Contents

Chapter 8

Next steps in regression

Alternative titles: Advanced regresson / Multiple regression and ANOVA.

The field of regression begins with evaluating the connection between two numerical variables, one called the predictor and one called the outcome (or response). Analyses becomes more complex but also tend to better represent relationships between variables when many variables are simultaneously used to predict the outcome ~~variable.~~

8.1 Introduction to multiple regression

Multiple regression is the extension of the two-variable framework of Chapter **??** to the case where many predictor variables are used. The method is motivated by scenarios where many variables may be simultaneously connected to an output.

We will consider Ebay auctions for a game called *Mario Kart* for the Nintendo Wii. We will consider the total price of an auction – the highest bid plus the shipping cost – as the outcome variable. Naturally, both buyers and sellers on Ebay are especially conscientious of the total auction price. But how is the total price related to other characteristics of an auction? For instance, does having a longer auction tend to correspond to a higher or lower price? Or how much more do folks tend to pay for additional Wii wheels in auctions? Multiple regression will help us answer these and other questions.

We will use a new data set called `marioKart`, which includes records for 143 auctions for the game *Mario Kart Wii* on Ebay. 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: the condition variable and the variable describing whether a stock photo was used for the auction. Two-level categorical variables are often coded

	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**
`totalPr`	the total of the final auction price and the shipping cost, in US dollars
`condNew`	a coded two-level categorical variable, which takes value `1` when the game is new and `0` if the game is used
`stockPhoto`	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
`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 Wii)

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

into `0`s and `1`s, which allows them to be used in a regression model in the same way as a numerical predictor:

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

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

$$\widehat{\texttt{totalPr}} = 42.87 + 10.90 * \texttt{condNew} \tag{8.1}$$

The `0`-`1` coding of the two-level categorical variable allows us to interpret the coefficient of `condNew`. When the game is used, the `condNew` variable takes a value of zero, so the model predicts the auction will have a total price of \$42.87. If the game is new, then the `condNew` variable takes value one and the total price is predicted to be \$42.87 + \$10.90 = \$53.77. The coefficient of `condNew` estimates the difference in the total auction price when the game is new versus used.

TIP: The coefficient of a two-level categorical variable
The coefficient of a binary variable corresponds to the estimated difference in the outcome under the two possible levels of the variable.

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

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

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

● **Example 8.4** In Exercise 8.2, you found that auctions whose primary photo was a stock photo tend 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. Does this mean that her auction will sell for about \$4.17 more than it otherwise would have if she used a unique photo?

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 predictor variables have an underlying structure. For instance, new games sold on Ebay tend to come with more ~~more~~ Wii wheels, leading 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 connection of a predictor variable with the outcome while correcting for the potential influence of other variables. This is the strategy used in **multiple regression**.

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:

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

where y represents the total price, x_1 the game's condition, x_2 whether a stock photo was utilized, x_3 indicates the duration of the auction, and x_4 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 variables. However, even while this linear model isn't perfect, we might find that it fits the data reasonably well.

We estimate the parameters β_0, β_1, ..., β_4 in the same way as we did in the case of a single predictor, by minimizing the sum of the squared errors (residuals):

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

We typically use a computer to minimize this sum and provide the estimates $\hat{\beta}_i$. Sample output is shown in Table 8.3. Using this output, we identify the point estimates 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: The 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 β_i using a computer.

⊙ **Exercise 8.7** Write out the model in Equation (8.5) using the point estimates from Table 8.3. What is p for this model? Answers in the footnote[1].

⊙ **Exercise 8.8** What does β_4, the coefficient of the x_4 variable (Wii wheels), represent? Answer in the footnote[2].

⊙ **Exercise 8.9** Compute the residual of the first observation in Table 8.1 on page 3. Hint: use the equation from Exercise 8.7: $\hat{y} = 36.21 + 5.13x_1 + 1.08x_2 - 0.03x_3 + 7.29x_4$. Answer in the footnote[3].

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

[2] It is the average difference in auction price for each additional Wii wheel included.

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

● **Example 8.10** The coefficients for x_1 (`condNew`) and x_2 (`stockPhoto`) are different than in the two separate models in Equations (8.1) and (8.3). Why might that be?

If we examined the data carefully, we would see that some predictors are correlated with each other. For instance, many auctions selling a new game also used a stock photo. When we looked at only one variable, such as `stockPhoto`, the predictor was also representing the lurking variable that was missing in the model. When we use both variables, this underlying and unintentional bias is reduced or eliminated, though there might be other variables that we have not taken into account.

Example 8.10 describes a common issue in multiple regression: correlation in predictor variables. We say the two predictor variables are **collinear** when they are correlated, and this collinearity complicates model estimation.

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

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

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

where $\hat{\epsilon}_i$ represents the residuals of the model and y_i the outcomes. This equation remains valid in the multiple regression framework.

⊙ **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 $\mathbf{R^2}$ as a tool for model assessment
The **adjusted $\mathbf{R^2}$** is computed as

$$R^2_{adj} = 1 - \frac{Var(\hat{\epsilon}_i)/(n-p-1)}{Var(y_i)/(n-1)} = 1 - \frac{Var(\hat{\epsilon}_i)}{Var(y_i)}\frac{n-1}{n-p-1}$$

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 variance[4].

⊙ **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 $R^2_{adj} = 0.711$ for the Mario Kart model.

8.2 Model selection

The best model is not always the largest. Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions. Additionally, collecting data can be expensive, so why spend money on collecting and reporting unimportant variables?

In this section we discuss model selection strategies, which will help us eliminate variables that are less important from the model. Next section we will assess whether the underlying assumptions for the fitted model are satisfied.

8.2.1 Using regression output to evaluate variable inclusion

Table 8.4 shows the summary of the parameter estimates in the model estimating total auction price based on four predictor variables. The last column of the table lists p-values that can be used to assess hypotheses of the following form:

H_0: $\beta_i = 0$ and the other parameters are included in the model.

H_A: $\beta_i \neq 0$ and the other parameters are included in the model.

The p-values provided in Table 8.4 can be used to assess the hypotheses above for each variable in the model. If there is not strong evidence favoring the alternative hypothesis for a coefficient, we should consider eliminating the corresponding variable from 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.

[4] In multiple regression, the degrees of freedom associated with the variance estimate of the residuals is $n-p-1$, not $n-1$. 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.

● **Example 8.13** 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 interpretted?

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 connection to the total auction price.

● **Example 8.14** Is there strong evidence that using a stock photo is connected 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 has a connection to the total price of the auction. We should consider removing the `stockPhoto` variable from the model.

⊙ **Exercise 8.15** Identify the p-value for both the `duration` and `wheels` variables in the model. Is there strong evidence supporting the inclusion of these variables in the model?

There is not statistically significant evidence that either `stockPhoto` or `duration` are meaningful in the model. Next we consider common strategies for pruning such variables from a model.

8.2.2 Two model selection strategies

There are two **stepwise** strategies for adding or removing variables in a multiple regression model. They are called *backward-elimination* and *forward-selection*.

The **backward-elimination** strategy starts with the model that includes all potential predictor variables. Then, one-by-one, variables are eliminated from the model until all variables have corresponding p-values that are statistically significant. In each elimination step, we drop the variable with the largest p-value, refit the model, and reassess the inclusion of all variables.

● **Example 8.16** The *full model* for the `marioKart` data with the total price as the outcome is summarized 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 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:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_4 x_4 = 36.78 + 5.58 x_1 + 7.23 x_4$$

	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 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 predictor variables are highly correlated.

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 8.17** 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, we stop adding variables.

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.

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 what 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. If both strategies are tried and they arrive at different models, one might use another criteria to select between the two competing models, such as choosing the model with the larger adjusted R^2.

(I'm not certain the following is true – can anyone verify?) There is also no guarantee that the forward-selection strategy will result in a model where all included variables have coefficients that are statistically different from zero. For instance, the first variable added may no longer be statistically significant after adding in other variables.

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.

TIP: Sometimes keep variables even when the p-value > 0.05
If we are not interested in a particular variable's connection to the outcome and it is only included the model because it is a potential lurking variable, then it is okay to keep it even if the p-value is a little larger than 0.05. Consider using a significance level of 0.10 or 0.15 for such variables.

8.3 Checking model assumptions using graphs

Multiple regression models take the following form:

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

where the errors (residuals) are independent, and nearly normal with constant variance. Based on these assumptions, we have four conditions to check:

1. the residuals 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.

We check the four assumptions using four types of plots:

Normal probability plot. A normal probability plot of the residuals is shown in Figure 8.7. While the plot shows some minor irregularities, there are no outliers that might be cause for concern. In a normal quantile plot for residuals, we tend to be most concerned about residuals that appear to be outliers, since this indicates 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.

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. perhaps the final price of auctions tend to be higher during some times and so consecutive auctions would tend to have similar residuals. Here we see no structure[5].

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 the variability doesn't fluctuate across groups. However, when we consider the residuals against the `wheels` variable, we see structure. There appear to be curvature in the residuals, indicating the relationship is probably not linear.

[5] 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.

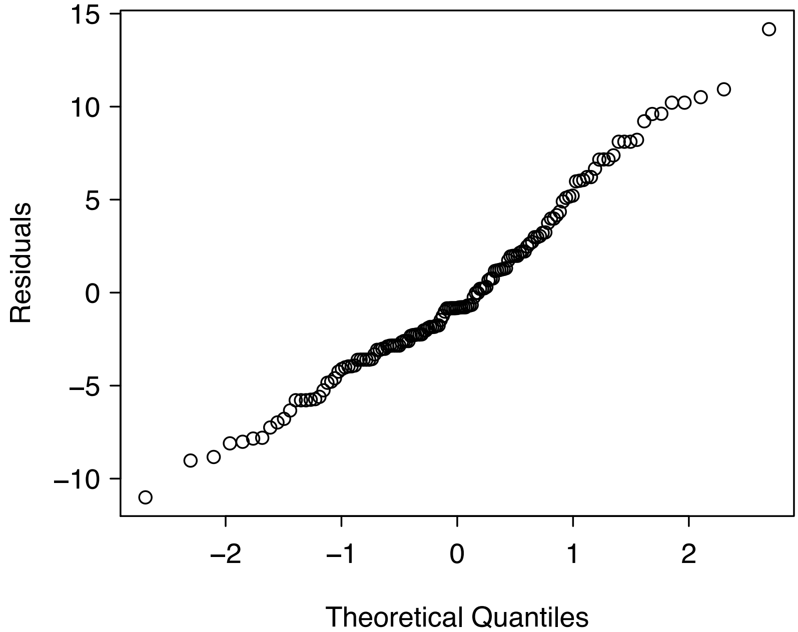


Figure 8.7: A normal quantile plot of the residuals is helpful in identifying observations that might be outliers.

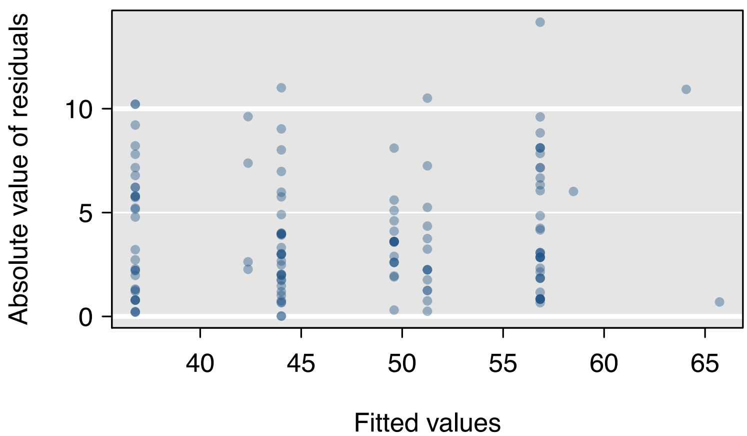


Figure 8.8: Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.

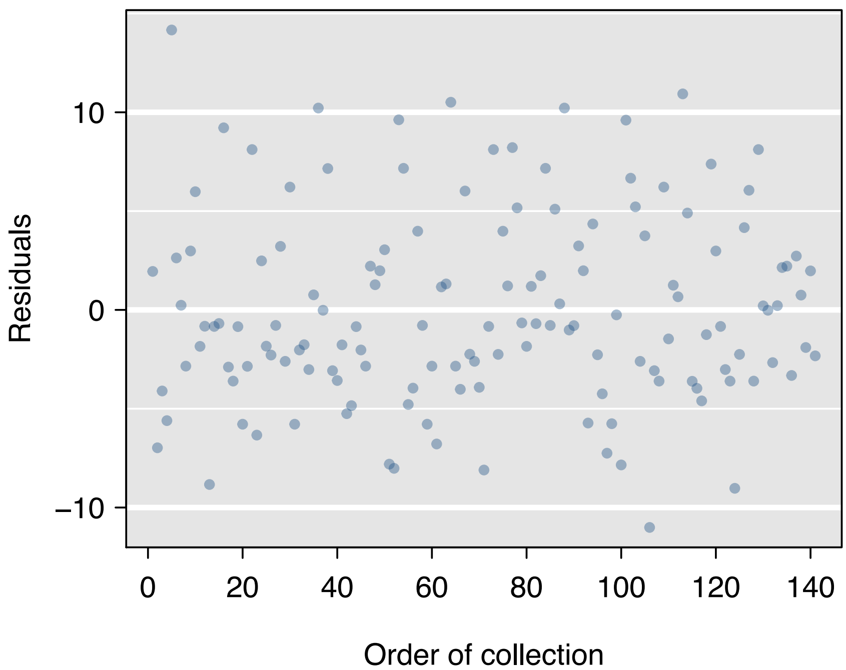


Figure 8.9: Plotting residuals in the order that their corresponding observations were collected helps identify connections between successive observations. If it seems that consecutive observations tend to be close to each other, this indicates the independence assumption of the observations would fail.

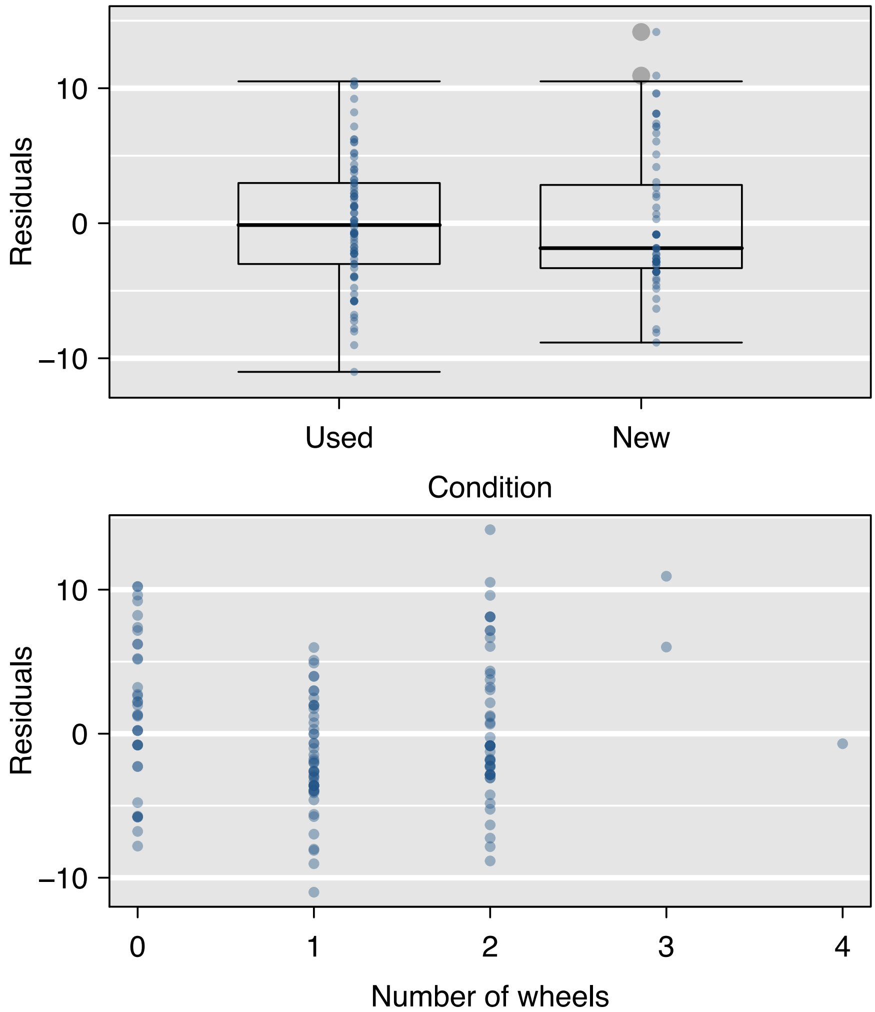


Figure 8.10: In the two-level variable for the game's condition, we check for differences in distribution shape or variability. For numerical predictors, we also check for trends or other structure. We see some slight bowing in the residuals against the `wheels` variable.

It is generally appropriate to summarize diagnostics for any model fit. If the diagnostics closely align with the model assumptions, we could report this to improve confidence 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, and to omit it 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 is often reasonable so long as we are upfront and report the model's shortcomings.

Caution: Don't report results where model assumptions are grossly violated
While there is a little leeway in model assumptions, don't go overboard. 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
Computing confidence intervals for coefficients in multiple regression uses the same formula as in the single predictor model:

$$\hat{\beta}_i \pm t^*_{df} SE_{\hat{\beta}_i}$$

where t^*_{df} is the appropriate t value corresponding to the confidence level and model degrees of freedom.

8.4 Transforming variables in a linear model

When we observe a nonlinear relationship between a predictor and outcome variable, we can sometimes try out transformations on either the predictor or the response variable. Generally we only attempt such transformations when a linear model is inadequate. However, before considering transformations, a word a caution.

Caution: Avoid transformations when possible
(1) Try fitting a linear model before considering any transformations unless it is very clear from the start that a transformation is necessary. (2) Limit the total number of variables in a model that have been transform. (3) Don't try transformations on any variables simply because those variables were eliminated in the model selection step. Never try out transformations without good reason.

One important purpose of model fitting is understanding how variables are connected. A model becomes more difficult to interpret or describe as soon as we begin transforming variables in the model. The complications continue expanding with the number of transformations. It is best to avoid transformations when possible.

Transformations that might be considered on predictor variables include the following:

Natural logarithm. If a few of observations are many times larger than the others (e.g. some observations are in the hundreds while most others are below 10), a natural logarithm may help restrain the impact of such large values. Note that the natural logarithm requires that all observations are greater than zero.

Square root. The square root is a helpful alternative to the natural logarithm when using count data or when observations take value zero.

Squaring a predictor. If there is curvature in the residuals against a predictor variable, then consider incorporating an additional variable into the model: the square of the predictor.

8.5 Regression with categorical variables

Fitting and interpreting models for categorical variables is very similar to what we have encountered in simple and multiple regression. We generally obtain estimates for model parameters and corresponding standard errors. However, one difference that a single categorical variable can have a parameter estimate and standard error associated with *each* of its levels, and it isn't obvious how to bring all these estimates together to assess the statistical significance for the categorical variable. (We may also call the levels of a category its *groups* or *categories*.) We need a new strategy, which is called **analysis of variance (ANOVA)** and is used to evaluate whether there are any statistically significant differences in the groups associated with the categorical variable. The ANOVA method will provide us with a new test statistic, called F, which is used to compute the p-value for a test and also a new summary table to assess the significance of variables.

We will consider the case where we have a single categorical variable as a predictor and the following hypotheses:

H_0 : The mean outcome is the same across the categories. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_j represents the mean for category j.

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

In effect we are evaluating the model

$$x_{i,j} = \mu_j + \epsilon_i$$

where observation $x_{i,j}$ belongs to group j and has error ϵ_i. Generally we assume the errors are independent and nearly normal, and we try to determine whether the data provide strong evidence against the null hypothesis that all the μ_j are equal.

● **Example 8.18** Examine Figure 8.11. Compare groups A, B, and C. Can you tell if the means for each group are actually different? Now compare groups D, E, and F. Can you observe any differences in the means of these three groups?

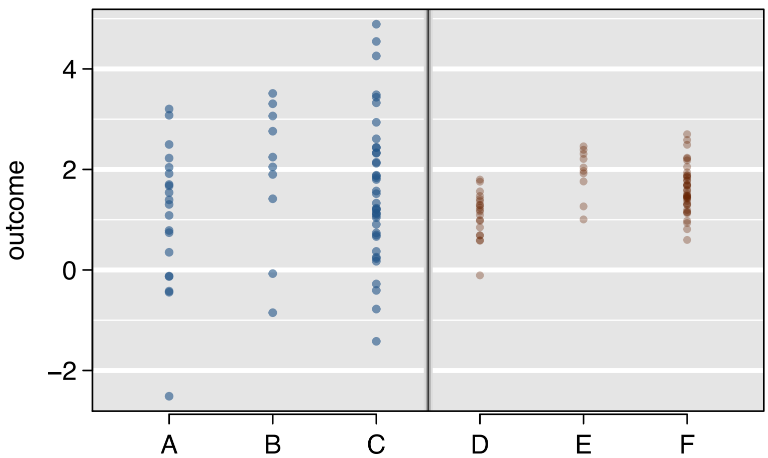


Figure 8.11: Side-by-side dot plot for the outcomes for six groups.

Any real difference in the means of groups A, B, and C is difficult to discern. The data are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in groups D, E, and F. For instance, group D appears to have a lower mean/center than that of the other two groups. In these last three groups, the difference in the groups' centers is noticeable because those differences are large *relative to the variability in the individual observations.*

8.5.1 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 baseball position (i.e. outfielder, infielder, catcher, and designated hitter). We will use a data set called `mlbBat10`, which includes batting records of 327 Major League Baseball players from the 2010

	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.

variable	**description**
`name`	Player name
`team`	Professional team, name abbreviated
`position`	The player's primary field position, where `OF` is for outfield, `IF` is for an infield position, `C` is for catcher, and `DH` is for designated hitter (i.e. doesn't play in the field).
`AB`	Number of at bats.
`H`	Hits.
`HR`	Home runs.
`RBI`	Runs batted in.
`batAverage`	Batting average, which is the proportion of at bats that the player gets a hit.

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

season. Six of the 327 cases represented in `mlbBat10` are shown in Table 8.12, and descriptions for each variable are in Table 8.13. The measure we will use for the player batting performance (the outcome variable) will be the on-base percentage (`OBP`). The on-base percentage roughly represents the fraction of the time a player gets on base without being called out.

⊙ **Exercise 8.19** The hypotheses under consideration are the following:

H_0 : $\mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{DH}} = \mu_{\texttt{C}}$
H_A : The average on-base percentage (μ_j) varies across some (or all) groups.

Write the null hypothesis in plain language. Answer in the footnote[6].

● **Example 8.20** The player positions have been broken 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, $\mu_{\texttt{OF}}$?

A good estimate of the batting average by outfielders would be the sample average of `batAverage` for just those players whose position is outfield: $\bar{x}_{OF} = 0.265$.

[6] H_0: The average on-base percentage is equal across the four positions.

	OF	IF	DH	C
Sample size (n_j)	120	154	14	39
Sample mean ($\bar{x}_j$)	0.334	0.332	0.348	0.323
Sample SD (s_j)	0.029	0.037	0.036	0.045

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

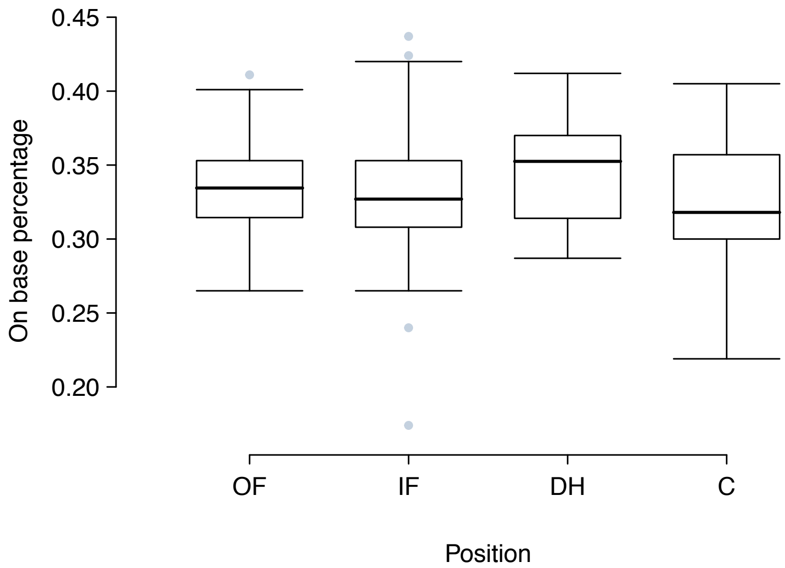


Figure 8.15: Side-by-side box plot of the on-base percentage for 327 players across four groups.

An initial look at the data might begin with the summary statistics for each group and a simple plot to compare the data. 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. Include a comment about roughly constant variance.

● **Example 8.21** The largest sample difference in the group means is between the designated hitter and the catcher positions. Consider again the original hypotheses:

$$H_0: \mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{DH}} = \mu_{\texttt{C}}$$
$$H_A: \text{The average on-base percentage } (\mu_j) \text{ varies across some (or all) groups.}$$

Why might it be inappropriate to run the test by simply estimating whether the difference of $\mu_{\texttt{DH}}$ and $\mu_{\texttt{C}}$ is statistically significant at the 0.05 significance level?

The primary issue here is that we are inspecting the data before picking the groups that will be compared. In reality, we are examining all data by eye (informal testing) and only afterwards deciding which parts to formally test. Naturally we pick the groups with the large differences for the formal test, and this can lead 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 are purely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different, we may incorrectly believe that there is something fishy going on in how students were assigned. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye. And if we examine enough pairs of classes, eventually we will find a pair that are quite different. If we are to proceed in this fashion, we need to employ methods for multiple comparisons.

We will examine the issue of **multiple comparisons** further in Section **??**. For additional reading on the ideas expressed in Example 8.21, we recommend reading about the **Prosecutor's fallacy**, which is summarized on a number of websites.

In the next section we will learn how to assess the means across many groups simultaneously, which will require use to use a new test statistic called F in the ANOVA framework.

8.5.2 The F test for the ANOVA framework

The method of analysis of variance (ANOVA) 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 hypothesis tests since we are looking at whether these estimates tend to, on the average, vary more than we would expect from natural variation. We call this variability the **mean square for the groups (**MSG**)**, and it has an associated degrees of freedom associated with it: $df_G = k - 1$. The MSG is sort of a scaled variance formula for means, and details of MSG calculations are provided in the footnote[7], though we typically use software for the computations. For example, for the baseball outcome and groups, the MSG is computed to be $MSG = 0.00252$ and have an associated $df_G = 3$.

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 we should expect to be associated with the sample means if the null hypothesis is true. To this end, we

[7] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square for the groups is computed as

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x}\right)^2$$

where SSG is called the **sum of squares for the groups** and n_j is the sample size corresponding to group j.

compute the **mean square of the errors** (**MSE**), which also has a degrees of freedom value associated with it: $df_E = n - k$. Details of the computations of the MSE are provided in the footnote[8] for the interested reader. For the baseball data, we have $MSE = 0.00127$ on $df_E = 327 - 4 = 323$ degrees of freedom.

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

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{MSG}{MSE} \tag{8.22}$$

The MSG represents a measure of the between-group variability, and MSE the variability within each of the groups (i.e. in the residuals).

⊙ **Exercise 8.23** For the baseball data, $MSG = 0.00252$ and $MSE = 0.00127$. Using this information and Equation (8.22), verify the F statistic is 1.994.

The hypothesis test using the F statistic to evaluate the hypotheses is called an **F test**. We compute a p-value from the F statistic using an F distribution, which has two associated parameters: df_{top} and df_{bottom}. For the F statistic, $df_{top} = df_G$ and $df_{bottom} = df_E$. 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 between-group variability is – i.e. the more variability observed in the sample means – the larger F will be and the more evidence there is against the null hypothesis. That is, a large value of F represents strong evidence against the null hypothesis.

Larger F values represent stronger evidence against the null hypothesis than what was observed, so we use the upper tail to compute the p-value.

---

[8]Let $\bar{x}$ represent the mean of outcomes across all groups. Then the **sum of squares total** (**SST**) is computed as

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

where the sum is over all observations in the data set. Then we compute the **sum of squares of the errors** (**MSE**) as

$$SSE = SST - SSG = (n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2 + \cdots + (n_k - 1) * s_k^2$$

We can compute SSE using SST with SSG (SSG is described in an earlier footnote) or by using the second expression that represents a weighted sum of the individual group variances. The MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

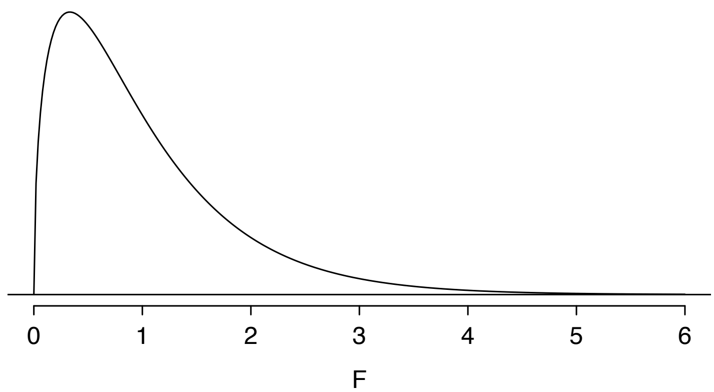


Figure 8.16: An F distribution with $df_1 = 3$ and $df_2 = 323$.

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 of observations between the groups relative to the variability of observations within the groups. The statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$, and the upper tail corresponding to the F statistic represents the p-value.

⊙ **Exercise 8.24** The test statistic for the baseball example is $F = 1.994$. Shade the area corresponding to the p-value in Figure 8.16. Answer in the footnote[9].

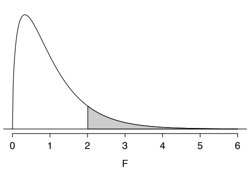
● **Example 8.25** The p-value pictured in the solution to Exercise 8.24 is equal to about 0.115. Does this provide strong evidence against the null hypothesis?

The p-value is larger than 0.05, indicating the evidence is not sufficiently strong to reject the null hypothesis for a significance level of 0.05. That is, the data do not provide strong evidence that the means of the groups are actually different.

8.5.3 Reading regression and ANOVA output from software

The calculuations required to run an ANOVA analysis by hand are tedious and prone to human error. For these reasons it is common to instead let a computer do the heavy lifting and compute the F statistic and p-value.

9



An ANOVA analysis is typically summarized in a table very similar to that of a regression summary. Table 8.17 shows an ANOVA summary for testing whether the mean on-base percentage is different for players of different positions.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	3.0000	0.0076	0.0025	1.9943	0.1147
Residuals	323.0000	0.4080	0.0013		

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

⊙ **Exercise 8.26** Earlier you verified the F statistic for this analysis was 1.994, and the p-value was provided as about 0.115. Identify these values in Table 8.17. Notice that both of these values are in the row labeled *position*, which corresponds to the categorical variable representing the player position variable.

8.5.4 Graphical diagnostics for an ANOVA analysis

There are three primary assumptions 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* and are *nearly normal* with *constant variance*.

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 13). For the baseball data, the data were not truly collected in a particular order since they were collected from a sorted table. However, we have little reason to believe player performance is dependent, and the assumption seems reasonable.

Nearly normal. The normality assumption for the residuals is especially important when the sample size is small but can be slightly relaxed for residuals of groups with larger sample sizes. We do see some deviations from normality in the lower tail. However, the three smallest residuals are from the infield and outfield groups. We are partially appeased in that these two groups each contain over 100 cases, so the potential outliers probably alright (but these observations are worth mentioning in a report of the analysis).

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 plotting the residuals by their groups in a box plot, as in Figure 8.19. In this case, the variability is similar in the three groups.

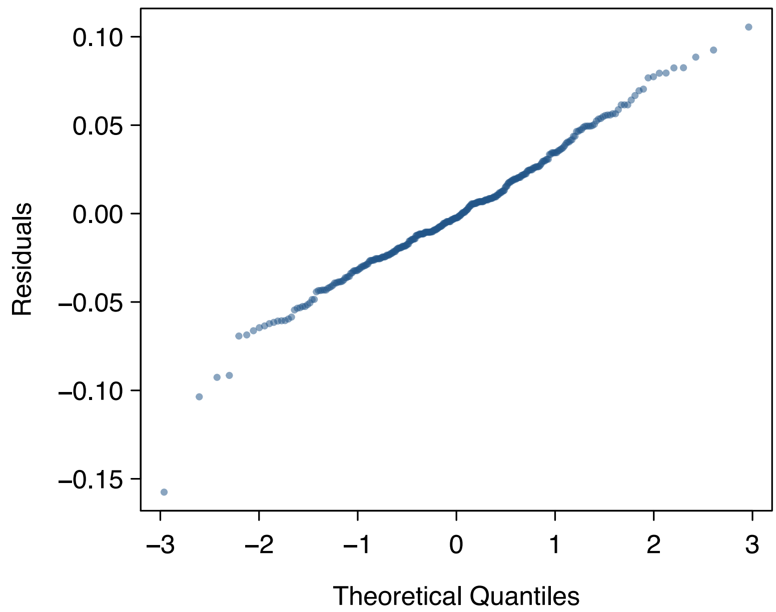


Figure 8.18: Normal probability plot of the residuals.

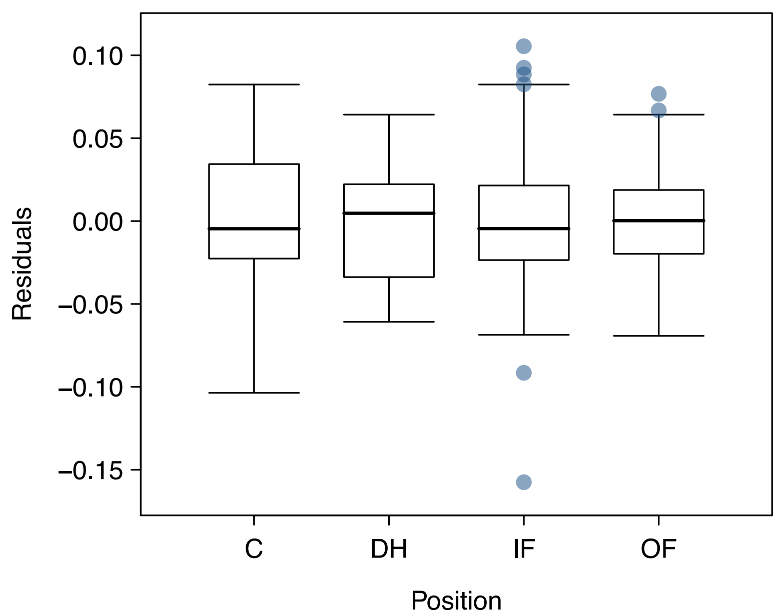


Figure 8.19: Side-by-side box plot of the residuals in their corresponding groups.

Caution: Diagnostics for 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.

8.5.5 Using ANOVA for multiple regression

The ANOVA methodology is a very powerful technique. Its use can be extended to that of multiple regression, where we simultaneously incorporate categorical and numerical predictors into a model. There are two uses for ANOVA that we 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, and this additional information can be used as a 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:

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.0078  -3.0754  -0.8254   2.9822  14.1646

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.7849     0.7066  52.062  < 2e-16 ***
condNew       5.5848     0.9245   6.041 1.35e-08 ***
wheels        7.2328     0.5419  13.347  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.887 on 138 degrees of freedom
Multiple R-squared: 0.7165,Adjusted R-squared: 0.7124
F-statistic: 174.4 on 2 and 138 DF,  p-value: < 2.2e-16
```

The main table with the estimates, standard errors, and other details should be familiar. The last three lines are new. They provide details about R^2, adjusted R^2, degrees of freedom, and also an F statistic with an associated p-value. This 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 the answer the following question: Is there strong evidence that the model is better than using no variables in predication at all? 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.

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 omit any further details of this method since it is an advanced technique requiring much attention.

8.6 Logistic regression for a binomial outcome

8.6.1 Evaluating the 2010 House election using 2008 results

8.6.2 Modeling the probability of an event

8.6.3 Fitting a model to

.1 t Distribution Table

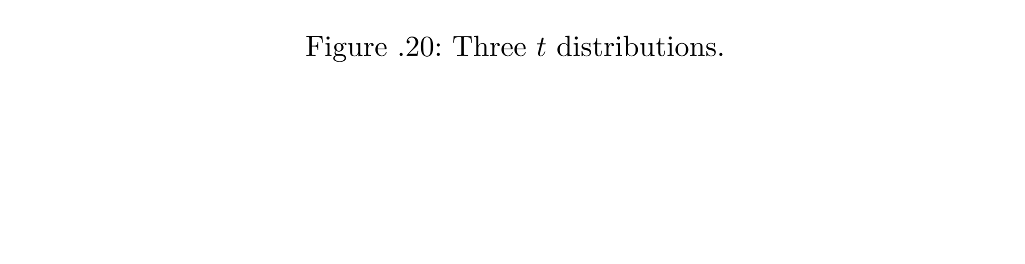


Figure .20: Three t distributions.

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 31	1.31	1.70	2.04	2.45	2.74
32	1.31	1.69	2.04	2.45	2.74
33	1.31	1.69	2.03	2.44	2.73
34	1.31	1.69	2.03	2.44	2.73
35	1.31	1.69	2.03	2.44	2.72
36	1.31	1.69	2.03	2.43	2.72
37	1.30	1.69	2.03	2.43	2.72
38	1.30	1.69	2.02	2.43	2.71
39	1.30	1.68	2.02	2.43	2.71
40	1.30	1.68	2.02	2.42	2.70
41	1.30	1.68	2.02	2.42	2.70
42	1.30	1.68	2.02	2.42	2.70
43	1.30	1.68	2.02	2.42	2.70
44	1.30	1.68	2.02	2.41	2.69
45	1.30	1.68	2.01	2.41	2.69
46	1.30	1.68	2.01	2.41	2.69
47	1.30	1.68	2.01	2.41	2.68
48	1.30	1.68	2.01	2.41	2.68
49	1.30	1.68	2.01	2.40	2.68
50	1.30	1.68	2.01	2.40	2.68
60	1.30	1.67	2.00	2.39	2.66
70	1.29	1.67	1.99	2.38	2.65
80	1.29	1.66	1.99	2.37	2.64
90	1.29	1.66	1.99	2.37	2.63
100	1.29	1.66	1.98	2.36	2.63
150	1.29	1.66	1.98	2.35	2.61
200	1.29	1.65	1.97	2.35	2.60
300	1.28	1.65	1.97	2.34	2.59
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58