**, MANUAL PINXU MASTER RACE APROVAR EXAMEN ADEI 2 DIAS ANTES SIN SABER NADA**

EXAMEN:

**1) Determine the most promigsing variables for forefasting pruposes.**

condes.

Per totes les variables que estan més correlacionadas son les primeres en el dataset.(les que tinguin el p-value més petit), per tant son les que estan mes associades al targets.

**2) Fil the blanks**

**t value→** Estimate/Std.Error

Estimate→ t value \*Std.Error

**Degrees of freedom →** Número de mostres - numero de parametres(files de la sortida, (intercept i accidents)

**p-value→** Molt petit sempre perque sempre rebutgem el test omnibus (2\*e^-16).

**Múltiple R^2→** 1-(RSS/TSS)

RSS=(S^2\*(n-p)) → Suma quadrats residuals

TSS= (s^2\*(n-1)) → Suma de quadrats totals

S = std error

(n-p)= deegres freedom

s =Std error var target (sd)

**3) Check de diagnostic plots xD**

**Residual vs Fitted** → Cas ideal = No hi hagi cap estructura. Els que queden per sobre de la linea vermella son grans residus i els de sota petits. Els que estan lluny tenen pinta outliers i per tant sembla que el model necessiti una modificació logaritmica.

**Normal Q-Q** → Cas ideal = Haurien de seguir la recta de regresió lineal, si aixo no es compleix no existeix normalitat i els que estan lluny son valors atípics dels residus.

**Scale Location** → Cas ideal = La linea vermella ha de ser constant, aleshores la variança es normal, si no veiem dispersió; i en aquest cas quan augmenten les prediccions augmenta les variances del residus.(Heterosticitat)

Idealment l'smoother ha de ser pla: una recta plana sense pendent

**Residual vs Leverage** → Els que estan entre dos corbes de nivell està lluny del centre de gravetat, per tant la distancia de Cook sera elevada i seran mostresinfluents. Els que estan per sobre de tot segons els criteri de R fairfield son molt influents totes tenen els average elevat.

**4) Pros and cons of models.**

Comparar amb l’anterior i veure els canvis a les gràfiques, veure que ha suposat millor o pitjor i extreure conclusions de quin model s’ajusta mes a la realitat. A mes el mes el Multiple R-Squared determina el % de representacio de la mostra.

**5) Write the prediccion equation**

Prediction Equation → Log(\*var\*+1) = \*PrimerEstimate\* + \*SegundoEstimate\* log(x);

Point prediction → \*var\* = (exp^(\*primerEstimate\*) x mean(\*var\* ^\*SegundoEstimate\*))-1;

**6) interactions of modeling**

El multiple R-squared ens dona la informació del % que ens pot explicar les seguents dades.

Per calcular les prediccions del nombre de claims es pot deduir a partir de les seguents formules, una formula per cada possible combinacio amb els valors de la variable:

log(Y+1) = 4.8035 + 0 +0 +0 (No, 1)

log(Y+1) = 4.8035 + (-1.827) + 0 + 0 (Si,1)

log(Y+1) = 4.8035 + 0+ 2.5875 + 0 (No, 2)

log(Y+1) = 4.8034 + (-1.0823) + 2.5875 + 1.7637(si, 2)

**7) Anova interpretation**

Segons el criteri AIC els que tinguin el P-value menos de 0.05 son les que més t’interesen, pero si estan en una interaccio s’ha de fer un nou model.

**Si en el Anova hi ha una interaccio que fem amb les variables implicades?**

**8) LGA…**

-

**9) El mateix que el 3**

**10) BUBBLE GRAPHICS**

La variable que esta mes a la dreta te un average molt alt.

El que te una CookDistance més alta es el que es mes remarcable.

StudRes = Residuals

Hat = Average

CookD = Distance