

## Lecture 2a: Model building I

### Index

#### Page

Predictors (X variables).....	2
Hierarchical dummy variables.....	5
Continuous predictors.....	7
Detecting confounding (VER 13.5).....	10
Detecting and modelling interaction.....	11
Causal interpretation (VER 14.7).....	15
Dataset: daisy2red - Stata code.....	16

● Home work for Friday:

★ Exercise 2 in "Linear Regression Exercises"

## Predictors (X variables)

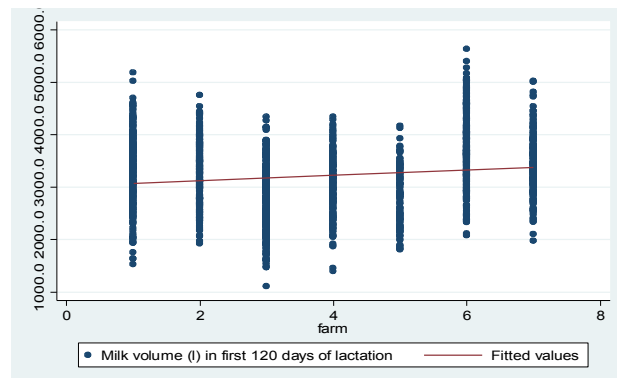
- categorical
  - ★ nominal / ordinal - must be recoded
    - ➔ indicator
    - ➔ hierarchical
- continuous
  - ★ scaling
  - ★ assumption of linearity

## Categorical predictors

- predictors with more than two levels should not be used as numeric

```
. table farm, c(mean milk120)
```

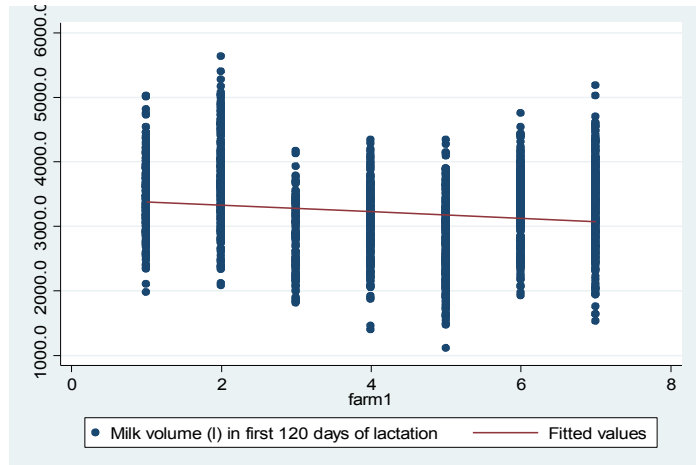
farm	mean(milk120)
1	3357.8
2	3279.3
3	2778.6
4	3032.5
5	2838.2
6	3730.6
7	3435.8



milk120	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
farm	50.90872	8.848643	5.75	0.000	33.552 68.26543
_cons	3026.143	37.27521	81.18	0.000	2953.028 3099.259

```
list farm milk120 farm1
```

	farm	milk120	farm1
1.	1	3357.8	7
2.	2	3279.3	6
3.	3	2778.6	5
4.	4	3032.5	4
5.	5	2838.2	3
6.	6	3730.6	2
7.	7	3435.8	1



milk120	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
farm1	-50.90872	8.848643	-5.75	0.000	-68.26543 -33.552
_cons	3433.413	41.84204	82.06	0.000	3351.339 3515.487

## Indicator variables

- convert nominal or ordinal variables to a set of dichotomous variables
- one level is referent (or baseline) level

herd	hrd1	hrd2	hrd3
1	1	0	0
2	0	1	0
3	0	0	1

- choice of referent level
  - ★ biological sense
  - ★ reasonable sample size
  - ★ ease of interpretation
  - ★ Stata by default selects first category
    - ➔ see *fvvarlist*
    - ➔ use indicator syntax
      - eg. regress milk120 **i.parity**
  
- all in / all out
  
- changing coding - no effect on overall model fit
  - ★ same  $R^2$  etc.

● Example - Parity as an ordinal variable

★ convert parity to 4 level categorical variable (parity\_c4)

★ regular coding

parity	parity_c4	parity_c4_0	parity_c4_1	parity_c4_2	parity_c4_3
1	0	1	0	0	0
2	1	0	1	0	0
3	2	0	0	1	0
4	3	0	0	0	1
5	3	0	0	0	1
6	3	0	0	0	1
7	3	0	0	0	1

★ *regress milk120 parity\_c4 (indicator variables)*

	Avg milk120	Regular coding	Hierarchical coding
parity_c4_0	2639.7		
parity_c4_1	3347.9	708.2	
parity_c4_2	3429.5	789.8	
parity_c4_3	3471.4	831.8	
intercept		2639.7	

## Hierarchical dummy variables

- ★ measure effect moving up one level
- ★ only applicable to ordinal variables

disease	dz0	dz1	dz2
negative	1	0	0
mild	1	1	0
severe	1	1	1

- ★ hierarchical coding

parity	parity_h0	parity_h1	parity_h2	parity_h3
1	1	0	0	0
2	1	1	0	0
3	1	1	1	0
4+	1	1	1	1

- ★ `regress milk120 parity_h1 parity_h2 parity_h3`

	Avg milk120	Regular coding	Hierarchical coding
parity_c4_0	2639.7		
parity_c4_1	3347.9	708.2	708.2
parity_c4_2	3429.5	789.8	81.6
parity_c4_3	3471.4	831.8	41.9
intercept		2639.7	2639.7

## Continuous predictors

### Improving “interpretability” of X variables

- scaling X variables

- ★ limited range of plausible values

- ➔ effects only the constant (not the coefficient for the variable - slope)

- ➔ subtract min. plausible value (e.g. parity)

```
. regress milk120 parity
...output omitted
```

milk120	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parity	178.347	11.01266	16.19	0.000	156.7455	199.9484
_cons	<b>2727.08</b>	34.33991	79.41	0.000	2659.722	2794.438

```
. gen parity_1=parity-1
```

```
. reg milk120 parity_1
....output omitted
```

milk120	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parity_1	178.347	11.01266	16.19	0.000	156.7455	199.9484
_cons	<b>2905.427</b>	25.23474	115.14	0.000	2855.928	2954.925

- ★ subtract mean (centring)

- ➔ helps with quadratic and interaction terms

```
. reg milk120 c.herd_size##c.herd_size, vsquish
```

Source	SS	df	MS	Number of obs = 1536			
Model	94747175.1	2	47373587.5	F( 2, 1533)	= 111.15		
Residual	653393017	1533	426218.537	Prob > F	= 0.0000		
Total	748140192	1535	487387.748	R-squared	= 0.1266		
				Adj R-squared	= 0.1255		
				Root MSE	= 652.85		

milk120	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
herd_size	28.73126	1.993023	14.42	0.000	24.82192	32.6406
c.herd_size#c.herd_size	-.0608255	.0041101	-14.80	0.000	-.0688875	-.0527634
_cons	66.06488	231.8877	0.28	0.776	-388.7858	520.9155

```
. reg milk120 c.hrdsz_ctr##c.hrdsz_ctr, vsquish
```

Source	SS	df	MS				
Model	94747175.1	2	47373587.5	Number of obs =	1536		
Residual	653393017	1533	426218.537	F( 2, 1533) =	111.15		
Total	748140192	1535	487387.748	Prob > F =	0.0000		
				R-squared =	0.1266		
				Adj R-squared =	0.1255		
				Root MSE =	652.85		

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
milk120						
hrdsz_ctr	-1.803116	.2847073	-6.33	0.000	-2.361573	-1.244659
c.hrdsz_ctr#c.hrdsz_ctr	-.0608255	.0041101	-14.80	0.000	-.0688875	-.0527634
_cons	3445.547	22.80494	151.09	0.000	3400.815	3490.279

★ scale of measurement (eg grams or kg)

➔ avoid very small regression coefficients

● Example: herd size (mean=251; min=125; max=333)

```
. reg milk120 herd_size
```

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
milk120						
herd_size	-.49048	.2891242	-1.70	0.090	-1.057601	.0766405
_cons	3338.164	74.69789	44.69	0.000	3191.643	3484.685

```
. gen herdsz_100=herd_size/100 /*rescale herd_size so coef are larger*/
```

```
. reg milk120 herdsz_100
```

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
milk120						
herdsz_100	-49.04796	28.91242	-1.70	0.090	-105.76	7.664098
_cons	3338.164	74.69789	44.69	0.000	3191.643	3484.685

## Linearity

- assumption about nature of relationship between X and Y

★ Note: the following are all discussed in more detail under Model Building (Chapter 15)

## Detecting non-linearity – in final model

- plot residuals vs fitted values
  - ★ simultaneous evaluation of all predictors
- plot of residuals vs predictor

## Detecting non-linearity – before / during model building

- smoothed scatter plot of outcome vs predictor
- explore polynomial functions of X (see below)
- transformation of X
- categorization of predictor
  - ★ indicator dummy variable
  - ★ hierarchical dummy variable
  - ★ compare categorical and linear variables

# Detecting confounding (VER 13.5)

## Vaginal Discharge

Retained placenta

WPC

### ● change in coefficient

```
. reg wpc vag_disch          /* vag_disch adds 10 days, P=0.04 */
```

wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
vag_disch	11.99647	5.846716	2.05	0.040	.5282858	23.46465
_cons	68.17426	1.334494	51.09	0.000	65.55669	70.79184

```
. reg wpc rp                /* rp adds 13 days, P=0.008 */
```

wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
rp	11.7344	4.434231	2.65	0.008	3.036767	20.43203
_cons	67.68842	1.364298	49.61	0.000	65.01239	70.36446

```
. reg wpc rp vag_disch     /* vag_disch no sig. , rp is a confounder */
```

wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
rp	10.2397	4.540774	2.26	0.024	1.333087	19.14632
vag_disch	9.066942	5.9819	1.52	0.130	-2.666406	20.80029
_cons	67.35756	1.381095	48.77	0.000	64.64857	70.06654

## Detecting and modelling interaction

- cross-product term
  - ★ eg. `retpla` and `vag_disch`
  - ★ `retpla*vag_disch`
  
- interpreting coefficients (VER 14.7)
- examples 14.9, 14.10 and 14.11

## Interaction between 2 dichotomous predictors

```
. reg wpc i.vag_disch##rp
```

Source	SS	df	MS			
Model	35915.9774	3	11971.9925	Number of obs =	1574	
Residual	4152174.58	1570	2644.69719	F( 3, 1570) =	4.53	
Total	4188090.56	1573	2662.48605	Prob > F =	0.0036	
				R-squared =	0.0086	
				Adj R-squared =	0.0067	
				Root MSE =	51.427	

wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.vag_disch	.5429296	7.265382	0.07	0.940	-13.70794	14.7938
1.rp	6.339794	4.914322	1.29	0.197	-3.299531	15.97912
vag_disch#rp						
1 1	26.34867	12.77367	2.06	0.039	1.293414	51.40392
_cons	67.66861	1.387883	48.76	0.000	64.94631	70.39091

```
. margins vag_disch#rp
```

```
Adjusted predictions      Number of obs   =      1574  
Model VCE      : OLS
```

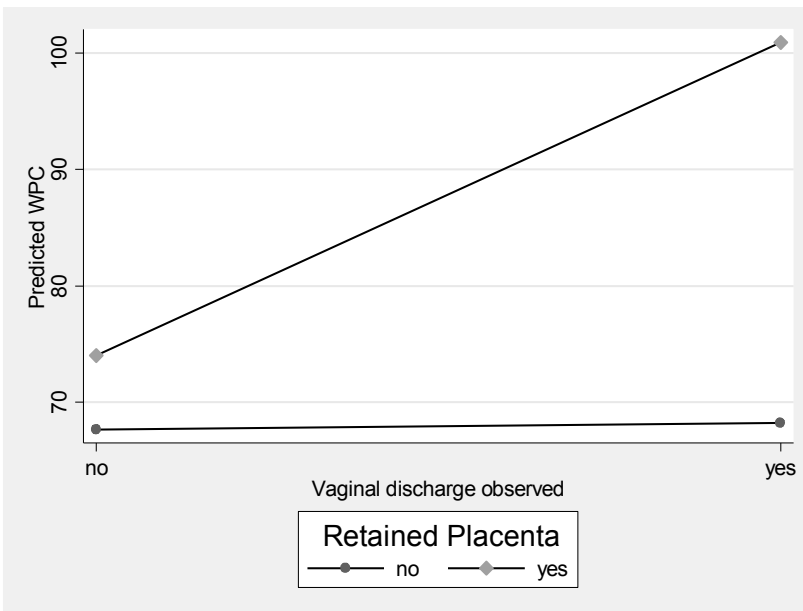
```
Expression      : Linear prediction, predict()
```

---

	Margin	Delta-method Std. Err.	t	P> t	[95% Conf. Interval]	
vag_disch#rp						
no#no	67.66861	1.387883	48.76	0.000	64.94631	70.39091
no#yes	74.0084	4.71427	15.70	0.000	64.76147	83.25533
yes#no	68.21154	7.131589	9.56	0.000	54.2231	82.19998
yes#yes	100.9	9.389173	10.75	0.000	82.48336	119.3166

---

```
. marginsplot, legend(title("Retained Placenta")) noci ytitle("Predicted WPC")  
title("")
```



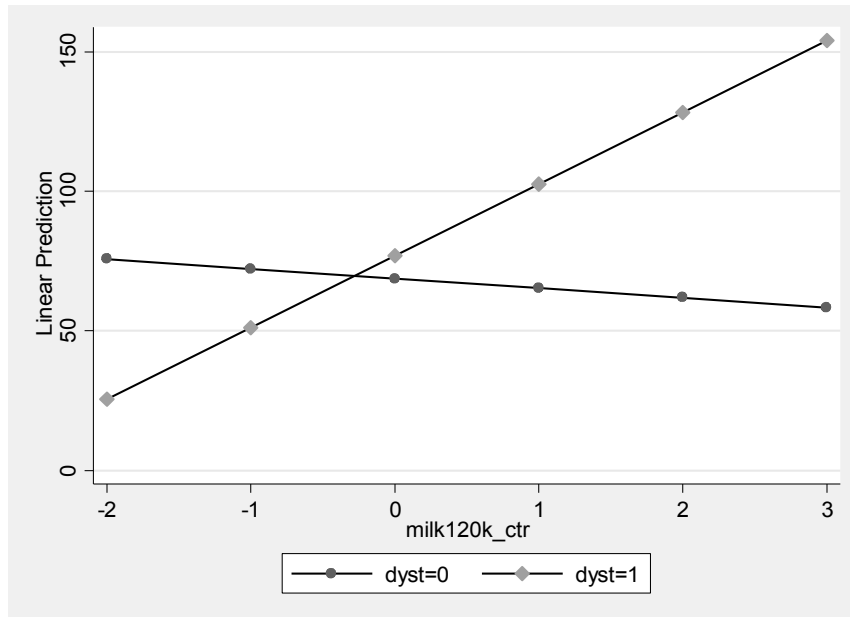
# Interaction between a dichotomous and a continuous predictor

```
. reg wpc i.dyst##c.milk120k_ctr
```

Source	SS	df	MS	Number of obs = 1536		
Model	30572.8752	3	10190.9584	F( 3, 1532)	=	3.83
Residual	4073791.78	1532	2659.13302	Prob > F	=	0.0095
Total	4104364.66	1535	2673.8532	R-squared	=	0.0074
				Adj R-squared	=	0.0055
				Root MSE	=	51.567

	wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dyst						
yes		8.20714	5.718528	1.44	0.151	-3.00983 19.42411
milk120k_ctr		-3.446531	1.928535	-1.79	0.074	-7.229379 .3363161
dyst#c.milk120k_ctr						
yes		29.14238	9.468101	3.08	0.002	10.57057 47.71419
_cons		68.75682	1.357147	50.66	0.000	66.09475 71.41888



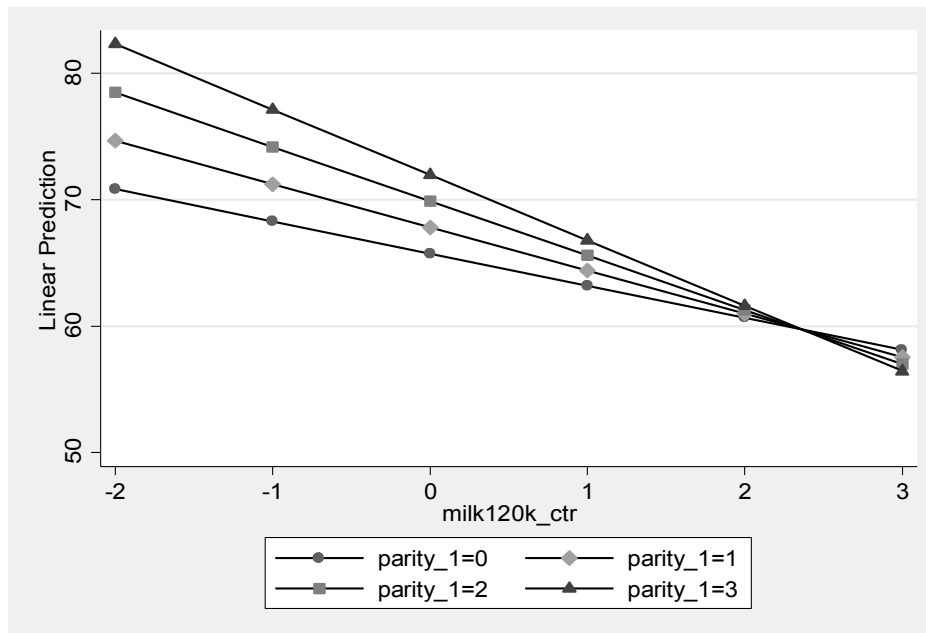
# Interaction between 2 continuous predictors

```
.reg wpc c.parity_1##c.milk120k_ctr
```

Source	SS	df	MS	Number of obs = 1536			
Model	18084.1899	3	6028.06329	F( 3, 1532)	=	2.26	
Residual	4086280.47	1532	2667.2849	Prob > F	=	0.0797	
Total	4104364.66	1535	2673.8532	R-squared	=	0.0044	
				Adj R-squared	=	0.0025	
				Root MSE	=	51.646	

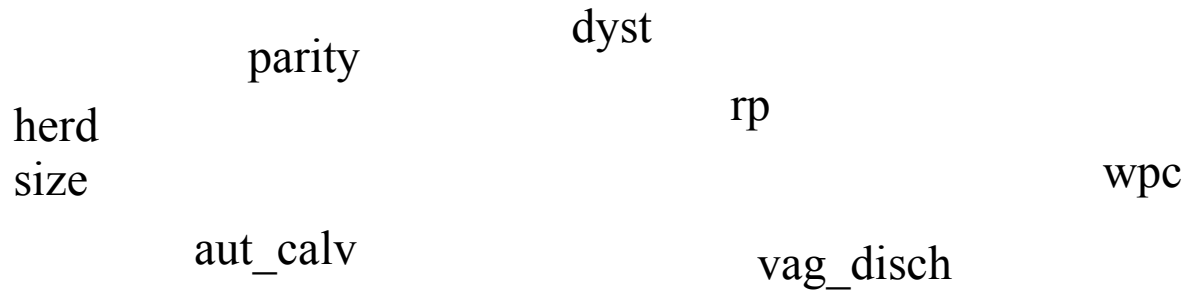
  

	wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
parity_1		2.072164	.9549532	2.17	0.030	.1990103 3.945318
milk120k_ctr		-2.542834	3.091716	-0.82	0.411	-8.607278 3.521609
c.parity_1#c.milk120k_ctr		-.8764358	1.363504	-0.64	0.520	-3.550968 1.798096
_cons		65.73655	2.207989	29.77	0.000	61.40555 70.06755



- Two way interactions between continuous predictors are difficult to interpret, and, whenever significant, should be evaluated by fitting a range of possible values for both predictors.

# Causal interpretation (VER 14.7)



```
. reg wpc hs100 hs100_2 parity_sc aut_calv twin dyst i.rp##vag_disch
```

Source	SS	df	MS	Number of obs = 1574		
Model	296062.713	9	32895.857	F( 9, 1564) = 13.22		
Residual	3892027.85	1564	2488.50885	Prob > F = 0.0000		
-----				R-squared = 0.0707		
Total	4188090.56	1573	2662.48605	Adj R-squared = 0.0653		
-----				Root MSE = 49.885		
wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hs100	-36.05708	15.05032	-2.40	0.017	-65.57801	-6.536147
hs100_2	11.13828	3.111145	3.58	0.000	5.035824	17.24073
parity_sc	1.13721	.8583103	1.32	0.185	-.5463501	2.82077
aut_calv	-8.263839	2.537751	-3.26	0.001	-13.24159	-3.286086
twin	20.68314	9.845164	2.10	0.036	1.37203	39.99425
dyst	11.70041	5.462576	2.14	0.032	.9856669	22.41516
1.rp	5.98687	4.811976	1.24	0.214	-3.451734	15.42547
1.vag_disch	1.228198	7.161395	0.17	0.864	-12.81875	15.27514
rp#vag_disch						
1 1	22.85194	12.51605	1.83	0.068	-1.698058	47.40194
_cons	84.66128	17.61671	4.81	0.000	50.10642	119.2161

- herd\_size
- parity
- aut\_calv
- twin
- dyst
- rp
- vag\_disch

## Dataset: daisy2red - Stata code

```
* VHM 812 - Winter 2014
* 2a1 - Linear regression: predictors, confounding and interaction

* change working directory and open a log file
cd "F:\javier\Presentations\TA\data_epi_teach"
* capture log close
* log using l2a_ch14_ovh.log, text replace
set more off

* open the DAISY dataset
use daisy2red.dta, clear

*meaningless use of a nominal variable
egen farm=group(herd)
label var farm farm
table farm, c(mean milk120)
tway (scatter milk120 farm) (lfit milk120 farm) , name(pos, replace)
regress milk120 farm

gen farm1=abs(farm-7)+1
preserve
collapse milk120 farm, by(farm1)
sort farm
list farm milk120 farm1
restore
tway (scatter milk120 farm1) (lfit milk120 farm1) , name(pos, replace)
regress milk120 farm1

* create a 4 level categorical variable for age
egen parity_c4=cut(parity), at(1 2 3 4 99) icodes
sort parity
br parity parity_c4
tab parity parity_c4

*INDICATOR VARIABLES
* create indicator variables "manually" (you never actually do this)
gen parity_c4_0=(parity_c4==0) if !missing(parity)
gen parity_c4_1=(parity_c4==1) if !missing(parity)
gen parity_c4_2=(parity_c4==2) if !missing(parity)
gen parity_c4_3=(parity_c4==3) if !missing(parity)

br parity parity_c4 parity_c4_0 parity_c4_1 parity_c4_2 parity_c4_3

* determine the average milk120 in each age group
format milk120 %4.2f
tab parity_c4, sum(milk120)

* regress -milk120- on the age groups
reg milk120 parity_c4_1 parity_c4_2 parity_c4_3
* or
reg milk120 i.parity_c4
* testing a group of categorical predictors
testparm i.parity_c4

* HIERARCHICAL DUMMY VARIABLES
* create hierarchical dummy variables "manually"
gen parity_h0=1 if !missing(parity)
gen parity_h1=(parity_c4>=1) if !missing(parity)
```

```

gen parity_h2=(parity_c4>=2) if !missing(parity)
gen parity_h3=(parity_c4>=3) if !missing(parity)

br parity parity_c4 parity_h0 parity_h1 parity_h2 parity_h3

* regress -milk120 on hierarchical dummy variables for parity
reg milk120 parity_h1 parity_h2 parity_h3

* SCALING X VARIABLES
* subtract min value
reg milk120 parity
gen parity_1=parity-1
reg milk120 parity_1

* centring
reg milk120 c.herd_size##c.herd_size, vsquish
margins, at(herd_size=(100(50)350))
marginsplot, name(herd_size, replace)

summ herd_size
gen hrdsz_ctr=herd_size - 251

reg milk120 c.hrdsz_ctr##c.hrdsz_ctr, vsquish
margins, at(hrdsz_ctr=(-150(25)100))
marginsplot, name(hrdsz_ctr, replace) xline(-14.82) noci

* scale of measurement
reg milk120 herd_size
capture drop herdsz_100
gen herdsz_100=herd_size/100 /*rescale herd_size so coef are larger*/
reg milk120 herdsz_100

* DETECTING CONFOUNDING
reg wpc vag_disch /* vag_disch adds 10 days, P=0.04 */
reg wpc rp /* rp adds 12 days, P=0.008 */
reg wpc rp vag_disch /* vag_disch no sig. , rp is a confounder */

tab vag_disch rp, chi2
cs vag_disch rp, or

* DETECTING INTERACTION
* 2 dichotomous variables
reg wpc vag_disch
reg wpc i.vag_disch i.rp vag_disch#rp
reg wpc vag_disch##rp //same as above
table vag_disch rp, c(mean wpc) // display wpc means by vag_dich and rp

reg wpc vag_disch#rp //total effects
margins vag_disch#rp //same as table command
marginsplot, legend(title("Retained Placenta")) noci ytitle("Predicted WPC")
title("")

* dichotomous and continuous variable
gen milk120k=milk120/1000
summ milk120k
gen milk120k_ctr=milk120k-r(mean)
reg wpc i.dyst##c.milk120k_ctr, vsquish
margins dyst, at(milk120k_ctr=(-2(1)3))
marginsplot, noci title("") legend(order(1 "dyst=0" 2 "dyst=1"))

* two continuous variables
reg wpc c.parity_1##c.milk120k_ctr

```

```

margins, at(milk120k_ctr=(-2(1)3) parity=(0(1)3))
marginsplot, noci title("")

* interaction between 2 cont. variables assumes both effect are linear
* is this true for -milk120k- ? - generate a smoothed scatterplot
tway (scatter wpc milk120k) (lowess wpc milk120k)
tway (scatter wpc parity) (lowess wpc parity)

* CAUSAL INTERPRETATION
use daisy2red.dta, clear

* model to evaluate effect of - rp and vag_disch
gen calv_mth=month(calv_dt)
gen aut_calv=(calv_mth>=2 & calv_mth<=7) if !missing(calv_mth)
gen hs100=(herd_size/100)
gen hs100_2=(herd_size/100)^2
gen parity_sc=parity-1

reg wpc hs100 hs100_2 parity_sc aut_calv twin dyst i.rp##vag_disch

predict yhat
tway scatter yhat herd_size || lowess yhat herd_size

* dyst conf for rp
reg wpc hs100 hs100_2 parity_sc aut_calv twin i.rp
reg wpc hs100 hs100_2 parity_sc aut_calv twin i.rp dyst

```