

Lab 9**ANCOVA --- ANalysis of COVAriance**

```
* To guide the steps to an ANCOVA analysis -- to accompany lec8.pdf
* ANCOVA -- ANalysis of COVAriance
```

```
*****
```

```
* chds.dta -- child birth weight data
```

```
clear
```

```
* load the dataset
```

```
use chds.dta
```

```
* print to screen
```

```
describe
```

```
* create the mother smoking groups (0=no smoke, 1=some smoke)
```

```
generate ms_gp=0 if msmoke == 0
```

```
replace ms_gp=1 if msmoke > 0
```

```
* create the father smoking groups (0=no smoke, 1=some smoke)
```

```
generate ps_gp=0 if psmoke == 0
```

```
replace ps_gp=1 if psmoke > 0
```

```
* table of means, standard deviations, and frequencies
```

```
tabulate ms_gp ps_gp, summarize(weight)
```

Means, Standard Deviations and Frequencies of Child's birth weight (lbs)

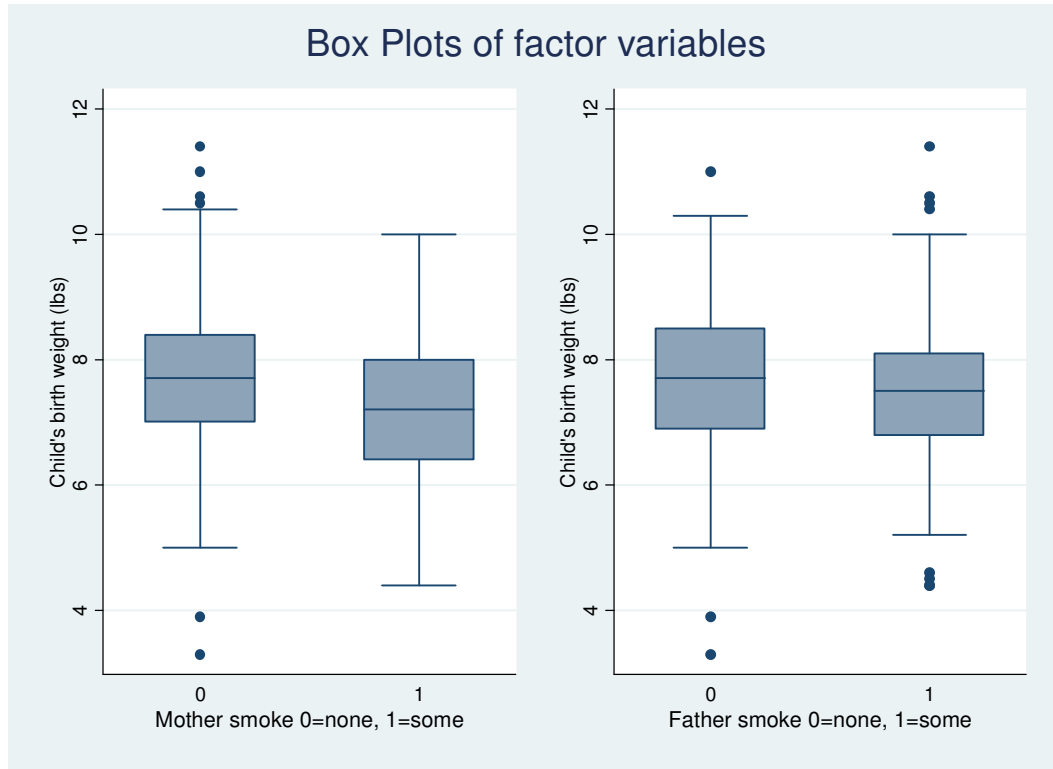
ms_gp	ps_gp		Total
	0	1	
0	7.6981132	7.7576577	7.7328084
	1.1336885	.99187192	1.0523406
	159	222	381
1	7.4727273	7.1885246	7.2408027
	1.1172808	1.0691866	1.081917
	55	244	299
Total	7.6401869	7.4596567	7.5164706
	1.1311932	1.0704853	1.0923455
	214	466	680

```
* boxplots of the two factor variables
```

```
quietly graph box weight, over(ms_gp) name(box1,replace) nodraw bltitle("Mother smoke  
0=none, 1=some")
```

```
quietly graph box weight, over(ps_gp) name(box2,replace) nodraw bltitle("Father smoke  
0=none, 1=some")
```

```
graph combine box1 box2, title(Box Plots of factor variables)
```



There doesn't appear to be much of a difference in mean birth weight for the different factor levels of mother smoking or father smoking.

If we fit a one-way ANOVA model to the mother smoking factor, we see that the mean birth weight is significantly different between mother smoking groups.

```
anova weight ms_gp
```

Source	Partial SS	df	MS	F	Prob > F
Model	40.5534298	1	40.5534298	35.72	0.0000
ms_gp	40.5534298	1	40.5534298	35.72	0.0000
Residual	769.642117	678	1.13516536		
Total	810.195546	679	1.19321877		

In the two-way ANOVA model with both mother and father smoking factors, and their interaction, the interaction is not significant, and neither is the father main effect (even if the interaction is removed).

```
anova weight ms_gp ps_gp ms_gp*ps_gp
```

Source	Partial SS	df	MS	F	Prob > F
Model	40.5534298	2	20.2767149	18.11	0.0000
ms_gp	40.5534298	1	40.5534298	35.72	0.0000
ps_gp	0.0000000	1	0.0000000	0.00	0.9833
ms_gp*ps_gp	0.0000000	1	0.0000000	0.00	0.9833
Residual	769.642117	678	1.13516536		
Total	810.195546	679	1.19321877		

Model	44.5071566	3	14.8357189	13.10	0.0000
ms_gp	19.0863111	1	19.0863111	16.85	0.0000
ps_gp	1.52601237	1	1.52601237	1.35	0.2462
ms_gp*ps_gp	3.57265918	1	3.57265918	3.15	0.0762
Residual	765.68839	676	1.13267513		
Total	810.195546	679	1.19321877		

What happens when we include the covariates in an ANCOVA model?

```

* fit the anova with covariates
* I doubly specify which are categorical and which are continuous
* if you specify one variable type for some of the variables (eg. cat),
* it will assume the other variables are the other (eg. cont)
* without any specification, anova assumes variables are categorical, cat.
anova weight ps_gp ms_gp ps_gp*ms_gp head length gest mage mheight mweight page ped
pheight, cat(ps_gp ms_gp) cont(head length gest mage mheight mweight page ped pheight)

```

Number of obs = 680 R-squared = 0.6565
 Root MSE = .645904 Adj R-squared = 0.6504

Source	Partial SS	df	MS	F	Prob > F
Model	531.928143	12	44.3273452	106.25	0.0000
ps_gp	.255533523	1	.255533523	0.61	0.4341
ms_gp	2.04644943	1	2.04644943	4.91	0.0271
ps_gp*ms_gp	.316166791	1	.316166791	0.76	0.3843
head	74.4881142	1	74.4881142	178.55	0.0000
length	126.647072	1	126.647072	303.57	0.0000
gest	19.3876485	1	19.3876485	46.47	0.0000
mage	.272245149	1	.272245149	0.65	0.4195
mheight	.761113209	1	.761113209	1.82	0.1773
mweight	3.02859461	1	3.02859461	7.26	0.0072
page	.001724666	1	.001724666	0.00	0.9488
ped	.464022037	1	.464022037	1.11	0.2920
pheight	.137995826	1	.137995826	0.33	0.5654
Residual	278.267403	667	.417192509		
Total	810.195546	679	1.19321877		

```

* same analysis using xi regress (different parameter constraints)
* specify the categorical variables with the prefix "i."
xi:regress weight i.ps_gp i.ms_gp i.ps_gp*i.ms_gp head length gest mage mheight mweight
page ped pheight

```

```

i.ps_gp      _Ips_gp_0-1      (naturally coded; _Ips_gp_0 omitted)
i.ms_gp      _Ims_gp_0-1      (naturally coded; _Ims_gp_0 omitted)
i.ps_gp*i.ms_gp  _Ips_Xms_#_#      (coded as above)

```

Source	SS	df	MS	Number of obs = 680
Model	531.928143	12	44.3273452	F(12, 667) = 106.25
Residual	278.267403	667	.417192509	Prob > F = 0.0000
Total	810.195546	679	1.19321877	R-squared = 0.6565

Adj R-squared = 0.6504
 Root MSE = .6459

weight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Ips_gp_1	-.0046965	.0687882	-0.07	0.946	-.139764 .130371
_Ims_gp_1	-.1842704	.1019	-1.81	0.071	-.3843538 .0158129
_Ips_gp_1	(dropped)				

```

    _Ims_gp_1 | (dropped)
  _Ips_Xms__~1 | .1035174 .1189114 0.87 0.384 -.1299684 .3370033
    head | .6051457 .0452882 13.36 0.000 .5162212 .6940702
    length | .5281956 .0303155 17.42 0.000 .4686702 .587721
    gest | .0974037 .0142883 6.82 0.000 .0693482 .1254592
    mage | -.0064265 .0079554 -0.81 0.419 -.0220472 .0091942
  mheight | .0162096 .012001 1.35 0.177 -.0073546 .0397738
  mweight | .0044385 .0016473 2.69 0.007 .0012039 .007673
    page | -.0004583 .0071287 -0.06 0.949 -.0144558 .0135391
    ped | .0126964 .0120387 1.05 0.292 -.0109419 .0363347
  pheight | -.0059028 .0102634 -0.58 0.565 -.0260553 .0142497
    _cons | -16.20254 1.042089 -15.55 0.000 -18.24871 -14.15637
  -----
  
```

* significance of individual parameters (same as in xi:regress parameter estimate table because only two levels)

```

testparm _Ims_gp_1
testparm _Ips_gp_1
testparm _Ips_Xms__1_1

( 1)  _Ims_gp_1 = 0
      F( 1, 667) = 3.27
      Prob > F = 0.0710

( 1)  _Ips_gp_1 = 0
      F( 1, 667) = 0.00
      Prob > F = 0.9456

( 1)  _Ips_Xms__1_1 = 0
      F( 1, 667) = 0.76
      Prob > F = 0.3843
  
```

With all the covariates in the model, it does not appear that the factor variables or their interaction are significant. This may change after we remove the nonsignificant covariates.

```

* Variable selection -- removing nonsignificant covariates and factors until all are
significant
* Because we are interested in ms_gp, retain that variable regardless of
significance
* Backward selection steps below
* Begin with full model
* remove page
* remove pheight
* remove ps_gp*ms_gp
* remove ps_gp
* remove ped
* remove mage
* remove mheight
anova weight ps_gp ms_gp ps_gp*ms_gp head length gest mage mheight mweight page ped
pheight, cat(ps_gp ms_gp) cont(head length gest mage mheight mweight page ped pheight)
anova weight ps_gp ms_gp ps_gp*ms_gp head length gest mage mheight mweight ped
pheight, cat(ps_gp ms_gp) cont(head length gest mage mheight mweight ped pheight)
anova weight ps_gp ms_gp ps_gp*ms_gp head length gest mage mheight mweight ped
, cat(ps_gp ms_gp) cont(head length gest mage mheight mweight ped )
anova weight ps_gp ms_gp head length gest mage mheight mweight ped
, cat(ps_gp ms_gp) cont(head length gest mage mheight mweight ped )
anova weight ms_gp head length gest mage mheight mweight ped
, cat( ms_gp) cont(head length gest mage mheight mweight ped )
anova weight ms_gp head length gest mage mheight mweight
, cat( ms_gp) cont(head length gest mheight mweight )
anova weight ms_gp head length gest mheight mweight
, cat( ms_gp) cont(head length gest mheight mweight )
anova weight ms_gp head length gest mweight
, cat( ms_gp) cont(head length gest mweight )
* all remaining factors and covariates significant at the 0.05 level
anova weight ms_gp head length gest mweight, cat(ms_gp) cont(head length gest mweight)
  
```

Source	Partial SS	df	MS	F	Prob > F
Model	529.559431	5	105.911886	254.37	0.0000
ms_gp	1.6179411	1	1.6179411	3.89	0.0491
head	74.2577655	1	74.2577655	178.34	0.0000
length	132.854298	1	132.854298	319.07	0.0000
gest	19.474804	1	19.474804	46.77	0.0000
mweight	5.41183929	1	5.41183929	13.00	0.0003
Residual	280.636115	674	.416374058		
Total	810.195546	679	1.19321877		

With the significant covariates in the model, the mother's smoking factor is marginally significant at the 0.05 significance level.

* same analysis using xi regress (different parameter constraints)
xi:regress weight i.ms_gp head length gest mweight

Source	SS	df	MS	Number of obs =	680
Model	529.559431	5	105.911886	F(5, 674) =	254.37
Residual	280.636115	674	.416374058	Prob > F =	0.0000
Total	810.195546	679	1.19321877	R-squared =	0.6536
				Adj R-squared =	0.6510
				Root MSE =	.64527

weight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Ims_gp_1	-.101244	.0513605	-1.97	0.049	-.2020899 -.0003981
head	.6018683	.0450684	13.35	0.000	.5133769 .6903597
length	.5293279	.0296332	17.86	0.000	.4711433 .5875124
gest	.0967248	.014143	6.84	0.000	.0689551 .1244945
mweight	.0051072	.0014166	3.61	0.000	.0023257 .0078887
_cons	-15.62452	.7174963	-21.78	0.000	-17.03332 -14.21573

* significance of individual parameters (same as in xi:regress parameter estimate table because only two levels)

```
testparm _Ims_gp_1
( 1) _Ims_gp_1 = 0
    F( 1, 674) = 3.89
    Prob > F = 0.0491
```

* need to run anova again for lincom statement below (doesn't like the regress for the ps_gp[1])
anova weight ms_gp head length gest mweight, cat(ms_gp) cont(head length gest mweight)
tabstat weight, by(ms_gp) stat(mean semean)

Summary for variables: weight
by categories of: ms_gp

ms_gp	mean	se(mean)
0	7.732808	.053913
1	7.240803	.0625689
Total	7.516471	.0418895

The unadjusted birth weight means for the levels of the mother's smoking factor are 7.73 for nonsmoking and 7.24 for smoking.

```
adjust head length gest mweight, by(ms_gp) se
```

```
-----
Dependent variable: weight      Command: anova
Covariates set to mean: head = 13.219118, length = 20.279411, gest = 39.770588, mweight = 126.89559
-----

ms_gp |          xb          stdp
-----+-----
    0 |    7.56099    (.033501)
    1 |    7.45974    (.037953)
-----

Key:  xb    = Linear Prediction
      stdp  = Standard Error
```

The adjusted birth weight means at the mean levels of the significant covariates in the model for the levels of the mother's smoking factor are 7.56 for nonsmoking and 7.46 for smoking.

```
* reproduce the adjust values
tabstat head length gest mweight
stats |      head      length      gest      mweight
-----+-----
mean | 13.21912  20.27941  39.77059  126.8956
-----
```

These means are used in the lincom statements below to replicate the adjust values above. If you wanted to estimate the mean birth weights at other levels of the covariates, you can do it in either the adjust or lincom statements. Notice, however, that the lincom statements are more complete, offering confidence intervals for the estimates.

```
lincom( _b[_cons] + _b[ms_gp[1]] + _b[head]*13.21912 + _b[length]*20.27941 +
_b[gest]*39.77059 + _b[mweight]*126.8956 )
( 1)  _cons + ms_gp[1] + 13.21912 head + 20.27941 length + 39.77059 gest + 126.8956
mweight = 0
```

```
-----
weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
(1) |    7.560989   .0335012    225.69   0.000     7.49521     7.626768
-----
```

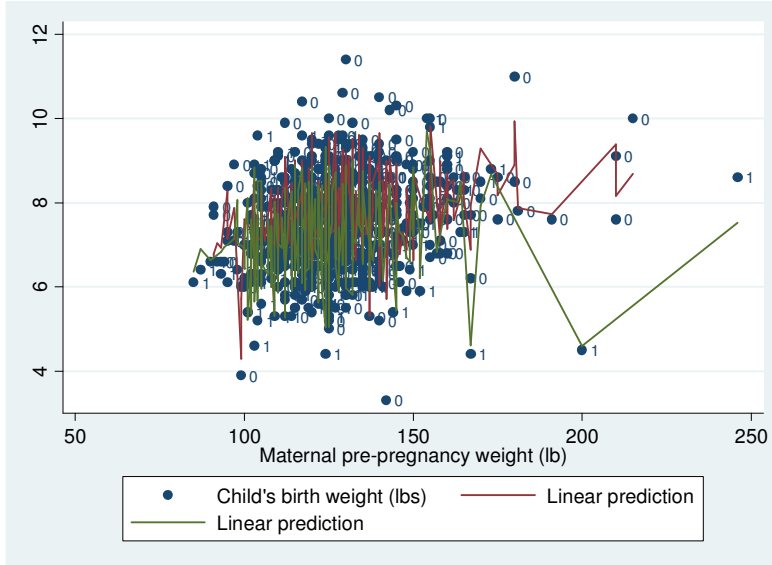
```
lincom( _b[_cons] + _b[ms_gp[2]] + _b[head]*13.21912 + _b[length]*20.27941 +
_b[gest]*39.77059 + _b[mweight]*126.8956 )
( 1)  _cons + ms_gp[2] + 13.21912 head + 20.27941 length + 39.77059 gest + 126.8956
mweight = 0
```

```
-----
weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
(1) |    7.459745   .037953    196.55   0.000     7.385225     7.534265
-----
```

```
* create fitted values, yhat
drop yhat
predict yhat, xb
* create residuals, residual
drop residual
```

```
predict residual,r

* fit lines to groups on scatter plots
* This plot looks strange (not straight best fit line) because there are many
* other covariates in the model other than mweight2.
* Because everything depends on everything else, this is not a meaningful plot.
twoway (scatter weight mweight, mlabel(ms_gp)) (line yhat mweight if ms_gp==0, sort)
(line yhat mweight if ms_gp==1, sort)
```



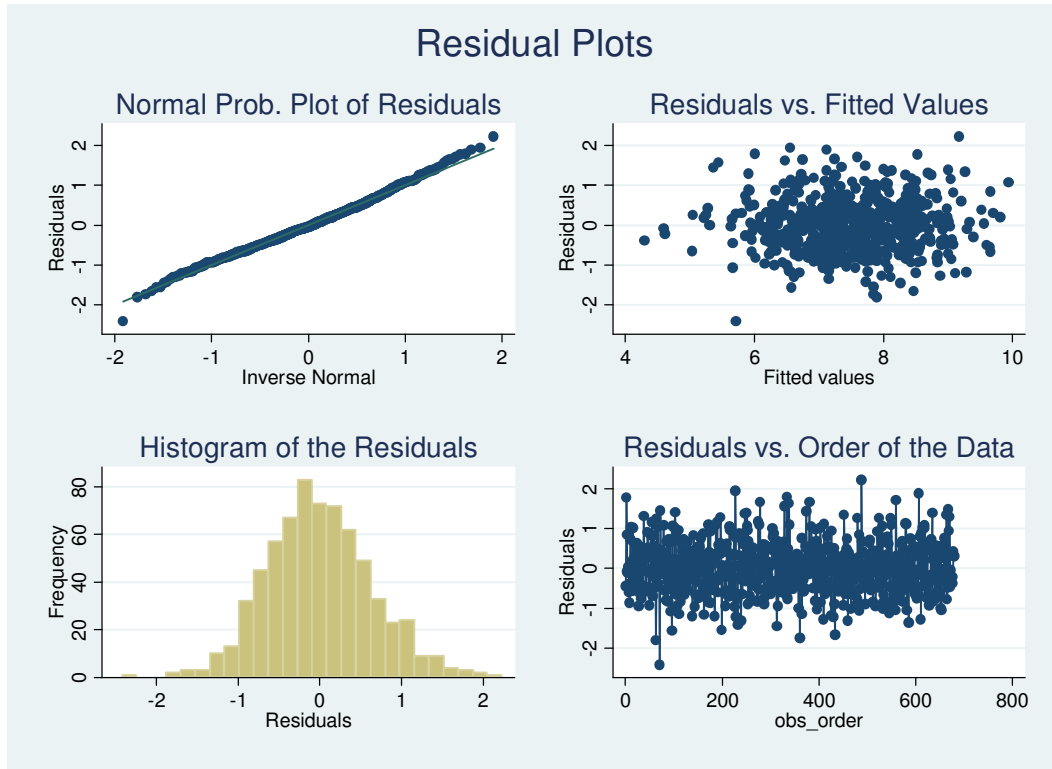
Note that when there is more than one covariate in the model, it makes no sense to plot the lsfit lines a single covariate.

Checking model assumptions.

```
* check for equal variances by ms_gp
robvar(weight), by(ms_gp)
```

Summary of Child's birth weight (lbs)			
ms_gp	Mean	Std. Dev.	Freq.
0	7.7328084	1.0523406	381
1	7.2408027	1.081917	299
Total	7.5164706	1.0923455	680
W0	= 1.6872892	df(1, 678)	Pr > F = .19440027
W50	= 1.6513707	df(1, 678)	Pr > F = .19921129
W10	= 1.6711662	df(1, 678)	Pr > F = .19654276

```
* Create a four-in-one plot
quietly qnorm residual, name(probplot,replace) nodraw title(Normal
Prob. Plot of Residuals)
quietly rvfplot, name(respredplot,replace) nodraw
title(Residuals vs. Fitted Values)
quietly hist residual, freq name(hist,replace) nodraw
title(Histogram of the Residuals)
generate obs_order = _n
quietly twoway connect residual obs_order, name(obs_order,replace) nodraw
title(Residuals vs. Order of the Data)
drop obs_order
graph combine probplot respredplot hist obs_order,
title(Residual Plots)
```



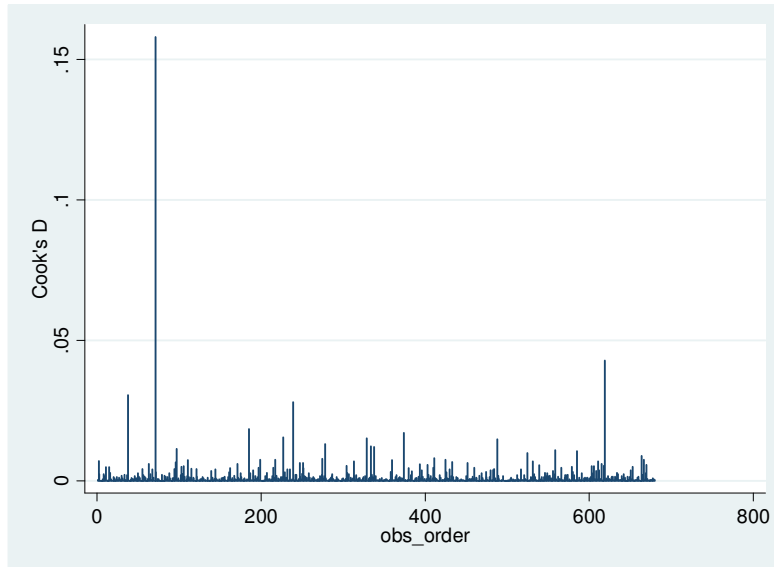
These residuals look really good. There is a tiny U curve to the normal plot, indicating a slight right-skewness.

```
* print the Shapiro-Wilks normality test results
swilk residual
```

Variable	Shapiro-Wilk W test for normal data				
	Obs	W	V	z	Prob>z
residual	680	0.99538	2.050	1.749	0.04012

Shapiro-Wilk rejects normality at the 0.05 level, but from the plots above the data are actually very close to normal.

```
* Cook's distance to examine if any observations have great influence on the
regression
predict cooks,cooks
gene obs_order = _n
tway spike cooks obs_order
drop obs_order cooks
```



There is one observation that is having much more influence on the model fit than any other. With more time, we should remove this observation and see what changes about our resulting model fit. We would go back to the very beginning of the analysis – it is possible to end up with a different model. That would be a big influence, indeed.

* END