14
Linear Models and Regression

Imagine you measured an output (y) given an input (x). You hypothesized a linear relationship between x and y, i.e.

$$y = \beta x$$

If $y = 2.4$ when $x = 2$, then you can easily calculate the value of the parameter β.

$$\beta = y/x = 2.4/2 = 1.2$$

You could plot the relationship between x and y, which is just the line $y = 1.2x$. Given an input x, it is easily to calculate the corresponding value y. All of this works because we've assumed our measurements of the input x and output y are exact. With only one unknown (β), we have sufficient information to calculate a value with only one set of observations. However, life is messy. Measurements are noisy and filled with error and uncertainty. Even if the *true* relationship between y and x was really a factor of 1.2, we would never observe a value of y that was exactly 1.2 times x.

To compensate for the imprecision of the real world, engineers make multiple observations of their systems. Imagine we made five “noisy" measurements of the input x and output y, which we will call x^{obs} and y^{obs}.

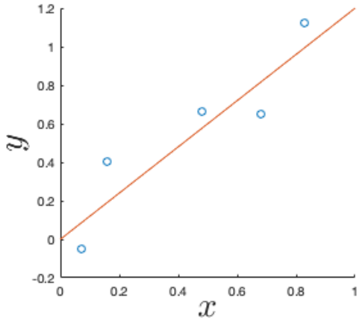

Figure 14.1: Five noisy observations (circles) of the linear relationship $y = 1.2\,x$.

x^{obs}	y^{obs}
0.07	-0.05
0.16	0.40
0.48	0.66
0.68	0.65
0.83	1.12

As a measure of uncertainty, we can compare the measured outputs (y^{obs}) and the predicted outputs (y^{pred}), which we calculate using our model and the observed inputs ($y^{\text{pred}} = 1.2\,x^{\text{obs}}$). The differences between the predicted outputs and the observed outputs ($y^{\text{pred}} - y^{\text{obs}}$) are called the *residuals*.

x^{obs}	y^{obs}	Prediction ($y^{\text{pred}} = 1.2\,x^{\text{obs}}$)	Residual ($y^{\text{pred}} - y^{\text{obs}}$)
0.07	-0.05	0.084	0.134
0.16	0.40	0.192	-0.208
0.48	0.66	0.576	-0.084
0.68	0.65	0.816	0.166
0.83	1.12	0.996	-0.124

From the above plot we can be reasonably certain that $y = 1.2x$ is a good representation of the relationship between y and x. But none of the five outputs y is exactly equal to $1.2x$. How are we able to recover our estimate that $\beta = 1.2$? We have five separate observations that can be used to estimate β.

1. $\beta = -0.05/0.07 = -0.7$
2. $\beta = 0.40/0.16 = 2.5$
3. $\beta = 0.66/0.48 = 1.3$
4. $\beta = 0.65/0.68 = 0.9$
5. $\beta = 1.12/0.83 = 1.3$

Our estimates for β vary wildly, from -0.7 to 2.5, and none of the estimates match the true value of 1.2. We could average the individual estimates to produce a single composite result ($\beta = 1.09$), but we would still miss the true value by nearly 9%.

Using the same five observations and the statistical techniques in this chapter, we can estimate β to be 1.21, an error of less than 1%.

In Chapter 4 we quantified exactly the amount of information we need to solve a linear system. In short, we need one linearly independent observation for each unknown. What if we have more observations, but the observations are noisy? When we have too much or too little information, we change our approach from solving linear systems directly to fitting linear models approximately. In this chapter we will create and fit models that are never exact by nonetheless useful. **The goal of model fitting is not to find exact values for the unknowns, but rather to minimize the error between the observed outputs and the outputs predicted by the model.**

14.1 Quantifying Error

We need a method to measure the error between the observed and predicted outputs of our model. There are two requirements for measuring error. First, we require that the error never be negative, i.e. $\text{Error}(y^{\text{obs}}, y^{\text{pred}}) \geq 0$. Second, we require that the error be zero if and only if the predicted value matches the observed value exactly:

$$\text{Error}(y^{\text{obs}}, y^{\text{pred}}) = 0 \iff y^{\text{obs}} = y^{\text{pred}}$$

Many functions satisfy these rules. The most common are the absolute value

$$\text{Error}(y^{\text{obs}}, y^{\text{pred}}) = |y^{\text{obs}} - y^{\text{pred}}|$$

and the *squared error*

$$\text{Error}(y^{\text{obs}}, y^{\text{pred}}) = (y^{\text{obs}} - y^{\text{pred}})^2$$

We will use the squared error because it has several advantages over the absolute value:

1. Quadratic functions like the squared error are continuously differentiable, while the derivative of the absolute value is discontinuous at zero. A continuous first derivative makes it easy to optimize functions involving the squared error.

2. The squared error more harshly penalizes predictions that are far from the observed values. If a prediction is twice as far from an observation, the squared error increases by a factor of four while the absolute value error only doubles. We will see shortly that assigning large penalties to far away points is more intuitive.

3. There is always a single solution when minimizing squared error, but there can be infinitely many solutions that minimize the absolute error. We prefer having a unique solution whenever possible.

14.2 Fitting Linear Models

Now that we've settled on a method for quantifying error, let's formalize the process of fitting linear models. There are three steps.

1. Choose a model that you think explains the relationship between inputs (x) and outputs (y). The models should contains unknown parameters (β) that you will fit to a set of observations. **The model should be linear with respect to the parameters (β). It does not need to be linear with respect to the inputs (x) or outputs (y).**

2. To find values for the unknown parameters (β), we will minimize the total error between the observed outputs (y^{obs}) and the outputs predicted from the model

$$\min_{\beta} \sum_{y^{\text{obs}}} \text{Error}(y^{\text{obs}}, y^{\text{pred}})$$

or, for the specific case when we choose the squared error

$$\min_{\beta} \sum_{y^{\text{obs}}} \left(y^{\text{obs}} - y^{\text{pred}} \right)^2$$

Substitute the model you selected in Step 1 in place of y^{pred} in the above minimization.

3. Minimize the function by taking the derivative of the sum squared error and setting it equal to zero. Solve for the unknown parameters β.

You should also check that your solution is a minimum, not a maximum or inflection point by checking that the second derivative is positive.

Fitting a single parameter, constant model $y = \beta_0$.

The simplest model we can fit has only a single parameter and no dependence on the inputs:

$$y^{\text{pred}} = \beta_0$$

This might seem like a silly model. Given an input observation x^{obs}, we ignore the input and predict that y will always be equal to β_0. For example, imagine we are predicting someone's height (y^{pred}) given their age (x^{obs}). Rather than make a prediction based on the person's age, we simply guess the same height for everyone (β_0).

Regardless of the utility of such a simple model, let's fit it to a set of n pairs of observations (x^{obs}, y^{obs}). We've already completed Step 1 by choosing the model $y^{\text{pred}} = \beta_0$. We begin Step 2 by writing our goal, which is to choose a value for β_0 using the n observations that minimizes the sum squared error.

$$\min_{\beta_0} \sum_{i=1}^{n} \left(y_i^{\text{obs}} - y_i^{\text{pred}} \right)^2$$

Now we substitute with our model $y^{\text{pred}} = \beta_0$.

$$\min_{\beta_0} \sum_{i=1}^{n} \left(y_i^{\text{obs}} - \beta_0 \right)^2$$

To minimize the error we find where the derivative of the error **with respect to the parameter** β_0 is zero. Remember that the values y^{obs} are constants in the error function. They are numbers given to us so we can choose an appropriate value for β_0.

Remember that the derivative of a sum is the sum of the derivative.

$$
\begin{aligned}
\frac{d}{d\beta_0}\left(\sum_{i=1}^{n}(y_i^{\text{obs}}-\beta_0)^2\right) &= 0\\
\sum_{i=1}^{n}\left(\frac{d}{d\beta_0}(y_i^{\text{obs}}-\beta_0)^2\right) &= 0\\
\sum_{i=1}^{n}\Big(2(y_i^{\text{obs}}-\beta_0)(-1)\Big) &= 0\\
-2\sum_{i=1}^{n}(y_i^{\text{obs}}-\beta_0) &= 0\\
\sum_{i=1}^{n}y_i^{\text{obs}}-\sum_{i=1}^{n}\beta_0 &= 0\\
\sum_{i=1}^{n}y_i^{\text{obs}}-n\beta_0 &= 0
\end{aligned}
$$

We can rearrange the final equation and discover that the optimal value for the parameter β_0 is

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n}y_i^{\text{obs}}$$

If our strategy is to always guess the same output value (β_0), the best value to guess is the mean of the observed outputs y^{obs}. Said another way, the mean is the best fit of a constant model to a set of data. If we need to represent a set of numbers with a single number, choosing the mean minimizes the squared error.

A couple interesting things come from this result. First, now you know where the mean comes from. It is the least squares estimate for a set of points. Second, the mean does not minimize the absolute error in a set of points – this is a common misconception! If we repeated the same calculation using the absolute error instead of the squared error we would discover that the least absolute estimator for a set of numbers is the median, not the mean.

The phase “least squares" is a convenient way of saying “minimizes the sum of the squared error".

Going further, if we make our error function binary (the error is zero if $y^{\text{obs}} = y^{\text{pred}}$ and one otherwise), the best estimator is called the *mode*, or the most frequent value in the set of observed outputs.

Fitting a two parameter model $y = \beta_0 + \beta_1 x$.

Let’s fit a more complicated model that uses the observed inputs x^{obs} when making output predictions y^{obs}. Our model has the form

$$y^{\text{pred}} = \beta_0 + \beta_1 x^{\text{obs}}$$

with two unknown parameters β_0 and β_1. This is a linear model with respect to the parameters. We discovered earlier that the functions of the form $y = \beta_0 + \beta_1 x$ are not linear with respect to x (they are affine).

But remember that x^{obs} is not an independent variable in the model. It is a known, observed value – a constant. The unknowns in a linear model are the parameters, not y or x.

Now that we've chosen our model, we write our goal to minimize the sum squared error.

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(y_i^{\text{obs}} - y_i^{\text{pred}}\right)^2$$

Notice we are minimizing over both parameters β_0 and β_1. Substituting our model for the value y^{pred} yields

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(y_i^{\text{obs}} - \beta_0 - \beta_1 x_i^{\text{obs}}\right)^2$$

The total error is minimized when the derivatives with respect to both β_0 and β_1 are zero. Let's start by taking the derivative or the error with respect to β_0.

We are using partial derivatives since our error is a function of more than one unknown parameter.

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i^{\text{obs}} - \beta_0 - \beta_1 x_i^{\text{obs}})^2 &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (y_i^{\text{obs}} - \beta_0 - \beta_1 x_i^{\text{obs}})^2 \\
&= -2 \sum_{i=1}^{n} (y_i^{\text{obs}} - \beta_0 - \beta_1 x_i^{\text{obs}}) \\
&= -2 \left(\sum_{i=1}^{n} y_i^{\text{obs}} - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i^{\text{obs}} \right) \\
&= -2 \left(\sum_{i=1}^{n} y_i^{\text{obs}} - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i^{\text{obs}} \right)
\end{aligned}$$

We set this derivative equal to zero and solve for β_0.

We call β_0 (or the affine parameter in a linear model) the *grand mean* since it equals the mean of the outputs when all inputs are zero.

$$\begin{aligned}
\beta_0 &= \frac{1}{n} \sum_{i=1}^{n} y_i^{\text{obs}} - \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i^{\text{obs}} \\
&= \text{mean}[y^{\text{obs}}] - \beta_1 \text{mean}[x^{\text{obs}}]
\end{aligned}$$

We see that β_0 depends on the mean input, the mean output, and the parameter β_1. Let's substitute our value for β_0 into the total error.

$$\begin{aligned}
\sum_{i=1}^{n} \left(y_i^{\text{obs}} - \beta_0 - \beta_1 x_i^{\text{obs}}\right)^2 &= \sum_{i=1}^{n} \left(y_i^{\text{obs}} - \text{mean}[y^{\text{obs}}] + \beta_1 \text{mean}[x^{\text{obs}}] - \beta_1 x_i^{\text{obs}}\right)^2 \\
&= \sum_{i=1}^{n} \left(y_i^{\text{obs}} - \text{mean}[y^{\text{obs}}] + \beta_1 (\text{mean}[x^{\text{obs}}] - x_i^{\text{obs}})\right)^2
\end{aligned}$$

Now we find the optimal value for the parameter β_1. First we differentiate the total error with respect to β_1.

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \left(y_i^{\text{obs}} - \text{mean}[y^{\text{obs}}] + \beta_1 (\text{mean}[x^{\text{obs}}] - x_i^{\text{obs}})\right)^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} \left(y_i^{\text{obs}} - \text{mean}[y^{\text{obs}}] + \beta_1(\text{mean}[x^{\text{obs}}] - x_i^{\text{obs}}) \right)^2 \\
&= 2 \sum_{i=1}^{n} \left(y_i^{\text{obs}} - \text{mean}[y^{\text{obs}}] + \beta_1(\text{mean}[x^{\text{obs}}] - x_i^{\text{obs}}) \right) (\text{mean}[x^{\text{obs}}] - x_i^{\text{obs}}) \\
&= -2 \left(\sum_{i=1}^{n} \left(y_i^{\text{obs}} - \text{mean}[y^{\text{obs}}] \right) \left(x_i^{\text{obs}} - \text{mean}[x^{\text{obs}}] \right) - \beta_1 \sum_{i=1}^{n} \left(x_i^{\text{obs}} - \text{mean}[x^{\text{obs}}] \right)^2 \right)
\end{aligned}
$$

We set the derivative equal to zero and solve for the parameter β_1.

$$
\beta_1 = \frac{\sum_{i=1}^{n} \left(y_i^{\text{obs}} - \text{mean}[y^{\text{obs}}] \right) \left(x_i^{\text{obs}} - \text{mean}[x^{\text{obs}}] \right)}{\sum_{i=1}^{n} \left(x_i^{\text{obs}} - \text{mean}[x^{\text{obs}}] \right)^2}
$$

Let's try fitting the expression $y^{\text{pred}} = \beta_0 + \beta_1 x^{\text{obs}}$ to the data from earlier in this chapter:

x^{obs}	y^{obs}
0.07	-0.05
0.16	0.40
0.48	0.66
0.68	0.65
0.83	1.12

First we calculate the means of the inputs and outputs.

$$
\begin{aligned}
\text{mean}[x^{\text{obs}}] &= (1/5)(0.07 + 0.16 + 0.48 + 0.68 + 0.83) = 0.44 \\
\text{mean}[y^{\text{obs}}] &= (1/5)(-0.05 + 0.40 + 0.66 + 0.65 + 1.12) = 0.56
\end{aligned}
$$

Now we can calculate the value for the parameter β_1. It's easiest to make a table.

x^{obs}	y^{obs}	$(x^{\text{obs}} - \text{mean}[x^{\text{obs}}])(y^{\text{obs}} - \text{mean}[y^{\text{obs}}])$	$(x^{\text{obs}} - \text{mean}[x^{\text{obs}}])^2$
0.07	-0.05	$(0.07 - 0.44)(-0.05 - 0.56) = 0.23$	$(0.07 - 0.44)^2 = 0.14$
0.16	0.40	$(0.16 - 0.44)(0.40 - 0.56) = 0.044$	$(0.16 - 0.44)^2 = 0.081$
0.48	0.66	$(0.48 - 0.44)(0.66 - 0.56) = 0.0037$	$(0.48 - 0.44)^2 = 0.0013$
0.68	0.65	$(0.68 - 0.44)(0.65 - 0.56) = 0.022$	$(0.68 - 0.44)^2 = 0.056$
0.83	1.12	$(0.83 - 0.44)(1.12 - 0.56) = 0.22$	$(0.83 - 0.44)^2 = 0.15$

$$
\begin{aligned}
\beta_1 &= \frac{\sum_{i=1}^{n} \left(y_i^{\text{obs}} - \text{mean}[y^{\text{obs}}] \right) \left(x_i^{\text{obs}} - \text{mean}[x^{\text{obs}}] \right)}{\sum_{i=1}^{n} \left(x_i^{\text{obs}} - \text{mean}[x^{\text{obs}}] \right)^2} \\
&= \frac{0.23 + 0.044 + 0.0037 + 0.022 + 0.22}{0.14 + 0.081 + 0.0013 + 0.056 + 0.15} \\
&= 1.21
\end{aligned}
$$

We can use the value of the parameter β_1 to find the other parameter β_0.

$$\begin{aligned}\beta_0 &= \text{mean}[y^{\text{obs}}] - \beta_1\text{mean}[x^{\text{obs}}]\\ &= 0.56 - (1.21)(0.44)\\ &= 0.020\end{aligned}$$

According to our five observations, the best fit least squares estimate is

$$y = 0.020 + 1.21x$$

This agrees well with our hypothesized relationship that $y = 1.2x$.

14.3 Matrix formalism for linear models

You might be thinking that there has to be an easier method for fitting linear models. Finding formulae for the parameters is unwieldy, and the problem only worsens as the number of parameters grows. Fortunately, linear algebra can help.

Let's return to our two parameter model $y = \beta_0 + \beta_1 x$. Using the five data points from the previous section, we can write five linear equations, one for each point

$$\begin{aligned}-0.05 &= \beta_0 + 0.07\beta_1 + \epsilon_1\\ 0.40 &= \beta_0 + 0.16\beta_1 + \epsilon_2\\ 0.66 &= \beta_0 + 0.48\beta_1 + \epsilon_3\\ 0.65 &= \beta_0 + 0.68\beta_1 + \epsilon_4\\ 1.12 &= \beta_0 + 0.83\beta_1 + \epsilon_5\end{aligned}$$

All we've done to write these equations is substituted the observed values for x and y and added an *error term* (ϵ_i). Remember that each observation is imprecise, so the observed value of y will never exactly equal the predicted value $\beta_0 + \beta_1 x$. Since our equations must be exact, we add a term to each equation to hold the error between the predicted and observed values. The same equations can be written in matrix form.

$$\begin{pmatrix}-0.05\\ 0.40\\ 0.66\\ 0.65\\ 1.12\end{pmatrix} = \begin{pmatrix}1 & 0.07\\ 1 & 0.16\\ 1 & 0.48\\ 1 & 0.68\\ 1 & 0.83\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} + \begin{pmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \epsilon_4\\ \epsilon_5\end{pmatrix}$$

The parameters β_0 and β_1 are the same for every equation, but each equation has its own error term.

Or, more succinctly,

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

There are several noteworthy things about the above expression.

- The variable $\mathbf{y}$ is a vector of the outputs (responses), $\boldsymbol{\epsilon}$ is a vector of errors, and $\boldsymbol{\beta}$ is a vector of the unknown parameters.

- The inputs (or predictor variables) form a matrix $\mathbf{X}$ called the *design matrix*.

- The first column in $\mathbf{X}$ is all ones. This column corresponds to the constant parameter in the model, β_0.

- The unknowns in the equation are the parameters in the vector $\boldsymbol{\beta}$, not the values in the matrix $\mathbf{X}$. The values in $\mathbf{X}$ are known inputs from our dataset.

The term design matrix comes from statistics. The nonzero coefficients in $\mathbf{X}$ correspond to inputs that are set to give the responses in $\mathbf{y}$. Thus $\mathbf{X}$ mimics the design of the experiment.

The errors $\boldsymbol{\epsilon}$ are also unknown, but we do not solve for these explicitly.

Fitting a linear model involves finding a set of values for the vector $\boldsymbol{\beta}$. There are a few complications to solving the linear system $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. First, our goal is not to find any values for the parameters in $\boldsymbol{\beta}$, but to find the values that minimize the square of the error terms in $\boldsymbol{\epsilon}$ (i.e. the least squares solution). Second, the matrix $\mathbf{X}$ is almost never square. We often have more rows (observations) that we have columns (parameters) to compensate for the noise in our measurements.

Fortunately, there is a tool from matrix theory – the pseudoinverse – that overcomes both these difficulties. The least squares solution to the problem $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is

$$\boldsymbol{\beta} = \mathbf{X}^{+}\mathbf{y}$$

We've seen the pseudoinverse before when discussing the singular value decomposition (see section 13.3). If we decompose a matrix by SVD into $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}$, then the pseudoinverse is $\mathbf{X}^{+} = \mathbf{V}\boldsymbol{\Sigma}^{+}\mathbf{U}^{\mathsf{T}}$. There is another formula for calculating the pseudoinverse that is useful for linear models:

$$\mathbf{X}^{+} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}$$

To understand this formula, consider the linear system

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

Let's multiply both sizes by the matrix $\mathbf{X}^{\mathsf{T}}$.

$$\mathbf{X}^{\mathsf{T}}\mathbf{y} = \mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}$$

We know that the matrix $\mathbf{X}$ is not square; however, the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is always square. Since $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is square, it is invertible provided it is full rank. While we won't get into the details here, the requirement that $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ be full rank is usually satisfied in problems that arise in engineering. Assuming $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ exists, let's multiply both sides of our equation by it.

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}$$

If matrix $\mathbf{X}$ has m rows and n columns, the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ has n rows and n columns.

Look carefully at the righthand side. We have the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ multiplied by its inverse, $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$. This is equal to the identity matrix, so

$$(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \boldsymbol{\beta}$$

We have solved the system $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ for the vector $\boldsymbol{\beta}$, so the quantity $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$ on the lefthand side must be the pseudoinverse of the matrix $\mathbf{X}$.

Let's use pseudoinversion to solve our example model:

$$\begin{pmatrix} -0.05 \\ 0.40 \\ 0.66 \\ 0.65 \\ 1.12 \end{pmatrix} = \begin{pmatrix} 1 & 0.07 \\ 1 & 0.16 \\ 1 & 0.48 \\ 1 & 0.68 \\ 1 & 0.83 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{pmatrix}$$

Using MATLAB's `pinv` function we can calculate the pseudoinverse of the design matrix

$$\mathbf{X}^+ = \begin{pmatrix} 0.59 & 0.50 & 0.16 & -0.046 & -0.20 \\ -0.88 & -0.67 & 0.084 & 0.55 & 0.91 \end{pmatrix}$$

and find the least squares estimates for the parameters

$$\boldsymbol{\beta} = \mathbf{X}^+\mathbf{y} = \begin{pmatrix} 0.59 & 0.50 & 0.16 & -0.046 & -0.20 \\ -0.88 & -0.67 & 0.084 & 0.55 & 0.91 \end{pmatrix} \begin{pmatrix} -0.05 \\ 0.40 \\ 0.66 \\ 0.65 \\ 1.12 \end{pmatrix} = \begin{pmatrix} 0.020 \\ 1.21 \end{pmatrix}$$

Again, we see that $\beta_0 = 0.020$ and $\beta_1 = 1.21$.

Calculating the pseudoinverse

Our new formula for the pseudoinverse ($\mathbf{X}^+ = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$) gives us intuition about solving nonsquare linear systems – we are actually solving a related system involving the special matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$. This form of the pseudoinverse requires calculating a matrix inverse, which we have all sworn never to do except in dire situations. Calculating the pseudoinverse using the SVD is far more efficient. The function `pinv` in MATLAB uses a variant of the SVD method to find pseudoinverses.

Dimensions of the design matrix

The pseudoinverse of $\mathbf{X}$ is part of the least square solution for the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Each row in $\mathbf{X}$ is an observation and each column corresponds to an unknown parameter. If $\mathbf{X}$ is square, we have one observation per parameter. We know that the pseudoinverse of a square, invertible matrix is identical to the ordinary inverse. We also know that linear systems with square coefficient matrices have a unique solution. There is no room to find a solution that minimizes the error when there is only a single unique solution. We can fit all of the parameters, but as we will see later, we have no information about how well we did minimizing error.

If we have more observations than parameters (**X** has more rows than columns), the extra information in the observations can be used to estimate how well our solution minimizes the sum squared error. The extra degrees of freedom can quantify our confidence in the model.

Finally, our system is *underdetermined* if we have fewer observations than parameters. The search space for parameters is simply too large, and we often cannot find a meaningful solution. Fitting these models requires special tools that we will discuss in a later chapter.