MVE190-MSG500 Linear Statistical Models, 25/04/2019 Solutions

Question 1

(i) Clearly gender is a categorical variable. Here I set baseline gender=male. We have the linear model $E(income|gender) = \beta_0 + \beta_1 \cdot X_{female}$ with dummy variable $X_{female} = 1$ for female subjects and 0 otherwise, β_0 the expected income for a male employee and $\beta_0+\beta_1$ the expected income for a female employee. Therefore β_1 represents the difference in income between genders. It is therefore sufficient to check the hypothesis $H_0 : \beta_1 = 0$ vs the alternative hypothesis $H_1 : \beta_1 \neq 0$. We can use a t-test for this task at some significance level α. If we turn out failing to accept H_0 then we conclude that there is a significant difference (at the specified level α).
Notice: the following hasn't been treated in the course. But for your personal knowledge, you can also test H_0 vs the "one-tailed" hypothesis $H_1 : \beta_1 < 0$ if it is clear from the data that a female employee can't earn more than a male. Lookup for "one-tailed" and "two-tailed" tests (the latter is the type we have treated in the course, where $H_1 : \beta_1 \neq 0$).

(ii) Here I choose the following baselines: baseline gender=male and baseline experience=junior. The model with main effects only is $E(income|gender, experience) = \beta_0 + \beta_1 \cdot X_{female} + \beta_2 \cdot X_{intermediate} + \beta_3 \cdot X_{senior}$. We have that β_0 is the expected income for a junior male (all introduced dummy variables are zero in this case). $\beta_0 + \beta_1 + \beta_3$ is the expected income for a senior female.

(iii) This means that, *marginally* with respect to gender (i.e. regardless of the gender), employees with intermediate experience and senior experience are earning on average more than junior employees. Also, marginally with respect to the experience level, there is no difference in income between different genders.

(iv) The intercept says that, for lifevents=0 (which is a positive thing), the expected mental impairment index equals 23.3. Then, for each additional unit in the lifeevent score, we expect an increase of about 0.09 in the mental impairment index.

Then we have

$$T = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.08983}{0.03633} = 2.473.$$

The t-quantile is $t_{0.975,40-2} = 2.024$ and $|T| > 2.024$ hence we fail to accept $H_0 : \beta_1 = 0$. There seem to be a positive association between the response and the covariate. Overall the model seems too simplistic, the R-squared is pretty low, meaning we fail at explaining most of variability (unsurprisingly given the complexity of the studied phenomenon).

Question 2

First of all, this question was not made up. Rather astonishingly, that research has been performed and (unfortunately) has even been published on the most prestigious medical journal in

the world[1], which says a lot...anyhow:

(i) Quite simply, some of the covariates that is correlated with chocolate consumption (gdp or gdponrd) is perhaps affecting a further variable (which is not necessarily in the dataset) and the latter really influences the number of Nobel prizes. It makes sense to assume that countries having a large economy and investing a lot on research can produce better education which, in turn, might help brilliant thinkers to emerge. Or might attract promising scientists from abroad, who then take the citizenship of the country where they performed research, and there you go. If such countries also are large consumers of chocolate, this creates a "circular" effect that can easily confuse the naive reader. Notice that `prizes` has non negligible positive correlation with gdp and gdponrd, and that chocolate consumption is positive correlated with the latter two.

(ii) We have p=8 parameters (including the intercept) and n=23 observations.

	dof	SS	MS=SS/dof	F
Regression	8-1	1579.17	225.59	5.12
Error	23-8	661	44.07	
Total	23-1	2240.17		

SS(Regression) is easy to find by difference, since SS(Tot) = SS(Regression)+SS(Error).

Then F= MS(Regression)/MS(Error) = 5.12.

This F is the Global-F-test, named this way because it tests $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, that is it tests that all the parameters (but the intercept) are null, against the hypothesis that $H_1 : \beta_j \neq 0$ for at least one $j = 1, ..., p-1$.

Under H_0 true, we fail to accept H_0 if $F > F_{p-1,n-p,1-\alpha}$. Let's take $\alpha = 0.05$ and we have $F_{7,15,0.95} = 2.70$, hence we fail to accept H_0. There is at least a covariate having a significant effect on the response.

(iii) We can compute both R^2 and the adjusted R^2. We have $R^2 = SS(Regr)/SS(Tot) = 0.705$ or 70.5%. The adjusted R^2 is $1 - MS(Error)/MS(tot) = 1 - (1-R^2)(n-1)/(n-p-1) = 1 - (1-0.705)22/14 = 0.536$ or 53.6%.

(iv) Having a good R^2 and a test that tells us that "at least a parameter is significant", is absolutely not enough to stop our enquiry. In fact the following things can happen (individually or simultaneously): (a) in some specific cases the presence of some outlier can "pull" the regression fit, thus inflating the value of the R^2; (b) We could have a spurious association, that is an association that exists numerically but not in reality, such as what we described in (i). This would make the global-F test happy, but not make us happy. (c) Some covariates could be unnecessary and should be removed, and here the battery of variable selection tools can be employed. (d) 5 model assumptions should be checked and the influence of outliers assessed, to see if some specific outliers are influencing the fit. (e)...

Question 3

(i) Potentially influential observations are those cases in the dataset that have high leverage values h_i. The leverage h_i is the i-th entry on the diagonal of the matrix $X(X'X)^{-1}X'$ and geometrically it represents a distance from the point $\bar{x} = (\bar{x}_1, ..., \bar{x}_{p-1})$, the point having coordinates given by the sample means of the columns of X. Therefore high leverage values identify outliers in the space spanned by the columns of X. These are cases to keep an

[1] "Chocolate Consumption, Cognitive Function, and Nobel Laureates", New England Journal of Medicine, 2012, 367:1562-1564.

eye on as they can (though not necessarily, hence the "potential") affect the regression coefficient estimates, the RSS (hence s^2) and hence the R^2, and their own fitted values as $\hat{y}_i = \cdots + h_i y_i + \cdots$. These cases can be spotted by plotting their leverage values vs i. Leverage values larger than 2p/n are considered potentially influential, though more in general it is best to keep an eye on those cases that have dramatically large leverage values compared to other cases.

(ii) we have $p-1$ covariates at disposal, we want to fit all $M = 2^{p-1}$ models while avoiding under/overfitting. This way we have a variable selection procedure.

We divide the data into training and testing datasets. We fit each of the M models on the training dataset, obtaining corresponding $\hat{\beta}(m)$ for $m = 1, ..., M$. Then construct predictions $\hat{y}_i(m) = x_i^{test}(m)\hat{\beta}(m)$, where the $x_i^{test}(m)$ are covariates for model m based on testing data. We compare those predictions with the responses y_i^{test} from the testing data, via the pMSE(m):

$$pMSE(m) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i^{test} - \hat{y}_i(m))^2$$

where n_{test} is the size of the testing data.

We iterate the above for all $m = 1, ..., M$. Finally we plot pMSE(m) vs m and choose as best model the one returning the smallest pMSE.

The procedure is useful for model selection as it considers the problem of overfitting/underfitting the data: the parameter estimates are obtained on a fraction of the data, and this way we can monitor the ability for each model to generalize to unseen observations (test data) that are not involved in the estimation of the coefficients. The pMSE indeed measures the quality of the model fit in predicting unseen data. See also the answer to the next question.

(iii) Figure 3 is the output of the all-subsets-regression and the six circles for models of size 2 are due to the 6 pMSEs : there are six models of size 2, that is models with an intercept and one covariate. Since we have a total of 6 possible covariates in the dataset, the result follows. Similarly, we have only one model of size 7 (intercept plus all six covariates).

As such, the purpose of the figure is, as given in the previous question, to point to a model having the best ability to predict unseen data without underfitting/overfitting (models that underfit data are those on the left of the smallest pMSE, those overfitting are the ones on the right of the minimal pMSE). The best model here has size 2. The figure itself does not tell exactly which model it is, but the other R output shows that this is a model with intercept and the fourth covariate.

Question 4

(i) we require responses to be all distributed according to the same member of the exponential family (in the course we only considered iid responses). If this is the case, then the following also follows: (a) we define a linear predictor $\eta = X\beta$; (b) we have a monotonous and differentiable (hence invertible) function $g(\cdot)$ such that $g(\mu) = \eta$, for $\mu = E(Y)$. This g is called link-function as it links the linear prediction to a transformation of the expected response. $\phi > 0$ is called dispersion parameter, the w_i are "weights" for each observation Y_i.

(ii) We use maximum likelihood. Therefore we first need to construct the likelihood function. This means that for regression parameters β we wish to construct $L(\beta) = p(Y_1, ..., Y_n|\beta)$ for data of size n and $p(\cdot)$ the joint density of the data. In the course we have worked under the assumption that the Y_i are independent, hence $L(\beta) = p(Y_1, ..., Y_n|\beta) = \prod_{i=1}^n p(Y_i|\beta)$, and $p(Y_i|\beta)$ is a known density or probability mass function (it's known because we decided

to pick a specific member of the exponential family). Then we can use Newton-Raphson to efficiently find a the $\hat{\beta}$ maximising the likelihood, or equivalently minimizing the negative-loglikelihood

$$\hat{\beta} = \arg\min_{\beta} -\log L(\beta).$$

This is the maximum-likelihood estimate and is asymptotically Gaussian (as $n \to \infty$) and unbiased

$$\hat{\beta} \sim_{n\to\infty} N(\beta, H^{-1})$$

where H is the Hessian matrix, the matrix of the second derivatives of the negative log-likelihood, evaluated at $\hat{\beta}$.

(iii) For $Y_i \sim Po(\mu_i)$ we have $\ln \mu_i = \beta_0 + \beta_1 x_{0,i}$, where $\mu_i = E(Y_i|x_0)$. Then $\hat{\mu}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{0,i}}$. The required confidence interval is asymptotically given by $I_{\mu_i} = e^{I_{\beta_0+\beta_1 x_{0,i}}} = = e^{(\hat{\beta}_0+\hat{\beta}_1 x_{0,i} \pm z_{\alpha/2} \cdot S.E.(\hat{\beta}_0+\hat{\beta}_1 x_{0,i}))}$, where z_α is the α- quantile of the standard Gaussian distribution. Then we have $S.E.(\hat{\beta}_0 + \hat{\beta}_1 x_{0,i}) = \sqrt{S.E.^2(\hat{\beta}_0) + x_{0,i}^2 S.E.^2(\hat{\beta}_1) + 2x_{0,i} Cov(\hat{\beta}_0, \hat{\beta}_1)}$, where the standard errors and other elements of the covariance matrix of the regression estimates are obtained from the entries of H^{-1} in (ii), the asymptotic covariance matrix of the estimates.