

Correlation and Regression

1

Correlations

- Bivariate correlation (ρ) is an indicator of the strength and direction of the relationship between two variables
- Correlation 'matrix' is frequently used to screen for important relationships
- Ranges from -1 (perfect inverse relationship) to $+1$ (perfect positive relationship)
- Types of Correlations
 - Parametric correlations – Pearson (requires normal distribution)
 - Nonparametric alternatives – Spearman's or Kendall's

2

Simple Linear Regression

- ANOVA was extension of T-Test to multiple group means
- Linear regression extends ANOVA to continuous predictor variables
 - Systolic blood pressure predicted by body mass index
 - Body mass index predicted by caloric intake
 - Caloric intake predicted by measure of stress
 - However, may relate to inactivity due to lack of visual acuity
- Important to specify a biologically plausible model
 - Systolic blood pressure predicted by eye color
 - Body mass index predicted by visual acuity
 - However, may relate to inactivity due to lack of visual acuity

3

Regression Models

- Assumptions are important in linear regression, but are not absolute
 - Predictor variables are 'fixed'; i.e., same meaning among individuals
 - Predictor variable measured 'without error'
 - For each value of the predictor variable, there is a normal distribution of outcomes (subpopulations) and the variance of these distributions are equal
 - The outcomes are independent of each other
- Regression model: $y = \alpha + \beta x + \epsilon$

4

Least squares fit

- Regression parameters ($\alpha + \beta$) are determined using method of least squares
- Minimizes the squared differences between each observation and the 'fitted' line in the multivariate 'plane'; i.e., minimizes the residuals
 - Can be dramatically affected by unusual values in the multivariate plane – multivariate 'outliers'

5

Interpretation of regression model

- Two parameters
 - α is the intercept (Y value) when the predictor is zero.
 - β is the 'slope' of the regression line and represents the change in Y for a unit change in X.
 - i.e., a slope of 0.58 would indicate that for a one unit change in X, there is a 0.58 unit change in Y
 - ϵ is the error term for each individual and is the residual for that individual
 - Residual is the difference between the fitted line (predicted value) and the observed value

6

Regression Analysis Results-I

- Fit of the model
 - Determine if there are any parameters that are significantly different from zero and, thus, explain some part of the variation
- Coefficient of determination (R^2)
 - Ratio of regression sums of squares (SSR) to the total sums of squares
 - Can be interpreted as the proportion of the total variation in the outcome that can be explained by the regression model
- Total Sums of Squares (SST)
 - SST is total SS in the outcome for the entire sample – analogous to the variance

7

Regression Analysis Results-II

- Error Sums of Squares (SSE)
 - Unexplained variation based on the sum of the squared deviations between the observed and predicted values
 - What is being minimized in least squares
- Regression Sums of Squares (SSR)
 - Represents the variability in the outcome accounted for by differences among the group means
 - This is the part of the Regression Table that is of most interest
 - It tells you if the factor that you are interested in explains a significant amount of the variance in the sample

8

Multiple Linear Regression

- ANOVA was extension of T-Test to multiple group means
- Linear regression extends ANOVA to continuous predictor variables
- Multiple linear regression extends simple linear regression to multiple predictor variables
 - Systolic blood pressure predicted by body mass index, sodium intake, and gender
 - Body mass index predicted by caloric intake and gender
 - Caloric intake predicted by measure of stress and gender

9

Reason for Multiple Regression Models

1. To assess effect of different factors on the subgroup means of an outcome
 - Since factors have differing effects on outcome and are interrelated themselves, necessary to look at a number of predictors simultaneously
 - We know that each predictor has an independent effect on systolic blood pressure, but how much of the information carried by each one is also carried by the others
2. To predict for an individual the value of the outcome based on the values of the predictor variables
 - e.g., based on body mass index, sodium intake, and race, what would the predicted value of systolic blood pressure for a person
 - Why? To identify those people who may have significantly higher systolic blood pressure than expected from similar people

10

Multiple Regression Models - Assumptions

- Predictor variables are 'fixed'; i.e., same meaning among individuals
- Predictor variable measured 'without error'
- For each value of the predictor variable, there is a normal distribution of outcomes (subpopulations) with equal variance
- The outcomes are independent
- Regression model: $y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_k x_{kj} + \varepsilon_j$

11

Interpretation of regression model

- α is the intercept (Y value) when the predictors are zero
- β_k is one of the 'slopes' of regression line and represents change in Y for a unit change in X_k with other predictors held constant. All 'slopes' interact to produce the overall slope
 - i.e., β_k is the average slope across all subgroups created by the X_k levels
- ε is the error term for each individual and is the residual for that individual
 - Residual is the difference between predicted and observed values

12

Multivariate and Partial Correlations

- Multivariate correlation is an indicator of strength of the relationship between the outcome and predictor variables
 - Can be tested with an F-test, but F-statistic will be the same as that from the test of the regression sums of squares
- Partial correlation is an indicator of strength and direction of relationship between the outcome and one predictor with the effect of the other variables removed
 - Can be calculated from simple correlations or produced by software

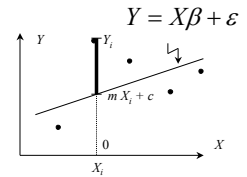
13

Linear Regression Models

Regression

Suppose modelling relationship between markers and putative genes

ENV	3	5	4	5	6	7
MARKER	0	1	2	3	4	5



Want straight line " $Y = mX + c$ " that best approximates the data. "Best" in this case is the line minimising the sum of squares of vertical deviations of points from the line:
 $SSQ = \sum (Y_i - [mX_i + c])^2$

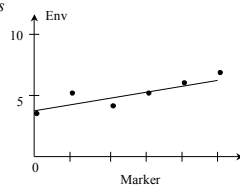
14

Linear (Regression) Models

Setting partial derivatives of SSQ w.r.t. m and c to zero \rightarrow Normal Equations

$$\sum_{i=1}^n Y_i = m \sum_{i=1}^n X_i + nc$$

$$\sum_{i=1}^n X_i Y_i = m \sum_{i=1}^n X_i^2 + c \sum_{i=1}^n X_i$$



X.X	X	X.Y	Y	Y.Y
0	0	0	3	9
1	1	5	5	25
4	2	8	4	16
9	3	15	5	25
16	4	24	6	36
25	5	35	7	49
55	15	87	30	160

15

Example contd.

• **Model Assumptions** - as for ANOVA (also a Linear Model)
 Calculations give:

X.X	X	X.Y	Y	Y.Y
0	0	0	3	9
1	1	5	5	25
4	2	8	4	16
9	3	15	5	25
16	4	24	6	36
25	5	35	7	49

55 15 87 30 160
 So Normal equations are:
 $30 = 15m + 6c \Rightarrow 150 = 75m + 30c$
 $87 = 55m + 15c \Rightarrow 174 = 110m + 30c$
 $\Rightarrow 24 = 35m \Rightarrow 30 = 15(24/35) + 6c \Rightarrow c = 23/7$

16

Example contd.

- Thus the regression line of Y on X is
 $Y = (24/35)X + (23/7)$
 and to plot the line we need two points, so
- $X = 0 \Rightarrow Y = 23/7$ and $X = 5 \Rightarrow Y = (24/35)5 + 23/7 = 47/7$.

It is easy to see that (X, Y) satisfies the normal equations, so that the regression line of Y on X passes through the "Centre of Gravity" of the data. By expanding terms, we also get

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \text{ with } \hat{Y}_i = mX_i + c$$

Total Sum of Squares	Error Sum of Squares	Regression Sum of Squares
SST	= SSE	+ SSR

X is the **independent**, Y the **dependent variable** and above can be represented in ANOVA table

17

Least Squares Estimation in general

- Suppose want to find relationship between group of markers and phenotype of a trait
 $Y = X\beta + \epsilon$ Y is an $N \times 1$ vector of observed trait values for N individuals in a mapping population, X is an $N \times k$ matrix of re-coded marker data, β is a $k \times 1$ vector of unknown parameters and ϵ is an $N \times 1$ vector of residual errors, expectation = 0.
- The Error SSQ is then $\epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta)$
 $= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta$
 all terms in matrix/vector form
- The Least Squares estimates of the unknown parameters β is $\hat{\beta}$ which minimises $\epsilon^T \epsilon$. Differentiating this SSQ w.r.t. β 's and setting =0 gives the normal equations

18

LSE in general contd.

So
$$\frac{\partial e^T \varepsilon}{\partial \beta} = -2X^T Y + 2X^T X \beta$$

$$X^T X \hat{\beta} = X^T Y$$

so L.S.E.
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- Hypothesis tests for parameters - use F-statistic - tests $H_0: \underline{\beta} = 0$ on k and $N-k-1$ dof (assuming Total SSQ "corrected for the mean")
- Hypothesis tests for sub-sets of X's, use F-statistic = ratio between residual SSQ for the reduced model and the full model.

$SSE_{full} = Y^T Y - \hat{\beta}^T X^T Y$ has $N-k$ dof, so to test $H_0: \beta_i = 0$ use
 $SSE_{reduced} = Y^T Y - \hat{\beta}^{RT} X^{RT} Y$ with dimensions $(k-1) \times 1$ and $N \times (k-1)$ for β
 and X reduced, so $SSE_{reduced}$ has $N-k+1$ dof
 $F_{N-k+1, N-k} = \frac{SSE_{reduced} / (N-k+1)}{SSE_{full} / (N-k)}$ tests that subset of X's adequate

19

Prediction, Tolerance, Residuals

- Prediction:** Given value (s) of $X(s)$, line (plane) substitute to predict Y . Both **point** and **interval** estimates - C.I. for "mean response" = line, prediction limits for new individual value (wider since $Y_{new} = \mu + \varepsilon$)
 General form same:

$$(\hat{\beta}_0 + \hat{\beta}_1 X) \pm t_{n-2, \alpha/2} \times SE(Estimate)$$

$$\equiv \bar{Y} - \hat{\beta}_1 (X_o - \bar{X}) \pm t_{n-2, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum (X_o - \bar{X})^2}}$$

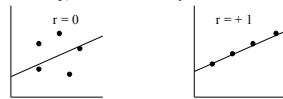
- Tolerance** - bands s.t. 95% of C.I. contain mean
- Residuals** ($Y_i - \hat{Y}_i$) = **Observed - Fitted (or Expected) values**
 Measures of goodness of fit, influence of outlying values of Y; used to investigate assumptions underlying regression, e.g. through plots.

20

Correlation, Determination, Collinearity

- Coefficient of Determination** r^2 (or R^2) ($0 \leq R^2 \leq 1$) proportion of **total variation** that is associated with the regression. (Goodness of Fit)
 $r^2 = SSR / SST = 1 - SSE / SST$
- Coefficient of correlation**, r or R ($0 \leq R \leq 1$) is degree of association of X and Y (strength of linear relationship). Mathematically

$$r = \frac{Cov(X, Y)}{\sqrt{Var X} \sqrt{Var Y}}$$



- Suppose r_{xy} close to 1, X is a function of Z and Y is a function of Z also. It does not follow that r_{xy} makes sense, as relation with Z may be hidden. Recognising hidden dependencies (**collinearity**) between distributions is difficult. A high r between heart disease deaths now and No. of cigarettes consumed twenty years earlier does *not* establish a cause-and-effect relationship.

21