1 Predicting the quality and prices of wines

Tool: Linear regression.

The Analytics Edge: The price of mature wines can be predicted from data available when grapes are picked. Using a linear regression model with weather variables it is possible to develop good predictive models of wine prices. Traditionally the quality of wines are predicted by wine experts, based on tasting samples. The analytics edge here is provided by identifying a new set of variables that were traditionally not used to infer the quality of wines.

1.1 Overview

Bordeaux is a region in France that is well-known for making wines. The major reason for the success of the wines made in this region is the excellent environment that is conducive for growing vines in Bordeaux. Roughly 90% of the wines produced in Bordeaux are red wines. Often these wines are recognized as some of the finest in the world.

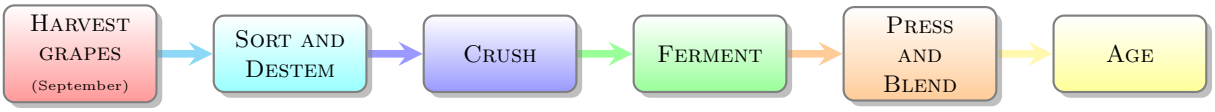

Figure 1: Schematic of wine making process

Much of the wine in the region has been produced in the same way for hundreds of years yet there are significant differences in the quality and prices of the wines from year to year. Dr. Orley Ashenfelter, a professor at Princeton University developed a simple but powerful approach to predict the quality and prices of Bordeaux wines. Bordeaux wines taste better when they are older and hence there is an incentive to store them till they come of age. The younger wines are typically not particularly pleasant to drink.

Key Question: Can one predict how good a wine will be when it matures?

This is useful since *en primeur* or *wine futures* give people an opportunity to buy wines early and invest in them before it is bottled. This is based on some tasting samples of wine within a year or two after it is made and much before it ages. Wine experts give scores (wine ratings), based on such tastings. One such wine that is valued very highly is the 1982 vintage wine of Chateau Latour that was sold at 250 pounds a case *en primeur* in 1983 and was valued at 9000 British pounds in 2007.

Some of the possible predictors of the quality of wine are:

1. Vineyard (chateau - location where wine is made)
2. Vintage (year - time when wine is made)

Ashenfelter focused on the vintage as a predictor for the quality of wine by averaging auction prices across chateaus. From the data, one can observe that:

1. Older the wine is, the greater is the value.
2. There is still significant variation in average prices that remains unexplained.

To explain the quality of wine better (as approximated by the price of wine), Ashenfelter proposed the use of weather variables as a good predictor of quality. In Bordeaux, the weather changes significantly from year to year that led him to believe it to be a possibly good predictor.

To study this approach, we will use data from the website `www.liquidasset.com` with the following variables:

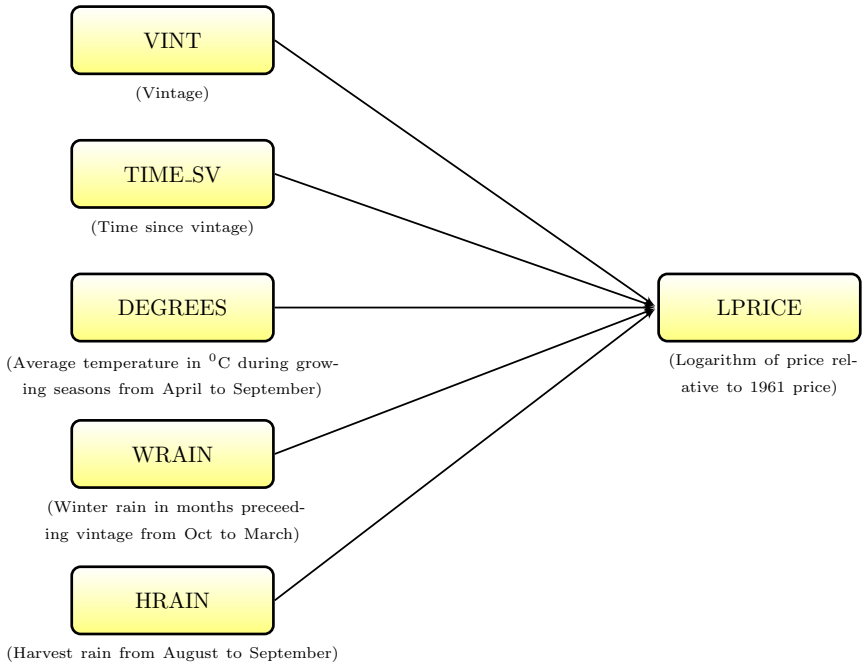

Figure 2: Variables used from data in `www.liquidasset.com`

We will build a linear regression model to predict the price of wine from these variables.

1.2 Summary

Data: Source is `http://www.liquidasset.com`. The data consists of the prices of wine from auctions and weather information for the vintage. The dataset was from years 1952 to 1989 (fairly small dataset).

Model: Linear regression is used to predict wine quality (price) in terms of vintage, summer temperature, winter rain and harvest rain.

Decision: The model develops a prediction on the quality of wine that is known only when it matures much later using weather information that is available at the time of making the wine. Such predictions are useful for people who invest in wine.

Value: The predictions are comparable to and sometimes beat the predictions of the best experts using an elegant model.

The results from the predictions indicate that 1989 wine would be of very high quality. How did the predictions compare with the predictions of the best wine critics?

1. Ashenfelter predicted 1986 wine to be mediocre due to a below average growing season temperature and above average harvest rainfall. Robert Parker on the other hand predicted this wine to be very good and sometimes exceptional.
2. Ashenfelter predicted the 1989 vintage to be excellent and 1990 even better. At first Robert Parker predicted this to be similar to the 1985 vintage but then later said it was the vintage of the century.
3. Ashenfelter's model and Parker's expert opinion both agree that the 2000-2001 vintage would be very high quality wine.

	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990	1989	1988	1987	1986
AUSTRIA RIESLING & GRÜNER VELTLINER		89-93E	90-94T	88-92E	95E	78E	91T	89T	89E	88I	89I	88R	90R	91I	87I	88I	89I	89T	88I	85C	95R	82C	96R	84C	90I	87C	89C	88C	84C	94R	NT	NT	NT	96R
FRANCE																																		
ALSACE		NT	NT	90-92E	90-94	89I	86T	87R	86E	89I	82I	89R	91R	79I	87R	86E	82R	86R	91R	90R	87R	90R	87R	87R	89R	90R	87R	85C	78C	93R	93R	88R	83C	82C
BORDEAUX: ST JULIEN/PAUILLAC/ST ESTEPHE		NT	91I	97T	94T	93T	81C	92E	88E	98T	99E	91E	86E	87E	95T	88T	88R	96T	88R	96T	88R	87T	84R	96T	92T	85C	78C	79C	75R	98E	90E	87R	82R	94T
BORDEAUX: MARGAUX		NT	90I	97T	94T	90T	80C	89E	87E	95T	97E	90E	86E	88E	98T	87T	88I	88T	89E	94T	89R	86T	82R	88T	88E	85C	77C	75C	74R	90E	86E	85R	76R	90T
BORDEAUX: GRAVES/PESSAC LEOGNAN		NT	92I	97T	96T	93T	81C	91E	86E	99T	98E	91E	87E	87E	96T	88T	88I	87T	88R	97T	88R	94T	86R	86E	89E	88E	86C	75C	74R	90R	89R	89R	84R	89E
BORDEAUX: POMEROL		NT	90I	97T	96T	94T	84R	94E	88R	95T	98E	96E	86I	90T	95T	88E	84E	85E	90E	95T	88R	96T	87R	85E	92T	89T	87C	82C	58C	96R	96R	89R	85C	87T
BORDEAUX: ST EMILION		NT	89I	95T	95T	92T	82R	93E	87R	94T	93E	92E	86I	88E	99T	88E	90I	87E	90E	96T	88R	96T	86R	87T	88E	86T	84C	75C	59C	98R	88R	88R	74C	88E
BORDEAUX: BARSAC/SAUTERNES		NT	91E	92E	95E	92E	92E	88E	93E	90R	97E	89R	94R	88R	96R	87R	89R	85R	93R	88E	88E	87E	89E	87E	85E	78E	70C	68C	70C	98R	90R	98R	70R	94R
BURGUNDY: COTE DE NUITS (RED)		93T	93E	97T	98T	92E	92E	95T	91E	96E	95E	88I	91E	89I	98T	83C	89R	94R	91R	87R	92R	84I	89R	89T	90R	72C	85C	68C	86R	93R	85C	84C	77C	65C
BURGUNDY: COTE DE BEAUNE (RED)		94T	92E	92E	96T	91E	89E	91E	90E	94E	95E	89I	90E	82I	96T	79C	87T	92R	88R	86R	93R	82C	88R	89R	85R	73C	80C	82R	77C	92R	86C	87C	79C	72C
BURGUNDY (WHITE)		92E	96E	87I	94E	97E	90E	92E	91E	94E	91E	94E	91I	90E	88E	91R	84R	92R	86C	88C	89C	84C	89C	92C	93C	77C	72C	90C	70C	87C	90C	82R	79C	82C
BURGUNDY: BEAUJOLAIS	94E	93E	95E	95I	90I	93R	86R	90R	91R	93R	97R	86R	85R	89R	95R	81C	93R	86C	75C	91R	89R	84C	87C	82C	87C	85C	80C	77C	88C	86C	92C	86C	85C	84C
CHAMPAGNE		NT	NT	NT	NT	92E	95T	96T	87R	89R	92R	99T	80R	86R	88I	90T	88I	95T	88R	92R	92R	93R	90I	97T	95T	NT	88E	NT	NT	93R	95R	95R	NT	NT
JURA		93E	91I	93T	94T	93E	90I	93E	91R	94E	90R	91T	91I	90E	96R	92R	94R	93R	85R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
LANGUEDOC		NT	91T	92E	93T	87E	88E	88I	91E	94T	91R	87R	93R	90I	88R	88R	87I	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
ROUSSILLON		NT	91T	92E	93T	88E	94T	89E	91E	94T	91R	87R	92R	90I	88R	88R	87I	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
LOIRE VALLEY (WHITE)		NT	NT	88-91E	NT	91R	80I	83E	86R	92T	88I	90I	84I	83E	94E	82C	82R	96R	82C	84R	84R	84C	88C	91R	88C	87C	86C	80C	75C	90R	92R	88C	87C	87R
LOIRE VALLEY (RED)		NT	NT	88-91E	NT	88E	78E	82E	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
RHONE: COTE ROTIE/HERMITAGE		NT	95T	92E	97T	87E	89T	91E	92E	97T	98T	79I	89E	92E	89R	85C	96R	78C	89R	87E	95T	90T	90E	86R	90T	88C	58C	78R	92R	92R	92T	92R	86R	84C
RHONE: CHATEAUNEUF DU PAPE		NT	94T	98E	93T	87E	88E	92E	88R	98T	93E	86R	98E	92R	95T	88R	90I	58C	96T	98E	90E	98E	82C	82C	90T	86C	85C	78C	65C	95R	94T	88R	60C	78C
GERMANY																																		
MOSEL SAAR RUWER		89-93	91-95T	93E	95R	78I	79I	NT	95R	89I	95I	82I	92R	95I	94R	92R	91I	91R	91R	76C	86E	92T	88R	91T	90R	94R	91R	87C	88R	96R	91C	92R	84C	85C
RHEINHESSEN (RIESLING)		NT	NT	90-92E	92R	81I	83I	90T	94R	87I	93R	88R	92R	86I	92R	93T	89I	92R	95R	69C	87E	93T	87R	91T	86C	87R	88R	85C	87R	96R	90C	90R	84R	84C
GERMANY (PINOT NOIR/SPÄTBURGUNDER)		92-95	91-92	88-91	NT	NT	91I	89E	90T	87T	91T	87E	88R	89R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
		2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990	1989	1988	1987	1986

	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990	1989	1988	1987	1986
USA/OREGON																																		
WILLAMETTE VALLEY PINOT NOIR	NT	NT	90E	94T	93R	92R	87R	92R	90R	88I	86R	94T	84T	91E	85T	86E	88E	92I	85E	86E	92E	89T	87C	83C	76C	92R	89R	88R	87C	90C	86C	88C	72C	85C
USA/WASHINGTON																																		
CABERNET SAUVIGNON / SYRAH	NT	NT	93T	90E	86I	94T	92T	95T	87E	92T	92T	91T	95R	91R	94R	91R	90R	89R	92R	89R	90T	90T	88T	88T	87R	90R	87R	89R	85C	87R	92R	88R	90R	78R
ARGENTINA																																		
ARGENTINA		95E	92E	96I	92I	92R	95R	91R	94R	92R	90R	91R	92R	94R	93R	91R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
AUSTRALIA																																		
SOUTH AUSTRALIA: BAROSSA / MCLAREN VALE		NT	89T	91T	94T	87R	94T	94T	79I	96T	89T	85I	85R	94R	96R	94R	90C	95R	95C	88C	88E	95E	88R	90E	87R	90R	87R	87R	89R	88R	88C	85C	87C	90R
WESTERN AUSTRALIA		NT	90T	92T	91T	90T	91T	91T	90T	92R	89R	92R	87R	82I	91R	88R	89R	90R	90C	88E	89R	90R	87R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
NEW SOUTH WALES		NT	88T	85I	89I	94T	81I	78I	82I	89R	85I	79I	90R	84R	87R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
VICTORIA / TASMANIA		NT	87T	91T	93T	90T	93E	94T	78I	92T	87I	85R	85I	93R	91R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
CHILE																																		
CHILE		96E	91E	93I	92E	91I	93R	87R	92R	90E	88R	90R	88R	89R	90R	89R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
NEW ZEALAND																																		
NEW ZEALAND			84E	85E	86E	90E	91R	82I	88R	91R	89R	81I	91R	90R	85C	83C	78C	87C	76C	81I	80I	90R	86I	89R	84I	87R	83C	86I	86C	85C	88C	80C	77C	85C
SOUTH AFRICA																																		
SOUTH AFRICA	NT	NT	96E	91R	96E	92E	93R	91E	89I	91R	93R	88R	90R	92R	87R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990	1989	1988	1987	1986

RATINGS	
96-100	Extraordinary
90-95	Outstanding
80-89	Above Average to Excellent
70-79	Average
60-69	Below Average
<59	Appalling

MATURITY	
C	Caution, may be too old
E	Early maturing and accessible
NV	Vintage not declared
I	Irregular, even among the best wines
NT	Not yet sufficiently tested to rate
R	Ready to drink
T	Still tannic, youthful, or slow to mature

1.3 Linear Regression

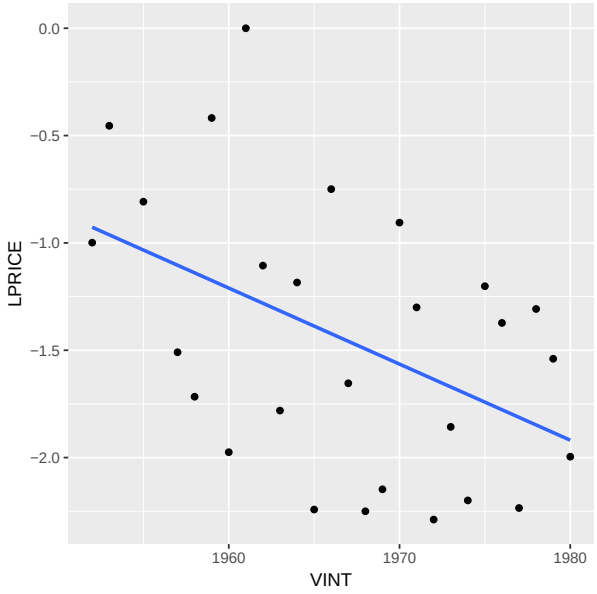

Figure 3: Linear model

Problem setup:

1. $n =$ Number of observations,
2. $p =$ Number of predictor variables (excluding the constant 1),
3. $y =$ Dependent variable in $\mathbb{R}$,
4. $x_1, \ldots, x_p =$ Independent variables (predictors).

We are interested in estimating the linear model:

$$y \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon,$$

where ϵ is the error term that models noise which is not captured by the predictor variables. The data consists of observations

$$\{y_i, x_{i1}, \ldots, x_{ip}\} \text{ for } i = 1, \ldots, n.$$

The coefficients in the multiple linear regression model are chosen to minimize the sum of squared of errors (residuals) given as:

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}\right)^2.$$

Key ideas:

1. Let us setup the optimization problem in vector and matrix notation as follows. Define:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

   Hence we have our model to be

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

   and we can rewrite the problem as:

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}).$$

   Now taking a derivative with respect to $\boldsymbol{\beta}$ we get (*normal equations*):

$$(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})\boldsymbol{\beta} - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{y} = 0$$

   Hence the optimal solution is given by:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}$$

   where the fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{x}\hat{\boldsymbol{\beta}}$. Hence we have the fitted values and errors as

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip},$$
$$e_i = y_i - \hat{y}_i.$$

   Also note that this is a convex (quadratic) minimization problem, and the solution actually gives a global minimum.

2. The estimates have standard errors associated with them. This is based on the frequentist interpretation that we are developing the linear regression estimates using an observed data set that is sampled from a true population distribution.

   A usual assumption is that the errors ϵ_i are independent and identically distributed mean 0 and variance σ^2. It is often also assumed that $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Moreover assume that $\boldsymbol{x}$ is fixed. Considering $\boldsymbol{Y}$ as the random variable representing $\boldsymbol{y}$, we have $\mathrm{Var}(\boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$.

   Note that for an $n \times p$ fixed matrix $\boldsymbol{A}$, $\mathrm{Var}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{Y})\boldsymbol{A}^{\mathsf{T}} = \sigma^2\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}$. Therefore,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}) = \sigma^2(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}.$$

   Since the true variance σ^2 is unknown, we estimate it using the data and the regression fit as follows:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2.$$

   The division by the number $n-p-1$ is to ensure that the estimator is unbiased such that $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$. The standard error of the coefficients is equal to the square root of the diagonal elements of the matrix $(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\sigma^2$.

3. One of our main goals is to check whether the jth variable in the model is useful. Therefore we want to test the null hypothesis

$$\mathbb{H}_0 : \beta_j = 0 \qquad \text{vs.} \qquad \mathbb{H}_1 : \beta_j \neq 0$$

We know that under assumption of $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$, we have $\beta_j \sim N(\beta_i, ((\boldsymbol{x}^\mathsf{T}\boldsymbol{x})^{-1})_{jj}\sigma^2)$ and if we estimate σ^2 by $\hat{\sigma^2}$ then

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})^{-1})_{jj}}}$$

follows a t-distribution with $n-p-1$ degrees of freedom. This is a t-statistic and if the absolute value of the t-statistic is high, the null hypothesis $\mathbb{H}_0$ will be rejected in favor of $\mathbb{H}_1$. This indicates statistically that the j-th variable is a significant predictor in the model and the p-value provides the probability of seeing a t-statistic as extreme as we observe under the null hypothesis.

1.3.1 Quality of fit

1. Let $\bar{y} = \sum_{i=1}^n y_i/n$. With the regression estimates $\hat{y}_i$ define:

$$\texttt{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \text{ (Sum of squared errors)}$$
$$\texttt{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \text{ (Sum of squares due to regression)}$$
$$\texttt{SST} = \sum_{i=1}^n (y_i - \bar{y})^2 \text{ (Total sum of squares)}$$

In linear regression, with the optimal estimates, we have:

$$\texttt{SST} = \texttt{SSE} + \texttt{SSR}.$$

The residual standard error is defined as

$$\sqrt{\texttt{SSE}/(n-p-1)}$$

and measures the lack of fit of the model. It is possible for models with more variables to have a higher residual standard error if the decrease in SSE is small relative to the increase in p.

The proportion of the variance in the dependent variable that can be accounted by for the variation in the independent variables is defined as R-squared or coefficient of determination:

$$R^2 = \frac{\texttt{SSR}}{\texttt{SST}} = 1 - \frac{\texttt{SSE}}{\texttt{SST}} \text{ (R-squared or Coefficient of determination)}$$

R^2 is always between 0 and 1 and provides information on the goodness of the fit of the model. For example:

(a) If the regression fit is a horizontal line then this implies that $R^2 = 0$ (the predictor variables have no explanatory power).

(b) If the regression fits perfectly all points on a straight line, this implies $R^2 = 1$ (the predictor variables have perfect explanatory power)

(c) All the values of y_i lie in the same vertical line implies R^2 cannot be computed.

As we increase the number of predictor variables in the model, R^2 will never decrease (it will stay the same or increase). Hence it is important to be careful in using this to do model selection as you might overfit data. Furthermore, a good value of R^2 might be very different for a variety of applications. For example in finance, it is hard to predict stock prices and so even a useful model might have a small value of R^2 because the problem is challenging. On the other hand, a less useful model for an easier problem such as predicting revenue from the number of items sold might have a high R^2.

For simple linear regression with a single variable:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon,$$

the R^2 value is simply the Correlation$(X_1, Y)^2$.

2. The adjusted R^2 statistic penalizes the R^2 statistic as more variables are added to the fit. The adjusted R^2 value can be negative and its value will always be lesser than or equal to R^2. The adjusted R^2 increases when a new explanatory variable is added such that the increase in the fit is more than that expected by chance. The adjusted R^2 is one of the useful measures in selecting predictor variables in the final model building. It is defined as:

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right)$$

3. The F-statistic is used to test joint hypotheses. Suppose we want to test

$$\mathbb{H}_0 : \ \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$\text{vs.}$$
$$\mathbb{H}_1 : \text{At least one of the } \beta_j \text{ is nonzero.}$$

This means we want to test whether the linear regression model is useful at all. The F-statistic is defined as:

$$\text{F-statistic} = \frac{\texttt{SSR}/p}{\texttt{SSE}/(n-p-1)}$$

When there is no relationship between the predictors and the predicted variable, F-statistic is expected to be close to 1. If $\mathbb{H}_1$ is true we expect the F-statistic to be greater than 1.

1.4 Summary of output from linear regression in R

1. Residuals - This provides a summary of the residuals from the linear regression model. To access these for a model, use `model$residuals`.

2. Coefficients - This provides estimates of coefficients, standard error of coefficients, t-value and p-value (P$>|t|$). To access these use `model$coefficients` or `coefficients`(*your_model*). You can access the standard error by `coefficients(summary(model))[,``Std. Error"]`.

3. Residual standard error - This provides the average amount the response will deviate from the true regression line. It provides a measure of the lack of the fit of a linear model to the data.

4. Multiple R-squared, Adjusted R-squared - R-squared is a measure between 0 and 1 to indicate the amount of variability explained by regression while adjusted R-squared accounts for number of predictors.

5. F-statistic and p-value - Test to see if at least one of the predictors is nonzero.

1.5 Additional results

1.5.1 The equality $\texttt{SST} = \texttt{SSE} + \texttt{SSR}$.

When we minimize $Q(\boldsymbol{\beta})$, taking derivative with respect to β_0 we have

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}) = 0.$$

Since the optimal solution $\hat{\boldsymbol{\beta}}$ satisfies the above, we have

$$0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}) = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i. \tag{1}$$

Similarly taking derivative with respect to $\beta_j, j = 1, \ldots, p$ and noticing that the optimal solution $\hat{\boldsymbol{\beta}}$ satisfy the normal equations, we have for $j = 1, \ldots, p$,

$$0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})x_{ij} = \sum_{i=1}^{n}(y_i - \hat{y}_i)x_{ij} = \sum_{i=1}^{n} e_i x_{ij}. \tag{2}$$

Now observe that

$$\begin{aligned}
\texttt{SST} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \texttt{SSE} + \texttt{SSR} + 2\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) \\
&= \texttt{SSE} + \texttt{SSR} - 2\bar{y}\sum_{i=1}^{n} e_i + 2\sum_{i=1}^{n} e_i(\hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}) \\
&= \texttt{SSE} + \texttt{SSR} - 0 + 2\hat{\beta}_0\sum_{i=1}^{n} e_i + 2\hat{\beta}_1\sum_{i=1}^{n} e_i x_{i1} + \ldots + +2\hat{\beta}_p\sum_{i=1}^{n} e_i x_{ip} \\
&= \texttt{SSE} + \texttt{SSR}.
\end{aligned}$$

The penultimate equality uses (1) and the final equality uses both (1) and (2).

1.5.2 Single variable regression: R^2 and empirical correlation

Suppose we have one independent variable and our data set is $(x_i, y_i), i = 1, \ldots, n$. Then the empirical correlation between the two variables is given by

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Now our model is $y = \beta_0 + \beta_1 x + \epsilon$ and setting up the quadratic sum of squares

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2,$$

by taking derivatives and using (1) and (2) we have

$$\begin{aligned} 0 &= \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}, \qquad \text{or,} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \text{and,} \\ 0 &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \end{aligned}$$

Therefore

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Now we know that

$$\begin{aligned} R^2 = \frac{\texttt{SSR}}{\texttt{SST}} &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \hat{\beta_1}^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= r(\boldsymbol{x}, \boldsymbol{y})^2. \end{aligned}$$

Since R^2 measures the strength of the linear model and correlation measures the strength of a linear relationship, this is consistent