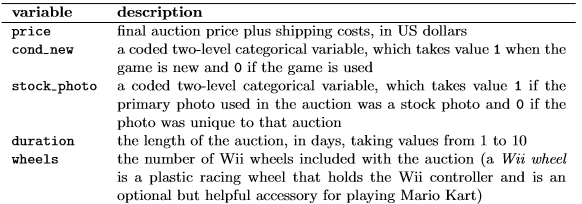
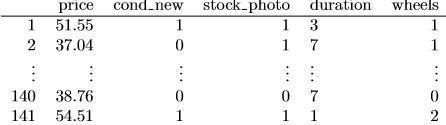
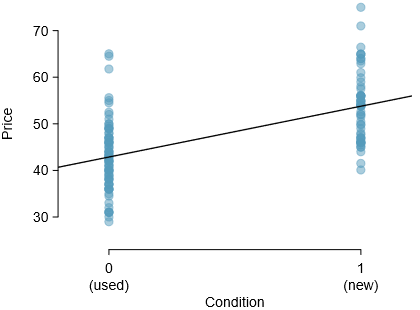
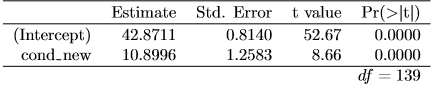
# Chapter 6: Multiple and logistic regression

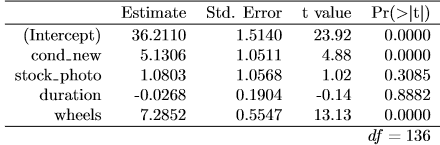
* Principles of simple linear regression lay foundation for more sophisticted regression methods used in a wide range of challenging settings
* **multiple regression** = more than 1 predictor
* **logistic regression** = predicting categorical outcomes w/ 2 possible categories.
* Multiple regression extends simple 2-variable regression w/ 1 response but many predictors (X = {x1, x2, x3, ... x\_n})
* Method motivated by scenarios where many variables may be simultaneously connected to output
* Consider Ebay auctions of Mario Kart for the Wii w/ outcome variable of interest = total price of an auction (highest bid + shipping cost) + try to determine how total price is related to each characteristic in an auction, *while simultaneously controlling for other variables.*
* Ex: “All other characteristics held constant, are longer auctions associated w/ higher or lower prices?”
* Ex: “On average, how much more do buyers tend to pay for additional Wii wheels in auctions?”
* Data set includes results from 141 auctions.



* condition + stock photo = **indicator variables** 🡺 cond = 1 if game new, 0 if used.
* Using indicator variables allows for these variables to be directly used in regression.
* Multiple regression also allows for categorical variables w/ many levels (do not have any such variables in this analysis)
* Fit a linear regression w/ condition as a predictor of auction price. 



* Guided Practice 6.1 Examine Figure 6.4. Does the linear model seem reasonable?
* Yes 🡺 Constant variability, nearly normal residuals, + linearity all appear reasonable.
* Guided Practice 6.2 Interpret the coeﬃcient for the game’s condition in the model. Is this coeﬃcient signiﬁcantly diﬀerent from 0?
* Yes, its significantly different from 0 (see p-value), so this model predicts a ~$10.90 increase in auction price for a new game compared to a used game
* Low p-value gives strong evidence the coeﬃcient != 0 when using this simple 1-variable model
* \*\*\*Sometimes there are **underlying structures or relationships BETWEEN predictor variables**\*\*\*
* For instance, new games on Ebay tend to come w/ more Wii wheels, which may have led to higher prices for those auctions.
* Fit a model that includes all potentially important variables simultaneously to help evaluate the relationship between a predictor + the outcome while controlling for potential inﬂuence of other variables = **multiple regression.**
* \*\*\*Remain cautious about making any causal interpretations using multiple regression, though such models = a common 1st step in providing evidence of a causal connection\*\*\*
* Want to construct a model that accounts for not only game condition but simultaneously accounts stock photo, duration, + wheels. 
* Just as w/ single predictor, multiple regression may be missing important components or might not precisely represent the relationship between outcome + available predictors.
* \*\*\*While **no model is perfect**, we wish to explore the possibility this one may ﬁt the data reasonably well\*\*\*
* Estimate parameters β0, β1, ..., β4 in the same way as w/ a single predictor 🡺 select b0, b1, ..., b4 to minimize the sum of the squared residuals (**OLS**): 
* Here = 141 residuals, 1 for each observation 🡺 ID the **point estimates** b\_i of each βi, just as w/ 1-predictor case:



* Guided Practice 6.5 Write out the model in Equation (6.3) using the point estimates from Table 6.5. How many predictors are there in this model?
* **y = 36.2110 + 5.1306\*cond\_new + 1.0803\*stock\_photo - .0268\*duration + 7.2852\*wheels**
* Guided Practice 6.6 What does β4, the coeﬃcient of variable x4 (Wii wheels), represent? What is the point estimate of β4?
* **It represents an \*\*\**average difference*\*\*\* of ~$7.29 in price for each Wii wheel included w/ the game, \*\*\*holding other variables constant\*\*\***
* 6.7 Compute the residual of the 1st observation in the dataset using the equation in Practice 6.5.
* **y = 36.2110 + 5.1306\*1 + 1.0803\*1 - .0268\*3 + 7.2852\*1**
* = 36.2110 + 5.1306 + 1.0803 - .0268\*3 + 7.2852 = **49.6267**
* e = 51.55 – 49.6267 = **1.9233**
* 6.8 We estimated a coeﬃcient for cond new in Section 6.1.1 of b1 = 10.90 w/ SE\_b1 = 1.26 when using simple linear regression. Why might there be a diﬀerence between that estimate + the one in the multiple regression setting?
* **Interactions** between different predictors b/c some may be correlated
* Ex: When estimating the connection of price + cond\_new using simple linear regression = unable to control for other variables like # of Wii wheels included in the auction.
* *\*\*\*That* model was **biased** by the **confounding** **variable** *wheels\*\*\**
* W/ **both** variables, this particular underlying + unintentional bias reduced/eliminated (\*\*\*bias from other confounding variables may still remain\*\*\*)
* **Correlation among predictor variables = common issue in multiple regression**
* 2 predictors = **collinear** if correlated = complicates model estimation.
* Impossible to prevent collinearity from arising in *observational* data, experiments = usually designed to prevent predictors from being collinear.
* 6.9 Estimated value of intercept = 36.21, one might be tempted to make some interpretation of this coeﬃcient, such as, “it is the model’s predicted price when each variable takes value = 0: used game, primary image = not a stock photo, auction duration = 0 days, no wheels included. Is there any value gained by making this interpretation?
* NO 🡺 intercept just serves to set the y-axis for the model
* While 3 variables (cond new, stock photo, wheels) do take value 0, auction duration = *always* 1+ days.
* If auction is not up for any days, no one can bid = total auction price would always be 0 for such an auction 🡺 **interpretation of the intercept in this setting is not insightful.**
* 1st used R2 in to determine amount of variability in the response explained by the model:
*  = 1 – (RSS/TSS)
* e\_i = residuals of the model + y\_i = outcomes.
* This equation remains valid in the multiple regression framework, but a small enhancement can often be even more informative. J Guided Practice 6.10 The variance of the residuals for the model given in Guided Practice 6.7 is 23.34, and the variance of the total price in all the auctions is 83.06. Calculate R2 for this model.7 5ei = yi − ˆ yi = 51.55−49.62 = 1.93, where 49.62 was computed using the variables values from the observation and the equation identiﬁed in Guided Practice 6.5. 6Three of the variables (cond new, stock photo, and wheels) do take value 0, but the auction duration is always one or more days. If the auction is not up for any days, then no one can bid on it! That means the total auction price would always be zero for such an auction; the interpretation of the intercept in this setting is not insightful. 7R2 = 1− 23.34 83.06 = 0.719.

266 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

This strategy for estimating R2 is acceptable when there is just a single variable. However, it becomes less helpful when there are many variables. The regular R2 is actually a biased estimate of the amount of variability explained by the model. To get a better estimate, we use the adjusted R2.

Adjusted R2 as a tool for model assessment The adjusted R2 is computed as R2adj = 1− V ar(ei)/(n−k−1) V ar(yi)/(n−1) = 1− V ar(ei) V ar(yi) ×

n−1 n−k−1 where n is the number of cases used to ﬁt the model and k is the number of predictor variables in the model.

Because k is never negative, the adjusted R2 will be smaller – often times just a little smaller – than the unadjusted R2. The reasoning behind the adjusted R2 lies in the degrees of freedom associated with each variance.8 J Guided Practice 6.11 There were n = 141 auctions in the mario kart data set and k = 4 predictor variables in the model. Use n, k, and the variances from Guided Practice 6.10 to calculate R2adj for the Mario Kart model.9 J Guided Practice 6.12 Suppose you added another predictor to the model, but the variance of the errors V ar(ei) didn’t go down. What would happen to the R2? What would happen to the adjusted R2? 10

6.2 Model selection

The best model is not always the most complicated. Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions. In this section we discuss model selection strategies, which will help us eliminate from the model variables that are less important. In this section, and in practice, the model that includes all available explanatory variables is often referred to as the full model. Our goal is to assess whether the full model is the best model. If it isn’t, we want to identify a smaller model that is preferable.

8In multiple regression, the degrees of freedom associated with the variance of the estimate of the residuals is n−k−1, not n−1. For instance, if we were to make predictions for new data using our current model, we would ﬁnd that the unadjusted R2 is an overly optimistic estimate of the reduction in variance in the response, and using the degrees of freedom in the adjusted R2 formula helps correct this bias. 9R2adj = 1− 23.34 83.06 × 141−1 141−4−1 = 0.711. 10The unadjusted R2 would stay the same and the adjusted R2 would go down.

6.2. MODEL SELECTION 267

6.2.1 Identifying variables in the model that may not be helpful

Table 6.6 provides a summary of the regression output for the full model for the auction data. The last column of the table lists p-values that can be used to assess hypotheses of the following form:

H0: βi = 0 when the other explanatory variables are included in the model. HA: βi 6= 0 when the other explanatory variables are included in the model.

Estimate Std. Error t value Pr(>|t|) (Intercept) 36.2110 1.5140 23.92 0.0000 cond new 5.1306 1.0511 4.88 0.0000 stock photo 1.0803 1.0568 1.02 0.3085 duration -0.0268 0.1904 -0.14 0.8882 wheels 7.2852 0.5547 13.13 0.0000 R2adj = 0.7108 df = 136

Table 6.6: The ﬁt for the full regression model, including the adjusted R2.

Example 6.13 The coeﬃcient of cond new has a t test statistic of T = 4.88 and a p-value for its corresponding hypotheses (H0 : β1 = 0, HA : β1 6= 0) of about zero. How can this be interpreted?

If we keep all the other variables in the model and add no others, then there is strong evidence that a game’s condition (new or used) has a real relationship with the total auction price. Example 6.14 Is there strong evidence that using a stock photo is related to the total auction price? The t test statistic for stock photo is T = 1.02 and the p-value is about 0.31. After accounting for the other predictors, there is not strong evidence that using a stock photo in an auction is related to the total price of the auction. We might consider removing the stock photo variable from the model. J Guided Practice 6.15 Identify the p-values for both the duration and wheels variables in the model. Is there strong evidence supporting the connection of these variables with the total price in the model?11

There is not statistically signiﬁcant evidence that either the stock photo or duration variables contribute meaningfully to the model. Next we consider common strategies for pruning such variables from a model.

11The p-value for the auction duration is 0.8882, which indicates that there is not statistically signiﬁcant evidence that the duration is related to the total auction price when accounting for the other variables. The p-value for the Wii wheels variable is about zero, indicating that this variable is associated with the total auction price.

268 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

TIP: Using adjusted R2 instead of p-values for model selection The adjusted R2 may be used as an alternative to p-values for model selection, where a higher adjusted R2 represents a better model ﬁt. For instance, we could compare two models using their adjusted R2, and the model with the higher adjusted R2 would be preferred. This approach tends to include more variables in the ﬁnal model when compared to the p-value approach.

6.2.2 Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called backward-selection and forward-selection. These techniques are often referred to as stepwise model selection strategies, because they add or delete one variable at a time as they “step” through the candidate predictors. We will discuss these strategies in the context of the p-value approach. Alternatively, we could have employed an R2adj approach. The backward-elimination strategy starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time from the model until only variables with statistically signiﬁcant p-values remain. The strategy within each elimination step is to drop the variable with the largest p-value, reﬁt the model, and reassess the inclusion of all variables. Example 6.16 Results corresponding to the full model for the mario kart data are shown in Table 6.6. How should we proceed under the backward-elimination strategy?

There are two variables with coeﬃcients that are not statistically diﬀerent from zero: stock photo and duration. We ﬁrst drop the duration variable since it has a larger corresponding p-value, then we reﬁt the model. A regression summary for the new model is shown in Table 6.7. In the new model, there is not strong evidence that the coeﬃcient for stock photo is diﬀerent from zero, even though the p-value decreased slightly, and the other pvalues remain very small. Next, we again eliminate the variable with the largest non-signiﬁcant p-value, stock photo, and reﬁt the model. The updated regression summary is shown in Table 6.8. In the latest model, we see that the two remaining predictors have statistically significant coeﬃcients with p-values of about zero. Since there are no variables remaining that could be eliminated from the model, we stop. The ﬁnal model includes only the cond new and wheels variables in predicting the total auction price:

ˆ y = b0 + b1x1 + b4x4 = 36.78 + 5.58x1 + 7.23x4

where x1 represents cond new and x4 represents wheels. An alternative to using p-values in model selection is to use the adjusted R2. At each elimination step, we reﬁt the model without each of the variables up for potential elimination. For example, in the ﬁrst step, we would ﬁt four models, where each would be missing a diﬀerent predictor. If one of these smaller models has a higher adjusted R2 than our current model, we pick the smaller model with the largest adjusted R2. We continue in this way until removing variables does not increase R2adj. Had we

6.2. MODEL SELECTION 269

used the adjusted R2 criteria, we would have kept the stock photo variable along with the cond new and wheels variables.

Estimate Std. Error t value Pr(>|t|) (Intercept) 36.0483 0.9745 36.99 0.0000 cond new 5.1763 0.9961 5.20 0.0000 stock photo 1.1177 1.0192 1.10 0.2747 wheels 7.2984 0.5448 13.40 0.0000 R2adj = 0.7128 df = 137

Table 6.7: The output for the regression model where price is the outcome and the duration variable has been eliminated from the model.

Estimate Std. Error t value Pr(>|t|) (Intercept) 36.7849 0.7066 52.06 0.0000 cond new 5.5848 0.9245 6.04 0.0000 wheels 7.2328 0.5419 13.35 0.0000 R2adj = 0.7124 df = 138

Table 6.8: The output for the regression model where price is the outcome and the duration and stock photo variables have been eliminated from the model.

Notice that the p-value for stock photo changed a little from the full model (0.309) to the model that did not include the duration variable (0.275). It is common for p-values of one variable to change, due to collinearity, after eliminating a diﬀerent variable. This ﬂuctuation emphasizes the importance of reﬁtting a model after each variable elimination step. The p-values tend to change dramatically when the eliminated variable is highly correlated with another variable in the model. The forward-selection strategy is the reverse of the backward-elimination technique. Instead of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot ﬁnd any variables that present strong evidence of their importance in the model. Example 6.17 Construct a model for the mario kart data set using the forwardselection strategy. We start with the model that includes no variables. Then we ﬁt each of the possible models with just one variable. That is, we ﬁt the model including just the cond new predictor, then the model including just the stock photo variable, then a model with just duration, and a model with just wheels. Each of the four models (yes, we ﬁt four models!) provides a p-value for the coeﬃcient of the predictor variable. Out of these four variables, the wheels variable had the smallest p-value. Since its p-value is less than 0.05 (the p-value was smaller than 2e-16), we add the Wii wheels variable to the model. Once a variable is added in forward-selection, it will be included in all models considered as well as the ﬁnal model. Since we successfully found a ﬁrst variable to add, we consider adding another. We ﬁt three new models: (1) the model including just the cond new and wheels variables (output in Table 6.8), (2) the model including just the stock photo and wheels

270 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

variables, and (3) the model including only the duration and wheels variables. Of these models, the ﬁrst had the lowest p-value for its new variable (the p-value corresponding to cond new was 1.4e-08). Because this p-value is below 0.05, we add the cond new variable to the model. Now the ﬁnal model is guaranteed to include both the condition and wheels variables. We must then repeat the process a third time, ﬁtting two new models: (1) the model including the stock photo, cond new, and wheels variables (output in Table 6.7) and (2) the model including the duration, cond new, and wheels variables. The p-value corresponding to stock photo in the ﬁrst model (0.275) was smaller than the p-value corresponding to duration in the second model (0.682). However, since this smaller p-value was not below 0.05, there was not strong evidence that it should be included in the model. Therefore, neither variable is added and we are ﬁnished. The ﬁnal model is the same as that arrived at using the backward-selection strategy. Example 6.18 As before, we could have used the R2adj criteria instead of examining p-values in selecting variables for the model. Rather than look for variables with the smallest p-value, we look for the model with the largest R2adj. What would the result of forward-selection be using the adjusted R2 approach?

Using the forward-selection strategy, we start with the model with no predictors. Next we look at each model with a single predictor. If one of these models has a larger R2adj than the model with no variables, we use this new model. We repeat this procedure, adding one variable at a time, until we cannot ﬁnd a model with a larger R2adj. If we had done the forward-selection strategy using R2adj, we would have arrived at the model including cond new, stock photo, and wheels, which is a slightly larger model than we arrived at using the p-value approach and the same model we arrived at using the adjusted R2 and backwards-elimination.

Model selection strategies The backward-elimination strategy begins with the largest model and eliminates variables one-by-one until we are satisﬁed that all remaining variables are important to the model. The forward-selection strategy starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found.

There is no guarantee that the backward-elimination and forward-selection strategies will arrive at the same ﬁnal model using the p-value or adjusted R2 methods. If the backwards-elimination and forward-selection strategies are both tried and they arrive at diﬀerent models, choose the model with the larger R2adj as a tie-breaker; other tie-break options exist but are beyond the scope of this book. It is generally acceptable to use just one strategy, usually backward-elimination with either the p-value or adjusted R2 criteria. However, before reporting the model results, we must verify the model conditions are reasonable.

6.3. CHECKING MODEL ASSUMPTIONS USING GRAPHS 271

−2 −1 0 1 2

−10

−5

0

5

10

15

Theoretical Quantiles

Residuals

Figure 6.9: A normal probability plot of the residuals is helpful in identifying observations that might be outliers.

6.3 Checking model assumptions using graphs

Multiple regression methods using the model

ˆ y = β0 + β1x1 + β2x2 +···+ βkxk generally depend on the following four assumptions:

1. the residuals of the model are nearly normal, 2. the variability of the residuals is nearly constant, 3. the residuals are independent, and 4. each variable is linearly related to the outcome.

Simple and eﬀective plots can be used to check each of these assumptions. We will consider the model for the auction data that uses the game condition and number of wheels as predictors. The plotting methods presented here may also be used to check the conditions for the models introduced in Chapter 5.

Normal probability plot. A normal probability plot of the residuals is shown in Figure 6.9. While the plot exhibits some minor irregularities, there are no outliers that might be cause for concern. In a normal probability plot for residuals, we tend to be most worried about residuals that appear to be outliers, since these indicate long tails in the distribution of residuals.

Absolute values of residuals against ﬁtted values. A plot of the absolute value of the residuals against their corresponding ﬁtted values (ˆ yi) is shown in Figure 6.10. This plot is helpful to check the condition that the variance of the residuals is approximately constant. We don’t see any obvious deviations from constant variance in this example.

272 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

●

●

●

●

●

●

●

●●

●

●

●

●

●●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

● ●

●

●

●

●

●●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●●

●

●

●

●●

●

Fitted values

Absolute value of residuals

40 45 50 55 60 65

0

5

10

Figure 6.10: Comparing the absolute value of the residuals against the ﬁtted values (ˆ yi) is helpful in identifying deviations from the constant variance assumption.

Residuals in order of their data collection. A plot of the residuals in the order their corresponding auctions were observed is shown in Figure 6.11. Such a plot is helpful in identifying any connection between cases that are close to one another, e.g. we could look for declining prices over time or if there was a time of the day when auctions tended to fetch a higher price. Here we see no structure that indicates a problem.12

Residuals against each predictor variable. We consider a plot of the residuals against the cond new variable and the residuals against the wheels variable. These plots are shown in Figure 6.12. For the two-level condition variable, we are guaranteed not to see any remaining trend, and instead we are checking that the variability doesn’t ﬂuctuate across groups. In this example, when we consider the residuals against the wheels variable, we see some possible structure. There appears to be curvature in the residuals, indicating the relationship is probably not linear.

It is necessary to summarize diagnostics for any model ﬁt. If the diagnostics support the model assumptions, this would improve credibility in the ﬁndings. If the diagnostic assessment shows remaining underlying structure in the residuals, we should try to adjust the model to account for that structure. If we are unable to do so, we may still report the model but must also note its shortcomings. In the case of the auction data, we report that there may be a nonlinear relationship between the total price and the number of wheels included for an auction. This information would be important to buyers and sellers; omitting this information could be a setback to the very people who the model might assist.

12An especially rigorous check would use time series methods. For instance, we could check whether consecutive residuals are correlated. Doing so with these residuals yields no statistically signiﬁcant correlations.

6.3. CHECKING MODEL ASSUMPTIONS USING GRAPHS 273

●

●

●

●

●

●

●

●

●

●

● ●

●

● ●

●

● ●

●

●

●

●

●

●

● ● ●

●

●

●

●

● ● ●

●

●

●

●

● ●

●

● ●

● ● ●

● ● ● ●

● ●

●

●

● ●

●

●

●

●

●

● ●

●

● ●

●

● ●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

● ●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

● ●

●

●

● ● ●

●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

●

● ●

●

●

●

●

●

●

Order of collection

Residuals

0 20 40 60 80 100 120 140

−10

0

10

Figure 6.11: Plotting residuals in the order that their corresponding observations were collected helps identify connections between successive observations. If it seems that consecutive observations tend to be close to each other, this indicates the independence assumption of the observations would fail.

274 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

Condition

Residuals

Used New

−10

0

10

●

●

● ●

●

●

●

●

●

●

● ●

●

●●

●

● ●

●

●

●

●

●

●

●● ●

●

●

●

●

● ● ●

●

●

●

●

●

●

●

● ●

●

●

●

● ●

●

●

● ●

●

●

●

●

●

●

●

●

●

●●

●

●

●

●

●● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

● ● ●

●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

●

●●

●

●

●

●

●

●

Number of wheels

Residuals

0 1 2 3 4

−10

0

10

Figure 6.12: In the two-level variable for the game’s condition, we check for diﬀerences in distribution shape or variability. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the wheels variable.

6.4. LOGISTIC REGRESSION 275

“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.

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.

TIP: Conﬁdence intervals in multiple regression Conﬁdence intervals for coeﬃcients in multiple regression can be computed using the same formula as in the single predictor model: bi ± t? dfSEbi where t? df is the appropriate t value corresponding to the conﬁdence level and model degrees of freedom, df = n−k−1.

6.4 Logistic regression

In this section we introduce logistic regression as a tool for building models when there is a categorical response variable with two levels. Logistic regression is a type of generalized linear model (GLM) for response variables where regular multiple regression does not work very well. In particular, the response variable in these settings often takes a form where residuals look completely diﬀerent from the normal distribution. GLMs can be thought of as a two-stage modeling approach. We ﬁrst model the response variable using a probability distribution, such as the binomial or Poisson distribution. Second, we model the parameter of the distribution using a collection of predictors and a special form of multiple regression. In Section 6.4 we will revisit the email data set from Chapter 1. These emails were collected from a single email account, and we will work on developing a basic spam ﬁlter using these data. The response variable, spam, has been encoded to take value 0 when a message is not spam and 1 when it is spam. Our task will be to build an appropriate model that classiﬁes messages as spam or not spam using email characteristics coded as predictor variables. While this model will not be the same as those used in large-scale spam ﬁlters, it shares many of the same features.

276 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

variable description spam Speciﬁes whether the message was spam. to multiple An indicator variable for if more than one person was listed in the To ﬁeld of the email. cc An indicator for if someone was CCed on the email. attach An indicator for if there was an attachment, such as a document or image. dollar An indicator for if the word “dollar” or dollar symbol ($) appeared in the email. winner An indicator for if the word “winner” appeared in the email message. inherit An indicator for if the word “inherit” (or a variation, like “inheritance”) appeared in the email. password An indicator for if the word “password” was present in the email. format Indicates if the email contained special formatting, such as bolding, tables, or links re subj Indicates whether “Re:” was included at the start of the email subject. exclaim subj Indicates whether any exclamation point was included in the email subject.

Table 6.13: Descriptions for 11 variables in the email data set. Notice that all of the variables are indicator variables, which take the value 1 if the speciﬁed characteristic is present and 0 otherwise.

6.4.1 Email data

The email data set was ﬁrst presented in Chapter 1 with a relatively small number of variables. In fact, there are many more variables available that might be useful for classifying spam. Descriptions of these variables are presented in Table 6.13. The spam variable will be the outcome, and the other 10 variables will be the model predictors. While we have limited the predictors used in this section to be categorical variables (where many are represented as indicator variables), numerical predictors may also be used in logistic regression. See the footnote for an additional discussion on this topic.13

6.4.2 Modeling the probability of an event

TIP: Notation for a logistic regression model The outcome variable for a GLM is denoted by Yi, where the index i is used to represent observation i. In the email application, Yi will be used to represent whether email i is spam (Yi = 1) or not (Yi = 0).

The predictor variables are represented as follows: x1,i is the value of variable 1 for observation i, x2,i is the value of variable 2 for observation i, and so on.

Logistic regression is a generalized linear model where the outcome is a two-level categorical variable. The outcome, Yi, takes the value 1 (in our application, this represents a spam message) with probability pi and the value 0 with probability 1 − pi. It is the probability pi that we model in relation to the predictor variables.

13Recall from Chapter 5 that if outliers are present in predictor variables, the corresponding observations may be especially inﬂuential on the resulting model. This is the motivation for omitting the numerical variables, such as the number of characters and line breaks in emails, that we saw in Chapter 1. These variables exhibited extreme skew. We could resolve this issue by transforming these variables (e.g. using a log-transformation), but we will omit this further investigation for brevity.

6.4. LOGISTIC REGRESSION 277

−6 −4 −2 0 2 4 6

0.0

0.2

0.4

0.6

0.8

1.0

logit(pi)

pi

● ( 6.0, 0.998)

● ( 5.0, 0.993)

● ( 4.0, 0.982)

● ( 3.0, 0.95)

●( 2.0, 0.88)

●( 1.0, 0.73)

●( 0.0, 0.50)

●(−1.0, 0.27)

● (−2.0, 0.12)

● (−3.0, 0.05)

● (−4.0, 0.018)

● (−5.0, 0.007)

Figure 6.14: Values of pi against values of logit(pi).

The logistic regression model relates the probability an email is spam (pi) to the predictors x1,i, x2,i, ..., xk,i through a framework much like that of multiple regression:

transformation(pi) = β0 + β1x1,i + β2x2,i +···βkxk,i (6.19)

We want to choose a transformation in Equation (6.19) that makes practical and mathematical sense. For example, we want a transformation that makes the range of possibilities on the left hand side of Equation (6.19) equal to the range of possibilities for the right hand side; if there was no transformation for this equation, the left hand side could only take values between 0 and 1, but the right hand side could take values outside of this range. A common transformation for pi is the logit transformation, which may be written as logit(pi) = logepi 1−pi

The logit transformation is shown in Figure 6.14. Below, we rewrite Equation (6.19) using the logit transformation of pi: logepi 1−pi= β0 + β1x1,i + β2x2,i +···+ βkxk,i

In our spam example, there are 10 predictor variables, so k = 10. This model isn’t very intuitive, but it still has some resemblance to multiple regression, and we can ﬁt this model using software. In fact, once we look at results from software, it will start to feel like we’re back in multiple regression, even if the interpretation of the coeﬃcients is more complex.

278 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION Example 6.20 Here we create a spam ﬁlter with a single predictor: to multiple. This variable indicates whether more than one email address was listed in the To ﬁeld of the email. The following logistic regression model was ﬁt using statistical software: logpi 1−pi= −2.12−1.81×to multiple If an email is randomly selected and it has just one address in the To ﬁeld, what is the probability it is spam? What if more than one address is listed in the To ﬁeld?

If there is only one email in the To ﬁeld, then to multiple takes value 0 and the right side of the model equation equals -2.12. Solving for pi: e−2.12 1+e−2.12 = 0.11. Just as we labeled a ﬁtted value of yi with a “hat” in single-variable and multiple regression, we will do the same for this probability: ˆ pi = 0.11. If there is more than one address listed in the To ﬁeld, then the right side of the model equation is −2.12−1.81×1 = −3.93, which corresponds to a probability ˆ pi = 0.02. Notice that we could examine -2.12 and -3.93 in Figure 6.14 to estimate the probability before formally calculating the value.

To convert from values on the regression-scale (e.g. -2.12 and -3.93 in Example 6.20), use the following formula, which is the result of solving for pi in the regression model:

pi =

eβ0+β1x1,i+···+βkxk,i 1 + eβ0+β1x1,i+···+βkxk,i As with most applied data problems, we substitute the point estimates for the parameters (the βi) so that we may make use of this formula. In Example 6.20, the probabilities were calculated as e−2.12 1 + e−2.12 = 0.11 e−2.12−1.81 1 + e−2.12−1.81 = 0.02 While the information about whether the email is addressed to multiple people is a helpful start in classifying email as spam or not, the probabilities of 11% and 2% are not dramatically diﬀerent, and neither provides very strong evidence about which particular email messages are spam. To get more precise estimates, we’ll need to include many more variables in the model. We used statistical software to ﬁt the logistic regression model with all ten predictors described in Table 6.13. Like multiple regression, the result may be presented in a summary table, which is shown in Table 6.15. The structure of this table is almost identical to that of multiple regression; the only notable diﬀerence is that the p-values are calculated using the normal distribution rather than the t distribution. Just like multiple regression, we could trim some variables from the model using the p-value. Using backwards elimination with a p-value cutoﬀ of 0.05 (start with the full model and trim the predictors with p-values greater than 0.05), we ultimately eliminate the exclaim subj, dollar, inherit, and cc predictors. The remainder of this section will rely on this smaller model, which is summarized in Table 6.16. J Guided Practice 6.21 Examine the summary of the reduced model in Table 6.16, and in particular, examine the to multiple row. Is the point estimate the same as we found before, -1.81, or is it diﬀerent? Explain why this might be.14

14The new estimate is diﬀerent: -2.87. This new value represents the estimated coeﬃcient when we are also accounting for other variables in the logistic regression model.

6.4. LOGISTIC REGRESSION 279

Estimate Std. Error z value Pr(>|z|) (Intercept) -0.8362 0.0962 -8.69 0.0000 to multiple -2.8836 0.3121 -9.24 0.0000 winner 1.7038 0.3254 5.24 0.0000 format -1.5902 0.1239 -12.84 0.0000 re subj -2.9082 0.3708 -7.84 0.0000 exclaim subj 0.1355 0.2268 0.60 0.5503 cc -0.4863 0.3054 -1.59 0.1113 attach 0.9790 0.2170 4.51 0.0000 dollar -0.0582 0.1589 -0.37 0.7144 inherit 0.2093 0.3197 0.65 0.5127 password -1.4929 0.5295 -2.82 0.0048

Table 6.15: Summary table for the full logistic regression model for the spam ﬁlter example.

Estimate Std. Error z value Pr(>|z|) (Intercept) -0.8595 0.0910 -9.44 0.0000 to multiple -2.8372 0.3092 -9.18 0.0000 winner 1.7370 0.3218 5.40 0.0000 format -1.5569 0.1207 -12.90 0.0000 re subj -3.0482 0.3630 -8.40 0.0000 attach 0.8643 0.2042 4.23 0.0000 password -1.4871 0.5290 -2.81 0.0049

Table 6.16: Summary table for the logistic regression model for the spam ﬁlter, where variable selection has been performed.

Point estimates will generally change a little – and sometimes a lot – depending on which other variables are included in the model. This is usually due to colinearity in the predictor variables. We previously saw this in the Ebay auction example when we compared the coeﬃcient of cond new in a single-variable model and the corresponding coeﬃcient in the multiple regression model that used three additional variables (see Sections 6.1.1 and 6.1.2). Example 6.22 Spam ﬁlters are built to be automated, meaning a piece of software is written to collect information about emails as they arrive, and this information is put in the form of variables. These variables are then put into an algorithm that uses a statistical model, like the one we’ve ﬁt, to classify the email. Suppose we write software for a spam ﬁlter using the reduced model shown in Table 6.16. If an incoming email has the word “winner” in it, will this raise or lower the model’s calculated probability that the incoming email is spam?

The estimated coeﬃcient of winner is positive (1.7370). A positive coeﬃcient estimate in logistic regression, just like in multiple regression, corresponds to a positive association between the predictor and response variables when accounting for the other variables in the model. Since the response variable takes value 1 if an email is spam and 0 otherwise, the positive coeﬃcient indicates that the presence of “winner” in an email raises the model probability that the message is spam.

280 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION Example 6.23 Suppose the same email from Example 6.22 was in HTML format, meaning the format variable took value 1. Does this characteristic increase or decrease the probability that the email is spam according to the model?

Since HTML corresponds to a value of 1 in the format variable and the coeﬃcient of this variable is negative (-1.5569), this would lower the probability estimate returned from the model.

6.4.3 Practical decisions in the email application

Examples 6.22 and 6.23 highlight a key feature of logistic and multiple regression. In the spam ﬁlter example, some email characteristics will push an email’s classiﬁcation in the direction of spam while other characteristics will push it in the opposite direction. If we were to implement a spam ﬁlter using the model we have ﬁt, then each future email we analyze would fall into one of three categories based on the email’s characteristics:

1. The email characteristics generally indicate the email is not spam, and so the resulting probability that the email is spam is quite low, say, under 0.05.

2. The characteristics generally indicate the email is spam, and so the resulting probability that the email is spam is quite large, say, over 0.95.

3. The characteristics roughly balance each other out in terms of evidence for and against the message being classiﬁed as spam. Its probability falls in the remaining range, meaning the email cannot be adequately classiﬁed as spam or not spam.

If we were managing an email service, we would have to think about what should be done in each of these three instances. In an email application, there are usually just two possibilities: ﬁlter the email out from the regular inbox and put it in a “spambox”, or let the email go to the regular inbox. J Guided Practice 6.24 The ﬁrst and second scenarios are intuitive. If the evidence strongly suggests a message is not spam, send it to the inbox. If the evidence strongly suggests the message is spam, send it to the spambox. How should we handle emails in the third category?15 J Guided Practice 6.25 Suppose we apply the logistic model we have built as a spam ﬁlter and that 100 messages are placed in the spambox over 3 months. If we used the guidelines above for putting messages into the spambox, about how many legitimate (non-spam) messages would you expect to ﬁnd among the 100 messages?16

Almost any classiﬁer will have some error. In the spam ﬁlter guidelines above, we have decided that it is okay to allow up to 5% of the messages in the spambox to be real messages. If we wanted to make it a little harder to classify messages as spam, we could use a cutoﬀ of 0.99. This would have two eﬀects. Because it raises the standard for what can be classiﬁed as spam, it reduces the number of good emails that are classiﬁed as spam.

15In this particular application, we should err on the side of sending more mail to the inbox rather than mistakenly putting good messages in the spambox. So, in summary: emails in the ﬁrst and last categories go to the regular inbox, and those in the second scenario go to the spambox. 16First, note that we proposed a cutoﬀ for the predicted probability of 0.95 for spam. In a worst case scenario, all the messages in the spambox had the minimum probability equal to about 0.95. Thus, we should expect to ﬁnd about 5 or fewer legitimate messages among the 100 messages placed in the spambox.

6.4. LOGISTIC REGRESSION 281

However, it will also fail to correctly classify an increased fraction of spam messages. No matter the complexity and the conﬁdence we might have in our model, these practical considerations are absolutely crucial to making a helpful spam ﬁlter. Without them, we could actually do more harm than good by using our statistical model.

6.4.4 Diagnostics for the email classiﬁer

Logistic regression conditions There are two key conditions for ﬁtting a logistic regression model: 1. Each predictor xi is linearly related to logit(pi) if all other predictors are held constant. 2. Each outcome Yi is independent of the other outcomes.

The ﬁrst condition of the logistic regression model is not easily checked without a fairly sizable amount of data. Luckily, we have 3,921 emails in our data set! Let’s ﬁrst visualize these data by plotting the true classiﬁcation of the emails against the model’s ﬁtted probabilities, as shown in Figure 6.17. The vast majority of emails (spam or not) still have ﬁtted probabilities below 0.5.

Predicted probability

0.0 0.2 0.4 0.6 0.8 1.0

0 (not spam)

1 (spam)

Figure 6.17: 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.

This may at ﬁrst seem very discouraging: we have ﬁt a logistic model to create a spam ﬁlter, but no emails have a ﬁtted probability of being spam above 0.75. Don’t despair; we will discuss ways to improve the model through the use of better variables in Section 6.4.5. We’d like to assess the quality of our model. For example, we might ask: if we look at emails that we modeled as having a 10% chance of being spam, do we ﬁnd about 10% of them actually are spam? To help us out, we’ll borrow an advanced statistical method called natural splines that estimates the local probability over the region 0.00 to 0.75 (the largest predicted probability was 0.73, so we avoid extrapolating). All you need to know about natural splines to understand what we are doing is that they are used to ﬁt ﬂexible lines rather than straight lines.

282 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

Predicted probability

Truth

What we expect if the logistic model is reasonable

0.0 0.2 0.4 0.6 0.8 1.0

0 (not spam)

0.2

0.4

0.6

0.8

1 (spam)

Locally−estimated probabilities with confidence bounds

The bounds become wide because not much data are found this far right

Figure 6.18: The solid black line provides the empirical estimate of the probability for observations based on their predicted probabilities (conﬁdence bounds are also shown for this line), which is ﬁt using natural splines. A small amount of noise was added to the observations in the plot to allow more observations to be seen.

The curve ﬁt using natural splines is shown in Figure 6.18 as a solid black line. If the logistic model ﬁts well, the curve should closely follow the dashed y = x line. We have added shading to represent the conﬁdence bound for the curved line to clarify what ﬂuctuations might plausibly be due to chance. Even with this conﬁdence bound, there are weaknesses in the ﬁrst model assumption. The solid curve and its conﬁdence bound dips below the dashed line from about 0.1 to 0.3, and then it drifts above the dashed line from about 0.35 to 0.55. These deviations indicate the model relating the parameter to the predictors does not closely resemble the true relationship. We could evaluate the second logistic regression model assumption – independence of the outcomes – using the model residuals. The residuals for a logistic regression model are calculated the same way as with multiple regression: the observed outcome minus the expected outcome. For logistic regression, the expected value of the outcome is the ﬁtted probability for the observation, and the residual may be written as

ei = Yi − ˆ pi We could plot these residuals against a variety of variables or in their order of collection, as we did with the residuals in multiple regression. However, since the model will need to be revised to eﬀectively classify spam and you have already seen similar residual plots in Section 6.3, we won’t investigate the residuals here.

6.4. LOGISTIC REGRESSION 283

6.4.5 Improving the set of variables for a spam ﬁlter

If we were building a spam ﬁlter for an email service that managed many accounts (e.g. Gmail or Hotmail), we would spend much more time thinking about additional variables that could be useful in classifying emails as spam or not. We also would use transformations or other techniques that would help us include strongly skewed numerical variables as predictors. Take a few minutes to think about additional variables that might be useful in identifying spam. Below is a list of variables we think might be useful:

(1) An indicator variable could be used to represent whether there was prior two-way correspondence with a message’s sender. For instance, if you sent a message to john@example.com and then John sent you an email, this variable would take value 1 for the email that John sent. If you had never sent John an email, then the variable would be set to 0.

(2) A second indicator variable could utilize an account’s past spam ﬂagging information. The variable could take value 1 if the sender of the message has previously sent messages ﬂagged as spam.

(3) A third indicator variable could ﬂag emails that contain links included in previous spam messages. If such a link is found, then set the variable to 1 for the email. Otherwise, set it to 0.

The variables described above take one of two approaches. Variable (1) is specially designed to capitalize on the fact that spam is rarely sent between individuals that have two-way communication. Variables (2) and (3) are specially designed to ﬂag common spammers or spam messages. While we would have to verify using the data that each of the variables is eﬀective, these seem like promising ideas. Table 6.19 shows a contingency table for spam and also for the new variable described in (1) above. If we look at the 1,090 emails where there was correspondence with the sender in the preceding 30 days, not one of these message was spam. This suggests variable (1) would be very eﬀective at accurately classifying some messages as not spam. With this single variable, we would be able to send about 28% of messages through to the inbox with conﬁdence that almost none are spam.

prior correspondence no yes Total spam 367 0 367 not spam 2464 1090 3554 Total 2831 1090 3921

Table 6.19: A contingency table for spam and a new variable that represents whether there had been correspondence with the sender in the preceding 30 days.

The variables described in (2) and (3) would provide an excellent foundation for distinguishing messages coming from known spammers or messages that take a known form of spam. To utilize these variables, we would need to build databases: one holding email addresses of known spammers, and one holding URLs found in known spam messages. Our access to such information is limited, so we cannot implement these two variables in this

284 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

textbook. However, if we were hired by an email service to build a spam ﬁlter, these would be important next steps. In addition to ﬁnding more and better predictors, we would need to create a customized logistic regression model for each email account. This may sound like an intimidating task, but its complexity is not as daunting as it may at ﬁrst seem. We’ll save the details for a statistics course where computer programming plays a more central role. For what is the extremely challenging task of classifying spam messages, we have made a lot of progress. We have seen that simple email variables, such as the format, inclusion of certain words, and other circumstantial characteristics, provide helpful information for spam classiﬁcation. Many challenges remain, from better understanding logistic regression to carrying out the necessary computer programming, but completing such a task is very nearly within your reach.

6.5. EXERCISES 285

6.5 Exercises

6.5.1 Introduction to multiple regression

6.1 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.17 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. (b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers. (c) Is there a statistically signiﬁcant relationship between the average birth weight and smoking?

6.2 Baby weights, Part II. Exercise 6.1 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. (b) Interpret the slope in this context, and calculate the predicted birth weight of ﬁrst borns and others. (c) Is there a statistically signiﬁcant relationship between the average birth weight and parity?

17Child Health and Development Studies, Baby weights data set.

286 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

6.3 Baby weights, Part III. We considered the variables smoke and parity, one at a time, in modeling birth weights of babies in Exercises 6.1 and 6.2. 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. (b) Interpret the slopes of gestation and age in this context. (c) The coeﬃcient for parity is diﬀerent than in the linear model shown in Exercise 6.2. Why might there be a diﬀerence? (d) Calculate the residual for the ﬁrst observation in the data set. (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.

6.5. EXERCISES 287

6.4 Absenteeism. Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled students in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.

eth sex lrn days 1 0 1 1 2 2 0 1 1 11 . . . . . . . . . . . . . . . 146 1 0 0 37

The summary table below shows the results of a linear regression model for predicting the average number of days absent based on ethnic background (eth: 0 - aboriginal, 1 - not aboriginal), sex (sex: 0 - female, 1 - male), and learner status (lrn: 0 - average learner, 1 - slow learner).18 Estimate Std. Error t value Pr(>|t|) (Intercept) 18.93 2.57 7.37 0.0000 eth -9.11 2.60 -3.51 0.0000 sex 3.10 2.64 1.18 0.2411 lrn 2.15 2.65 0.81 0.4177

(a) Write the equation of the regression line. (b) Interpret each one of the slopes in this context. (c) Calculate the residual for the ﬁrst observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school. (d) The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the R2 and the adjusted R2. Note that there are 146 observations in the data set.

6.5 GPA. A survey of 55 Duke University students asked about their GPA, number of hours they study at night, number of nights they go out, and their gender. Summary output of the regression model is shown below. Note that male is coded as 1. Estimate Std. Error t value Pr(>|t|) (Intercept) 3.45 0.35 9.85 0.00 studyweek 0.00 0.00 0.27 0.79 sleepnight 0.01 0.05 0.11 0.91 outnight 0.05 0.05 1.01 0.32 gender -0.08 0.12 -0.68 0.50

(a) Calculate a 95% conﬁdence interval for the coeﬃcient of gender in the model, and interpret it in the context of the data. (b) Would you expect a 95% conﬁdence interval for the slope of the remaining variables to include 0? Explain

18W. N. Venables and B. D. Ripley. Modern Applied Statistics with S. Fourth Edition. Data can also be found in the R MASS package. New York: Springer, 2002.

288 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

6.6 Cherry trees. Timber yield is approximately equal to the volume of a tree, however, this value is diﬃcult to measure without ﬁrst cutting the tree down. Instead, other variables, such as height and diameter, may be used to predict a tree’s volume and yield. Researchers wanting to understand the relationship between these variables for black cherry trees collected data from 31 such trees in the Allegheny National Forest, Pennsylvania. Height is measured in feet, diameter in inches (at 54 inches above ground), and volume in cubic feet.19

Estimate Std. Error t value Pr(>|t|) (Intercept) -57.99 8.64 -6.71 0.00 height 0.34 0.13 2.61 0.01 diameter 4.71 0.26 17.82 0.00

(a) Calculate a 95% conﬁdence interval for the coeﬃcient of height, and interpret it in the context of the data. (b) One tree in this sample is 79 feet tall, has a diameter of 11.3 inches, and is 24.2 cubic feet in volume. Determine if the model overestimates or underestimates the volume of this tree, and by how much.

6.5.2 Model selection

6.7 Baby weights, Part IV. Exercise 6.3 considers a model that predicts a newborn’s weight using several predictors. Use the regression table below, which summarizes the model, to answer the following questions. If necessary, refer back to Exercise 6.3 for a reminder about the meaning of each variable. 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) Determine which variables, if any, do not have a signiﬁcant linear relationship with the outcome and should be candidates for removal from the model. If there is more than one such variable, indicate which one should be removed ﬁrst. (b) The summary table below shows the results of the model with the age variable removed. Determine if any other variable(s) should be removed from the model. Estimate Std. Error t value Pr(>|t|) (Intercept) -80.64 14.04 -5.74 0.0000 gestation 0.44 0.03 15.28 0.0000 parity -3.29 1.06 -3.10 0.0020 height 1.15 0.20 5.64 0.0000 weight 0.05 0.03 2.00 0.0459 smoke -8.38 0.95 -8.82 0.0000

19D.J. Hand. A handbook of small data sets. Chapman & Hall/CRC, 1994.

6.5. EXERCISES 289

6.8 Absenteeism, Part II. Exercise 6.4 considers a model that predicts the number of days absent using three predictors: ethnic background (eth), gender (sex), and learner status (lrn). Use the regression table below to answer the following questions. If necessary, refer back to Exercise 6.4 for additional details about each variable. Estimate Std. Error t value Pr(>|t|) (Intercept) 18.93 2.57 7.37 0.0000 eth -9.11 2.60 -3.51 0.0000 sex 3.10 2.64 1.18 0.2411 lrn 2.15 2.65 0.81 0.4177

(a) Determine which variables, if any, do not have a signiﬁcant linear relationship with the outcome and should be candidates for removal from the model. If there is more than one such variable, indicate which one should be removed ﬁrst. (b) The summary table below shows the results of the regression we reﬁt after removing learner status from the model. Determine if any other variable(s) should be removed from the model. Estimate Std. Error t value Pr(>|t|) (Intercept) 19.98 2.22 9.01 0.0000 eth -9.06 2.60 -3.49 0.0006 sex 2.78 2.60 1.07 0.2878

6.9 Baby weights, Part V. Exercise 6.3 provides regression output for the full model (including all explanatory variables available in the data set) for predicting birth weight of babies. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted R2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model ﬁrst?

variable gestation parity age height weight smoke p-value 2.2×10−16 0.1052 0.2375 2.97×10−12 8.2×10−8 2.2×10−16 R2adj 0.1657 0.0013 0.0003 0.0386 0.0229 0.0569

6.10 Absenteeism, Part III. Exercise 6.4 provides regression output for the full model, including all explanatory variables available in the data set, for predicting the number of days absent from school. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted R2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model ﬁrst?

variable ethnicity sex learner status p-value 0.0007 0.3142 0.5870 R2adj 0.0714 0.0001 0

290 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

6.5.3 Checking model assumptions using graphs

6.11 Baby weights, Part V. Exercise 6.7 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.

−3 −2 −1 0 1 2 3

−60

−40

−20

0

20

40

Theoretical Quantiles

Residuals

Fitted values

Residuals

80 120 160

−40

0

40

Order of collection

Residuals

0 400 800 1200

−40

0

40

Length of gestation

Residuals

150 200 250 300 350

−40

0

40

Parity

Residuals

0 1

−40

0

40

Height of mother

Residuals

55 60 65 70

−40

0

40

Weight of mother

Residuals

100 150 200 250

−40

0

40

Smoke

Residuals

0 1

−40

0

40

6.5. EXERCISES 291

6.12 GPA and IQ. A regression model for predicting GPA from gender and IQ was ﬁt, and both predictors were found to be statistically signiﬁcant. Using the plots given below, determine if this regression model is appropriate for these data.

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●●

●

●●

●

●

●

●

●

●

●

−2 −1 0 1 2

−6

−4

−2

0

2

Theoretical Quantiles

Residuals

●

●●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

● ●

●

●

●

●

●

●

●●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●●

● ●

●

●

● ●

Fitted values

Absolute values of residuals

4 6 8 10

0

2

4

6

●

●

●

● ● ● ●

●

●

●

● ●

● ●

●

● ●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

● ● ● ●

●

● ●

●

●

●

●

● ●

●● ●

●

●

●

●

●

●

● ●

●

●

●

●

● ● ● ●

●

●

●● ● ●●

● ●

● ●

●

● ●

Order of collection

Residuals

0 40 80

−6

−2

2

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

● ●

●

●

●

●

●

●

●●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●●

●

● ●

●

●

● ●

●

●

●

IQ

Residuals

75 100 125

−6

−2

2

●

●

●

●

●

● ●

●

●

●

● ●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

● ●

●

●

●

● ●

●

●

●

●

● ●

● ● ●

●

●

●

●

●

●

● ●

●

●

●

●

● ●

●

●

●

●

● ● ● ● ●

●

●

● ●

●

●

●

Gender

Residuals

0 1

−6

−2

2

292 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

6.5.4 Logistic regression

6.13 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 5.5 on page 222). 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.

Frequency

sex\_male

0 (Female)

1 (Male)

0

20

40

60

head\_length (in mm)

Frequency

85 90 95 100

0

5

10

15

skull\_width (in mm)

Frequency

50 55 60 65

0

5

10

15

total\_length (in cm)

Frequency

75 80 85 90 95

0

5

10

tail\_length (in cm)

Frequency

32 34 36 38 40 42

0

5

10

15

20

Frequency

0 (Not Victoria)

1 (Victoria)

population

0

20

40

60

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.

6.5. EXERCISES 293

6.14 Challenger disaster, Part I. On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the ﬂight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. The table below summarizes observational data on O-rings for 23 shuttle missions, where the mission order is based on the temperature at the time of the launch. Temp gives the temperature in Fahrenheit, Damaged represents the number of damaged O-rings, and Undamaged represents the number of O-rings that were not damaged.

Shuttle Mission 1 2 3 4 5 6 7 8 9 10 11 12 Temperature 53 57 58 63 66 67 67 67 68 69 70 70 Damaged 5 1 1 1 0 0 0 0 0 0 1 0 Undamaged 1 5 5 5 6 6 6 6 6 6 5 6

Shuttle Mission 13 14 15 16 17 18 19 20 21 22 23 Temperature 70 70 72 73 75 75 76 76 78 79 81 Damaged 1 0 0 0 0 1 0 0 0 0 0 Undamaged 5 6 6 6 6 5 6 6 6 6 6

(a) Each column of the table above represents a diﬀerent shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings. (b) Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was ﬁt to these data. A summary of this model is given below. Describe the key components of this summary table in words. Estimate Std. Error z value Pr(>|z|) (Intercept) 11.6630 3.2963 3.54 0.0004 Temperature -0.2162 0.0532 -4.07 0.0000

(c) Write out the logistic model using the point estimates of the model parameters. (d) Based on the model, do you think concerns regarding O-rings are justiﬁed? Explain.

6.15 Possum classiﬁcation, Part II. A logistic regression model was proposed for classifying common brushtail possums into their two regions in Exercise 6.13. Use the results of the summary table for the reduced model presented in Exercise 6.13 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?

294 CHAPTER 6. MULTIPLE AND LOGISTIC REGRESSION

6.16 Challenger disaster, Part II. Exercise 6.14 introduced us to O-rings that were identiﬁed as a plausible explanation for the breakup of the Challenger space shuttle 73 seconds into takeoﬀ in 1986. The investigation found that the ambient temperature at the time of the shuttle launch was closely related to the damage of O-rings, which are a critical component of the shuttle. See this earlier exercise if you would like to browse the original data.

50 55 60 65 70 75 80

0.0

0.2

0.4

0.6

0.8

1.0

Probability of damage

Temperature (Fahrenheit)

(a) The data provided in the previous exercise are shown in the plot. The logistic model ﬁt to these data may be written as logˆ p 1− ˆ p= 11.6630−0.2162×Temperature where ˆ p is the model-estimated probability that an O-ring will become damaged. Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

ˆ p57 = 0.341 ˆ p59 = 0.251 ˆ p61 = 0.179 ˆ p63 = 0.124 ˆ p65 = 0.084 ˆ p67 = 0.056 ˆ p69 = 0.037 ˆ p71 = 0.024

(b) Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities. (c) Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model’s validity.