MAREB | Econometrics | Instructors: Dr.Gabriel Facchini & Dr.David Castells-Quintana

Students:

Shayan Abbasi (1596105), Kiana Keshavarz (1625840), Carolin Maassen(1549089), Shatha Sameen(1615336)

Note 1: All the tests are carried on with a 95% confidence interval.

1. Estimate the following regression model and ask the following questions:

$$rdi\_exp_i = \alpha + \beta_1 n\_workers_i + \beta_2 sales_i + \beta_3 firm\_age_i + u_i$$

```
. regress rdi_exp n_workers sales firm_age
```

Source	SS	df	MS	Number of obs	=	4,000
				F(3, 3996)	=	431.29
Model	1.9357e+17	3	6.4525e+16	Prob > F	=	0.0000
Residual	5.9784e+17	3,996	1.4961e+14	R-squared	=	0.2446
				Adj R-squared	=	0.2440
Total	7.9141e+17	3,999	1.9790e+14	Root MSE	=	1.2e+07

rdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
n_workers	-699.8289	166.4643	-4.20	0.000	-1026.192	-373.4661
sales	.0152678	.0005279	28.92	0.000	.0142329	.0163027
firm_age	-2063.889	11307.52	-0.18	0.855	-24232.94	20105.17
_cons	1067442	358160	2.98	0.003	365248.9	1769636

1.1. Discuss the individual significance of the estimated coefficients.

Sales and Number of workers (n_workers) have significant a high t-value (-4.2, 28.92) above the value for 95% confidence interval which is also shown as a p-value (0.000, 0.000) lower than 0.05, and thus, statistically significant. The coefficient of firm_age, on the other hand, has a low t-value (or high p-value, 0.855), showing a failure in rejecting the null hypothesis (coefficient is equal to zero), and thus, statistically not significant.

Let's discuss how economically significant is the sales coefficient in question 1.4. For now, for every additional worker, R&D expenditure decreases by 699.8289. Comparing it to the mean value of R&D expenditure ($\frac{699.8289}{2207427}$ = 0.000317). Considering the range of n_workers from 2 to 39591 among firms, the difference in the number of workers can have an economically significant effect on the R&D expenditure in the next period.

1.2. What is the fraction of variability in R&D expenditure that can be explained considering the number of workers, sales, and firm's age as explanatory variables?

With R2 = 0.2446, the independent variables n_workers, sales, and firm_age explain 24,46% of the variability of the dependent variable rdi_exp.

In social systems and natural systems, a dependent variable usually depends on many variables. Therefore, when we focus on explaining a dependent variable with only 3 variables, it is normal to have low R2. The important procedure after this is to show the robustness of the model.

1.3. Test the null hypothesis that the coefficients of all the explanatory variables are jointly equal to zero. What is the conclusion of this test?

High F-Value (431.29) and accordingly low P-value (0.0000) in the regression results show the rejection of the null hypothesis that the coefficients are jointly equal to zero.

1.4. What is the interpretation of the coefficient of the variable sales?

For a 1 Euro increase in Sales in year t, the R&D Expenditure in year t+1 increases by approximately 0.015 Euro.

When we look at the low value for the coefficient, we may misinterpret its economic significance. Considering that the range of the Sales (Min: 317, Max: 1.06e+10) shows that sales can potentially vary a lot across firms, the coefficient for sales is Economically significant, with the level effect of approximately 160 million euro when the sales value goes from min to max. Maybe to show a higher value for coefficient, it's better to change sales scale from units of euros to thousand euros.

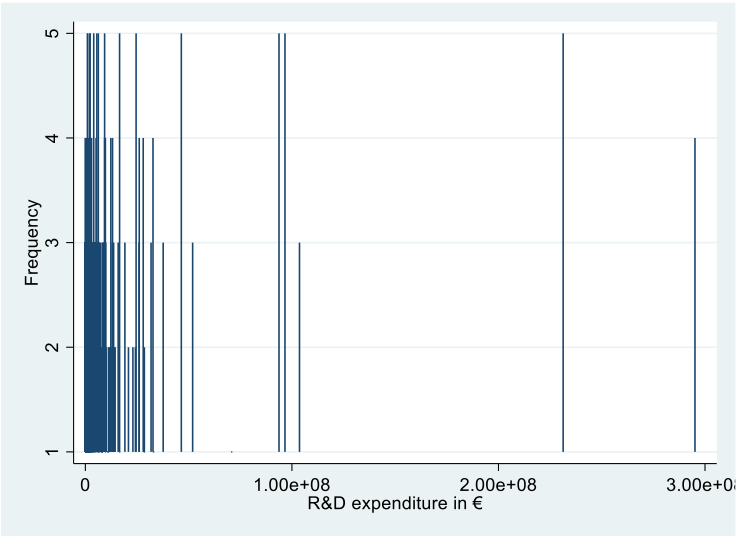
1.5. Compute the prediction (i.e. the expected value) of R&D expenditure of two hypothetical firms, one with 100 workers, 10 years of life, and sales equal to 10,000,000€, and another with the median number of workers, the median age, and sales equal to 10,000,000€ (hint: you should compute the median of the number of workers and firm's age).

$$rdi\_exp_i = \alpha + \beta_1 n\_workers_i + \beta_2 sales_i + \beta_3 firm\_age_i + u_i$$

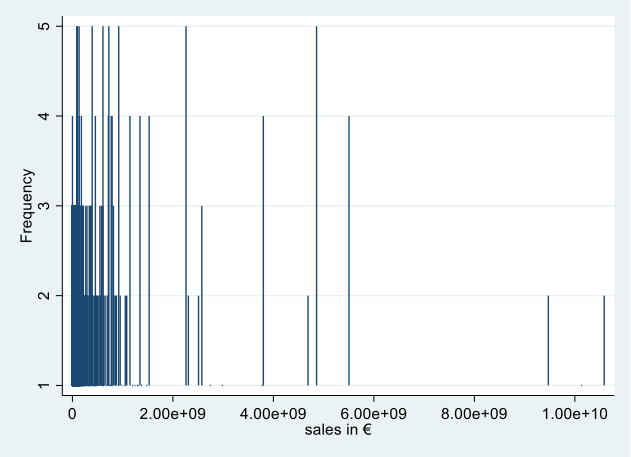
Thus,

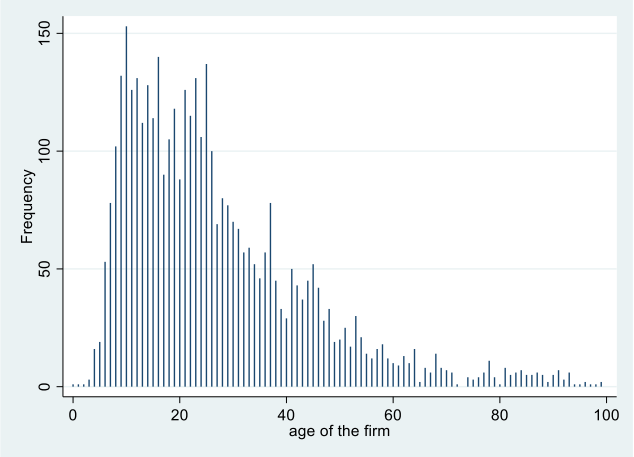
$$rdi\_exp_1 = 1067442 - 699.8289 * (100) + 0.0152678 * (10000000) - 2063.889 * (10)$$
$$= 1{,}129{,}498.22\ Euros$$
$$rdi\_exp_2 = 1067442 - 699.8289 * (64) + 0.0152678 * (10000000) - 2063.889 * (23)$$
$$= 1{,}127{,}861.5034\ Euros$$

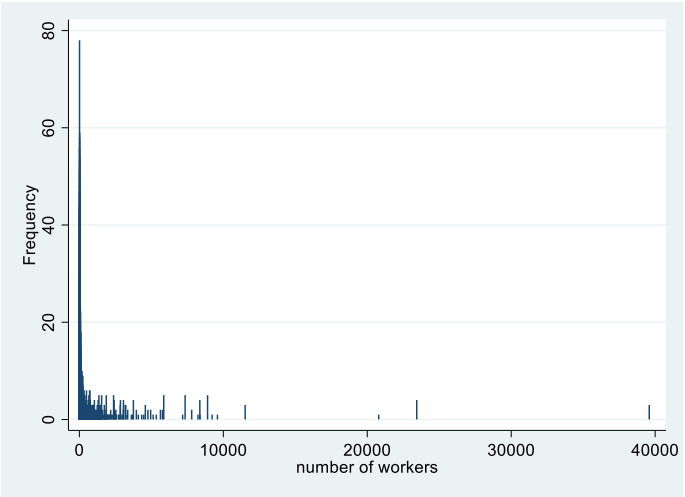
1.6. Check for the presence and the potential impact of outliers in the estimated model. Is there any real problem? Why?



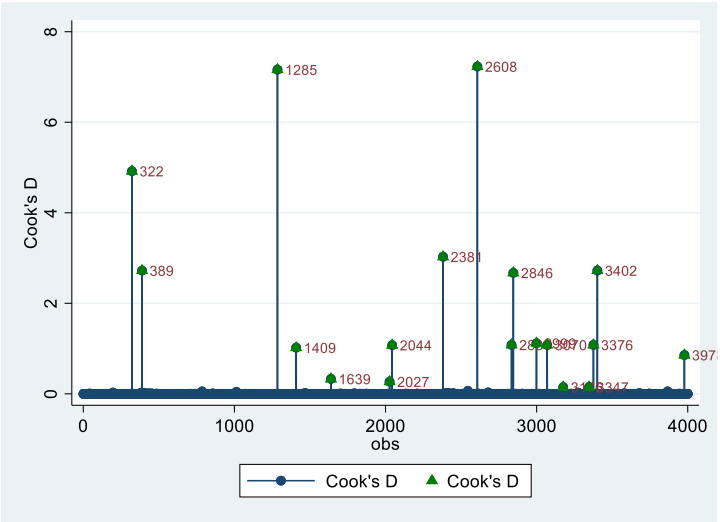
When we look at the firms, we see that a few of the firms have high R&D expenditures (Shown in forms spike plot). To see potential influences of an outlier, we should first see if there are observations with a high value of an independent variable (spike plot for independent variable). Do we see that they Are the high values of R&D expenditure follow the trend in independent variables? If yes, there is no problem. If not, the outliers are a problem for the regression model and we need to Winsorize. Since we have a multivariate regression model, a scatter plot may seem not enough. So let us do Cook's distance, DFFIT, DFBETA.







We observe extreme values for all independent variables. More severe cases for Number of workers and sales. Thus, we carry on the tests to see the real influence of outliers:



Since the number of obs is 4000, we can choose the criteria value equal to 1. As we can see, some observations have a cook's distance higher than 1, showing their real influence on the global fit. To keep the main part of the report brief, we included DFIT and DFBETA in appendix I.

1.7. Check for the presence of multicollinearity in the model. Is there an excessive degree of relationship between explanatory variables? Why?

```
. * correlation matrix between regressors
. corr n_workers sales firm_age
(obs=4,000)
```

	n_work~s	sales	firm_age
n_workers	1.0000		
sales	0.6831	1.0000	
firm_age	0.1290	0.1340	1.0000

The correlation between sales and the number of workers is quite high. Possible multicollinearity caused by them can invalidate the inference of the coefficients that we have.

Variable	VIF	1/VIF
sales	1.88	0.531218
n_workers	1.88	0.531938
firm_age	1.02	0.979416
Mean VIF	1.59	

The VIF values are way below 5.26, therefore, there is no evidence of multicollinearity.

2. Adopt a log-linear model that explains the logarithm of R&D expenditure as a function of the above variables, that is:

$$\ln(rdi\_exp_i) = \alpha + \beta_1 n\_workers_i + \beta_2 sales_i + \beta_3 firm\_age_i + u_i$$

```
. regress log_rdi_exp n_workers sales firm_age
```

Source	SS	df	MS		
				Number of obs	= 4,000
				F(3, 3996)	= 177.59
Model	1195.55967	3	398.51989	Prob > F	= 0.0000
Residual	8967.43069	3,996	2.24410177	R-squared	= 0.1176
				Adj R-squared	= 0.1170
Total	10162.9904	3,999	2.54138294	Root MSE	= 1.498

log_rdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
n_workers	.0000705	.0000204	3.46	0.001	.0000306	.0001105
sales	8.04e-10	6.46e-11	12.44	0.000	6.77e-10	9.31e-10
firm_age	.0108367	.0013849	7.83	0.000	.0081215	.0135518
_cons	12.37588	.0438651	282.13	0.000	12.28988	12.46188

2.1. What is the interpretation of the coefficient associated with firms' age in this case? Is it statistically different from zero?

For a 1 unit increase in a firm's age, the R&D expenditure increases by 1.08% on average. This coefficient is also statistically significant indicating the rejection of the null hypothesis (statistically different than zero)

2.2. Test for the presence of non-linearity in the model using the RESET test. What is the conclusion that can be derived from this test?

```
Ramsey RESET test using powers of the fitted values of log_rdi_exp
       Ho:  model has no omitted variables
                 F(3, 3993) =     161.28
                   Prob > F =     0.0000
```

RAMSEY RESET test tests the model to see if the omitted variables are causing any misspecification. The p-value is less than .05, rejecting the null hypothesis and showing evidence on omitted variable causing misspecification problem.

```
. regress log_rdi_exp n_workers sales firm_age yhat2 yhat3 yhat4
```

Source	SS	df	MS		
				Number of obs	= 4,000
				F(6, 3993)	= 180.12
Model	2164.74276	6	360.790461	Prob > F	= 0.0000
Residual	7998.2476	3,993	2.00306727	R-squared	= 0.2130
				Adj R-squared	= 0.2118
Total	10162.9904	3,999	2.54138294	Root MSE	= 1.4153

log_rdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
n_workers	.0168828	.0023236	7.27	0.000	.0123272	.0214384
sales	1.93e-07	2.64e-08	7.29	0.000	1.41e-07	2.45e-07
firm_age	2.55589	.3552112	7.20	0.000	1.859478	3.252302
yhat2	-19.92803	3.015263	-6.61	0.000	-25.83963	-14.01644
yhat3	.7281857	.1209935	6.02	0.000	.490971	.9654005
yhat4	-.0098673	.0017906	-5.51	0.000	-.0133778	-.0063568
_cons	1915.409	274.6148	6.97	0.000	1377.011	2453.807

Moreover, on other things that we can perform since the coefficients of the powered variables in the unrestricted model are statistically significant, we misspecified our model. We test whether the powered variables are jointly (or in-pairs) zero with an F-test. The P-value is lower than 5%, thus, we reject the hypothesis meaning that the model

is misspecified.

```
. test yhat2=yhat3=yhat4=0

 ( 1)  yhat2 - yhat3 = 0
 ( 2)  yhat2 - yhat4 = 0
 ( 3)  yhat2 = 0
       Constraint 3 dropped

       F(  2,  3993) =  101.97
            Prob > F =    0.0000
```

3. Adopt a log-log specification in which the log of R&D expenditure is regressed against the log of the explanatory variables (except age that can be zero for firms created in the same year of the survey):

$$\ln(rdi\_exp_i) = \alpha + \beta_1 \ln(n\_workers_i) + \beta_2 ln(sales_i) + \beta_3 firm\_age_i + u_i$$

```
. regress lrdi_exp ln_workers lsales firm_age
```

Source	SS	df	MS	Number of obs	=	4,000
				F(3, 3996)	=	698.03
Model	3494.57613	3	1164.85871	Prob > F	=	0.0000
Residual	6668.41424	3,996	1.66877233	R-squared	=	0.3439
				Adj R-squared	=	0.3434
Total	10162.9904	3,999	2.54138294	Root MSE	=	1.2918

lrdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_workers	.4733833	.0285361	16.59	0.000	.4174365	.52933
lsales	.1409174	.0221107	6.37	0.000	.0975682	.1842667
firm_age	-.0063806	.0012855	-4.96	0.000	-.0089009	-.0038604
_cons	8.62893	.2531465	34.09	0.000	8.132621	9.125238

3.1. What is the interpretation of the coefficients β_1 and β_2 in the current specification?

For a 1% increase in the number of workers(sales), the R&D expenditure (for the following period) increases by 0.478% (0.151%) on average. Both coefficients are statistically significant (not zero, p-value < 0.05) and economically significant.

3.2. Test the null hypothesis that the effect of the (logged) number of workers on (logged) R&D expenditure is equal to three times the effect of (logged) sales. What is the conclusion obtained from this test?

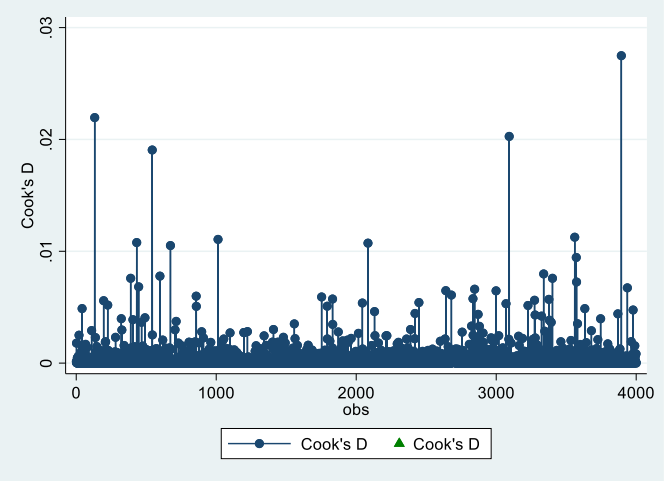
```
. test ln_workers - 3*lsales = 0

 ( 1)  ln_workers - 3*lsales = 0

       F(  1,  3996) =    0.30
            Prob > F =    0.5824
```

The F-test is carried on to see if the (logged) number of workers on (logged) R&D expenditure is equal to three times the effect of (logged) sales. The f-value is not high enough (P-Value is not below 5%) to reject the hypothesis, thus, it can still be the case the effect of num. of workers is 3 times the effect of sales on R&D expenditure.

3.3. Repeat the analysis for the presence and effect of outliers considering this specification. Is there any problem? Why? How can you explain the differences obtained concerning the results obtained in point 1?



Based on Cook's distance and criteria value of 1, the outliers do not seem to be a major issue anymore. The reason is that **logarithm reduces the effect of outliers**. We also use log-log models to have a better interpretation of the coefficients, deal with non-linearity, and a couple of more positive aspects.

4. Generate a new variable capturing a quadratic effect of a firm's age and add it to the model. Therefore, the new model will be:

$$\ln(rdi\_exp_i) = \alpha + \beta_1 \ln(n\_workers_i) + \beta_2 ln(sales_i) + \beta_3 firm\_age_i + \beta_4 firm\_age_i^2 + u_i$$

```
.
. regress lrdi_exp ln_workers lsales firm_age firm_age_2
```

Source	SS	df	MS	Number of obs	=	4,000
				F(4, 3995)	=	546.86
Model	3595.8395	4	898.959875	Prob > F	=	0.0000
Residual	6567.15086	3,995	1.64384252	R-squared	=	0.3538
				Adj R-squared	=	0.3532
Total	10162.9904	3,999	2.54138294	Root MSE	=	1.2821

lrdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_workers	.4597212	.0283756	16.20	0.000	.4040891	.5153533
lsales	.1584281	.022058	7.18	0.000	.1151821	.2016742
firm_age	-.0350375	.0038677	-9.06	0.000	-.0426203	-.0274547
firm_age_2	.0003647	.0000465	7.85	0.000	.0002736	.0004558
_cons	8.803285	.2522287	34.90	0.000	8.308776	9.297794

4.1. Is the inclusion of a quadratic effect of age justified?

The quadratic age has a coefficient significantly not zero. We can even see that the age effect wears off at some level, showing a quadratic relation. This is justified by theory as well, which we elaborated on in the next section. It is good to mention that the structural multicollinearity is also justified for three reasons:

1- We are not going to interpret the firm's age coefficient alone, without considering the quadratic form effect.
2- Multicollinearity occurs when two separate independent variables have the intersection of information represented by them at a very high level. While here we are trying to capture two streams of information validated by different signs for coefficients of age and age squared.
3- There is a good theoretical reason to have the squared independent variable.

4.2. What can be said about the relationship between the age of the firm and R&D expenditure from the estimates obtained from this new specification?

Age seems two have two different effects on R&D expenditure. Firms invest less and less as they grow older (high investment in innovation in early years and low proportional investment in innovation in later years), but as they reach their optimum scale (as they grow old), more R&D expenditure is needed to maintain the growth and stay competitive. Thus, the positive coefficient for age squared tries to offset the negative age effect on R&D expenditure.

4.3. After how many years of existence, the effect of age on a firm's R&D expenditure reaches its minimum? How can you interpret this result?

If the specification were correct, the minimum expected reaches its minimum when the derivative of y with respect to the firm's age reaches zero.

$$\ln(rdi\_exp_i) = \alpha + \beta_1 \ln(n_{workers_i}) + \beta_2 \ln(sales_i) + \beta_3 firm_{age_i} + \beta_4 firm_{age_i}^2$$

$$\frac{\partial \ln(rdi\_exp_i)}{\partial firm_{age_i}} = \beta_3 + 2\beta_4 firm_{age_i} = 0 \rightarrow firm^*_{age} = -\frac{\beta_3}{2\beta_4} = 48.036\ years$$

Since, log will not change the max poing, after 48 years of existence, the effect of ages on a firm´s R&D expenditure reaches its minimum as the calculation of the derivative shows. This implies that in the early years' firms spend more on R&D, however, there is a long-term decline in expenditure reaching its minimum at the year 48. After this, expenditure on R&D goes up again, giving a U-shaped quadratic function on R&D expenditure over the years.

5. We want to account for a separate effect of the number of workers involved in R&D activities, so we first generate a new variable that captures the number of workers "not" involved in R&D activities such as "nworkers_notrd = n_workers - rdi_nworkers" and second, we estimate a model that contains both the number of workers in R&D and the number of workers not involved in R&D (both in logs):

$$\ln(rdi\_exp_i) = \alpha + \beta_1 \ln(nworkers\_notrd_i) + \beta_2 \ln(rdi\_nworkers_i) + \beta_3 ln(sales_i) + \beta_4 firm\_age_i + \beta_5 firm\_age_i^2 + u_i$$

```
. regress lrdi_exp lnworkers_notrd lrdi_nworkers lsales firm_age firm_age_2
```

Source	SS	df	MS	Number of obs	=	4,000
				F(5, 3994)	=	1934.17
Model	7192.52702	5	1438.5054	Prob > F	=	0.0000
Residual	2970.46334	3,994	.743731433	R-squared	=	0.7077
				Adj R-squared	=	0.7074
Total	10162.9904	3,999	2.54138294	Root MSE	=	.8624

lrdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnworkers_notrd	-.0355454	.016582	-2.14	0.032	-.0680554	-.0030354
lrdi_nworkers	.9917227	.0134766	73.59	0.000	.9653011	1.018144
lsales	.1799436	.0151175	11.90	0.000	.1503049	.2095822
firm_age	-.0092814	.0026302	-3.53	0.000	-.0144381	-.0041248
firm_age_2	.000112	.0000315	3.56	0.000	.0000503	.0001737
_cons	8.127911	.1833243	44.34	0.000	7.768493	8.487329

5.1. Is this new specification better than the previous one in terms of the explanatory power of the model?

The new model explains 70.77% of the variability in the dependent variable. So, in terms of goodness of fit, it's a better model. As we add more variables to the model, we reach a higher R^2.

It's good to mention that developing model in social sciences is not focused on reaching higher R^2, rather show a robust representation of statistical model which can lead to a generalized finding of the relationship between the dependent variable and the variable of interest. In many cases (let's say papers), R^2 remains at low levels (e.g., 0.15), while the models are in good shape and generate insights on data.

5.2. How can you interpret the coefficients β_1 and β_2 in the current specification?

For a 1% increase in non-R&D workers (R&D workers) at time t, the R&D expenditure at time t+1 changes by -0.035% (+0.99%). This makes much more sense since R&D workers bring more costs related to R&D to the company. Now, we have separated the different effects of recruits, R&D workers increase the R&D expenditure, while non-R&D workers seem to reduce it. In the previous specification of the model, the non_R&D workers influence reduced the level of total worker's effect R&D expenditure to approximately 0.46%. Now, with the new specification, interpretation of the coefficients can give us better insights.

5.3. Test the null hypothesis that the coefficients for the (logged) number of workers in R&D and the (logged) number of workers not in R&D are jointly equal to zero. What is the result of this test?

```
. test lnworkers_notrd = lrdi_nworkers = 0

 ( 1)  lnworkers_notrd - lrdi_nworkers = 0
 ( 2)  lnworkers_notrd = 0

       F(  2,  3994) = 2708.08
            Prob > F =    0.0000
```

Using an F-test, we test the null hypothesis that the (logged) number of workers in R&D and the (logged) number of workers not in R&D are jointly equal to zero. P-value <0.05 rejects the null hypothesis. Therefore, both variables have effects on R&D expenditure.

6. Include the qualitative information regarding a) whether the firm operates in the international market and b) the type of the firm, where for the latter variable private firms with 100% national capital should be taken as reference categories. The new equation to be estimated takes the form,

$$\ln(rdi\_exp_i) = \alpha + \beta_1 \ln(nworkers\_notrd_i) + \beta_2 \ln(rdi\_nworkers_i) + \beta_3 ln(sales_i) + \beta_4 firm\_age_i + \beta_5 firm\_age_i^2 + \beta_6 inter\_mkt_i + \sum_{j=2}^{4} \delta_j I(firm\_type_i = j) + u_i$$

Where I(firm_type = j) represents dummy variables that take the value 1 if the firm i is of type j (j = 2 if private firm, <50% foreign capital", 3 if "private firm, >=50% foreign capital" and 4 if "public firm or other institution, excluding j =1 if "private firm, with 100% national capital" which is the reference category), and δ_j are the associated coefficients.

```
. regress lrdi_exp lnworkers_notrd lrdi_nworkers lsales firm_age firm_age_2 inter_mkt i.firm_ty
> pe
```

Source	SS	df	MS		
				Number of obs	= 4,000
				F(9, 3990)	= 1104.22
Model	7251.56967	9	805.729964	Prob > F	= 0.0000
Residual	2911.42069	3,990	.729679371	R-squared	= 0.7135
				Adj R-squared	= 0.7129
Total	10162.9904	3,999	2.54138294	Root MSE	= .85421

lrdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnworkers_notrd	-.0391566	.0166059	-2.36	0.018	-.0717134	-.0065998
lrdi_nworkers	.9716446	.0136708	71.07	0.000	.9448423	.9984469
lsales	.1721123	.0155275	11.08	0.000	.1416698	.2025548
firm_age	-.0092292	.0026327	-3.51	0.000	-.0143908	-.0040676
firm_age_2	.0001124	.0000314	3.58	0.000	.0000509	.0001739
inter_mkt	.0931103	.0344763	2.70	0.007	.0255175	.160703
firm_type						
private firm, <50% foreig..	.2831642	.0637892	4.44	0.000	.1581017	.4082267
private firm, >=50% forei..	.2039092	.0426971	4.78	0.000	.1201989	.2876194
public firm or other inst..	.4707934	.0705628	6.67	0.000	.3324509	.6091359
_cons	8.174862	.1869361	43.73	0.000	7.808363	8.541361

6.1. What is the interpretation of the coefficient for the dummy of operating in the international market?

Based on Halvorsen_palmquist's correction (1980), the coefficients' effect on the dependent variable (instead of representing it as a change in log(y) as taught in class) is $100 * \left(e^{\beta_6} - 1\right) = 9.76\%$. Thus,

switching the dummy variable to 1 from 0 increases the expected R&D expenditure by approximately 9.76%. Moreover, the mentioned coefficient is statistically significant.

In other words, based on **Interpretation given in class,** we could just say that the β_6 is the impact of being an internationally operating firm on a firm's logged expenditure regardless of firm type, which is equal to 0.0931103.

6.2. What is the interpretation of the coefficients associated with firm type?

In this estimation, the benchmark is "private firm, 100% national". Therefore, $\delta_2 = 0.2832$ is the expected impact of being a private firm with less than 50% foreign capital compared to "100% national private firms" on logged R&D expenditure. The same interpretation works for $\delta_4\ and\ \delta_3$. $\delta_3 = 0.204$ is the expected impact of being a "private firm with more than 50% foreign capital" compared to "100% national private firms" on logged R&D expenditure. Moreover, $\delta_4 = 0.4708$ is the expected impact of being a "public firm or other institution" compared to "100% national private firms" on logged R&D expenditure. It's good to mention that the intercept (β_0) is equal to the impact of "firm operating domestic" ($\beta_{6'}$) plus the impact of being a "private and 100% national firm" (δ_1), ($\beta_0 = \delta_1 + \beta_{6'}$) now represents the average impact of being a "private firm and 100% national + operating domestic". All coefficients $\delta_2, \delta_3, and\ \delta_4$ interpretations are regardless of the firm's operating market status (international or domestic). To calculate the percentage effect of R&D expenditure, we can use **Halvorsen_palmquist's correction (1980**) as it was calculated above. Also, all the coefficients mentioned are statistically significant.

6.3. Test the null hypothesis that firm type does not affect R&D expenditure.

This null hypothesis is rejected with the t-test carried on in the regression estimation (p-value 0.007 < 0.05).

6.4. What is the difference in (logged) R&D expenditure between public firms and firms with more than 50% of foreign capital?

difference in (logged) R&D expenditure between public firms and firms with more than 50% of foreign capital is equal to $\delta_4 - \delta_3 = 0.2668$. Public firms and other institutions have 0.2668 more effects on logged R&D expenditure than firms with more than 50% foreign capital.

6.5. Change the specification and use firms with more than 50% of foreign capital as the reference group. Is there any real difference in the results?

```
. regress lrdi_exp lnworkers_notrd lrdi_nworkers lsales firm_age firm_age_2 inter_mkt ib3.firm_
> type
```

Source	SS	df	MS	Number of obs	=	4,000
				F(9, 3990)	=	1104.22
Model	7251.56967	9	805.729964	Prob > F	=	0.0000
Residual	2911.42069	3,990	.729679371	R-squared	=	0.7135
				Adj R-squared	=	0.7129
Total	10162.9904	3,999	2.54138294	Root MSE	=	.85421

lrdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnworkers_notrd	-.0391566	.0166059	-2.36	0.018	-.0717134	-.0065998
lrdi_nworkers	.9716446	.0136708	71.07	0.000	.9448423	.9984469
lsales	.1721123	.0155275	11.08	0.000	.1416698	.2025548
firm_age	-.0092292	.0026327	-3.51	0.000	-.0143908	-.0040676
firm_age_2	.0001124	.0000314	3.58	0.000	.0000509	.0001739
inter_mkt	.0931103	.0344763	2.70	0.007	.0255175	.160703
firm_type						
private firm, 100% national	-.2039092	.0426971	-4.78	0.000	-.2876194	-.1201989
private firm, <50% foreig..	.079255	.0718349	1.10	0.270	-.0615816	.2200916
public firm or other inst..	.2668843	.0800123	3.34	0.001	.1100155	.423753
_cons	8.378771	.1986893	42.17	0.000	7.989229	8.768313

Other than the reference, nothing meaningful has changed. In the first specification, "Private firms with more than 50% foreign capital" had an expected impact of 0.204 on R&D expenditure in comparison to "Private firm and 100% national". In this specification, "Private firm and 100% national" has a -0.204 impact on R&D expenditure in comparison to "Private firms with more than 50% foreign capital" regardless of the firm's operating market status. The same interpretation applies to other coefficients. The only difference is that we have "Private firms with more than 50% foreign capital" as the reference now.

It's good to mention that the reference is usually the variable that has the most frequency, and sometimes, it's the variable we are not that much interested in.

7. Consider a new model in which the effect of (logged) sales in the previous year on (logged) R&D expenditure is different according to whether the firm operates in the international market or not, that is:

$$\ln(rdi\_exp_i) = \alpha + \beta_1 \ln(nworkers\_notrd_i) + \beta_2 \ln(rdi\_nworkers_i) + \beta_3 ln(sales_i) + \beta_4 firm\_age_i + \beta_5 firm\_age_i^2 + \beta_6 inter\_mkt_i + \beta_7 (inter\_mkt_i \times ln(sales_i)) + \sum_{j=2}^{4} \delta_j I(firm\_type_i = j) + u_i$$

```
. gen interaction = inter_mkt*lsales

. regress lrdi_exp lnworkers_notrd lrdi_nworkers lsales firm_age firm_age_2 inter_mkt interaction i.firm_type
```

Source	SS	df	MS	Number of obs	=	4,000
				F(10, 3989)	=	996.33
Model	7257.36288	10	725.736288	Prob > F	=	0.0000
Residual	2905.62748	3,989	.728409998	R-squared	=	0.7141
				Adj R-squared	=	0.7134
Total	10162.9904	3,999	2.54138294	Root MSE	=	.85347

lrdi_exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnworkers_notrd	-.038456	.0165933	-2.32	0.021	-.0709881	-.0059238
lrdi_nworkers	.9675428	.0137361	70.44	0.000	.9406124	.9944732
lsales	.1436095	.0185157	7.76	0.000	.1073084	.1799106
firm_age	-.009191	.0026305	-3.49	0.000	-.0143481	-.0040338
firm_age_2	.0001105	.0000313	3.52	0.000	.000049	.0001719
inter_mkt	-.5833714	.2423353	-2.41	0.016	-1.058484	-.1082588
interaction	.04333	.0153644	2.82	0.005	.0132071	.0734529
firm_type						
private firm, <50% foreign capital	.2772685	.063768	4.35	0.000	.1522476	.4022893
private firm, >=50% foreign capital	.192094	.0428652	4.48	0.000	.1080542	.2761338
public firm or other institution	.473861	.0705098	6.72	0.000	.3356224	.6120996
_cons	8.617367	.2439355	35.33	0.000	8.139117	9.095616

7.1. Is it possible to argue that the effect of sales is constant regardless of whether the firm operates in the international market or not?

Since the interaction's coefficient is significantly other than zero, it's not possible to consider the effect of sales on expected R&D expenditure just equal to the coefficient (a constant). The following derivative shows the effect we observe in the new model:

$$\frac{d(lrdi\_epx_i)}{d(lsales_i)} = \beta_3 + \beta_7 . inter\_mkt_i$$

Thus, it depends on the firm's operating market status as well.

7.2. What is the elasticity of R&D expenditure with respect to sales for firms that operate in the international market and firms that operate only in the national market?

In log-log equations, elasticity is the coefficient of the respected variable. When we use interactions with dummies, it's like we are considering two different groups with different elasticities, elaborated on as the following:

For the firms operating in international markets:

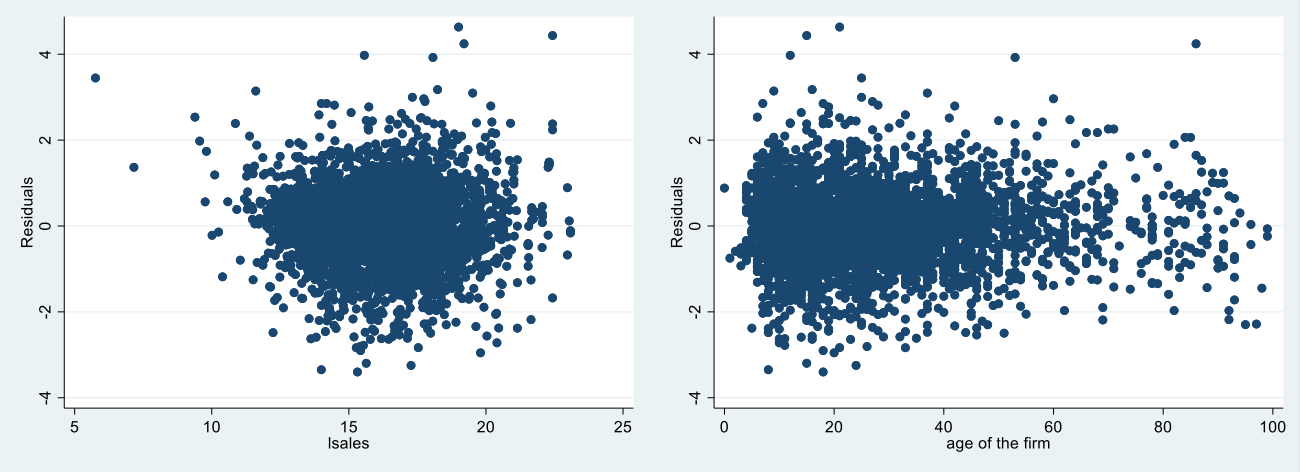
$$\frac{d(lrdi\_epx_i)}{d(lsales_i)} = \beta_3 + \beta_7$$

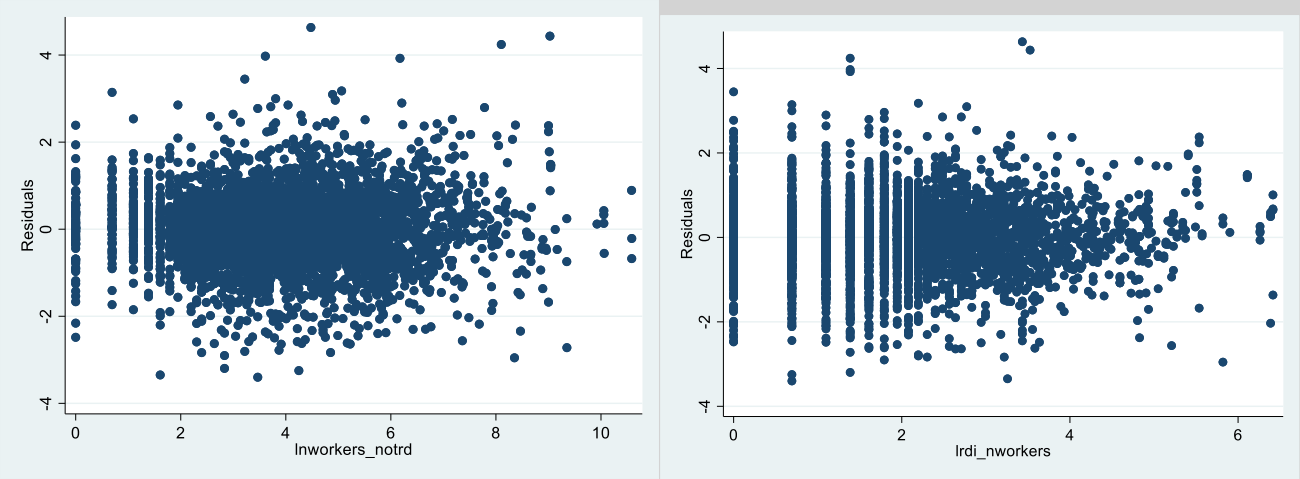
For the firms operating in the domestic market:

$$\frac{d(lrdi\_epx_i)}{d(lsales_i)} = \beta_3$$

8. Check for the presence of heteroscedasticity:

8.1. Do it graphically: for instance, by plotting residuals against independent variables





The graphs could suggest a possible heteroskedasticity. However, the dispersion in the late levels of the independent variable could just be a data availability problem. For example, maybe there are not many firms with age more than 80, which could lead to possible lower dispersion tan firms with lower ages.

8.2. Perform some tests

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lrdi_exp

         chi2(1)      =     0.33
         Prob > chi2  =   0.5631


White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(54)     =    321.32
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	321.32	54	0.0000
Skewness	45.14	10	0.0000
Kurtosis	28.33	1	0.0000
Total	394.79	65	0.0000

To begin, the Breusch-Pagan test tests whether we have a constant variance for different levels of independent variables. Since the P-value is more than 5% we cannot reject the null hypothesis and we cannot be sure if we have heteroscedasticity.

Moreover, the White test rejects the null hypothesis that the OLS model satisfies the homoscedasticity. So, why are we having different tests asserting contrary results?

Breusch-Pagan tests try to find a linear form of heteroscedasticity. It also assumes normal distribution for the error term. White's general test is more powerful and designed to overcome those assumptions if we can provide a large dataset for the test. Since we have 4000 observations, we will consider the white's test as the right test for detecting heteroscedasticity.

8.3. Depending on the results of the tests, implement a heteroscedasticity-robust estimation, and interpret new results.

```
. xi: regress lrdi_exp lnworkers_notrd lrdi_nworkers lsales firm_age firm_age_2 inter_mkt inter
> action i.firm_type, vce(robust)
i.firm_type       _Ifirm_type_1-4     (naturally coded; _Ifirm_type_1 omitted)

Linear regression                               Number of obs     =      4,000
                                                F(10, 3989)       =     889.48
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7141
                                                Root MSE          =     .85347
```

lrdi_exp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnworkers_notrd	-.038456	.0198303	-1.94	0.053	-.0773345	.0004226
lrdi_nworkers	.9675428	.0150728	64.19	0.000	.9379917	.9970939
lsales	.1436095	.0263356	5.45	0.000	.0919769	.1952421
firm_age	-.009191	.0027959	-3.29	0.001	-.0146726	-.0037093
firm_age_2	.0001105	.0000356	3.10	0.002	.0000406	.0001803
inter_mkt	-.5833714	.3089104	-1.89	0.059	-1.189008	.0222656
interaction	.04333	.0201351	2.15	0.031	.003854	.082806
_Ifirm_type_2	.2772685	.0704362	3.94	0.000	.1391742	.4153627
_Ifirm_type_3	.192094	.0449496	4.27	0.000	.1039677	.2802204
_Ifirm_type_4	.473861	.08188	5.79	0.000	.3133305	.6343914
_cons	8.617367	.345645	24.93	0.000	7.939709	9.295024

The heteroscedasticity-robust estimation adjusts the standard error and therefore, calculates a new p-value. The adjusted p-values in the heteroscedasticity-robust estimation show that the variables lnworkers_notrd and inter_mkt are not statistically significant anymore.

One possible consideration could be to look for more data and better datasets if our variable of interest is one of these two variable. If these are control variable, maybe we wouldn't have a major problem.

Appendix I)

```
.
. list rdi_exp n_workers sales firm_age dfit if abs(dfit)>2*sqrt(3/4000)
```

	rdi_exp	n_work~s	sales	firm_age	dfit
42.	9.38e+07	729	3.46e+07	12	.1703009
196.	9.66e+07	1545	5.57e+08	74	.3403151
208.	4.65e+07	5861	9.22e+08	31	.1726068
281.	3.27e+07	1475	2.69e+08	4	.0699897
322.	1558751	39591	9.47e+09	41	-4.496428
382.	612534	2207	2.51e+09	92	-.3192268
389.	2.95e+08	8369	5.50e+09	15	3.441953
405.	9.66e+07	1215	1.34e+09	36	.3086036
432.	91025	9222	2.31e+09	18	-.2277136
452.	1660186	4952	2.26e+09	81	-.2021798
484.	9327342	7346	2.26e+09	13	-.139523
491.	2.46e+07	1868	6.08e+08	93	.0814014
653.	2534159	2854	1.53e+09	40	-.0791638
714.	3.77e+07	4475	1.47e+09	90	.1018087
764.	3807658	7346	1.37e+09	99	-.0975171
786.	1.66e+07	3222	3.80e+09	10	-.469045
857.	5.19e+07	918	3.57e+08	83	.2041392
878.	3.27e+07	1556	2.64e+08	54	.0737403
1013.	2.95e+08	119	1.79e+08	21	.4587651
1074.	2.86e+07	3959	2.73e+09	55	-.0848019
1285.	2.42e+07	5630	1.06e+10	81	-5.451221
1409.	7.06e+07	20794	1.01e+10	84	-2.029933
1468.	1622286	135	2.31e+09	24	-.2823094
1639.	4294028	8369	5.50e+09	12	-1.154454
1702.	1.91e+07	4952	2.97e+09	17	-.1904473
1793.	9.66e+07	759	7.12e+08	13	.2223799
1809.	990289	985	8.57e+08	86	-.0615509
1832.	4.65e+07	1774	9.29e+07	10	.1119001
1872.	9.38e+07	288	9.87e+07	17	.1376977
2016.	5.19e+07	1449	2.88e+08	12	.0957466
2027.	8795646	915	4.69e+09	46	-1.040425
2044.	2.31e+08	8904	4.86e+09	52	2.1234
2134.	5.19e+07	1545	1.34e+09	48	.1278302
2149.	2.80e+07	23439	2.58e+09	5	.090933
2157.	3009800	2069	1.08e+09	79	-.0600661
2209.	9327342	2444	2.26e+09	24	-.1592532
2232.	2235166	3959	1.23e+09	64	-.0616469
2381.	2.60e+07	39591	9.47e+09	70	-3.509438
2404.	647706	1626	2.51e+09	68	-.3070364
2418.	9.38e+07	303	9.87e+07	24	.1208656
2420.	1.04e+08	2375	7.16e+08	9	.2564236
2448.	774161	23439	4.44e+08	39	.2581014
2545.	1.66e+07	1610	3.80e+09	18	-.5210583
2608.	2.42e+07	5630	1.06e+10	24	-5.477629
2617.	9327342	7346	2.26e+09	22	-.1357918
2671.	1780374	1122	1.14e+09	53	-.0607224
2680.	9.66e+07	1545	7.12e+08	77	.3575486
2808.	2.61e+07	3753	3.33e+08	54	.09
2834.	2.31e+08	8904	4.86e+09	60	2.129138
2846.	2.95e+08	8369	5.50e+09	27	3.409847
2860.	4.65e+07	2177	9.22e+08	17	.0889824
2864.	3.27e+07	817	9.73e+07	53	.0739461
2906.	9.38e+07	729	9.87e+07	22	.1309761
2969.	4.65e+07	2511	9.22e+08	22	.0868875
2999.	2.31e+08	8904	4.86e+09	15	2.164478
3070.	2.31e+08	8904	4.86e+09	36	2.127591
3176.	2381662	287	3.80e+09	12	-.7733629
3227.	3.77e+07	1545	5.57e+08	89	.1431491
3267.	1.23e+07	506	1.53e+09	13	-.0617807
3274.	9.66e+07	1013	1.34e+09	13	.3326851
3339.	1.04e+08	2375	7.16e+08	87	.4550235
3347.	313039	418	3.77e+09	18	-.7757565
3355.	2.80e+07	23439	2.58e+09	19	.0908304
3376.	2.31e+08	8904	4.86e+09	59	2.128136
3402.	2.95e+08	8369	5.50e+09	15	3.441953
3534.	3.18e+07	1001	6.03e+08	58	.0650774
3572.	55076	115	1.53e+09	5	-.1394551
3581.	1.04e+08	2375	7.16e+08	21	.2171022
3633.	3.18e+07	1435	1.28e+09	91	.0726015
3680.	4.65e+07	5861	2.72e+08	40	.270191
3746.	9.38e+07	729	9.87e+07	8	.1840503
3867.	1.66e+07	3222	3.80e+09	10	-.469045
3869.	2.80e+07	3307	2.18e+08	86	.1412432
3884.	2.80e+07	23439	2.58e+09	83	.0937069
3946.	9327342	7346	2.26e+09	17	-.1376858
3978.	1181979	39591	4.69e+09	44	-1.854476

The DFIT shows the individual fit to the model. As we can see, the same observations in Cook's difference have high positive or negative DFIT values. For example, obs 322 has a Cook's distance near 5 and approximately a DFIT value of -4.5, asserting that it lies way below the regression line.

```
. list rdi_exp n_workers sales firm_age _dfbeta_1 if abs(_dfbeta_1)>2/sqrt(4000)
```

	rdi_exp	n_work~s	sales	firm_age	_dfbeta_1
42.	9.38e+07	729	3.46e+07	12	.0595434
132.	201992	8377	1116051	0	.046822
208.	4.65e+07	5861	9.22e+08	31	.148001
322.	1558751	39591	9.47e+09	41	-2.959575
382.	612534	2207	2.51e+09	92	.1481079
389.	2.95e+08	8369	5.50e+09	15	-.919577
405.	9.66e+07	1215	1.34e+09	36	-.1566007
432.	91025	9222	2.31e+09	18	-.1388733
484.	9327342	7346	2.26e+09	13	-.0567701
764.	3807658	7346	1.37e+09	99	-.0576465
786.	1.66e+07	3222	3.80e+09	10	.2294173
956.	3009800	7802	1.19e+08	11	.0470636
1013.	2.95e+08	119	1.79e+08	21	-.1326647
1140.	2174227	7346	8.77e+08	33	-.0420407
1285.	2.42e+07	5630	1.06e+10	81	3.179782
1409.	7.06e+07	20794	1.01e+10	84	.1127501
1468.	1622286	135	2.31e+09	24	.1922003
1639.	4294028	8369	5.50e+09	12	.3066636
1702.	1.91e+07	4952	2.97e+09	17	.0412238
1793.	9.66e+07	759	7.12e+08	13	-.0817451
1832.	4.65e+07	1774	9.29e+07	10	.0756748
2016.	5.19e+07	1449	2.88e+08	12	.0396506
2027.	8795646	915	4.69e+09	46	.6855944
2044.	2.31e+08	8904	4.86e+09	52	-.3329314
2134.	5.19e+07	1545	1.34e+09	48	-.0546069
2149.	2.80e+07	23439	2.58e+09	5	.0857851
2209.	9327342	2444	2.26e+09	24	.0696796
2381.	2.60e+07	39591	9.47e+09	70	-2.308324
2404.	647706	1626	2.51e+09	68	.1705795
2420.	1.04e+08	2375	7.16e+08	9	.0802167
2448.	774161	23439	4.44e+08	39	.2574534
2545.	1.66e+07	1610	3.80e+09	18	.313849
2608.	2.42e+07	5630	1.06e+10	24	3.149697
2617.	9327342	7346	2.26e+09	22	-.0559555
2808.	2.61e+07	3753	3.33e+08	54	.0710836
2834.	2.31e+08	8904	4.86e+09	60	-.3383264
2846.	2.95e+08	8369	5.50e+09	27	-.9300284
2906.	9.38e+07	729	9.87e+07	22	.0411337
2999.	2.31e+08	8904	4.86e+09	15	-.3084714
3070.	2.31e+08	8904	4.86e+09	36	-.3222597
3176.	2381662	287	3.80e+09	12	.5151142
3242.	1466446	5126	9.51e+08	50	-.0333733
3267.	1.23e+07	506	1.53e+09	13	.0372983
3274.	9.66e+07	1013	1.34e+09	13	-.1665727
3279.	839292	11518	7.25e+08	66	-.0340422
3339.	1.04e+08	2375	7.16e+08	87	.0517434
3347.	313039	418	3.77e+09	18	.515514
3355.	2.80e+07	23439	2.58e+09	19	.0860737
3376.	2.31e+08	8904	4.86e+09	59	-.3376497
3402.	2.95e+08	8369	5.50e+09	15	-.919577
3572.	55076	115	1.53e+09	5	.0873819
3581.	1.04e+08	2375	7.16e+08	21	.0757998
3667.	1.32e+07	3062	1.53e+08	37	.0342389
3680.	4.65e+07	5861	2.72e+08	40	.2593805
3746.	9.38e+07	729	9.87e+07	8	.0461869
3867.	1.66e+07	3222	3.80e+09	10	.2294173
3869.	2.80e+07	3307	2.18e+08	86	.0728848
3884.	2.80e+07	23439	2.58e+09	83	.0877414
3894.	16775	11518	7.25e+08	10	-.0452493
3946.	9327342	7346	2.26e+09	17	-.0564068
3978.	1181979	39591	4.69e+09	44	-1.742294

DFBETA shows how the outlier influences the coefficients. For example, obs 322 have a negative effect on coefficients. Thus, excluding it will increase the coefficients.