Unit 05: Percy Weasley and linear regression

Applied AI with R

Ferdinand Ferber and Wolfgang Trutschnig

Paris Lodron Universität Salzburg

4/20/24

Table of contents I

Percy Weasley and linear regression

AI generated image for the prompt "Percy Weasley with a large ruler in his hand in a hallway in Howgards."

Percy Weasley and linear regression

- What is linear regression?
  - We want to predict one variable (the *outcome*) from all other variables (the *predictors*)...
  - ...and assume a linear/affine relationship between them
- Why linear regression?
  - Interpretable
  - Statistically understood
  - Performs surprisingly well in many situations
  - Very fast (time complexity of $\mathcal{O}(np^2 + p^3)$ for n datapoints and p predictors)

Percy Weasley and linear regression

Linear model

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_p X_p + \epsilon$$

- ...thereby Y is the outcome, $X_1, \ldots, X_p$ are the predictors, $c_1, \ldots, c_p$ are the model parameters (to be learned) and ϵ is some random (unobservable) noise.
- We will view $X_1, \ldots, X_n$ and ϵ as random variables.
- To study linear regression, we first need some basics on correlation.

Section 1

Variance, Covariance and Correlation

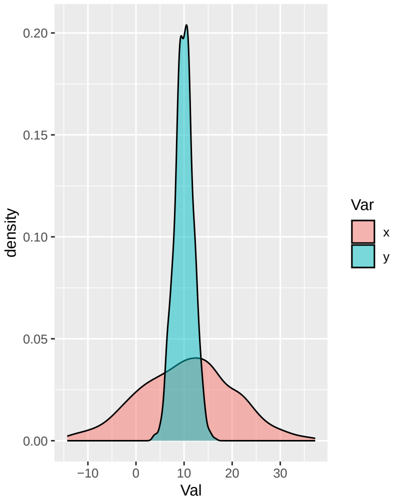
Variance

- *Variance* quantifies the dispersion/spread of a random variable
- It is defined as the expected squared deviation from the mean
- In layman's terms: "On average, how far is a point away from the mean?"

Variance

The *variance* of a random variable X is defined as

$$\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$



Empirical Variance

- In practice we don't know the variance of a random variable
- But we can estimate it

Empirical variance

Let X be a random variable and $x_1, \ldots, x_n$ be a random sample of X. Then

$$\hat{\mathbb{E}}[X] := \overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$\hat{\mathbb{V}}[X] := s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mathbb{E}}[X])^2$$

are unbiased estimators for the expectation and the variance of X

Empirical Variance

- The following equality is easy to derive:

$$\hat{\mathbb{V}}[X] = s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \overline{x}_n^2\right)$$

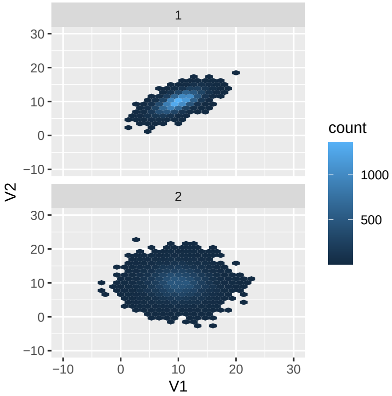
- Advantage: The sums $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ can both be computed in the same traversal of the data and from them both mean and variance are computable

Exercise

- Assume you have a random variable X and collected 8 samples: $1, 2, 3, 4, 5, 6, 7, 8$. Estimate the mean and the variance of X.
- Assume you have a random variable Y and collected 8 samples: $1, 3, 2, 3, 4, 3, 5, 6$. Estimate the mean and the variance of Y.

Covariance matrix

- If we have two random real-valued variables X and Y, we might are naturally interested in the pair (X, Y).
- This is now a 2d random variable.
- We can still ask for the amount of *variance* or the *spread* of it.
- But now we have two magnitudes and a direction of the variance.
- The *covariance matrix* captures all the information



Covariance matrix

Covariance matrix

Let X and Y be two random variables. Set $Z := (X, Y)$. Then the *covariance matrix* is defined as

$$S_{X,Y} := \mathbb{E}\left[(Z - \mathbb{E}[Z])\,(Z - \mathbb{E}[Z])^{\top}\right]$$

- It turns out that the diagonal entries of $S_{X,Y}$ are the variances of X resp. Y.
- The off-diagonal entries are called *covariances*:

$$\Sigma_{X,Y} =: \begin{pmatrix} \mathbb{V}[X] & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathbb{V}[Y] \end{pmatrix}$$

Estimating the covariances

The covariances can be estimated by

$$\begin{aligned}\hat{\Sigma}_{X,Y} = \widehat{\mathsf{Cov}}(X,Y) &:= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mathbb{E}}[X])(y_i - \hat{\mathbb{E}}[Y]) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]\right)\end{aligned}$$

Exercise

- Assume you have a pair (X, Y), where you collected 8 samples $(1,1),(2,3),(3,2),(4,4),(5,3),(6,3),(7,5),(8,6)$
- Compute the (estimated) covariance matrix of (X, Y).

Pearson correlation

- The covariance can be interpreted as a measure of linear dependence of the two random variables
- But its value depend on the variances of the two underlying random variables
- Normalizing the covariance yields the *Pearson correlation coefficient*:

Pearson correlation

Given two real-valued random variables X and Y, the *Pearson correlation* between them is defined as

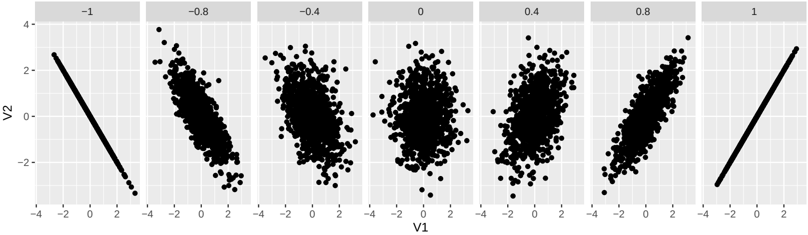
$$\rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}[X]} \cdot \sqrt{\mathbb{V}[Y]}} \in [-1,1]$$

Exercise

- Calculate the Pearson correlation $\hat{\rho}_n \in [-1, 1]$ for the sample of the pair (X, Y) from the last exercise.
- Consider the sample version $\hat{\rho}_n \in [-1, 1]$ for a general sample (i.e., use the sample versions for the covariance and the variances). Can you prove that we always have $\hat{\rho}_n \in [-1, 1]$?

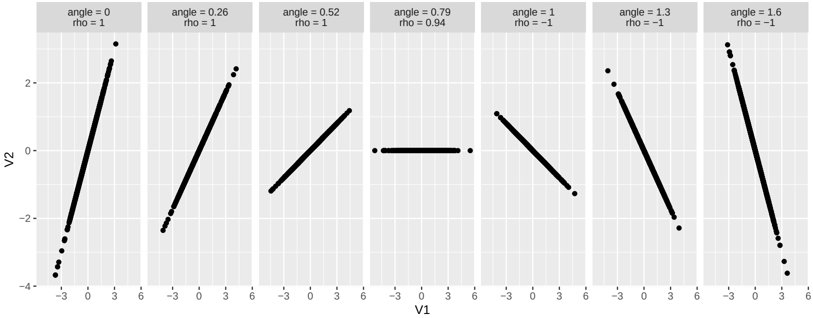
Interpretation of the Pearson correlation

- The Pearson correlation $\rho_{X,Y}$ measures the (extent of) *linear* dependence between X and Y
  - If $\rho_{X,Y} = +1$, then the data points lie perfectly on a straight line with positive slope
  - If $\rho_{X,Y} = 0$, then there is no *linear* dependence between X and Y
  - If $\rho_{X,Y} = -1$, then the data points lie perfectly on a straight line with negative slope



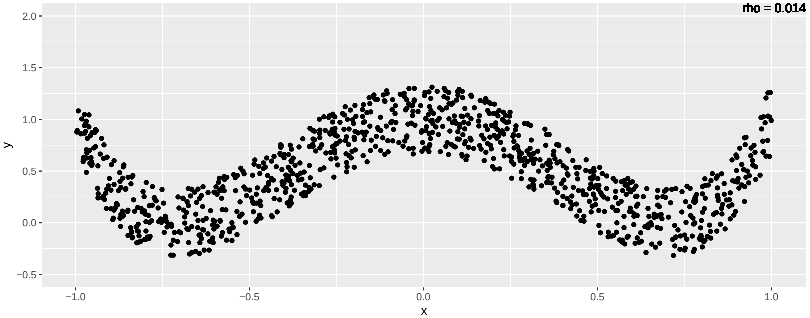
Interpretation of the Pearson correlation

- Notice that the Pearson correlation does not provide detailed information on the slope (other than "upwards" or "downwards"):



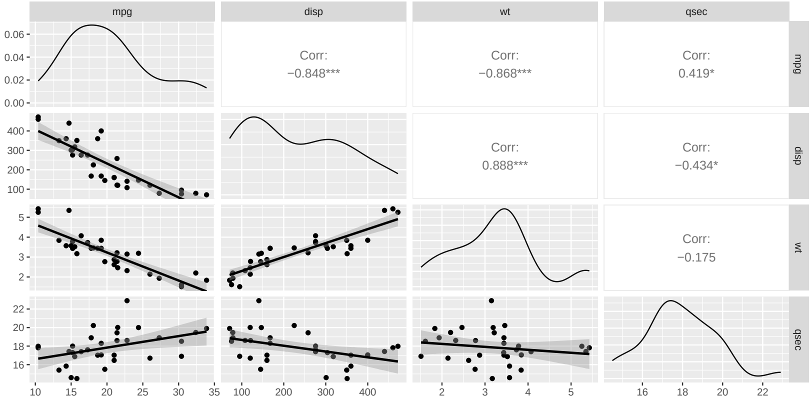
Limitations of Pearson correlation

- Also notice that the Pearson correlation only measures *linear* dependence and that it has no direction (i.e., it is symmetric):



Exercise

- Use the `ggpairs()` function of the `{GGally}` package to visualize the pairwise correlations of a dataset:



Section 2

Univariate Linear Regression

Linear Regression

- In regression we have a couple of random variables $X_1, \dots, X_n, Y, \epsilon$ and assume the following relationship to hold:
$$Y = f(X_1, \dots, X_n) + \epsilon$$
- The function f (called regression function) is unknown (to be estimated) and we generally assume the error ϵ to satisfy $\mathbb{E}[\epsilon] = 0$
- In *linear regression* we additionally assume that f has the form $f(X_1, \dots, X_n) := a_0 + a_1 X_1 + \dots + a_n X_n$, where the $a_1, \dots, a_n$ are unknown parameters

Univariate linear regression

- We start with the simples case, univariate linear setting, i.e.

$$f(X) = a + bX$$

  for a real-valued random variable X
- General idea:
  - Collect samples $(x_1, y_1) \dots, (x_n, y_n)$ from (X, Y)
  - For given parameters $\hat{a}$ and $\hat{b}$ we estimate the error as $L(\hat{a}, \hat{b}) := \sum_{i=1}^{n} (\hat{a} + \hat{b} x_i - y_i)^2$
  - Among all possible $\hat{a}$ and $\hat{b}$, choose those ones that minimize $L(\hat{a}, \hat{b})$
- Open questions
  - How can we solve the optimization problem efficiently?
  - How good are our estimates for a and b?

Reminder: Function optimization

- Let $A \subseteq \mathbb{R}^m$ and $f : A \to \mathbb{R}$ be a function. Candidates for (local) extrema are:
- Points $x \in \mathbb{R}^m$ where $\nabla f(x) = 0$
- Points $x \in \mathbb{R}^m$ where $\nabla f'(x)$ is undefined (in particular $x \in \partial A$)
- Remember that $\nabla f(x) := \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x)\right)$ is the *gradient* of f.

Univariate linear regression

- The loss function for the univariate linear regression was

$$F(\hat{a},\hat{b}) = \sum_{i=1}^{n} (\hat{a} + \hat{b}x_i - y_i)^2$$

- and the partial derivates can be easily seen as

$$\frac{\partial F}{\partial \hat{a}}(\hat{a},\hat{b}) = \sum_{i=1}^{n} 2(\hat{a} + \hat{b}x_i - y_i)$$

$$\frac{\partial F}{\partial \hat{b}}(\hat{a},\hat{b}) = \sum_{i=1}^{n} 2(\hat{a} + \hat{b}x_i - y_i)x_i$$

Univariate linear regression

- Setting these two equations to zero yields the following system of linear equations:

Univariate linear regression

Let $(x_1, y_1), \dots, (x_n, y_n)$ be some data points. Then the best line of fit through the data points is given by $\hat{a} + \hat{b}x$, where $\hat{a}$ and $\hat{b}$ have to fullfill

$$\underbrace{\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}}_{=:C} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Univariate linear regression

- It is not hard to calculate the following quantities (use Cramer's rule):

$$\det(C) = n(n-1)\hat{\mathbb{V}}[X]$$

$$\hat{b} = \frac{\widehat{\mathsf{Cov}}(X,Y)}{\hat{\mathbb{V}}[X]} = \hat{\rho}_{X,Y}\frac{\sqrt{\hat{\mathbb{V}}[Y]}}{\sqrt{\hat{\mathbb{V}}[X]}}$$

$$\hat{a} = \mathbb{E}[\hat{Y}] - \hat{b}\cdot\hat{\mathbb{E}}[X]$$

- So the regression line can be fully determined by statistical measures of X and Y.

Exercise

- Again consider the samples of (X, Y) from the previous exercise.
- You already computed the empirical covariance matrix and the empirical Pearson correlation for (X, Y)
- Compute the linear regression coefficients for the model $Y = a_0 + a_1 X$.

Coefficient of Determination (R-squared)

- Assume we have samples $(x_1, y_1), \dots, (x_n, y_n)$ of (X, Y)
- Compute $\hat{a}, \hat{b}$ as the parameters of the linear regression line
- Then we can use the model to predict the data points: $\hat{y}_i := \hat{a} + \hat{b}x_i$ for every i
- The *coefficient of determination*, also called *R-squared*, of the model is defined as

$$R^2 := 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{\mathbb{E}}[Y])^2}$$

- Interpretation: R^2 is the proportion of y-variance explained by the linear model

Exercise

- Again consider the samples of (X, Y) from the previous exercises.
- You already computed the linear regression coefficients for the model $Y = a_0 + a_1 X$.
- Now compute the R-squared metric for this model
- If not done yet, construct a dataframe containing the sample (columns x and y) and use the lm command in R to calculate everything without effort.

Coefficient of Determination (R-squared)

- We always have $0 \leq R^2 \leq 1$ (for general models we may also obtain negative values)
- If $R^2 \approx 1$, then the linear model explains the data very well
- If $R^2 \approx 0$, then the linear model does not help much to explain the data

Section 3

Multivariate Linear Regression

Multivariate linear regression

- So far: $Y = a + bX + \epsilon$ (univariate)
- Now: $Y = a_0 + a_1 X_1 + \cdots + a_m X_m + \epsilon$ (multivariate)
- We can formulate the loss function as

$$F(\hat{a}) := (X\hat{a} - y)^\top (X\hat{a} - y)$$

where

$$X := \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix} \quad \hat{a} := \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_m \end{pmatrix} \quad y := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- X collects the observed predictors, y collects the observed outcomes and $\hat{a}$ collects the estimated model coefficients

Multivariate linear regression

- One can show (tedious/ugly, but not hard) that

$$\nabla F(\hat{a}) = 2X^\top X\hat{a} - 2X^\top y$$

- Setting this to zero and rearranging yields

$$\hat{a} = \left(X^\top X\right)^{-1} X^\top y$$

- This is the solution to our problem of estimating the model coefficients $\hat{a}$, given the data X and y.

Reminder: Linear regression in R

- In the {tidymodels} framework we can use (multivariate) linear regression as follows:

```
data_split <- initial_split(mtcars, prop = 3/4)
model <- linear_reg()
fitted_model <- model |> fit(
  mpg ~ hp + wt, data = data_split |> training()) |>
  extract_fit_engine()
```

Reminder: Linear regression in R

```
summary(fitted_model)
```

```
Call:
stats::lm(formula = mpg ~ hp + wt, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3192 -1.1228  0.0191  0.5908  4.6776 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.631428   1.413716  25.204  < 2e-16 ***
hp          -0.035951   0.008776  -4.096 0.000516 ***
wt          -3.244213   0.536678  -6.045 5.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.945 on 21 degrees of freedom
Multiple R-squared:  0.8815,	Adjusted R-squared:  0.8702 
F-statistic: 78.12 on 2 and 21 DF,  p-value: 1.876e-10
```