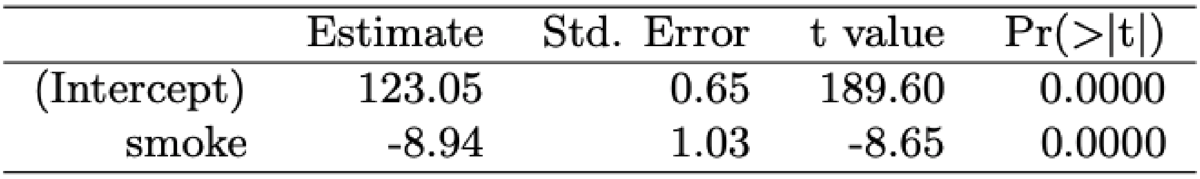
Unidad 5 - Actividad 5B Regresión múltiple y logística
Alumno: **Luis Fernando Izquierdo Berdugo**
Materia: **Estadística**
Fecha de Entrega: **21 de Noviembre de 2024**

Realiza los siguientes ejercicios del capítulo 4 Fundamentos para la inferencia.

Página	Ejercicios
395	8.1
395	8.2
396	8.3
398	8.7
400	8.13
402	8.15
403	8.17

8.1 Baby weights, Part I. The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between smoking and weight of the baby. The variable smoke is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.05	0.65	189.60	0.0000
smoke	-8.94	1.03	-8.65	0.0000

The variability within the smokers and non-smokers are about equal and the distributions are symmetric. With these conditions satisfied, it is reasonable to apply the model. (Note that we don't need to check linearity since the predictor has only two levels.)
(a) Write the equation of the regression line.

La ecuación general de una recta de regresión es:

$$Y = a + bX$$

Donde:

- Y: Variable dependiente (peso al nacer)
- a: Intercepto (peso promedio al nacer de bebés de madres no fumadoras)
- b: Pendiente (efecto del hábito de fumar en el peso al nacer)
- X: Variable independiente (1 si la madre fuma, 0 si no)

Pasándolo a términos del problema:

$Peso\ al\ nacer\ =\ 123.05\ -\ 8.94\ *\ (si\ fuma)$

(b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.

Cada vez que el valor de "si fuma" aumenta en 1 (es decir, la madre fuma), el peso al nacer disminuye en 8.94 onzas en promedio.

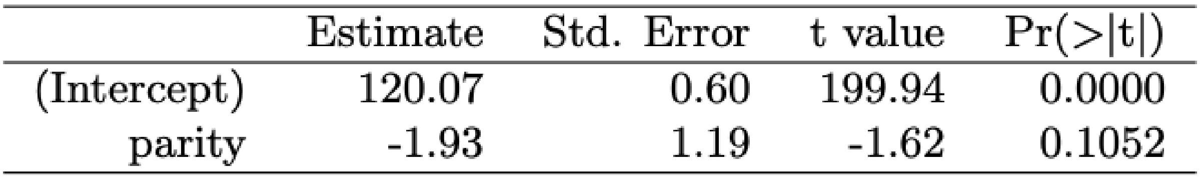
Para madres no fumadoras se estima:
$Peso\ al\ nacer\ =\ 123.05\ -\ 8.94\ *\ 0\ =\ 123.05\ onzas$

Para madres fumadoras se estima:
$Peso\ al\ nacer\ =\ 123.05\ -\ 8.94\ *\ 1\ =\ 114.11\ onzas$

(c) Is there a statistically significant relationship between the average birth weight and smoking?

El valor de p para la variable "smoke" es 0.0000, lo cual es mucho menor que el nivel de significancia típico de 0.05, por lo cual se puede rechazar la hipótesis nula y asumir que si existe una relación estadística entre los hábitos de fumar de la madre y el peso al nacer del bebé.

8.2 Baby weights, Part II. Exercise 8.1 introduces a data set on birth weight of babies. Another variable we consider is parity, which is 0 if the child is the first born, and 1 otherwise. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, from parity.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	120.07	0.60	199.94	0.0000
parity	-1.93	1.19	-1.62	0.1052

(a) Write the equation of the regression line.

La ecuación sería:

$Peso\ al\ nacer\ =\ 120.07\ -\ 1.93\ *\ paridad$

(b) Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.

Cada vez que el valor de "paridad" aumenta en 1 (es decir, el bebé no es el primero), el peso al nacer disminuye en 1.93 onzas en promedio.

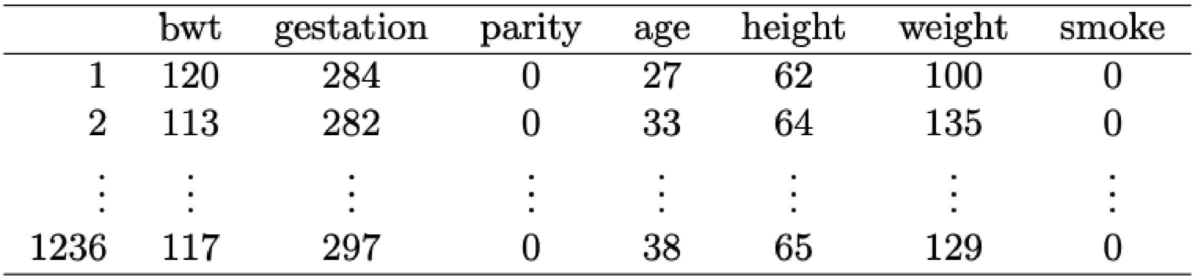
Para primeros hijos, se estima:
$Peso\ al\ nacer\ =\ 120.07\ -\ 1.93\ *\ 0\ =\ 120.07\ onzas$

Para bebés que no son primer hijo, se estima:
$Peso\ al\ nacer\ =\ 120.07\ -\ 1.93\ *\ 1\ =\ 118.14\ onzas$

(c) Is there a statistically significant relationship between the average birth weight and parity?

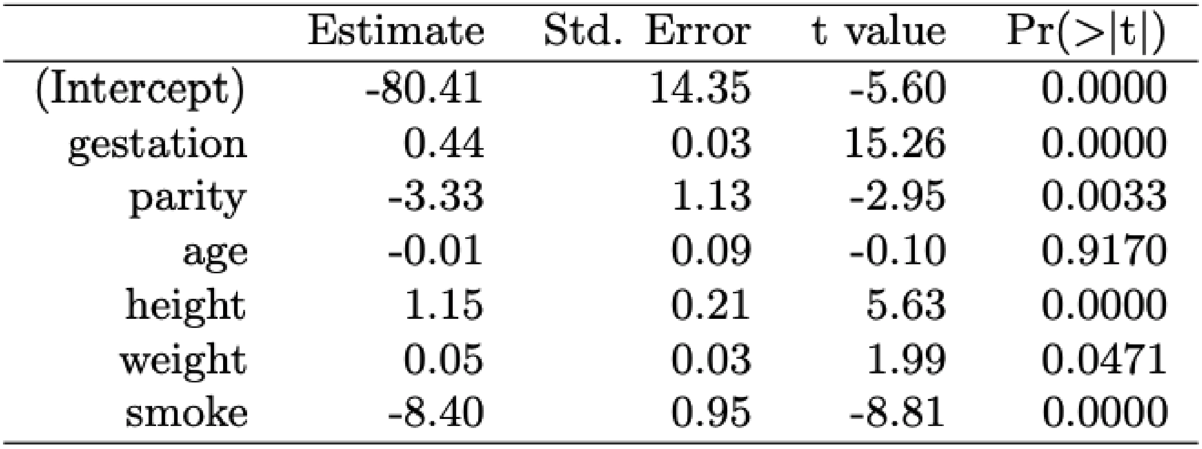
El valor de p para la variable "parity" es 0.1052. Este valor es mayor que el nivel de significancia típico de 0.05. Esto significa que no podemos rechazar la hipótesis nula de que no hay relación entre la paridad y el peso al nacer.

8.3 Baby weights, Part III. We considered the variables smoke and parity, one at a time, in modeling birth weights of babies in Exercises 8.1 and 8.2. A more realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (gestation), mother's age in years (age), mother's height in inches (height), and mother's pregnancy weight in pounds (weight). Below are three observations from this data set.



	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

(a) Write the equation of the regression line that includes all of the variables.

La ecuación de la recta de regresión múltiple es:

$$Peso\ al\ nacer\ =\ -\ 80.41\ +\ (0.44\ *\ gestación)\ -\ (3.33\ *\ paridad)\ -\ (0.01\ *\ edad)\ +\ (1.15\ *\ altura)\ +\ (0.05\ *\ peso)\ -\ (8.40\ *\ si\ fuma)$$

(b) Interpret the slopes of gestation and age in this context.

Por cada día adicional de gestación, el peso al nacer aumenta en promedio 0.44 onzas y por cada año adicional de edad de la madre, el peso al nacer disminuye en promedio 0.01 onzas.

(c) The coefficient for parity is different than in the linear model shown in Exercise 8.2. Why might there be a difference?

En el ejercicio 8.2 únicamente se consideraba la paridad como variable, al añadir más variables en este ejercicio, se puede modificar el efecto de la paridad en el peso al nacer.

(d) Calculate the residual for the first observation in the data set.

$$peso\ predicho\ =\ -\ 80.41\ +\ (0.44\ *\ 284)\ -\ (3.33\ *\ 0)\ -\ (0.01\ *\ 27)\ +\ (1.15\ *\ 62)\ +\ (0.05\ *\ 100)\ -\ (8.40\ *\ 0)\ =\ 120.58$$

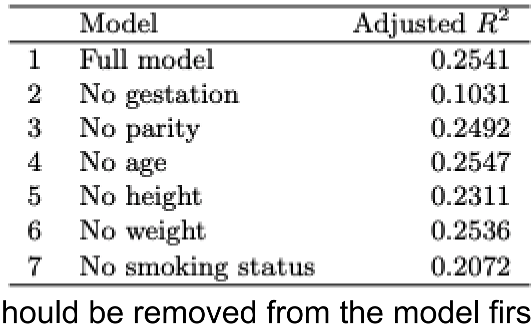
$$residuo\ =\ 120\ -\ peso\ predicho\ =\ 120\ -\ 120.58 =\ -\ 0.58$$

(e) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R2 and the adjusted R2. Note that there are 1,236 observations in the data set.

$$R2\ =\ 1 - \frac{Varianza_{Residuales}}{Varianza_{pesos}} = 1\ -\ \frac{249.28}{332.57} = 0.2504$$

$$R2_{ajustada}\ =\ 1 - \frac{Varianza_{Residuales}}{Varianza_{pesos}} * \frac{n-1}{n-k-1} = 1\ -\ \frac{249.28}{332.57} - \frac{1236-1}{1236-6-1} = 0.2467$$

8.7 Baby weights, Part IV. Exercise 8.3 considers a model that predicts a newborn's weight using several predictors (gestation length, parity, age of mother, height of mother, weight of mother, smoking status of mother). The table below shows the adjusted R-squared for the full model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

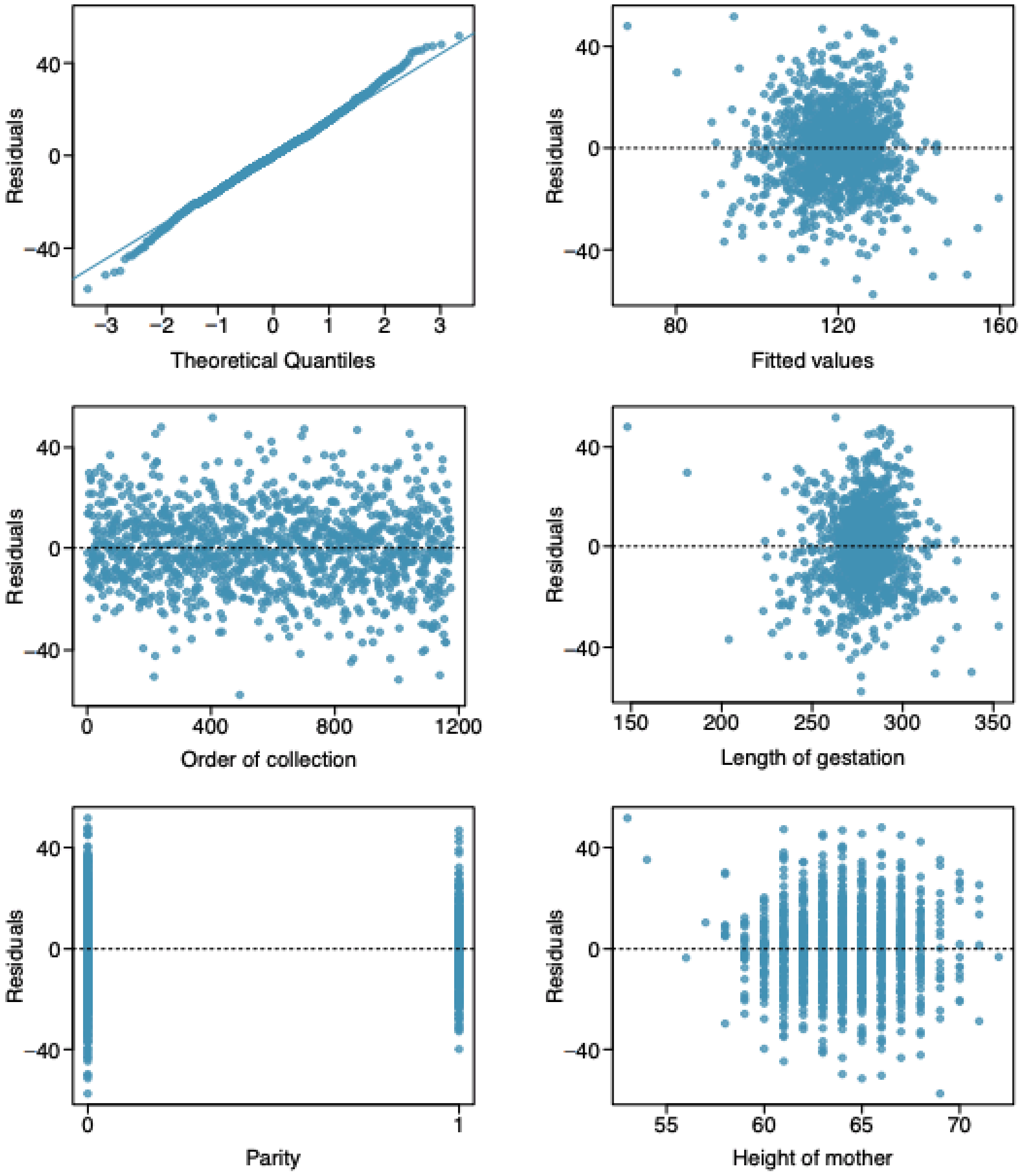


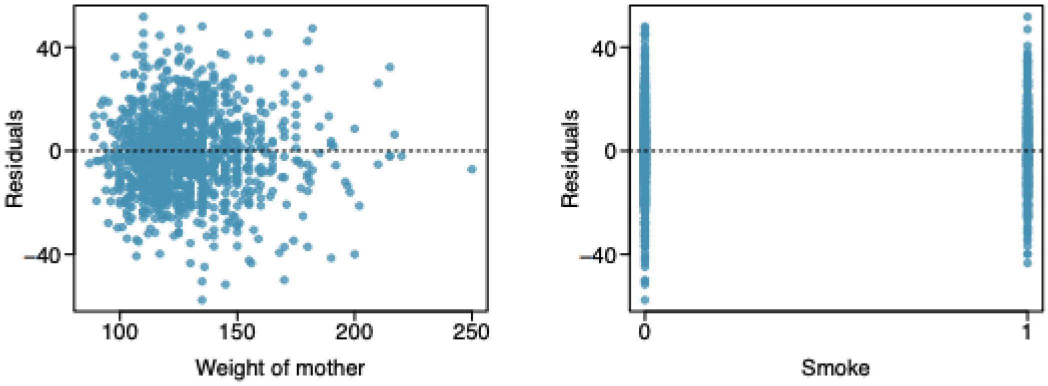
	Model	Adjusted R^2
1	Full model	0.2541
2	No gestation	0.1031
3	No parity	0.2492
4	No age	0.2547
5	No height	0.2311
6	No weight	0.2536
7	No smoking status	0.2072

Which, if any, variable should be removed from the model first?

Al eliminar la variable edad, el R2 ajustado aumenta ligeramente, lo que sugiere que la variable "edad" no está aportando una mejora significativa al modelo, entonces esa variable sería la que se debe eliminar del modelo

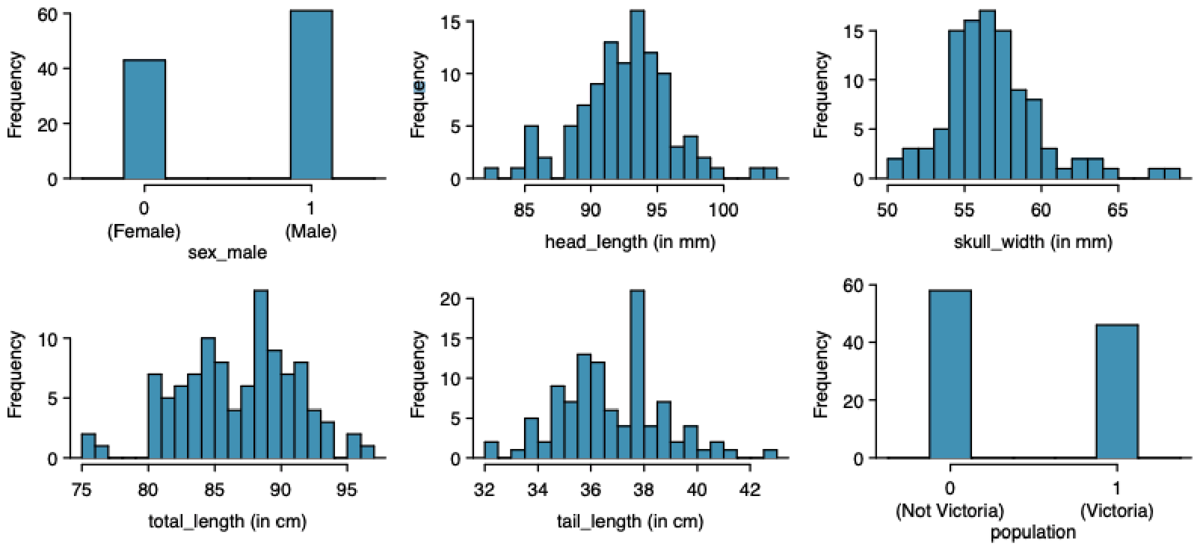
8.13 Baby weights, Part V. Exercise 8.3 presents a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoking status of the mother. Determine if the model assumptions are met using the plots below. If not, describe how to proceed with the analysis.





Con base en los gráficos las suposiciones del modelo de regresión lineal se cumplen. Esto se puede observar, por ejemplo, en la relación entre los residuales y los valores ajustados, ya que estas deben ser aleatorias sin ningún patrón claro, lo cual se puede observar en la segunda gráfica, en esta misma se puede observar que los residuales tienen varianza constante en los valores ajustados.

8.15 Possum classification, Part I. The common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum (see Figure 7.5 on page 334). We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from the population. The first region is Victoria, which is in the eastern half of Australia and traverses the southern coast. The second region consists of New South Wales and Queensland, which make up eastern and northeastern Australia. We use logistic regression to differentiate between possums in these two regions. The outcome variable, called population, takes value 1 when a possum is from Victoria and 0 when it is from New South Wales or Queensland. We consider five predictors: sex male (an indicator for a possum being male), head length, skull width, total length, and tail length. Each variable is summarized in a histogram. The full logistic regression model and a reduced model after variable selection are summarized in the table.



	Full Model				*Reduced Model*			
	Estimate	SE	Z	Pr(>|Z|)	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	39.2349	11.5368	3.40	0.0007	33.5095	9.9053	3.38	0.0007
sex_male	-1.2376	0.6662	-1.86	0.0632	-1.4207	0.6457	-2.20	0.0278
head_length	-0.1601	0.1386	-1.16	0.2480				
skull_width	-0.2012	0.1327	-1.52	0.1294	-0.2787	0.1226	-2.27	0.0231
total_length	0.6488	0.1531	4.24	0.0000	0.5687	0.1322	4.30	0.0000
tail_length	-1.8708	0.3741	-5.00	0.0000	-1.8057	0.3599	-5.02	0.0000

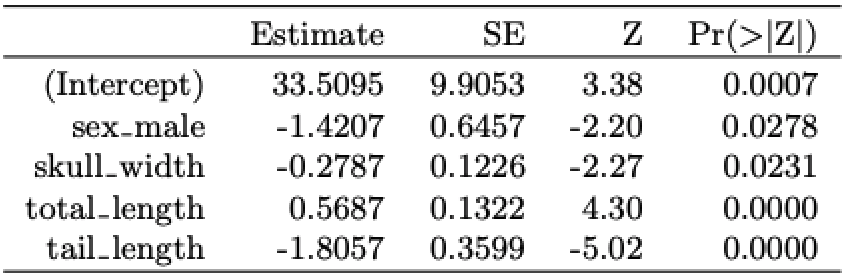
(a) Examine each of the predictors. Are there any outliers that are likely to have a very large influence on the logistic regression model?

Solamente en los gráficos de “sex_male” y “population” no se observan ciertos outliers que puedan afectar el modelo de regresión logística. En “head_length” vemos uno con valor de 15 y otro parece estar en 13, para “skull_width” vemos algunos valores de 16 y 17, en “total_length” se observa un valor de 15 y otro de 10 y para “tail_length” podemos observar de un valor de más de 20.

(b) The summary table for the full model indicates that at least one variable should be eliminated when using the p-value approach for variable selection: head length. The second component of the table summarizes the reduced model following variable selection. Explain why the remaining estimates change between the two models.

Al eliminar "head_length", las otras variables tuvieron que absorber parte de la variabilidad de esta variable, lo cual puede llevar a cambios en los coeficientes de las demás variables. Con este cambio los coeficientes de "skull_width" y "total_length" se volvieron más significativos.

8.17 Possum classification, Part II. A logistic regression model was proposed for classifying common brushtail possums into their two regions in Exercise 8.15. The outcome variable took value 1 if the possum was from Victoria and 0 otherwise.



	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	33.5095	9.9053	3.38	0.0007
sex_male	-1.4207	0.6457	-2.20	0.0278
skull_width	-0.2787	0.1226	-2.27	0.0231
total_length	0.5687	0.1322	4.30	0.0000
tail_length	-1.8057	0.3599	-5.02	0.0000

(a) Write out the form of the model. Also identify which of the variables are positively associated when controlling for other variables.

El modelo de regresión logística tiene la siguiente forma:

$log(\frac{P(Victoria)}{1-P(Victoria)}) = β_0 + β_1 * sexo + β_2 * (ancho\ del\ cráneo) + β_3 * (longitud\ total) + β_4 * (longitud\ cola)$

Donde:

$β_0$, $β_1$, $β_2$, $β_3$ y $β_4$ son los coeficientes que indican la importancia de cada variable.

En este modelo solamente la longitud total tiene una asociación positiva, lo que podría indicar que las zarigüeyas más largas tienen más probabilidades de ser Victoria.

(b) Suppose we see a brushtail possum at a zoo in the US, and a sign says the possum had been captured in the wild in Australia, but it doesn't say which part of Australia. However, the sign does indicate that the possum is male, its skull is about 63 mm wide, its tail is 37 cm long, and its total length is 83 cm. What is the reduced model's computed probability that this possum is from Victoria? How confident are you in the model's accuracy of this probability calculation?

Despejando el modelo propuesto anteriormente y sustituyendo con los datos de la tabla y de la instrucción:

$$\frac{e^{β_0+β_1*sexo + β_2*(ancho\ del\ cráneo) + β_3*(longitud\ total) + β_4*(longitud\ cola)}}{1+e^{β_0+β_1*sexo + β_2*(ancho\ del\ cráneo) + β_3*(longitud\ total)+ β_4*(longitud\ cola)}}$$

Resolviendo el coeficiente

$β_0 + β_1 * sexo + β_2 * (ancho\ del\ cráneo) + β_3 * (longitud\ total) + β_4 * (longitud\ cola)$

$33.5095 - 1.4207 * 1 - 0.2787 * 63 + 0.5687 * 83 - 1.8057 * 37 = - 5.0781$

$$P(Victoria) = \frac{e^{-5.0781}}{1+e^{-5.0781}} = 0.0062$$

Entonces la probabilidad de que la zarigüeya haya sido de Victoria es de 0.0062