# Regression and Stats Primer

D. Alex Hughes

dhughes@ucsd.edu

March 1, 2012

The bulk of the course has focused on theorizing, conceptualizing, hypothesizing, and operationalizing. Today's section deals with Inference-izing.

- Given we have established a valid research-design strategy; or that we are minimally aware of the threats to validity.
- And... we have asked a question that others are interested in hearing the answer to – theoriz-ation
- And...we have measured it validly – operationalization
- And... we have gathered the data validly – randomization
- And... we have a clear, strong causal pathway – Internal Validity/Experimentation
- **Then...** how can we draw a conclusions about our data?

Frequently, how to design data-collection is influenced by available analysis tools.

- Early AIDS research
- Freud's Psychoanalytic Theories
- The Bohr model of the Atom

Feasible data-gathering frequently influences the types of questions that people ask.

- What is the role of the brain in influencing behavior?
- What is the role of International Organizations in influencing Global Policy?
- How do Electoral Systems influence Representation?

# Mechanics of OLS Regression

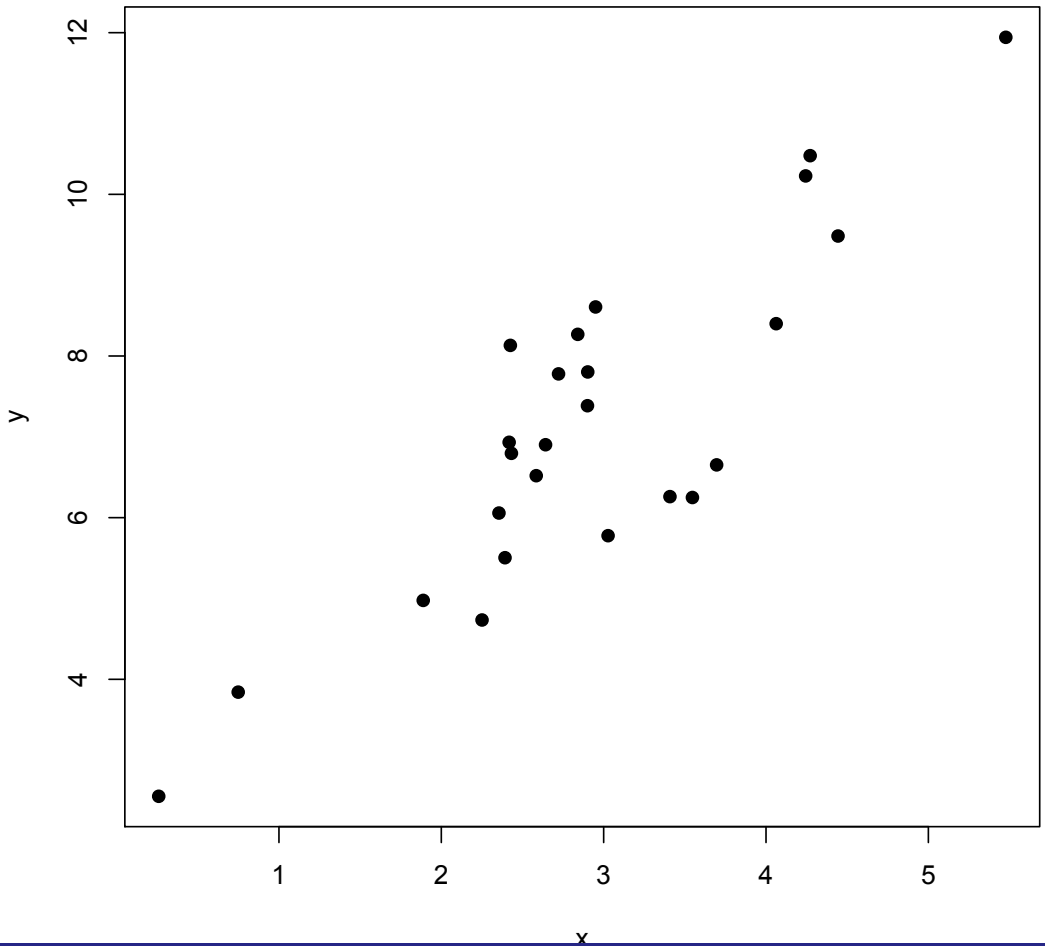
What is regression doing?

- In linear regression, the goal is to fit the best line (or plane, or hyperplane) to the data.
- To do so, we set up a measure of distance between a line and the data, and try and make that distance as small as possible.
- Regression mechanics are just doing this for us.

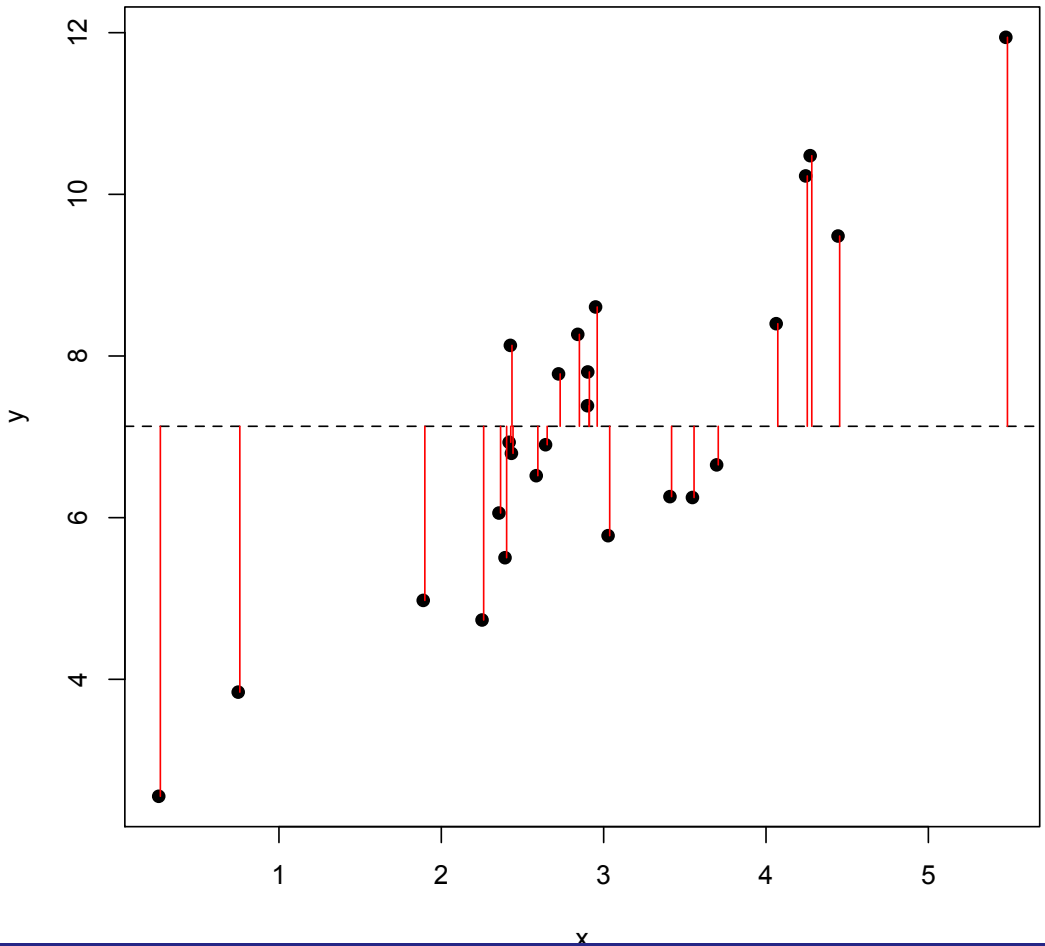
How does it do that, Alex?

- Imagine you have data, and you array it in two dimensional space.
- Now fit a line into that two-dimensional space that make the distances between the data points and the line you fit the smallest.

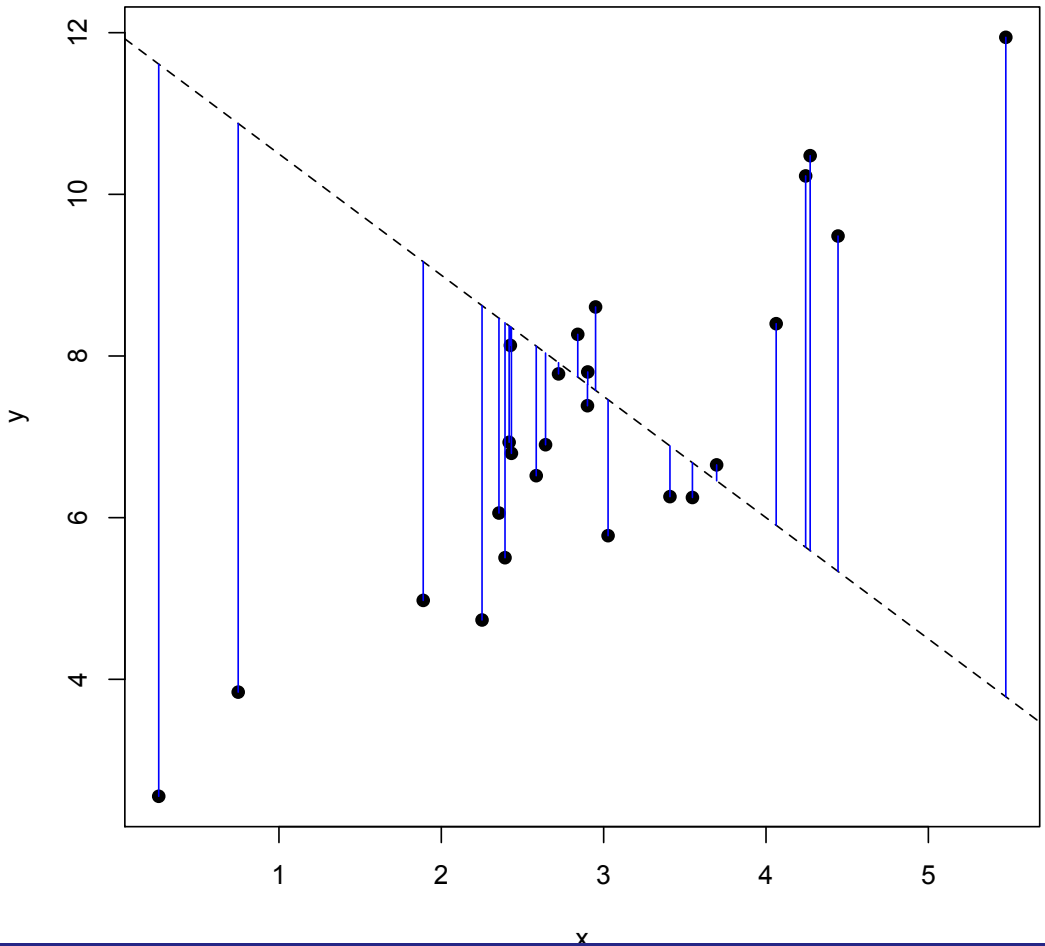
# The Data



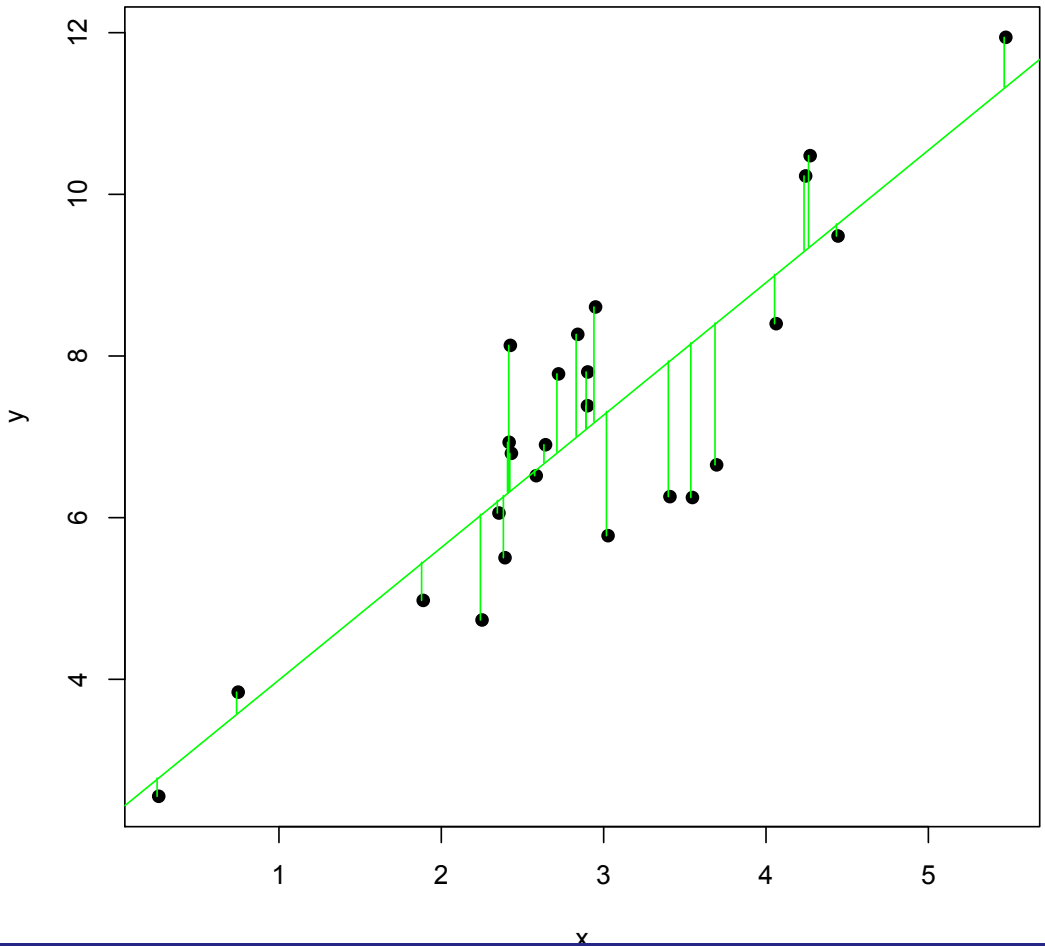
# The Null Model



# A Very Bad Model



# The Best Model



# Replication Code

```
# Fit of models
# compare null and alt models

  n<-25
  x<-rnorm(n)+3
  y<-2+1.5*x + rnorm(n,.5)*1
  q <- rnorm(n)
##Plot Null Model
  plot(x,y,pch=19,cex=1)
  abline(mean(y),0,lty=2)
  yhat<-lm(y~x)$fitted.values
  for(i in 1:n){lines(c(x[i]+.01,x[i]+.01),c(y[i],mean(y)),col="red")}
## Plot Very Bad Model
  plot(x,y,pch=19, cex = 1)
  abline(12,-1.5,lty=2)
  yhat<-lm(y~x)$fitted.values
  for(i in 1:n){lines(c(x[i],x[i]),c(y[i],12-1.5*x[i]), col = "blue")}
## Plot Best Fitting Model
  plot(x,y,pch=19,cex=1)
  abline(lm(y~x)$coef, col = "green")
# abline(mean(y),0,lty=2)
  yhat<-lm(y~x)$fitted.values
  for(i in 1:n){lines(c(x[i]-.01,x[i]-.01),c(y[i],yhat[i]), col = "green")}
```

ModelFits.R

# More Dimensions

```
1  library(rgl)
2  example(plot3d)
3
4  n<-2000
5  x<-rnorm(n, sd=1)
6  y<-x+rnorm(n, sd=.5)
7  z<- -x+y^2+rnorm(n,sd=.5)
8
9
10 plot(x,y, pch = 19, cex = .5)
11 plot3d(x,y,z,size=4, cex = 0.5)
```

plot3d.R

# Total Sidebar – Correlation

- Correlation is a measure of the strength of a **linear** association.
- Note, that this does not measure the strength of the relationship.
- Remember, if the association is non-linear, simple correlation will not be a good fit
- May Range from -1 to 1.

$$\text{“}r\text{”} = \rho = \sum_{i=1}^{n} \frac{(Z_x) * (Z_y)}{(n-1)}$$

# Correlation Interpretation

- Make a scatter plot, IV on X-axis, DV on Y-axis.
- If the plot looks "cloudy" – say there is little meaningful correlation
- If the plot looks "line-y" – say there is probably meaningful correlation
- If the plot looks liney, note the slope of the line. If it is a negative slope, we have negative correlation
- The stronger the line pattern, the stronger the correlation $\rightarrow$ if we calculated the correlation, it would be near either 1 or -1.

# Correlation and Explanation of Variance

- The Square of correlation is just the amount of variance in data that is explained by the best fitting straight line.
- $R^2 = r * r = r^2$
- Just say: " $R^2$ is the amount of variance in Y that is explained by X.
- **Saying that $R^2$ is a function of the "best fitting straight line" begs the question that we can find the best straight line. We can, and we use Ordinary Least Squares regression to do so.**

# The Nuts and Bolts

The Math behind what the Computer is doing:

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

# Assumptions

Predicated on a series of Assumptions about the data. Imagine we were linguists:

- We would need words to follow some basic patterns if we were to discern a grammar – the rules by which language is governed.
- If they didn't follow that pattern, we could make up a grammar, but it isn't clear what it would mean.

In the same way, for us to know anything about what our regression is saying, we need to know the rules that make it a meaningful estimate of relationships.

# Gauss-Markov Assumptions

Linear regression is the best – or at least as good as others "*Which others Alex?*" – estimate of the relationship between variables if:

- The Average of the Errors is 0
  - $E[\epsilon_i] = 0$
- There is no correlation in the Errors
  - $COV[\epsilon_i, \epsilon_j] = 0$
- The variance of the data is *finite, known, and constant*
  - $VAR(\epsilon_i) = \sigma^2 < \infty$
- The data are described by a linear relationship
  - The DV is described by a *linear combination* of IVs, even if those IVs are themselves non-linear
  - $Y = mX + b$
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \epsilon$

# Practical Regression

# Interpretation of Coefficients

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- A 1-unit change in $X_1$ is associated with a $\beta_1$-unit change in Y, *all else equal*.
- A 1-unit change in $X_2$ is associated with a $\beta_2$-unit change in Y, *all else equal*.
- This is the controlled comparison bit.

$$Income = \$20,000 + \$2,000[Years\_Edu] - \$5,000[Female] + \epsilon$$

- A one-year increase in Education is associated with ...
- A college graduate is expected to earn how much *more* than a HS-Grad?

# Predicting Values

Not only can we predict marginal contributions to the dependent variable, but we can use our equation to *predict* outcomes.

$Income = \$20,000 + \$2,000[Years\_Edu] - \$5,000[Female] + \epsilon$

- What would we expect a college educated male to earn?
- What would we expect a HS-Educated female to earn?

$P(Turnout) = \beta_0 + \beta_1[AGE] + \beta_2[TeaParty] + \beta_3[Urban] + \epsilon$

# Interpreting Interactions

- Dummy-Variable Interactions
- Continuous-Variable Interactions

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon + \beta_3 X_1 X_2$, but let $X_2$ be a $\{0,1\}$ dummy-variable

- A 1-unit change in $X_1$ is associated with a $\beta_1 + \beta_3 * X_2$-unit change in Y
  - If $X_2 = 0$, then a 1-unit change in $X_1$ is associated with a $\beta_1$ change in Y.
  - If $X_2 = 1$, then a 1-unit change in $X_1$ is associated with a $[\beta_1 + \beta_3]$ change in Y.
- A 1-unit change in $X_2$ is associated with a $[\beta_2 + \beta_3 * X_1]$-unit change in Y.

# Inference about Individual Coefficients

Are the coefficients that we generated actually different than zero?

- Skipping discussion of sampling error...
- **Using stats, are we pretty sure that our coefficient-estimates mean something?**

To test this, we set up a standard t-test – is the magnitude of the coefficient that we estimate sufficiently larger than the uncertainty we have about that estimate?

- $t = \frac{\hat{\beta}}{\sqrt{\widehat{Var}(\hat{\beta})}}$

# Selecting Models – Analysis of Variance, AIC, BIC

It is possible that we set up an equation, and we find that many of the predictors are significantly related to the Dependent Variable.

- Does the whole equation matter?
- Is there another equation that does better?

A series of hypothesis tests have been developed that address just these qusetions.

- Whole Equation? F-test, which is a test-statistic defined by the Residual Sum of Squares (remember those red, blue, and green lines), divided by a measure of the total variance in the DV.

# Selecting Models – Analysis of Variance, AIC, BIC

- Improvements on the Whole Equation? Are there parts of the equation that aren't "pulling their weight?" Set up a *linear restriction* condition where we restrict a regressor to be zero, and then see if there is a *significant* decrease in the explanatory power of the regression.

# Practical Model Selection – Step-wise Deletion

- Establish a "kitchen-sink" model of the terms that theory & literature include – call this model $\Omega$
- Establish a smaller model where you have restricted some of the terms in the larger model to be 0 – call this model $\omega$
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \Omega$
  - $Y = \beta_0 + \beta_1 X_1 + -X_2 = \omega$
- Does the smaller model explain significantly *less* of the variance in Y?
- $F = R^2_\Omega / R^2_\omega$

# Practical Model Selection – More Complex Models

AIC and BIC are different types of *Information Criteria* and are used in more complex cases than the linear regression case. They are closely related to another concept called the log-likelihood. in essence, all three concepts are used to compare one model to another.

- For AIC and BIC, the better fitting model is the model with lower AIC or BIC.
- For ln(L) the better fitting model is the model with a higher ln(L).

Get our your Computers

Many of the questions we are interested in do not meet the assumptions of the Gauss-Markov Theorem.

- Especially that the expectation of the errors is zero, and that the Variance in the errors is constant.
- Especially War, Voting, Democratization – these outcomes either *are* or *aren't*.
- There are estimation techniques that outperform OLS, namely Logit/ Probit regression.

On our front end, the process looks essentially the same, but we make slightly different assumptions about the data-generating process, and the interpretation of regression coefficients is slightly different.

# Dichotomous Regression

We observe 0s and 1s, but we believe that there is an unobserved, continuous distribution driving the outcome.

- What would an OLS model fit? What are some problems?
- How could we fit something more appropriate?

$Yz = \frac{1}{1 + e^{-z}}$, where:
$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

# Difference in Differences

- The regression form for many kinds of experimental data.
- For pure experiment, we include a Dummy Variable for being in the Treatment Condition.
- For policy experiments, we are concerned about History Threats to validity.
- We need to "back out" what we think would have happened in the counterfactual case.

# Difference in Differences

Take for example the Mariel Boat-lift. Between May and September 1980 Miami's labor force grew by 7%. Where did they end up? Why?

- What were the effects on the Economy?
- What are the threats to the validity of our estimates?

Unemployment Rates of Whites
(Standard errors in parentheses)

| | Before (1979) | After (1981) | Difference |
|---|---|---|---|
| Miami | 5.1 (1.1) | 3.9 (0.9) | -1.2 (1.4) |
| Control Cities | 4.4 (0.3) | 4.3 (0.3) | -0.1 (0.4) |
| Difference | 0.7 (1.1) | -0.4 (0.95) | -1.1 (1.5) |

# Difference in Differences

There is a straightforward regression analogue that allows us to analyze this form of DGP.

- Let $G_i = 1$ if an individual is in the Treatment Group (Miami)
- Let $T_i = 1$ if *after* the policy change (for both Treatment and Control Groups)
- Then with time-series cross section data (uuungh...): We estimate
  $Y_i = \beta_0 + \beta_1 T_i + \beta_2 G_i + \beta_3 G_i T_i + Z\beta + u_i$
- $\beta_3$ is our coefficient of interest, the difference in differences of outcomes.

# Duration Models

These are a class of models used to describe the time until an event happens. Think time until war breaks out, think duration of Peace, think time until Democratic Transition, and so on.

- Hazard Models – `http://en.wikipedia.org/wiki/Hazard_model`
- Cox-Proportional Hazard Models – `http://en.wikipedia.org/wiki/Cox_proportional_hazards_model`
- Negative Binomial Models – `http://cran.r-project.org/web/packages/Zelig/vignettes/negbin.pdf`

Useful because the process described is typically non-linear in the predicting variables, so to fit a line doesn't make sense.