1 Bivariate data

1.1 What is Bivariate data?

A dataset with two variables contains what is called bivariate data. For example, the heights and weights of people (i.e. for the purposes of determining the extent to which taller people weigh more). Common bivariate statistical analyses include

- Correlation
- Simple Linear Regression

1.2 Scatter Plot

A scatter plot of two variables shows the values of one variable on the Y axis and the values of the other variable on the X axis. Scatter plots are well suited for revealing the relationship between two variables.

- Scatterplots can be implemented in `R` using the command `plot()`

Exercise: Let us construct scatter-plots for the **immer** and **iris** data sets.

```
plot(immer$Y1,immer$Y2)

plot(iris[,1],iris[,3])
```

More complex scatterplots, with better visual aesthetics, can be constructed. We will look at this more later on in the module.

1.3 Correlation

- Recall that correlation describes the strength of a relationship between two numeric variables, and that the Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables.
- It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.
- The symbol for Pearson's correlation is "ρ" when it is measured in the population and `r` when it is measured in a sample.
- As we will be dealing almost exclusively with samples, we will use `r` to to represent Pearson's correlation unless otherwise noted.
- Pearson's rho can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an `r` of 0 indicates no linear relationship between variables, and an `r` of 1 indicates a perfect positive relationship between variables.

- Importantly it is assumed that the relationship in question is supposed to be linear. Some variables will in fact have a non-linear relationship (more on that latet)
- The relevant R command is `cor()`.

```
cor(immer$Y1,immer$Y2)

cor(iris[,1],iris[,3])
```

- The strength of the relation is represented in a numeric value known at the correlation coefficient. This coefficient can take a value between -1 and1. Additionally there are no units.
- Getting a correlation coefficient is generally only half the story; you will want to know if the relationship is significant. There is a more complex command called `cor.test()`. This command additionally provides a hypothesis test for the correlation estimate.

```
cor.test(immer$Y1,immer$Y2)

cor.test(iris[,1],iris[,3])
```

Ho : The correlation coefficient for the population of values is zero. (i.e. No linear relationship.)

Ha : The coefficient is not zero. (Linear relationship exists.)

- A confidence interval for the coefficient is provided for in the `R` output. If the interval includes 0 then we fail to reject the null hypothesis.
- Simple linear regression is used to describe the relationship between two variables x and y.
- For example, you may want to describe the relationship between age and blood pressure or the relationship between scores in a midterm exam and scores in the final exam, etc.
- x is the independent (i.e. predictor) variable.
- y is the dependent (i.e. response) variable.

That is to say x is said to cause or influence y.

Necessarily both x and y should be of equal length. One of the first steps in a regression analysis is to determine if any kind of relationship exists between x and y.

A scatterplot can created and can initially be used to get an idea about the nature of the relationship between the variables, e.g. if the relationship is linear, curvilinear, or no relationship exists.

To make a simple scatter-plot, we simply use the `plot()` command. The independent variable (the variable to go along the x-axis) is always specified first.

```
X=c(5.98, 8.80, 6.89, 8.49, 8.48, 7.47, 7.97,5.94, 7.32, 6.64, 6.94, 3.51)

Y=c(5.56, 7.80, 6.13, 8.15, 7.95, 7.87, 8.03, 5.67, 7.11, 6.65, 7.02, 3.88)

plot(X,Y)
cor(X,Y)
```

In this case here, we can see from the scatter-plot that there is a linear relationship between x and y. Simple linear regression is only useful when there is evidence of a linear relationship. In other cases, such as quadratic relationships, other types of regression may be more appropriate.

1.4 Linear Regression Model

A linear relationship can be defined by the simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

The intercept β_0 describes the point at which the line intersects the y axis. The slope β_1 describes the change in y for every unit increase in the predictor variable x.

From the data set, we determine the regression coefficients, i.e. estimates for slope and intercept. (N.B. There are variations on this notation in textbooks).

- b_0 : the intercept estimate.
- b_1 : the slope estimate.

Therefore the fitted model can be expressed as

$$\hat{y} = b_0 + b_1 x$$

Recall $\hat{y}$ denotes the predicted value for y, given some value x.

1.5 Fitting a Model with R

The `R` command `lm()` is used to fit linear models. Firstly the response variable y is specified, then the predictor variable x.

The tilde sign is used to denote the dependent relationship (i.e. y depends on x). The regression coefficients are then determined.

```
lm(Y~X) # y depends on X
```

The output will include the formula, and two coefficient terms

- The intercept estimate is recorded under (*Intercept*)
- The slope estimate is recorded under the name of the predictor variable (here : X).

```
Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
     0.7812       0.8581
```

A more detailed data output (i.e. more than just the coefficients) is generated in the form of a data object, using the `summary()` command.

We can give a name to the model (e.g. $FIT1$), and view all of the results of the calculation, including the regression coefficients, hypothesis test results and information on the residuals (i.e. the differences between the estimated y values and the observed y values).

In common with all data structures we can use the `names()` function and $ to access components.

```
FIT1 = lm(Y~X)
summary(FIT1)
names(FIT1)
names(summary(FIT1))
FIT1$coefficients
class(FIT1)
```

1.6 Confidence Interval for Regression Estimate

To compute the confidence intervals for both estimates, we use the confint() command, specifying the name of the fitted model.

```
C=c(0,2,4,6,8,10,12)
F=c(2.1,5.0,9.0,12.6,17.3,21.0,24.7)
Fit1=lm(F~C)
coef(Fit1)
# (Intercept)        Conc
     1.517857    1.930357

confint(Fit1)
#                2.5 %   97.5 %
# (Intercept) 0.75970 2.276014
# Conc        1.82522 2.035495
```

1.7 The Coefficient of Determination

The coefficient of determination R^2 is the proportion of variability in a data set that is accounted for by the linear model.

Equivalently R^2 provides a measure of how well future outcomes are likely to be predicted by the model.

(For simple linear regression, it canbe computed by squaring the correlation coefficient.)

```
summary(fit1)$r.squared
```

2 Assessing Model Assumptions

2.1 Residuals

The difference between the predicted value (based on the regression equation) and the actual, observed value. In simple linear regression models, the matter of whether or not residuals are normally distributed often arises.

Additionally the expected value of the residuals should be zero.

We have seen previously two methodologies for determining whether or not a data set is normally distributed;

- Shapiro-Wilk tests (or Anderson-Darling test)
- QQ plots

We will explore this more in a forthcoming example.

2.2 Influence Analysis

2.2.1 Outlier

In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

2.2.2 Leverage

An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. These leverage points can have an effect on the estimate of regression coefficients.

2.2.3 Influence

An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness.

2.3 Example

A new hotel is built 15 miles from the location of a prominent annual sporting event. A study of the number of enquiries received by a random sample of 9 established hotels in the area showed that the number of enquiries and the distance in miles between the hotel and event. Here the independent variable is distance (x) and the dependent variable is number of enquiries.

Lets looks at the residuals, and assess whether they are normally distributed.

```
#enquiries
y=c(35,61,74,92,113,159,188,217,328)

#distance from hotel
x=c(28,20,17,12,16,8,2,3,1)
#

#fit the linear model
fit2=lm(y~x)
resid(fit2)
res.fit=resid(fit2)

# test the residuals for normality.
# Normal if p.value is high.
shapiro.test(res.fit)

qqnorm(res.fit) #QQ plot
qqline(res.fit) #Add Trendline


#Do all your analyses agree?
```

Lets look at the scatterplot of x and y (`plot(x,y)`). Does the first covariate seem to be an outlier, given that a linear model is assumed? Lets omit the first element of both data sets and run the analysis again.

```
fit2=lm(y[-1]~x[-1])
resid(fit2)
res.fit2=resid(fit2)

shapiro.test(res.fit2)

#test the residuals for normality. Normal if p.value is high.
qqnorm(res.fit2);  qqline(res.fit2)

# compare the coefficients of both models.
coef(fit1)
coef(fit2)
```

Does the covariate in question have high leverage or high influence?

Remark: Arguably it is a case that this problem is not best described by a simple linear regression model, and that a non-linear model would be more suitable.

2.4 Diagnostic Plots

Homoscedascity (constant variance) is one of the assumptions required in a regression analysis in order to make valid statistical inferences about population relationships.

Homoscedasticity requires that the variance of the residuals are constant for all fitted values, indicated by a uniform scatter or dispersion of data points about the trend line (i.e. "The Zero Line").

From the above plot, we can conclude that the constant variance assumption is valid. We can see that the mean value of the residuals is zero.

```
plot(fit1)
#Four Diagnostic Plots are printed to screen sequentially.
```

3 Multiple Linear Regression

In your future studies, you will come across multiple linear regression (MLR). This is a linear model uses multiple independent variables to explain a single dependent variable.

The implementation is very similar to simple linear regression (SLR). All that is required is to specify the additional independent variables.

```
Fit.slr =lm(y~x)   # SLR: y explained by predictor x
Fit.mlr=lm(y~x+z)  # MLR: y explained by predictors x and z
```

For this case, a linear relationship can be defined by the regression model

$$y = \beta_0 + \beta + 1x + \beta_2 z + \epsilon$$

.

Again, we determine the regression coefficients, i.e. estimates for slopes and intercept. (N.B. There are variations on this notation).

- b_0 : the intercept estimate.
- b_1 : the slope estimate for X
- b_2 : the slope estimate for z

In many project datasets it is possible to implement a MLR model. For the moment, we will just look at slope and intercept estimates, their p-values and the coefficient of determination.

Let try this out using the ***iris*** data set. (This is not be a valid statistical analysis in practice. However we are focussing on the mechanics, so we shall proceed nonetheless).

```
lm(Sepal.Length ~ Sepal.Width + Petal.Width)
```

3.1 Model Selection

There are many important methodologies for determining which combination of predictor variables bests describes a response variable. You will meet this in future modules. We will use two simple ones for this module only.

- Adjusted Rsquared value
- The Akaike Information Criterion (AIC)

The adjusted R-square value is found on the summary output for a fitted model. It is called ***adjusted*** because it takes into account the number of predictor variables being used. The law of parsimony states the simplest model that adequately explains the outcomes is the best. The candidate model with the higher adjusted R squared is considered preferable.

The AIC is a model selection metric often used in statistics.It is computed using the R command `AIC()`.The candidate model with the smallest AIC value is considered preferable.

```
fitA = lm(Sepal.Length ~ Sepal.Width + Petal.Width)
fitB = lm(Sepal.Length ~ Sepal.Width + Petal.Length)

summary(fitA)$adj.r.squared
summary(fitB)$adj.r.squared

AIC(fitA)
AIC(fitB)
```