Multiple Linear Regression: Collinearity and Categories

Author: Nicholas G Reich, Jeff Goldsmith

This material is part of the ***statsTeachR*** *project*

Recap: Least squares for MLR

As in simple linear regression, we want to find the $\boldsymbol{\beta}$ that minimizes the residual sum of squares.

$$RSS(\boldsymbol{\beta}) = \sum_i \epsilon_i^2 = \epsilon^T \epsilon$$

After taking the derivative, setting equal to zero, we obtain:

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

Hat matrix

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

Some properties of the hat matrix:

- It is a projection matrix: $\mathbf{HH} = \mathbf{H}$
- It is symmetric: $\mathbf{H}^T = \mathbf{H}$
- The residuals are $\hat{\epsilon} = (\mathbf{I} - \mathbf{H})\mathbf{y}$
- The inner product of $(\mathbf{I} - \mathbf{H})\mathbf{y}$ and $\mathbf{Hy}$ is zero (predicted values and residuals are uncorrelated).

Projection space interpretation

The hat matrix projects $\mathbf{y}$ onto the column space of $\mathbf{X}$.
Alternatively, minimizing the $RSS(\beta)$ is equivalent to minimizing the Euclidean distance between $\mathbf{y}$ and the column space of $\mathbf{X}$.

Lung Data Example (con't from previous clas)

```
mlr2 <- lm(disease ~ crowding + education + airqual,
           data=dat, x=TRUE, y=TRUE)
coef(mlr2)

## (Intercept)    crowding   education     airqual
##     -7.7505      1.3128      1.4377      0.2881

X = mlr2$x
y = mlr2$y
(betaHat = solve( t(X) %*% X) %*% t(X) %*% y )

##                [,1]
## (Intercept) -7.7505
## crowding     1.3128
## education    1.4377
## airqual      0.2881
```

Key points so far

- Our model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\epsilon \sim (0, \sigma^2 \mathbf{I})$
- The design matrix $\mathbf{X}$ contains the terms included in the model
- We have least squares solutions under some conditions

Least squares estimates

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

A condition on $\left(\mathbf{X}^T\mathbf{X}\right)$

- If $\left(\mathbf{X}^T\mathbf{X}\right)$ is singular, there are infinitely many least squares solutions, making $\hat{\boldsymbol{\beta}}$ non-identifiable (can't choose between different solutions)

Non-identifiability

- Can happen if $\mathbf{X}$ is not of full rank, i.e. the columns of $\mathbf{X}$ are linearly dependent (for example, including weight in Kg and lb as predictors)
- Can happen if there are fewer data points than terms in $\mathbf{X}$: $n < p$ (having 100 predictors and only 50 observations)
- Generally, the $p \times p$ matrix $\left(\mathbf{X}^T\mathbf{X}\right)$ is invertible if and only if it has rank p.

Infinite solutions

Suppose I fit a model $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$.

- I have estimates $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2$
- I put in a new variable $x_2 = x_1$
- My new model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
- Possible least squares estimates that are equivalent to my first model:
  - $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2, \hat{\beta}_2 = 0$
  - $\hat{\beta}_0 = 1, \hat{\beta}_1 = 0, \hat{\beta}_2 = 2$
  - $\hat{\beta}_0 = 1, \hat{\beta}_1 = 1002, \hat{\beta}_2 = -1000$
  - $\ldots$

Non-identifiablity

- Often due to data coding errors (variable duplication, scale changes)
- Pretty easy to detect and resolve
- Can be addressed using *penalties* (might come up much later)
- A bigger problem is near-unidentifiability (collinearity)

Causes of collinearity

- Arises when variables are highly correlated, but not exact duplicates
- Commonly arises in data (perfect correlation is usually there by mistake)
- Might exist between several variables, i.e. a linear combination of several variables exists in the data
- A variety of tools exist (correlation analyses, multiple R^2, eigen decompositions)

Effects of collinearity

Suppose I fit a model $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$.

- I have estimates $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2$
- I put in a new variable $x_2 = x_1 + \mathit{error}$, where *error* is pretty small
- My new model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
- Possible least squares estimates that are nearly equivalent to my first model:
  - $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2, \hat{\beta}_2 = 0$
  - $\hat{\beta}_0 = 1, \hat{\beta}_1 = 0, \hat{\beta}_2 = 2$
  - $\hat{\beta}_0 = 1, \hat{\beta}_1 = 1002, \hat{\beta}_2 = -1000$
  - $\ldots$
- A unique solution exists, but it is hard to find

Effects of collinearity

- Collinearity results in a "flat" RSS
- Makes identifying a unique solution difficult
- Dramatically inflates the variance of LSEs

Non-identifiability example: lung data

```
mlr3 <- lm(disease ~ airqual, data=dat)
coef(mlr3)

## (Intercept)     airqual
##     35.4445      0.3537

dat$x2 <- dat$airqual/100
mlr4 <- lm(disease ~ airqual + x2, data=dat, x=TRUE)
coef(mlr4)

## (Intercept)     airqual          x2
##     35.4445      0.3537          NA

X = mlr4$x
solve( t(X) %*% X)

## Error:  system is computationally singular:
reciprocal condition number = 3.57906e-20
```

Collinearity example: lung data

```
dat$crowd2 <- dat$crowding + rnorm(nrow(dat), sd=.1)
mlr5 <- lm(disease ~ crowding, data=dat)
summary(mlr5)$coef

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   12.992     3.4750   3.739 3.130e-04
## crowding       1.509     0.1394  10.826 2.232e-18

mlr6 <- lm(disease ~ crowding + crowd2, data=dat)
summary(mlr6)$coef

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   12.642      3.518  3.5937 0.0005166
## crowding      -4.015      7.719 -0.5202 0.6041320
## crowd2         5.537      7.736  0.7158 0.4758717
```

Some take away messages

- Collinearity can (and does) happen, so be careful
- Often contributes to the problem of variable selection, which we'll touch on later

Categorical predictors

- Assume X is a categorical / nominal / factor variable with k levels
- With only one categorical X, we have classic one-way ANOVA design
- Can't use a single predictor with levels $1, 2, \ldots, K$ – this has the wrong interpretation
- Need to create *indicator* or *dummy* variables

Indicator variables

- Choose one group as the baseline
- Create 0/1 terms to include in the model $x_1, x_2, \ldots x_{k=1}$
- Pose the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{k-1} x_{i,k-1} + \epsilon_i$$

  and estimate parameters using least squares
- Note distinction between *predictors* and *terms*

Categorical predictor design matrix

Which of the following is a "correct" design matrix for a categorical predictor with 3 levels?

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{or} \quad \mathbf{X}_2 = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \quad \text{or} \quad \mathbf{X}_3 = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix}$$

ANOVA model interpretation

Using the model $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{k-1} x_{i,k-1} + \epsilon_i$, interpret

$\beta_0 =$

$\beta_1 =$

Equivalent model

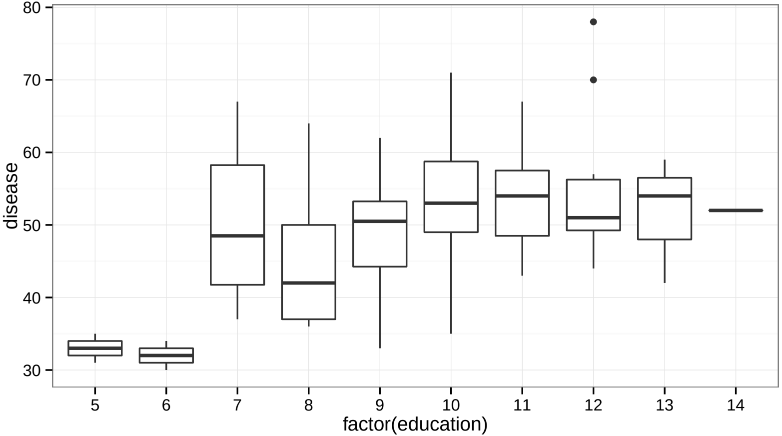
Define the model $y_i = \beta_1 x_{i1} + \ldots + \beta_k x_{i,k} + \epsilon_i$ where there are indicators for each possible group

$\beta_1 =$

$\beta_2 =$

Categorical predictor example: lung data

```
require(ggplot2)
qplot(factor(education), disease, geom="boxplot", data=dat)
```



Categorical predictor example: lung data

```
mlr7 <- lm(disease ~ factor(education), data=dat)
summary(mlr7)$coef

##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)             33.00      4.913  6.7173 1.689e-09
## factor(education)6      -1.00      7.768 -0.1287 8.979e-01
## factor(education)7      17.33      6.017  2.8808 4.969e-03
## factor(education)8      11.18      5.329  2.0975 3.879e-02
## factor(education)9      15.50      5.353  2.8953 4.765e-03
## factor(education)10     20.38      5.188  3.9289 1.683e-04
## factor(education)11     20.53      5.382  3.8155 2.505e-04
## factor(education)12     22.20      5.601  3.9633 1.489e-04
## factor(education)13     18.67      6.948  2.6868 8.609e-03
## factor(education)14     19.00      9.825  1.9338 5.632e-02
```

Categorical predictor example: lung data

```
mlr8 <- lm(disease ~ factor(education) - 1, data=dat)
summary(mlr8)$coef

##                      Estimate Std. Error t value  Pr(>|t|)
## factor(education)5     33.00      4.913   6.717 1.689e-09
## factor(education)6     32.00      6.017   5.318 7.716e-07
## factor(education)7     50.33      3.474  14.489 3.846e-25
## factor(education)8     44.18      2.064  21.406 7.303e-37
## factor(education)9     48.50      2.127  22.799 6.282e-39
## factor(education)10    53.38      1.669  31.991 1.359e-50
## factor(education)11    53.53      2.197  24.366 3.801e-41
## factor(education)12    55.20      2.691  20.514 1.713e-35
## factor(education)13    51.67      4.913  10.517 2.758e-17
## factor(education)14    52.00      8.509   6.111 2.561e-08
```

Today's big ideas

- Multiple linear regression models, projections, collinearity, categorical variables