Multiple Linear Regression

* Simple linear regression tries to fit a simple line between two variables Y and X.

* If X is linearly related to Y this explains some of the variability in Y.

* In most cases, there is still a lot of variability about the line remaining.

* Some of this variability may be explained by including other covariates in the model.

Adding a Second Covariate

Example: University staff are interested in the relationship between how well students do in High School with their performance in first year of university. In one US study, 224 Computer Science students in a large were followed for their first 3 semesters of university. The data are called `CSData.txt`. Measured variables include the students' GPA after first year university, High school average grades (coded from 1–10) in math, science and English and the students' Math and Verbal SAT scores.

Adding a Second Covariate

Example continued:

An initial analysis looked at how well average high school math grade (coded from 1–10) predicted their GPA.

The fitted least squares line was

$$\text{GPA} = 0.908 + 0.208\text{HSM}$$

A test of whether $\beta_1 = 0$ gave a p-value of < 0.0001.

The R^2 value, however was only 0.1905.

Can we do better and account for more of the variability?

Adding a Second Covariate

Example continued:
Maybe the Math SAT scores are a better predictor.

The fitted least squares line for this is

$$\text{GPA} = 1.284 + 0.0023\text{SATM}$$

The test of $\beta_1 = 0$ has a p-value of 0.0001 and $R^2 = 0.0634$

Conclusion: Math SAT scores are a useful predictor of first year university GPA but not as good as high school math grade.

Adding a Second Covariate

Example continued:
Now suppose that we try to include both High School and SAT Math scores in the model.

The least squares estimated model is now

$$\text{GPA} = 0.666 + 0.193\text{HSM} + 0.0006\text{SATM}$$

The p-value for testing $\beta_1 = 0$ (coefficient of HSM) is < 0.0001 as before. That for testing $\beta_2 = 0$ (coefficient of SATM) is now 0.319

The R^2 for this model with 2 covariates is 0.1942, only slightly higher than it was for the simple regression with the covariate HSM alone.

What is going on?

The Two Covariate Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

* The interpretation of the parameters β_1 and β_2 is different than in the simple model.

* β_2 is now the effect of X_2 after adjusting for X_1.

* Similarly β_1 is the effect of X_1 after adjusting for the effect of X_2.

* If X_1 and X_2 are uncorrelated then the estimates will be the same as the estimates in the simple models but in general this is not true.

The Two Covariate Model

Consider the following sequence of models.

1. Fit a simple linear model between $y_1, \ldots, y_n$ and the covariate X_1 with observed values $x_{11}, \ldots, x_{n1}$. Find the residuals $e_{11}, \ldots, e_{n1}$.

2. Fit a regression with the variable X_2 as response (observed values $x_{12}, \ldots, x_{n2}$) and $x_{11}, \ldots, x_{n1}$ as covariate values and find the residuals $e_{12}, \ldots, e_{n2}$.

3. Fit a regression with $e_{11}, \ldots, e_{n1}$ as response values and $e_{12}, \ldots, e_{n2}$ as covariate values.

The Two Covariate Model

* The residuals from the first model can be considered the values of Y adjusted for the values of X_1.

* Similarly the residuals from the second model can be considered the values of the covariate X_2 adjusted for the values of X_1.

* Hence the third model regresses the adjusted response on the adjusted second covariate.

* The estimated slope in this third model will be exactly the estimated coefficient of X_2 in the two covariate model.

* The standard errors, test statistics and p-values will also be identical (but may be subject to rounding error).

Adding a Second Covariate

* Note that it is not always the case that the R^2 for the two covariate model is less than the sum of the R^2 for each of the individual simple models.

* In some cases, neither covariate alone is a good predictor of the response but the two together do a very good job of explaining the variability in the response.

* If X_1 and X_2 are highly negatively correlated to each other but Y is related to both in the same direction this can happen.

* In a 1987 paper David Hamilton (Dalhousie) showed this quite nicely.

Multiple Regression

* The assumed model is of the form

$$Y \;=\; \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

* The error terms ε are assumed to have mean 0 for every value of $\boldsymbol{x} = (x_1, \ldots, x_p)^t$.

* The variance of ε is assumed constant (equal to σ^2) for all covariate vectors $\boldsymbol{x}$.

* As before we wish to estimate the $p+2$ parameters and make inference about the coefficients in the model.

Estimating the Coefficients

* As in simple regression we will estimate the coefficients using least squares.

* That is we will find values of $\beta_0, \beta_1, \ldots, \beta_p$ which minimize

$$S(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

* On taking partial derivatives and setting them equal to 0 we get the $p + 1$ normal equations which must be solved simultaneously to get the estimators.

The Model in Matrix Notation

* Let $\boldsymbol{Y} = (y_1, \ldots, y_n)^t$ be the response vector.
* The error vector is similarly defined as $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^t$.
* Define the vector of coefficients as $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$.
* The Design Matrix $\boldsymbol{X}$ is the $n \times (p+1)$ matrix with all elements in the first column equal to 1 and column $r+1$ equal to $(x_{1r}, \ldots, x_{nr})^t$.
* Then the linear regression model can be written as

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

Estimation in Matrix Notation

* For given $\boldsymbol{\beta}$ the residuals can be written as

$$e_i \;=\; y_i - \boldsymbol{x}_i\boldsymbol{\beta}$$

where $\boldsymbol{x}_i$ is the i^{th} row of $\boldsymbol{X}$.

* Then we can write

$$S(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 \;=\; (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

* On taking derivatives and setting them equal to 0 we get the least squares estimates

$$\hat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y}$$

Estimation of σ^2

* The fitted values from the regression are $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

* The residuals are

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

* As in the simple model, we can find an unbiased estimator of σ^2 by looking at the sum of squared residuals

$$\text{SSE} = \boldsymbol{e}^t\boldsymbol{e} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

* The estimator is

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-(p+1)}.$$

Simple Regression in Matrix Notation

* Simple regression is a special case of multiple regression with $p = 1$ and can be formulated in the same matrix framework.

* $\boldsymbol{Y}$ is still the vector of observed response variables.

* $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

* The $\boldsymbol{X}$ matrix is

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

Simple Regression in Matrix Notation

* We can then see that

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum x_i^2 \end{pmatrix}$$

* The inverse of this matrix is

$$(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \frac{1}{nSxx}\begin{pmatrix} \sum x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix}$$

Simple Regression in Matrix Notation

* We also have

$$\boldsymbol{X}^t\boldsymbol{Y} \;=\; \begin{pmatrix} n\overline{y} \\ \sum x_i y_i \end{pmatrix}$$

* Hence we get

$$\widehat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \;=\; \frac{1}{nSxx}\begin{pmatrix} n\overline{y}\sum x_i^2 - n\overline{x}\sum x_i y_i \\ n\sum x_i y_i - n^2\overline{xy} \end{pmatrix}$$

* A little algebra shows that these estimates agree with those given for the simple regression model.

Simple Regression in Matrix Notation

* Note that we can write

$$(\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2 = \frac{1}{nSxx}\begin{pmatrix} \sum x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix}\sigma^2$$

$$= \begin{pmatrix} \left(\frac{1}{n}+\frac{\overline{x}^2}{Sxx}\right)\sigma^2 & -\frac{\overline{x}}{Sxx}\sigma^2 \\ -\frac{\overline{x}}{Sxx}\sigma^2 & \frac{\sigma^2}{Sxx} \end{pmatrix}$$

* The diagonal elements of this matrix give $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1)$ as stated in Theorem 2.

* The off-diagonal element gives $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ again as in Theorem 2.

Properties of the Estimators

Theorem 6

The estimators described on the previous slides satisfy

1. $\mathsf{E}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}.$

2. $\mathsf{Var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}.$

3. $\mathsf{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) = \sigma^2.$

Standard Errors and Sampling Distributions

* The estimated Variance-Covariance matrix for $\boldsymbol{\beta}$ is given by

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) \;=\; \hat{\sigma}^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} \;=\; \sigma^2\boldsymbol{C}.$$

* The standard errors can be found by looking at the diagonal elements of this matrix.

$$\text{se}(\hat{\beta}_j) \;=\; \hat{\sigma}\sqrt{c_{jj}}$$

* These standard errors are always output as part of the SAS output from a regression.

* The Sampling Distributions are

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \;\sim\; t_{n-(p+1)} \qquad j = 0, 1, \ldots, p$$

Individual Confidence Intervals

* Confidence intervals for individual parameters are given by

$$\widehat{\beta}_j \pm t_{(n-p-1;\alpha/2)}\mathrm{se}(\widehat{\beta}_j) \qquad j = 0, \ldots, p$$

* As with the estimates, these are to be interpreted as confidence intervals for the additional effect of X_j when all of the other covariates are included in the model.

Individual Hypothesis Tests

* Similarly we can test if $\beta_j = \beta_j^0$ using the test statistic

$$t_j = \frac{\beta_j - \beta_j^0}{\text{se}(\widehat{\beta}_j)}.$$

* The p-value for this test can be found by comparing the test statistic to the t_{n-p-1} distribution.

* It is important to remember that this is a test of the additional contribution of X_j when all of the other covariates are in the model.

Analysis of Variance Table

* As in the simple model we can write the total sum of squares as the error sum of squares plus a sum of squares related to the model.

* These are generally displayed in an ANOVA table exactly as in the simple case.

* The degrees of freedom for the model is now p whereas that for the error is $n - p - 1$.

* The mean squares are the sums of squares divided by the degrees of freedom.

The Coefficient of Determination

* As in the simple model we can define the coefficient of determination

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

* This is still a measure of the proportion of variability in Y which is accounted for by the fitted model.

* It can also be shown that R^2 is the square of the correlation coefficient between Y and the fitted values $\hat{Y}$.

* In multiple regression, a plot of Y against $\hat{Y}$ is also very useful since it incorporates all of the covariates at once.

Testing All Coefficients Equal to 0

* We will usually want to test if the fitted model is useful in predicting the response.

* The null hypothesis here is

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

* The alternative is that at least one of the $\beta_j \neq 0$.

Testing All Coefficients Equal to 0

* Under the null hypothesis it can be shown that

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F_{p,n-p-1}$$

* This value is given in the `F Value` column of the ANOVA table.

* The final column gives the p-value for the test.

* Note that this is a simultaneous test for all of the covariates. Rejecting H_0 does not imply that all of the covariates are important, just that some subset of them are important.

Testing a Subset of Coefficients Equal to 0

* In some cases we may want to test if some but not all of the coefficients are 0.

* We define the full model to be

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

* Suppose the null hypothesis we want to test is

$$H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$$

against the alternative that one of these $\beta_j \neq 0$.

* Then we can define the reduced model to be

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k + \varepsilon$$

Testing a Subset of Coefficients Equal to 0

* From the full model we can get the SSE and MSE for this model. Lets call them SSE_{full} and MSE_{full}.

* As always the degrees of freedom will be $n - p - 1$

* Similarly we can find SSE_{red} from the reduced model with $n - k - 1$ degrees of freedom.

* Then we can define a test statistic to be

$$F = \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(p-k)}{\text{MSE}_{\text{full}}}$$

which has an $F_{p-k,n-p-1}$ distribution if H_0 is true so we can find a p-value.

Other Tests

* Suppose that H_0 specifies some reduced model which can be thought of as a special case of the full model.

* For example we may have a model that specifies some subset of coefficients are all equal but not necessarily 0.

* Then we fit the full and reduced model to get the SSE from each and the df from each.

* The test statistic is then

$$F = \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{SSE}_{\text{full}}/df_{\text{full}}}$$

  which has an F distribution with $df_{\text{red}} - df_{\text{full}}$ and df_{full} degrees of freedom when H_0 is true.

Prediction

* As before we often wish to do inference for a new vector of covariate values $\boldsymbol{x}_0 = (1, x_{1,0}, x_{2,0}, \ldots, x_{p,0})^t$.

* The point estimator of $\mu_0 = \mathsf{E}[Y \mid \boldsymbol{x}_0]$ is

$$\hat{\mu}_0 = \boldsymbol{x}_0^t \hat{\boldsymbol{\beta}}$$

* The standard error of this estimator is

$$\mathsf{se}(\hat{\mu}_0) = \hat{\sigma}\sqrt{\boldsymbol{x}_0^t (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_0}$$

* A $100(1-\alpha)\%$ confidence interval is given by

$$\hat{\mu}_0 \pm t_{(n-p-1,\alpha/2)} \mathsf{se}(\hat{\mu}_0)$$

Prediction

* We may also wish to predict the value for an individual with covariate values $\boldsymbol{x}_0$.

* The point predictor is the same as the point estimator of μ_0

$$\widehat{y}_0 \;=\; \boldsymbol{x}_0^t \widehat{\boldsymbol{\beta}}$$

* In the standard error, however, we need to account for the extra error term so we get the standard error of prediction

$$\mathsf{se}(\widehat{y}_0) \;=\; \widehat{\sigma}\sqrt{1 + \boldsymbol{x}_0^t (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_0}$$

* A $100(1-\alpha)\%$ prediction interval is given by

$$\widehat{y}_0 \;\pm\; t_{(n-p-1,\alpha/2)} \mathsf{se}(\widehat{y}_0)$$