**GRAU D’ENGINYERIA INFORMÀTICA (UPC).**

**CURS 19-20 Q2** –**QUIZ 2**

**Anàlisi de Dades i Explotació de la Informació (ADEI) .**

**(Data: 29/5/2020 10:00-12:00 h On-line** [**https://meet.google.com/uzh-kvbr-uus**](https://meet.google.com/uzh-kvbr-uus)

**Professor**: Lídia Montero Mercadé

**Rules for the quiz:** Emailing and chatting is strictly forbidden. Mobile phones should be switched off. PC camara should be turned on to invigilate you. You have to deliver 1 Name.FamilyName.pdf file containing answers to the questions, used commands and R output results needed to justify your answers.

**Duration:** 1h 45 min

**Marks**: Before 5/6/20 Subject ATENEA website.

**Open Office- online**: 5/6/20 10:000

**Problem 1: All questions account for 1 point**

1793 choices by 561 individuals of a transport mode from/to Freetown airport (Sierra Leone) to downtown. This problem exploits an unusual transportation setting to generate some of the first revealed preference value of a statistical life (VSL) estimates from a low-income setting. Four alternatives are available: ferry, helicopter, water-taxi and hovercraft. A striking characteristic of the study is that all these alternatives experienced fatal accidents in recent years, so that the fatality risk is non-negligible and differs much from an alternative to another. For example, the probabilities of dying using the water taxi and the helicopter are respectively of 2.55 and 18.41 out of 100,000 passenger-trips.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| ***id*** | Individual id *(not to be used in this exercise)* |
| **choice** | 1 for the chosen mode |
| **mode** | One of Helicopter, *(not to be used in this exercise)*WaterTaxi (a small craft for 12 to 18 pax), Ferry, and Hovercraft |
| **cost** | the generalised cost of the transport mode (US$) – *numeric target* |
| **risk** | The fatality rate, numbers of death per 100,000 trips for the selected mode |
| ***weight*** | Weights *(not to be used in this exercise)* |
| **seats** | Level of seat availability - comfort (Likert scale 1 to 5, transformed to 0 to 1 scale) |
| **noise** | Level for less noise disturbance (Likert scale 1 to 5, transformed to 0 to 1 scale) |
| **crowdness** | Level for less crowdedness (Likert scale 1 to 5, transformed to 0 to 1 scale) |
| **convloc** | Level of convenience location for the transfer (Likert scale 1 to 5, transformed to 0 to 1 scale) |
| **clientele** | Level of quality of ‘trip makers’ (Likert scale 1 to 5, transformed to 0 to 1 scale) |
| ***chid*** | Choice situation id *(not to be used in this exercise)* |
| **african** | yes if born in Africa, no otherwise |
| **lifeExp** | declared life expectancy |
| **dwage** | declared hourly wage |
| **iwage** | imputed hourly wage |
| **educ** | level of education, one of low and high |
| **fatalism** | self-ranking of the degree of fatalism |
| **gender** | gender, one of female and male |
| **age** | age |
| **haveChildren** | yes if the traveler has children, no otherwise |
| **swim** | yes if the traveler knows how to swim, 'no’, otherwise |
| **noalt** | Number of available alternatives for the selected choice |

The trade-offs that individuals are willing to make between mortality risk and cost as they travel to and from the international airport in Sierra Leone are estimated. The setting and original dataset allow us to address some typical variable concerns, and also to compare VSL estimates for travelers from different countries, all facing the same choice situation. The average VSL estimate for African travelers in the sample is US$ 577,000 compared to US$ 924,000 for non-Africans. The two covariates of interest are cost (the generalized cost in $PPP unit, not *leones*) and risk (mortality per 100,000 passenger-trips). The risk variable being purely alternative specific, intercepts for the alternatives cannot therefore be estimated. To avoid endogeneity problems, the authors introduce as covariates marks the individuals gave to 5 attributes of the alternatives: comfort, noise level, crowdedness, convenience and transfer location and the "quality" of the clientele.

### Source

**data(“RiskyTransport”)** # mlogit package – long format dataset

[American Economic Association data archive.](http://aeaweb.org/aer/" \t "_blank)

### References

León, Gianmarco, and Miguel, Edward. *Risky Transportation Choices and the Value of a Statistical Life*. Nashville, TN: American Economic Association [publisher], 2017. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12. https://doi.org/10.3886/E113686V1.

**Let us focus on travel cost (cost variable). Firstly, restrict your active data set to observations involving 4 available alternatives (noalt=4) and actual choice (choice=1). Secondly, define a new binary factor containing WaterTaxi choice versus Others.**

1. Indicate by data exploration tools which are globally the most associated variables with the **response variable (cost)**.

A condes() method in FactoMineR package can be used. Only global association has to be addressed. Global association of cost with numeric variables is shown using Pearson correlation coefficient and pvalues of the null hypothesis ‘correlation coefficient equal 0’. Positively correlated with high intensity are dwage, iwage and less intensity is shown for numeric scores crowdness, noise, convloc. An inverse relation indicated by a negative coefficient of correlation is shown for lifeExp and risk, but is not very intense.

Factor variables globally related to cost are the selected transportation mode (low intensity) and almost negligible are swimming capability (swim) and the binary factor WaterTaxi.

**> names(df4)**

**[1] "id" "choice" "mode" "cost"**

**[5] "risk" "weight" "seats" "noise"**

**[9] "crowdness" "convloc" "clientele" "chid"**

**[13] "african" "lifeExp" "dwage" "iwage"**

**[17] "educ" "fatalism" "gender" "age"**

**[21] "haveChildren" "swim" "f.wtaxi"**

**> res.con<-condes(df4,num.var=4)**

**> res.con$quanti**

**correlation p.value**

**dwage 0.7742782 3.742985e-65**

**iwage 0.7663530 8.960149e-82**

**crowdness 0.3399495 9.698704e-13**

**noise 0.2747916 1.165168e-08**

**convloc 0.2357815 1.120744e-06**

**seats 0.2152384 9.242447e-06**

**age 0.1507996 2.016119e-03**

**clientele 0.1335785 6.298648e-03**

**weight -0.1325241 6.727160e-03**

**risk -0.1695825 5.055468e-04**

**lifeExp -0.1723265 4.078928e-04**

**> res.con$quali**

**R2 p.value**

**mode 0.20847078 8.082973e-21**

**f.wtaxi 0.07666578 8.924484e-09**

**swim 0.01098157 3.240303e-02**

1. Calculate the linear model that explains the cost of the transfer from the imputed wage (iwage) and factor mode: interpret the regression lines and assess its global quality. What is the percentage of the cost variability that is explained by the transportation mode?

The complete Ancova model (main effects and interactions) has 8 parameters and according to Anova() tests for net-effects the interactions are significant once the main effects for iwage and mode have already been included in the model. Goodness of fit can be assessed using R2 80.46% of target’s variability is explained by the model. Transportation mode has to be introduced in the model as main effect and interaction with iwage. The model containing only iwage has an R2 of 58.73%, so almost 21% of target’s variability is explained by mode. The additive model is not the solution : interactions are needed.

Model interpretation:

* For mode==Helicopter Y= (37.14+0)+(2.33+0)\*iwage
* For mode==WaterTaxi Y= ( 37.14+23.16 )+( 2.33-1.49 )\*iwage
* For mode==Ferry Y= ( 37.14 – 33.57 )+( 2.33 – 0.66 )\*iwage
* For mode==Hovercraft Y= (37.14 + 55.95)+(2.33 - 1.85)\*iwage

**> m1<-lm(cost~mode\*iwage, data=df4)**

**> summary(m1)**

**Call: lm(formula = cost ~ mode \* iwage, data = df4)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 37.1415 32.1249 1.156 0.248291**

**modeWaterTaxi 23.1560 32.2502 0.718 0.473162**

**modeFerry -33.5647 32.2196 -1.042 0.298144**

**modeHovercraft 55.9495 32.3770 1.728 0.084732 .**

**iwage 2.3285 0.6251 3.725 0.000223 \*\*\***

**modeWaterTaxi:iwage -1.4847 0.6270 -2.368 0.018347 \***

**modeFerry:iwage -0.6589 0.6276 -1.050 0.294390**

**modeHovercraft:iwage -1.8493 0.6292 -2.939 0.003479 \*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 23.77 on 409 degrees of freedom**

**Multiple R-squared: 0.8046, Adjusted R-squared: 0.8012**

**F-statistic: 240.5 on 7 and 409 DF, p-value: < 2.2e-16**

**> Anova(m1)**

**Anova Table (Type II tests)**

**Response: cost**

**Sum Sq Df F value Pr(>F)**

**mode 139841 3 82.532 < 2.2e-16 \*\*\***

**iwage 587609 1 1040.399 < 2.2e-16 \*\*\***

**mode:iwage 116970 3 69.034 < 2.2e-16 \*\*\***

**Residuals 231000 409**

**---**

**> summary(m2)**

**Call: lm(formula = cost ~ iwage, data = df4)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 39.52579 2.42510 16.3 <2e-16 \*\*\***

**iwage 1.12119 0.04614 24.3 <2e-16 \*\*\***

**---**

**Residual standard error: 34.28 on 415 degrees of freedom**

**Multiple R-squared: 0.5873, Adjusted R-squared: 0.5863**

**F-statistic: 590.6 on 1 and 415 DF, p-value: < 2.2e-16**

1. Calculate a linear model for the target cost using all available numeric variables. Are there any collinearity issues in the model? Justify the solution to remove collinearity.

The model using numeric variables has to contain risk, fatalism, age, lifeExp as characteristics of the trip maker and numeric scores seats, noise, crowdness, convloc and clientele. The model explains 65.88% of cost variability. Only crowdness and iwage net-effects are significant at the 5% usual threshold, but noise pvalue is not so far and has to be also included as a remarkable variable. Using vif() method in library car, noise and crowdness pair seem to be correlated and age and lifeExp pair also. You have to retain one variable in each pair, either the most correlated, or the reliable: I choose crowdness to solve the first pair problem and age for the second (more objective variable than lifeExp). You can see that m4 containing all numeric except noise and lifeExp has solved collinearity problems. Removing non-significant variables, only iwage and crowdness are retained.

**> summary(m3)**

**Call: lm(formula = cost ~ risk + seats + noise + crowdness + convloc +**

**clientele + lifeExp + iwage + fatalism + age, data = df4)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 32.48380 23.93538 1.357 0.175489**

**risk 0.87010 1.25178 0.695 0.487397**

**seats 5.68797 10.75632 0.529 0.597232**

**noise 19.93784 10.98507 1.815 0.070262 .**

**crowdness 35.67569 10.12037 3.525 0.000471 \*\*\***

**convloc 3.48669 8.65774 0.403 0.687362**

**clientele -12.70834 12.08796 -1.051 0.293736**

**lifeExp -0.28703 0.26061 -1.101 0.271389**

**iwage 1.08426 0.04381 24.748 < 2e-16 \*\*\***

**fatalism -0.37977 0.54686 -0.694 0.487794**

**age -0.35538 0.28439 -1.250 0.212165**

**---**

**Residual standard error: 31.52 on 406 degrees of freedom**

**Multiple R-squared: 0.6588, Adjusted R-squared: 0.6504**

**F-statistic: 78.38 on 10 and 406 DF, p-value: < 2.2e-16**

**> vif(m3)**

**risk seats noise crowdness convloc clientele lifeExp**

**1.201182 1.674560 3.260330 3.442236 1.589556 1.533807 4.357885**

**iwage fatalism age**

**1.067008 1.038980 4.407140**

**> m4<-lm(cost~risk+seats+crowdness+convloc+clientele+age+iwage+fatalism, data=df4)**

**> vif(m4)**

**risk seats crowdness convloc clientele age iwage**

**1.197047 1.551178 2.140106 1.574705 1.510840 1.081610 1.058039**

**fatalism**

**1.037831**

**> m5<-step(m4,k=log(nrow(df4)))**

**Start: AIC=2925.65**

**cost ~ risk + seats + crowdness + convloc + clientele + age +**

**iwage + fatalism**

**Step: AIC=2892.2**

**cost ~ crowdness + iwage**

**…**

**Df Sum of Sq RSS AIC**

**<none> 410617 2892.2**

**- crowdness 1 77194 487811 2958.0**

**- iwage 1 634775 1045392 3275.8**

**> summary(m5)**

**Call: lm(formula = cost ~ crowdness + iwage, data = df4)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 9.23976 4.09239 2.258 0.0245 \***

**crowdness 48.38858 5.48492 8.822 <2e-16 \*\*\***

**iwage 1.07896 0.04265 25.298 <2e-16 \*\*\***

**---**

**Residual standard error: 31.49 on 414 degrees of freedom**

**Multiple R-squared: 0.6526, Adjusted R-squared: 0.6509**

**F-statistic: 388.9 on 2 and 414 DF, p-value: < 2.2e-16**

1. Once the best model for target cost using explanatory numeric variables has been proposed, are there any significant main factor effects to be included? And interactions? Justify your answer.

Transformations to explanatory variables are not considered in the exercise, but they should be tested in a real study. Model m7<-lm(cost~crowdness+iwage+mode+gender+african+educ+haveChildren+swim, data=df4) is considered and Anova(m7) shows that some variables are redundant, being only crowdness, iwage, mode, gender, African and educ those with net significant effects. Interactions between factors and covariates are included: some aliased coefficients message indicates an specification problem: mode and crowdness interactions cannot be calculated, thus mode and crowdness interaction is not considered. After m8 model calculation and reduction using step() method with BIC monitoring, a final model (m9) containing:

**cost ~ iwage + mode + crowdness + gender + african + educ + iwage:mode +**

**crowdness:african + crowdness:educ + iwage:african is obtained.**

It is a complex model that explains 84% of cost variability.

**> m7<-lm(cost~crowdness+iwage+mode+gender+african+educ+haveChildren+swim, data=df4)**

**> #summary(m7)**

**> Anova(m7)**

**Anova Table (Type II tests)**

**Response: cost**

**Sum Sq Df F value Pr(>F)**

**crowdness 1282 1 1.6312 0.202272**

**iwage 601717 1 765.8140 < 2.2e-16 \*\*\***

**mode 72832 3 30.8980 < 2.2e-16 \*\*\***

**gender 3198 1 4.0696 0.044320 \***

**african 14831 1 18.8759 1.765e-05 \*\*\***

**educ 10802 1 13.7485 0.000238 \*\*\***

**haveChildren 120 1 0.1528 0.696102**

**swim 708 1 0.9008 0.343141**

**Residuals 319003 406**

**---**

**> m8<-lm(cost~(crowdness+iwage)\*(mode+gender+african+educ+haveChildren+swim), data=df4) # Some crwodness:mode parameters can not be estimated**

**> m8<-lm(cost~iwage\*mode+(crowdness+iwage)\*(gender+african+educ+haveChildren+swim), data=df4)**

**> Anova(m8)**

**Anova Table (Type II tests)**

**Response: cost**

**Sum Sq Df F value Pr(>F)**

**iwage 584296 1 1244.2909 < 2.2e-16 \*\*\***

**mode 61569 3 43.7047 < 2.2e-16 \*\*\***

**crowdness 465 1 0.9894 0.320508**

**gender 2730 1 5.8139 0.016358 \***

**african 15220 1 32.4121 2.444e-08 \*\*\***

**educ 4612 1 9.8217 0.001854 \*\***

**haveChildren 8 1 0.0175 0.894685**

**swim 130 1 0.2766 0.599210**

**iwage:mode 97741 3 69.3819 < 2.2e-16 \*\*\***

**crowdness:gender 2543 1 5.4162 0.020458 \***

**crowdness:african 4063 1 8.6515 0.003461 \*\***

**crowdness:educ 4864 1 10.3585 0.001396 \*\***

**crowdness:haveChildren 138 1 0.2930 0.588603**

**crowdness:swim 1292 1 2.7514 0.097968 .**

**iwage:gender 42 1 0.0888 0.765818**

**iwage:african 4201 1 8.9457 0.002956 \*\***

**iwage:educ 61 1 0.1298 0.718820**

**iwage:haveChildren 38 1 0.0820 0.774819**

**iwage:swim 44 1 0.0935 0.759894**

**Residuals 184546 393**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**> m9<-step(m8,k=log(nrow(df4)))**

**Start: AIC=2685.39**

**cost ~ iwage \* mode + (crowdness + iwage) \* (gender + african +**

**educ + haveChildren + swim)**

**…**

**Step: AIC=2640.01**

**cost ~ iwage + mode + crowdness + gender + african + educ + iwage:mode +**

**crowdness:african + crowdness:educ + iwage:african**

**Df Sum of Sq RSS AIC**

**<none> 188533 2640.0**

**- crowdness:african 1 2752 191285 2640.0**

**- gender 1 3026 191560 2640.6**

**- iwage:african 1 5711 194245 2646.4**

**- crowdness:educ 1 5910 194444 2646.8**

**- iwage:mode 3 105401 293934 2807.1**

**> Anova(m9)**

**Anova Table (Type II tests)**

**Response: cost**

**Sum Sq Df F value Pr(>F)**

**iwage 599777 1 1278.8731 < 2.2e-16 \*\*\***

**mode 61614 3 43.7918 < 2.2e-16 \*\*\***

**crowdness 519 1 1.1067 0.2934391**

**gender 3026 1 6.4526 0.0114544 \***

**african 15244 1 32.5049 2.305e-08 \*\*\***

**educ 4720 1 10.0633 0.0016286 \*\***

**iwage:mode 105401 3 74.9135 < 2.2e-16 \*\*\***

**crowdness:african 2752 1 5.8671 0.0158679 \***

**crowdness:educ 5910 1 12.6023 0.0004309 \*\*\***

**iwage:african 5711 1 12.1776 0.0005372 \*\*\***

**Residuals 188533 402**

**---**

**> summary(m9)**

**Call:**

**lm(formula = cost ~ iwage + mode + crowdness + gender + african +**

**educ + iwage:mode + crowdness:african + crowdness:educ +**

**iwage:african, data = df4)**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 38.01555 32.22528 1.180 0.238825**

**iwage 2.27626 0.57501 3.959 8.91e-05 \*\*\***

**modeWaterTaxi 7.74194 29.74325 0.260 0.794772**

**modeFerry -54.83830 29.97680 -1.829 0.068087 .**

**modeHovercraft 35.81018 29.81342 1.201 0.230403**

**crowdness 22.73403 15.13510 1.502 0.133863**

**gendermale -6.61084 2.60249 -2.540 0.011454 \***

**africanAfr.Yes 7.34304 6.67988 1.099 0.272305**

**educhigh 23.56247 9.83545 2.396 0.017047 \***

**iwage:modeWaterTaxi -1.29623 0.57760 -2.244 0.025366 \***

**iwage:modeFerry -0.48275 0.57717 -0.836 0.403429**

**iwage:modeHovercraft -1.64087 0.57875 -2.835 0.004811 \*\***

**crowdness:africanAfr.Yes 20.52763 8.47476 2.422 0.015868 \***

**crowdness:educhigh -46.28645 13.03856 -3.550 0.000431 \*\*\***

**iwage:africanAfr.Yes -0.21790 0.06244 -3.490 0.000537 \*\*\***

**Residual standard error: 21.66 on 402 degrees of freedom**

**Multiple R-squared: 0.8405, Adjusted R-squared: 0.8349**

**F-statistic: 151.3 on 14 and 402 DF, p-value: < 2.2e-16**

1. Select the best model available so far. Let us assume an observation on the median of numeric variables and reference levels for the factors. Estimate a 90% confidence interval for predicted transfer cost.

This question can be easily answered using predict() method. My best model is m1: the one using iwage\*mode, since it explains almost 80% of the target and it is simpler than m9 (explaining 84%). Answers including the best model obtained after Question 5 have been also considered correct.

Median of iwage is 27.96236 and reference level for mode is ‘Helicopter’ then

For mode==Helicopter Y= (37.14+0)+(2.33+0)\*iwage= 37.14+2.33\*27.96=102.2525 $ is the point estimate. 90% confidence interval for the predicted cost can not be easily calculated without using predict(model, newdata=.) method in R.

**> predict(m1,newdata=data.frame(iwage=median(df4$iwage),mode="Helicopter"),interval="prediction",level=0.9)**

**fit lwr upr**

**1 102.2525 51.56798 152.9369**

1. Graphically assess the best model obtained so far. Assess the presence of outliers in the studentized residuals at 95% confidence level. Indicate which those observations are and why they are showing lack of fit.

Again m1 is my best model so far, but the best model obtained at Question 4 can be also used. Diagnostic show that the model is not good. Since this is a question in an exam, you have to answer lack of fit issues. Absolute studentized residuals over 3.0 are considered outliers and these correspond to observations 45 and 46 in df4 register order or rownames “627” and “631”. These registers belong to young women that have paid a lot of money for a Hovercraft service to downtown. Influent data is present and transformations would be needed for explanatory variables and outcome, but this is not the aim for this exam.

**qnorm(0.975)**

**[1] 1.959964**

**> ll<-which(abs(rstudent(m1))>qnorm(0.975));ll;length(ll)**

**570 574 606 613 617 627 631 644 907 928 932 947 951 962**

**36 37 42 43 44 45 46 47 62 64 65 68 69 71**

**966 993 1010 1304 1339 1364 1377 1597 1897 3082 3103 3117 3121 3128**

**72 75 77 94 100 104 105 121 143 236 238 240 241 242**

**3132 3558 3569 3593 3603 3971 4643 4647 4773**

**243 263 265 268 269 297 347 348 356**

**[1] 37**

**> #df4[ll,]**

**> ll<-which(abs(rstudent(m1))>3.0);ll**

**627 631**

**45 46**

**> df4[ll,]**

**id choice mode cost risk weight seats noise**

**627 8290608 1 Hovercraft 170.9406 3.881836 1.215615 0.8 1**

**631 8290608 1 Hovercraft 205.1287 3.881836 1.215615 0.8 1**

**crowdness convloc clientele chid african lifeExp dwage iwage educ**

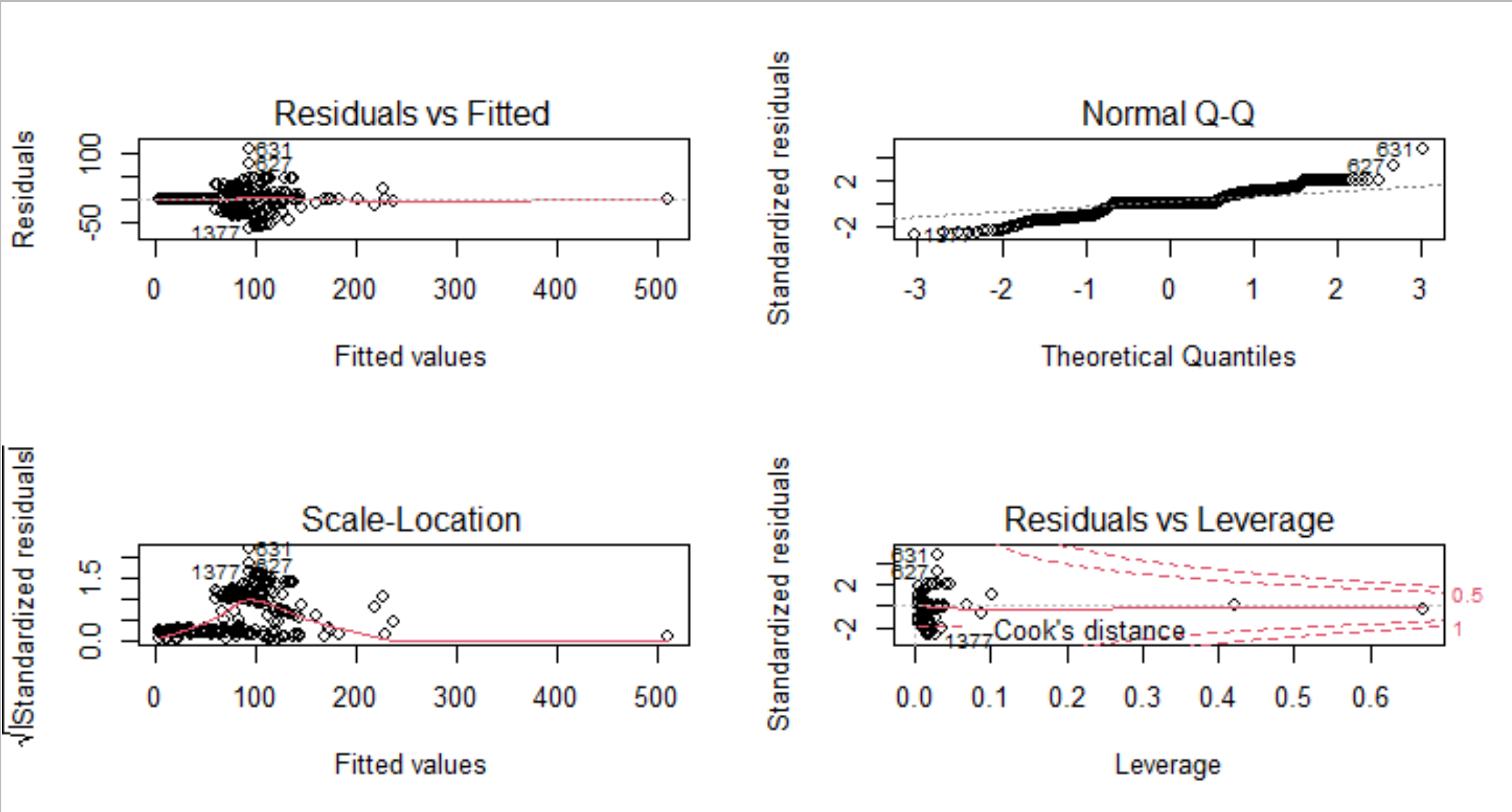
**627 1 0.6 0.8 303 Afr.Yes 66 0 0 low**

**631 1 0.6 0.8 304 Afr.Yes 66 0 0 low**

**fatalism gender age haveChildren swim f.wtaxi**

**627 1 female 19 chil.Yes swim.No WTaxi.No**

**631 1 female 19 chil.Yes swim.No WTaxi.No**



1. Study the presence of *a priori and a posteriori* influential data observations. Indicate thresholds to be applied to the statistic involved in the diagnostic.

Easily done using influencePlot(model). Helicopter users are just 2 in the sample and those are the influent data: there is not model that can deal with 4 modes given the low market share for Helicopter. These observations should be removed and the exercise has to be repeated again.

|  |
| --- |
| **> influencePlot(m1)**  **StudRes Hat CookD**  **627 3.365672 0.02879271 0.04094436**  **631 4.917400 0.02879271 0.08480268**  **779 NaN 1.00000000 NaN**  **4785 NaN 1.00000000 NaN**  **> df4[c("779","4785"),]**  **id choice mode cost risk weight seats noise**  **779 8300204 1 Helicopter 76.53285 18.4082 0.3863821 1.0 1.0**  **4785 9160602 1 Helicopter 201.73314 18.4082 0.3863821 0.4 0.4**  **crowdness convloc clientele chid african lifeExp dwage iwage**  **779 1.0 1.0 1.0 397 Afr.No 36 16.9169 16.91690**  **4785 0.6 0.8 0.6 2326 Afr.No 25 NA 70.68501**  **educ fatalism gender age haveChildren swim f.wtaxi**  **779 high 2 female 49 chil.Yes swim.Yes WTaxi.No**  **4785 high 6 male 60 chil.Yes swim.Yes WTaxi.No**  **> table(df4$mode)**  **Helicopter WaterTaxi Ferry Hovercraft**  **2 180 174 61** |

1. **WaterTaxi binary choice factor is the new target to be addressed.** Estimate a logit model including seats, crowdness, convloc covariates and educ and swim factors. Discuss model fit taking into account marginal trends and residual plots.

Some lack of fit is shown in the marginal plots for seats and mainly for crowdness scores, anyway residualPlots do show a fat smoother for the global fit (last plot, right below). All factors and covariates have significant net-effects according to Anova() method. No collinearity is present in the model.

Since residual deviance is 376.95 on 411 degrees of freedom and disaggregated data is the type of this dataset, using the practical ‘rule of thumb’ that indicates that residual deviance should not be less than d.ll. and this holds as shown in the output.

**> m20<-glm(f.wtaxi~seats+crowdness+convloc+educ+swim, family=binomial, data=df4)**

**> summary(m20)**

**Call:**

**glm(formula = f.wtaxi ~ seats + crowdness + convloc + educ +**

**swim, family = binomial, data = df4)**

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

**(Intercept) -6.5213 0.8205 -7.948 1.90e-15 \*\*\***

**seats -2.4709 0.8743 -2.826 0.00471 \*\***

**crowdness 4.1474 0.6661 6.226 4.78e-10 \*\*\***

**convloc 5.4724 0.8427 6.494 8.38e-11 \*\*\***

**educhigh 1.1201 0.3351 3.343 0.00083 \*\*\***

**swimswim.Yes 0.7489 0.2577 2.906 0.00366 \*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for binomial family taken to be 1)**

**Null deviance: 570.27 on 416 degrees of freedom**

**Residual deviance: 376.95 on 411 degrees of freedom**

**AIC: 388.95**

**> Anova(m20,test="LR")**

**Analysis of Deviance Table (Type II tests)**

**Response: f.wtaxi**

**LR Chisq Df Pr(>Chisq)**

**seats 8.371 1 0.0038127 \*\***

**crowdness 47.420 1 5.729e-12 \*\*\***

**convloc 52.660 1 3.966e-13 \*\*\***

**educ 11.838 1 0.0005803 \*\*\***

**swim 8.646 1 0.0032782 \*\***

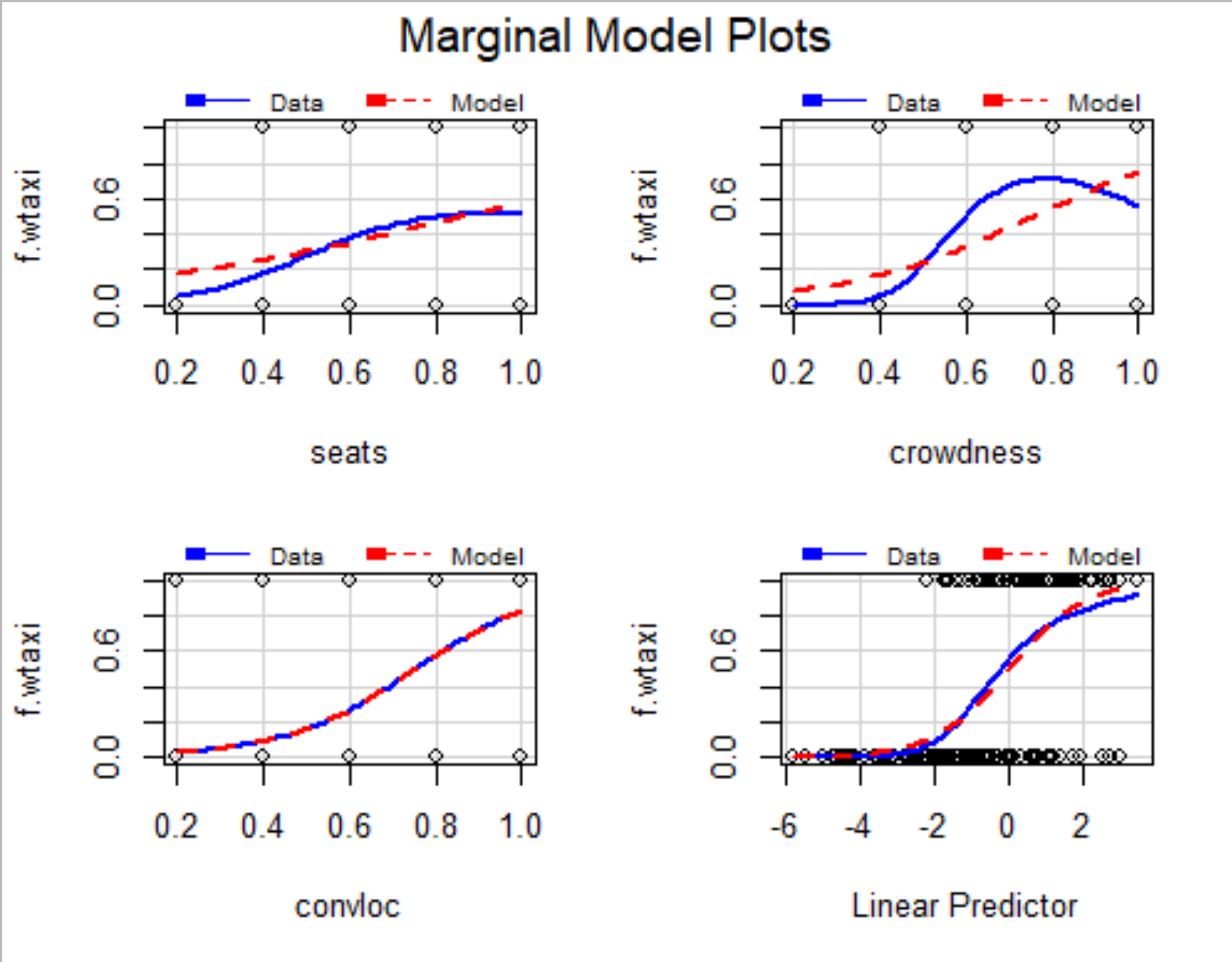
**---**

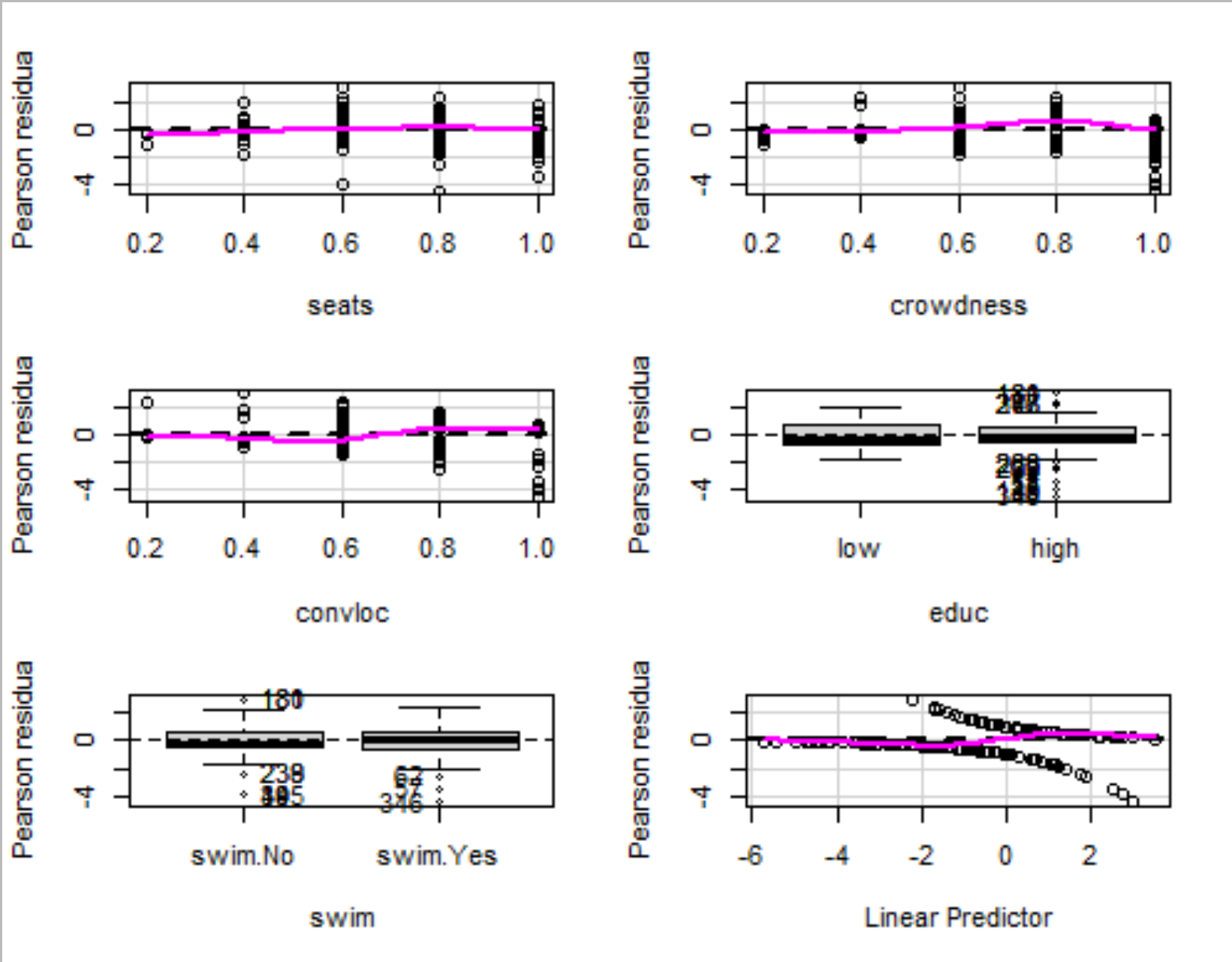
**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**> vif(m20)**

**seats crowdness convloc educ swim**

**1.482731 1.430404 1.141454 1.033825 1.023785**





1. Interpret model equations and the effects in the odds scale of involved factors.

where

,

* for factor educ, where level 1 is education-low and 2 is education-high.
* for factor swim, where level 1 is swim-No and 2 is swim-Yes
* There are as many model equations as 2 x 2= 4 (product of number of levels for factors educ and swim)

Interpretation of the model in the odds scale:

Increasing by 0.1 units seats scored then exp(-2.47\*0.1)= 0.7811407 -> 100\*(1-0.7811)=22%, the odds of the probability of choosing WaterTaxi decreases by 22%, all else being equal.

Increasing by 0.1 units seats scored then exp(4.15\*0.1)= 1.514371

-> 100\*(1.514371-1)=51%, the odds of the probability of choosing WaterTaxi increases by 51%, all else being equal.

Increasing by 0.1 units seats scored then exp(5.47\*0.1)= 1.728061

-> 100\*( 1.728061-1)=72%, the odds of the probability of choosing WaterTaxi increases by 72%, all else being equal.

The odds of the probability of choosing WaterTaxi for high educated people increases by exp(1.1201)=3.065 -> 100\*(3.065 -1)= 206% the probability of choosing WaterTaxi in the reference level education-low all else being equal.

The odds of the probability of choosing WaterTaxi for people that can swim increases by exp(0.7489)= 2.114-> 100\*(2.114 -1)= 111% the probability of choosing WaterTaxi in the reference level of people that cannot swim, all else being equal.

**> summary(m20)**

**Call: glm(formula = f.wtaxi ~ seats + crowdness + convloc + educ +**

**swim, family = binomial, data = df4)**

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

**(Intercept) -6.5213 0.8205 -7.948 1.90e-15 \*\*\***

**seats -2.4709 0.8743 -2.826 0.00471 \*\***

**crowdness 4.1474 0.6661 6.226 4.78e-10 \*\*\***

**convloc 5.4724 0.8427 6.494 8.38e-11 \*\*\***

**educhigh 1.1201 0.3351 3.343 0.00083 \*\*\***

**swimswim.Yes 0.7489 0.2577 2.906 0.00366 \*\***

**> exp(coef(m20))**

**(Intercept) seats crowdness convloc educhigh**

**1.471806e-03 8.450815e-02 6.327141e+01 2.380306e+02 3.065202e+00**

**swimswim.Yes**

**2.114570e+00**

1. What would be the expected probability of using a ‘WaterTaxi’ for a high education and swimmer trip maker when numeric explanatory variables are set to their sample minimum?

**> predict(m20,newdata=data.frame(seats=min(df4$seats),crowdness=min(df4$crowdness),convloc=min(df4$convloc),educ="high",swim="swim.No"),type="response",se.fit=T,level=0.95)**

**$fit**

**1**

**0.01849899**

**$se.fit**

**1**

**0.01022678**

**$residual.scale**

**[1] 1**