Predicting the value of Y

- **Y** is a RV with known pmf or pdf
- Problem: Predict (or estimate) what value of **Y** will be observed on the next trial
- What value should we predict?
- What is a good prediction?
- We need to specify some criterion that determines what is a good/reasonable estimate
- Else any estimate is just as good as any other estimate

Minimize probability of error – I

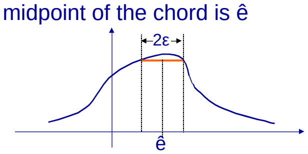
- ê denotes our estimate of the value of **Y**
- ê is a number that we get to choose
- Minimum-probability-of-error criterion: choose α so as to minimize P{**Y** ≠ ê}
- If **Y** is a discrete random variable, then P{**Y** ≠ ê} = 1 – P{**Y** = ê} = 1 – p_Y(ê)
- Choose ê to be the location of the maximum of the pmf $p_Y(u)$
- Our estimate is wrong with probability 1 – p_Y(ê)

Minimize probability of error – II

- If **Y** is a continuous random variable, then P{**Y** ≠ ê} = 1 – P{**Y** = ê} = 1 no matter what we number we choose as ê
- Alternative: choose ê to minimize
  P{|**Y**–ê| > ε}
  for some suitable (small) choice of ε
- P{|**Y**–ê| > ε} = 1 – P{|**Y**–ê| ≤ ε} and so we want to maximize P{|**Y**–ê| ≤ ε}
- P{|**Y**–ê| ≤ ε} = P{ê – ε ≤ **Y** ≤ ê + ε} = F_Y(ê + ε) – F_Y(ê – ε)

Minimize probability of error – III

- P{|**Y**–ê| ≤ ε} = F_Y(ê + ε) – F_Y(ê – ε) has derivative f_Y(ê + ε) – f_Y(ê – ε) = 0 if ê is chosen such that f_Y(ê + ε) = f_Y(ê – ε)
- Graphically, find a horizontal “chord” of length 2ε under the “peak” of the pdf: the midpoint of the chord is ê



What’s apple pie à la mode?

- In the limit as ε → 0, ê approaches the location of the maximum value of the pdf
- For both continuous and discrete RVs, we get the location of the maximum of the pdf or the pmf
- The location of the maximum of the pdf or the pmf is called the mode of the pdf/pmf
- It is the value of **Y** that has the “maximum probability” of occuring
- Mode = most fashionable or most frequent

Large errors are worse than small

- If our estimate is ê, then we make an estimation error of **Y** – ê
- Cost of making this error is |**Y** – ê|
- Large estimation errors cost us more than small estimation errors
- The absolute estimation error is |**Y** – ê|
- Average absolute estimation error (or average cost) is E[|**Y** – ê|]
- E[|**Y** – ê|] is minimized if ê is chosen to be the median value of **Y**

What's the median of Y?

- The median of **Y** is usually defined to be the number m such that $F_Y(m) = 1/2$
- Not satisfactory: It is possible $F_Y(u) \neq 1/2$ for any choice of u, e.g. if **Y** is a Bernoulli RV with parameter $p \neq 1/2$, or it might be that $F_Y(u) = 1/2$ for all u in some interval
- Median = m such that $P\{\mathbf{Y} \leq m\} \geq 1/2$ and $P\{\mathbf{Y} \geq m\} \geq 1/2$ for discrete RV
- Median = m such that $F_Y(m) = 1/2$ for continuous RV (or midpoint of interval)

The median minimizes E[|Y–α|]

- $E[|\mathbf{Y}-\hat{e}|] = \int_{v=-\infty}^{\hat{e}} (\hat{e}-v) f_Y(v)dv + \int_{v=\hat{e}}^{\infty} (v-\hat{e}) f_Y(v)dv$
- The derivative of $E[|\mathbf{Y} - \hat{e}|]$ with respect to $\hat{e}$ is given by
- $\frac{d}{d\hat{e}} E[|\mathbf{Y}-\hat{e}|] = \int_{v=-\infty}^{\hat{e}} f_Y(v)dv + \int_{v=\hat{e}}^{\infty} -f_Y(v)dv$
  $= F_Y(\hat{e}) - [1 - F_Y(\hat{e})] = 0$ when $F_Y(\hat{e}) = 1/2$
- Exercises: compute the derivative yourself and verify that you get a minimum, not a maximum, on setting derivative = 0

Large errors are worser than small

- The absolute estimation error is $|\mathbf{Y} - \hat{e}|$
- Large estimation errors cost us more if the cost function is $(\mathbf{Y} - \hat{e})^2$
- But, since $x^2 < |x|$ for $|x| < 1$, small estimation errors are underpenalized in comparison to the cost function $|\mathbf{Y} - \hat{e}|$
- The mean-square estimation error (more simply, mean-square error) is $E[(\mathbf{Y} - \hat{e})^2]$
- $E[(\mathbf{Y} - \hat{e})^2]$ is minimized if $\hat{e}$ is chosen to be the mean value of **Y**

The mean minimizes E[(Y– ê)²]

- Let $E[\mathbf{Y}] = \mu$
- $E[(\mathbf{Y} - \hat{e})^2] = E[(\mathbf{Y} - \mu + \mu - \hat{e})^2]$
  $= E[(\mathbf{Y} - \mu)^2 + 2(\mathbf{Y} - \mu)(\mu - \hat{e}) + (\mu - \hat{e})^2]$
  $= \mathrm{var}(\mathbf{Y}) + 2(\mu - \hat{e})E[\mathbf{Y} - \mu] + (\mu - \hat{e})^2$
  $= \mathrm{var}(\mathbf{Y}) + (\mu - \hat{e})^2 > \mathrm{var}(\mathbf{Y})$ if $\hat{e} \neq \mu$
- Choosing $\hat{e} = \mu$ minimizes the mean-square error of the estimate
- $\hat{e} = \mu$ is called minimum- (or least-) mean-square error (MMSE or LMSE) estimate
- The minimum mean-square error is var(**Y**)

Upping the ante

- Let **X** and **Y** denote random variables with known joint distribution
- When the experiment is performed, we observe random point (**X**, **Y**) in the plane
- Now suppose that the value of **X** becomes known to us, but not the value of **Y**
- What is the MMSE estimate of **Y**?
- Ostrich's answer: Ignore **X** and, as before, estimate the value of **Y** as $\hat{e} = \mu_Y = E[\mathbf{Y}]$
- The minimum mean-square error is var(**Y**)

Using all the available information

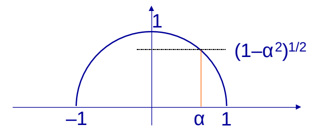
- **X** and **Y**: RVs with known joint distribution
- We know that $\mathbf{X} = \alpha$ on this trial and want to find MMSE estimate of **Y** on this trial
- Knowing that $\mathbf{X} = \alpha$ on this trial, **Y** has conditional pdf $f_{Y|X}(v|\alpha) = f_{X,Y}(\alpha, v)/f_X(\alpha)$ or conditional pmf $p_{Y|X}(v|\alpha) = p_{X,Y}(\alpha, v)/p_X(\alpha)$
- MMSE estimate of **Y** is the mean of this conditional pdf/pmf
- $\hat{e} = E[\mathbf{Y}|\mathbf{X}=\alpha] = \int v \cdot f_{Y|X}(v|\alpha)$ or $\sum v \cdot p_{Y|X}(v|\alpha)$

Using all the available information

- **X** and **Y**: RVs with known joint distribution
- Given that **X** = α on this trial, the MMSE estimate of **Y** on this trial is ê = mean of the conditional pdf/pmf of **Y** given **X** = α
- Mean-square error on this trial = conditional variance of **Y** given **X** = α
- $\text{var}(Y|X=\alpha) = \int (v-\hat{e})^2 \cdot f_{Y|X}(v|\alpha)\, dv$ or $\sum (v-\hat{e})^2 \cdot p_{Y|X}(v|\alpha)$
- Remember that the mean is ê = E[**Y**|**X**=α]

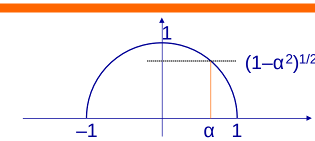
An example

- The random point (**X**, **Y**) is uniformly distributed on the semicircle shown



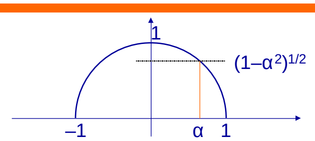
- Joint pdf has value 2/π on the semicircle
- Conditional pdf of **Y** given that **X** = α is a uniform density on $[0, (1-\alpha^2)^{1/2}]$

Conditional mean and variance



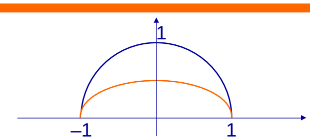
- Conditional pdf of **Y** given that **X** = α is a uniform density on $[0, (1-\alpha^2)^{1/2}]$
- Hence, $\hat{e} = E[Y|X=\alpha] = (1/2)\cdot(1-\alpha^2)^{1/2}$ and this estimate achieves the least possible MSE of $\text{var}(Y|X=\alpha) = (1-\alpha^2)/12$

Does it all make any sense?



- $\hat{e} = E[Y|X=\alpha] = (1/2)\cdot(1-\alpha^2)^{1/2}$
- MSE is $\text{var}(Y|X=\alpha) = (1-\alpha^2)/12$
- If |α| is nearly 1, the MSE is small
- If |α| is nearly 0, the MSE is large
- Makes sense to me!

The regression curve of Y on X



- ê = E[**Y**|**X**=α] as a function of α is a curve called the regression curve of **Y** on **X**
- Graph of $(1/2)\cdot(1-\alpha^2)^{1/2}$ is a half-ellipse
- Given **X** value, the MMSE estimator of **Y** can be "read off" from the regression curve

MMSE estimate and MSE are RVs

- Given that **X** = α, the MMSE estimate of **Y** is ê = E[**Y**|**X** = α] and it achieves the least possible MSE of var(**Y**|**X** = α)
- If **X** had taken on value β, then MMSE estimate would have had a different value E[**Y**|**X** = β] and different MSE var(**Y**|**X** = β)
- The MMSE estimate ê and the MSE that it achieves both are functions of **X**
- The MMSE estimate and its MSE are random variables that are functions of **X**

The RVs E[Y|X] and var(Y|X)

- The MMSE estimate of Y and the MSE achieved by this estimate are random variables that are functions of **X**
- E[**Y**|**X**] = g(**X**) is a random variable whose value is E[**Y**|**X** = α] whenever **X** = α
- var(**Y**|**X**) = h(**X**) is a random variable with value var(**Y**|**X** = α) whenever **X** = α
- E[**Y**|**X**] and var(**Y**|**X**) look like constants but they are not: they are functions g(**X**) and h(**X**) of **X**. Note:They are not functions of **Y**

Professor, this is so confusing…

- You told us that expected values are constants and now you tell us that E[**Y**|**X**] and var(**Y**|**X**) are not constants but functions g(**X**), h(**X**) of **X**? and not of **Y**?
- “All the randomness due to **Y**” was “integrated out” when we found the mean and variance of **Y**
- But, we have not taken expectations with respect to **X** as yet…
- E[**Y**|**X**] and var(**Y**|**X**) are random variables

Sometimes you feel like a nut …

- Given that **X** = α, the MMSE estimate of **Y** is ê = E[**Y**|**X** = α] and it achieves the least possible MSE of var(**Y**|**X** = α)
- If **X** had taken on value β, then the MMSE estimate would have a different value E[**Y**|**X** = β] and different MSE var(**Y**|**X** = β)
- $P\{\alpha - \varepsilon/2 \le \mathbf{X} \le \alpha + \varepsilon/2\} \approx f_X(\alpha)\bullet\varepsilon$
- $P\{\beta - \varepsilon/2 \le \mathbf{X} \le \beta + \varepsilon/2\} \approx f_X(\beta)\bullet\varepsilon$
- What is the average MMSE estimate of **Y**?

… sometimes you don’t …

- On average, the MMSE estimator has value $\int E[\mathbf{Y}|\mathbf{X}=\alpha]\bullet f_X(\alpha)\,d\alpha = \int g(\alpha)\bullet f_X(\alpha)\,d\alpha$
- But, $g(\alpha) = E[\mathbf{Y}|\mathbf{X}=\alpha] = \int v\bullet f_{Y|X}(v|\alpha)\,dv$
- $E[g(\mathbf{X})] = E[E[\mathbf{Y}|\mathbf{X}]] = \int g(\alpha)\bullet f_X(\alpha)\,d\alpha$

$= \int\int v\bullet f_{Y|X}(v|\alpha)\,dv\,f_X(\alpha)\,d\alpha$

$= \int\int v\bullet[f_{X,Y}(\alpha, v)/f_X(\alpha)]\bullet f_X(\alpha)\,dv\,d\alpha$

$= \int\int v\bullet f_{X,Y}(\alpha, v)\,dv\,d\alpha = E[\mathbf{Y}]$!!

Let’s make sure we understand…

- Given that **X**=α, the MMSE estimate of **Y** is $E[\mathbf{Y}|\mathbf{X}=\alpha] = \int v\bullet f_{Y|X}(v|\alpha)\,dv$
- As we repeat the experiment over and over, we observe different values of **X** and obtain different MMSE estimates of **Y**
- The statement E[g(**X**)] = E[E[**Y**|**X**]] = E[**Y**] is saying that the average of our MMSE estimates of **Y** is just E[**Y**], the MMSE estimate if we didn’t know the value of **X**

… how we are improving things

- We can always ignore the information that **X** = α, use the MMSE estimate E[**Y**] of **Y**, and achieve MSE var(**Y**)
- Using the estimate E[**Y**|**X** = α] instead of E[**Y**] achieves MSE var(**Y**|**X** = α)
- Average of MMSE estimates E[**Y**|**X** = α] is E[**Y**], the MMSE estimate we would have been using if we didn’t know the value of **X**
- But, average MSE is smaller than var(**Y**)

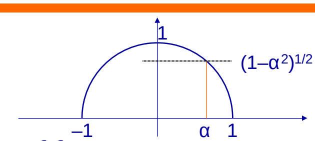
What's the average MSE?

- $\mathrm{var}(\mathbf{Y}|\mathbf{X} = \alpha) = E[\mathbf{Y}^2|\mathbf{X} = \alpha] - (E[\mathbf{Y}|\mathbf{X} = \alpha])^2$
- $\mathrm{var}(\mathbf{Y}|\mathbf{X}) = E[\mathbf{Y}^2|\mathbf{X}] - (E[\mathbf{Y}|\mathbf{X}])^2$
- Recall that for any random variable **Z**, $\mathrm{var}(\mathbf{Z}) = E[\mathbf{Z}^2] - (E[\mathbf{Z}])^2$ and $E[E[\mathbf{Z}|\mathbf{X}]] = E[\mathbf{Z}]$
- $E[\mathrm{var}(\mathbf{Y}|\mathbf{X})] = E[E[\mathbf{Y}^2|\mathbf{X}]] - E[(g(\mathbf{X}))^2]$
  $= E[\mathbf{Y}^2] - E[(g(\mathbf{X}))^2]$
  $= \{E[\mathbf{Y}^2] - (E[\mathbf{Y}])^2\} - \{E[(g(\mathbf{X}))^2] - (E[\mathbf{Y}])^2\}$
  $= \mathrm{var}(\mathbf{Y}) - \mathrm{var}(g(\mathbf{X})) = \mathrm{var}(\mathbf{Y}) - \mathrm{var}(E[\mathbf{Y}|\mathbf{X}])$
  because $g(\mathbf{X}) = E[\mathbf{Y}|\mathbf{X}]$ is a RV whose mean is $E[E[\mathbf{Y}|\mathbf{X}]] = E[\mathbf{Y}]$

The average MSE is smaller

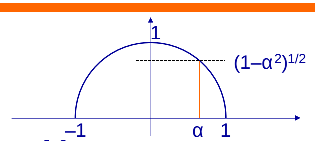
- For each observed value α of **X**, the MMSE estimator $\hat{e} = E[\mathbf{Y}|\mathbf{X} = \alpha]$ achieves the MSE $\mathrm{var}(\mathbf{Y}|\mathbf{X} = \alpha)$
- On average, the MSE achieved is $E[\mathrm{var}(\mathbf{Y}|\mathbf{X})] = \mathrm{var}(\mathbf{Y}) - \mathrm{var}(E[\mathbf{Y}|\mathbf{X}]) \le \mathrm{var}(\mathbf{Y})$
- The average of the MMSE estimates is $E[\mathbf{Y}]$, the MMSE estimate if we did not know the value of **X**
- The average MSE is smaller than $\mathrm{var}(\mathbf{Y})$, the MSE of the ostrich's estimate $E[\mathbf{Y}]$

Example: E[E[Y|X] = E[Y]



- $E[\mathbf{Y}] = \int\int v \cdot f_{\mathbf{X},\mathbf{Y}}(u, v)\, dv\, du = 4/3\pi$ (switch to polar coordinates…)
- $E[\mathbf{Y}|\mathbf{X}] = (1/2) \cdot (1-\mathbf{X}^2)^{1/2}$
- $f_{\mathbf{X}}(u) = (2/\pi) \cdot (1-u^2)^{1/2}, |u| \le 1$
- $E[E[\mathbf{Y}|\mathbf{X}]] = \int (1/\pi) \cdot (1-u^2)\, du = 4/3\pi$

Example: E[var(Y|X)]



- $E[\mathbf{Y}^2] = \int\int v^2 \cdot f_{\mathbf{X},\mathbf{Y}}(u, v)\, dv\, du = 0.25$ giving $\mathrm{var}(\mathbf{Y}) = 0.25 - (4/3\pi)^2 \approx 0.070\ldots$
- $\mathrm{var}(\mathbf{Y}|\mathbf{X}) = (1-\mathbf{X}^2)/12$
- $f_{\mathbf{X}}(u) = (2/\pi) \cdot (1-u^2)^{1/2}, |u| \le 1$
- $E[\mathrm{var}(\mathbf{Y}|\mathbf{X})] = \int (1-u^2)^{3/2}/6\pi\, du = 1/16 \approx .063$

The conditional variance formula

- The conditional variance formula says that $\mathrm{var}(\mathbf{Y}) = E[\mathrm{var}(\mathbf{Y}|\mathbf{X})] + \mathrm{var}(E[\mathbf{Y}|\mathbf{X}])$
- We have seen this in the context of discrete random variables already, where we noted that the unconditional variance of a random variable is the mean of the conditional variance plus the variance of the conditional mean
- See Lecture 15, Slides 15-18

Saying it again and again…

- If we don't know the value of **X**, the MMSE estimate is $E[\mathbf{Y}]$ with an MSE of $\mathrm{var}(\mathbf{Y})$
- If we know the value of **X**, the MMSE estimate of **Y** is $E[\mathbf{Y}|\mathbf{X}]$ and it achieves a MSE of $\mathrm{var}(\mathbf{Y}|\mathbf{X})$
- Average estimate = $E[\mathbf{Y}]$ in either case
- $\mathrm{var}(\mathbf{Y}) = E[\mathrm{var}(\mathbf{Y}|\mathbf{X})] + \mathrm{var}(E[\mathbf{Y}|\mathbf{X}])$
- By changing our estimate of **Y** based on knowledge of **X**, we reduce the MSE from $\mathrm{var}(\mathbf{Y})$ to $E[\mathrm{var}(\mathbf{Y}|\mathbf{X})]$

When does all this not work?

- Suppose **X** and **Y** are independent RVs
- Knowing **X** tell us nothing about **Y**
- If **X** and **Y** are independent RVs, the conditional pdf/pmf is the same as the unconditional pdf/pmf
- E[**Y**|**X**] = constant = E[**Y**] does not depend on value of **X**; var(E[**Y**|**X**]) = 0
- MMSE estimate of Y is E[**Y**] and has an MSE of var(**Y**), just as if we did not know the value of **X**

Linear MMSE estimation — I

- Suppose that we wish to estimate **Y** as a linear function of the observation **X**
- The linear MMSE estimate of **Y** is a**X** + b where a and b are chosen to minimize the mean-square error $E[(\mathbf{Y} - a\mathbf{X} - b)^2]$
- Let **Z** = **Y** – a**X** – b
- $E[(\mathbf{Y} - a\mathbf{X} - b)^2] = E[\mathbf{Z}^2] = \mathrm{var}(\mathbf{Z}) + (E[\mathbf{Z}])^2$

$= \mathrm{var}(\mathbf{Y}) + a^2\mathrm{var}(\mathbf{X}) - 2a\bullet\mathrm{cov}(\mathbf{X},\mathbf{Y}) + (E[\mathbf{Z}])^2$

- What should we choose a and b to be?

Linear MMSE estimation — II

- $E[(\mathbf{Y} - a\mathbf{X} - b)^2] = E[\mathbf{Z}^2] = \mathrm{var}(\mathbf{Z}) + (E[\mathbf{Z}])^2$

$= \mathrm{var}(\mathbf{Y}) + a^2\mathrm{var}(\mathbf{X}) - 2a\bullet\mathrm{cov}(\mathbf{X},\mathbf{Y}) + (E[\mathbf{Z}])^2$

- Choose a and b to make E[**Z**] = 0
- A quadratic attains its maximum/minimum at the average of its roots
- $a = \mathrm{cov}(\mathbf{X},\mathbf{Y})/\mathrm{var}(\mathbf{X}) = \rho\sigma_X\sigma_Y/(\sigma_X)^2 = \rho\sigma_Y/\sigma_X$
- $E[\mathbf{Z}] = E[\mathbf{Y} - a\mathbf{X} - b] = 0$ if $b = \mu_Y - a\bullet \mu_X = \mu_Y - \mu_X\bullet\rho\sigma_Y/\sigma_X$
- $a\mathbf{X} + b = (\rho\sigma_Y/\sigma_X)\mathbf{X} + (\mu_Y - \mu_X\bullet\rho\sigma_Y/\sigma_X)$

Linear MMSE estimation — III

- The linear MMSE estimate L is best remembered in the “symmetric” form

$$(L - \mu_Y)/\sigma_Y = \rho\bullet(\mathbf{X} - \mu_X)/\sigma_X$$

- The MSE achieved is $E[(\mathbf{Y} - a\mathbf{X} - b)^2]$

$= \mathrm{var}(\mathbf{Y}) + a^2\mathrm{var}(\mathbf{X}) - 2a\bullet\mathrm{cov}(\mathbf{X},\mathbf{Y})$

$= (\sigma_Y)^2 + (\rho\sigma_Y/\sigma_X)^2(\sigma_X)^2 - 2\ \rho\sigma_Y/\sigma_X\bullet\rho\sigma_Y\sigma_X$

$= (\sigma_Y)^2(1 - \rho^2) \le (\sigma_Y)^2$

- Linear MMSE estimate also reduces the MSE: from var(**Y**) to $\mathrm{var}(\mathbf{Y})\bullet(1 - \rho^2)$

When does it not work?

- If **X** and **Y** are uncorrelated RVs, then ρ = 0 and

$$(L - \mu_Y)/\sigma_Y = \rho\bullet(\mathbf{X} - \mu_X)/\sigma_X$$

reduces to $L = \mu_Y = E[\mathbf{Y}]$ just as if we did not know the value of **X**

- The MSE achieved is $(\sigma_Y)^2(1 - \rho^2) = (\sigma_Y)^2 = \mathrm{var}(\mathbf{Y})$ just as if we did not know the value of **X**
- Reminder: Independent RVs are uncorrelated RVs

When does it work perfectly?

- If **X** and **Y** are perfectly correlated RVs, then ρ = ±1 and

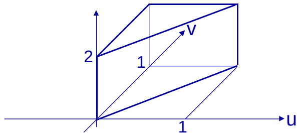
$$(L - \mu_Y)/\sigma_Y = \rho\bullet(\mathbf{X} - \mu_X)/\sigma_X$$

reduces to $(L - \mu_Y)/\sigma_Y = \pm1\bullet(\mathbf{X} - \mu_X)/\sigma_X$ and the MSE achieved is $(\sigma_Y)^2(1 - \rho^2) = 0$

- The mean-square error in estimating **Y** is 0 which means that we have predicted the value of **Y** exactly!
- The probability mass lies on the straight line $(v - \mu_Y)/\sigma_Y = \pm1\bullet(u - \mu_X)/\sigma_X$

Linear MMSE versus MMSE

- In general, the linear MMSE estimate a**X**+b has a higher MSE than the (usually nonlinear) MMSE estimate E[**Y**|**X**]
- Sometimes, both estimates are the same
- E[**Y**|**X** = α] = (1 + α)/2 = linear estimate!



Gaussian MMSE = Linear MMSE

- If **X** and **Y** are jointly Gaussian RVs, then the conditional pdf of **Y** given **X** = α is a Gaussian pdf with mean

  $\mu_Y + (\rho\sigma_Y/\sigma_X)(\alpha - \mu_X)$

  and variance $(\sigma_Y)^2(1 - \rho^2)$
- Hence, E[**Y**|**X** = α] = $\mu_Y + (\rho\sigma_Y/\sigma_X)(\alpha - \mu_X)$ is the same as the linear MMSE and has MSE var(**Y**|**X** = α) = $(\sigma_Y)^2(1 - \rho^2)$ which is the same as that of the linear MMSE

Least-squares straight line fitting

- If **X** and **Y** are discrete RVs taking on n values (u_i, v_i) with equal probability, then the linear MMSE estimate of **Y** given **X** is the “least-squares straight-line fit” to the “data”
- The linear MMSE estimate of **X** given **Y** is also a “least-squares straight-line fit” to the “data”
- In one case, v is the dependent variable; while in the other case, u is …

Summary

- The MMSE estimate E[**Y**|**X** = α] achieves MSE of var(**Y**|**X** = α)
- The average MSE is E[var(**Y**|**X**)] ≤ var(**Y**) which is the MSE of the estimate E[**Y**] that does not use knowledge of the value of **X**
- Linear MMSE estimate

  $\mu_Y + (\rho\sigma_Y/\sigma_X)(\alpha - \mu_X)$

  has MSE $(\sigma_Y)^2(1 - \rho^2)$
- For jointly Gaussian RVs, MMSE estimate = linear MMSE estimate

