Chapter 1

Multiple regression and model building

This chapter corresponds to Chapter 12 in McClave and Sincich's ”Statistics”, with supplements from Gareth James, Daniela Witten, Trevor Hastie, Robert Tibhshirani's ”Introduction to Statistical Learning with Applications in Python”. Sections 12.5 and 12.6 are skipped.

In this chapter, we focus on probabilistic models with more than one independent variable, called **multiple-regression models**, with the general form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

where $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ is the deterministic portion of the model and each β_i determines the contribution of each X_i.

To analyse a MRM, we use the following steps:

1. Hypothesise the deterministic component of the model, which relates the mean $E(Y)$ to the independent variables X_i.
2. Use the sample data to estimate the unknown parameters β_i in the model.
3. Specify the probability distribution of the random-error term ϵ and estimate the standard deviation σ of this distribution.
4. Check that all assumptions about ϵ are satisfied and make modifications where necessary.
5. Evaluate the model statistically.

Note:-

The assumptions about ϵ for multiple linear regression is similar to the case of linear regression: that is $\epsilon \sim N(0, \sigma^2)$ and that random errors are independent. If ε_i have constant variance, they are said to be homoscedastic.

1.1 First-order models with quantitative independent variables

Definition 1.1.1: First-order models

A **first-order model** is a model that includes only terms denoting quantitative independent variables which are not functions of other independent variables. The general form is given by

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n.$$

As with the straight-line model, we will use the least squares method to fit first-order models.

1.1.1 Estimating and making inferences about the β parameters

Using the least squares method, we obtain the estimated model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_n X_k$$

such that

1. the average error of prediction is zero: $\sum(Y - \hat{Y}) = 0$;
2. $\sum(Y - \hat{Y})^2$ is minimised.

As with the case of linear regression, the parameter estimates are obtained as a solution of a set of simultaneous equations, which becomes increasingly difficult to compute with more variables. Hence we will use software to obtain these estimates.

Note:-

The estimator of σ^2 for a MRM with k independent variables is given by

$$s^2 = \frac{\text{SSE}}{n - (k+1)}$$

where $k+1$ is the number of estimated parameters.

After obtaining the least squares prediction equation, we now want to interpret the β estimates. When the independent variables are quantitative, the parameters have similar interpretations to that of simple linear regression, except that when we look at the coefficient of one variable, we need to keep the other variables fixed.

For example, consider the first-order model

$$E(Y) = 1 + 2X_1 + X_2.$$

If we graph $E(Y)$ against X_1, keeping X_2 constant, we will obtain a set of straight line with slope equals 2 for each constant value of X_2. This implies that the effect of one X_i on $E(Y)$ is independent of all the other X_i's, and this effect is measure by β_i.

Example 1.1.1

Given the least squares prediction equation for the auction price Y

$$\hat{Y} = -1339 + 12.74X_1 + 85.95X_2,$$

where X_1 represents the age of an antique clock and X_2 is the number of bidders, we interpret the estimates of the β parameters as follows:

1. $\hat{\beta}_1 = 12.74$: We estimate the mean auction price $E(Y)$ of an antique clock to increase \$12.74 for every 1-year increase in age (X_1) when the number of bidders is fixed.
2. $\hat{\beta}_2 = 85.95$: We estimate the mean auction price $E(Y)$ of an antique clock to increase \$85.95 for every 1-bidder increase in the number of bidders (X_2) when age (X_1) is fixed.

In this particular example, $\hat{\beta}_0$ does not have a meaningful interpretation, since when we set all the independent variables to zero, we end up with a negative estimated mean auction price. Moreover, an antique clock with these characteristics is not practical.

Inferences about β parameters

A $100(1-\alpha)\%$ confidence interval for a β parameter is given by

$$\hat{\beta}_i \pm (t_{\frac{\alpha}{2}}) s_{\hat{\beta}_i}$$

where $t_{\frac{\alpha}{2}}$ has $\nu = n-(k+1)$, n is the number of observations and $k+1$ is the number of β parameters in the model.

For a hypothesis test, we use the test statistic

$$t_c = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

with the four assumptions about the probability distribution for the random error.

Example 1.1.2

Referring to previous examples, the collector hypothesises that the auction price of the clocks increases linearly with the number of bidders (X_2).

1. Test the hypothesis that mean auction price of a clock increases as the number of bidders increases when age is held constant. Use $\alpha = 0.05$.
2. Find a 90% confidence interval for β_1 and interpret the result.

From the software, we have seen that $\hat{\beta}_1 = 12.74$, $\hat{\beta}_2 = 85.953$, $s_{\hat{\beta}_1} = 0.905$, and $s_{\hat{\beta}_2} = 8.729$.

Solution:

1. Test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 > 0$. We calculate the test statistic using information given by software ($\hat{\beta}_2 = 85.953$, $s_{\hat{\beta}_2}$):

   $$t = \frac{85.953}{8.729} = 9.85$$

   At $\alpha = 0.05$ with $32 - 3 = 29$ degress of freedom, the critical value is 1.699. Since $t > 1.699$, the test statistic lies in the rejection region. Hence we reject H_0 and conclude that there is significant evidence that the mean auction price of a clock increases as the number of bidders increase when age is held constant.

2. A 90% confidence interval for β_1 is

   $$\hat{\beta}_1 \pm t_{29,0.05} s_{\hat{\beta}_1} = 12.74 \pm 1.699(0.905) = 12.74 \pm 1.54 = (11.20, 14.28).$$

   Hence we are 90% confident that β_1 falls between 11.20 and 14.28, and conclude that the price increases between \$11.20 and \$14.28 for every 1-year increase in age, holding the number of bidders constant.

1.1.2 Evaluating overall model utility

We have used t-tests to make inferences about the β parameters in a MRM. However, doing this to determine which variables are useful for predicting Y has some caveats we have to take note of. If we fail to reject the null hypothesis that $\beta_i = 0$, there are several possible conclusions:

1. there is no relationship between Y and X_i;
2. there is a straight-line relationship between Y and X_i, but a **Type II error** occured;
3. there is a nonlinear relationship between Y and X_i.

The inference that we make from accepting H_0 is that there is no significant evidence that Y and X_i have a **linear** relationship.

Additionally, conducting one test for each parameter make it increasingly likely for us to commit a Type I error as the number of independent variables increase, resulting in us including insignificant variables and excluding useful ones. Hence, to test the overall utility of a MRM, we use a **global test**.

Definition 1.1.2: Multiple coefficient of determination

The multiple coefficient of determination, R^2, is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SS}_y} = \frac{\text{SS}_{yy} - \text{SSE}}{\text{SS}_{yy}} = \frac{\text{Explained variability}}{\text{Total variability}}.$$

Like r^2 in the simple linear model, R^2 represents the proportion of the total sample variation in Y that can be explained by the MRM, and serves as a measure of the usefulness of the entire model. Note that $\text{SSE} = \sum(y_i - \hat{y}_i)$, $_{yy} = \sum(y_i - \bar{y})$. There is a common identity, where the regression sum of squares, RegSS, which is the amount of variability explained by the regression model, is equal to TSS+RSS.

$R^2 = 0$ implies a complete lack of fit of the model to the data, $R^2 = 1$ implies a perfect fit, any other values for R^2 should lie between this range.

Example 1.1.3

In the auction price example, the value of R^2 calculated by the software is 0.8923. This means that using age and number of bidders as the independent variables explains 89.2% of the total sample variation (SS_{yy}) in auction price Y.

Note that a large R^2 does not necessarily mean that the model will fit well to the population. Moreover, we can always obtain a perfect fit by using the same number of parameters as the number of data points by overfitting.

Definition 1.1.3: Adjusted multiple coefficient of determination

The **adjusted multiple coefficient of determination** is given by

$$\begin{aligned} R_a^2 &= 1 - \left[\frac{n-1}{n-(k+1)}\right]\left(\frac{\text{SSE}}{\text{SS}_{yy}}\right) \\ &= 1 - \left[\frac{n-1}{n-(k+1)}\right]\left(1 - R^2\right), \end{aligned}$$

where $k+1$ is the number of β parameters in the model, noting that $R_a^2 \leq R^2$

R_a^2 has similar interpretations to R^2, except that this measure adjusts for both the sample size n and the number of β parameters in the model. This version cannot be forced to 1 by adding more variables and so is preferred by analysts.

Note that unlike R^2, R_a^2 can in fact be negative whenever we have:

$$\begin{aligned} 1 - \left[\frac{n-1}{n-(k+1)}\right]\left(1 - R^2\right) &< 0 \\ \left[\frac{n-1}{n-(k+1)}\right]\left(1 - R^2\right) &> 1 \\ 1 - R^2 &> \frac{n-(k+1)}{n-1} \\ R^2 &< 1 - \frac{n-(k+1)}{n-1} = \frac{k}{n-1}. \end{aligned}$$

Notice that if we used n variables, this ratio would be greater than 1 which is greater than all R^2, hence R_a^2 would give a negative value.

However, these are only sample statistics. Before we make any conclusions about the global usefulness, we conduct a hypothesis test with

$$H_0 : \beta_1 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \text{ At least one of the coefficients is nonzero.}$$

The test statistic used is given by several equivalent versions as follows:

$$\begin{aligned} F &= \frac{(_{yy} - SSE)/k}{\text{SSE}/[n-(k+1)]} = \frac{\text{Mean square (Model)}}{\text{Mean square (Error)}} \\ &= \frac{R^2/k}{(1-R^2)/[n-(k+1)]} \\ &= \frac{R^2(n-k-1)}{k(1-R^2)} \end{aligned}$$

which is the ratio of the explained variability divided by the d.f. in the model, divided by the unexplained variability divided by the d.f. associated with the error. Hence this test is often called the ”analysis-of-variance” F-test (ANOVA). The larger the proportion of the total variability accounted for by the model, the larger the F-statistic.

We reject H_0 when F becomes larger than the tabulated F-value with k d.f. in the numerator and $[n-(k+1)]$ d.f. in the denominator. Also, we make the same assumptions about the random error component.

Note:-

Reject H_0 implies that the model is useful, but not necessarily the best. We use this global F-test to check if we should consider this model further.

Example 1.1.4

Referring to previous examples,

1. find and interpret the adjusted coefficient of determination, R_a^2;
2. conduct the global F-test of model usefulness at the $\alpha = 0.05$ level of significance.

Solution:

1. The R_a^2 value was given by the software to be 0.8849. This means that the least squares model has explained about 88.5% of the total sample variation in y values, **after adjusting for sample size and number of independent variables in the model**.
2. Test $H_0 : \beta_1 = \beta_2 = 0$ vs. H_1 : At least one of the two model coefficients is nonzero.

$$\begin{aligned} F &= \frac{\text{MS(Model)}}{MSE} \\ &= \frac{2,141,531}{17,818} = 120.19 \end{aligned}$$

We use software to calculate the p-value since the degree of freedom in the denominator is too large. We obtain $p < 0,0001 < 0.05$, so we reject H_0 and conclude that there is strong evidence that at least one of the model coefficients is nonzero. The overal model appears to be statistically useful in predicting auction prices.

1.1.3 Using the model for estimation and prediction

The least squares line yields the same value for both the estimate of $E(Y)$ and the prediction of some future value of y. However, the confidence interval for the $E(Y)$ is narrower than the prediction interval for y because of the additional uncertainty added by ϵ when predicting a future value of y, which is not present for $E(Y)$ since the expected value of the random error is equal to 0.

Example 1.1.5 (Estimating $E(Y)$ and predicting y (Auction price model))

Referring to the auction price example, we have the first-order model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where Y is the price of a grandfather clock, X_1 is the age of the clock, X_2 is the number of bidders.

1. Estimate the **average** auction price **for all** 150-year-old clocks sold at an auction with 10 bidders. Use a 95% **confidence interval**. Interpret the result.
2. Predict the auction price for a **single** 150-year-old clock sold at an auction with 10 bidder. Use a 95% **prediction interval**. Interpret the result.
3. Suppose that we want to predict the auction price for one clock that is 50 years old and 2 bidders. How should we proceed?

Solution:

1. We can do this using softwere. The 95% confidence interval for the mean auction price of the clock when $X_1 = 150$ and $X_2 = 10$ is $(1381.4, 1481.9)$. Hence we are 95% confident that the mean auction price for all 150-year-old clocks sold at an auction with 10 bidders lies between $1,381.4 and $1,481.9.
2. The 95% prediction interval for y when $X_1 = 150$ and $X_2 = 10$ is $(1154.1, 1709.3)$. Hence, we are 95% confident that the auction price for a single 150-year-old clock sold at an auction with 10 bidders falls between $1,154.1 and $1,709.3.
3. We can construct a predict interval, **however** we observe that the range for age in the sample is $108 \leq X_1 \leq 194$ and the range for number of bidders is $5 \leq X_2 \leq 15$. Thus, both the selected values fall outside of these ranges, using the model to predict Y for these values may lead to an unreliable prediction.

1.2 Model building in multiple regression

1.2.1 Qualitative (dummy) variable models

MRMs can also include **qualitative** or **categorical** independent variables by coding the **levels**, i.e. values of the variable as numbers before we can fit the model. These coded qualitative variables are called **dummy** (or **indicator) variables** since the number assigned are arbitrary.

One convenient way of coding values of a qualitative variable at two levels is by using 0-1 coding, for example

$$x = \begin{cases} 1 & \text{if male;} \\ 0 & \text{if female.} \end{cases}$$

Then, we can compare the means for males and females in the first-order model easily:

$$\text{Males } (x = 1): \quad E(Y) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$
$$\text{Females } (x = 0): \quad E(Y) = \beta_0 + \beta_1(0) = \beta_0$$

The level of the QlV assigned 0 is called the **base level** or **baseline** and the mean response associated with it will always be β_0. β_1 on the other hand represents the difference between the responses associated with level assigned value 1 and the base level.

$$\beta_0 = \mu_0, \qquad \beta_1 = \mu_1 - \mu_0.$$

We then extend this idea to any finite number of levels. We let one level A be the base level, so we have

$$\mu_A = \beta_0.$$

Then we code the dummy variable of any particular level to equal 1 and the others to equal 0, so that we have

$$\mu_B = \beta_0 + \beta_1,$$

$$\mu_C = \beta_0 + \beta_2,$$

and so on. Then we can interpret each β_i the same way as before: as a difference between the means for that level and the base level:

$$\beta_1 = \mu_B - \mu_A,$$

$$\beta_2 = \mu_C - \mu_A.$$

And so we have our model relating $E(Y)$ to one qualitiative independent variable with k levels:

$$E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1},$$

where X_i is the dummy variable for level $i+1$ defined as

$$X_i = \begin{cases} 1 & \text{if } y \text{ is observed at level } i+1; \\ 0 & \text{otherwise.} \end{cases}$$

Example 1.2.1 (Model with one qualitaative independent variable: Gold ball driving distances)

USGA wants to compare the mean driving distances of four different golf ball brands (A,B,C,D). Iron Byron, the USGA's robotic golfer, is used to hit a sample of 10 balls of each brand. The distance data are as follows:

Brand A	Brand B	Brand C	Brand D
251.2	263.2	269.7	251.6
245.1	262.9	263.2	248.6
248.0	265.0	277.5	249.4
251.1	254.5	267.4	242.0
260.5	264.3	270.5	246.5
250.0	257.0	265.5	251.3
253.9	262.8	270.7	261.8
244.6	264.4	272.9	249.0
254.6	260.0	275.6	247.1
248.8	255.9	266.5	245.9

1. Hypothesise a regression model for driving distance Y, using Brand as an independent variable.
2. Interpret the β's in the model.
3. Fit the model to the data and give the least squares prediction equation. Show that the β-estimates can also be obtained from the sample means.
4. Use the model to determine whether the mean driving distances for the four brands are significantly different at $\alpha = 0.05$.

Solution:

1. Choosing brand A to be the base level, we define the dummy variables as follows:

$$X_1 = \begin{cases} 1 & \text{if Brand B;} \\ 0 & \text{otherwise.} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if Brand C;} \\ 0 & \text{otherwise.} \end{cases} \quad X_3 = \begin{cases} 1 & \text{if Brand D;} \\ 0 & \text{otherwise.} \end{cases}$$

   so the model relating $E(Y)$, the mean distance driven to the single qualitative variable, gold ball brand, is

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

2. - β_0 is the mean distance driven by Iron Byron using gold ball Brand A, μ_A;
   - $\beta_1 = \mu_B - \mu_A$;
   - $\beta_2 = \mu_C - \mu_A$;
   - $\beta_3 = \mu_D - \mu_A$.

   where μ_i is the mean distance for brand i.

3. Using software, we obtain the least squares prediction equation:

$$\hat{Y} = 250.78 + 10.28X_1 + 19.17X_2 - 1.46X_3.$$

   We observe that when using the interpretation in (2), we can compute the same values for the estimates using the sample means.

4. This is equivalent to testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ against H_1 : at least one of the parameters is not 0, since if each population mean is equal then we would have zero values for each parameter.

   Conducting a global F-test on the model, we obtain an F-statistic of $F = 43.99$ using software which gives a p-value of 0.000 which is less than 0.05. Hence, we reject H_0 and conclude that there is sufficient evidence to indicate that the mean driving distance does vary from one fold ball brand to another.

Note:-

THe number of 0-1 dummy variables in a single qualitative independent variable will always be one less than the number of levels of the variable.

1.2.2 Models with both quantitative and qualitative variables

Suppose we have a response which is a function of one quantitative variable X_1 and one qualitative variable. After plotting the lines for each value of the qualitative variables, we want to know if the lines should be considered different, and to that, we need to perform a test on the model. Additionally. if the variance of the random error is the same for each type of medium, the pooled estimated variance is superior to using three separate estimates by fitting separate models.

There are three cases to consider:

1. The straight-line relationship between $E(Y)$ and X_1 is the same for all the values of the qualitative variable. Hence,

$$E(Y) = \beta_0 + \beta_1 X_1.$$

2. The straight lines relating $E(Y)$ to X_1 have different intercepts between values of the qualitative variable but same gradient. Then, the model is a combination of a first-order model and a model witha single qualitative variable:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

   where $X_2, \ldots, X_k$ are 0-1 coded dummy variables for the qualitative variables with $k + 1$ levels. Since the lines have equal gradient, there is no interaction and the terms corresponding to each of the indepedent variables are called **main-effect terms**.

3. The straight lines relating $E(Y)$ to X_1 differ for all values of the qualitative variable. This model has interaction due to presence of cross-product terms between the two independent variables:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$$

   for three levels in the qualitative variable.

Note:-

The cross-product terms are first-order because X_2 and X_3 are dummy variables that introduces or deletes terms in the model. The order of a model is only determined by the quantitative variables.

Example 1.2.2 (Testing for two different slopes - worker productivity data)

In an experiment, the productivity y per worker was measured by recording the number of machined castings that a worker would produce in a four-week period of 40 hours per week. The incentive was the amount x_1 of bonus (cents per casting) paid for all castings produced after the 1000th.

9 workers were selected from two plants. 3 from each plant was given a 20 cent bonus per casting, 3 given 30 cents bonus, and 3 a 40 cents bonus. The productivity data for 18 workers are as follows:

Management style	20 cents	30 cents	40 cents
Traditional	1435, 1512, 1491	1583, 1529, 1610	1601, 1574, 1636
DIsciplined	1575, 1512, 1488	1635, 1589, 1661	1645, 1616, 1689

1. Write a model for mean productivity $E(Y)$, assuming that the relationship between $E(Y)$ and incentive X_1 is first order.
2. Fit the model, and graph the prediction equations for the tranditional and siciplined plants.
3. Do the data provide sufficient evidence to indicate that the rate of increase in worker productivity is different for disciplined and traditional plants? Test at $\alpha = 0.10$.

Solution:

1. The model which describes the productivity as a function of incentive and management style has the form
$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$
where X_1 is the amount of incentive and
$$X_2 = \begin{cases} 1 & \text{if disciplined management style;} \\ 0 & \text{if traditional management style.} \end{cases}$$
2. By using software, we obtain the parameter estimates and least squares regression line
$$\hat{Y} = 1365.83 + 6.217X_1 + 47.78X_2 + 0.033X_1X_2.$$
To obtain the predict equations for both plants, we substitute $X_2 = 0$ and $X_2 = 1$ and obtain
$$\hat{Y} = 1365.83 + 6.217X_1 \quad \text{and} \quad \hat{Y} = 1365.83 + 6.217X_1 + 47.78 + 0.033X_1 = 1413.61 + 6.250X_1$$
respectively. Graphing these two lines shows that their gradients are nearly identical due to the coefficient of the X_1X_2 term being close to zero. However, we cannot make any conclusions before conducting a test.
3. If the rate of increase in worker productivity differs between plants, then the interaction β_3 with differ from 0. So we test $H_0 : \beta_3 = 0$ vs. $\beta_3 \neq 0$. This is simply the t-test from Section 1.1.1.

   By using software, we obtain a test statistic of 0.014 with a p-value of 0.989>0.10. Hence we fail to reject H_0 and conclude that there is insufficient evidence to indicate that the two management styles differ. The test supports our observation of two nearly identical slopes, hence the interaction is not significant and we can drop the X_1X_2 term from the model.

1.2.3 Comparing nested models

We need a statistical method to help us determine with a high degree of confidence which one among a set of candidate models best fits the data. In this section, we try to do this for nested models.

Definition 1.2.1: Nested models

Two models are **nested** is one model contains all the terms of the second model and at least one additional term. The more complex model is the **complete model** and the simpler model is the **reducted model**.

Example 1.2.3

Consider the straight-line interaction model for the mean auction price $E(Y)$ as a function of two quantitative variables: age of the clock, X_1, and the number of bidders, X_2:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

If assume that the relationship between Y and X_1 and X_2 is curvilinear, then the complete second-order model is more appropriate:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2.$$

The curvilinear model and the linear model are nested. The curvilinear model is the complete model and the linear model is the reducted model.

Now we may want to know if the curvilinear model contributes more information relevant to the prediction of Y than the straight-line interaction model does, i.e. whether β_4 and β_5 should be retained in the model. We use an F-test to test

$$H_0 : \beta_4 = \beta_5 = 0 \quad \text{vs.} \quad H_1 : \text{ at least one of the two parameters is nonzero.}$$

1. Check assumptions about random error.
2. Fit the reduced and complete model using the least squares method.
3. Compute the sum of squares for error for the reduced and complete model, SSE_R and SSE_C.
4. Calculate the difference $(\mathrm{SSE}_R - \mathrm{SSE}_C)$. If the additional terms in the complete model is significant, then this difference should be large. We calculate the test statistic by

   $$F_c = \frac{(\mathrm{SSE}_R - \mathrm{SSE}_C)/(k-g)}{\mathrm{SSE}_C/[n-(k+1)]} = \frac{(\mathrm{SSE}_R - \mathrm{SSE}_C)/(k-g)}{\mathrm{MSE}_C},$$

   where MSE_C denotes the mean square error for the complete model, g and k are the number of independent variables in the reduced and complete model respectively, $k-g$ is number of β parameters tested, n is the total sample size.
5. Reject H_0 if $F_c > F_\alpha$, based on d.f. of $\nu_1 = k - g$ and $\nu_2 = n - (k+1)$.

Example 1.2.4 (Analysing a complete second-order model - carnation growth data)

An experiment is conducted to study the growth of carnations in terms of height in cm (Y) as a function of the temperature in degrees Fahrenheit (X_1) and the amount of fertilizer in kg (X_2) applied to the soil.

27 plots of equal size were treated with varying amounts of fertilizer and kept at constant temperatures between 80 and $100°F$. Small carnations with initial height approximately 15cm were planted and height measured after 6 weeks. The data are as follows:

X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y
80	50	50.8	90	50	63.4	100	50	46.6
80	50	50.7	90	50	61.6	100	50	49.1
80	50	49.4	90	50	63.4	100	50	46.4
80	55	93.7	90	55	93.8	100	55	69.8
80	55	90.9	90	55	92.1	100	55	72.5
80	55	90.9	90	55	97.4	100	55	73.2
80	60	74.5	90	60	70.9	100	60	38.7
80	60	73.0	90	60	68.8	100	60	42.5
80	60	71.2	90	60	71.3	100	60	41.4

1. Fit a complete second-order model to the data.
2. Do the data provide sufficient evidence to indicate that the second-order terms $\beta_3, \beta_4, \beta_5$ contribute information relecant to the prediction of Y?

Solution:

1. The complete second-order model is

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2.$$

   Fitting this model using the data given using software, we obtain the least squares prediction equation:

$$\hat{Y} = -5127.90 + 31.10X_1 + 139.74X_2 - 0.145X_1X_2 - 0.133X_1^2 - 1.14X_2^2.$$

2. To determine if the second-order terms $\beta_3, \beta_4, \beta_5$ contribute information relevant to the prediction of y, we perform an F-test of

$$H_0 : \beta_3 = \beta_4 = \beta_5 \quad \text{vs.} \quad H_1 : \text{ at least one of the parameters is not zero.}$$

   First we fit the reduced model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. The least squares prediction equation is

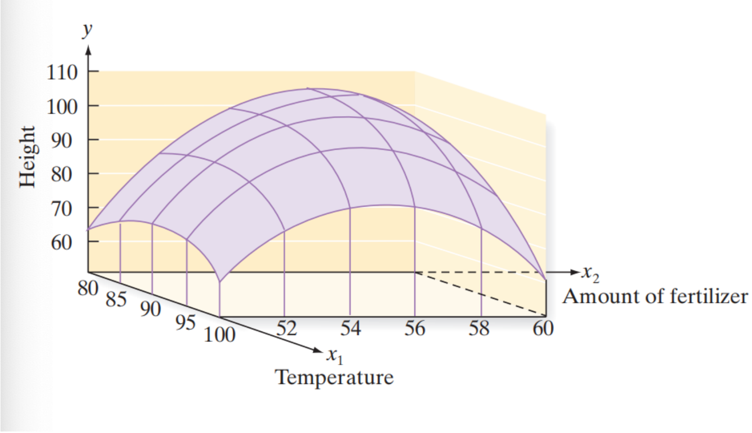
$$\hat{Y} = 106.09 - 0.92X_1 + 0.79X_2.$$

   The sum of squares for error for the complete model and reduced model is 59.17843 and 6671.50852 respectively, and the MSE for the complete model is 2.81802.
   Now we compute the $F - statistic$:

$$F_c = \frac{(6671.50852 - 59.17843)/3}{2.81802} = 782.15.$$

   Comparing this value of F with the tabulated value based on $\nu_1 = 3$ and $\nu_2 = 27 - (5 + 1) = 21$, $F_{0.05} = 3.07$, the rejection region is $F > 3.07$. Since our test statistic lies in the rejection region, we reject H_0 and conclude that there is at least one second-order term that is significant to the prediction of y.

A 3D graph of the prediction model, called a **response surface** is as follows.



The nested-model F-test can be used to determine whether *any* subset of terms should be included in a complete model. Although we are usually cautious about accepting H_0, usually we adopt the principle of parsimony in regression analysis, that is, in situations where two competing models are found to have essentially the same predictive power, the model with the lesser number of β (the more parsimonious model) is selected.

Note that this test is only appropriate when the candidate models are nested models. If the models are not nested, we decide which model to use based on R_a^2 and s, though without performing a test, such decisions are highly subjective in nature.

1.2.4 Stepwise regression

Consider the problem of predicting the salary Y of an executive. The biggest problem is choosing the important independent variables to be included, which may be age, experience, tenure, education level, etc. Hence we need an objective way of deciding which ones to include in our model.

A systematic approach to doing this for a large number of independent variables is difficult because at high orders, the interpretations of multivariable interactions is tedious, hence we use a screening prodedure known as **stepwise regression**.

1. Identify the response Y and the set of potentially important indepedent variables $X_1, \dots, X_k$, where k is generally large and the set may contain higher order terms. (But we usually only include first-order)
2. Software program fits all possible one-variable models of the form
$$E(Y) = \beta_0 + \beta_1 X_i$$
to the data, where X_i is the ith independent variable, for $i = 1, \dots, k$. For each model, the t-test (or equivalent F-test) for a single β parameter is conducted to test the
$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad \beta_1 \neq 0.$$
The X_i that produces the largest absolute t-value is the best one-variable predictor of Y, labelled X_1.
3. The stepwise program searches through the remaining $(k-1)$ independent variables for the best two-variable model of the form
$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_i$$
by fitting all such models to the data. It then performs similar tests and chooses the X_i with the greatest t value. Ideally, the software should go back and check the t-value for $\hat{\beta}_1$ after $\hat{\beta}_2 X_2$ has been added to the model because its significance might (in fact often) change at this step because the meaning of the coefficient $\hat{\beta}_1$ changes. If the t-value *has* become nonsignificant at some specified α, then it removes X_1 and searches for another variable with a β parameter that will give the greatest t-value in the presence of $\hat{\beta}_2 X_2$.

4. Step 2 is repeated for the third variable, checking the t value of β_1 and β_2 and replacing the variables that give nonsignificant t-values. The procedure is repeated until we can no longer add independent variables that yield significant t-values at a specified α in the presence of the current variables.

Note that we have done a very large number of tests, so there is a very high probability that one or more errors have been made in including or excluding variables in the model. Additionally, initially we have omitted high-order terms to keep the number of variables manageable, which may have been important terms. Hence, we should be careful when we use the results of stepwise regression to make inferences, and we say that stepwise regression is actually an **objective variable-screening procedure**.

Ideally, we want to consider second-order terms for quantitative variables and other interactions among variables screened by the stepwise procedure by developing a response surface model using an independent set of data, so that we can verify both results. However, in many situations only a small amount of data is available.

Example 1.2.5

An international management consulting company develops MRMs for executive salaries of its client firms. The consulting company has found that models that use the natural logarithm of salary as the dependent variable have better predictive power than those using salary as the dependent variable (because salaries tend to be incremented in percentages rather than dollar values).

To construct these models, we first need to determine the most important independent variables. For one firm, 7 quantitative and 3 qualitative variables were measured in a sample of 100 executives. Use stepwise regression to decide which variables should be included in the building of the final model.

Solution: Using software, we find the 5 independent variables we should concentrate on. Models with second-order terms and interactions should be proposed and evaluated to determine the best model for predicting executive salaries.

1.3 Model building in multiple regression

1.3.1 Residual analysis: checking the regression assumptions

Recall that for any given set of $x_1, \ldots, x_k$, we assume that the random error ϵ are probabilistically independent and have the distribution $N(0, \sigma^2)$, where σ^2 is a constant. However, these assumptions are rarely satisfied exactly in practical applications, but the least squares regression analysis still produes reliable results as long as the departures are not too great. We discuss how we determine whether or not it is so in this section.

First we want to estimate the random error. The actual random error associated with a particular value of y is the difference between y and its unknown mean (there is a true underlying function describing the relationship, but the observed value does not lie on the line due to added error), so we estimate it by taking the difference between y and the **estimated mean**.

Definition 1.3.1: Residual

A **regression residual** $\hat{\epsilon}$ is the difference between an observed y-value and its corresponding predicted value, i.e. the estimated error:

$$\hat{\epsilon} = (y - \hat{y}) = y - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k).$$

Two useful properties of residuals are as follows:

1. $\sum \hat{\epsilon}_i = \sum (y_i - \hat{y}_i) = 0$, since we are using least squares to calculate $\hat{y}_i$;

2. The standard deviation of the residuals is equal to the standard deviations of the fitted regression model,

i.e.

$$s = \sqrt{\frac{\sum(\hat{\epsilon})^2}{n-(k+1)}} = \sqrt{\frac{\text{SSE}}{n-(k+1)}} = \sqrt{\text{MSE}}.$$

This measure is called the **residual standard error** and measures the lack of fit of the model to the data. It represents how far the predicted values deviate from the true response values on average.

Checking assumption 1: Mean $\epsilon = 0$

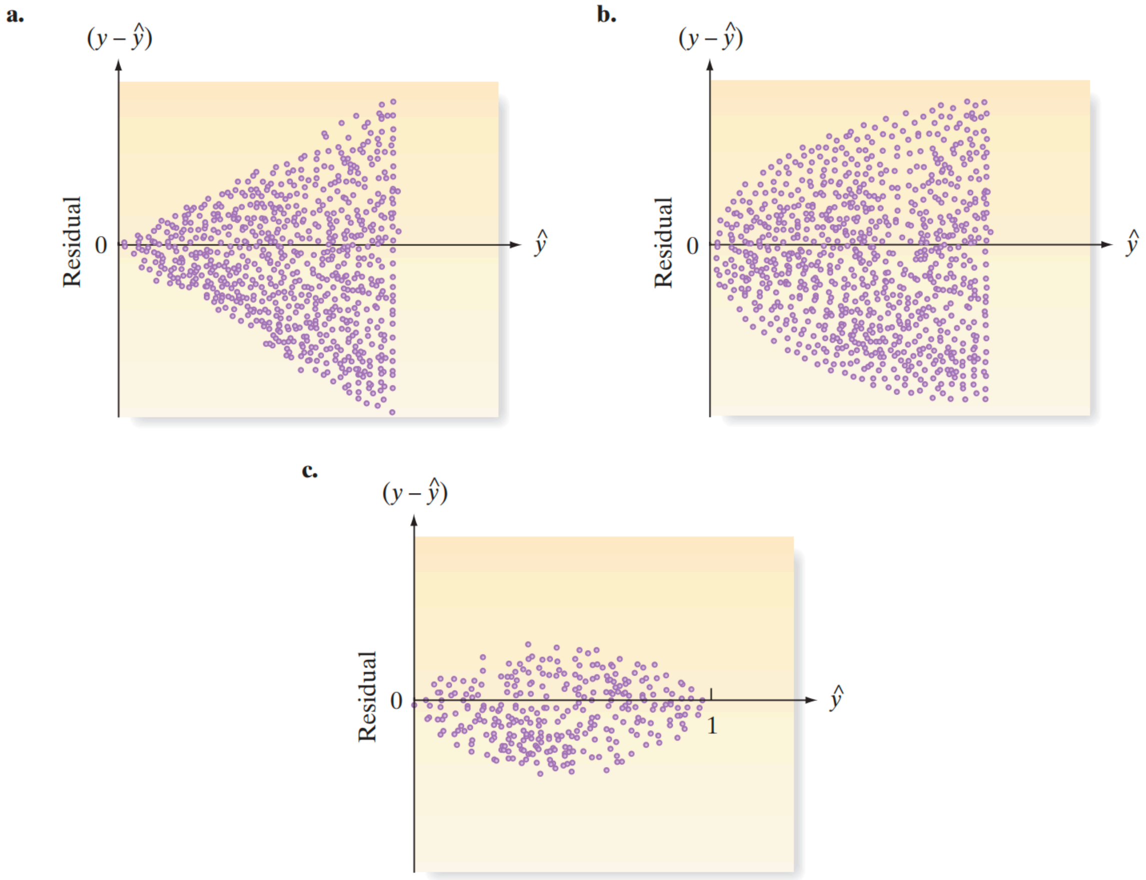
If a model is misspecified, the hypothesised mean of Y, $E_h(Y)$ will not equal the true mean of Y, $E(Y)$, then

$$\begin{aligned} Y &= E_h(Y) + \epsilon \\ \epsilon &= Y - E_h(Y) \\ E(\epsilon) = E(Y - E_h(Y)) &= E(Y) - E_h(Y) \neq 0 \end{aligned}$$

We can check if $E(\epsilon) = 0$ by plotting the residuals of the data against the independent variable and observe its distribution around the zero horizontal line. If the plot exhibits a curved shape, this indicates that curvature needs to be added to the model. Otherwise, we expect the residuals to randomly distribute around the 0 line.

Checking assumption 2: Constant error variance

To check if the error variance is constant, we plot the residuals against the predicted values $\hat{y}$ and check for the patterns shown below:



Notice how the range in values of the residuals increase as $\hat{y}$ increases, implying that the variance of the random error becomes larger as the estimate of $E(Y)$ (which depends on the x values) increases, so the variance of ϵ is not constant for all x.

Example 1.3.1

By using software, we found that the first-order model provides adequent fit to some data of salaries and years of experience of 50 social workers.

- The R^2-value indicates that the model explains about 78.7% of the sample variation in salaries.
- The t-value for testing β_1, $t = 13.31$ is highly significant as it gives a p-value of approximately zero. This indicates that the model contributes information for the prediction of Y.

However, after plotting the residuals against the predicted salaries, the plot indicates a cone shape, implying that non-constant variance. To stabilise it, we refit the model using a **variance-stabilising transformation** on the dependent variable Y. Particularly, we often use $\ln(Y)$ for economic data:

$$\ln(Y) = \beta_0 + \beta_1 X + \epsilon$$

After fitting the new model, there is no longer any apparent tendency of the residual variance to increase as mean salary increases.

Note:-

Other transformations such as $\sqrt{Y}$ or $\arcsin\sqrt{Y}$ are possible. With transformed models, be careful with interpreting model statistics like $\hat{\beta_1}$ and s, noting that the dependent variable is not some function of Y.

For example, in the $\ln(Y)$ model, Y represents the percentage change in salary for every 1 year increase in experience.

Checking assumption 3: Errors normally distributed

We use residuals to check for outliers before assessing whether errors are normally distributed using stem-and-leaf plots, histograms, normal probability plots, QQ-plots, or tests. We consider any residual that lie more than 3 standard deviations away from the mean of 0 to be **regression outliers**.

Once we identify them, we try to find out the cause for these outliers, e.g. human error, and see if we can correct them. In the more often case where we cannot identify the cause, we will want to refit the model excluding the outlier to see how that affects the results of the analysis. Do note that if the outlier was in fact from the same population, we may remove important information and end up with a misleading model.

This assumption is actually the least restrictive in practice as moderate departures from a normal distribution have very little effect on the validity of the statistical tests, CIs, and PIs.

Checking assumptions 4: Errors independent

This assumption is usually violated when the data for the dependent and independent variables are observed sequentially, i.e. **time-series data**, where the experimental unit represents a unit of time. To check graphically, we plot the residuals against time:
If the residuals tend to group alternately into positive and negative clusters as shown above, it is likely that the errors are correlated and the assumption is violated. The methods to solve this problem, such as constructing a time-series model for $E(Y)$, is beyond the scope the reference text.

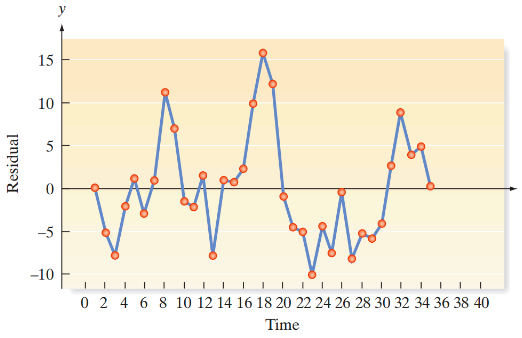
1.3.2 High leverage points

In multiple regression, we can express the model using matrix notation:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \epsilon$$

with predicted responses represented by $\hat{\boldsymbol{Y}} = \boldsymbol{Xb}$ where we obtain $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, so we can rewrite the predicted response as

$$\hat{\boldsymbol{Y}} = \boldsymbol{Hy}$$



where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Hence, each predicted response $\hat{y}_i$ is a linear combination of the observed responsed y_i. The weights, i.e. the diagonal entries of matrix $\boldsymbol{H}$ is called the **leverages**, quantifies the influence of the observed response y_i on the predicted response $\hat{y}_i$. We only consider the diagonal entries because they measure how much each observed y contributes to its predicted y. To compute each h_{ii} we simply take

$$h_{ii} = \boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}^T.$$

In the simple linear regression case, our data/design matrix is

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Then computing h_{ii} gives

$$\begin{aligned}
\boldsymbol{X}^T\boldsymbol{X} &= \begin{pmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{pmatrix} \\
(\boldsymbol{X}^T\boldsymbol{X})^{-1} &= \frac{1}{n\sum x_j^2 - (\sum x_j)^2} \begin{pmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{pmatrix} \\
&= \frac{1}{n\left(\sum x_j^2 - \frac{(\sum x_j)^2}{n}\right)} \begin{pmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{pmatrix} \\
h_{ii} &= \boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i^T \\
&= \frac{1}{n\sum(x_j - \bar{x})^2} \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{pmatrix} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\
&= \frac{1}{n\sum(x_j - \bar{x})^2} \left(\sum x_j^2 - 2x_i \sum x_j + n x_i\right) \\
&= \frac{1}{n\sum(x_j - \bar{x})^2} \left(\sum (x_j - \bar{x})^2 + \frac{(\sum x_j)^2}{n} - 2x_i \sum x_j + n x_i\right) \\
&= \frac{1}{n} + \frac{1}{\sum(x_j - \bar{x})^2} \left(\bar{x}^2 - 2x_i\bar{x} + x_i^2\right) \\
&= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_j - \bar{x})^2}.
\end{aligned}$$

1.3.3 Some pitfalls: estimability, multicollinearity, and extrapolation

Problem 1: Parameter estimability

Imagine you're fitting a first-order model with one independent variable and all the data points form a vertical line. To fit a first-order model we require a minimum of two points with distinct x values. In general, the number of level observed x-values must be one more than the order of the polynomial in x that we want to fit.

Problem 2: Multicollinearity

Multicollinearity exists when two or more independent variables used in a regression are correlated. This is a problem because

1. high correlations among the independent variables increase the likelihood of rounding errors in the calculations of the β estimates, standard errors, etc.;
2. the regression results may be confusing and misleading, e.g. for the model

   $$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

   after fitting to the data, we may find that the t-tests for testing β_1 and β_2 are both insignificant at $\alpha = 0.05$, but the global F-test for $H_0 : \beta_1 = \beta_2 = 0$ is highly significant.

   The t-tests indicate that the contribution of one variable is insignificant after the effect of the other has been accounted for. The F-test indicates that at least one of the two variables is making a contribution to the prediction of Y, in this case probably both, but the contribution of one overlaps with that of the other.

   For another example, multicollinearity may have an impact on the signs of the parameter estimates, which will cause us to interpret the parameters incorrectly.

To avoid this problem, one may conduct a **designed experiment**, which may not be feasible due to time and cost constraints. Hence, we can only collect observational data, which frequently consist of correlated independent variables, so we need to detect when multicollinearity is present and make adjustments accordingly.

One simple way is to conduct a test of correlation between each pair of independent variables in the model. Though note that any number of variables may be highly correlated as a group, but may not exhibit large pairwise correlations. Hence, even if all pairwise correlations are not significantly different from 0, multicollinearity might still be present.

Note:-

Using the correlation coefficient r to detect multicollinearity:

- extreme: $|r| \geq 0.8$;
- moderate: $0.2 \leq |r| < 0.8$;
- low: $|r| < 0.2$.

Then, if multicollinearity is detected, there several things we can do:

1. Drop one or more of the correlated independent variables from the model. One way to decide which variables to drop is by using stepwise regression.
2. If not, then avoid making inferences about the individual β parameters on the basis of the t-tests and restrict inferences about $E(Y)$ and future Y values to the values of the X's within the sample range.

Problem 3: Prediction outside the experimental region

Interesting story about how economists developed a lot of fancy models to predict the state of the economy which all failed to predict multiple economic recessions in history because the inflation rate when the models were developed ranged between 6-8% and so failed to predict future growth in GDP when double-digit inflation rates became a thing.

Chapter 2

Linear model selection and regularisation

In this chapter, we look at some ways to improve the simple linear model by replacing OLS with some alternative fitting procedures which can yield prediction accuracy and model interpretability.

- **Prediction accuracy**: By constraining or shrinking the estimated coefficients, we can often reduce the variance of the least squares estimates at the cost of a negligible increase in bias if our sample size is not large enough.
- **Model interpretability**: We will introduce *feature selection* or *variable selection* to exclude irrelevant variables from a multiple regression model, leading to a more interpretable model.

2.1 Subset selection

2.1.1 Best subset selection

2.1.2 Stepwise selection

2.1.3 Choosing the optimal model

2.2 Shrinkage methods

In this section, we attempt to fit to model while constraining or regularising the coefficient estimates. It turns out that these method scan significantly reduce the variance of the coefficient estimates.

2.2.1 Ridge regression

Recall that in OLS, we estimate β_j by minimising

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

In ridge regression, we instead minimise

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a **tuning parameter**, determined separately, controls the relative impact of the two terms on the regression coefficient estimates. The second term, $\lambda \sum_{j=1}^{p} \beta_j^2$ is a **shrinking penalty** and is small when the true β_j are close to zero. Note that this penalty is not applied to the intercept.

In ridge regression, we produce a set of coefficient estimates, $\hat{\beta}_\lambda^R$ for each λ. It is critical that we select a good value for λ which we will discuss in later sections.

Application: Credit data

We apply ridge regression to the Credit dataset:

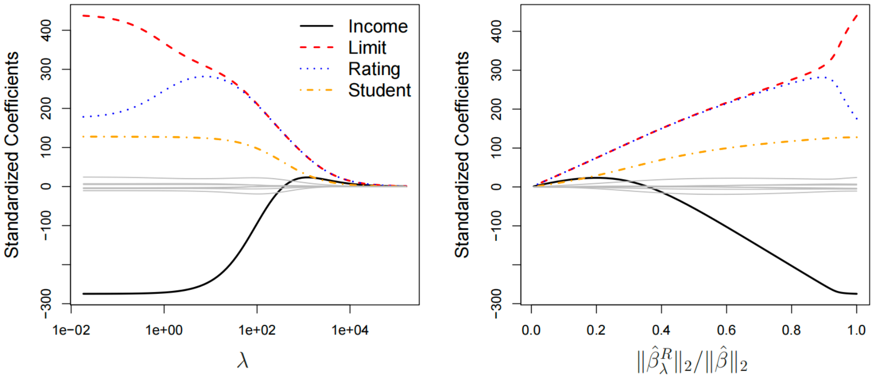


Figure 2.1: The standardised ridge regression coefficients are displayed for the Credit dataset, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.

In the left panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ. When $\lambda = 0$, the shrinking penalty is removed so the coefficient estimates obtained is the same as the least squares estimates. As we increase λ to infinity, all the estimates are shrunken to zero, leading to the null model.

In the right panel, the x-axis represents the ratio of how much the estimates has been shrinken. Starting from $\lambda = 0$, $\|\hat{\beta}_\lambda^R\|_2 = \|\hat{\beta}\|_2$ so the ratio is 1. As λ approaches infinity, the ridge estimate shrinks to 0, so the ratio becomes 0.

Note:-

There are different types of norms:

1. The l_0 norm gives the number of nonzero entries within a vector, and is technically not a norm as it does not fulfill the definition of a norm. Nevertheless, it has some applications.
2. The l_1 norm is also known as Manhattan distance. For example, $\|(2,3,4)\|_1 = 2+3+4 = 9$. Consider this geometrically!
3. The l_2 norm is also known as the Euclidean norm. For example, $\|(3,4)\|_2 = \sqrt{3^2+4^2} = 5$.

Standardising predictors before apply ridge regression

The least squares coefficient estimates are **scale equivariant**. For example, we can measure the `income` variable by dollars or thousands of dollars and we would just have to rescale the corresponding coefficient estimate accordingly.

In the case of ridge regression, it is not that simple, and it may even affect other predictors. Hence, before we perform ridge regression we need to standardise the predictors by

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}},$$

so that they are on the same scale and all have a standard deviation of one. Hence, the final fit will not depend on the scale on which the predictors are measured.

The bias-variance tradeoff

Bias associates with accuracy while variance is associated with precision. There is a tradeoff between the two when considering the number of predictors used in the model: more predictors leads to lower bias but greater variance and vice versa.

This tradeoff also occurs in the context of ridge regression: as λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias. We illustrate this using a simulated dataset of $p = 45$ predictors and $n = 50$ observations.

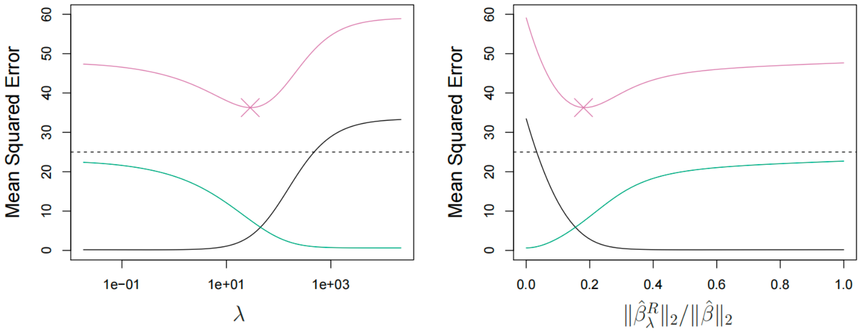


Figure 2.2: Squared bias (black), variance (green), test mean squared (magenta) for the ridge regression predictions on a simulated dataset, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. Horizontal dashed lines indicate the minimum possible MSE. The purple corsss indicate the ridge regression models for which the MSE is smallest.

From the left panel, we see that as variance (green) decrease with increasing λ, bias (black) increases. The magenta line represents the test MSE which is the sum of the variance, bias, and irreducible error. We can find the value of λ which minimises the test MSE by using cross validation. As a result, we greatly reduce variance at a cost of negligible bias.

Note:-

We can decompose the test MSE at a point x into square bias, variance, and irreducible error:

$$\mathbb{E}[(Y - \hat{f}(x))^2] = (\mathbb{E}[\hat{f}(x)] - f(x))^2 + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + \sigma^2 \,.$$

Derivation:

$$\begin{aligned}
\mathbb{E}[(Y - \hat{f}(x))^2] &= \mathbb{E}[(f(x) + \varepsilon - \hat{f}(x))^2] \\
&= \mathbb{E}[(f(x) - \hat{f}(x))^2 + 2\varepsilon(f(x) - \hat{f}(x)) + \varepsilon^2] \\
&= \mathbb{E}[(f(x) - \hat{f}(x))^2] + \mathbb{E}[2\varepsilon(f(x) - \hat{f}(x))] + \mathbb{E}[\varepsilon^2] \text{ by linearity,} \\
&= \mathbb{E}[(f(x) - \hat{f}(x))^2] + 0 + \sigma^2 \;\text{ since } \varepsilon \text{ is independent of } \hat{f}(x), \\
&= \mathbb{E}[(f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2] + \sigma^2 \\
&= \mathbb{E}[[(f(x) - \mathbb{E}[\hat{f}(x)]) - (\hat{f}(x) - \mathbb{E}[\hat{f}(x)])]^2] + \sigma^2 \\
&= \mathbb{E}[(f(x) - \mathbb{E}[\hat{f}(x)])^2 - 2(f(x) - \mathbb{E}[\hat{f}(x)])(\hat{f}(x) - \mathbb{E}[\hat{f}(x)]) + (\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + \sigma^2 \\
&= \mathbb{E}[(f(x) - \mathbb{E}[\hat{f}(x)])^2] - \mathbb{E}[2(f(x) - \mathbb{E}[\hat{f}(x)])(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])] + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + \sigma^2 \\
&= (f(x) - \mathbb{E}[\hat{f}(x)])^2 - 2(f(x) - \mathbb{E}[\hat{f}(x)])\mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])] + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + \sigma^2 \\
&= \mathbb{E}[\hat{f}(x)] - (f(x))^2 + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + \sigma^2
\end{aligned}$$

Note: we have $\mathbb{E}(\varepsilon^2) = \sigma^2$ because $\mathrm{Var}[\boldsymbol{X}] = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T) - \mathbb{E}(\boldsymbol{X})[\mathbb{E}(\boldsymbol{X})]^T$ and $\mathbb{E}(\varepsilon) = 0$.

2.2.2 Lasso

Ridge regression has one main disadvantage: it shrinks every coefficient estimate towards zero as you increase λ but it will never eliminate any irrelevant variables. This can give rise to difficulties in interpreting the model when the number of predictors is large.

To overcome this disadvantage, we use **lasso**. The lasso coefficients, $\hat{\beta}^L_\lambda$ minimise the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Notice that it is essentially similar to ridge regression except instead of using l_2, we are using the l_1 norm. It also shrinks the coefficient estimates towards zero like ridge regression, except that the l_1 penalty has the effect of forcing some of the estimates to zero when λ is large enough.

Hence, lasso performs variable selection and is said to yield **sparce models**, that is, models that involve a subset of the variables.

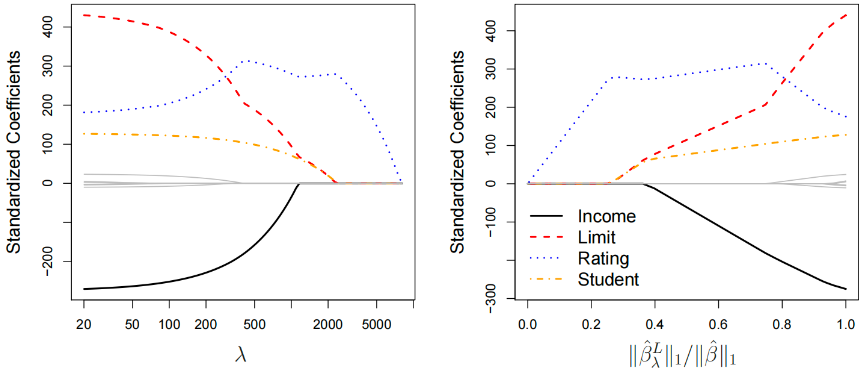


Figure 2.3: The standarised lasso coefficients on the Credit dataset are shown as a function of λ and $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Let $\boldsymbol{A} = \begin{pmatrix} \boldsymbol{u} & k\boldsymbol{u} & \boldsymbol{v} \end{pmatrix}$ where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ are arbitrarily linearly independent column vectors containing n observations of a particular feature with $\boldsymbol{u}^T = \begin{pmatrix} u_1 & \dots & u_n \end{pmatrix}$ and $\boldsymbol{v}^T = \begin{pmatrix} v_1 & \dots & v_n \end{pmatrix}$, $k \in \mathbb{R}$ is such that the column $k\boldsymbol{u}$ is a scalar multiple of the first column. Then we consider

$$
\begin{aligned}
\boldsymbol{A}^T\boldsymbol{A} &= \begin{pmatrix} \boldsymbol{u}^T \\ k\boldsymbol{u}^T \\ \boldsymbol{v}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{u} & k\boldsymbol{u} & \boldsymbol{v} \end{pmatrix} \\
&= \begin{pmatrix} u_1 & \dots & u_n \\ ku_1 & \dots & ku_n \\ v_1 & \dots & v_n \end{pmatrix} \begin{pmatrix} u_1 & ku_1 & v_1 \\ \vdots & \vdots & \vdots \\ u_n & ku_n & v_n \end{pmatrix} \\
&= \sum_{i=1}^{n} \begin{pmatrix} u_i \\ ku_i \\ v_i \end{pmatrix} \begin{pmatrix} u_i & ku_i & v_i \end{pmatrix} \\
&= \sum_{i=1}^{n} \begin{pmatrix} u_i^2 & ku_i^2 & u_i v_i \\ ku_i^2 & k^2 u_i^2 & ku_i v_i \\ u_i v_i & ku_i v_i & v_i^2 \end{pmatrix}.
\end{aligned}
$$

Since for every i, the second row of $\boldsymbol{A}^T\boldsymbol{A}$ is k times the first row, adding each rank 1 matrix, the second row of the final matrix will also be k times the first, so $\boldsymbol{A}^T\boldsymbol{A}$ is rank-deficient and not full rank, and hence is singular.

Note that writing matrix-matrix products in this form, each term in the sum is a rank 1 matrix, however, under no collinearity, the rows in every term will not necessarily be dependent on some other rows **by the same factor**. Notice that actually in our formulation, the third row differs from the first row by a multiplicative factor of v_i/u_i (this is consistent with the case of collinear columns since $ku_i/u_i = k$). Since under no collinearity, this fraction will differ for varying i, so after adding each term up, there is no common factor between the first and third row.

Alternatively for a simpler(?) approach, we look at each entry in the product matrix as multiplying rows with columns:

$$
\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} \boldsymbol{u} \cdot \boldsymbol{u} & \boldsymbol{u} \cdot (k\boldsymbol{u}) & \boldsymbol{u} \cdot \boldsymbol{v} \\ (k\boldsymbol{u}) \cdot \boldsymbol{u} & (k\boldsymbol{u}) \cdot (k\boldsymbol{u}) & (k\boldsymbol{u}) \cdot \boldsymbol{v} \\ \boldsymbol{v} \cdot \boldsymbol{u} & \boldsymbol{v} \cdot (k\boldsymbol{u}) & \boldsymbol{v} \cdot \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u} \cdot \boldsymbol{u} & k\boldsymbol{u} \cdot \boldsymbol{u} & \boldsymbol{u} \cdot \boldsymbol{v} \\ k(\boldsymbol{u} \cdot \boldsymbol{u}) & k(k\boldsymbol{u} \cdot \boldsymbol{u}) & k(\boldsymbol{u} \cdot \boldsymbol{v}) \\ \boldsymbol{v} \cdot \boldsymbol{u} & k(\boldsymbol{v} \cdot \boldsymbol{u}) & \boldsymbol{v} \cdot \boldsymbol{v} \end{pmatrix}.
$$

We observe that the second row is k times the first row as expected. The third row again is not collinear with the first two. Suppose they are, then we solve for some $\lambda \in \mathbb{R}$ such that

$$
\boldsymbol{v} \cdot \boldsymbol{u} = \lambda(\boldsymbol{u} \cdot \boldsymbol{u}) \quad \text{and} \quad \boldsymbol{v} \cdot \boldsymbol{v} = \lambda(\boldsymbol{u} \cdot \boldsymbol{v}).
$$

It then follows that

$$
\begin{aligned}
\|\boldsymbol{v}\|^2 &= \boldsymbol{v} \cdot \boldsymbol{v} \\
&= \lambda(\boldsymbol{v} \cdot \boldsymbol{u}) \\
&= \lambda^2(\boldsymbol{u} \cdot \boldsymbol{u}) \\
&= \lambda^2\|\boldsymbol{u}\|^2 \\
\|\boldsymbol{v}\| &= \lambda\|\boldsymbol{u}\| \\
&= \frac{\boldsymbol{v} \cdot \boldsymbol{u}}{\|\boldsymbol{u}\|^2}\|\boldsymbol{u}\| \\
\boldsymbol{v} &= \frac{\boldsymbol{v} \cdot \boldsymbol{u}}{\|\boldsymbol{u}\|^2}\boldsymbol{u}
\end{aligned}
$$

i.e. $\boldsymbol{v}$ is its own orthogonal projection on vector $\boldsymbol{u}$, implying that they are collinear, which is a contradiction. Hence the third row is linear independent of the first two rows.