COLUMBIA UNIVERSITY — Prof. Marco Avella
Statistical Inference and Modeling — TA: Casey Bradshaw, Leon Fernandes
Graded Homework 3 — Spring 2020

You are free to present your solutions to the exercises below in *groups of at most three*. This homework is due on April 16 at 11pm. Bonne chance!

Exercise 1

The data **milk.txt** reports monthly milk production (pounds per cow) from January 1962 to December 1975.

1. Fit a linear model to the data and comment the estimated coefficients. Do you see some structure in the residuals?
2. Plot the correlogram and partial correlogram of the residuals obtained in the previous point and comment them.
3. Try fitting an AR(1) to the data. Does it provides a better model? what about an AR(2), MA(1) or MA(2) model?
4. Try different ARMA models and present two models that best fit the data in your analysis.

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.

[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.

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

- `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?

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

3. Look at parameters estimates. Are all the parameters significant? How would you test whether the `sector` variable is significant or not?

4. Use some model selection criteria to find 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

Consider the linear model

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

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

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

   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.

Exercise 5 (Optional bonus question)

The dataset `pollution` (located in the `R` package SMPracticals; original data courtesy of McDonald and Schwing, 1973) contains data on weather (variables 1-3, 15), socio-economic factors (variables 4-11), and pollution (variables 12-14) for 60 Standard Metropolitan Statistical Areas in the USA. The response (variable 16) is the age-adjusted mortality rate from all causes, expressed as deaths per 100,000 persons.

1. For an initial look at the data, generate the following plots:

   ```
   pairs(pollution)
   pairs(pollution[,c(1:3,15:16)]) # association of mortality with weather
   pairs(pollution[,c(4:11,16)])   # and social factors
   pairs(pollution[,c(12:14,16)])  # and pollution measures
   ```

   Examine these plots carefully, and comment. Are there outliers? Should covariates and/or the response be transformed? What difficulties might arise in accounting for the effect of air pollution on mortality?

   Note- this is open-ended and has many possible correct answers; focus on explaining your observations and your reasoning

2. Try using `step` to eliminate weather and social variables from the regression:

   ```
   fit <- step(glm(mort˜.-hc-nox-so,data=pollution))
   boxcox(fit)
   plot.glm.diag(fit) # model adequate?
   fit <- update(fit,log(mort)˜.) # try log transform of response
   plot.glm.diag(fit) # model adequate?
   ```

   Should all variables be included? Try various models, choose one or perhaps a few that you think are similarly adequate, give careful interpretations of the covariate effects, and discuss their plausibility. Check the adequacy of your model (*for instance, this could include looking at metrics such as R squared or AIC, as well as considering some of the standard linear model assumptions and exploring whether they seem to be satisfied for your model*).

3. For an initial assessment of the relation between the pollution variables and mortality, after adjustment for the other variables, we use `resid` and `lm` to make added variable plots:

   ```
   pairs(resid(lm(cbind(log(mort),hc,nox,so)~.,data=pollution)))
   ```

   The top line of this scatterplot matrix contains the added variable plots for log mortality and the pollution variables. What difficulties do you foresee for regression on all three pollution variables? Are outliers present? Try adding in these variables, or suitable transformations of them, to your chosen best model (or models) from above, and discuss the interpretation and fit of the various models.

   Again, there is no single correct approach to this exercise; just be sure to clearly explain your reasoning for the modeling choices you make.

4. One possible approach to dealing with some of the problems above would be to use ridge regression. Try using the `ridge.lm` function, for example by:

   ```
   rfit <- lm.ridge(mort~.-hc-nox,data=pollution,lambda=seq(0,20,0.01))
   plot(rfit)
   select(rfit)
   ```

   Discuss the interpretability of the resulting parameter estimates.

5. Try using the functions `lqs` in `library(lqs)` for least trimmed squares regression, and `rlm` in `library(MASS)` for robust M-estimation, and see if your conclusions change (*For instance, do you still identify the same variables as important for modelling the outcome? Do these new models result in a better fit? Worse?*) .