1

Stat 104: Quantitative Methods for Economists

Class 31: Regression Redux

2

Stat 104 - Roadmap

Data

Intro
- Population vs. sample
- Parameters vs. statistics

Graphs
- Dotplots and Histograms

Descriptive Stats
- Central tendency
- Variability
- Relative standing
- 2 variable stats

Basic Probability
- Marginal prob.
- Conditional probability
- Laws of probability
- Probability tables

Inference

Discrete Distributions
- Random variable
- E[X], Var[X] & Laws of expectation
- Binomial

Continuous Distributions
- Density functions
- Uniform distribution
- Normal distribution
- T-distribution

Central Limit Theorem
- $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
- $\hat{p} \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

Estimation
- Point estimators
- Confidence intervals
- Levels of confidence: $1-\alpha$

Analysis

Hypothesis Testing
- Testing parameters: mu and pi
- Left tailed, right tailed, two-tailed
- Rejection regions approach
- P-values
- Levels of significance
- Type I and Type II errors

Two Sample Tests
- Compare mean across two populations
- Compare two proportions

Linear Regression
- Least squares lines: scatterplot
- Multivariate regression
- Dummy variables: 0/1
- Regression Diagnostics

3

Regression Analysis…

Regression analysis is used to predict the value of one variable (the ***dependent variable***) on the basis of other variables (the ***independent variables***).

Dependent variable: denoted **Y**

Independent variables: denoted $\mathbf{X_1, X_2, \ldots, X_k}$

dependent variable

independent variables

$$\boxed{y} = \boxed{\beta_0} + \boxed{\beta_1}\boxed{x_1} + \boxed{\beta_2}\boxed{x_2} + \ldots + \boxed{\beta_k}\boxed{x_k} + \boxed{\varepsilon}$$

error variable

coefficients

4

Some Notes and Terms

- In Simple Linear Regression, one X variable is used to explain the variable Y
- In Multiple Regression, more than one X variable is used to explain the variable Y.
- For now we will concentrate on simple regression.

5

Basically, we want to fit a line to our data set.

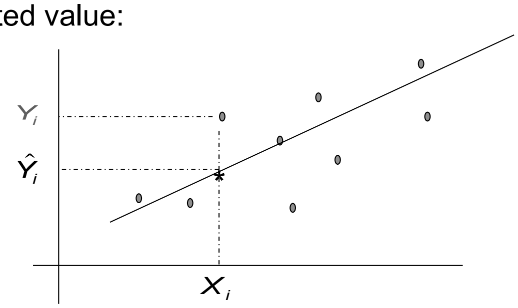
The equation of ***our*** line is given by

$$\hat{Y} = b_0 + b_1 X$$

we use the symbol $\hat{Y}$ to stand for the fitted line; Y will always stand for the observed observations.

6

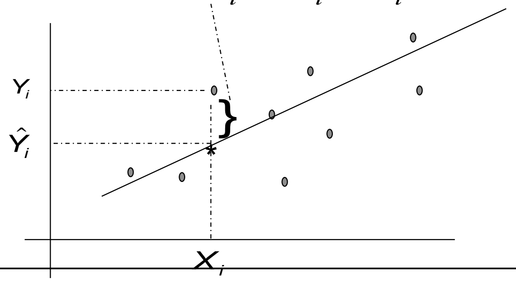
The fitted value:



$$\hat{Y}_i = b_0 + b_1 X_i$$

For the ith observation the **residual** is defined to be:

residual $e_i = Y_i - \hat{Y}_i$



The most popular criterion for **fitting a line** is called the *least squares method.* This method says to

Find b_0 and b_1 ← These two values define a line

that makes this sum as small as possible

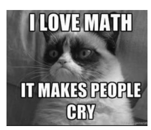
$$\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}e_i^2$$

The farther away a point is from the estimated line, the more serious the error. By squaring the errors, we "penalize" large residuals so that we can avoid them.

The values of b_0 and b_1 which minimize the residual sum of squares are:

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

$$b_0 = \overline{Y} - b_1\overline{X}$$



These formulas can be derived using calculus- we pass.

These formulas are the intercept and slope for the "best fitting line".

Example

- Suppose we want to **predict** the sale price of used Honda Accords.
- Many factors influence the price of a used car; model year, condition, transmission type, 2 or 4 door, color, mileage, how badly owner wants to sell, etc....
- We will choose just the variable mileage and see if price can be predicted from the mileage of the car.

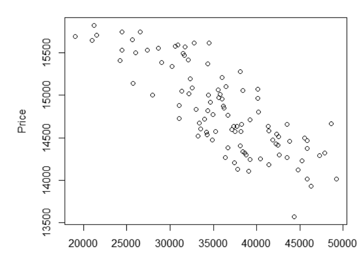
Load in the data

- For completeness, don't really need to see this.

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/accordprices.csv")
> names(mydata)
 [1] "Price"    "Odometer" "Color"    "X"        "X.1"      "X.2"
 [7] "X.3"      "X.4"      "X.5"      "X.6"      "X.7"
> attach(mydata)
```


Scatter Plot of Car Data

What's going on?




```
plot(Odometer,Price)
```

Performing Regression in R

Y modelled as X

```
> fit=lm(Price~Odometer)
> summary(fit)

Call:
lm(formula = Price ~ Odometer)

Residuals:
    Min      1Q  Median      3Q     Max
-730.32 -235[illegible]  [illegible]  287.75  691.25

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.707e+04  1.690e+02  100.97   <2e-16 ***
Odometer    -6.232e-02  4.618e-03  -13.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303.1 on 98 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.6501,     Adjusted R-squared:  0.6466
F-statistic: 182.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

b_0 (Intercept)

b_1 Odometer

Ignore

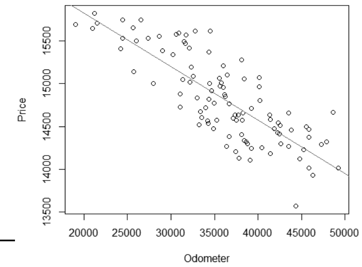
Ignore

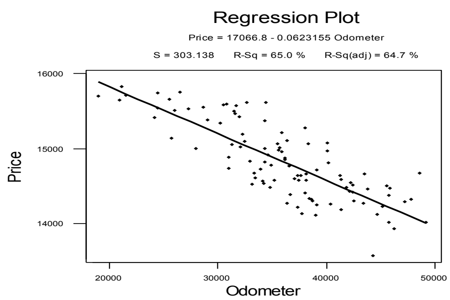
```
> coef(fit)
 (Intercept)      Odometer
17066.766070     -0.062315
```

We will eventually explain this complete printout-ignore most of it for now.

Fitted Line Plot in R

```
> plot(Odometer,Price)
> abline(fit,col="red")
```



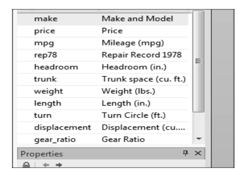


Interpretation of the slope: For each additional mile on the odometer, the price *decreases* by an average of \$0.062

R-sq : 65% of the variation in the selling price is explained by the variation in odometer reading. The rest (35%) remains unexplained by this model.

Do not interpret the intercept as cars that have not been driven cost \$17066.8

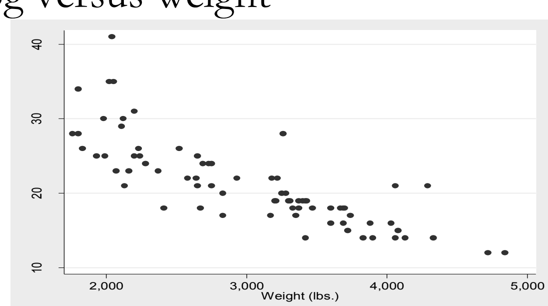
Example: 1978 auto data set



make	Make and Model
price	Price
mpg	Mileage (mpg)
rep78	Repair Record 1978
headroom	Headroom (in.)
trunk	Trunk space (cu. ft.)
weight	Weight (lbs.)
length	Length (in.)
turn	Turn Circle (ft.)
displacement	Displacement (cu....
gear_ratio	Gear Ratio

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/cars10.csv")
> names(mydata)
 [1] "make"         "price"        "mpg"          "headroom"     "trunk"
 [6] "weight"       "length"       "turn"         "displacement" "gear_ratio"
[11] "foreign"
```


Mpg versus weight

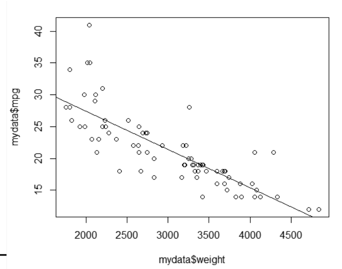


Interpret

```
> fit=lm(mydata$mpg~mydata$weight)
> coef(fit)
  (Intercept) mydata$weight
   39.4402835    -0.0060087
```

Fitted Line Plot

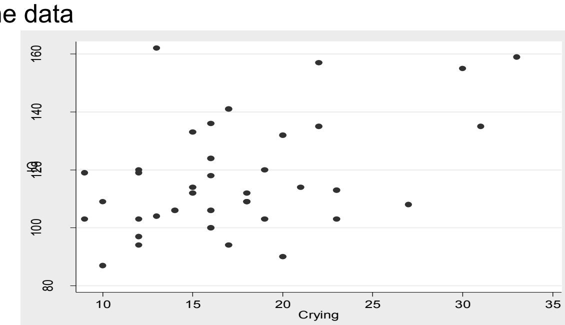
```
> plot(mydata$weight,mydata$mpg)
> abline(fit)
```



Example: Crying Babies

- Babies who cry a lot may be more easily stimulated than other babies, and this may be an indication of higher IQ. Karelitz, et al. (1964) studied the association between IQ and crying frequency with 37 babies.
- The researchers caused the babies to cry by snapping a rubber band on the sole of their foot (bastards...).
- They recorded the frequency of cries as the number of peak cries (example: WAAAHHHHH-WAAAAHHHH is two peaks) in the most active 20 seconds of crying. Three years later, they measured the babies' IQs.

The data



Regression Output

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/crying.csv")
> names(mydata)
[1] "Crying" "IQ"
> coef(fit)
  (Intercept) mydata$Crying
      86.6898        1.6751
```

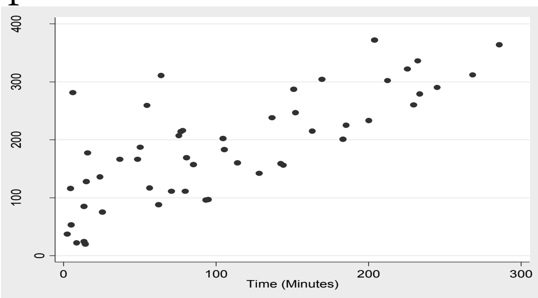
Interpretation?

Example: Online Purchases

- We have data on 48 randomly chosen customers who made purchases last quarter from an online retailer.
- The file contains information related to the time each customer spent viewing the online catalog and the dollar amount of purchases made.
- The retailer would like to analyze the sample data to determine whether a relationship exists between the time spent viewing the online catalog and the dollar amount of purchases.

25

Graph



26

R Output

```
> fit=lm(mydata$purchase~mydata$time)
> summary(fit)

Call:
lm(formula = mydata$purchase ~ mydata$time)

Residuals:
   Min     1Q Median     3Q    Max
 -88.0  -46.4  -12.4   34.9  180.1

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)   95.739     14.321    6.69 0.00000002041 ***
mydata$time    0.865      0.107    8.10 0.00000000014 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61.3 on 49 degrees of freedom
Multiple R-squared:  0.572,     Adjusted R-squared:  0.563
F-statistic: 65.5 on 1 and 49 DF,  p-value: 0.000000000137
```

27

Properties of the Residuals and Fitted Values

The residuals and fitted values obtained from the least squares line have special properties.

Let's go back to the Accord data and check them out.

28

Obtaining the residuals and fits

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/accordprices.csv")
> fit=lm(Price~Odometer)
> e=residuals(fit)
> yhat=predict(fit)
```

29

```
temp=cbind(Price,yhat,e)
View(temp)
```

Y $\hat{Y}$ e

Data: temp

	Price	yhat	e
1	14636	14737	-100.9150
2	14122	14278	-155.6499
3	14016	14211	-194.6608
4	15590	15144	446.4142
5	15568	15091	476.9461
6	14718	14947	-229.4167
7	14470	14209	260.6478
8	15690	15879	-189.2200
9	15072	14565	507.1380
10	14802	14559	242.6218
11	15190	15050	139.7005
12	14660	14354	306.0136
13	15612	15026	585.6919
14	15610	14919	691.2484
15	14634	14716	-82.2263
16	14632	14490	141.9789
17	15740	15542	198.0313
18	15008	14837	170.9440
19	14666	14037	628.5762

30

What can R tell us about the residuals ?

```
> sum(e)
[1] 0.00000000000066258
> mean(e)
[1] 0.0000000000000066003
```

Hmm. The mean of the residuals is 0.

What does that imply about the sum of the residuals ?

Does this make sense ?

31

Another interesting result is that the mean of the fitted values is the same as the mean of original Y values.

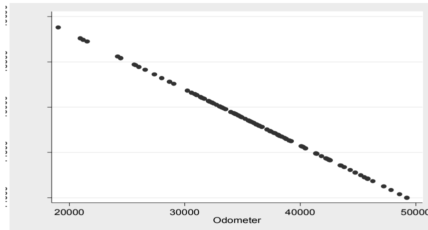
```
> mean(Price)
  [1] 14823
> mean(yhat)
  [1] 14823
```

The average of the observed value is average of the predicted value

32

Let's check out these "yhat" values:

Plot of $\hat{Y}$ versus odometer



Is there a linear relationship between yhat and X ?

$corr(\hat{Y}, X) = ?$

-1

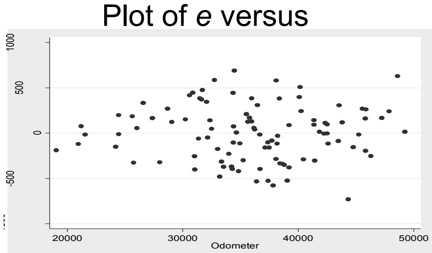
Y' = b0 = b1*X

Y' is made up X stuff so there is an exact linear relationship

33

Let's get a handle on these "e" things.

Plot of *e* versus



Is there a linear relationship between *e* and X ?

$corr(e, X) = ?$ 0

34

Basic Algebra:

$$Y = \hat{Y} + (Y - \hat{Y})$$

or equivalently

$$Y = \hat{Y} + e$$

this is an important decomposition of Y

35

To summarize:

We have the decomposition of our observation

$$Y = \hat{Y} + e$$

Related to X

$[corr(\hat{Y}, X) = 1]$

Unrelated to X

$[corr(e, X) = 0]$

36

A Summary of the fit: R^2

We have:

$corr(\hat{Y}, X) = 1$ $\qquad$ $corr(e, X) = 0$

What is

$corr(\hat{Y}, e)$??

37

We have

$$Y = \hat{Y} + e$$

and $$corr(\hat{Y}, e) = 0$$

Also: $$Var(Y) = Var(\hat{Y} + e)$$
$$= Var(\hat{Y}) + Var(e) + 2Cov(\hat{Y}, e)$$

But Cov = 0 since Corr = 0

38

So, $$Var(Y) = Var(\hat{Y}) + Var(e)$$

or,

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \frac{1}{n-1}\sum_{i=1}^{n} e_i^2$$

or,

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{total sum of squares SST}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{regression ss SSR}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{error ss SSE}}$$

39

$$Var(Y) = Var(\hat{Y}) + Var(e)$$

SST = SSR + SSE

Decomposing information

Trying to explain variation in Y

40

Recap

- SST tells us how much variation there is in the dependent variable Y
- SSR tells us how much of the variation in the dependent variable our model explained. How much of the variation is explained by X
- SSE tells us how much of the variation in the dependent variable our model did not explain. How much of the variation is not explained by X
- SST does not depend on number of X's in the model.

We're hoping SSR is really large and SSE is really small

Ideally, SSE is 0

SST does not depend on Xs in the model while SSR and SSE does

It is a fixed value

41

R^2: A measure of fit:

We have a "good fit" if SSR is big and SSE is small.
If SST=SSR we have a perfect fit.
To summarize how close SSR is to SST we define the *coefficient of determination*

$$R^2 = \frac{SSR}{SST}$$

the proportion of variation in Y explained by the regression i.e. the Xs

R^2 is between 0 and 1, and the closer R^2 is to 1, the better the fit.

42

Range of R-sq

- R-squared is always between 0 and 100%:
- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.
- R-sq of 0 or 1 is not a good result.

1 does not mean predictions are good. It means you are modelling the variation

Caution about R-sq

- R-squared does not indicate whether a regression model is adequate
- You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!
- We will learn other techniques to check the adequacy of the model.

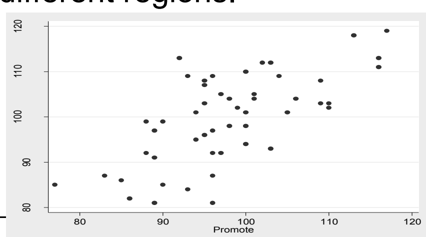
The Accord Data Again

the R squared value

```
> 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
```


Example: Pharmex Pharmaceuticals

- We have data on sales and promotion costs for 50 different regions.

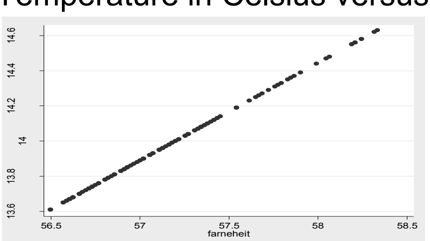


R-sq

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

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

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$Promo    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
```


Example: Global Temperature

- Temperature in Celsius versus Fahrenheit.

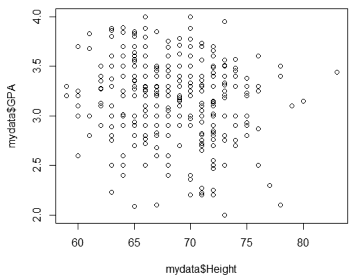


What should the R-sq be?

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: Height versus GPA



What should the R-sq be?

Example: Height versus GPA

- Regression output

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

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

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1053 -0.2514  0.0434  0.2873  0.8611 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    3.92886    0.36104   10.88   <2e-16 ***
mydata$Height -0.01128    0.00527   -2.14    0.033 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.396 on 336 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.0135,    Adjusted R-squared:  0.0105 
F-statistic: 4.59 on 1 and 336 DF,  p-value: 0.0328
```


A note about R^2:

Some people spend a lot of time worrying about R^2 . They think that the higher the value of R^2 , the better the fit of the regression.

Well, we will see later that there are problems with R^2, including the fact that when you add variables to a model (perform multiple regression), the value always increases.

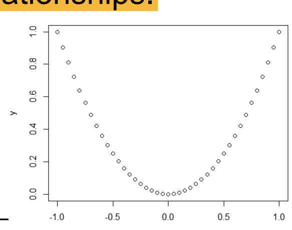
More about this later.

Low R-sq could be a Nobel Prize

```
. regress anfret spyret

      Source |       SS           df       MS      Number of obs   =        36
                                                   F(1, 34)        =      2.24
                                                   Prob > F        =    0.1436
                                                   R-squared       =    0.0618
                                                   Adj R-squared   =    0.0343
                                                   Root MSE        =    .09308

                                                  ------------------------------
                                                  P>|t|     [95% Conf. Interval]
                                                  ------------------------------
                                                  0.144    -.2787692    1.838798
                                                  0.290    -.0516163    .0159197
                                                  ------------------------------
```

Regression is for linear relationships

- Remember that regression is only suitable for linear relationships.



Regression Output

```
> fit=lm(y~x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max 
-0.350 -0.287 -0.100  0.212  0.650 

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)    
(Intercept) 3.50e-01   5.01e-02    6.99 0.000000022 ***
x           1.04e-16   8.47e-02    0.00           1    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.321 on 39 degrees of freedom
Multiple R-squared:  4.07e-32,  Adjusted R-squared:  -0.0256 
F-statistic: 1.59e-30 on 1 and 39 DF,  p-value: 1
```

R^2 Criticism

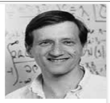
WORKSHOP

*How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science**

Gary King, *New York University*

Gary King

Political Scientist



Gary M. King is an American political scientist and quantitative methodologist. He is the Albert J. Weatherhead III University Professor and Director for the Institute for Quantitative Social Science at Harvard University. Wikipedia

Born: December 8, 1958 (age 56), Madison, WI

Education: University of Wisconsin-Madison (1984), State University of New York at New Paltz

Awards: Guggenheim Fellowship for Social Sciences, US & Canada

R^2 Criticism

The Race (3): Coefficient of Determination?

R^2 is often called the "coefficient of determination." The result (or cause) of this unfortunate terminology is that the R^2 statistic is sometimes interpreted as a measure of the influence of X on y. Others consider it to be a measure of the fit between the statistical model and the true model. A high R^2 is considered to be proof that the correct model has been specified or that the theory being tested is correct. A higher R^2 in one model is taken to mean that that model is better.

All these interpretations are wrong. R^2 is a measure of the spread of points around a regression line, and it is a poor measure of even that (Achen, 1982). Taking all variables as deviations from their means, R^2 can

Worse, however, is that there is no statistical theory behind the R^2 statistic. Thus, R^2 is not an estimator because there exists no relevant population parameter. All calculated values of R^2 refer only to the particular sample from which they come. This is clear from the standardized coefficient example in preceding paragraphs, but it is more graphically

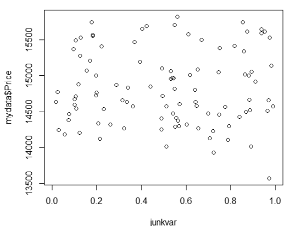
Q: But do you really want me to stop using R^2? After all, my R^2 is higher than that of all my friends and higher than those in all the articles in the last issue of the *APSR!*

A: If your goal is to get a big R^2, then your goal is not the same as that for which regression analysis was designed. The purpose of regression analysis and all of parametric statistical analyses is to estimate interesting population parameters (regression coefficients in this case). *The best regression model usually has an* R^2 *that is lower than could be obtained otherwise.*

Example: Adding Junk to a Model

- We can generate random data in R

```
junkvar=runif(length(mydata$Price))
plot(junkvar,mydata$Price)
```



There is no relationship between price and this junk variable.

Original Model

```
> 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
```


However, what happens to R^2 ?

```
> fit=lm(mydata$Price~mydata$Odometer+junkvar)
> summary(fit)

Coefficients:
                  Estimate  Std. Error t value Pr(>|t|)
(Intercept)     17045.57912   177.03317   96.28   <2e-16 ***
mydata$Odometer    -0.06235     0.00464  -13.44   <2e-16 ***
junkvar            43.44618   103.13553    0.42     0.67
---
Residual standard error: 304 on 97 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.651,     Adjusted R-squared:  0.644 
F-statistic: 90.4 on 2 and 97 DF,  p-value: <2e-16
```

- The value of R-squared went up, even though this isn't a better model!
- **Looking towards the next lecture, there is info on this output that tells us we don't need junkvar in the model.**

Things you should know

- The least squares estimates
- Properties of residuals
- Information decomposition (SST=SSR+SSE)
- R^2
- Criticism of R^2