Modulo 5: Técnicas Avanzadas de Predicción

Modelos Lineales Generalizados

Leandro Gutierrez

11/10/2024

Descripción de la tarea

Descripción de la tarea

1. Propón un modelo lineal logit en el que la variable respuesta (crédito bueno=0, crédito malo=1), lo expliquen el resto de variables.
2. Interpreta la variable duration. ¿Es significativa? ¿A partir de qué nivel de significación deja de ser significativa?
3. Si eliminamos la variable amount del modelo, ¿crees que alguna otra variable incrementaría el sesgo provocado por la falta de amount en el modelo? Es decir, identifica el sesgo en otra variable producido por eliminar la variable amount.
4. Identifica efectos no lineales en la variable duration y amount. Interpreta los nuevos resultados después de meter, en el modelo, estas no linealidades.
5. ¿Cuál es la probabilidad estimada media de que el crédito sea malo para mayores de 50 años?
6. ¿Crees que hay discriminación de género en este último modelo creado?
7. Propón un modelo Ridge para modelizar el fenómeno crediticio. ¿Cuál es el lambda que minimiza el error? Compara este modelo con el logit que teníamos, anteriormente, con la curva ROC.

Solución

Carga de los datos

```
# cargamos el dataset
df <- read.table('https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data')

colnames(df) <- c("chk_acct", "duration", "credit_his", "purpose", "amount",
"saving_acct", "present_emp", "installment_rate", "sex", "other_debtor", "present_resid",
"property", "age", "other_install", "housing", "n_credits", "job", "n_people", "telephone",
"foreign", "response")


df$response <- df$response - 1

df$response.old <- df$response

df$response  <- as.factor(df$response)

summary(df)
```

```
##    chk_acct            duration     credit_his          purpose
```

```
##  Length:1000        Min.   : 4.0   Length:1000        Length:1000
##  Class :character   1st Qu.:12.0   Class :character   Class :character
##  Mode  :character   Median :18.0   Mode  :character   Mode  :character
##                     Mean   :20.9
##                     3rd Qu.:24.0
##                     Max.   :72.0
##      amount      saving_acct        present_emp        installment_rate
##  Min.   :  250   Length:1000        Length:1000        Min.   :1.000
##  1st Qu.: 1366   Class :character   Class :character   1st Qu.:2.000
##  Median : 2320   Mode  :character   Mode  :character   Median :3.000
##  Mean   : 3271                                         Mean   :2.973
##  3rd Qu.: 3972                                         3rd Qu.:4.000
##  Max.   :18424                                         Max.   :4.000
##      sex            other_debtor       present_resid     property
##  Length:1000        Length:1000        Min.   :1.000   Length:1000
##  Class :character   Class :character   1st Qu.:2.000   Class :character
##  Mode  :character   Mode  :character   Median :3.000   Mode  :character
##                                        Mean   :2.845
##                                        3rd Qu.:4.000
##                                        Max.   :4.000
##       age         other_install        housing            n_credits
##  Min.   :19.00   Length:1000        Length:1000        Min.   :1.000
##  1st Qu.:27.00   Class :character   Class :character   1st Qu.:1.000
##  Median :33.00   Mode  :character   Mode  :character   Median :1.000
##  Mean   :35.55                                         Mean   :1.407
##  3rd Qu.:42.00                                         3rd Qu.:2.000
##  Max.   :75.00                                         Max.   :4.000
##      job               n_people      telephone          foreign
##  Length:1000        Min.   :1.000   Length:1000        Length:1000
##  Class :character   1st Qu.:1.000   Class :character   Class :character
##  Mode  :character   Median :1.000   Mode  :character   Mode  :character
##                     Mean   :1.155
##                     3rd Qu.:1.000
##                     Max.   :2.000
##  response  response.old
##  0:700     Min.   :0.0
##  1:300     1st Qu.:0.0
##            Median :0.0
##            Mean   :0.3
##            3rd Qu.:1.0
##            Max.   :1.0
```

Podemos observar que contamos con un set de credit de **1000 observaciones**, con **21 variables**. No se observan valores nulos.

Apartado 1

En primer lugar utilizaremos un modelo con todas nuestras variables independientes:

$$\hat{respònse} = \beta_0 + \beta_1 * chk\_acct + \beta_2 * duration + \beta_3 * credit\_his + \beta_4 * purpose + ... + \epsilon$$

```
# creamos modelo completo con todas las variables excepto la variable auxiliar response.old
modelo.1 <- glm(response~. - response.old, data=df, family=binomial(link="logit"))
```

```
# visualizamos el summary del modelo
pander(summary(modelo.1))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0,4005	1,084	0,3693	0,7119
chk_acctA12	-0,3749	0,2179	-1,72	0,0854
chk_acctA13	-0,9657	0,3692	-2,616	0,008905
chk_acctA14	-1,712	0,2322	-7,373	1,664e-13
duration	0,02786	0,009296	2,997	0,002724
credit_hisA31	0,1434	0,5489	0,2612	0,7939
credit_hisA32	-0,5861	0,4305	-1,362	0,1733
credit_hisA33	-0,8532	0,4717	-1,809	0,07047
credit_hisA34	-1,436	0,4399	-3,264	0,001099
purposeA41	-1,666	0,3743	-4,452	8,508e-06
purposeA410	-1,489	0,7764	-1,918	0,05516
purposeA42	-0,7916	0,261	-3,033	0,002421
purposeA43	-0,8916	0,2471	-3,609	0,0003078
purposeA44	-0,5228	0,7623	-0,6858	0,4928
purposeA45	-0,2164	0,55	-0,3934	0,694
purposeA46	0,03628	0,3965	0,09152	0,9271
purposeA48	-2,059	1,212	-1,699	0,0893
purposeA49	-0,7401	0,3339	-2,216	0,02667
amount	0,0001283	4,444e-05	2,887	0,003894
saving_acctA62	-0,3577	0,2861	-1,25	0,2111
saving_acctA63	-0,3761	0,4011	-0,9376	0,3485
saving_acctA64	-1,339	0,5249	-2,551	0,01073
saving_acctA65	-0,9467	0,2625	-3,607	0,00031
present_empA72	-0,06691	0,427	-0,1567	0,8755
present_empA73	-0,1828	0,4105	-0,4454	0,656
present_empA74	-0,831	0,4455	-1,866	0,06211
present_empA75	-0,2766	0,4134	-0,6691	0,5034
installment_rate	0,3301	0,08828	3,739	0,0001846
sexA92	-0,2755	0,3865	-0,7127	0,476
sexA93	-0,8161	0,3799	-2,148	0,03172
sexA94	-0,3671	0,4537	-0,8091	0,4184
other_debtorA102	0,436	0,4101	1,063	0,2877
other_debtorA103	-0,9786	0,4243	-2,307	0,02107
present_resid	0,004776	0,08641	0,05527	0,9559
propertyA122	0,2814	0,2534	1,111	0,2666
propertyA123	0,1945	0,236	0,8243	0,4097
propertyA124	0,7304	0,4245	1,721	0,08531
age	-0,01454	0,009222	-1,576	0,115
other_installA142	-0,1232	0,4119	-0,2991	0,7649
other_installA143	-0,6463	0,2391	-2,703	0,006871
housingA152	-0,4436	0,2347	-1,89	0,05871
housingA153	-0,6839	0,477	-1,434	0,1517
n_credits	0,2721	0,1895	1,436	0,1511
jobA172	0,5361	0,6796	0,7889	0,4302
jobA173	0,5547	0,6549	0,847	0,397
jobA174	0,4795	0,6623	0,724	0,4691
n_people	0,2647	0,2492	1,062	0,2882
telephoneA192	-0,3	0,2013	-1,491	0,1361
foreignA202	-1,392	0,6258	-2,225	0,02609

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1221.7 on 999 degrees of freedom
Residual deviance:	895.8 on 951 degrees of freedom

```
AIC(modelo.1)
```

```
## [1] 993.8178
```

Para este primer modelo podemos observar que obtuvimos un AIC (Criterio de Información de Akaike) de **993.82**.

Vamos a intentar mejorarlo, para ello utilizaremos la función `stepAIC` del paquete `MASS` el cual realiza una reducción de nuestro modelo, descartando variables menos significativas generando la menor perdida de información posible, en este caso utilizaremos el metodo hibrido (**both**) para la selección de las variables

```
# utilizamos stepAIC para encontrar un modelo simplificado
modelo.aic <- stepAIC(modelo.1, trace=FALSE, direction="both", scope=respuesta~.)

# visualizamos la formula del modelo propuesto
pander(formula(modelo.aic))
```

response ~ chk_acct + duration + credit_his + purpose + amount + saving_acct + installment_rate + sex + other_debtor + age + other_install + housing + telephone + foreign

```
# vemos summary del modelo propuesto
pander(summary(modelo.aic))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1,75	0,7392	2,367	0,01794
chk_acctA12	-0,39	0,2121	-1,839	0,06593
chk_acctA13	-1,024	0,3626	-2,824	0,004739
chk_acctA14	-1,718	0,2281	-7,531	5,05e-14
duration	0,02568	0,00894	2,872	0,004074
credit_hisA31	-0,1188	0,5268	-0,2255	0,8216
credit_hisA32	-0,8303	0,4106	-2,022	0,04317
credit_hisA33	-0,9097	0,4657	-1,954	0,05075
credit_hisA34	-1,492	0,4324	-3,45	0,0005612
purposeA41	-1,607	0,3657	-4,395	1,108e-05
purposeA410	-1,435	0,7613	-1,885	0,05946
purposeA42	-0,7405	0,2534	-2,922	0,003475
purposeA43	-0,9195	0,2438	-3,772	0,0001619
purposeA44	-0,5251	0,7369	-0,7126	0,4761
purposeA45	-0,1424	0,5381	-0,2647	0,7912
purposeA46	0,1436	0,3916	0,3666	0,7139
purposeA48	-2,164	1,22	-1,774	0,0761
purposeA49	-0,7827	0,3272	-2,392	0,01675
amount	0,0001294	4,221e-05	3,066	0,002169
saving_acctA62	-0,3282	0,2767	-1,186	0,2355
saving_acctA63	-0,4304	0,3933	-1,094	0,2739
saving_acctA64	-1,289	0,5072	-2,542	0,01101
saving_acctA65	-0,9628	0,257	-3,746	0,0001794
installment_rate	0,3299	0,08554	3,857	0,0001148
sexA92	-0,2872	0,3763	-0,7632	0,4453
sexA93	-0,8228	0,3664	-2,246	0,02472

	Estimate	Std. Error	z value	Pr(>|z|)
sexA94	-0,4169	0,4449	-0,9372	0,3487
other_debtorA102	0,4874	0,3997	1,22	0,2226
other_debtorA103	-1,04	0,419	-2,483	0,01303
age	-0,01309	0,008398	-1,559	0,119
other_installA142	-0,07864	0,4033	-0,195	0,8454
other_installA143	-0,6995	0,235	-2,976	0,002916
housingA152	-0,4415	0,222	-1,989	0,04671
housingA153	-0,1497	0,3411	-0,4388	0,6608
telephoneA192	-0,2794	0,1842	-1,516	0,1294
foreignA202	-1,382	0,6207	-2,227	0,02593

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1221.7 on 999 degrees of freedom
Residual deviance:	910.5 on 964 degrees of freedom

```
AIC(modelo.aic)
```

```
## [1] 982.498
```

Vemos que el AIC para este nuevo modelo es de **982.5**, una mejora de mas de 11 puntos. Por lo que continuaremos con este nuevo modelo como el propuesto. También podemos notar que aumentó la suma de los residuos al cuadrado, pasando de un valor original de **895.82** a **910.5** con el nuevo modelo propuesto.

Utilizaremos el modelo que nos entrega el algortimo Stepwise (modelo.aic) como base para el resto del trabajo.

Para analizar la calidad de nuestro modelo primero veamos la matriz de confusión para nuestro modelo utilizando un punto de corte (threshold) del 0.5 como primera aproximación

```
# generamos las predicciones sobre el mismo dataset que tenemos
predicciones <- predict(modelo.aic, newdata = df, type = "response")

# utilizamos un threshold de 0.5
predicciones <- ifelse(predicciones > 0.5, 1, 0)

# creamos la tabla de confusión
tabla_confusion <- table(Predicción = predicciones, Real = df$response)

# visualizamos resultados
pander(tabla_confusion)
```

	0	1
0	627	143
1	73	157

```
# obtenemos tp tn fp y fn
tp <- tabla_confusion[2, 2]
tn <- tabla_confusion[1, 1]
fp <- tabla_confusion[2, 1]
fn <- tabla_confusion[1, 2]
```

```r
# calculamos sensibilidad
sensibilidad <- tp / (tp + fn)

sensibilidad
```

```
## [1] 0.5233333
```

```r
# calculamos especificidad
especificidad <- tn / (tn + fp)

especificidad
```

```
## [1] 0.8957143
```

Podemos observar que contamos con una sensibilidad demasiado baja, del ordel de los **0.52** y siendo que nos interesa **minimizar la cantidad de crédito calificado como bueno cuando en realidad es malo**, intentaremos minimizar los **falsos negativos** modificando iterativamente nuestro punto de corte. Como resultado de este proceso nuestra **sensibilidad** debe incrementar, en detrimento de la **especificidad**.

```r
# generamos las predicciones sobre el mismo dataset que tenemos
predicciones <- predict(modelo.aic, newdata = df, type = "response")

# utilizamos un threshold de 0.3
predicciones <- ifelse(predicciones > 0.3, 1, 0)

# creamos la tabla de confusión
tabla_confusion <- table(Predicción = predicciones, Real = df$response)

# obtenemos tp tn fp y fn
tp <- tabla_confusion[2, 2]
tn <- tabla_confusion[1, 1]
fp <- tabla_confusion[2, 1]
fn <- tabla_confusion[1, 2]

# visualizamos resultados
pander(tabla_confusion)
```

	0	1
0	520	74
1	180	226

```r
# calculamos sensibilidad
sensibilidad <- tp / (tp + fn)

sensibilidad
```

```
## [1] 0.7533333
```

```r
# calculamos especificidad
especificidad <- tn / (tn + fp)

especificidad
```

```
## [1] 0.7428571
```

```r
# veamos el AUC del modelo
auc(df$response, predicciones)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.7481
```

Podemos considerar que con un AUC de **0.74** estamos ante un modelo aceptable.

Apartado 2

Para nuestro modelo de regresión lineal generalizado la variable `duration` posee un efecto marginal de **2.568e-02**, con un error estandard de **8.940e-03**. Además podemos notar que posee un z-value de **2.872**, y un p-value igual a **0.004074**, lo que indica que es estadisticamente significativa, con un **nivel de significación de 0.4%**. Recordemos que el p-value nos indica las probabilidades de obtener el valor calculado del coeficiente a partir de una hipotesis nula (H_0) cierta, donde el coeficiente de la variable es cero, es decir donde la variable no es significativa para el modelo. En este caso al obtener una probabilidad de 0.004074, nos está marcando lo poco probable de que el coeficiente sea despreciable para el modelo. Además que para que la variable deje de ser significativa tendríamos que ir a buscar valores de significación del orden de los **0.30%**

Apartado 3

A partir de la formula que nos entrega el modelo generado con `stepAIC` generamos la nueva, donde quitamos la variable `amount`

```r
# definimos la formula partir de la que entrega stepAIC excepto amount
formula.2 <- as.formula(response ~ chk_acct + duration + credit_his + purpose +
    saving_acct + installment_rate + sex + other_debtor + age +
    other_install + housing + telephone + foreign)

# creamos modelo
modelo.2 <- glm(formula.2, data=df, family=binomial(link="logit"))

# visualizamos el summary del modelo
pander(summary(modelo.2))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2,121	0,7247	2,926	0,003433
chk_acctA12	-0,338	0,2097	-1,612	0,1071
chk_acctA13	-1,066	0,3607	-2,957	0,00311
chk_acctA14	-1,685	0,2263	-7,448	9,485e-14
duration	0,0418	0,007312	5,716	1,089e-08
credit_hisA31	-0,2436	0,5213	-0,4673	0,6403
credit_hisA32	-0,9266	0,4059	-2,283	0,02245
credit_hisA33	-0,9914	0,462	-2,146	0,0319
credit_hisA34	-1,58	0,4283	-3,689	0,0002255
purposeA41	-1,431	0,3552	-4,029	5,597e-05
purposeA410	-1,181	0,702	-1,682	0,09257
purposeA42	-0,7347	0,2525	-2,909	0,00362
purposeA43	-0,9507	0,2425	-3,921	8,818e-05
purposeA44	-0,6097	0,7351	-0,8294	0,4069
purposeA45	-0,1659	0,5382	-0,3081	0,758
purposeA46	0,1231	0,3878	0,3173	0,751

	Estimate	Std. Error	z value	Pr(>|z|)
purposeA48	-2,153	1,204	-1,788	0,0738
purposeA49	-0,8172	0,3245	-2,518	0,01179
saving_acctA62	-0,3569	0,2754	-1,296	0,195
saving_acctA63	-0,4791	0,391	-1,225	0,2205
saving_acctA64	-1,296	0,5016	-2,583	0,009789
saving_acctA65	-0,9247	0,2547	-3,63	0,0002835
installment_rate	0,2259	0,0777	2,908	0,003642
sexA92	-0,3002	0,3728	-0,8054	0,4206
sexA93	-0,7794	0,3625	-2,15	0,03156
sexA94	-0,4852	0,4408	-1,101	0,2709
other_debtorA102	0,5788	0,3955	1,464	0,1433
other_debtorA103	-1,086	0,4162	-2,61	0,009052
age	-0,01279	0,008326	-1,536	0,1245
other_installA142	-0,08474	0,4033	-0,2101	0,8336
other_installA143	-0,6853	0,2338	-2,932	0,003371
housingA152	-0,4416	0,2206	-2,002	0,04526
housingA153	-0,1115	0,3397	-0,3283	0,7427
telephoneA192	-0,14	0,1766	-0,7924	0,4281
foreignA202	-1,267	0,5998	-2,113	0,03464

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1221.7 on 999 degrees of freedom
Residual deviance:	920.1 on 965 degrees of freedom

```
AIC(modelo.2)
```

```
## [1] 990.0677
```

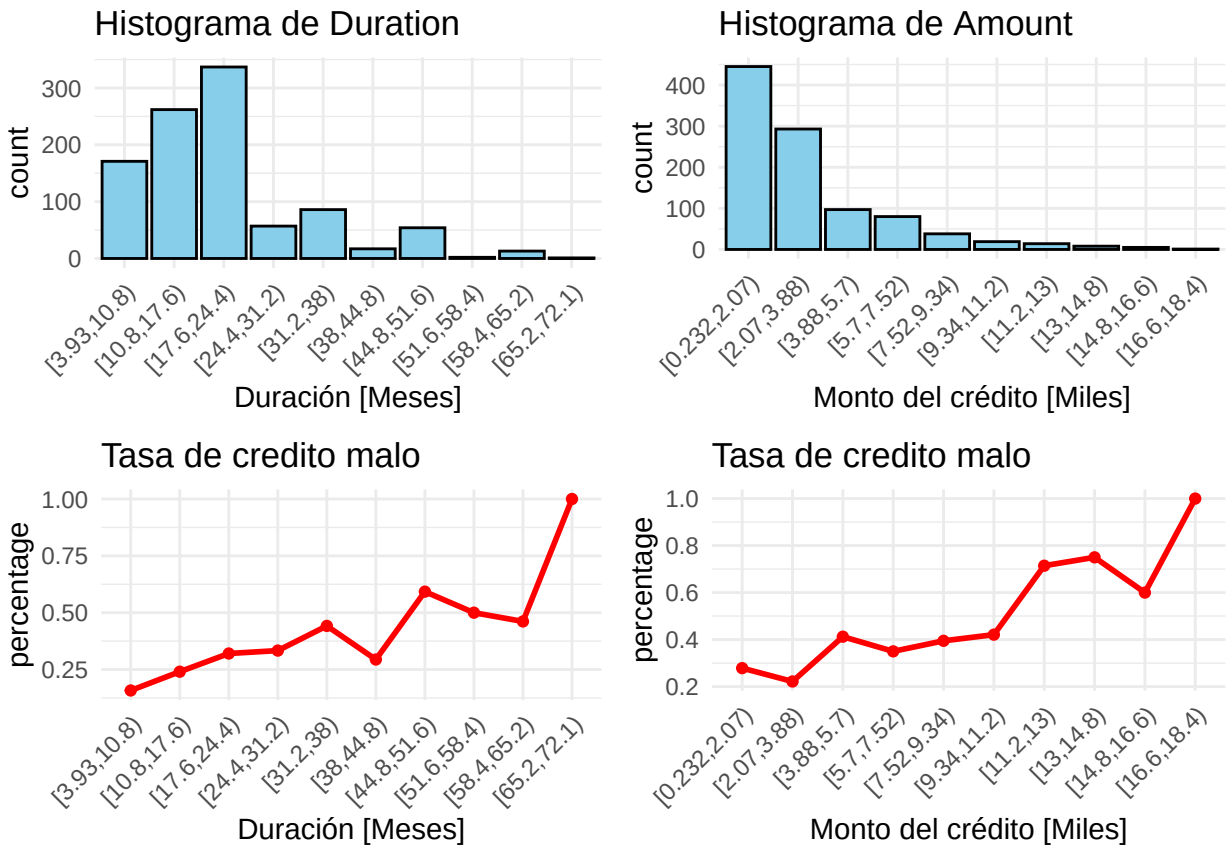
Se puede apreciar una pérdida en la calidad del modelo al eliminar la variable `amount`, esto lo notamos en el **AIC** del modelo, el cual pasó de un valor de **982.5** para el **modelo.aic** (resultado de la función stepAIC) con la variable `amount` incorporada, a un valor de **990.07** en este modelo donde se la excluye. Tambien se percibe un aumento en la suma de los residuos, que pasaron de un valor de **910.5** a **920.07** con la exclusión de la variable. Se puede observar que todos los betas se ven modificados por la eliminación de la variable `amount`.

Apartado 4

En primer lugar vamos a visualizar los histogramas de las variables que queremos analizar, en este caso `duration` y `amount`. Acompañando los histogramas anexamos las lines de tendencia correspondientes a los porcentajes de **creditos originalmente calificado como malo de cada intervalo**

```
# calculamos los bines para cada variable a analizar
df$duration.bin <- cut(df$duration, breaks = 10, right = FALSE)
df$amount.bin <- cut(df$amount/1000, breaks = 10, right = FALSE)

percentage.by.duration <- df %>%
  group_by(duration.bin) %>%
  summarise(percentage = sum(response.old)/n())

percentage.by.amount <- df %>%
  group_by(amount.bin) %>%
  summarise(percentage = sum(response.old)/n())
```

```r
# plot histograma de duration
plot1 <- ggplot(df, aes(x = duration.bin)) +
  geom_bar(aes(y = ..count..), fill = "skyblue", color = "black") +
  labs(title = "Histograma de Duration", x = "Duración [Meses]") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# plot histograma de amount
plot2 <- ggplot(df, aes(x = amount.bin)) +
  geom_bar(aes(y = ..count..), fill = "skyblue", color = "black") +
  labs(title = "Histograma de Amount", x = "Monto del crédito [Miles]") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# geomline del crédito calificado como malo en función de la duracion
plot3 <- ggplot(df, aes(x = duration.bin)) +
  geom_line(data = percentage.by.duration, aes(y = percentage), group = 1, color = "red", size = 1) +
  geom_point(data = percentage.by.duration, aes(y = percentage), group = 1, color = "red") +
  labs(title = "Tasa de credito malo", x = "Duración [Meses]") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# geomline del crédito calificado como malo en función del monto
plot4 <- ggplot(df, aes(x = amount.bin)) +
  geom_line(data = percentage.by.amount, aes(y = percentage), group = 1, color = "red", size = 1) +
  geom_point(data = percentage.by.amount, aes(y = percentage), group = 1, color = "red") +
  labs(title = "Tasa de credito malo", x = "Monto del crédito [Miles]") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

plot_grid(plot1, plot2, plot3, plot4, ncol = 2)
```



Podemos observar que ambas variables parecen guardar una relación no estrictamente lineal con la clasificación del crédito. Existen tendencias crecientes en la tasa de credito calificado como malo a medida que aumentan ambas variables, a partir de cierto punto la tasa parece incrementarse drasticamente, probablemente sugiriendo una relación cuadrática entre las variables.

Incorporaremos terminos cuadraticos para las variables `duration` y `amount` en nuestra formula y analizaremos los resultados del modelo

```
# definimos la formula partir de la que entrega stepAIC
formula.3 <- as.formula(response ~ chk_acct + duration +  I(duration^2) +
    amount + I(amount^2) +credit_his + purpose +
    saving_acct + installment_rate + sex + other_debtor + age +
    other_install + housing + telephone + foreign)

# creamos modelo
modelo.3 <- glm(formula.3, data=df, family=binomial(link="logit"))

# visualizamos el summary del modelo
pander(summary(modelo.3))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1,61	0,8074	1,994	0,04617
chk_acctA12	-0,3797	0,2129	-1,784	0,07447
chk_acctA13	-1,042	0,3632	-2,869	0,004124
chk_acctA14	-1,71	0,2286	-7,48	7,42e-14
duration	0,07161	0,02871	2,494	0,01261
I(duration^2)	-0,0007113	0,0004601	-1,546	0,1221

	Estimate	Std. Error	z value	Pr(>|z|)
amount	-4,909e-05	0,0001159	-0,4234	0,672
I(amount^2)	1,324e-08	7,997e-09	1,656	0,09771
credit_hisA31	-0,1113	0,5301	-0,21	0,8337
credit_hisA32	-0,8327	0,4142	-2,01	0,04439
credit_hisA33	-0,9068	0,4699	-1,93	0,05365
credit_hisA34	-1,475	0,4359	-3,384	0,0007139
purposeA41	-1,577	0,3731	-4,225	2,391e-05
purposeA410	-1,649	0,8238	-2,002	0,04527
purposeA42	-0,7125	0,2565	-2,777	0,005484
purposeA43	-0,8985	0,2441	-3,681	0,0002322
purposeA44	-0,5156	0,7366	-0,7	0,4839
purposeA45	-0,1663	0,5431	-0,3062	0,7595
purposeA46	0,1395	0,3918	0,356	0,7218
purposeA48	-2,18	1,226	-1,778	0,07542
purposeA49	-0,7739	0,3273	-2,365	0,01804
saving_acctA62	-0,3147	0,2767	-1,137	0,2555
saving_acctA63	-0,4604	0,3935	-1,17	0,242
saving_acctA64	-1,295	0,5052	-2,563	0,01039
saving_acctA65	-0,9558	0,2587	-3,694	0,0002206
installment_rate	0,2766	0,08967	3,085	0,002037
sexA92	-0,2842	0,38	-0,7479	0,4545
sexA93	-0,7814	0,3705	-2,109	0,03494
sexA94	-0,4203	0,4481	-0,9379	0,3483
other_debtorA102	0,4973	0,4034	1,233	0,2176
other_debtorA103	-1,042	0,4195	-2,484	0,013
age	-0,01231	0,008437	-1,459	0,1445
other_installA142	-0,07383	0,4044	-0,1826	0,8551
other_installA143	-0,6767	0,2363	-2,864	0,004185
housingA152	-0,4494	0,2234	-2,012	0,04423
housingA153	-0,125	0,3434	-0,3639	0,7159
telephoneA192	-0,2806	0,1851	-1,516	0,1294
foreignA202	-1,404	0,6472	-2,169	0,03007

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1221.7 on 999 degrees of freedom
Residual deviance:	906.3 on 962 degrees of freedom

```
AIC(modelo.3)
```

```
## [1] 982.2881
```

Al agregarle los terminos no lineales `amount^2` y `duration^2` no se aprecia una diferencia significativa en la calidad del modelo respecto al modelo base propuesto por el algortimo Stepwise. Para el modelo original propuesto (modelo.aic) obtuvimos un AIC **982.5**, mientras que para el modelo con los términos cuadráticos obtuvimos un AIC de **982.29** una mejora mejor a un punto, lo que podríamos considerar no significativa. Además en terminos de la **suma de los cuadrados de los residuos** pasamos de un valor de **910.5** en el modelo base, a un valor de **906.29** para el modelo con terminos cuadráticos.

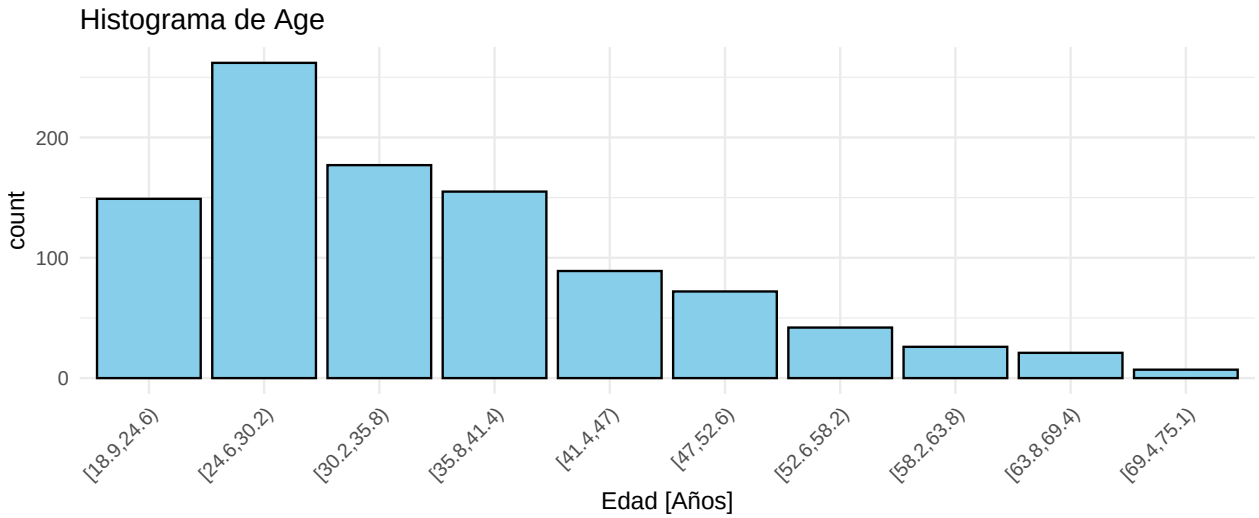
Apartado 5

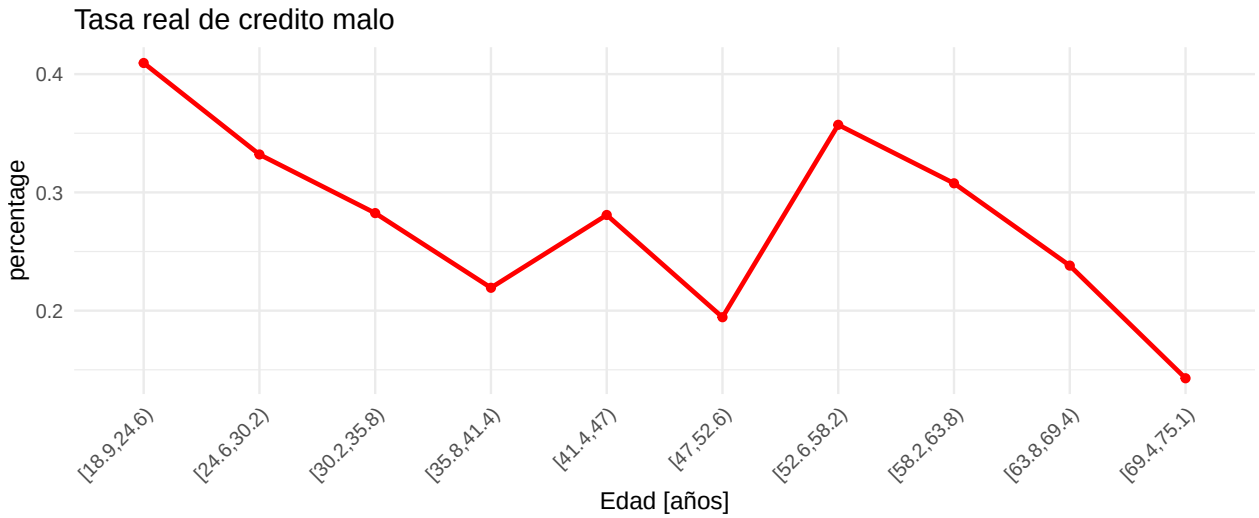
Para desarrollar este enunciado primero obtendremos las predicciones de nuestro modelo para el dataset de estudio y lo compararemos visualmente con las respuestas reales

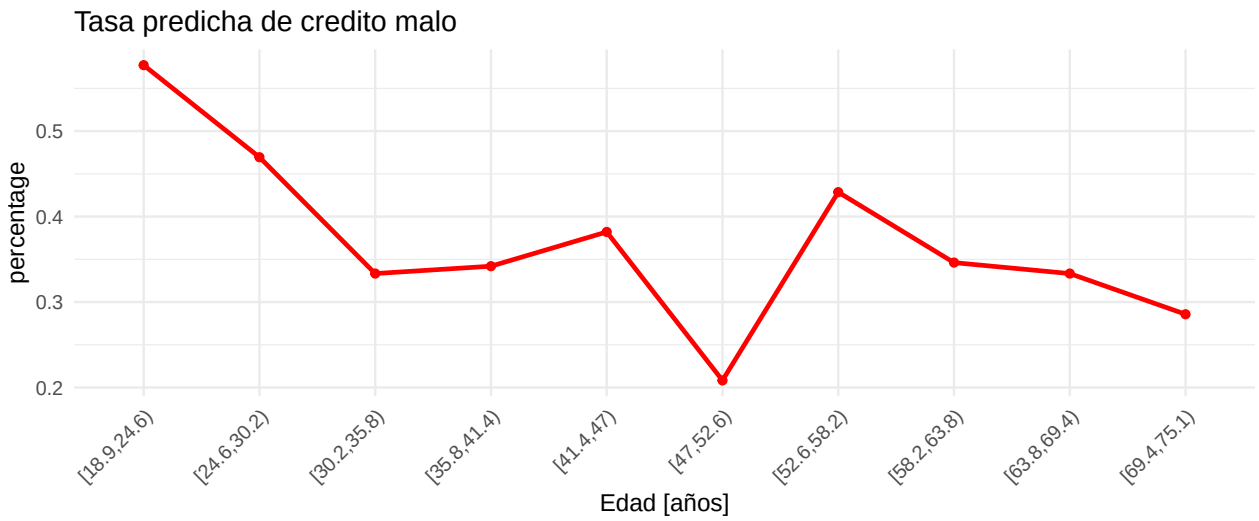
```
# calculamos las predicciones de nuestro modelo
predicciones <- predict(modelo.aic, newdata = df, type = "response")

# utilizamos el punto de corte determinado en el ejercicio anterior
predicciones <- ifelse(predicciones > 0.3, 1, 0)

# anexamos la columna de predicciones al dataframe original
df <- cbind(df, predicciones)
```

Ahora visualicemos las predicciones vs las respuestas reales en función de la edad de los postulantes al crédito

```
# creamos bines para el estudio visual
df$age.bin <- cut(df$age, breaks = 10, right = FALSE)

# calculamos la tasa de aceptación del crédito predicho
percentage.by.age <- df %>%
  group_by(age.bin) %>%
  summarise(percentage = sum(predicciones)/n())

# calculamos la tasa de aceptación real de crédito
real.percentage.by.age <- df %>%
  group_by(age.bin) %>%
  summarise(percentage = sum(response.old)/n())

# plot histograma de duration
plot5 <- ggplot(df, aes(x = age.bin)) +
  geom_bar(aes(y = ..count..), fill = "skyblue", color = "black") +
  labs(title = "Histograma de Age", x = "Edad [Años]") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# plot linea de tendencia para predicción
plot6 <- ggplot(df, aes(x = age.bin)) +
  geom_line(data = percentage.by.age, aes(y = percentage), group = 1, color = "red", size = 1) +
  geom_point(data = percentage.by.age, aes(y = percentage), group = 1, color = "red") +
  labs(title = "Tasa predicha de credito malo", x = "Edad [años]") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# plot linea de tendencia para credito real
plot7 <- ggplot(df, aes(x = age.bin)) +
  geom_line(data = real.percentage.by.age, aes(y = percentage), group = 1, color = "red", size = 1) +
  geom_point(data = real.percentage.by.age, aes(y = percentage), group = 1, color = "red") +
  labs(title = "Tasa real de credito malo", x = "Edad [años]") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

plot_grid(plot5, plot7, plot6, ncol = 1)
```








```r
# obtenemos subset de mayores de 50
df.over50 <- df %>% filter(age > 50)

# calculamos la media de las predicciones para el subset
```

```r
prob.over.50 <- mean(df.over50$predicciones)

# vemos resultado
prob.over.50
```

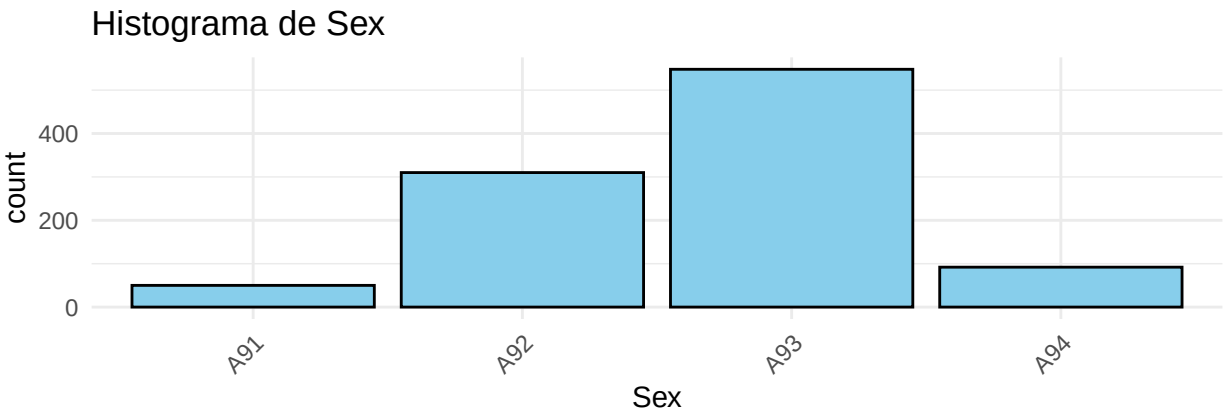
```
## [1] 0.3274336
```

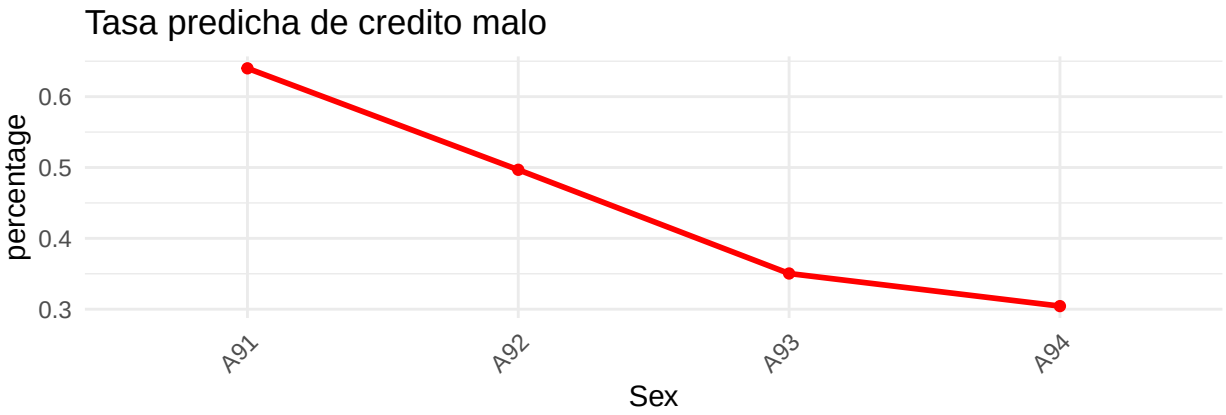
La probabilidad media de clasificación de un crédito como malo dado que el cliente tenga más de 50 años es de **0.3274**.

Apartado 6

Primero visualicemos las lineas de tendencia de las predicciones en función del sexo y el estado civil

```r
# calculamos la tasa de aceptación del crédito predicho
percentage.by.sex <- df %>%
  group_by(sex) %>%
  summarise(percentage = sum(predicciones)/n())

# plot histograma de duration
plot8 <- ggplot(df, aes(x = sex)) +
  geom_bar(aes(y = ..count..), fill = "skyblue", color = "black") +
  labs(title = "Histograma de Sex", x = "Sex") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# plot linea de tendencia para predicción
plot9 <- ggplot(df, aes(x = sex)) +
  geom_line(data = percentage.by.sex, aes(y = percentage), group = 1, color = "red", size = 1) +
  geom_point(data = percentage.by.sex, aes(y = percentage), group = 1, color = "red") +
  labs(title = "Tasa predicha de credito malo", x = "Sex") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

plot_grid(plot8, plot9, ncol = 1)
```





Sabiendo que la variable sex tiene el siguiente significado

Personal status and sex:

- A91: male - divorced/separated
- A92: female - divorced/separated/married
- A93: male - single
- A94: male - married/widowed
- A95: female - single

Vemos que paras las categorias A93 y A94 que representan **hombres solteros** y **hombres casados/viudos**, respectivameente, la tasa de credito malo predicho es menor respecto a la categoria A92 que representa **mujeres casadas/divorsiadas/viudas**, lo que podría indicar un sesgo vinculado al sexo

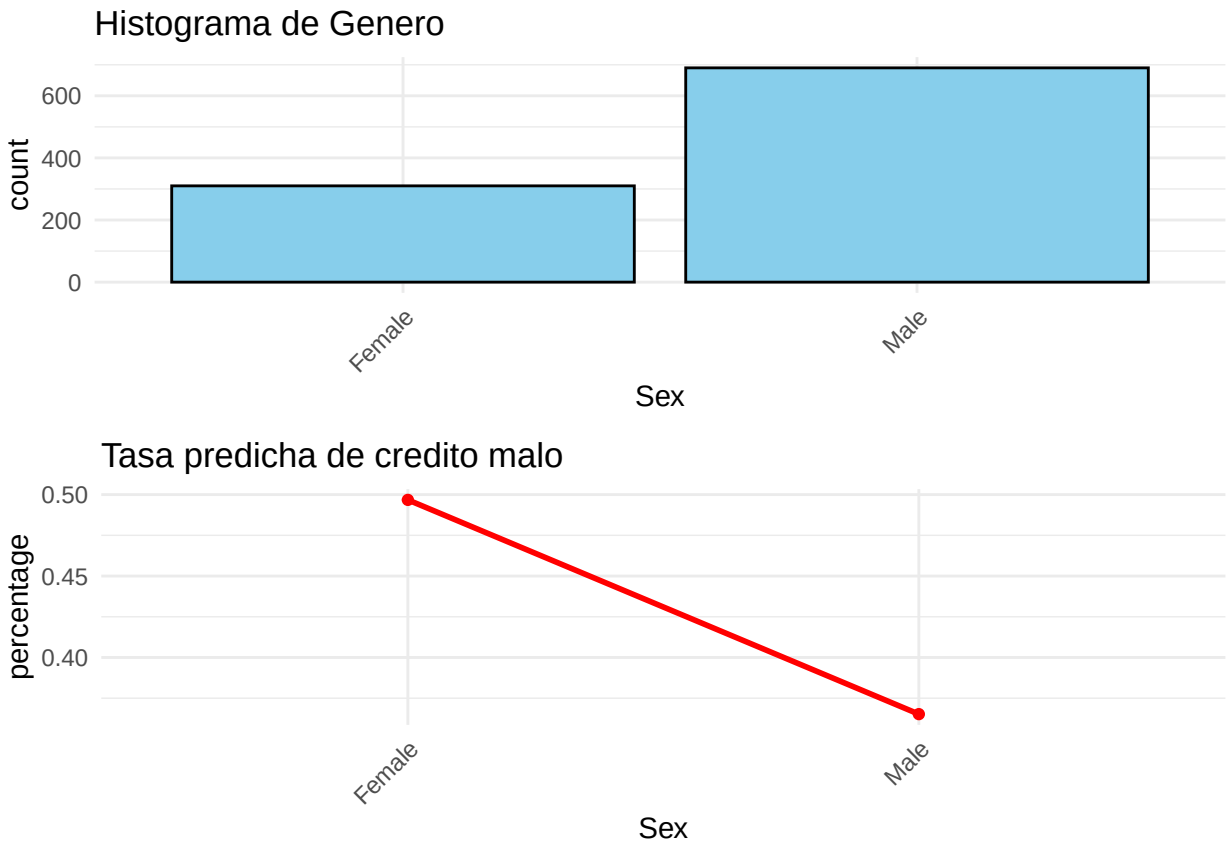
Agrupemos segun la definición que tenemos arriba solo en dos categorias **Male** y **Female**, para analizar si efectiavamente nuestro modelo está sesgado

```
# definimos nuestra variable dicotomica sex.binary
df <- df %>% mutate(sex.binary = if_else(sex %in% c("A91", "A93", "A94"), "Male", "Female"))

# calculamos la tasa de aceptación del crédito predicho
percentage.by.sex <- df %>%
  group_by(sex.binary) %>%
  summarise(percentage = sum(predicciones)/n())

pander(percentage.by.sex)
```

sex.binary	percentage
Female	0,4968

sex.binary	percentage
Male	0,3652

```r
# plot histograma de duration
plot8 <- ggplot(df, aes(x = sex.binary)) +
  geom_bar(aes(y = ..count..), fill = "skyblue", color = "black") +
  labs(title = "Histograma de Genero", x = "Sex") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# plot linea de tendencia para predicción
plot9 <- ggplot(df, aes(x = sex.binary)) +
  geom_line(data = percentage.by.sex, aes(y = percentage), group = 1, color = "red", size = 1) +
  geom_point(data = percentage.by.sex, aes(y = percentage), group = 1, color = "red") +
  labs(title = "Tasa predicha de credito malo", x = "Sex") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

plot_grid(plot8, plot9, ncol = 1)
```



Con estos resultados se nota la leve tendencia a calificar como bueno al sexo Masculino (Male) por sobre el Femenino (Female), con una diferencia porcuentual de al rededor del 13%, resultados que podríam indicar un modelo sesgado respecto al género del tomador del crédito.