Multiple linear regression

Stats 203, Winter 2024

Lecture 3 [Last update: January 17, 2024]

1 Linear model with multiple predictors

- In many situations, we have more than one – and possibly many – predictors at our disposal. A (linear) regression analysis with multiple predictors is called **multiple (linear) regression**.
- The linear model easily accommodates multiple predictors. In the multiple linear model each predictor $X_i = (X_{i1}, \ldots, X_{ip})$ is a vector of random variables rather than a single random variable. The model supposes that independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are each distributed according to

$$\begin{aligned} X_i &\text{ comes from an arbitrary distribution } P_i, \\ Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i, \end{aligned} \tag{1}$$

  where as before $\epsilon_i \perp\!\!\!\perp X_i$, $\mathbb{E}[\epsilon_i] = 0$, and $\text{Var}[\epsilon_i] = \sigma^2$. Notice that there is no assumption that the different predictors X_{ij}, X_{ik} are independent of one another. The assumption on the predictors is that they each contribute to the expectation of Y in an additive way.

2 Interpreting the model: correlation not causation

- What do the parameters of the linear model (1) mean? An appealingly simple interpretation is to say that β_j represents the expected change in Y if X_j were increased by one unit, with all other predictors X_k held fixed. Unfortunately this is dead wrong. Consider the simple linear model $GradeLevel = \beta_0 + \beta_1 ShoeSize + \epsilon$. It is intuitively clear that grade level and shoe size are positively correlated – hence, β_1 should be positive – but if you went out and bought a larger pair of shoes, you would not expect your grade level to change.

  This mistake is so common that is has gained popular awareness through the expression "correlation doesn't equal causation."

- The correct way to interpret the β_js comes from the mathematical fact that

$$\beta_j = \mathbb{E}[Y | X_1 = x_1, \ldots, X_j = x_j + 1, X_{j+1} = x_{j+1}, \ldots] - \mathbb{E}[Y | X_1 = x_1, \ldots, X_j = x_j, X_{j+1} = x_{j+1}, \ldots].$$

  That is, β_j is the expected difference in the response between observations whose value of X_j differs by 1, and whose values of X_k are equal for all $k \neq j$. This is not as simple as the causal interpretation, but it is correct.

3 Some of what we need of linear algebra and multivariable calculus

- We begin by recalling matrices and vectors. A length-n vector v is an $n \times 1$ array of numbers, and an $n \times p$ matrix $\mathbf{A}$ is an $n \times p$ array of numbers:

$$v = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{21} & \dots & a_{np} \end{bmatrix}.$$

  Two matrices $\mathbf{A} \in \mathbb{R}^{n \times p}, \mathbf{B} \in \mathbb{R}^{p \times q}$ multiply like

$$\mathbf{AB} = \begin{bmatrix} \sum_{i=1}^p a_{1i} b_{i1} & \sum_{i=1}^p a_{1i} b_{i2} & \dots & \sum_{i=1}^p a_{1i} b_{iq} \\ \sum_{i=1}^p a_{2i} b_{i1} & \sum_{i=1}^p a_{2i} b_{i2} & \dots & \sum_{i=1}^p a_{2i} b_{iq} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^p a_{ni} b_{i1} & \sum_{i=1}^p a_{ni} b_{i2} & \dots & \sum_{i=1}^p a_{ni} b_{iq} \end{bmatrix}.$$

  The **transpose** of a matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ swaps the role of rows and columns, so $\mathbf{A}^\top_{ij} = \mathbf{A}_{ji}$. The angle between two vectors $u \in \mathbb{R}^n, v \in \mathbb{R}^n$ is described by the **dot product** $u^\top v = \sum_{i=1}^n u_i v_i$. The length of a vector $v \in \mathbb{R}^n$ is given by the squared norm $\|v\|^2 = v^\top v = \sum_{i=1}^n v_i^2$. If $u^\top v = 0$ then the vectors are said to be **orthogonal**.

- We use the notation $\mathbf{A}_{\cdot j}$ for the jth column of a matrix $\mathbf{A}$, and $\mathbf{A}_{i\cdot}$ for the ith row of $\mathbf{A}$. The **rank** of $\mathbf{A}$ is the number of linearly independent columns of $\mathbf{A}$, where we recall that vectors $v_1, \dots, v_q$ are linearly independent if

$$c_1 v_1 + c_2 v_2 + \ldots + c_q v_q = 0 \Longrightarrow c_1 = c_2 = \ldots = c_q = 0.$$

  $\mathbf{A} \in \mathbb{R}^{n \times p}$ is said to be full rank if $\text{rank}(\mathbf{A}) = p$.

- A square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ is **invertible** if there exists a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_p$. Some useful criteria for invertibility: (i) a square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ is invertible if and only if it is full-rank, (ii) if $\mathbf{A} \in \mathbb{R}^{n \times p}$ is full rank, then $\mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{p \times p}$ is also full-rank, and as a result $\mathbf{A}^\top \mathbf{A}$ is invertible.

- The gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ collects the partial derivatives,

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

  The gradient of a linear function $L(x) = a^\top x$ is $\nabla L(x) = a$. The gradient of a quadratic function $Q(x) = x^\top \mathbf{A} x$ is $\nabla Q(x) = 2\mathbf{A}x$.

4 Ordinary least squares

- Now we want to estimate the parameters of the multiple linear model using data $(x_1, y_1), \dots, (x_n, y_n)$, $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$. As in the case of simple linear regression, this can be done by the method of least squares, i.e. by minimizing

$$\text{MSE}_n(b_0, b_1, \dots, b_p) = \frac{1}{n} \sum_{i=1}^n \big(y_i - (b_0 + x_{i1} b_1 + \ldots + x_{ip} b_p)\big)^2.$$

- In solving for the least-squares estimates it will be extremely convenient to introduce some matrix-vector notation. Collect the responses into a vector $y = (y_1 \cdots y_n)^\top \in \mathbb{R}^n$, the predictors into a matrix

$\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$, and the arguments of MSE_n into a vector $b = (b_0\ b_1\ \cdots\ b_p)^\top \in \mathbb{R}^p$. That is,

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots x_{1p} \\ 1 & x_{21} & x_{22} & \dots x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots x_{np}. \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$$

Notice that the first column of $\mathbf{X}$ is the all-ones vector. This allows us to write MSE concisely as $\text{MSE}_n(b) = \frac{1}{n}\|y - \mathbf{X}b\|^2 = \frac{1}{n}(y - \mathbf{X}b)^\top(y - \mathbf{X}b)$.

- The least-squares estimates can be found by setting the gradient of MSE_n equal to 0. The gradient of MSE_n with respect to b is

$$\nabla_b \text{MSE}_n(b) = \frac{2}{n}\mathbf{X}^\top(y - \mathbf{X}b) \tag{2}$$

The first-order condition for optimality is $\nabla_b \text{MSE}_n(\hat{\beta}) = 0$, giving the *normal equations*

$$\mathbf{X}^\top(y - \mathbf{X}\hat{\beta}) = 0 \iff \mathbf{X}^\top y = \mathbf{X}^\top \mathbf{X}\hat{\beta}. \tag{3}$$

Left-multiplying both sides of the equation by $(\mathbf{X}^\top\mathbf{X})^{-1}$, we have that

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

- The previous derivation assumed that $\mathbf{X}^\top\mathbf{X}$ was invertible, which turns out to be necessary for the least-squares solution to be well-defined. As stated above, $\mathbf{X}^\top\mathbf{X}$ is invertible if and only if $\mathbf{X}$ is full rank, i.e. $\text{rank}(\mathbf{X}) = p + 1$. This means $\mathbf{X}$ cannot have more columns than rows, so $\mathbf{X}$ must have at most n columns. In other words, there must be more observations than there are predictors, $n > p$.

  Even if $n > p$, $\mathbf{X}$ may not be full rank. For instance, if any j of the columns of $\mathbf{X}$ can be used to perfectly reconstruct a $(j+1)$st column then $\mathbf{X}$ will not be full rank. In this case we say the predictors are (perfectly) **multicollinear**. Later in the quarter we will learn about ways to deal with multicollinear data, or regression problems where $p \geq n$.

- The estimates $\hat{\beta}$ are called the **ordinary least squares** (OLS) estimates. The extra word "ordinary" is used to distinguish from other least-squares estimates: for example, there is also weighted least squares, in which each observation can receive a different weight. We will discuss some other least squares estimates later in the quarter.

- The OLS line is the linear function $\hat{m} : \mathbb{R}^p \to \mathbb{R}$ defined by

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p.$$

As in simple linear regression, we can think of this as an estimate of the conditional mean $m(x) = \mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. At a new value of predictors x, our prediction for the response is $\hat{m}(x)$.

5 The geometry of least squares

- For each value of the predictors $x_i, i = 1, \ldots, n$, the OLS line predicts that the response will be

$$\hat{y}_i := \hat{m}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}.$$

The $\hat{y}_i$s are known as the **fitted values**. (The term predicted value is reserved for $\hat{m}(x)$ evaluated at a new value of x.) The difference between observed and fitted values, $e_i := y_i - \hat{y}_i$, are known as the **residuals**. We will collect the fitted values and residuals into vectors $\hat{y} = (\hat{y}_1 \ldots \hat{y}_n)^\top$ and $e = (e_1 \ldots e_n)^\top$.

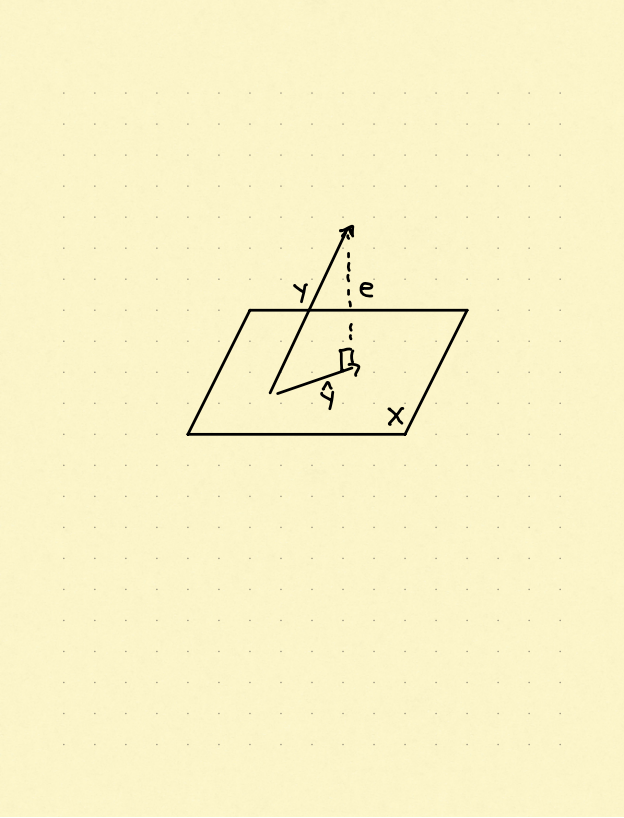


Figure 1: Geometric diagram of least squares. There are $n = 3$ observations, so the vector $y \in \mathbb{R}^3$. The predictor matrix has two columns (so $p = 1$) whose span is visualized as a 2d plane in 3d space. The fitted values $\hat{y}$ lie in the plane. The residuals e are normal to the plane, hence orthogonal to everything in it include $\hat{y}$. The sum of $\hat{y}$ and e is y.

- By construction
$$y_i = \hat{y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip} + e_i.$$
Warning: do not confuse this equation with the modeling equation $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i$. The former is an algebraic identity, the latter an assumption. Likewise, do not confuse residuals with random errors of the linear model!

- Notice that the normal equation (3) can be rearranged to read $\mathbf{X}^\top(y - \mathbf{X}\hat{\beta}) = 0$. But $\mathbf{X}\hat{\beta} = \hat{y}$, so $y - \mathbf{X}\hat{\beta} = e$ and consequently
$$\mathbf{X}^\top e = 0.$$
This simple equation turns out to be at the heart of the theory of linear regression. Geometrically, it means that *the residuals are orthogonal to any linear combination of the columns of* $\mathbf{X}$:
$$v \in \text{span}(\mathbf{X}_{\cdot 1}, \ldots, \mathbf{X}_{\cdot p+1}) \Longrightarrow v^\top e = 0.$$
For example, the all-ones vector is the first column of the predictor matrix, $\mathbf{1} = (1 \ldots 1)^\top = \mathbf{X}_{\cdot 1}$, and so $\mathbf{1}^\top e = 0$: *the residuals must sum up to zero.* Notice that $\hat{y} = \mathbf{X}\hat{\beta}$ is a linear combination of the columns of $\mathbf{X}$. Thus $\hat{y}^\top e = 0$: *the residuals are orthogonal to the fitted values.*

- For a diagram of these relations, see Figure **??**.

6 Multiple R^2

- As with simple linear regression, after having computed the least-squares estimates, it seems natural to ask how good the estimates are. The most common way measuring the quality of OLS estimates is called **(multiple)** R^2, which works by comparing the predictions of the OLS line to the actual ys:

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

  R^2 is an example of a measure of **goodness-of-fit**, since it measures how well least-squares estimates fit the data. You need to know what R^2 is because it is so commonly used in practice; but later, we will see better ways of measuring goodness-of-fit.

- R^2 satisfies similar properties to sample correlation. For example,
  - R^2 is unitless: any linear transformation of the xs or ys does not change the R^2.
  - R^2 is between 0 and 1. Larger values of R^2 indicate a stronger linear relationship between predictors and response. If the observed xs perfectly predict the observed ys – meaning $\hat{y}_i = y_i$ for $i = 1, \ldots, n$ – then $R^2 = 1$. If the observed xs predict no better than sample mean – meaning $\hat{y}_i = \bar{y}$ for $i = 1, \ldots, n$ – then $R^2 = 0$.

  It is not immediately obvious that $R^2 > 0$ (the other properties should be obvious). Let's derive this, using the orthogonality of residuals and predictors. Focus on the term $\sum_{i=1}^n (y_i - \bar{y})^2$ which is the denominator of $1 - R^2$. Add and subtract by the fitted values, and expand the square:

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2e_i(\hat{y}_i - \bar{y}).
\end{aligned}$$

  The sum of the cross-terms can be written in matrix-vector notation as $2e^\top(\hat{y} - \bar{y}\mathbf{1})$. But we have shown that $e^\top \hat{y} = 0$ and $e^\top \mathbf{1} = 0$. So the sum of the cross-terms equals 0, and

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \tag{4}$$

  We conclude that

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \tag{5}$$

  Thus R^2 must be non-negative.

- There is another interpretation of R^2, that follows from (5). Notice that $e^\top \mathbf{1} = 0 \Leftrightarrow \hat{y}^\top \mathbf{1} = y^\top \mathbf{1}$, meaning the sample mean of the fitted values and responses is the same, $\bar{\hat{y}} = \bar{y}$. So we further have

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{s_{\hat{y}}^2}{s_y^2}.$$

  In words, R^2 is the *proportion of variance in y that is explained by x*. For this reason R^2 is sometimes called the **fraction of variance explained**. Freedman 4.3 has a nice rant against this terminology.

- There are a number of subtleties involved in using R^2.

- Goodness-of-fit means something different than the model being correct. When the error variance σ^2 is large, the R^2 will be small even if the fitted model is exactly correct. On other hand, there are cases where the true model is *not* linear but R^2 is (arbitrarily) close to 1.
- Adding more predictors will almost always increase the R^2 (and it will *never* decrease the R^2) whether or not they have any true relationship with the response.
- R^2 says nothing about whether the model will predict well at a new value of the predictors.
- R^2 only measures the strength of the linear relationship between x and y. It is possible for x to perfectly predict y – meaning $y = f(x)$ is a deterministic function of x – but for the R^2 to equal 0.

For these reasons and others it is challenging to interpret what R^2 means and what constitutes a "large" value of R^2.

Computing OLS with delivery data

We will work through an example from Chapter 3 of *Introduction to Regression Analysis* (Montgomery, Peck and Vining). Excerpting from that book:

> A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. It has been suggested that the two most important variables affecting delivery time are the number of cases of product stocked and the distance walked by the route driver.

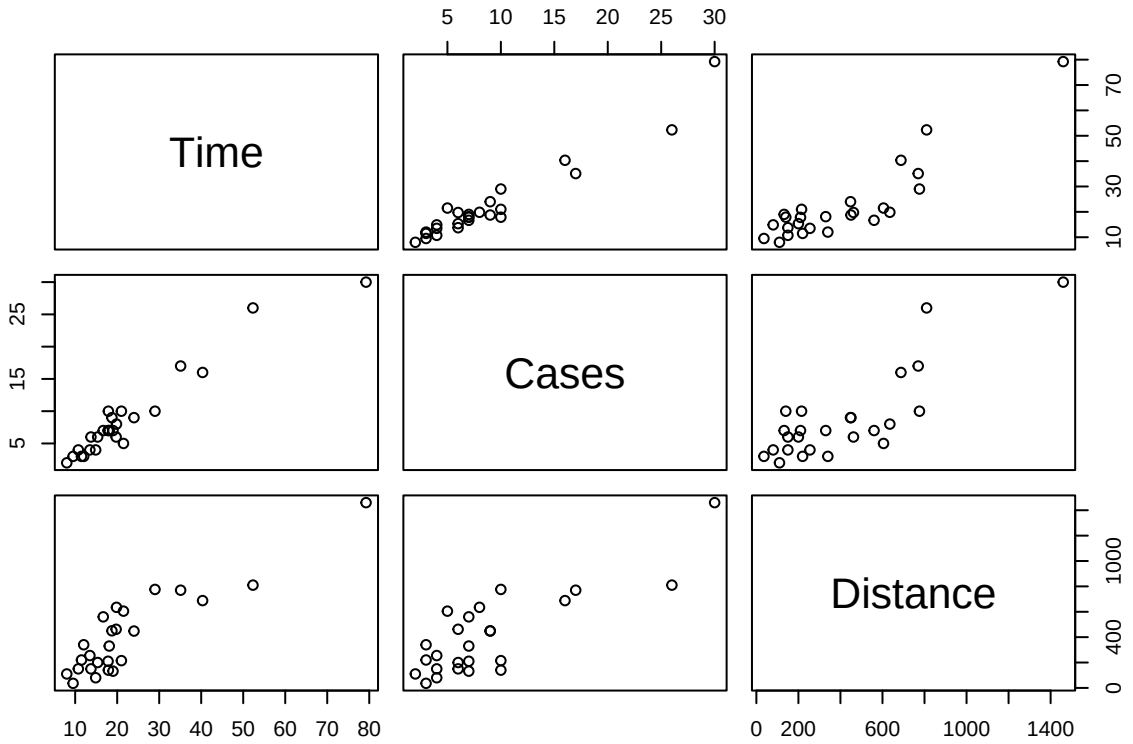
We start by loading the data.

```
delivery = read.table("delivery.txt")
head(delivery) # Looks at all columns, first 6 rows.
```

```
##    Time Cases Distance
## 1 16.68     7      560
## 2 11.50     3      220
## 3 12.03     3      340
## 4 14.88     4       80
## 5 13.75     6      150
## 6 18.11     7      330
```

As always we start with some EDA to assess whether a linear model is plausible. When applied to a data frame the `plot` function makes bivariate scatterplots of all pairs of variables. These are known as **pairs plots**.

```
plot(delivery)
```



From the first row – which shows the relationship between the response (`Time`) and the predictors (`Cases` and `Distance`) – the linear model

$$\text{Time} = \beta_0 + \beta_1 \text{cases} + \beta_2 \text{distance} + \epsilon$$

seems reasonable.

So we will compute the ordinary least squares estimate, in two ways: (1) using the formula, (2) using the `lm` function.

(Note: this is not really the proper way to check whether the assumptions of the linear model are met. We will treat this issue later in the quarter when we talk about diagnostics.)

Compute the OLS estimate using the formula

First we save the predictors and response as matrices/vectors. It would be reasonable to exclude the intercept term but we will include an intercept to show how it works.

```
delivery$Intercept = 1
X = cbind(delivery$Intercept,delivery$Cases,delivery$Distance)
y = delivery$Time
```

Compute the matrix $\mathbf{X}^\top \mathbf{X}$.

```
A = t(X) %*% X
A
```

```
##       [,1]   [,2]    [,3]
## [1,]    25    219   10232
## [2,]   219   3055  133899
## [3,] 10232 133899 6725688
```

Compute the vector $\mathbf{X}^\top y$.

```
b = t(X) %*% y
b
```

```
##           [,1]
## [1,]    559.60
## [2,]   7375.44
## [3,] 337071.69
```

Solve for the least squares estimate.

```
betahat = solve(A) %*% b # solve(A) computes A^{-1}
betahat
```

```
##            [,1]
## [1,] 2.34123115
## [2,] 1.61590721
## [3,] 0.01438483
```

Compute the OLS using lm

Note the syntax for writing formulas in R. Use `help(lm)` for more information and examples.

```
delivery.lm = lm(Time ~ Cases + Distance,data = delivery)
delivery.lm$coef
```

```
## (Intercept)       Cases    Distance
##  2.34123115  1.61590721  0.01438483
```

The two solutions are identical up to numerical error.

Verifying the geometry

We look at the angle between fitted values and residuals. The vectors are perpendicular (up to numerical error.)

```
yhat = X %*% betahat
e = y - yhat
t(yhat) %*% e
```

```
##               [,1]
## [1,] 3.072032e-11
```

Computing the R-squared

The `lm` function computes a lot of information besides the least squares estimates, including the R^2. Look at the `Multiple R-squared` entry in the output below:

```
summary(delivery.lm)
```

```
##
## 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
```

We compare this with R^2 as computed using the formula.

```
var(yhat)/var(y)
```

```
##           [,1]
## [1,] 0.9595937
```

To four digits, they agree.