Tarea 1

Juan Cantero Jimenez

4/26/2022

Modelización banco de datos titanic::titanic_train

a) Identificación de variables y principales características

Es necesario destacar que se han eliminado las variables Ticket, Cabin y Name pues aportan información del individuo que no permite relacionarlos con el resto. Sin embargo, esta última además del nombre contiene información sociocultural sobre este en forma de un titulo honorífico, ej. Dr, Sir etc., que puede ser interesante de cara a la predicción de la mortalidad. Este ha sido extraido de la variable Name haciendo uso de expresiones regulares. Tambien se han retirado las observaciones con datos faltantes

Table 1: Variables usadas en el modelo

Variables	Tipo	Subtipo	Rol	Descripción
Survived	Cuantitativa	Discreta	respuesta	Indicacor de supervivencia del pasajero
Pclass	Categórica	ordinal	explicativa	clase del pasajero
Sex	Categórica	nominal	explicativa	sexo del individuo
Age	Cuantitativa	continua	explicativa	edad del individuo
SibSp	Cuantitativa	discreta	explicativa	Numero de hermanas/esposas
Parch	Cuantitativa	discreta	explicativa	Numero de padres/hijos
Fare	Cuantitativa	continua	explicativa	Tarifa pagada
Embarked	Categórica	nominal	explicativa	Puerto de embarque
Title	Categórica	nominal	explicativa	Titulo honorífico del individuo

b) Análisis descriptivo de los datos

Table 2: Descriptiva numérica de los datos

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Survived	1	712	0.40	0.49	0.00	0.38	0.00	0.00	1.00	1.00	0.39	-1.85	0.02
Pclass*	2	712	2.24	0.84	2.00	2.30	1.48	1.00	3.00	2.00	-0.47	-1.41	0.03
Sex*	3	712	1.64	0.48	2.00	1.67	0.00	1.00	2.00	1.00	-0.57	-1.68	0.02
Age	4	712	29.64	14.49	28.00	29.22	12.97	0.42	80.00	79.58	0.39	0.17	0.54
SibSp	5	712	0.51	0.93	0.00	0.30	0.00	0.00	5.00	5.00	2.50	6.93	0.03
Parch	6	712	0.43	0.85	0.00	0.24	0.00	0.00	6.00	6.00	2.60	8.72	0.03
Fare	7	712	34.57	52.94	15.65	23.00	12.02	0.00	512.33	512.33	4.65	30.69	1.98
Embarked*	8	712	2.60	0.78	3.00	2.74	0.00	1.00	3.00	2.00	-1.48	0.28	0.03
Title*	9	712	11.21	1.90	12.00	11.42	0.00	1.00	17.00	16.00	-1.48	3.46	0.07

Si se atiende a la figura 2, se podrá apreciar la fuerte relación entre la tasa de mortalidad y el sexo, la clase en la que se viaje, así como del número de esposas y hermanas.

Figura 1: Histogramas variables cuantitativas

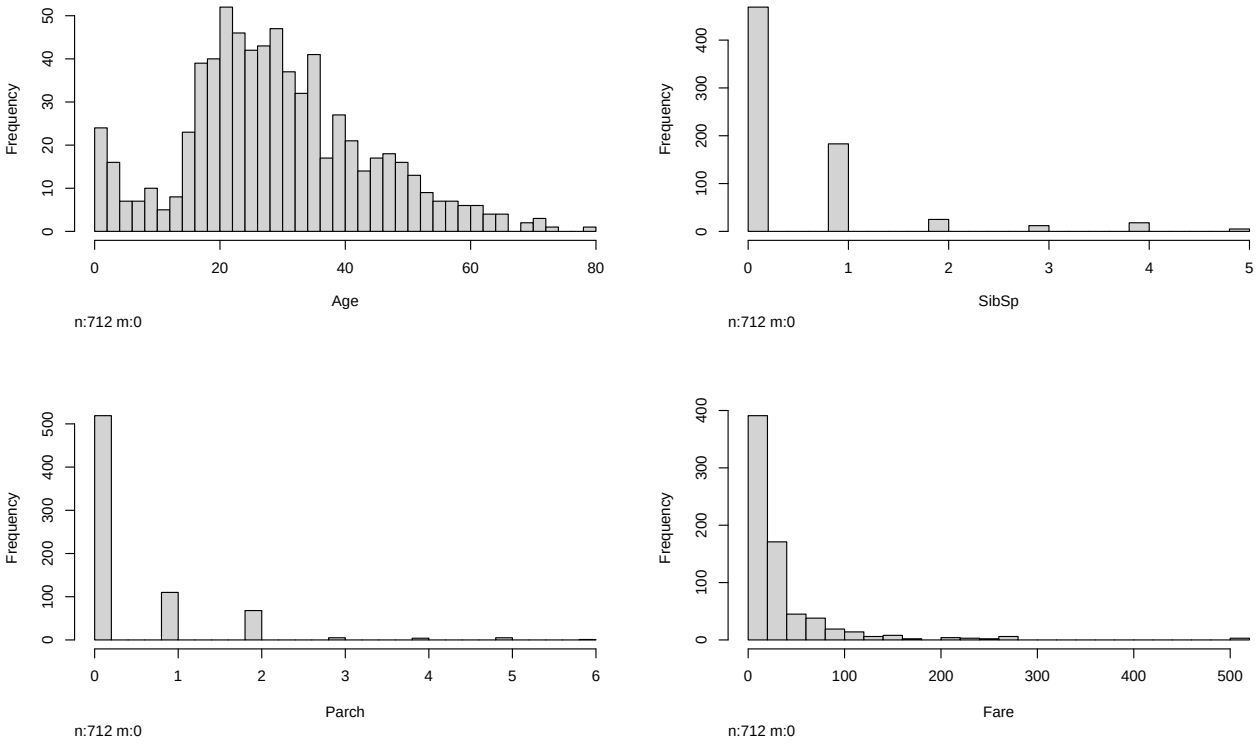
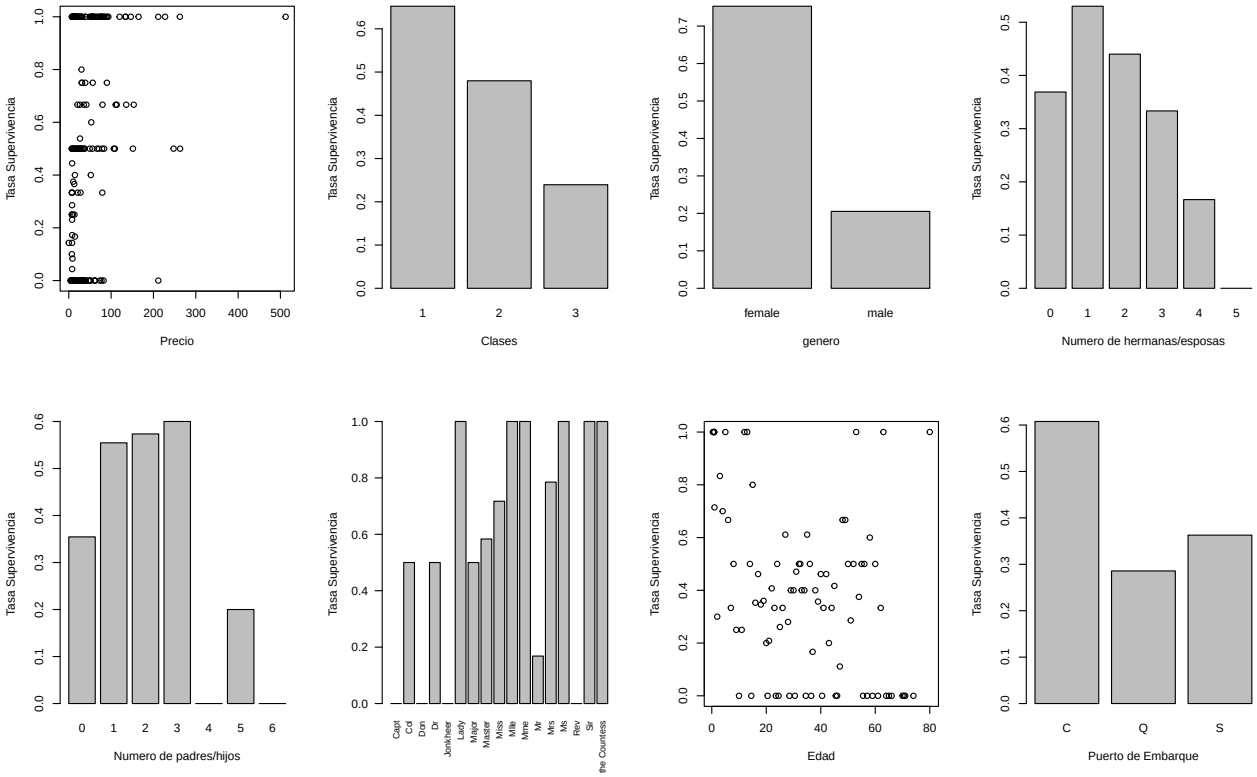


Figura 2: Relación entre variables explicativas y probabilidad de sobrevivir



c) Describe la distribución de probabilidad que se pueda asumir para la variable respuesta ¿Cuál es el parámetro de interés?

Se utilizarán dos distribuciones similares para describir los datos. La primera se trata de una distribución Bernoulli, que describe la probabilidad de exito en una prueba, que modelizará la variable Survived, notese que cada observación se denota como s:

$$- \textit{Función de probabilidad} : f(s \mid \pi) = \pi^s(1-\pi)^{1-s} \; si \; s = 0, 1 \tag{1}$$

$$- \textit{Rango del parámetro} : 0 \leq \pi \leq 1 \tag{2}$$

$$- MediayVarianza : E(Survived) = \pi; Var(Survived) = \pi(1-\pi) \tag{3}$$

Además se discretizan las variables Age y Fare hacia las variables Age.g y Fare.g permitiendo agrupar los individuos que sobreviven, y así poder ser modelizada mediante una distribución binomial, que describe el número de supervivientes en n individuos que pueden ser entendidos como eventos Bernoulli. Notese que la variable Survived pasara a ser Survived.g y cada evento se denota como sg:

$$- \textit{Función de probabilidad} : f(sg \mid n, \pi) = \binom{n}{sg}\pi^{sg}(1-\pi)^{1-sg} \; si \; sg = 0, 1, ..., n \tag{4}$$

$$- \textit{Rango del parámetro} : 0 \leq \pi \leq 1 \tag{5}$$

$$- MediayVarianza : E(Survived.g) = n\pi; Var(Survived.g) = n\pi(1-\pi) \tag{6}$$

d) Modelos lineales generalizados propuestos y validación de estos.

Table 3: Modelos propuesto

Nombre	GLM_call
aj1.ber.logit	glm(Survived ~ ., family = binomial(link="logit"), data=titanic.t) *
aj1.bi.logit	glm(cbind(Total, Survived.g) ~ ., family = binomial(link="logit"), data=titanic.t.g) †
aj1.ber.probit	glm(Survived ~ ., family = binomial(link="probit"), data=titanic.t) *
aj1.bi.probit	glm(cbind(Total, Survived.g) ~ ., family = binomial(link="probit"), data=titanic.t.g) †
aj1.ber.cloglog	glm(Survived ~ ., family = binomial(link="cloglog"), data=titanic.t) *
aj1.bi.cloglog	glm(cbind(Total, Survived.g) ~ ., family = binomial(link="cloglog"), data=titanic.t.g) †

* El data.frame titanic.t contiene las variables: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked y Title.

† De igual forma el data.frame titanic.t.g contiene las variables: Survived.g, Total, Pclass, Sex, Age.g, SibSp, Parch, Fare.g, Embarked y Title

Table 4: Validación modelos propuestos

Modelo	Chi.sq	Deviance	Null.Deviance	Test.Hosmer.Lemeshow	Shapiro.test	Tendencia	Residuos.extremos	AIC
aj1.ber.logit	1	575.76	960.90	0.0059 **		Si	Si	627.76
aj1.bi.logit	1	64.51	201.82		0.1021	No	No	361.52
aj1.ber.probit	1	577.60	960.90	0.0181 *		Si	Si	629.60
aj1.bi.probit	1	63.45	201.82		0.079	No	No	360.46
aj1.ber.cloglog	1	568.02	960.90	0.0019 **		Si	Si	620.02
aj1.bi.cloglog	1	60.83	201.82		0.0434 *	No	No	357.84

Note:

* P.value <= 0.05, ** P.value <= 0.01. Los residuos extremos implican un valor absoluto mayor a 2

Atendiendo a la tabla 4, se descartan la aplicabilidad de los modelos lineales generalizados con distribución Bernoulli (aj1.ber.logit, aj1.ber.probit, aj1.ber.cloglog) pues se rechaza la hipótesis nula de un buen modelo según el test de Hosmer y Lemeshow, además estos poseen residuos extremos y tendencia. Se consideran validos los modelos basados en distribución binomial (aj1.bi.logit, aj1.bi.probit, aj1.bi.cloglog), de estos el único que no cumple de forma estricta las condiciones de aplicabilidad es el aj1.b1.cloglog pues se rechaza la hipótesis de normalidad para los residuos Deviance.

e) Mejora de los modelos propuestos.

Se aplicará la función step, con el agrumento direction="backward", sobre los modelos escogidos en el apartado anterior. Cabe destacar que la función step escoge las variables Pclass, Sex y Age.g, así, se ha

visto combeniente reagrupar los datos en función de estas variables. Esto tiene como consecuencia directa un aumento del número de individuos presentes en cada grupo. Los datos reagrupados se encuentran en el data.frame titanic.t.g2. A además para cada uno de estos se presenta un modelo con interacciones. La interacción ha sido seleccionado mediante ensayo y error, comparando con la función anova, test="Chi", el modelo sin interacción y el modelo con interacción. Aquellas en las que la interacción es significativa con un nivel de confianza del 5 % se muestran en la tabla 5.

Table 5: Nuevos modelos propuesto

Nombre	GLM_call
aj2.bi.logit	glm(cbind(Total, Survived.g) ~ Pclass + Sex + Age.g, family = binomial(link = "logit"), data=titanic.t.g2) †
aj3.bi.logit	glm(cbind(Total, Survived.g) ~ Pclass + Sex + Age.g + Sex:Pclass , family = binomial(link = "logit"), data=titanic.t.g2) †
aj2.bi.probit	glm(cbind(Total, Survived.g) ~ Pclass + Sex + Age.g, family = binomial(link = "probit"), data=titanic.t.g2) †
aj3.bi.probit	glm(cbind(Total, Survived.g) ~ Pclass + Sex + Age.g + Sex:Pclass , family = binomial(link = "probit"), data=titanic.t.g2) †
aj2.bi.cloglog	glm(cbind(Total, Survived.g) ~ Pclass + Sex + Age.g, family = binomial(link = "cloglog"), data=titanic.t.g2) †
aj3.bi.cloglog	glm(cbind(Total, Survived.g) ~ Pclass + Sex + Age.g + Sex:Pclass, family = binomial(link = "cloglog"), data=titanic.t.g2) †

* El data.frame titanic.t contiene las variables: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked y Title.

† De igual forma el data.frame titanic.t.g2 contiene las variables: Survived.g, Total, Pclass, Sex, Age.g, SibSp, Parch, Fare.g, Embarked y Title

Table 6: Validación modelos propuestos

Modelo	Chi.sq	Deviance	Null.Deviance	Shapiro.test	Tendencia	Residuos.extremos	AIC
aj2.bi.logit	0.20	9.86	117.69	0.234795	No	No	68.99
aj3.bi.logit	0.55	3.98	117.69	0.998905	No	No	67.11
aj2.bi.probit	0.22	9.46	117.69	0.241068	No	No	68.58
aj3.bi.probit	0.60	3.68	117.69	0.999903	No	No	66.81
aj2.bi.cloglog	0.29	8.56	117.69	0.364982	No	No	67.69
aj3.bi.cloglog	0.71	2.95	117.69	0.998646	No	No	66.08

Note:

** P.value <= 0.05, **** P.value <= 0.0001. Los residuos extremos implican un valor absoluto mayor a 2

En base a los resultados mostrados en la tabla 6, todos los modelos propuestos cumplen las condiciones de aplicabilidad. De entre estos el más optimo parece ser el modelo aj3.bi.cloglog pues posee el menor AIC y Deviance.

f) Validación cruzada.

Table 7: Resultados de validación cruzada mediante Leave-one-Out

Modelo	MSEP	MSEP.corregido
aj2.bi.logit	0.00985521	0.00953838
aj3.bi.logit	0.01081207	0.01042545
aj2.bi.probit	0.00947141	0.00918217
aj3.bi.probit	0.00998223	0.00963860
aj2.bi.cloglog	0.00864231	0.00841692
aj3.bi.cloglog	0.00776880	0.00753225

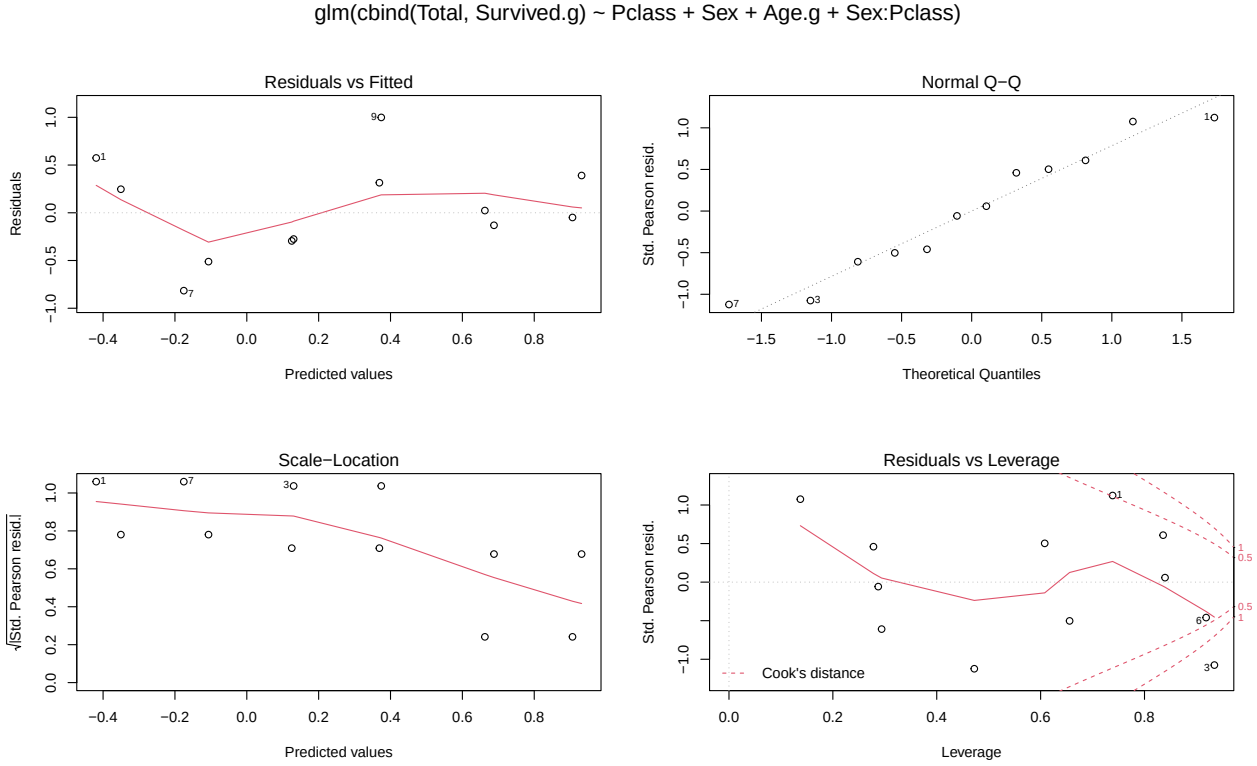
Si se observa la tabla número 7 se verá como el modelo, con un menor MSEP es el aj3.bi.cloglog que ya habia sido escogido como modelo optimo en base a su AIC y Devianze así como por respetar las condiciones de aplicabilidad.

g) Modelo final e interpretación de los coeficientes.

Así la modelización del dataset titanic::titanic_train da como resultado un modelo lineal generalizado que sigue una distribución binomial, con un link de tipo cloglog $log(-log(1-\pi))$. La figura 3 muestran una serie de plots de los residuos que pueden ser usados para diagnosticar el modelo. Si se observa el panel Residuals vs Fitted se podra apreciar la carencia de una tendencia marcada así como de residuos extremos, valor absoluto

mayor a 2. A traves del panel Normal Q-Q podemos garantizar la normalidad de los residuos. Además en el panel Residual vs Leverage se puede apreciar como las observaciones 1 y 3, que se corresponden con las mujeres de primera clase en un rango de edad de los 0 a los 40 años así como las de tercera clase de la misma edad.

Figura 3: Diagnoístico del modelo escogido



Si se atiende ahora a los coeficientes, estos son:

```
aj3.bi.cloglog$coefficients
```

```
##     (Intercept)         Pclass2         Pclass3         Sexmale    Age.g(40,80]
##     -0.41917125      0.06875460      0.54957318      0.54432997      0.24384067
## Pclass2:Sexmale Pclass3:Sexmale
##      0.46873457      0.01327022
```

La interpretación de estos radica en que a mayor clase, entiendose el máximo en la primera clase, menor probabilidad de morir. A su vez a mayor edad menor probabilidad de superviviencia. Tambien es interesante como el hecho de ser hombre aumenta tambien las probabilidades de morir en el Titanic. Si realizamos predicciones esto resulta aun más aparente si se observan las predicciones.

```
predict(aj3.bi.cloglog, newdata = data.frame(Pclass="3", Sex="male", Age.g="(40,80]"), type="response")
```

```
##         1
## 0.9210692
```

```
predict(aj3.bi.cloglog, newdata = data.frame(Pclass="3", Sex="female", Age.g="(40,80]"), type="response
```

```
##         1
## 0.7663404
```

```
predict(aj3.bi.cloglog, newdata = data.frame(Pclass="1", Sex="male", Age.g="(40,80]"), type="response")
```

```
##         1
```

```
## 0.7645571
```

```
predict(aj3.bi.cloglog, newdata = data.frame(Pclass="1", Sex="female", Age.g="(40,80]"), type="response
```

```
##         1
## 0.5679351
```

Modelización banco de datos quejas.dat

a) Identificación de variables y principales características

Se han creado dos datasets quejas.fi1 y quejas.fi2, en cada uno de ellos se ha eliminado la variable consultas y horas respectivamente, sustituyéndose esta por una variable nueva que sera el cociente entre la variable hora y consultas. Esto se ha realizado con el objetivo de minimizar posibles efectos de colinealidad puesto que la variable consultas y horas se encuentran altamente correlacionadas (correlación de Pearson 0.83)

Table 8: Variables usadas en el modelo

Variables	Tipo	Subtipo	Rol	Descripción
quejas	Cuantitativa	Discreta	respuesta	Quejas que ha recebidio el médico
residente	Categórica	Nominal	explicativa	Si ha sido residente o no en urgencias
sexo	Categórica	Nominal	explicativa	sexo del individuo
ingresos	Cuantitativa	Continua	explicativa	ingresos del individuo
horas	Cuantitativa	continua	explicativa	horas de trabajo realizadas por el médico
tiempo.consulta	Cuantitativa	Continua	explicativa	media de horas dedicadas a cada consulta
consultas	Cuantitativa	discreta	explicativa	consultas realizadas por el médico

b) Análisis descriptivo de los datos

Table 9: Descriptiva numérica de los datos

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
quejas	1	44	3.34	2.77	2.00	2.94	1.48	0.00	11.00	11.00	1.14	0.24	0.42
residente*	2	44	1.45	0.50	1.00	1.44	0.00	1.00	2.00	1.00	0.18	-2.01	0.08
sexo*	3	44	1.27	0.45	1.00	1.22	0.00	1.00	2.00	1.00	0.99	-1.05	0.07
ingresos	4	44	260.14	32.64	258.49	258.87	36.26	206.42	334.94	128.52	0.28	-0.76	4.92
horas	5	44	1417.40	326.98	1512.00	1440.88	264.09	589.00	1917.25	1328.25	-0.72	-0.40	49.29
tiempo.consulta	6	44	0.60	0.09	0.61	0.60	0.07	0.44	0.81	0.37	0.32	-0.22	0.01
consultas	7	44	2385.57	627.27	2384.50	2397.00	553.75	879.00	3763.00	2884.00	-0.15	-0.14	94.56

Si se observa la figura 5, se podrá apreciar como existe una fuerte relación entre el número de quejas entre el hecho de que el facultativo sea residente o no, en el número de horas trabajadas, con el número de consultas, así como de la variable tiempo.consulta.

Figura 4: Histogramas variables cuantitativas

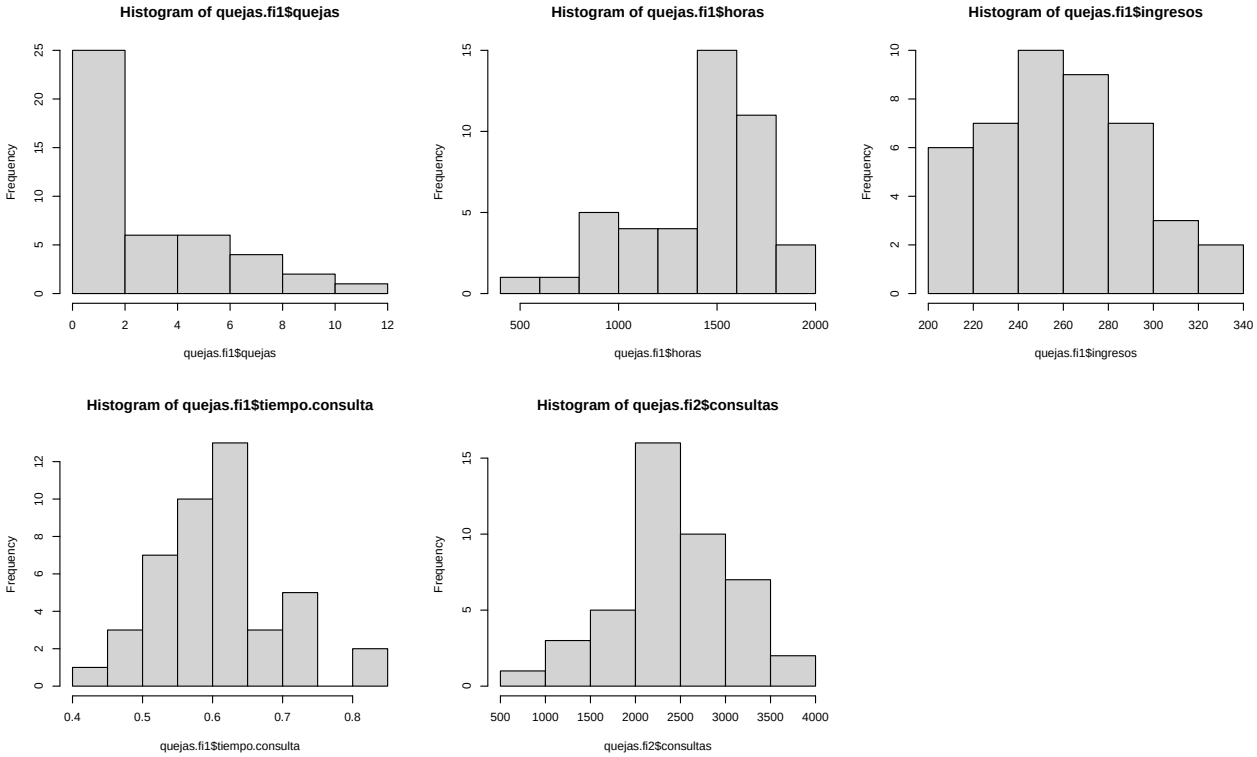
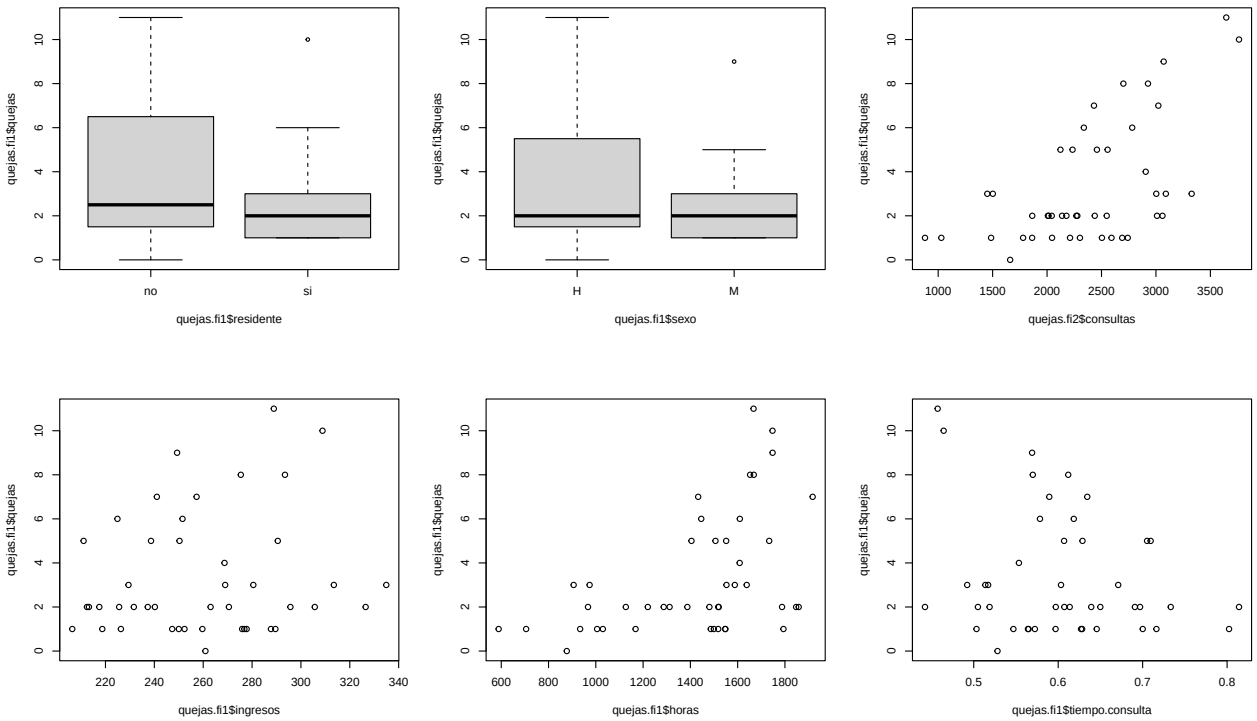


Figura 5: Relación entre variables y número de quejas



c) Describe la distribución de probabilidad que se pueda asumir para la variable respuesta ¿Cuál es el parámetro de interés?

Si se asume que la información recogida para cada observación se ha realizado en la misma cantidad de tiempo, se puede asumir que la variable quejas, sigue una distribución Poisson, siendo q cada evento:

$$- \textit{Función de probabilidad} : f(q \mid \lambda) = \frac{e^{\lambda}\lambda^{q}}{q!} \; si \; q = 0, 1, ... \tag{7}$$

$$- \textit{Rango del parámetro} : \lambda > 0 \tag{8}$$

$$- MediayVarianza : E(quejas) = Var(quejas) = \lambda \tag{9}$$

Se utiliza el logaritmo como link entre el predictor lineal y la variable respuesta.

d) Modelo lineal generalizado propuesto y validación de este.

Table 10: Modelos propuesto

Nombre	GLM_call
aj1.poi.log.horas	glm(quejas ~ ., family = poisson, data = quejas.fi1) *
aj1.poi.log.consultas	glm(quejas ~ ., family = poisson, data = quejas.fi2) †
aj1.poi.log.original	glm(quejas ~ ., family = poisson, data=quejas) ‡

* El data.frame quejas.fi1 contiene las variables: quejas, residente, sexo, ingresos, horas y tiempo.consulta.

† De igual forma el data.frame quejas.fi2 contiene las variables: quejas, residente, sexo, ingresos, consultas y tiempo.consulta.

‡ Por último el data.frame quejas contiene las variables: quejas, residente, sexo, ingresos, horas y consultas.

Table 11: Validación modelos propuestos

Modelo	Chi.sq	Deviance	Null.Deviance	Shapiro.test	Tendencia	Residuos.extremos	AIC	Dispersion.test
aj1.poi.log.horas	0.06	52.57	89.45	0.0404 *	Si	No	187.35	0.19
aj1.poi.log.consultas	0.09	50.03	89.45	0.0444 *	S1	No	184.81	0.25
aj1.poi.log.original	0.09	49.99	89.45	0.0519	Si	No	184.77	0.24

Note:

* P.value <= 0.05, ** P.value <= 0.01. Los residuos extremos implican un valor absoluto mayor a 2

Atendiendo a la Tabla 11, se ve como los modelos aj1.poi.log.horas y aj1.poi.log.consultas poseen residuos en los que se encuentra evidencia como para descartar la hipótesis nula de normalidad en estos. El único que cumple los requisitos de aplicabilidad, aj1.poi.log.original, es además el mejor modelo en términos de AIC y Deviance.

e) Mejora de los modelos propuestos

Partiendo de los modelos completos mostrados en la tabla 10, se aplicará la función step con el algoritmo backward, con el objetivo de encontrar el mejor modelo según esta metodología. Además sobre los modelos resultantes, que son denotados como aj2.poi.***, con la ayuda de la función anova, con test="Chi", se han introducido interacciones así como transformaciones de las variables. Los resultados se muestran en la tabla 12 y su validación en la tabla 13. Es necesario destacar que el modelo ofrecido por la función step sobre aj1.poi.log.consultas y aj1.poi.log.original es el mismo. Así se elimina la "rama" original.

Table 12: Modelos propuesto

Nombre	GLM_call
aj2.poi.horas	glm(quejas ~ residente + horas + tiempo.consulta, poisson, quejas.fi1)
aj3.poi.horas	glm(quejas ~ residente + horas + tiempo.consulta + horas:tiempo.consulta, poisson, quejas.fi1)
aj4.poi.horas	glm(quejas ~ residente + horas + tiempo.consulta + horas:tiempo.consulta + horas:residente , poisson, quejas.fi1)
aj2.poi.consultas	glm(quejas ~ residente + consultas, poisson, quejas.fi2)
aj3.poi.consultas	glm(quejas ~ residente * consultas, poisson, quejas.fi2)
aj4.poi.consultas	glm(quejas ~ residente * log(consultas), poisson, quejas.fi2)

Table 13: Validación modelos propuestos

Modelo	Chi.sq	Deviance	Null.Deviance	Shapiro.test	Tendencia	Residuos.extremos	AIC	Dispersion.test
aj2.poi.horas	0.08	53.07	89.45	0.0783	Si	No	183.85	0.18
aj3.poi.horas	0.18	46.97	89.45	0.2036	No	No	179.75	0.47
aj4.poi.horas	0.36	40.63	89.45	0.3955	No	No	175.41	0.60
aj2.poi.consultas	0.14	50.88	89.45	0.1401	No	No	179.66	0.23
aj3.poi.consultas	0.22	46.51	89.45	0.075	No	No	177.29	0.32
aj4.poi.consultas	0.22	46.68	89.45	0.3218	No	Si	177.46	0.33

Note:

* P.value <= 0.05, ** P.value <= 0.01. Los residuos extremos implican un valor absoluto mayor a 2

A excepción de aj2.poi.horas y aj4.poi.consultas todos los modelos mostrados en la tabla 13 cumplen las condiciones de aplicabilidad. En términos de AIC y Deviance el mejor modelo es aj4.poi.horas.

f) Validación cruzada.

Table 14: Resultados de validación cruzada mediante Leave-one-Out

Modelo	MSEP	MSEP.corregido
aj2.poi.horas	5.121773	5.107597
aj3.poi.horas	4.866277	4.845248
aj4.poi.horas	7.939652	7.843684
aj2.poi.consultas	4.487112	4.478116
aj3.poi.consultas	5.094280	5.075845
aj4.poi.consultas	4.769272	4.758171

Si se presta atención a la tabla 14, se podrá ver como el mejor modelo en términos de su calidad predictiva, evaluada mediante Leav-one-Out, es el aj2.poi.consultas. Resulta relevante la observación de que el mejor modelo en términos de calidad de ajuste, aj4.poi.horas es el peor en términos de calidad predictiva, seguramente debido a sobreajuste.

g) Modelo final e interpretación de los coeficientes.

Así la modelización del dataset quejas.dat da como resultado un modelo lineal generalizado que sigue una distribución Poisson, con link logarítmico, el modelo aj2.poi.consultas. La figura 6 muestra una serie de plots de los residuos que pueden ser usados para diagnosticar el modelo. Si se observa el panel Residuals vs Fitted se podrá apreciar la carencia de una tendencia marcada así como de residuos extremos, valor absoluto mayor a 2. A traves del panel Nomral Q-Q podemos garantizar la normalidad de los residuos. Por último en el panel Residual vs Leverage no se pueden identificar observaciones que posean una gran influencia sobre el modelo atendiendo al criterio de la distancia de Cook.

Si se presta atención a los coeficientes, estos son:

```
##   (Intercept)   residentesi     consultas
## -0.7274573833 -0.3121729222  0.0008101315
```

Así el hecho de ser residente disminuye en exp(-0.3121) unidades el número de quejas recibidas por el facultativo, además el aumento en una consulta aumenta el número de quejas recibidas por el médico en exp(0.0008101). Estos resultados son lógicos pues la resistencia implica más experiencia y es razonable asumir que un médico con mayor experiencia será capaz de lidiar mejor con los pacientes. Si atendemos al coeficiente de consultas también resulta necesario pues a mayor consultas mas cansanció y mayor probabilidad de dar un mal serivicio al paciente. Esto resulta evidente si realizamos predicciones:

```
predict(aj2.poi.consultas, newdata = data.frame(residente="si", consultas=1700), type="response")
```

```
##       1
## 1.40157
```

```
predict(aj2.poi.consultas, newdata = data.frame(residente="no", consultas=1700), type="response")
```

```
##        1
## 1.915093
```

```
predict(aj2.poi.consultas, newdata = data.frame(residente="si", consultas=3000), type="response")
```

```
##       1
## 4.01792
```

```
predict(aj2.poi.consultas, newdata = data.frame(residente="no", consultas=3000), type="response")
```

```
##        1
## 5.490049
```

Figura 6: Diagnoístico del modelo escogido