Advanced Methods in Biostatistics I

Lecture 6

Martin Lindquist

September 14, 2017

Conceptual examples of least squares

- Today we discuss some conceptual examples of least squares.
- This will illustrate the flexibility of the approach and how it allows us to effectively analyze different types of data.
- First, we revisit some of the models previously used, before moving on to introduce some new ones.

Mean only regression

- First let us revisit mean only regression, which can be expressed as as $\mathbf{y} = \mathbf{J}_n\mu + \epsilon$.
- Placing this into the multivariate least-squares framework, our design matrix is $\mathbf{X} = \mathbf{J}_n$.
- Our coefficient estimate is therefore:

$$\hat{\mu} = (\mathbf{J}_n'\mathbf{J}_n)^{-1}\mathbf{J}_n'\mathbf{y} = \bar{y}.$$

Regression through the origin

- Next, we revisit the regression through the origin problem, i.e., $\mathbf{y} = \mathbf{x}\beta + \epsilon$.
- Here the design matrix is $\mathbf{X} = \mathbf{x}$.
- Our coefficient estimate is therefore:

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{||\mathbf{x}||^2}.$$

Regression through the origin - Comments

- In this setting the residuals do not need to sum to 0, as the only constraint is that $\mathbf{e}'\mathbf{x} = 0$.
- In addition, R^2 has no clear meaning for regression through origin
- It is possible for $SS_{Res} > SS_{Tot}$.

Regression through the origin - Comments

- Regression through the origin should not be forced unless there are compelling reasons.
- If the line does go through the origin, little is lost by fitting a line with both intercept and slope

Linear regression

- Finally, we revisit simple linear regression, i.e.

$$\mathbf{y} = \mathbf{J}_n\beta_0 + \mathbf{x}\beta_1 + \epsilon.$$

- Here we can write the design matrix as $\mathbf{X} = [\mathbf{J}_n \ \mathbf{x}]$.
- Now, the estimate of $\beta = [\beta_0 \ \beta_1]'$ can be obtained through the equations:

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

Linear regression

- Here the term $\mathbf{X}'\mathbf{X}$ is a 2×2 matrix and easily invertible.
- It is thus relatively easy to show that this solution corresponds to the one we have previously obtained.
- We leave this as an exercise.

ANOVA

- Analysis of Variance (ANOVA) is a technique for comparing the means across multiple groups.
- For example, we may be interested in determining whether the cholesterol levels (y) differ between subjects in a drug group and a control group.

ANOVA

- There are several ways to formulate this model, including:

$$y_{ij} = \alpha_i + \epsilon_{ij} \qquad \text{for } j = 1, \ldots n_i; \quad i = 1, 2$$

or

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix}$$

Dummy variables

- Here the first column codes whether an observation belongs to the first group, and the second column whether it belongs to the second group.
- These types of indicator variables, or 'dummy variables' are often used to denote values of a categorical variable.

ANOVA

- To estimate α_1 and α_2, first note that:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}$$

and

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} n_1\bar{\mathbf{y}}_1 & 0 \\ 0 & n_2\bar{\mathbf{y}}_2 \end{pmatrix}.$$

- The solution is obtained by computing $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$:

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{y}}_1 \\ \bar{\mathbf{y}}_2 \end{pmatrix}.$$

ANOVA

- We can generalize this to more than two treatment groups.
- Assume J groups, each with K observations.
- Denote the outcome vector, $\mathbf{y}$, as comprised of y_{ij} for $i = 1, \ldots K$ and $j = 1, \ldots, J$ stacked in the relevant order, i.e. $\mathbf{y} = [y_{11}, \ldots, y_{1K}, y_{2,1}, \ldots, y_{JK}]'$.

Kronecker product

Definition

The Kronecker product of the $p \times q$ matrix $\mathbf{A}$ with the $r \times s$ matrix $\mathbf{B}$ is defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1q}\mathbf{B} \\ \vdots & & \vdots \\ a_{p1}\mathbf{B} & \dots & a_{pq}\mathbf{B} \end{bmatrix}$$

ANOVA

- The design matrix can be expressed as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \mathbf{I}_K \otimes \mathbf{J}_n,$$

where $\otimes$ is the Kronecker product.

ANOVA

- Let $\bar{y}_j$ be the mean of the **y** measurements in group j.
- Then it is straightforward to show that

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} K\bar{y}_1 \\ \vdots \\ K\bar{y}_J \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{X} = K\mathbf{I}.$$

ANOVA

- Therefore, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\bar{y}_1 \ldots, \bar{y}_J)'$.
- Thus, if our design matrix parcels $\mathbf{y}$ into groups, the coefficients are the group means.

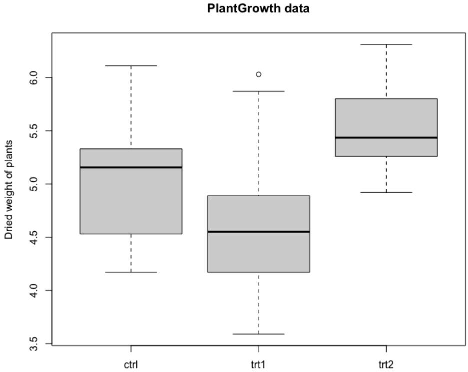
R code

The data set `PlantGrowth` in `R` contains results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions.

```
A data frame of 30 cases on 2 variables.

[, 1]   weight     numeric
[, 2]   group      factor

The levels of group are 'ctrl', 'trt1', and 'trt2'.
```

R code

```
> boxplot(weight ~ group, data = PlantGrowth,
+          main = "PlantGrowth data",
+          ylab = "Dried weight of plants", col = "lightgray")
```



R code

```
> fit = lm(weight ~ group -1, data = PlantGrowth)
> summary(fit)

Call:
lm(formula = weight ~ group - 1, data = PlantGrowth)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
groupctrl   5.0320     0.1971   25.53   <2e-16 ***
grouptrt1   4.6610     0.1971   23.64   <2e-16 ***
grouptrt2   5.5260     0.1971   28.03   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared:  0.9867,Adjusted R-squared:  0.9852
F-statistic: 665.5 on 3 and 27 DF,  p-value: < 2.2e-16
```

ANCOVA

- Next, we consider analysis of covariance, or ANCOVA.
- This approach allows us to compare differences in means between two or more groups while taking into account the variability of other variables, called covariates.

- Suppose we have data on two variables **x** and **y** collected for two separate groups (A and B).
- Let $\mathbf{x} = (\mathbf{x}_1 \ \mathbf{x}_2)'$, where $\mathbf{x}_1$ are the observations associated with group A and $\mathbf{x}_2$ those associated with group B.

ANCOVA

- We can write this model as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & x_{11} \\ 1 & 0 & x_{12} \\ \vdots & \dots & \vdots \\ 1 & 0 & x_{1n} \\ 0 & 1 & x_{21} \\ 0 & 1 & x_{22} \\ \vdots & \dots & \dots \\ 0 & 1 & x_{2n} \end{bmatrix} = [\mathbf{I}_2 \otimes \mathbf{J}_n \;\; \mathbf{x}].$$

ANCOVA

- In this setting we seek to project $\mathbf{y}$ onto the space spanned by two groups and a regression variable.
- This is equivalent to fitting two parallel lines to the data.

ANCOVA

- Let $\boldsymbol{\beta} = (\mu_1 \ \mu_2 \ \beta)' = (\boldsymbol{\mu}' \ \beta)'$.
- Denote the outcome vector, $\mathbf{y}$, as comprised of y_{ij} for $i = 1, 2$ and $j = 1, \ldots, n$ stacked in the relevant order.
- Begin by holding β fixed.
- We now want to solve:

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = ||\mathbf{y} - \mathbf{x}\beta - (\mathbf{I}_2 \otimes \mathbf{1}_n)\boldsymbol{\mu}||^2 \tag{1}$$

ANCOVA

- This is equivalent to the ANOVA problem, and the best estimate of $\boldsymbol{\mu}$ are the 'group means', which can be written:

$$\frac{1}{n}(\mathbf{I}_2 \otimes \mathbf{J}_n)'(y - \mathbf{x}\beta) = (\bar{y}_1 \;\; \bar{y}_2)' - (\bar{x}_1 \;\; \bar{x}_2)'\beta$$

where $\bar{y}_i$ and $\bar{x}_i$ are the group means of $\mathbf{y}$ and $\mathbf{x}$, respectively.

ANCOVA

- Now, it holds that (1) satisfies:

$$\begin{aligned}(1) \quad &\geq \quad ||\mathbf{y} - \mathbf{x}\beta - (\mathbf{I}_2 \otimes \mathbf{1}_n)\{(\bar{y}_1 \ \bar{y}_2)' - (\bar{x}_1 \ \bar{x}_2)'\beta\}||^2 \\ &= \quad ||\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\beta||^2\end{aligned}$$

  where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ are the group centered versions of $\mathbf{y}$ and $\mathbf{x}$, respectively.

- For example, $\tilde{y}_{ij} = y_{ij} - \bar{y}_i$.

ANCOVA

- This is now equivalent to the regression through the origin problem, yielding the solution:

$$\hat{\beta} = \frac{\sum_{ij}(y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)}{\sum_{ij}(x_{ij} - \bar{x}_i)^2} = p\hat{\beta}_1 + (1 - p)\hat{\beta}_2$$

where

$$p = \frac{\sum_j (x_{1j} - \bar{x}_1)^2}{\sum_{ij}(x_{ij} - \bar{x}_i)^2}$$

and

$$\hat{\beta}_i = \frac{\sum_j (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)}{\sum_j (x_{ij} - \bar{x}_i)^2}.$$

ANCOVA

- This implies that the estimated slope is a convex combination of the group-specific slopes weighted by the variability in the **x**'s within the group.
- Furthermore, $\hat{\mu}_i = \bar{y}_i - \bar{x}_i\hat{\beta}$ and thus

$$\hat{\mu}_1 - \hat{\mu}_2 = (\bar{y}_1 - \bar{y}_2) - (\bar{x}_1 - \bar{x}_2)\hat{\beta}.$$

R code

- We illustrate simple linear regression using the `mtcars` data set that is directly available in R. The data was extracted from the 1974 Motor Trend US magazine, and consists of gas consumption (`mpg`) and 10 other aspects of automobile design and performance for a total of 32 cars.

- Here we focus on how mileage depends upon transmission type (`am`) (0 = automatic, 1 = manual), controlling for the weight of the car (`wt`).

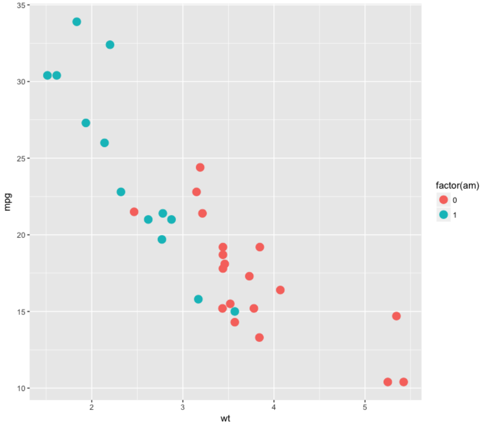
R code

First fit ANOVA, ignoring weight.

```
> fit = lm(mpg~factor(am) - 1,data = mtcars)
> summary(fit)

Call:
lm(formula = mpg ~ factor(am) - 1, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3923 -3.0923 -0.2974  3.2439  9.5077

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
factor(am)0   17.147      1.125   15.25 1.13e-15 ***
factor(am)1   24.392      1.360   17.94  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 4.902 on 30 degrees of freedom
Multiple R-squared:  0.9487,Adjusted R-squared:  0.9452
F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

R code

```
> install.packages("ggplot2")
> library(ggplot2)
> ggplot(mtcars, aes(x=wt, y=mpg, color=factor(am))) + geom_point(size=4)
```



R code

Now fit ANCOVA, controlling for weight.

```
> fit = lm(mpg~factor(am) +wt - 1,data = mtcars)
> summary(fit)

Call:
lm(formula = mpg ~ factor(am) + wt - 1, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5295 -2.3619 -0.1317  1.4025  6.8782

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
factor(am)0  37.3216     3.0546  12.218 5.84e-13 ***
factor(am)1  37.2979     2.0857  17.883  < 2e-16 ***
wt           -5.3528     0.7882  -6.791 1.87e-07 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 3.098 on 29 degrees of freedom
Multiple R-squared:  0.9802,Adjusted R-squared:  0.9781
F-statistic: 478.1 on 3 and 29 DF,  p-value: < 2.2e-16
```

Including interactions

- In the ANCOVA model we used an indicator variable to model differences in the intercept between groups.
- Sometimes we also want the slopes of the regression model to differ between groups.
- This can be done by including an interaction term together with an indicator variable in the model.

Including interactions

- Suppose we have data on two variables $\mathbf{z}$ and $\mathbf{y}$ collected for two groups (A and B).
- Let $\mathbf{x}_1$ be equal to 1 if the observation belongs to group A and 0 if it belongs to group B.
- Let $\mathbf{x}_2$ be equal to 1 if the observation belongs to group B and 0 if it belongs to group A.
- Let $\mathbf{z} = (\mathbf{z}_1 \ \mathbf{z}_2)'$, where $\mathbf{z}_1$ are the observations associated with group A and $\mathbf{z}_2$ those associated with group B.

Including interactions

- Consider the following model with interactions:

$$\mathbf{y} = \mathbf{J}_n\mu_1 + \mathbf{x}_2\mu_2 + \mathbf{z}\beta_1 + \mathbf{z} * \mathbf{x}_2\beta_2 + \epsilon$$

- We can fit this model using the following design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & z_{11} & 0 \\ 1 & 0 & z_{12} & 0 \\ \vdots & \dots & \vdots & \dots \\ 1 & 0 & z_{1n} & 0 \\ 1 & 1 & z_{21} & z_{21} \\ 1 & 1 & z_{22} & z_{22} \\ \vdots & \dots & \dots & \dots \\ 1 & 1 & z_{2n} & z_{2n} \end{bmatrix}.$$

Including interactions

- The model allows both the slopes and intercepts to vary between groups.
- It can be fit in the same manner as described above.

R code

```
> fit = lm(mpg ~ factor(am) + wt + factor(am)*wt, data = mtcars)
> summary(fit)

Call:
lm(formula = mpg ~ factor(am) + wt + factor(am) * wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6004 -1.5446 -0.5325  0.9012  6.0909

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    31.4161     3.0201  10.402 4.00e-11 ***
factor(am)1    14.8784     4.2640   3.489  0.00162 **
wt             -3.7859     0.7856  -4.819 4.55e-05 ***
factor(am)1:wt -5.2984     1.4447  -3.667  0.00102 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.591 on 28 degrees of freedom
Multiple R-squared:  0.833,Adjusted R-squared:  0.8151
F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

R code

```
> install.packages("ggplot2")
> library(ggplot2)
> ggplot(mtcars, aes(x=wt, y=mpg, color=factor(am)))
+ geom_point(size=4)
+ geom_smooth(aes(group=factor(am)), method="lm", se=FALSE, lty="dashed")
```