Correlation and Regression

Eleisa Heron

07/10/09
Introduction

• Correlation and regression – for quantitative variables
  - Correlation: assessing the association between quantitative variables
  - Simple linear regression: description and prediction of one quantitative variable from another

• Only considering linear relationships

• When considering correlation or carrying out a regression analysis between two variables always plot the data on a scatter plot first
Scatter Plots

Correlation and Regression
Scatter Plots

Non-linear

No relationship
The usual measure of correlation between two continuous variables is the Pearson correlation coefficient also known as the ‘product moment’ correlation coefficient.

Suppose we have $n$ pairs of measurements $(x_i, y_i)$
- The correlation coefficient is calculated as:

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

- $r$ takes values between -1 and +1
  - An $r$ value of -1 or +1 indicates perfect correlation
  - Values close to -1 indicate high negative correlation
  - Values close to +1 indicate high positive correlation
Pearson Correlation Coefficient

- The correlation coefficient is a measure of the scatter of the data points around an underlying LINEAR trend
- Interchanging variables will not change the value of the correlation coefficient
- The value of r does not depend on the unit of measurement for either variable
- Always plot the data on a scatter plot before calculating the correlation coefficient
- There is an associated hypothesis test for the correlation coefficient (null hypothesis that $r = 0$), for this to be valid:
  - the two variables must be observed on a random sample of individuals
  - the data for at least one of the variables must be Normally distributed in the population
  - in practice usually prefer both variables to be approximately Normal (needed to calculate a valid confidence interval for r)
  - all the observations should be independent
Correlation – Linear Relationship

Corr(X1, Y1) = 0.779
Corr(X2, Y2) = 0.779
Correlation – Linear Relationship

Corr(X3, Y3) = 0.779

Corr(X4, Y4) = 0.779
Some Problems/Misuses of Correlation

Large number of correlations:
• Measuring many variables and calculating correlations for all possible pair combinations of variables – multiple testing issues

Recording variables over time:
• Suppose you have two variables that you recorded repeatedly over time, calculation of correlation can be misleading – need to remove the time trends before looking at correlation

Restricted sampling:
• Restricting the sampling of individuals can affect the correlation – correlation assumes a random sample has been collected

Subgroups in data:
• Calculating correlation when there are subgroups in the sample – may miss correlations, or identify false correlations, need to identify the subgroups

Assessing agreement:
• Comparing two methods of measuring the same quantity. The correlation coefficient measures association between the quantities, does not measure how closely they agree
Correlation Does Not Imply Causation

• Correlation does not mean causation

• If we observe high correlation between two variables, this does not necessarily imply that because one variable has a high value it causes the other to have a high value

• There may be a third variable causing a simultaneous change in both variables

• Example:
  – Suppose we measured children’s shoe size and reading skills
  – There would be a high correlation between these two variables, as the shoe size increases so too do the child’s reading abilities
  – But one does not cause the other, the underlying variable is age
  – As age increases so too does shoes size and reading ability
Non-Parametric Correlation

- Rank correlation may be used whatever type of pattern is seen in the scatter diagram, doesn’t specifically assess linear association but more general association

- Spearman’s rank correlation rho
  - Non-parametric measure of correlation – doesn’t make any assumptions about the particular nature of the relationship between the variables, doesn’t assume a linear relationship
  - rho is a special case of Pearson’s r in which the two sets of data are converted to rankings
  - can test null hypothesis that the correlation is zero and calculate confidence intervals

- Kendall’s tau rank correlation
  - Non-parametric statistic used to measure the degree of correspondence between two rankings and assess the significance of this correspondence
Simple Linear Regression

- Data on two quantitative variables

- Aim is to *describe* the relationship between the two variables and/or to *predict* the value of one variable when we only know the other variable

- Interested in a linear relationship between the two variables $X$ and $Y$
  - $Y =$ predicted variable = dependent variable = response variable = outcome variable
  - $X =$ predictor variable = independent variable = carrier variable = input variable

- Simple linear regression - when there is only one predictor variable, which we will consider here

- Multiple or multivariate regression - when there is more than one predictor variable
Simple Linear Regression

- The aim is to fit a straight line to the data that predicts the mean value of the dependent variable (Y) for a given value of the independent variable (X).

- Intuitively this will be a line that minimizes the distance between the data and the fitted line.

- Standard method is least squares regression.

- Notation: \( n \) pairs of data points, \((x_i, y_i)\).
Two Quantitative Variables
Two Quantitative Variables, Regression Line
Two Quantitative Variables, Regression Line

Predicted or dependent variable

Predictor or independent variable
Linear Regression Assumptions

- The values of the dependent variable Y should be Normally distributed for each value of the independent variable X (needed for hypothesis testing and confidence intervals)

- The variability of Y (variance or standard deviation) should be the same for each value of X (homoscedasticity)

- The relationship between the two variables should be linear

- The observations should be independent

- Do not have to have both variables random, values of X do not have to be random and they don’t have to be Normally distributed
Linear Regression Assumptions

- The straight line or linear relationship is described by the equation for a straight line:

\[ y = a + bx \]

- **Dependent variable**: Value of \( y \) when \( x = 0 \), \( a = 5 \), \( y = 5 + b \times 0 \)

- **Intercept**: \( b = 2 \) here

- **Slope of the line**: \( b = 2 \) here

\[ y = 5 + 2x \]
Least Squares Regression

• No line could pass through all the data points

• We want the best “average” equation (regression equation) that would represent a line through the middle of the data, this is the regression line:

\[ \hat{y}_i = \hat{a} + \hat{b}x_i \]

• The constants \( a \), the intercept and \( b \), the slope or regression coefficient are computed using the method of least squares

• Fitted value = value given by the line for any value of the variable \( X \) (remember what a fitted value was in ANOVA)

• Residual = difference between the observed value and the fitted value (again, remember what a residual was in ANOVA)

• Least squares aim: to minimize the sum of squares of the residuals
Fitted Values and Residuals

![Graph showing fitted values and residuals](image-url)
Least Squares Regression

• At any point $x_i$, the corresponding point on the line is given by $a + bx_i$

Regression equation:

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

Residuals (errors):

$$e_i = y_i - (\hat{a} + \hat{b}x_i)$$

Linear model:

$$y_i = \hat{a} + \hat{b}x_i + e_i, \quad e_i \sim \text{Normal}(0, \sigma^2)$$

• Note: if the errors/residuals are correlated or have unequal variances then least squares is not the best way to estimate the regression coefficient
Least Squares Regression

• Minimize the sum of squares ($S$) of the vertical distances of the observations from the fitted line (residuals)

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

• In order to find the intercept and regression coefficient that minimize $S$:
  
  – Partially differentiate with respect to the intercept and the regression coefficient, giving two equations
  – Set these two equations equal to zero
  – Solve the two simultaneous equations (sometimes referred to as normal equations) to obtain the estimates of the intercept and regression coefficients that minimize the sum of squares
Least Squares Regression

- The solution for these two equations results in the following two formulae for the estimates of the intercept and regression coefficients respectively:

\[
\hat{a} = \bar{y} - \hat{b} \bar{x}, \quad \hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})}
\]

- For the simulated data in the previous plots:

<table>
<thead>
<tr>
<th>(x_i)</th>
<th>(y_i)</th>
<th>(\hat{a})</th>
<th>(\hat{b})</th>
<th>(\hat{y}_i)</th>
<th>(\epsilon_i = y_i - \hat{y}_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.79</td>
<td>13.48</td>
<td>4.296</td>
<td>2.69</td>
<td>16.3</td>
<td>1.67</td>
</tr>
<tr>
<td>3.05</td>
<td>12.22</td>
<td>4.296</td>
<td>2.69</td>
<td>17.41</td>
<td>-0.29</td>
</tr>
<tr>
<td>3.21</td>
<td>13.66</td>
<td>4.296</td>
<td>2.69</td>
<td>18.09</td>
<td>0.73</td>
</tr>
<tr>
<td>3.61</td>
<td>13.15</td>
<td>4.296</td>
<td>2.69</td>
<td>19.81</td>
<td>-0.86</td>
</tr>
<tr>
<td>2.67</td>
<td>11.11</td>
<td>4.296</td>
<td>2.69</td>
<td>15.76</td>
<td>-0.37</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>
Residuals

- Checking assumptions:
  - Residuals should have a Normal distribution with zero mean
  - Plot X against residuals, looking for even scatter at all X values
  - Consider transformations of the data if these are not satisfied (eg., log, square root)
Regression Coefficient \( b \)

- Regression coefficient:
  - this is the slope of the regression line
  - indicates the strength of the relationship between the two variables
  - interpreted as the expected change in \( y \) for a one-unit change in \( x \)
  - can calculate a standard error for the regression coefficient
  - can calculate a confidence interval for the coefficient
  - can test the hypothesis that \( b = 0 \), i.e., that there is no relationship between the two variables

- Standard error for \( b \) is given by:

\[
se(b) = \frac{s_{res}}{\sqrt{\sum(x - \bar{x})^2}}
\]

where, the residual standard error is given by

\[
s_{res} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}
\]
Regression Coefficient $b$

- To test the hypothesis that $b = 0$, testing the hypothesis that there is no relationship between the X and Y variables, the test statistic is given by:

$$ t = \frac{b}{se(b)} $$

comparing this ratio with a t distribution with $n-2$ degrees of freedom

- Can also calculate a confidence interval for $b$:

$$ b \pm t_{0.975} se(b) $$
Intercept a

- Intercept:
  - the estimated intercept $a$ gives the value of $y$ that is expected when $x = 0$
  - often not very useful as in many situations it may not be realistic or relevant to consider $x = 0$
  - it is possible to get a confidence interval and to test the null hypothesis that the intercept is zero and most statistical packages will report these
Coefficient of Determination, R-Squared

- The coefficient of determination or R-squared is the amount of variability in the data set that is explained by the statistical model.

- Used as a measure of how good predictions from the model will be.

- In linear regression R-squared is the square of the correlation coefficient.

- The regression analysis can be displayed as ANOVA table, many statistical packages present the regression analysis in this format.

\[
R^2 = \frac{\text{sum of squares regression}}{\text{sum of squares total}}
\]

- Often expressed as a percentage
- High R-squared says that the majority of the variability in the data is explained by the model (good!)
• Adjusted R-squared

  – Sometimes an adjusted R-squared will be presented in the output as well as the R-squared

  – Adjusted R-squared is a modification to the R-squared to compensate for the number of explanatory or predictor variables in the model (more relevant when considering multiple regression)

  – The adjusted R-squared will only increase if the addition of the new predictor improves the model more than would be expected by chance
Residual Standard Deviation

- The residual standard deviation measures the variability not explained or accounted for by the regression line

$$s_{res} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$$

- It can be used as a measure of goodness-of-fit
Interpolation and Extrapolation

• \textit{Interpolation}
  – Making a prediction for $Y$ within the range of values of the predictor $X$ in the sample used in the analysis
  – Generally this is fine

• \textit{Extrapolation}
  – Making a prediction for $Y$ outside the range of values of the predictor $X$ in the sample used in the analysis
  – No way to check linearity outside the range of values sampled, not a good idea to predict outside this range
Correlation and Regression

- Correlation only indicates the strength of the relationship between two variables, it does not give a description of the relationship or allow for prediction.

- The t-test of the null hypothesis that the correlation is zero is exactly equivalent to that for the hypothesis of zero slope in the regression analysis.

- For correlation both variables must be random, for regression X does not have to be random.

- Correlation is often over used.

- One role for correlation is in generating hypotheses, remember correlation is based on one number, limit to what can be said.
# To carry out a linear regression:

```r
> fit <- lm(y~x)
> summary(fit)
```

**Call:**
```
lm(formula = y ~ x)
```

**Residuals:**
```
  Min 1Q Median 3Q Max
-1.77453 -0.58215 -0.07382 0.56110 1.76854
```

**Coefficients:**
```
            Estimate     Std. Error   t value     Pr(>|t|)
(Intercept)  4.2957         1.0530         4.079        0.000339 ***
x            2.6898        0.3467         7.758        1.88e-08 ***
```

---

**Signif. codes:**
```
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Residual standard error:** 0.9305 on 28 degrees of freedom

**Multiple R-squared:** 0.6825  |  **Adjusted R-squared:** 0.6712

**F-statistic:** 60.19 on 1 and 28 DF,  **p-value:** 1.884e-08

**R-squared** says that 68% of the variability in the data is explained by the model.
# ANOVA table for the regression

> anova(fit)

Analysis of Variance Table

Response: y[, 1]

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x[, 1]</td>
<td>1</td>
<td>52.115</td>
<td>52.115</td>
<td>60.194</td>
<td>1.884e-08 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>28</td>
<td>24.242</td>
<td>0.866</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$\frac{52.115}{24.242 + 52.115} = 0.6825$, which is the same as the R-squared, previous slide
# To test for Pearson correlation

> cor.test(y,x)

Pearson's product-moment correlation

data:  y and x
t = 7.7585, df = 28, p-value = 1.884e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.663299 0.914291
sample estimates:
cor
 0.826147

Square of correlation = coefficient of determination, R-squared:
0.826147 x 0.826147 = 0.6825
Relationships Between NES Scores, IQ Scores, and Treatment With Psychotropic Medications

We utilized simple linear regression to investigate whether total NES, FSIQ, PIQ, or VIQ predicted treatment status (treatment versus no treatment) with each of the classes of psychotropic medication described above. Total NES scores did not predict treatment status with any class of medication (all p>0.26). Initial analyses suggested that FSIQ predicts antipsychotic medication treatment status (adjusted $R^2=0.23$, $p<0.03$) and that VIQ predicts antipsychotic and antidepressant medication treatment status (adjusted $R^2=0.23$, $p<0.03$, and adjusted $R^2=0.17$, $p<0.05$, respectively). However, these findings failed to remain significant after Bonferroni correction for multiple comparisons.

NES: Neurological evaluation scale,  FSIQ: Full scale IQ,  PIQ: Performance IQ,  VIQ: Verbal IQ
“.....analysis to determine if differences in brain density were associated with behavioural abnormality within people with autistic-spectrum disorder. To do this, we related (using Pearson product-moment correlation coefficients) severity of clinical symptoms within people with the disorder as measured by the ADI–R to the density of brain regions, which differed significantly from controls.”

Fig. 1  Negative correlation between the amount of grey matter in the right limbic region, including the right anterior cingulate (centoid) extending into the posterior cingulate, parahippocampal gyrus and uncus, and abnormal reciprocal social interaction measured on the Autism Diagnostic Interview (ADI–S); $r=-0.767$, $n=7$, $P=0.04$. 
Further Reading

BOOKS:

*Interpretation and Uses of Medical Statistics*
by Leslie E Daly and Geoffrey J. Bourke, Blackwell science, 2000

*Practical Statistics for Medical Research*
by Douglas G. Altman, Chapman and Hall, 1991