Lecture 11: Linear Regression

wbg231

December 2022

1 introduction

- again the prof stopped posting videos so i am just to type notes from the lecture

2 regression

regression

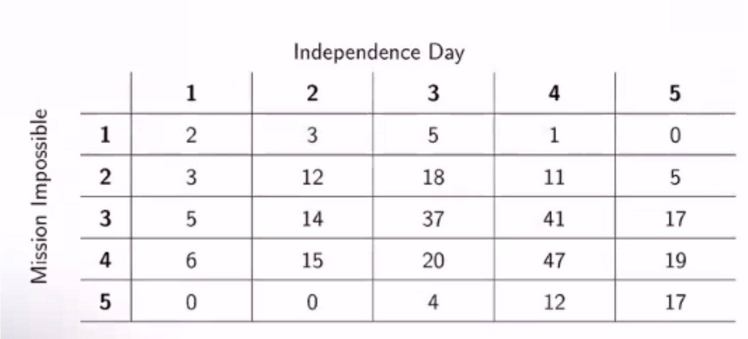
- the goal is given an input feature $\tilde{X} \in \mathbb{R}^d$ estimate a response $\tilde{y}$
- so we need an estimator $h(x)$ which approximates $\tilde{y}$ when $\tilde{x} = x$
- how do we evaluate an approximation?
- the evauluation depends in general on the application but we can use mean square error in general

  $$MSE(y, h(x)) = E[(\tilde{y} - h(\tilde{x}))^2]$$

  that is the squared average loss of our estimation

example

- suppose we have the following data



		Independence Day				
		1	2	3	4	5
Mission Impossible	1	2	3	5	1	0
	2	3	12	18	11	5
	3	5	14	37	41	17
	4	6	15	20	47	19
	5	0	0	4	12	17

- so we want to get the rating for mission imposable what is the raining for mission imposable
- there is one dimension in this task
- so what is the minium mean squared error estimator
- the minium mean squared error estimator is the conditional mean of $\tilde{y}$ given $\tilde{x}$

are we done here

- given we know that the conditional mean will optimize mean squared error, are we done?
- no we are not done, the curse of dimensionality that would require us to estimate parameters for every possible combinations of features. so this requires a lot data very quickly.
- so there will be combinations of data we have not seen unless the dimensionality of the data is very small

3 MMSE linear regression

linear regression

- overall the mmse is the conditional mean, however this is a non-linear estimator
- we can simplify this problem by narrowing our hypothesis space to just be linear functions (or an affine function of) the input features

$$\tilde{y} \approx \ell(\tilde{x}) := \Sigma_{i=1}^{d} \beta[i]\tilde{x}[i] + \alpha = \beta^t \tilde{x} + \alpha$$

- so we are overcoming the curse of dimensionality by reducing our hypothesis space and only need to learn d parameters
- so our goal is to find the linear mean squared error estimator (LMMSE) that is

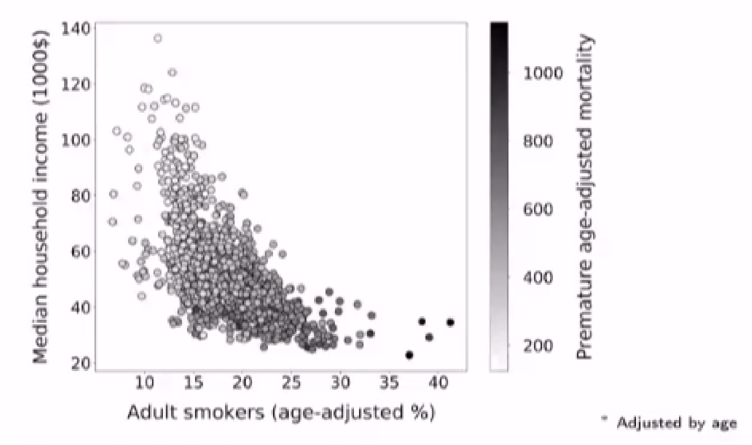
$$(\beta', \alpha') = argmin_{\beta,\alpha} E[(\tilde{y} - \beta^t \tilde{x} - \alpha)^2]$$

  and

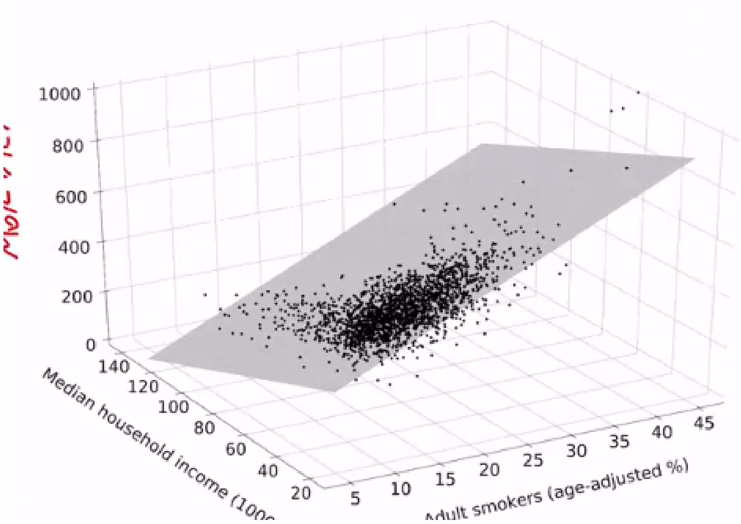
$$\ell'(\tilde{x}) = \beta'^t \tilde{x} + \alpha'$$

counties in the united states

- this is going to be our running example



- we can see there is interfeature corelation
- we can see that in income is negatively associated with mortality
- we can see that smoking is associated positively with mortality
- keep in mind graphically that a linear model in $\tilde{x} \in \mathbb{R}^d$ and $\tilde{y} \in \mathbb{R}$ can be understood as fitting a hyperplane in $\mathbb{R}^d$ to the total space $\mathbb{R}^{d+1}$



- what is this LMMSE estimator

zero mean feautre and response

- assume that the mean of all features and response is zero then we can just think of our LMMSE as

$$\beta_{mmse} = argmin_{\beta} E[(\tilde{y} - \beta^t \tilde{x})^2]$$

- we can put back in the mean by adding the offset or bias term α

geometry

review of geometric intuition

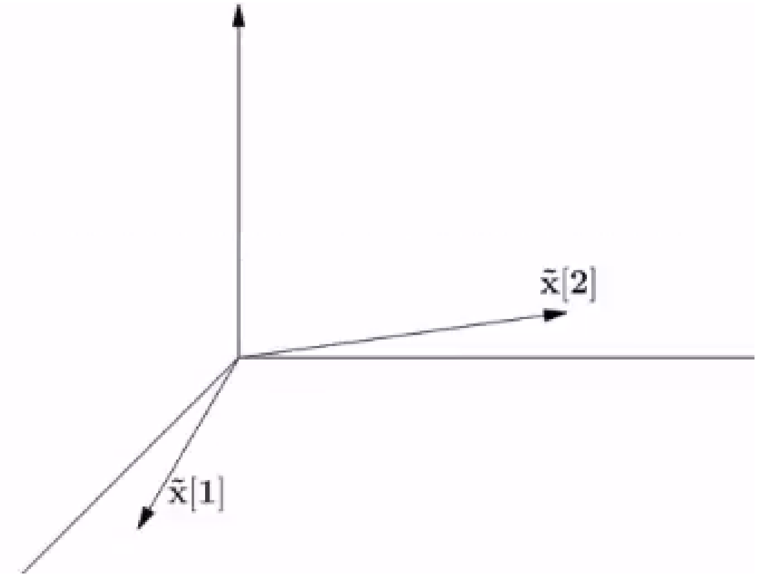
- zero mean random variables can be thought of as a vector space
- we did this when we defined corelation the inner product is the covariance that is for two vectors x,y in this space

$$< x, y >= cov(x, y)$$

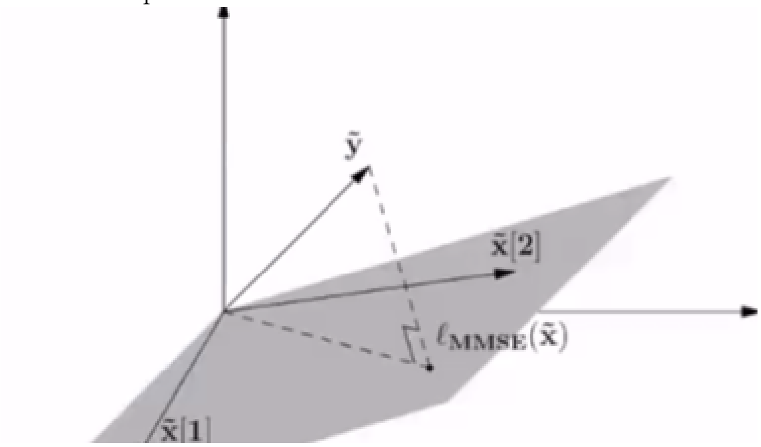
- the squared norm is the variance that is for a vector x in this space

$$\sqrt{< x, x >}^2 = \sqrt{cov(x, x)}^2 = \sqrt{var(x)}^2 = var(x)$$

- let our true response be y and let our prediction be $\tilde{y}$ so our goal is to minimize mean squared error $(y - \tilde{y})^2 = var(y - \hat{y})^2 = ||y - \hat{y}||^2$ that is the squared distance
- here is our space



- our estimator is a linear combination of our features $\hat{y} = \beta^T\tilde{x}$ that implies we only can make predictions along the plane these two features span ie $\hat{y} = span(x[1], x[2])$ which is a plane
- then we want to take our response which is a third random variable and find the point taht is closest on the plane which is the projection of y onto the plane spanned by x[1],x[2]
- here is our space



- so we know by defection $\forall\beta \in \mathbb{R}^2$ the vectors $\beta^t\tilde{x}$ are the vectors on our plane and further as ℓ_{mmse} is projected on the plane we know that $< \beta^t\tilde{x}, (\tilde{y} - \ell_{mmse}(\tilde{x})) >= 0$
- that is the residual of our model must be orthogonal to the plane containing all linear combinations of our input features (this makes sense because if this were not the case we could get a combination of our input features that would further reduce the residual of our estimator)
- this can be expressed as

$$0 =< \beta^t\tilde{x}, \tilde{y} - \ell_{mmse}(\tilde{x}) >=< \beta^t\tilde{x}, \tilde{y} - \tilde{x}^t\beta_{mmse} >= cov(\beta^t\tilde{x}, \tilde{y} - \tilde{x}^t\beta_{mmse})$$
$$= E[\beta^t\tilde{x}\tilde{y} - \tilde{x}^t\beta_{mmse}] - E[\beta^t\tilde{x}]E[\tilde{y} - \tilde{x}^t\beta_{mmse}] = E[\beta^t\tilde{x}(\tilde{y} - \tilde{x}^t\beta_{mmse})]$$
$$= \beta^t(E[\tilde{x}\tilde{y}] - E[\tilde{x}\tilde{x}^t]\beta_{mmse})\forall\beta \in \mathbb{R}^2$$

- this condition only holds when $E[\tilde{x}\tilde{y}] = E[\tilde{x}\tilde{x}^t] \iff cov[\tilde{x}\tilde{y}] = cov[\tilde{x}, \tilde{x}^t]$

cross covariance

- the cross covariance is defined as $cov(\tilde{x}, \tilde{y}) = \Sigma_{\tilde{x}\tilde{y}}$ where

$$\Sigma_{\tilde{x},\tilde{y}} : E[ct(\tilde{x})ct(\tilde{y})] = \begin{pmatrix} cov(x\tilde{[1]}, \tilde{y}) \\ \cdots \\ cov(\tilde{x}[d], \tilde{y}) \end{pmatrix}$$

- this is the covariance between each feature and the response as a vector
- so we have $0 =< \beta^t \tilde{x}, \tilde{y} - \ell_{mmse}(\tilde{x}) >= \beta^t [cov(\tilde{x}, \tilde{y}) - cov(\tilde{x}, \tilde{x}\beta_{mmse})] = \beta^t(\Sigma_{\tilde{x}\tilde{y}} - \Sigma_{\tilde{x}}\beta_{mmse}) \iff \beta\Sigma_{\tilde{x}\tilde{y}} = \Sigma_{\tilde{x}}\beta_{mmse}$
- and finally solving this yields

$$\beta_{mmse} = \Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}}$$

- this will make the linear estimator as close as possible while staying on this plane (in this vector space of mean

  zero random vectors)

intercept term

- the way we solved for β_{mmmse} assumed that our data was mean zero, this may not be true for our data.
- the most recent and third to last pictures are planes that are more or less shifted version of one another

general case

- what if we do not have centered data then we are looking for

$$\ell_{mmse} = \beta_{mmse}^t \tilde{x} + \alpha_{mmse}$$

- and we want to optimize this thing with respect to mean squared error
- so we have

$$(\beta_{mmse}, \alpha_{mmse}) = argmin_{\alpha,\beta} E[(\tilde{y} - \beta^t \tilde{x} - \alpha)^2]$$

minium constant estimate or rv

- what is the minium constant estimate of the random varible $\tilde{a}$? well we want

$$argmin_{c \in \mathbb{R}} E[(c - \alpha)^2] = argmin E[c^2 - 2\tilde{a}c + \tilde{a}^2] = c^2 - 2cE[\tilde{a}] + E[\tilde{a}^2]$$

  then we can differentiate this wrt to c to get $2^c - 2E[\tilde{a}] \rightarrow c^* = E[\tilde{a}]$
- so yeah the best constant estimator of a random variable with respect to mean squared error is the mean

additive constant

- so we have more or less the same problem

$$\alpha^* = argmin_{\alpha\in\mathbb{R}} E[(\tilde{y} - \beta^t\tilde{x} - \alpha)^2] = E[\tilde{y}] - \beta^t E[\tilde{x}] = \mu_y - \beta^t\mu_x$$

- so we know that our alpha is more or less correcting for the mean. (it is kind of adding the portion of the mean missed by our β)

linear coefficients

- so we have a way to find the optimal α for any β

$$\alpha^* = argmin_{\alpha\in\mathbb{R}} E[(\tilde{y} - \beta^t\tilde{x} - \alpha)^2] = E[\tilde{y}] - \beta^t E[\tilde{x}] = \mu_y - \beta^t\mu_x$$

- so we have $mse(\beta, \alpha) \geq mse(\beta, \alpha^*(\beta))$ this allows us to reduce this optimization problem to 1 dimension
- so the question becomes $\beta_{mse} = argmin_\beta MSE(\beta, \alpha^*(\beta))$
- this is equivlent to centering our linear regression problem

$$MSE(\beta, \alpha^*(\beta)) = E[(\tilde{y} - \beta^t\tilde{x} - \alpha^*(\beta))^2] = E[(\tilde{y} - \beta^t\tilde{x} - \mu_y + \beta^T\mu_x)^2]$$

$$= E[((\tilde{y} - \mu_y) - \beta^t(\tilde{x} - \mu_x))^2] = E[ct(\tilde{y} - \beta^t ct(\tilde{x}))]$$

$$= E[ct(\tilde{y})] + \beta^t E[ct(\tilde{x}\tilde{x}^t)]\beta - 2\beta^t E[ct(\tilde{x})ct(\tilde{y})] = \Sigma_y^2 + \beta^t\Sigma_x\beta - 2\beta^t\Sigma_{x,y} = q(\beta)$$

- we need to check the hessian to make sure we have the right convexity

$$\nabla^2(\beta) = 2\Sigma_{\tilde{x}}$$

which is a covariance matrix and thus positive semi definite meaning our function is convex and that first order condition point is a global minimum

covariance matrix is PSD

- a matrix A is PSD if $\forall a \in \mathbb{R}^d$ we have

$$a^t A a < geq 0$$

- so for an arbitrary vector $a \in \mathbb{R}^d$ and our covariance matrix Σ_x we have

$$a^t\Sigma_x a = var(a^t\tilde{x}) \geq 0$$

by definition of variance meaning the covariance matrix is PSD

optimality

- now taking the darivative of q wrt β and setting equal to zero gives us
$$0 = 2\beta^t \Sigma_{\tilde{x}} - 2\Sigma_{x,y} \Rightarrow \beta^* = \Sigma_x^{-1}\Sigma_{x,y}$$
- and this will be a minium as the function is convex
- and this is the same as the mmmse estimator for β using geometry
- so finally we get
$$\beta_{mmse} = \Sigma_x^{-1}\Sigma_{x,y}$$
$$\alpha_{mmse} = \alpha(\beta_{mmse}) = \mu_y - \beta_{mmse}^t \mu_x$$
- and this gives us our mmse linear estimator as
$$\ell_{mmse}(\tilde{x}) = \beta_{mmse}^t \tilde{x} + \alpha_{mmse} = \Sigma_{x,y}^t \Sigma_x^{-1}(\tilde{x}) + \mu_y - \Sigma_x^{-1}\Sigma_{x,y}\mu_x = \Sigma_x^{-1}\Sigma_{x,y}(\tilde{x} - \mu_x) + \mu_{\tilde{y}}$$
- so this is equivalent to taking our feature vector $\tilde{x}$ centering it, multiplying by the inverse of it's covariance matrix, multiplying by the cross covariance with $\tilde{y}$ and adding the mean of $\tilde{y}$ to our prediction

uncorrelated features

- what if the feature of $\tilde{x}$ are uncorrelated and response are zero mean and also uncorrelated
- $$\ell_{mmse}(\tilde{x}) = \beta_{mmse}^t \tilde{x} + \alpha_{mmse} = \Sigma_{x,y}^t \Sigma_x^{-1}(\tilde{x}) + \mu_y$$
- note in this case our covariance matrix Σ_x will be diagonal since the features are uncorrelated
- $$\ell_{mmse}(\tilde{x}) = \beta_{mmse}^t \tilde{x} + \alpha_{mmse} = \Sigma_{x,y}^t \Sigma_x^{-1}(\tilde{x}) + \mu_y = \begin{pmatrix} cov(x_1, y) \\ \cdots \\ cov(x_n, y) \end{pmatrix} \begin{pmatrix} var(x_1) & \cdots & 0 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & var(x_n) \end{pmatrix}^{-1} \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix}$$
$$= \begin{pmatrix} cov(x_1, y) \\ \cdots \\ cov(x_n, y) \end{pmatrix} \begin{pmatrix} \frac{1}{var(x_1)} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \frac{1}{var(x_n)} \end{pmatrix} \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \Sigma_{i=1}^d \tilde{x}_i \frac{cov(x_i, y_i)}{var(x_i)} = \Sigma_{i=1}^d cov(x_i, y) * var(x_1)^{-1} x_i$$
$$= \Sigma_{i=1}^d \tilde{x}_i \ell_{mmse}(x[i])$$
- so geometrically we are pretty much just assign orthogonal components separably and taking the sum since the features are orthogonal
- the inverse covariance matrix accounts for the covariance between our input features

4 OLS

OLS

- we are going back to the example we looked at earlier with mortality as y and x as smoking and median house hold income
- so we have a dataset $D = ((x_1, y_1) \cdots (x_n, y_n))$
- we can basically shift from random vectors to data by just using the sample statistic equivalents
- that is

$$\ell_{mmse}(\tilde{x}) \approx \ell_{ols}(x_i) = \Sigma_{X,Y}^T \Sigma_X^{-1}(x_i - m(X)) + m(Y) = \beta_{ols}^t + \alpha_{ols}$$

- this is called ordinary least squares estimator since we can view this as ERM where our loss function is least squares (which is just sample square error)

data

- we are going back to the example we looked at earlier with mortality as y and x as smoking and median house hold income
- we can calculate the ols estimator on our data
- this is a plane bisecting our data
- we could also look at the level curves of our plane over the original data

interpreting the coefficients

- suppose we learned a model $\ell_{ols}(x_i) = 15.7\text{tobaco} + 3\text{income} + 282$
- so these coefficients are effect on the output (y) of a 1 unit change of x_i holding all other features of x constant

5 explained variance

decomposition of variance

- recall that $< \ell_{mmse}(x), y - \ell_{mmse}(x) >= 0 = cov(\ell_{mmse}(x), y - \ell_{mmse}(x))$ ie our estimator and residual are orthogonal in this vector space and thus uncorrelated

- given this we can write

$$var(\tilde{y}) = Var(\ell_{mmse}(x) + y - \ell_{mmse}(x)) = ||\ell_{mmse}(x) + y - \ell_{mmse}(x)||^2$$

$$= ||\ell_{mmse}(x)||^2 + ||y - \ell_{mmse}(x)||^2 = var(\ell_{mmse}(x)) + var(y - \ell_{mmse}(x))$$

- we know the residual is zero mean. this must be the case since, if it is not true we could have a better estimator
- we can show this quickly $E[y - \ell_{mse}(x)] = E[y - \beta_{mmse}(x) - \alpha_{mmse}] = E[y] - \beta_{mmse}E[x] - \alpha_{mmse} = E[y] - e[y] = 0$
- this allows us to write

$$MSE(y - \ell_{mse}) = E[(y - \ell_{mse})^2] = E[(y - \ell_{mse})^2 - 0] = E[(y - \ell_{mse})^2 - E[y - \ell_{mse}]] = var(y - \ell_{mse})$$

- so we can write

$$var(\tilde{y}) = Var(\ell_{mmse}(x) + y - \ell_{mmse}(x))$$

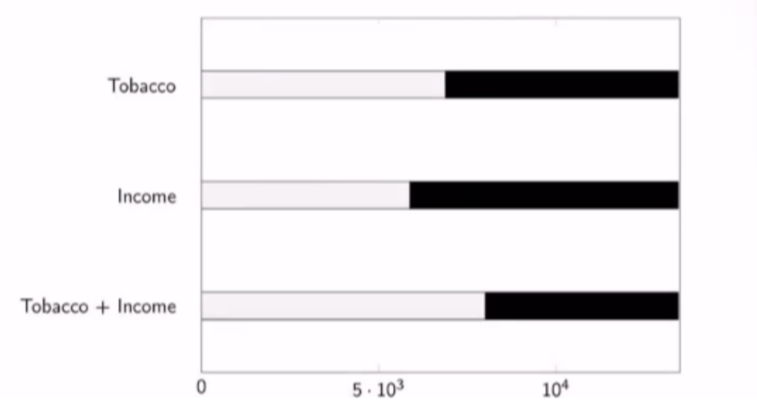
$$= var(\ell_{mmse}(x)) + var(y - \ell_{mmse}(x)) = var(\ell_{mmse}(x)) + mse(y - \ell_{mmse}(x))$$

- var(y) is fixed.
- so the larger $var(\ell_{mmse}(x))$ the lower mse so we want to max this
- we can further write it as the coefficients of determination is

$$R^2 = \frac{var(\ell_{mmse}(x))}{var(y)} = 1 - \frac{mse}{var(y)}$$

feature importance

- look at this chart



- where the black section is the mse of the model, the white section is the explained variance ie $var(\ell_{mse}(x))$

- and the total length of the bar is $var(y)$ which is fixed
- so we can compute the R^2 for each variable by deviling our bar by var(y) ie normalizing the bar
- we can see that tobacco alone counts for a lot of variance
- we can see that income accounts for a bit less than half
- then both models together account for only a bit more than each feature alone because they are highly corelate
- so we can get feature importance by looking at models built of different subsets of the features
- also note that a models r^2 can never be greater than zero and 1.

6 casual inference

interpreting features

- when we interpreting features we say that a one unit increase in one feature would lead to a certain increase in the output variable holding other features constant
- this seems to suggest causality but that is not the case necessarily
- linear regression looks at corelation between features not causality so there may be confounding factors
- this was a big deal when trying to prove that smoking caused cancer (look into fisher and how he defended smoking)

casual inference

- the question in casual inference is if we had the same person and they had smoked versus had not smoked how would the outcome change it is all about potential outcomes
- we can only really prove this if we have randomized control trials, which we can not do in the case of smoking
- what happens if we modified a single feature holding all else constant
- observed outcome $\tilde{y}$
- treatment $\tilde{t}$
- our goal is to understand the average casual effect of the treatment on the outcome

potential outcome

- potential outcomes capture what happens to the same person if the treatment had changed
- we get one observed potential outcome in the data, and the others are counter factual
- this allows us to control for spurious corelation
- a linear mode on observed data does not capture this since we are not controling

example

- have data set on unemployment and temperature in spain
- there is a corelation in the data
- so we can build a linear model
- but we can not view it is a casual since there is a confounding factors of tourism
- if we account for tourism the casual effect is more or less reduced to zero.

guinea-pig example

- we want to fatten the g pigs, and we want to know if a supplement helps them gain weight
- looking at the raw data there is a positive corelation if we give the supplement before they eat food
- we can do the same experiment but give them the supplement after they eat and there is in fact a negative corelation
- so there is a potential confounder of how much food they are before hand
- we can try to do a randomized control trial on this data

assumptions

- if we just have observed data
- we define potential outcomes that depend on a treatment and confounder

- so we are saying the distribution of $\tilde{po}_{t,c}$ =the distribution of $\tilde{po}_{t,c}$ conditioned on $\tilde{t} = 1 \cap \tilde{c}c$ that is the potential outcomes are conditionally independent of treatment and outcome
- so given the confounder there is once we take into account the confounder nothing else effects potential outcome
- and we assume that the average casual effect is linear $E[\tilde{po}_{t,c}] = \beta^* t + \gamma c$
- this can be generalized to more features and confounder
- so now we can fit a linear model that includes the confounder

features

- we assume that the features and response are centered $\tilde{x} = \begin{pmatrix} \tilde{t} \\ \tilde{c} \end{pmatrix}$ can be generalized to more features and confounders
- and we have $\Sigma_x = \begin{pmatrix} \sigma_t^2 & \sigma_{t,c} \\ \sigma_{t,c} & \sigma_c^2 \end{pmatrix}$
- what is the cross covariance $\Sigma_{x,y} = cov(y,t) = E[yt] - E[y]E[t] = E[yt] = E[\mu_{\tilde{y}\tilde{c}|\tilde{y}\tilde{c}}(\tilde{t}, \tilde{c})]$
- so this is iterated expectations which allows us to use our assumptions
- $\mu_{\tilde{y}\tilde{c}|\tilde{y}\tilde{c}}(t,x) = \int_y yt f_{y|t,c}(y|t,c)dy = \int_y yt f_{po_{t,c}|t,c}(po_{t,c}|t,c)dy = t\int_y y f_{po_{t,c}|t,c}(po_{t,c}|t,c)dy = t\int_y y f_y(y)dy = tE[\tilde{po}_{t,x}] = \beta t^2 + \gamma c * t$
- thus e $\Sigma_{t,y} = cov(y,t) = E[yt] - E[y]E[t] = E[yt] = E[\mu_{\tilde{y}\tilde{c}|\tilde{y}\tilde{c}}(\tilde{t}, \tilde{c})] = E[\beta t^2 + \gamma c * t] = \beta^* E[\tilde{t}^2] + \gamma E[\tilde{t}\tilde{c}]$
- then by the same logic
- $\Sigma_{c,y} = cov(y,c) = E[yc] - E[y]E[c] = E[yc] = E[\mu_{\tilde{y}\tilde{t}|\tilde{y}\tilde{c}}(\tilde{t}, \tilde{c})] = E[\beta c^2 + \gamma c * t] = \gamma^* E[\tilde{c}^2] + \beta E[\tilde{t}\tilde{c}]$
- so now we have the terms of our cross covariance vector
- so from here we can just use our normal formulas for $\beta_{mmse}, \alpha_{mmse}$
- so doing this out we see that

$$\beta_{mmse} = \Sigma_x^{-1}\Sigma_{x,y} = \begin{pmatrix} \sigma_t^2 & \sigma_{t,c} \\ \sigma_{t,c} & \sigma_c^2 \end{pmatrix} \begin{pmatrix} \beta^* E[\tilde{t}^2] + \gamma E[\tilde{t}\tilde{c}]^{-1} \\ \gamma^* E[\tilde{c}^2] + \beta E[\tilde{t}\tilde{c}] \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_t^2 & \sigma_{t,c} \\ \sigma_{t,c} & \sigma_c^2 \end{pmatrix}^{-1} \begin{pmatrix} \beta^* \sigma_t^2 + \gamma \sigma_{t,c} \\ \gamma^* \sigma_c^2 + \beta \sigma_{t,c} \end{pmatrix} = \begin{pmatrix} \beta^* \\ \gamma \end{pmatrix}$$

- under the assumptions we end up getting $\beta_{mmse} = \begin{pmatrix} \beta^* \\ \gamma \end{pmatrix}$ = that is under these assumptions we are able to understand the true casual effect between our feature and effect (x,y) is β

- then we can see that if the assumptions are right this will model the casual effect correctly, where as if we do not control for causality we will be modeling the spurious corelation not the true effect.