Chapter 17: Regression

1

Correlation vs regression

- Correlation measures the aspects of the linear relationship between two numerical variables
- **Regression** is a method that predicts values of one numerical variable from values of another numerical variable
- Fits a line through the data
  - Used for prediction
  - Measures how steeply one variable changes with the other

2

Linear regression

- The most common type of regression
  - Although there are non-linear models (e.g., quadratic, logistic)

- Draws a straight line through the data to predict the response variable (Y, vertical axis) from the explanatory variable (X, horizontal axis)

3

Ex 17.1: The lion's nose

- Trophy hunting of African lions typically aims to remove older males

- Is there a way to predict the age?

- Research has shown that the amount of black pigmentation on the nose of male lions increase with age

4

Ex 17.1: The lion's nose

- Study used proportion black on nose (explanatory) to predict age (response)

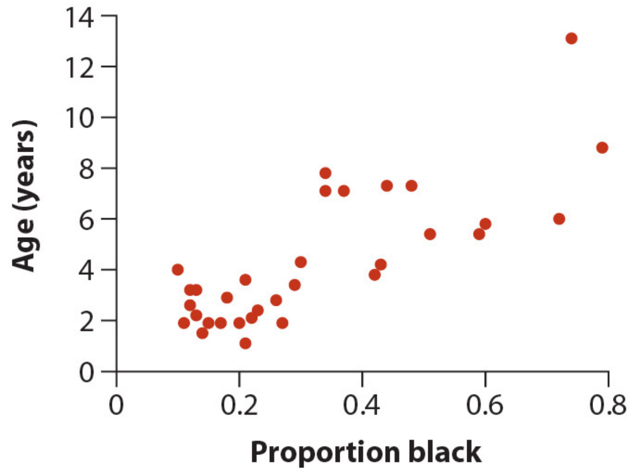


Fig 17.1-1

5

Fitting the "best" line

- You want a line that gives the most accurate predictions of Y from X
- Least-squares regression: line for which the sum of all the squared deviations in Y is smallest

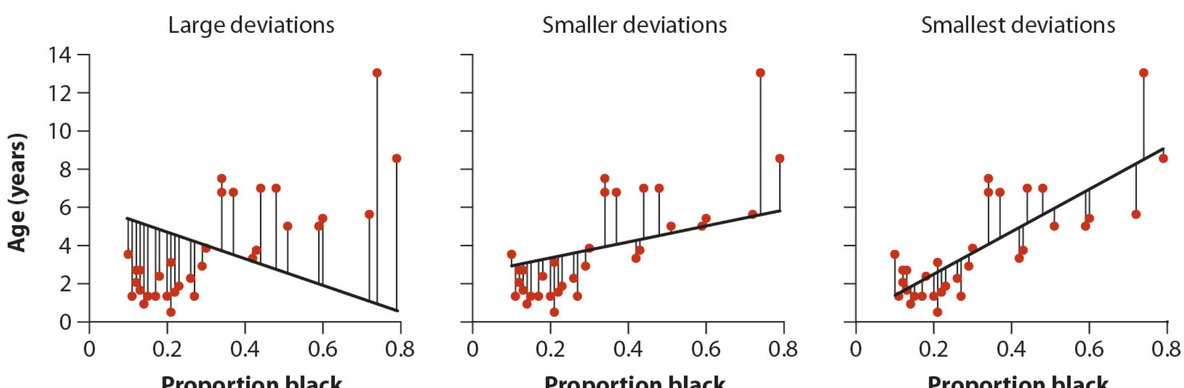


Fig 17.1-2

6

Formula for the line

$$Y = a + bX$$

- a is the Y-**intercept**; b is the **slope**
- The **slope** of a linear regression is the rate of change in Y per unit X
- Also measures direction of prediction
  - Positive: as X increases Y increases
  - Negative: as X increases Y decreases

7

Slopes and intercepts

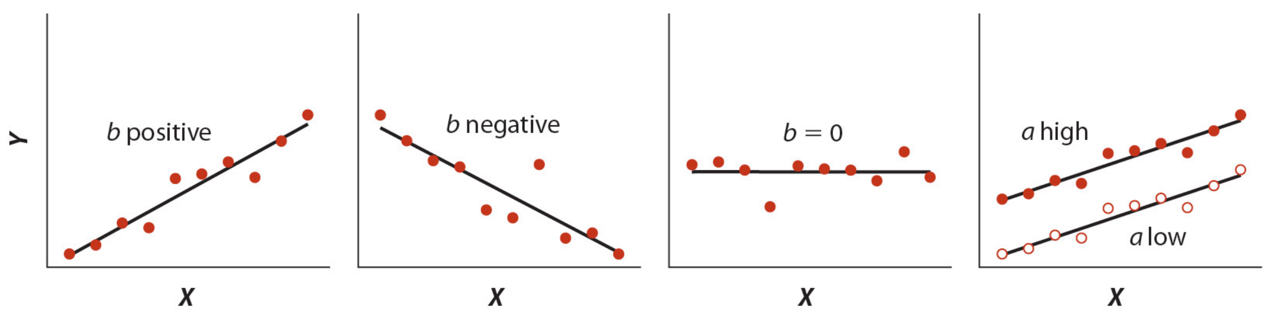


Fig 17.1-3

8

Calculating slope

Measures how deviations in X and Y vary together

$$b = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_i \left(X_i - \overline{X}\right)^2}$$

Sum of squares for X

9

Calculating intercept

- Once slope is calculated, getting intercept is straightforward because the least-squares regression **always** goes through point $(\overline{X}, \overline{Y})$

Plug mean values into line formula: $\overline{Y} = a + b\overline{X}$

Rearrange to solve for intercept: $a = \overline{Y} - b\overline{X}$

10

Ex 17.1: The lion's nose

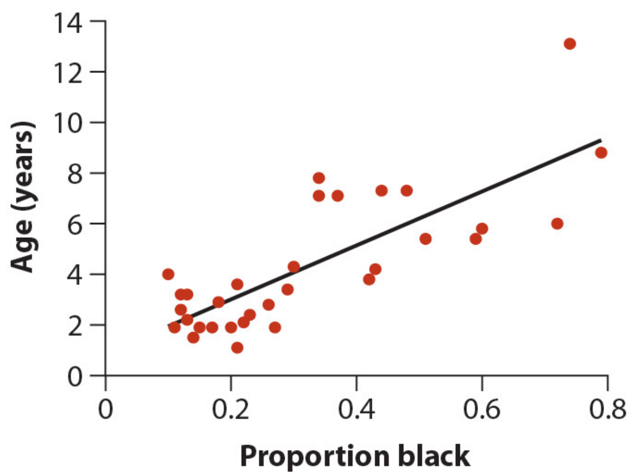
$$b = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} = \frac{13.0123}{1.2221} = 10.647$$

$$a = \overline{Y} - b\overline{X} = 4.3094 - 10.647(0.3222) = 0.879$$

11

Ex 17.1: The lion's nose

$$Y = 0.88 + 10.65X$$



- On average, lion age increases by 10.65 years per unit of change in proportion of nose that is black
- Or, 1.065 years for every 0.1 increase of proportion of black on nose

Fig 17.1-4

12

Samples vs populations

- The slope (b) and intercept (a) are estimated from a sample of measurements, hence these are estimates/statistics
- The true population slope (β) and intercept (α) are parameters
- Regression assumes that there is a population for every value of X, and the mean Y for each of these populations lies on the regression line

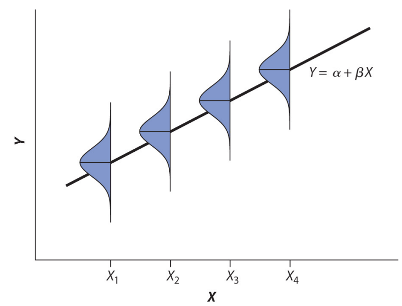


Fig 17.5-1

13

Predicting values

- Now that you have the regression line you can predict values of Y for any specified value of X
- Predictions are mean Y for all individuals with value X
- Designated $\hat{Y}$, or "Y-hat"
- How old are lions with a proportion of black of 0.5?

$$\hat{Y} = 0.88 + 10.65(0.5) = 6.2$$

14

15

How well do data fit line?

- The **residual** of a point is the difference between its measured *Y* value and the value of *Y* predicted by the regression line

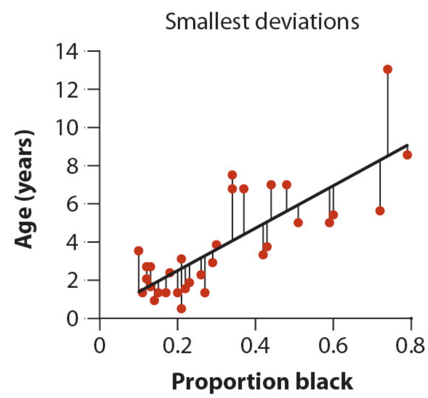


Fig 17.1-2

16

How well do data fit line?

- **Residuals** measure the scatter of points above and below the least-squares regression line
- Can be positive or negative
- Variance in residuals ($MS_{residual}$) quantifies the spread of the scatter
  - **Residual mean square**
  - Analogous to error mean square in ANOVA
- Used to quantify the uncertainty of the slope

17

Residual mean square

$$MS_{residual} = \frac{\sum_i (Y_i - \overline{Y})^2 - b \sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{n - 2}$$

$$= \frac{222.0872 - 10.647(13.0123)}{32 - 2} = 2.785$$

18

Standard error of slope

- Uncertainty (precision) with the sample estimate (b) of the population slope (β)

$$SE_b = \sqrt{\frac{MS_{residual}}{\sum_i \left(X_i - \overline{X}\right)^2}} = \sqrt{\frac{2.785}{1.2221}} = 1.510$$

19

Confidence interval of the slope

$$b - t_{\alpha(2),df} SE_b > \beta > b + t_{\alpha(2),df} SE_b$$

$$10.647 - 2.042(1.510) > \beta > 10.647 + 2.042(1.510)$$

$$7.56 > \beta > 13.73$$

20

21

Two types of predictions

- Predict **mean** Y for a given X
  - e.g., what is the mean age of all male lions whose noses are 60% black?
- Predict **single** Y for a given X
  - e.g., how old is that lion over there with a 60% black nose?
- Both predictions give the same value of $\hat{Y}$, but they **differ in precision**
  - Can predict mean with more certainty than a single value

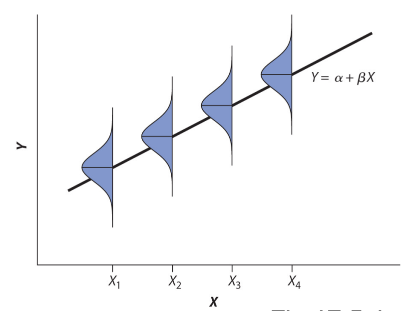


Fig 17.5-1

22

Two types of predictions

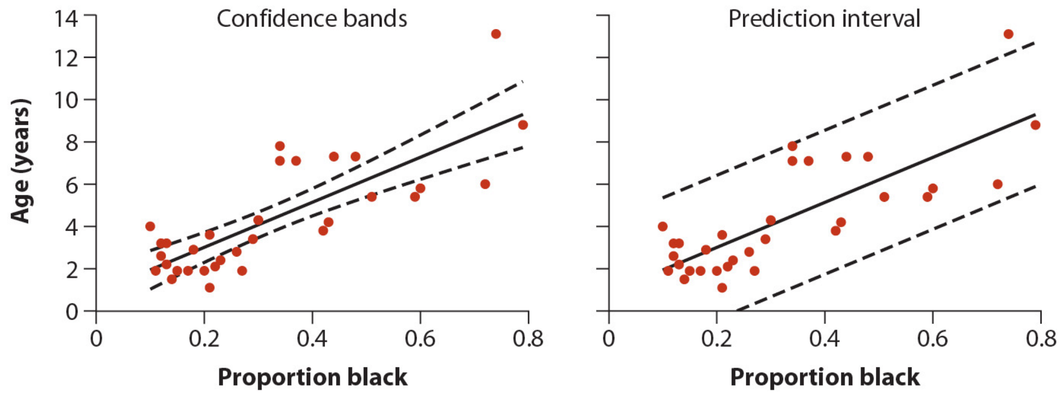


Fig 17.2-1

- **Confidence bands** measure the precision of the predicted mean Y for each value of X. **Prediction intervals** measure the precision of the predicted single Y-values for each X

23

Extrapolation

- Regressions should be used to predict Y for any value of X lying between the smallest and largest values of X (**interpolation**)

- **Extrapolation** is the prediction of the value of a response variable outside the range of X-values in the data

- No way to ensure the relationship continues to be linear beyond the range of the data!

24

25

Testing hypotheses about a slope

- H_0: age cannot be predicted by proportion of black on nose ($\beta = 0$)
- H_A: age can be predicted by proportion of black on nose ($\beta \neq 0$)

26

t-test of regression slope

- t-statistic $t = \frac{b - \beta_0}{SE_b}$ $df = n - 2$

$$t = \frac{10.65 - 0}{1.510} = 7.053 \qquad df = 32 - 2 = 30$$

- P-value for $t_{30} = 7.053$?
- P = 7.68 x 10^{-8} (reject the null hypothesis that $\beta = 0$)

27

ANOVA (F) approach

- Recall two source of variation in ANOVA
  - Among groups (MS_{groups})
  - Within groups (MS_{error})
- In regression framework:
- Deviations between the predicted values $\hat{Y}_i$ and $\overline{Y}$
  - Analogous to MS_{groups}
- Deviations between each Y_i and its predictive value $\hat{Y}_i$
  - Analogous to MS_{error}

28

ANOVA (F) approach

- Using ANOVA approach will generate the same P-value as the t-test approach

- Can be used to measure R^2: the fraction of the variation in Y that is "explained" by X

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

29

summary(regression) in R

```
summary(lionRegression)
##
## Call:
## lm(formula = ageInYears ~ proportionBlack, data = lion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5449 -1.1117 -0.5285  0.9635  4.3421
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.8790     0.5688   1.545    0.133
## proportionBlack  10.6471     1.5095   7.053 7.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.669 on 30 degrees of freedom
## Multiple R-squared:  0.6238, Adjusted R-squared:  0.6113
## F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

Y-intercept

slope

R^2

P-value
H_0: $\alpha = 0$

P-value
H_0: $\beta = 0$

30

31

Regression toward the mean

- **Regression toward the mean** results when two variables measured on a sample of individuals have a correlation less than one. Individuals that are far from the mean for one of the measurements will, on average, lie closer to the mean for the other measurement

32

Regression toward the mean

- Cholesterol measurements before and after drug
- Solid line: linear regression
- Dashed line: one-to-one line with slope of 1

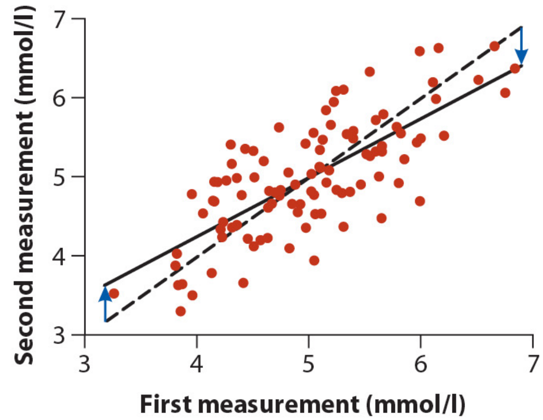


Fig 17.4-1

33

Assumptions of linear regression

- At each value of X:
  - there is a population of Y-values whose mean lies on the regression line
  - the distribution of possible Y-values is normal (with same variance)
  - The variance of Y-values is the same at all values of X
  - the Y-measurements represent a random sample from the possible Y-values

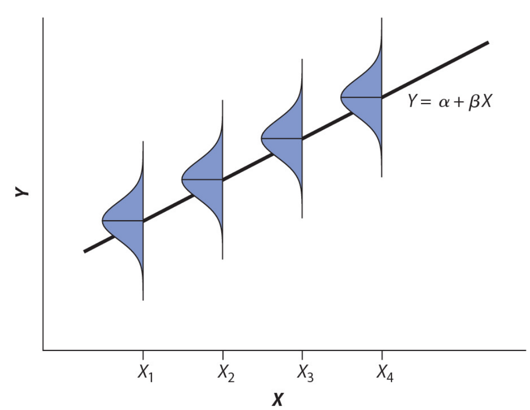


Fig 17.5-1

34

Detecting issues

- **Outliers**
- If only one (or a low number) then it may be reasonable to report regression with and without outlier

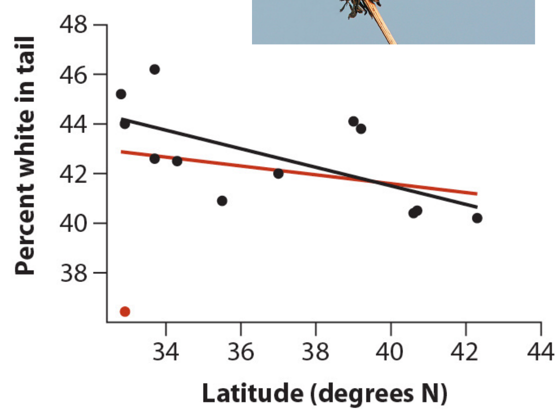


Fig 17.5-2

35

Detecting issues

- **Nonlinearity** can be detected by inspecting graphs

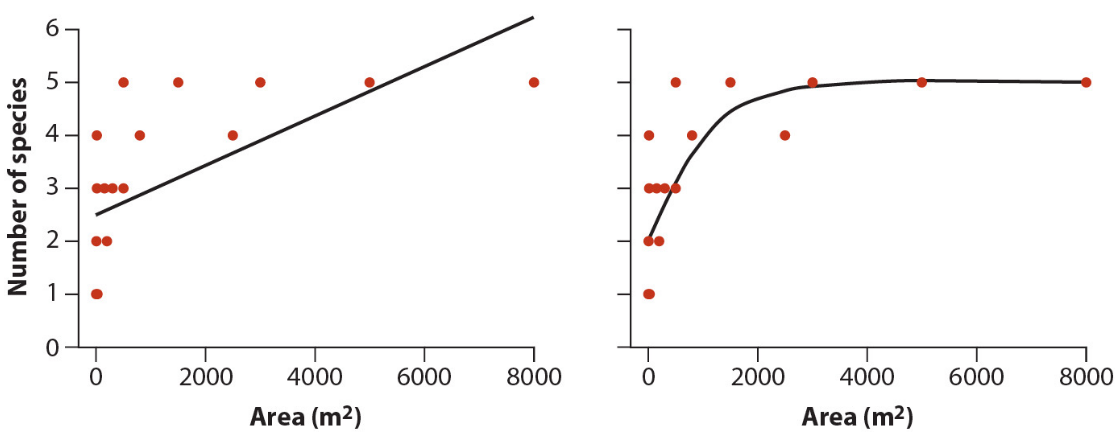


Fig 17.5-3

36

Detecting issues

- **Non-normality and unequal variances** can be inspected with a residual plot
- Residual plot: residual of every data point ($Y_i - \hat{Y}_i$) is plotted against X_i
- If assumptions of normality and equal variances are met then there should be a roughly symmetric cloud above/below horizontal line at 0

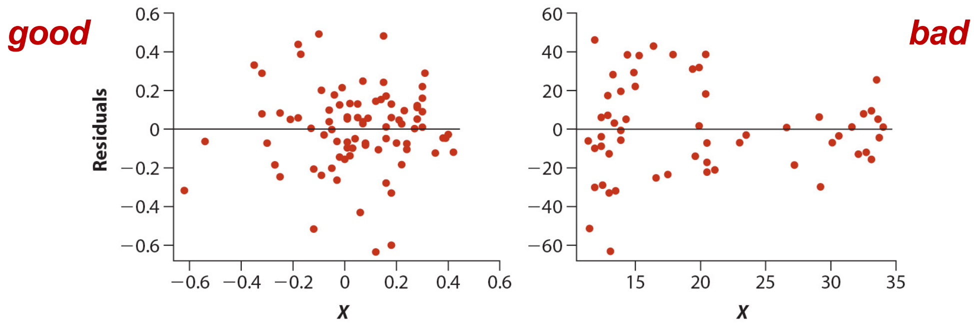


Fig 17.5-4

37

Notes

- Skipping sections 17.6 through 17.9

38