Stat 104: Quantitative Methods

Class 32: Understanding the Regression Output

Introduction

- All we have done so far is learned how to fit a line to some (X,Y) data. There really hasn't been very much statistics involved at all.
- In this section we detail some statistical theory necessary to discuss how good a line we have and how accurate predictions using the fitted line will be.

Exact versus Inexact Relationships:

Many relationships in science are exact:

- distance = rate x time
- $E = mc^2$
- Force = mass x acceleration

But most relationships in business are inexact:

How do we express uncertainty in our relationships ?

Here is the Accord regression output again:

```
> summary(fit)

Call:
lm(formula = Price ~ Odometer)

Residuals:
   Min     1Q Median     3Q    Max
-730.3 -235.0    1.3  187.7  691.2

Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept) 17066.76607   169.02464   101.0   <2e-16 ***
Odometer       -0.06232     0.00462   -13.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303 on 98 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.65,      Adjusted R-squared:  0.647
F-statistic:  182 on 1 and 98 DF,  p-value: <2e-16
```


For example, we know that there isn't an **exact relationship** between mileage of a car and its price (how do we know this, by the way?)

price = \$17067-\$.06(odometer)

That is, not every Accord with 30000 miles will sell for \$15267. Some will sell for more, and some houses will sell for less.

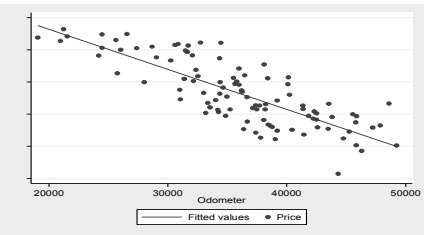
A more realistic statement is that

Average Car Price = \$17067-\$.06(odometer)

This is a main point about regression: we model the **average of something** rather than the something itself.

The Average Line

- The regression line should be viewed as the average value of Y for a given X, or in symbols E(Y|X).



7

In symbols, this line of averages can be expressed as:

$$E(Y|X) = \beta_0 + \beta_1 X$$

where E(Y|X) is the expected value (average) of Y for a given X value.

In our Accord example, we might think of

$$E(Y \mid X)$$

as the average price of cars with mileage X.

8

Another way of explaining this concept is to write the model as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$or$$

$$Y = E(Y|X) + \varepsilon$$

where ε refers to some random "noise"

if there were no noise in the system, what would be the relationship between X and Y ?

9

To describe what the likely value of ε may be, we let it be a normally distributed random variable:

$$\varepsilon \sim N(0, \sigma^2)$$

mean of ε is 0.

Sometimes *Y* will be above the line, sometimes below.

variance of ε is σ^2

If σ is small, ε will tend to be small (close to 0).
If σ is big, ε could be big (far from 0).

10

ASSUMPTIONS of the Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Bo + B1 X — the part of *Y* related to *X*

ε — the part of *Y* unrelated to *X*: $\varepsilon \sim N(0, \sigma^2)$

Note: the distribution of ε does not depend on *X*

ε is *independent* of *X*.

11

Estimates of the Parameters

We estimate β_0 by b_0.

We estimate β_1 by b_1.

$$b_0 = \bar{Y} - b_1 \bar{X} \qquad b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

the formulas show us how the estimates depend on the data.

12

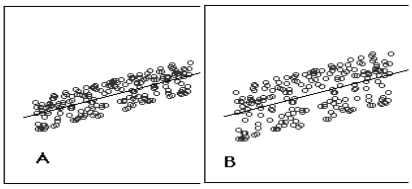
Note !!!!

b_0 is not β_0 $\quad b_1$ is not β_1 $\quad e_i$ is not ε_i

Greek letters refer to population quantities-they are unobserved and we are trying to estimate them.

Estimating σ

- It is important to get a handle on σ because it is directly related to how well our line fits and how good our predictions are.



Var(ε) = σ2

Estimating σ

How do we estimate σ?

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$$

s_e^2 is our estimate of σ^2

$s_e = \sqrt{s_e^2}$ is our estimate of σ

Accord Data


```
> summary(fit)

Call:
lm(formula = Price ~ Odometer)

Residuals:
   Min     1Q Median     3Q    Max
-730.3 -235.0    1.3  187.7  691.2

Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) 17066.76607   169.02464   101.0   <2e-16 ***
Odometer       -0.06232     0.00462   -13.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303 on 98 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.65,      Adjusted R-squared:  0.647
F-statistic:  182 on 1 and 98 DF,  p-value: <2e-16
```

s_e

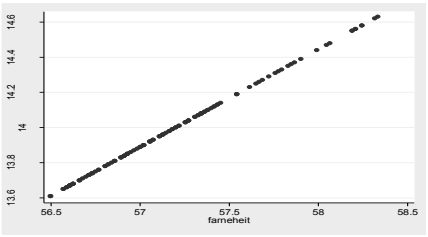
Pharmex Data


```
> fit=lm(mydata$Sales~mydata$Promote)
> summary(fit)

Call:
lm(formula = mydata$Sales ~ mydata$Promote)

Residuals:
    Min      1Q  Median      3Q     Max
-17.307  -4.954  -0.454   5.121  17.742

Coefficients:
             Estimate Std. Error t value    Pr(>|t|)
(Intercept)    25.126     11.883    2.11        0.04 *
mydata$Promote  0.762      0.121    6.30 0.000000086 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.39 on 48 degrees of freedom
Multiple R-squared:  0.453,     Adjusted R-squared:  0.442
F-statistic: 39.7 on 1 and 48 DF,  p-value: 0.000000086
```

s_e

Example: Global Temperature

- Temperature in Celsius versus Fahrenheit.

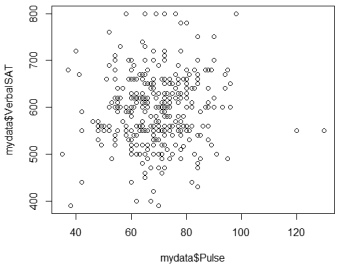


What should the R-sq be?

Example: Temp Regression Output


```
> fit=lm(mydata$celsius~mydata$farneheit)
> summary(fit)

Call:
lm(formula = mydata$celsius ~ mydata$farneheit)

Residuals:
      Min        1Q    Median        3Q       Max
-2.67e-15 -9.75e-16 -1.70e-17  8.22e-16  7.90e-15

Coefficients:
                Estimate Std. Error           t value Pr(>|t|)
(Intercept)    -1.78e+01   1.50e-14 -1184739660600820   <2e-16 ***
mydata$farneheit  5.56e-01   2.62e-16  2116671180489917   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.38e-15 on 129 degrees of freedom
Multiple R-squared:     1,      Adjusted R-squared:     1
F-statistic: 4.48e+30 on 1 and 129 DF,  p-value: <2e-16
```

Example: Verbal SAT versus Pulse



What should the R-sq be?

Example: Verbal SAT versus Pulse


```
> fit=lm(mydata$VerbalSAT~mydata$Pulse)
> summary(fit)

Call:
lm(formula = mydata$VerbalSAT ~ mydata$Pulse)

Residuals:
    Min      1Q  Median      3Q     Max
-204.38  -47.94    1.17   46.51  210.96

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   563.207     22.564   24.96   <2e-16 ***
mydata$Pulse    0.445      0.319    1.39     0.16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.1 on 360 degrees of freedom
Multiple R-squared:  0.00537,   Adjusted R-squared:  0.00261
F-statistic: 1.94 on 1 and 360 DF,  p-value: 0.164
```


Approximate Prediction Intervals

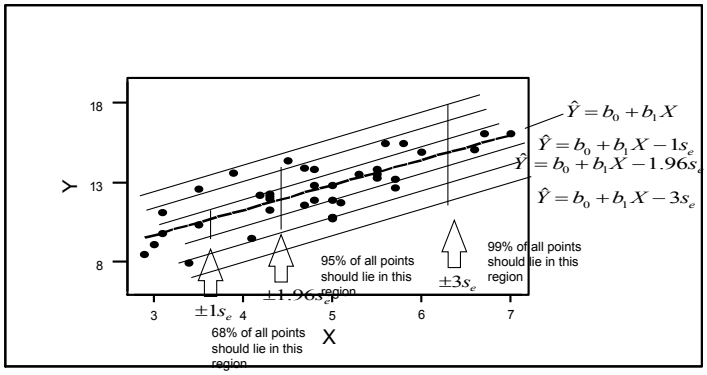
We can use the value of s_e to form bands around the regression line:

68% of the Y values should lie within the interval $b_0 + b_1 X \pm 1 s_e$

95% of the Y values should lie within the interval $b_0 + b_1 X \pm 1.96 s_e$

99% of the Y values should lie within the interval $b_0 + b_1 X \pm 3 s_e$

these intervals are sometimes called "plug-in" intervals



The Accord data again


```
> summary(fit)

Call:
lm(formula = Price ~ Odometer)

Residuals:
   Min     1Q Median     3Q    Max
-730.3 -235.0    1.3  187.7  691.2

Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) 17066.76607   169.02464   101.0   <2e-16 ***
Odometer       -0.06232     0.00462   -13.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303 on 98 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.65,      Adjusted R-squared:  0.647
F-statistic:  182 on 1 and 98 DF,  p-value: <2e-16
```

We are roughly 95% confident that the (average) price of an Accord with 50,000 miles is in the interval

$$17066 - 0.06(50000) \pm 1.96(303) = (13472, 14660)$$

Price and MPG


```
> fit=lm(mydata$price~mydata$mpg)
> summary(fit)

Call:
lm(formula = mydata$price ~ mydata$mpg)

Residuals:
   Min     1Q Median     3Q    Max
 -3184  -1887   -960   1360   9670

Coefficients:
             Estimate Std. Error t value           Pr(>|t|)
(Intercept)  11253.1     1170.8    9.61 0.000000000000015 ***
mydata$mpg    -238.9       53.1   -4.50 0.000025461312051 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2620 on 72 degrees of freedom
Multiple R-squared:  0.22,      Adjusted R-squared:  0.209
F-statistic: 20.3 on 1 and 72 DF,  p-value: 0.0000255
```

Reviewing s_e Intervals

- IF we didn't have any x variables and had to guess a new value of Y for a new value of X, we would probably use the mean and our prediction interval

$$\bar{y} \pm 2s_Y$$

- IF we have X data, we should have better predictions, so our interval would be

$$\hat{y} \pm 2s_e$$

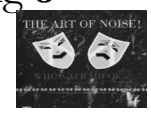
compare

Yhat

For a good regression model, Se <<< Sy

The art of noise – understanding ε

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



•What is the role of noise in the regression model ?

•If there was no noise, what type of relationship would exist ?

•As more and more noise is added, what do you think happens to our estimates ?

Match the plot with the model

$Y = 1 + 5X + \varepsilon$

$\varepsilon \sim N(0,0)$

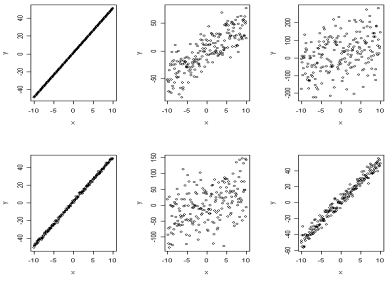
$\varepsilon \sim N(0,1)$

$\varepsilon \sim N(0,5)$

$\varepsilon \sim N(0,20)$

$\varepsilon \sim N(0,50)$

$\varepsilon \sim N(0,100)$



How Sure Are We ? The Standard Errors

We now have an estimator for all three of our model parameters.

The estimates depend on the data. They try to "guess" the true parameter values from the information in the data.

We need to know if our estimates are "good".

Can we be fairly sure that the true values are not too far from our guesses?

Standard Errors for the Accord Data

```
> summary(fit)

Call:
lm(formula = Price ~ Odometer)

Residuals:
    Min      1Q  Median      3Q     Max
-730.3 -235.0    1.3   187.7   691.2

Coefficients:                    s_b0
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) 17066.76607   169.02464   101.0   <2e-16 ***
Odometer       -0.06232     0.00462   -13.5   <2e-16 ***
---                      s_b1
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303 on 98 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.65,      Adjusted R-squared:  0.647
F-statistic:  182 on 1 and 98 DF,  p-value: <2e-16
```

Nomenclature review

b_0 is a guess of β_0

b_1 is a guess of β_1

these are sometimes called *point estimates*

A point estimate without some idea of precision is useless- famous proverb

s_{b_0} amount of uncertainty in our estimate of β_0

s_{b_1} amount of uncertainty in our estimate of β_1

Martha Stewart does Statistics

s_{b_0} { small, GOOD, lots of info in data about β_0
Large, BAD, little info about β_0

s_{b_1} { small, GOOD, lots of info in data about β_1
Large, BAD, little info about β_1

Match output with model

$Y = 1 + 5X + \varepsilon$

$\varepsilon \sim N(0,0)$
$\varepsilon \sim N(0,1)$
$\varepsilon \sim N(0,5)$
$\varepsilon \sim N(0,20)$
$\varepsilon \sim N(0,50)$
$\varepsilon \sim N(0,100)$

Estimate of error standard deviation: NaN
Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	1	NaN	199	NaN	NaN
Slope	5	NaN	199	NaN	NaN

Estimate of error standard deviation: 0.97156674
Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	1.0539687	0.06852903	199	15.379885	<0.0001
Slope	5.0075765	0.011810671	199	423.98752	<0.0001

Estimate of error standard deviation: 4.851036
Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	1.4927968	0.3421657	199	4.3627687	<0.0001
Slope	5.0336647	0.05897072	199	85.35871	<0.0001

Estimate of error standard deviation: 20.000694
Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	1.1275297	1.4107403	199	0.7992469	0.4251
Slope	5.060032	0.24313474	199	20.811636	<0.0001

Estimate of error standard deviation: 48.82322
Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	2.8995328	3.4437242	199	0.84197587	0.4008
Slope	5.181193	0.5935104	199	8.729743	<0.0001

Estimate of error standard deviation: 97.64781
Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	-8.000299	6.8875456	199	-1.1615602	0.2468
Slope	3.8545618	1.1870375	199	3.2472115	0.0014

As more noise enters into the model, what happens to our estimates ? In terms of precision and actual value ?

Estimates get worse. Standard errors get larger.

How close is b_1 to the truth ?

A confidence interval for β_1 is given

$$b_1 \pm 1.96(s_{b_1})$$

Called the standard error

where

$$Var(b_1) = s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2}.$$

Examine the formula:

$$Var(b_1) = s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2}.$$

says

n small $\Rightarrow$ bad

s_e big $\Rightarrow$ bad

s_x^2 small $\Rightarrow$ bad

Given a choice, you want the Xs spread out.
So we want a large Sx i.e. variation in X

How close is b_0 to the truth ?

A confidence interval for β_0 is given

$$b_0 \pm 1.96(s_{b_0})$$

Called the standard error

where

$$Var(b_0) = s_{b_0}^2 = s_e^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}\right)$$

Review : 3 Step Plan

1) Model: $Y = \beta_0 + \beta_1 X + \varepsilon$ inexact relationship

$\varepsilon \sim N(0, \sigma^2)$ Noise

2) Data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

3) Estimate: β_0, β_1, σ Truth

b_0, b_1, s Guesses

Review : Interval Estimates

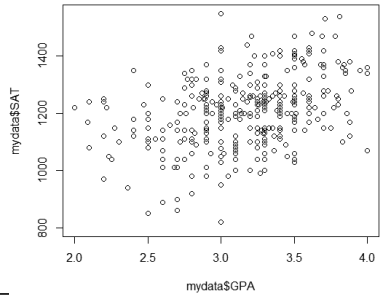
$$b_1 \pm 1.96(s_{b_1})$$

is a 95% confidence interval for β_1

We are 95% confident the true value of β_1 is in the interval

$$(b_1 - 1.96s_{b_1}, b_1 + 1.96s_{b_1})$$

SAT and High School GPA



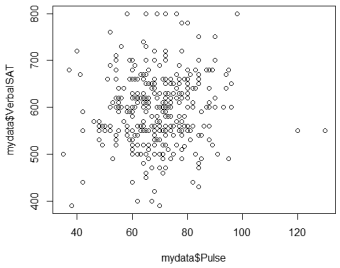
R Ouput

```
> fit=lm(mydata$SAT~mydata$GPA)
> summary(fit)

Call:
lm(formula = mydata$SAT ~ mydata$GPA)

Residuals:
    Min      1Q  Median      3Q     Max
-367.2  -71.6   -1.6   69.8  362.8

Coefficients:
            Estimate Std. Error t value          Pr(>|t|)
(Intercept)    845.2       48.3   17.52           < 2e-16 ***
mydata$GPA     114.0       15.2    7.52 0.00000000000048 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112 on 343 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.142,     Adjusted R-squared:  0.139
F-statistic: 56.6 on 1 and 343 DF,  p-value: 0.00000000000482
```

```
> confint(fit)
                2.5 % 97.5 %
(Intercept) 750.303 940.11
mydata$GPA   84.188 143.82
```

$$(114 - 1.96(15.2), 114 + 1.96(15.2))$$

Example: Verbal Sat and Pulse



Regression Output

```
> fit=lm(mydata$VerbalSAT~mydata$Pulse)
> confint(fit)
                2.5 %    97.5 %
(Intercept)  518.83294 607.5808
mydata$Pulse  -0.18289   1.0736
```

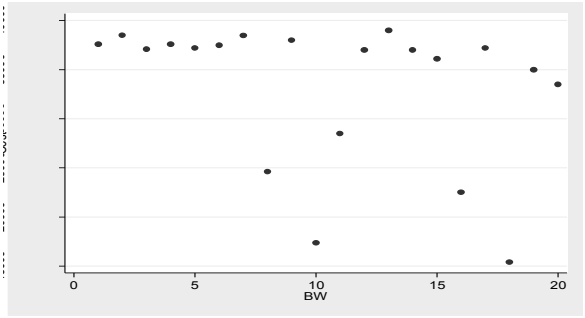
Interpretation?

True B1 is between -0.18289 and 1.0736
There's a chance the true B1 is 0
Then Y = B0 + ε
i.e. Y is simply noise.

Another example: 1992 BW Data

School	BW	UNSNWR	bystudents	byfirms	Cost	Salary
Northwester	1	4	3	1	37600	70200
Chicago	2	6	10	4	38500	68600
Harvard	3	2	12	3	37100	84960
Wharton	4	3	15	2	37600	72200
Michigan	5	7	9	6	37200	58110
Dartmouth	6	10	1	12	37500	74260
Stanford	7	1	5	7	38480	82860
Indiana	8	18	6	8	24600	49070
Columbia	9	8	18	5	38000	66620
N.Carolina	10	16	8	11	17360	55500
Virginia	11	11	2	15	28500	65280
Duke	12	9	7	14	37000	59870
MIT	13	5	14	10	39000	73000
Cornell	14	12	4	17	37000	59940
NYU	15	17	16	13	36100	56730
UCLA	16	14	11	16	22500	64540
Carnegie	17	15	23	9	37200	56980
Berkeley	18	13	13	19	15400	65500
Vanderbilt	19	19	19	20	35000	47320
Washington	20	20	24	18	33500	48200

Cost versus BW Rank



Regression Output


```
> confint(fit)
                2.5 %    97.5 %
(Intercept) 30828.4 44726.95
mydata$BW   -1010.7   149.57
> summary(fit)

Call:
lm(formula = mydata$Cost ~ mydata$BW)

Residuals:
   Min     1Q Median     3Q    Max
-16112   -946   1944   4487   6819

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)    37778       3308   11.42 0.0000000011 ***
mydata$BW       -430        276   -1.56         0.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7120 on 18 degrees of freedom
Multiple R-squared:  0.119,     Adjusted R-squared:  0.0701
F-statistic: 2.43 on 1 and 18 DF,  p-value: 0.136
```

Any conclusions ?

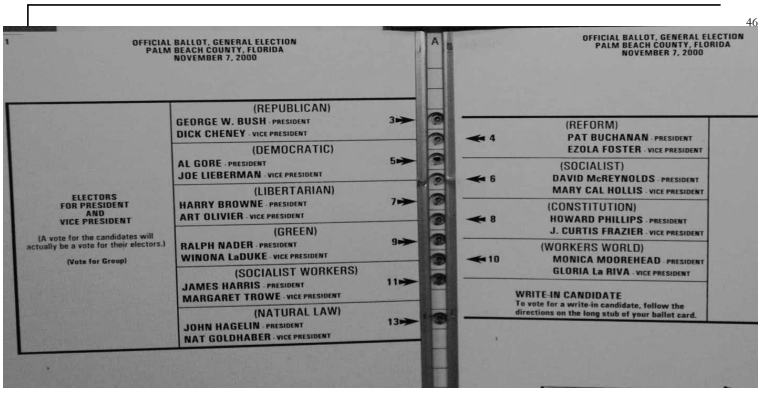
Example: The 2000 Florida Vote

In this example, we examine county-by-county data on presidential voting during the 2000 election.

We take a look at the two variables Buchanan and Bush, defined as:

Buchanan the # of votes for Buchanan in the 2000 election (in a given county)

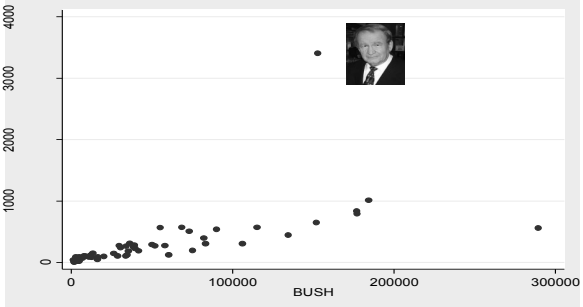
Bush the # of votes Bush received (in a given county)



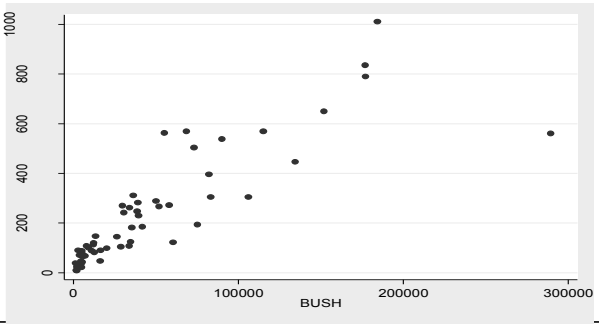
The dataset contains data on a total of 67 counties in Florida. One contention about the 2000 Florida vote is that due to the (allegedly) confusing design of the butterfly ballot, Buchanan received a lot more votes in Palm Beach county then were intended.

To investigate this, we will first examine the relationship between the variables Buchanan and Bush for the **66 other counties in florida**

All Counties (including Palm Beach)



Data without Palm Beach



Regression Output (wo Palm Beach)


```
> fit=lm(mydata$Buchanan~mydata$Bush,subset = -50)
> summary(fit)

Call:
lm(formula = mydata$Buchanan ~ mydata$Bush, subset = -50)

Residuals:
   Min     1Q Median     3Q    Max
-513.1  -48.3  -13.9   41.7  305.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.081280  17.293760    3.82   0.0003 ***
mydata$Bush  0.003478   0.000249   13.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112 on 64 degrees of freedom
Multiple R-squared:  0.753,     Adjusted R-squared:  0.749
F-statistic:  195 on 1 and 64 DF,  p-value: <2e-16

> confint(fit)
                  2.5 %      97.5 %
(Intercept) 31.5330240 100.6295368
mydata$Bush  0.0029812   0.0039757
```


Now there were 152846 votes for Bush in Palm Beach County. According to our regression model, (assuming that Palm Beach County voters behaved like all other voters), we should have seen about

66.08 + 0.00348 (152846) = 597 votes for Buchanan.

Of course, we need to bound our guess:

597 +/- 1.96(112) = (377,817)

So, if Palm Beach County was like the other counties in Florida, Buchanan should have received between 377 and 817 votes.

The actual number of Buchanan votes was 3407!

Some people argued that Buchanan did extraordinarily well in Palm County because there were a lot of registered “Independent” voters and Buchanan had done well there in prior elections. To allow for this possibility, we need to do a multiple regression on number of registered Republicans, number of registered Independents, and number of votes that Buchanan received in the 1996 Republican primary in Florida. We will show how to do this in a few weeks.