UNAP
FINESI

IV Semestre

Modelos Discretos

Regresión Ordinal

Mtr. Alcides Ramos Calcina

Regresión Ordinal

Modelo de ODDs proporcionales

Mtr. Alcides Ramos Calcina

Introducción

- Cuando la variable de respuesta es cualitativa y toma valores en diferentes grupos o modalidades, además estos están configurados de forma **ordinal** surge el problema que se tratará en este capítulo.

- Los datos ordinales son datos que toman valores que tienen un **orden natural**, para los cuales los intervalos entre valores no tienen por que tener un significado.

- Por ejemplo, estado de salud: excelente, muy bueno, bueno, regular; las encuestas de satisfacción: mucho, poco, nada, etc. En este caso tendríamos una variable respuesta con una distribución ordinal, y los modelos que tendremos que ajustar van a ser los modelos de regresión ordinal.

1. Modelo de Odds Proporcionales

El modelo logístico ordinal que se va a desarrollar en este capítulo es el llamado de **Odds Proporcionales** o **modelo de odds proporcionales**, también conocido como el ***Modelo Logit Acumulado.***

Se ilustra la idea del modelo odds proporcional asumiendo que se tiene una variable respuesta con cinco categorías y considerando las cuatro posibles formas de dividirlas en sólo dos categorías respetando el orden natural.

0	1	2	3	4
0	1	2	3	4
0	1	2	3	4
0	1	2	3	4

1. Modelo de Odds Proporcionales

Generalmente, si una variable respuesta ordinal Y tiene K categorías ($Y = 0, 1. 2, \ldots, K\text{-}1$), entonces hay K - 1 formas de dicotomizar la respuesta:

$$Y \geq 1 \text{ ó } Y < 1,\ Y \geq 2 \text{ ó } Y < 2,\ \ldots,\ Y \geq K\text{-}1 \text{ ó } Y < K - 1$$

Para un suceso aleatorio S, se define su "odds" o "ventaja" como la razón entre la probabilidad de ocurrencia y la probabilidad de no ocurrencia.

Con la categorización de Y, se puede definir la "odds" de que $Y \geq k$ dividida por la probabilidad de que $Y < k$, es decir:

$$odds\left(Y \geq k\right) = \frac{P\left(Y \geq k\right)}{P\left(Y < k\right)}, \text{ donde } k = 1, 2, 3, \ldots, K\text{-}1$$

1. Modelo de Odds Proporcionales

Supongamos que tenemos una variable respuesta con cinco niveles y una variable explicativa dicotómica ($X = 0, X = 1$).

Entonces, bajo la suposición de *odds proporcionales*, el *odds ratio* que compara categorías iguales o mayores que 1 y categorías menores que 1 es el mismo que el que compara categorías mayores o iguales a 4 con categorías menores que 4. Formalmente:

$$OR(Y \geq 1) = \frac{odds(Y \geq 1 \mid X = 1)}{odds(Y \geq 1 \mid X = 0)} = \frac{odds(Y \geq 4 \mid X = 1)}{odds(Y \geq 4 \mid X = 0)} = OR(Y \geq 4)$$

En otras palabras, el odds ratio (OR) es invariante al punto utilizado para la dicotomización.

1. Modelo de Odds Proporcionales

Esto implica que:

- Si hay K categorías en la respuesta, solo hay un parámetro (β) para cada una de las variables predictoras o explicativas.
- Sin embargo, sigue habiendo constantes separadas (β_{k0}) para cada una de las K -1 comparaciones.

En Resumen

Comparación de parámetros

Variable	Parámetro
Constante	$\beta_{10}, \beta_{20}, \cdots, \beta_{(K-1)0}$
X_1	β_1

a) Ordinal

Variable	Parámetro
Constante	$\beta_{10}, \beta_{20}, \cdots, \beta_{(K-1)0}$
X_1	$\beta_{11}, \beta_{21}, \cdots, \beta_{(K-1)1}$

b) Policotómica

1.1. Modelo de regresión ordinal simple

Procedemos ahora a presentar la forma del *modelo odds proporcional* con una respuesta Y de K niveles ($Y = 0, 1, 2, \ldots, K - 1$) y una variable explicativa X_1.

$$P\left(Y \geq k \mid X_1\right) = \frac{1}{1 + e^{\beta_{k0} + \beta_1 X_1}} \quad , \qquad k = 1, 2, \ldots, K\text{-}1$$

Por tanto, la probabilidad de que la variable respuesta esté en una categoría inferior a k es:

$$P\left(Y < k \mid X_1\right) = \frac{e^{\beta_{k0} + \beta_1 X_1}}{1 + e^{\beta_{k0} + \beta_1 X_1}}$$

1.1. Modelo de regresión ordinal simple

El modelo en términos del odds de una desigualdad. Si sustituimos la formula $P(Y \geq k \mid X_1)$ por la expresión para el odds entonces:

$$odds(Y \geq k \mid X_1) = \frac{P(Y \geq k \mid X_1)}{1 - P(Y \geq k \mid X_1)} = \frac{P(Y \geq k \mid X_1)}{P(Y < k \mid X_1)}$$
$$= e^{\beta_{k0} + \beta_1 X_1} = e^{\beta_{k0}} e^{\beta_1 X}$$

El modelo se formula como la probabilidad de una desigualdad, esto es, que la variable respuesta Y sea mayor o igual a k.

- Modelo Odd Proporcional : $P(Y \geq k \mid X)$
- Modelo Logístico Estándar : $P(Y = k \mid X)$

1.2. Odd Ratio e Intervalo de Confianza

- **Odd Ratio (OR)**

Para evaluar el efecto de la variable explicativa sobre la variable respuesta formulamos el llamado odds ratio de $Y \geq k$ para comparar $X_1 = 0$ y $X_1 = 1$ (es decir, el *odds ratio* para $X_1 = 0$ vs. $X_1 = 1$).

$$OR\left(Y \geq k \mid X_1\right) = \frac{odds\left(Y \geq 1 \mid X_1 = 1\right)}{odds\left(Y \geq 1 \mid X_1 = 0\right)} = \frac{e^{\beta_{k0} + \beta_1}}{e^{\beta_{k0}}} = e^{\beta_1}$$

Es decir, la *odds ratio* es constante para cualquier punto de corte k considerado. Además, el coeficiente β_1 es:

$$\beta_1 = \log OR\left(Y \geq k \mid X_1\right) \quad \forall k$$

1.2. Odd Ratio e Intervalo de Confianza

- **Intervalo de Confianza (IC)**

$$IC_{95\%} = e^{\left\{\hat{\beta}_1 \pm 1.96 s_{\hat{\beta}_1}\right\}}$$

$$IC_{95\%} = \left[e^{\left\{\hat{\beta}_1 - 1.96 s_{\hat{\beta}_1}\right\}} \leq e^{\beta_1} \leq e^{\left\{\hat{\beta}_1 + 1.96 s_{\hat{\beta}_1}\right\}} \right] = 0.95$$

Donde:

$\hat{\beta}$: es el estimador de máxima-verosimilitud del modelo.

$s_{\hat{\beta}}$: es el error de estimación del mismo.

2. MODELO DE REGRESIÓN ORDINAL MÚLTIPLE

Expandir el modelo para añadir más variables explicativas se obtiene de forma directa, basta expandir el predictor lineal.

Representando por **X** el vector aleatorio de variables explicativas, el modelo se puede expresar por:

$$P\left(Y \geq k \mid \mathrm{X}\right) = \frac{1}{1 + e^{\beta_{k0} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_r X_r}} \quad , k = 1, 2, \ldots, K\text{-}1$$

También,

$$P\left(Y \geq k \mid \mathrm{X}\right) = \frac{1}{1 + e^{\beta_{k0} + \sum_{i=1}^{r} \beta_i X_i}}$$

2. MODELO DE REGRESIÓN ORDINAL MÚLTIPLE

El odds para la respuesta mayor o igual al nivel *k* sería el siguiente:

$$odds\left(Y \geq k \mid \mathrm{X}\right) = \frac{P\left(Y \geq k \mid \mathrm{X}\right)}{1 - P\left(Y \geq k \mid \mathrm{X}\right)} == e^{\beta_{k0} + \sum_{i=1}^{r} \beta_i X_i} = e^{\beta_{k0}} e^{\sum_{i=1}^{r} \beta_i X_i}$$

Como en la regresión logística estándar, el uso de múltiples variables independientes permite la estimación del odds ratio para una variable controlando los efectos de las demás variables explicativas del modelo.

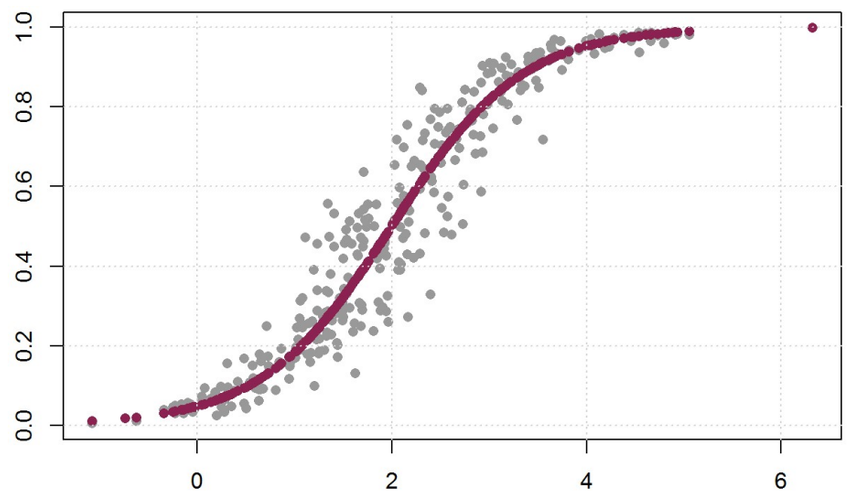
$$OR = e^{\beta_i} \qquad X_i \in \{0,1\}$$

3. LA FUNCION POLR EN R

Esta función forma parte de la librería **MASS**. Es importante tener en cuenta que esta función usa una parametrización que implica que los odds ratio son de estar en una categoría superior.

EJEMPLOS

$$P(Y \geq k \mid X_1) = \frac{1}{1 + e^{\beta_{k0} + \beta_1 X_1}}$$



Ejemplo 6

Vamos a considerar los datos de acidez de ciertos vinos. El conjunto de datos representa un experimento sobre ciertos factores que determinan la amargura del vino; donde:

1 = “amargura muy baja”
2 = “amargura baja”
3 = “amargura media”
4 = “amargura alta”
5 = “amargura muy alta”,

dos factores de tratamiento (temperatura y contacto) cada una con dos niveles, la temperatura y el contacto entre el zumo y las pieles de las uvas cuando se extrae de ellas. Nueve jueces evaluaron cada vino de dos botellas para cada una de las cuatro condiciones de tratamiento, por lo tanto, hay 72 observaciones en total. La variable objetivo Y a estudiar será la variable “**rating**” $\in \{1,...,5\}$ que es una categorización de la variable “**response**” la cual califica la acidez de los vinos.

Ejemplo 6

Los datos se encuentran el el archivo **wine.txt**, como se muestra a continuación:

Ejemplo 6

Solución

Vamos a ajustar el siguiente modelo acumulado para los datos wine:

$$\text{logit}\left[P\left(Y_i \leq k\right)\right] = \beta_{k0} - \beta_1 temp_i$$

$$i = 1, 2, \ldots, n \text{ y } k = 1, 2, \ldots, K\text{-}1$$

Este es un modelo para la probabilidad acumulada de que la clasificación i-ésima caiga sobre la categoría k-ésima o superior, donde i indica cada observación (n = 72) y los índices k = 1, 2, ..., K reflejan la categoría respuesta (K = 5). El parámetro β_{k0} es el punto de corte para el k-ésimo modelo acumulado, logit($P(Y_i \leq k)$).

Ejemplo 6

Antes de ajustar el modelo es necesario considerar que las variables categóricas deberán estar consideradas como factor; así mismo, la variable dependiente (**rating**) debe ser de tipo factor ordenado.

```
datos$temp <- factor(datos$temp)
datos$contact <- factor(datos$contact)
datos$rating <- factor(datos$rating, ordered = T)
attach(datos)
```

Lo que se puede verificar

```
str(datos[2:4])
'data.frame': 72 obs. of  3 variables:
 $ rating : Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 2 3 3 4 4 4 5 5 1 2
...
 $ temp   : Factor w/ 2 levels "cold","warm": 1 1 1 1 2 2 2 2 1 1 ...
 $ contact: Factor w/ 2 levels "no","yes": 1 1 2 2 1 1 2 2 1 1 ...
```

Ejemplo 6

Estimación del modelo. Modelizamos con el comando **polr**. Ahora veremos si las variables individuales son significativas.

```
Modelo1 <- polr(rating ~ temp, data = datos)
summary(modelo1)
Call:
polr(formula = rating ~ temp, data = datos)
Coefficients:
         Value Std. Error t value
tempwarm 2.287     0.5129   4.458

Intercepts:
    Value   Std. Error t value
1|2 -1.9361  0.4843    -3.9979
2|3  0.4351  0.3313     1.3133
3|4  2.4325  0.4576     5.3156
4|5  3.8270  0.5736     6.6720

Residual Deviance: 184.0269
AIC: 194.0269
```

Los parámetros se pueden interpretar de manera similar al caso de **regresión logística**, pero ahora hay dos transiciones en vez de una (que sería el caso de una variable dicotómica).

Ejemplo 6

El primer resultado es la tabla de coeficientes con estimación del parámetro, error estándar y el p-valor basado en el método de Wald. Las estimaciones mediante el método de máxima verosimilitud para los parámetros son:

$$\hat{\beta}_1 = 2.287 \quad \text{y} \quad \left\{\hat{\beta}_{k0}\right\} = \{-1.94, \quad 0.44, \quad 2.43, \quad 3.83\}$$

El coeficiente para la temperatura es positivo, lo que indica que una temperatura más alta aumenta la amargura del vino, es decir, la calificación en categorías superiores es más probable.

Ejemplo 6

Ahora estimamos el segundo modelo individual

```
modelo2 <- polr(rating ~ contact, data = datos)
summary(modelo2)

Call:
polr(formula = rating ~ contact, data = datos)

Coefficients:
           Value Std. Error t value
contactyes 1.207     0.4499   2.683

Intercepts:
    Value   Std. Error t value
1|2 -2.1393  0.4898    -4.3676
2|3  0.0426  0.3206     0.1328
3|4  1.7145  0.3864     4.4375
4|5  2.9788  0.5021     5.9330

Residual Deviance: 199.9118
AIC: 209.9118
```

El parámetro de **contact** también es positivo, lo que indica que a más contacto aumenta la amargura del vino.

Ejemplo 6

Calculamos la razón de verosimilitud para ver la significancia individual de cada variable.

```
modelo0 <- polr(rating ~ 1, data = datos)
anova(modelo0, modelo1)
Likelihood ratio tests of ordinal regression models

Response: rating
  Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
1     1         68   207.4382
2  temp         67   184.0269 1 vs 2     1  23.4113 1.308076e-06
```

El valor de probabilidad asociado al estadístico Chi-cuadrada $Pr(chi) = 0.0000013 < 0.05$, entonces rechazamos la hipótesis nula de que $\beta_1 = 0$. Es decir, que la variable **temp** es estadísticamente significativa.

Ejemplo 6

Modelo 2.

```
anova(modelo0, modelo2)
Likelihood ratio tests of ordinal regression models

Response: rating
    Model Resid. df Resid. Dev   Test    Df LR stat.     Pr(Chi)
1       1        68   207.4382
2 contact        67   199.9118 1 vs 2     1 7.526333 0.006080355
```

Del mismo modo, el p-valor Chi-cuadrada $Pr(chi) = 0.00608 < 0.05$, entonces rechazamos la hipótesis nula de que $\beta_2 = 0$. Es decir, que la variable **contact** influye significativamente en la variable de respuesta **rating**.

Ejemplo 6

Si introducimos las variables juntas, tenemos:

```
modelo3 <- polr(rating ~ temp + contact, data = datos)
summary(modelo3)
Call:
polr(formula = rating ~ temp + contact, data = datos)

Coefficients:
            Value Std. Error t value
tempwarm    2.503     0.5287   4.735
contactyes 1.528     0.4766   3.205

Intercepts:
    Value   Std. Error t value
1|2 -1.3444  0.5171    -2.5998
2|3  1.2508  0.4379     2.8565
3|4  3.4669  0.5978     5.7998
4|5  5.0064  0.7309     6.8496
Residual Deviance: 172.9838
AIC: 184.9838
```

Los coeficientes para la temperatura y el contacto siguen siendo positivos en el modelo múltiple, lo que indica que una temperatura más alta y más contacto aumenta la amargura del vino.

Ejemplo 6

Verificción a través del intervalo de confianza para los parámetros.

```
confint(modelo3, level = 0.95)
Waiting for profiling to be done...

Re-fitting to get Hessian

                2.5 %    97.5 %
tempwarm    1.5097621 3.595211
contactyes 0.6158072 2.492428
```

Se observa que los β para temperatura y contacto no contemplan el valor de cero.

Ejemplo 6

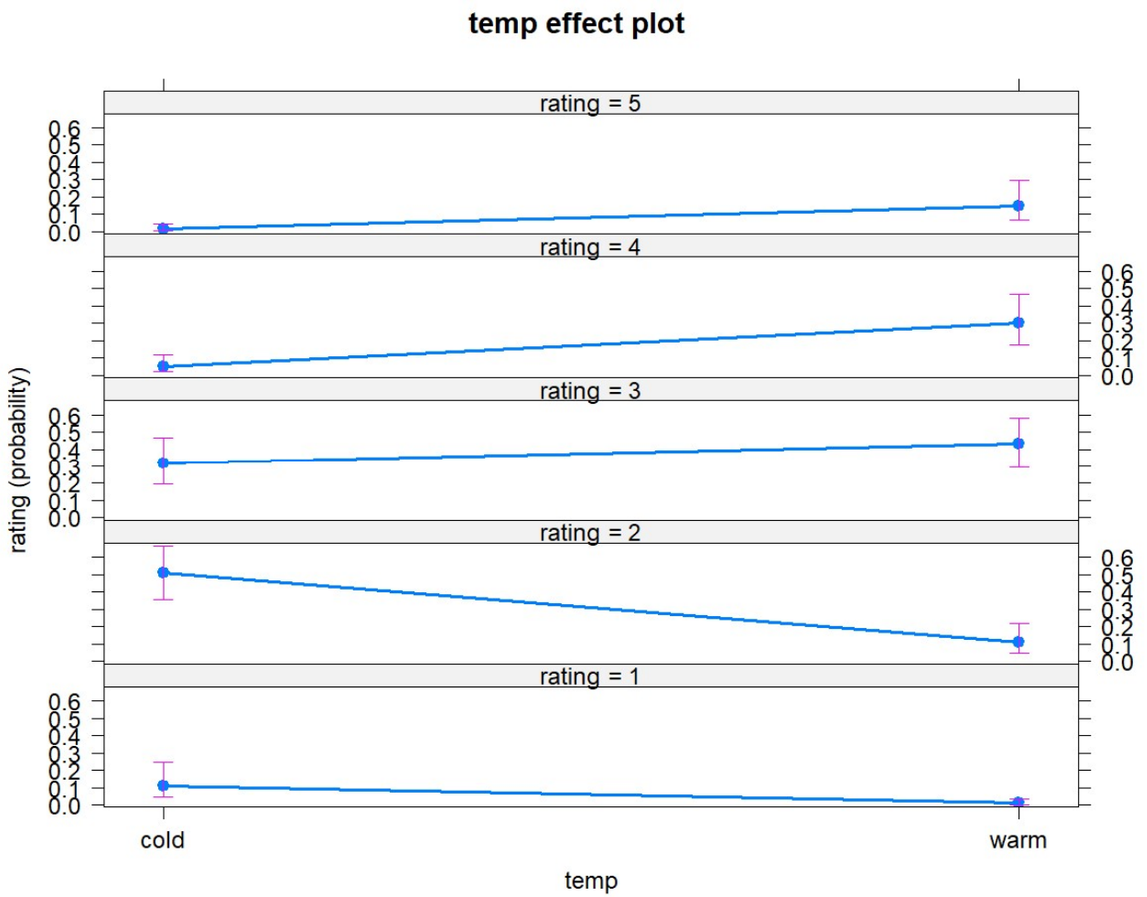
La odds ratio del suceso $Y \geq k$ es $exp(\beta_{\text{tratamiento}})$, por lo que la odds ratio de acidez que clasifica en la categoría k o superior a temperaturas templadas frente a las frías es:

```
exp(coef(modelo3))
  tempwarm contactyes
 12.219985   4.607962
```

Entonces, la posibilidad de pasar de **amargura muy baja** a **amargura baja**, es 12.2 veces mayor para temperaturas **templadas** que para las temperaturas **frías** (también sería 12.2 veces más probable para amargura **baja** a amargura **media**). Algo similar ocurre con el contacto que aumenta el amargo del vino. Las ordenadas en el origen corresponderían a los umbrales de corte de la variable.

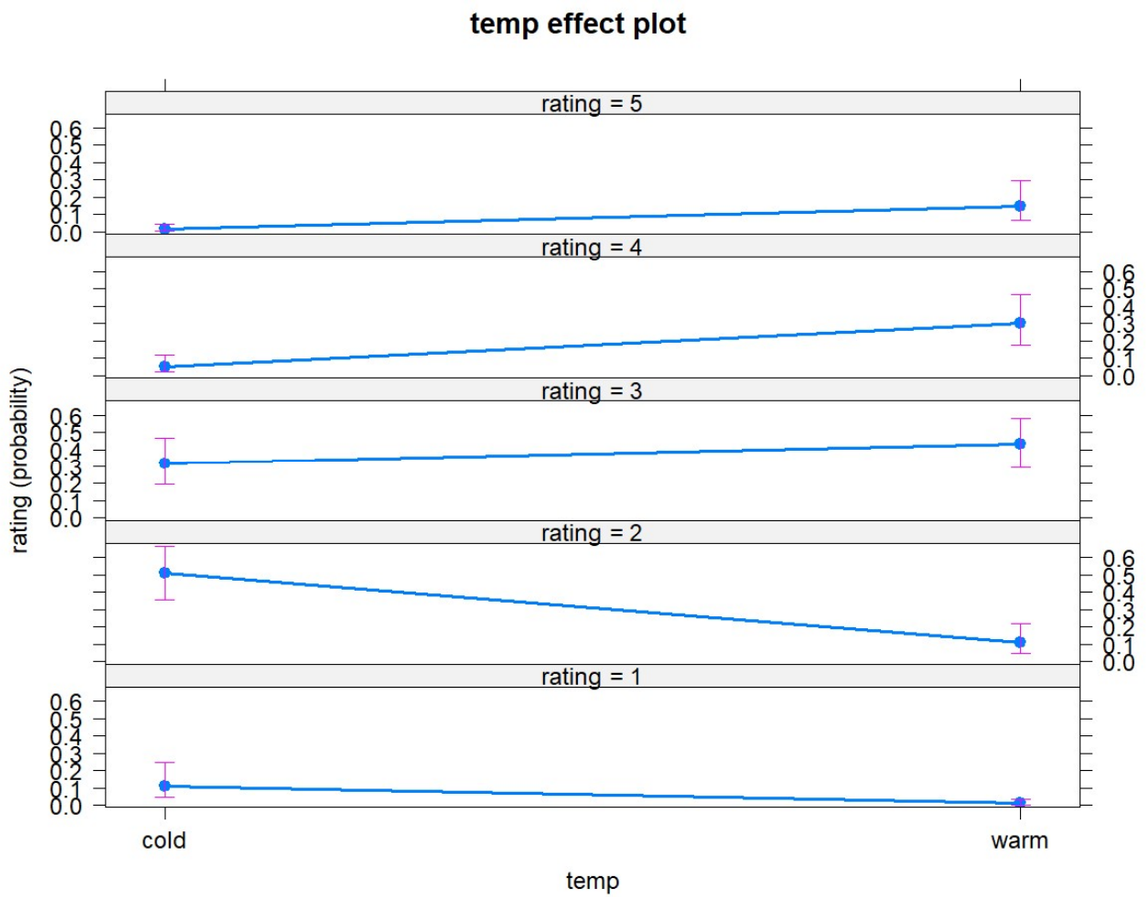
Ejemplo 6

Podemos obtener y dibujar las probabilidades ajustadas:



Ejemplo 6

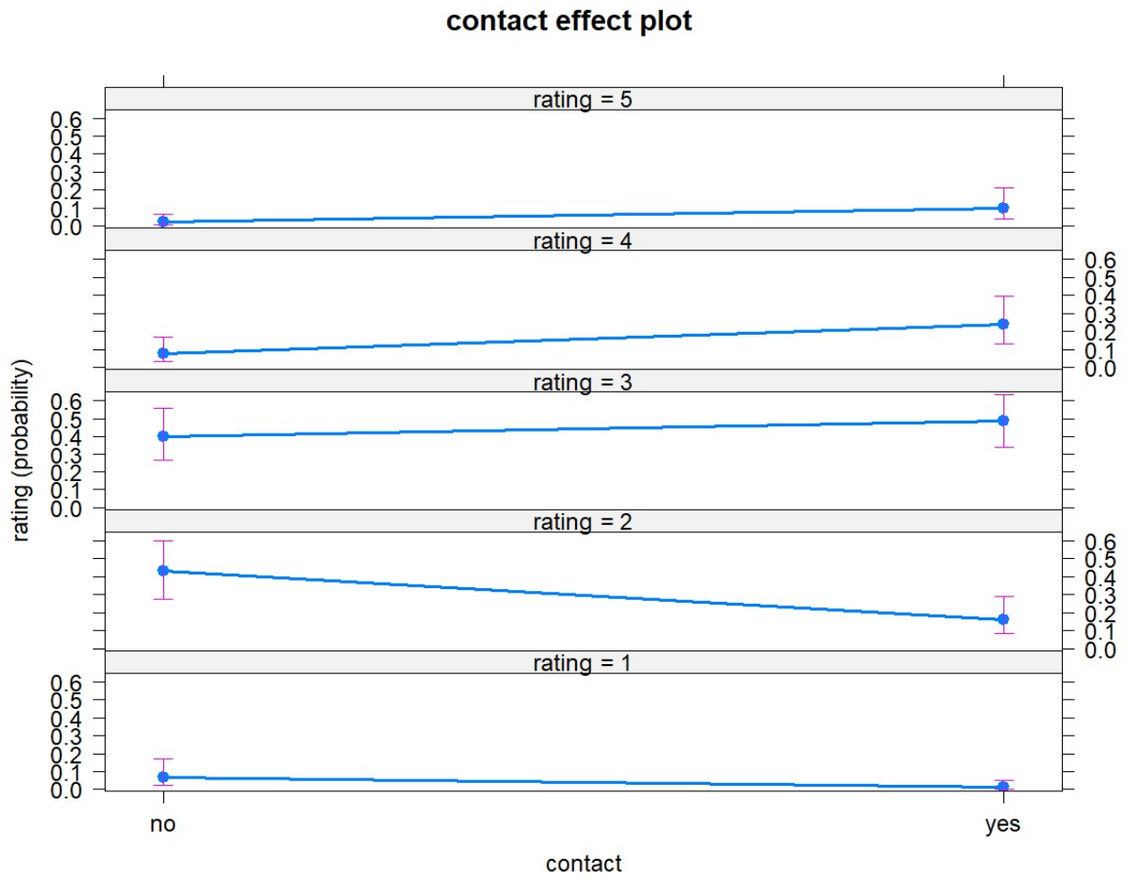
Podemos obtener y dibujar las probabilidades ajustadas:



Visualización de los efectos de rating (amargura de vino) respecto a la **temperatura**.

Ejemplo 6

Podemos obtener y dibujar las probabilidades ajustadas:



Visualización de los efectos de rating (amargura de vino) respecto al **contacto**.

FINESI

Modelos Discretos

IV Semestre

https://aulavirtual2.unap.edu.pe/

GRACIAS