Chapter 9

Multiple and logistic regression

The principles of simple linear regression lay the foundation for more sophisticated regression models used in a wide range of challenging settings. In Chapter 9, we explore multiple regression, which introduces the possibility of more than one predictor in a linear model, and logistic regression, a technique for predicting categorical outcomes with two levels.

For videos, slides, and other resources, please visit
www.openintro.org/os

9.1 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 data about loans from the peer-to-peer lender, Lending Club, which is a data set we first encountered in Chapters 1 and 2. The loan data includes terms of the loan as well as information about the borrower. The outcome variable we would like to better understand is the interest rate assigned to the loan. For instance, all other characteristics held constant, does it matter how much debt someone already has? Does it matter if their income has been verified? Multiple regression will help us answer these and other questions.

The data set `loans` includes results from 10,000 loans, and we'll be looking at a subset of the available variables, some of which will be new from those we saw in earlier chapters. The first six observations in the data set are shown in Figure 9.1, and descriptions for each variable are shown in Figure 9.2. Notice that the past bankruptcy variable (`bankruptcy`) is an indicator variable, where it takes the value 1 if the borrower had a past bankruptcy in their record and 0 if not. Using an indicator variable in place of a category name allows for these variables to be directly used in regression. Two of the other variables are categorical (`income_ver` and `issued`), each of which can take one of a few different non-numerical values; we'll discuss how these are handled in the model in Section 9.1.1.

	interest_rate	income_ver	debt_to_income	credit_util	bankruptcy	term	issued	credit_checks
1	14.07	verified	18.01	0.55	0	60	Mar2018	6
2	12.61	not	5.04	0.15	1	36	Feb2018	1
3	17.09	source_only	21.15	0.66	0	36	Feb2018	4
4	6.72	not	10.16	0.20	0	36	Jan2018	0
5	14.07	verified	57.96	0.75	0	36	Mar2018	7
6	6.72	not	6.46	0.09	0	36	Jan2018	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 9.1: First six rows from the `loans` data set.

variable	description
`interest_rate`	Interest rate for the loan.
`income_ver`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels `verified`, `source_only`, and `not`.
`debt_to_income`	Debt-to-income ratio, which is the percentage of total debt of the borrower divided by their total income.
`credit_util`	Of all the credit available to the borrower, what fraction are they utilizing. For example, the credit utilization on a credit card would be the card's balance divided by the card's credit limit.
`bankruptcy`	An indicator variable for whether the borrower has a past bankruptcy in her record. This variable takes a value of `1` if the answer is "yes" and `0` if the answer is "no".
`term`	The length of the loan, in months.
`issued`	The month and year the loan was issued, which for these loans is always during the first quarter of 2018.
`credit_checks`	Number of credit checks in the last 12 months. For example, when filing an application for a credit card, it is common for the company receiving the application to run a credit check.

Figure 9.2: Variables and their descriptions for the `loans` data set.

9.1.1 Indicator and categorical variables as predictors

Let's start by fitting a linear regression model for interest rate with a single predictor indicating whether or not a person has a bankruptcy in their record:

$$\widehat{rate} = 12.33 + 0.74 \times bankruptcy$$

Results of this model are shown in Figure 9.3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3380	0.0533	231.49	<0.0001
bankruptcy	0.7368	0.1529	4.82	<0.0001
				$df = 9998$

Figure 9.3: Summary of a linear model for predicting interest rate based on whether the borrower has a bankruptcy in their record.

EXAMPLE 9.1

Interpret the coefficient for the past bankruptcy variable in the model. Is this coefficient significantly different from 0?

(E) The `bankruptcy` variable takes one of two values: 1 when the borrower has a bankruptcy in their history and 0 otherwise. A slope of 0.74 means that the model predicts a 0.74% higher interest rate for those borrowers with a bankruptcy in their record. (See Section 8.2.8 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Figure 9.3, we can see that the p-value for `bankruptcy` is very close to zero, indicating there is strong evidence the coefficient is different from zero when using this simple one-predictor model.

Suppose we had fit a model using a 3-level categorical variable, such as `income_ver`. The output from software is shown in Figure 9.4. This regression output provides multiple rows for the `income_ver` variable. Each row represents the relative difference for each level of `income_ver`. However, we are missing one of the levels: `not` (for *not verified*). The missing level is called the **reference level**, and it represents the default level that other levels are measured against.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.0995	0.0809	137.18	<0.0001
income_ver: *source_only*	1.4160	0.1107	12.79	<0.0001
income_ver: *verified*	3.2543	0.1297	25.09	<0.0001
				$df = 9998$

Figure 9.4: Summary of a linear model for predicting interest rate based on whether the borrower's income source and amount has been verified. This predictor has three levels, which results in 2 rows in the regression output.

EXAMPLE 9.2

How would we write an equation for this regression model?

(E) The equation for the regression model may be written as a model with two predictors:

$$\widehat{rate} = 11.10 + 1.42 \times \texttt{income\_ver}_{\texttt{source\_only}} + 3.25 \times \texttt{income\_ver}_{\texttt{verified}}$$

We use the notation $\texttt{variable}_{\texttt{level}}$ to represent indicator variables for when the categorical variable takes a particular value. For example, $\texttt{income\_ver}_{\texttt{source\_only}}$ would take a value of 1 if `income_ver` was `source_only` for a loan, and it would take a value of 0 otherwise. Likewise, $\texttt{income\_ver}_{\texttt{verified}}$ would take a value of 1 if `income_ver` took a value of `verified` and 0 if it took any other value.

The notation used in Example 9.2 may feel a bit confusing. Let's figure out how to use the equation for each level of the `income_ver` variable.

EXAMPLE 9.3

Using the model from Example 9.2, compute the average interest rate for borrowers whose income source and amount are both unverified.

When `income_ver` takes a value of `not`, then both indicator functions in the equation from Example 9.2 are set to zero:

E

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 0 + 3.25 \times 0 \\ &= 11.10\end{aligned}$$

The average interest rate for these borrowers is 11.1%. Because the `not` level does not have its own coefficient and it is the reference value, the indicators for the other levels for this variable all drop out.

EXAMPLE 9.4

Using the model from Example 9.2, compute the average interest rate for borrowers whose income source is verified but the amount is not.

When `income_ver` takes a value of `source_only`, then the corresponding variable takes a value of 1 while the other ($\texttt{income\_ver}_{\texttt{verified}}$) is 0:

E

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 1 + 3.25 \times 0 \\ &= 12.52\end{aligned}$$

The average interest rate for these borrowers is 12.52%.

GUIDED PRACTICE 9.5

G

Compute the average interest rate for borrowers whose income source and amount are both verified.[1]

PREDICTORS WITH SEVERAL CATEGORIES

When fitting a regression model with a categorical variable that has k levels where $k > 2$, software will provide a coefficient for $k - 1$ of those levels. For the last level that does not receive a coefficient, this is the **reference level**, and the coefficients listed for the other levels are all considered relative to this reference level.

[1] When `income_ver` takes a value of `verified`, then the corresponding variable takes a value of 1 while the other ($\texttt{income\_ver}_{\texttt{source\_only}}$) is 0:

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 0 + 3.25 \times 1 \\ &= 14.35\end{aligned}$$

The average interest rate for these borrowers is 14.35%.

GUIDED PRACTICE 9.6

Interpret the coefficients in the `income_ver` model.[2]

The higher interest rate for borrowers who have verified their income source or amount is surprising. Intuitively, we'd think that a loan would look *less* risky if the borrower's income has been verified. However, note that the situation may be more complex, and there may be confounding variables that we didn't account for. For example, perhaps lender require borrowers with poor credit to verify their income. That is, verifying income in our data set might be a signal of some concerns about the borrower rather than a reassurance that the borrower will pay back the loan. For this reason, the borrower could be deemed higher risk, resulting in a higher interest rate. (What other confounding variables might explain this counter-intuitive relationship suggested by the model?)

GUIDED PRACTICE 9.7

How much larger of an interest rate would we expect for a borrower who has verified their income source and amount vs a borrower whose income source has only been verified?[3]

9.1.2 Including and assessing many variables in a model

The world is complex, and it can be helpful to consider many factors at once in statistical modeling. For example, we might like to use the full context of borrower to predict the interest rate they receive rather than using a single variable. This is the strategy used in **multiple regression**. While we remain cautious about making any causal interpretations using multiple regression on observational data, such models are a common first step in gaining insights or providing some evidence of a causal connection.

We want to construct a model that accounts not only for any past bankruptcy or whether the borrower had their income source or amount verified, but simultaneously accounts for all the variables in the data set: `income_ver`, `debt_to_income`, `credit_util`, `bankruptcy`, `term`, `issued`, and `credit_checks`.

$$\begin{aligned}\widehat{\texttt{rate}} = \beta_0 &+ \beta_1 \times \texttt{income\_ver}_{\texttt{source\_only}} + \beta_2 \times \texttt{income\_ver}_{\texttt{verified}} + \beta_3 \times \texttt{debt\_to\_income} \\ &+ \beta_4 \times \texttt{credit\_util} + \beta_5 \times \texttt{bankruptcy} + \beta_6 \times \texttt{term} \\ &+ \beta_7 \times \texttt{issued}_{\texttt{Jan2018}} + \beta_8 \times \texttt{issued}_{\texttt{Mar2018}} + \beta_9 \times \texttt{credit\_checks}\end{aligned}$$

This equation represents a holistic approach for modeling all of the variables simultaneously. Notice that there are two coefficients for `income_ver` and also two coefficients for `issued`, since both are 3-level categorical variables.

We estimate the parameters β_0, β_1, β_2, ..., β_9 in the same way as we did in the case of a single predictor. We select b_0, b_1, b_2, ..., b_9 that minimize the sum of the squared residuals:

$$SSE = e_1^2 + e_2^2 + \cdots + e_{10000}^2 = \sum_{i=1}^{10000} e_i^2 = \sum_{i=1}^{10000} (y_i - \hat{y}_i)^2 \tag{9.8}$$

where y_i and $\hat{y}_i$ represent the observed interest rates and their estimated values according to the model, respectively. 10,000 residuals are calculated, one for each observation. We typically use a computer to minimize the sum of squares and compute point estimates, as shown in the sample output in Figure 9.5. Using this output, we identify the point estimates b_i of each β_i, just as we did in the one-predictor case.

[2] Each of the coefficients gives the incremental interest rate for the corresponding level relative to the `not` level, which is the reference level. For example, for a borrower whose income source and amount have been verified, the model predicts that they will have a 3.25% higher interest rate than a borrower who has not had their income source or amount verified.

[3] Relative to the `not` category, the `verified` category has an interest rate of 3.25% higher, while the `source_only` category is only 1.42% higher. Thus, `verified` borrowers will tend to get an interest rate about $3.25\% - 1.42\% = 1.83\%$ higher than `source_only` borrowers.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9251	0.2102	9.16	<0.0001
income_ver: *source_only*	0.9750	0.0991	9.83	<0.0001
income_ver: *verified*	2.5374	0.1172	21.65	<0.0001
debt_to_income	0.0211	0.0029	7.18	<0.0001
credit_util	4.8959	0.1619	30.24	<0.0001
bankruptcy	0.3864	0.1324	2.92	0.0035
term	0.1537	0.0039	38.96	<0.0001
issued: *Jan2018*	0.0276	0.1081	0.26	0.7981
issued: *Mar2018*	-0.0397	0.1065	-0.37	0.7093
credit_checks	0.2282	0.0182	12.51	<0.0001
				$df = 9990$

Figure 9.5: Output for the regression model, where `interest_rate` is the outcome and the variables listed 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 always estimate the β_i parameters using statistical software.

EXAMPLE 9.9

Write out the regression model using the point estimates from Figure 9.5. How many predictors are there in this model?

The fitted model for the interest rate is given by:

$$\widehat{\texttt{rate}} = 1.925 + 0.975 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.537 \times \texttt{income\_ver}_{\texttt{verified}} + 0.021 \times \texttt{debt\_to\_income} + 4.896 \times \texttt{credit\_util} + 0.386 \times \texttt{bankruptcy} + 0.154 \times \texttt{term} + 0.028 \times \texttt{issued}_{\texttt{Jan2018}} - 0.040 \times \texttt{issued}_{\texttt{Mar2018}} + 0.228 \times \texttt{credit\_checks}$$

If we count up the number of predictor coefficients, we get the *effective* number of predictors in the model: $k = 9$. Notice that the `issued` categorical predictor counts as two, once for the two levels shown in the model. In general, a categorical predictor with p different levels will be represented by $p - 1$ terms in a multiple regression model.

GUIDED PRACTICE 9.10

What does β_4, the coefficient of variable `credit_util`, represent? What is the point estimate of β_4?[4]

[4] β_4 represents the change in interest rate we would expect if someone's credit utilization was 0 and went to 1, all other factors held even. The point estimate is $b_4 = 4.90\%$.

EXAMPLE 9.11

Compute the residual of the first observation in Figure 9.1 on page 343 using the equation identified in Guided Practice 9.9.

(E) To compute the residual, we first need the predicted value, which we compute by plugging values into the equation from Example 9.9. For example, $\texttt{income\_ver}_{\texttt{source\_only}}$ takes a value of 0, $\texttt{income\_ver}_{\texttt{verified}}$ takes a value of 1 (since the borrower's income source and amount were verified), `debt_to_income` was 18.01, and so on. This leads to a prediction of $\widehat{rate}_1 = 18.09$. The observed interest rate was 14.07%, which leads to a residual of $e_1 = 14.07 - 18.09 = -4.02$.

EXAMPLE 9.12

We estimated a coefficient for `bankruptcy` in Section 9.1.1 of $b_4 = 0.74$ with a standard error of $SE_{b_1} = 0.15$ when using simple linear regression. Why is there a difference between that estimate and the estimated coefficient of 0.39 in the multiple regression setting?

(E) If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome `interest_rate` and predictor `bankruptcy` using simple linear regression, we were unable to control for other variables like whether the borrower had her income verified, the borrower's debt-to-income ratio, and other variables. That original model was constructed in a vacuum and did not consider the full context. When we include all of the variables, underlying and unintentional bias that was missed by these other variables is reduced or eliminated. Of course, bias can still exist from other confounding variables.

Example 9.12 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.

GUIDED PRACTICE 9.13

(G) The estimated value of the intercept is 1.925, 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: income source is not verified, the borrower has no debt (debt-to-income and credit utilization are zero), and so on. Is this reasonable? Is there any value gained by making this interpretation?[5]

[5]Many of the variables do take a value 0 for at least one data point, and for those variables, it is reasonable. However, one variable never takes a value of zero: `term`, which describes the length of the loan, in months. If `term` is set to zero, then the loan must be paid back immediately; the borrower must give the money back as soon as she receives it, which means it is not a real loan. Ultimately, the interpretation of the intercept in this setting is not insightful.

9.1.3 Adjusted R^2 as a better tool for multiple regression

We first used R^2 in Section 8.2 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 make it even more informative when comparing models.

GUIDED PRACTICE 9.14

G The variance of the residuals for the model given in Guided Practice 9.9 is 18.53, and the variance of the total price in all the auctions is 25.01. Calculate R^2 for this model.[6]

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 a biased estimate of the amount of variability explained by the model when applied to a new sample of data. To get a better estimate, we use the adjusted R^2.

ADJUSTED R^2 AS A TOOL FOR MODEL ASSESSMENT

The **adjusted R^2** is computed as

$$R^2_{adj} = 1 - \frac{s^2_{\text{residuals}}/(n-k-1)}{s^2_{\text{outcome}}/(n-1)} = 1 - \frac{s^2_{\text{residuals}}}{s^2_{\text{outcome}}} \times \frac{n-1}{n-k-1}$$

where n is the number of cases used to fit the model and k is the number of predictor variables in the model. Remember that a categorical predictor with p levels will contribute $p-1$ to the number of 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, which is equal to $n-k-1$ for the multiple regression context. If we were to make predictions for *new data* using our current model, we would find that the unadjusted R^2 would tend to be slightly overly optimistic, while the adjusted R^2 formula helps correct this bias.

GUIDED PRACTICE 9.15

G There were $n = 10000$ auctions in the `loans` data set and $k = 9$ predictor variables in the model. Use n, k, and the variances from Guided Practice 9.14 to calculate R^2_{adj} for the interest rate model.[7]

GUIDED PRACTICE 9.16

G 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?[8]

Adjusted R^2 could have been used in Chapter 8. However, when there is only $k = 1$ predictors, adjusted R^2 is very close to regular R^2, so this nuance isn't typically important when the model has only one predictor.

[6] $R^2 = 1 - \frac{18.53}{25.01} = 0.2591$.

[7] $R^2_{adj} = 1 - \frac{18.53}{25.01} \times \frac{10000-1}{1000-9-1} = 0.2584$. While the difference is very small, it will be important when we fine tune the model in the next section.

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

Exercises

9.1 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.[9]

	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 model.

(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?

9.2 Baby weights, Part II. Exercise 9.1 introduces a data set on birth weight of babies. Another variable we consider is `parity`, which is 1 if the child is the first born, and 0 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 model.

(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?

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

9.3 Baby weights, Part III. We considered the variables `smoke` and `parity`, one at a time, in modeling birth weights of babies in Exercises 9.1 and 9.2. 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 model 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 9.2. 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.

9.4 Absenteeism, Part I. 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).[10]

	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 model.

(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.

9.5 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

9.6 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.[11]

	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.

[10] 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.

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

9.2 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 variables from the model that are found to be less important. It's common (and hip, at least in the statistical world) to refer to models that have undergone such variable pruning as **parsimonious**.

In practice, the model that includes all available explanatory variables is often referred to as the **full model**. The full model may not be the best model, and if it isn't, we want to identify a smaller model that is preferable.

9.2.1 Identifying variables in the model that may not be helpful

Adjusted R^2 describes the strength of a model fit, and it is a useful tool for evaluating which predictors are adding value to the model, where *adding value* means they are (likely) improving the accuracy in predicting future outcomes.

Let's consider two models, which are shown in Tables 9.6 and 9.7. The first table summarizes the full model since it includes all predictors, while the second does not include the `issued` variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9251	0.2102	9.16	<0.0001
income_ver: *source_only*	0.9750	0.0991	9.83	<0.0001
income_ver: *verified*	2.5374	0.1172	21.65	<0.0001
debt_to_income	0.0211	0.0029	7.18	<0.0001
credit_util	4.8959	0.1619	30.24	<0.0001
bankruptcy	0.3864	0.1324	2.92	0.0035
term	0.1537	0.0039	38.96	<0.0001
issued: *Jan2018*	0.0276	0.1081	0.26	0.7981
issued: *Mar2018*	-0.0397	0.1065	-0.37	0.7093
credit_checks	0.2282	0.0182	12.51	<0.0001
$R^2_{adj} = 0.25843$				$df = 9990$

Figure 9.6: The fit for the full regression model, including the adjusted R^2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9213	0.1982	9.69	<0.0001
income_ver: *source_only*	0.9740	0.0991	9.83	<0.0001
income_ver: *verified*	2.5355	0.1172	21.64	<0.0001
debt_to_income	0.0211	0.0029	7.19	<0.0001
credit_util	4.8958	0.1619	30.25	<0.0001
bankruptcy	0.3869	0.1324	2.92	0.0035
term	0.1537	0.0039	38.97	<0.0001
credit_checks	0.2283	0.0182	12.51	<0.0001
$R^2_{adj} = 0.25854$				$df = 9992$

Figure 9.7: The fit for the regression model after dropping the `issued` variable.

EXAMPLE 9.17

Which of the two models is better?

E

We compare the adjusted R^2 of each model to determine which to choose. Since the first model has an R^2_{adj} smaller than the R^2_{adj} of the second model, we prefer the second model to the first.

Will the model without `issued` be better than the model with `issued`? We cannot know for sure, but based on the adjusted R^2, this is our best assessment.

9.2.2 Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called *backward elimination* and *forward selection*. These techniques are often referred to as **stepwise** model selection strategies, because they add or delete one variable at a time as they "step" through the candidate predictors.

Backward elimination starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time from the model until we cannot improve the adjusted R^2. The strategy within each elimination step is to eliminate the variable that leads to the largest improvement in adjusted R^2.

EXAMPLE 9.18

Results corresponding to the *full model* for the `loans` data are shown in Figure 9.6. How should we proceed under the backward elimination strategy?

Our baseline adjusted R^2 from the full model is $R^2_{adj} = 0.25843$, and we need to determine whether dropping a predictor will improve the adjusted R^2. To check, we fit models that each drop a different predictor, and we record the adjusted R^2:

Exclude ...				
	`income_ver` $R^2_{adj} = 0.22380$	`debt_to_income` $R^2_{adj} = 0.25468$	`credit_util` $R^2_{adj} = 0.19063$	`bankruptcy` $R^2_{adj} = 0.25787$
	`term` $R^2_{adj} = 0.14581$	`issued` $R^2_{adj} = 0.25854$	`credit_checks` $R^2_{adj} = 0.24689$	

The model without `issued` has the highest adjusted R^2 of 0.25854, higher than the adjusted R^2 for the full model. Because eliminating `issued` leads to a model with a higher adjusted R^2, we drop `issued` from the model.

(E) Since we eliminated a predictor from the model in the first step, we see whether we should eliminate any additional predictors. Our baseline adjusted R^2 is now $R^2_{adj} = 0.25854$. We now fit new models, which consider eliminating each of the remaining predictors in addition to `issued`:

Exclude `issued` and ...			
	`income_ver` $R^2_{adj} = 0.22395$	`debt_to_income` $R^2_{adj} = 0.25479$	`credit_util` $R^2_{adj} = 0.19074$
	`bankruptcy` $R^2_{adj} = 0.25798$	`term` $R^2_{adj} = 0.14592$	`credit_checks` $R^2_{adj} = 0.24701$

None of these models lead to an improvement in adjusted R^2, so we do not eliminate any of the remaining predictors. That is, after backward elimination, we are left with the model that keeps all predictors except `issued`, which we can summarize using the coefficients from Figure 9.7:

$$\begin{aligned}\widehat{rate} = \; & 1.921 + 0.974 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.535 \times \texttt{income\_ver}_{\texttt{verified}} \\ & + 0.021 \times \texttt{debt\_to\_income} + 4.896 \times \texttt{credit\_util} + 0.387 \times \texttt{bankruptcy} \\ & + 0.154 \times \texttt{term} + 0.228 \times \texttt{credit\_check}\end{aligned}$$

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 improve the model (as measured by adjusted R^2).

EXAMPLE 9.19

Construct a model for the `loans` 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 `income_ver`, then the model including just `debt_to_income`, then a model with just `credit_util`, and so on. Then we examine the adjusted R^2 for each of these models:

Add ...				
	`income_ver` $R^2_{adj} = 0.05926$	`debt_to_income` $R^2_{adj} = 0.01946$	`credit_util` $R^2_{adj} = 0.06452$	`bankruptcy` $R^2_{adj} = 0.00222$
	`term` $R^2_{adj} = 0.12855$	`issued` $R^2_{adj} = -¡0.00018$	`credit_checks` $R^2_{adj} = 0.01711$	

In this first step, we compare the adjusted R^2 against a baseline model that has no predictors. The no-predictors model always has $R^2_{adj} = 0$. The model with one predictor that has the largest adjusted R^2 is the model with the `term` predictor, and because this adjusted R^2 is larger than the adjusted R^2 from the model with no predictors ($R^2_{adj} = 0$), we will add this variable to our model.

We repeat the process again, this time considering 2-predictor models where one of the predictors is `term` and with a new baseline of $R^2_{adj} = 0.12855$:

Add `term` and ...			
	`income_ver` $R^2_{adj} = 0.16851$	`debt_to_income` $R^2_{adj} = 0.14368$	`credit_util` $R^2_{adj} = 0.20046$
	`bankruptcy` $R^2_{adj} = 0.13070$	`issued` $R^2_{adj} = 0.12840$	`credit_checks` $R^2_{adj} = 0.14294$

The best second predictor, `credit_util`, has a higher adjusted R^2 (0.20046) than the baseline (0.12855), so we also add `credit_util` to the model.

Since we have again added a variable to the model, we continue and see whether it would be beneficial to add a third variable:

Add `term`, `credit_util`, and ...			
	`income_ver` $R^2_{adj} = 0.24183$	`debt_to_income` $R^2_{adj} = 0.20810$	
	`bankruptcy` $R^2_{adj} = 0.20169$	`issued` $R^2_{adj} = 0.20031$	`credit_checks` $R^2_{adj} = 0.21629$

The model adding `income_ver` improved adjusted R^2 (0.24183 to 0.20046), so we add `income_ver` to the model.

We continue on in this way, next adding `debt_to_income`, then `credit_checks`, and `bankruptcy`. At this point, we come again to the `issued` variable: adding this variable leads to $R^2_{adj} = 0.25843$, while keeping all the other variables but excluding `issued` leads to a higher $R^2_{adj} = 0.25854$. This means we do not add `issued`. In this example, we have arrived at the same model that we identified from backward elimination.

MODEL SELECTION STRATEGIES

Backward elimination begins with the model having the largest number of predictors and eliminates variables one-by-one until we are satisfied that all remaining variables are important to the model. Forward selection starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found.

Backward elimination and forward selection sometimes arrive at different final models. If trying both techniques and this happens, it's common to choose the model with the larger R^2_{adj}.

9.2.3 The p-value approach, an alternative to adjusted R^2

The p-value may be used as an alternative to R^2_{adj} for model selection:

Backward elimination with the p-value approach. In backward elimination, we would identify the predictor corresponding to the largest p-value. If the p-value is above the significance level, usually $\alpha = 0.05$, then we would drop that variable, refit the model, and repeat the process. If the largest p-value is less than $\alpha = 0.05$, then we would not eliminate any predictors and the current model would be our best-fitting model.

Forward selection with the p-value approach. In forward selection with p-values, we reverse the process. We begin with a model that has no predictors, then we fit a model for each possible predictor, identifying the model where the corresponding predictor's p-value is smallest. If that p-value is smaller than $\alpha = 0.05$, we add it to the model and repeat the process, considering whether to add more variables one-at-a-time. When none of the remaining predictors can be added to the model and have a p-value less than 0.05, then we stop adding variables and the current model would be our best-fitting model.

GUIDED PRACTICE 9.20

Examine Figure 9.7 on page 353, which considers the model including all variables except the variable for the month the loan was issued. If we were using the p-value approach with backward elimination and we were considering this model, which of these variables would be up for elimination? Would we drop that variable, or would we keep it in the model?[12]

While the adjusted R^2 and p-value approaches are similar, they sometimes lead to different models, with the R^2_{adj} approach tending to include more predictors in the final model.

ADJUSTED R^2 VS P-VALUE APPROACH

When the sole goal is to improve prediction accuracy, use R^2_{adj}. This is commonly the case in machine learning applications.

When we care about understanding which variables are statistically significant predictors of the response, or if there is interest in producing a simpler model at the potential cost of a little prediction accuracy, then the p-value approach is preferred.

Regardless of whether you use R^2_{adj} or the p-value approach, or if you use the backward elimination of forward selection strategy, our job is not done after variable selection. We must still verify the model conditions are reasonable.

[12] The `bankruptcy` predictor is up for elimination since it has the largest p-value. However, since that p-value is smaller than 0.05, we would still keep it in the model.

Exercises

9.7 Baby weights, Part IV. Exercise 9.3 considers a model that predicts a newborn's weight using several predictors (gestation length, parity, age of mother, height of mother, weight of mother, smoking status of mother). The table below shows the adjusted R-squared for the full model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

	Model	Adjusted R^2
1	Full model	0.2541
2	No gestation	0.1031
3	No parity	0.2492
4	No age	0.2547
5	No height	0.2311
6	No weight	0.2536
7	No smoking status	0.2072

Which, if any, variable should be removed from the model first?

9.8 Absenteeism, Part II. Exercise 9.4 considers a model that predicts the number of days absent using three predictors: ethnic background (`eth`), gender (`sex`), and learner status (`lrn`). The table below shows the adjusted R-squared for the model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

	Model	Adjusted R^2
1	Full model	0.0701
2	No ethnicity	-0.0033
3	No sex	0.0676
4	No learner status	0.0723

Which, if any, variable should be removed from the model first?

9.9 Baby weights, Part V. Exercise 9.3 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

9.10 Absenteeism, Part III. Exercise 9.4 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

9.11 Movie lovers, Part I. Suppose a social scientist is interested in studying what makes audiences love or hate a movie. She collects a random sample of movies (genre, length, cast, director, budget, etc.) as well as a measure of the success of the movie (score on a film review aggregator website). If as part of her research she is interested in finding out which variables are significant predictors of movie success, what type of model selection method should she use?

9.12 Movie lovers, Part II. Suppose an online media streaming company is interested in building a movie recommendation system. The website maintains data on the movies in their database (genre, length, cast, director, budget, etc.) and additionally collects data from their subscribers (demographic information, previously watched movies, how they rated previously watched movies, etc.). The recommendation system will be deemed successful if subscribers actually watch, and rate highly, the movies recommended to them. Should the company use the adjusted R^2 or the p-value approach in selecting variables for their recommendation system?

9.3 Checking model conditions 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 conditions:

1. the residuals of the model are nearly normal (less important for larger data sets),
2. the variability of the residuals is nearly constant,
3. the residuals are independent, and
4. each variable is linearly related to the outcome.

9.3.1 Diagnostic plots

Diagnostic plots can be used to check each of these conditions. We will consider the model from the Lending Club loans data, and check whether there are any notable concerns:

$$\begin{aligned}\widehat{rate} = \ & 1.921 + 0.974 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.535 \times \texttt{income\_ver}_{\texttt{verified}} \\ & + 0.021 \times \texttt{debt\_to\_income} + 4.896 \times \texttt{credit\_util} + 0.387 \times \texttt{bankruptcy} \\ & + 0.154 \times \texttt{term} + 0.228 \times \texttt{credit\_check}\end{aligned}$$

Check for outliers. In theory, the distribution of the residuals should be nearly normal; in practice, normality can be relaxed for most applications. Instead, we examine a histogram of the residuals to check if there are any outliers: Figure 9.8 is a histogram of these outliers. Since this is a very large data set, only particularly extreme observations would be a concern in this particular case. There are no extreme observations that might cause a concern.

If we intended to construct what are called **prediction intervals** for future observations, we would be more strict and require the residuals to be nearly normal. Prediction intervals are further discussed in an online extra on the OpenIntro website:

www.openintro.org/d?id=stat_extra_linear_regression_supp

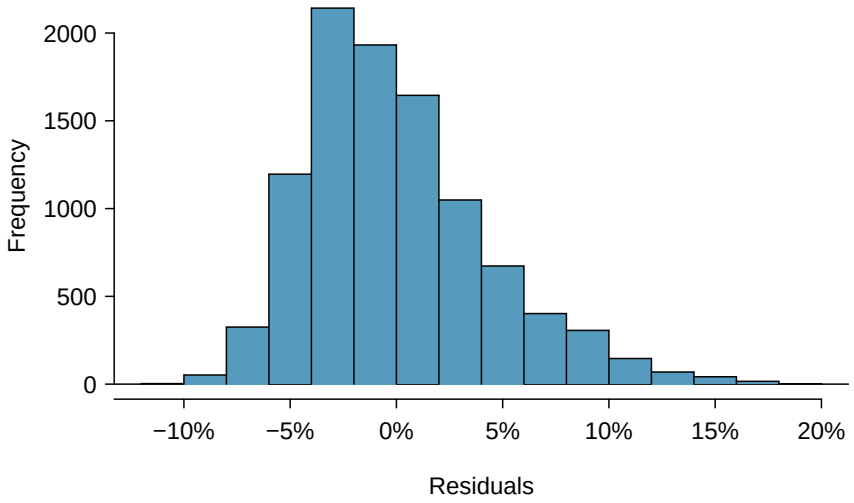


Figure 9.8: A histogram of the residuals.

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 9.9. This plot is helpful to check the condition that the variance of the residuals is approximately constant, and a smoothed line has been added to represent the approximate trend in this plot. There is more evident variability for fitted values that are larger, which we'll discuss further.

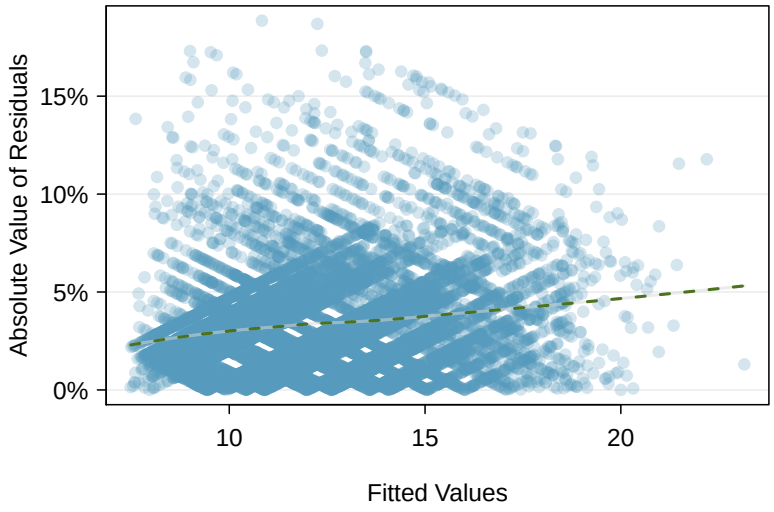


Figure 9.9: Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.

Residuals in order of their data collection. This type of plot can be helpful when observations were collected in a sequence. Such a plot is helpful in identifying any connection between cases that are close to one another. The loans in this data set were issued over a 3 month period, and the month the loan was issued was not found to be important, suggesting this is not a concern for this data set. In cases where a data set does show some pattern for this check, **time series** methods may be useful.

Residuals against each predictor variable. We consider a plot of the residuals against each of the predictors in Figure 9.10. For those instances where there are only 2-3 groups, box plots are shown. For the numerical outcomes, a smoothed line has been fit to the data to make it easier to review. Ultimately, we are looking for any notable change in variability between groups or pattern in the data.

Here are the things of importance from these plots:

- There is some minor differences in variability between the verified income groups.
- There is a very clear pattern for the debt-to-income variable. What also stands out is that this variable is very strongly right skewed: there are few observations with very high debt-to-income ratios.
- The downward curve on the right side of the credit utilization and credit check plots suggests some minor misfitting for those larger values.

Having reviewed the diagnostic plots, there are two options. The first option is to, if we're not concerned about the issues observed, use this as the final model; if going this route, it is important to still note any abnormalities observed in the diagnostics. The second option is to try to improve the model, which is what we'll try to do with this particular model fit.

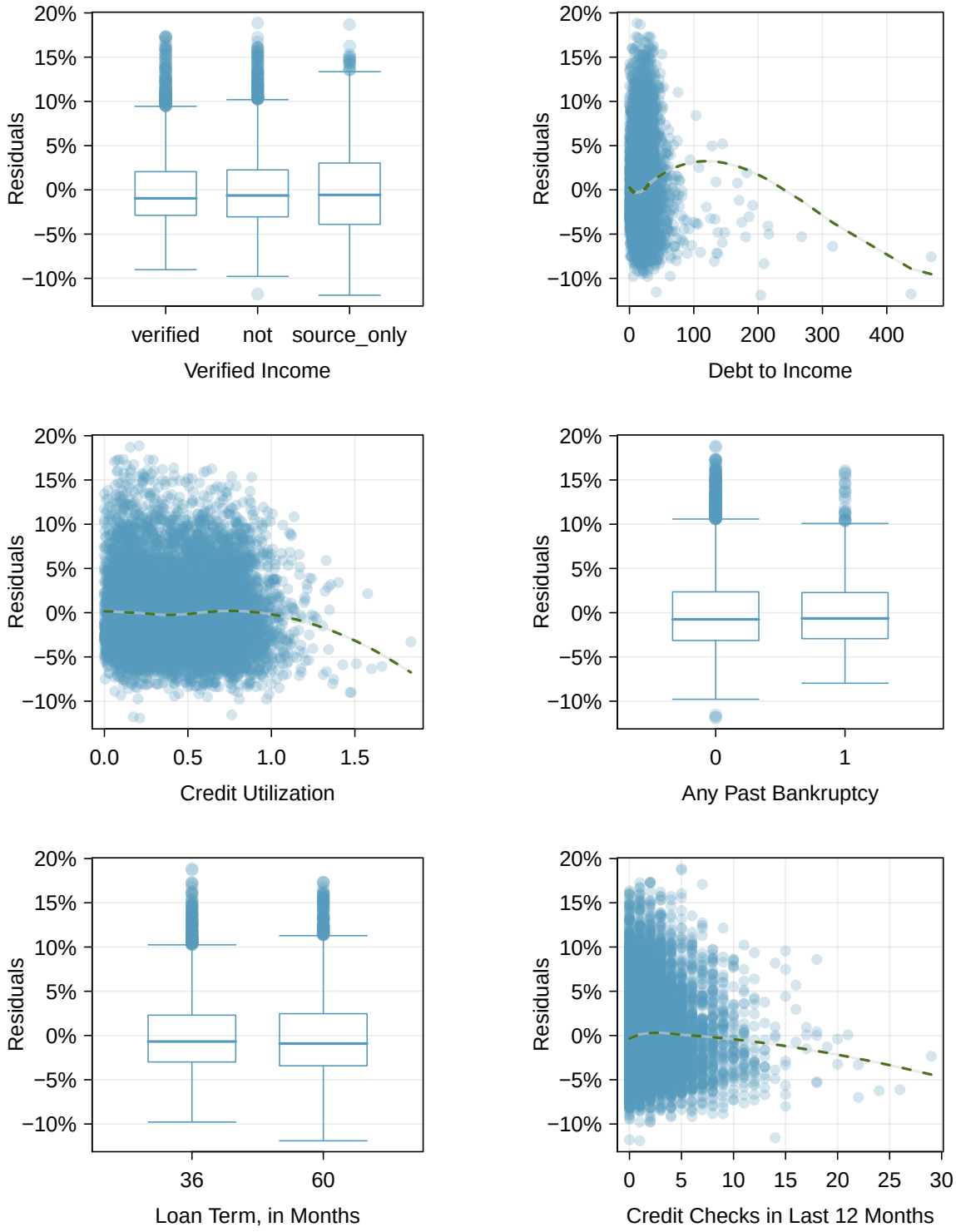


Figure 9.10: Diagnostic plots for residuals against each of the predictors. For the box plots, we're looking for notable differences in variability. For numerical predictors, we also check for trends or other structure in the data.

9.3.2 Options for improving the model fit

There are several options for improvement of a model, including transforming variables, seeking out additional variables to fill model gaps, or using more advanced methods that would account for challenges around inconsistent variability or nonlinear relationships between predictors and the outcome.

The main concern for the initial model is that there is a notable nonlinear relationship between the debt-to-income variable observed in Figure 9.10. To resolve this issue, we're going to consider a couple strategies for adjusting the relationship between the predictor variable and the outcome.

Let's start by taking a look at a histogram of `debt_to_income` in Figure 9.11. The variable is extremely skewed, and upper values will have a lot of leverage on the fit. Below are several options:

- log transformation ($\log x$),
- square root transformation ($\sqrt{x}$),
- inverse transformation ($1/x$),
- truncation (cap the max value possible)

If we inspected the data more closely, we'd observe some instances where the variable takes a value of 0, and since $\log(0)$ and $1/x$ are undefined when $x = 0$, we'll exclude these transformations from further consideration.[13] A square root transformation is valid for all values the variable takes, and truncating some of the larger observations is also a valid approach. We'll consider both of these approaches.

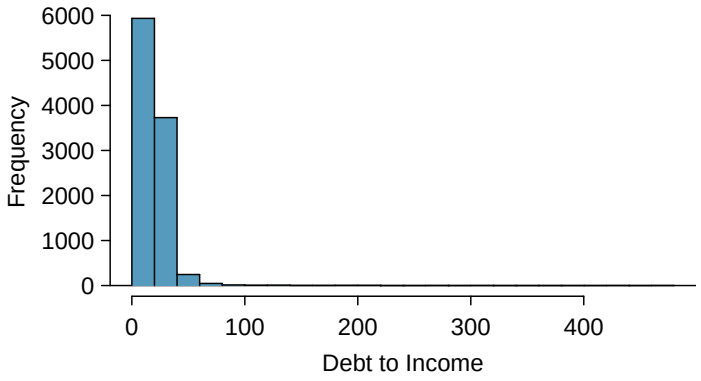


Figure 9.11: Histogram of `debt_to_income`, where extreme skew is evident.

To try transforming the variable, we make two new variables representing the transformed versions:

Square root. We create a new variable, `sqrt_debt_to_income`, where all the values are simply the square roots of the values in `debt_to_income`, and then refit the model as before. The result is shown in the left panel of Figure 9.12. The square root pulled in the higher values a bit, but the fit still doesn't look great since the smoothed line is still wavy.

Truncate at 50. We create a new variable, `debt_to_income_50`, where any values in `debt_to_income` that are greater than 50 are shrunk to exactly 50. Refitting the model once more, the diagnostic plot for this new variable is shown in the right panel of Figure 9.12. Here the fit looks much more reasonable, so this appears to be a reasonable approach.

The downside of using transformations is that it reduces the ease of interpreting the results. Fortunately, since the truncation transformation only affects a relatively small number of cases, the interpretation isn't dramatically impacted.

[13] There are ways to make them work, but we'll leave those options to a later course.

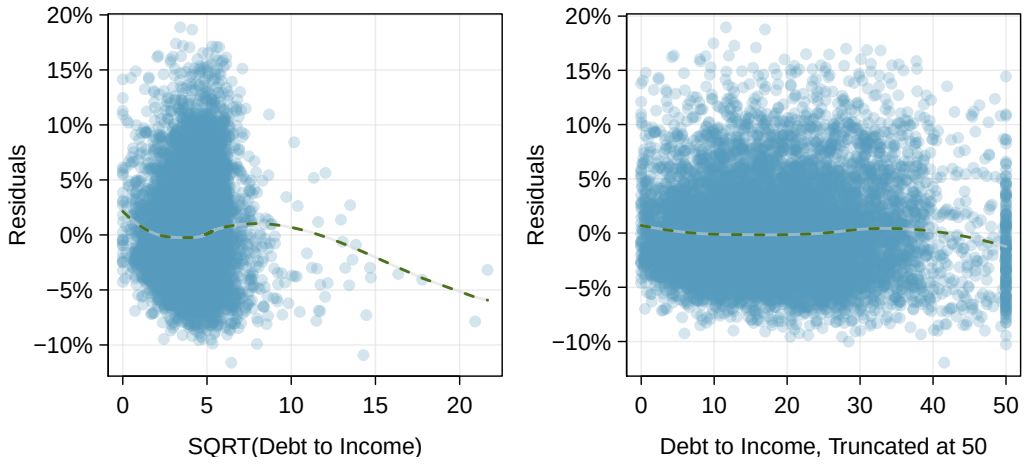


Figure 9.12: Histogram of `debt_to_income`, where extreme skew is evident.

As a next step, we'd evaluate the new model using the truncated version of `debt_to_income`, we would complete all the same procedures as before. The other two issues noted while inspecting diagnostics in Section 9.3.1 are still present in the updated model. If we choose to report this model, we would want to also discuss these shortcomings to be transparent in our work. Depending on what the model will be used for, we could either try to bring those under control, or we could stop since those issues aren't severe. Had the non-constant variance been a little more dramatic, it would be a higher priority. Ultimately we decided that the model was reasonable, and we report its final form here:

$$\begin{aligned}\widehat{rate} = \; & 1.562 + 1.002 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.436 \times \texttt{income\_ver}_{\texttt{verified}} \\ & + 0.048 \times \texttt{debt\_to\_income\_50} + 4.694 \times \texttt{credit\_util} + 0.394 \times \texttt{bankruptcy} \\ & + 0.153 \times \texttt{term} + 0.223 \times \texttt{credit\_check}\end{aligned}$$

A sharp eye would notice that the coefficient for `debt_to_income_50` is more than twice as large as what the coefficient had been for the `debt_to_income` variable in the earlier model. This suggests those larger values not only were points with high leverage, but they were influential points that were dramatically impacting the coefficient.

"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.

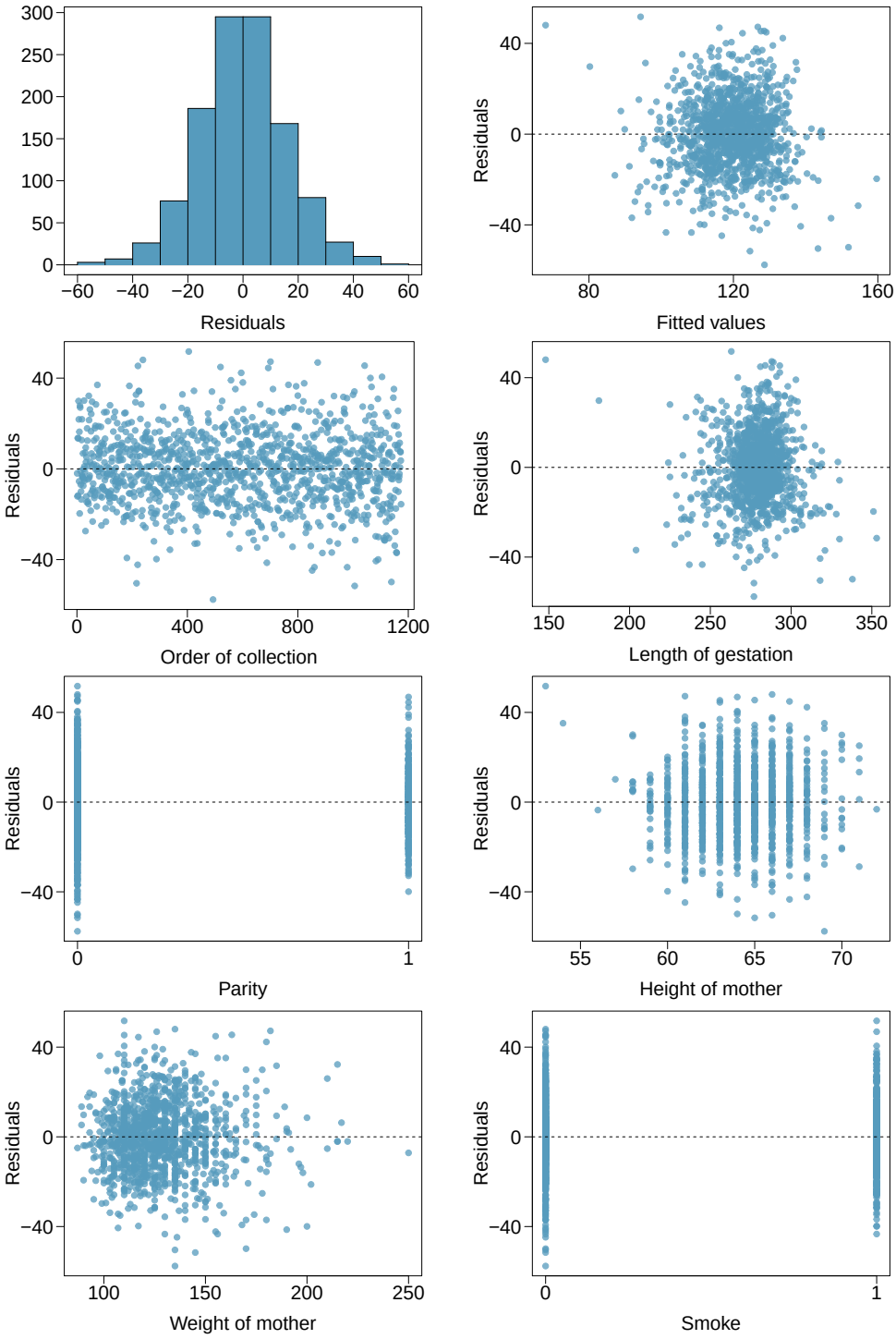
Don't report results when conditions are grossly violated. While there is a little leeway in model conditions, don't go too far. If model conditions are very clearly violated, consider a new model, even if it means learning more statistical methods or hiring someone who can help. To help you get started, we've developed a couple additional sections that you may find on OpenIntro's website. These sections provide a light introduction to what are called **interaction terms** and to fitting **nonlinear curves** to data, respectively:

www.openintro.org/d?file=stat_extra_interaction_effects

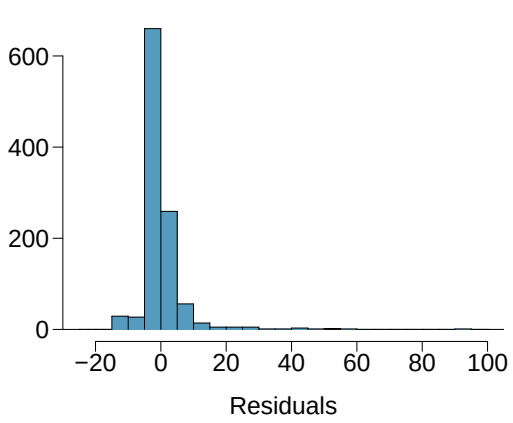
www.openintro.org/d?file=stat_extra_nonlinear_relationships

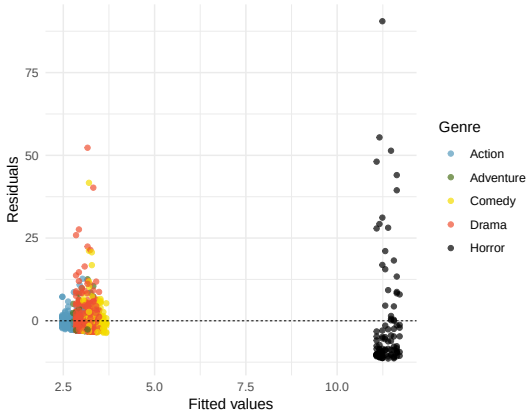
Exercises

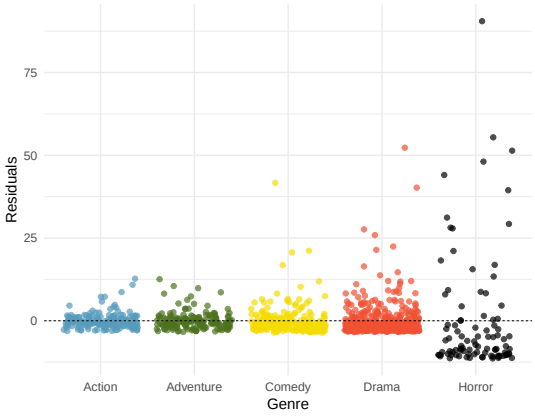
9.13 Baby weights, Part VI. Exercise 9.3 presents a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoking status of the mother. Determine if the model assumptions are met using the plots below. If not, describe how to proceed with the analysis.

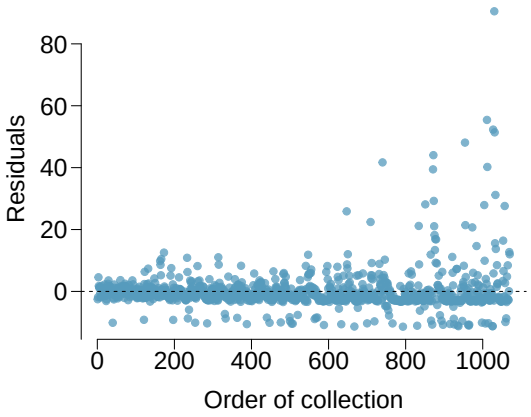


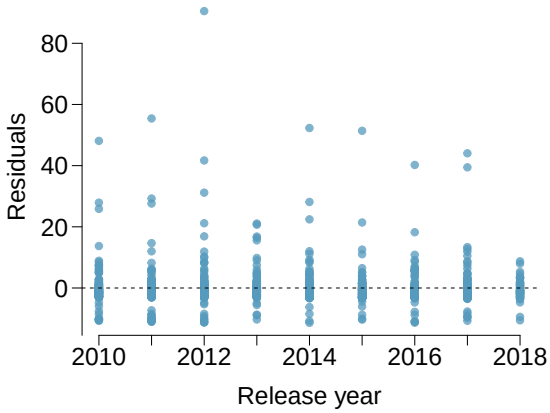
9.14 Movie returns, Part I. A FiveThirtyEight.com article reports that "Horror movies get nowhere near as much draw at the box office as the big-time summer blockbusters or action/adventure movies ... but there's a huge incentive for studios to continue pushing them out. The return-on-investment potential for horror movies is absurd." To investigate how the return-on-investment compares between genres and how this relationship has changed over time, an introductory statistics student fit a model predicting the ratio of gross revenue of movies from genre and release year for 1,070 movies released between 2000 and 2018. Using the plots given below, determine if this regression model is appropriate for these data.[14]











[14] FiveThirtyEight, Scary Movies Are The Best Investment In Hollywood.

9.4 Multiple regression case study: Mario Kart

We'll 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.

9.4.1 Data set and the full model

The `mariokart` data set includes results from 141 auctions. Four observations from this data set are shown in Figure 9.13, and descriptions for each variable are shown in Figure 9.14. Notice that the condition and stock photo variables are indicator variables, similar to `bankruptcy` in the `loan` data set.

	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

Figure 9.13: Four observations from the `mariokart` data set.

variable	description
`price`	Final auction price plus shipping costs, in US dollars.
`cond_new`	Indicator variable for if the game is new (`1`) or used (`0`).
`stock_photo`	Indicator variable for if the auction's main photo is a stock photo.
`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 an optional steering wheel accessory that holds the Wii controller.

Figure 9.14: Variables and their descriptions for the `mariokart` data set.

GUIDED PRACTICE 9.21

We fit a linear regression model with the game's condition as a predictor of auction price. Results of this model are summarized below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.8711	0.8140	52.67	<0.0001
cond_new	10.8996	1.2583	8.66	<0.0001
				$df = 139$

Write down the equation for the model, note whether the slope is statistically different from zero, and interpret the coefficient.[15]

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.

We want to construct a model that accounts for not only the game condition, as in Guided Practice 9.21, but simultaneously accounts for three other variables:

$$\widehat{\texttt{price}} = \beta_0 + \beta_1 \times \texttt{cond\_new} + \beta_2 \times \texttt{stock\_photo} + \beta_3 \times \texttt{duration} + \beta_4 \times \texttt{wheels}$$

Figure 9.15 summarizes the full model. Using this output, we identify the point estimates of each coefficient.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.2110	1.5140	23.92	<0.0001
cond_new	5.1306	1.0511	4.88	<0.0001
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	<0.0001
				$df = 136$

Figure 9.15: Output for the regression model where `price` is the outcome and `cond_new`, `stock_photo`, `duration`, and `wheels` are the predictors.

GUIDED PRACTICE 9.22

Write out the model's equation using the point estimates from Figure 9.15. How many predictors are there in this model?[16]

GUIDED PRACTICE 9.23

What does β_4, the coefficient of variable x_4 (Wii wheels), represent? What is the point estimate of β_4?[17]

[15] The equation for the line may be written as

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

Examining the regression output in Guided Practice 9.21, 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.

The `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. This means the 10.90 model coefficient predicts an extra $10.90 for those games that are new versus those that are used.

[16] $\widehat{price} = 36.21 + 5.13 \times \texttt{cond\_new} + 1.08 \times \texttt{stock\_photo} - 0.03 \times \texttt{duration} + 7.29 \times \texttt{wheels}$, with the $k = 4$ predictors.

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

GUIDED PRACTICE 9.24

Compute the residual of the first observation in Figure 9.13 using the equation identified in Guided Practice 9.22.[18]

EXAMPLE 9.25

We estimated a coefficient for `cond_new` in Section 9.21 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 there is collinearity among some predictors. 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).

9.4.2 Model selection

Let's revisit the model for the Mario Kart auction and complete model selection using backwards selection. Recall that the full model took the following form:

$$\widehat{price} = 36.21 + 5.13 \times \texttt{cond\_new} + 1.08 \times \texttt{stock\_photo} - 0.03 \times \texttt{duration} + 7.29 \times \texttt{wheels}$$

EXAMPLE 9.26

Results corresponding to the full model for the `mariokart` data were shown in Figure 9.15 on the facing page. For this model, we consider what would happen if dropping each of the variables in the model:

Exclude ...	`cond_new`	`stock_photo`	`duration`	`wheels`
	$R^2_{adj} = 0.6626$	$R^2_{adj} = 0.7107$	$R^2_{adj} = 0.7128$	$R^2_{adj} = 0.3487$

For the full model, $R^2_{adj} = 0.7108$. How should we proceed under the backward elimination strategy?

The third model without `duration` has the highest R^2_{adj} of 0.7128, so we compare it to R^2_{adj} for the full model. Because eliminating `duration` leads to a model with a higher R^2_{adj}, we drop `duration` from the model.

GUIDED PRACTICE 9.27

In Example 9.26, we eliminated the `duration` variable, which resulted in a model with $R^2_{adj} = 0.7128$. Let's look at if we would eliminate another variable from the model using backwards elimination:

Exclude `duration` and ...	`cond_new`	`stock_photo`	`wheels`
	$R^2_{adj} = 0.6587$	$R^2_{adj} = 0.7124$	$R^2_{adj} = 0.3414$

Should we eliminate any additional variable, and if so, which variable should we eliminate?[19]

[18] $e_i = y_i - \hat{y}_i = 51.55 - 49.62 = 1.93$, where 49.62 was computed using the variables values from the observation and the equation identified in Guided Practice 9.22.

[19] Removing any of the three remaining variables would lead to a decrease in R^2_{adj}, so we should not remove any additional variables from the model after we removed `duration`.

GUIDED PRACTICE 9.28

After eliminating the auction's duration from the model, we are left with the following reduced model:

$$\widehat{price} = 36.05 + 5.18 \times \texttt{cond\_new} + 1.12 \times \texttt{stock\_photo} + 7.30 \times \texttt{wheels}$$

How much would you predict for the total price for the Mario Kart game if it was used, used a stock photo, and included two wheels and put up for auction during the time period that the Mario Kart data were collected?[20]

GUIDED PRACTICE 9.29

Would you be surprised if the seller from Guided Practice 9.28 didn't get the exact price predicted?[21]

9.4.3 Checking model conditions using graphs

Let's take a closer look at the diagnostics for the Mario Kart model to check if the model we have identified is reasonable.

Check for outliers. A histogram of the residuals is shown in Figure 9.16. With a data set well over a hundred, we're primarily looking for major outliers. While one minor outlier appears on the upper end, it is not a concern for this large of a data set.

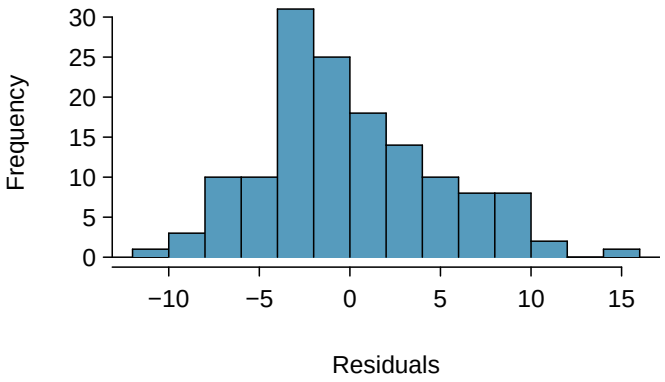


Figure 9.16: Histogram of the residuals. No clear outliers are evident.

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 9.17. 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 9.18. Here we see no structure that indicates a problem.

Residuals against each predictor variable. We consider a plot of the residuals against the `cond_new` variable, the residuals against the `stock_photo` variable, and the residuals against the `wheels` variable. These plots are shown in Figure 9.19. 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, which it does not. However, looking at the stock

[20] We would plug in `0` for `cond_new` `1` for `stock_photo`, and `2` for `wheels` into the equation, which would return \$51.77, which is the total price we would expect for the auction.

[21] No. The model provides the *average* auction price we would expect, and the price for one auction to the next will continue to vary a bit (but less than what our prediction would be without the model).

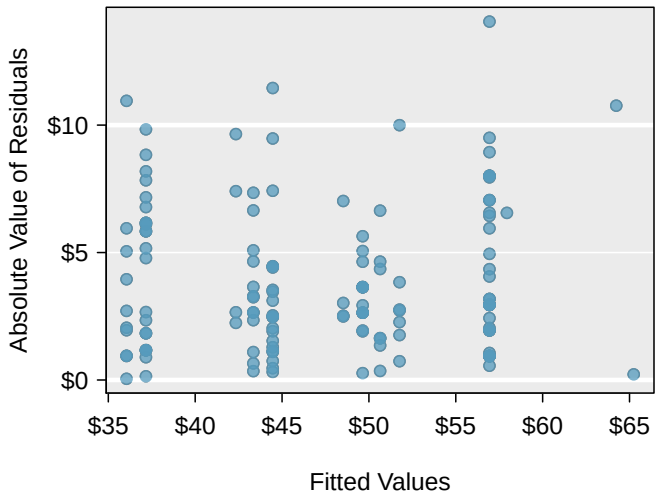


Figure 9.17: Absolute value of the residuals against the fitted values. No patterns are evident.

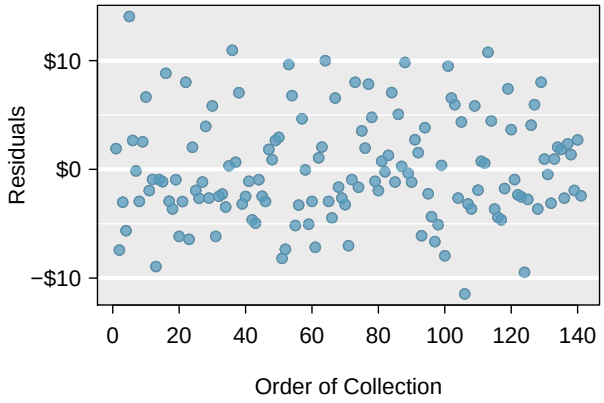


Figure 9.18: Residuals in the order that their corresponding observations were collected. There are no evident patterns.

photo variable, we find that there is some difference in the variability of the residuals in the two groups. Additionally, 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.

As with the `loans` analysis, we would summarize diagnostics when reporting the model results. In the case of this auction data, we would report that there appears to be non-constant variance in the stock photo variable and 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 who may review the analysis, and omitting this information could be a setback to the very people who the model might assist.

Note: there are no exercises for this section.

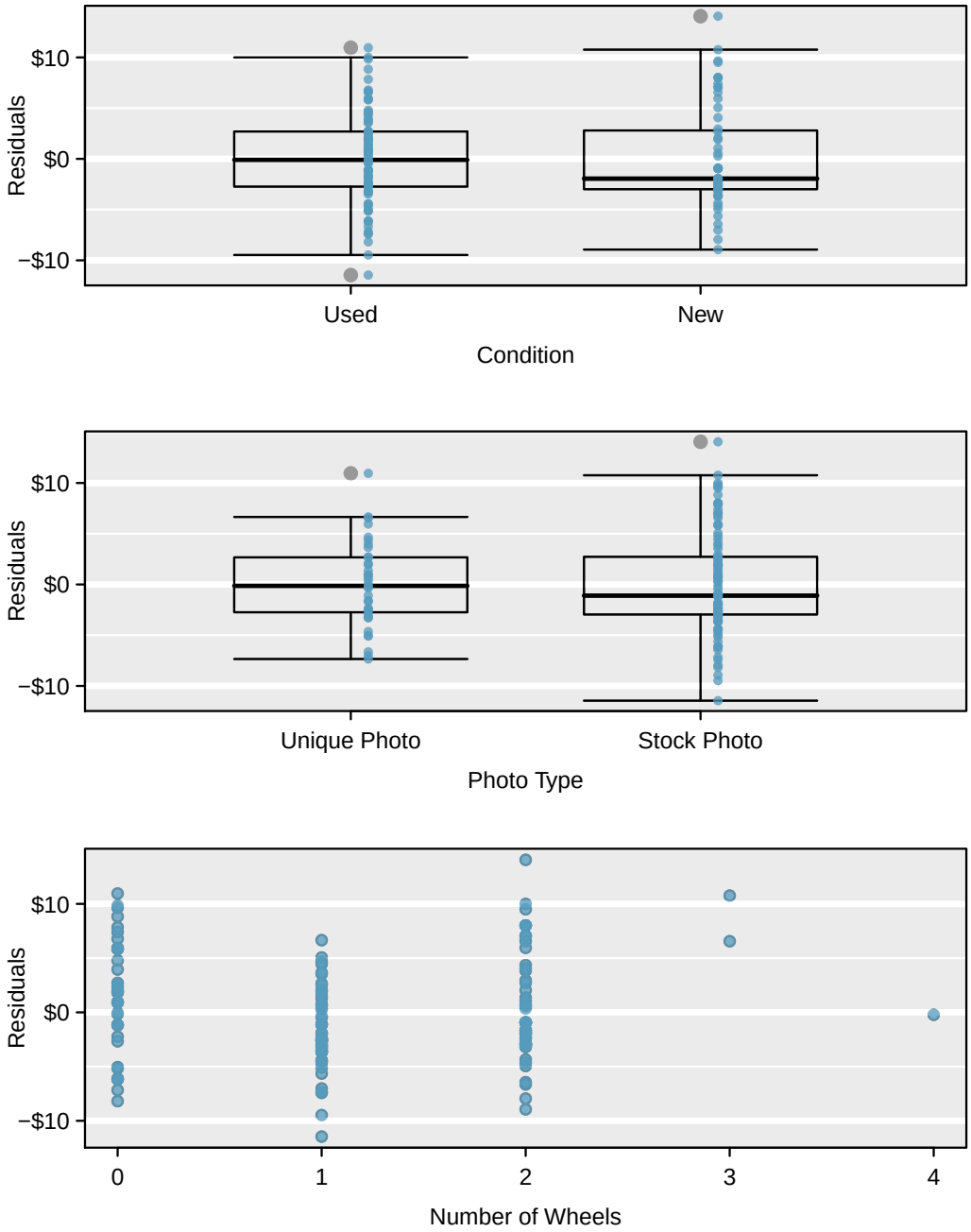


Figure 9.19: For the condition and stock photo variables, we check for differences in the distribution shape or variability of the residuals. In the case of the stock photos variable, we see a little less variability in the unique photo group than the stock photo group. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the `wheels` variable in the bottom plot.

9.5 Introduction to 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, e.g. yes and no. 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. Ultimately, the application of a GLM will feel very similar to multiple regression, even if some of the details are different.

9.5.1 Resume data

We will consider experiment data from a study that sought to understand the effect of race and sex on job application callback rates; details of the study and a link to the data set may be found in Appendix B.9. To evaluate which factors were important, job postings were identified in Boston and Chicago for the study, and researchers created many fake resumes to send off to these jobs to see which would elicit a callback. The researchers enumerated important characteristics, such as years of experience and education details, and they used these characteristics to randomly generate the resumes. Finally, they randomly assigned a name to each resume, where the name would imply the applicant's sex and race.

The first names that were used and randomly assigned in this experiment were selected so that they would predominantly be recognized as belonging to Black or White individuals; other races were not considered in this study. While no name would definitively be inferred as pertaining to a Black individual or to a White individual, the researchers conducted a survey to check for racial association of the names; names that did not pass this survey check were excluded from usage in the experiment. You can find the full set of names that did pass the survey test and were ultimately used in the study in Figure 9.20. For example, Lakisha was a name that their survey indicated would be interpreted as a Black woman, while Greg was a name that would generally be interpreted to be associated with a White male.

first_name	race	sex
Aisha	black	female
Allison	white	female
Anne	white	female
Brad	white	male
Brendan	white	male
Brett	white	male
Carrie	white	female
Darnell	black	male
Ebony	black	female
Emily	white	female
Geoffrey	white	male
Greg	white	male

first_name	race	sex
Hakim	black	male
Jamal	black	male
Jay	white	male
Jermaine	black	male
Jill	white	female
Kareem	black	male
Keisha	black	female
Kenya	black	female
Kristen	white	female
Lakisha	black	female
Latonya	black	female
Latoya	black	female

first_name	race	sex
Laurie	white	female
Leroy	black	male
Matthew	white	male
Meredith	white	female
Neil	white	male
Rasheed	black	male
Sarah	white	female
Tamika	black	female
Tanisha	black	female
Todd	white	male
Tremayne	black	male
Tyrone	black	male

Figure 9.20: List of all 36 unique names along with the commonly inferred race and sex associated with these names.

The response variable of interest is whether or not there was a callback from the employer for the applicant, and there were 8 attributes that were randomly assigned that we'll consider, with special interest in the race and sex variables. Race and sex are **protected classes** in the United States, meaning they are not legally permitted factors for hiring or employment decisions. The full set of attributes considered is provided in Figure 9.21.

variable	description
`callback`	Specifies whether the employer called the applicant following submission of the application for the job.
`job_city`	City where the job was located: Boston or Chicago.
`college_degree`	An indicator for whether the resume listed a college degree.
`years_experience`	Number of years of experience listed on the resume.
`honors`	Indicator for the resume listing some sort of honors, e.g. employee of the month.
`military`	Indicator for if the resume listed any military experience.
`email_address`	Indicator for if the resume listed an email address for the applicant.
`race`	Race of the applicant, implied by their first name listed on the resume.
`sex`	Sex of the applicant (limited to only `male` and `female` in this study), implied by the first name listed on the resume.

Figure 9.21: Descriptions for the `callback` variable along with 8 other variables in the `resume` data set. Many of the variables are indicator variables, meaning they take the value 1 if the specified characteristic is present and 0 otherwise.

All of the attributes listed on each resume were randomly assigned. This means that no attributes that might be favorable or detrimental to employment would favor one demographic over another on these resumes. Importantly, due to the experimental nature of this study, we can infer causation between these variables and the callback rate, if the variable is statistically significant. Our analysis will allow us to compare the practical importance of each of the variables relative to each other.

9.5.2 Modeling the probability of an event

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 callback for the resume) with probability p_i and the value 0 with probability $1 - p_i$. Because each observation has a slightly different context, e.g. different education level or a different number of years of experience, the probability p_i will differ for each observation. Ultimately, it is this probability that we model in relation to the predictor variables: we will examine which resume characteristics correspond to higher or lower callback rates.

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 resume application, Y_i will be used to represent whether resume i received a callback ($Y_i = 1$) 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.

The logistic regression model relates the probability a resume would receive a callback (p_i) to the predictors $x_{1,i}$, $x_{2,i}$, ..., $x_{k,i}$ through a framework much like that of multiple regression:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} \tag{9.30}$$

We want to choose a transformation in the equation that makes practical and mathematical sense. For example, we want a transformation that makes the range of possibilities on the left hand side of the equation 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 9.22. Below, we rewrite the equation relating Y_i to its

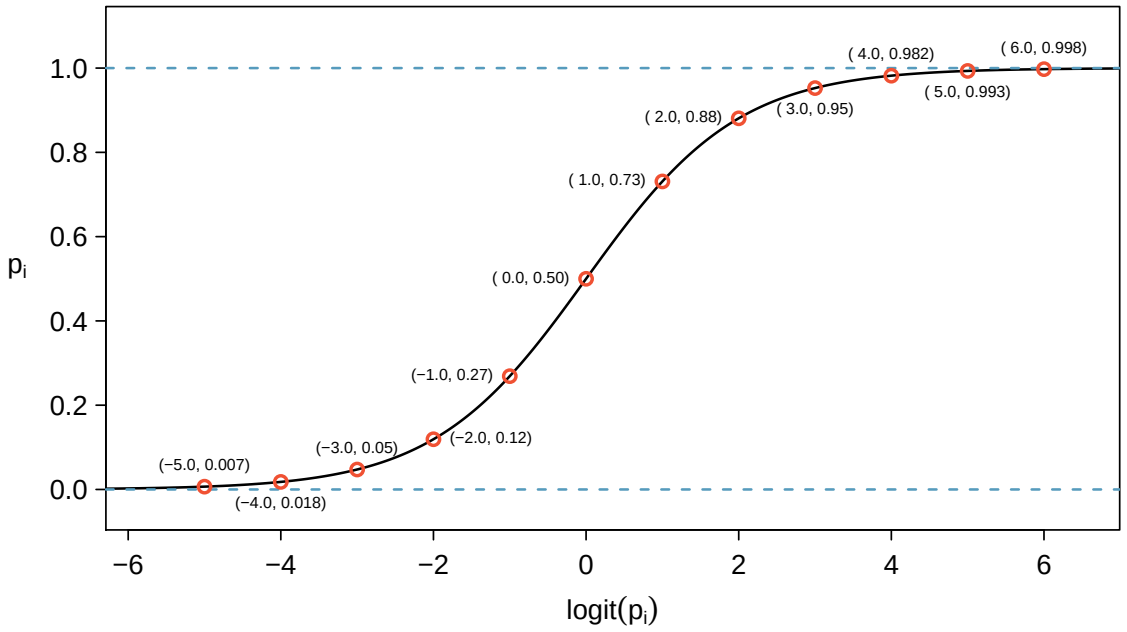


Figure 9.22: Values of p_i against values of $logit(p_i)$.

predictors 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 resume example, there are 8 predictor variables, so $k = 8$. While the precise choice of a logit function isn't intuitive, it is based on theory that underpins generalized linear models, which is beyond the scope of this book. Fortunately, once we fit a model using software, it will start to feel like we're back in the multiple regression context, even if the interpretation of the coefficients is more complex.

EXAMPLE 9.31

We start by fitting a model with a single predictor: `honors`. This variable indicates whether the applicant had any type of honors listed on their resume, such as employee of the month. The following logistic regression model was fit using statistical software:

$$\log_e\left(\frac{p_i}{1-p_i}\right) = -2.4998 + 0.8668 \times \texttt{honors}$$

(a) If a resume is randomly selected from the study and it does not have any honors listed, what is the probability resulted in a callback?

(E) (b) What would the probability be if the resume did list some honors?

(a) If a randomly chosen resume from those sent out is considered, and it does not list honors, then `honors` takes value 0 and the right side of the model equation equals -2.4998. Solving for p_i: $\frac{e^{-2.4998}}{1+e^{-2.4998}} = 0.076$. Just as we labeled a fitted value of y_i with a "hat" in single-variable and multiple regression, we do the same for this probability: $\hat{p}_i = 0.076$.

(b) If the resume had listed some honors, then the right side of the model equation is $-2.4998 + 0.8668 \times 1 = -1.6330$, which corresponds to a probability $\hat{p}_i = 0.163$.

Notice that we could examine -2.4998 and -1.6330 in Figure 9.22 to estimate the probability before formally calculating the value.

To convert from values on the logistic regression scale (e.g. -2.4998 and -1.6330 in Example 9.31), 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 can make use of this formula. In Example 9.31, the probabilities were calculated as

$$\frac{e^{-2.4998}}{1 + e^{-2.4998}} = 0.076 \qquad \frac{e^{-2.4998+0.8668}}{1 + e^{-2.4998+0.8668}} = 0.163$$

While knowing whether a resume listed honors provides some signal when predicting whether or not the employer would call, we would like to account for many different variables at once to understand how each of the different resume characteristics affected the chance of a callback.

9.5.3 Building the logistic model with many variables

We used statistical software to fit the logistic regression model with all 8 predictors described in Figure 9.21. Like multiple regression, the result may be presented in a summary table, which is shown in Figure 9.23. 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.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.6632	0.1820	-14.64	<0.0001
job_city: *Chicago*	-0.4403	0.1142	-3.85	0.0001
college_degree	-0.0666	0.1211	-0.55	0.5821
years_experience	0.0200	0.0102	1.96	0.0503
honors	0.7694	0.1858	4.14	<0.0001
military	-0.3422	0.2157	-1.59	0.1127
email_address	0.2183	0.1133	1.93	0.0541
race: *white*	0.4424	0.1080	4.10	<0.0001
sex: *male*	-0.1818	0.1376	-1.32	0.1863

Figure 9.23: Summary table for the full logistic regression model for the resume callback example.

Just like multiple regression, we could trim some variables from the model. Here we'll use a statistic called **Akaike information criterion (AIC)**, which is an analog to how we used adjusted R-squared in multiple regression, and we look for models with a lower AIC through a backward elimination strategy. After using this criteria, the `college_degree` variable is eliminated, giving the smaller model summarized in Figure 9.24, which is what we'll rely on for the remainder of this section.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7162	0.1551	-17.51	<0.0001
job_city: *Chicago*	-0.4364	0.1141	-3.83	0.0001
years_experience	0.0206	0.0102	2.02	0.0430
honors	0.7634	0.1852	4.12	<0.0001
military	-0.3443	0.2157	-1.60	0.1105
email_address	0.2221	0.1130	1.97	0.0494
race: *white*	0.4429	0.1080	4.10	<0.0001
sex: *male*	-0.1959	0.1352	-1.45	0.1473

Figure 9.24: Summary table for the logistic regression model for the resume callback example, where variable selection has been performed using AIC.

EXAMPLE 9.32

The `race` variable had taken only two levels: `black` and `white`. Based on the model results, was race a meaningful factor for if a prospective employer would call back?

(E) We see that the p-value for this coefficient is very small (very nearly zero), which implies that race played a statistically significant role in whether a candidate received a callback. Additionally, we see that the coefficient shown corresponds to the level of `white`, and it is positive. This positive coefficient reflects a positive gain in callback rate for resumes where the candidate's first name implied they were White. The data provide very strong evidence of racism by prospective employers that favors resumes where the first name is typically interpreted to be White.

The coefficient of $\texttt{race}_{\texttt{white}}$ in the full model in Figure 9.23, is nearly identical to the model shown in Figure 9.24. The predictors in this experiment were thoughtfully laid out so that the coefficient estimates would typically not be much influenced by which other predictors were in the model, which aligned with the motivation of the study to tease out which effects were important to getting a callback. In most observational data, it's common for point estimates to change a little, and sometimes a lot, depending on which other variables are included in the model.

EXAMPLE 9.33

Use the model summarized in Figure 9.24 to estimate the probability of receiving a callback for a job in Chicago where the candidate lists 14 years experience, no honors, no military experience, includes an email address, and has a first name that implies they are a White male.

We can start by writing out the equation using the coefficients from the model, then we can add in the corresponding values of each variable for this individual:

(E)
$$
\begin{aligned}
&log_e\left(\frac{p}{1-p}\right) \\
&\quad = -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} + 0.0206 \times \texttt{years\_experience} + 0.7634 \times \texttt{honors} \\
&\qquad - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} + 0.4429 \times \texttt{race}_{\texttt{white}} - 0.1959 \times \texttt{sex}_{\texttt{male}} \\
&\quad = -2.7162 - 0.4364 \times 1 + 0.0206 \times 14 + 0.7634 \times 0 \\
&\qquad - 0.3443 \times 0 + 0.2221 \times 1 + 0.4429 \times 1 - 0.1959 \times 1 \\
&\quad = -2.3955
\end{aligned}
$$

We can now back-solve for p: the chance such an individual will receive a callback is about 8.35%.

EXAMPLE 9.34

Compute the probability of a callback for an individual with a name commonly inferred to be from a Black male but who otherwise has the same characteristics as the one described in Example 9.33.

(E) We can complete the same steps for an individual with the same characteristics who is Black, where the only difference in the calculation is that the indicator variable $\texttt{race}_{\texttt{white}}$ will take a value of `0`. Doing so yields a probability of 0.0553. Let's compare the results with those of Example 9.33.

In practical terms, an individual perceived as White based on their first name would need to apply to $\frac{1}{0.0835} \approx 12$ jobs on average to receive a callback, while an individual perceived as Black based on their first name would need to apply to $\frac{1}{0.0553} \approx 18$ jobs on average to receive a callback. That is, applicants who are perceived as Black need to apply to 50% more employers to receive a callback than someone who is perceived as White based on their first name for jobs like those in the study.

What we've quantified in this section is alarming and disturbing. However, one aspect that makes this racism so difficult to address is that the experiment, as well-designed as it is, cannot send us much signal about which employers are discriminating. It is only possible to say that discrimination is happening, even if we cannot say which particular callbacks – or non-callbacks – represent discrimination. Finding strong evidence of racism for individual cases is a persistent challenge in enforcing anti-discrimination laws.

9.5.4 Diagnostics for the callback rate model

LOGISTIC REGRESSION CONDITIONS

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

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

The first logistic regression model condition – independence of the outcomes – is reasonable for the experiment since characteristics of resumes were randomly assigned to the resumes that were sent out.

The second condition of the logistic regression model is not easily checked without a fairly sizable amount of data. Luckily, we have 4870 resume submissions in the data set! Let's first visualize these data by plotting the true classification of the resumes against the model's fitted probabilities, as shown in Figure 9.25.

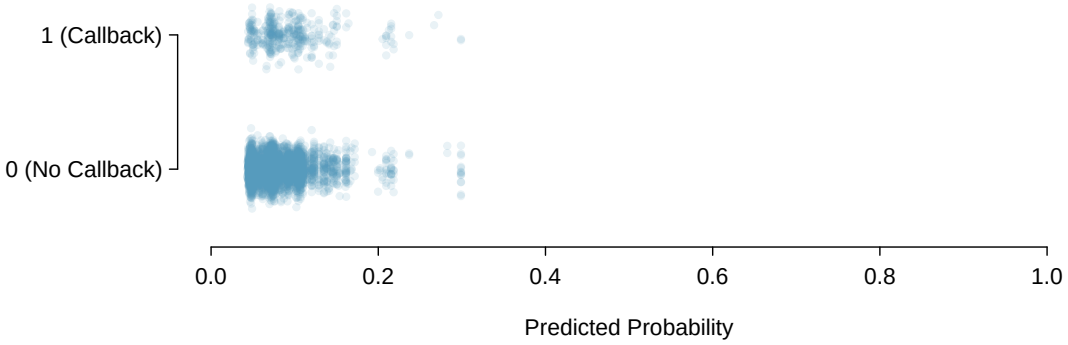


Figure 9.25: The predicted probability that each of the 4870 resumes results in a callback. Noise (small, random vertical shifts) have been added to each point so points with nearly identical values aren't plotted exactly on top of one another.

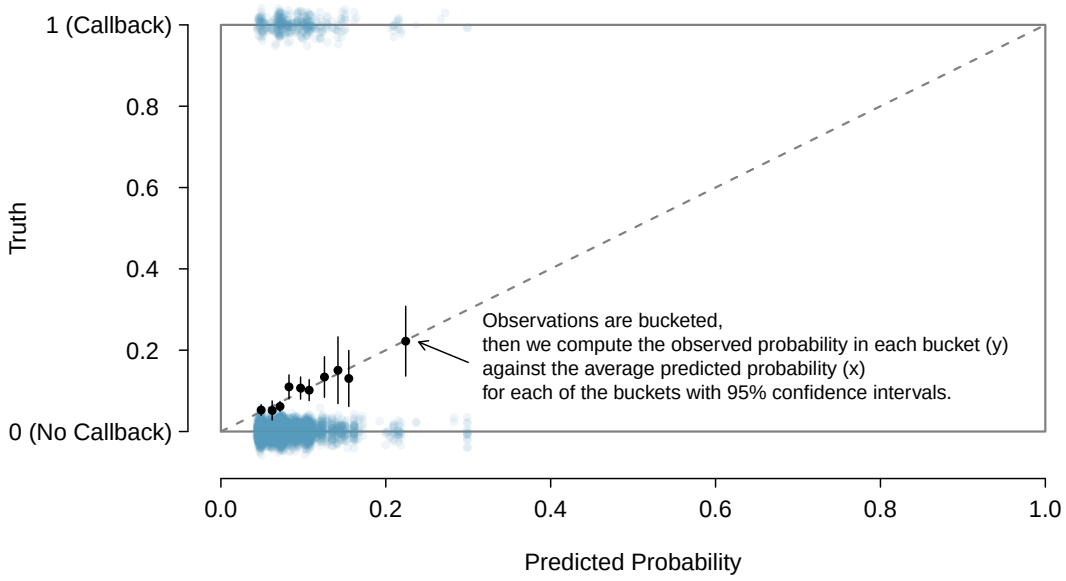


Figure 9.26: The dashed line is within the confidence bound of the 95% confidence intervals of each of the buckets, suggesting the logistic fit is reasonable.

We'd like to assess the quality of the model. For example, we might ask: if we look at resumes that we modeled as having a 10% chance of getting a callback, do we find about 10% of them actually receive a callback? We can check this for groups of the data by constructing a plot as follows:

1. Bucket the data into groups based on their predicted probabilities.
2. Compute the average predicted probability for each group.
3. Compute the observed probability for each group, along with a 95% confidence interval.
4. Plot the observed probabilities (with 95% confidence intervals) against the average predicted probabilities for each group.

The points plotted should fall close to the line $y = x$, since the predicted probabilities should be similar to the observed probabilities. We can use the confidence intervals to roughly gauge whether anything might be amiss. Such a plot is shown in Figure 9.26.

Additional diagnostics may be created that are similar to those featured in Section 9.3. For instance, we could compute residuals as the observed outcome minus the expected outcome ($e_i = Y_i - \hat{p}_i$), and then we could create plots of these residuals against each predictor. We might also create a plot like that in Figure 9.26 to better understand the deviations.

9.5.5 Exploring discrimination between groups of different sizes

Any form of discrimination is concerning, and this is why we decided it was so important to discuss this topic using data. The resume study also only examined discrimination in a single aspect: whether a prospective employer would call a candidate who submitted their resume. There was a 50% higher barrier for resumes simply when the candidate had a first name that was perceived to be from a Black individual. It's unlikely that discrimination would stop there.

EXAMPLE 9.35

Let's consider a sex-imbalanced company that consists of 20% women and 80% men,[22] and we'll suppose that the company is very large, consisting of perhaps 20,000 employees. Suppose when someone goes up for promotion at this company, 5 of their colleagues are randomly chosen to provide feedback on their work.

Now let's imagine that 10% of the people in the company are prejudiced against the other sex. That is, 10% of men are prejudiced against women, and similarly, 10% of women are prejudiced against men.

Who is discriminated against more at the company, men or women?

(E)

Let's suppose we took 100 men who have gone up for promotion in the past few years. For these men, $5 \times 100 = 500$ random colleagues will be tapped for their feedback, of which about 20% will be women (100 women). Of these 100 women, 10 are expected to be biased against the man they are reviewing. Then, of the 500 colleagues reviewing them, men will experience discrimination by about 2% of their colleagues when they go up for promotion.

Let's do a similar calculation for 100 women who have gone up for promotion in the last few years. They will also have 500 random colleagues providing feedback, of which about 400 (80%) will be men. Of these 400 men, about 40 (10%) hold a bias against women. Of the 500 colleagues providing feedback on the promotion packet for these women, 8% of the colleagues hold a bias against the women.

Example 9.35 highlights something profound: even in a hypothetical setting where each demographic has the same degree of prejudice against the other demographic, the smaller group experiences the negative effects more frequently. Additionally, if we would complete a handful of examples like the one above with different numbers, we'd learn that the greater the imbalance in the population groups, the more the smaller group is disproportionately impacted.[23]

Of course, there are other considerable real-world omissions from the hypothetical example. For example, studies have found instances where people from an oppressed group also discriminate against others within their own oppressed group. As another example, there are also instances where a majority group can be oppressed, with apartheid in South Africa being one such historic example. Ultimately, discrimination is complex, and there are many factors at play beyond the mathematics property we observed in Example 9.35.

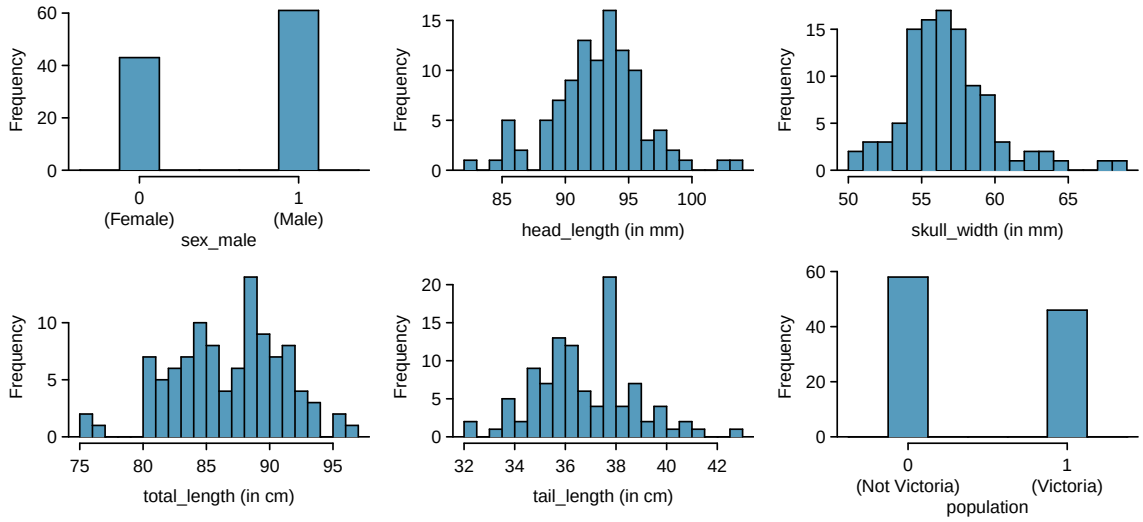
We close this book on this serious topic, and we hope it inspires you to think about the power of reasoning with data. Whether it is with a formal statistical model or by using critical thinking skills to structure a problem, we hope the ideas you have learned will help you do more and do better in life.

[22] A more thoughtful example would include non-binary individuals.

[23] If a proportion p of a company are women and the rest of the company consists of men, then under the hypothetical situation the ratio of rates of discrimination against women vs men would be given by $\frac{1-p}{p}$; this ratio is always greater than 1 when $p < 0.5$.

Exercises

9.15 Possum classification, Part I. The common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum (see Figure 8.4 on page 307). 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.

9.16 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.

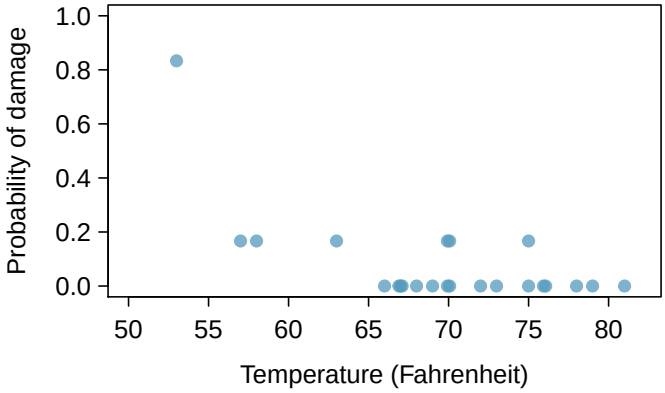
9.17 Possum classification, Part II. A logistic regression model was proposed for classifying common brushtail possums into their two regions in Exercise 9.15. The outcome variable took value 1 if the possum was from Victoria and 0 otherwise.

	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	33.5095	9.9053	3.38	0.0007
sex_male	-1.4207	0.6457	-2.20	0.0278
skull_width	-0.2787	0.1226	-2.27	0.0231
total_length	0.5687	0.1322	4.30	0.0000
tail_length	-1.8057	0.3599	-5.02	0.0000

(a) Write out the form of the model. Also identify which of the variables are positively associated when controlling for other variables.

(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?

9.18 Challenger disaster, Part II. Exercise 9.16 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.

Chapter exercises

9.19 Multiple regression fact checking. Determine which of the following statements are true and false. For each statement that is false, explain why it is false.

(a) If predictors are collinear, then removing one variable will have no influence on the point estimate of another variable's coefficient.

(b) Suppose a numerical variable x has a coefficient of $b_1 = 2.5$ in the multiple regression model. Suppose also that the first observation has $x_1 = 7.2$, the second observation has a value of $x_1 = 8.2$, and these two observations have the same values for all other predictors. Then the predicted value of the second observation will be 2.5 higher than the prediction of the first observation based on the multiple regression model.

(c) If a regression model's first variable has a coefficient of $b_1 = 5.7$, then if we are able to influence the data so that an observation will have its x_1 be 1 larger than it would otherwise, the value y_1 for this observation would increase by 5.7.

(d) Suppose we fit a multiple regression model based on a data set of 472 observations. We also notice that the distribution of the residuals includes some skew but does not include any particularly extreme outliers. Because the residuals are not nearly normal, we should not use this model and require more advanced methods to model these data.

9.20 Logistic regression fact checking. Determine which of the following statements are true and false. For each statement that is false, explain why it is false.

(a) Suppose we consider the first two observations based on a logistic regression model, where the first variable in observation 1 takes a value of $x_1 = 6$ and observation 2 has $x_1 = 4$. Suppose we realized we made an error for these two observations, and the first observation was actually $x_1 = 7$ (instead of 6) and the second observation actually had $x_1 = 5$ (instead of 4). Then the predicted probability from the logistic regression model would increase the same amount for each observation after we correct these variables.

(b) When using a logistic regression model, it is impossible for the model to predict a probability that is negative or a probability that is greater than 1.

(c) Because logistic regression predicts probabilities of outcomes, observations used to build a logistic regression model need not be independent.

(d) When fitting logistic regression, we typically complete model selection using adjusted R^2.

9.21 Spam filtering, Part I. Spam filters are built on principles similar to those used in logistic regression. We fit a probability that each message is spam or not spam. We have several email variables for this problem: `to_multiple`, `cc`, `attach`, `dollar`, `winner`, `inherit`, `password`, `format`, `re_subj`, `exclaim_subj`, and `sent_email`. We won't describe what each variable means here for the sake of brevity, but each is either a numerical or indicator variable.

(a) For variable selection, we fit the full model, which includes all variables, and then we also fit each model where we've dropped exactly one of the variables. In each of these reduced models, the AIC value for the model is reported below. Based on these results, which variable, if any, should we drop as part of model selection? Explain.

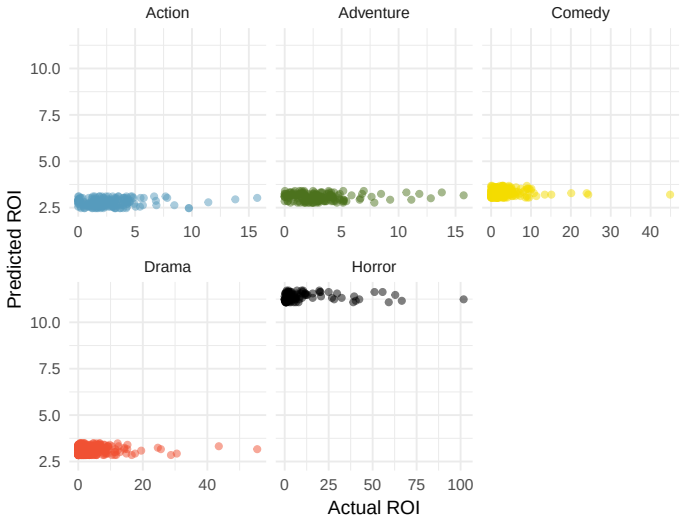
Variable Dropped	AIC
None Dropped	1863.50
`to_multiple`	2023.50
`cc`	1863.18
`attach`	1871.89
`dollar`	1879.70
`winner`	1885.03
`inherit`	1865.55
`password`	1879.31
`format`	2008.85
`re_subj`	1904.60
`exclaim_subj`	1862.76
`sent_email`	1958.18

See the next page for part (b).

(b) Consider the following model selection stage. Here again we've computed the AIC for each leave-one-variable-out model. Based on the results, which variable, if any, should we drop as part of model selection? Explain.

Variable Dropped	AIC
None Dropped	1862.41
`to_multiple`	2019.55
`attach`	1871.17
`dollar`	1877.73
`winner`	1884.95
`inherit`	1864.52
`password`	1878.19
`format`	2007.45
`re_subj`	1902.94
`sent_email`	1957.56

9.22 Movie returns, Part II. The student from Exercise 9.14 analyzed return-on-investment (ROI) for movies based on release year and genre of movies. The plots below show the predicted ROI vs. actual ROI for each of the genres separately. Do these figures support the comment in the FiveThirtyEight.com article that states, "The return-on-investment potential for horror movies is absurd." Note that the x-axis range varies for each plot.



9.23 Spam filtering, Part II. In Exercise 9.21, we encountered a data set where we applied logistic regression to aid in spam classification for individual emails. In this exercise, we've taken a small set of these variables and fit a formal model with the following output:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8124	0.0870	-9.34	0.0000
to_multiple	-2.6351	0.3036	-8.68	0.0000
winner	1.6272	0.3185	5.11	0.0000
format	-1.5881	0.1196	-13.28	0.0000
re_subj	-3.0467	0.3625	-8.40	0.0000

(a) Write down the model using the coefficients from the model fit.

(b) Suppose we have an observation where `to_multiple` = 0, `winner` = 1, `format` = 0, and `re_subj` = 0. What is the predicted probability that this message is spam?

(c) Put yourself in the shoes of a data scientist working on a spam filter. For a given message, how high must the probability a message is spam be before you think it would be reasonable to put it in a *spambox* (which the user is unlikely to check)? What tradeoffs might you consider? Any ideas about how you might make your spam-filtering system even better from the perspective of someone using your email service?