

Stat 539 - Biostatistics I – Spring 2006

Homework 7 due March 9 **Solutions**

(assignment for these solutions can be found on the last page)

The code for this analysis comes directly from the labs website
http://www.stat.unm.edu/~erike/courses/stat539/stat539_lab7.do
essentially just doing a find/replace of the three variables used.

I run each section of code, then discuss the results.

```
* ANCOVA -- ANalysis of COVariance
*****
* stat539_hw7_birth_weight.dta -- child birth weight data

clear
* load the dataset
use stat539_hw7_birth_weight.dta
* print to screen
list

* 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

* table of means, standard deviations, and frequencies
tabulate sex ms_gp, summarize(birthwt)
```

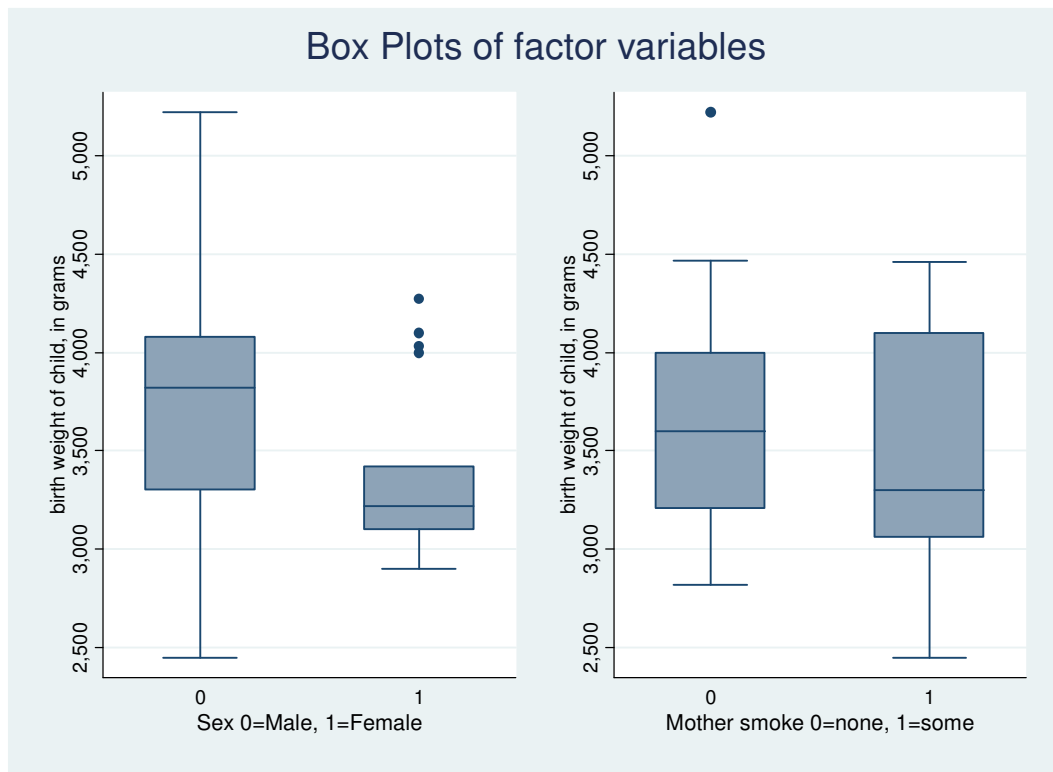
Means, Standard Deviations and Frequencies
of birth weight of child, in grams

sex of child, 0=male 1=female	ms_gp		Total
	0	1	
0	3753.5417 539.8983 24	3606 777.22584 5	3728.1034 573.54383 29
1	3389.6667 395.04279 15	3340 532.66625 4	3379.2105 411.21859 19
Total	3613.5897 515.836 39	3487.7778 654.28927 9	3590 538.94696 48

There appears to be a decrease of the birthwt when going from male to female, and hardly any decrease when going from nonsmoking to smoking. There does not appear to

be an interaction (in a two-by-two interaction occurs when diagonals are large or small together).

```
quietly graph box birthwt, over(sex) name(box1,replace) nodraw
blttitle("Sex 0=Male, 1=Female")
quietly graph box birthwt, over(ms_gp) name(box2,replace) nodraw
blttitle("Mother smoke 0=none, 1=some")
graph combine box1 box2, title(Box Plots of factor variables)
```



The boxplots indicate a likely difference in means for sex, but not for smoking.

```
* 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 birthwt sex ms_gp sex*ms_gp lgest mheight mweight1 mweight2, cat(sex ms_gp)
cont(lgest mheight mweight1 mweight2)
```

Source	Partial SS	df	MS	F	Prob > F
Model	5149273.06	7	735610.437	3.46	0.0054
sex	1141488.28	1	1141488.28	5.37	0.0257
ms_gp	205507.359	1	205507.359	0.97	0.3314
sex*ms_gp	10025.7343	1	10025.7343	0.05	0.8292
lgest	1642428.53	1	1642428.53	7.73	0.0083

Number of obs = 48 R-squared = 0.3772
 Root MSE = 461.046 Adj R-squared = 0.2682

mheight		834229.194	1	834229.194	3.92	0.0545
mweight1		209691.097	1	209691.097	0.99	0.3266
mweight2		1169577.54	1	1169577.54	5.50	0.0240
Residual		8502526.94	40	212563.173		

Total		13651800	47	290463.83		

Some covariates are significant, but the smoking group factor is not, most notably the factor interaction is not significant.

```
* same analysis using xi regress (different parameter constraints)
* specify the categorical variables with the prefix "i."
xi:regress birthwt i.sex i.ms_gp i.sex*i.ms_gp lgest mheight mweight1 mweight2
```

```
i.sex          _Isex_0-1      (naturally coded; _Isex_0 omitted)
i.ms_gp        _Ims_gp_0-1    (naturally coded; _Ims_gp_0 omitted)
i.sex*i.ms_gp  _IsexXms__#_#  (coded as above)
```

Source	SS	df	MS	Number of obs =	48
Model	5149273.06	7	735610.437	F(7, 40) =	3.46
Residual	8502526.94	40	212563.173	Prob > F =	0.0054
-----				R-squared =	0.3772
-----				Adj R-squared =	0.2682
Total	13651800	47	290463.83	Root MSE =	461.05

birthwt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Isex_1	-406.7851	156.7076	-2.60	0.013	-723.5029 -90.06724
_Ims_gp_1	-139.6303	237.8336	-0.59	0.560	-620.3099 341.0493
_Isex_1	(dropped)				
_Ims_gp_1	(dropped)				
_IsexXms__~1	-79.77047	367.3061	-0.22	0.829	-822.1237 662.5828
lgest	160.0929	57.59341	2.78	0.008	43.69226 276.4935
mheight	-30.10616	15.19696	-1.98	0.054	-60.82037 .6080395
mweight1	-17.25892	17.37671	-0.99	0.327	-52.37856 17.86073
mweight2	31.88233	13.59188	2.35	0.024	4.412115 59.35255
_cons	934.1229	3109.37	0.30	0.765	-5350.148 7218.394

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

```
testparm _Ims_gp_1
testparm _Isex_1
testparm _IsexXms__1_1
```

```
( 1)  _Ims_gp_1 = 0
      F( 1, 40) = 0.34
      Prob > F = 0.5604
```

```
( 1)  _Isex_1 = 0
      F( 1, 40) = 6.74
      Prob > F = 0.0131
```

```
( 1)  _IsexXms__1_1 = 0
      F( 1, 40) = 0.05
      Prob > F = 0.8292
```

The testparm command can be used to perform similar tests as appear in the regression table.

```

* Variable selection -- removing nonsignificant covariates and factors until all are
significant
*   Because we are interested in sex, retain that variable regardless of
significance
* Backward selection steps below
*   Begin with full model
*   sex*ms_gp factor interaction least significant (p-value = 0.6744)
*   ms_gp factor least significant (p-value = 0.3365)
*   mweight1 covariate least significant (p-value = 0.3382)
anova birthwt sex ms_gp sex*ms_gp lgest mheight mweight1 mweight2, cat(sex ms_gp)
cont(lgest mheight mweight1 mweight2)
anova birthwt sex ms_gp lgest mheight mweight1 mweight2, cat(sex ms_gp)
cont(lgest mheight mweight1 mweight2)
anova birthwt sex lgest mheight mweight1 mweight2, cat(sex )
cont(lgest mheight mweight1 mweight2)
anova birthwt sex lgest mheight mweight2, cat(sex )
cont(lgest mheight mweight2)
* all remaining factors and covariates significant at the 0.05 level
anova birthwt sex lgest mheight mweight2, cat(sex) cont(lgest mheight mweight2)

```

```

Number of obs = 48 R-squared = 0.3478
Root MSE = 455.036 Adj R-squared = 0.2871

```

Source	Partial SS	df	MS	F	Prob > F
Model	4748307.17	4	1187076.79	5.73	0.0009
sex	2367275.37	1	2367275.37	11.43	0.0015
lgest	1929166.3	1	1929166.3	9.32	0.0039
mheight	846953.27	1	846953.27	4.09	0.0494
mweight2	1429450.75	1	1429450.75	6.90	0.0119
Residual	8903492.83	43	207057.973		
Total	13651800	47	290463.83		

After backward selection, we arrive at a model where sex is a significant factor, and lgest,mheight and mweight2 are significant covariates for predicting birthwt.

```

* same analysis using xi regress (different parameter constraints)
xi:regress birthwt i.sex lgest mheight mweight2
* significance of individual parameters (same as in xi:regress parameter estimate
table because only two levels)
testparm _Isex_1

* need to run anova again for lincom statement below (doesn't like the regress for the
sex[1])
anova birthwt sex lgest mheight mweight2, cat(sex) cont(lgest mheight mweight2)
tabstat birthwt, by(sex) stat(mean semean)

```

Summary for variables: birthwt
 by categories of: sex (sex of child, 0=male 1=female)

sex	mean	se(mean)
0	3728.103	106.5044
1	3379.211	94.34002
Total	3590	77.79029

These are the means of birthwt by sex. They appear different. Are they still different after adjusting for the covariates in our model above?

```
adjust lgest mheight mweight2, by(sex) se
```

```
-----
Dependent variable: birthwt      Command: anova
Covariates set to mean: lgest = 40.041668, mheight = 165.28542, mweight2 = 76.945831
-----

sex of |
child, |
0=male |
1=female |          xb          stdp
-----+-----
      0 |      3774.37   (85.3628)
      1 |      3308.6   (106.019)
-----

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

They still appear different after adjusting for our three covariate, setting each covariate to their mean values. Below I do this manually using the lincom command and the parameter estimates “_b” from the anova command.

```
tabstat lgest mheight mweight2
```

stats	lgest	mheight	mweight2
mean	40.04167	165.2854	76.94583

```
lincom( _b[_cons] + _b[sex[1]] + _b[lgest]*40.04167 + _b[mheight]*165.2854 +
_b[mweight2]*76.94583 )
lincom( _b[_cons] + _b[sex[2]] + _b[lgest]*40.04167 + _b[mheight]*165.2854 +
_b[mweight2]*76.94583 )
```

```
( 1)  _cons + sex[1] + 40.04167 lgest + 165.2854 mheight + 76.94583 mweight2 = 0
```

birthwt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	3774.368	85.36283	44.22	0.000	3602.217 3946.519

```
( 1)  _cons + sex[2] + 40.04167 lgest + 165.2854 mheight + 76.94583 mweight2 = 0
```

birthwt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	3308.599	106.0186	31.21	0.000	3094.792 3522.406

The benefit of using the lincom is the 95% CI, which can be quickly used to see if the birthwts are different by whether the CIs overlap or not.

After adjusting for significant covariates (lgest, mheight, and mweight2), the effect of sex on birthwt is that males tend to be 3774.368-3308.599=465.769 grams heavier than females.

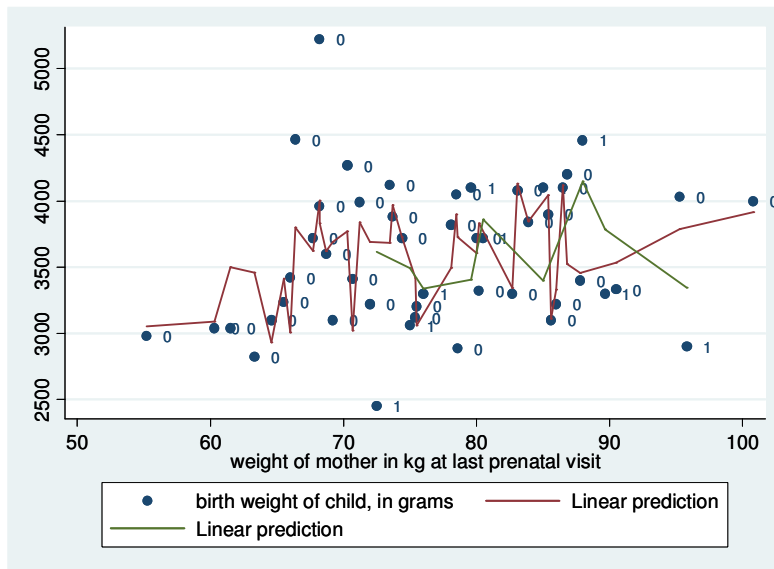
Checking assumptions...

```

* 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.
tway (scatter birthwt mweight2, mlabel(ms_gp)) (line yhat mweight2 if ms_gp==0, sort)
(line yhat mweight2 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.

```

* check for equal variances by sex
robvar(birthwt), by(sex)

```

sex of child,	Summary of birth weight of child, in grams		
0=male	Mean	Std. Dev.	Freq.
0	3728.1034	573.54383	29
1	3379.2105	411.21859	19
Total	3590	538.94696	48

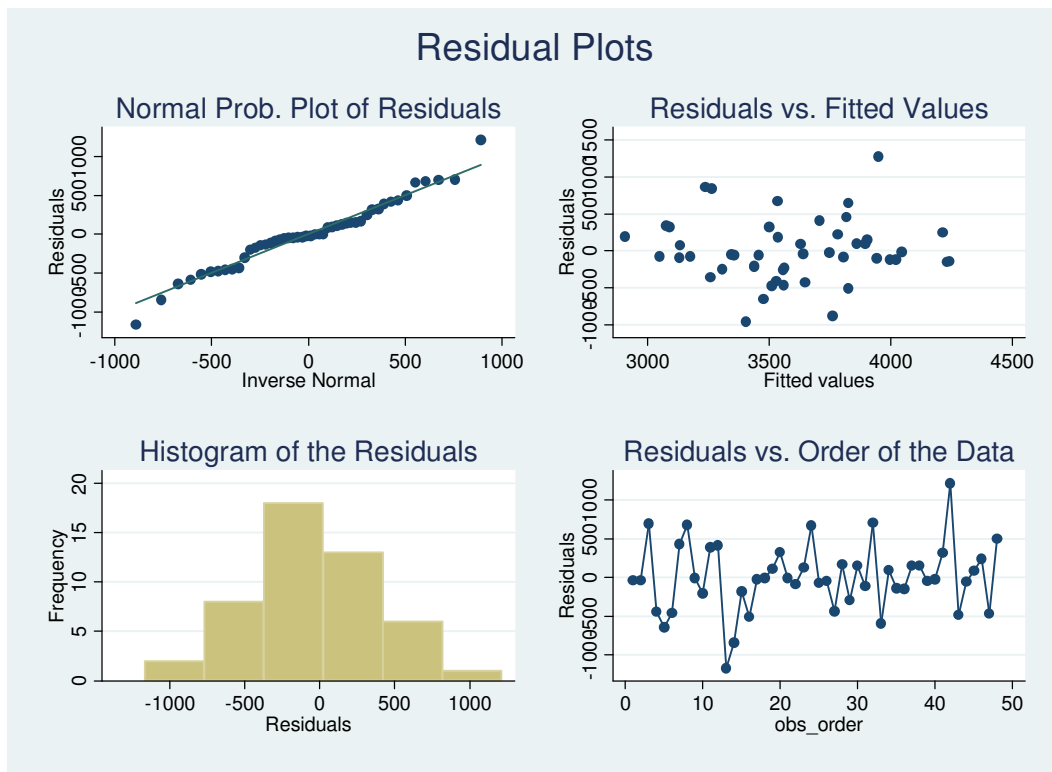
```

W0 = 1.495372 df(1, 46) Pr > F = .22761308
W50 = 1.6532681 df(1, 46) Pr > F = .20494988
W10 = 1.6338694 df(1, 46) Pr > F = .20758101

```

There does not appear to be any problem with the assumption of equal variance between the two sex groups.

```
* 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)
```



The residual plots appear fine.

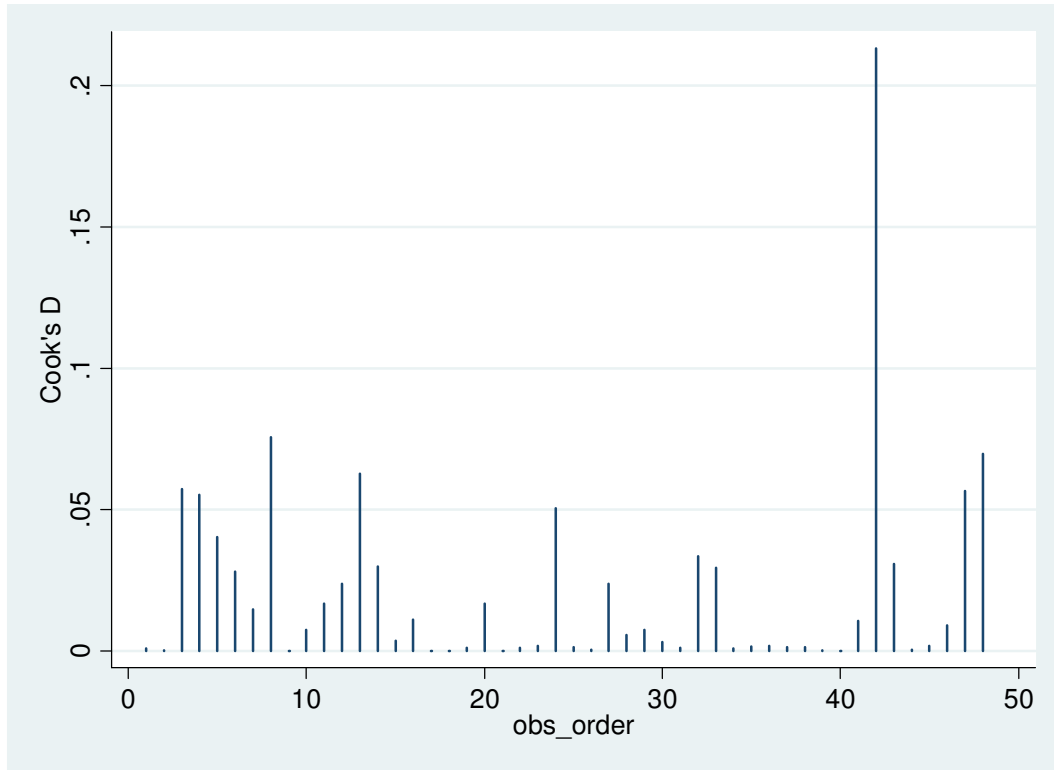
```
* print the Shapiro-Wilks normality test results
swilk residual
```

Shapiro-Wilk W test for normal data					
Variable	Obs	W	V	z	Prob>z
residual	48	0.97771	1.015	0.032	0.48710

Residuals appear normal.

```
* Cook's distance to examine if any observations have great influence on the
regression
predict cooks,cooks
```

```
gene obs_order = _n  
twoway spike cooks_d obs_order  
drop obs_order cooks_d
```



There appears to be one observation with a large influence on the model. I SHOULD remove the observation, and see what effect that observation is having on the model. Starting at the full model, are the same factors and covariates retained? Is the effect and degree of each retained factor and covariate roughly the same without the influential observation?

Stat 539: Biostatistics Methods II
Homework assignment 7, due Thursday March 30

Recall the data from the second HW, given below:

The following data were collected from 48 women who were at least 40 years old when they gave birth to their first child. The data concern the gestation period of that pregnancy, and related variables on the child and mother.

The columns are, from left to right:

- 1) ID
- 2) The child's gestation period, in weeks
- 3) Sex of the child (0=Male, 1=Female)
- 4) Birth Weight of child, in grams
- 5) Number of cigarettes smoked per day (on average) by the mother
- 6) Height of mother in cm
- 7) Weight of mother in kilograms at first prenatal visit
- 8) Weight of mother in kilograms at final prenatal visit

The goal of the analysis is to determine whether there is an effect of child's sex on the birth weight, after adjusting for other features that might impact birth weight. Devise an analysis to answer this question. Treat smoking as a categorical variable as in the CHDS data, but just define two groups (NS/S).

Now you have two categorical variables (although your main interest is in child's sex), and several covariates. This then is a two-way ANOVA with covariates, so you need to worry about interaction between the Sex and Smoking variables (maybe smoking affects the weight of one gender differently from another). Any such effects need to be adjusted for the covariates.

If you find a significant sex effect then quantify the direction and size of the effect (i.e. present and interpret appropriate mean differences based on the selected model). Make sure to clearly summarize your findings.

DATA

```
----  
1 36 0 3300 0 160.0 67.3 82.7  
2 38 0 3300 60 167.6 52.7 76.0  
3 38 0 4100 20 167.6 64.2 79.6  
4 38 1 2900 10 163.9 72.7 95.8  
5 39 0 2820 0 161.3 50.0 63.3  
6 39 0 3040 0 158.8 49.1 61.5  
7 39 0 4120 0 160.0 57.7 73.5  
8 39 0 4200 0 174.0 68.0 86.8  
9 39 1 3100 0 171.5 67.3 85.6  
10 39 1 3330 0 160.0 74.0 90.5  
11 39 1 3410 0 165.1 55.9 70.7  
12 39 1 3420 0 162.6 52.3 66.0  
13 40 0 2450 20 167.6 61.4 72.5  
14 40 0 2885 0 167.7 60.0 78.6  
15 40 0 3235 0 170.2 50.0 65.5  
16 40 0 3320 0 165.1 63.6 80.2  
17 40 0 3600 0 165.1 53.2 68.7  
18 40 0 3720 0 165.0 57.7 74.4  
19 40 0 3720 0 172.7 61.4 80.0  
20 40 0 3820 0 175.3 60.8 78.1  
21 40 0 3840 0 167.0 60.5 83.9  
22 40 0 3880 0 156.2 57.3 73.7  
23 40 0 3960 0 157.5 52.7 68.2  
24 40 0 4465 0 157.5 51.4 66.4  
25 40 1 2980 0 160.0 47.7 55.2  
26 40 1 3040 0 162.0 49.0 60.3  
27 40 1 3060 20 157.5 61.0 75.0  
28 40 1 3100 0 170.2 55.5 64.6  
29 40 1 3120 0 160.3 56.8 75.4  
30 40 1 3205 0 172.7 58.2 75.5  
31 40 1 3220 0 170.0 64.6 86.0  
32 40 1 4100 40 167.0 67.0 85.0  
33 41 0 3100 0 168.9 61.4 69.2  
34 41 0 3720 0 170.2 57.7 67.7  
35 41 0 3720 20 170.2 57.7 80.5  
36 41 0 3900 0 167.0 68.0 85.4  
37 41 0 3990 0 165.1 52.3 71.2  
38 41 0 4050 0 167.6 61.0 78.5  
39 41 0 4080 0 162.6 59.1 83.1  
40 41 0 4100 0 165.1 60.5 86.5  
41 41 0 4460 20 165.1 56.8 88.0  
42 41 0 5220 0 157.5 56.8 68.2  
43 41 1 3300 40 162.6 74.1 89.7  
44 41 1 3400 0 172.7 71.4 87.8  
45 41 1 4000 0 165.1 90.0 100.8  
46 41 1 4030 0 166.0 63.0 95.3  
47 43 1 3220 0 166.4 60.9 72.0  
48 43 1 4270 0 162.6 54.5 70.3
```