Multiple and Logistic Regression IV

Dajiang Liu

@PHS 525

Apr-21st-2016

Review of Last Two Classes

- Linear regression model:

- Logistic regression model
  - Deal with binary outcomes
  - Why is multiple (simple) linear regression model inadequate?
  - What is the procedure of logistic regression model
  - What does the parameter estimate mean?

Hands On Exercise

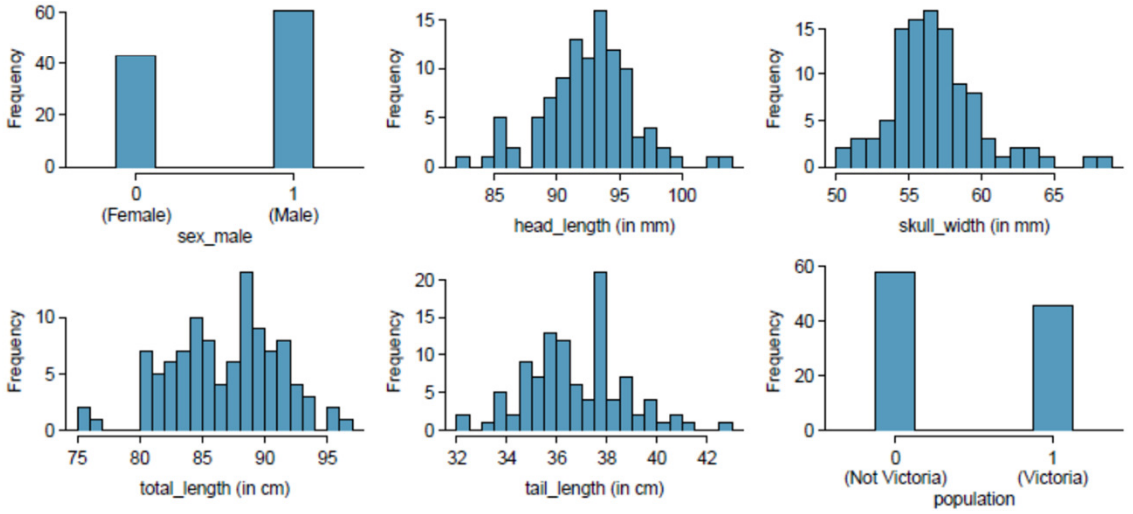
- Perform logistic regression

analyses for population

- Which predictors should be dropped
- What happens if the predictors are dropped?

8.13 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 318). 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.



population

	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?

(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.

8.15 Possum classification, Part II. A logistic regression model was proposed for classifying common brushtail possums into their two regions in Exercise 8.13. Use the results of the summary table for the reduced model presented in Exercise 8.13 for the questions below. The outcome variable took value 1 if the possum was from Victoria and 0 otherwise.

(a) Write out the form of the model. Also identify which of the following variables are positively associated (when controlling for other variables) with a possum being from Victoria: `skull_width`, `total_length`, and `tail_length`.

(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?

Exemplar Code

- Load data:
  - dat=read.table('Ch 8 Exercise Data/possum.txt',header=T,sep='\t');
- Analyze relations between response and predictors
  - res=glm(as.factor(dat$pop) ~ dat$sex + dat$headL + dat$skullW + dat$tailL + dat$totalL,family='binomial')
  - summary(res)
- Examine if the variables have outliers:
  - boxplot(dat$totalL)

Checking Model Assumptions

- Valid linear (logistic) regression analyses require valid model assumptions
  - If assumptions violated, the results can be invalid

- Model assumptions:
  - The residuals for models are nearly normal
  - Variability of the residuals are nearly constant;
  - The residuals are independent
  - Each predictor variables are linearly related to the responses

Checking Model Assumptions

- Important to validate model assumptions

> **Caution: Don't report results when assumptions are grossly violated**
> While there is a little leeway in model assumptions, don't go too far. If model assumptions are very clearly violated, consider a new model, even if it means learning more statistical methods or hiring someone who can help.

- But not too much because

> **"All models are wrong, but some are useful" -George E.P. Box**
> The truth is that no model is perfect. However, even imperfect models can be useful. Reporting a flawed model can be reasonable so long as we are clear and report the model's shortcomings.

Diagnostic Plots

- Normal probability plot
- Absolute values of residuals against fitted value
- Residuals in the order of data collection
- Residuals against each predictor variable

Normal Probability Plot

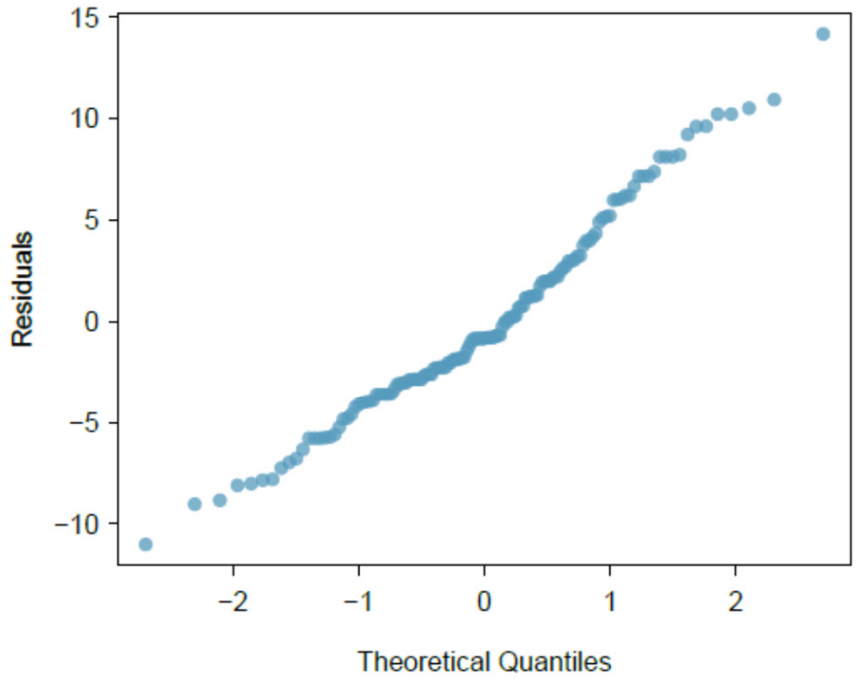


Figure 8.9: A normal probability plot of the residuals is helpful in identifying observations that might be outliers.

Fitted Values v.s. Absolute Values of Residuals

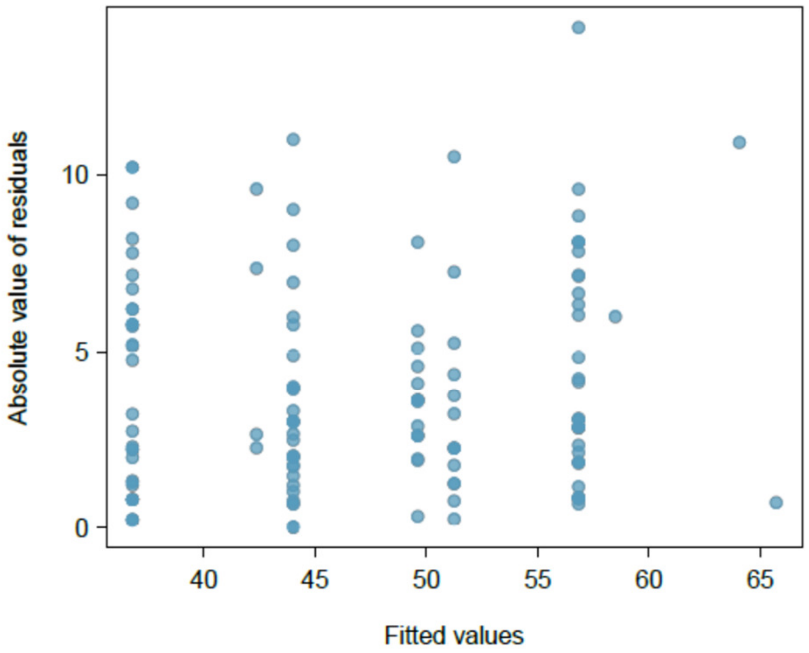
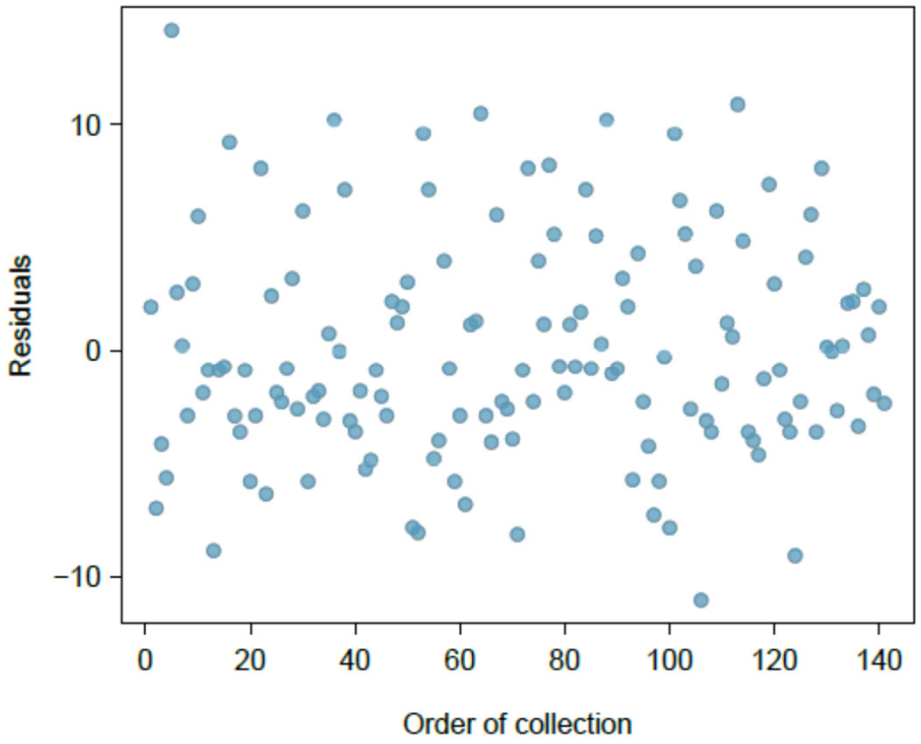


Figure 8.10: Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.

Plot Residuals in the Order of Data Collection



Diagnostics for Logistic Regression

Logistic regression conditions

There are two key conditions for fitting a logistic regression model:

1. Each predictor x_i is linearly related to $\text{logit}(p_i)$ if all other predictors are held constant.
2. Each outcome Y_i is independent of the other outcomes.

Diagnostics for Logistic Regression

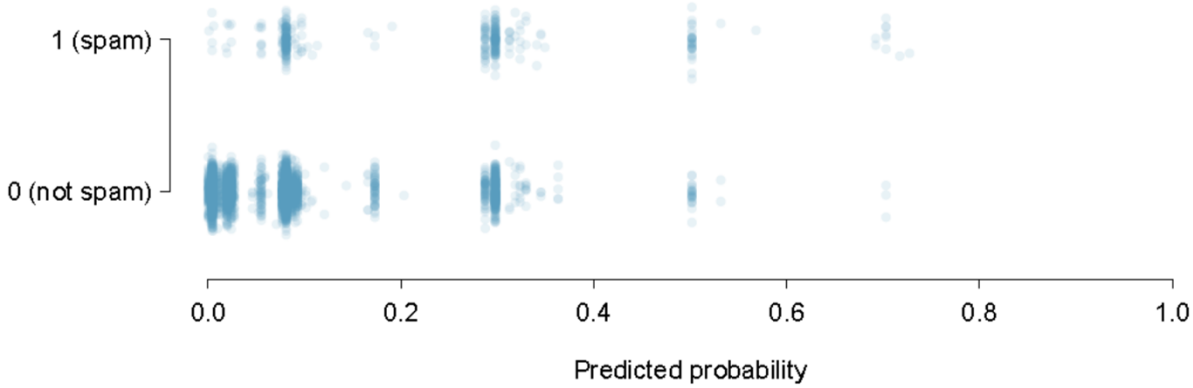


Figure 8.17: The predicted probability that each of the 3,912 emails is spam is classified by their grouping, spam or not. Noise (small, random vertical shifts) have been added to each point so that points with nearly identical values aren't plotted exactly on top of one another. This makes it possible to see more observations.

Practical Exercise

- Examine the four plots using the birth weight datasets