Chapter 8

Multiple and logistic regression

The principles of simple linear regression lay the foundation for more sophisticated regression methods used in a wide range of challenging settings. In Chapter 8, we explore multiple regression, which introduces the possibility of more than one predictor, and logistic regression, a technique for predicting categorical outcomes with two possible categories.

8.1 Introduction to multiple regression

Multiple regression extends simple two-variable regression to the case that still has one response but many predictors (denoted x_1, x_2, x_3, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

We will consider Ebay auctions of a video game called *Mario Kart* for the Nintendo Wii. The outcome variable of interest is the total price of an auction, which is the highest bid plus the shipping cost. We will try to determine how total price is related to each characteristic in an auction while simultaneously controlling for other variables. For instance, all other characteristics held constant, are longer auctions associated with higher or lower prices? And, on average, how much more do buyers tend to pay for additional Wii wheels (plastic steering wheels that attach to the Wii controller) in auctions? Multiple regression will help us answer these and other questions.

The data set `mario_kart` includes results from 141 auctions.[1] Four observations from this data set are shown in Table 8.1, and descriptions for each variable are shown in Table 8.2. Notice that the condition and stock photo variables are indicator variables. For instance, the `cond_new` variable takes value 1 if the game up for auction is new and 0 if it is used. Using indicator variables in place of category names allows for these variables to be directly used in regression. See Section 7.2.7 for additional details. Multiple regression also allows for categorical variables with many levels, though we do not have any such variables in this analysis, and we save these details for a second or third course.

[1] Diez DM, Barr CD, and Çetinkaya-Rundel M. 2012. *openintro*: OpenIntro data sets and supplemental functions. cran.r-project.org/web/packages/openintro.

	price	cond_new	stock_photo	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	7	1
⋮	⋮	⋮	⋮	⋮	⋮
140	38.76	0	0	7	0
141	54.51	1	1	1	2

Table 8.1: Four observations from the `mario_kart` data set.

variable	description
`price`	final auction price plus shipping costs, in US dollars
`cond_new`	a coded two-level categorical variable, which takes value `1` when the game is new and `0` if the game is used
`stock_photo`	a coded two-level categorical variable, which takes value `1` if the primary photo used in the auction was a stock photo and `0` if the photo was unique to that auction
`duration`	the length of the auction, in days, taking values from 1 to 10
`wheels`	the number of Wii wheels included with the auction (a *Wii wheel* is a plastic racing wheel that holds the Wii controller and is an optional but helpful accessory for playing Mario Kart)

Table 8.2: Variables and their descriptions for the `mario_kart` data set.

8.1.1 A single-variable model for the Mario Kart data

Let's fit a linear regression model with the game's condition as a predictor of auction price. The model may be written as

$$\widehat{price} = 42.87 + 10.90 \times cond\_new$$

Results of this model are shown in Table 8.3 and a scatterplot for price versus game condition is shown in Figure 8.4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.8711	0.8140	52.67	0.0000
cond_new	10.8996	1.2583	8.66	0.0000
				$df = 139$

Table 8.3: Summary of a linear model for predicting auction price based on game condition.

⊙ **Exercise 8.1** Examine Figure 8.4. Does the linear model seem reasonable?[2]

● **Example 8.2** Interpret the coefficient for the game's condition in the model. Is this coefficient significantly different from 0?

Note that `cond_new` is a two-level categorical variable that takes value 1 when the game is new and value 0 when the game is used. So 10.90 means that the model

[2]Yes. Constant variability, nearly normal residuals, and linearity all appear reasonable.

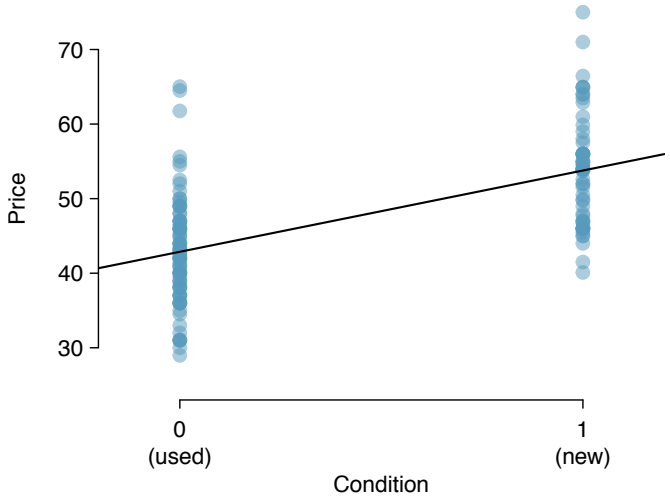


Figure 8.4: Scatterplot of the total auction price against the game's condition. The least squares line is also shown.

predicts an extra $10.90 for those games that are new versus those that are used. (See Section 7.2.7 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Table 8.3, we can see that the p-value for `cond_new` is very close to zero, indicating there is strong evidence that the coefficient is different from zero when using this simple one-variable model.

8.1.2 Including and assessing many variables in a model

Sometimes there are underlying structures or relationships between predictor variables. For instance, new games sold on Ebay tend to come with more Wii wheels, which may have led to higher prices for those auctions. We would like to fit a model that includes all potentially important variables simultaneously. This would help us evaluate the relationship between a predictor variable and the outcome while controlling for the potential influence of other variables. This is the strategy used in **multiple regression**. While we remain cautious about making any causal interpretations using multiple regression, such models are a common first step in providing evidence of a causal connection.

We want to construct a model that accounts for not only the game condition, as in Section 8.1.1, but simultaneously accounts for three other variables: `stock_photo`, `duration`, and `wheels`.

$$\begin{aligned}\widehat{\texttt{price}} &= \beta_0 + \beta_1 \times \texttt{cond\_new} + \beta_2 \times \texttt{stock\_photo} \\ &\quad + \beta_3 \times \texttt{duration} + \beta_4 \times \texttt{wheels} \\ \hat{y} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \end{aligned} \tag{8.3}$$

In this equation, y represents the total price, x_1 indicates whether the game is new, x_2 indicates whether a stock photo was used, x_3 is the duration of the auction, and x_4 is the number of Wii wheels included with the game. Just as with the single predictor case, a multiple regression model may be missing important components or it might not precisely represent the relationship between the outcome and the available explanatory variables.

While no model is perfect, we wish to explore the possibility that this one may fit the data reasonably well.

We estimate the parameters β_0, β_1, ..., β_4 in the same way as we did in the case of a single predictor. We select b_0, b_1, ..., b_4 that minimize the sum of the squared residuals:

$$SSE = e_1^2 + e_2^2 + \cdots + e_{141}^2 = \sum_{i=1}^{141} e_i^2 = \sum_{i=1}^{141} (y_i - \hat{y}_i)^2 \tag{8.4}$$

Here there are 141 residuals, one for each observation. We typically use a computer to minimize the sum in Equation (8.4) and compute point estimates, as shown in the sample output in Table 8.5. Using this output, we identify the point estimates b_i of each β_i, just as we did in the one-predictor case.

	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
				$df = 136$

Table 8.5: Output for the regression model where `price` is the outcome and `cond_new`, `stock_photo`, `duration`, and `wheels` are the predictors.

Multiple regression model

A multiple regression model is a linear model with many predictors. In general, we write the model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

when there are k predictors. We often estimate the β_i parameters using a computer.

⊙ **Exercise 8.5** Write out the model in Equation (8.3) using the point estimates from Table 8.5. How many predictors are there in this model?[3]

⊙ **Exercise 8.6** What does β_4, the coefficient of variable x_4 (Wii wheels), represent? What is the point estimate of β_4?[4]

⊙ **Exercise 8.7** Compute the residual of the first observation in Table 8.1 on page 355 using the equation identified in Exercise 8.5. [5]

[3] $\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.03x_3 + 7.29x_4$, and there are $k = 4$ predictor variables.

[4] It is the average difference in auction price for each additional Wii wheel included when holding the other variables constant. The point estimate is $b_4 = 7.29$.

[5] $e_i = y_i - \hat{y}_i = 51.55 - 49.62 = 1.93$, where 49.62 was computed using the variables values from the observation and the equation identified in Exercise 8.5.

Example 8.8 We estimated a coefficient for `cond_new` in Section 8.1.1 of $b_1 = 10.90$ with a standard error of $SE_{b_1} = 1.26$ when using simple linear regression. Why might there be a difference between that estimate and the one in the multiple regression setting?

If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome `price` and predictor `cond_new` using simple linear regression, we were unable to control for other variables like the number of Wii wheels included in the auction. That model was biased by the confounding variable `wheels`. When we use both variables, this particular underlying and unintentional bias is reduced or eliminated (though bias from other confounding variables may still remain).

Example 8.8 describes a common issue in multiple regression: correlation among predictor variables. We say the two predictor variables are **collinear** (pronounced as *co-linear*) when they are correlated, and this collinearity complicates model estimation. While it is impossible to prevent collinearity from arising in observational data, experiments are usually designed to prevent predictors from being collinear.

Exercise 8.9 The estimated value of the intercept is 36.21, and one might be tempted to make some interpretation of this coefficient, such as, it is the model's predicted price when each of the variables take value zero: the game is used, the primary image is not a stock photo, the auction duration is zero days, and there are no wheels included. Is there any value gained by making this interpretation?[6]

8.1.3 Adjusted R^2 as a better estimate of explained variance

We first used R^2 in Section 7.2 to determine the amount of variability in the response that was explained by the model:

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

where e_i represents the residuals of the model and y_i the outcomes. This equation remains valid in the multiple regression framework, but a small enhancement can often be even more informative.

Exercise 8.10 The variance of the residuals for the model given in Exercise 8.7 is 23.34, and the variance of the total price in all the auctions is 83.06. Calculate R^2 for this model.[7]

This strategy for estimating R^2 is acceptable when there is just a single variable. However, it becomes less helpful when there are many variables. The regular R^2 is actually a biased estimate of the amount of variability explained by the model. To get a better estimate, we use the adjusted R^2.

[6] Three 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.

[7] $R^2 = 1 - \frac{23.34}{83.06} = 0.719$.

Adjusted R^2 as a tool for model assessment
The **adjusted R^2** is computed as

$$R^2_{adj} = 1 - \frac{Var(e_i)/(n-k-1)}{Var(y_i)/(n-1)} = 1 - \frac{Var(e_i)}{Var(y_i)} \times \frac{n-1}{n-k-1}$$

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 R^2 will be smaller – often times just a little smaller – than the unadjusted R^2. The reasoning behind the adjusted R^2 lies in the **degrees of freedom** associated with each variance.[8]

⊙ **Exercise 8.11** 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 Exercise 8.10 to calculate R^2_{adj} for the Mario Kart model.[9]

⊙ **Exercise 8.12** 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 R^2? What would happen to the adjusted R^2?[10]

8.2 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.

8.2.1 Identifying variables in the model that may not be helpful

Table 8.6 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:

H_0: $\beta_i = 0$ when the other explanatory variables are included in the model.
H_A: $\beta_i \neq 0$ when the other explanatory variables are included in the model.

[8] In 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 R^2 is an overly optimistic estimate of the reduction in variance in the response, and using the degrees of freedom in the adjusted R^2 formula helps correct this bias.

[9] $R^2_{adj} = 1 - \frac{23.34}{83.06} \times \frac{141-1}{141-4-1} = 0.711$.

[10] The unadjusted R^2 would stay the same and the adjusted R^2 would go down.

	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
$R^2_{adj} = 0.7108$				$df = 136$

Table 8.6: The fit for the full regression model, including the adjusted R^2.

Example 8.13 The coefficient of `cond_new` has a t test statistic of $T = 4.88$ and a p-value for its corresponding hypotheses ($H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$) 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 8.14 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.

Exercise 8.15 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.

TIP: Using adjusted R^2 instead of p-values for model selection
The adjusted R^2 may be used as an alternative to p-values for model selection, where a higher adjusted R^2 represents a better model fit. For instance, we could compare two models using their adjusted R^2, and the model with the higher adjusted R^2 would be preferred. This approach tends to include more variables in the final model when compared to the p-value approach.

8.2.2 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

[11] The 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.

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 R^2_{adj} 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 8.16 Results corresponding to the *full model* for the `mario_kart` data are shown in Table 8.6. 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 8.7.

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 p-values 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 8.8.

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:

$$\begin{aligned} \hat{y} &= b_0 + b_1 x_1 + b_4 x_4 \\ &= 36.78 + 5.58 x_1 + 7.23 x_4 \end{aligned}$$

where x_1 represents `cond_new` and x_4 represents `wheels`.

An alternative to using p-values in model selection is to use the adjusted R^2. 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 R^2 than our current model, we pick the smaller model with the largest adjusted R^2. We continue in this way until removing variables does not increase R^2_{adj}. Had we used the adjusted R^2 criteria, we would have kept the `stock_photo` variable along with the `cond_new` and `wheels` variables.

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.

	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
$R^2_{adj} = 0.7128$				$df = 137$

Table 8.7: 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
$R^2_{adj} = 0.7124$				$df = 138$

Table 8.8: The output for the regression model where `price` is the outcome and the duration and stock photo variables have been eliminated from the model.

Example 8.17 Construct a model for the `mario_kart` data set using the forward-selection 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 8.8), (2) the model including just the `stock_photo` and `wheels` 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 8.7) 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 8.18 As before, we could have used the R^2_{adj} 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 R^2_{adj}. What would the result of forward-selection be using the adjusted R^2 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 R^2_{adj} 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 R^2_{adj}. If we had done the forward-selection strategy using R^2_{adj}, 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 R^2 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 R^2 methods. If the backwards-elimination and forward-selection strategies are both tried and they arrive at different models, choose the model with the larger R^2_{adj} 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 R^2 criteria. However, before reporting the model results, we must verify the model conditions are reasonable.

8.3 Checking model assumptions using graphs

Multiple regression methods using the model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

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.

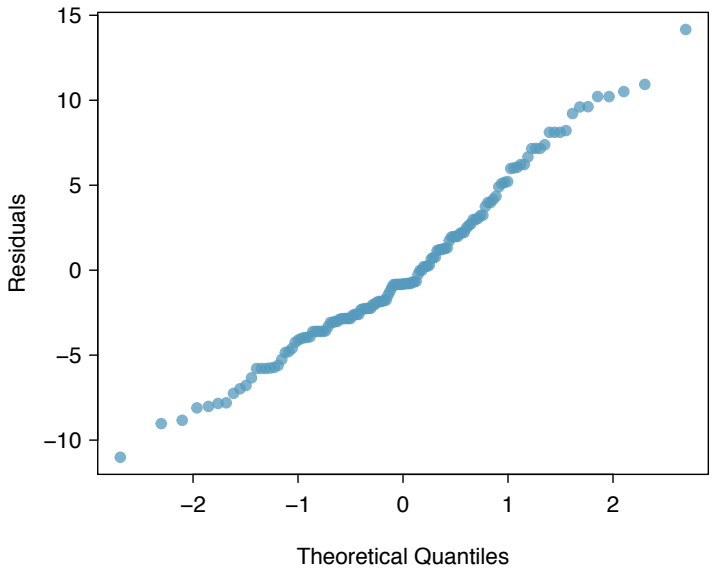


Figure 8.9: A normal probability plot of the residuals is helpful in identifying observations that might be outliers.

Normal probability plot. A normal probability plot of the residuals is shown in Figure 8.9. 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 ($\hat{y}_i$) is shown in Figure 8.10. 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.

Residuals in order of their data collection. A plot of the residuals in the order their corresponding auctions were observed is shown in Figure 8.11. 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 8.12. 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.

[12] An 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.

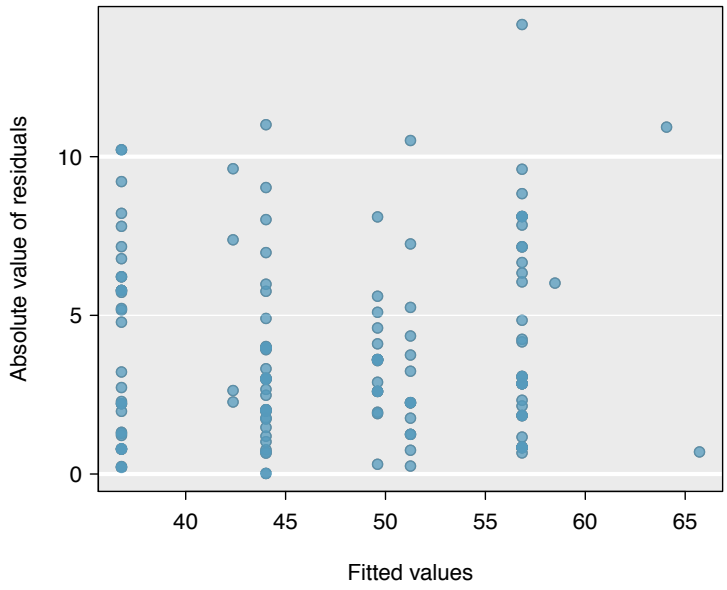


Figure 8.10: Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.

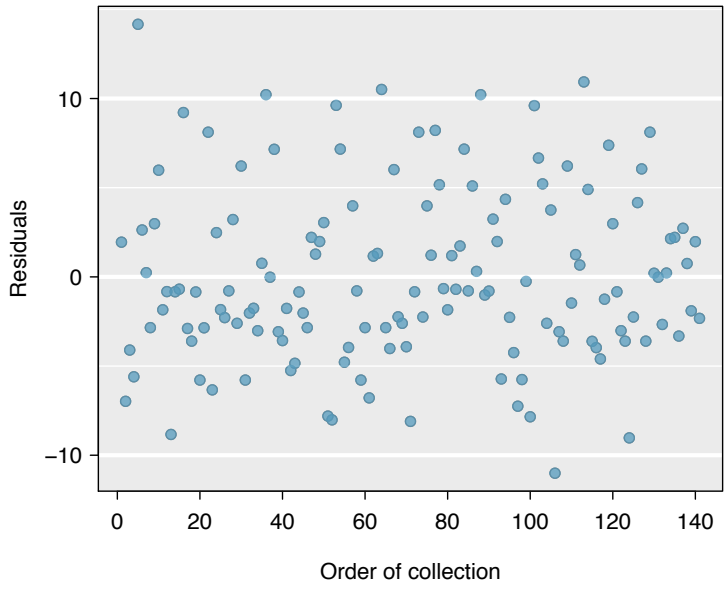


Figure 8.11: Plotting residuals in the order that their corresponding observations were collected helps identify connections between successive observations. If it seems that consecutive observations tend to be close to each other, this indicates the independence assumption of the observations would fail.

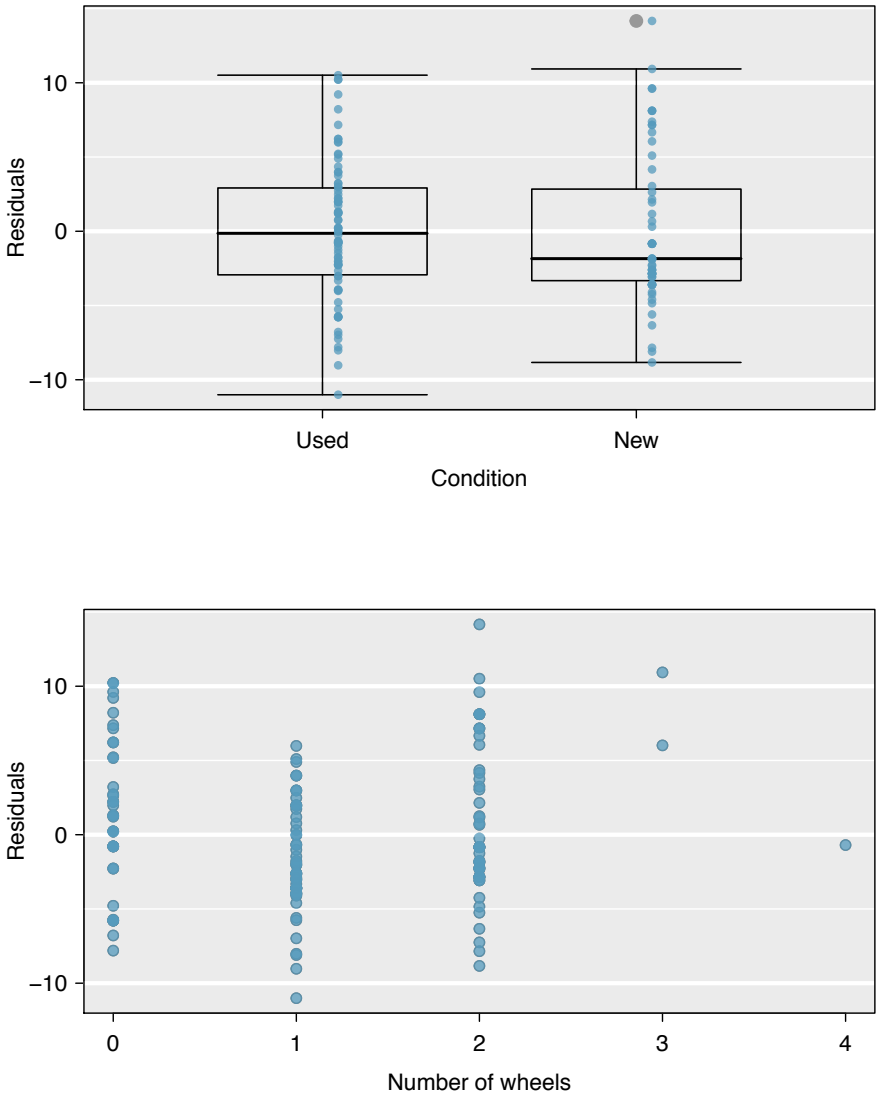


Figure 8.12: In the two-level variable for the game's condition, we check for differences in distribution shape or variability. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the `wheels` variable.

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.

"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:

$$b_i \pm t^\star_{df} SE_{b_i}$$

where $t^\star_{df}$ is the appropriate t value corresponding to the confidence level and model degrees of freedom, $df = n - k - 1$.

8.4 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 8.4 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.

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 8.13: 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.

8.4.1 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 8.13. 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]

8.4.2 Modeling the probability of an event

TIP: Notation for a logistic regression model

The outcome variable for a GLM is denoted by Y_i, where the index i is used to represent observation i. In the email application, Y_i will be used to represent whether email i is spam ($Y_i = 1$) or not ($Y_i = 0$).

The predictor variables are represented as follows: $x_{1,i}$ is the value of variable 1 for observation i, $x_{2,i}$ 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, Y_i, takes the value 1 (in our application, this represents a spam message) with probability p_i and the value 0 with probability $1 - p_i$. It is the probability p_i that we model in relation to the predictor variables.

[13] Recall from Chapter 7 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.

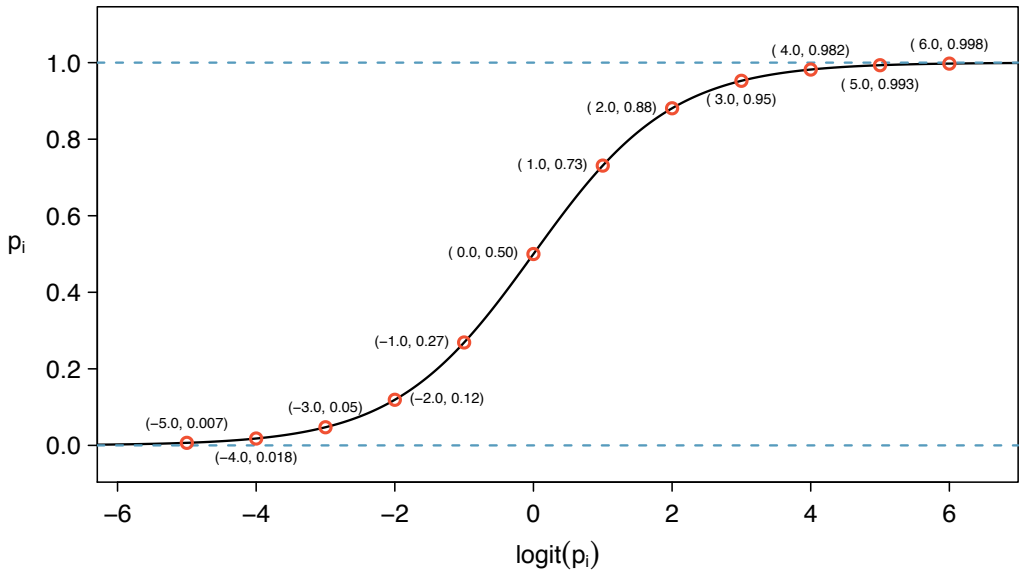


Figure 8.14: Values of p_i against values of $logit(p_i)$.

The logistic regression model relates the probability an email is spam (p_i) to the predictors $x_{1,i}$, $x_{2,i}$, ..., $x_{k,i}$ through a framework much like that of multiple regression:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots \beta_k x_{k,i} \tag{8.19}$$

We want to choose a transformation in Equation (8.19) 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 (8.19) 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 p_i is the **logit transformation**, which may be written as

$$logit(p_i) = \log_e \left(\frac{p_i}{1 - p_i} \right)$$

The logit transformation is shown in Figure 8.14. Below, we rewrite Equation (8.19) using the logit transformation of p_i:

$$\log_e \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

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.

Example 8.20 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:

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.12 - 1.81 \times \texttt{to\_multiple}$$

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 p_i: $\frac{e^{-2.12}}{1+e^{-2.12}} = 0.11$. Just as we labeled a fitted value of y_i with a "hat" in single-variable and multiple regression, we will do the same for this probability: $\hat{p}_i = 0.11$.

If there is more than one address listed in the *To* field, then the right side of the model equation is $-2.12 - 1.81 \times 1 = -3.93$, which corresponds to a probability $\hat{p}_i = 0.02$.

Notice that we could examine -2.12 and -3.93 in Figure 8.14 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 8.20), use the following formula, which is the result of solving for p_i in the regression model:

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1 \ + \ e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}$$

As with most applied data problems, we substitute the point estimates for the parameters (the β_i) so that we may make use of this formula. In Example 8.20, the probabilities were calculated as

$$\frac{e^{-2.12}}{1 \ + \ e^{-2.12}} = 0.11 \qquad\qquad \frac{e^{-2.12-1.81}}{1 \ + \ e^{-2.12-1.81}} = 0.02$$

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 8.13. Like multiple regression, the result may be presented in a summary table, which is shown in Table 8.15. 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 8.16.

Exercise 8.21 Examine the summary of the reduced model in Table 8.16, 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]

[14] The 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.

	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 8.15: 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 8.16: 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 8.1.1 and 8.1.2).

Example 8.22 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 8.16. 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.

Example 8.23 Suppose the same email from Example 8.22 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.

8.4.3 Practical decisions in the email application

Examples 8.22 and 8.23 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.

Exercise 8.24 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]

Exercise 8.25 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.

[15] In 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.

[16] First, 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.

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.

8.4.4 Diagnostics for the email classifier

Logistic regression conditions

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

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

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 8.17. The vast majority of emails (spam or not) still have fitted probabilities below 0.5.

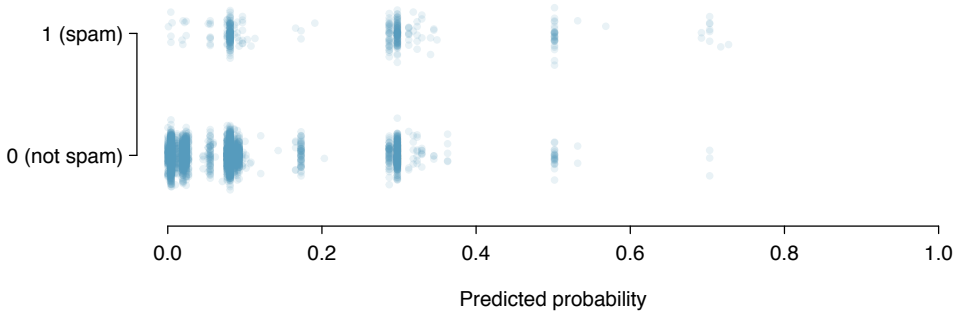


Figure 8.17: The predicted probability that each of the 3,912 emails is spam is classified by their grouping, spam or not. Noise (small, random vertical shifts) have been added to each point so that points with nearly identical values aren't plotted exactly on top of one another. This makes it possible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

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 8.4.5.

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.

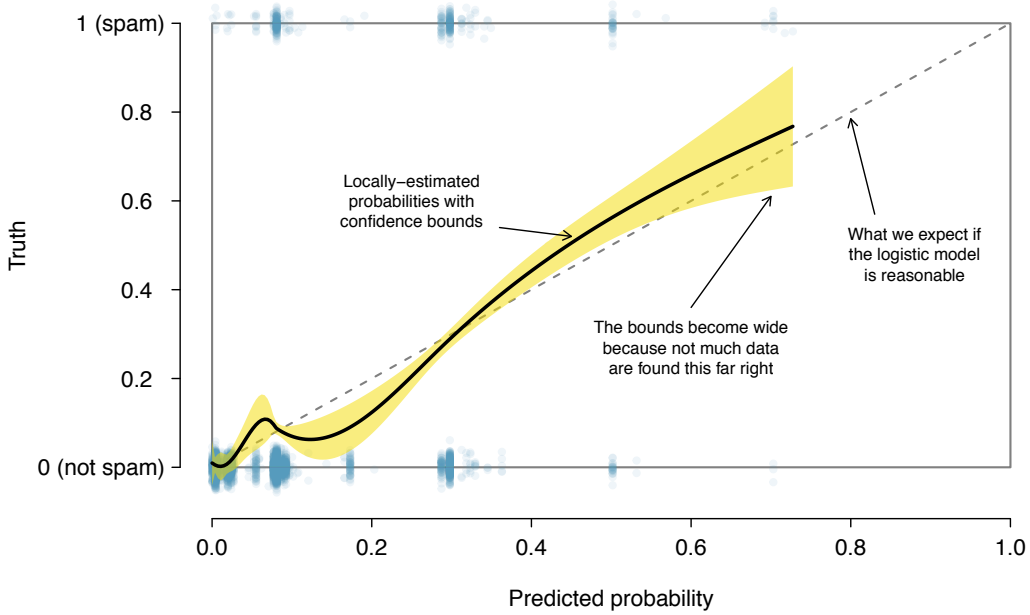


Figure 8.18: The solid black line provides the empirical estimate of the probability for observations based on their predicted probabilities (confidence bounds are also shown for this line), which is fit using natural splines. A small amount of noise was added to the observations in the plot to allow more observations to be seen.

The curve fit using natural splines is shown in Figure 8.18 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

$$e_i = Y_i - \hat{p}_i$$

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 8.3, we won't investigate the residuals here.

8.4.5 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 8.19 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 8.19: 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

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.

8.5 Exercises

8.5.1 Introduction to multiple regression

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

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

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

(a) Write the equation of the regression line.

(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?

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

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

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

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

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

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

(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 8.2. 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 R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

8.4 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 R^2 and the adjusted R^2. Note that there are 146 observations in the data set.

8.5 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

[18] W. 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.

8.6 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.

8.5.2 Model selection

8.7 Baby weights, Part IV. Exercise 8.3 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 8.3 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

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

8.8 Absenteeism, Part II. Exercise 8.4 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 8.4 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

8.9 Baby weights, Part V. Exercise 8.3 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 R^2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

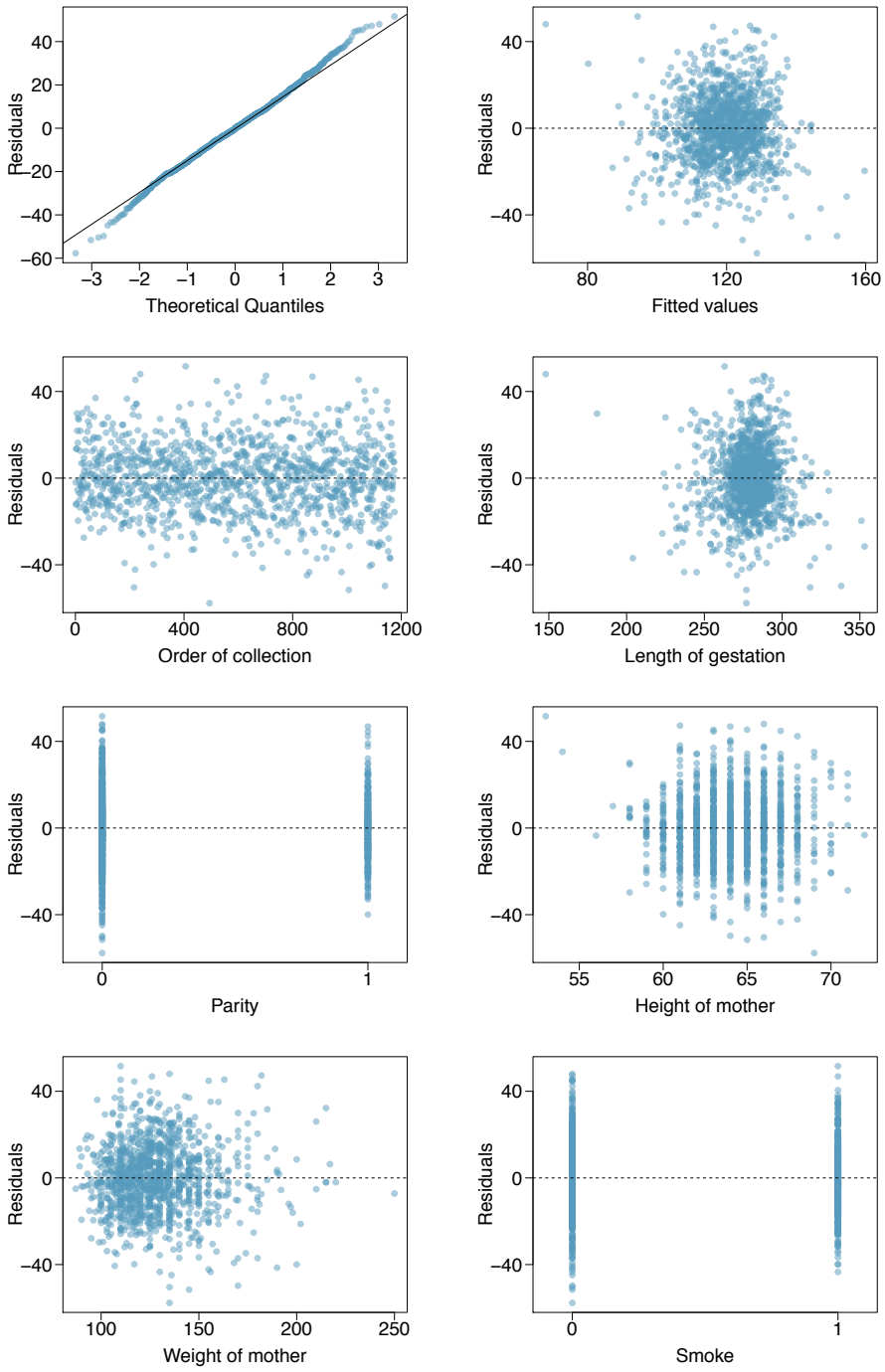
variable	gestation	parity	age	height	weight	smoke
p-value	2.2×10^{-16}	0.1052	0.2375	2.97×10^{-12}	8.2×10^{-8}	2.2×10^{-16}
R^2_{adj}	0.1657	0.0013	0.0003	0.0386	0.0229	0.0569

8.10 Absenteeism, Part III. Exercise 8.4 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 R^2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

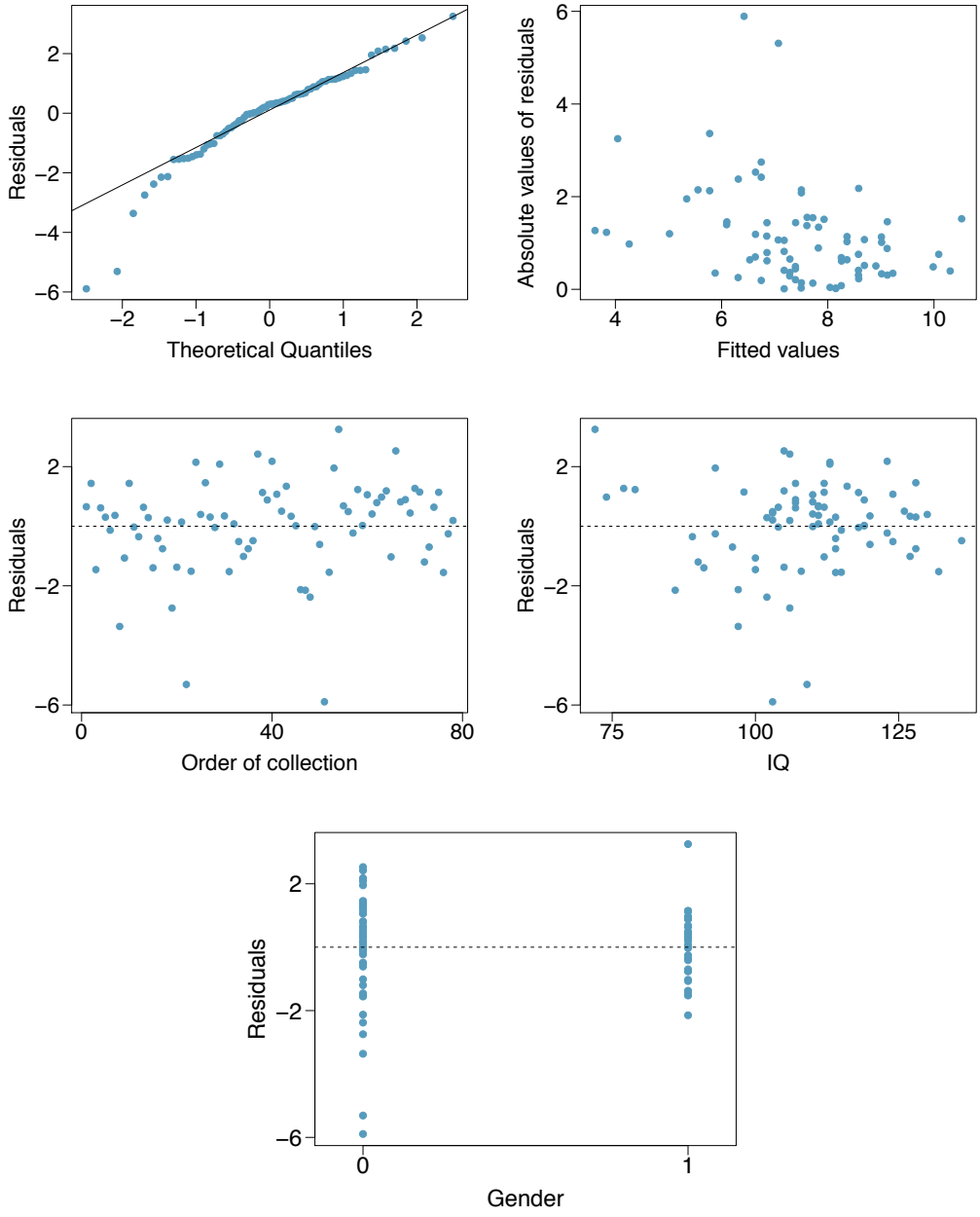
variable	ethnicity	sex	learner status
p-value	0.0007	0.3142	0.5870
R^2_{adj}	0.0714	0.0001	0

8.5.3 Checking model assumptions using graphs

8.11 Baby weights, Part V. Exercise 8.7 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.



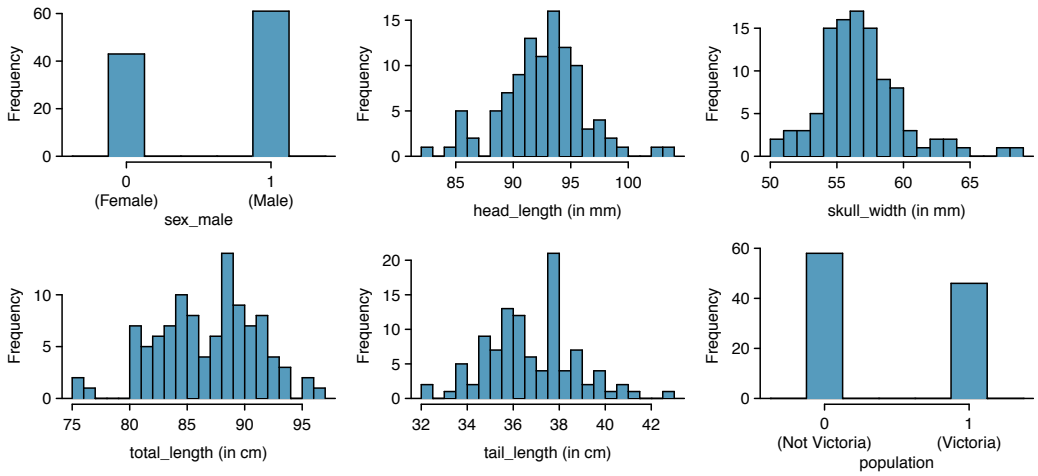
8.12 GPA and IQ. A regression model for predicting GPA from gender and IQ was fit, and both predictors were found to be statistically significant. Using the plots given below, determine if this regression model is appropriate for these data.



8.5.4 Logistic regression

8.13 Possum classification, Part I. The common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum (see Figure 7.5 on page 318). We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from the population. The first region is Victoria, which is in the eastern half of Australia and traverses the southern coast. The second region consists of New South Wales and Queensland, which make up eastern and northeastern Australia.

We use logistic regression to differentiate between possums in these two regions. The outcome variable, called `population`, takes value 1 when a possum is from Victoria and 0 when it is from New South Wales or Queensland. We consider five predictors: `sex_male` (an indicator for a possum being male), `head_length`, `skull_width`, `total_length`, and `tail_length`. Each variable is summarized in a histogram. The full logistic regression model and a reduced model after variable selection are summarized in the table.



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

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

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

8.14 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.

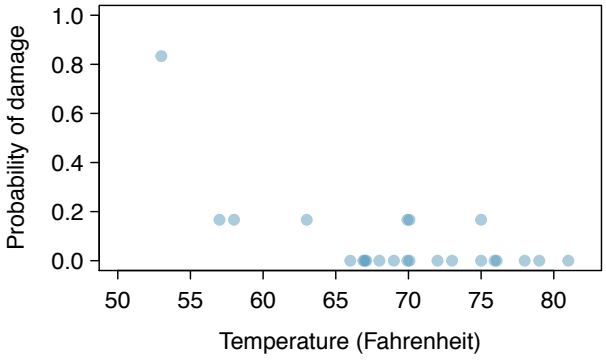
(d) Based on the model, do you think concerns regarding O-rings are justified? Explain.

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

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

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

8.16 Challenger disaster, Part II. Exercise 8.14 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.



(a) The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times Temperature$$

where $\hat{p}$ 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:

$$\hat{p}_{57} = 0.341 \qquad \hat{p}_{59} = 0.251 \qquad \hat{p}_{61} = 0.179 \qquad \hat{p}_{63} = 0.124$$
$$\hat{p}_{65} = 0.084 \qquad \hat{p}_{67} = 0.056 \qquad \hat{p}_{69} = 0.037 \qquad \hat{p}_{71} = 0.024$$

(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.