

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

Winter 2009

Week 2--IES 612-STA 4-573-STA 4-576-17jan09.doc

OUTLINE of discussions ...

Inference in Simple Linear Regression

i) confidence intervals and hypothesis tests for regression coefficients (β s)

[include a model comparison F test that will generalize to multiple regression context]

[digression: applying linear regression to nonlinear relationships after suitable transformation]

ii) confidence intervals for mean responses

iii) prediction intervals for new responses

Describing the strength of association between two variables - coefficients of correlation (r) and determination (R^2)

Confidence Interval for $\beta_1 \rightarrow b_1 \pm t_{\alpha/2, n-2} SE(b_1)$

$t_{\alpha/2, n-2}$ = critical value from a t-distribution with "n-2" degrees of freedom that cuts off " $\alpha/2$ " of the distribution above it (or equivalently, $1 - \alpha/2$ below it).

In R, the "qt(lower tail area, degrees of freedom)" function can be used to find this quantity. For example,

```
qt(1-.10/2,12) # find  $t_{12}$  value with 0.05 to the right of it (0.95 to the left of this value)
[1] 1.782288
```

Example: Manatee data - 90% CI for the SLOPE

```
manatee.df
  nboats killed
1     447     13
2     460     21
3     481     24
```

4	498	16
5	513	24
6	512	20
7	526	15
8	559	34
9	585	33
10	614	33
11	645	39
12	675	43
13	711	50
14	719	47

Using R

```
> Manatee.lmfit <- lm(killed ~ nboats, data=manatee.df)
> summary(Manatee.lmfit)

...
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 ***
```

And by hand we obtain: 90% CI $\Rightarrow \alpha=0.10 \Rightarrow \alpha/2=0.05$ and $n=14 \Rightarrow n-2 = 12 \Rightarrow t_{.05,12} = 1.782$

$$b_1 = 0.1249 \quad SE(b_1) = 0.0129 \quad 0.1249 \pm (1.782)(0.129)$$

$$0.1249 \pm .023$$

$$0.102 < \beta_1 < 0.148$$

Or (kinda by hand with R assistance)

```
0.1248617 + c(-1,1)* qt(1-.10/2,12)*0.01290497
[1] 0.1018613 0.1478621
```

Or using R and a built-in function of R, `confint()`, we obtain:

```
> confint(Manatee.lmfit, level=0.90)
              5 %          95 %
(Intercept) -54.6411415 -28.2197364
Number.Boats  0.1018613  0.1478621
```

Interpretation?

Based on this confidence interval for β_1 , could you draw any conclusions related to a hypothesis test for this parameter?

Why are you generally more interested in inference about β_1 as opposed to β_0 ?

Using Rcmdr

1. SELECT "Manatee.df" FOR USE

(since Manatee.df is an object that has been previously saved in my workspace)

Data > Active data set > Select active data set ... >

2. FIT THE LINEAR REGRESSION MODEL

Statistics > Fit models > Linear regression ...

3. REQUEST THE CONFIDENCE INTERVAL CONSTRUCTION

Type the "confint" function request in either

i) R Console

```
Model: Manatee.lmfit  
Models > Confidence Intervals...  
Confidence Level: 0.90
```

```
> confint(Manatee.lmfit, level=.90)  
              5 %           95 %  
(Intercept) -54.6411415 -28.2197364  
Number.Boats  0.1018613  0.1478621
```

ii) Rcmdr Script window and the "Submit" the function

Using SAS

```
proc glm;  
  model manatees = nboats / clparm alpha=0.10;
```

Parameter	Estimate	Standard Error	t Value	Pr > t	90% Confidence Limits	
Intercept	-41.43043895	7.41221723	-5.59	0.0001	-54.64114147	-28.21973642
nboats	0.12486169	0.01290497	9.68	<.0001	0.10186132	0.14786207

F Test of β_1 (more interesting when we get to multiple regression and anova models)

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_a: \beta_1 \neq 0$$

TS: $F_{\text{obs}} = [SS(\text{Reg})/1] / [SS(\text{Resid})/(n-2)]$, where:

$$SS(\text{Reg}) = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SS(\text{Resid}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

RR: Reject H_0 if $F_{\text{obs}} > F_{\alpha, 1, n-2}$

P-value: Reject H_0 if $P(F_{1, n-2} > F_{\text{obs}}) < \alpha$

Conclusions

Example: Manatee data

Using R

```
> Manatee.lmfit = lm(Manatees.Killed ~ Number.Boats)
> anova(Manatee.lmfit)
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
```

$F_{\text{obs}} = 93.615$ with associated P-value = $5.109e-07 = 0.0000005109$

$SS(\text{Reg}) = 1711.98$ and $SS(\text{Resid}) = 219.45$

$\sigma\text{-hat}^2 = s^2 = \text{MSE} = 18.29$

Using RCommander

Model: **Manatee.lmfit**

Models > Hypothesis Tests > ANOVA Table

```
> Anova(Manatee.lmfit)
Anova Table (Type II tests)

Response: Manatees.Killed
      Sum Sq Df F value    Pr(>F)
Number.Boats 1711.98  1  93.615 5.109e-07 ***
Residuals    219.45 12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using SAS

We obtain the same results as with R.

Notes and comments about an ANOVA table.

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			

1. ANOVA = ANalysis Of Variance
2. "Sum of Squares" represents a partitioning of the TOTAL variation into variability "explained" by a model (the linear regression model here) and the variability NOT explained (residual error)
3. SS(Total) [Corrected Total SS= 1931.43 above] is "partitioned" into the SS(Regression) [Model SS =1711.98 above] and SS(Residual) [Error SS = 219.45].
4. Mean Squares (MS) are defined as SS/(degrees of freedom) AND are "VARIANCES!"
5. A good regression model will have $SS(\text{Regression}) > SS(\text{Residual})$ which often translates into a large value of F_{obs} .
6. Alternative interpretation:

$SS(\text{Residual})$ = error in predicting response "y" when using the linear regression model.

$SS(\text{Total})$ = error in predicting response "y" when using Y_{BAR} .

$SS(\text{Regression}) = SS(\text{Total}) - SS(\text{Residual})$ measures how much better the YHAT prediction model is when compared to Y_{BAR} . (more to come later)

Alternatively, T Test of β_1 :

WHY DO WE NEED THIS TEST? LOOK AT H_A of the F Test!

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0 \text{ [some assoc.]}$$

$$H_a: \beta_1 < 0 \text{ [negative assoc.]}$$

$$H_a: \beta_1 > 0 \text{ [positive assoc.]}$$

$$TS: t_{obs} = \frac{b_1 - 0}{s/\sqrt{S_{XX}}}$$

RR: Reject H_0 if $|t_{obs}| > t_{\alpha/2, n-2}$

P-value: $2 * P(t_{n-2} > |t_{obs}|)$

Conclusions: Reject/Fail-to-reject H_0 ?

$$t_{obs} < -t_{\alpha, n-2}$$

$$P(t_{n-2} < t_{obs})$$

$$t_{obs} > t_{\alpha, n-2}$$

$$P(t_{n-2} > t_{obs})$$

Using R or RCommander

Model: *Manatee.lmfit*

Models > Summarize Model

```
> summary(Manatee.lmfit)
```

Stuff deleted...

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	***

Stuff deleted...

Using SAS

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

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$ [some assoc.]

TS: $t_{obs} = \frac{b_1 - 0}{s/\sqrt{S_{XX}}} = \frac{0.12486}{0.01290} = 9.68$ and with a P-value = $5.109e-07 = 0.0000005109 < 0.0001$

Decision/Conclusion: REJECT H_0 and conclude that there is a linear relationship between the number of manatees killed and the number of boats registered in Florida.

Comments: Always write your conclusions in the words of the problem. Translate the symbol representation back to the real world.

A confidence interval demonstrates the magnitude of the linear effect. This corresponds to an "effect size" idea which is often more interesting than a reject/fail to reject hypothesis testing decision.

Tests and Confidence intervals are related. For example, if a $100(1-\alpha)\%$ confidence interval for a parameter, say β_1 , does NOT contain 0 (e.g. $0.102 < \beta_1 < 0.148$), then you would reject $H_0: \beta_1 = 0$ in favor of $H_a: \beta_1 \neq 0$ at significance level α .

Notes and Comments

* Hypothesis tests / Confidence intervals for the intercept, β_0 , are similar.

* Can you select design points to have more precision when estimating the slope? I.e. how should you space your Xs if you want the most precision when estimating the slope?

$$\text{SE}(b_1) = \frac{S}{\sqrt{S_{xx}}} = \frac{S}{\sqrt{\sum (x_i - \bar{x})^2}}$$
$$\text{SE}(b_0) = S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

Remedial Measures and Transformations

RECALL: 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)

Formal assumptions:

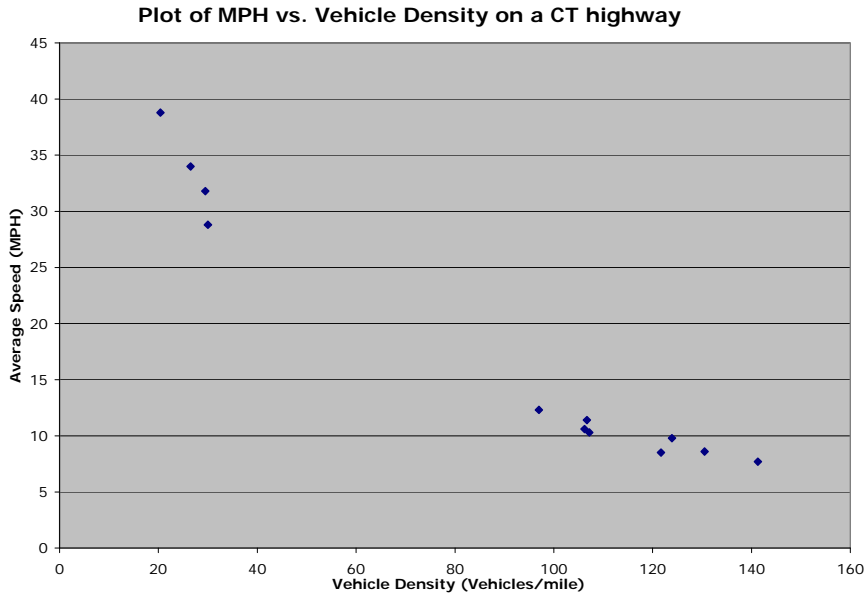
1. relation linear - on average error = 0 [$E(\varepsilon_i) = 0$] $\rightarrow E(Y_i) = \beta_0 + \beta_1 X_i$
2. Constant variance - $V(\varepsilon_i) = \sigma^2 \rightarrow V(Y_i) = \sigma^2$
3. ε_i independent
4. $\varepsilon_i \sim \text{Normal}$

We will talk more about model adequacy. Now, a few remarks about a special case when the first assumption might be violated.

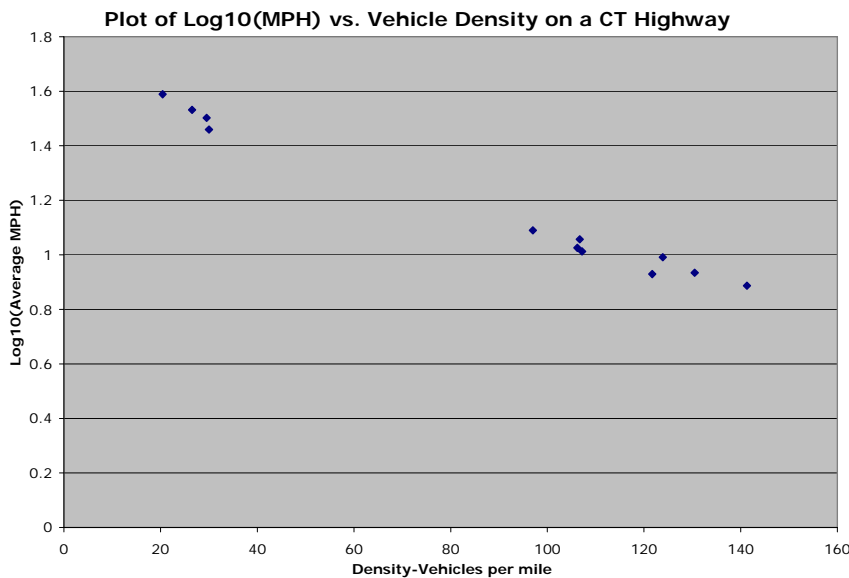
There may be times when a nonlinear relationship might be modeled by linear regression.

Example: MPH and Vehicle Density on a Connecticut Highway

Ref: <http://lib.stat.cmu.edu/DASL/Datafiles/transformationdat.html> and B.D. Greenshields and F.M. Weida, Statistics with Applications to Highway Traffic Analysis, Eno Foundation, 1978, 129-131. (DENS, MPH below)



What if we plot the $\text{Log}(\text{MPH})$ vs. Vehicle Density? Could we use a SLR Model to model $\text{Log}(\text{MPH})$ as a LINEAR function of Vehicle Density?



- * other common examples- exponential growth and decay AND power relationships as encountered in allometric scaling ($Y=aX^b$)
- * LOG10 transformations are also commonly used when the range of the response or predictor variables span many orders of magnitude (e.g. per capita gnp, population size, geographic area).

- * SQRT transformations are often encountered with COUNT data
- * More esoteric transformations can be encountered with proportion data (e.g. arc-sine-sqrt transformation). With logistic regression available, is it worth using this?

See the text on page 537 for steps in choosing an appropriate transformation.

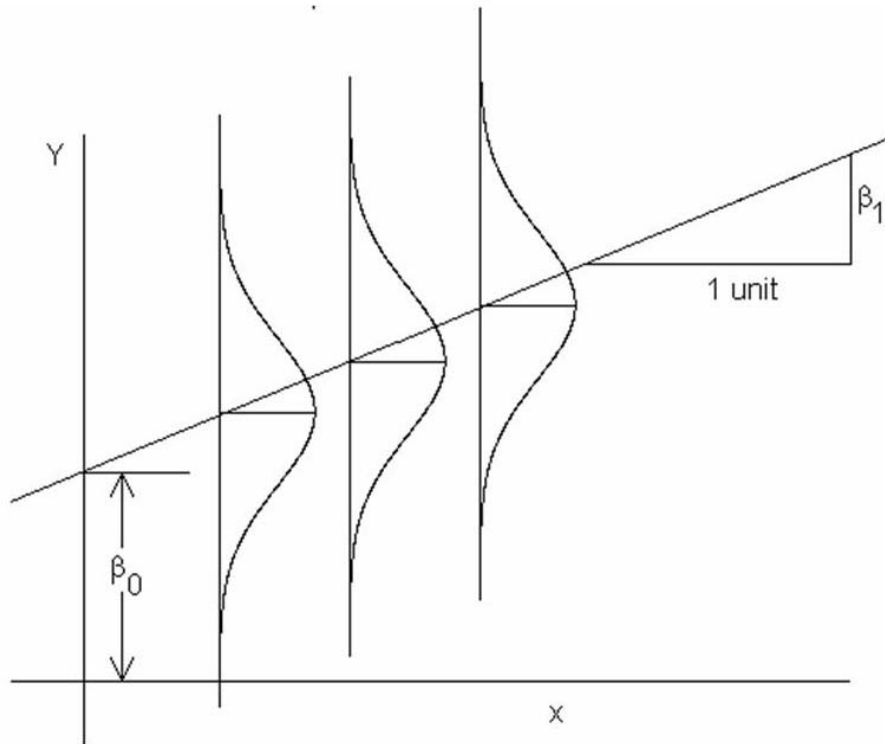
Aside: **Estimating a mean versus Predicting a y value**

Add a plot of a Normal with mean known ask about predicting a value from this popln

Add in her a plot of a normal distribution ask about the mean get a CI

Then add same plot but now ask about predicting a single observation from this population

Other Inference in Regression - **Estimating** the Average response at a particular value of x or **Predicting** a new observation at a particular value of x



X values in the dataset - x_1, \dots, x_n

Denote new value of X: x_{n+1}

Prediction of the mean response (or new response) at this x value: $\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$

SE of this prediction: $s(\hat{y}_{n+1}) = s \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$

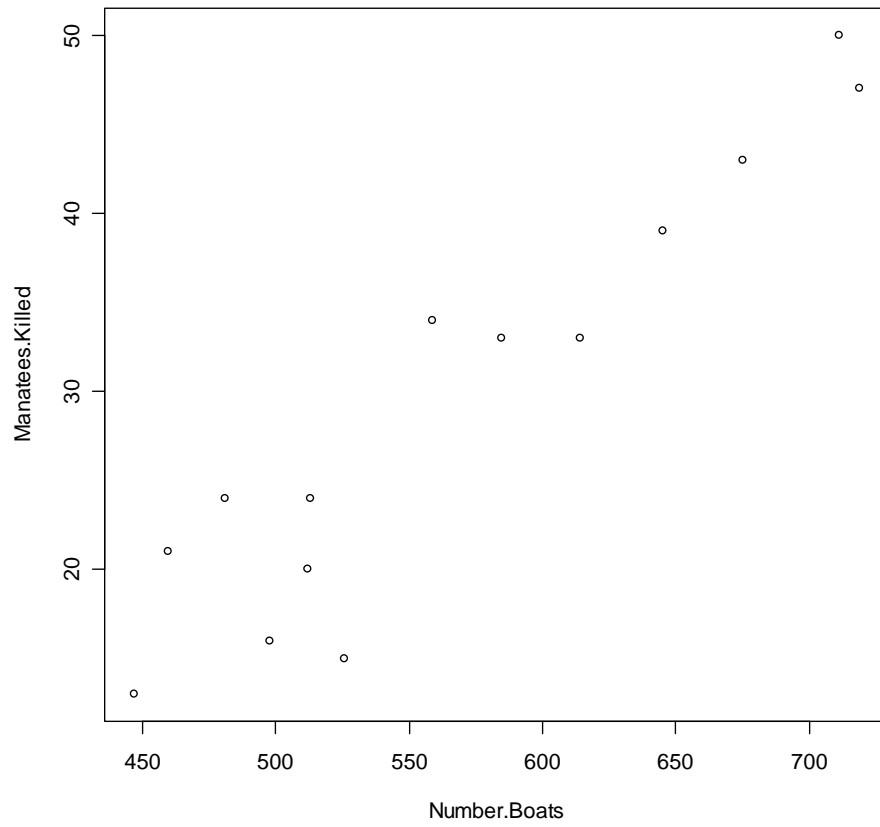
Confidence Interval for the Mean Response

$$\hat{y}_{n+1} \pm t_{\alpha/2, n-2} s \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

Observation: As x_{n+1} get farther from \bar{x} , the SE of the prediction increases (an "extrapolation" penalty)

EXAMPLE

```
> plot(killed~nboats, data=manatee.df)
```



Prediction Interval for a New Response

Both **Uncertainty** in the location of the MEAN RESPONSE and **variability** associated with individual value given the mean response must be considered.

$$\hat{Y}_{n+1} \pm t_{\alpha/2, n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

Using R-----Note that there is no option to obtain these CI's and PI's in RCommander!

```
> predict(Manatee.lmfit, interval="confidence", level=0.95)
      fit      lwr      upr
1  14.38274 10.17790 18.58758
2  16.00594 12.08964 19.92224
3  18.62804 15.14716 22.10891
4  20.75068 17.58527 23.91610
5  22.62361 19.69969 25.54753
6  22.49875 19.55999 25.43750
7  24.24681 21.49678 26.99684
8  28.36725 25.86561 30.86888
9  31.61365 29.07531 34.15199
10 35.23464 32.42207 38.04721
11 39.10535 35.79634 42.41437
12 42.85120 38.93491 46.76750
13 47.34622 42.60479 52.08766
14 48.34512 43.41085 53.27939
> predict(Manatee.lmfit, interval="prediction", level=0.95)
      fit      lwr      upr
1  14.38274  4.160431 24.60504
2  16.00594  5.898901 26.11298
3  18.62804  8.681612 28.57446
4  20.75068 10.910222 30.59115
5  22.62361 12.858151 32.38907
6  22.49875 12.728838 32.26866
7  24.24681 14.532002 33.96162
8  28.36725 18.719811 38.01468
9  31.61365 21.956631 41.27067
10 35.23464 25.501944 44.96734
11 39.10535 29.217763 48.99294
12 42.85120 32.744165 52.95824
13 47.34622 36.891750 57.80070
14 48.34512 37.801786 58.88845
Warning message:
In predict.lm(Manatee.lmfit, interval = "prediction", level = 0.95) :
  Predictions on current data refer to _future_ responses
```

Suppose our new $x_{n+1} = 559$ (which happens to corresponds to the 8th observation)

$25.87 < E(Y_{n+1}) < 30.87$ With 95% confidence we _____

$18.72 < Y_{n+1} < 38.01$ With 95% confidence we _____

Using SAS

```
proc reg;  
  model manatees = nboats / p r cli clm;
```

From Manatee SAS output

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

MEASURES of ASSOCIATION AND STRENGTH

Correlation and Coefficient of Determination

Recall the slope was estimated as:

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Correlation Coefficient:

$$r_{yx} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$
$$r_{yx} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sum_{i=1}^n (x_i - \bar{x})^2 \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$
$$r_{yx} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$$

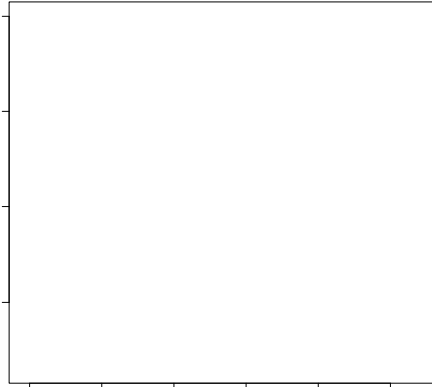
So r_{yx} = (Estimated slope) TIMES [SD(X) / SD(Y)] = "rescaled" slope estimate

Observations:

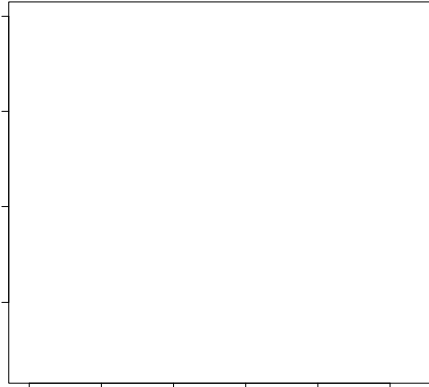
1. Pearson product-moment correlation (other types of correlation coefficients defined - e.g. Spearman's rho)
2. $-1 \leq r_{yx} \leq 1$
3. $r_{yx} = 0$ IMPLIES no LINEAR relationship
4. correlation coefficient tends to increase as range increases
5. test of population correlation coefficient = 0 given but not discussed since equivalent to the test of slopes

SKETCH various scatterplots associated with $r = +1.0$, $r = 0.9$, $r = 0.3$, $r = 0$, $r = -0.3$, $r = -0.9$, and $r = -1.0$.

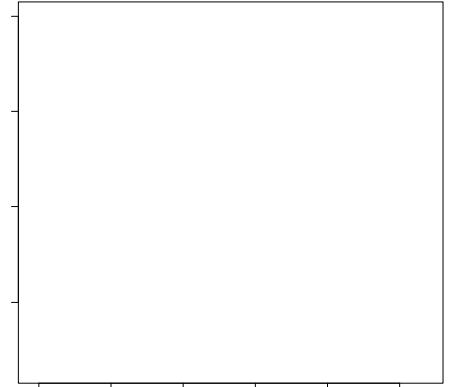
$r = 0.3$



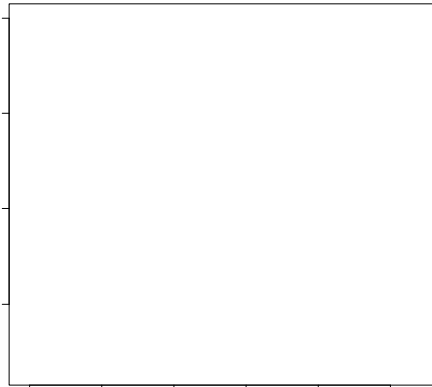
$r = 0.9$



$r = +1.0$



$r = 0$



$r = -0.3$



$r = -0.9$



$r = -1.0$



Coefficient of Determination "R-square"

Coefficient of Determination "R-square":

$$r_{yx}^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{SS(\text{Total}) - SS(\text{Resid})}{SS(\text{Total})} = R^2$$

"proportionate reduction in prediction error when using YHAT instead of YBAR to predict y"

"proportion of total variability accounted for/explained by the linear regression model"

Comments:

- * Coefficient of determination = $(r_{yx})^2 = (\text{correlation coefficient})^2$ for simple linear regression - NOT for multiple regression!
- * When people report a significant correlation coefficient of 0.40 between two variables X and Y, recognize that this means that 16% ($.4 \times .4$) of the variation in one variable is accounted for by its linear association with some other variable.

Example: Manatee deaths and boats registered

Using R

```
> summary(Manatee.lmfit)
. . .
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
> cor(manatee.df)
      nboats      killed
nboats 1.0000000 0.9414773
killed 0.9414773 1.0000000

> cor(manatee.df$killed, manatee.df$nboats)
[1] 0.9414773
```

$r^2 = 0.8864$ so approx. 89% of the variation in the number of manatees killed is explained by a linear relationship with the number of boats registered.

$r = 0.9414773$ implies that there is a STRONG, POSITIVE, LINEAR relation between Manatees killed and the number of boats.

Note that $r^2 = 0.9414773^2 = 0.8864$.

PREVIEW of coming attractions: what if you have multiple predictors? Will correlation coefficients or coefficients of determination be more useful?

Using RCommander

Model: **Manatee.lmfit**
Models > Summarize Model

```
> summary(Manatee.lmfit)

Stuff deleted...

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 ***

Stuff deleted...
```

Statistics > Summaries > Correlation Matrix...

```
> cor(Manatee[,c("Manatees.Killed", "Number.Boats")], use="complete.obs")
              Manatees.Killed Number.Boats
Manatees.Killed  1.0000000    0.9414773
Number.Boats    0.9414773    1.0000000
```

Using SAS

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

* SAS Proc CORR can be used to determine the correlation between variables