86-631/42-631 Neural Data Analysis
Lecture 09: Linear Regression

1) Correlation describes linear dependence
2) Ordinary linear regression
3) Multiple linear regression
4) R^2

We've spoken quite a bit about neural codes and information – we now know that neurons can carry information about stimuli in a number of different ways. Now comes perhaps a more important question: given a particular stimulus, can we predict the firing of the neuron? Or perhaps the alternate question: given that a neuron is firing in a particular way, how can we predict the stimulus? We are going to spend today's lecture talking about one particular prediction method: linear prediction. Linear regression is one of the most useful tools in all of statistics, and that remains as true for neuroscience as it does for any other field. Today we'll introduce linear regression, its relationship to correlation, how to estimate regression parameters, and how to characterize the goodness of fit.

Correlation describes linear dependence

Suppose we have two random variables, X and Y, and we are interested in building the best linear predictor we can of Y from the values of X. That is, by constructing a linear function of X [$f(X)=\alpha+\beta X$], we'd like to get the best possible prediction of Y.

$f(X) = \alpha + \beta X$

One way of restating this is that we would like to choose α and β such that f(X) is as close to Y as possible. Or, stated yet another way, we would like to solve the following:

find error across all α, β — expectation

$$\underset{\alpha,\beta}{\operatorname{argmin}}\{E[(Y-\alpha-\beta X)^2]\} = \underset{\alpha,\beta}{\operatorname{argmin}}\{MSE\}$$

real — prediction

Note: this notation means that we would like to find the values of α and β such that the Expectation is minimized. We are minimizing the mean squared error (MSE)!

$$\underset{\alpha,\beta}{\operatorname{argmin}}\{MSE\} = \underset{\alpha,\beta}{\operatorname{argmin}}\{E[(Y-\alpha-\beta X)^2]\}$$

$$= \underset{\alpha,\beta}{\operatorname{argmin}}\{E[Y^2+\alpha^2+\beta^2X^2-2\alpha Y-2\beta XY+2\alpha\beta X]\}$$

linearity of expectation property

$$= \underset{\alpha,\beta}{\operatorname{argmin}}\{E[Y^2]+\alpha^2+\beta^2E[X^2]-2\alpha\mu_Y-2\beta E[XY]+2\alpha\beta\mu_X\}$$

Recalling that $Var[Y] = E[Y^2]-E^2[Y]$, we see that $E[Y^2] = \sigma_Y^2+\mu_Y^2$ and $E[X^2]=\sigma_X^2+\mu_X^2$.
Also, remember that the correlation, ρ, is defined as $\rho=Cov[X,Y]/\sigma_X\sigma_Y$, so

$\rho\sigma_X\sigma_Y = Cov[X,Y] = E[XY]-E[X]E[Y]$, and therefore $E[XY] = \rho\sigma_X\sigma_Y+\mu_X\mu_Y$.

So…

$Var[Y] = E[Y^2] - E^2[Y]$ $\quad$ $\rho = \frac{cov[X,Y]}{\sigma_X\sigma_Y}$

$E[Y^2] = \sigma_y^2 + \mu_y^2$ $\quad$ $E[X^2] = \sigma_x^2 + \mu_x^2$ $\quad$ $cov[X,Y] = E[XY] - E[X]E[Y]$

$E[XY] = \rho\sigma_X\sigma_Y + \mu_X\mu_Y$

$$\underset{\alpha,\beta}{\operatorname{argmin}}\{MSE\}$$

$$= \underset{\alpha,\beta}{\operatorname{argmin}}\{\sigma_Y^2+\mu_Y^2+\alpha^2+\beta^2\sigma_X^2+\beta^2\mu_X^2-2\alpha\mu_Y-2\beta\rho\sigma_X\sigma_Y-2\beta\mu_X\mu_Y+2\alpha\beta\mu_X\}$$

To solve this, we first take the derivative of the equation with respect to α and set equal to 0:

$$\frac{\delta MSE}{\delta \alpha} = 0 \rightarrow 2\alpha - 2\mu_Y + 2\beta\mu_X = 0$$

$$\alpha = \mu_Y - \beta\mu_X$$

$$\frac{\partial MSE}{\partial \alpha} = 0 \quad \xrightarrow{yields} \quad 2\alpha - 2\mu_Y + 2\beta\mu_X = 0$$

Or

$$\alpha = \mu_Y - \beta\mu_X$$

Substituting this back in to the original equation yields:

$$\underset{\alpha,\beta}{\operatorname{argmin}}\{MSE\}$$

$$= \underset{\alpha,\beta}{\operatorname{argmin}}\{{\sigma_Y}^2 + {\mu_Y}^2 + (\mu_Y - \beta\mu_X)^2 + \beta^2{\sigma_X}^2 + \beta^2{\mu_X}^2 - 2(\mu_Y - \beta\mu_X)\mu_Y - 2\beta\rho\sigma_X\sigma_Y - 2\beta\mu_X\mu_Y + 2(\mu_Y - \beta\mu_X)\beta\mu_X\}$$

Cancelling some terms yields:

$$\underset{\alpha,\beta}{\operatorname{argmin}}\{MSE\}$$

$$= \underset{\alpha,\beta}{\operatorname{argmin}}\{{\sigma_Y}^2 + 2{\mu_Y}^2 - 2\beta\mu_X\mu_Y + \beta^2{\mu_X}^2 + \beta^2{\sigma_X}^2 + \beta^2{\mu_X}^2 - 2{\mu_Y}^2 + 2\beta\mu_X\mu_Y - 2\beta\rho\sigma_X\sigma_Y - 2\beta\mu_X\mu_Y + 2\beta\mu_X\mu_Y - 2\beta^2{\mu_X}^2\}$$

$$= \underset{\alpha,\beta}{\operatorname{argmin}}\{{\sigma_Y}^2 + \beta^2{\sigma_X}^2 - 2\beta\rho\sigma_X\sigma_Y\}$$

$\frac{\delta MSE}{\delta \beta} = 0 \rightarrow 2\beta\sigma_x^2 - 2\rho\sigma_x\sigma_y$

To solve for β, we take the derivative of the equation with respect to β and set equal to 0:

$$\frac{\partial MSE}{\partial \beta} = 0 \quad \xrightarrow{yields} \quad 2\beta{\sigma_X}^2 - 2\rho\sigma_X\sigma_Y = 0$$

Or

$$\beta = \frac{\rho\sigma_Y}{\sigma_X}$$

(ρ: correlation; σ_Y, σ_X: stddev)

$2\beta\sigma_x^2 = 2\rho\sigma_x\sigma_y$

$\beta = \frac{2\rho\sigma_x\sigma_y}{2\sigma_x^2}$

$= \frac{\rho\sigma_y}{\sigma_x}$

OK, so what does this all mean? The slope of the linear fit that gives the best prediction of Y from X is proportional to the correlation between X and Y. So, positive correlation yields positive slopes, and negative correlation means negative slope.

proportion to ρ (linear relationship)

Further, recall that to get α and β, we were minimizing the MSE, which boiled down to:

+ρ → + slope

−ρ → − slope

$$MSE = {\sigma_Y}^2 + \beta^2{\sigma_X}^2 - 2\beta\rho\sigma_X\sigma_Y$$

Plugging our solution for β back in yields:

$$MSE = {\sigma_Y}^2 + \left(\frac{\rho\sigma_Y}{\sigma_X}\right)^2 {\sigma_X}^2 - 2\left(\frac{\rho\sigma_Y}{\sigma_X}\right)\rho\sigma_X\sigma_Y$$

$$MSE = {\sigma_Y}^2 + \rho^2{\sigma_Y}^2 - 2\rho^2{\sigma_Y}^2$$

$$MSE = {\sigma_Y}^2 - \rho^2{\sigma_Y}^2$$

$$\boxed{MSE = {\sigma_Y}^2(1-\rho^2)}$$

If $\rho = 0$, $MSE = {\sigma_Y}^2$
As $\rho \uparrow$, $MSE \downarrow$
If $\rho = 1, -1$, $MSE = 0$ (perfect line)

The MSE is zero whenever the correlation is +/- 1: perfect prediction! Meanwhile, the MSE is maximized when the correlation is zero: it's as bad as if we hadn't used X at all. This is why we say that correlation measures the linear relationship between two random variables.

Ordinary linear regression

So, how do we get these terms when we're actually dealing with data? Assume we have a bunch of measurements y_i (for $i \in [1,N]$) in response to some covariate x_i. For example, y_i could be the number of spikes recorded in response to a grating with contrast xi, and you record a bunch of responses to gratings of different contrasts.

To set up this problem again, we want to find the parameters α and β such that we minimize the mean squared error between our measurement y_i and our predicted measurement $\hat{y}_i$, where

estimate

$$\hat{y}_i = \alpha + \beta x_i$$

x_i = orientation on trial i
y_i = fr on trial i
N = # of trials

So, you want to minimize:

MSE when dealing w/ actual data

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \alpha - \beta x_i)^2$$

Solve this on your own!!!

over α and β.

Solving this in the standard way (taking derivatives w.r.t. α and β and setting them equal to 0 yields solutions for α and β exactly as we derived, except using the sample means and variances as opposed to the theoretical means and variances. Thus:

$$\alpha = \bar{y} - \beta\bar{x} \qquad \text{and} \qquad \beta = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

Sample mean

(To see this last equation, realize that $Cov[X,Y]=\rho\sigma_X\sigma_Y$, so $\beta=\rho\sigma_Y/\sigma_X$ becomes $\beta=Cov[X,Y]/Var[X]$.)

(A useful exercise is to see if you can prove these two equations are correct.)

Multiple linear regression

Often we are in a situation where we have several covariates to regress against. For example, we could regress our measured firing rate of our motor cortical neuron against both the speed of the movement and the distance of the movement. In this case, we call the problem a multiple linear regression problem:

$$y^{(i)} = \alpha + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_p x_p^{(i)} + \epsilon^{(i)}$$

If we gather all n measurements into a single vector $\underline{y} \in \Re^{n\times 1}$, we can rewrite this as

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

Where $\mathbf{X} \in \Re^{n\times(p+1)}$ is:

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots X_{1p} \\ 1 & X_{21} & X_{22} & \cdots X_{2p} \\ & & \cdots & \\ & & \cdots & \\ 1 & X_{n1} & X_{n2} & \cdots X_{np} \end{pmatrix}$$

$\underline{\beta} \in \Re^{(p+1)\times 1}$ is:

$$\beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \beta_p \end{pmatrix}$$

and $\underline{\varepsilon}$ is the noise vector that accounts for deviations from the perfect line.

The solution to this problem is:

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

Note how this reduces to our solution for α and β in the one-dimensional case:

$$X = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \text{so} \quad X'X = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

Recall that the inverse of a 2x2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $\frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

That makes

$$(X'X)^{-1} = \frac{1}{n\sum X_i^2 - (\sum X_i)^2}\begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$$

$$(X'X)^{-1} = \frac{n}{n\sum X_i^2 - (\sum X_i)^2}\begin{bmatrix} \sum X_i^2/n & -\sum X_i/n \\ -\sum X_i/n & 1 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2}\begin{bmatrix} \sum X_i^2/n & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{\sum X_i^2 - n\left(\overline{X}\right)^2}\begin{bmatrix} \sum X_i^2/n & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}$$

Note that the sample variance s^2 is:

$$s_X^2 = \frac{1}{n-1}\sum\left(X_i - \overline{X}\right)^2$$

so

$$(n-1)s_X^2 = \sum\left(X_i^2 + \overline{X}^2 - 2X_i\overline{X}\right) = \sum\left(X_i^2\right) + n\overline{X}^2 - 2\overline{X}\sum X_i = \sum\left(X_i^2\right) + n\overline{X}^2 - 2n\overline{X}^2$$

and

$$(n-1)s_X^2 = \sum\left(X_i^2\right) - n\overline{X}^2$$

Therefore:

$$(X'X)^{-1} = \frac{1/(n-1)}{s_X^2}\begin{bmatrix} \sum X_i^2/n & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} n\overline{Y} \\ \sum X_iY_i \end{bmatrix} = \begin{bmatrix} n\overline{Y} \\ (n-1)C_{XY} + n\overline{XY} \end{bmatrix}$$

This latter equation comes about because the sample Covariance C_{XY} is

$$C_{XY} = \frac{1}{n-1}\sum\left(X_iY_i - \overline{XY}\right) \xrightarrow{yields} (n-1)C_{XY} = \sum(X_iY_i) - n\overline{XY}$$

Therefore:

$$(X'X)^{-1}X'Y = \frac{1}{(n-1)s_X^2}\begin{bmatrix} \sum X_i^2/n & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}\begin{bmatrix} n\overline{Y} \\ (n-1)C_{XY} + n\overline{XY} \end{bmatrix}$$

$$\frac{1}{(n-1)s_X^2}\begin{bmatrix} \overline{Y}\sum X_i^2 - \overline{X}\left((n-1)C_{XY} + n\overline{XY}\right) \\ -n\overline{XY} + (n-1)C_{XY} + n\overline{XY} \end{bmatrix} = \begin{bmatrix} \dfrac{\overline{Y}\left(\sum X_i^2 + n\overline{X}^2\right)}{(n-1)s_X^2} - \dfrac{\overline{X}C_{XY}}{s_X^2} \\ \dfrac{C_{XY}}{s_X^2} \end{bmatrix}$$

And so finally we have

$$(X'X)^{-1}X'Y = \begin{bmatrix} \overline{Y} - \overline{X}C_{XY}/s_X^2 \\ C_{XY}/s_X^2 \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}!$$

R^2: The coefficient of determination (goodness of fit)

Often, R^2 (called the coefficient of determination) is used as the measurement of the "goodness of fit" of the regression line. The R^2 is defined as:

$$R^2 = \frac{SSR}{SST} \qquad SST = SSE + SSR$$

Here, SSR is equal to

Sum Squared Residuals → $$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ (mean) sum squared deviations of estimated data from mean

SST is known as the Total Sum of Squares, and is equal to

Total Sum of squares → $$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ (mean) sum squared deviations of sample data from mean

An interesting (and occasionally useful) fact is that the total sum of squares is equal to the SSR plus another term known as the sum squared errors or SSE:

Sum squared errors → $$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ sum squared deviations of sample data from estimated data

That is, SST = SSE + SSR. To see this, we have the following proof: Sum of squares

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

$(a-b)^2 = (a-c+c-b)^2$
$= (a-c)^2 + (c-b)^2 + 2(a-c)(c-b)$

$$SST = SSE + SSR + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

So, it remains to prove that the last term is zero.
Recall that

$$\hat{y}_i = a + bx_i \quad \text{where} \quad a = \bar{y} - b\bar{x}$$

Substitution yields: $2\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$

$$2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 2\sum_{i=1}^{n} (\bar{y} - b\bar{x} + bx_i - \bar{y})(y_i - \hat{y}_i) = 2\sum_{i=1}^{n} b(x_i - \bar{x})(y_i - \hat{y}_i)$$

Substitution again yields

$$2\sum_{i=1}^{n} b(x_i-\bar{x})(y_i-\hat{y}_i) = 2\sum_{i=1}^{n} b(x_i-\bar{x})\big(y_i-(\bar{y}-b\bar{x}+bx_i)\big)$$

$$= 2\sum_{i=1}^{n} b(x_i-\bar{x})\big(y_i-\bar{y}-b(x_i-\bar{x})\big) = 2\sum_{i=1}^{n} b(x_i-\bar{x})(y_i-\bar{y}) - 2\sum_{i=1}^{n} b^2(x_i-\bar{x})^2$$

$$2b\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) - b\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

Now, again recall that

$$b = \frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}$$

So

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = b\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Substitution yields:

$$2b\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) - b\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = 2b\left(b\sum_{i=1}^{n}(x_i-\bar{x})^2 - b\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = 0$$

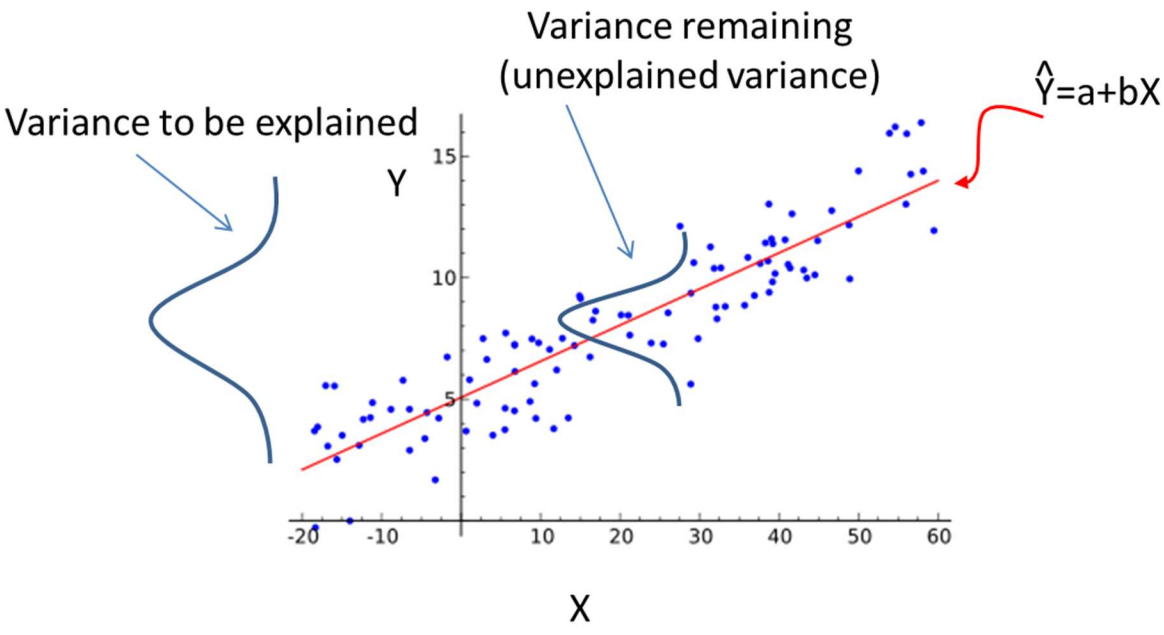
And we indeed have

SST=SSE+SSR

This is another example of a variance decomposition. SST is proportional to the variance in the data (it is the sample variance multiplied by n-1). SSE is proportional to the noise variance (it is multiplied by n-2), and SSR can be interpreted as proportional to signal variance. For this reason, R^2 is sometimes thought of as the fraction of explained variance:

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

When R^2 is 1, there is no error in the fit – all data lie perfectly on a line, and all of the original variation in the data have been "explained". When R^2 is zero, you have as much variation in the residuals as you did in the original data! Thus, you've explained nothing. Could R^2 be negative? Yes. If you, for example, fit the regression on a training set of data, and then measure the residual variation on a testing set, it's possible for your fit to be so bad that it does worse than just guessing the mean. In other words, the variation in the residuals is actually *greater than* the variation in the data.


Variance remaining
(unexplained variance)
Ŷ=a+bX
Variance to be explained
Y
15
10
5
-20
-10
10
20
30
40
50
60
X