LINEAR MODELS AND GLM: ASSIGNMENT 2

Exercise 1. *We have defined multivariate Normality as follows:* $\boldsymbol{x} = \boldsymbol{\mu} + \mathbf{A}\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}')$ *if components of* $\boldsymbol{z}$ *are i.i.d. standard Normal.*

(a) *(1 point) Show that if* $\boldsymbol{x}$ *follows a multivariate Normal distribution by this definition, then* $\boldsymbol{\ell}'\boldsymbol{x}$ *is univariate normal for all* $\boldsymbol{\ell}$*.*

(b) *(3 points) Show the converse, that is, if* $\boldsymbol{x}$ *is a random vector with the property that* $\boldsymbol{\ell}'\boldsymbol{x}$ *is univariate normal for all* $\boldsymbol{\ell}$*, then* $\boldsymbol{x}$ *has a multivariate Normal distribution.*

You may assume $\boldsymbol{\mu} = \mathbf{0}$*.*

Exercise 2. *(2 points) Prove that for a real symmetric matrix, eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Exercise 3. *(2 points) For a linear model* $(\boldsymbol{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$*, write* $H_0 : \boldsymbol{\beta} \in \mathcal{C}(\mathbf{B})$ *as* $H_0 : \mathbf{H}'\boldsymbol{\beta} = \mathbf{0}$ *for some* $\mathbf{H}'$*.*

Exercise 4. *The theory of likelihood ratio tests tells us that under certain regularity conditions,* $-2\log\Lambda(\boldsymbol{y})$ *has an asymptotic* χ^2_k *distribution under the null hypothesis, where* $\Lambda(\boldsymbol{y})$ *is the likelihood ratio and* k *is the difference in the degrees of freedom of the full and restricted models.*

(a) *(4 points) Show that this result holds in the linear model case (without invoking the general LRT result).*

(b) *(1 point) Is the result exact for finite sample size?*

Your solution does not have to be mathematically rigorous.

Exercise 5. *Let* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$*. Consider the problem of estimating* σ^2*.*

(a) *(1 point) Is the maximum likelihood estimate of* σ^2 *unbiased? If not, can it be made unbiased by multiplying it with a scalar?*

(b) *(4 points) What is the MSE of the MLE? How does it compare to the MSE of the unbiased estimate?*

The mean squared error of an estimator $\hat{\theta}$ *of* θ *is defined as* $MSE(\hat{\theta}) := E((\hat{\theta} - \theta)^2)$

Exercise 6. *Consider the* `VADeaths` *dataset available in R, which is a two-way classification with one observation per cell (see* `?VADeaths` *for details).*

```
> names(dimnames(VADeaths)) <- c("AgeGroup", "PopGroup")
> VADeaths
        PopGroup
AgeGroup Rural Male Rural Female Urban Male Urban Female
   50-54       11.7          8.7       15.4          8.4
   55-59       18.1         11.7       24.3         13.6
   60-64       26.9         20.3       37.0         19.3
   65-69       41.0         30.9       54.6         35.1
   70-74       66.0         54.3       71.1         50.0
```

Convert it into a data frame as follows:

```
> VADeaths.df <- as.data.frame(as.table(VADeaths), responseName = "DeathRate")
```

Figure 1 gives a plot of the data.

Date: 12 February, 2010.

```
> library(lattice)
> dotplot(DeathRate ~ reorder(PopGroup, DeathRate), data = VADeaths.df,
          groups = AgeGroup, type = c("p", "a"), auto.key = list(columns = 3))
```

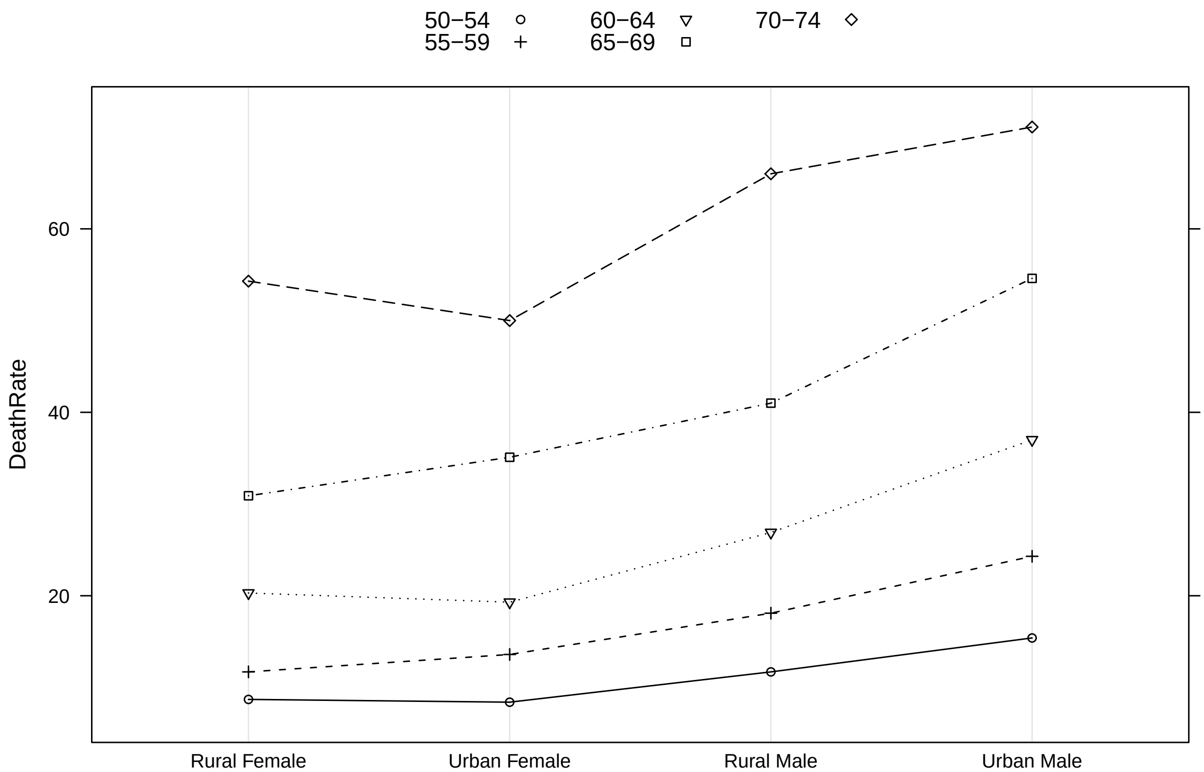

FIGURE 1. A plot of the `VADeaths` data.

(a) *(2 points) Compute the ANOVA table for a two-way classification without interaction (either by hand, or using* R*). Which factors are significant?*
(b) *(3 points) Create a plot corresponding to Figure 1, but replace the observed values with the fitted values. This is roughly what an additive model would have looked like (except for i.i.d. errors).*
(c) *(3 points) Does the plot suggest the presence of an interaction? How can we test for the interaction? How else can we improve the fit?*

Exercise 7. *(4 points) Consider the following linear model:*

$$y_i = \alpha + \beta f(x_{i,n}) + \varepsilon_i, \quad 1 \le i \le n$$

where $f\colon [0,1] \to \mathbb{R}$ *is a function,* $x_{i,n} := i/n, 1 \le i \le n$ *and* $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ *are uncorrelated with mean zero and variance* σ^2*.*

(a) *Find a sufficient condition for both* α *and* β *to be estimable for every* $n \ge 2$*.*
(b) *Find LSEs of* α *and* β *when they are estimable.*
(c) *Find the residual sum of squares for this model.*

Exercise 8. *(0 points) Verify the following results (assuming all relevant inverses exist):*

$$\begin{aligned}\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}^{-1} &= \begin{bmatrix}\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}\mathbf{E}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{E}^{-1}\\ -\mathbf{E}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{E}^{-1}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{F}^{-1} & -\mathbf{F}^{-1}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{F}^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}\mathbf{F}^{-1}\mathbf{B}\mathbf{D}^{-1}\end{bmatrix}\end{aligned}$$

where $\mathbf{E} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ *and* $\mathbf{F} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$.

Exercise 9. *Consider the problem of overfitting in a linear model. Suppose that the observations* $\boldsymbol{y}$ *is correctly modeled by* $(\boldsymbol{y}, \mathbf{X}_1\boldsymbol{\beta}_1, \sigma^2\mathbf{I})$, *but one has fit the model* $(\boldsymbol{y}, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \sigma^2\mathbf{I})$ *with some extra terms given by* $\mathbf{X}_2$. *Assume that both* $\mathbf{X}_1$ *and* $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$ *are of full column rank.*

(a) *(2 points) Show that the least squares estimates of* $\boldsymbol{\beta}_1$ *and* σ^2 *obtained from the overfitted model are unbiased.*

(b) *(4 points) Show that the variance-covariance matrix of the least square estimate of* $\boldsymbol{\beta}_1$ *obtained from the overfitted model is larger than that of the estimate obtained from the correct model (larger in the sense that the difference is n.n.d.).*