MATH 531 Homework 8 March 29, 2024
Hypothesis testing for the linear model

(Problem/points): (1/20), (2/20), (3/40), (4/30)

For convenience you can submit this assignments as two pdfs.
Part A: Problems 1 and 2 from Latex/handwritten and
Part B: Problems 3 and 4 as computations from R markdown.

Introduction

Throughout assume a linear model

$$\mathbf{y} = X\beta + \mathbf{e} \tag{1}$$

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

1. This is a riff on problem 5 in Exercises 4b from Seber and Lee. Note that I built in the constraint in the parameters to make these simpler.

   Suppose you just have a model with four observations: $\mathbf{y} \in \Re^4$ and $\beta \in \Re^3$ with

   $$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{bmatrix}$$

   Show the F statistic for testing $H_0 : \beta_1 = \beta_3$ vs. $H_a : \beta_1 \neq \beta_3$. Can be simplified to the form

   $$F = \frac{2(\mathbf{y}_1 - \mathbf{y}_3)^2}{(\mathbf{y}_1 + \mathbf{y}_2 + \mathbf{y}_3 + \mathbf{y}_4)^2}$$

   Hint: it is easily verified that

   $$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

   Has inverse

   $$A^{-1} = \frac{1}{4} \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

   and you might want to use R for some of the other arithmetic.

2. In class we derived, and then used, the identity that is a decomposition of a sum of squares into two parts:

   $$(\mathbf{y} - X\beta_0)^T(\mathbf{y} - X\beta_0) = (\mathbf{y} - X\hat{\beta})^T(\mathbf{y} - X\hat{\beta}) + (\hat{\beta} - \beta_0)^T(X^TX)(\hat{\beta} - \beta_0) \tag{2}$$

   Here $\hat{\beta}$ is the usual OLS estimate of the parameters for the linear model and β_0 can be any other choice for β (even the true values in the model). Note that the first term is the residual sum of squares based on OLS. Now assume that instead of the errors being independent with constant variance suppose that $\mathbf{e} \sim MN(0, \sigma^2\Omega)$ where Ω is a known matrix. In this case we want to take $\hat{\beta}$ as the generalized least squares estimates. E.g.

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

or the OLS estimate from the "star " model (see HW04).

Formulate and then prove an identity similar to (2) but using $\hat{\beta}_{GLS}$.

Hint: Work with $(\mathbf{y} - X\beta_0)^T\Omega^{-1}(\mathbf{y} - X\beta_0)$

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

   Just to make the numbers easier to work with divide both the mileage and the price by 1000. Also create two (0,1) variables for age of the cars: "old" and "new" based on the model year being less than or equal to 2016 or greater than 2016. I have also included these categorical variables in the more standard way as a single factor with "old" and "new" as the categories – but you will not need to use this in the homework. See the file `HW08.Rmd` to create the new data frame.

```
head( HW08AudiA4)
```

```
   price mileage old new age
1 39.649   3.848   0   1 new
2 43.175   3.962   0   1 new
3 43.675   5.316   0   1 new
4 40.649   5.417   0   1 new
5 42.175   5.846   0   1 new
6 45.675   6.539   0   1 new
```

Consider a model to predict `price` based on a linear function of `mileage`, and the `old` and `new` covariates. Suppose you want to test if old and new are statistically significant. There are two way to do this.

(a) `old` and `new` are linearly dependent with the constant term in the linear model. Compute the F statistic to test the hypothesis that H_0 : `old` = `new` by using a linear model where the constant (aka intercept) is omitted.

Note: To fit the linear model with `lm` and omit the constant term use the `-1` trick.

```
fit <- lm( price ~ mileage + old + new -1,  data= HW08AudiA4)
```

(b) Alternatively omit `old` from your model but include the constant term and compute the F statistic for the hypothesis that H_0 : `new` = 0 . Verify that this F statistic is *identical* to the the one above.

(c) Show how to obtain this F statistic for this second model using the `lm` function in R.

(d) The denominator of the F statistic is $\hat{\sigma}^2$ for the full model. Suppose you just assume that that this is a good estimate and it is basically the same as σ^2, the true variance. Explain why the distribution of your F statistic under H_0 is now approximated by a chi squared distribution. Find the 95% quantile for this chi -squared distribution and the F. How different are they?

4. For the Audi data example one might consider two simple models for price to improve on just using mileage as a linear predictor. For clarity I have also written these models in the R `lm` formula syntax.

   Model A:

   ```
   price = cubic polynomial in mileage
   ```

   ```
     fitA<- lm( price ~ mileage + I(mileage^2) +I(mileage^3) , data = HW08AudiA4)
   ```

   Model B:

   ```
   price = constant + mileage + old
   ```

   ```
    fitB<- lm( price ~ mileage + old, data = HW08AudiA4)
   ```

   Suppose you fit these two models – both reasonable choices – and find the residual sums of squares, RSS_A and RSS_B

   Not thinking too much about correctness you find the F statistic:

   $$F = \frac{(RSS_B - RSS_A)/(4-3)}{RSS_B/(n-4)}$$

   (a) Explain why F above is not distributed according to an F distribution (!).

   (b) How should you modify Model A so that F has an F distribution. Based on your modification, compute the F statistic and test at the .05 α level. Which model is preferred?

   (c) Compare the original Model A and Model B directly using a Monte Carlo version of 10 fold cross-validation. Refer to the R code in `HW08.Rmd` to get started. Based on this computation which model is preferred. (In the R code I suggest how to find the root mean squared prediction errors but it is up to you to decide what to do with them.)