2 Multiple Linear Regression

Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable. However, in practice we often have more than one predictor. For example, in the `Advertising` data, we have examined the relationship between sales and TV advertising. We also have data for the amount of money spent advertising on the radio and in newspapers, and we may want to know whether either of these two media is associated with sales. How can we extend our analysis of the advertising data in order to accommodate these two additional predictors?

One option is to run three separate simple linear regressions, each of which uses a different advertising medium as a predictor. For instance, we can fit a simple linear regression to predict sales on the basis of the amount spent on radio advertisements. Results are shown in Table 5 (top table). We find that a $1,000 increase in spending on radio advertising is associated with an increase in sales by around 203 units. Table 5 (bottom table) contains the least squares coefficients for a simple linear regression of sales onto newspaper advertising budget. A $1,000 increase in newspaper advertising budget is associated with an increase in sales by approximately 55 units.

However, the approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory. First of all, it is unclear how to make a single prediction of sales given levels of the three advertising media budgets, since each of the budgets is associated with a separate regression equation. Second, each of the three regression equations ignores the other two media in forming estimates for the regression coefficients. We will see shortly that if the media budgets are correlated with each other in the 200 markets that constitute our data set, then this can lead to very misleading estimates of the individual media effects on sales.

Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model (5) so that it can directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model. In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon, \tag{15}$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response. We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, (15) becomes

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{TV} + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \epsilon. \tag{16}$$

2.1 Estimating the Regression Coefficients

As was the case in the simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ in (15) are unknown, and must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p. \tag{17}$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.0012

Table 5: More simple linear regression models for the Advertising data. Coefficients of the simple linear regression model for number of units sold on Top: radio advertising budget and Bottom: newspaper advertising budget. A \$1,000 increase in spending on radio advertising is associated with an average increase in sales by around 203 units, while the same increase in spending on newspaper advertising is associated with an average increase in sales by around 55 units (Note that the `sales` variable is in thousands of units, and the `radio` and `newspaper` variables are in thousands of dollars).

The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression. We choose $\beta_0, \beta_1, \ldots, \beta_p$ to minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip} \right)^2 . \end{aligned} \tag{18}$$

The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize (18) are the multiple least squares regression coefficient estimates. Unlike the simple linear regression estimates given in (4), the multiple regression coefficient estimates have somewhat complicated forms that are most easily represented using matrix algebra. For this reason, we do not provide them here. Any statistical software package can be used to compute these coefficient estimates, and later in this chapter we will show how this can be done in `R`. Figure 7 illustrates an example of the least squares fit to a toy data set with $p = 2$ predictors.

Table 6 displays the multiple regression coefficient estimates when TV, radio, and newspaper advertising budgets are used to predict product sales using the `Advertising` data.We interpret these results as follows: for a given amount of TV and newspaper advertising, spending an additional \$1,000 on radio advertising leads to an increase in sales by approximately 189 units. Comparing these coefficient estimates to those displayed in Tables 3 and 5, we notice that the multiple regression coefficient estimates for `TV` and `radio` are pretty similar to the simple linear regression coefficient estimates. However, while the `newspaper` regression coefficient estimate in Table 5 was significantly non-zero, the coefficient estimate for `newspaper` in the multiple regression model is close to zero, and the corresponding p-value is no longer significant, with a value around 0.86. This illustrates that the simple and multiple regression coefficients can be quite different. This difference stems from the fact that in the simple regression case, the slope term represents the average effect of a \$1,000 increase in newspaper advertising, ignoring other predictors such as `TV` and `radio`. In contrast, in the multiple regression setting, the coefficient for `newspaper` represents the average effect of increasing `newspaper` spending by \$1,000 while holding `TV` and `radio` fixed.

Does it make sense for the multiple regression to suggest no relationship between `sales` and `newspaper` while the simple linear regression implies the opposite? In fact it does. Consider the correlation matrix for the three predictor variables and response variable, displayed in Table 7. Notice that the correlation between `radio` and `newspaper` is 0.35. This reveals a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising. Now suppose that the multiple regression is

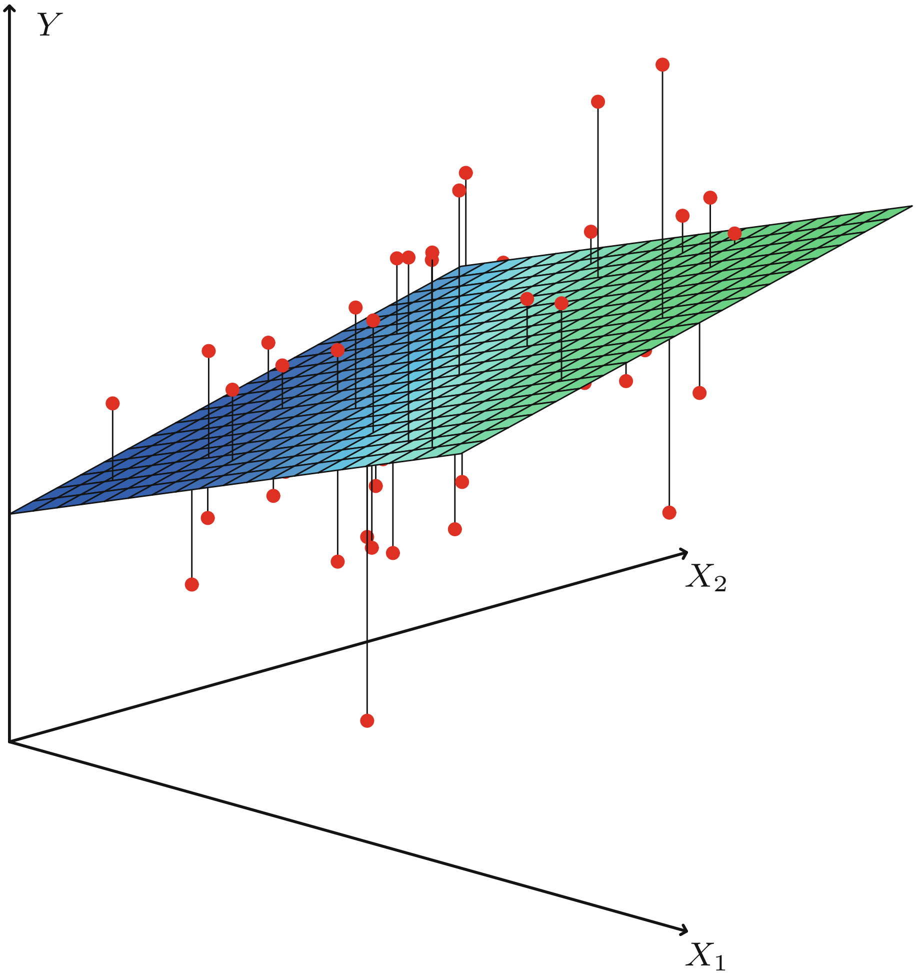

Figure 7: In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Table 6: For the `Advertising` data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Table 7: Correlation matrix for `TV`, `radio`, `newspaper`, and `sales` for the `Advertising` data.

correct and newspaper advertising has no direct impact on sales, but radio advertising does increase sales. Then in markets where we spend more on radio our sales will tend to be higher, and as our correlation matrix shows, we also tend to spend more on newspaper advertising in those same markets. Hence, in a simple linear regression which only examines `sales` versus `newspaper`, we will observe that higher values of `newspaper` tend to be associated with higher values of `sales`, even though newspaper advertising does not actually affect sales. So `newspaper` sales are a surrogate for `radio` advertising; `newspaper` gets "credit" for the effect of `radio` on `sales`.

This slightly counter-intuitive result is very common in many real life situations. Consider an absurd example to illustrate the point. Running a regression of shark attacks versus ice cream sales for data collected at a given beach community over a period of time would show a positive relationship, similar to that seen between `sales` and `newspaper`. Of course no one (yet) has suggested that ice creams should be banned at beaches to reduce shark attacks. In reality, higher temperatures cause more people to visit the beach, which in turn results in more ice cream sales and more shark attacks. A multiple regression of attacks versus ice cream sales and temperature reveals that, as intuition implies, the former predictor is no longer significant after adjusting for temperature.

2.1.1 Some Important Questions

When we perform multiple linear regression, we usually are interested in answering a few important questions.

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

We now address each of these questions in turn.

1. Is There a Relationship Between the Response and Predictors?

Recall that in the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta_1 = 0$. In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. As in the simple linear regression setting, we use a hypothesis test to answer this question. We test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

Table 8: More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the `Advertising` data.

versus the alternative

$$H_a\text{: at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the F-statistic,

$$\mathrm{F} = \frac{(\mathrm{TSS} - \mathrm{RSS})\,/p}{\mathrm{RSS}/\,(n-p-1)}, \tag{19}$$

where, as with simple linear regression, $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. If the linear model assumptions are correct, one can show that

$$\mathrm{E}\left\{\mathrm{RSS}/(n-p-1)\right\} = \sigma^2$$

and that, provided H_0 is true,

$$\mathrm{E}\left\{[\mathrm{TSS} - \mathrm{RSS}]\,/p\right\} = \sigma^2.$$

Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, then $\mathrm{E}\left\{[\mathrm{TSS} - \mathrm{RSS}]\,/p\right\} > \sigma^2$, so we expect F to be greater than 1.

The F-statistic for the multiple linear regression model obtained by regressing `sales` onto `radio`, `TV`, and `newspaper` is shown in Table 8. In this example the F-statistic is 570. Since this is far larger than 1, it provides compelling evidence against the null hypothesis H_0. In other words, the large F-statistic suggests that at least one of the advertising media must be related to `sales`. However, what if the F-statistic had been closer to 1? How large does the F-statistic need to be before we can reject H_0 and conclude that there is a relationship? It turns out that the answer depends on the values of n and p. When n is large, an F-statistic that is just a little larger than 1 might still provide evidence against H_0. In contrast, a larger F-statistic is needed to reject H_0 if n is small. When H_0 is true and the errors ϵ_i have a normal distribution, the F-statistic follows an F-distribution. For any given value of n and p, any statistical software package can be used to compute the p-value associated with the F-statistic using this distribution. Based on this p-value, we can determine whether or not to reject H_0. For the advertising data, the p-value associated with the F-statistic in Table 8 is essentially zero, so we have extremely strong evidence that at least one of the media is associated with increased `sales`.

In (19) we are testing H_0 that all the coefficients are zero. Sometimes we want to test that a particular subset of q of the coefficients are zero. This corresponds to a null hypothesis

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0,$$

where for convenience we have put the variables chosen for omission at the end of the list. In this case we fit a second model that uses all the variables except those last q. Suppose that the residual sum of squares for that model is RSS_0. Then the appropriate F-statistic is

$$\mathrm{F} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})\,/q}{\mathrm{RSS}/\,(n-p-1)}. \tag{20}$$

Notice that in Table 6, for each individual predictor a t-statistic and a p-value were reported. These provide information about whether each individual predictor is related to the response, after adjusting for the other predictors. It turns out that each of these are exactly equivalent to the F-test that omits that single variable from the model, leaving all the others in - i.e. $q = 1$ in (20). So it reports the partial effect of adding that variable to the model. For instance, as we discussed earlier, these p-values indicate that `TV` and `radio` are related to `sales`, but that there is no evidence that `newspaper` is associated with `sales`, in the presence of these two.

Given these individual p-values for each variable, why do we need to look at the overall F-statistic? After all, it seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response. However, this logic is flawed, especially when the number of predictors p is large.

For instance, consider an example in which $p = 100$ and $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ is true, so no variable is truly associated with the response. In this situation, about 5% of the p-values associated with each variable (of the type shown in Table 6) will be below 0.05 by chance. In other words, we expect to see approximately five small p-values even in the absence of any true association between the predictors and the response. In fact, we are almost guaranteed that we will observe at least one p-value below 0.05 by chance! Hence, if we use the individual t-statistics and associated p-values in order to decide whether or not there is any association between the variables and the response, there is a very high chance that we will incorrectly conclude that there is a relationship. However, the F-statistic does not suffer from this problem because it adjusts for the number of predictors. Hence, if H_0 is true, there is only a 5% chance that the F-statistic will result in a p-value below 0.05, regardless of the number of predictors or the number of observations.

The approach of using an F-statistic to test for any association between the predictors and the response works when p is relatively small, and certainly small compared to n. However, sometimes we have a very large number of variables. If $p > n$ then there are more coefficients β_j to estimate than observations from which to estimate them. In this case we cannot even fit the multiple linear regression model using least squares, so the F-statistic cannot be used, and neither can most of the other concepts that we have seen so far in this chapter. When p is large, some of the approaches discussed in the next section, such as forward selection, can be used.

2. Deciding on Important Variables

As discussed in the previous section, the first step in a multiple regression analysis is to compute the F-statistic and to examine the associated p-value. If we conclude on the basis of that p-value that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones! We could look at the individual p-values as in Table 6, but as discussed, if p is large we are likely to make some false discoveries. It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only related to a subset of the predictors. The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as variable selection. Here we will provide a brief outline selection of some classical approaches.

Ideally, we would like to perform variable selection by trying out a lot of different models, each containing a different subset of the predictors. For instance, if $p = 2$, then we can consider four models: (1) a model containing no variables, (2) a model containing X_1 only, (3) a model containing X_2 only, and (4) a model containing both X_1 and X_2. We can then select the best model out of all of the models that we have considered. How do we determine which model is best? Various statistics can be used to judge the quality of a model. These include Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2. We can also determine which model is best by plotting various model outputs, such as the residuals, in order to search for patterns.

Unfortunately, there are a total of 2^p models that contain subsets of p variables. This means that even for moderate p, trying out every possible subset of the predictors is infeasible. For instance, we saw that if $p = 2$, then there are $2^2 = 4$ models to consider. But if $p = 30$, then we must consider $2^{30} = 1,073,741,824$ models! This is not practical. Therefore, unless p is very small, we cannot consider all 2^p models, and instead we need an automated and efficient approach to choose a smaller set of models to consider. There are three classical approaches for this task:

- Forward selection. We begin with the null model - a model that contains an intercept but no predictors. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. We then add to that model the variable that results in the lowest RSS for the new two-variable model. This approach is continued until some stopping rule is satisfied.
- Backward selection. We start with all variables in the model, and remove the variable with the largest p-value - that is, the variable selection that is the least statistically significant. The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed. This procedure continues until a stopping rule is reached. For instance, we may stop when all remaining variables have a p-value below some threshold.
- Mixed selection. This is a combination of forward and backward selection. We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variables one-by-one. Of course, as we noted with the `Advertising` example, the p-values for variables can become larger as new predictors are added to the model. Hence, if at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Backward selection cannot be used if $p > n$, while forward selection can always be used. Forward selection is a greedy approach, and might include variables early that later become redundant. Mixed selection can remedy this.

3. Model Fit

Two of the most common numerical measures of model fit are the RSE and R^2. These quantities are computed and interpreted in the same fashion as for simple linear regression. Recall that in simple regression, R^2 is the square of the correlation of the response and the variable. In multiple linear regression, it turns out that it equals $\mathrm{Cor}(Y, \hat{Y})^2$, the square of the correlation between the response and the fitted linear model; in fact one property of the fitted linear model is that it maximizes this correlation among all possible linear models. An R^2 value close to 1 indicates that the model explains a large portion of the variance in the response variable. As an example, we saw in Table 8 that for the `Advertising` data, the model that uses all three advertising media to predict `sales` has an R^2 of 0.8972. On the other hand, the model that uses only `TV` and `radio` to predict `sales` has an R^2 value of 0.89719. In other words, there is a small increase in R^2 if we include newspaper advertising in the model that already contains TV and radio advertising, even though we saw earlier that the p-value for newspaper advertising in Table 6 is not significant. It turns out that R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. This is due to the fact that adding another variable to the least squares equations must allow us to fit the data more accurately. Thus, the R^2 statistic, which is also computed on the same data, must increase. The fact that adding newspaper advertising to the model containing only TV and radio advertising leads to just a tiny increase in R^2 provides additional evidence that newspaper can be dropped from the model. Essentially, newspaper provides no real improvement in the model fit to the training samples, and its inclusion will likely lead to poor results on independent test samples due to overfitting.

In contrast, the model containing only TV as a predictor had an R^2 of 0.61 (Table 4). Adding radio to the model leads to a substantial improvement in R^2. This implies that a model that uses TV and radio expenditures to predict sales is substantially better than one that uses only TV advertising. We could further quantify this improvement by looking at the p-value for the radio coefficient in a model that contains only TV and radio as predictors.

The model that contains only TV and radio as predictors has an RSE of 1.681, and the model that also contains newspaper as a predictor has an RSE of 1.686 (Table 8). In contrast, the model that contains only TV has an RSE of 3.26 (Table 4). This corroborates our previous conclusion that a model that uses TV and radio expenditures to predict sales is much more accurate than one that only uses TV spending. Furthermore, given that TV and radio expenditures are used as predictors, there is no point in also using newspaper spending as a predictor in the model. The observant reader may wonder how RSE can increase when newspaper is added to the model given that RSS must decrease. In general RSE is defined as

$$\text{RSE} = \sqrt{\frac{1}{(n-p-1)}\text{RSS}}, \tag{21}$$

which simplifies to (12) for a simple linear regression. Thus, models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p.

4. Predictions

Once we have fit the multiple regression model, it is straightforward to apply (17) in order to predict the response Y on the basis of a set of values for the predictors $X_1, X_2, \dots, X_p$. However, there are three sorts of uncertainty associated with this prediction.

1. The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, \dots, \beta_p$. That is, the least squares plane
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$
is only an estimate for the true population regression plane
$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$
The inaccuracy in the coefficient estimates is related to the reducible error (we can potentially improve the accuracy of the fitted model by using the most appropriate statistical model). We can compute a confidence interval in order to determine how close $\hat{Y}$ will be to $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

2. Of course, in practice assuming a linear model is almost always an approximation of reality, so there is an additional source of potentially reducible error which we call model bias. So when we use a linear model, we are in fact estimating the best linear approximation to the true surface. However, here we will ignore this discrepancy, and operate as if the linear model were correct.

3. Even if we knew the model - that is, even if we knew the true values for $\beta_0, \beta_1, \dots, \beta_p$ - the response value cannot be predicted perfectly because of the random error ϵ. We refer to this as the irreducible error. How much will Y vary from $\hat{Y}$? We use prediction intervals to answer this

question. Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for the model (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

We use a confidence interval to quantify the uncertainty surrounding the average `sales` over a large number of cities. For example, given that \$100,000 is spent on `TV` advertising and \$20,000 is spent on `radio` advertising in each city, the 95% confidence interval is [10,985, 11,528]. We interpret this to mean that 95% of intervals of this form will contain the true value of $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$. On the other hand, a prediction interval can be used to quantify the uncertainty surrounding `sales` for a particular city. Given that \$100,000 is spent on `TV` advertising and \$20,000 is spent on `radio` advertising in that city the 95% prediction interval is [7,930, 14,580]. We interpret this to mean that 95% of intervals of this form will contain the true value of Y for this city. Note that both intervals are centered at 11,256, but that the prediction interval is substantially wider than the confidence interval, reflecting the increased uncertainty about `sales` for a given city in comparison to the average `sales` over many locations.

2.1.2 Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

In practice, identifying and overcoming these problems is as much an art as a science. We will provide a brief summary of some key points.

1. Non-linearity of the Data

The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect. In addition, the prediction accuracy of the model can be significantly reduced.

Residual plots are a useful graphical tool for identifying non-linearity. Given a simple linear regression model, we can plot the residuals, $e_i = y_i - \hat{y}_i$, versus the predictor x_i. In the case of a multiple regression model, since there are multiple predictors, we instead plot the residuals versus the predicted (or fitted) values $\hat{y}_i$. Ideally, the residual plot will show no discernible pattern. The presence of a pattern may indicate a problem with some aspect of the linear model.

The left panel of Figure 8 displays a residual plot from the linear regression of `mpg` onto `horsepower` on the `Auto` data set. The red line is a smooth fit to the residuals, which is displayed in order to make it easier to identify any trends. The residuals exhibit a clear U-shape, which provides a strong indication of non-linearity in the data. In contrast, the right-hand panel of Figure 8 displays the residual plot that results from a model which contains a quadratic term. There appears to be little pattern in the residuals, suggesting that the quadratic term improves the fit to the data.

If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as $\log X$, $\sqrt{X}$ and X^2, in the regression model. In the later chapters of this notes, we will discuss other more advanced non-linear approaches for addressing this issue.

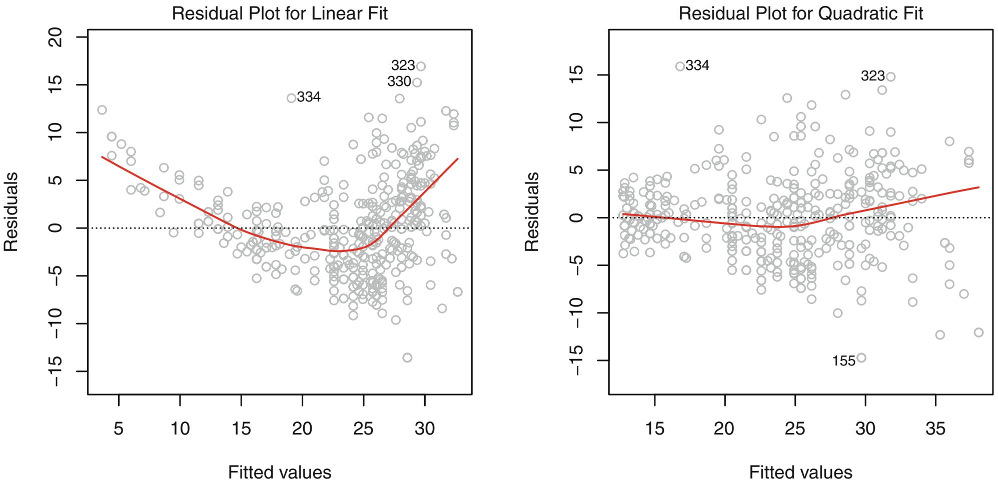

Figure 8: Plots of residuals versus predicted (or fitted) values for the `Auto` data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. Left: A linear regression of `mpg` on `horsepower`. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of `mpg` on `horsepower` and `horsepower`^{2}. There is little pattern in the residuals.

2. Correlation of Error Terms

An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, are uncorrelated. What does this mean? For instance, if the errors are uncorrelated, then the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}. The standard errors that are computed for the estimated regression coefficients or the fitted values are based on the assumption of uncorrelated error terms. If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be. For example, a 95% confidence interval may in reality have a much lower probability than 0.95 of containing the true value of the parameter. In addition, p-values associated with the model will be lower than they should be; this could cause us to erroneously conclude that a parameter is statistically significant. In short, if the error terms are correlated, we may have an unwarranted sense of confidence in our model.

Why might correlations among the error terms occur? Such correlations frequently occur in the context of time series data, which consists of observations for which measurements are obtained at discrete points in time. In many cases, observations that are obtained at adjacent time points will have positively correlated errors. In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time. If the errors are uncorrelated, then there should be no discernible pattern. On the other hand, if the error terms are positively correlated, then we may see tracking in the residuals - that is, adjacent residuals may have similar values. Figure 9 provides an illustration. In the top panel, we see the residuals from a linear regression fit to data generated with uncorrelated errors. There is no evidence of a time-related trend in the residuals. In contrast, the residuals in the bottom panel are from a data set in which adjacent errors had a correlation of 0.9. Now there is a clear pattern in the residuals - adjacent residuals tend to take on similar values. Finally, the center panel illustrates a more moderate case in which the residuals had a correlation of 0.5. There is still evidence of tracking, but the pattern is less clear. Many methods have been developed to properly take account of correlations in the error terms in time series data. Correlation among the error terms can also occur outside of time series data. For instance, consider a study in which individuals heights are predicted from their weights. The assumption of uncorrelated errors could be violated if some of the individuals in the study are members of the same family, or eat the same diet, or have been exposed to the same environmental factors. In general, the assumption of uncorrelated errors is extremely important for linear regression as well as for other statistical methods, and good experimental design is crucial in order to mitigate the risk of such correlations.

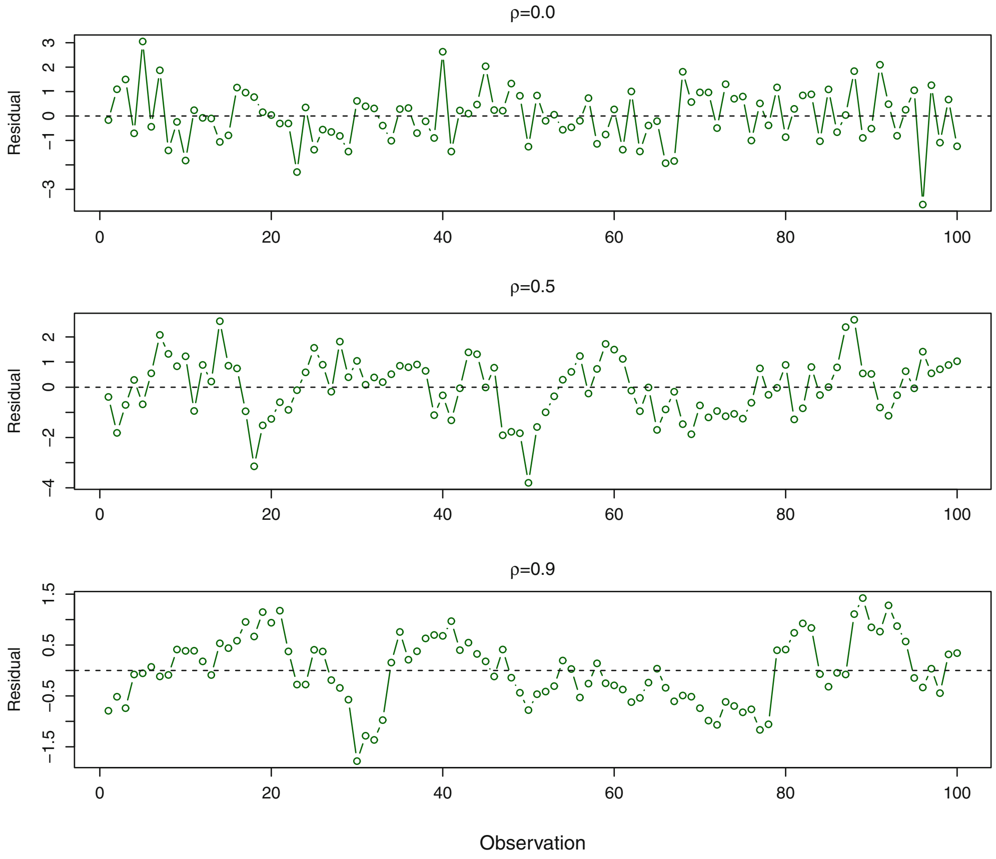

Figure 9: Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

3. Non-constant Variance of Error Terms

Another important assumption of the linear regression model is that the error terms have a constant variance, $\mathrm{Var}(\epsilon_i) = \sigma^2$. The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon this assumption.

Unfortunately, it is often the case that the variances of the error terms are non-constant. For instance, the variances of the error terms may increase with the value of the response. One can identify non-constant variances in the errors, or heteroscedasticity, from the presence of a funnel shape in the residual plot. An example is shown in the left-hand panel of Figure 10, in which the magnitude of the residuals tends to increase with the fitted values. When faced with this problem, one possible solution is to transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity. The right-hand panel of Figure 10 displays the residual plot after transforming the response using $\log Y$. The residuals now appear to have constant variance, though there is some evidence of a slight non-linear relationship in the data.

Sometimes we have a good idea of the variance of each response. For example, the ith response could be an average of n_i raw observations. If each of these raw observations is uncorrelated with variance σ^2, then their average has variance $\sigma_i^2 = \sigma^2/n_i$. In this case a simple remedy is to fit our model by weighted least squares, with weights proportional to the inverse variances - i.e. $w_i = n_i$ in this case. Most linear regression software allows least squares for observation weights.

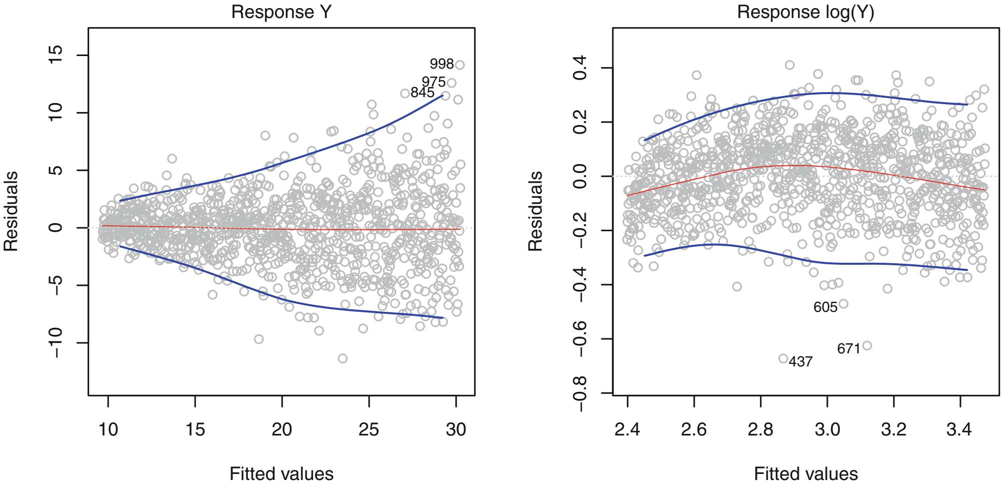

Figure 10: Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: The funnel shape indicates heteroscedasticity. Right: The response has been log transformed, and there is now no evidence of heteroscedasticity.

4. Outliers An outlier is a point for which y_i is far from the value predicted by the model. Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection.

The red point (observation 20) in the left-hand panel of Figure 11 illustrates a typical outlier. The red solid line is the least squares regression fit, while the blue dashed line is the least squares fit after removal of the outlier. In this case, removing the outlier has little effect on the least squares line: it leads to almost no change in the slope, and a miniscule reduction in the intercept. It is typical for an outlier that does not have an unusual predictor value to have little effect on the least squares fit. However, even if an outlier does not have much effect on the least squares fit, it can cause other problems. For instance, in this example, the RSE is 1.09 when the outlier is included in the regression, but it is only 0.77 when the outlier is removed. Since the RSE is used to compute all confidence intervals and p-values, such a dramatic increase caused by a single data point can have implications for the interpretation of the fit. Similarly, inclusion of the outlier causes the R^2 to decline from 0.892 to 0.805.

Residual plots can be used to identify outliers. In this example, the outlier is clearly visible in the

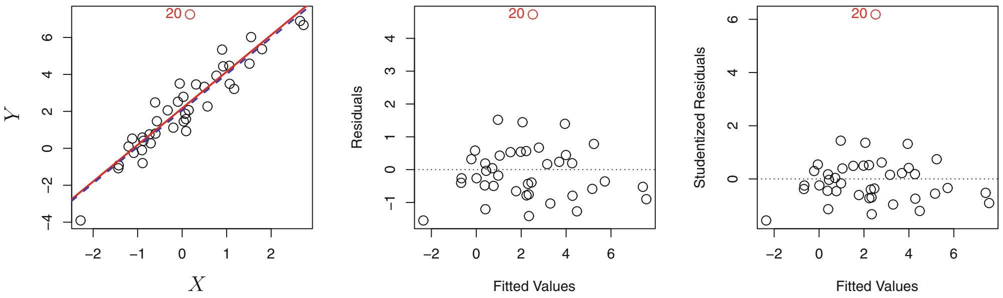

Figure 11: Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. Center: The residual plot clearly identifies the outlier. Right: The outlier has a studentized residual of 6; typically we expect values between 3 and 3.

residual plot illustrated in the center panel of Figure 11. But in practice, it can be difficult to decide how large a residual needs to be before we consider the point to be an outlier. To address this problem, instead of plotting the residuals, we can plot the standardized residuals, computed by dividing each residual e_i by its estimated standard error. Observations whose standardized residuals are greater than 3 in absolute value are possible outliers. In the right-hand panel of Figure 11, the outlier's standardized residual exceeds 6, while all other observations have standardized residuals between 2 and 2.

If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation. However, care should be taken, since an outlier may instead indicate a deficiency with the model, such as a missing predictor.

5. High Leverage Points

We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i. In contrast, observations with high leverage have an unusual value for x_i. For example, observation 41 in the left-hand panel of Figure 12 has high leverage, in that the predictor value for this observation is large relative to the other observations. (Note that the data displayed in Figure 12 are the same as the data displayed in Figure 11, but with the addition of a single high leverage observation.) The red solid line is the least squares fit to the data, while the blue dashed line is the fit produced when observation 41 is removed. Comparing the left-hand panels of Figures 11 and 12, we observe that removing the high leverage observation has a much more substantial impact on the least squares line than removing the outlier. In fact, high leverage observations tend to have a sizable impact on the estimated regression line. It is cause for concern if the least squares line is heavily affected by just a couple of observations, because any problems with these points may invalidate the entire fit. For this reason, it is important to identify high leverage observations. In a simple linear regression, high leverage observations are fairly easy to identify, since we can simply look for observations for which the predictor value is outside of the normal range of the observations. But in a multiple linear regression with many predictors, it is possible to have an observation that is well within the range of each individual predictor's values, but that is unusual in terms of the full set of predictors. An example is shown in the center panel of Figure 12, for a data set with two predictors, X_1 and X_2. Most of the observations' predictor values fall within the blue dashed ellipse, but the red observation is well outside of this range. But neither its value for X_1 nor its value for X_2 is unusual. So if we examine just X_1 or just X_2, we will fail to notice this high leverage point. This problem is more pronounced in multiple regression settings with more than two predictors, because then there is no simple way to plot all dimensions of the data simultaneously.

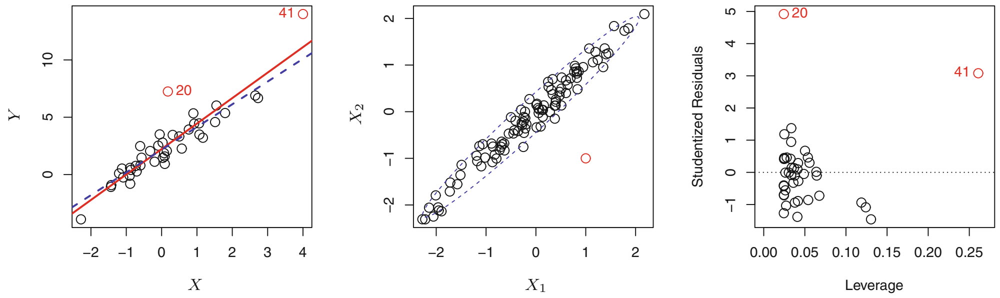

Figure 12: Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center: The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage. Right: Observation 41 has a high leverage and a high residual.

In order to quantify an observation's leverage, we compute the leverage statistic. A large value of this statistic indicates an observation with high leverage. For a simple linear regression,

$$h_i = \frac{1}{n} \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \bar{x})^2}. \tag{22}$$

It is clear from this equation that h_i increases with the distance of x_i from $\bar{x}$. There is a simple extension of h_i to the case of multiple predictors, though we do not provide the formula here. The leverage statistic h_i is always between $1/n$ and 1, and the average leverage for all the observations is always equal to $(p+1)/n$. So if a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, then we may suspect that the corresponding point has high leverage.

The right-hand panel of Figure 12 provides a plot of the studentized residuals versus h_i for the data in the left-hand panel of Figure 12. Observation 41 stands out as having a very high leverage statistic as well as a high standardized residual. In other words, it is an outlier as well as a high leverage observation. This is a particularly dangerous combination! This plot also reveals the reason that observation 20 had relatively little effect on the least squares fit in Figure 11: it has low leverage.

6. Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another. The concept of collinearity is illustrated in Figure 13 using the `Credit` data set. In the left-hand panel of Figure 13, the two predictors `limit` and `age` appear to have no obvious relationship. In contrast, in the right-hand panel of Figure 13, the predictors `limit` and `rating` are very highly correlated with each other, and we say that they are collinear. The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response. In other words, since `limit` and `rating` tend to increase or decrease together, it can be difficult to determine how each one separately is associated with the response, balance. Figure 14 illustrates some of the difficulties that can result from collinearity. The left-hand panel of Figure 14 is a contour plot of the RSS associated with different possible coefficient estimates for the regression of `balance` on `limit` and `age`. Each ellipse represents a set of coefficients that correspond to the same RSS, with ellipses nearest to the center taking on the lowest values of RSS. The black dots and associated dashed lines represent the coefficient estimates that result in the smallest possible RSS - in other words, these are the least squares estimates. The axes for `limit` and `age` have been scaled so that the plot includes possible coefficient estimates that are up to four standard errors on either side of the least squares estimates. Thus the plot includes all plausible values for the coefficients. For example, we see that the true `limit` coefficient is almost certainly somewhere between 0.15 and 0.20.

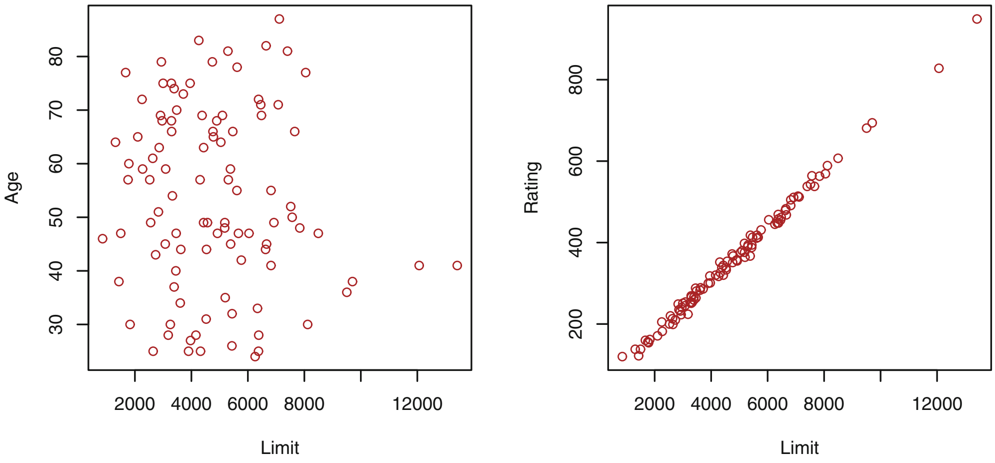

Figure 13: Scatterplots of the observations from the `Credit` data set. Left: A plot of `age` versus `limit`. These two variables are not collinear. Right: A plot of `rating` versus `limit`. There is high collinearity.

In contrast, the right-hand panel of Figure 14 displays contour plots of the RSS associated with possible coefficient estimates for the regression of `balance` onto `limit` and `rating`, which we know to be highly collinear. Now the contours run along a narrow valley; there is a broad range of values for the coefficient estimates that result in equal values for RSS. Hence a small change in the data could cause the pair of coefficient values that yield the smallest RSS - that is, the least squares estimates - to move anywhere along this valley. This results in a great deal of uncertainty in the coefficient estimates. Notice that the scale for the limit coefficient now runs from roughly -0.2 to 0.2; this is an eight-fold increase over the plausible range of the limit coefficient in the regression with age. Interestingly, even though the limit and rating coefficients now have much more individual uncertainty, they will almost certainly lie somewhere in this contour valley. For example, we would not expect the true value of the `limit` and `rating` coefficients to be 0.1 and 1 respectively, even though such a value is plausible for each coefficient individually.

Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow. Recall that the t-statistic for each predictor is calculated by dividing $\hat{\beta}_j$ by its standard error. Consequently, collinearity results in a decline in the t-statistic. As a result, in the presence of collinearity, we may fail to reject $H_0 : \beta_j = 0$. This means that the power of the hypothesis test - the probability of correctly detecting a non-zero coefficient - is reduced by collinearity.

Table 9 compares the coefficient estimates obtained from two separate multiple regression models. The first is a regression of `balance` on `age` and `limit`, and the second is a regression of `balance` on `rating` and `limit`. In the first regression, both `age` and `limit` are highly significant with very small p-values. In the second, the collinearity between `limit` and `rating` has caused the standard error for the limit coefficient estimate to increase by a factor of 12 and the p-value to increase to 0.701. In other words, the importance of the limit variable has been masked due to the presence of collinearity. To avoid such a situation, it is desirable to identify and address potential collinearity problems while fitting the model.

A simple way to detect collinearity is to look at the correlation matrix of the predictors. An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and

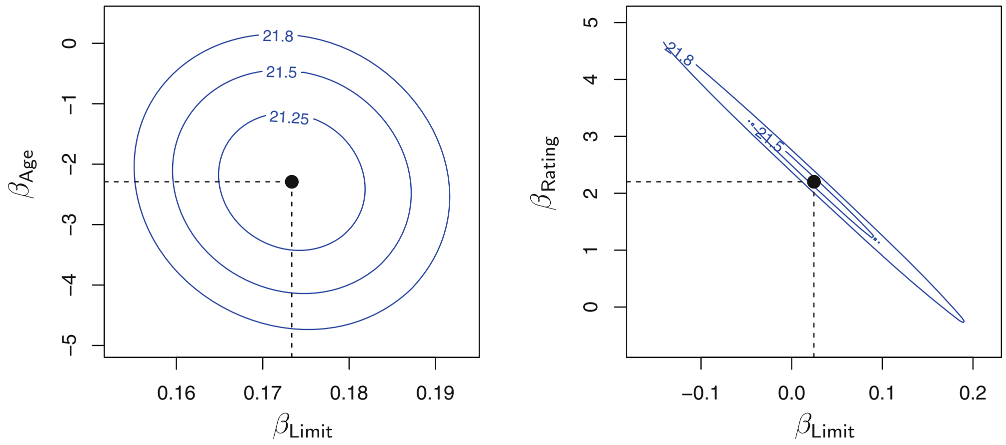

Figure 14: Contour plots for the RSS values as a function of the parameters β for various regressions involving the `Credit` data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS. Left: A contour plot of RSS for the regression of `balance` onto `age` and `limit`. The minimum value is well defined. Right: A contour plot of RSS for the regression of `balance` onto `rating` and `limit`. Because of the collinearity, there are many pairs $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$ with a similar value for RSS.

	Coefficient	Std. error	t-statistic	p-value	
Model 1	`Intercept`	-173.411	43.828	-3.957	< 0.0001
	`age`	-2.292	0.672	-3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	-377.537	45.254	-8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

Table 9: The results for two multiple regression models involving the `Credit` data set are shown. Model 1 is a regression of `balance` on `age` and `limit`, and Model 2 a regression of `balance` on `rating` and `limit`. The standard error of $\hat{\beta}_{\text{limit}}$ increases 12-fold in the second regression, due to collinearity.

therefore a collinearity problem in the data. Unfortunately, not all collinearity problems can be detected by inspection of the correlation matrix: it is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation. We call this situation multicollinearity.

Instead of inspecting the correlation matrix, a better way to assess multicollinearity is to compute the variance inflation factor (VIF). The VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own. The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. Typically in practice there is a small amount of collinearity among the predictors. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. The VIF for each variable can be computed using the formula

$$\text{VIF}\left(\hat{\beta}_j\right) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors. If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.

In the `Credit` data, a regression of `balance` on `age`, `rating`, and `limit` indicates that the predictors have VIF values of 1.01, 160.67, and 160.59. As we suspected, there is considerable collinearity in the data! When faced with the problem of collinearity, there are two simple solutions. The first is to drop one of the problematic variables from the regression. This can usually be done without much compromise to the regression fit, since the presence of collinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables. For instance, if we regress `balance` onto `age` and `limit`, without the `rating` predictor, then the resulting VIF values are close to the minimum possible value of 1, and the R^2 drops from 0.754 to 0.75. So dropping rating from the set of predictors has effectively solved the collinearity problem without compromising the fit. The second solution is to combine the collinear variables together into a single predictor. For instance, we might take the average of standardized versions of `limit` and `rating` in order to create a new variable that measures credit worthiness.

2.2 Lab: Multiple Linear Regression

2.2.1 Multiple Linear Regression I

In order to fit a multiple linear regression model using least squares, we again use the `lm()` function. The syntax `lm(y ∼ x1 + x2 + x3)` is used to fit a model with three predictors, `x1`, `x2`, and `x3`. The `summary()` function now outputs the regression coefficients for all the predictors.

```
library(MASS)
lm.fit <- lm(medv ~ lstat + age, data = Boston)
summary(lm.fit)
```

The Boston data set contains 13 variables, and so it would be cumbersome to have to type all of these in order to perform a regression using all of the predictors. Instead, we can use the following short-hand:

```
lm.fit <- lm(medv ~., data = Boston)
summary(lm.fit)
```

We can access the individual components of a summary object by name (type `?summary.lm` to see what is available). Hence `summary(lm.fit)$r.sq` gives us the R^2, and `summary(lm.fit)$sigma` gives us the RSE. The `vif()` function, part of the `car` package, can be used to compute variance inflation factors. Most VIF's are low to moderate for this data. The `car` package is not part of the base `R` installation so it must be downloaded the first time you use it via the `install.packages` option in `R`.

```
library(car)
vif(lm.fit)
```

What if we would like to perform a regression using all of the variables but one? For example, in the above regression output, `age` has a high p-value. So we may wish to run a regression excluding this predictor. The following syntax results in a regression using all predictors except `age`.

```
lm.fit1 <- lm(medv ~. -age, data = Boston)
summary(lm.fit1)
```

Alternatively, the `update()` function can be used.

```
lm.fit1 <- update(lm.fit, ~. -age)
```

2.2.2 Multiple Linear Regression II

An estimate is required of the percentage yield of petroleum spirit from crude oil, based upon certain rough laboratory determinations of properties of the crude oil. The following table shows actual percentage yields of petroleum spirit, Y, and four properties, X_1, X_2, X_3, X_4, of the crude oil, for samples from 32 different crudes.

Data on yields of petroleum spirit

y	x_1	x_2	x_3	x_4
6.9	38.4	6.1	220	235
14.4	40.3	4.8	231	307
7.4	40.0	6.1	217	212
8.5	31.8	0.2	316	365
8.0	40.8	3.5	210	218
2.8	41.3	1.8	267	235
5.0	38.1	1.2	274	285
12.2	50.8	8.6	190	205
10.0	32.2	5.2	236	267
15.2	38.4	6.1	220	300
26.8	40.3	4.8	231	367
14.0	32.2	2.4	284	351
14.7	31.8	0.2	316	379
6.4	41.3	1.8	267	275
17.6	38.1	1.2	274	365
22.3	50.8	8.6	190	275
24.8	32.2	5.2	236	360
26.0	38.4	6.1	220	365
34.9	40.3	4.8	231	395
18.2	40.0	6.1	217	272
23.2	32.2	2.4	284	424
18.0	31.8	0.2	316	428
13.1	40.8	3.5	210	273
16.1	41.3	1.8	267	358
32.1	38.1	1.2	274	444
34.7	50.8	8.6	190	345
31.7	32.2	5.2	236	402
33.6	38.4	6.1	220	410
30.4	40.0	6.1	217	340
26.6	40.8	3.5	210	347
27.8	41.3	1.8	267	416
45.7	50.8	8.6	190	407

The variables recorded are as follows.

Y:	percentage yield of petroleum spirit
X_1:	specific gravity of the crude
X_2:	crude oil vapour pressure, measured in pounds per square inch
X_3:	the ASTM 10% distillation point, in °F
X_4:	the petroleum fraction end point, in °F

It is required to use these data to provide an equation for predicting Y from measurements of the four explanatory variables, X_1, X_2, X_3, X_4, (or some subset of them).

The data have been read into `R` and stored as the *data frame* `oil`. The function `names` is used to assign names to the five variables. The function `lm` is then used to carry out a multiple linear regression of the response variable `spirit` upon the four regressor variables, `gravity`, `pressure`, `distil` and `endpoint`, the results of which are stored in the object `oil.lm`.

```
oil <- read.table("oil.txt")
names(oil) <- c("spirit","gravity","pressure","distil","endpoint")
oil.lm <- lm(spirit ~ gravity + pressure + distil + endpoint, data = oil)
summary(oil.lm)
```

The fitted regression equation is

$$\widehat{\texttt{spirit}} = -6.8208 + 0.2272\,\texttt{gravity} + 0.5537\,\texttt{pressure} - 0.1495\,\texttt{distil} + 0.1547\,\texttt{endpoint}$$

The corresponding p-values show that, in the presence of the other regressor variables, only the effect of the variable `pressure` is not significant at the 5% level. Thus, according to the test statistic that we have described, the removal of this variable from the regression does not lead to a significant reduction in the goodness of the fit, and we might consider omitting it.

The `Residual standard error: 2.234` is $\hat{\sigma}$. The value 171.7 of the overall F-statistic, with its corresponding P-value of 0, shows that there is highly significant evidence of a linear relationship between the response variable and the four regressor variables.

In view of the fact that `pressure` does not have a significant p-value, we repeat the regression analysis without it.

```
oil3.lm <- lm(spirit ~ gravity + distil + endpoint, data = oil)
summary(oil3.lm)
```

All the regressor variables are now significant, so we might very reasonably decide to adopt this model. However, the variable `gravity` is not very highly significant, so we might also consider the model after removing `gravity` as well.

```
oil2.lm <- lm(spirit ~ distil + endpoint, data = oil)
summary(oil2.lm)
```

Now we have a model for which both the remaining regressor variables are highly significant, so we could not remove either of them without a serious loss of fit of the model.

To summarize our results so far and to extend them somewhat, consider the following table, where m denotes the number of regressor variables being used.

m	$\hat{\sigma}$	R^2	$\bar{R}^2$
4	2.234	0.962	0.957
3	2.283	0.959	0.955
2	2.426	0.952	0.949
1	7.659	0.506	0.490

For each value of m, we have chosen the best set of m regressor variables, in the sense that it provides the best fit, giving the smallest value of σ and the largest value of R^2 (and $\bar{R}^2$). For $m = 2$, as may be checked by carrying out regressions on all possible pairs of regressor variables, `endpoint` and `distil` is the best pair of regressor variables to use. For $m = 1$, the best single regressor variable to use is `endpoint`.

From inspection of the table, we see that there is relatively little difference in the fit of the model, whether 2, 3 or 4 regressor variables are used. However, the use of only one regressor variable gives a much poorer fit. In choosing between the models with 2, 3 or 4 regressor variables, it is a matter of judgement whether we prefer a more complex model that give us a slightly better fit or a simpler model that gives a slightly poorer fit.

On this basis we may, at least for the present, opt for the model that uses the two regressor variables `endpoint` and `distil`. From the output for `oil2.lm`, we see that the fitted regression equation is

$$\widehat{\texttt{spirit}} = 18.4676 - 0.2093\,\texttt{distil} + 0.1558\,\texttt{endpoint}.$$

To compare two nested models constructed using the `lm()` function, `mod1` and `mod2`, say, we use

```
> anova(mod1,mod2).
```

For example:

```
oil0.lm <- lm(spirit ~ 1, data = oil)
anova(oil0.lm, oil.lm)
```

Since `oil0.lm` is the model with intercept only, this output results in the same F-statistic from the *full* model summary.

We use the function `predict` in `R` to obtain predicted values and their standard errors. We shall continue to use the model with two regressor variables. We construct a data frame `x` whose variable names are those of our regressor variables, `distil` and `endpoint`, and which contains the values of these regressor variables for which we wish to make predictions. In the present case, we shall use a single pair of values, (200, 400). The first argument of the `predict` function is the object `oil2.lm` that corresponds to our chosen model and the second argument is the data frame `x` that contains the values of the regressor variables for which we wish to make predictions. The arguments `se.fit = TRUE` and `interval = c("confidence", "prediction")`, are required so that we obtain confidence and prediction intervals, repsectively.

We may, if desired, calculate the confidence and prediction intervals.

```
x <- data.frame(distil = 200, endpoint = 400)
confidence <-predict(oil2.lm, x, se.fit = T, interval = c("confidence"))
confidence

prediction <-predict(oil2.lm, x, se.fit = T, interval = c("prediction"))
prediction
```

so that, for a crude oil with an ASTM 10% distillation point of 200°F and a petroleum fraction endpoint of 400°F, the model predicts a percentage yield of petroleum spirit of 38.92%, with 95% confidence interval (37.01, 40.85) and 95% prediction interval (33.61, 44.25).

The `stepAIC` from `MASS` package or `step` from `stats` are used to carry out a stepwise regression procedure which, by sequentially deleting or adding regressor variables, attempts to find a "best" set of regressor variables. Note that `step` is a minimal implementation and `stepAIC` could be used for a wider range of object classes.

In the following output for our example, the object `oil0.lm` is the result of fitting no regressor variables to the response variable `spirit`. It is used to specify the starting point for the iteration of no regressor variables present. Thus the `stepAIC` function has `oil0.lm` as its first argument. The second argument is the *scope* argument, which specifies the set of regressor variables to be considered for inclusion.

At each stage, for each candidate model, the output lists the change in the regression sum of squares due to the deleted or added variable, the resulting RSS and AIC.

Finally, a brief summary is provided for the chosen model, in the present case the model with the four regressor variables `endpoint`, `distil`, `gravity` and `pressure`.

```
library(MASS)
oil0.lm <- lm(spirit ~ 1, data = oil)
stepAIC(oil0.lm, ~ gravity + pressure + distil + endpoint, data = oil)
```

We carry out the stepwise procedure with all regressor variables present, which corresponds to the object `oil.lm`. In this case we find that the procedure deletes no variables, so that the suggested model is again the one with all four regressor variables.

```
stepAIC(oil.lm, ~ gravity + pressure + distil + endpoint, data = oil)
```

We end this lab by looking at some residual plots for the model `oil2.lm`, the *chosen* model using the regressors `distil` and `endpoint`. The generic `plot` function in `R` will produce *default* residual plots, using say

```
par(mfrow = c(2, 2))
plot(oil2.lm)
```

The top left plot of Residuals vs Fitted, does not show any indication of increasing variance with mean, which means that the constant variance assumption holds here. The other feature to look for is a pattern in the average value of the residuals as the fitted values change. The solid curve shows a running average of the residuals to help judging this: there is a slight pattern here, which is not extreme however. The lower left plot shows the square root of the absolute value of the standardized residuals against fitted value (again with a running average curve). If all is well the points should be evenly spread with respect the vertical axis here, with no trend in their average value. A trend in average value is indicative of a problem with the constant variance assumption, and is not a concern in this case. The top right plots the ordered standardized residuals against quantiles of a standard normal: a systematic deviation from a straight line is not present here indicating no departure from normality in the residuals. The lower right plot is looking at leverage and influence of residuals, by plotting standardized residuals against a measure of leverage. A combination of high residuals and high leverage indicates a point with substantial influence on the fit. A standard way of measuring this is via Cook's distance (which measures the change in all model fitted values on omission of the data point in question). It turns out that Cook's distance is a function of leverage and the standardized residuals, so contours of Cook distance values are shown on the plot. Cook distances over 0.5 are considered borderline problematic, while over 1 is usually considered highly influential. Although a couple of points have rather high leverage, their actual influence on the fit is not unduly large.

Finally we show plots of the studentized from `oil2.lm` against each of the four possible explanatory variables. There are no obvious patterns against any, whether included in the model or not, so that the adequacy of the model is not questioned.

```
stres <- rstudent(oil2.lm)
par(mfrow = c(2, 2))
plot(oil$gravity, stres)
plot(oil$distil, stres)
plot(oil$pressure, stres)
plot(oil$endpoint, stres)
```