BIO4300 and ENV4300

Stats fest 2007

Regression analysis

murray.logan@sci.monash.edu.au

MONASH University Science

Simple linear regression

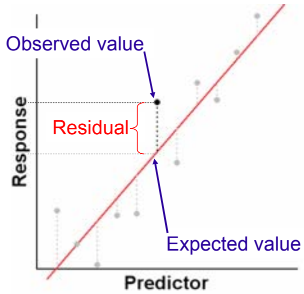
- **Aims**
  - Description
    - Linear relationship between response variable (Y) and predictor variable (X)
  - Explanation
    - How much of the variation in response variable (Y) is explained by linear relationship with predictor variable (X)
  - Prediction
    - New Y values from new X values

Simple linear regression

- **Data**
  - Dependent (response) variable
    - Continuous
    - Normally distributed
  - Independent (predictor) variable
    - Continuous
    - Uniform across a range
  - Each recorded from *n* sampling units (replicates)

Estimating regression parameters

- **y = *b*x + a** $\quad y = \beta_0 + \beta_1 x + \varepsilon$
  - *b* = slope
  - a = y-intercept
- **Ordinary least squares (OLS) regression line**
  - Minimizes residuals
  - Observed – expected
  - Minimizes sum of squared residuals



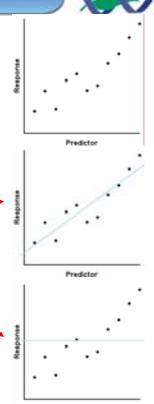
Null hypotheses

- $y = \beta_0 + \beta_1 x + \varepsilon$
- **Null hypotheses (H_0)**
  - Parameter based
    - Population intercept = 0 (β_0=0)
    - Population slope = 0 (β_1=0)
    - Use t-tests

Null hypotheses

- $y = \beta_0 + \beta_1 x + \varepsilon$
- **Null hypotheses (H_0)**
  - Model based (variance based)

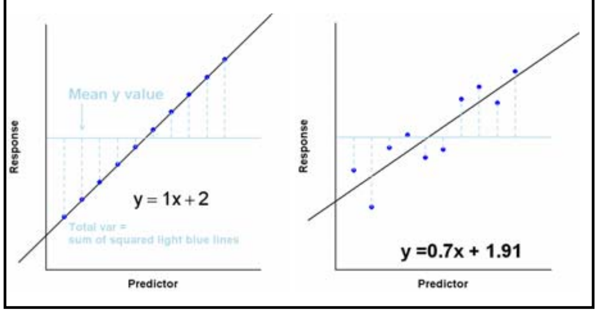
Compare fit $\left\{ \begin{array}{l} y = \beta_0 + \beta_1 x + \varepsilon \\ y = \beta_0 + \varepsilon \end{array} \right.$



- Generate a statistic based on the ratio of fit of the full and reduced models
  - F-ratio

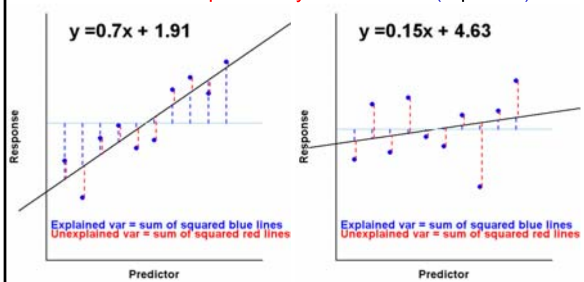
Regression

- **Partitioning of total variance**
  - Does the model (equation) explain the data?



Regression

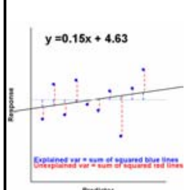
- **Partitioning total variance**
  - Variance explained by linear model (equation)
  - Variance not explained by linear model (equation)



Regression

F-ratio = $\frac{\text{Variation explained}}{\text{Variation unexplained}}$ **(Accounting for *df*)**

- **When H_0 is true F-ratio is expected to be close to zero**
  - Amount explained by the model (equation) is substantially less than the amount not explained



Analysis of Variance Table

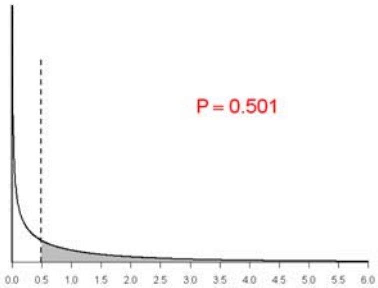
Response: y

	Df	Sum Sq	Mean Sq	F value
x	1	40.806	40.806	0.4925
Residuals	8	29.846	1.351	

Regression

10

- **F-distribution (1, ?)**
- **F-ratio = 30.125**
  - P-value = 0.001
  - Reject H_0
- **F-ratio = 0.4959**
  - P-value =0.501
  - Not reject H_0

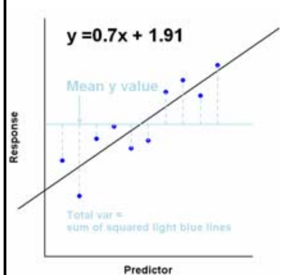


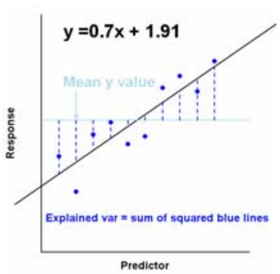
Regression

11

- **Strength of relationship (r^2)**

$$r^2 = \frac{\text{Explained variance}}{\text{Total variance}} = 0.09 (100\%)$$





Regression

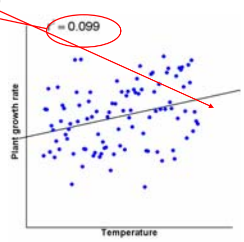
12

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.38993	0.41407	13.017	< 0.001
Slope	0.22319	0.06795	3.285	0.00142

Residual standard error: 1.783 on 98 degrees of freedom
Multiple R-Squared: 0.09918, Adjusted R-squared:0.08999

- **Puts result into perspective**

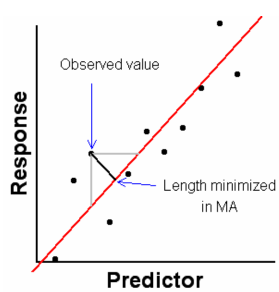


Model II regression

- **When uncertainty in both response and predictor variables**
- **Rather than select levels of the predictor variable to be uniform throughout a range**
  - Measure predictor variable
  - Predictor variable normally distributed
- **E.g. relationship between tree height and DBH**

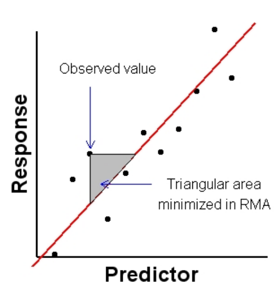
Model II regression

- **Major axis (MA) regression**
  - Minimize perpendicular spread to regression line
  - Assumes degree of uncertainty in X and Y same
- **Normality**
- **Homogeneity of variance**



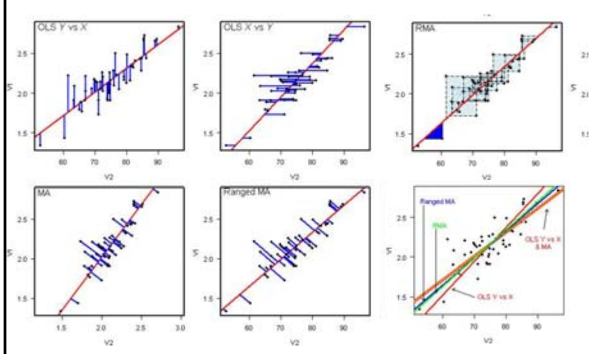
Model II regression

- **Reduced major axis (RMA) regression**
  - Minimize the sum of triangular areas from observed points to regression line
  - Slope = average of slope of Y on X and 1/slope of X on Y
- **Normality**
- **Homogeneity of variance**



Model II regression

16



Model II regression

17

- **Rarely used – why?**
  - Hypothesis tests unaffected
  - No good for predictive formula as we have no measure of uncertainty in new predictor values
  - Only used if need an accurate estimation of the nature of a relationship
    - Size scaling applications
    - Comparing relationship slopes

Simple linear regression

18

- **Linear model**

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

- **Reduced model (when H_0 is true, β_1=0)**

$$y = \beta_0 + \varepsilon$$

- **H_0:**
  - Population slope equals 0 (β_1=0)
  - Population y-intercept equals 0 (β_0=0)
  - Linear model fits better than reduced model

Simple linear regression

19

- **Assumptions**
  - Independent observations
  - Normality (residuals)
    - Boxplot of response variable
  - Homogeneity of variance (residuals)
    - Spread of observations around regression line
    - Residual plot
  - Linearity
    - Scatterplot
      - Lowess smoother

```
> scatterplot(RESPONSE ~ PREDICTOR, data=DATA)
```

Simple linear regression

20

- **Fit linear model**

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

```
> *.lm <- lm(RESPONSE ~ PREDICTOR, data=DATA)
```

Simple linear regression

21

- **Final checks (influence measures)**
  - Residual
    ```
    > resid(*.lm)
    ```
    - How much each Y value differs from expected
  - Leverage
    ```
    > influence.measures(*.lm)
    ```
    - How much of an outlier in X space the observation is
    - Influence of each X value on predicted Y
  - Cook's D
    ```
    > influence.measures(*.lm)
    ```
    - Incorporates residual and leverage
    - Influence of each point on slope
    - Values near or > 1 bad

Simple linear regression

- **Analysis sequence**
  - Design experiment/survey
  - Collect data
  - Test assumptions
  - Fit linear model
    - Estimate parameters
    - Full vs reduced
      - Partition variability into explained & unexplained
      - r^2

Simple linear regression

- **Analysis sequence cont.**
  - Test H_0's

```
> summary(*.lm)
```

    - $\beta_0=0$
      - t-statistic = b_0 / SE(b_0)
      - t-distribution ($df=n-2$)
    - $\beta_1=0$
      - t-statistic = b_1 / SE(b_1)
      - t-distribution ($df=n-2$)
    - Full vs Reduced (explained vs unexplained)
      - F-ratio statistic = $MS_{Regression}/MS_{Residual}$
      - F-distribution (df=1, n-2)
  - Conclusions
    - Reject or not reject H_0

Multiple linear regression

- **Aims**
  - Linear relationship between a response variable and two or more predictor variables
  - Predictions
  - Model selection
- **Data**
  - One response variable (Y)
  - Multiple predictor variables (X_1, X_2, ….)
  - Each variable measured from each sampling unit (n)

Multiple linear regression

25

- **Linear model**

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$

- **Reduced models**

  $y = \beta_0 + \varepsilon$

  $y = \beta_0 + \beta_2 x_2 + \ldots + \varepsilon$

  $y = \beta_0 + \beta_1 x_1 + \ldots + \varepsilon$

- **H_0:**
  - Partial population slope 1 equals 0 (β_1=0)
  - Partial population slope 2 equals 0 (β_2=0)
  - ….
  - Population y-intercept equals 0 (a=0)
  - Linear model fits better than reduced model(s)
    - All partial population slopes = 0

Multiple linear regression

26

- **Assumptions**
  - Independent observations
  - Normality (residuals)
    - Boxplot of variables
  - Homogeneity of variance (residuals)
    - Residual plot
  - Linearity
    - Scatterplot matrix (SPLOM)
    - Partial regression plots

```
> scatterplot.matrix(~RESPONSE+PRED1+PRED2+.., data=DATA)
```

Multiple linear regression

27

- **Assumptions cont**
  - No collinearity – predictors correlated
    - Each predictor variable must be independent
    - If not estimates of partial slopes unreliable
    - Variance-inflation
      - Values > 5 not good, >10 very bad

      ```
      > vif(*.lm)
      ```
    - Correlations between predictor pairs (or SPLOM)

      ```
      > cor(~RESPONSE+PRED1+PRED2+.., data=DATA)
      ```
    - Remove one of correlated variables
    - Center variables
    - Combine via PCA

Multiple linear regression

- **Analysis sequence**
  - Design experiment/survey
  - Collect data
  - Test assumptions
  - Fit linear model
    - Estimate parameters
    - Full vs reduced

```
> *.lm <- lm(RESPONSE~PRED1+PRED2+…, data=DATA)
```

Multiple linear regression

- Test H_0's

```
> summary(*.lm)
```

  - $\beta_0=0$
  - $\beta_1=0$, $\beta_2=0$, …
  - Full vs Reduced (explained vs unexplained)
    - Many competing models

Multiple linear regression

- **Model selection**
  - Selecting the 'best model'
    - Adjusted r^2 — `> summary(*.lm)$'adj.r.squared'`
    - AIC — `> extractAIC(*.lm)[2]`
    - BIC — `> BIC(lm)`
  - Predictor importance
    - Adjusted r^2, AIC, BIC
    - Hierarchical partitioning

```
> hier.part(RESPONSE,data.frame(PRED1,PRED2,…)
```

  - Conclusions
    - Reject or not reject H_0