Statistics

Stat 104: Quantitative Methods for Economists

Class 33 Regression Hypothesis Testing

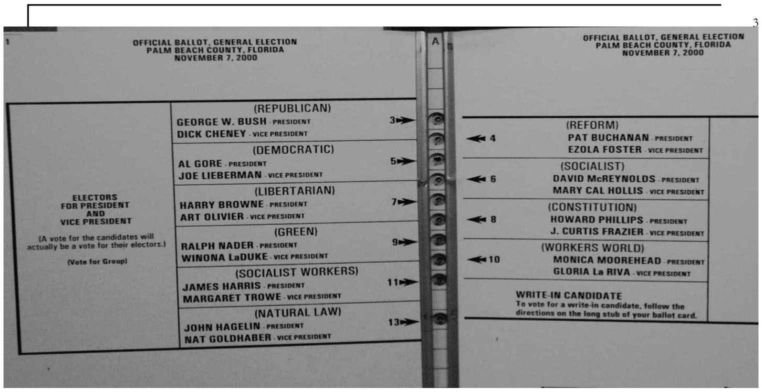
Example: The 2000 Florida Vote

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)

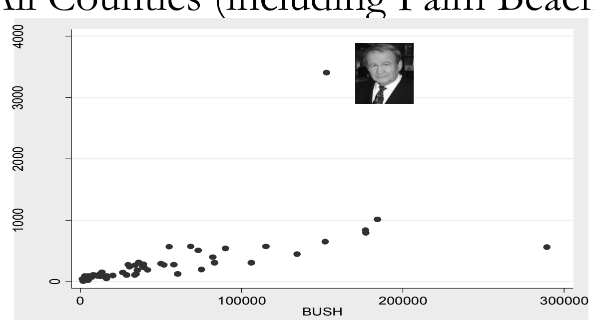
Bush the # of votes Bush received (in a given county)



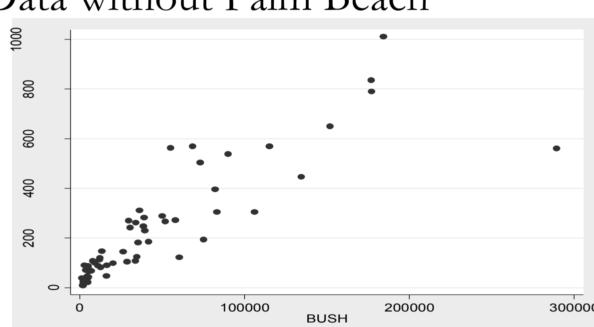
The dataset contains data on a total of 67 counties in Florida. One contention about the 2000 Florida vote is that due to the (allegedly) confusing design of the butterfly ballot, Buchanan received a lot more votes in Palm Beach county then were intended.

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.

i.e. other explanatory variables were not accounted for

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

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

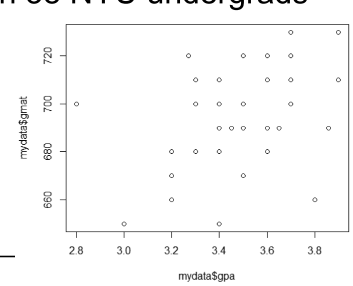
is a 95% confidence interval for β_1

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

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

GMAT and Undergrad GPA

- Data on 38 NYU undergrads



R Ouput

13

```
> fit=lm(mydata$gmat~mydata$gpa)
> summary(fit)

Call:
lm(formula = mydata$gmat ~ mydata$gpa)

Residuals:
   Min     1Q Median     3Q    Max
-44.86 -10.02   3.23  11.14  31.52

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    587.4       46.4   12.66    8e-15 ***
mydata$gpa      30.9       13.4    2.31    0.027 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.3 on 36 degrees of freedom
Multiple R-squared:  0.129,     Adjusted R-squared:  0.105
F-statistic: 5.35 on 1 and 36 DF,  p-value: 0.0265

> confint(fit)
               2.5 %  97.5 %
(Intercept) 493.3374 681.487
mydata$gpa    3.8109  58.004
```

$$(30.9 - 1.96(13.4), 30.9 + 1.96(13.4))$$

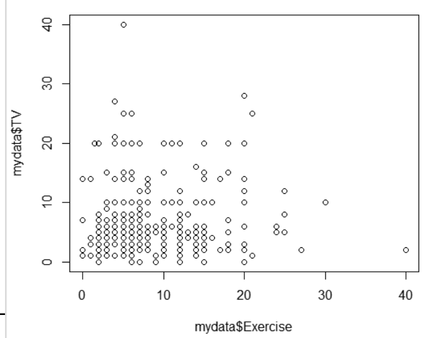
Price and MPG

14

```
> 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

> confint(fit)
             2.5 %   97.5 %
(Intercept) 8919.1 13587.03
mydata$mpg  -344.7  -133.09
```

Example: Exercise and TV Watching

15



Regression Output

16

```
> fit=lm(mydata$TV~mydata$Exercise)
> summary(fit)

Call:
lm(formula = mydata$TV ~ mydata$Exercise)

Residuals:
   Min     1Q Median     3Q    Max
 -6.62  -3.52  -1.54   2.73  33.53

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.42448    0.55083   11.66   <2e-16 ***
mydata$Exercise  0.00957    0.05139    0.19     0.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.6 on 358 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  9.69e-05,  Adjusted R-squared:  -0.0027
F-statistic: 0.0347 on 1 and 358 DF,  p-value: 0.852

> confint(fit)
                    2.5 %  97.5 %
(Intercept)      5.341214 7.50775
mydata$Exercise -0.091496 0.11064
```

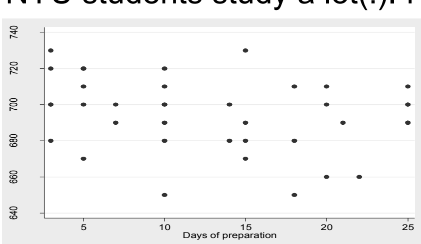
Interpretation?

β1 = 0.00957 so seems like theres a positive relationship
But that does not mean much.
When we analyze the confidence interval and get -0.091496 to 0.11064,
we realize there isn't really a relationship. Could be positive, negative or even zero

GMAT and Number of Prep Days

17

- These NYU students study a lot(!). Not.



Interpret: the power of studying

18

```
> fit=lm(mydata$gmat~mydata$prep)
> coef(fit)
(Intercept) mydata$prep
  703.30864    -0.67961
> confint(fit)
               2.5 %    97.5 %
(Intercept) 689.9623 716.6549
mydata$prep  -1.5729   0.2137
```

Again here, confidence interval paints a clearer picture about the relationship than the β1

Market Model (again)

19

In the simplest sense, the "market model" assumes that

α β i.e. risk

$$Stockreturn_t = \beta_0 + \beta_1 Indexreturn_t + \varepsilon_t$$

The finance people call β_1 Beta (go figure).

Beta=0 : cash under the mattress

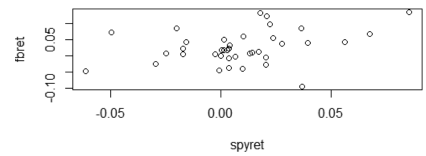
Beta=1 : same risk as the market

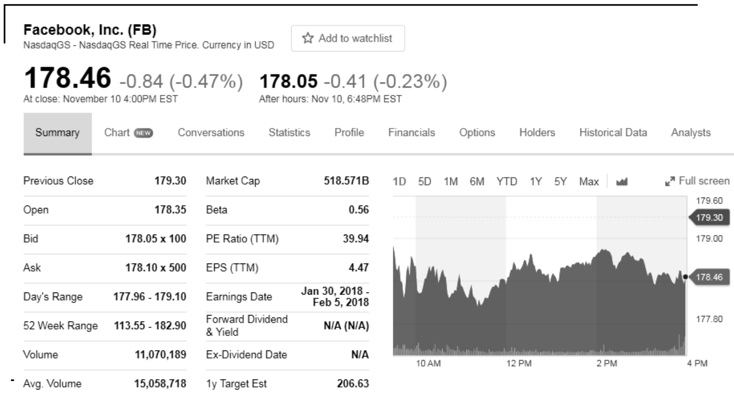
Beta<1 : safer than the market

Beta >1: riskier than the market

20

We will examine the market model for the stock Facebook (FB), using the S&P 500 as a proxy for the market. The returns are monthly from the last three years.





Facebook, Inc. (FB)
NasdaqGS - NasdaqGS Real Time Price. Currency in USD
Add to watchlist

178.46 -0.84 (-0.47%) **178.05** -0.41 (-0.23%)
At close: November 10 4:00PM EST After hours: Nov 10, 6:48PM EST

Summary | Chart | Conversations | Statistics | Profile | Financials | Options | Holders | Historical Data | Analysts

Previous Close	179.30	Market Cap	518.571B
Open	178.35	Beta	0.56
Bid	178.05 x 100	PE Ratio (TTM)	39.94
Ask	178.10 x 500	EPS (TTM)	4.47
Day's Range	177.96 - 179.10	Earnings Date	Jan 30, 2018 - Feb 5, 2018
52 Week Range	113.55 - 182.90	Forward Dividend & Yield	N/A (N/A)
Volume	11,070,189	Ex-Dividend Date	N/A
Avg. Volume	15,058,718	1y Target Est	206.63

Based on the β 0.605 you'd think FB was safer than the market

From R,

22

```
> getSymbols("FB",from="2014-10-31")
[1] "FB"
> getSymbols("SPY",from="2014-10-31")
[1] "SPY"
> spyret=monthlyReturn(Ad(SPY))
> fbret=monthlyReturn(Ad(FB))

> fit=lm(fbret~spyret)

> coef(fit)
(Intercept)      spyret
 0.01915543  0.60589069

> confint(fit)
                  2.5 %    97.5 %
(Intercept) 0.002313057 0.0359978
spyret      0.030077416 1.1817040
```

What can we say about the Beta for FB?

The confidence interval paints a whole other picture.
β could be as low as 0.03 i.e. no relation to market
or as high as 1.18 i.e. riskier than the market.

Hypothesis Tests for the Regression Model

23

We will discuss tests about β_1. Tests on β_0 work in exactly the same way.

Suppose you want to test whether β_1 equals a proposed value:

Null Alternative

$H_0: \beta_1 = \beta_1^*$ $H_a: \beta_1 \neq \beta_1^*$

For example, if we want test whether *X* affects Y, we would test whether β_1=0.

Huh??

Decision Rules for Testing the Slope:

24

$$T = \frac{b_1 - \beta_1^*}{s_{b_1}}$$

$H_0: \beta_1 = \beta_1^*$ If $|T| > 1.96$ reject H_o

$H_a: \beta_1 \neq \beta_1^*$

$H_0: \beta_1 = \beta_1^*$ If $T < -1.64$ reject H_o

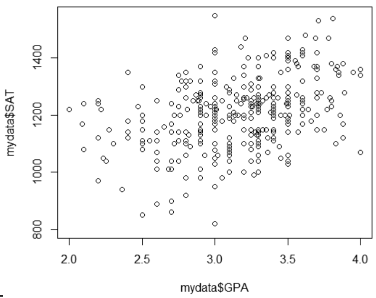
$H_a: \beta_1 < \beta_1^*$

$H_0: \beta_1 = \beta_1^*$ If $T > 1.64$ reject H_o

$H_a: \beta_1 > \beta_1^*$

Caution! If # obs < 30 must use t distribution **so use p-values!**

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: 0.000000000000482
```


$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$$

The test statistic is

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{114 - 0}{15.2} = 7.5$$

```
            Estimate Std. Error t value
(Intercept)    845.2       48.3   17.52
mydata$GPA     114.0       15.2    7.52
```

and

$$7.5 > 1.96$$

so we reject the null hypothesis.

Note: the hypothesis that the slope equals zero is tested so often that R **automatically** prints out the appropriate t statistic. The t for testing the intercept equal to 0 is also printed.

$H_0 : \beta_0 = 0 \quad \frac{b_0}{s_{b_0}}$

if | t | > 1.96 → Reject Ho

```
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 ***
```

$H_0 : \beta_1 = 0 \quad \frac{b_1}{s_{b_1}}$

P-values

the R commands are for Ho : β1 = 0

We now test the hypothesis that the effect is 120 points for each 1 unit increase in GPA:

$$H_0 : \beta_1 = 120 \qquad H_a : \beta_1 \neq 120$$

The t statistic is

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{114 - 120}{15.2} = -0.39$$

Now |-0.39| is less than 1.96 so we fail to reject the null hypothesis; the effect of a unit rise of GPA on SAT score might be 120.

Some Notes

- There is a routine in the `FSA` package to do hypothesis testing on the regression coefficients:

```
> library(FSA)
> hoCoef(fit)
 term Ho Value Estimate Std. Error      T  df             p value
    2        0      114     15.159 7.5205 343 0.0000000000004823

> hoCoef(fit,bo=120)
 term Ho Value Estimate Std. Error       T  df p value
    2      120      114     15.159 -0.3955 343 0.69272
```

Note on Hypothesis Testing

- We will make life easy for this and the regression hypothesis will be always two sided of the form

$$H_0 : \beta_1 = ? \qquad H_a : \beta_1 \neq ?$$

- There are then three ways one could test this hypothesis; get familiar with at least one:
  - Test statistic — | t | > 1.96 — reject
  - P-value — p-val < 0.05. — reject
  - Confidence Interval — if 0 not in conf. interval reject

Recap: Regression Modeling So Far

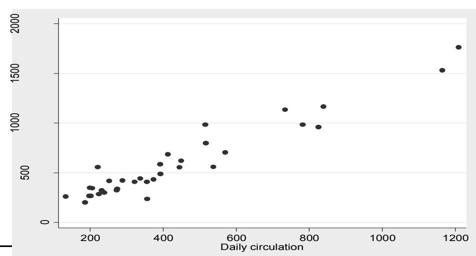
- In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands

Newspaper	Daily	Sunday
Baltimore Sun	391.952	488.506
Boston Globe	516.981	798.298
Boston Herald	355.628	235.084
Charlotte Observer	238.555	299.451
Chicago Sun Times	537.780	559.093
Chicago Tribune	733.775	1133.249
Cincinnati Enquirer	198.832	348.744
Denver Post	252.624	417.779
Des Moines Register	206.204	344.522
Hartford Courant	231.177	323.084

Data snapshot

Recap: Regression Modeling So Far

- Start with data where you think a linear relationship exists



Examine the Regression Output

- What is the value of R-squared? Is it low or high? 0.918 - It seems high
- If we used this model for predictions, how accurate would we be? +/- 2s = ± 2 * 109

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.8356    35.8040    0.39      0.7
mydata$daily    1.3397     0.0708   18.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109 on 32 degrees of freedom
Multiple R-squared:  0.918,     Adjusted R-squared:  0.915
F-statistic:  359 on 1 and 32 DF,  p-value: <2e-16

> sd(mydata$sunday)
[1] 376.42
```

Do we even need "x" in the model?

- How do we determine if we need the x variable in the model?
- Check to see if the |t| > 1.96 or p-value < .05, or CI does not span 0.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.8356    35.8040    0.39      0.7
mydata$daily    1.3397     0.0708   18.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109 on 32 degrees of freedom
Multiple R-squared:  0.918,     Adjusted R-squared:  0.915
F-statistic:  359 on 1 and 32 DF,  p-value: <2e-16

> confint(fit)
                2.5 %  97.5 %
(Intercept)  -59.0947 86.7660
mydata$daily   1.1956  1.4838
```

Make a prediction

- The particular newspaper that is being considered as a candidate for a Sunday edition has a Daily circulation of 600,000. Provide an interval estimate for the predicted Sunday circulation of this newspaper.
- Prediction=13.835+1.34(600)=817.835 (thousand)
- We can make a prediction interval as follows
- 817.835 +/- 1.96(109.42)=(603.37,1032.3)

37

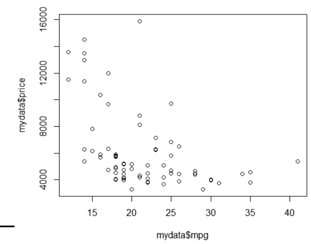
Reporting your results:

- A regression on the basis of a random sample of 34 newspapers indicates a strong relationship between daily circulation and Sunday edition sales. Each additional daily circulation of 1000 copies results in an increase of Sunday sales by 1340 copies.
- This effect is substantial and statistically significant.
- The regression line explains 91.8% of the variation in the Sunday circulation.
- A newspaper with daily circulation of 600,000 is expected to have a Sunday circulation of 817,835.

38

Example: Auto Data Again

- Price versus MPG
- Is regression appropriate?

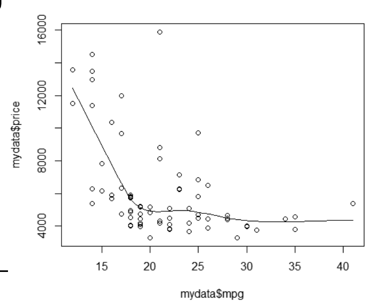


39

Looking Ahead: Lowess

- Scatterplot Smoothing

```
scatter.smooth(mydata$mpg,mydata$price)
```



40

Regression Output (ignore issues)

```
> 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
```

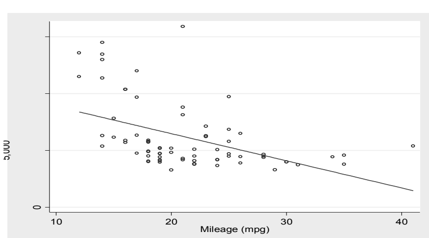
Interpretation?

Ho : β1 = 0
H1: β1 ≠ 0
p-value = 0.00002546 < 0.05 reject
| t | = 4.5 > 1.96 reject

41

Will Soon Learn Diagnostics

- Not a great fitting line to the data



42

Looking ahead (multiple regression)

- Add a quadratic..better model (why??)

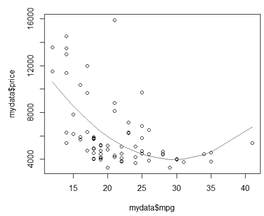
```
> fit=lm(mydata$price~mydata$mpg+I(mydata$mpg^2))
> summary(fit)


Coefficients:
                Estimate Std. Error t value     Pr(>|t|)
(Intercept)     22716.48    3366.58    6.75 0.0000000034 ***
mydata$mpg      -1265.19     289.54   -4.37 0.0000416736 ***
I(mydata$mpg^2)    21.36       5.94    3.60      0.00059 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2430 on 71 degrees of freedom
Multiple R-squared:  0.34,     Adjusted R-squared:  0.321
F-statistic: 18.3 on 2 and 71 DF,  p-value: 0.000000395
```

Because R-squared went from 22% to 34%? Wrong.
Because Se went down so smaller confidence interval of noise

Better looking fitted line

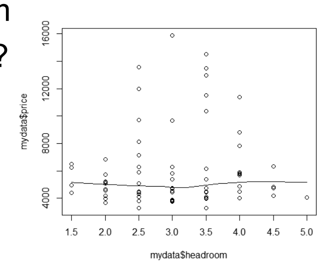
- Note-data has to be sorted

```
> plot(mydata$mpg,mydata$price)
> ord=order(mydata$mpg)
> lines(mydata$mpg[ord],predict(fit)[ord],col="red")
```



Same Dataset: Price vs Headroom

- Price versus Headroom
- Any linear relationship?



Regression Output

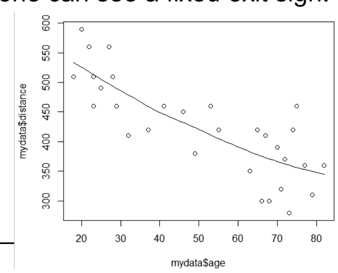
- Do we need headroom in the model? Explain

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

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

Residuals:
   Min     1Q Median     3Q    Max
 -3077  -1868   -939    577   9738

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)         4970       1269    3.92   0.0002 ***
mydata$headroom      399        408    0.98   0.3313
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2950 on 72 degrees of freedom
Multiple R-squared:  0.0131,    Adjusted R-squared:  -0.000595
F-statistic: 0.957 on 1 and 72 DF,  p-value: 0.331

> sd(mydata$price)
[1] 2949.5
```

t = 0.98 < 1.96 → fail to reject

Regression Example

- Data was collected by insurance company on age and distance one can see a fixed exit sign.

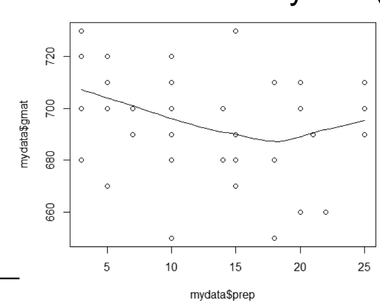


Interpret the Output


```
> fit=lm(mydata$distance~mydata$age)
> summary(fit)

Call:
lm(formula = mydata$distance ~ mydata$age)

Residuals:
   Min     1Q Median     3Q    Max
-78.23 -41.71   7.65  33.55 108.83

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  576.682     23.471   24.57   < 2e-16 ***
mydata$age    -3.007      0.424   -7.09 0.0000001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.8 on 28 degrees of freedom
Multiple R-squared:  0.642,     Adjusted R-squared:  0.629
F-statistic: 50.2 on 1 and 28 DF,  p-value: 0.000000104

> confint(fit)
               2.5 %   97.5 %
(Intercept) 528.6040 624.7599
mydata$age   -3.8761  -2.1376
```

GMAT and Number of Prep Days

- These NYU students study a lot(!). Not.



Interpret: the power of studying

```
> fit=lm(mydata$gmat~mydata$prep)
> summary(fit)

Call:
lm(formula = mydata$gmat ~ mydata$prep)

Residuals:
   Min     1Q Median     3Q    Max
-46.51 -12.60   2.47  13.68  36.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   703.31       6.58  106.87   <2e-16 ***
mydata$prep    -0.68       0.44   -1.54     0.13
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20 on 36 degrees of freedom
Multiple R-squared:  0.062,     Adjusted R-squared:  0.036
F-statistic: 2.38 on 1 and 36 DF,  p-value: 0.132

> confint(fit)
                2.5 %   97.5 %
(Intercept) 689.9623 716.6549
mydata$prep  -1.5729   0.2137
```


Things you should know

- Be comfortable examining regression output and determining if there is a significant relationship between x and y.
- Confidence intervals for the regression parameters
- Hypothesis tests for the regression parameters