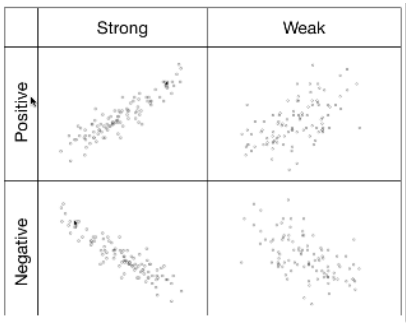
1

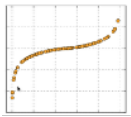
Correlation: compares two numeric variables

Pearson: assumes linearity




```
cor.test(x,y)
```

Spearman: rank-based



```
cor.test(x,y,method="spearman")
```

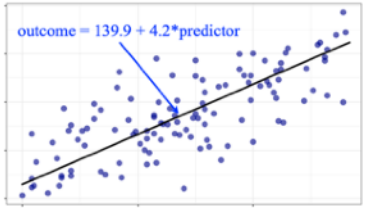
2

Regression: compares two or more numeric variables (and evaluates which of them most influence an outcome)

Linear regression: finds line of best fit to the data for one predictor X and one outcome Y, each datapoint i is given by:

$$Y_i = b_1 X_i + b_0 + \varepsilon_i$$

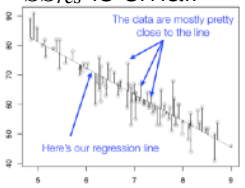
b_1 = slope of the line
b_0 = intercept
ε_i = residual, i.e., deviation from line



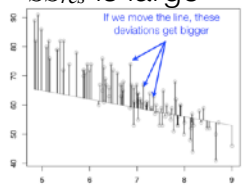
this is the line that minimises the deviations between the line and the data (SS_{res})

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

SS_{res} is small



SS_{res} is large



multiple regression: more than two predictors

$$Y_i = b_k X_{ki} + \ldots + b_1 X_{1i} + b_0 + \varepsilon_i$$

for each predictor k, figures out its slope and how much it affects the outcome

Interactions included by multiplying factors:
$X_{ki} * X_{ji}$ is an interaction between X_k and X_j

3

```
myModel <- lm(y ~ x1 + x2 + x1:x2, dataset)
```

this command returns a model object which also prints out slope of each factor (coefficients) as well as the intercept

```
summary(myModel)
```

4

Hypothesis test 1: Is the overall model significantly different to what you'd expect if H0 were true?

H0: No relationship between X and Y

H1: Relationship between X and Y as in the best-fitting regression equation

- look at *overall* F-statistic and p-value

Hypothesis test 2: Is a specific predictor (or interaction term) significant?

- look at p-value and slope for each predictor
- R does t-tests for each predictor against outcome

Effect size: given by R^2, which captures proportion of the variance in Y accounted for by the model

- the same as the correlation coefficient r squared (for the model with only that predictor)

5

standardized coefficients β: allow you to compare weight of variables

- calculated doing regression on z-scores