

IES 612/STA 4-573/STA 4-576

Winter 2008

Week 1--IES 612-STA 4-573-STA 4-576.doc

Review Notes: [OL] = Ott & Longnecker *Statistical Methods and Data Analysis*, 5th edition.
[Handouts based on notes prepared by J. Bailer and R. Schaefer]

Ch 20	Communicating and documenting results of analyses
2.5	Data Management issues
Describing data	
3.3	Graphical methods - histograms, stemplots, boxplots, dotplots
3.4	Central tendency - mode, median, mean
3.5	Variability - range, percentiles, IQR, var, SD, CV
3.6	Distributions
Representing Experimental Results	
4.1	Probability - definition (details 4.2-4.5)
4.6	Random variables
4.7	Probability Distributions - discrete RVs
4.8	Binomial
4.9	Probability Distributions - continuous RVs
4.10	Normal
Describing Populations based on Data	
4.11	Random Sample
4.12	Sampling Distribution - SE, CLT
5.2	Estimating population mean μ (sample size 5.3)
5.4	Tests involving μ - (H_0 , H_A , TS, RR), Type I/II errors
5.5	Sample size
5.6	P-value
5.7	Inference for μ with σ unknown ***
5.8	Inference for median
Comparing 2 population central values	
6.1, 6.2	$\mu_1 - \mu_2$ independent samples (6.3 nonparametric alternative)
6.4	$\mu_1 - \mu_2$ Paired samples (6.5 nonparametric alternative)
6.6	Sample size planning
This Semester	
11, 12, 13	Regression
8, 9, 14, 15, (16), 17, 18, 19	Experimental Designs

* CONCEPTS:

POPULATION = collection of all units of interest

SAMPLE = subset of population selected to represent the population

DISTRIBUTION = characteristics of a "population" of values (CENTER, SPREAD, SHAPE)

PARAMETERS = characteristic of the population (μ , σ^2 , ρ , β_0)

STATISTICS = characteristic of the sample (\bar{y} , s^2 , r , b_0)

Sampling - selecting elements from a population into a sample

Inference - making statements about a population based on information in a sample

Hypothesis Tests

H_0 - null/no-effect hypothesis

H_A (or H_1 or H_a) - research or alternative hypothesis

Test statistic (TS)

Rejection Region / P-value

Conclusion

Errors? Type I (False Positive); Type II (False Negative)

α , β

Confidence Intervals

(Point Estimate) +/- (Multiplier) (Standard Error)

* other ways to forms confidence intervals but this general form applies in many general cases

We are moving from DESCRIPTIVE STATISTICS and simple HYPOTHESIS TESTS towards **MODELS** for describing **ASSOCIATION** and **PREDICTION**

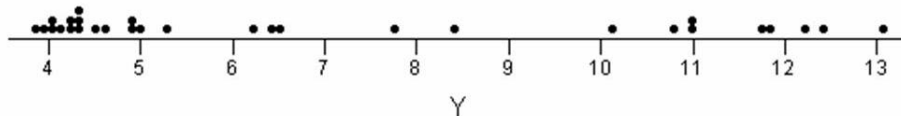
Basic concepts for regression models

- There is a probability distribution of response variable “Y” for each value of regressor variable “x”
- The means of these probability distributions vary in some systematic fashion with “x”.

$$\mu_{Y|x} = f(x)$$

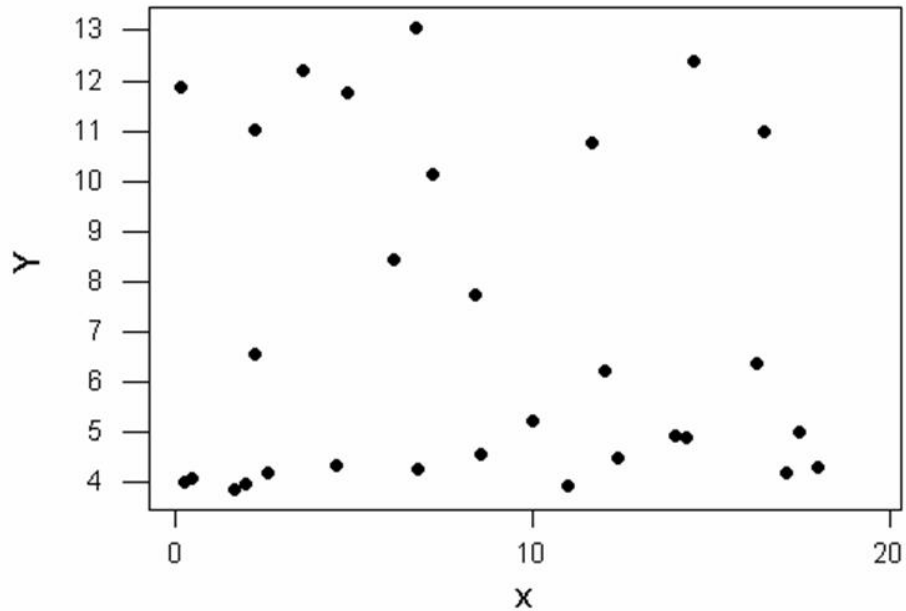
Goals of regression models

- To predict new values of the response variable for given values of the regressor variable(s)
- To describe the relationship between the response variable and the regressor variable(s)

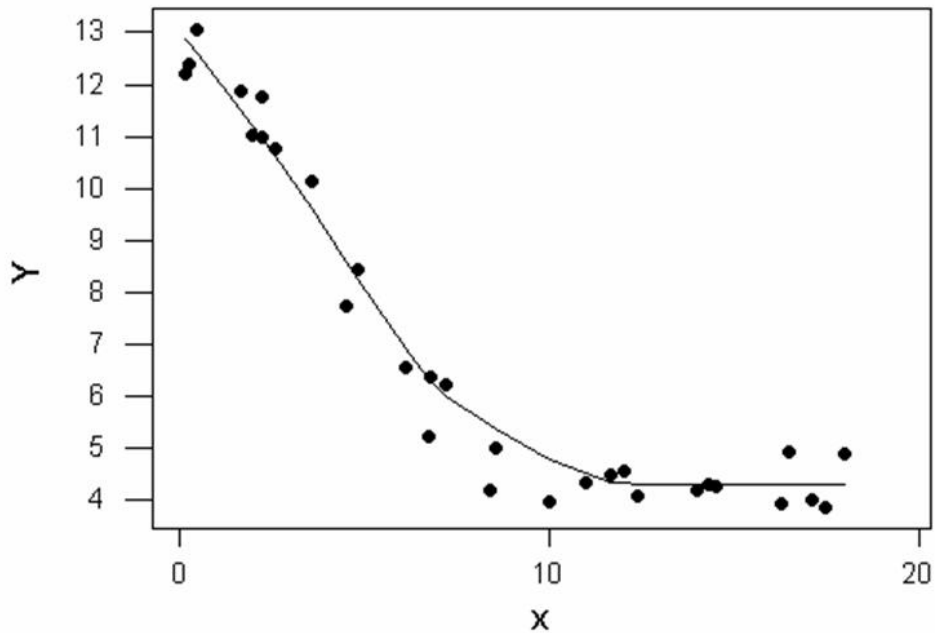


No relationship between variables x and Y

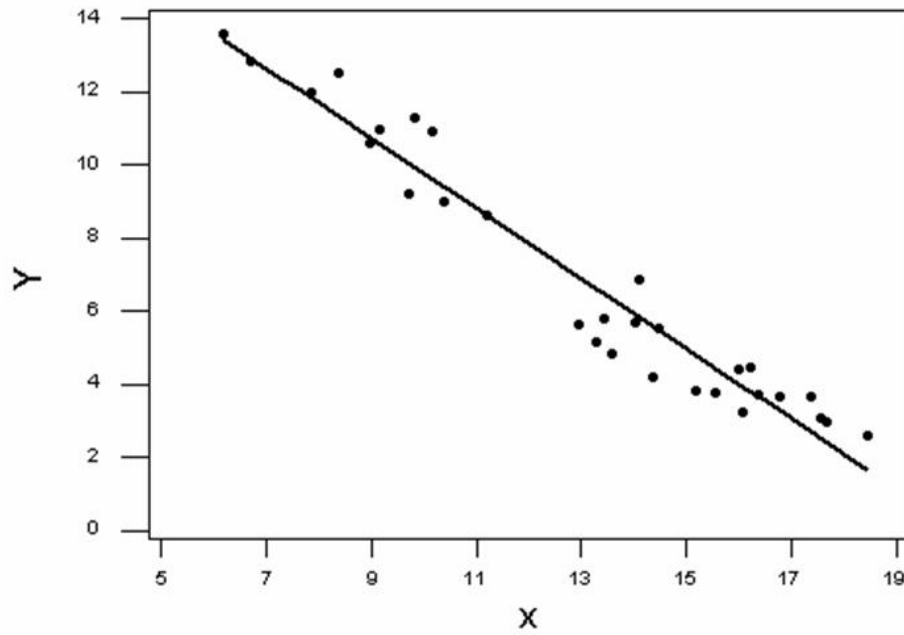
$$\mu_{Y|x} = \mu_Y$$



$$\mu_{Y|x} = \frac{\beta_2 \exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} + \beta_3$$



$$\mu_{Y|x} = \beta_0 + \beta_1 x$$



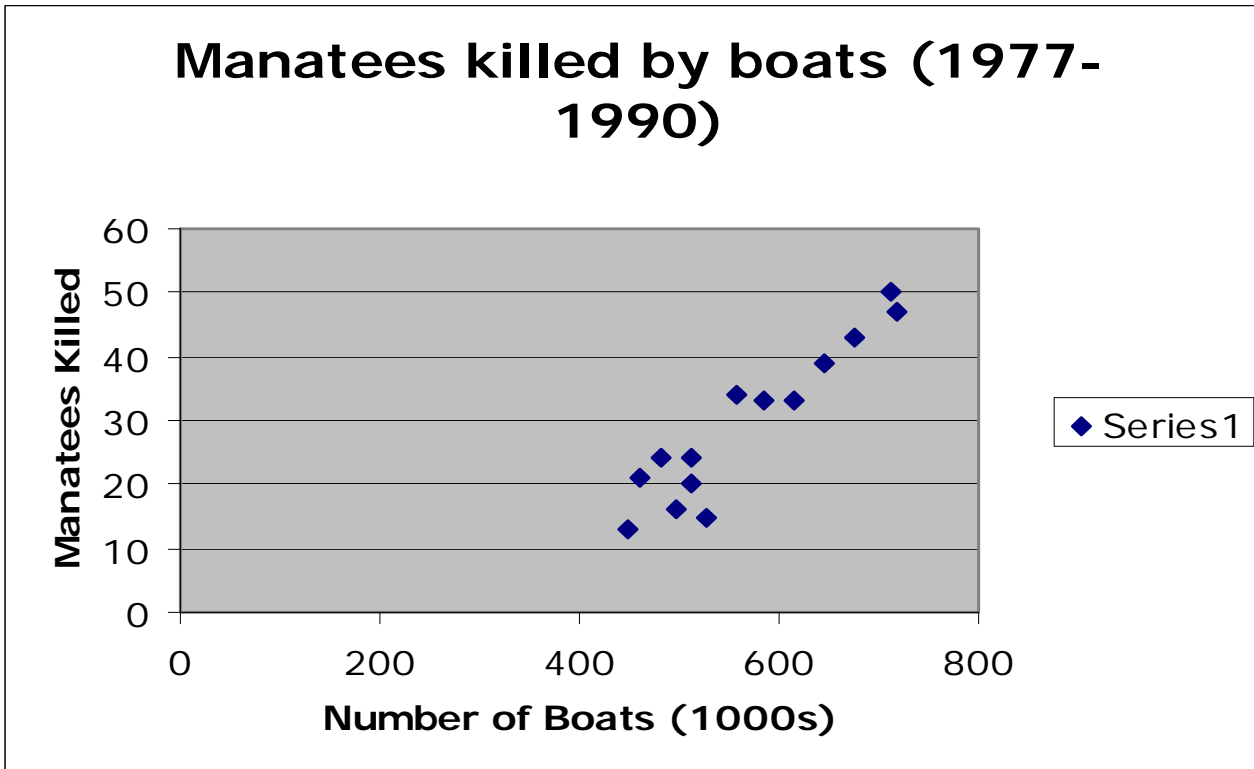
What three elements or features are evident in the above?

-
-
-

Example: Manatee deaths due to motorboats in Florida

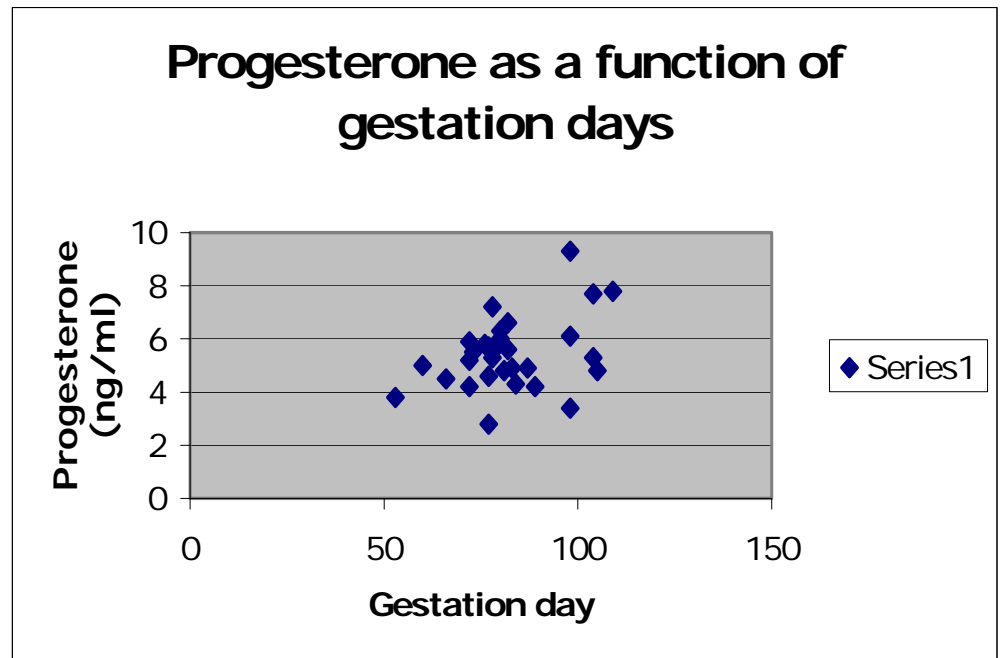
YEAR	Number Boats (1000s)	Manatees Killed
77	447	13
78	460	21
79	481	24
80	498	16
81	513	24
82	512	20
83	526	15
84	559	34
85	585	33
86	614	33
87	645	39
88	675	43
89	711	50
90	719	47

Graphical display? Scatterplot or scatterdiagram



Example: Progesterone level as a function of gestation day in sheep pregnant with singletons

Singleton Gestation Days	Singleton Progesterone
53	3.8
60	5
66	4.5
72	4.2
73	5.5
76	5.8
77	4.6
78	5.3
78	7.2
79	5.7
80	6
80	6.3
81	4.8
82	5.6
83	4.9
84	4.3
87	4.9
89	4.2
98	3.4
105	4.8
72	5.2
72	5.9
77	5.7
77	2.8
82	6.6
98	6.1
98	9.3
104	7.7
104	5.3
109	7.8



Regression data

$(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ or in shorthand, $(x_i, y_i) \ i = 1, \dots, n$

Basic Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{[“simple linear regression”]}$$

Y = response variable (dependent variable)

X = predictor variable (independent variable, covariate)

Why not $y = mx + b$? Form above can be more easily generalized to more than one predictor variable.

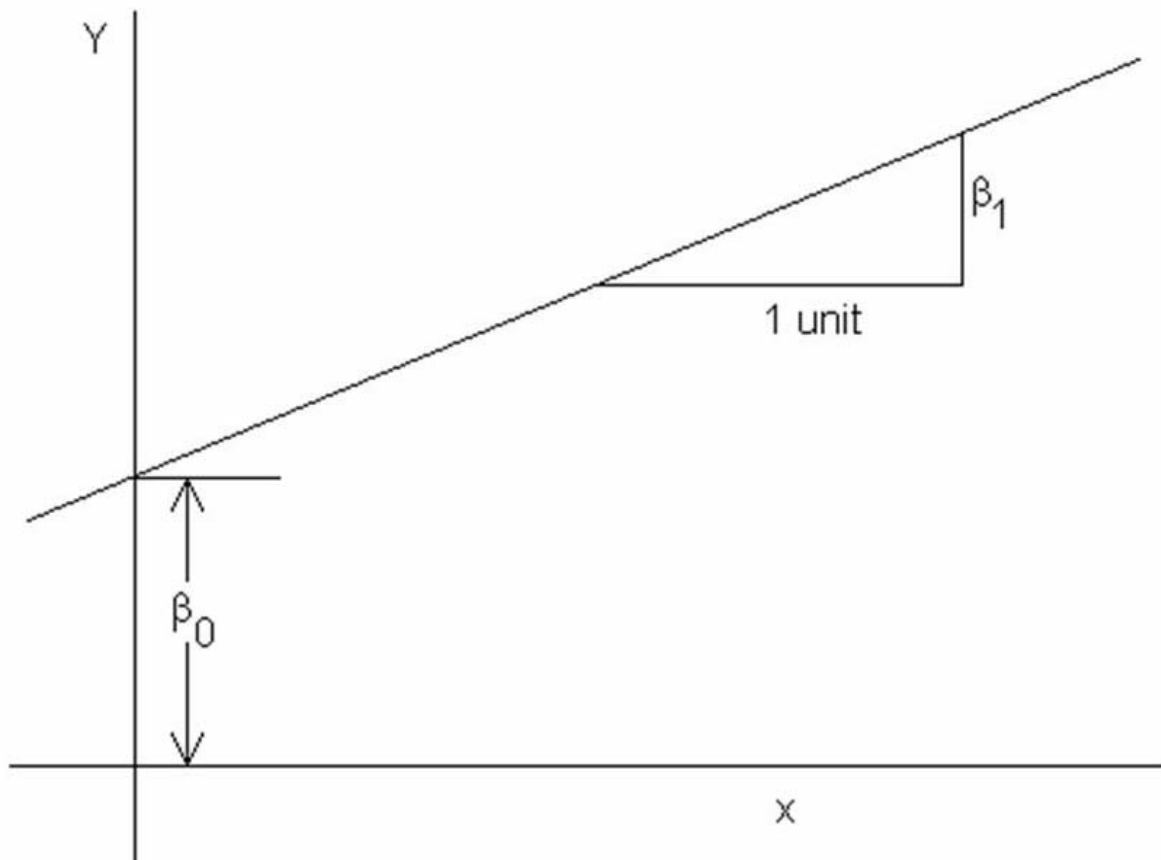
β_0 = y-intercept, value of “Y” at “X=0”

β_1 = slope, how “Y” changes with unit change in “X”

β_0 and β_1 are parameters. What does this mean?

Which parameter is generally of more interest? Why?

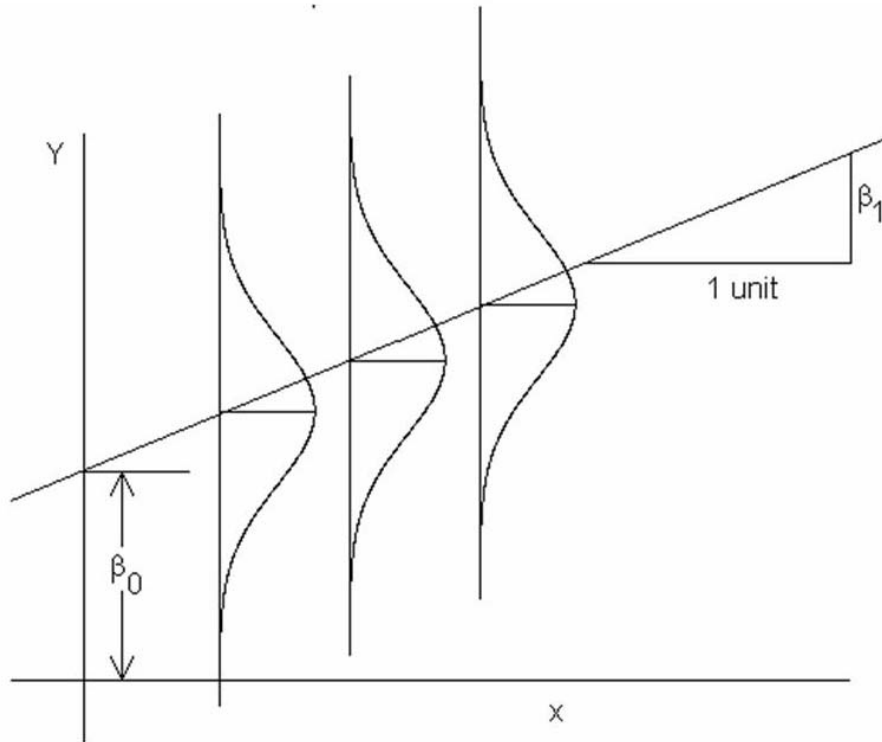
β_1 = contains information about the relationship between the two variables.



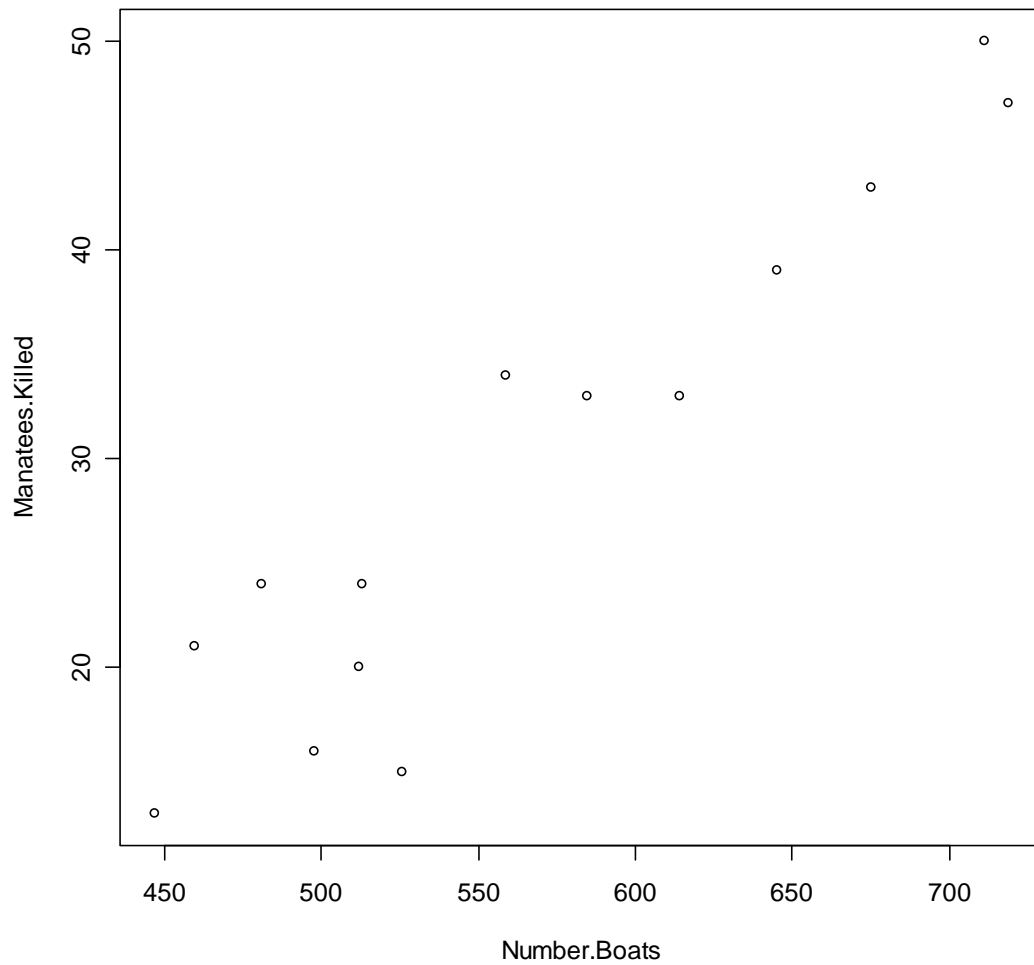
Formal assumptions:

1. relation linear - on average error = 0 [$E(\varepsilon_i) = 0$] $\rightarrow E(Y_i) = \beta_0 + \beta_1 X_i$
2. Constant variance - $V(\varepsilon_i) = \sigma^2 \rightarrow V(Y_i) = \sigma^2$
3. ε_i independent
4. $\varepsilon_i \sim \text{Normal}$

Pictorially This is What the Model is Assuming



Least squares - minimize : $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$



Solution:

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

Interpretation: Units?

SCOPE OF THE MODEL:

Example (Manatee): $b_0 = -41.43$ and $b_1 = 0.125$

Interpretation:

Intercept: When no boats were registered, predict -41.4 manatee death ??? Notice that $x=0$ is well outside the SCOPE of the model.

Slope: For each additional $x=1$ (1000) boats, predict an increase of 0.1 manatee deaths. Maybe a better interpretation, for each additional $x=10$ (10,000) boats, predict an additional manatee death.

How do you deal with the intercept? Reparameterize the model by rescaling the X variable.

[intercept is the average response at the mean X level]

[intercept is the average response at $X=447$]

Issues

Leverage = points with high/low values of the predictor variable X ("outliers" in the X direction)

Influential = omitting point causes estimates of the regression coefficients to change dramatically

Outlier = point with a large residual (more to come!)

Recall from your first stat class, $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ with "n-1" degrees of freedom

Pay penalty b/c mean unknown and estimated by \bar{y}

How about in regression?

Mean at any value of "x" is estimated by $\hat{y} = b_0 + b_1x$

So in regression, we estimate the variance by $s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n-2}$

"mean squared residual"

"mean squared error"

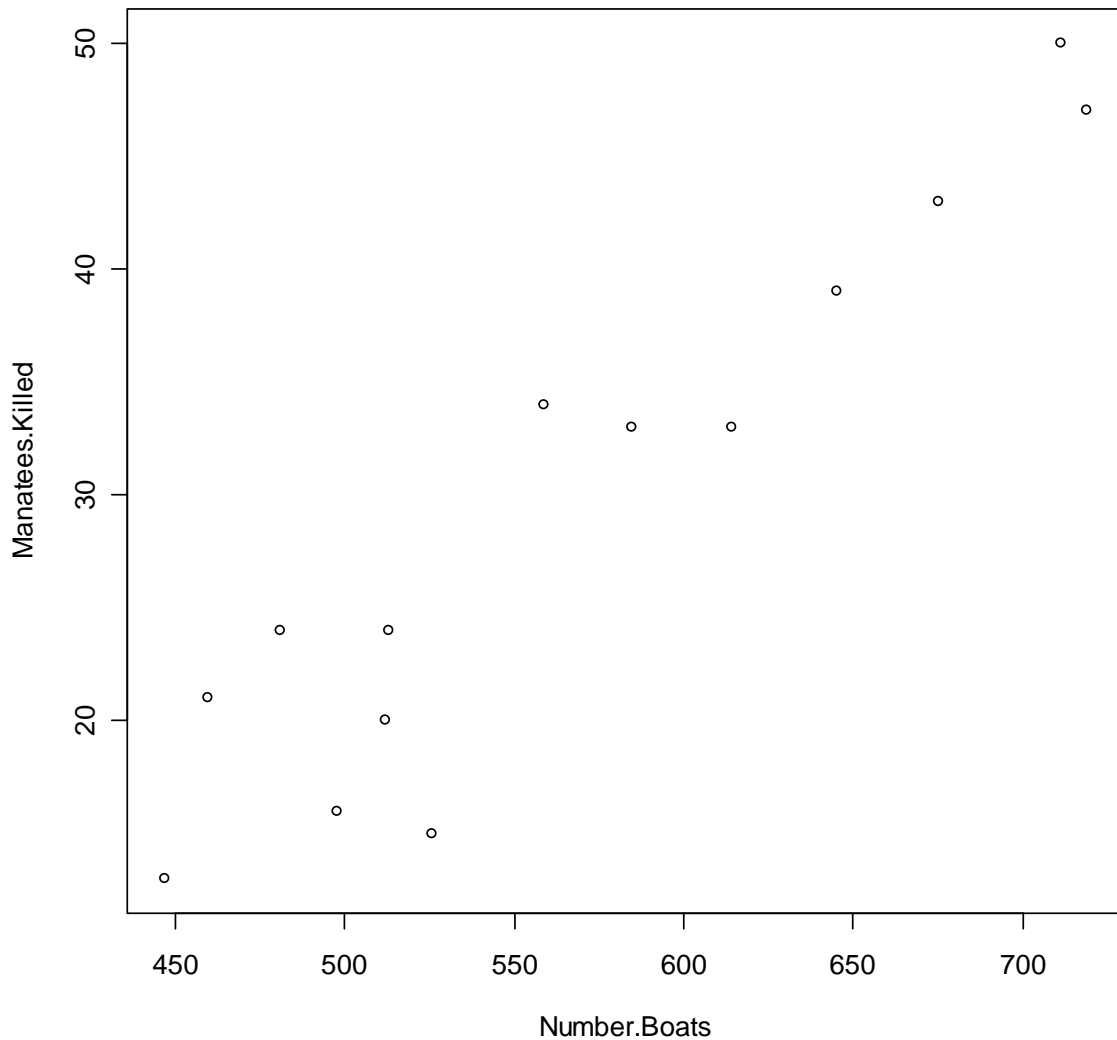
"s" = sample std. dev. around the regression line/ std. error of estimate/residual std. dev.

How do we use the estimate of σ^2 ?

1. If $\varepsilon \sim N$, then expect approx. 95% of residuals to be within +/- 2 s of 0 (more to come)
2. Used in inference for the regression coefficients

Using R to fit the simple regression model

```
> #  
> # Manatee and Sheep Class Notes SLR R Example  
> #  
> Manatee = as.data.frame(scan(what=list(Year=0, Number.Boats=0,  
Manatees.Killed=0), sep=",", dec="."))  
1: 77, 447, 13  
2: 78, 460, 21  
3: 79, 481, 24  
4: 80, 498, 16  
5: 81, 513, 24  
6: 82, 512, 20  
7: 83, 526, 15  
8: 84, 559, 34  
9: 85, 585, 33  
10: 86, 614, 33  
11: 87, 645, 39  
12: 88, 675, 43  
13: 89, 711, 50  
14: 90, 719, 47  
15:  
Read 14 records  
> attach(Manatee)  
> plot(Number.Boats, Manatees.Killed)
```



```

> Manatee.Regression = lm(Manatees.Killed ~ Number.Boats)
> print(Manatee.Regression)

Call:
lm(formula = Manatees.Killed ~ Number.Boats)

Coefficients:
(Intercept)  Number.Boats
  -41.4304      0.1249

> summary(Manatee.Regression)

Call:
lm(formula = Manatees.Killed ~ Number.Boats)

Residuals:
    Min       1Q   Median       3Q      Max
-9.24681 -2.02166  0.02172  2.33692  5.63275

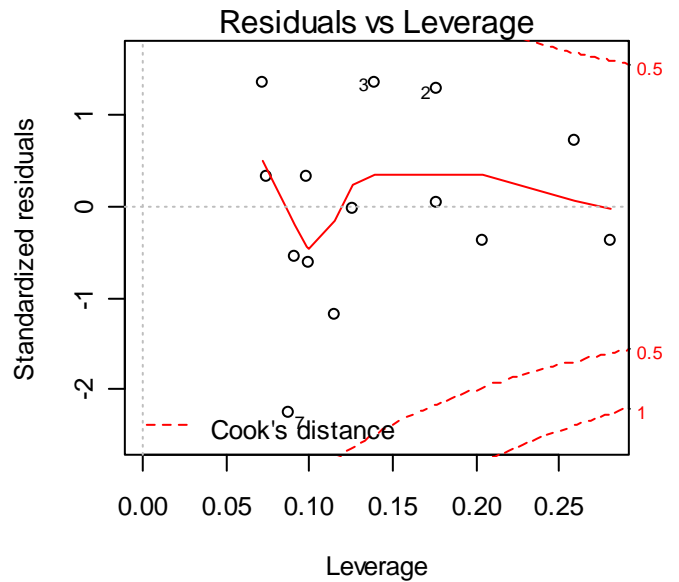
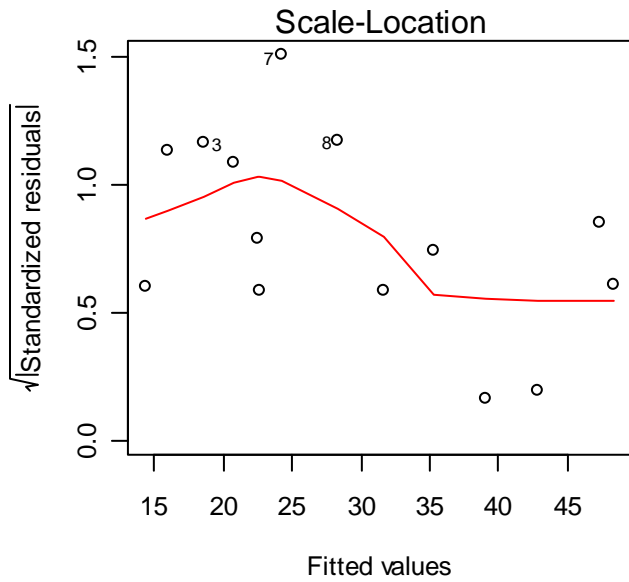
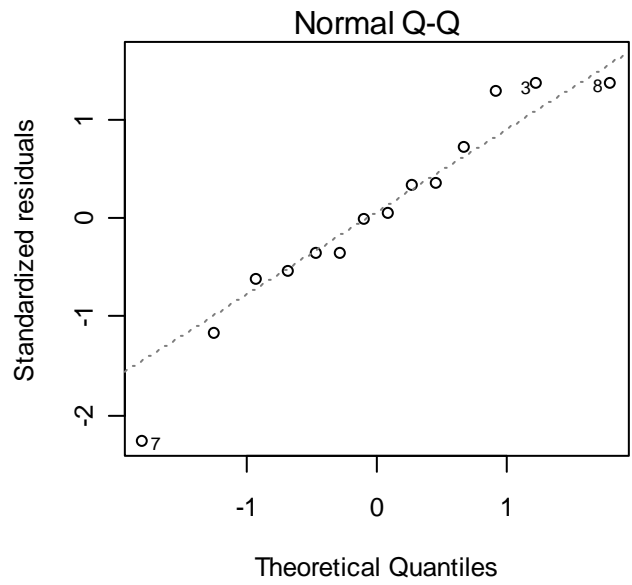
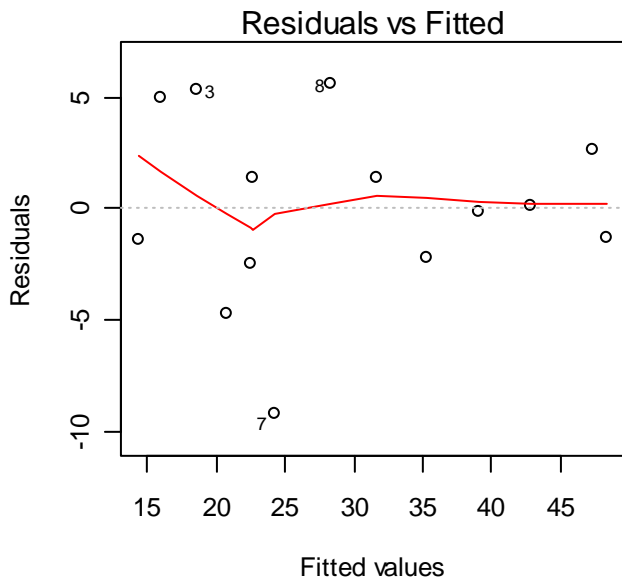
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -41.4304     7.4122  -5.589 0.000118 ***
Number.Boats   0.1249     0.0129   9.675 5.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.276 on 12 degrees of freedom
Multiple R-Squared:  0.8864,    Adjusted R-squared:  0.8769
F-statistic: 93.61 on 1 and 12 DF,  p-value: 5.109e-07

> anova(Manatee.Regression)
Analysis of Variance Table

Response: Manatees.Killed
      Df Sum Sq Mean Sq F value    Pr(>F)
Number.Boats  1 1711.98  1711.98   93.615 5.109e-07 ***
Residuals    12   219.45    18.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> par(mfrow=c(2,2))
> plot(Manatee.Regression)
> par(mfrow=c(1,1))

```



Using SAS to fit the simple regression model

```

/*
  example sas program that does simple linear regression
*/

options ls=75;

data example1;
  input year nboats manatees;
  cards;
77 447 13
78 460 21
79 481 24
80 498 16
81 513 24
82 512 20
83 526 15
84 559 34
85 585 33
86 614 33
87 645 39
88 675 43
89 711 50
90 719 47
;

ODS RTF file='D:\baileraj\Classes\Fall 2003\sta402\SAS-programs\linreg-output.rtf';

proc reg;
title 'Number of Manatees killed regressed on the number of boats registered in
Florida';
  model manatees = nboats / p r cli clm;
  plot manatees*nboats="o" p.*nboats="+" / overlay;
  plot r.*nboats r.*p.;
run;

ODS RTF CLOSE;

```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1711.97866	1711.97866	93.61	<.0001
Error	12	219.44991	18.28749		
Corrected Total	13	1931.42857			

Root MSE	4.27639	R-Square	0.8864
Dependent Mean	29.42857	Adj R-Sq	0.8769
Coeff Var	14.53141		

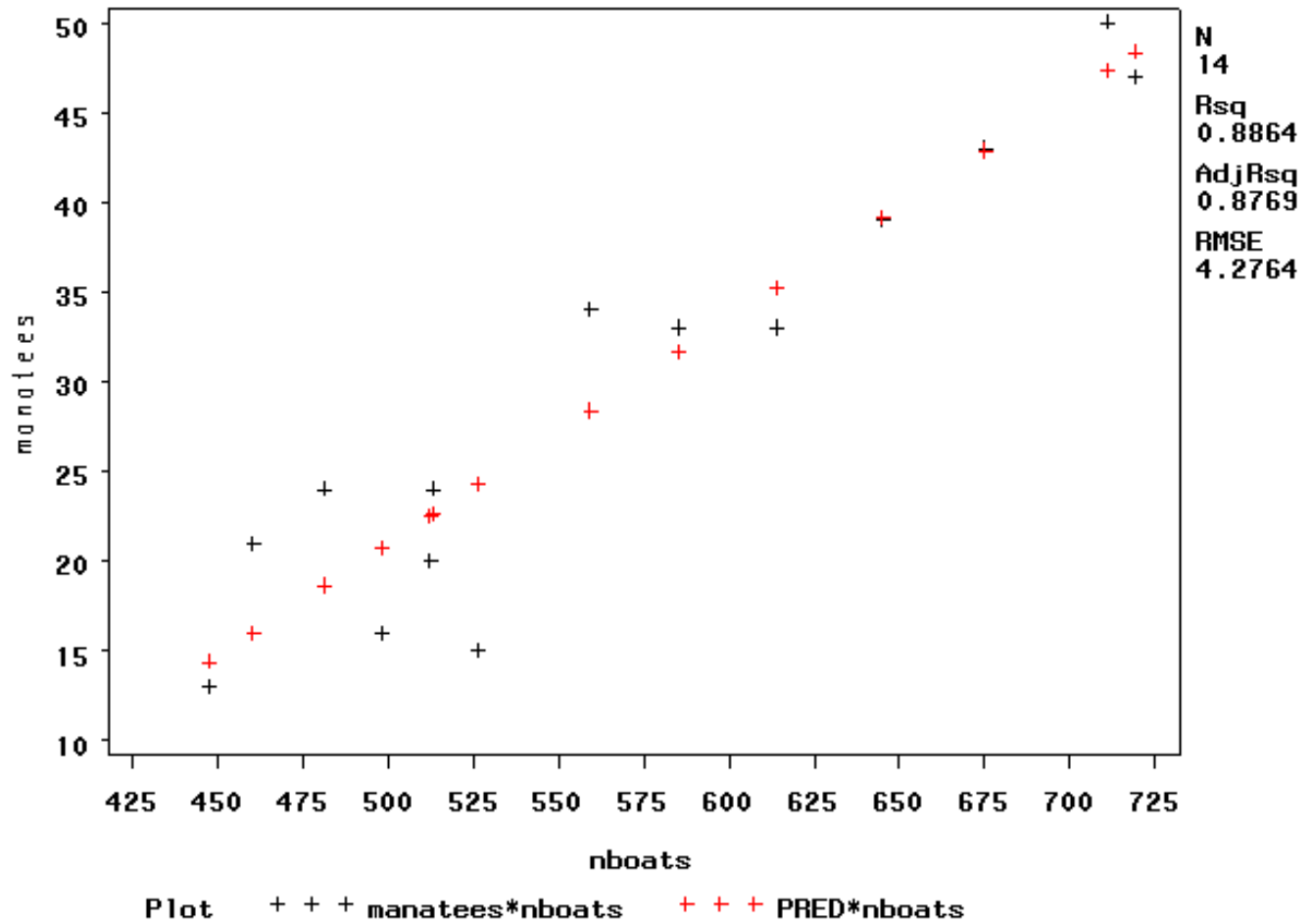
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-41.43044	7.41222	-5.59	0.0001
nboats	1	0.12486	0.01290	9.68	<.0001

Output Statistics										
Obs	Dep Var manatees	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual
1	13.0000	14.3827	1.9299	10.1779	18.5876	4.1604	24.6050	-1.3827	3.816	-0.362
2	21.0000	16.0059	1.7974	12.0896	19.9222	5.8989	26.1130	4.9941	3.880	1.287
3	24.0000	18.6280	1.5976	15.1472	22.1089	8.6816	28.5745	5.3720	3.967	1.354
4	16.0000	20.7507	1.4528	17.5853	23.9161	10.9102	30.5911	-4.7507	4.022	-1.181
5	24.0000	22.6236	1.3420	19.6997	25.5475	12.8582	32.3891	1.3764	4.060	0.339
6	20.0000	22.4987	1.3488	19.5600	25.4375	12.7288	32.2687	-2.4987	4.058	-0.616
7	15.0000	24.2468	1.2622	21.4968	26.9968	14.5320	33.9616	-9.2468	4.086	-2.263
8	34.0000	28.3672	1.1482	25.8656	30.8689	18.7198	38.0147	5.6328	4.119	1.367
9	33.0000	31.6137	1.1650	29.0753	34.1520	21.9566	41.2707	1.3863	4.115	0.337
10	33.0000	35.2346	1.2909	32.4221	38.0472	25.5019	44.9673	-2.2346	4.077	-0.548
11	39.0000	39.1054	1.5187	35.7963	42.4144	29.2178	48.9929	-0.1054	3.998	-0.0264
12	43.0000	42.8512	1.7974	38.9349	46.7675	32.7442	52.9582	0.1488	3.880	0.0383
13	50.0000	47.3462	2.1762	42.6048	52.0877	36.8917	57.8007	2.6538	3.681	0.721
14	47.0000	48.3451	2.2647	43.4109	53.2794	37.8018	58.8884	-1.3451	3.628	-0.371

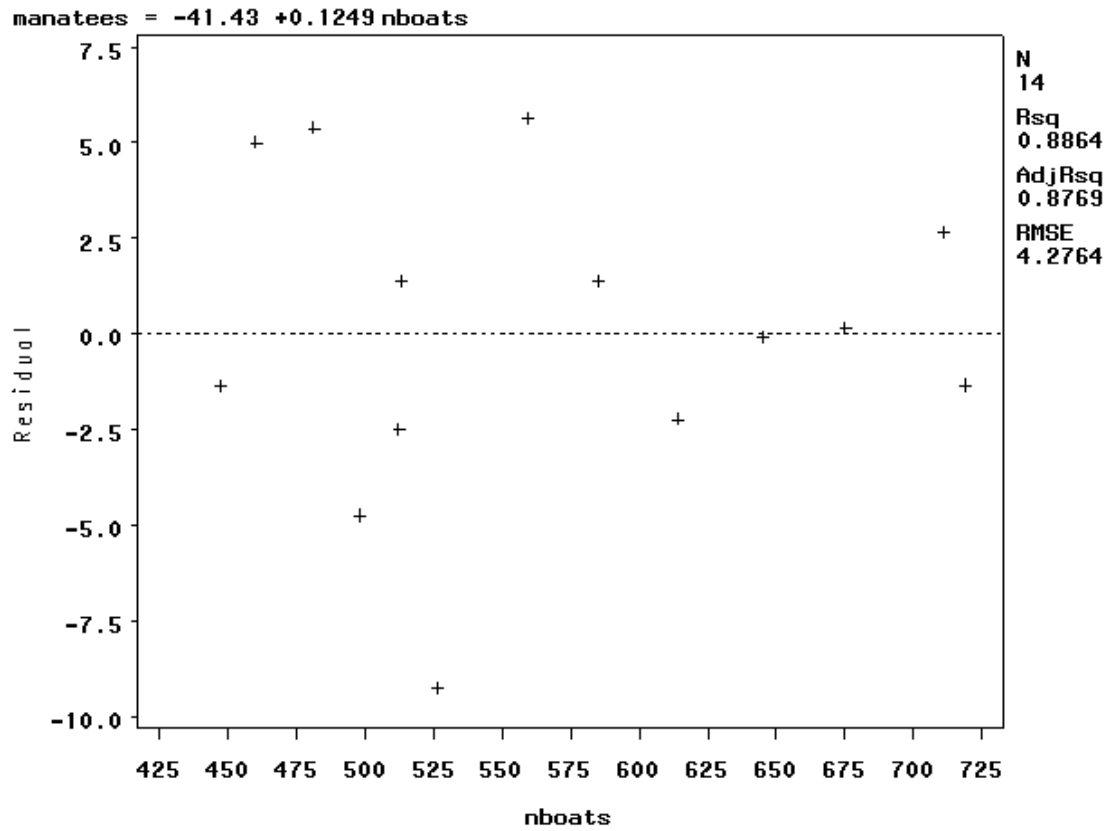
Output Statistics				
Obs	-2	-1	0 1 2	Cook's D
1				0.017
2		**		0.178
3		**		0.149
4		**		0.091
5				0.006
6		*		0.021
7		****		0.244
8		**		0.073
9				0.005
10		*		0.015
11				0.000
12				0.000
13		*		0.091
14				0.027
Sum of Residuals				0
Sum of Squared Residuals				219.44991
Predicted Residual SS (PRESS)				281.76275

Number of Manatees killed regressed on the number of boats registered in Florida

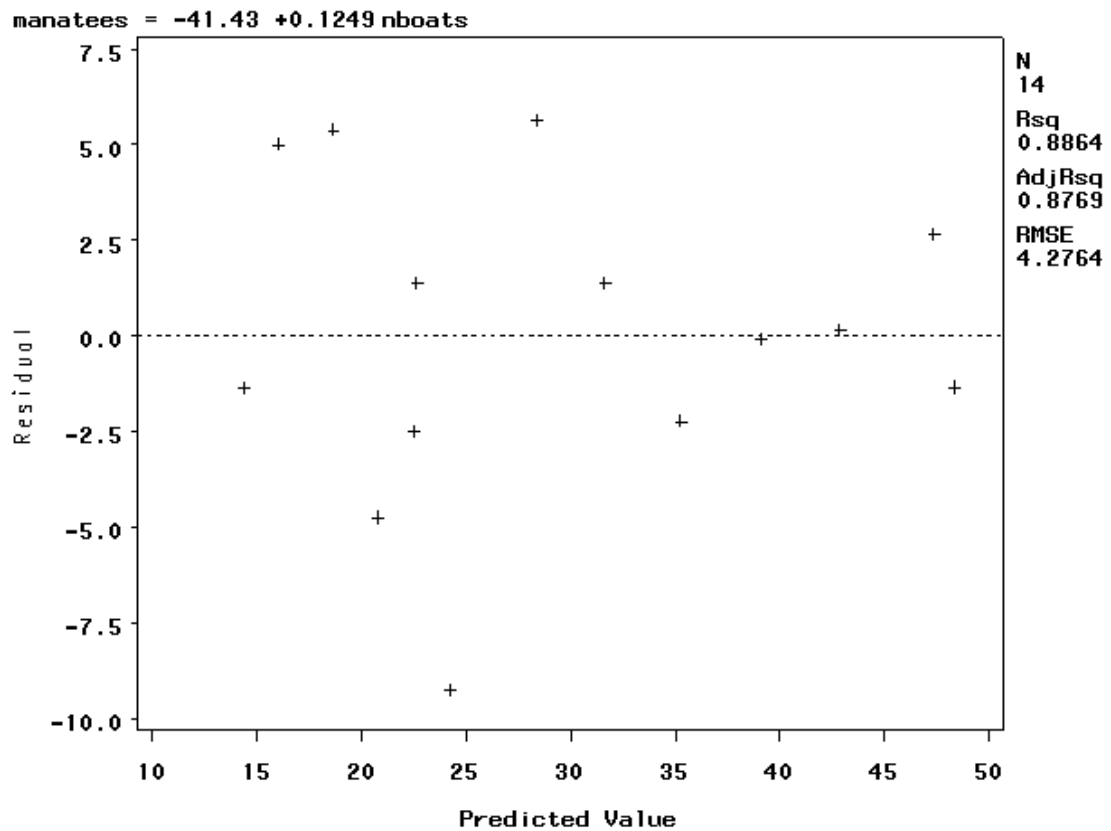
$$\text{manatees} = -41.43 + 0.1249 \text{ nboats}$$



Number of Manatees killed regressed on the number of boats registered in Florida



Number of Manatees killed regressed on the number of boats registered in Florida



*** see Rcmdr handout in Blackboard documents

Using R to decide if a "linear" relationship is appropriate

