# OpenIntro Statistics

## Third Edition

David M Diez
*Quantitative Analyst*
*david@openintro.org*

Christopher D Barr
*Graduate Student*
*Yale School of Management*
*chris@openintro.org*

Mine Çetinkaya-Rundel
*Assistant Professor of the Practice*
*Department of Statistics*
*Duke University*
*mine@openintro.org*

# Chapter 8

# Multiple and logistic regression

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

## 8.1 Introduction to multiple regression

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

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

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

[1] Diez DM, Barr CD, Çetinkaya-Rundel M. 2015. `openintro`: OpenIntro data sets and supplement functions. github.com/OpenIntroOrg/openintro-r-package.

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

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

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

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

### 8.1.1 A single-variable model for the Mario Kart data

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

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

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

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

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

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

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

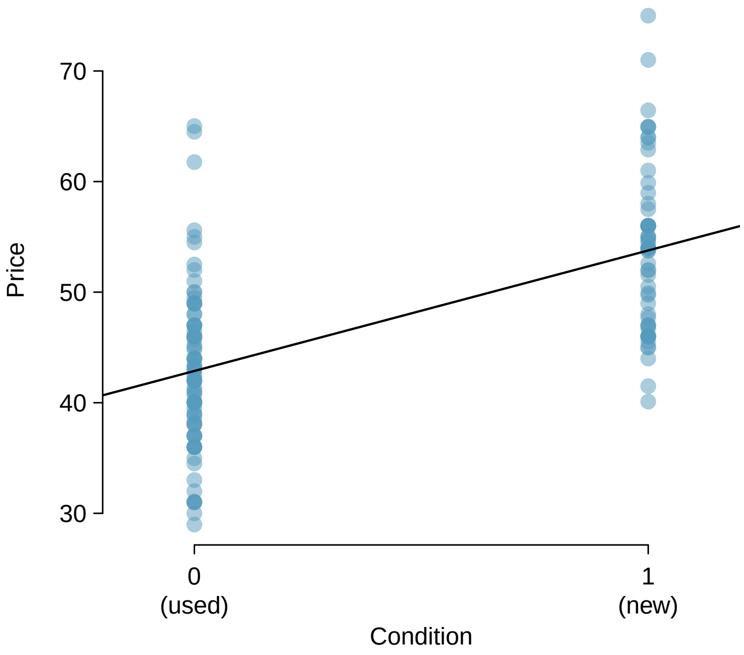


Figure 8.4: Scatterplot of the total auction price against the game's condition. The least squares line is also shown.

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

---

Note that `cond_new` is a two-level categorical variable that takes value 1 when the game is new and value 0 when the game is used. So 10.90 means that the model predicts an extra $10.90 for those games that are new versus those that are used. (See Section 7.2.7 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Table 8.3, we can see that the p-value for `cond_new` is very close to zero, indicating there is strong evidence that the coefficient is different from zero when using this simple one-variable model.

### 8.1.2 Including and assessing many variables in a model

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

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

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

In this equation, $y$ represents the total price, $x_1$ indicates whether the game is new, $x_2$ indicates whether a stock photo was used, $x_3$ is the duration of the auction, and $x_4$ is the number of Wii wheels included with the game. Just as with the single predictor case, a multiple regression model may be missing important components or it might not precisely represent the relationship between the outcome and the available explanatory variables. While no model is perfect, we wish to explore the possibility that this one may fit the data reasonably well.

We estimate the parameters $\beta_0$, $\beta_1$, ..., $\beta_4$ in the same way as we did in the case of a single predictor. We select $b_0$, $b_1$, ..., $b_4$ that minimize the sum of the squared residuals:

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

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

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 36.2110 | 1.5140 | 23.92 | 0.0000 |
| cond_new | 5.1306 | 1.0511 | 4.88 | 0.0000 |
| stock_photo | 1.0803 | 1.0568 | 1.02 | 0.3085 |
| duration | -0.0268 | 0.1904 | -0.14 | 0.8882 |
| wheels | 7.2852 | 0.5547 | 13.13 | 0.0000 |
| | | | | $df = 136$ |

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

### Multiple regression model

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

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

when there are $k$ predictors. We often estimate the $\beta_i$ parameters using a computer.

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

⊙ **Guided Practice 8.6** What does $\beta_4$, the coefficient of variable $x_4$ (Wii wheels), represent? What is the point estimate of $\beta_4$?[4]

⊙ **Guided Practice 8.7** Compute the residual of the first observation in Table 8.1 on page 373 using the equation identified in Guided Practice 8.5.[5]

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

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

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

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

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

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

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

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

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

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

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

[6] Three of the variables (`cond_new`, `stock_photo`, and `wheels`) do take value 0, but the auction duration is always one or more days. If the auction is not up for any days, then no one can bid on it! That means the total auction price would always be zero for such an auction; the interpretation of the intercept in this setting is not insightful.

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

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

### Adjusted $\mathbf{R^2}$ as a tool for model assessment

The **adjusted $\mathbf{R^2}$** is computed as

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

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

Because $k$ is never negative, the adjusted $R^2$ will be smaller – often times just a little smaller – than the unadjusted $R^2$. The reasoning behind the adjusted $R^2$ lies in the **degrees of freedom** associated with each variance.[8]

⊙ **Guided Practice 8.11** There were $n = 141$ auctions in the `mario_kart` data set and $k = 4$ predictor variables in the model. Use $n$, $k$, and the variances from Guided Practice 8.10 to calculate $R^2_{adj}$ for the Mario Kart model.[9]

⊙ **Guided Practice 8.12** Suppose you added another predictor to the model, but the variance of the errors $Var(e_i)$ didn't go down. What would happen to the $R^2$? What would happen to the adjusted $R^2$?[10]

Adjusted $R^2$ could have been used in Chapter 7. 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 considering only one predictor.

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

[8] In multiple regression, the degrees of freedom associated with the variance of the estimate of the residuals is $n-k-1$, not $n-1$. For instance, if we were to make predictions for new data using our current model, we would find that the unadjusted $R^2$ is an overly optimistic estimate of the reduction in variance in the response, and using the degrees of freedom in the adjusted $R^2$ formula helps correct this bias.

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

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

## 8.2 Model selection

The best model is not always the most complicated. Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions. In this section we discuss model selection strategies, which will help us eliminate variables from the model that are found to be less important.

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.

### 8.2.1 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 8.6 and 8.7. The first table summarizes the full model since it includes all predictors, while the second does not include the `duration` variable.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 36.2110 | 1.5140 | 23.92 | 0.0000 |
| cond_new | 5.1306 | 1.0511 | 4.88 | 0.0000 |
| stock_photo | 1.0803 | 1.0568 | 1.02 | 0.3085 |
| duration | -0.0268 | 0.1904 | -0.14 | 0.8882 |
| wheels | 7.2852 | 0.5547 | 13.13 | 0.0000 |
| $R^2_{adj} = 0.7108$ | | | | $df = 136$ |

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

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 36.0483 | 0.9745 | 36.99 | 0.0000 |
| cond_new | 5.1763 | 0.9961 | 5.20 | 0.0000 |
| stock_photo | 1.1177 | 1.0192 | 1.10 | 0.2747 |
| wheels | 7.2984 | 0.5448 | 13.40 | 0.0000 |
| $R^2_{adj} = 0.7128$ | | | | $df = 137$ |

Table 8.7: The fit for the regression model for predictors `cond_new`, `stock_photo`, and `wheels`.

**Example 8.13** Which of the two models is better?

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 `duration` be better than the model with `duration`? We cannot know for sure, but based on the adjusted $R^2$, this is our best assessment.

### 8.2.2 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 8.14** Results corresponding to the *full model* for the `mario_kart` data are shown in Table 8.6. How should we proceed under the backward elimination strategy?

Our baseline adjusted $R^2$ from the full model is $R^2_{adj} = 0.7108$, and we need to determine whether dropping a predictor will improve the adjusted $R^2$. To check, we fit four models that each drop a different predictor, and we record the adjusted $R^2$ from each:

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

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

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 $R^2_{adj} = 0.7128$. We now fit three new models, which consider eliminating each of the three remaining predictors:

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

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 `cond_new`, `stock_photos`, and `wheels`, which we can summarize using the coefficients from Table 8.7:

$$\begin{aligned} \hat{y} &= b_0 + b_1 x_1 + b_2 x_2 + b_4 x_4 \\ \widehat{price} &= 36.05 + 5.18 \times \texttt{cond\_new} + 1.12 \times \texttt{stock\_photo} + 7.30 \times \texttt{wheels} \end{aligned}$$

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 8.15** Construct a model for the `mario_kart` data set using the forward selection strategy.

We start with the model that includes no variables. Then we fit each of the possible models with just one variable. That is, we fit the model including just `cond_new`, then the model including just `stock_photo`, then a model with just `duration`, and a model with just `wheels`. Each of the four models provides an adjusted $R^2$ value:

| Add ... | `cond_new` | `stock_photo` | `duration` | `wheels` |
|---|---|---|---|---|
| | $R^2_{adj} = 0.3459$ | $R^2_{adj} = 0.0332$ | $R^2_{adj} = 0.1338$ | $R^2_{adj} = 0.6390$ |

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 `wheels` 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 `wheels` and with a new baseline of $R^2_{adj} = 0.6390$:

| Add `wheels` and ... | `cond_new` | `stock_photo` | `duration` |
|---|---|---|---|
| | $R^2_{adj} = 0.7124$ | $R^2_{adj} = 0.6587$ | $R^2_{adj} = 0.6528$ |

The best predictor in this stage, `cond_new`, has a higher adjusted $R^2$ (0.7124) than the baseline (0.6390), so we also add `cond_new` 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 `wheels`, `cond_new`, and ... | `stock_photo` | `duration` |
|---|---|---|
| | $R^2_{adj} = 0.7128$ | $R^2_{adj} = 0.7107$ |

The model adding `stock_photo` improved adjusted $R^2$ (0.7124 to 0.7128), so we add `stock_photo` to the model.

Because we have again added a predictor, we check whether adding the last variable, `duration`, will improve adjusted $R^2$. We compare the adjusted $R^2$ for the model with `duration` and the other three predictors (0.7108) to the model that only considers `wheels`, `cond_new`, and `stock_photo` (0.7128). Adding `duration` does not improve the adjusted $R^2$, so we do not add it to the model, and we have arrived at the same model that we identified from backward elimination.

### Model selection strategies

Backward elimination begins with the largest model 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.

There is no guarantee that backward elimination and forward selection will arrive at the same final model. If both techniques are tried and they arrive at different models, we choose the model with the larger $R^2_{adj}$; other tie-break options exist but are beyond the scope of this book.

### 8.2.3 The p-value approach, an alternative to adjusted $R^2$

The p-value may be used as an alternative to adjusted $R^2$ for model selection.

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.

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 8.16** Examine Table 8.7 on page 378, which considers the model including the `cond_new`, `stock_photo`, and `wheels` predictors. If we were using the p-value approach with backward elimination and we were considering this model, which of these three variables would be up for elimination? Would we drop that variable, or would we keep it in the model?[11]

While the adjusted $R^2$ and p-value approaches are similar, they sometimes lead to different models, with the adjusted $R^2$ approach tending to include more predictors in the final model. For example, if we had used the p-value approach with the auction data, we would not have included the `stock_photo` predictor in the final model.

**When to use the adjusted $R^2$ and when to use the p-value approach**

When the sole goal is to improve prediction accuracy, use adjusted $R^2$. 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 adjusted $R^2$ 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.

[11] The `stock_photo` predictor is up for elimination since it has the largest p-value. Additionally, since that p-value is larger than 0.05, we would in fact eliminate `stock_photo` from the model.

## 8.3 Checking model assumptions using graphs

Multiple regression methods using the model

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

generally depend on the following four assumptions:

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

**Diagnostic plots** can be used to check each of these assumptions. We will consider the model from the Mario Kart auction data, and check whether there are any notable concerns:

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

**Normal probability plot.** A normal probability plot of the residuals is shown in Figure 8.8. While the plot exhibits some minor irregularities, there are no outliers that might be cause for concern. In a normal probability plot for residuals, we tend to be most worried about residuals that appear to be outliers, since these indicate long tails in the distribution of residuals.

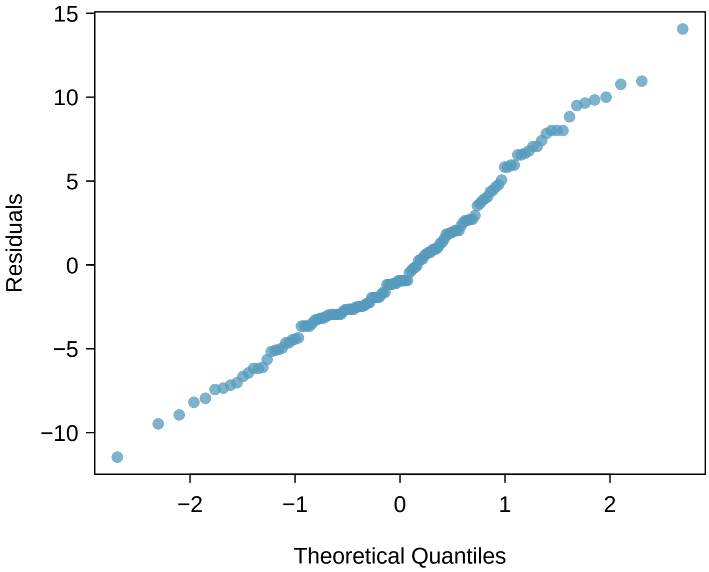


Figure 8.8: A normal probability plot of the residuals is helpful in identifying observations that might be outliers.

**Absolute values of residuals against fitted values.** A plot of the absolute value of the residuals against their corresponding fitted values ($\hat{y}_i$) is shown in Figure 8.9. This plot is helpful to check the condition that the variance of the residuals is approximately constant. We don't see any obvious deviations from constant variance in this example.

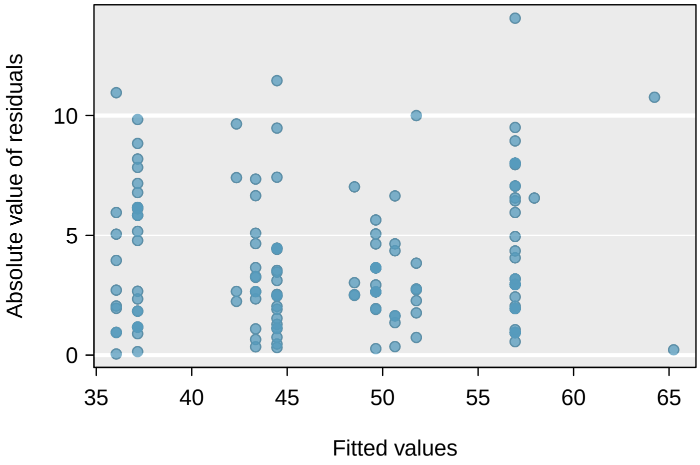


Figure 8.9: Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.

**Residuals in order of their data collection.** A plot of the residuals in the order their corresponding auctions were observed is shown in Figure 8.10. Such a plot is helpful in identifying any connection between cases that are close to one another, e.g. we could look for declining prices over time or if there was a time of the day when auctions tended to fetch a higher price. Here we see no structure that indicates a problem.[12]

**Residuals against each predictor variable.** We consider a plot of the residuals against the `cond_new` variable, the residuals against the `stock_photo` variable, and the residuals against the `wheels` variable. These plots are shown in Figure 8.11. 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 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.

It is necessary to summarize diagnostics for any model fit. If the diagnostics support the model assumptions, this would improve credibility in the findings. If the diagnostic assessment shows remaining underlying structure in the residuals, we should try to adjust the model to account for that structure. If we are unable to do so, we may still report the model but must also note its shortcomings. In the case of the auction data, we report that there 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.

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

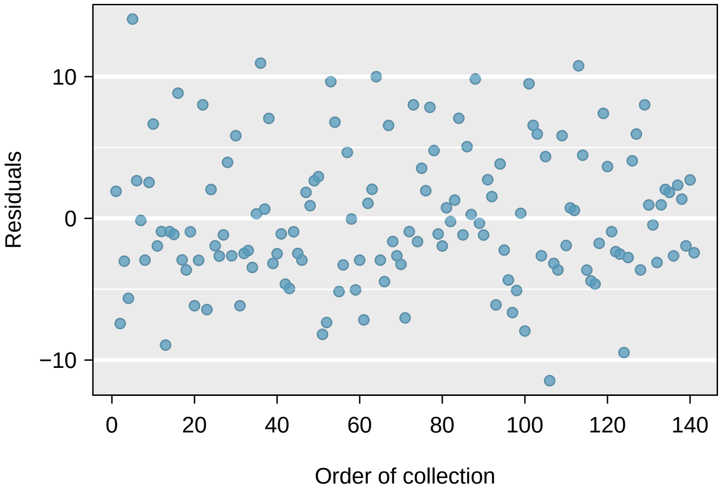


Figure 8.10: Plotting residuals in the order that their corresponding observations were collected helps identify connections between successive observations. If it seems that consecutive observations tend to be close to each other, this indicates the independence assumption of the observations would fail.

**“All models are wrong, but some are useful” -George E.P. Box**

The truth is that no model is perfect. However, even imperfect models can be useful. Reporting a flawed model can be reasonable so long as we are clear and report the model’s shortcomings.

**Caution: Don’t report results when assumptions are grossly violated**

While there is a little leeway in model assumptions, don’t go too far. If model assumptions are very clearly violated, consider a new model, even if it means learning more statistical methods or hiring someone who can help.

**TIP: Confidence intervals in multiple regression**

Confidence intervals for coefficients in multiple regression can be computed using the same formula as in the single predictor model:

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

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

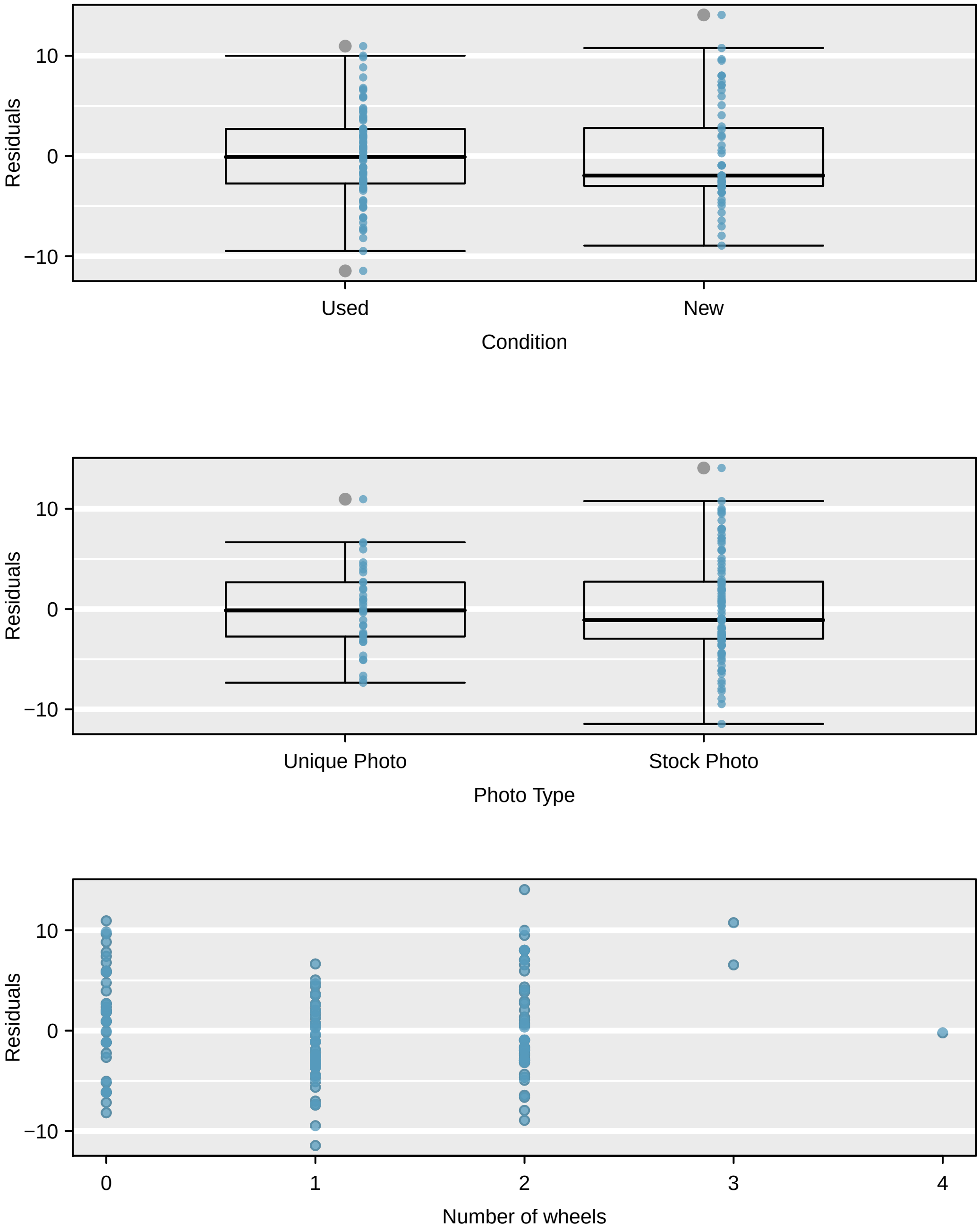


Figure 8.11: For the condition and stock photo variables, we check for differences in the distribution shape or variability of the residuals. In the case of the stock photos variable, we see a little less variability in the unique photo group than the stock photo group. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the `wheels` variable in the bottom plot.

## 8.5 Exercises

### 8.5.1 Introduction to multiple regression

**8.1 Baby weights, Part I.** The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between smoking and weight of the baby. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother.[17]

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 123.05 | 0.65 | 189.60 | 0.0000 |
| smoke | -8.94 | 1.03 | -8.65 | 0.0000 |

The variability within the smokers and non-smokers are about equal and the distributions are symmetric. With these conditions satisfied, it is reasonable to apply the model. (Note that we don't need to check linearity since the predictor has only two levels.)

(a) Write the equation of the regression line.

(b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.

(c) Is there a statistically significant relationship between the average birth weight and smoking?

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

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

(a) Write the equation of the regression line.

(b) Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.

(c) Is there a statistically significant relationship between the average birth weight and parity?

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

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

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

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

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

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

(b) Interpret the slopes of gestation and age in this context.

(c) The coefficient for parity is different than in the linear model shown in Exercise 8.2. Why might there be a difference?

(d) Calculate the residual for the first observation in the data set.

(e) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the $R^2$ and the adjusted $R^2$. Note that there are 1,236 observations in the data set.

**8.4 Absenteeism, 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).[18]

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 18.93 | 2.57 | 7.37 | 0.0000 |
| eth | -9.11 | 2.60 | -3.51 | 0.0000 |
| sex | 3.10 | 2.64 | 1.18 | 0.2411 |
| lrn | 2.15 | 2.65 | 0.81 | 0.4177 |

(a) Write the equation of the regression line.

(b) Interpret each one of the slopes in this context.

(c) Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school.

(d) The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the $R^2$ and the adjusted $R^2$. Note that there are 146 observations in the data set.

**8.5 GPA.** A survey of 55 Duke University students asked about their GPA, number of hours they study at night, number of nights they go out, and their gender. Summary output of the regression model is shown below. Note that male is coded as 1.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.45 | 0.35 | 9.85 | 0.00 |
| studyweek | 0.00 | 0.00 | 0.27 | 0.79 |
| sleepnight | 0.01 | 0.05 | 0.11 | 0.91 |
| outnight | 0.05 | 0.05 | 1.01 | 0.32 |
| gender | -0.08 | 0.12 | -0.68 | 0.50 |

(a) Calculate a 95% confidence interval for the coefficient of gender in the model, and interpret it in the context of the data.

(b) Would you expect a 95% confidence interval for the slope of the remaining variables to include 0? Explain

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

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

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

(a) Calculate a 95% confidence interval for the coefficient of height, and interpret it in the context of the data.

(b) One tree in this sample is 79 feet tall, has a diameter of 11.3 inches, and is 24.2 cubic feet in volume. Determine if the model overestimates or underestimates the volume of this tree, and by how much.

## 8.5.2 Model selection

**8.7 Baby weights, Part IV.** Exercise 8.3 considers a model that predicts a newborn's weight using several predictors (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?

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

**8.8 Absenteeism, Part II.** Exercise 8.4 considers a model that predicts the number of days absent using three predictors: ethnic background (`eth`), gender (`sex`), and learner status (`lrn`). 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?

**8.9 Baby weights, Part V.** Exercise 8.3 provides regression output for the full model (including all explanatory variables available in the data set) for predicting birth weight of babies. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted $R^2$ of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

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

**8.10 Absenteeism, Part III.** Exercise 8.4 provides regression output for the full model, including all explanatory variables available in the data set, for predicting the number of days absent from school. In this exercise we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted $R^2$ of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

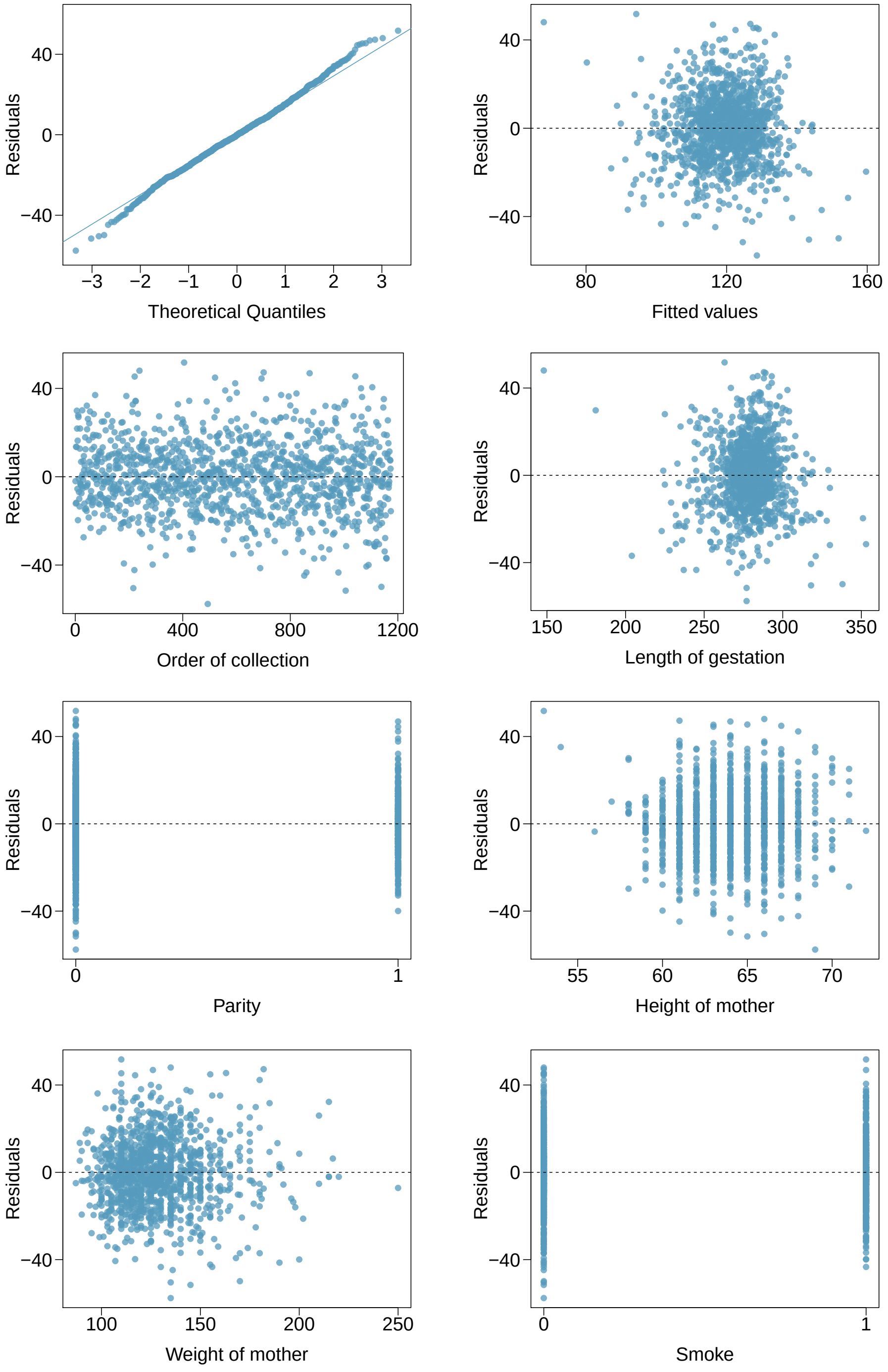
| variable | ethnicity | sex | learner status |
|---|---|---|---|
| p-value | 0.0007 | 0.3142 | 0.5870 |
| $R^2_{adj}$ | 0.0714 | 0.0001 | 0 |

**8.11 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?

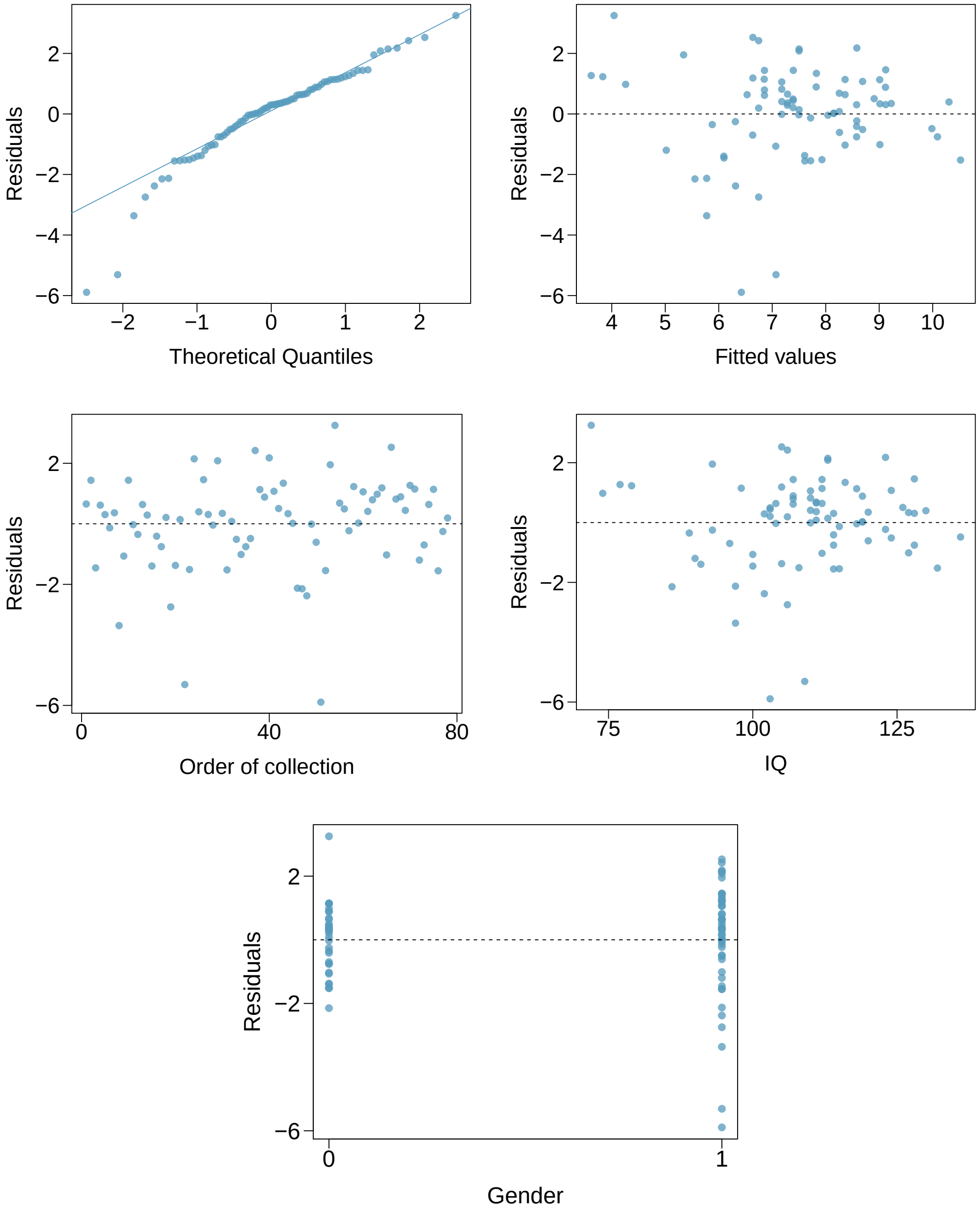
**8.12 Movie lovers, Part II.** Suppose an online media streaming company is interested in building a movie recommendation system. The website maintains data on the movies in their database (genre, length, cast, director, budget, etc.) and additionally collects data from their subscribers (demographic 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?

### 8.5.3 Checking model assumptions using graphs

**8.13 Baby weights, Part V.** Exercise 8.3 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.



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



# Appendix A

# End of chapter exercise solutions

in husband's height, the average wife's height is expected to be an additional 0.2863 inches on average. Intercept: Men who are 0 inches tall are expected to have wives who are, on average, 43.5755 inches tall. The intercept here is meaningless, and it serves only to adjust the height of the line. (d) The slope is positive, so $r$ must also be positive. $r = \sqrt{0.09} = 0.30$. (e) 63.2612. Since $R^2$ is low, the prediction based on this regression model is not very reliable. (f) No, we should avoid extrapolating.

**7.39** (a) $r = \sqrt{0.28} \approx -0.53$. We know the correlation is negative due to the negative association shown in the scatterplot. (b) The residuals appear to be fan shaped, indicating non-constant variance. Therefore a simple least squares fit is not appropriate for these data.

**7.41** (a) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$ (b) The p-value for this test is approximately 0, therefore we reject $H_0$. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate. (c) $n = 20, df = 18, T^*_{18} = 2.10$; $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$; For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million. (d) Yes, we rejected $H_0$ and the confidence interval does not include 0.

**7.43** This is a one-sided test, so the p-value should be half of the p-value given in the regression table, which will be approximately 0. Therefore the data provide convincing evidence that poverty percentage is positively associated with murder rate.

### 8 Multiple and logistic regression

**8.1** (a) $\widehat{baby\_weight} = 123.05 - 8.94 \times smoke$ (b) The estimated body weight of babies born to smoking mothers is 8.94 ounces lower than babies born to non-smoking mothers. Smoker: $123.05 - 8.94 \times 1 = 114.11$ ounces. Non-smoker: $123.05 - 8.94 \times 0 = 123.05$ ounces. (c) $H_0$: $\beta_1 = 0$. $H_A$: $\beta_1 \neq 0$. $T = -8.65$, and the p-value is approximately 0. Since the p-value is very small, we reject $H_0$. The data provide strong evidence that the true slope parameter is different than 0 and that there is an association between birth weight and smoking. Furthermore, having rejected $H_0$, we can conclude that smoking is associated with lower birth weights.

**8.3** (a) $\widehat{baby\_weight} = -80.41 + 0.44 \times gestation - 3.33 \times parity - 0.01 \times age + 1.15 \times height + 0.05 \times weight - 8.40 \times smoke$. (b) $\beta_{gestation}$: The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day of pregnancy, all else held constant. $\beta_{age}$: The model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant. (c) Parity might be correlated with one of the other variables in the model, which complicates model estimation. (d) $\widehat{baby\_weight} = 120.58$. $e = 120 - 120.58 = -0.58$. The model over-predicts this baby's birth weight. (e) $R^2 = 0.2504$. $R^2_{adj} = 0.2468$.

**8.5** (a) (-0.32, 0.16). We are 95% confident that male students on average have GPAs 0.32 points lower to 0.16 points higher than females when controlling for the other variables in the model. (b) Yes, since the p-value is larger than 0.05 in all cases (not including the intercept).

**8.7** Remove age.

**8.9** Based on the p-value alone, either gestation or smoke should be added to the model first. However, since the adjusted $R^2$ for the model with gestation is higher, it would be preferable to add gestation in the first step of the forward-selection algorithm. (Other explanations are possible. For instance, it would be reasonable to only use the adjusted $R^2$.)

**8.11** She should use p-value selection since she is interested in finding out about significant predictors, not just optimizing predictions.

**8.13** Nearly normal residuals: The normal probability plot shows a nearly normal distribution of the residuals, however, there are some minor irregularities at the tails. With a data set so large, these would not be a concern.

Constant variability of residuals: The scatterplot of the residuals versus the fitted values does not show any overall structure. However, values that have very low or very high fitted values appear to also have somewhat larger outliers. In addition, the residuals do appear to have constant variability between the two parity and smoking status groups, though these items are relatively minor.

Independent residuals: The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that there is no apparent structures related to the order the data were collected.

Linear relationships between the response variable and numerical explanatory variables: The residuals vs. height and weight of mother are randomly distributed around 0. The residuals vs. length of gestation plot also does not show any clear or strong remaining structures, with the possible exception of very short or long gestations. The rest of the residuals do appear to be randomly distributed around 0.

All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor.

**8.15** (a) There are a few potential outliers, e.g. on the left in the `total_length` variable, but nothing that will be of serious concern in a data set this large. (b) When coefficient estimates are sensitive to which variables are included in the model, this typically indicates that some variables are collinear. For example, a possum's gender may be related to its head length, which would explain why the coefficient (and p-value) for `sex_male` changed when we removed the `head_length` variable. Likewise, a possum's skull width is likely to be related to its head length, probably even much more closely related than the head length was to gender.