Linear Regression

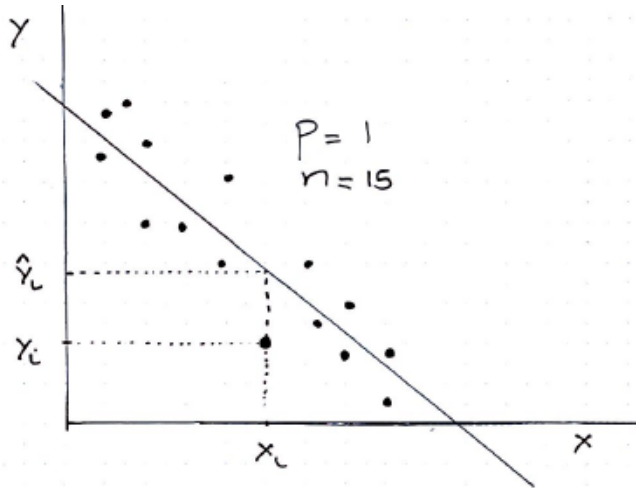

Figure 0.4: Linear model

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.