COLUMBIA UNIVERSITY
Statistical Inference and Modeling
Graded Homework 3

Prof. Marco Avella
TA: Ian Kinsella
Fall 2019

You are free to present your solutions to the exercises below in *groups of at most three*. This homework is due on April 9. Bonne chance!

Exercise 1

Follow the instructions of Exercise 6 of Chapter 8 in:
`http://statwww.epfl.ch/davison/SM/SM_Practicals.1.pdf`.

Exercise 2

Let us start by considering an example about modeling the relationship between stopping distance of a car and its speed at the moment that the driver is signalled to stop. Data on this are provided in `R` data frame `cars` available in the library of the package `mgcv`. For this data we ask you to make the following assumptions:

a). A car's kinetic energy is proportional to the square of its speed. In addition, the brakes dissipate the kinetic energy at a constant rate per unit distance traveled. This assumption implies that we would expect the distance traveled between the application of the breaks and coming to a complete stop to be proportional to the square of the vehicle's speed when the brakes were applied.

b). Drivers have a fixed "reaction time" (time between receiving the signal to stop and actually applying the brakes). The implication is that this "reaction time" should contribute to an increase in stopping distance proportional to the vehicle's speed at the time when the signal was sent.

c). The velocity of the car between the time when the braking signal is sent and the brakes are applied does not change (ie. the speed used for estimating the contribution to stopping distance from the driver's "reaction time" should be the same as the speed used for estimating the contribution from the physics of braking).

Use these assumptions to answer the following questions.

1. Fit a model to the data in `cars` of the form

$$\texttt{dist}_i = \beta_0 + \beta_1 \texttt{speed}_i + \beta_2 \texttt{speed}_i^2 + \epsilon_i$$

Using this "complete" model as a starting point, select the most appropriate model for the data using both AIC and hypothesis testing methods.

2. Use the parameter estimates from the selected model to provide an estimate of the fixed "reaction time" for the drivers in this experiment (Hint: you will need to use the fact that there are 5280 feet in a mile).

Let's now turn to the question of how to calculate estimates and associated quantities for the linear model, $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$, where the ϵ_i are i.i.d. $N(0, \sigma^2)$. This question relates to the material discussed in Chapter 7 of Simon Wood's book "Core Staistics" [1] The goal of the exercise is to code your own linear regression function using the QR decomposition of a matrix, and review some basic distributional theory for the linear model in the process.

3. Write an R function which will take a vector of response variables, y, and a model matrix, $\mathbf{X}$, as arguments, and compute the least squares estimates of associated parameters, $\boldsymbol{\beta}$, based on QR decomposition of $\mathbf{X}$. Note that you can form the QR decomposition of $\mathbf{X}$ in R as follows

   ```
   qrx <- qr(X)  ## returns a QR decomposition object
   Q <- qr.Q(qrx,complete=TRUE)  ## extract Q
   R <- qr.R(qrx)  ## extract R
   ```

4. Test your function by using it to estimate the parameters of the "complete" model from part 1 of this exerise. Note that you can use:

   ```
   X <- model.matrix(dist ~speed + I(speed^2),cars)
   ```

   to generate suitable model matrix. Validate your answers against those produced by the `lm` function.

5. Extend your function to also return the estimated standard errors of the parameter estimators, and the estimated residual variance. Again, check your answers against what `lm` produces, using the cars model. Note that `solve(R)` or more efficiently `backsolve(R,diag(ncol(R)))` will produce the inverse of an upper triangular matrix $\mathbf{R}$.

6. Use the R function `pt` to produce p-values for testing the null hypothesis that each β_i is zero in the "complete" model (against a two sided alternative). Once again, check your answers against a summary of an equivalent `lm` fit.

Exercise 3

The dataset (`CpsWages.txt`) consists of a random sample of 534 persons (no missing data) from the Current Population Survey, with information on wages (`wage`, in dollars per hour) and other characteristics of the workers, including

[1] The pdf of the book is available in Simon Wood's webpage `https://people.maths.bris.ac.uk/~sw15190/core-statistics-nup.pdf`. The pages relevant for this exercise can be found in the Homework 3 folder on courseworks.

- `sex` coded 1=female and 0=male,
- `age` in years,
- `race` coded 1=other, 2=hispanic and 3=white,
- `marr`: marital status coded 1=married and 0=unmarried,
- `education`: number of years of education,
- `experience`: number of years of work experience,
- `occupation`: occupational status coded 1=management, 2=sales, 3=clerical, 4=service, 5=professional and 6=other,
- `sector`: work sector coded 0=other, 1=manufacturing and 2=construction,
- `south`: region of residence coded 1=lives in the South and 0=lives in the North,
- `union`: union membership coded 1=union member and 0=not a member.

We wish to determine whether wages are related to these characteristics and specifically whether there is a gender gap in wages[2].

1. Suggest a model specification for this dataset. Why is it not a good idea to include at the same time `age`, `education` and `experience` ?
2. Fit the proposed model and perform diagnostic plots. Do you observe any departure from the hypotheses?
3. Look at parameters estimates. Are all the parameters significant? How would you test whether the `sector` variable is significant or not?
4. Use a global approach to select a simpler model.
5. Estimate the final model and check its features.
6. Would your conclusions be altered if you remove the 171st and 200th observations?

Exercise 4

Health surveys are often used to study health care expenditures and especially the impact of income and insurance level on the frequency of health care services

[2] Reference: Berndt, ER. The Practice of Econometrics. 1991. NY:Addison-Wesley.

use. Here, we are interested to see if a high insurance level (with high premium) "causes" individuals to consume more health care services.
Import the `docvisits.asc` data file, which is part of the 1977-1978 Australian Health Survey and contains information on the number of consultations with a doctor or specialist, denoted `dvisits`, in the two-weeks period before an interview. These counts represent our response variable and we consider the following thirteen explanatory variables:

- `sex` coded 1 for female,
- `age` in years/100,
- `agesq`, $(\text{age})^2/1000$,
- `income`, annual income/1000,
- `illness`, number of illnesses in past two weeks, with five or more coded as 5,
- `actdays`, number of reduced activity days in past two weeks due to illness or injury,
- `hscore`, general health questionnaire score, with high scores indicating bad health,
- `chcond1` coded 1 if chronic condition(s) but not limited in activity,
- `chcond2` coded 1 if chronic condition(s) and limited in activity,
- `levyplus`, `freepoor` and `freerepa`, three dummy variables for levels of health insurance, where "levyplus" represents a higher level of insurance cover while "freepoor" and "freerepa" are basic levels of state-provided insurances.

1. Fit a Poisson GLM to this data. Does it fit well this dataset? Do you see signs of over dispersion or underdispersion?

2. One possible way to account for the large number of zeros in the responses is to formulate the model

$$\mathbb{P}(Y_i = y_i | \mathbf{X}_i = \mathbf{x}_i) = \begin{cases} 1 - \pi(\mathbf{x}_i), & y_i = 0 \\ \pi(\mathbf{x}_i) \frac{\exp\{-\lambda(\mathbf{x}_i)\}\lambda(\mathbf{x}_i)^{y_i}}{y_i![1-\exp\{-\lambda(\mathbf{x}_i)\}]}, & y_i = 1, 2, \dots, \end{cases}$$

   where $\pi(\cdot)$ and $\lambda(\cdot)$ are two unknown functions. Interpret this model and write down its log-likelihood function.

3. Propose a GLM formulation for estimating $\pi(\mathbf{x}_i)$ and $\lambda(\mathbf{x}_i)$ as two functions of the form $\pi(\mathbf{x}_i^T\boldsymbol{\beta})$ and $\lambda(\mathbf{x}_i^T\boldsymbol{\gamma})$ respectively. Write down the corresponding log-likelihood equations.

4. Use the previous point to derive the asymptotic distribution of t $(\hat{\boldsymbol{\beta}}_{ML}^T, \hat{\boldsymbol{\gamma}}_{ML}^T)^T$ and comment the result

5. Source the `truncpoisson.R` to create the `truncpoisson` GLM family object. Fit manually the two steps of the above Poisson hurdle model to the data by using the `family=truncpoisson` argument in the `glm` command. Compare your results to the output from the `hurdle` command of the `pscl` package.

6. Did the model from last point provide a better fit? Do high insurance levels make the number of consultations increase?

Exercise 5 (Optional bonus question)

Consider the linear model

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \varepsilon_i, \ i = 1, \ldots, n$$

where ε_i, are i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$, $\text{var}(\varepsilon_i) = \sigma^2$ and some *fixed* covariates $\mathbf{X}_i \in \mathbb{R}^p$. The goal of this exercise is to get some insights into the theoretical properties of model selection criteria such as the C_p, AIC and BIC

1. Consider the prediction error

$$\Gamma = \frac{1}{\sigma^2}\mathbb{E}\Big[\sum_{i=1}^{n}(\hat{Y}_i - \mathbb{E}[Y_i])^2\Big]$$

where $\hat{Y}_i$ are some predicted values of Y_i (not necessarily least squares). Show that the

$$\sigma^2\Gamma = \mathbb{E}\Big[\text{RSS}(\hat{\mathbf{Y}}) - \sum_{i=1}^{n}\text{var}(\hat{\varepsilon}_i) + \sum_{i=1}^{n}\text{var}(\delta_i)\Big],$$

where $\text{RSS}(\hat{\mathbf{Y}}) = \sum_i (Y_i - \hat{Y}_i)^2$, $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ and $\delta_i = \hat{Y}_i - \mathbb{E}[Y_i]$

2. Suppose now that we have a linear predictor of the form $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$, where $\mathbf{S}$ is a $n \times n$ matrix. Show that in this case

$$C = \frac{1}{\sigma^2}\text{RSS}(\hat{\mathbf{Y}}) + 2\text{tr}(\mathbf{S}) - n$$

is an unbiased estimator of Γ.

3. Use the previous results to show that if we take the least squares predictor $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, then
$$C_p = \frac{1}{\sigma^2}\mathrm{RSS}(\hat{\mathbf{Y}}) + 2p - n$$
is an unbiased estimator of the prediction error. Interpret the AIC in light of its connection to the C_p.

4. Suppose now that the true parameter $\boldsymbol{\beta}$ has only q non-zero entries. Show that when the sample size tends to infinity, the probability that the AIC chooses a larger model than the true model M_q is strictly positive. For this, consider a model that adds on parameter to M_q and give an approximation to the probability of the event $\mathrm{AIC}(\hat{\boldsymbol{\beta}}_{q+1}) < \mathrm{AIC}(\hat{\boldsymbol{\beta}}_q)$. This shows that the AIC will tend to overfit.

5. Show that the BIC corrects the inconsistency in model selection pointed out in the previous question. In particular, show that when $n \to \infty$, the probability to choose larger models than the true model M_q goes to 0.