MATH 531 Homework 7 March 12, 2024

OLS distribution theory

Each subsection counts for 10 points

Introduction

Throughout assume a linear model

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

where $\mathbf{y}$ is an vector of length n and X is a full rank, $n \times p$ matrix. $\mathbf{e}_i$ are independent $N(0, \sigma^2 I)$.

1. (a) Following the notation and results in Section 3.8.1 of Seber and Lee let $\hat{\beta}_H$ be the OLS estimate under the constraint $A\beta = \mathbf{c}$ for some known A and $\mathbf{c}$ (here A is $q \times p$ and of course $q < p$.) Refer to the formula (3.38) for $\hat{\beta}_H$ and define the matrix P_H as satisfying $P_H\mathbf{y} = X\hat{\beta}_H$ (Yes, P_H is pretty complicated!).
   Show explicitly that P_H is a projection matrix.

   (b) Seber and Lee also give a different derivation of the constrained OLS estimate in section 3.8.2. In their development what is the role of β_0? How could this idea be used in section 3.8.1 to simplify the derivation?

2. Following the linear model displayed above let $X = QR$ be the "QR" decomposition for X Where the columns of Q are orthogonal and R is upper triangular. This decomposition always exists for X full rank. Now reparametrize the model as $\gamma = R\beta$ and so

   $$\mathbf{y} = Q\gamma + \mathbf{e}$$

   and set $\hat{\gamma}$ as the OLS estimate for γ.

   (a) Show that $\{\hat{\gamma}_j\}$ are independent, and normal with variance σ^2

   (b) Let $\hat{\sigma}^2$ be the usual unbiased estimate of σ^2 based on the residuals and let $\gamma_{T,j}$ be the true value of the parameter.
   Show that

   $$\frac{(\hat{\gamma}_j - \gamma_{T,j})^2}{\hat{\sigma}^2}$$

   is distributed as an F distribution, $F(1, n-p)$.

   (c) Show that $R^{-1}\hat{\gamma}$ are the usual OLS estimates for β

3. Consider the AudiA4 data in the R binary file `AudiA4.rda`.

```
> load("AudiA4.rda" )
> head( AudiA4)
  year price mileage distance
1 2018 28999   21991        3
2 2017 29389   40138        3
3 2014    NA   43500        3
4 2017 25863   35064        3
5 2017 25749   50934        3
6 2017 25999   44139        3
```

Just to make the numbers easier to work with divide both the mileage and the price by 1000.

```
mileage<- AudiA4$mileage/1000
price<- AudiA4$price/1000
```

Here is some R code to fit a piecewise linear function to price as a function of mileage where the break in the lines is at 30 (i.e 30,000 miles). Also a simple plot to see the fit.

```
brk<- 30
  X<- cbind( ifelse( mileage<= brk, 1, 0),
             ifelse( mileage<= brk, mileage, 0),
             ifelse( mileage > brk, 1, 0),
             ifelse( mileage > brk, mileage, 0)
             )
fit<-   lm( price~ X - 1)
#
plot( mileage, price, pch=16)
# note lines works because data is sorted by mileage ...
lines( mileage, fit$fitted.values, col="red", lwd=2)
```

(a) Note that in the full OLS fit the broken line is not continuous at 30. What is the A matrix in this case to enforce the constraint that the fit is continuous? Note in this case you can write the constraint with $\mathbf{c} = 0$.

(b) Code up the constrained estimate of $\hat{\beta_H}$ using (3.38) from Seber and Lee and add the predicted values for this fit onto a scatterplot of the data and the unconstrained OLS fit.

(c) Here is a trick to fit a broken line that is continuous without using a constraint. (This is a simple case of the the more useful B-spline models for curve fitting.)

```
brk<- 30
  X2<- cbind( 1, mileage,
              ifelse( mileage > brk,  (mileage - brk), 0)
              )
fit2<-   lm( price ~ X2 - 1)
```

Explain why this will give the same predicted values as your constrained fit above.