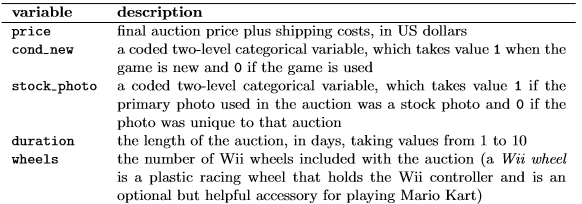
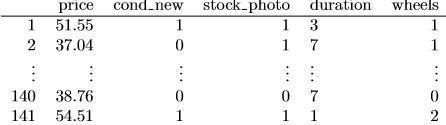
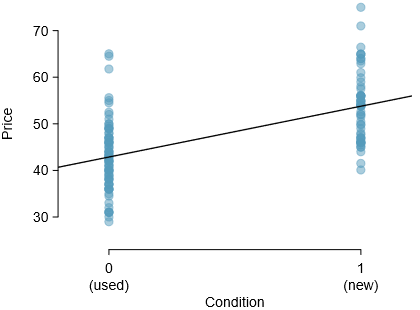
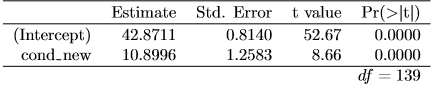
# Chapter 6: Multiple regression

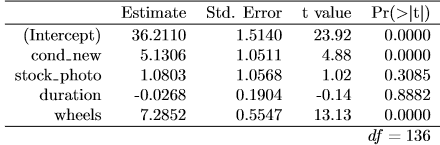
* Principles of simple linear regression lay foundation for more sophisticted regression methods used in a wide range of challenging settings
* **multiple regression** = more than 1 predictor
* **logistic regression** = predicting categorical outcomes w/ 2 possible categories.
* Multiple regression extends 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.



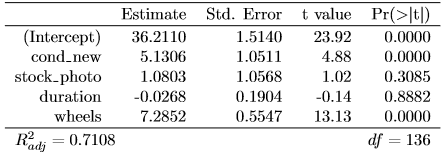
* condition + stock photo = **indicator variables** 🡺 cond = 1 if game new, 0 if used.
* Using indicator variables allows for these variables to be directly used in regression.
* Multiple regression also allows for categorical variables w/ many levels (do not have any such variables in this analysis)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



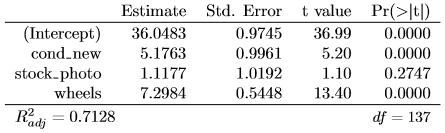
* Guided Practice 6.1 Examine Figure 6.4. Does the linear model seem reasonable?
* Yes 🡺 Constant variability, nearly normal residuals, + linearity all appear reasonable.
* Guided Practice 6.2 Interpret the coeﬃcient for the game’s condition in the model. Is this coeﬃcient signiﬁcantly diﬀerent from 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:



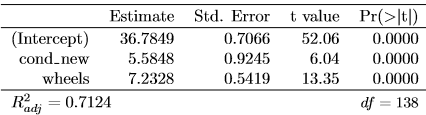
* Guided Practice 6.5 Write out the model in Equation (6.3) using the point estimates from Table 6.5. How many predictors are there in this model?
* **y = 36.2110 + 5.1306\*cond\_new + 1.0803\*stock\_photo - .0268\*duration + 7.2852\*wheels**
* Guided Practice 6.6 What does β4, the coeﬃcient of variable 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.
* 6.10 Variance of the residuals for the model in Guided Practice 6.7 = 23.34, + the variance of the total price in all auctions = 83.06. Calculate R2 for this model.
* **R2 = 1 – (23.34/83.06) = .719**
* R2 = acceptable for just 1 variable but less helpful w/ many variables b/c R2 = **biased estimate** of the amount of variability explained by the model.
* Better estimate = use **adjusted R2** = 
* n = # of cases used to ﬁt model, k = # of predictors in model
* B/c k is *never negative*, adjusted R2 = smaller (often times just a little) than R2.
* We adjust R2 b/c of the dF associated w/ each variance.
* *Multiple* regression, dF associated w/ variance of the estimate of the residuals = **n−k−1**, NOT n−1
* If making predictions for new data using our current model, R2 = overly optimistic estimate of reduction in variance in the response, + using dF in *adjusted* R2 helps correct this bias
* 6.11: There were n = 141 auctions in the Mario Kart data set + k = 4 predictors in the model. Use n, k, + the variances from 6.10 to calculate R2adj for the Mario Kart model
* **AdjR2 = 1 – ((23.34/(141-4-1)) / (83.06/141-1)) = .7107**
* 6.12 Suppose you added another predictor to the model, but the variance of the errors Var(e\_i) didn’t go down. What would happen to the R2? What would happen to the adjusted R2?
* R2 would not be affected, but AdjR2 would decrease
* Best model = NOT always most complicated
* Sometimes including variables that’re not evidently important can reduce prediction accuracy
* 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.
* Regression output for full model for auction data.

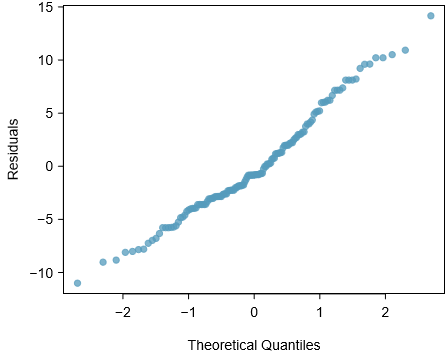
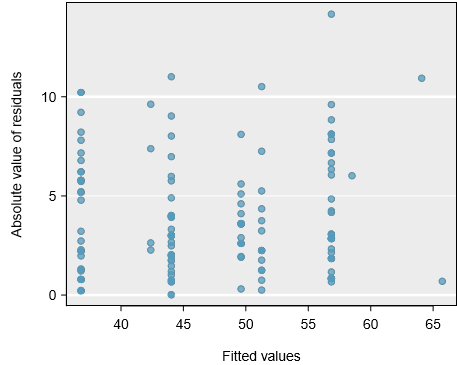


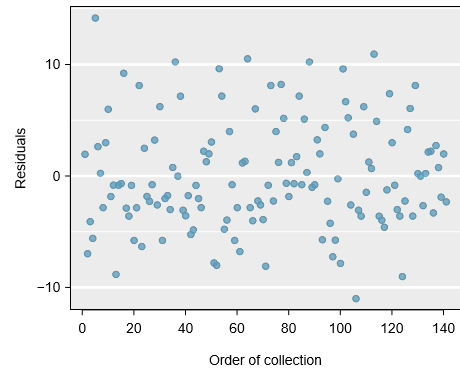
* H0: βi = 0 when the other predictors are included in the model.
* HA: βi != 0 when the other predictors are included in the model.
* 6.13 The coeﬃcient of **cond new** has a t-test statistic = 4.88 + a p-value for its corresponding hypotheses (H0 : β1 = 0, HA : β1 6= 0) of ~0. How can this be interpreted?
* The chances of observing the relationship between cond\_new and the outcome price if H0 is true is close to null
* I.e. There is strong evidence of a relationship between cond\_new and the outcome price, **all other predictors held constant**
* 6.14 Is there strong evidence that using a stock photo is related to the total auction price?
* stock\_photo has a small t-statistic = 1.02 + a large p-value = ~0.31.
* This suggests the chances of observing the relationship between stock\_photo + the outcome price if H0 is true is close to null
* I.e. There NOT is strong evidence of a relationship between stock\_photo + price, **all other predictors held constant**
* 6.15 Identify the p-values for both duration + wheels. Is there strong evidence supporting the connection of these variables with the total price in the model?
* Yes for **wheels**, VERY not so for **duration**
* **TIP: Using adjusted R2 instead of p-values for model selection**
* Adjusted R2 may be used as an alternative to p-values for model selection, where higher adjusted R2 = a better model ﬁt.
* Could compare 2 models using adjusted R2 + model w/ higher adjusted R2 would be preferred.
* This approach tends to include *more* variables in a ﬁnal model when compared to the p-value approach.
* **Backward-selection and forward-selection** = **stepwise model selection strategies**
* 6.16 How should we proceed under the backward-elimination strategy?
* Remove **duration +** refit model



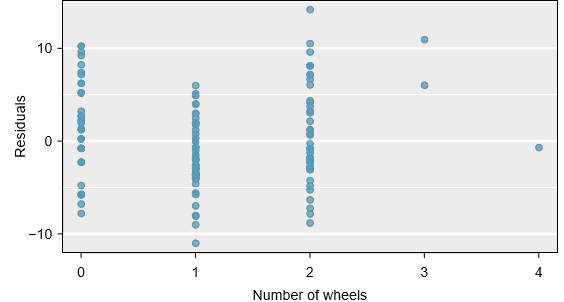
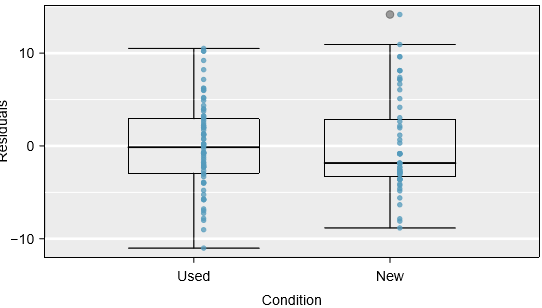
* Rempove **stock photo** + refit



* 2 remaining predictors have statistically significant coeﬃcients w/ p-values = ~0 🡺 Stop.
* Final model includes only **cond new** + **wheels** in predicting the total auction **price**:
* yˆ = b0 + b1x1 + b4x4 = 36.78 + 5.58x1 + 7.23x4
* Alternative model selection = use adjusted R2.
* At each elimination step, reﬁt model w/out each predictor up for potential elimination.
* 1st step, ﬁt 4 models, where each would be missing a diﬀerent predictor.
* If one of these smaller models has a higher adjusted R2 than current model, pick smaller model w/ largest adjusted R2.
* Continue until removing variables does not increase R2adj.
* Had we used adjusted R2 criteria, we would’ve kept **stock photo** along w/ **cond new** + **wheels**
* Common for p-values of a variable to change, due to **collinearity**, after eliminating another variable
* This ﬂuctuation emphasizes **importance of reﬁtting a model after each variable elimination step**.
* p-values tend to change dramatically when eliminated variable = highly correlated w/ another in the model.
* Forward-selection = add variables 1-at-a-time until we cannot ﬁnd any variables that present strong evidence of importance in model.
* Start w/ model w/ no variables 🡺 ﬁt each possible models w/ just 1 variable.
* Each of the 4 models provides a p-value for coeﬃcient of the predictor
* Out of these 4 variables, wheels = smallest p-value 🡺 add wheels to the model.
* Fit 3 new models 🡺 Of these models, cond\_new + wheels = lowest p-value
* Repeat process 🡺 neither of remaining variables is added = ﬁnished.
* Final model = same as backward-selection strategy.
* 6.18 As before, could’ve used R2adj criteria instead of examining p-values in selecting variables for the model. Rather than look for variables w/ smallest p-value, look for model w/ largest R2adj. What would the result of forward-selection be using the adjusted R2 approach?
* If we had done forward-selection strategy using R2adj, would’ve arrived at model including cond\_new, stock photo, + wheels = slightly larger model than if using p-value approach
* = same model from using adjusted R2 backwards-elimination.
* **NO GUARANTEE** backward-elimination + forward-selection strategies will arrive at same ﬁnal model using the p-value OR adjusted R2 methods.
* If both are tried + arrive at diﬀerent models, choose model w/ larger R2adj as a tie-breaker (other tie-break options exist)
* Generally acceptable to use just 1 strategy, usually backward-elimination w/ either the p-value or adjusted R2 criteria.
* However, before reporting model results, **must verify the model conditions are reasonable.**
* Multiple regression methods using model yˆ = β0 + β1x1 + β2x2 +···+ βkxk generally depend on the following 4 assumptions:
* **1. Residuals of model = nearly normal**
* **2. Variability of residuals = nearly constant**
* **3. Residuals = identically and independently distributed (IID)**
* **4. Each variable = *linearly* related to the outcome.**
* **Normal probability plot of residuals:** 
* plot exhibits some minor irregularities, but there are no outliers that might be cause for concern
* Tend to be most worried about residuals that appear to be outliers, which indicate long tails in the distribution of residuals.
* **Absolute values of residuals against ﬁtted values (**yˆi): 
* helpful to check if variance of residuals = approximately constant.
* Don’t see any obvious deviations from constant variance in this example.
* **Residuals in order of their data collection** (order of corresponding auctions)



* Such a plot = helpful in IDing any connection between cases that’re close to one another,
* could look for declining prices over time or if there was a time of day when auctions tended to fetch a higher price.
* Here we see no structure that indicates a problem
* *Especially* rigorous check would use **time series methods** to, for examples, check whether consecutive residuals = correlated.
* Doing so w/ these residuals yields no statistically signiﬁcant correlations.
* **Residuals against each predictor variable.**



* For the 2-level **condition** variable 🡺 guaranteed not to see any remaining trend, + instead are checking that variability doesn’t ﬂuctuate across groups.
* When we consider residuals against **wheels**, we *see some possible structure*.
* appears to be curvature in the residuals, indicating relationship is probably not linear.
* **It is necessary to summarize diagnostics for any model ﬁt.**
* If diagnostics *support* the model assumptions, this improves credibility in the ﬁndings.
* **If diagnostic assessment shows remaining underlying structure in the residuals, try to adjust the model to account for that structure**.
* If unable to do so, may still report the model **but must also note its shortcomings.**
* In the case of the auction data, report there may be a nonlinear relationship between **price** + # of **wheels** included for an auction.
* This info would be important to buyers + sellers, as omitting this info could be a setback to the very people who the model might assist.

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.