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.

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.

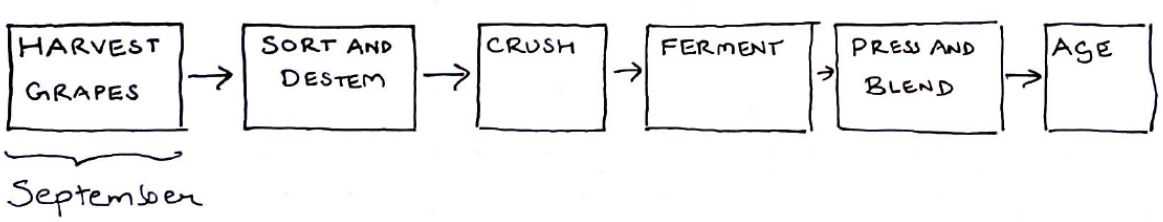


Figure 0.1: Schematic of wine making process

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. 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 more unpleasant 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 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 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 make use of data from the website www.liquidasset.com with the following variables:

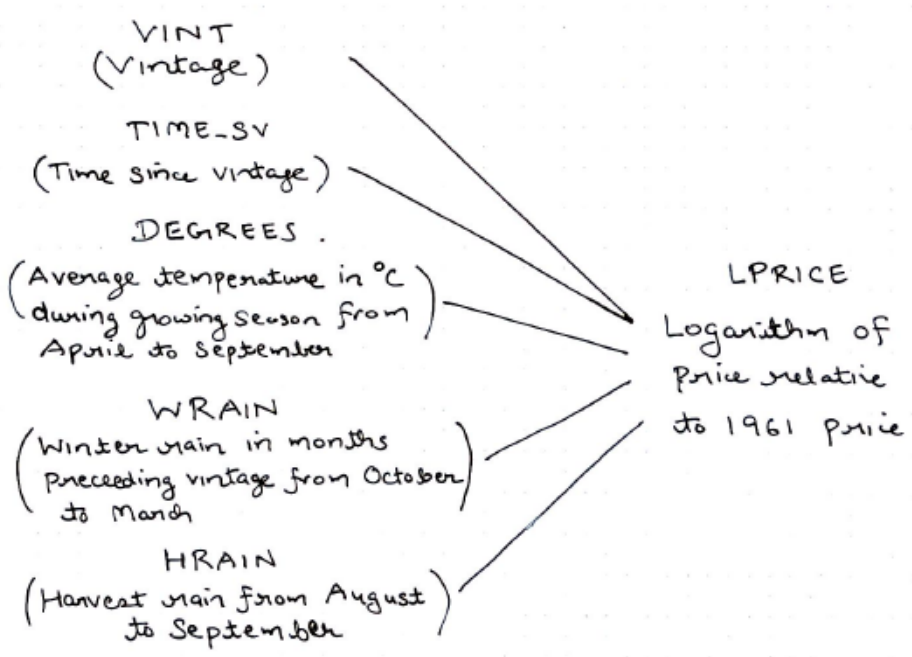


Figure 0.2: Data from liquidasset.com

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

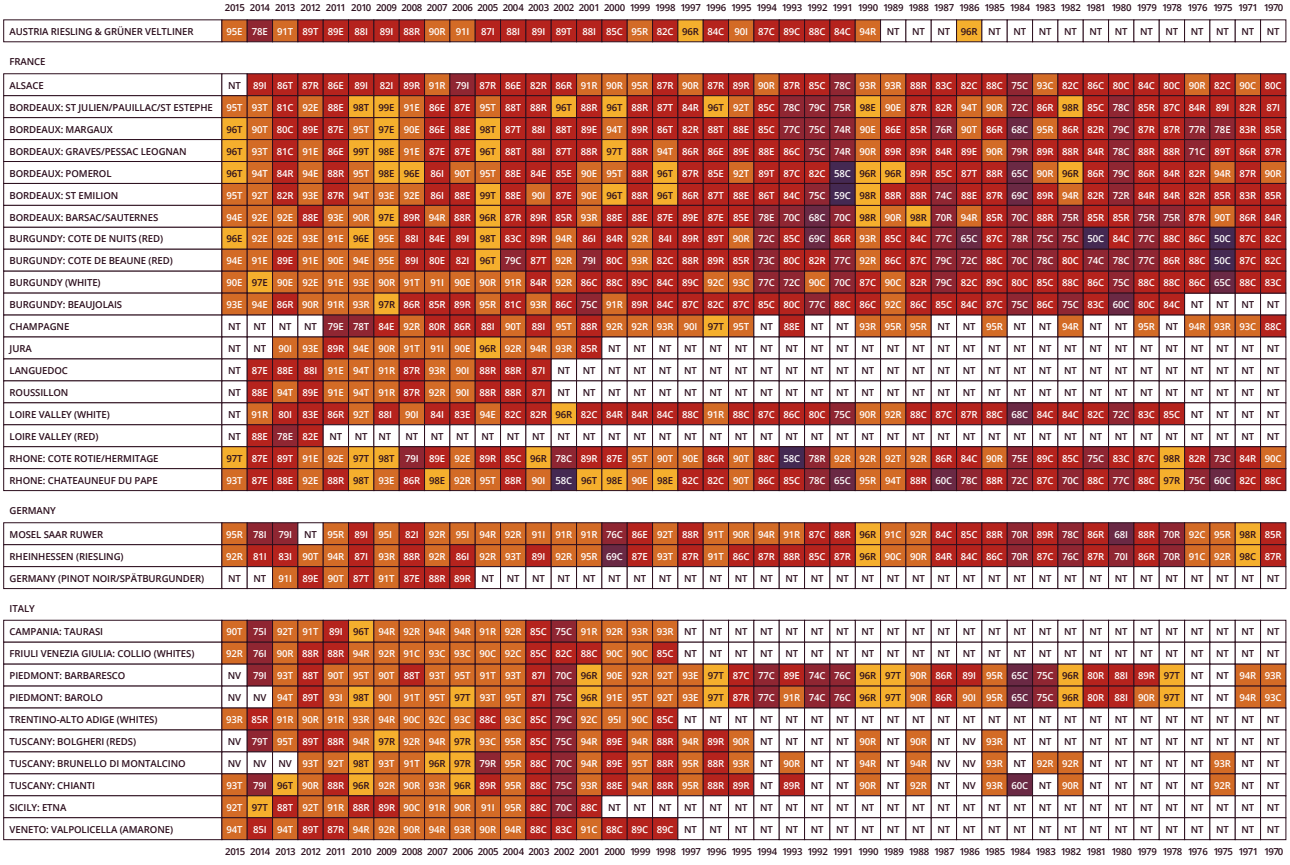
Summary

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.



	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	1985	1984	1983	1982	1981	1980	1979	1978	1976	1975	1971	1970
AUSTRIA RIESLING & GRÜNER VELTLINER	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	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
FRANCE																																											
ALSACE	NT	89I	86T	87R	86E	89I	82I	89R	91R	79I	87R	86E	82R	86R	91R	90R	95R	87R	90R	87R	89R	90R	87R	85C	78C	93R	93R	88R	83C	82C	88C	75C	93C	82C	86C	80C	84C	80C	90R	82C	90C	80C
BORDEAUX: ST JULIEN/PAUILLAC/ST ESTEPHE	95T	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	90R	72C	86R	98R	85C	78C	85R	87C	84R	89I	82R	87I
BORDEAUX: MARGAUX	96T	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	86R	68C	95R	86R	82R	79C	87R	87R	77R	78E	83R	85R
BORDEAUX: GRAVES/PESSAC LEOGNAN	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	90R	79R	89R	88R	84R	78C	88R	88R	71C	89T	86R	87R
BORDEAUX: POMEROL	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	88R	65C	90R	96R	86R	79C	86R	84R	82R	94R	87R	90R
BORDEAUX: ST EMILION	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	87R	69C	89R	94R	82R	72R	84R	84R	82R	85R	83R	85R
BORDEAUX: BARSAC/SAUTERNES	94E	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	85R	70C	88R	75R	85R	85R	75R	75R	87R	90T	86R	84R
BURGUNDY: COTE DE NUITS (RED)	96E	92E	92E	93E	91E	96E	95E	88I	84E	89I	98T	83C	89R	94R	86I	84R	92R	84I	89R	89T	90R	72C	85C	69C	86R	93R	85C	84C	77C	65C	87C	78R	75C	75C	50C	84C	77C	88C	86C	50C	87C	82C
BURGUNDY: COTE DE BEAUNE (RED)	94E	91E	89E	91E	90E	94E	95E	89I	80E	82I	96T	79C	87T	92R	79I	80C	93R	82C	88R	89R	85R	73C	80C	82R	77C	92R	86C	87C	79C	72C	88C	70C	78C	80C	74C	78C	77C	86R	88C	50C	87C	82C
BURGUNDY (WHITE)	90E	97E	90E	92E	91E	93E	90R	91T	91I	90E	90R	91R	84R	92R	86C	88C	89C	84C	89C	92C	93C	77C	72C	90C	70C	87C	90C	82R	79C	82C	89C	80C	85C	88C	86C	75C	88C	88C	86C	65C	88C	83C
BURGUNDY: BEAUJOLAIS	93E	94E	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	87C	75C	86C	75C	83C	60C	80C	84C	NT	NT	NT	NT
CHAMPAGNE	NT	NT	NT	NT	79E	78T	84E	92R	80R	86R	88I	90T	88I	95T	88R	92R	92R	93R	90I	97T	95T	NT	88E	NT	NT	93R	95R	95R	NT	NT	95R	NT	NT	94R	NT	NT	95R	NT	94R	93R	93C	88C
JURA	NT	NT	90I	93E	89R	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	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
LANGUEDOC	NT	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	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
ROUSSILLON	NT	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	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
LOIRE VALLEY (WHITE)	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	88C	68C	84C	84C	82C	72C	83C	85C	NT	NT	NT	NT
LOIRE VALLEY (RED)	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	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
RHONE: COTE ROTIE/HERMITAGE	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	90R	75E	89C	85C	75C	83C	87C	98R	82R	73C	84R	90C
RHONE: CHATEAUNEUF DU PAPE	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	88R	72C	87C	70C	88C	77C	88C	97R	75C	60C	82C	88C
GERMANY																																										
MOSEL SAAR RUWER	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	88R	70R	89R	78C	86R	68I	88R	70R	92C	95R	98R	85R
RHEINHESSEN (RIESLING)	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	86C	70R	87C	76C	87R	70I	86R	70R	91C	92R	98C	87R
GERMANY (PINOT NOIR/SPÄTBURGUNDER)	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	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
ITALY																																										
CAMPANIA: TAURASI	90T	75I	92T	91T	89I	96T	94R	92R	94R	94R	91R	92R	85C	75C	91R	92R	93R	93R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
FRIULI VENEZIA GIULIA: COLLIO (WHITES)	92R	76I	90R	88R	88R	94R	92R	91C	93C	93C	90C	92C	85C	82C	88C	90C	90C	85C	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
PIEDMONT: BARBARESCO	NV	79I	93T	88T	90T	95T	90T	88T	93T	95T	91T	93T	87I	70C	96R	90E	92R	92T	93E	97T	87C	77C	89E	74C	76C	96R	97T	90R	86R	89I	95R	65C	75C	96R	80R	88I	89R	97T	NT	NT	94R	93R
PIEDMONT: BAROLO	NV	NV	94T	89T	93I	98T	90I	91T	95T	97T	93T	95T	87I	75C	96R	91E	95T	92T	93E	97T	87R	77C	91R	74C	76C	96R	97T	90R	86R	90I	95R	65C	75C	96R	80R	88I	90R	97T	NT	NT	94R	93C
TRENTINO-ALTO ADIGE (WHITES)	93R	85R	91R	90R	91R	93R	94R	90C	92C	93C	88C	93C	85C	79C	92C	95I	90C	85C	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
TUSCANY: BOLGHERI (REDS)	NV	79T	95T	89T	88R	94R	97R	92R	94R	97R	93C	95R	85C	75C	94R	89E	94R	88R	94R	89R	90R	NT	NT	NT	NT	90R	NT	90R	NT	NV	93R	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
TUSCANY: BRUNELLO DI MONTALCINO	NV	NV	NV	93T	92T	98T	93T	91T	96R	97R	79R	95R	88C	70C	94R	89E	95T	88R	95R	88R	93R	NT	90R	NT	NT	94R	NT	94R	NV	NV	93R	NT	92R	92R	NT	NT	NT	NT	NT	93R	NT	NT
TUSCANY: CHIANTI	93T	79I	96T	90R	88R	96R	92R	90R	93R	96R	89R	95R	88C	75C	93R	88E	94R	88R	95R	88R	89R	NT	89R	NT	NT	90R	NT	92R	NT	NV	93R	60C	NT	90R	NT	NT	NT	NT	NT	92R	NT	NT
SICILY: ETNA	92T	97T	88T	92T	91R	88R	89R	90C	91R	90R	91I	95R	88C	70C	88C	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	NT
VENETO: VALPOLICELLA (AMARONE)	94T	85I	94T	89T	87R	94R	92R	90R	94R	93R	90R	94R	88C	83C	91C	88C	89C	89C	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
	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	1985	1984	1983	1982	1981	1980	1979	1978	1976	1975	1971	1970

Figure 0.3: Vintage chart from Robert Parker's website

Linear Regression

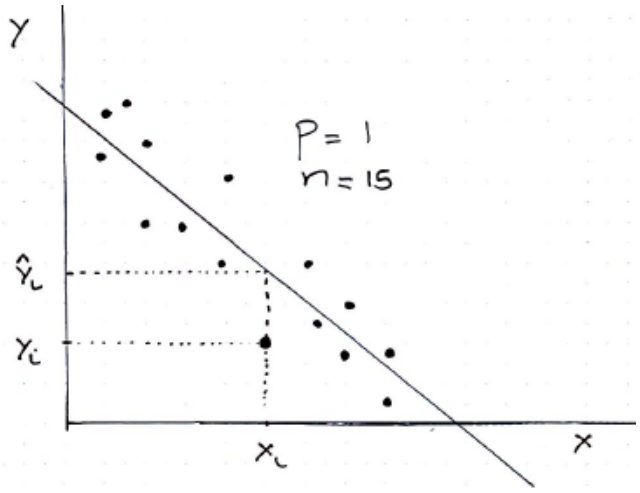


Figure 0.4: Linear model

Problem setup:

1. n = Number of observations
2. p = Number of predictor variables (excluding the constant 1)
3. y = Dependent variable in $\Re$ (outcome)
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 $\{y_i, x_{i1}, \ldots, x_{ip}\}$ 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:

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

We can rewrite the problem as:

$$\min_{\beta} (y - X\beta)'(y - X\beta).$$

Note that this is a convex quadratic minimization problem. The optimal solution is given by:

$$\hat{\beta} = (X'X)^{-1}X'y$$

where the fitted values are $\hat{y} = X\hat{\beta}$.

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. Assume that the random observations y_i are independent of each other, have a constant variance denoted by σ^2 and X is fixed. We obtain:

$$\text{Variance}(\beta) = \text{Variance}((X'X)^{-1}X'y) = (X'X)^{-1}\sigma^2.$$

Since the true variance σ^2 is unknown, we estimate it using the sample variance as follows:

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

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

Under the null hypothesis H_0: $\beta_i = 0$, the t-statistic is given as:

$$\text{t-statistic} = \frac{\hat{\beta}_i}{\text{Standard error}(\hat{\beta}_i)}.$$

If the absolute value of the t-statistic is high, the null hypothesis will be rejected in favor of H_1: $\beta_i \neq 0$. This indicates statistically that x_i 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.

Quality of fit

1. Let $\bar{y} = \sum_{i=1}^{n} y_i/n$. Define:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \text{ (Sum of squared errors)}$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \text{ (Sum of squares due to regression)}$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ (Total sum of squares)}$$

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

$$SST = SSE + SSR.$$

The residual standard error is defined as $\sqrt{SSE/(n-p-1)}$ 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:

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

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

(a) Regression fit is a horizontal line implies $R^2 = 0$ (the predictor variables have no explanatory power).

(b) Regression fits perfectly all points on a straight line 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 hypothesis. Let:

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

$$H_1 : \text{ At least one of the } \beta_j \text{ is nonzero}$$

The F-statistic is defined as:

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

When there is no relationship between the predictors and the predicted variable, F-statistic is expected to be close to 1. If H_1 is true we expect the F-statistic to be greater than 1.

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(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.