula specifying model 1 are also present in the R formula specifying model 2.

Either of these implies the technical condition $V_1 \subset V_2$, but the technical condition is more general.

The model with formula `y ~ x1 + x2` is nested within the model with formula `y ~ x1 + x2 + x3 + x4 + x5`.

The model with formula `y ~ poly(x1, x2, degree = 2)` is nested within the model with formula `y ~ poly(x1, x2, degree = 3)`.

Nested Models (cont.)

Lemma. If model 1 is nested within model 2, then

$$\mathbf{H}_1\mathbf{H}_2 = \mathbf{H}_2\mathbf{H}_1 = \mathbf{H}_1$$
$$(\mathbf{I} - \mathbf{H}_1)(\mathbf{I} - \mathbf{H}_2) = (\mathbf{I} - \mathbf{H}_2)(\mathbf{I} - \mathbf{H}_1) = (\mathbf{I} - \mathbf{H}_2)$$

Proof. We only need to prove

$$\mathbf{H}_2\mathbf{H}_1 = \mathbf{H}_1 \qquad (*)$$
$$(\mathbf{I} - \mathbf{H}_1)(\mathbf{I} - \mathbf{H}_2) = (\mathbf{I} - \mathbf{H}_2) \qquad (**)$$

since these imply the other two (as the next slide shows).

Nested Models (cont.)

One of these implications

$$\begin{aligned}(\mathbf{I}-\mathbf{H}_2)(\mathbf{I}-\mathbf{H}_1) &= \mathbf{I}-\mathbf{H}_2-\mathbf{H}_1+\mathbf{H}_2\mathbf{H}_1 \\ &= \mathbf{I}-\mathbf{H}_2\end{aligned}$$

uses $(*)$, and the other

$$\begin{aligned}\mathbf{H}_1\mathbf{H}_2 &= [\mathbf{I}-(\mathbf{I}-\mathbf{H}_1)][\mathbf{I}-(\mathbf{I}-\mathbf{H}_2)] \\ &= \mathbf{I}-(\mathbf{I}-\mathbf{H}_1)-(\mathbf{I}-\mathbf{H}_2)+(\mathbf{I}-\mathbf{H}_1)(\mathbf{I}-\mathbf{H}_2) \\ &= \mathbf{I}-(\mathbf{I}-\mathbf{H}_1) \\ &= \mathbf{H}_1\end{aligned}$$

uses $(**)$.

Nested Models (cont.)

For any $\mathbf{y}$ the projection $\mathbf{H}_1\mathbf{y}$ is in V_1 hence in V_2. The projection $\mathbf{H}_2\mathbf{H}_1\mathbf{y}$ of this point on V_2 does nothing (the nearest point in V_2 to a point already in V_2 is that point itself). Hence $\mathbf{H}_2\mathbf{H}_1\mathbf{y} = \mathbf{H}_1\mathbf{y}$ for all $\mathbf{y}$, which is $(*)$.

We already know (slide 38) that the dimensions of V_i and $V_i^\perp$ add up to n, hence an orthogonal basis for V_i combined with an orthogonal basis for $V_i^\perp$ makes an orthogonal basis for $\mathbb{R}^n$. Thus $V_i^\perp$ consists of all vectors perpendicular to all elements of V_i. From this we see that $V_1 \subset V_2$ implies $V_2^\perp \subset V_1^\perp$.

Hence the proof of $(**)$ is exactly like the proof of $(*)$ above, except with $\mathbf{I} - \mathbf{H}_i$ replacing $\mathbf{H}_i$ and $\mathbf{V}_i^\perp$ replacing $\mathbf{V}_i$. That finishes the proof of the lemma.

Hypothesis Tests of Model Comparison

Theorem. Suppose the nested models setup on the preceding five slides, and suppose $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, where $\boldsymbol{\mu} = \mathbf{M}_1\boldsymbol{\beta}_1$. Then $(\mathbf{I}-\mathbf{H}_2)\mathbf{Y}$ and $(\mathbf{H}_2-\mathbf{H}_1)\mathbf{Y}$ are independent random vectors, and

$$\frac{\|(\mathbf{I}-\mathbf{H}_2)\mathbf{Y}\|^2}{\sigma^2} \sim \text{chi}^2(n-q_2)$$

$$\frac{\|(\mathbf{H}_2-\mathbf{H}_1)\mathbf{Y}\|^2}{\sigma^2} \sim \text{chi}^2(q_2-q_1)$$

$$\frac{\|(\mathbf{H}_2-\mathbf{H}_1)\mathbf{Y}\|^2/(q_2-q_1)}{\|(\mathbf{I}-\mathbf{H}_2)\mathbf{Y}\|^2/(n-q_2)} \sim F(q_2-q_1, n-q_2)$$

where q_i is the dimension of V_i and the rank of $\mathbf{M}_i$.

Hypothesis Tests of Model Comparison (cont.)

The first assertion follows from the theorem on slide 31. The last assertion follows from the preceding ones and the definition of an F random variable as the ratio of independent chi-squares each divided by its degrees of freedom.

To prove the independence assertion, we show the random vectors in question are uncorrelated

$$\begin{aligned}
&\text{cov}\{(\mathbf{I}-\mathbf{H}_2)\mathbf{Y}, (\mathbf{H}_2-\mathbf{H}_1)\mathbf{Y}\} \\
&\qquad = E\{(\mathbf{I}-\mathbf{H}_2)(\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}^T-\boldsymbol{\mu})^T(\mathbf{H}_2-\mathbf{H}_1)\} \\
&\qquad\qquad = \sigma^2(\mathbf{I}-\mathbf{H}_2)(\mathbf{H}_2-\mathbf{H}_1) \\
&\qquad\qquad\qquad = \sigma^2(\mathbf{H}_2-\mathbf{H}_1-\mathbf{H}_2^2+\mathbf{H}_2\mathbf{H}_1)
\end{aligned}$$

is zero because $\mathbf{H}_2^2 = \mathbf{H}_2$ and $\mathbf{H}_2\mathbf{H}_1 = \mathbf{H}_1$.

Hypothesis Tests of Model Comparison (cont.)

The argument on slides 35–38 generalizes to prove the following.

Lemma. Suppose $\mathbf{W} \sim \mathcal{N}(0, \mathbf{H})$, where $\mathbf{H}$ is a projection matrix. Then $\|\mathbf{W}\|^2$ has a chi-square distribution with degrees of freedom equal to the rank of $\mathbf{H}$.

It only remains to be shown that $\mathbf{H}_2 - \mathbf{H}_1$ is symmetric and idempotent and has rank $q_2 - q_1$. Symmetry is obvious. Idempotence is

$$\begin{aligned} (\mathbf{H}_2 - \mathbf{H}_1)^2 &= \mathbf{H}_2^2 - \mathbf{H}_1\mathbf{H}_2 - \mathbf{H}_2\mathbf{H}_1 + \mathbf{H}_1^2 \\ &= \mathbf{H}_2 - \mathbf{H}_1 - \mathbf{H}_1 + \mathbf{H}_1 \\ &= \mathbf{H}_2 - \mathbf{H}_1 \end{aligned}$$

Hypothesis Tests of Model Comparison (cont.)

Thus $\mathbf{H}_2 - \mathbf{H}_1$ is a projection matrix. Since

$$\mathbf{H}_2 - \mathbf{H}_1 = \mathbf{H}_2(\mathbf{I} - \mathbf{H}_1)$$

this matrix projects on the vector subspace

$$V_2 \cap V_1^{\perp}$$

where $\mathbf{H}_2$ projects on V_2 and $\mathbf{H}_1$ projects on V_1. Hence the rank of $\mathbf{H}_2 - \mathbf{H}_1$ is the dimension of this subspace.

Consider an orthogonal basis for V_1, which has q_1 vectors. Extend this basis by adding $q_2 - q_1$ vectors to make an orthogonal basis for V_2. These additional $q_2 - q_1$ vectors are an orthogonal basis for $V_2 \cap V_1^{\perp}$. That finishes the proof of the theorem.

Categorical Covariates and Dummy Variables

Suppose, as in the computer example, we have one quantitative covariate `x` and one categorical covariate `color` which takes values `"red"`, `"green"`, and `"blue"`.

How do we deal with that? It depends on what model we want to fit. If we write down the equation giving means as a function of regression coefficients, then that implicitly defines the model matrix.

Suppose we want to fit parallel regression lines. The regression lines for different colors have the same slope but different intercepts

$$\mu_i = \beta_{\text{color}_i} + \gamma x_i$$

Categorical Covariates and Dummy Variables (cont.)

There are four regression coefficients, three betas, one for each color, and γ. In order to see the model matrix clearly we need to re-express the formula specifying the model so that all of the betas occur as coefficients

$$\mu_i = \sum_{c \in \text{colors}} d_{ic}\beta_c + x_i\gamma$$

Where

$$d_{ic} = \begin{cases} 1, & \text{individual } i \text{ is color } c \\ 0, & \text{otherwise} \end{cases}$$

Think of d_{ic} as the elements of a matrix, then each of its columns is a column of the model matrix, and the model matrix has one additional column whose elements are x_i.

Categorical Covariates and Dummy Variables (cont.)

The general principle follows. Whenever one has a categorical covariate with k categories, replace it with k indicator variables, the j-th of which is one if individual i is in category j and zero otherwise.

These k new variables are sometimes called *dummy variables*.

Dummy Variables and Intercepts

By default the R formula mini-language includes an "intercept", that is, it includes a column of ones in the model matrix.

If an intercept is included, then one of the dummy variables must be dropped, because

$$\sum_{j=1}^{k} d_{ij} = 1$$

(every individual is in one and only one category). Thus the dummy variables for a category add up to the the "intercept" variable.

Thus the dummy variables and the "intercept" variable are not linearly independent and any model matrix that contains all of them cannot be full rank.

Dummy Variables and Intercepts (cont.)

If there are multiple categories, then one dummy variable must be dropped from each group if an "intercept" variable is included in the model. This is the default strategy of the R formula mini-language.

> *A categorical covariate with k categories corresponds to $k - 1$ dummy variables, which are columns of the model matrix.*

Which category is "dropped" (not one of the $k - 1$ dummy variables included) is arbitrary. This is another aspect of "regression coefficients are meaningless" (slides 19–23).

Dummy Variables and Intercepts (cont.)

If one chooses not to include an intercept — which, somewhat confusingly is indicated by adding either "- 1" or "+ 0" to the R model formula — then there is no reason to drop a dummy variable if there is only one categorical predictor. All must be included.

If there is more than one categorical covariate and no "intercept", then one dummy variable must be dropped from all but one group. Which group keeps all of its dummy variables is arbitrary. This is another aspect of "regression coefficients are meaningless" (slides 19–23).

Categorical Covariates and Dummy Variables (cont.)

The R terminology for a categorical covariate is `factor`.

R automatically assumes any non-numeric variable appearing in a model formula is a factor, and it automatically assumes any numeric variable appearing in a model formula is not a factor.

If you want a numeric variable to be treated as a factor, you must explicitly say so.

```
out <- lm(y ~ factor(fred) + x)
```

or

```
fred <- factor(fred)
out <- lm(y ~ fred + x)
```

One-Way ANOVA

When all covariates are categorical this is called *analysis of variance* (ANOVA). When there is only one covariate and it is categorical, this is called one-way ANOVA.

The linear model that includes all of the dummy variables and no intercept has the form

$$\mu_i = \beta_{c_i}$$

where c_i is the category for the i-th individual. In effect, we fit a separate mean for each category.

One-Way ANOVA (cont.)

Because we assume (as always in linear models) all components of the response have the same variance, this generalizes the two-sample t procedures that assume equality of variance (deck 2, slides 130–135 and the hypothesis test using the same pivotal quantity).

One-Way ANOVA (cont.)

Because there are several parameters of interest and no single parameter is of much interest by itself, the primary interest is on F tests of model comparison. That is what all the computer examples are about.

Nevertheless, ANOVA are linear models just like any other. All of the theory and practice in this deck of slides is applicable to them. In particular, one can do confidence intervals for regression coefficients or means or prediction intervals for new data. The R function `predict` works the same way for linear models in which all predictors are categorical as it does for any other linear models.

Two-Way ANOVA

When there are exactly two covariates and both are categorical, a linear model is called two-way ANOVA. Again, the primary interest is in F tests of model comparison.

We say the model has only main effects if the R formula specifying the model is `y ~ c + d`, where `y` is the response and `c` and `d` are the categorical predictors.

As discussed above this leads to a model matrix that includes one column of all ones (the "intercept") $k - 1$ dummy variables for the first predictor and $m - 1$ dummy variables for the second predictor if they have k and m categories, respectively.

Two-Way ANOVA (cont.)

This results in the mean being the sum of two terms

$$\mu_i = \beta_{c_i} + \gamma_{d_i} \qquad (*)$$

the mean for individual i is the sum of the main effect for the first predictor and the main effect for the second predictor, in both cases for the categories containing individual i.

If we actually used the parametrization $(*)$, the model matrix would not be full rank and the regression coefficients would not have unique least squares estimates.

Two-Way ANOVA (cont.)

There are many ways to rewrite the model so that it has $k+m-1$ parameters and the model matrix is full rank. The theoretically elegant way, found in many textbooks is to write

$$\mu_i = \alpha + \beta_{c_i} + \gamma_{d_i}$$

where the constraints

$$\sum_{j=1}^{k} \beta_j = 0$$

$$\sum_{j=1}^{m} \gamma_j = 0$$

are imposed. These constraints allow the elimination of two parameters, leaving $k + m - 1$.

Two-Way ANOVA (cont.)

Theoretically elegant this may be, but R does something simpler. It puts in an intercept and drops one of each group of dummy variables. In effect, it uses the model

$$\mu_i = \begin{cases} \alpha + \beta_{c_i} + \gamma_{d_i}, & c_i \neq 1 \text{ and } d_i \neq 1 \\ \alpha + \beta_{c_i}, & c_i \neq 1 \text{ and } d_i = 1 \\ \alpha + \gamma_{d_i}, & c_i = 1 \text{ and } d_i \neq 1 \\ \alpha, & c_i = 1 \text{ and } d_i = 1 \end{cases}$$

For some purposes, it is o. k. to be a bit vague about how the model is parametrized. For other purposes, particularly if regression coefficients are being interpreted, it is crucial to understand the parametrization completely.

Two-Way ANOVA (cont.)

The next step in model complication is to add an "interaction" term.

We say the model having `y` as the response and `c` and `d` as the categorical predictors, has main effects and interaction if the R formula specifying the model is `y ~ c * d`.

Two-Way ANOVA (cont.)

This results in the mean being different for each different pair of categories. Essentially the model is

$$\mu_i = \beta_{c_i d_i}$$

the mean for individual i is the same as for all individuals classified the same way by both categorical covariates.

Two-Way ANOVA (cont.)

The same model can also be written

$$\mu_i = \alpha + \beta_{c_i} + \gamma_{d_i} + \delta_{c_i d_i}$$

imposing the constraints

$$\sum_{j=1}^{k} \beta_j = 0$$

$$\sum_{j=1}^{m} \gamma_j = 0$$

$$\sum_{j=1}^{k} \delta_{ju} = 0, \qquad u = 1, \ldots, m$$

$$\sum_{j=1}^{m} \delta_{uj} = 0, \qquad u = 1, \ldots, k$$

Two-Way ANOVA (cont.)

The same model can also be written

$$\mu_i = \begin{cases} \alpha + \beta_{c_i} + \gamma_{d_i} + \delta_{c_i d_i}, & c_i \neq 1 \text{ and } d_i \neq 1 \\ \alpha + \beta_{c_i}, & c_i \neq 1 \text{ and } d_i = 1 \\ \alpha + \gamma_{d_i}, & c_i = 1 \text{ and } d_i \neq 1 \\ \alpha, & c_i = 1 \text{ and } d_i = 1 \end{cases}$$

This is the parametrization R uses by default.