|  |  |
| --- | --- |
| **ASSIGNMENT COVER SHEET** | Research School of Finance, Actuarial Studies and Statistics ANU College of Business and Economics  Australian National University Canberra ACT 0200 Australia www.anu.edu.au |
| Submission and assessment is anonymous where appropriate and possible. Please do not write your name on this coversheet.  This coversheet must be attached to the front of your assessment when submitted in hard copy.  All assessment items submitted in hard copy are due by 12:00 pm. |

|  |  |  |  |
| --- | --- | --- | --- |
| Student ID |  | | |
| Course Code |  | | |
| Course Name |  | | |
| Assignment number |  | | |
| Assignment Topic |  | | |
| Lecturer |  | | |
| Tutor |  | | |
| Tutorial (day and time) |  | | |
| Word count |  | Due Date: Oct 18 |  |
| Date Submitted |  | Extension Granted |  |

I declare that this work:

* upholds the principles of academic integrity, as defined in the ANU Policy[: Code of Practice for Student Academic Integrity](https://hkxprd0610.outlook.com/owa/redir.aspx?C=pkUS4AqeVkC0OHXUsRYzk8JcJE65y9AI4r3Mqfll_bLO9DXo_dFgmbuC6N5TOcnRwCb-AmVT460.&URL=https%3a%2f%2fpolicies.anu.edu.au%2fppl%2fdocument%2fANUP_000392);
* is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the course outline and/or Wattle site;
* is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
* gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
* in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

|  |  |
| --- | --- |
| **Initials** |  |

|  |  |
| --- | --- |
| **ANSWER SHEET** | Research School of Finance, Actuarial Studies and Statistics ANU College of Business and Economics  Australian National University Canberra ACT 0200 Australia www.anu.edu.au |

Please input your answers of the questions in Assignment 2 on the right side of the table.

|  |  |
| --- | --- |
| **Question 1 (2.5 points)** | |
| a) | **Please paste the plot in the following:** |
| b) | Based on the Cp statistics, the variables **Age, Sex, MaritalStatus, EdCode, INortheast, ISouth, INotMetropolitan, IFedGov, IStateGov, Sex:MaritalStatus** are necessary to predict the logarithm of WeeklyEarnings. |
| c) | > vif(fit)  Age IMale IMarried EdCode IMidwest INortheast  1.085647 2.486618 2.196868 1.063858 1.583150 1.522163  ISouth IMetropolitan INotMetropolitan IFedGov ILocalGov IStateGov  1.617833 23.222433 23.303701 1.021278 1.039201 1.047618  IMale:IMarried  3.845590  We noticed that the VIFs of **IMetropolitan** and **INotMetropolitan** are both greater than 10, which is the “rule of thumb” cut-off value for VIF, thus indicating the multicolinearity exists. |
| d) | If we drop the variable **IMetropolitan**, the deviance of the model is 3043.996.  If we drop the variable **INotMetropolitan**, the deviance of the model is 3045.14.  Therefore, we should drop **IMetropolitan**, i.e. we use all the original variables except **IMetropolitan**, and the VIFs of the remaining variables are all below 10. |
| e) | # ------------------------------- Q1 -------------------------------  library(Sleuth3)  library(leaps)  library(car)  data <- ex1225  attach(data)  IMale <- ifelse(Sex=="Male",1,0)  IMarried <- ifelse(MaritalStatus=="Married",1,0)  IMidwest=ifelse(Region=="Midwest",1,0)  INortheast=ifelse(Region=="Northeast",1,0)  ISouth=ifelse(Region=="South",1,0)  IMetropolitan=ifelse(MetropolitanStatus=="Metropolitan",1,0)  INotMetropolitan=ifelse(MetropolitanStatus=="Not Metropolitan",1,0)  IFedGov <- ifelse(JobClass=="FedGov",1,0)  ILocalGov <- ifelse(JobClass=="LocalGov",1,0)  IStateGov <- ifelse(JobClass=="StateGov",1,0)  # ------------------------------- (a) -------------------------------  Y <- log(WeeklyEarnings)  X <- cbind(Age,IMale,IMarried,EdCode,  IMidwest,INortheast,ISouth,  IMetropolitan,INotMetropolitan,  IFedGov,IStateGov,ILocalGov,IMale\*IMarried)  Cpsel <- leaps(X,Y,method="Cp",nbest=2)  plot(Cpsel$size,Cpsel$Cp,main="Cp plot",xlab="size",ylab="Cp")  # ------------------------------- (b) -------------------------------  cbind(Cpsel$which, Cpsel$size, Cpsel$Cp)  # ------------------------------- (c) -------------------------------  fit=lm(Y~Age+IMale+IMarried+EdCode+IMidwest+INortheast+ISouth+  IMetropolitan+INotMetropolitan+IFedGov+ILocalGov+IStateGov+  IMale\*IMarried)  vif(fit)  # ------------------------------- (d) -------------------------------  # dropping IMetropolitan  X1 <- X[,-8]  fit1 <- lm(Y~.,data=as.data.frame(X1))  deviance(fit1) # 3043.996  # dropping INotMetropolitan  X2 <- X[,-9]  fit2 <- lm(Y~.,data=as.data.frame(X2))  deviance(fit2) # 3045.14  vif(fit1) # problem solved  detach(data) |
| **Question 2 (2.0 points)** | |
| a) | Yes, we can use “null deviance” and “residual deviance” to construct a drop-in-deviance chi-square test.  **Null hypothesis**: all the parameters s are zero.  **Alternative hypothesis**: at least one of the parameter  is not zero. |
| b) | The test statistic is **79.802**. According to the drop-in-chisq test, we reject null hypothesis in favor of alternative hypothesis. So at least one of the parameter is not zero. |
| c) | According to forward selection based on BIC, the variables we should keep in the model are **TL**, **HL**, **WT** and **KL.** |
| d) | # ------------------------------- Q2 -------------------------------  sparrow <- ex2016  head(sparrow)  sparrow$AG <- factor(sparrow$AG)  # ------------------------------- (a) -------------------------------  sparrow.glm <- glm(Status~AG\*(TL+AE+WT+BH+HL+FL+TT+SK+KL),  family=binomial(link=logit),  data=sparrow)  summary(sparrow.glm)  # ------------------------------- (b) -------------------------------  sparrow.null.glm <- glm(Status~1,family=binomial(link=logit),data=sparrow)  anova(sparrow.null.glm,sparrow.glm,test="Chisq")  # the TS is 79.802  # ------------------------------- (c) -------------------------------  library(RcmdrMisc)  (sparrow.glm2 <- stepwise(sparrow.glm,direction="forward",criterion="BIC",trace=F))  # Coefficients:  # (Intercept) TL HL WT KL  # 49.9861 -0.6573 72.3327 -0.7896 27.3775  library(MASS)  sparrow.glm22 <- stepAIC(sparrow.null.glm,  scope=list(lower=~1,upper=~AG\*(TL+AE+WT+BH+HL+FL+TT+SK+KL)),  direction="forward",  k=log(length(sparrow$Status)),trace=F)  # Coefficients:  # (Intercept) TL HL WT KL  # 49.9861 -0.6573 72.3327 -0.7896 27.3775 |
| **Question 3 (4.0 points)** | |
| a) | There are in total 8+1=**9** unknown parameters in this ordinal response regression model. |
| b) | The 95% confidence interval for the coefficient of “fat” based on the fitted ordinal response above is  **(-0.6211,2.9352).** |
| c) | is not needed.  is needed.  The p-value is **0.2022** > 0.05, indicating “fat” is not significant. |
| d) | Now we have 82=**16** unknown parameters in the nominal response regression model. |
| e) | is not needed.  is needed.  The p-value is **0.2197** > 0.05, indicating “fat” is also not significant even in a nominal response regression model. |
| f) | The PCF for the ordinal response regression model is **0.24**. |
| g) | The PCF for the nominal response regression model is **0.20**.  Based on the two PCF values, the **ordinal response regression model is better**. This makes sense, as rating has an order that provides more information. |
| h) | # ------------------------------- Q3 -------------------------------  train <- read.csv("ice\_cream1.csv",header=T)  test <- read.csv("ice\_cream2.csv",header=T)  # ------------------------------- (a) -------------------------------  levels(factor(train$rating))  library(MASS)  rate.ord <- polr(formula=factor(rating)~fat,data=train,method="logistic")  summary(rate.ord)  # 1+8=9 unknown parameters  # ------------------------------- (b) -------------------------------  # round(confint(rate.ord, 'fat', level=0.95),4)  round(summary(rate.ord)$coefficient[1,1]-qnorm(0.975)\*summary(rate.ord)$coefficient[1,2],4)  round(summary(rate.ord)$coefficient[1,1]+qnorm(0.975)\*summary(rate.ord)$coefficient[1,2],4)  # (-0.6216,2.9369)  # ------------------------------- (c) -------------------------------  Anova(rate.ord)  # p-value 0.2022>0.05, fat not needed  # ------------------------------- (d) -------------------------------  library(nnet)  rate.nom <- multinom(formula=factor(rating)~fat,data=train,method="logistic")  summary(rate.nom)  # 2\*8=16 unknown parameters  # ------------------------------- (e) -------------------------------  Anova(rate.nom)  # p-value 0.2197>0.05, fat not needed as well  # ------------------------------- (f) -------------------------------  test$rating  pred.ord <- predict(rate.ord,test,type="class")  (pcf.ord <- sum(pred.ord==test$rating)/length(test$rating))  # 24%  # ------------------------------- (g) -------------------------------  pred.nom <- predict(rate.nom,test,type="class")  (pcf.nom <- sum(pred.nom==test$rating)/length(test$rating))  # 20% |
| **Question 4 (1.5 points)** | |
| a) | # ------------------------------- Q4 -------------------------------  # ------------------------------- (a) -------------------------------  set.seed(7001)  beta0 <- 2; beta1 <- 1  n <- 1000  X <- seq(0.001,1,0.001)  pii <- exp(beta0+beta1\*X)/(1+exp(beta0+beta1\*X))  k <- 0  Yi.df <- NULL  for (i in 1:1000) {  Yi <- rbinom(1000,1,pii)  Yi.df <- rbind(Yi.df,Yi)  }  coef.df <- NULL  for (i in 1:1000) {  fit <- glm(Yi.df[i,]~X,family="binomial")  coef.df <- rbind(coef.df,fit$coefficients)  }  # ------------------------------- (b) -------------------------------  round(mean(coef.df[,2]),4) # 1.0306  # ------------------------------- (c) -------------------------------  hist(coef.df[,2],main="Histogram of 1000 simulated beta1.hat",xlab="beta1.hat",  col="#7B84FC") |
| b) | The sample average of 1,000 estimates of  is **1.0306**, which is pretty close to the true value of =1. |
| c) | **Please paste the plot in the following:**    The values of simulated are generally **normally** distributed. |