Stat Models (02): Correlation and Regression

Dale Barr
University of Glasgow

1: Regression & multilevel models

lecture	topic
1	introduction
2	correlation & regression
3	multiple regression
4	interactions
5	multilevel models

correlation

notation

- Latin alphabet ($X, Y, r, \ldots$): observed variables associated with the sample ("statistics")
- Greek alphabet (β, σ, ρ): unobserved variables associated with the population ("parameters")
  - estimated parameter value ($\hat{\beta}$, $\hat{\sigma}$, $\hat{\rho}$)
- summation notation (capital "sigma")
  - ΣX : add up all X values

univariate statistics

- mean (μ, $\bar{X}$): $\bar{X} = \frac{\Sigma X}{N}$
- deviation score: $X - \bar{X}$
- standard deviation (σ, S):[1]

$$S = \sqrt{\frac{\Sigma\left(X - \bar{X}\right)\left(X - \bar{X}\right)}{N}}$$

- variance (σ^2, S^2):

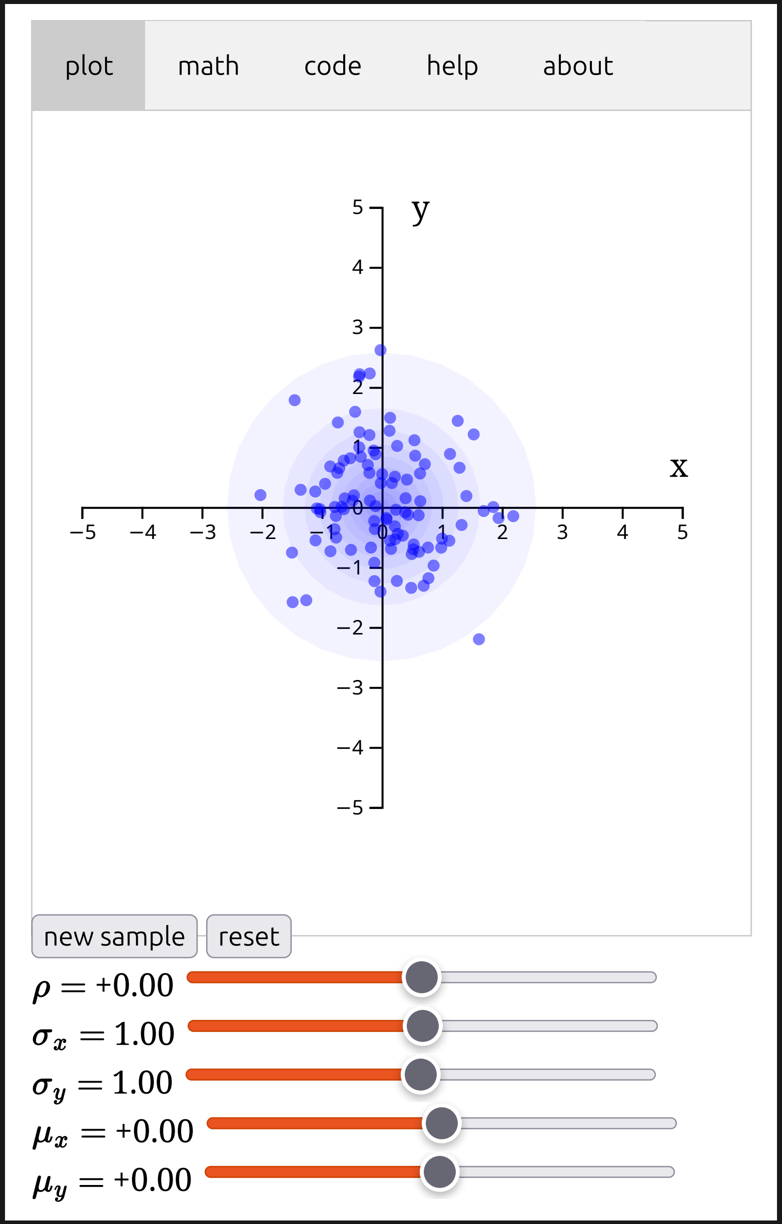
$$S^2 = \frac{\Sigma\left(X - \bar{X}\right)\left(X - \bar{X}\right)}{N}$$

- z-score:

$$z = \frac{X - \bar{X}}{S_X}$$

1. *usually divided by $N - 1$ instead of N when estimating population value*

bivariate data



- scatterplot
- covariance (cov_{XY}):

$$cov_{XY} = \frac{\Sigma\,(X - \bar{X})\,(Y - \bar{Y})}{N}$$

- correlation (ρ_{XY}, r_{XY})

$$r_{XY} = \frac{cov_{XY}}{S_X S_Y} = \frac{\Sigma z_x z_y}{N}$$

$$cov_{XY} = r_{XY} S_X S_Y \text{ or } \rho_{XY} \sigma_X \sigma_Y$$

*N is number of **pairs** of observations*

correlation coefficient

Typicaly denoted as ρ (Greek symbol 'rho') or r

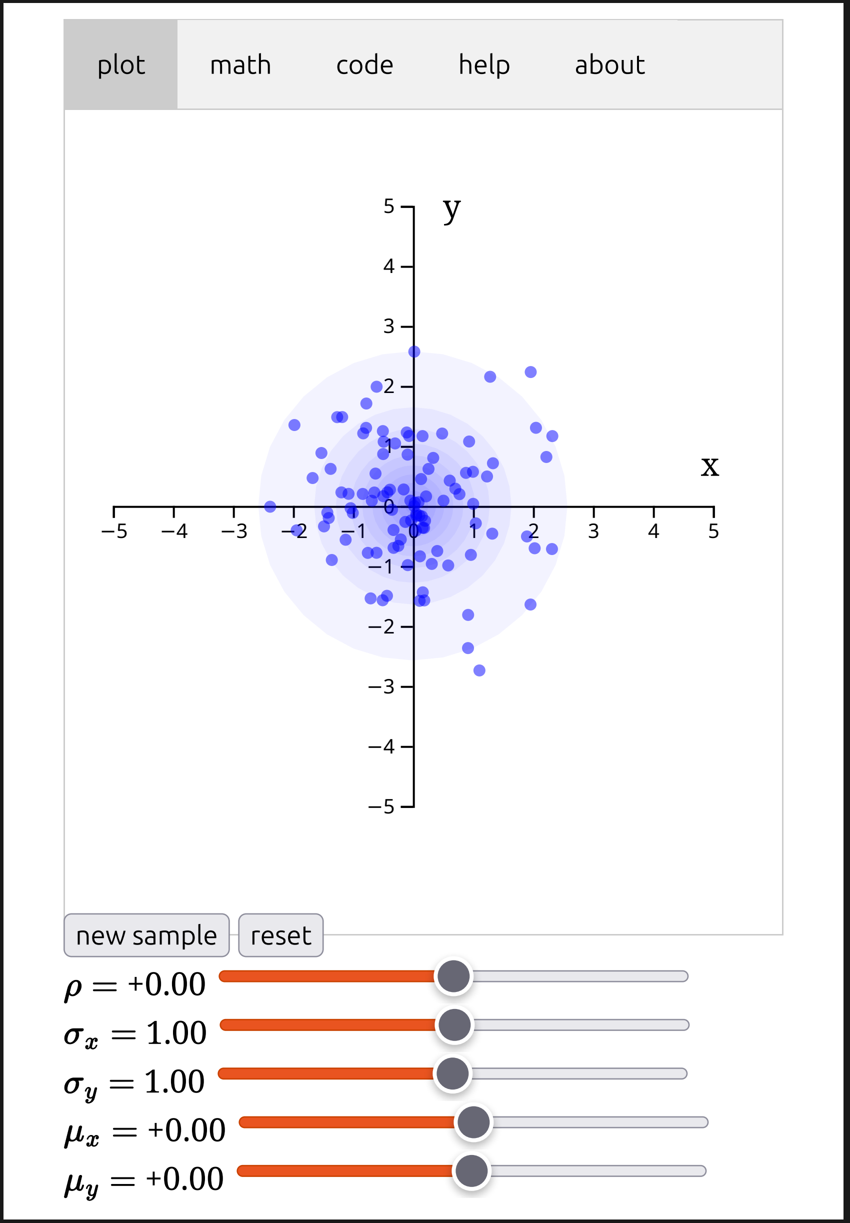
$-1 \geq r \leq 1$

- $r > 0$: positive relationship
- $r < 0$: negative relationship
- $r = 0$: no relationship

Estimated using Pearson or Spearman (rank) method

- `cor()`, `cor.test()`, `corrr::correlate()`

covariance matrices & simulation



regression

univariate analyses

$$Y = ???$$

- predicting from the mean
  - mean height of a 16-24 y.o. Scot: 170cm (about 5’7”)
- using other knowledge
  - 16-24 y.o. man: $\bar{X}$ = 176.2 (~5’9”), S_X = 6.9cm (~2.7”)
  - 16-24 y.o. woman: $\bar{X}$ = 163.8 (~5’5”), S_X = 5.6cm (~2.2”)

Scottish height statistics from the 2008 Scottish Health Survey

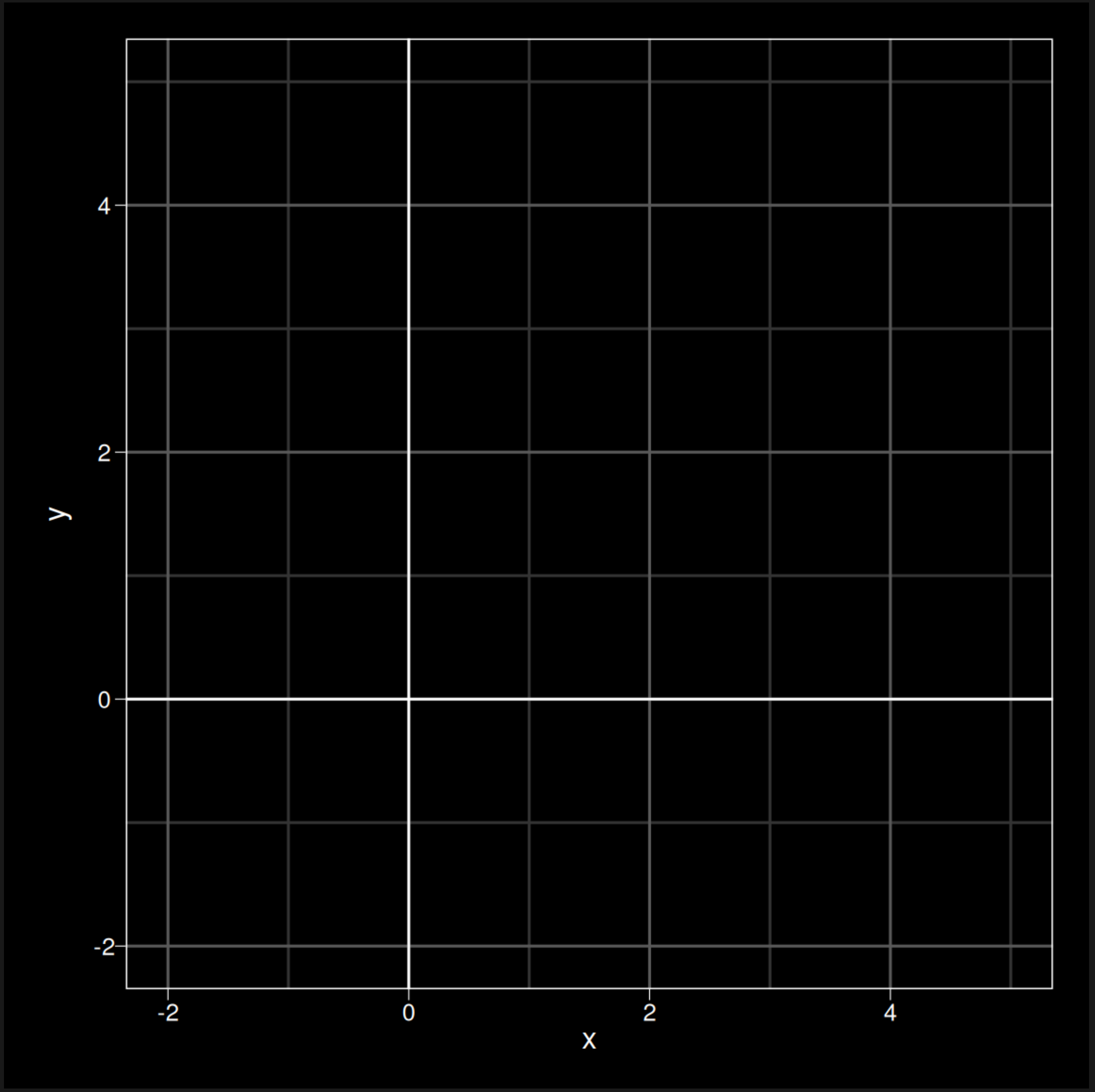
lines

$$y = mx + b$$

$$Y_i = \beta_0 + \beta_1 X_i$$

- y-intercept: value of Y where the line cuts through the vertical axis ($X = 0$)
- slope: effect of 1 unit increase of X on the value of Y

$$m = \frac{\Delta_Y}{\Delta_X}$$



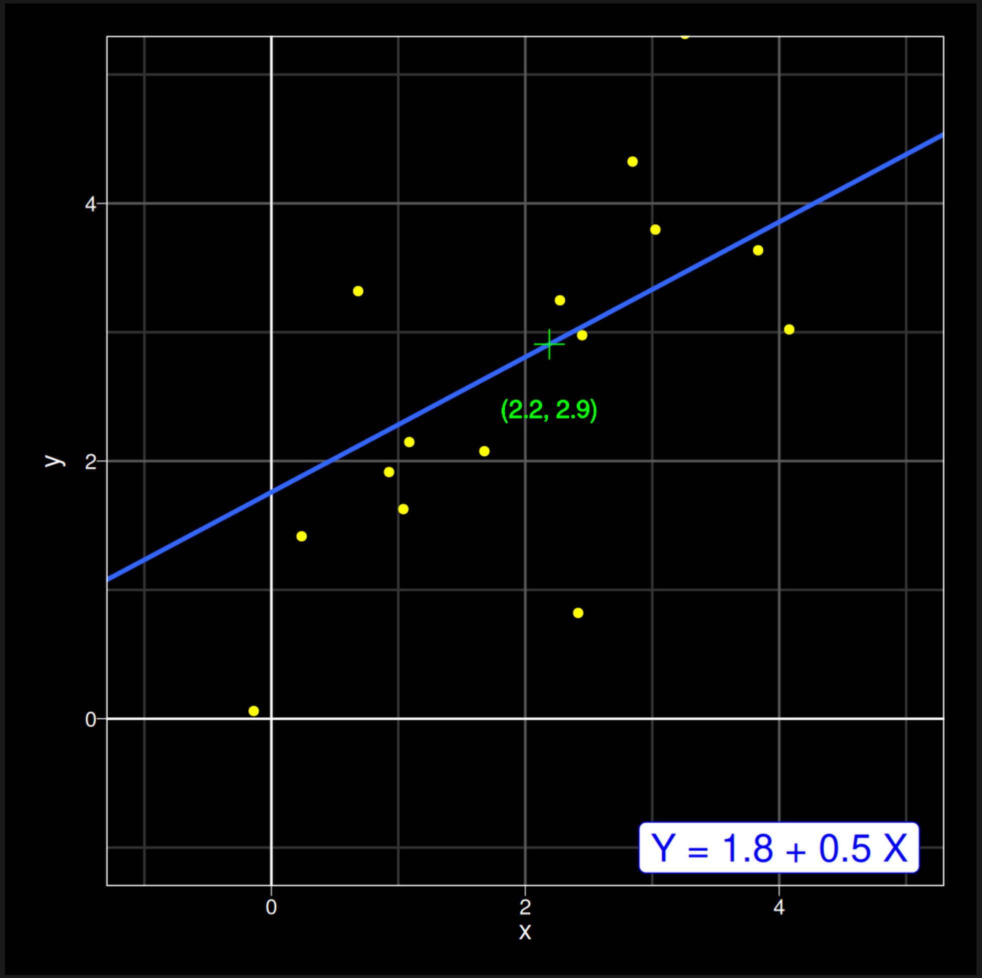
"least squares" regression

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

- Y_i: response variable (criterion, DV)
- $\hat{Y}_i$: fitted value
- X_i: predictor variable (IV)
- β_0, β_1: coefficients
- e_i: error; $\hat{e}_i = Y_i - \hat{Y}_i$: residual

line of best fit minimizes "sum squared error"; passes through $(\bar{X}, \bar{Y})$



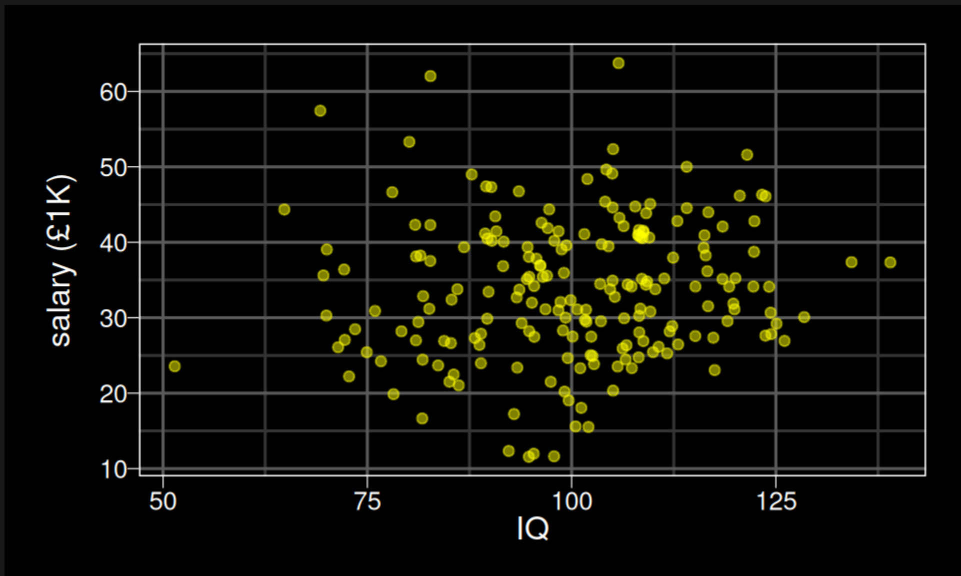
Fitting in R with `lm()`

```r
mod <- lm(y ~ x)

summary(mod)
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2046 -0.6478 -0.1568  0.9184  1.8506 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.7582     0.4229   4.158 0.000741 ***
x             0.5245     0.1471   3.566 0.002577 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.163 on 16 degrees of freedom
Multiple R-squared:  0.4429,	Adjusted R-squared:  0.408 
F-statistic: 12.72 on 1 and 16 DF,  p-value: 0.002577
```

mean-centering predictor values

predicting annual salary from IQ



Uncentered | Centered

```
## note y-intercept meaningless:
## predicted salary for IQ=0

summary(lm(salary ~ IQ, data = IQ_dat))
```

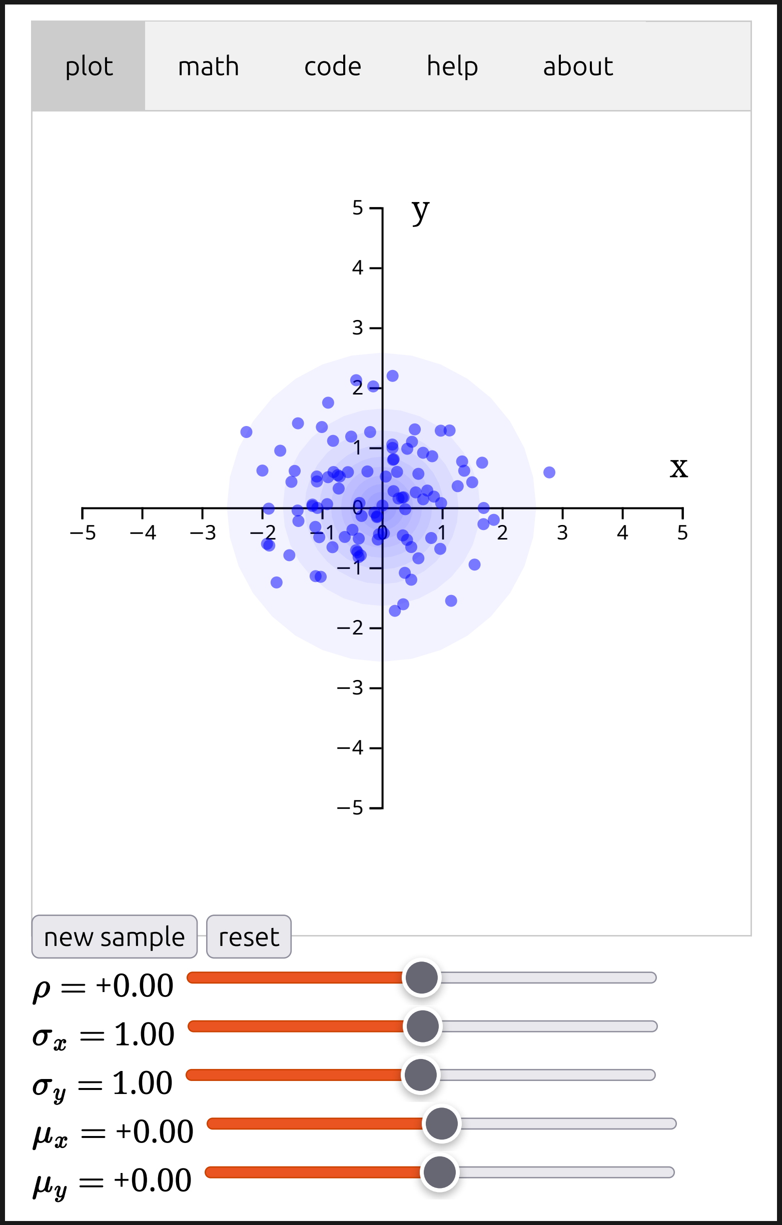
```
Call:
lm(formula = salary ~ IQ, data = IQ_dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-21.9480  -6.2419  -0.4654   6.6332  29.6683 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 27.31845    4.49689   6.075 6.24e-09 ***
IQ           0.06399    0.04446   1.439    0.152    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 9.212 on 198 degrees of
freedom
Multiple R-squared:  0.01036,	Adjusted R-squared:
0.005357 
F-statistic: 2.072 on 1 and 198 DF,  p-value: 0.1516
```

categorical predictors

- are cat owners happier than dog owners?
- define a ‘dummy’ predictor
  - `has_dog` (0 for cat, 1 for dog)
- NOTE: sign of the slope is arbitrary!

NB: we will deal with categorical variables having more than 2 categories when we get to multiple regression

relationship between correlation & regression



- $\beta_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$
- $\beta_0 = \mu_Y - \mu_X \beta_1$

Note: standard deviations can never be negative.

So:

- $\beta_1 = 0$ is the same as $\rho = 0$
- $\beta_1 > 0$ implies $\rho > 0$
- $\beta_1 < 0$ implies $\rho < 0$
- Rejecting the null hypothesis that $\beta_1 = 0$ is the same as rejecting the null hypothesis that $\rho = 0$

assumptions

When calculating a Pearson product-moment correlation coefficient between two variables X and Y, or performing a regression, we assume:

- **linearity**: the relationship between X and Y is linear
- **normality of residuals**: deviations from line of best fit are normally distributed
- **homogeneity of variance** of Y across values of X
- **independence of residuals**

other things to worry about

- restricted-range effects
- outliers
- DV is not continuous
    - see “generalized linear models”