Multiple Linear Regression

So far, we have seen the concept of simple linear regression where a single predictor variable X was used to model the response variable Y. In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables.

Examples:

- The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage of the lot and a number of other factors.
- The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.
-

Note: We will reserve the term MULTIPLE REGRESSION for models with two or more predictors and one response. There are also regression models with two or more response variables. These models are usually called MULTIVARIATE REGRESSION MODELS.

In this chapter, we will introduce a new (linear algebra based) method for computing the parameter estimates of multiple regression models. This more compact method is convenient for models for which the number of unknown parameters is large.

Example: A multiple linear regression model with k predictor variables $X_1, X_2, ..., X_k$ and a response Y, can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon.$$

As before, the ϵ are the residual terms of the model and the distribution assumption we place on the residuals will allow us later to do inference on the remaining model parameters. Interpret the meaning of the REGRESSION COEFFICIENTS $\beta_0, \beta_1, \beta_2, ..., \beta_k$ in this model.

More complex models may include higher powers of one or more predictor variables, e.g.,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \qquad (1)$$

or interaction effects of two or more variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \qquad (2)$$

Note: Models of this type can be called LINEAR REGRESSION MODELS as they can be written as linear combinations of the β-parameters in the model. The x-terms are the weights and it does not matter, that they may be non-linear in x. Confusingly, models of type (1) are also sometimes called NON-LINEAR REGRESSION MODELS or POLYNOMIAL REGRESSION MODELS, as the regression curve is not a line. Models of type (2) are usually called linear models with interaction terms.

It helps to develop a little geometric intuition when working with regression models. Models with two predictor variables (say x_1 and x_2) and a response variable y can be understood as a two-dimensional surface in space. The shape of this surface depends on the structure of the model. The observations are points in space and the surface is "fitted" to best approximate the observations.

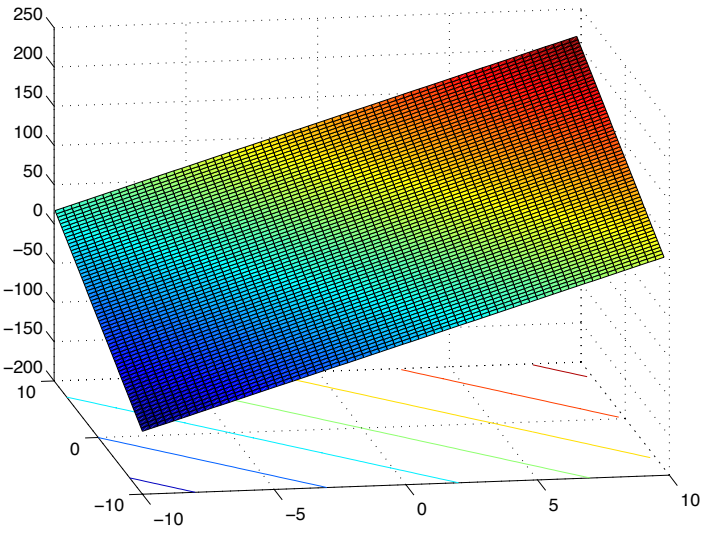
Example: The simplest multiple regression model for two predictor variables is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

The surface that corresponds to the model

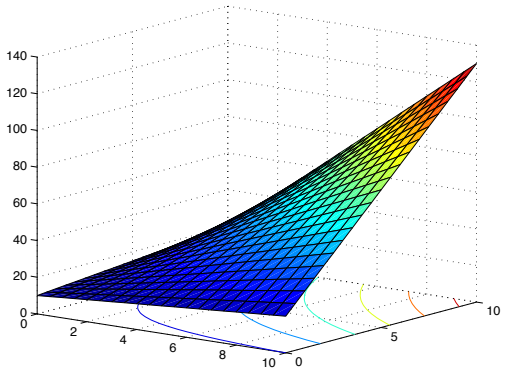
$$y = 50 + 10x_1 + 7x_2$$

looks like this. It is a plane in $\mathbb{R}^3$ with different slopes in x_1 and x_2 direction.



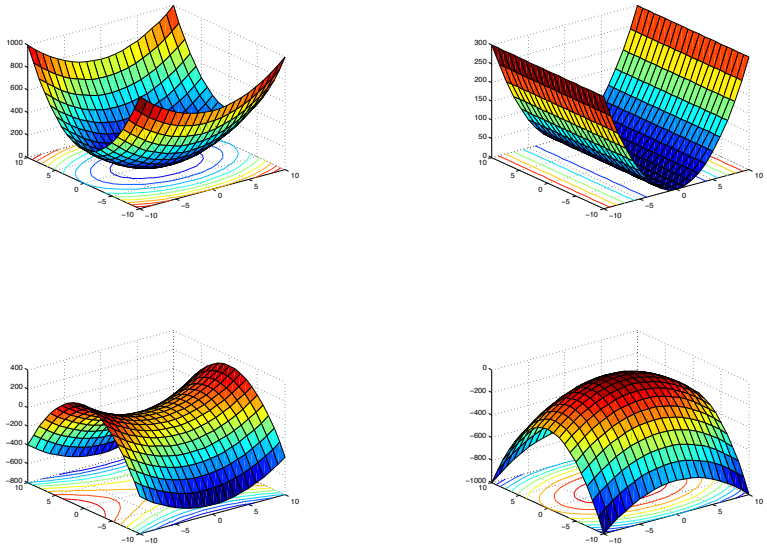
Example: For a simple linear model with two predictor variables and an interaction term, the surface is no longer flat but curved.

$$y = 10 + x_1 + x_2 + x_1 x_2$$



Example: Polynomial regression models with two predictor variables and interaction terms are quadratic forms. Their surfaces can have many different shapes depending on the values of the model parameters with the contour lines being either parallel lines, parabolas or ellipses.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$



Estimation of the Model Parameters

While it is possible to estimate the parameters of more complex linear models with methods similar to those we have seen in chapter 2, the computations become very complicated very quickly. Thus, we will employ linear algebra methods to make the computations more efficient.

The setup: Consider a multiple linear regression model with k independent predictor variables $x_1, \ldots, x_k$ and one response variable y.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

Suppose, we have n observations on the $k+1$ variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

n should be bigger than k. Why?

You can think of the observations as points in $(k+1)$-dimensional space if you like. Our goal in least-squares regression is to fit a hyper-plane into $(k+1)$-dimensional space that minimizes the sum of squared residuals.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

As before, we could take derivatives with respect to the model parameters $\beta_0, \ldots, \beta_k$, set them equal to zero and derive the LEAST-SQUARES NORMAL EQUATIONS that our parameter estimates $\hat{\beta}_0, \ldots, \hat{\beta}_k$ would have to fulfill.

$$\begin{array}{ccccccc}
n\hat{\beta}_0 & +\hat{\beta}_1 \sum_{i=1}^{n} x_{i1} & +\hat{\beta}_2 \sum_{i=1}^{n} x_{i2} & +\cdots & +\hat{\beta}_k \sum_{i=1}^{n} x_{ik} & = & \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} & +\hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 & +\hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} & +\cdots & +\hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} & = & \sum_{i=1}^{n} x_{i1}y_i \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} & +\hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} & +\hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} & +\cdots & +\hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 & = & \sum_{i=1}^{n} x_{ik}y_i
\end{array}$$

These equations are much more conveniently formulated with the help of vectors and matrices.

Note: Bold-faced lower case letters will now denote vectors and bold-faced upper case letters will denote matrices. Greek letters cannot be bold-faced in Latex. Whether a Greek letter denotes a random variable or a vector of random variables should be clear from the context, hopefully.

Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

With this compact notation, the linear regression model can be written in the form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

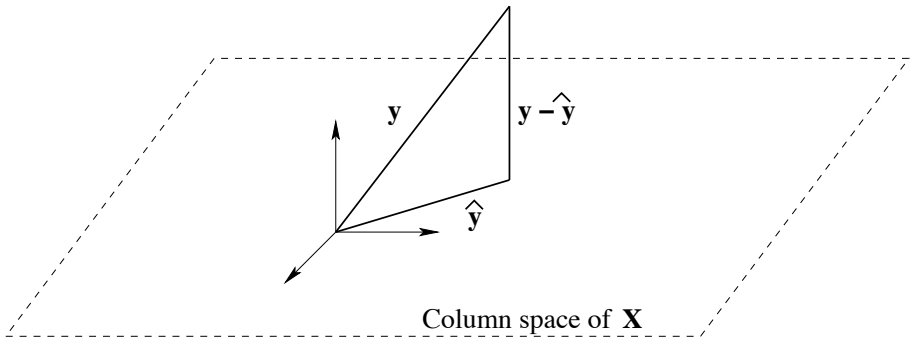
In linear algebra terms, the least-squares parameter estimates β are the vectors that minimize

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Any expression of the form $\mathbf{X}\beta$ is an element of a (at most) $(k+1)$-dimensional hyperspace in $\mathbb{R}^n$ spanned by the $(k+1)$ columns of X. Imagine the columns of $\mathbf{X}$ to be fixed, they are the data for a specific problem, and imagine β to be variable. We want to find the "best" β in the sense that the sum of squared residuals is minimized. The smallest that the sum of squares could be is zero. If all ϵ_i were zero, then

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

Here $\hat{\mathbf{y}}$ is the projection of the n-dimensional data vector $\mathbf{y}$ onto the hyperplane spanned by $\mathbf{X}$.



The $\hat{\mathbf{y}}$ are the predicted values in our regression model that all lie on the regression hyper-plane. Suppose further that $\hat{\beta}$ satisfies the equation above. Then the residuals $\mathbf{y} - \hat{\mathbf{y}}$ are orthogonal to the columns of $\mathbf{X}$ (by the Orthogonal Decomposition Theorem) and thus

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$$
$$\Leftrightarrow \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta} = 0$$
$$\Leftrightarrow \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

These vector normal equations are the same normal equations that one could obtain from taking derivatives. To solve the normal equations (i.e., to find the parameter estimates $\hat{\beta}$), multiply both sides with the inverse of $\mathbf{X}'\mathbf{X}$. Thus, the least-squares estimator of β is (in vector form)

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

This of course works only if the inverse exists. If the inverse does not exist, the normal equations can still be solved, but the solution may not be unique. The inverse of $\mathbf{X}'\mathbf{X}$ exists, if the columns of $\mathbf{X}$ are linearly independent. That means that no column can be written as a linear combination of the other columns.

The vector of fitted values $\hat{\mathbf{y}}$ in a linear regression model can be expressed as

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

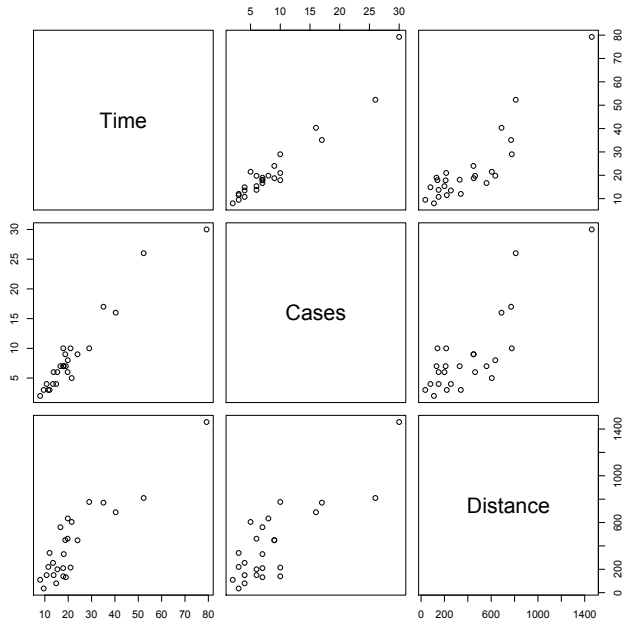
The $n \times n$ matrix $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is often called the HAT-MATRIX. It maps the vector of observed values $\mathbf{y}$ onto the vector of fitted values $\hat{\mathbf{y}}$ that lie on the regression hyper-plane. The regression residuals can be written in different ways as

$$\epsilon = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Example: The Delivery Times Data

A soft drink bottler is analyzing the vending machine serving routes in his distribution system. He is interested in predicting the time required by the distribution driver to service the vending machines in an outlet. This service activity includes stocking the machines with new beverage products and performing minor maintenance or housekeeping. It has been suggested that the two most important variables influencing delivery time (y in min) are the number of cases of product stocked (x_1) and the distance walked by the driver (x_2 in feet). 25 observations on delivery times, cases stocked and walking times have been recorded and are available in the file "DeliveryTimes.txt".

Before you begin fitting a model, it makes sense to check that there is indeed a (somewhat) linear relationship between the predictor variables and the response. The easiest way to do this is with the `plot()` command in R. If your object is a data file where each column corresponds to a variable (predictor or response), you will automatically obtain a matrix of scatterplots.



Look at the panels that describe the relationship between the response (here time) and the predictors. Make sure that the pattern is somewhat linear (look for obvious curves in which case the simple linear model without powers or interaction terms would not be a good fit).

Caution: Do not rely too much on a panel of scatterplots to judge how well a multiple linear regression really works. It can be very hard to see. A perfectly fitting model can look like a random "confetti" plot if the predictor variables are themselves correlated.

If a regression model has only two predictor variables, it is also possible to create a three-dimensional plot of the observations.

Computing the parameter estimates of this linear regression model "by-hand" in R, means to formulate the $\mathbf{X}$ matrix and the $\mathbf{y}$-vector and to use the equations derived on the previous pages to compute $\hat{\beta}$.

```
> X <- as.matrix(cbind(1,delivery$Cases, delivery$Distance))
> head(X)
     [,1] [,2] [,3]
[1,]    1    7  560
[2,]    1    3  220
[3,]    1    3  340
[4,]    1    4   80
[5,]    1    6  150
[6,]    1    7  330
> y <- as.matrix(delivery$Time)
> head(y)
      [,1]
[1,] 16.68
[2,] 11.50
[3,] 12.03
[4,] 14.88
[5,] 13.75
[6,] 18.11
> beta_hat <- solve(t(X)%*%X)%*%t(X)%*%y
> beta_hat
             [,1]
[1,] 2.34123115
[2,] 1.61590721
[3,] 0.01438483
```

Thus,

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2.34123115 \\ 1.61590721 \\ 0.01438483 \end{bmatrix}$$

With this, the estimated multiple regression equation becomes:

$$\hat{y} = 2.341 + 1.616x_1 + 0.0144x_2$$

where y is the delivery time, x_1 is the number of cases and x_2 is the distance walked by the driver. We can get more details about the fitted regression model, such as the estimated residual variance and hypothesis tests for both slopes.

```
> fit <- lm(Time ~ Cases + Distance , data = delivery)
> summary(fit)

Call:
lm(formula = Time ~ Cases + Distance, data = delivery)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7880 -0.6629  0.4364  1.1566  7.4197

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
Cases       1.615907   0.170735   9.464 3.25e-09 ***
Distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared: 0.9596, Adjusted R-squared: 0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16

> anova(fit)
Analysis of Variance Table

Response: Time
          Df Sum Sq Mean Sq F value    Pr(>F)
Cases      1 5382.4  5382.4 506.619 < 2.2e-16 ***
Distance   1  168.4   168.4  15.851 0.0006312 ***
Residuals 22  233.7    10.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example: Read off the estimated residual variance from the output shown above.

Properties of the Least Squares Estimators

Example: The least squares estimate vector $\hat{\beta}$ in the multiple linear regression model is unbiased.

Example: Find the covariance matrix of the least squares estimate vector $\hat{\beta}$.

The estimate of the residual variance can still be found via the residual sum of squares SS_{Res} which has the same definition as in the simple linear regression case.

$$SS_{Res} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \epsilon'\epsilon$$

It can also be expressed in vector form:

$SS_{Res} =$

If the multiple regression model contains k predictors, then the degree of freedom of the residual sum of squares is $n-k$ (we lose one degree of freedom for the estimation of each slope and the intercept). Thus

$$MS_{Res} = \frac{SS_{Res}}{n-k-1} = \hat{\sigma}^2$$

The residual variance is model dependent. Its estimate changes if additional predictor variables are included in the model or if predictors are removed. It is hard to say which one the "correct" residual variance is. We will learn later how to compare different models with each other. In general, a smaller residual variance is preferred in a model.

Maximum Likelihood Estimation

As in the simple linear regression model, the maximum likelihood parameter estimates are identical to the least squares parameter estimates in the multiple regression model.

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where the ϵ are assumed to be iid $\mathrm{N}(0, \sigma^2)$. Or short, $\epsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The likelihood function can be written in vector form. Maximizing the likelihood function leads to the ML parameter estimates $\tilde{\beta}$ and $\tilde{\sigma}^2$.

$L(\epsilon, \beta, \sigma^2) =$

Thus,

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \qquad \tilde{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{n}$$

Hypothesis Testing for Multiple Regression

After fitting a multiple linear regression model and computing the parameter estimates, we have to make some decisions about the model:

- Is the model a good fit for the data?
- Do we really need all the predictor variables in the model? (Generally, a model with fewer predictors and about the same "explanatory power" is better).

There are several hypothesis tests that we can utilize to answer these questions. Their results are usually reported in the coefficients and ANOVA tables that are produced as routine output in multiple regression analysis. But the tests can also be conducted "by-hand", if necessary.

TESTING FOR SIGNIFICANCE OF REGRESSION: This very pessimistic test asks whether any of the k predictor variables in the model have any relationship with the response.

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{vs.} \quad H_a : \beta_j \neq 0 \text{ for at least one } j$$

The test statistic function for this test is based on the sums of squares that we have previously defined for the simple linear regression model (the definitions are still the same).

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$$

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

The test statistic function then becomes

$$F = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}} \sim F_{k,n-k-1}$$

If the value of this test statistic is large, then the regression "works well" and at least one predictor in the model is relevant for the response. The F-test statistic and p-value are reported in the regression ANOVA table (columns `F value` and `Pr(>F)`).

Example: Read off and interpret the result of the F-test for significance of regression in the Delivery Time Example.

ASSESSING MODEL ADEQUACY: There are several ways in which to judge how well a specific model fits. We have already seen that in general, a smaller residual variance is desirable. Other quantities that describe the "goodness of fit" of the model are R^2 and adjusted R^2. Recall, that in the simple linear regression model, R^2 was simply the square of the correlation coefficient between the predictor and the response. This is no longer true in the multiple regression model. But there is another interpretation for R^2. In general, R^2 is the proportion of variation in the response that is explained through the regression on all the predictors in the model.

Including more predictors in a multiple regression model will always bring up the value of R^2. But using more predictors is not necessarily better. To weigh the proportion of variation explained with the number of predictors, we can use the ADJUSTED R^2.

$$R^2_{\text{Adj}} = 1 - \frac{SS_R/(n-k-1)}{SS_T/(n-1)}$$

Here, k is the number of predictors in the current model and $SS_R/(n-k)$ is actually the estimated residual variance of the model with k predictors. The adjusted R^2 does not automatically increase when more predictors are added to the model and it can be used as one tool in the arsenal of finding the "best" model for a given data set. Higher adjusted R^2 indicates a better fitting model.

Example: For the Delivery Time data, find R^2 and the adjusted R^2 for the model with both predictor variables in the R-output.

TESTING INDIVIDUAL REGRESSION COEFFICIENTS: As in the simple linear regression model, we can formulate individual hypothesis tests for each slope (or even the intercept) in the model. For instance

$$H_0 : \beta_j = 0, \quad \text{vs.} \quad H_A : \beta_j \neq 0$$

tests whether the slope associated with the j^{th} predictor is significantly different from zero. The test statistics for this test is

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t(df = n-k-1)$$

Here, $se(\hat{\beta}_j)$ is the square root of the j^{th} diagonal entry of the covariance matrix $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ of the estimated parameter vector $\hat{\beta}$. This test is a MARGINAL TEST. That means that the test statistic (and thus the p-value of the test) depends not just on the j^{th} predictor but also on all other predictors that are included in the model at the same time. Thus, if any predictor is added or removed from a regression model, hypothesis tests for individual slopes need to be repeated. If this test's null hypothesis is rejected, we can conclude that the j^{th} predictor has a significant influence on the response, given the other regressors in the model at the same time.

Example: Read off test statistic values and p-values for the two regressors CASES and DISTANCE in the Delivery Time data example. Formulate conclusions for both predictors.

Note: As we've seen before, every two-sided hypothesis test for a regression slope can also be reformulated as a confidence interval for the same slope. The 95% confidence intervals for the slopes can also be computed by R (command `confint()`.

TEST FOR A GROUP OF PREDICTORS: Consider the multiple linear regression model with k predictors. Suppose that we can partition the predictors into two groups $(x_1, \ldots, x_{k-p})$ and $(x_{k-p+1}, \ldots, x_k)$. We want to simultaneously test, whether the latter group of p predictors can be removed from the model. Suppose we partition the vector of regression slopes accordingly into two parts

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

where β_1 contains the intercept and the slopes for the first $k-p$ predictors and β_2 contains the remaining p slopes. We want to test

$$H_0 : \beta_2 = \mathbf{0} \quad \text{vs.} \quad H_A : \beta_2 \neq \mathbf{0}$$

We will compare two alternative regression models to each other:

(Full Model)	$\mathbf{y} = \mathbf{X}\beta + \epsilon$	with $SS_R(\beta) = \hat{\beta}\mathbf{X}'\mathbf{y}$ (k degrees of freedom).
(Reduced Model)	$\mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon$	with $SS_R(\beta_1) = \hat{\beta}_1\mathbf{X}_1'\mathbf{y}$ ($k-p$ degrees of freedom)

With this notation, the regression sum of squares that describes the contribution of the slopes in β_2 given that β_1 is already in the model becomes

$$SS_R(\beta_2|\beta_1) = SS_R(\beta_1, \beta_2) - SS_R(\beta_1)$$

The test statistic that tests the hypotheses described above is

$$F = \frac{SS_R(\beta_2|\beta_1)/p}{MS_{Res}} \sim F_{p,n-k-1}$$

Caution: Under certain circumstances (when there is multicollinearity in the data), the power of this test is very low. That means that even if the predictors in β_2 may be important, this test may fail to reject the null hypothesis and consequently exclude these predictors.

Note: Tests like the above play an important role in MODEL BUILDING. Model building is the task of selecting a subset of relevant predictors from a larger set of available predictors to build a good regression model. This kind of test is well suited for this task, because it tests whether additional predictors contribute significantly to the quality of the model, given the predictors that are already included.

Example: For the Delivery Time data, test whether there is a significant contribution from including the `Distance` variable, if the `Cases` variable is already included in the model.

General Linear Hypotheses

So far, we have discussed how to test hypotheses that state that single slopes or sets of slopes are all equal to zero. There are more general ways in which statements about the slopes in a multiple regression model could be phrased. For instance, consider the null hypothesis $H_0 : \mathbf{T}\beta = \mathbf{0}$, where $\mathbf{T}$ is some $r \times k$ matrix of constants. Effectively, this defines r linear equations in the k slope (and intercept) parameters. Assume that the equations are independent (not redundant).

The FULL MODEL for this problem is $\mathbf{y} = \mathbf{X}\beta + \epsilon$ with $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ with residual sum of squares

$$SS_{Res}(FM) = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} \quad (n - k - 1 \text{ degrees of freedom})$$

To obtain the reduced model (the model under the null hypothesis), the r independent equations in $\mathbf{T}\beta = \mathbf{0}$ are used to solve for r of the regression coefficients in terms of the remaining $k-r$ coefficients. This leads to the reduced model $\mathbf{y} = \mathbf{Z}\gamma + \epsilon$ in which the estimate of γ is $\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ with residual sum of squares

$$SS_{Res}(RM) = \mathbf{y}'\mathbf{y} - \hat{\gamma}'\mathbf{Z}'\mathbf{y} \quad (n - k - 1 + r \text{ degrees of freedom})$$

The full model contains more parameters than the reduced model and thus has higher explanatory power: $SS_{Res}(RM) \geq SS_{Res}(FM)$. Compute

$$SS_H = SS_{Res}(RM) - SS_{Res}(FM) \quad (n-k-1+r-(n-k-1) = r \text{ degrees of freedom})$$

and use the test statistic

$$F = \frac{SS_H/r}{SS_{Res}/(n-k-1)} \sim F_{r,n-k-1}$$

for the general linear hypothesis phrased above.

Example: Consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

and phrase the test statistic for the (simultaneous) test $H_0 : \beta_1 = \beta_3, \beta_2 = 0$.

Simultaneous Confidence Intervals on Regression Coefficients

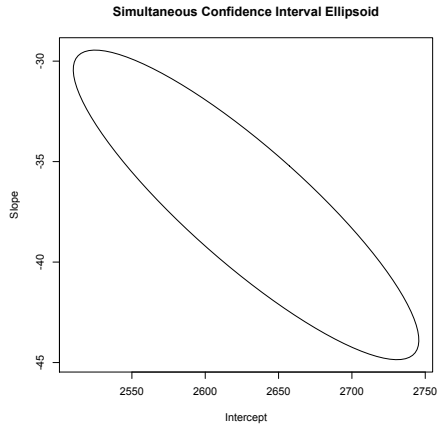
The confidence intervals that we have computed so far, were confidence intervals for a single slope or intercept. In multiple regression models it is not uncommon to compute several confidence intervals (one for each slope, for example) from the same set of data. The confidence level $100(1-\alpha)\%$ refers to each individual confidence interval. A set of confidence intervals that *all* contain their respective population parameters 95% of the time (with respect to repeated sampling) are called SIMULTANEOUS CONFIDENCE INTERVALS.

Definition: The JOINT CONFIDENCE REGION for the multiple regression parameter vector β can be formulated as follows:

$$\frac{(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)}{kMS_{Res}} \leq F_{\alpha,k,n-k-1}$$

This inequality described an elliptical region in k dimensional space. For simple linear regression ($k=2$) this is a two-dimensional ellipse.

Example: Construct the confidence ellipse for $\hat{\beta}_0$ and $\hat{\beta}_1$ for the Rocket Propellant data from Chapter 2.



Other methods for constructing simultaneous confidence intervals include the Bonferroni method which effectively splits the α into as many equal portions as confidence intervals need to be computed (say p) and then computes each interval individually at level $(1-\alpha/p)$.

Prediction in Multiple Linear Regression

Just as in simple linear regression, we may be interested to produce prediction intervals for specific or for general new observations. For a specific set of values of the predictor

$$\mathbf{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$$

a POINT ESTIMATE FOR A FUTURE OBSERVATION y at $\mathbf{x}_0$ is

$$\hat{y}_0 = \mathbf{x}_0'\hat{\beta}$$

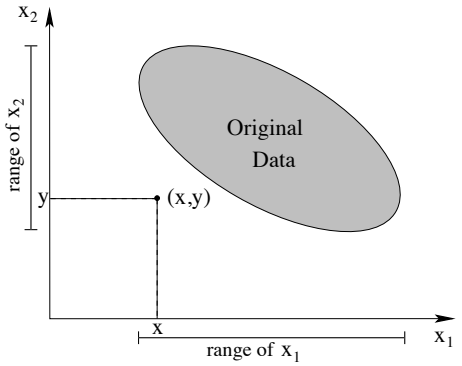
a $100(1-\alpha)\%$ prediction interval for this future observation is

$$\hat{y}_0 \pm t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)}$$

Example: For the Delivery Time data, calculate a 95% prediction interval for the time it takes to restock a vending machine with $x_1 = 8$ cases if the driver has to walk $x_2 = 275$ feet.

Note: In an introductory regression class, you may have learned that it is dangerous to predict new observations outside of the range of data you have collected. For instance, if you have data on the ages and heights of young girls, all between age 2 and 12, it would not be a good idea to use that linear regression model to predict the height of a 25 year old young woman. This concept of "outside the range" has to be extended in multiple linear regression.

Consider a regression problem with two predictor variables in which the collected data all falls within the ellipse in the picture shown on the right. The point (x, y) has coordinates that are each within the ranges of the observed variables individually, but it would still not be a good idea to predict the value of the response at this point, because we have no data to check the validity of the model in the vicinity of the point.



Standardized Regression Coefficients

The value of a slope $\hat{\beta}_j$ in a multiple regression problem depends on the units in which the corresponding predictor x_j is measured. This makes it difficult to compare slopes with each other. Both within the same model and across different models. To make slope estimates comparable, it is sometimes advantageous to scale them (make them unit less). These dimensionless regression coefficients are usually called STANDARDIZED REGRESSION COEFFICIENTS. There are different techniques for scaling the coefficients.

UNIT NORMAL SCALING: Subtract the sample mean and divide by the sample standard deviation both the predictor variables and the response:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad y_i^* = \frac{y_i - \bar{y}}{s_y}$$

where s_j is the estimated sample standard deviation of predictor x_j and s_y is the estimated sample standard deviation of the response. Using these new standardized variables, our regression model becomes

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

Question: What happened to the intercept?

The least squares estimator $\hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}^*$ is the standardized coefficient estimate.

UNIT LENGTH SCALING: Subtract the mean again, but now divide by the root of the sum of squares for each regressor:

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}, \qquad y_i^0 = \frac{y_i - \bar{y}}{\sqrt{SS_T}}$$

where $S_{jj} = \sum(x_{ij} - \bar{x}_j)^2$ is the corrected sum of squares for regressor x_j. In this case the regression model becomes

$$y_i^0 = b_1 w_{i1} + b_2 w_{i2} + \cdots + b_k w_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

and the vector of scaled least-squares regression coefficients is $\hat{\mathbf{b}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}^0$. The $\mathbf{W}'\mathbf{W}$ matrix is the correlation matrix for the k predictor variables. I.e., $\mathbf{W}'\mathbf{W}_{ij}$ is simply the correlation between x_i and x_j.

The matrices $\mathbf{Z}$ in unit normal scaling and $\mathbf{W}$ in unit length scaling are closely related and both methods will produce the exact same standardized regression coefficients $\hat{\mathbf{b}}$. The relationship between the original and scaled coefficients is

$$\hat{\beta}_j = \hat{b}_j \left(\frac{SS_T}{S_{jj}} \right)^{1/2}, \quad j = 1, 2, \ldots, k$$

Multicollinearity

In theory, one would like to have predictors in a multiple regression model that each have a different influence on the response and are independent from each other. In practice, the predictor variables are often correlated themselves. Multicollinearity is the prevalence of near-linear dependence among the regressors.

If one regressor were a linear combination of the other regressors, then the matrix $\mathbf{X}$ (whose columns are the regressors) would have linearly dependent columns, which would make the matrix $(\mathbf{X}'\mathbf{X})$ singular (non-invertible). In practice, it would mean that the predictor that can be expressed through the other predictors *cannot* contribute any new information about the response. But, worse than that, the linear dependence of the predictors makes the estimated slopes in the regression model arbitrary.

Example: Consider a regression model in which somebody's height (in inches) is expressed as a function of arm-span (in inches). Suppose the true regression equation is

$$y = 12 + 1.1x$$

Now, suppose further that when measuring the arm span, two people took independent measurements in inches (x_1) and in centimeters (x_2) of the same subjects and both variables have erroneously been included in the same linear regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

We know that in this case, $x_2 = 0.394x_1$ and thus we should have $\beta_1 + 0.394\beta_2 = 1.1$, in theory. But since this is a single equation with two unknowns, there are infinitely many possible solutions - some quite nonsensical. For instance, we could have $\beta_1 = -2.7$ and $\beta_2 = 9.645$. Of course, these slopes are not interpretable in the context of the original problem. The computer used to fit the data and to compute parameter estimates cannot distinguish between sensible and nonsensical estimates.

How can you tell whether you have multicollinearity in your data? Suppose your data have been standardized, so that $\mathbf{X}'\mathbf{X}$ is the correlation matrix for the k predictors in the model. The main diagonal elements of the inverse of the predictor correlation matrix are called the variance inflation factors (VIF). The larger these factors are, the more you should worry about multicollinearity in your model. On the other extreme, VIF's of 1 mean that the predictors are all orthogonal.

In general, the variance inflation factor for the j^{th} regressor coefficient can be computed as

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of multiple determination obtained from regressing x_j on the remaining predictor variables. We will discuss how to diagnose (and fix) effects of multicollinearity in more detail in Chapter 11 towards the end of the course.