

Lecture 2a: Model building I

Index	Page
Predictors (X variables).....	2
Categorical predictors.....	2
Indicator variables.....	3
Continuous predictors.....	5
Detecting confounding (VER 13.5).....	8
Confounding and collinearity.....	11
Detecting and modeling interaction.....	12
Causal interpretation (VER 14.7).....	16
Assessing linearity.....	18
Stata Factor-Notation basics.....	19

● Exercises

★ today - questions exercise 1 linear reg.

★ home work for Tuesday

➔ exercise 2 linear regression

● Lab session Monday - small computer lab - 225N

★ bring your laptop (if you want to work on Stata)

Predictors (X variables)

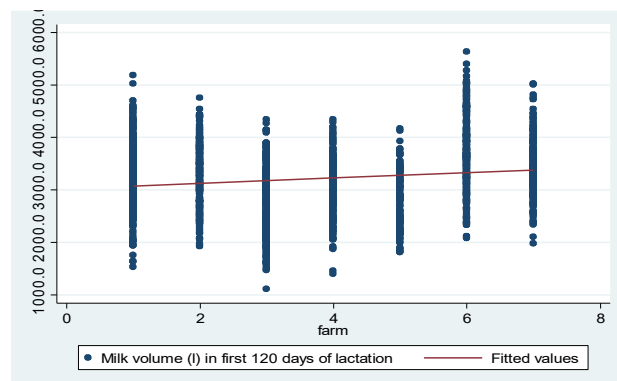
- Categorical
 - ★ nominal – values represent “levels” (no numerical meaning)
 - ★ ordinal – values represent ordered levels
 - ★ must be recoded
 - ➔ indicator or dummy variables
- Continuous (quantitative)
 - ★ scaling
 - ★ assumption of linearity

Categorical predictors

- Nominal (and sometimes ordinal) predictors with more than two levels should not be used as numeric

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

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

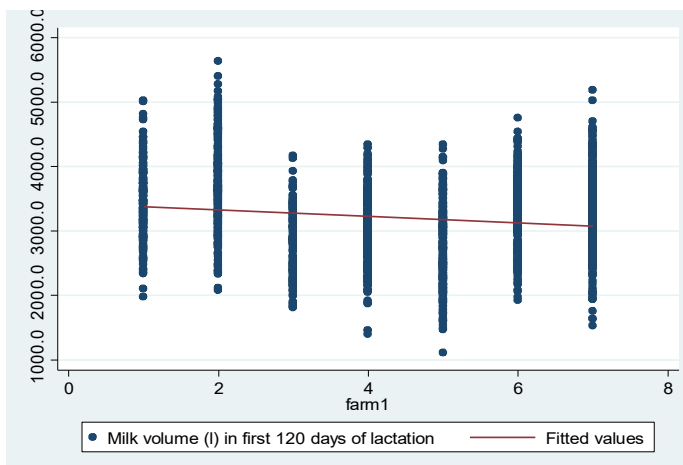


```
. regress milk120 farm
```

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_id milk120 farm1_id
```

	farm_jd	milk120	farm1_id
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 or indicator variables
 - ★ assign observations to one of two categories (usually 0 and 1)
 - ★ $j-1$ indicator variables are required in the regression model
- One level is referent (or baseline) level

Obs. #	farm_id	farm1	farm2	farm3
1	1	1	0	0
2	2	0	1	0
3	3	0	0	1

- Choice of referent level
 - ★ biological sense
 - ★ reasonable sample size
 - ★ ease of interpretation
 - ★ Stata by default selects first category (see *fvvarlist* page #8)

- All in / all out
- Changing coding – no effect on overall model fit
 - ★ same R^2 etc.
- Example – parity as an ordinal variable
 - ★ new categorical variable with 4 levels – parity_c4
 - ★ convert parity to 4 indicator variables (parity_c4)

		Indicator variables			
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 i.parity_c4

	Avg milk120	Indicator variables
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

Continuous predictors

Improving the “interpretation” 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 (eg. 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	2727.08	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	2905.427	25.23474	115.14	0.000	2855.928 2954.925

★ subtract the central value (eg. mean) (centring)

➔ helps to reduce collinearity with quadratic and interaction terms

```
. reg milk120 c.herd_size##c.herd_size
```

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		

	milkl20	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

```
. estat vce, corr -> correlation = -0.9907
```

```
. summ herd_size
```

```
. gen hrdsz_ctr=herd_size - 251
```

```
. reg milk120 c.hrdsz_ctr##c.hrdsz_ctr
```

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		

	milkl20	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
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

```
. estat vce, corr -> correlation = 0.3115
```

★ 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
```

milk120	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
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
```

milk120	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
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

Detecting confounding (VER 13.5)

● Review

★ components: Y (outcome), E (exposure) and Z (measured or unmeasured confounder)

★ criteria

➔ Z must be a risk factor of Y (in E-, because risk must not be caused by E->Y))

➔ Z must be associated with E

- cohort – start follow up period

- if constant during follow up -> look for unconditional assoc. Z->E

- case-control – in controls (represent source population if no selection bias)

➔ Z must not be the result of E or the result of Y

★ causal model

Vaginal
Discharge

Herd

WPC

● Assessment of confounding

★ when three criteria are met

★ difference between crude and adjusted effect/association changes substantially

➔ $(\text{crude-adjusted})/\text{crude}$

➔ eg. 20-30%

● Example - change in regression coefficient

```
. reg wpc i.vag_disch /* vag_disch adds 12 days, P=0.04 */
```

wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
vag_disch						
yes	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 i.herd if vag_disch==0 /*herd is associated with WPC*/
```

Source	SS	df	MS	Number of obs	=	1,492
Model	223706.812	6	37284.4686	F(6, 1485)	=	14.92
Residual	3711781.88	1,485	2499.51642	Prob > F	=	0.0000
				R-squared	=	0.0568
				Adj R-squared	=	0.0530
Total	3935488.69	1,491	2639.4961	Root MSE	=	49.995

wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
herd						
2	-7.455598	4.572715	-1.63	0.103	-16.42527	1.51407
3	12.30396	4.184586	2.94	0.003	4.095636	20.51229
4	-20.0733	4.70619	-4.27	0.000	-29.30479	-10.84181
5	-21.78125	5.489853	-3.97	0.000	-32.54994	-11.01256
106	-15.40129	4.618509	-3.33	0.001	-24.46079	-6.341796
119	-17.26021	5.05679	-3.41	0.001	-27.17942	-7.341
_cons	75.1583	3.106548	24.19	0.000	69.06461	81.25199

```
. tab herd vag_disch, chi row /* herd is associated with vag_disc*/
```

Herd Number	Vaginal discharge observed		Total
	no	yes	
3	318 98.76	4 1.24	322 100.00
...output omitted			
106	214 84.58	39 15.42	253 100.00
Total	1,492 94.79	82 5.21	1,574 100.00

Pearson chi2(6) = 74.2267 Pr = 0.000

```
. reg wpc i.vag_disch i.herd
```

Source	SS	df	MS	Number of obs	=	1,574
Model	252509.59	7	36072.7985	F(7, 1566)	=	14.35
Residual	3935580.97	1,566	2513.14238	Prob > F	=	0.0000
				R-squared	=	0.0603
				Adj R-squared	=	0.0561
Total	4188090.56	1,573	2662.48605	Root MSE	=	50.131

wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
vag_disch						
yes	17.81936	5.825177	3.06	0.002	6.393393	29.24533
herd						
2	-8.178615	4.509228	-1.81	0.070	-17.02338	.6661455
3	11.71303	4.133611	2.83	0.005	3.605039	19.82103
4	-20.74627	4.653654	-4.46	0.000	-29.87432	-11.61823
5	-22.15954	5.359351	-4.13	0.000	-32.6718	-11.64728
106	-18.18881	4.422295	-4.11	0.000	-26.86305	-9.514562
119	-17.85657	4.920293	-3.63	0.000	-27.50763	-8.205518
_cons	75.97555	3.052377	24.89	0.000	69.98837	81.96272

Confounding and collinearity

- Confounding => collinearity

- ★ example: bwt – smoking (only for example, not for model building)

- ➔ cig_2, cig_3 on bwt

- ★ two models

- ➔ 1) $E = \text{cig_2}$

- ➔ 2) $E = \text{cig_3}$

- ★ causal diagrams / criteria

- Model 1

- ★ cig_3 is not a confounder

- ★ cig_3 intervening variable

- ★ cig_3 highly correlated with cig_2 -> keep cig_2

- Model 2

- ★ cig_2 meets (partially) confounding criteria

- ★ change of coefficient?

Detecting and modeling 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##i.rp
```

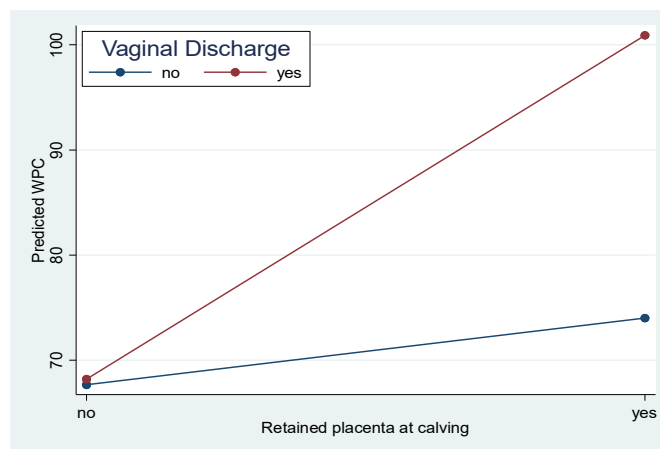
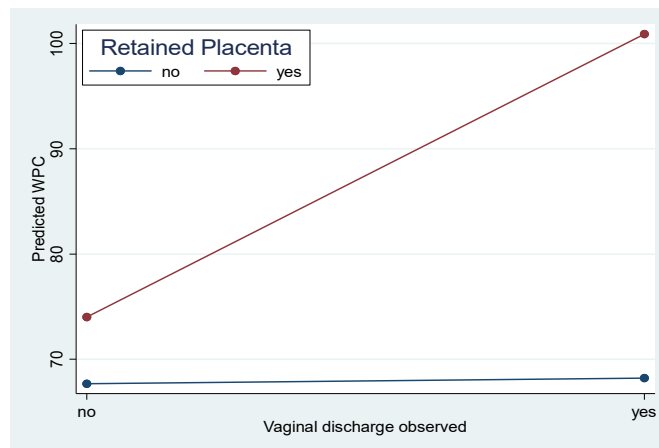
Source	SS	df	MS			
Model	35915.9774	3	11971.9925	Number of obs =	1574	
Residual	4152174.58	1570	2644.69719	F(3, 1570) =	4.53	
Total	4188090.56	1573	2662.48605	Prob > F =	0.0036	
				R-squared =	0.0086	
				Adj R-squared =	0.0067	
				Root MSE =	51.427	

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

```
. table vag_disch rp, c(mean wpc) // display wpc means by vag_dich and rp
```

Vaginal discharge observed	Retained placenta at calving	
	no	yes
no	67.66861	74.0084
yes	68.21154	100.9

● Interaction plots (margins command - later)

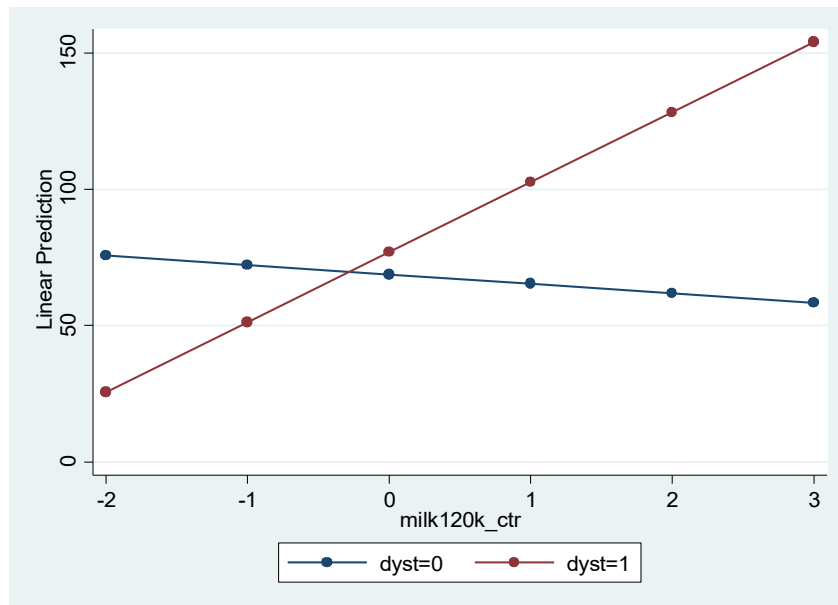


Interaction between a dichotomous and a continuous predictor

```
. 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
				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

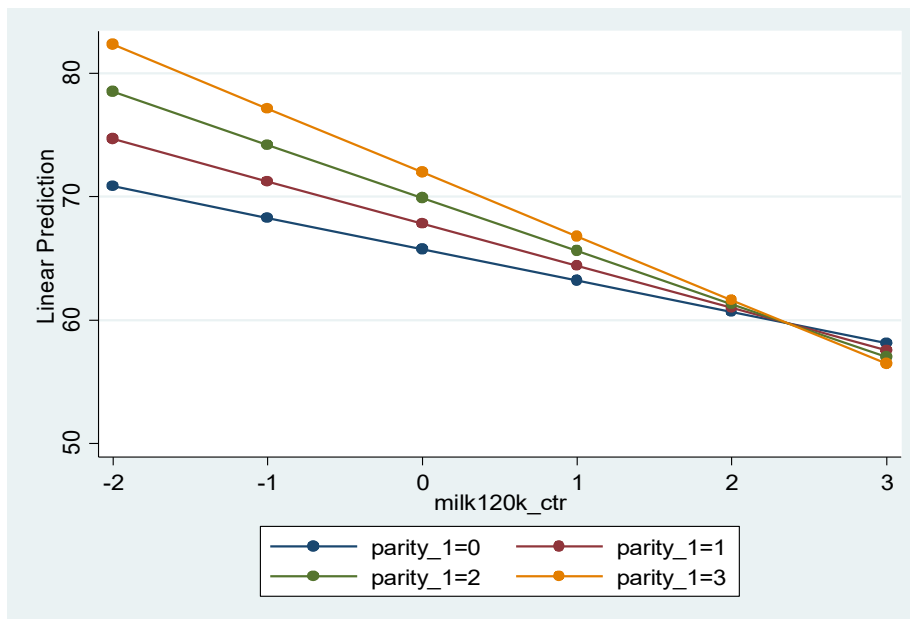


Interaction between 2 continuous predictors

```
.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	

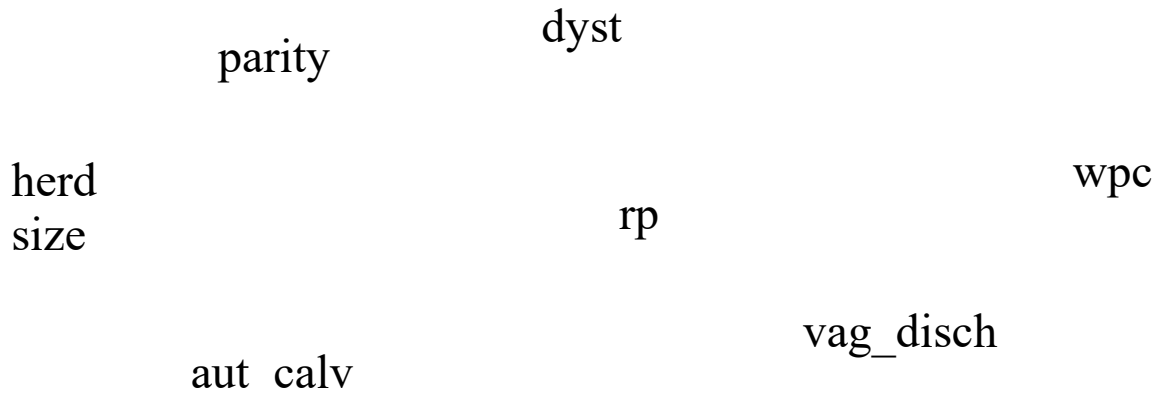
	wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
parity_1		2.072164	.9549532	2.17	0.030	.1990103 3.945318
milk120k_ctr		-2.542834	3.091716	-0.82	0.411	-8.607278 3.521609
c.parity_1#c.milk120k_ctr		-.8764358	1.363504	-0.64	0.520	-3.550968 1.798096
_cons		65.73655	2.207989	29.77	0.000	61.40555 70.06755



- 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.

★ be sure that predictions are within range of the data

Causal interpretation (VER 14.7)



● Graphical assessment confounding

★ identify potential confounder variables in complex causal models

★ steps

➔ draw causal diagram

- nodes-> variables; arrows-> causal relations
- time on horizontal axis (right most recent)
- intermediate variables

➔ delete all arrows from E->Y

➔ identify potential confounders

- unblocked paths E->Y

➔ identify collider variables

- controlling for the effect of 2 variables (eg. dyst) will create an spurious association between them (eg. between herd size and parity)

```
.reg wpc c.hs100#c.hs100 parity_1 i.aut_calv twin dyst i.rp##vag_disch
```

Source	SS	df	MS	Number of obs	=	1,574
Model	296062.681	9	32895.8535	F(9, 1564)	=	13.22
Residual	3892027.88	1,564	2488.50887	Prob > F	=	0.0000
				R-squared	=	0.0707
				Adj R-squared	=	0.0653
Total	4188090.56	1,573	2662.48605	Root MSE	=	49.885

wpc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hs100	-36.05705	15.05032	-2.40	0.017	-65.57798 -6.53612
c.hs100#c.hs100	11.13827	3.111145	3.58	0.000	5.035818 17.24073
parity_1	1.13721	.8583103	1.32	0.185	-.54635 2.82077
1.aut_calv	-8.263839	2.537751	-3.26	0.001	-13.24159 -3.286086
twin	20.68314	9.845165	2.10	0.036	1.37203 39.99425
dyst	11.70041	5.462576	2.14	0.032	.9856659 22.41516
rp					
yes	5.98687	4.811976	1.24	0.214	-3.451734 15.42547
vag_disch					
yes	1.228195	7.161395	0.17	0.864	-12.81875 15.27514
rp#vag_disch					
yes#yes	22.85194	12.51605	1.83	0.068	-1.698056 47.40194
_cons	84.66125	17.61671	4.81	0.000	50.10639 119.2161

➤ herd_size

➤ parity

➤ aut_calv

➤ twin

➤ dyst

➤ rp

➤ vag_disch

Assessing 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 (see L1a)
 - ➔ simultaneous evaluation of all predictors
 - ➔ plot of residuals vs predictor (see L1a)
- **Detecting non-linearity – before / during model building**
 - ★ smoothed scatter plot of outcome vs predictor
 - ★ explore polynomial functions of X
 - ★ transformation of X
 - ★ categorization of predictor
 - ➔ indicator dummy variable
 - ➔ compare categorical and linear variables

Stata Factor-Notation basics

- For predictors x , z and outcome y
 - ★ $i.x$ = categorical effect of x (x must be integer)
 - ★ $c.x$ = continuous effect (slope) of x (x must be numerical)
 - ★ Stata default depends on command
 - ➔ $\text{reg } y \ x = \text{reg } y \ c.x$ (default is $c.x$)
 - ➔ $\text{anova } y \ x = \text{anova } y \ i.x$ (default is $i.x$)

- Combined effects
 - ★ $c.x##c.x$ = continuous terms for x and x^2
 - ★ interaction
 - ➔ $x\#z$ = interaction $x * z$;
 - ➔ in all commands the default is $x\#z = i.x\#i.z$
 - ➔ $c.x\#c.z$ = need to use “c” for continuous variables
 - ➔ always $x##z = x \ z \ x\#z$

- Factor terms can be used in tests
 - ★ $\text{testparm } i.x$
 - ★ $\text{testparm } c.x##c.x$ (test $c.x$ and $c.x\#c.x$)
 - ★ $\text{testparm } c.x\#c.x$ (test $c.x\#c.x$)