Estadística

U5 5B

David Aarón Ramírez Olmeda

Abril 2023

Multiple and logistic regression

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.

y = 123.05 - 8.94*x

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

In this context, the slope of -8.94 means that, on average, babies born to smoking mothers weigh 8.94 ounces less than babies born to non-smoking mothers.

To calculate the predicted birth weight of babies born to smoking and non-smoking, respectively, we can use the equation of the regression line:

```
y = 123.05 - 8.94*1
y
```

```
## [1] 114.11
```

```r
y = 123.05 - 8.94*0
y
```

```
## [1] 123.05
```

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

There is, as indicated by the t-value of -8.65 and the p-value of 0.0000 in the summary table. The p-value is less than 0.05, indicating strong evidence against the null hypothesis of no relationship between birth weight and smoking.

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.

y = 120.07 - 1.93*x

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

The slope of -1.93 means that, on average, babies who are not first borns (parity = 1) weigh 1.93 ounces less than first-born babies (parity = 0)

To calculate the predicted birth weight of first borns and non first borns, respectively, we can substitute the values for parity in the regression equation:

```
y = 120.07 - 1.93*1
y
```

```
## [1] 118.14
```

```
y = 120.07 - 1.93*0
y
```

```
## [1] 120.07
```

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

The p-value for the coefficient of parity is 0.1052. Since this p-value is greater than the conventional significance level of 0.05, we can conclude that there is not strong evidence of a statistically significant relationship between the average birth weight and parity

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**
120	284	0	27	62	100	0
113	282	0	33	64	135	0
...	...	...	...	...	...	...
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(>
(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.

y = -80.41 + 0.44 x **gestation** - 3.33 x **parity** - 0.01 x **age** + 1.15 x **height** + 0.05 x **weight** - 8.40 x **smoke**

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

The slope of gestation means that for each additional day of gestation, we can expect an increase in birth weight by 0.44 ounces, holding all other variables constant.

The slope of age means that for each additional year of mother's age, we can expect a decrease in birth weight by 0.01 ounces, holding all other variables constant.

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

The difference in the coefficient for parity between the two models may be due to the presence of other variables that are also related to birth weight and may partially explain the relationship between parity and birth weight. In the multiple regression model, the coefficient for parity represents the effect of parity on birth weight after controlling for the other variables in the model.

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

```
predicted = -80.41 + (0.44 * 284) + (-3.33 * 0) + (-0.01 * 27) + (1.15 * 62) + (0.05 * 100) + (-8.40 * 0
predicted
```

```
## [1] 120.58
```

```
residual = 120 - y
residual
```

```
## [1] -0.07
```

(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 -(249.28/332.57)
r2
```

```
## [1] 0.2504435
```

```
r2adj <- 1 -(249.28/332.57) * (1235/1229)
r2adj
```

```
## [1] 0.2467842
```

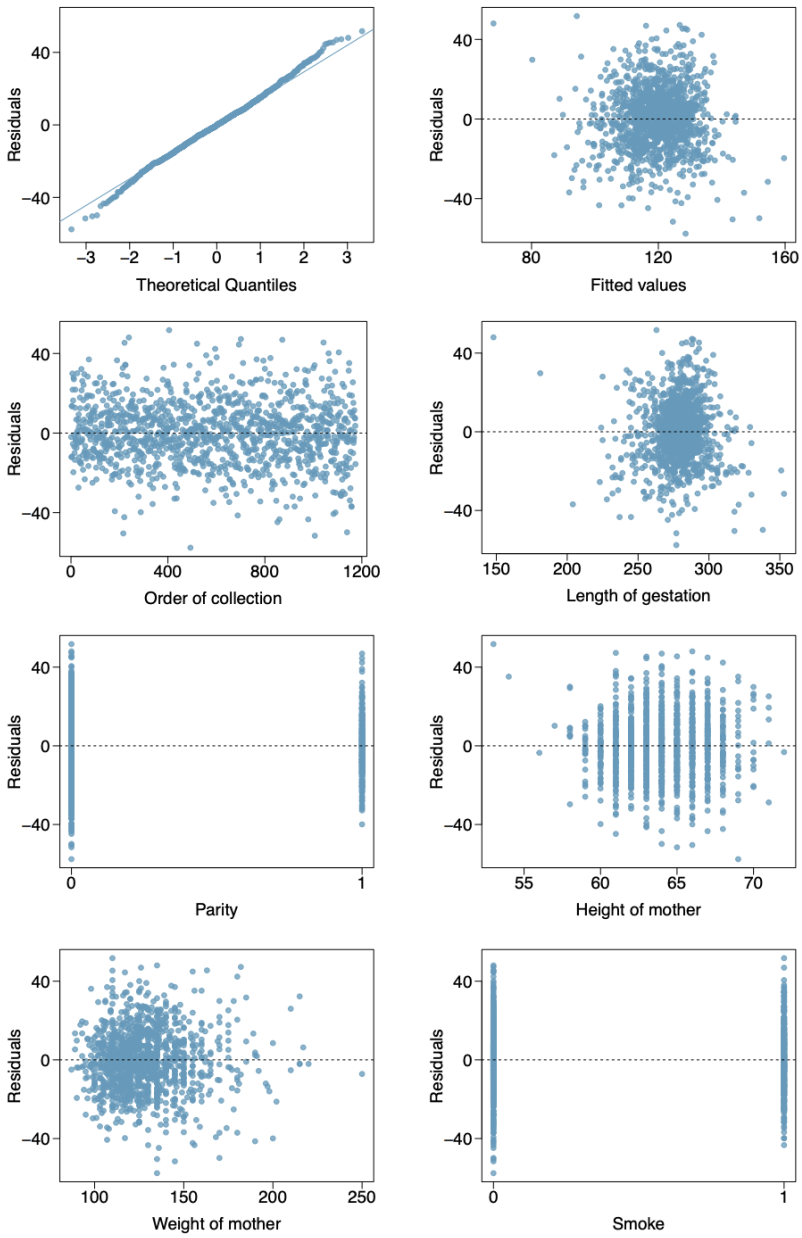
8.7 Baby weights, Part IV. Exercise 8.3 considers a model that predicts a newborn's weight using several predictors (gestation length, parity, age of mother, height of mother, weight of mother, smoking status of mother). The table below shows the adjusted R-squared for the full model as well as adjusted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

Model	**Adjusted R2**
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?

The variable with the lowest adjusted R-squared value should be removed first, as it contributes the least to the overall variability explained by the model. In this case, "No gestation"

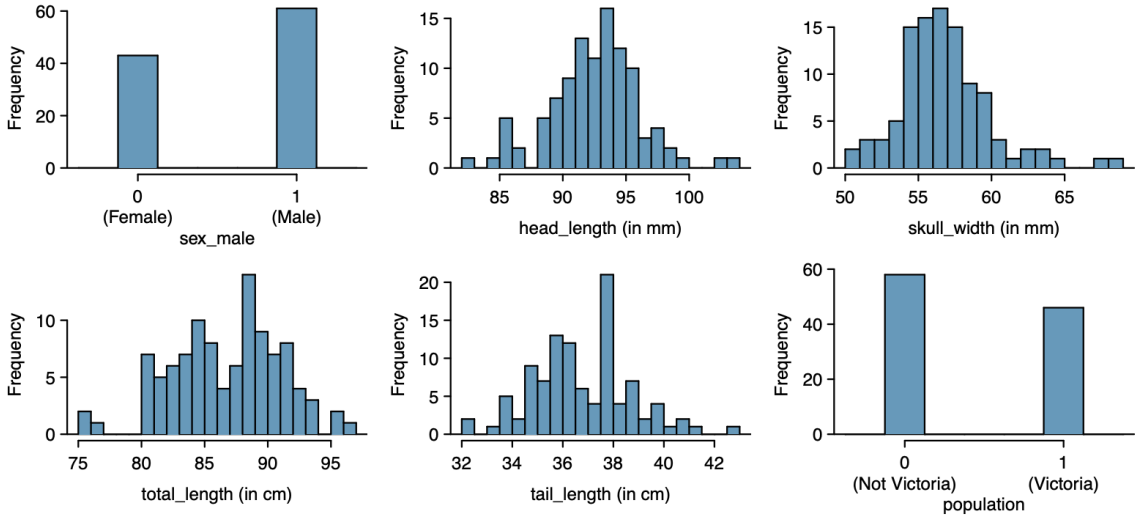
8.13 Baby weights, Part V. Exercise 8.3 presents a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoking status of the mother. Determine if the model assumptions are met using the plots below. If not, describe how to proceed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



The results appear to be in order. The residuals graph shows expected linearity, while the randomly distributed values appear as expected, the variances appear to be normal upon visual inspection. Overall, it seems that our sample data is producing appropriate and non-abnormal graphs.

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?

There are several outliers in head_length, in skull_width, and in total_length. The sample size if big enough so outliers will not have large influence on the model.

(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.

The exclusion of the variable head_length led to modifications in the remaining model due to its correlation with other variables. The P-values for sex_male and skull_width decreased, indicating a potential relationship between a possum's head length, its gender, and skull width.

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.

$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 sexmale + \beta_2 skullwidth + \beta_3 totallength + \beta_4 taillength$

total_length is positively associated with the outcome variable (being from Victoria) when controlling for the other variables

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?

```
sex_male <- 1
skull_width <- 63
tail_length <- 37
total_length <- 83

log = 33.5095 - 1.4207 * sex_male - 0.2787 * skull_width + 0.5687 * total_length - 1.8057 * tail_length
log
```

```
## [1] -5.0781
```

```
p <- exp(log) / (1 + exp(log))
p
```

```
## [1] 0.006193144
```

So the model's computed probability that this possum is from Victoria is approximately 0.6%. Therefore, we can't be that confident about the possum origin.