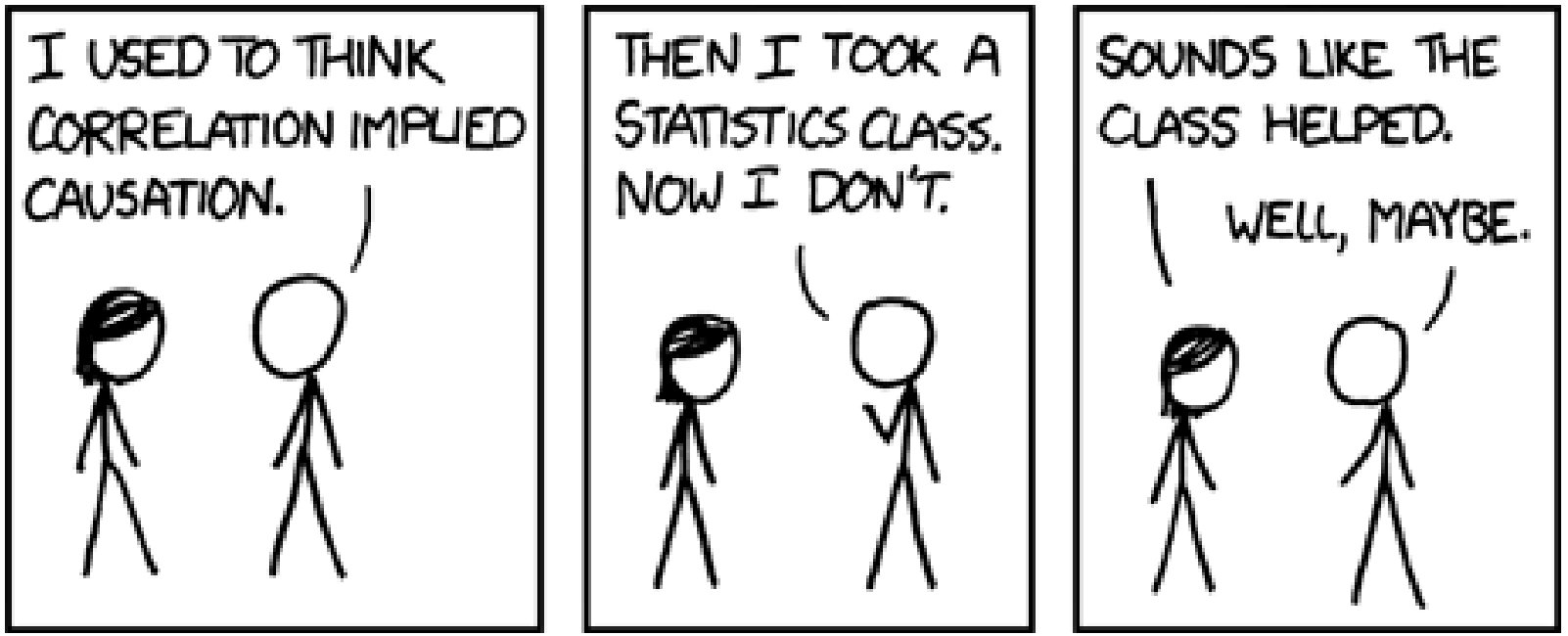
Correlation



https://xkcd.com/552/

Covariance and Correlation

Lecture 22, CS70 Summer 2025

Covariance and Correlation

Conditional Expectation

Prediction

Review – Covariance

Last time we showed that:

$$\mathrm{Var}(X+Y) = E[(X+Y)^2] - (E[X+Y])^2$$

$$= \mathrm{Var}(X) + 2E[XY] - 2E[X]E[Y] + \mathrm{Var}(Y)$$

If X and Y are independent, then these black terms canceled out.

If they are not independent?

- In that case, the variance of the sum will depend on this term. This term (excluding the factor of 2) is called the covariance.

$$\mathrm{cov}(X,Y) = E[XY] - E[X]\,E[Y]$$

Covariance Properties

We previously showed three properties of variance $E[(X - E[X])^2]$:

- Can rewrite as $\text{Var}(X) = E[X^2] - E[X]^2$
- $\text{Var}(cX) = c^2\text{Var}(X)$
- $\text{Var}(c + X) = \text{Var}(X)$

The covariance $E[XY] - E[X]\,E[Y]$ has similar properties:

- Equal to $E[(X - E[X])(Y - E[Y])]$
- $\text{cov}(X + a, Y + b) = \text{cov}(X, Y)$
- $\text{cov}(cX, dY) = c \cdot d \cdot \text{cov}(X, Y)$

Proofs are similar to those for variance. Good exercise to work through!

Correlation

Recall "interpretation" of covariance sign:

- $\text{cov}(X, Y) > 0$: X and Y generally move in the same direction
- $\text{cov}(X, Y) < 0$: X and Y generally move in opposite directions

What about magnitude? What are the units?

- Units of covariance are the product of the units of X times the units of Y.
  - *Good luck making sense out of that….*

If X and Y are random variables with standard deviations $\sigma_X > 0$ and $\sigma_Y > 0$ respectively, then the correlation of X and Y is given by:

$$\text{Corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Benefit? Always lies between -1 and +1. More intuitive to interpret.

Related concept: Standardizing a Random Variable

If X is a random variable with mean μ and stdev σ, then the standardized version of this random variable is given by $\tilde{X}$:

$$\tilde{X} = \frac{X - \mu}{\sigma}$$

Questions:

Type equation here.

- What is $E[\tilde{X}]$?

$$E[\tilde{X}] = \frac{1}{\sigma}(E[X - \mu]) = \frac{1}{\sigma}(E[X] - \mu) = 0$$

- What is $\mathrm{Var}(\tilde{X})$?

$$\mathrm{Var}(\tilde{X}) = \frac{1}{\sigma^2}\mathrm{Var}(X - \mu) = \frac{1}{\sigma^2}\mathrm{Var}(X) = 1$$

So $\tilde{X}$ is a shifted and scaled version of X to have

- *Zero mean*
- *Unit standard deviation*

Note: "Standardizing" is the idea behind using "z-Scores" for exam grades!

Proof of Correlation Range

Claim: The correlation is always between -1 and 1.

$$-1 \leq \mathrm{Corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \leq +1$$

Lemma: Let $E[X] = \mu_X$, $E[Y] = \mu_Y$. Let $\tilde{X}$ and $\tilde{Y}$ be the standardized versions of X and Y. Then $\mathrm{Corr}(X,Y) = E[\tilde{X}\tilde{Y}]$

Proof of Lemma:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = E\left[\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right] = E[\tilde{X}\tilde{Y}]$$

Proof of Correlation Range

Claim: The correlation is always between -1 and 1.

$$-1 \leq \mathrm{Corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \leq +1$$

Lemma: Let $E[X] = \mu_X$, $E[Y] = \mu_Y$. Let $\tilde{X}$ and $\tilde{Y}$ be the standardized versions of X and Y. Then $\mathrm{Corr}(X,Y) = E[\tilde{X}\tilde{Y}]$

Proof that $\mathrm{Corr}(X,Y) \leq 1$: Consider the quantity $E\left[\left(\tilde{X} - \tilde{Y}\right)^2\right]$, which is ≥ 0.

$$\begin{aligned} E\left[\left(\tilde{X} - \tilde{Y}\right)^2\right] &= E[\tilde{X}^2] - 2E[\tilde{X}\tilde{Y}] + E[\tilde{Y}^2] \\ &= 1 - 2E[\tilde{X}\tilde{Y}] + 1 \\ &= 2 - 2\mathrm{Corr}(X,Y) = 2\big(1 - \mathrm{Corr}(X,Y)\big) \geq 0 \end{aligned}$$

Thus, $1 - \mathrm{Corr}(X,Y) \geq 0$ So: $\mathrm{Corr}(X,Y) \leq 1$

Proof of Correlation Range

Claim: The correlation is always between -1 and 1.

$$-1 \leq \mathrm{Corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \leq +1$$

Lemma: Let $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y$. Let $\tilde{X}$ and $\tilde{Y}$ be the standardized versions of X and Y. Then $\mathrm{Corr}(X,Y) = \mathbb{E}[\tilde{X}\tilde{Y}]$

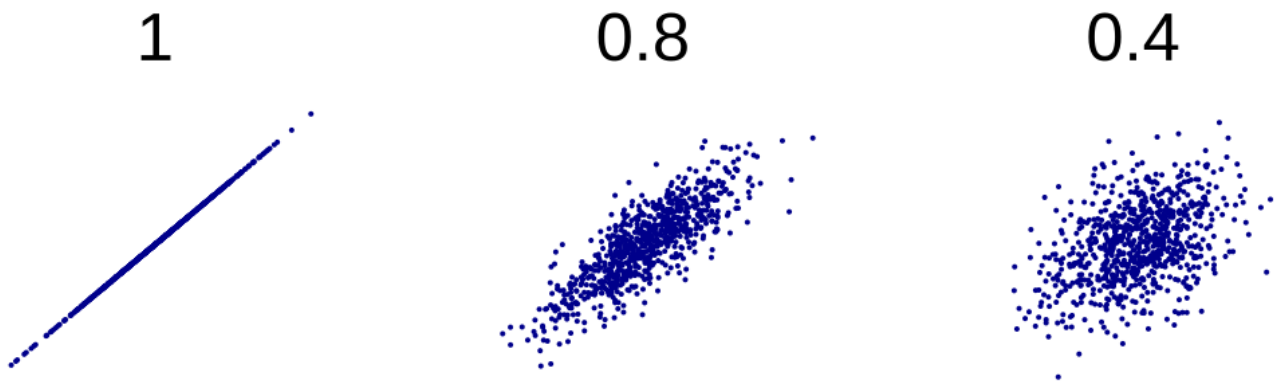
Proof that $\mathbf{Corr}(X,Y) \geq -1$: Consider the quantity $\mathbb{E}\left[\left(\tilde{X} + \tilde{Y}\right)^2\right]$, which is ≥ 0.

$$\begin{aligned}\mathbb{E}\left[\left(\tilde{X} + \tilde{Y}\right)^2\right] &= \mathbb{E}[\tilde{X}^2] + 2\mathbb{E}[\tilde{X}\tilde{Y}] + \mathbb{E}[\tilde{Y}^2] \\ &= 1 + 2\mathbb{E}[\tilde{X}\tilde{Y}] + 1 \\ &= 2 + 2\mathrm{Corr}(X,Y) = 2\big(1 + \mathrm{Corr}(X,Y)\big) \geq 0\end{aligned}$$

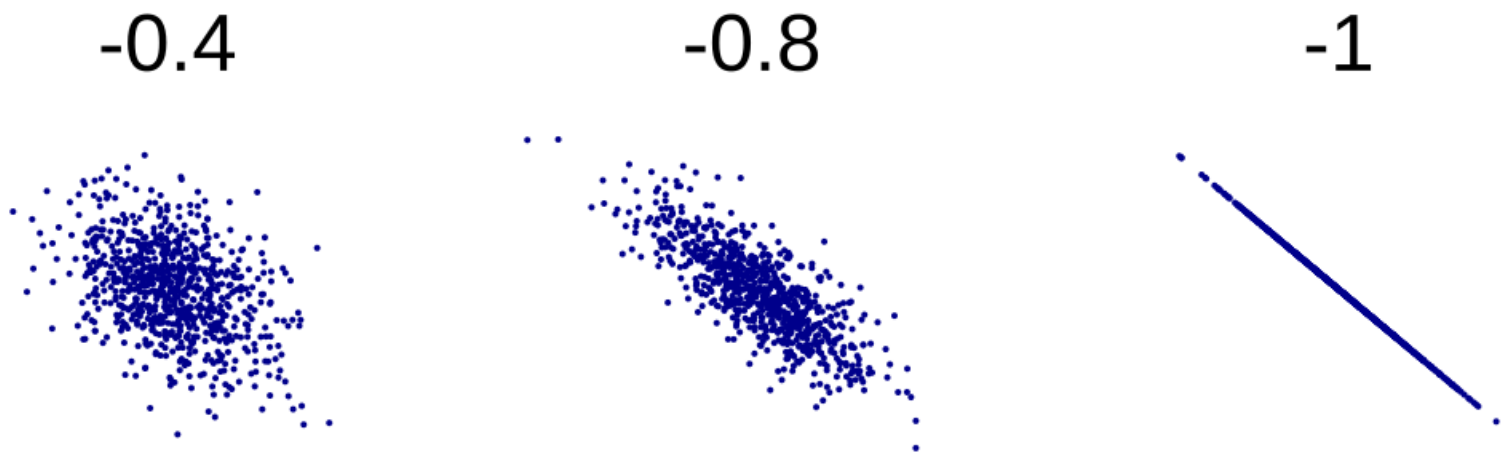
Thus, $2 + 2\mathrm{Corr}(X,Y) \geq 0$ So: $\mathrm{Corr}(X,Y) \geq -1$

Correlation Examples

Examples of random samples of correlated random variables, $\mathrm{Corr}(X, Y) > 0$:



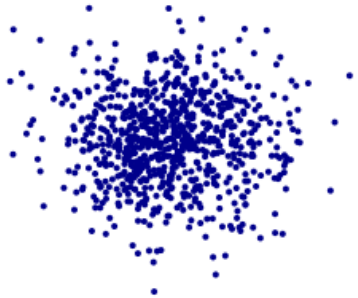
Examples of negative correlated random variables, $\mathrm{Corr}(X, Y) < 0$:



Correlation Of Zero

Examples of independent RV with $\text{Corr}(X, Y) = 0$:

0



Example of dependent RV with $\text{Corr}(X, Y) = 0$. Why are these not independent?

0

Observations about Extreme Cases

Suppose $\text{Corr}(X, Y) = 1$. In that case, we have that $\tilde{X} = \tilde{Y}$. Why?

- Because $E\left[\left(\tilde{X} - \tilde{Y}\right)^2\right] = E[\tilde{X}^2] - 2E[\tilde{X}\tilde{Y}] + E[\tilde{Y}^2] = \text{Var}\left(\tilde{X}\right) + \text{Var}\left(\tilde{Y}\right) - 2\text{cov}(\tilde{X}, \tilde{Y}))$

$$= 2 - 2\ \text{Corr}(\tilde{X}, \tilde{Y})$$

- If $\text{Corr}(\tilde{X}, \tilde{Y}) = 1$, then $E\left[\left(\tilde{X} - \tilde{Y}\right)^2\right] = 0$, so $\tilde{X} = \tilde{Y}$

Implication: if $\text{Corr}(X, Y) = \text{Corr}(\tilde{X}, \tilde{Y}) = 1$, then random variables X and Y are just rescaled and shifted versions of each other, i.e., there exist $a > 0$ and b such that:

$$Y = aX + b$$

Can show a similar result if $\text{Corr}(X, Y) = -1$, in which case $a < 0$.

Conditional Expectation

Lecture 22, CS70 Summer 2025

Covariance and Correlation

Conditional Expectation

Prediction

Intuitive Question

Suppose we flip a biased coin that comes up heads ¾ of the time tails ¼ of the time.

- If it comes up heads, then we roll a fair six-sided dice.
- If it comes up tails, we roll a fair four-sided die.
- Let X be the outcome of the die roll.

What is $E[X]$?

Observation: Can think of this the expectation of one RV dependent on another.

- Answer coming later

Conditional Probability and Events

Let X be a random variable over Ω, and let A be an event. We saw previously that the usual conditional probability rules apply to events involving random variables, e.g.,

$$P(X = x|A) = \frac{P\big((X = x) \cap A\big)}{P(A)}$$

This is because $X = x$ and A are just events, i.e., subsets of Ω.

- Which outcomes are in the event $X = x$? Those outcomes $\omega \in \Omega$ which the function X maps to x.

Conditional Probability and Events

Let X be a random variable over Ω, and let A be an event. We saw previously that the usual conditional probability rules apply to events involving random variables, e.g.,

$$P(X = x|A) = \frac{P\big((X = x) \cap A\big)}{P(A)}$$

Naturally then, we also have the other usual statements about conditional probability. For example, for n disjoint events A_1 through A_n that partition the sample space, we have:

$$P(X = x) = P\big((X = x) \cap A_1\big) + \cdots + P\big((X = x) \cap A_n\big)$$

$$P(X = x) = P(A_1) \cdot P\big((X = x)|A_1\big) + \cdots + P(A_n) \cdot P\big((X = x)|A_n\big)$$

Conditional Distributions and Expectations

We can also think about conditional distributions.

- Example X is the roll of a six-sided die, and A is the event that the die roll is even.

The conditional distribution $P(X|A)$ is:

x	$P(\boldsymbol{X} = x\|A)$
1	0
2	1/3
3	0
4	1/3
5	0
6	1/3

$$E[X|A] = 4$$

We can also define the conditional expectation $E[\mathrm{X}|A]$ naturally as:

$$E[\mathrm{X}|A] = \sum_{x \in \mathrm{range}(X)} x \cdot P(X = x|A)$$

Law of Total Expectation (Binary Event)

For a binary event, we also have:

$$E[\mathrm{X}] = P(A) \cdot E[\mathrm{X}|A] + P(\bar{A}) \cdot E[\mathrm{X}|\bar{A}]$$

Note: This is intuitive but we haven't proven this. We will at the end of this section.

Let's see how we can use this to more formally solve our dice and coin problem.

More Formal Answer to Intuitive Question

Suppose we flip a biased coin that comes up heads ¾ of the time tails ¼ of the time.

- If it comes up heads, then we roll a fair six-sided dice.
- If it comes up tails, we roll a fair four-sided die.
- Let X be the outcome of the die roll.

What is $E[\mathrm{X}]$? More formally, let A be the event that the coin comes up heads.

- $E[\mathrm{X}] = E[\mathrm{X}|A] \cdot P(A) + E[\mathrm{X}|\bar{A}] \cdot P(\bar{A})$

$$= \frac{7}{2} \cdot \frac{3}{4} + \frac{5}{2} \cdot \frac{1}{4}$$

$$= \frac{26}{8} = \frac{13}{4} \qquad = 3.25$$

Example 2 – Rolling A Six-Side Die Twice

Random variables: R_1 is value of first roll, R_2 is second, and $S = R_1 + R_2$

$$E[R_1|S=7] = \sum_{x \in \text{range}(R_1)} x \cdot P(R_1 = x|S=7)$$

Six outcomes with $S = 7$, one for each R_1 value,
so $P(R_1 = x|S=7) = \frac{1}{6}$ for all x, and

$$E[R_1|S=7] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

Knowing $S = 7$ does not add information!

6	7	8	9	10	11	12
5	6	7	8	9	10	11
4	5	6	7	8	9	10
3	4	5	6	7	8	9
2	3	4	5	6	7	8
1	2	3	4	5	6	7
	1	2	3	4	5	6

$E[R_1|S=2]$:

Only one outcome with S=2, so

$$E[R_1|S=2] = 1 \cdot P(R_1 = 1|S=2) = 1 \cdot 1 = 1$$

Less Trivial Example: Expectation of a Geometric Random Variable

Let $X \sim \text{Geometric}(p)$. Let A be the event that the first flip is heads. Let's see how we can use conditional expectation to (for a third time) compute $E[X]$.

$$E[X] = P(A) \cdot E[X|A] + P(\bar{A}) \cdot E[X|\bar{A}]$$

What is $P(A) \cdot E[X|A]$?

Less Trivial Example: Expectation of a Geometric Random Variable

Let $X \sim \text{Geometric}(p)$. Let A be the event that the first flip is heads. Let's see how we can use conditional expectation to (for a third time) compute $E[X]$.

$$E[X] = P(A) \cdot E[X|A] + P(\bar{A}) \cdot E[X|\bar{A}]$$

What is $P(A) \cdot E[X|A]$?

- $P(A) = p$
- And if A is true, $X = 1$, so $E[X|A] = 1$.
- Answer: $p \cdot 1 = p$

Less Trivial Example: Expectation of a Geometric Random Variable

Let $X \sim \text{Geometric}(p)$. Let A be the event that the first flip is heads. Let's see how we can use conditional expectation to (for a third time) compute $E[X]$.

$$E[X] = p + P(\bar{A}) \cdot E[X|\bar{A}]$$

What is $P(\bar{A}) \cdot E[X|\bar{A}]$?

- $P(\bar{A}) = 1 - p$
- Give $E[X|\bar{A}]$ answer in terms of $E[X]$. (Remember memoryless property!)
  - If $Y = X|\bar{A}$ then $Y \sim 1 + \text{Geometric}(p)$, so $E[X|\bar{A}] = 1 + E[X]$

So $P(\bar{A}) \cdot E[X|\bar{A}] = (1 - p)(1 + E[X])$

Less Trivial Example: Expectation of a Geometric Random Variable

Let $X \sim \text{Geometric}(p)$. Let A be the event that the first flip is heads. Let's see how we can use conditional expectation to (for a third time) compute $E[X]$.

$$E[X] = p + (1-p) \cdot (1 + E[X])$$

Now we can solve for $E[X]$:

$$\begin{aligned} E[X] &= p + 1 + E\,[X] - p - p\,E[X] \\ &= 1 + E[X] - p\,E[X] \\ p\,E[X] &= 1 \\ E[X] &= \frac{1}{p} \end{aligned}$$

Law of Total Expectation Proof

More general version: Let X be a random variable over Ω. Let $A_1, \ldots, A_n$ be disjoint events that partition Ω. Then:

$$E[X] = \sum_{i=1}^{n} P(A_i) \cdot E[X|A_i]$$

Proof:

$$E[X] = \sum_{x \in \text{range}(X)} x \cdot P(X = x)$$

$$= \sum_{x \in \text{range}(X)} x \sum_{i=1}^{n} P(A_i) \cdot P(X = x|A_i)$$

$$= \sum_{i=1}^{n} P(A_i) \cdot \underbrace{\sum_{x \in \text{range}(X)} x\, P(X = x|A_i)}_{E[X|A_i]}$$

Prediction

Lecture 22, CS70 Summer 2025

Prediction

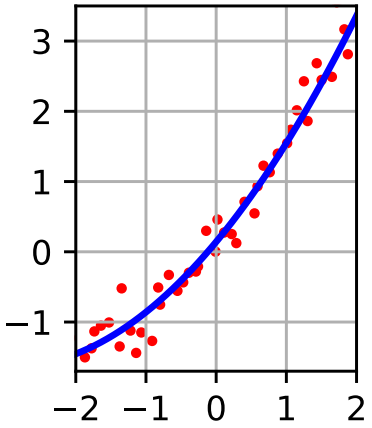
Problem: For correlated random variables X and Y, can we create a *prediction* function $f(X)$ that gives a good prediction for Y?

Similar mathematical idea for data points – curve-fitting:

Given points $(x_1, y_1), \ldots, (x_k, y_k)$ find a polynomial of degree d for these points

Some instances:

- $d = k - 1$: interpolation
- $d > k - 1$, with conditions on derivative: splines
- $d < k - 1$: approximation (minimize some error function)



Random variables are distributions though – not points...

For Random Variables

Common error function for approximating points: Least Square Error

Find f that minimizes $\sum_{i=1}^{k} \left(y_k - f(x_k)\right)^2$

For probability distribution – same idea, but weight outcomes by probability:

$$\sum_{\omega \in \Omega} P(\omega) \cdot \left(Y(\omega) - f\left(X(\omega)\right)\right)^2$$

This is just the expected square of distance between $f(X)$ and Y:

$$E[\left(Y - f(X)\right)^2]$$

Minimize f? Restrict to a class of functions (i.e., degree d polynomials).

Degree 0: Constant approximation

Simplest possible "function": use $f(X) = c$ as approximation function.

Goal: Minimize $E[(Y - f(X))^2] = E[(Y - c)^2]$ (X is ignored!)

$$\text{err}(c) = E[(Y - c)^2] = E[Y^2 - 2cY + c^2] = E[Y^2] - 2cE[Y] + c^2$$

Consider as a polynomial in c, take derivative to find ...

$$\text{err}'(c) = 2c - 2E[Y]$$

Setting $\text{err}'(c) = 0$, we get $c = E[Y]$

Double-check second derivative: $err''(c) = 2 > 0$ so this is indeed a minimum

Conclusion: The best degree 0 approximation to Y is $f(X) = E[Y]$

This makes intuitive sense, but it's good to see that the math "works."

Using knowledge of a correlated variable – two six-sided die rolls

Random variables: R_1 is value of first roll, R_2 is second, and $S = R_1 + R_2$

Given a value s for S and want the best degree-0 (constant) predictor for R_1.

What do you think this is?

Using knowledge of a correlated variable – two six-sided die rolls

Random variables: R_1 is value of first roll, R_2 is second, and $S = R_1 + R_2$

Given a value s for S and want the best degree-0 (constant) predictor for R_1.

What do you think this is? Unsurprisingly, it is $E[R_1|S = s]$

s	$E[R_1 \mid S = s]$
2	1
3	1.5
4	2
5	2.5
6	3
7	3.5
8	4
9	4.5
10	5
11	5.5
12	6

Recall from earlier:

$$E[R_1|S = 7] = \frac{7}{2}$$
$$E[R_1|S = 2] = 1$$

So if you know $S = 2$, estimate that $R_1 = 1$,

if you know $S = 7$, estimate $R_1 = \frac{7}{2}$,

if you know $S = 4$, estimate $R_1 = 2$, ...

Down-side: Need a table of conditional expectations (not necessarily structured)

Degree 1 with mean-zero random variables – Part 1

Assume $\boldsymbol{E[X] = 0}$ and $\boldsymbol{E[Y] = 0}$. Let's use $f(X) = mX + b$ to predict Y.

Goal: Minimize $E[(Y - mX - b)^2]$

Focus on b first:

$$\text{err}(m, b) = E[(Y - mX - b)^2] = E[(Y - mX)^2 - 2(Y - mX)b + b^2]$$

$$= E[(Y - mX)^2] - 2bE[Y] + 2mbE[X] + b^2$$

These are 0 since we assume mean-zero for X and Y

$$= E[(Y - mX)^2] + b^2$$

Partial derivative wrt b: $\frac{\partial}{\partial b} \text{err}(m, b) = 2b$

Set to $0 \implies b = 0$ ... so to minimize error, don't shift!

Degree 1 with mean-zero random variables – Part 2

Assume $\boldsymbol{E[X] = 0}$ and $\boldsymbol{E[Y] = 0}$ and using $\boldsymbol{b = 0}$: Let's use $f(X) = mX$ to predict Y.

Goal: Minimize $E[(Y - mX)^2]$

$$\text{err}(m) = E[(Y - mX)^2] = E[Y^2 - 2mXY + m^2X^2]$$

$$= E[Y^2] - 2mE[XY] + m^2E[X^2]$$

Take derivative: $\text{err}'(m) = 2mE[X^2] - 2E[XY]$

... and set to zero: $m = \frac{E[XY]}{E[X^2]}$ (and verify with 2nd derivative that this is a min)

Since $E[X] = E[Y] = 0$, $E[XY] = \text{cov}(X, Y)$ and $E[X^2] = \text{Var}(X)$,

... so best predictor for Y is $f(X) = \frac{\text{cov}(X,Y)}{\text{Var}(X)} X$

Degree 1 with arbitrary random variables

Let's use $f(X) = mX + b$ to predict Y.

Step 1: Shift X and Y to mean-zero random variables

$$\tilde{X} = X - E[X] \quad \text{and} \quad \tilde{Y} = Y - E[Y]$$

Note: $\text{Var}(\tilde{X}) = \text{Var}(X)$ $\quad \text{Var}(\tilde{Y}) = \text{Var}(Y)$ $\quad \text{cov}(\tilde{X}, \tilde{Y}) = \text{cov}(X, Y)$

Step 2: Best predictor for $\tilde{Y}$ using $\tilde{X}$

Predictor for $\tilde{Y}$: $\frac{\text{cov}(\tilde{X},\tilde{Y})}{\text{Var}(\tilde{X})}\tilde{X} = \frac{\text{cov}(X,Y)}{\text{Var}(X)}\tilde{X} = \frac{\text{cov}(X,Y)}{\text{Var}(X)}(X - E[X])$

Step 3: To estimate Y (instead of $\tilde{Y}$), add $E[Y]$.

$$f(X) = \frac{\text{cov}(X,Y)}{\text{Var}(X)}(X - E[X]) + E[Y]$$

Called the Linear Least Squares Estimate (LLSE) of Y given X

Degree 1 with arbitrary random variables – sanity check!

LLSE estimator: $f(X) = \frac{\text{cov}(X,Y)}{\text{Var}(X)}\ (X - E[X]) + E[Y]$

Think through some cases:

- If X and Y are not correlated:
  Then $\text{cov}(X,Y) = 0$, so $f(X) = E[Y]$
  In other words: ignore X and just use the expected value of Y

- If $X = Y$:
  Then $\text{cov}(X,Y) = \text{Var}(X)$, and $E[X] = E[Y]$ so $f(X) = X$

Both make sense!

Summary

Correlation

- Covariance sign is meaningful – magnitude has weird units
- Correlation normalizes to range $-1 \,..\, +1$: $\mathrm{Corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$
  - $\mathrm{Corr}(X,Y) = +1$ perfect positive correlation
  - $\mathrm{Corr}(X,Y) = -1$ perfect negative correlation

Conditional expectation

- Restrict attention to part of sample space ("given" condition)
- Probabilities and expectations work the same here: Notation $E[A|B]$

Prediction

- Idea: To predict random variable Y, use correlated random variable X
- Error function: Squared error
- Best linear approximator has slope $\frac{\mathrm{cov}(X,Y)}{\mathrm{Var}(X)}$