Lecture 1

Introduction to the course

In this second semester of Statistics with R course, we are going to focus mainly on Regression: relating two or more (usually numerical) variables in order to predict one of the variable values. We will look at several types of regression including simple linear regression, and also multiple regression, non-linear regression, logistic regression and the relationship to machine learning. We'll also spend some time at the end of the semester discussing time series. To get started, we'll review some of the calculus we discussed last semester that is relevant to our discussion of simple linear regression. In the coming weeks we'll also briefly discuss some linear algebra basics so that we can discuss regression formulas in that context.

Let's get started!

Review of joint probability distributions and covariance

Last semester, we discussed joint probability distributions and they are newly relevant now as we move into a discussion of regression. So, let's review what we discussed before proceeding further.

Just as with the single random variable case, we must deal with both the discrete case and the continuous case with probability density functions.

In the discrete case

$$P(X = x, Y = y) = p(x, y)$$

And which follows the usual rules that all values of $0 \leq p(x, y) \leq 1$, and

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) = 1$$

Discrete probabilities can be expressed with piecewise function notation, but is often expressed in the form of a table (this becomes more difficult when there are more than two variables).

$p(x, y)$		y = 0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

In the continuous case, the pdf is a function of two (or more) variables. In the two-variable case

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dydx = 1$$

And $0 \leq f(x, y) \leq 1$ for all allowable values of x and y.
To find the probability that (X, Y) is in some region, we integrate the function within those limits.

Note: If the function has more than two variables, it will need one integral for each variable, but we will stick with the 2D case here.

To break the probabilities down into their single variable cases, these are called marginal probabilities. In the discrete case

$p_X(x) = \sum_y p(x, y)$ for each value of x
$p_Y(y) = \sum_x p(x, y)$ for each value of y

In the continuous case

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx$$

Notice that the ideas are similar, we just replace the summations with integrals.

We'll do a complete integration example later, so let's look at the marginal distributions for our discrete case.

Our $p_X(x, y)$ becomes

$\boldsymbol{x}$	**100**	**250**
$\boldsymbol{p_X(x)}$	0.5	0.5

Our $p_Y(x, y)$ becomes

$\boldsymbol{y}$	**0**	**100**	**200**
$\boldsymbol{p_Y(y)}$	0.25	0.25	0.5

If the probabilities are independent, then the product of the marginal probabilities is the same as the original probabilities.

$$p(x, y) = p_X(x)p_Y(y)$$
$$f(x, y) = f_X(x)f_Y(x)$$

Our discrete case is not independent. $P(X = 100) = 0.5, P(Y = 0) = 0.25$, but $P(X = 100, Y = 0) = 0.20 \neq (0.5)(0.25)$.

We can talk about conditional probabilities in the joint case.

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

Thus, the conditional probabilities are the two-variable distribution divided by the marginal distribution. We can generalize this to the $x|y$ case, and the discrete case.

The expected values are found similarly to the one variable case.

$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)dydx$$

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)dydx$$

In the discrete case

$$E(X) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i p(x_i, y_j)$$

$$E(Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} y_j p(x_i, y_j)$$

The variances are calculated similarly as before, by multiplying our pdf by $(x-\mu_X)^2$ or $(y-\mu_Y)^2$ respectively.

Covariance

One new idea we have with two variables is the covariance, which measures how two the variables are related to each other. This is calculated as

$$Cov(X,Y) = E[(x-\mu_X)(y-\mu_Y)]$$

As with variance, we have an alternative formulation that produces an equivalent result.

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y$$

Covariance leads us to an idea that will be important when we tackle regression next semester: correlation coefficient.

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

Correlation is a value that can only take on values in the interval [-1,1]. Values closer to 1 or -1 show a stronger relationship, while values closer to 0 show a weaker linear relationship.

Linear transformations of X and Y do not change the correlation value which is useful since it is scale invariant.

Let's look at a complete continuous example.

Consider the joint pdf $f(x,y) = Kxy, 0 \le x \le 1, 0 \le y \le 1$.

1. Find the value of K that makes this a valid probability distribution.
$$\int_0^1 \int_0^1 Kxydydx = \int_0^1 K\left(\frac{1}{2}xy^2\right]_0^1 = \int_0^1 \frac{K}{2}xdx = \frac{K}{4}x^2\Big|_0^1 = \frac{1}{4}K = 1$$
So, we set $K = 4$.

2. Find the marginal pdf for x and y respectively.
$$f_X(x) = \int_0^1 4xydy = 2xy^2|_0^1 = 2x$$
$$f_Y(y) = \int_0^1 4xydx = 2x^2y|_0^1 = 2y$$

3. Are the variables independent?
Yes, since $f_X(x)f_Y(y) = (2x)(2y) = 4xy = f(x,y)$

4. What is the expected value of x and y respectively?
$$E(X) = \int_0^1 \int_0^1 x(4xy)dydx = \int_0^1 \int_0^1 4x^2ydydx = \int_0^1 2x^2dx = \frac{2}{3}x^3\Big|_0^1 = \frac{2}{3}$$

The math for $E(Y)$ is exactly the same.

5. Let's calculate the variance.
$$E(X^2) = \int_0^1 \int_0^1 x^2(4xy)dydx = \int_0^1 \int_0^1 4x^3ydydx = \int_0^1 2x^3dx = \frac{1}{2}$$
$$V(X) = E(X^2) - [E(X)]^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$$

The math for $V(Y)$ is exactly the same.

6. Let's find the covariance.
$$E(XY) = \int_0^1 \int_0^1 xy(4xy)dydx = \int_0^1 \int_0^1 4x^2y^2dydx = \int_0^1 \frac{4}{3}x^2dx = \frac{4}{9}$$
$$Cov(X) = E(XY) - \mu_X\mu_y = \frac{4}{9} - \frac{2}{3}\left(\frac{2}{3}\right) = 0$$

7. The correlation is therefore also 0.

As with other statistical measures, the theoretical parameter value uses the Greek letter (here ρ), and the descriptive statistic that we measure from data uses the Latin equivalent (here r). The correlation describes the relationship between two variables. As we noted above, correlation values fall in the range of $[-1,1]$. A -1 correlation is a perfect negative linear relationship (the relationship has a negative slope and all the data falls on a straight line). A 1 correlation value is a perfect positive linear correlation (the relationship has a positive slope and all the data falls on a straight line). A zero correlation occurs then the slope of the regression line is zero, which can happen when there is no relationship, or in certain types of nonlinear relationships. The best way to determine which situation applies is to look at a scatterplot. We'll look at how those relate to correlation estimation in the next class. Today, we are going to look at how to calculate the correlation from data. For the Pearson correlation, it's similar to the formula we'd use for a discrete distribution.

The formula for correlation is often broken down into variances. S_{xy} is the covariance, and S_{xx} and S_{yy} is the variance for x and y separately. So, we can write the correlation as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

But we can write this out in a bit more detail so that we can compute it if need be:

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

For even a small dataset, this is awful to calculate by hand, but fortunately, R can do it quite easily even for large data sets. This is the standard formula for the Pearson correlation coefficient. This is the typical default method of computing correlation.

We make several assumptions when we compute this correlation value. Among them is that there is a linear relationship between the variables, and that the errors relative to that line are normally distributed with constant variance. We generally use the t-distribution to model those errors and make inferences on the coefficient values of the regression line (which we'll look at in greater detail in future lectures).

We can conduct hypothesis tests on ρ as we can with other inferences. Suppose that we want to test $H_0: \rho = \rho_0, H_a: \rho \neq 0$, we can construct a confidence interval to test this hypothesis using the following formula using the Fisher transformation.

$$V = \frac{1}{2}\ln\left(\frac{1 + R}{1 - R}\right)$$

has approximately a normal distribution with mean and variance

$$\mu_V = \frac{1}{2}\ln\left(\frac{1 + \rho}{1 - \rho}\right) \quad \sigma_V^2 = \frac{1}{n - 3}$$

The variance here is dependent on the sample size, but not ρ itself. After performing the Fisher transformation, we build a confidence interval in this transformation, in the traditional way.

$$\left(v - \frac{z_{\alpha/2}}{\sqrt{n-3}}, v + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right)$$

We can then convert this back to the original correlation measure:

$$\left(\frac{e^{2c_1} - 1}{e^{2c_1} + 1}, \frac{e^{2c_2} - 1}{e^{2c_2} + 1} \right)$$

Where $c_1 = v - \frac{z_{\alpha/2}}{\sqrt{n-3}}$, and $c_2 = v + \frac{z_{\alpha/2}}{\sqrt{n-3}}$, the endpoints of our interval in the Fisher transformation. These formulas can easily be obtained by replacing the endpoint values into the Fisher transformation formula and solving for ρ.

We can also conduct a traditional hypothesis test with test statistic:

$$Z = \frac{V - \frac{1}{2}\ln[(1 + \rho_0)/(1 - \rho_0)]}{1/\sqrt{n-3}}$$

In most cases, we are testing whether the correlation is 0 or not, which simplifies this calculation significantly since the second term in the numerator reduces to 0.

The Devore text has more information on the bivariate normal distribution at the heart of our assumptions about correlation in the context of regression, although correlation is a more general idea.

In the next lecture, we'll look at some distribution-free measures of correlation and how correlation is reflected in scatterplots.

References:

1. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
2. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
3. https://www.spss-tutorials.com/pearson-correlation-coefficient/