# Advanced Methods in Biostatistics I

## Lecture 5

Martin Lindquist

September 12, 2017

# General linear model

- Recall we seek to develop least squares for the general linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

# Design matrix

- Let **X** be a design matrix, notationally its elements and column vectors are given by:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = [\mathbf{x}_1 \dots \mathbf{x}_p].$$

- We are assuming that $n \geq p$ and **X** is of full (column) rank.

# Least squares

- Consider the ordinary least squares criteria:

$$f(\beta) \quad = \quad ||\mathbf{y} - \mathbf{X}\beta||^2$$

- We showed last time that it has the following solution:

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

# Fitted values

- The vector of fitted values is given by

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

- Here the matrix **H** is called the hat matrix.
- **H** is idempotent and symmetric.

# Residuals

- The vector of residuals is given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

- $\mathbf{I} - \mathbf{H}$ is idempotent and symmetric.

# Residuals

- Note that $(\mathbf{I} - \mathbf{H})\mathbf{X} = 0$, making the residuals orthogonal to any vector, $\mathbf{X}\gamma$, in the space spanned by the columns of $\mathbf{X}$.
- Hence, if an intercept term is included in the model, the residuals must sum to 0.
- Specifically, since the residuals are orthogonal to any column of $\mathbf{X}$, $\mathbf{e}'\mathbf{J}_n = 0$.

# Geometrical perspective

- Consider the column space of the design matrix,

$$\Gamma = \{\mathbf{X}\beta \mid \beta \in \mathbb{R}^p\}.$$

- This $p$-dimensional space belongs to $\mathbb{R}^n$.

# Geometrical perspective

- Consider the vector $\mathbf{y} \in \mathbb{R}^n$.
- Multiplication by the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ projects $\mathbf{y}$ into $\Gamma$.
- That is,

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

  is the linear projection map between $\mathbb{R}^n$ and $\Gamma$.

# Geometrical perspective

- The vector $\hat{\mathbf{y}}$ is the point in $\Gamma$ that is closest to $\mathbf{y}$ and $\hat{\beta}$ is the specific linear combination of the columns of $\mathbf{X}$ that yields $\hat{\mathbf{y}}$.
- The vector $\mathbf{e}$ is the vector connecting $\mathbf{y}$ and $\hat{\mathbf{y}}$, and is orthogonal to all elements in $\Gamma$, i.e. it lies in $\Gamma^{\perp}$.
- It represents the projection of $\mathbf{y}$ onto $\Gamma^{\perp}$.

# Geometrical perspective

- Note that if $\mathbf{W}$ is any $p \times p$ invertible matrix, then the fitted values, $\hat{\mathbf{y}}$ will be the same for the design matrix $\mathbf{XW}$.
- This holds because the spaces $\{\mathbf{X}\beta \mid \beta \in \mathbb{R}^p\}$ and $\{\mathbf{XW}\gamma \mid \gamma \in \mathbb{R}^p\}$ are the same, since if $\mathbf{a} = \mathbf{X}\beta$ then $\mathbf{a} = \mathbf{X}\gamma$ via the relationship $\gamma = \mathbf{W}\beta$.

# Geometrical perspective

- Thus, any element of the first space lies in the second.
- The same argument implies in the other direction, thus the two spaces are the same.
- Any linear reorganization of the columns of $\mathbf{X}$ results in the same column space and the same fitted values.

# Full row rank case

- In the case where $\mathbf{X}$ is $n \times n$ of full rank, then the columns of $\mathbf{X}$ form a basis for $\mathbb{R}^n$.
- In this case, $\hat{\mathbf{y}} = \mathbf{y}$, since $\mathbf{y}$ lives in the space spanned by the columns of $\mathbf{X}$.
- All this linear model accomplishes is a lossless linear reorganization of $\mathbf{y}$.

# Full row rank case

- This is surprisingly useful, especially when the columns of $\mathbf{X}$ are orthonormal ($\mathbf{X}'\mathbf{X} = \mathbf{I}$).
- In this case, the function that takes the outcome vector and converts it to the coefficients is called a "transform".
- The most well known versions of transforms are Fourier and wavelet.

# Another approach

- Next let's look at the problem from another perspective.
- Let $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$ be two submatrices of dimension $n \times p_1$ and $n \times p_2$, respectively, and let $\beta = (\beta_1' \ \beta_2')'$.
- Consider minimizing:

$$||\mathbf{y} - \mathbf{X}_1\beta_1 - \mathbf{X}_2\beta_2||^2.$$

# Another approach

- If we hold $\beta_2$ fixed, this would be minimized when

$$\hat{\beta}_1(\beta_2) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\beta_2).$$

- Plugging this result back into the least squares criteria we obtain:

$$\begin{aligned} ||\mathbf{y} &- \mathbf{X}_1\beta_1 - \mathbf{X}_2\beta_2||^2 \\ &\leq ||(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y} - (\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2\beta_2||^2 \end{aligned}$$

# Another approach

- This is equivalent to the least squares problem where the response variable is the residual of $\mathbf{y}$ having regressed out $\mathbf{X}_1$, and the explanatory variables the residual matrix of $\mathbf{X}_2$ having regressed $\mathbf{X}_1$ out of every column.
- Our estimate of $\beta_2$ will be the regression of these two sets of residuals.

# Another approach

- This illustrates how the estimate of $\beta_2$ has been adjusted for $\mathbf{X}_1$, both the outcome and the $\mathbf{X}_2$ predictors have been orthogonalized to the space spanned by the columns of $\mathbf{X}_1$.
- This example helps our interpretation of the regression coefficients and how they are "adjust" for the other variables.

# Another approach

- The estimate of $\beta_2$ represents the effect of the explanatory variables, $\mathbf{X}_2$, while controlling for the effects of the other explanatory variables in the model, i.e. $\mathbf{X}_1$.
- Ultimately the interpretation of a coefficient depends on which other variables are included in the model.
- An exception is when variables are orthogonal.

# R code

Recall the Swiss fertility data.

```
> y = swiss$Fertility
> X = as.matrix(swiss[,-1])
> dim(X)
[1] 47  5
> X1 = X[,1:3]
> X2 = X[,4:5]

> ytilde = (I - X1%*%solve(t(X1)%*%X1)%*%t(X1))%*%y
> X2tilde = (I - X1%*%solve(t(X1)%*%X1)%*%t(X1))%*%X2
> beta2 = solve(t(X2tilde)%*%X2tilde)%*%t(X2tilde)%*%ytilde
> beta2
                      [,1]
Catholic         0.1170662
Infant.Mortality 2.9836617

> beta1 = solve(t(X1)%*%X1)%*%t(X1)%*%(y - X2%*%beta2)
> beta1
                  [,1]
Agriculture  0.1110005
Examination  0.4440591
Education   -0.7067362
```

# R code

Soultion using `lm`.

```
> summary(lm(y ~ X - 1))$coef
                  Estimate Std. Error   t value     Pr(>|t|)
Agriculture      0.1110005 0.07423536  1.495250 1.423257e-01
Examination      0.4440591 0.31435258  1.412615 1.651367e-01
Education       -0.7067362 0.25008979 -2.825930 7.186594e-03
Catholic         0.1170662 0.04859619  2.408958 2.046207e-02
Infant.Mortality 2.9836617 0.31682721  9.417315 6.528210e-12
```

# Mean centering

- Before continuing, it is useful to note that the mean centered version of $\mathbf{y}$, $\mathbf{y} - \mathbf{J}_n \bar{y}$ can be written as follows:

$$
\begin{aligned}
\tilde{\mathbf{y}} &= \mathbf{y} - \mathbf{J}_n \bar{y} \\
&= \mathbf{y} - \mathbf{J}_n (\mathbf{J}_n' \mathbf{J}_n)^{-1} \mathbf{J}_n' \mathbf{y} \\
&= (\mathbf{I} - \mathbf{J}_n (\mathbf{J}_n' \mathbf{J}_n)^{-1} \mathbf{J}_n') \mathbf{y}.
\end{aligned}
$$

# Mean centering

- In other words, multiplication by the matrix $(\mathbf{I} - \mathbf{J}_n(\mathbf{J}_n'\mathbf{J}_n)^{-1}\mathbf{J}_n')$ centers vectors.
- This can be very useful for centering matrices as well.
- For example, if $\mathbf{X}$ is an $n \times p$ matrix then the matrix $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{J}_n(\mathbf{J}_n'\mathbf{J}_n)^{-1}\mathbf{J}_n')\mathbf{X}$ is the matrix with every column mean centered.

# Partitioning the variability

- Using this result, we now seek to partition the variation in the data into various components.
- For convenience, let us define two projection matrices:

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

and

$$\mathbf{H_J} = \mathbf{J}_n(\mathbf{J}_n'\mathbf{J}_n)^{-1}\mathbf{J}_n'.$$

# Total sums of squares

- Let us define the total sum of squares as

$$SS_{Tot} = ||\mathbf{y} - \bar{y}\mathbf{J}_n||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H}_\mathbf{J})\mathbf{y}.$$

- This is an unscaled measure of the total variability in the data.

# Residual & Regression sums of squares

- Similarly, given a design matrix, $\mathbf{X}$, we can define the residual sums of squares as

$$\mathrm{SS}_{Res} = ||\mathbf{y} - \hat{\mathbf{y}}||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}})\mathbf{y}$$

and the regression sums of squares as

$$\mathrm{SS}_{Reg} = ||\hat{\mathbf{y}} - \mathbf{J}_n \bar{y}||^2 = \mathbf{y}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{J}})\mathbf{y}.$$

# Regression sums of squares

- To show the later result first note that $(\mathbf{I} - \mathbf{H_X})\mathbf{J}_n = 0$ since $\mathbf{X}$ contains an intercept.
- Thus, it holds that $\mathbf{H_X}\mathbf{J}_n = \mathbf{J}_n$ and $\mathbf{H_X}\mathbf{H_J} = \mathbf{H_J}$ and $\mathbf{H_J} = \mathbf{H_J}\mathbf{H_X}$.
- Also, note that $\mathbf{H_X}$ is symmetric and idempotent.

# Regression sums of squares

- Now we can perform the following manipulation

$$
\begin{aligned}
||\hat{\mathbf{y}} - \mathbf{J}_n \bar{y}||^2 &= ||\mathbf{H}_{\mathbf{X}}\mathbf{y} - \mathbf{J}_n(\mathbf{J}_n'\mathbf{J}_n)^{-1}\mathbf{J}_n'\mathbf{y}||^2 \\
&= ||\mathbf{H}_{\mathbf{X}}\mathbf{y} - \mathbf{H}_{\mathbf{J}}\mathbf{y}||^2 \\
&= \mathbf{y}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{J}})'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{J}})\mathbf{y} \\
&= \mathbf{y}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{J}})(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{J}})\mathbf{y} \\
&= \mathbf{y}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{J}}\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{X}}\mathbf{H}_{\mathbf{J}} + \mathbf{H}_{\mathbf{J}})\mathbf{y} \\
&= \mathbf{y}'(\mathbf{H}_{\mathbf{X}} - \mathbf{H}_{\mathbf{J}})\mathbf{y}.
\end{aligned}
$$

# Partitioning the variability

- Using this identity we can now show that

$$\begin{aligned} SS_{Tot} &= \mathbf{y}'(\mathbf{I} - \mathbf{H_J})\mathbf{y} \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{H_X} + \mathbf{H_X} - \mathbf{H_J})\mathbf{y} \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{H_X})\mathbf{y} + \mathbf{y}'(\mathbf{H_X} - \mathbf{H_J})\mathbf{y} \\ &= SS_{Res} + SS_{Reg} \end{aligned}$$

- Thus the total sum of squares partition into the residual and regression sums of squares.

# $R^2$

- Using this result, we can now define the coefficient of determination

$$R^2 = \frac{SS_{Reg}}{SS_{Tot}} = 1 - \frac{SS_{Res}}{SS_{Tot}}.$$

- This represents the proportion of the total variability explained by our model.
- This is guaranteed to be between 0 and 1.
- High values imply that the explanatory variables are useful in explaining the response and low values imply that the explanatory variables are not useful.

# Problems with $R^2$

- Note that $SS_{Tot}$ only depends on the response variable and not on the model formulation.
- Hence, it is equal for all regression models.
- Adding additional explanatory variables to a multiple regression model can only lower $SS_{Reg}$.

# Problems with $R^2$

- Thus, including additional explanatory variables will always lead to an increase in the value of $R^2$.
- Since $R^2$ can be made large by including more (and sometimes unimportant) explanatory variables, it is sometimes modified to adjust for the number of variables included in the model.
- This allows us to balance model parsimony with explanatory power.

# Mean squares

- The ratio of the sum of squares to the ‘degrees of freedom’ (corresponding to the dimensions of the respective subspaces) gives the mean squares:

$$\mathrm{MS}_{Tot} = \frac{\mathrm{SS}_{Tot}}{n-1}$$

$$\mathrm{MS}_{Res} = \frac{\mathrm{SS}_{Res}}{n-p}$$

$$\mathrm{MS}_{Reg} = \frac{\mathrm{SS}_{Reg}}{p-1}$$

# Adjusted $R^2$

- The adjusted coefficient of multiple determination, uses the mean squares instead of the sums of square, i.e.

$$R_a^2 = 1 - \frac{\mathrm{MS}_{Res}}{\mathrm{MS}_{Tot}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\mathrm{SS}_{Res}}{\mathrm{SS}_{Tot}}.$$

- Since the term includes the number of model parameters, $p$, it penalizes for model complexity.

# R code

```
> fit = lm(y ~ X)
> summary(fit)

Call:
lm(formula = y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2743  -5.2617   0.5032   4.1198  15.3213 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       66.91518   10.70604   6.250 1.91e-07 ***
XAgriculture      -0.17211    0.07030  -2.448  0.01873 *  
XExamination      -0.25801    0.25388  -1.016  0.31546    
XEducation        -0.87094    0.18303  -4.758 2.43e-05 ***
XCatholic          0.10412    0.03526   2.953  0.00519 ** 
XInfant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,Adjusted R-squared:  0.671 
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

# R code

Computing the sums of square.

```
> anova(fit)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          5 5072.9 1014.58  19.761 5.594e-10 ***
Residuals 41 2105.0   51.34
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> SSreg = anova(fit)[1,2]
> SSres = anova(fit)[2,2]
> SStot = SSres + SSreg

> 1-SSres/SStot
[1] 0.706735
```