

Chapter 8

Multiple regression and ANOVA

The principles of simple linear regression with one numerical predictor and one numerical response 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. We will also consider methods for analysis of variance (ANOVA), a tool useful both in practice and when learning about the mechanics of regression.

8.1 Introduction to multiple regression

Multiple regression extends the simple bivariate regression (two variables: x and y) to the case that still has one response but may have many predictors (denoted x_1, x_2, x_3, \dots). 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 - the highest bid plus the shipping cost. But how is the total price related to characteristics of an auction? For instance, are longer auctions associated with a 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 `marioKart` includes results from 143 auctions¹. Four observations from this data set are shown in Table 8.1, and descriptions for each variable are shown in Table 8.2.

8.1.1 Using categorical variables with two levels as predictors

There are two predictor variables in the `marioKart` data set that are inherently categorical: a variable describing the condition of the game and the variable describing whether a stock photo was used for the auction. Two-level categorical variables are often coded using 0's

¹Diez DM, Barr CD, and Çetinkaya M. 2011. *openintro*: OpenIntro data sets and supplemental functions. R package Version 1.2.

	totalPr	condNew	stockPhoto	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	7	1
⋮	⋮	⋮	⋮	⋮	⋮
142	38.76	0	0	7	0
143	54.51	1	1	1	2

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

variable	description
<code>totalPr</code>	the total of the final auction price and the shipping cost, in US dollars
<code>condNew</code>	a coded two-level categorical variable, which takes value 1 when the game is new and 0 if the game is used
<code>stockPhoto</code>	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
<code>duration</code>	the length of the auction, in days
<code>wheels</code>	the number of Wii wheels included with the auction (a <i>Wii wheel</i> is a plastic racing wheel that holds the Wii controller and is an optional but helpful accessory for playing Mario Kart Wii)

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

and 1's, which allows them to be incorporated into a regression model in the same way as a numerical predictor:

$$\widehat{\text{totalPr}} = \beta_0 + \beta_1 * \text{condNew}$$

If we fit this model for total price and game condition using simple linear regression, we obtain the following regression line estimate:

$$\widehat{\text{totalPr}} = 42.87 + 10.90 * \text{condNew} \quad (8.1)$$

The 0-1 coding of the two-level categorical variable allows for a simple interpretation of the coefficient of `condNew`. When the game is in `used` condition, the `condNew` variable takes a value of zero, and the total auction price predicted from the model would be $\$42.87 + \$10.90 * (0) = \$42.87$. If the game is in `new` condition, then the `condNew` variable takes value one and the total price is predicted to be $\$42.87 + \$10.90 * (1) = \$53.77$. We now see clearly that the coefficient of `condNew` estimates the difference ($\$10.90$) in the total auction price when the game is new ($\$53.77$) versus used ($\42.87).

TIP: The coefficient of a two-level categorical variable

The coefficient of a binary variable corresponds to the estimated difference in the outcome between the two levels of the variable.

- ⊙ **Exercise 8.2** The best fitting linear model for the outcome `totalPr` and predictor `stockPhoto` is

$$\widehat{\text{totalPr}} = 44.33 + 4.17 * \text{stockPhoto} \quad (8.3)$$

where the variable `stockPhoto` takes value 1 when a stock photo is being used and 0 when the photo is unique to that auction. Interpret the coefficient of `stockPhoto`.

- **Example 8.4** In Exercise 8.2, you found that auctions whose primary photo was a stock photo tended to sell for about \$4.17 more than auctions that feature a unique photo. Suppose a seller learns this and decides to change her Mario Kart Wii auction to have its primary photo be a stock photo. Will modifying her auction in this way earn her, on average, an additional \$4.17?

No, we cannot infer a causal relationship. It might be that there are inherent differences in auctions that use stock photos and those that do not. For instance, if we sorted through the data, we would actually notice that many of the auctions with stock photos tended to also include more Wii wheels. In this case, Wii wheels is a potential lurking variable.

8.1.2 Including and assessing many variables in a model

Sometimes there is underlying structure or relationship between the 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 included all potentially important variables simultaneously, which 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.

Earlier we had constructed a simple linear model using `condNew` as a predictor and `totalPr` as the outcome. We also constructed a separate model using only `stockPhoto` as a predictor. Next, we want a model that uses both of these variables simultaneously and, while we're at it, we'll include the `duration` and `wheels` variables described Table 8.2:

$$\begin{aligned}\widehat{\text{totalPr}} &= \beta_0 + \beta_1 * \text{condNew} + \beta_2 * \text{stockPhoto} \\ &\quad + \beta_3 * \text{duration} + \beta_4 * \text{wheels} \\ \hat{y} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4\end{aligned}\tag{8.5}$$

where y represents the total price, x_1 is the game's condition, x_2 is 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, this model may be missing important components or it might not properly represent the relationship between the total price and the available explanatory variables. However, while no model is perfect, we wish to explore the possibility that this one may fit the data reasonably well.

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

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2\tag{8.6}$$

We typically use a computer to minimize this sum and compute point estimates, as shown in the sample output in Table 8.3. Using this output, we identify the point estimates b_i of each β_i , just as we did in the one-predictor case.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.2110	1.5140	23.92	0.0000
condNew	5.1306	1.0511	4.88	0.0000
stockPhoto	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 8.3: Output for the regression model where `totalPr` is the outcome and `condNew`, `stockPhoto`, `duration`, and `wheels` are the predictors.

Multiple regression model

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

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

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

- ⊙ **Exercise 8.7** Write out the model in Equation (8.5) using the point estimates from Table 8.3. How many predictors are there in this model? Answers in the footnote².
- ⊙ **Exercise 8.8** What does β_4 , the coefficient of variable x_4 (Wii wheels), represent? What is the point estimate of β_4 ? Answers in the footnote³.
- ⊙ **Exercise 8.9** Compute the residual of the first observation in Table 8.1 on page 310. Hint: use the equation from Exercise 8.7. Answer in the footnote⁴.
- **Example 8.10** The coefficients for x_1 (`condNew`) and x_2 (`stockPhoto`) are different than in the two simple linear models shown in Equations (8.1) and (8.3). Why might there be a difference?

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

Example 8.10 describes a common issue in multiple regression: correlation among predictor variables. We say the two predictor variables are **collinear** 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 correlated.

² $\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.03x_3 + 7.29x_4$, and there are $p = 4$ predictor variables.

³It is the average difference in auction price for each additional Wii wheel included when holding the other variables constant. The point estimate is $b_4 = 7.29$.

⁴ $e_i = y_i - \hat{y}_i = 51.55 - 49.62 = 1.93$, where 49.62 was computed using the predictor values for the observation and the equation identified in Exercise 8.7.

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:

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

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

- ⊙ **Exercise 8.11** The variance of the residuals for the model given in Exercise 8.9 is 23.34, and the variance of the total price in all the auctions is 83.06. Verify the R^2 for this model is 0.719.

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

Adjusted R^2 as a tool for model assessment

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

$$R_{adj}^2 = 1 - \frac{\text{Var}(e_i)/(n-p-1)}{\text{Var}(y_i)/(n-1)} = 1 - \frac{\text{Var}(e_i)}{\text{Var}(y_i)} \times \frac{n-1}{n-p-1}$$

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

Because p 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 with the **degrees of freedom** associated with each variance⁵.

- ⊙ **Exercise 8.12** There were $n = 141$ auctions in the `marioKart` data set and $p = 4$ predictor variables in the model. Use n , p , and the variances from Exercise 8.11 to verify $R_{adj}^2 = 0.711$ for the Mario Kart model.
- ⊙ **Exercise 8.13** Suppose you added another predictor to the model, but the variance of the errors $\text{Var}(e_i)$ didn't go down. What would happen to the R^2 ? What would happen to the adjusted R^2 ? Answers in the footnote⁶.

The idea that a predictor that doesn't explain any extra variance would actually “hurt” the adjusted R^2 highlights a common sentiment in statistics: avoid making a model more complicated than it needs to be.

⁵In multiple regression, the degrees of freedom associated with the variance of the estimate of the residuals is $n-p-1$, 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.

⁶The 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 that are less important from the model.

In this section, and in practice, the model that includes all available explanatory variables is often referred to as the **full model**. Our goal is assess whether the full model is the best model. If it isn't, we want to identify a smaller model that is preferable.

8.2.1 Identifying variables that may not be helpful in the model

Table 8.4 provides a summary of the regression output for the full model. The last column of the table lists p-values that can be used to assess hypotheses of the following form:

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

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

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.2110	1.5140	23.92	0.0000
condNew	5.1306	1.0511	4.88	0.0000
stockPhoto	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 8.4: The fit for the full regression model. This table is identical to Table 8.3.

- **Example 8.14** The coefficient of `condNew` has a t test statistic of $T = 4.88$ and a p-value for its corresponding hypotheses ($H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$) of about zero. How can this be interpreted?

If we keep all the other variables in the model and add no others, then there is strong evidence that a game's condition (new or used) has a real relationship with the total auction price.

- **Example 8.15** Is there strong evidence that using a stock photo is related to the total auction price?

The t test statistic for `stockPhoto` is $T = 1.02$ and the p-value is about 0.31. There is not strong evidence that using a stock photo in an auction is related to the total price of the auction. We might consider removing the `stockPhoto` variable from the model.

- ⊙ **Exercise 8.16** Identify the p-values for both the `duration` and `wheels` variables in the model. Is there strong evidence supporting the connection of these variables with the total price in the model?

There is not statistically significant evidence that either `stockPhoto` or `duration` are meaningfully contributing to the model. If the coefficients of these variables are not zero, their association with the outcome variable is probably weak. Next we consider common strategies for pruning such variables from a model.

TIP: Using adjusted R^2 instead of p-values for model selection

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

8.2.2 Two model selection strategies

Two common strategies for adding or removing variables in a multiple regression model are called *backward-selection* and *forward-selection*. These techniques are often referred to as **stepwise** model selection strategies, because they add or delete one variable at a time as they “step” through the candidate predictors. We will discuss these strategies in the context of the p-value approach, however, the adjusted R^2 approach may be employed as an alternative.

The **backward-elimination** strategy starts with the model that includes all potential predictor variables. One-by-one variables are eliminated from the model until only variables with statistically significant p-values remain. The strategy within each elimination step is to drop the variable with the largest p-value, refit the model, and reassess the inclusion of all variables.

- **Example 8.17** Results corresponding to the *full model* for the `marioKart` data are shown in Table 8.4. How should we proceed under the backward-elimination strategy?

There are two variables with coefficients that are not statistically different from zero: `stockPhoto` and `duration`. We first drop the `duration` variable since it has a larger corresponding p-value, *then we refit the model*. A regression summary for the new model is shown in Table 8.5.

In the new model, there is not strong evidence that the coefficient for `stockPhoto` is different from zero (even though the p-value dropped a little) and the other p-values remain very small. So again we eliminate the variable with the largest non-significant p-value, `stockPhoto`, and refit the model. The updated regression summary is shown in Table 8.6.

In the latest model, we see that the two remaining predictors have statistically significant coefficients with p-values of about zero. Since there are no variables remaining that could be eliminated from the model, we stop. The final model includes only the `condNew` and `wheels` variables in predicting the total auction price:

$$\hat{y} = b_0 + b_1x_1 + b_4x_4 = 36.78 + 5.58x_1 + 7.23x_4$$

As an alternative description of how we could have performed this model selection strategy using adjusted R^2 , please see the footnote⁷.

⁷At each elimination step, we refit the model without each of the variables up for potential elimination (e.g. in the first step, we would fit four models, where each would be missing a different predictor). If one of these smaller models has a higher adjusted R^2 than our current model, we pick the smaller model with the largest adjusted R^2 . Had we used the adjusted R^2 criteria, we would have kept the `stockPhoto` variable in this backwards-elimination example.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.0483	0.9745	36.99	0.0000
condNew	5.1763	0.9961	5.20	0.0000
stockPhoto	1.1177	1.0192	1.10	0.2747
wheels	7.2984	0.5448	13.40	0.0000

df = 137

Table 8.5: The output for the regression model where `totalPr` is the outcome and the duration variable has been eliminated from the model.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.7849	0.7066	52.06	0.0000
condNew	5.5848	0.9245	6.04	0.0000
wheels	7.2328	0.5419	13.35	0.0000

df = 138

Table 8.6: The output for the regression model where `totalPr` is the outcome and the duration and stock photo variables have been eliminated from the model.

Notice that the p-value for `stockPhoto` changed a little from the full model (0.309) to the model that did not include the `duration` variable (0.275). It is common for p-values of one variable to change, due to collinearity, after eliminating a different variable. This fluctuation emphasizes the importance of refitting a model after each variable elimination step. The p-values tend to change dramatically when the eliminated variable is highly correlated with another variable in the model.

The **forward-selection** strategy is the reverse of the backward-elimination technique. Instead of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot find any variables that present strong evidence of their importance in the model.

- **Example 8.18** Construct a model for the `marioKart` data set using the forward-selection strategy.

We start with the model that includes no variables. Then we fit each of the possible models with just one variable. That is, we fit the model including just the `condNew` predictor, then the model just including the `stockPhoto` variable, then a model with just `duration`, and a model with just `wheels`. Each of the four models (yes, we fit four models!) provides a p-value for the coefficient of the predictor variable. Out of these four variables, the `wheels` variable had the smallest p-value. Since its p-value is less than 0.05 (the p-value was smaller than $2e-16$), we add the Wii wheels variable to the model. Once a variable is added in forward-selection, it will be included in all models considered and in the final model.

Since we successfully found a first variable to add, we consider adding another. We fit three new models: (1) the model including just the `condNew` and `wheels` variables (output in Table 8.6), (2) the model including just the `stockPhoto` and `wheels` variables, and (3) the model including only the `duration` and `wheels` variables. Of these models, the first had the lowest p-value for its new variable (the p-value corresponding to `condNew` was $1.4e-08$). Because this p-value is below 0.05, we add the `condNew` variable to the model. Now the final model is guaranteed to include both the condition and Wii wheels variables.

We repeat the process a third time, fitting two new models: (1) the model including the `stockPhoto`, `condNew`, and `wheels` variables (output in Table 8.5) and (2) the model including the `duration`, `condNew`, and `wheels` variables. The p-value corresponding to `stockPhoto` in the first model (0.275) was smaller than the p-value corresponding to `duration` in the second model (0.682). However, since this smaller p-value was not below 0.05, there was not strong evidence that it should be included in the model. Therefore, neither variable is added and we are finished.

The final model is the same as that arrived at using the backward-selection strategy: we include the `condNew` and `wheels` variables into the final model. See the footnote for how we would have proceeded had we used the R_{adj}^2 criteria instead of examining p-values⁸.

Model selection strategies

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

There is no guarantee that the backward-elimination and forward-selection strategies will arrive at the same final model regardless of whether we are using the p-value or R_{adj}^2 criteria. If the backwards-elimination and forward-selection strategies are both tried and they arrive at different models, one option is to choose between the models using the R_{adj}^2 criteria (other options exist but are beyond the scope of this book).

It is generally acceptable to use just one strategy, usually backward-elimination, and report the final model after verifying the conditions for fitting a linear model are reasonable.

8.3 Checking model assumptions using graphs

Multiple regression methods using the model

$$\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$$

generally depend on the following four assumptions:

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

Simple and effective plots can be used to check each of these assumptions.

⁸Rather than look for variables with the smallest p-value, we look for the model with the largest R_{adj}^2 . Using the forward-selection strategy, we start with the model with no predictors. Next we look at each model with a single predictor. If one of these models has a larger R_{adj}^2 than the model with no variables, we use this new model. We repeat this procedure, adding one variable at a time, until we cannot find a model with a smaller R_{adj}^2 . If we had done the forward-selection strategy using R_{adj}^2 , we would have arrived at the model including `condNew`, `stockPhoto`, and `wheels`, which is a slightly larger model than we arrived at using the p-value approach.

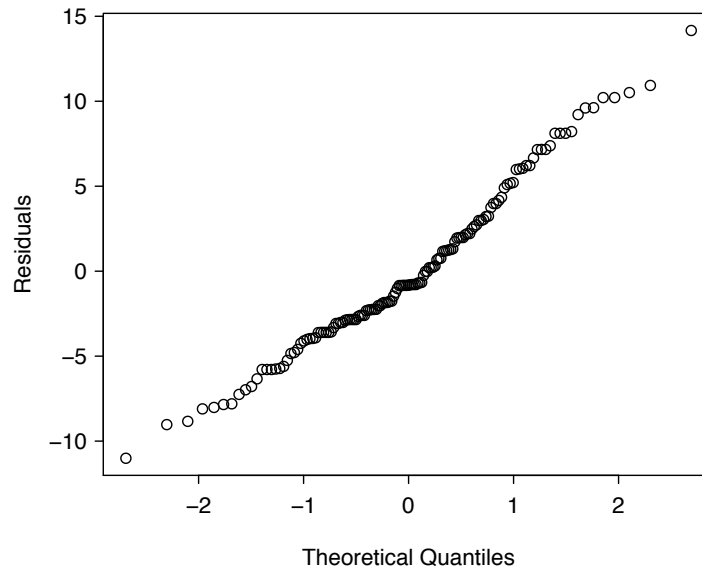


Figure 8.7: A normal probability plot of the residuals is helpful in identifying observations that might be outliers.

Normal probability plot. A normal probability plot of the residuals is shown in Figure 8.7. While the plot exhibits some minor irregularities, there are no outliers that might be cause for concern. In a normal probability plot for residuals, we tend to be most worried about residuals that appear to be outliers, since these indicate long tails in the distribution of residuals.

Absolute values of residuals against fitted values. A plot of the absolute value of the residuals against their corresponding fitted values (\hat{y}_i) is shown in Figure 8.8. This plot is helpful to check the condition that the variance of the residuals is approximately constant. We don't see any obvious deviations from constant variance in this example.

Residuals in order of their data collection. A plot of the residuals in the order their corresponding auctions were observed is shown in Figure 8.9. 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 problem⁹.

Residuals against each predictor variable. We consider a plot of the residuals against the `condNew` variable and the residuals against the `wheels` variable. These plots are shown in Figure 8.10. For the two-level condition variable, we are guaranteed not to see a trend, and instead we are verifying that the variability doesn't fluctuate across groups. In this example, when we consider the residuals against the `wheels` variable, we see structure. There appears to be curvature in the residuals, indicating the relationship is probably not linear.

⁹An 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.

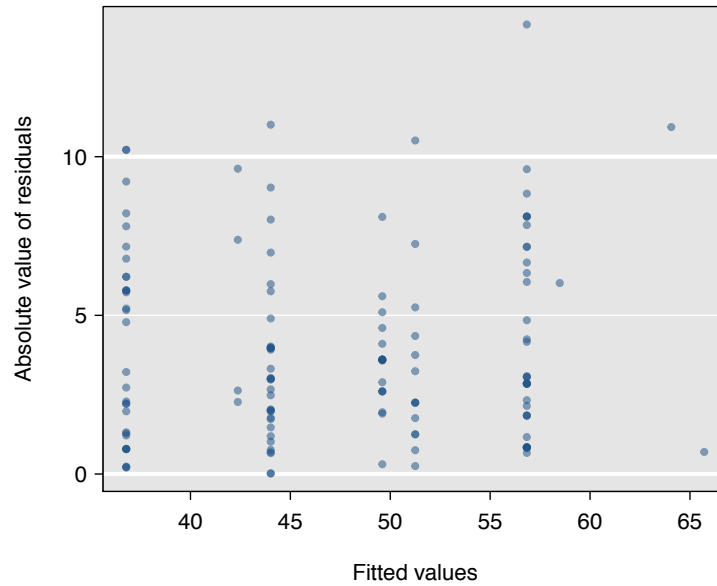


Figure 8.8: Comparing the absolute value of the residuals against the fitted values (\hat{y}_i) is helpful in identifying deviations from the constant variance assumption.

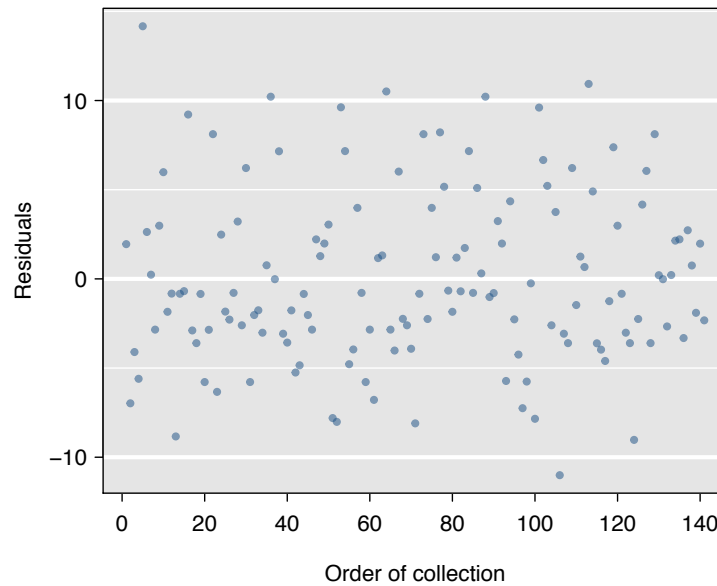


Figure 8.9: Plotting residuals in the order that their corresponding observations were collected helps identify connections between successive observations. If it seems that consecutive observations tend to be close to each other, this indicates the independence assumption of the observations would fail.

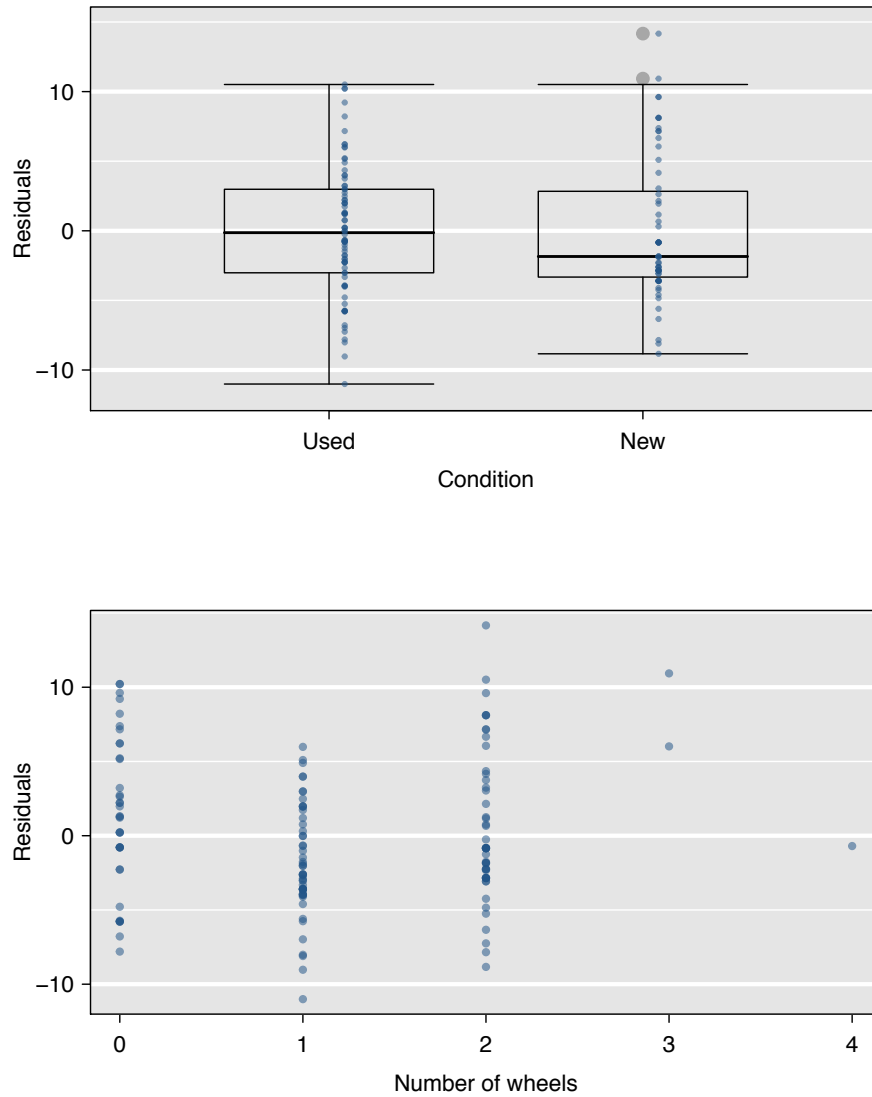


Figure 8.10: In the two-level variable for the game's condition, we check for differences in distribution shape or variability. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the `wheels` variable.

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

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

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

Caution: Don’t report results when assumptions are heavily violated

While there is a little leeway in model assumptions, don’t go too far. If model assumptions are grossly 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:

$$b_i \pm t_{df}^* SE_{b_i}$$

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

8.4 ANOVA and regression with categorical variables

Fitting and interpreting models using categorical variables as predictors is similar to what we have encountered in simple and multiple regression. However, there is a twist: a single categorical variable will have multiple corresponding parameter estimates. To be precise, if the variable has C categories, then there will be $C - 1$ parameter estimates. Furthermore, it is not appropriate to use a Z or T score to determine the significance of the categorical variable as a predictor unless it only has $C = 2$ levels.

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called F . ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable:

H_0 : The mean outcome is the same across all categories. In statistical notation, $\mu_1 = \mu_2 = \dots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i .

H_A : The mean of the outcome variable is different for some (or all) groups.

These hypotheses are used to evaluate a model of the form

$$y_{i,j} = \mu_i + \epsilon_j \tag{8.19}$$

where an observation $y_{i,j}$ belongs to group i and has error ϵ_j . Generally we make three assumptions in applying this model:

- the errors are independent,
- the errors are nearly normal, and
- the errors have nearly constant variance.

These conditions probably look familiar: they are the same conditions we used for multiple regression. When these three assumptions are reasonable, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

TIP: Level, category, and group are synonyms

We sometimes call the levels of a categorical variable its categories or its groups.

- **Example 8.20** College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences in first exam scores in these three classes (A , B , and C). Describe how the model and hypotheses above could be used to determine whether there are any differences between the three classes.

The hypotheses may be written in the following form:

H_0 : The average score is identical in all lectures. Any observed difference is due to chance. Notationally, we write $\mu_A = \mu_B = \mu_C$.

H_A : The average score varies by class. We would reject the null hypothesis in favor of this hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

We could label students in the first class as $y_{A,1}$, $y_{A,2}$, $y_{A,3}$, and so on. Students in the second class would be labeled $y_{B,1}$, $y_{B,2}$, etc. And students in the third class: $y_{C,1}$, $y_{C,2}$, etc. Then we could estimate the true averages (μ_A , μ_B , and μ_C) using the group averages: \bar{y}_A , \bar{y}_B , and \bar{y}_C .

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. We will soon learn that assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

- **Example 8.21** Examine Figure 8.11. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do these differences appear to be due to chance?

Any real difference in the means of groups I, II, and III is difficult to discern, because the data within each group are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in the centers of groups IV, V, and VI. For instance, group IV appears to have a lower mean than that of the other two groups. Investigating groups IV, V, and VI, we see the differences in the groups' centers are noticeable because those differences are large *relative to the variability in the individual observations within each group*.

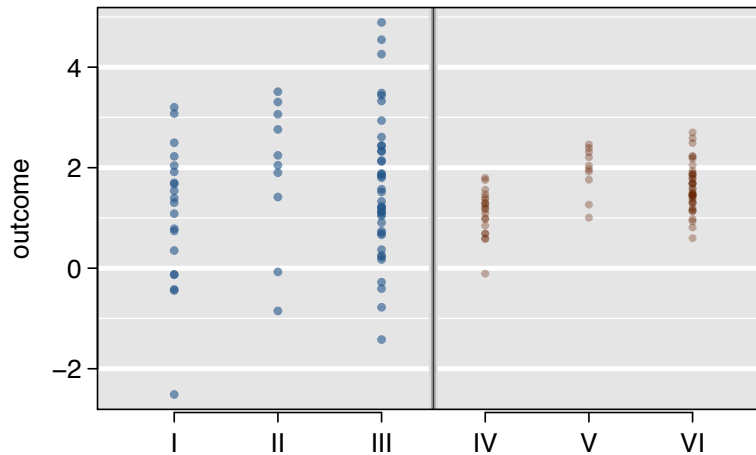


Figure 8.11: Side-by-side dot plot for the outcomes for six groups.

8.4.1 Is batting performance related to player position in MLB?

We would like to discern whether there are real differences between the batting performance of baseball players according to their position: outfielder (OF), infielder (IF), designated hitter (DH), and catcher (C). We will use a data set called `mlbBat10`, which includes batting records of 327 Major League Baseball (MLB) players from the 2010 season. Six of the 327 cases represented in `mlbBat10` are shown in Table 8.12, and descriptions for each variable are provided in Table 8.13. The measure we will use for the player batting performance (the outcome variable) is on-base percentage (OBP). The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

	name	team	position	AB	H	HR	RBI	AVG	OBP
1	I Suzuki	SEA	OF	680	214	6	43	0.315	0.359
2	D Jeter	NYY	IF	663	179	10	67	0.270	0.340
3	M Young	TEX	IF	656	186	21	91	0.284	0.330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
325	B Molina	SF	C	202	52	3	17	0.257	0.312
326	J Thole	NYM	C	202	56	3	17	0.277	0.357
327	C Heisey	CIN	OF	201	51	8	21	0.254	0.324

Table 8.12: Six cases from the `mlbBat10` data matrix.

- ⊙ **Exercise 8.22** The null hypothesis under consideration is the following: $\mu_{\text{OF}} = \mu_{\text{IF}} = \mu_{\text{DH}} = \mu_{\text{C}}$. Write the null and corresponding alternative hypotheses in plain language. Answers in the footnote¹⁰.

¹⁰ H_0 : The average on-base percentage is equal across the four positions. H_A : The average on-base percentage varies across some (or all) groups.

variable	description
name	Player name
team	The player's team, where the team names are abbreviated
position	The player's primary field position (OF, IF, DH, C)
AB	Number of opportunities at bat
H	Number of hits
HR	Number of home runs
RBI	Number of runs batted in
batAverage	Batting average, which is equal to H/AB

Table 8.13: Variables and their descriptions for the `mlbBat10` data set.

- **Example 8.23** The player positions have been divided into four groups: outfield (OF), infield (IF), designated hitter (DH), and catcher (C). What would be an appropriate point estimate of the batting average by outfielders, μ_{OF} ?

A good estimate of the batting average by outfielders would be the sample average of `batAverage` for just those players whose position is outfield: $\bar{y}_{OF} = 0.334$.

Table 8.14 provides summary statistics for each group. A side-by-side box plot for the batting average is shown in Figure 8.15. Notice that the variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied before we consider the ANOVA approach.

	OF	IF	DH	C
Sample size (n_i)	120	154	14	39
Sample mean (\bar{y}_i)	0.334	0.332	0.348	0.323
Sample SD (s_i)	0.029	0.037	0.036	0.045

Table 8.14: Summary statistics of on-base percentage, split by player position.

- **Example 8.24** The largest difference between the sample means is between the designated hitter and the catcher positions. Consider again the original hypotheses:

$$H_0: \mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_C$$

H_A : The average on-base percentage (μ_i) varies across some (or all) groups.

Why might it be inappropriate to run the test by simply estimating whether the difference of μ_{DH} and μ_C is statistically significant at a 0.05 significance level?

The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally we would pick the groups with the large differences for the formal test, leading to an unintentional inflation in the Type 1 Error rate. To understand this better, let's consider a slightly different problem.

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably

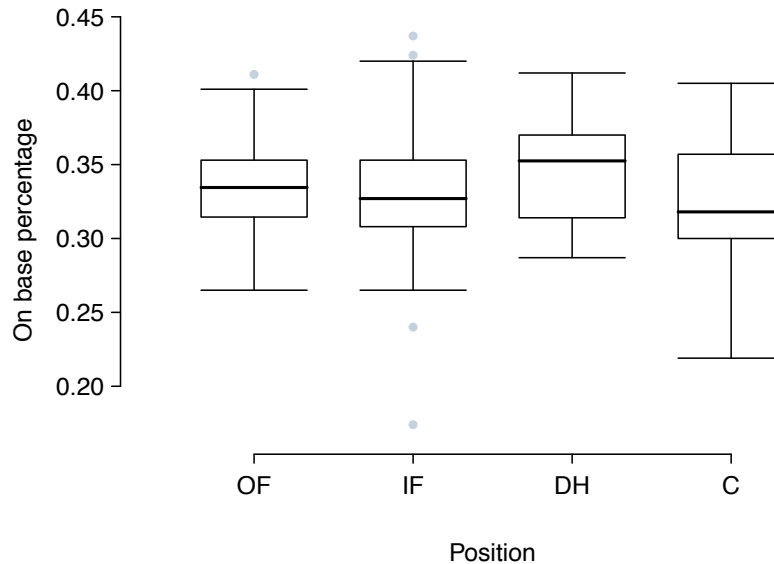


Figure 8.15: Side-by-side box plot of the on-base percentage for 327 players across four groups.

observe a few groups that look rather different from each other. If we select only these classes that look so different, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison.

For additional reading on the ideas expressed in Example 8.24, we recommend reading about the **prosecutor's fallacy**¹¹.

In the next section we will learn how to use the F statistic and ANOVA to test whether differences in means could have happened just by chance.

8.4.2 Analysis of variance (ANOVA) and the F test

The method of analysis of variance focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups** (MSG), and it has an associated degrees of freedom, $df_G = k - 1$ when there are k groups. The MSG is sort of a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations

¹¹See, for example, http://www.stat.columbia.edu/~cook/movabletype/archives/2007/05/the_prosecutors.html.

are provided in the footnote¹², however, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute the mean of the squared errors, often abbreviated as the **mean square error** (MSE), which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability of the residuals. Details of the computations of the MSE are provided in the footnote¹³ for the interested reader.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE :

$$F = \frac{MSG}{MSE} \quad (8.25)$$

The MSG represents a measure of the between-group variability, and MSE the variability within each of the groups.

- ⊙ **Exercise 8.26** For the baseball data, $MSG = 0.00252$ and $MSE = 0.00127$. Identify the degrees of freedom associated with each mean square and verify the F statistic is 1.994.

We use the F statistic to evaluate the hypotheses in what is called an **F test**. We compute a p-value from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2 . For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. An F distribution with 3 and 323 degrees of freedom, corresponding to the F statistic for the baseball hypothesis test, is shown in Figure 8.16.

The larger the observed variability in the sample means (MSG) relative to the residuals (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

¹²Let \bar{y} represent the mean of outcomes across all groups. Then the mean square between groups is computed as

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

where SSG is called the **sum of squares between groups** and n_i is the sample size of group i .

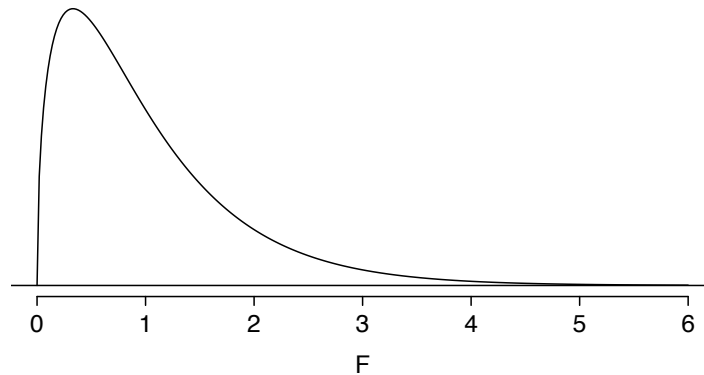
¹³Let \bar{y} represent the mean of outcomes across all groups. Then the **sum of squares total** (SST) is computed as

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

where the sum is over all observations in the data set. Then we compute the **sum of squared errors** (SSE) in one of three equivalent ways:

$$\begin{aligned} SSE &= SST - SSG \\ &= (n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2 + \dots + (n_k - 1) * s_k^2 \\ &= \sum_{j=1}^n e_j^2 \end{aligned}$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i , and the last expression represents the sum of the squared residuals across all groups. Then the MSE is the standardized form of SSE : $MSE = \frac{1}{df_E} SSE$.

Figure 8.16: An F distribution with $df_1 = 3$ and $df_2 = 323$.**The F statistic and the F test**

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F , which represents a standardized ratio of variability in the sample means relative to the variability of the residuals. If H_0 is true and the model assumptions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

- ⊙ **Exercise 8.27** The test statistic for the baseball example is $F = 1.994$. Shade the area corresponding to the p-value in Figure 8.16.
- **Example 8.28** The p-value corresponding to the solution for Exercise 8.27 is equal to about 0.115. Does this provide strong evidence against the null hypothesis?

The p-value is larger than 0.05, indicating the evidence is not sufficiently strong to reject the null hypothesis at a significance level of 0.05. That is, the data do not provide strong evidence that the average on-base percentage varies by player's primary field position.

8.4.3 Reading regression and ANOVA output from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. For these reasons it is common to use a statistical software to calculate the F statistic and p-value.

An ANOVA can be summarized in a table very similar to that of a regression summary. Table 8.17 shows an ANOVA summary to test whether the mean of on-base percentage varies by player positions in the MLB.

- ⊙ **Exercise 8.29** Earlier you verified that the F statistic for this analysis was 1.994, and the p-value of 0.115 was provided. Circle these values in Table 8.17 and notice the corresponding column name. Notice that both of these values are in the row labeled *position*, which corresponds to the categorical variable representing the player position variable.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	3	0.0076	0.0025	1.9943	0.1147
Residuals	323	0.4080	0.0013		

$s_{pooled} = 0.036$ on $df = 323$

Table 8.17: ANOVA summary for testing whether the average on-base percentage differs across player positions.

- ⊙ **Exercise 8.30** The $s_{pooled} = 0.036$ on $df = 323$ describes the estimated standard deviation associated with the residuals. Verify that s_{pooled} equals the square root of the *MSE* for the *Residuals* row.

8.4.4 Graphical diagnostics for an ANOVA analysis

There are three primary conditions we must check for an ANOVA analysis, all related to the residuals (errors) associated with the model. Recall that we assume the errors are independent, nearly normal, and have nearly constant variance across the groups.

Independence. If observations are collected in a particular order, we should plot the residuals in the order the corresponding observations were collected (e.g. see Figure 8.9 on page 319). For the baseball data, the data were collected from a sorted table, making such a review impossible. However, we can consider the nature of the data: Do we have reason to believe players are not independent? There are not obvious reasons why independence should not hold, so we will assume independence is reasonable in lieu of being able to examine this condition using data.

Approximately normal. The normality assumption for the residuals is especially important when the sample size is quite small. Figure 8.18 shows a normal probability plot for the residuals from the baseball data. We do see some deviation from normality at the low end, where there is a longer tail than what we would expect if the residuals were truly normal. While we should report this finding with the results of the hypothesis test, this slight deviation probably has little impact on the test results since there are so many players included in the sample and they are not spread thinly across many groups.

Constant variance. The last assumption is that the variance associated with the residuals is nearly constant from one group to the next. This assumption can be checked by examining a side-by-side box plot of the outcomes, as in Figure 8.15. In this case, the variability is similar in the four groups but not identical. We see in Table 8.14 on page 324 that the standard deviation varies a bit from one group to the next. Whether these differences are from natural variation is unclear, so we should report this uncertainty with the final results.

Caution: Diagnostics for an ANOVA analysis

Independence is always important to an ANOVA analysis. The normality condition is very important when the sample sizes for each group are relatively small. The constant variance condition is especially important when the sample sizes differ between groups.

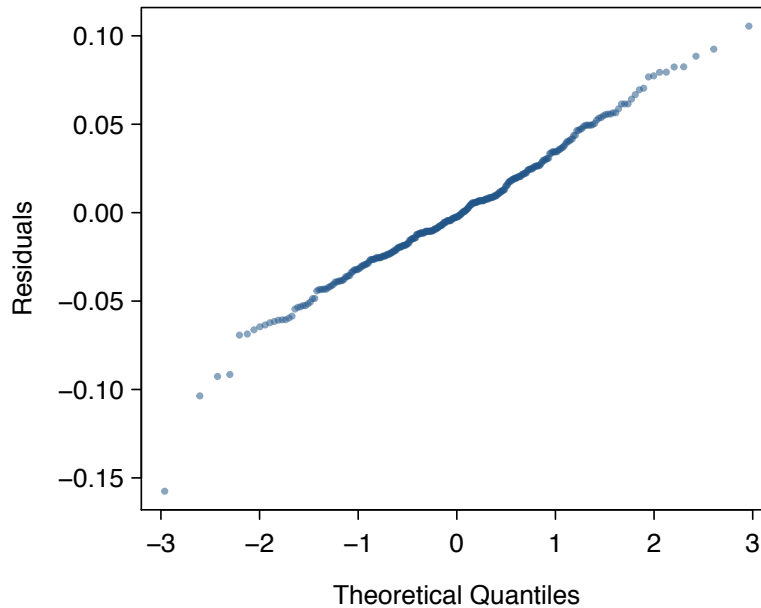


Figure 8.18: Normal probability plot of the residuals.

8.4.5 Multiple comparisons and controlling Type 1 Error rate

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3. These comparisons can be accomplished using a two-sample t test, but we must use a modified significance level and a pooled estimate of the standard deviation across groups.

- **Example 8.31** Example 8.20 on page 322 discussed three statistics lectures, all taught during the same semester. Table 8.19 shows summary statistics for these three courses, and a side-by-side box plot of the data is shown in Figure 8.20. We would like to conduct an ANOVA for these data. Do you see any deviations from the three conditions for ANOVA?

In this case (like many others) it is difficult to check independence in a rigorous way. Instead, the best we can do is use common sense to consider reasons the assumption of independence may not hold. For instance, the independence assumption may not be reasonable if there is a star teaching assistant that only half of the students may access; such a scenario would divide a class into two subgroups. After carefully considering the data, we believe that assuming independence may be acceptable.

The distributions in the side-by-side box plot appear to be roughly symmetric and show no noticeable outliers.

The box plots show approximately equal variability, which can be verified in Table 8.19, supporting the constant variance assumption.

Class i	A	B	C
n_i	58	55	51
\bar{y}_i	75.1	72.0	78.9
s_i	13.9	13.8	13.1

Table 8.19: Summary statistics for the first midterm scores in three different lectures of the same course.

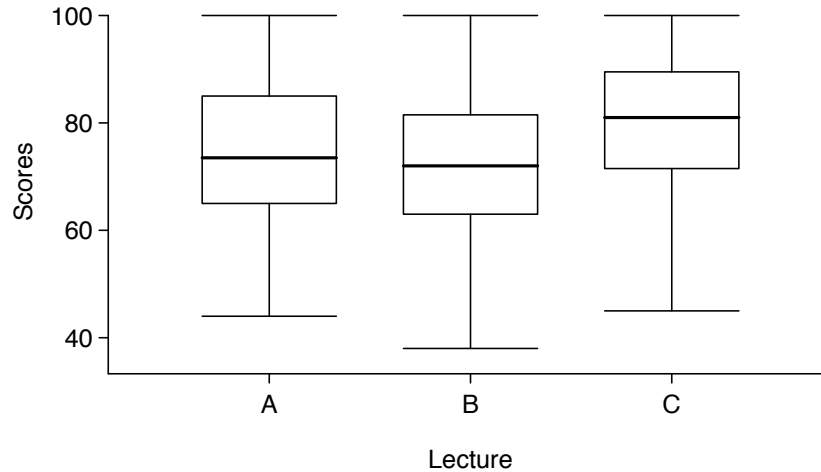


Figure 8.20: Side-by-side box plot for the first midterm scores in three different lectures of the same course.

- ⊙ **Exercise 8.32** An ANOVA was conducted for the midterm data, and a summary is shown in Table 8.21. What should we conclude?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture	2	1290.11	645.06	3.48	0.0330
Residuals	161	29810.13	185.16		

$s_{pooled} = 13.61$ on $df = 161$

Table 8.21: ANOVA summary table for the midterm data.

There is strong evidence that the different means in each of the three classes is not simply due to chance. We might wonder, which of the classes are actually different? As discussed in earlier chapters, a two-sample t test could be used to test for differences in each possible pair of groups. However, one pitfall was discussed in Example 8.24 on page 324: when we run so many tests, the Type 1 Error rate increases. This issue is resolved by using a modified significance level.

Multiple comparisons and the Bonferroni correction for α

The scenario of testing many pairs of groups is called **multiple comparisons**. The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha/K$$

where K is the number of comparisons being considered (formally or informally). If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

- **Example 8.33** In Exercise 8.32, you found that the data showed strong evidence of differences in the average midterm grades between the three lectures. Complete the three possible pairwise comparisons using the Bonferroni correction and report any differences.

We use a modified significance level of $\alpha^* = 0.05/3 = 0.0167$. Additionally, we use the pooled estimate of the standard deviation: $s_{pooled} = 13.61$ on $df = 161$.

Lecture A versus Lecture B: The estimated difference and standard error are, respectively,

$$\bar{y}_A - \bar{y}_B = 75.1 - 72 = 3.1 \quad SE = \sqrt{\frac{13.61^2}{58} + \frac{13.61^2}{55}} = 2.56$$

(See Section 6.2.4 on page 6.2.4 for additional details.) This results in a T score of 1.21 on $df = 161$ (we use the df associated with s_{pooled}) and a two-tailed p-value of 0.228. This p-value is larger than $\alpha^* = 0.0167$, so there is not strong evidence of a difference in the means of lectures A and B.

Lecture A versus Lecture C: The estimated difference and standard error are 3.8 and 2.61, respectively. This results in a T score of 1.46 on $df = 161$ and a two-tailed p-value of 0.1462. This p-value is larger than α^* , so there is not strong evidence of a difference in the means of lectures A and C.

Lecture B versus Lecture C: The estimated difference and standard error are 6.9 and 2.65, respectively. This results in a T score of 2.60 on $df = 161$ and a two-tailed p-value of 0.0102. This p-value is smaller than α^* . Here we find strong evidence of a difference in the means of lectures B and C.

We might summarize the findings of the analysis from Example 8.33 using the following notation:

$$\mu_A \stackrel{?}{=} \mu_B \quad \mu_A \stackrel{?}{=} \mu_C \quad \mu_B \neq \mu_C$$

The midterm mean in lecture A is not statistically distinguishable from those of lectures B or C. However, there is strong evidence that lectures B and C are different. In the first two pairwise comparisons, we did not have sufficient evidence to reject the null hypothesis. Recall that failing to reject H_0 does not imply H_0 is true.

Caution: Sometimes an ANOVA will reject the null but no groups will have statistically significant differences

It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, *this does not invalidate the ANOVA conclusion*. It only means we have not been able to successfully identify which groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high confidence a specific difference as statistically significant is more difficult.

Consider the following analogy: we observe a Wall Street firm that makes large quantities of money based on predicting mergers. Mergers are generally difficult to predict, and if the prediction success rate is extremely high, that may be considered sufficiently strong evidence to warrant investigation by the Securities and Exchange Commission (SEC). While the SEC may be quite certain that there is insider trading taking place at the firm, the evidence against any single trader may not be very strong. It is only when the SEC considers all the data that they identify the pattern. This is effectively the strategy of ANOVA: stand back and consider all the groups simultaneously.

8.4.6 Using ANOVA for multiple regression

The ANOVA methodology can be extended to multiple regression, where we simultaneously incorporate categorical and numerical predictors into a model. The methods discussed so far – an outcome for a single categorical variable – is called **one-way ANOVA**. There are two extensions that we briefly discuss here: evaluating all variables in a model simultaneously, and using ANOVA in model selection where some variables are numerical and others categorical.

Some software will supply additional information about a multiple regression model fit beyond the regression summaries described in this textbook. This additional information can be used in an assessment of the utility of the full model. For instance, below is the full regression summary for the Mario Kart Wii game analysis from Section 8.2 (implemented with R statistical software¹⁴) using all four predictors:

```
Residuals:
      Min       1Q   Median       3Q      Max
-11.3788  -2.9854  -0.9654   2.6915  14.0346

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.21097    1.51401  23.917 < 2e-16 ***
condNew      5.13056    1.05112   4.881 2.91e-06 ***
stockPhoto   1.08031    1.05682   1.022  0.308
duration    -0.02681    0.19041  -0.141  0.888
wheels       7.28518    0.55469  13.134 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

¹⁴R is free and can be downloaded at www.r-project.org.

```
Residual standard error: 4.901 on 136 degrees of freedom
Multiple R-squared: 0.719, Adjusted R-squared: 0.7108
F-statistic: 87.01 on 4 and 136 DF, p-value: < 2.2e-16
```

The main output labeled `Coefficients` should be familiar as the multiple regression summary. The last three lines are new and provide details about

- the standard deviation associated with the residuals (4.901),
- degrees of freedom (136),
- R^2 (0.719) and adjusted R^2 (0.7108), and
- also an F statistic (174.4 with $df_1 = 4$ and $df_2 = 136$) with an associated p-value ($< 2.2e-16$, i.e. about zero).

The F statistic and p-value in the last line can be used for a test of the entire model. The p-value can be used to answer the following question: Is there strong evidence that the model as a whole is significantly better than using no variables? In this case, with a p-value of less than 2.2×10^{-16} , there is extremely strong evidence that the variables included are helpful in prediction. Notice that the p-value does not verify that all variables are actually important in the model; it only considers the importance of all of the variables simultaneously. This is similar to how ANOVA was earlier used to assess differences across all means without saying anything about the difference between a particular pair of means.

The second setting for ANOVA in the general multiple regression framework is one that is more delicate: model selection. We could compare the variability in the residuals of two models that differ by just one predictor using ANOVA as a tool to evaluate whether the data support the inclusion of that variable in the model. We postpone further details of this method to a later course.

8.5 Exercises

8.5.1 Introduction to multiple regression

8.1 In Chapter 6 you were introduced to a data set from an experiment to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types in the data set called `chickwts`. We are specifically interested in the effect of casein feed on the weights of these chicks, so we have created a variable called `casein` and coded chicks who were on casein feed as 1 and those who were on other diets as 0. The summary table below shows the results of a simple linear regression model for predicting `weight` from `casein`. [38]

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	248.64	9.54	26.06	0.0000
casein	74.94	23.21	3.23	0.0019

- Write the equation of the regression line.
- Interpret the slope in context, and calculate the predicted weight of chicks who are and who not are on another feed.
- Is there a statistically significant relationship between feed type (casein or other) and the average weight of chicks? State the hypotheses and include any information used to conduct the test. Note that if we look back at Exercise 6.19 on page 268, we would see that the variability within the casein group and the variability across the other groups are about equal and the distributions symmetric. With these conditions satisfied, it is reasonable to proceed with the test. (Note also that we don't need to check linearity since the predictor has only two levels.)

8.2 Vitamin C is believed to help promote dental health. One common way to get Vitamin C is by drinking orange juice. Another option is to take ascorbic acid tablets. An experiment was conducted to test if one source is more effective than the other. 60 guinea pigs were randomly assigned to these two delivery methods for Vitamin C, 30 in each group. The length of teeth in millimeters are given along with delivery methods in the data set called `ToothGrowth`. We created a variable called `OJ` and coded guinea pigs who were given orange juice as 1 and those who were given ascorbic acid as 0. The summary table below shows the results of a simple linear regression model for predicting the average tooth length, `len`, from `OJ`. [51]

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	16.96	1.37	12.42	0.0000
oj	3.70	1.93	1.92	0.0604

- Write the equation of the regression line.
- Interpret the slope in context, and calculate the predicted tooth length for guinea pigs who were given orange juice and those who were given ascorbic acid.
- Is there a statistically significant relationship between the average tooth length and delivery method of Vitamin C in guinea pigs? State the hypotheses and include any information used to conduct the test. Note that the variability within the orange juice and the ascorbic acid groups are about equal and the distributions symmetric. With these conditions satisfied, it is reasonable to proceed with the test.

8.3 The Child Health and Development Studies (CHDS) is a collection of studies, one of which considers all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. A random sample of these data are given in a data set called `babies`. We consider 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 simple linear regression model for predicting the average birth weight of babies, measured in ounces (`bwt`), from `smoke`. [52]

	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 symmetric. With these conditions satisfied, it is reasonable to proceed with the test. (Note that we don't need to check linearity since the predictor has only two levels.)

- Write the equation of the regression line.
- Interpret the slope in context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.
- Is there a statistically significant relationship between the average birth weight and smoking? State the hypotheses and include any information used to conduct the test.

8.4 Exercise 8.3 introduces a data set on birth weight of babies. Another variable we consider is `parity`, where 0 is first born, and 1 is otherwise. The summary table below shows the results of a simple 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

- Write the equation of the regression line.
- Interpret the slope in context, and calculate the predicted birth weight of first borns and others.
- Is there a statistically significant relationship between the average birth weight and parity? State the hypotheses and include any information used to conduct the test.

8.5 The `babies` dataset used in Exercises 8.3 and 8.4 includes information on 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`), in addition to the `smoking` and `parity` variables considered earlier. 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
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
1236	117	297	0	38	65	129	0

The summary table below shows the results of a linear 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

- Write the equation of the regression line that includes all of the variables.
- Interpret the slopes of `gestation` and `age` in context.
- The coefficient for `parity` is different than in the simple linear model shown in Exercise 8.4. Why might there be a difference?
- Calculate the residual for the first observation in the data set.

8.6 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 in a particular school year. These data are given in a data set called `quine`. Below are three observations from this data set.

	eth	sex	lrn	days
1	0	1	1	2
2	0	1	1	11
\vdots	\vdots	\vdots	\vdots	\vdots
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). [53]

	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 context.
- (c) Calculate the residual for the first observation in the data set.

8.7 The variance of the residuals for the model given in Exercise 8.5 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 1236 observations in the data set.

8.8 The variance of the residuals for the model given in Exercise 8.6 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.

8.5.2 Model selection

8.9 Exercise 8.5 presents summary output for a regression model for predicting the average birth weight of babies based on six explanatory variables.

- (a) Determine which variable(s) do not have a significant relationship with the outcome and should be candidates for removal from the model. If there is more than one such model, indicate which one should be removed first.
- (b) The summary table below shows the results of the regression we refit after removing `age` from the model. Determine if any other variable(s) should be removed from the model.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-80.64	14.04	-5.74	0.0000
gestation	0.44	0.03	15.28	0.0000
parity	-3.29	1.06	-3.10	0.0020
height	1.15	0.20	5.64	0.0000
weight	0.05	0.03	2.00	0.0459
smoke	-8.38	0.95	-8.82	0.0000

8.10 Exercise 8.6 presents summary output for a regression model for predicting the average number of days absent based on three explanatory variables.

- (a) Determine which variable(s) do not have a significant relationship with the outcome and should be candidates for removal from the model. If there is more than one such model, indicate which one should be removed first.
- (b) The summary table below shows the results of the regression we refit after removing learner status from the model. Determine if any other variable(s) should be removed from the model.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	19.98	2.22	9.01	0.0000
eth	-9.06	2.60	-3.49	0.0006
sex	2.78	2.60	1.07	0.2878

8.11 Exercise 8.5 provides regression output for the full model (including all explanatory variables available in the data set) for predicting birth weight of babies. In this exercise, we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows the p-value and adjusted R^2 of each model where we include only the corresponding predictor. Based on this table, which variable should be added to the model first?

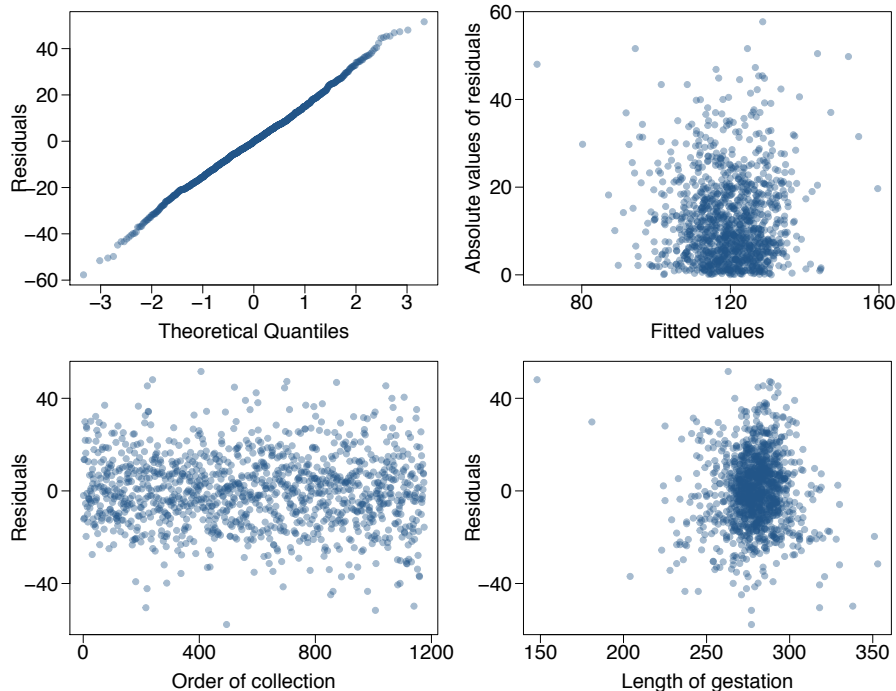
variable	gestation	parity	age	height	weight	smoke
p-value	2.2×10^{-16}	0.1052	0.2375	2.97×10^{-12}	8.2×10^{-8}	2.2×10^{-16}
R^2_{adj}	0.1657	0.0013	0.0003	0.0386	0.0229	0.0569

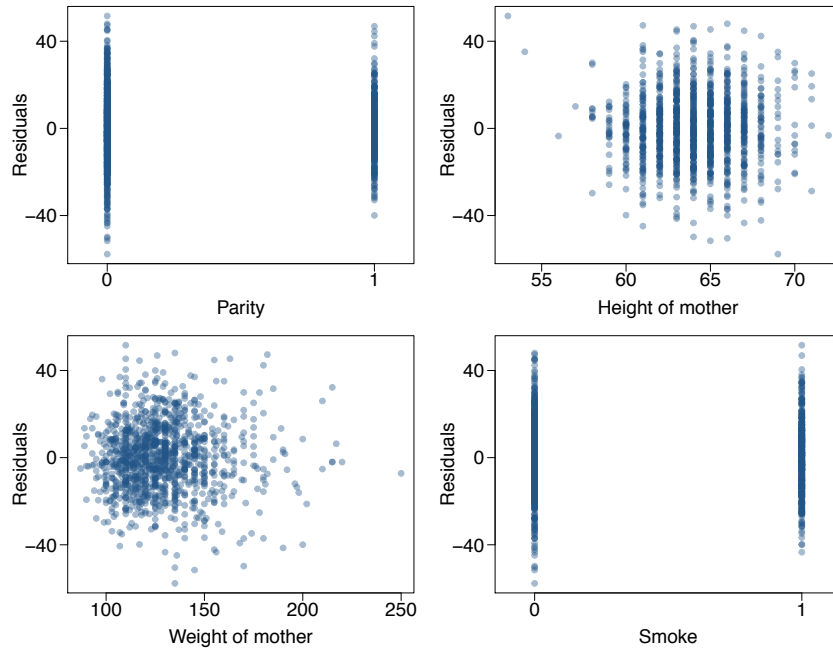
8.12 Exercise 8.6 provides regression output for the full model (including all explanatory variables available in the data set) for predicting number of days absent from school. In this exercise, we consider a forward-selection algorithm and add variables to the model one-at-a-time. The table below shows 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	leaner status
p-value	0.0007	0.3142	0.5870
R^2_{adj}	0.0714	0.0001	0

8.5.3 Checking model assumptions using graphs

8.13 Exercise 8.9 presents a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoke. Determine if the model assumptions are met using the plots below. If not, describe how to proceed with the analysis.

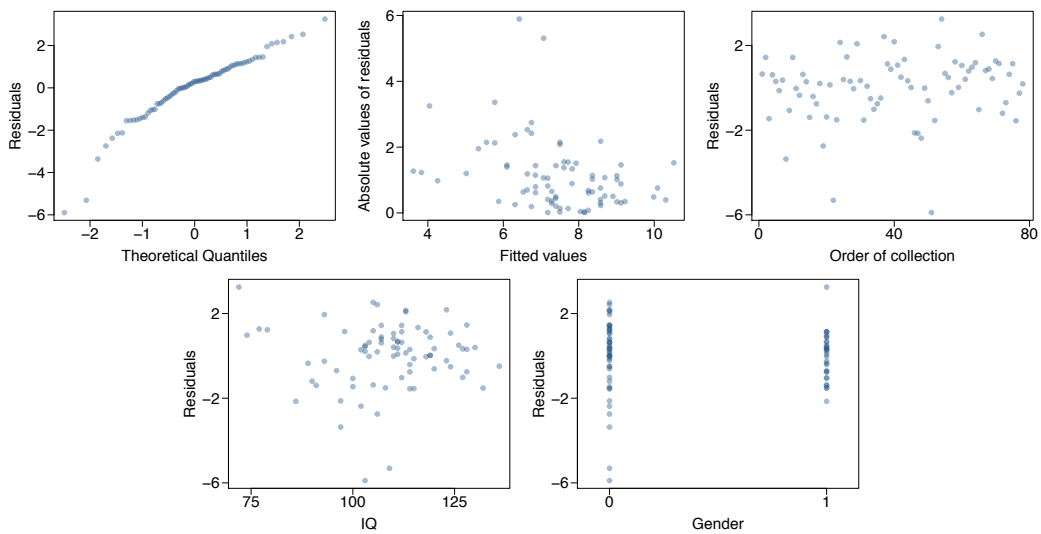




8.14 The table below presents summary output for a regression model for predicting the average GPA based IQ and gender, where 0 represents a female and 1 represents a male.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.70	1.56	-3.01	0.0035
iq	0.11	0.01	7.77	0.0000
gender	0.97	0.37	2.60	0.0111

The p-values in this table suggest a significant relationship between GPA and the predictors, IQ and gender. Using the plots given below, determine if this regression model is appropriate for these data.



8.5.4 ANOVA and regression with categorical variables

8.15 In Exercise 8.1, we considered the effect of casein feed on chicks' weight. Instead of categorizing feed type as casein or other, we might also want to consider all feed types at once: casein, horsebean, linseed, meat meal, soybean, and sunflower. The ANOVA output below can be used to test for differences between the average weights of chicks on different diets.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	231129.16	46225.83	15.36	0.0000
Residuals	65	195556.02	3008.55		

Conduct a hypothesis test to determine if these data provide strong evidence that the average weight of chicks varies across some (or all) groups. Refer to Exercise 6.19 on page 268 to assist in checking ANOVA conditions.

8.16 A professor who teaches a large introductory statistics class with eight discussion sections would like to test if student performance differs by discussion section. Each discussion section has a different teaching assistant. The summary table below shows the average final exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.

	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
n_i	33	19	10	29	33	10	32	31
\bar{x}_i	92.94	91.11	91.80	92.45	89.30	88.30	90.12	93.35
s_i	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57

The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
section	7	525.01	75.00	1.87	0.0767
Residuals	189	7584.11	40.13		

Conduct a hypothesis test to determine if these data provide strong evidence that the average score varies across some (or all) groups. Check conditions and describe any assumptions you must make to conduct the test.