Week 7: Diagnostics II

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

- Roadmap for the rest of the semester
- Important concepts to remember
- Reminder: why we don't check that y distributes normal
- Digression: low R^2
- Measures of influence: Cook's distance and DFITS

Roadmap for the rest of the semester

- I need to cover regression diagnostics or how to detect some violations of model assumptions
- Then we will finally cover qualitative variables since models are interpreted in slightly different ways
- With the tools covered, we will be ready to make sense of how Analysis of Variance (ANOVA) is just linear regression parametrized in different ways
- For example, "two-way factorial ANOVA" is just a model with interactions. But there is a large number of statistical tests one can do, and that's when things get confusing
- Although it's called analysis of **variance** because the tests use "variances" (SST = SSE + SSR), it's all about testing if **MEANS** are the same. The tests happen to be **ratios of variances**, so we know they will be F tests

Roadmap for the rest of the semester

- After the Spring break we are going to spend a lot of time interpreting all sorts of models, from easy to complex
- I need to change lecture notes (got myself Spring break chores). You need to practice model building and interpretation
- We will use derivatives to interpret models because life it's a lot easier that way. And we will spend time **writing down** models
- For example in the model $y_i = \beta_0 + \beta_1 age_i + \beta_2 male_i + \beta_3 age * male_i + \epsilon_i$ the effect of age is easy to interpret taking the derivative: $\frac{\partial E[y_i]}{\partial age} = \beta_1 + \beta_3 male_i$
- So how *age* influences $E[y]$ depends on being male or female. For males the effect is $\beta_1 + \beta_3$. For females it's just β_1

Roadmap for the rest of the semester

- We are going to spend **several weeks playing with models**
- For some models, the analytical derivative is not an easy expression. For example, remember that the derivative of $y = e^x$ is just $y' = e^x$
- In the probability scale, the logistic model is $p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}$
- The effect of, say, X_1 is $\frac{\partial p}{\partial X1}$. But without even trying to solve the derivative we can see that it won't be an easy expression. In the probability scale, the effect of X_1 depends on the **value of X_1 and the value of all other covariates**
- We will instead compute a **numerical derivative** as opposed to an **analytical derivative**
- That's what economists and Stata call **marginal effects**. SAS has some procedures that calculate marginal effects. R has a package too. Sometimes they are called, in the context of logistic regression, **average predicted probabilities**
- *But you can use marginal effects for any type of model*

Don't forget the basics

- **Forgetting the basics of statistics comes back to bite you in the face**. Trust me, I'm older and have been doing this for a while
- Two probability density functions (pdf) (not cumulative):
  $f(x; p) = p^x(1-p)^{(1-x)}$, where $x \in \{0, 1\}$ and
  $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ where $x \in \{-\infty, \infty\}$
- These functions **describe all the possible values a random variable can take and their probability**. They have parameters (p and μ, σ)
- If we know the pdf then we can do hypothesis testing and probabilistic statements like what is the probability that x will be 0, 10, etc? Or, say, 95% percent of the time the value of x will be between $[a, b]$
- Why is this important? Because model parameters are random variables that have a theoretical distribution. That's why we can do hypothesis testing

Don't forget the basics

- Estimators have a distribution and also variability (standard error)
- They have properties that are important: bias, efficiency, and consistency
- Consistency is an **asymptotic** property **in probability**
- Next semester you will see different models that have different properties and you will get lost if you don't understand these concepts
- The only relatively tricky question was 8. Everybody forgot about power and most forgot about the other covariate, X_2
- The correlation coefficient is not that great, only measures a linear relationship not any relationship. That's why we do simple linear models. More flexibility
- Zero conditional mean assumption, ignorability, exchangeability, selection on observables, exogeneity, they **all mean the same thing**. We need to know about the data generating mechanism. You need to know the assignment mechanism to figure out of these assumptions are met. **Different names because of different fields**

Digression

- The reg command uses OLS to estimate models, not MLE
- Rather than plugging in the betas in to the likelihood function, Stata uses a formula that maps SSE to the log-likelihood
- The formula is $-0.5N(ln(2\pi) + ln(\frac{SSE}{N}) + 1)$

The regression command again

- Stata uses a formula to go from SSE to log likelihood. Remember, SSE is Stata is stored in the scalar r(rss)

```
sysuse auto, clear
qui reg price weight mpg

* Save sample size and SSE
local N   = e(N)
local rss = e(rss)

* Use formula
local ll  = -0.5*‘N’*(ln(2*_pi)+ln(‘rss’/‘N’)+1)

display %20.6f ‘ll’
-682.863688
display %20.6f e(ll)
-682.863688
```

Where we are

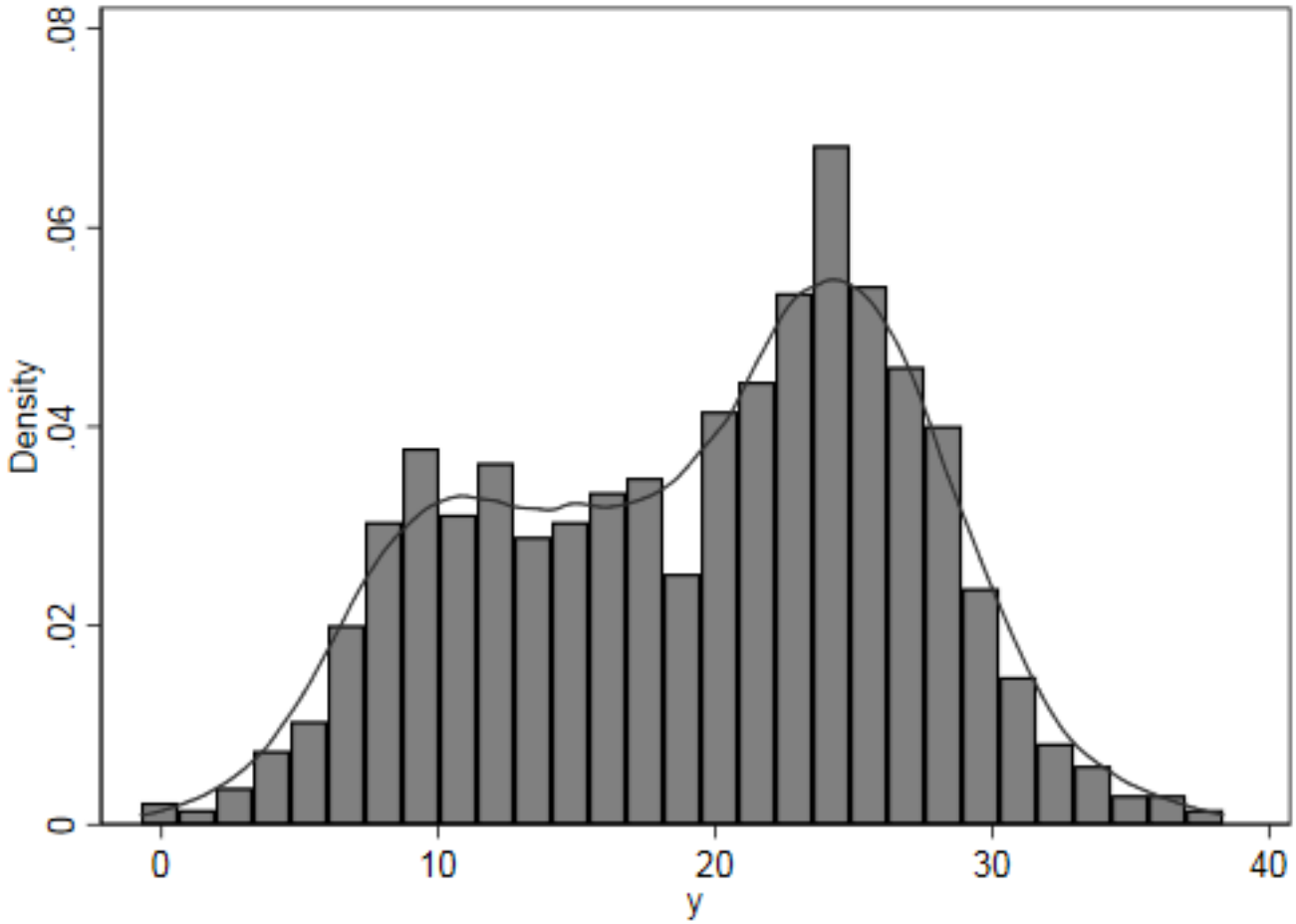
- Recall that when I say residuals I mean *standardized* residuals
- We saw that scatter plots of residuals against x helps to detect non-linearity (functional form) or non-normal errors
- It also helps with heteroskedasticy (funnel like plots, more on that later)
- Your textbook (Chatterjee) has more on this; please read the chapter carefully

Reminder: why we don't check that y distributes normal?

- I told you several times that in the linear model we assume that $\epsilon \sim N(0, \sigma^2)$, which is equivalent to saying that $Y \sim N(\mathbf{X}\beta, \sigma^2)$
- So why don't we just verify that Y distributes normal? I said several times that it 's because the we assume that the outcome distributes normal conditional on covariates
- Now, a more developed example (because the idea also illustrates another issue about using graphs instead of theory)

Distribution of an outcome

- Suppose the distribution below is from an outcome Y. It doesn't look like normally distributed data



If you showed the above graph to a friend...

- If you showed the above graph to about 100 well-educated friends who took a stats class, about 95 percent will tell you that you should not use a linear regression because the outcome Y does not appear to be normally distributed
- If some of your friends knew about mixture models, they would be tempted to use a mixture model
- BUT...
- **Don't forget this**: The assumption is not that *unconditional* Y is normal, the assumption is that the **conditional** Y is normal or that the error is normally distributed
- When we covered MLE, we saw that we made μ a function of explanatory variables: $\mu = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$

Actual data generating process

- I simulated the graph using an indicator variable to shift the distribution; the error is normally distributed: $\epsilon \sim N(0, 4)$
- I created an indicator variable that shifts Y by 3 units

```
clear
set obs 1000
gen x1 = uniform()
gen _x2 = uniform()

* Indicator variable
gen      x2 = 0
replace  x2 = 1 if _x2 > 0.4

capture drop y
gen      y = 10 + 3*x1 + 10*x2 + rnormal(0,4) if x2 ==0
replace  y = 10 + 3*x1 + 13*x2 + rnormal(0,4) if x2 ==1
```

Checking the residuals

- Let's run a model that uses $x1$ and $x2$ as predictors
- Note the value of the parameters and Root MSE

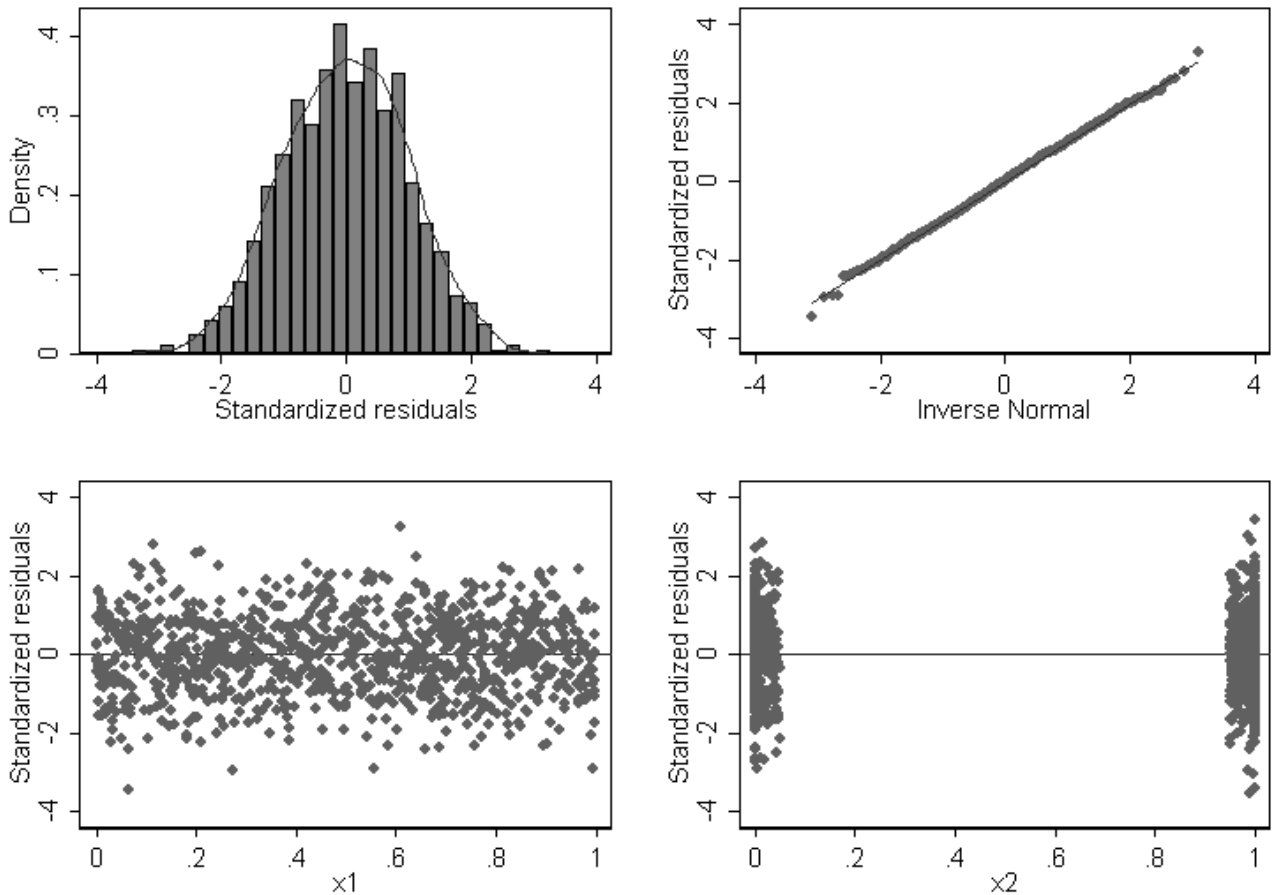
```
reg y x1 x2

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =   1260.54
       Model |  39228.3782         2  19614.1891   Prob > F        =    0.0000
    Residual |  15513.4921       997  15.5601727   R-squared       =    0.7166
-------------+----------------------------------   Adj R-squared   =    0.7160
       Total |  54741.8703       999   54.796667   Root MSE        =    3.9446

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   3.086334   .4247581     7.27   0.000     2.252811    3.919856
          x2 |   12.69668   .2550823    49.77   0.000     12.19612    13.19724
       _cons |   10.01805   .2886578    34.71   0.000     9.451601    10.58449
------------------------------------------------------------------------------

. predict yres10, rstandard
```

Checking the residuals

- We can do the recommended diagnostic plots from last class

```
hist yres10, kdensity saving(yres10.gph, replace)

qnorm yres10, saving(yres101.gph, replace)

scatter yres10 x1, yline(0) saving(yresx1.gph, replace)
* I added the option jitter to see where is the mass of points

scatter yres10 x2, jitter(5) yline(0) saving(yresx2.gph, replace) ///
    title("Jitter = 5")

graph combine yres10.gph yres101.gph yresx1.gph yresx2.gph, ///
              saving(yres10_a.gph, replace)
graph export yres10_a.png, replace
```

Examining the residuals

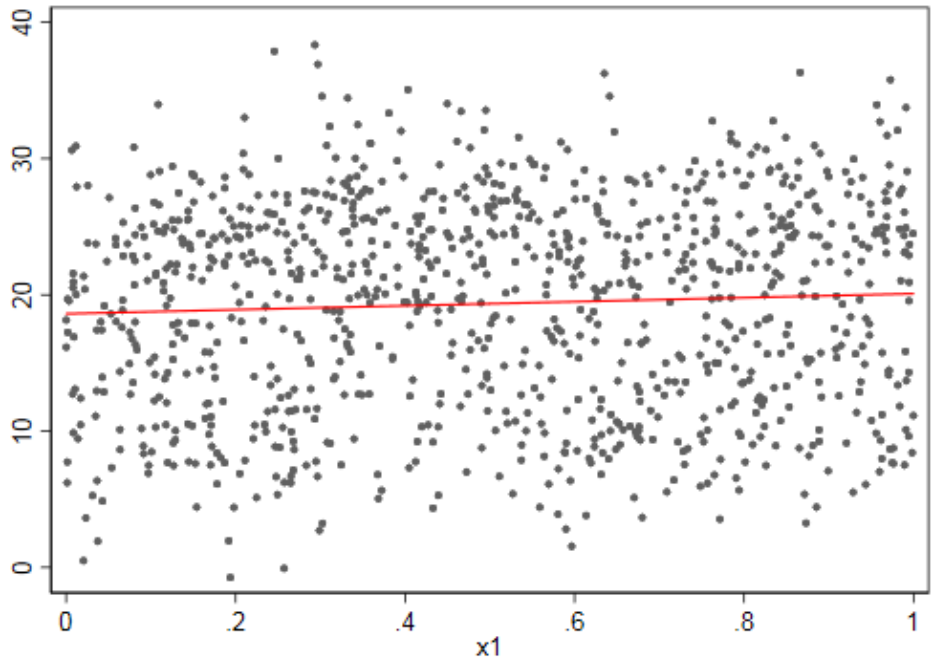


- As expected, all looks perfect, but we started with an outcome variable that didn't look very normal

The previous example illustrates another warning

- I have told you that you **should always do graphs to see how variables are related along with basic statistics**
- Do scatter plots, lowess, non-linear graphs, correlations, etc
- But be aware that simple comparisons **might be misleading**
- It could be that a graph of Y versus X looks nonlinear or linear but that could be due to a third variable
- After building your model, you may want to do an **adjusted plot** (we will see those later; keep that thought in mind)
- In other words, the function form relating X to Y could be, like in the example, influence by a third variable Z
- We really need a conceptual understanding of the problem...

In previous example, suppose x_1 is the variable of interest



- $\beta_1 = 2.5, p - value = 0.002$ and $R^2 = 0.0096$
- What happened? Wrong model? Nope

Is a model with low R^2 a bad model?

- Not necessarily; this is often a common source of confusion
- Remember, the R^2 is the percent of explained variance by the model and also the square of $cor(y, \hat{y})$
- In HSR or social sciences, it is very unlikely that we can **predict outcomes very well**
- A low R^2 **does not imply that we haven't capture correctly** the relationship between Y and one X
- In the previous simulated example, we know Y and x_1 are related and p-value was close to zero
- Not uncommon to see papers with R^2 around 10%. We do worry about omitted variables and causality (see Wooldridge, page 192)
- If you're **building a model for prediction**, a low R^2 would be a concern

Back to diagnostics: Influential observations

- An observation is **influential** if it "affects" the model
- Another way: an observation is influential if its **deletion** causes **substantial** changes in the fitted model
- It could be changes in the estimated parameters $\hat{\beta}_j$ or standard errors, which implies that Wald tests and confidence intervals change
- **Substantial changes are relative and subjective**. It depends on the study question. For sure, if the study conclusions change, that's a substantial change
- Intuitively, **the more data you have the less influential a single observation can be**
- Influential observations tend to be **outliers** in Y or some X

What is an outlier?

- **Not a lot of agreement** on what is an outlier
- For sure, it is a value that is larger of smaller than most of the other values of a variable
- It might be a "mistake" but it could be a valid measurement
- In the linear model (OLS or MLE), we minimize the sum of squares; large errors influence the minimization
- Your textbook has a rather simplistic definition: observations with large standardized residuals are outliers in Y
- Observations with high leverage are outliers in X
- Regardless, the bottom line is that we want to identify points that have a substantial or meaningful impact on results, where "large" is rather subjective. My point of cautionis that "outlier" denotes a problem, but it might not be

Measures of influence

- (**Important**: one key reason for doing an exploratory analysis of the data before running a model is that you will learn about points that you want to check later – after you have a model with covariates)
- **Cook's distance**: It is the the difference between the regression coefficients obtained using all the data and the coefficients obtained deleting the ith observation:

  $$C_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)}$$

- $\hat{y}_{j(i)}$ is the predicted y when observation ith has been removed
- Note that we will have **one Cook's distance for each data point** and we have to estimate n models

DFITS

- Not that different from Cook's distance:
- $DFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{p_{ii}}}$
- Again, one DFITS per observation; it's just the difference in prediction between the model with all the data and the model removing the ith observation
- The numerator scales the distance to take into account the standard error and the leverage of the observation
- There are some rules about what is a large DFITS but better to ignore them
- $|DFITS| > 2\sqrt{\frac{p+1}{n-p-1}}$ are considered influential (rule of thumb really)

Example

- Data from Wooldridge on research and development (RD)
- RD, measured as a percentage of sales (rdintens), is explained by sales (in millions) and profits as a percentage of sales (profmarg)
- The model we want to estimate is:
  $rdintens_i = \beta_0 + \beta_1 sales_i + \beta_2 profmarg_i + \epsilon_i$
  where $\epsilon_i \sim N(0, \sigma^2)$ and i indexes the firms for $i = 1, ..., 32$
- Some basic stats

RD data

```
sum rdintens profmarg sales

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    rdintens |         32     3.26621    1.873701   1.026694   9.421906
    profmarg |         32    9.823068    7.241944  -3.218563   27.18747
       sales |         32    3797.013    7587.992         42      39709

 reg rdintens profmarg sales

      Source |       SS           df       MS      Number of obs   =        32
-------------+----------------------------------   F(2, 29)        =      1.19
       Model |  8.28423732         2  4.14211866   Prob > F        =    0.3173
    Residual |  100.549233        29  3.46721493   R-squared       =    0.0761
-------------+----------------------------------   Adj R-squared   =    0.0124
       Total |   108.83347        31  3.51075711   Root MSE        =     1.862

------------------------------------------------------------------------------
    rdintens |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    profmarg |   .0446166   .0461805     0.97   0.342    -.0498332    .1390664
       sales |   .0000534   .0000441     1.21   0.236    -.0000368    .0001435
       _cons |   2.625261   .5855328     4.48   0.000     1.427712     3.82281
------------------------------------------------------------------------------
```

Cook's distance and DFITS in Stata

- Cook's distance and DFITS are options of predict after regress
- Note the e(sample) in the predict statement

```
* Get Cook's and absoluate value of DFITS
predict ycook if e(sample), cooksd
predict ydfits if e(sample), dfits
gen ydifts_a = abs(ydfits)

* Create id to identify observations
gen id = _n

* Sort by DFITS
gsort -ydifts_a

* Display data top 5
list id rdintens profmarg sales ydifts_a ycook in 1/5
```

Cook's distance and DFITS in Stata

- Since data sorted by DFITS, the top 5 observations are the most influential

```
list id rdintens profmarg sales ydifts_a ycook in 1/5, noc
```

	id	rdintens	profmarg	sales	ydifts_a	ycook
1.	10	3.60	10.46	39709	3.184468	3.13136
2.	1	9.42	4.09	4570.2	1.06444	.2246774
3.	22	6.80	8.99	8995	.4193123	.0538159
4.	9	5.75	12.96	19773	.412281	.057109
5.	17	2.74	27.19	1066.3	.3710729	.0467259

- Observation 10 is more influential that any other. Note the sales value; it's way above any other value

Removing the observation

- We know that removing the observation with id $= 10$ will change the model somehow; that's how DFITS and Cook's distance were calculated

```
. * Drop observation
. reg rdintens profmarg sales if id ~= 10

      Source |       SS           df       MS      Number of obs   =        31
-------------+----------------------------------   F(2, 28)        =      2.92
       Model |  18.7880289         2  9.39401445   Prob > F        =    0.0702
    Residual |  89.9330615        28  3.21189505   R-squared       =    0.1728
-------------+----------------------------------   Adj R-squared   =    0.1137
       Total |   108.72109        30  3.62403635   Root MSE        =    1.7922

------------------------------------------------------------------------------
    rdintens |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    profmarg |   .0478411   .0444831     1.08   0.291    -.0432784    .1389605
       sales |   .0001856   .0000842     2.20   0.036     .0000131    .0003581
       _cons |   2.296851   .5918045     3.88   0.001     1.084594    3.509107
------------------------------------------------------------------------------
```

- Note that sales is now significant; the R^2 is higher

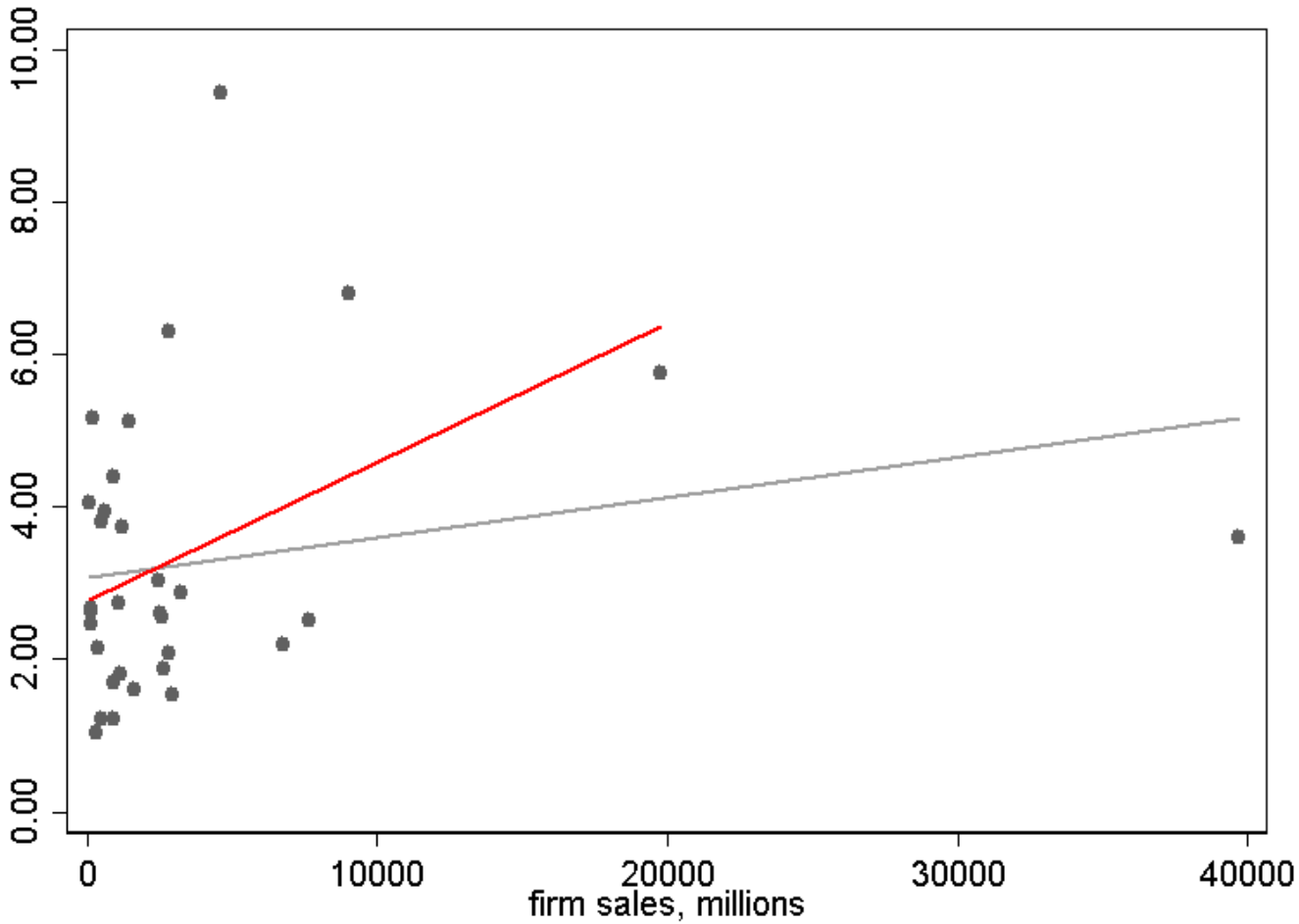
We can see the same in a graph (**ignoring profmarg**)

- Three graphs in one:

```
scatter rdintens sales || lfit rdintens sales || ///
        lfit rdintens sales if id ~= 10, color(red) legend(off) ///
saving(dropone.gph, replace)

graph export dropone.png, replace
```

- Lfit is the same as estimating a model, getting the predicted values, and plotting them. Saves typing

Graph



- Now, a pickle: Is the firm with sales of \$20K an influential observation now?

Big picture

- Understanding the observations that might change results is of course important
- One advantage of DFITS and Cook's distance is that they use the residuals
- The residuals are the unexplained part; the effect of explanatory variables were taken into account
- In exploratory data analysis, we often work with two variables at a time
- **The more data you have, the less influential each observation is**, which sort of explains why with so much observational data we don't often use these measures of influence...

Summary

- It is important to check for influential observations but with large samples less important
- It is even more important to perform an exploratory analysis before running any model
- Use summarize, plots, histograms, tabulations...
- It is a gray area. I would say that most researchers in HSR do not report dropped observations
- In some cases knowledge about the subject will help you evaluate if a value can be considered an outlier