Chapter 5: Linear Regression

Study relationship between 2 quantitative variables.

One variable is the response variable, denoted by y.

Measures the outcome of the study.

Also called the dependent/predicted variable.

Other variable is the explanatory variable, denoted by x.

Thought to explain changes in the response.

Also called the independent/predictor variable.

Modeling a relationship with regression

The linear regression model suggests the relationship that predicts the value of y for a given value of x can be expressed as:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

y is the observed value of the dependent variable Y when the value of the independent variable is $X = x$.

β_0 is the y-intercept; the mean of Y when $x = 0$.

Modeling a relationship with regression

The linear regression model suggests the relationship that predicts the value of y for a given value of x can be expressed as:

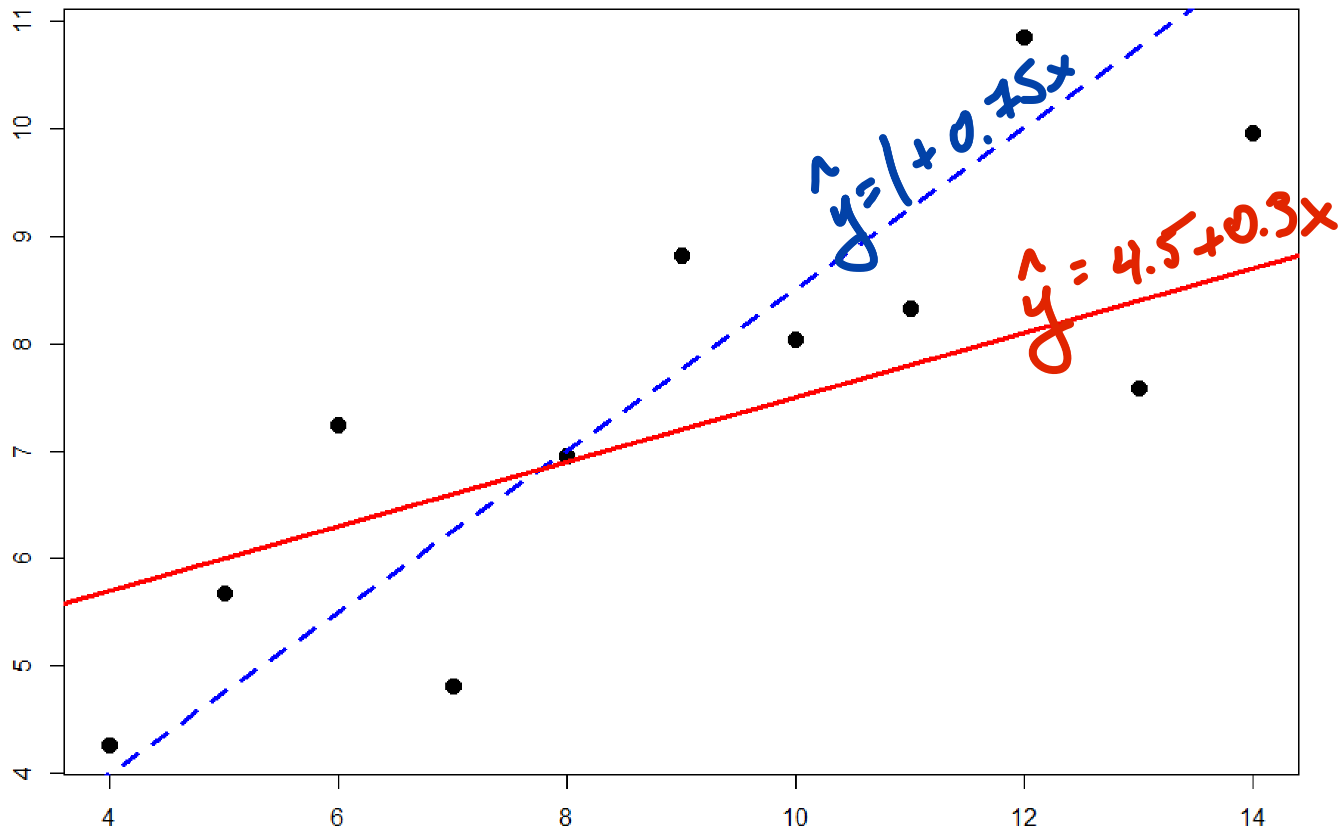
$$Y = \beta_0 + \beta_1 x + \epsilon$$

β_1 is the slope; the change in the mean of Y per unit change in X.

ε is an error term that describes the effect on Y of all factors other than X.

Example: a fictitious (but famous) data set

Var	1	2	3	4	5	6	7	8	9	10	11	*mean*	sd	*r*
X	10	8	13	9	11	14	6	4	12	7	5	9	3.3167	0.8164
Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68	7.5	2.0316	



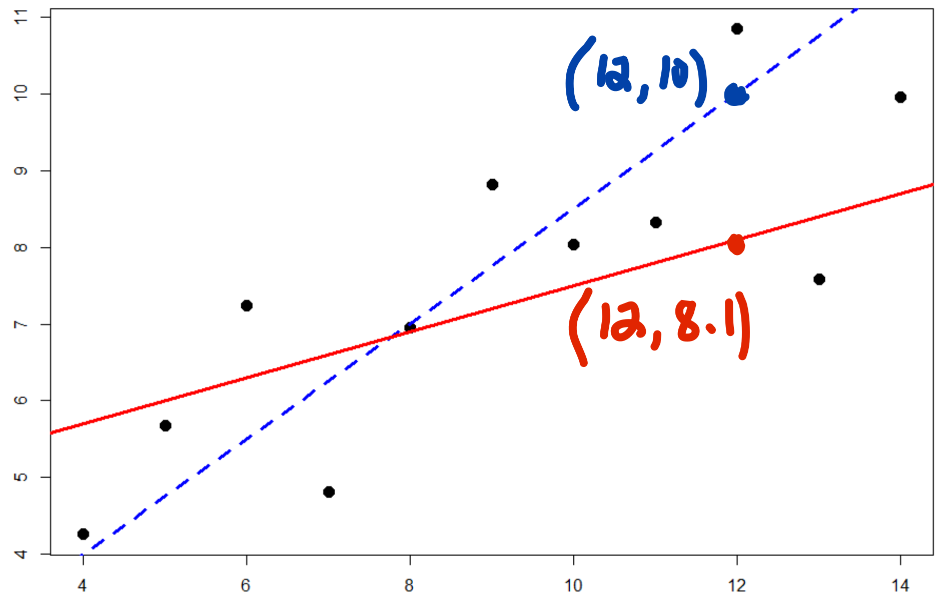
Which line better describes the relationship between x & y?

Interpreting the slope!

b. Suppose an observation has a predictor-value of $x = 12$? What value of y would you predict it had? [Get a prediction from both lines.]

$\hat{y} = 1 + 0.75(12) = 10$

$\hat{y} = 4.5 + 0.3(12) = 8.1$

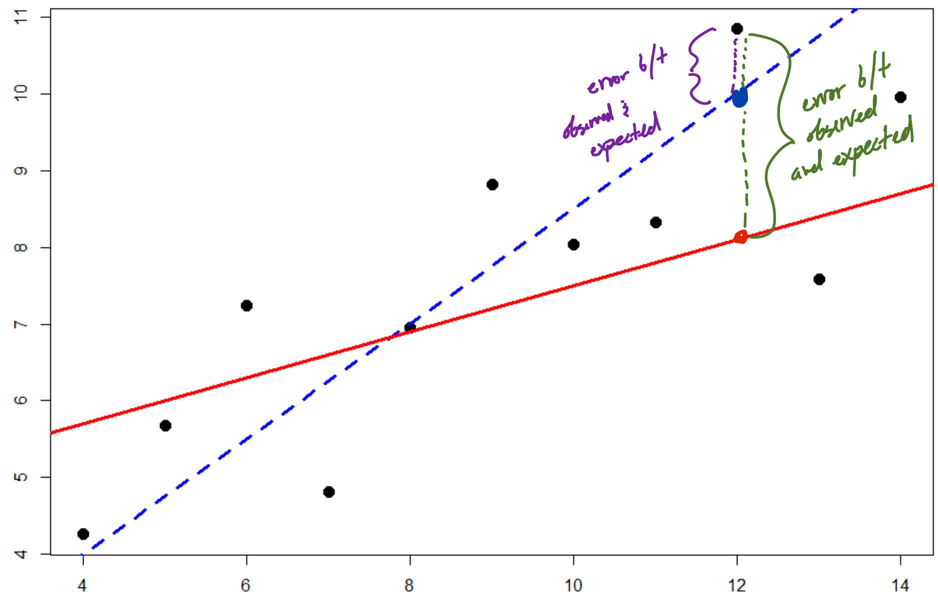


Interpreting the slope!

c. How far off are these estimates from *observed* y-value of the case in the collected data with $x = 12$?

$e = 10.84 - 10 = 0.84$

$e = 10.84 - 8.1 = 2.71$



These errors are called residuals.

Residuals

Residuals are the leftover variation in the data after accounting for the model fit. A good way of thinking about residuals is:

$$Data = \underline{\text{value expected by model} + \text{residual}}$$

Equivalently, we can say ...

$$e = \underline{\text{observed} - \text{expected}}$$

$$(y - \hat{y})$$

Fitting a line by OLS regression

A line that fits the data “**best**” will be the one for which the sum of squared residuals is smallest.

X	Y	$\hat{y}_{dashed}$	$\hat{y}_{solid}$	e_{dashed}	e_{solid}	e^2_{dashed}	e^2_{dashed}
10.00	8.04	8.50	7.50	-0.46	0.54	0.21	0.29
8.00	6.95	7.00	6.90	-0.05	0.05	0.00	0.00
13.00	7.58	10.75	8.40	-3.17	-0.82	10.05	0.67
9.00	8.81	7.75	7.20	1.06	1.61	1.12	2.59
11.00	8.33	9.25	7.80	-0.92	0.53	0.85	0.28
14.00	9.96	11.50	8.70	-1.54	1.26	2.37	1.59
6.00	7.24	5.50	6.30	1.74	0.94	3.03	0.88
4.00	4.26	4.00	5.70	0.26	-1.44	0.07	2.07
12.00	10.84	10	8.1	0.84	2.71	0.71	7.34
7.00	4.82	6.25	6.60	-1.43	-1.78	2.04	3.17
5.00	5.68	4.75	6.00	0.93	-0.32	0.86	0.10

Equation of Ordinary Least Squares (OLS) line

d. Which equation has the smaller sum of squared residuals $\sum e^2$ [i.e., which line better describes the relationship between X and Y?

Blue $\sum_i e_i^2 = 21.31$

Red $\sum_i e_i^2 = 18.98$

So red is superior model.

The OLS regression

KEY IDEA: ordinary least-squares (OLS) regression line will produce the **smallest sum of squared residuals** mathematically possible.

Property 1: An estimate of the slope of the OLS regression is

$b_1 = r\left(\frac{s_y}{s_x}\right)$.

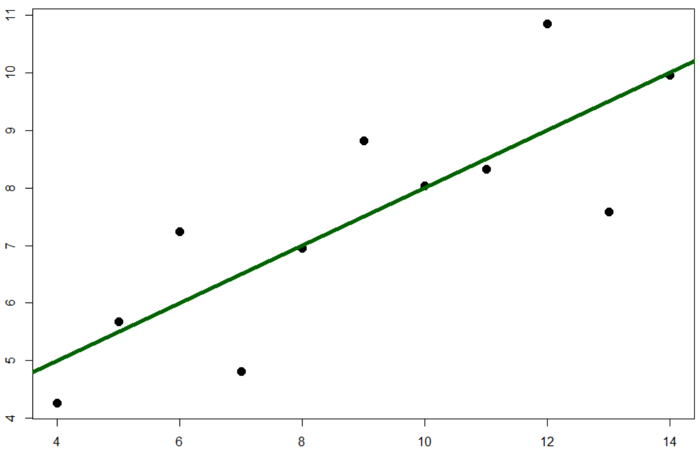
Property 2: The OLS line *must* pass through the point

$(\bar{x}, \bar{y})$, which means an estimate of the y-intercept of the OLS regression is

$b_0 = \bar{y} - b_1\bar{x}$.

The OLS regression

e. Use the summary statistics below to compute the equation of the OLS regression line, plotted with the original data below:



Step 1: Compute b_1, the slope

$$b_1 = 0.8164\left(\frac{2.0316}{3.3167}\right) = 0.5$$

Step 2: Compute b_0, the intercept

$$b_0 = 7.5 - (0.5)9 = 3$$

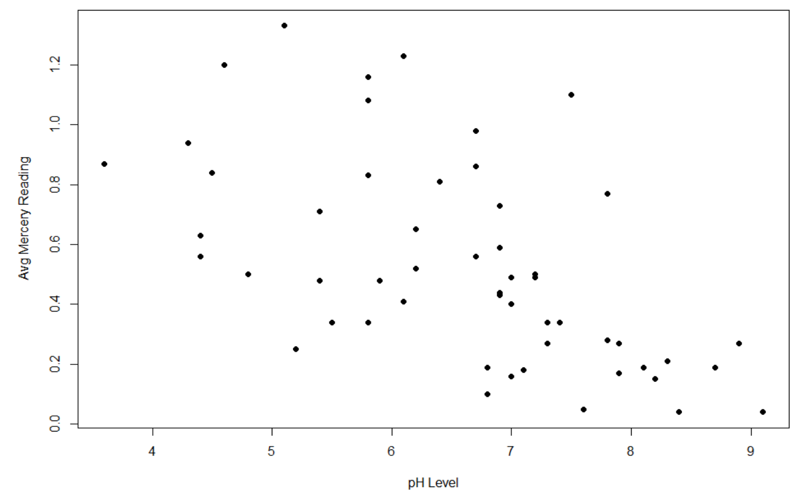
$\hat{y} =$ $3 + 0.5x$

Example 5.1: Predicting Mercury levels from Alkalinity

Page 121

The scatterplot below describes characteristics of water samples taken at $n = 53$ Florida lakes. The acidity (pH) was recorded as well as the average mercury level (in parts-per-million ppm) for a sample of fish (largemouth bass) from each lake.

Variable	$mean$	sd	r
pH level	6.5906	1.288	−0.5754
Avg Mercury	0.5272	0.3410	



Use the summary statistics provided in the table above to compute the equation of the OLS regression line.

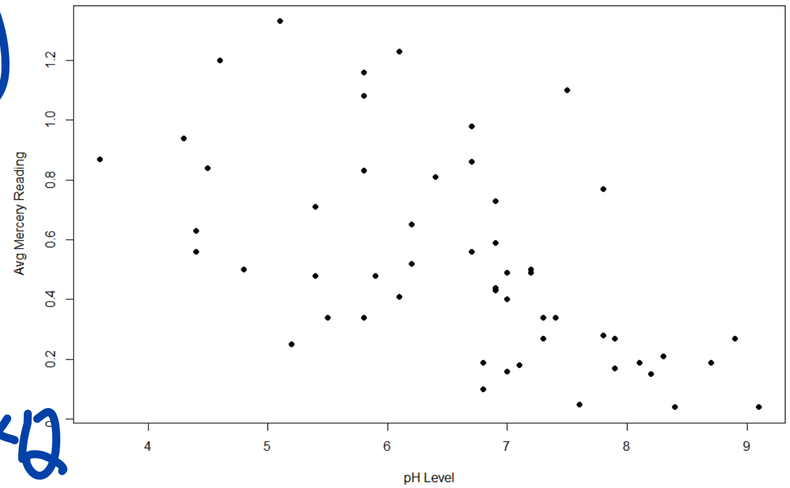
$\hat{y} =$ 1.5309 + -0.1523 $* x$

Example 5.1: Predicting Mercury levels from Alkalinity

b. One of the lakes sampled had a pH level of 5.1 and an average mercury reading of 1.23 ppm. What was the residual for this lake?

$$\hat{y} = 1.5309 - 0.1523(5.1)$$

$$= 0.7542$$



$$e = y - \hat{y} = 1.23 - 0.7542$$

$$= 0.4758 \text{ ppm.}$$

Lecture 5-2: Evaluating OLS Regression

In the previous lecture we learned how to compute the Ordinary Least Squares regression line which, under certain conditions, is the single best-fitting line statistics can produce to summarize a relationship between two quantitative variables.

The next logical question is

how well does the model fit ?

To find the standard error for our estimates we first calculate the

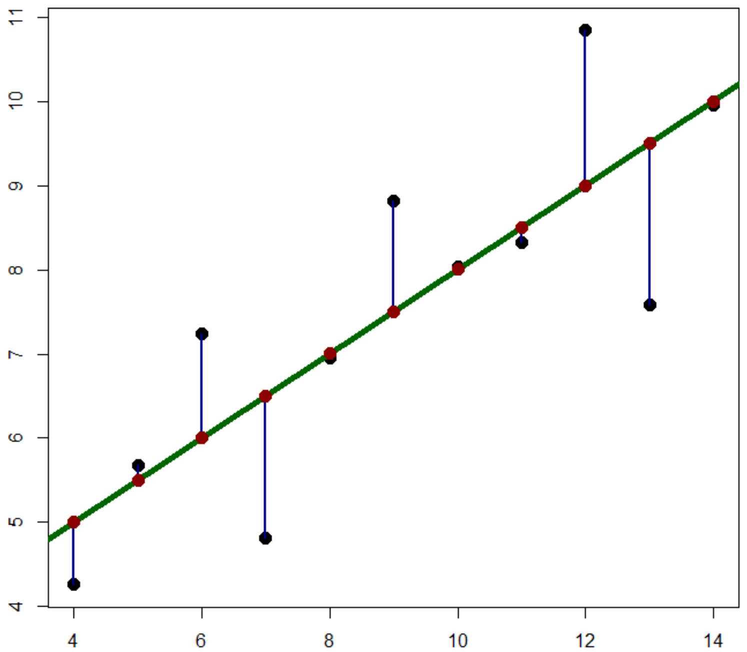
Sum of squared errors:

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

Taking the square root of the average gives s, the

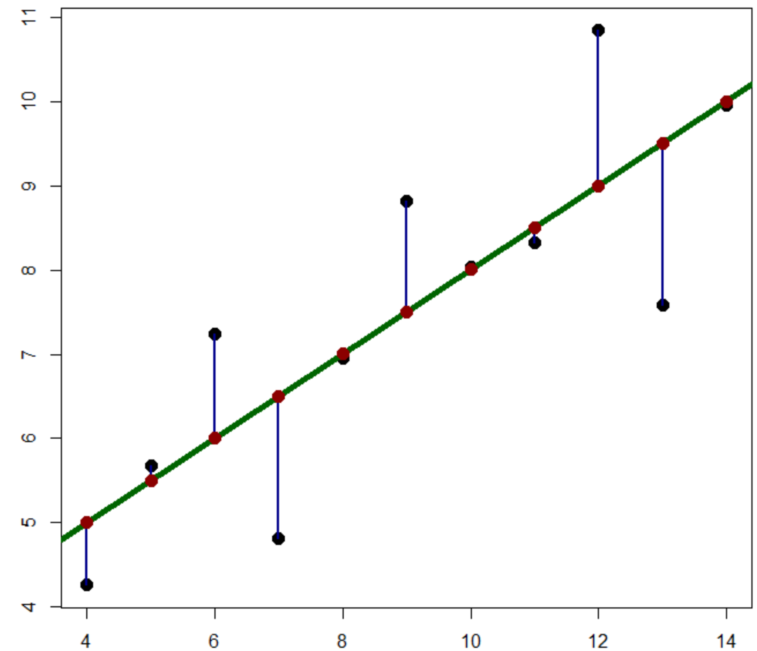
residual std. error:

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$



Lecture 5-2: Evaluating OLS Regression

x	y	$\hat{y}$	e	e^2
10.00	8.04	8.00	0.04	0.0014
8.00	6.95	7.00	-0.05	0.0027
13.00	7.58	9.50	-1.92	3.6952
9.00	8.81	7.50	1.31	1.7111
11.00	8.33	8.50	-0.17	0.0296
14.00	9.96	10.00	-0.04	0.0018
6.00	7.24	6.00	1.24	1.5336
2.00	4.26	4.00	0.26	0.0670
12.00	10.84	9.00	1.84	3.3775
7.00	4.82	6.50	-1.68	2.8281
5.00	5.68	5.50	0.18	0.0319
9.00	7.50	7.50	0.00	0.0000
			SSE	13.2799
			MSE	1.32799
			s	1.1524



Interpreting residual standard error

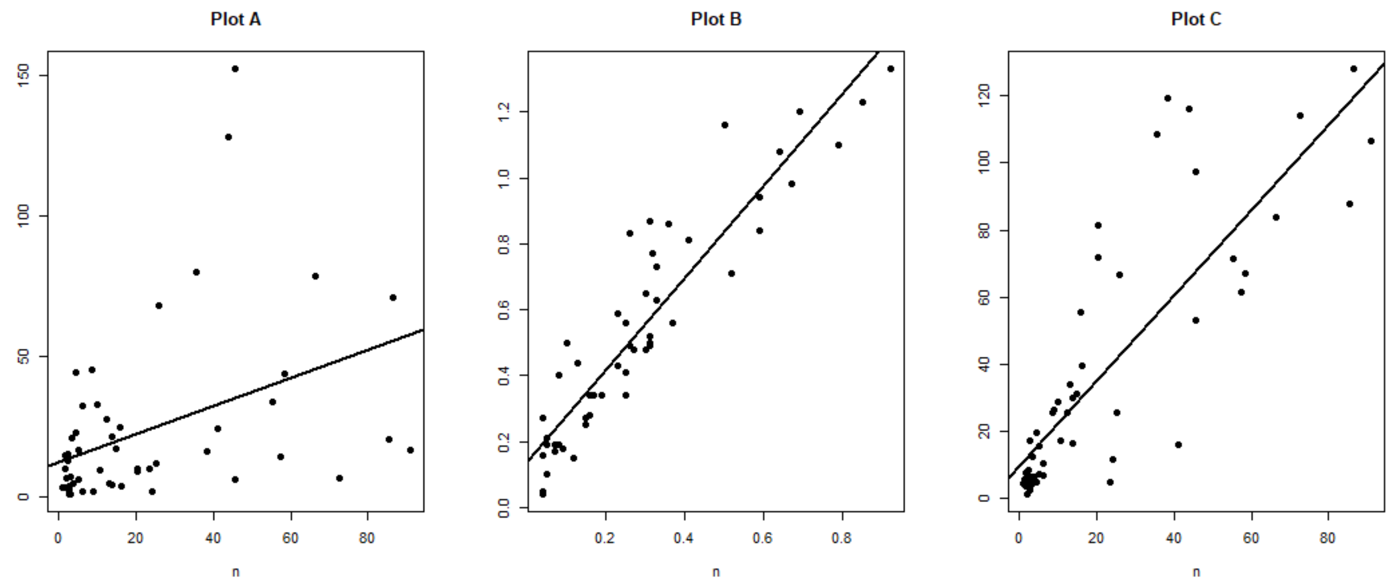
The residual standard error, s, measures the typical scatter or spread of data around the regression line.

If s is very large (i.e., of similar size to s_y), the standard deviation of the response variable, then the regression model does not help us make more accurate prediction for a particular x-value than simply guessing the mean, $\bar{y}$.

If s is very small (much smaller than s_y). then we are getting more predictive power from our model. Thus, s is one of the ways we evaluate the usefulness of the regression model.

Interpreting residual standard error

a. Can you match each value of s and s_y to their corresponding scatterplots?



Scatterplot	Residual Standard Error (average residual size) s	Standard deviation of response s_y
Plot A	$s = 21.37$	$s_y = 38.2035$
Plot C	$s = 28.38$	$s_y = 30.82632$
Plot B	$s = 0.2116$	$s_y = 0.3410$

Coefficient of Determination

Often, it is valuable to more formally compare the relationship between s and s_y.

Statisticians typically use the coefficient of determination, which is just the square of correlation.

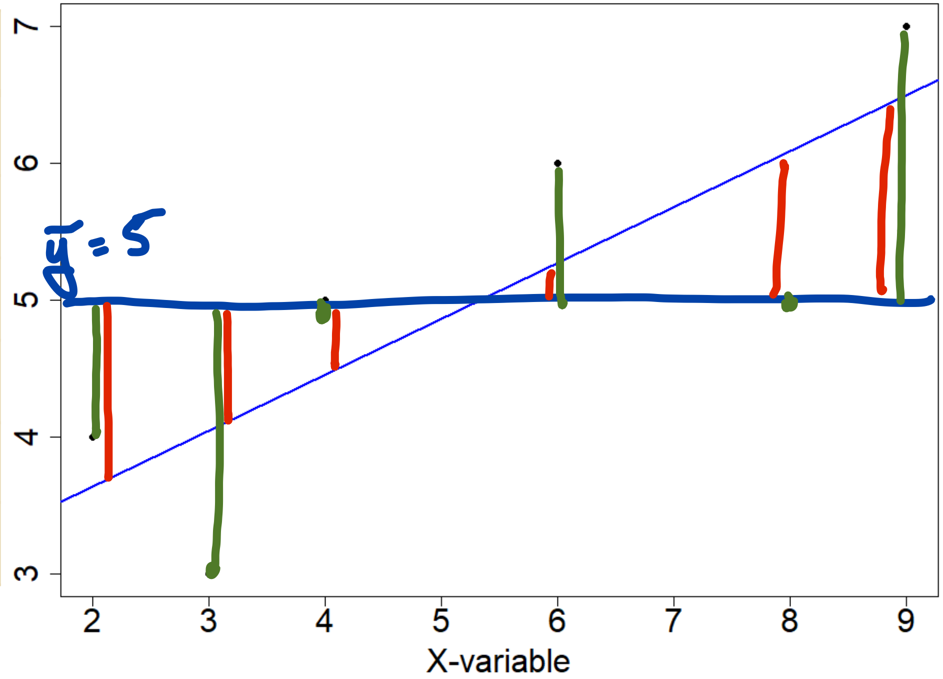
Definition: The **coefficient of determination R^2** quantifies the percent of variation in the response variable accounted for by its linear relationship with the explanatory variable.

Coefficient of Determination

Visualizing R^2: Let's visualize R^2 using a simple example. Below we plot some manufactured data long with its OLS regression.

OLS regression line: $\hat{y} = 2.8305 + 0.4068x$

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$
2	4	3.64	1	1.8496
3	3	4.05	4	0.9025
4	5	4.46	0	0.2916
6	6	5.27	1	0.0729
8	5	6.08	0	1.1664
9	7	6.49	4	2.2201



Coefficient of Determination

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$
2	4	3.64	1	1.8496
3	3	4.05	4	0.9025
4	5	4.46	0	0.2916
6	6	5.27	1	0.0729
8	5	6.08	0	1.1664
9	7	6.49	4	2.2201

a. What is the overall variability $\sum(y-\bar{y})^2$ in the response y? This value is called the **Total Sum of Squares or SST.**

Total green lengths is 10

b. The amount of variability that is explained by the relationship between the two variables is called the **Model Sum of Squares or SSM.** Use the table to calculate this, i.e., what is $\sum(\hat{y}-\bar{y})^2$?

Total red lengths is 6.508

c. What percentage of this variability does our OLS account for? The ratio of SSM/SST is the **coefficient of determination, R²**. Calculate it for this example.

$R^2 = \frac{6.508}{10} = 0.6508$

Coefficient of Determination

d. How do we interpret the coefficient of determination R^2 computed in (c) above?

Interpretation: We were able to account for

___65.08%___ of the variability in the response

variable by its ___linear relationship___ with the

sampled cases of the predictor variable.

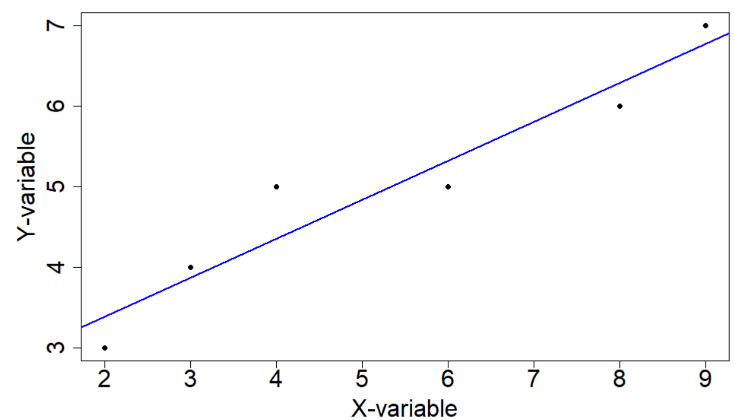
Coefficient of Determination

e. Calculate the coefficient of determination R^2 for this example. Compare how well the regression line models the data in this example to the example above.

New OLS regression line: $\hat{y} = 2.42373 + 0.48305x$

x	y	$\hat{y}$	$(y - \bar{y})^2$	$(\hat{y} - \bar{y})^2$
2	3	3.39	1	2.5921
3	4	3.87	4	1.2769
4	5	4.36	0	0.4096
6	5	5.32	1	0.1024
8	6	6.29	0	1.6641
9	7	6.77	4	3.1329

10 9.178



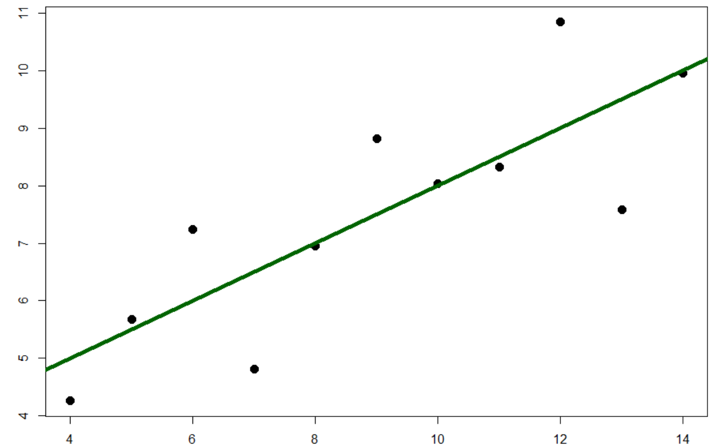
$$R^2 = \frac{9.178}{10} = 0.9178$$

A note on computation

Most lecture examples thus far asked us to compute many aspects of this regression by-hand, but it is more typical to view the results of a regression performed by computational software.

Consider the R output below, which computes the ordinary least squares (OLS) regression for our toy data set originally introduced in Lecture 5-1.

A note on computation




```
Call:
lm(formula = y ~ x)
```

b_1 b_0

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x             0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
1 ' ' 1
```

s R^2

```
Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,	Adjusted R-squared:  0.6
295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Diagnostics

Conditions for OLS to be optimal regression method:

Relationship b/t x & y is linear. The data should show a linear trend. If there is a nonlinear trend, an advanced regression method from another book or later course should be applied.

True errors are normally dist. Generally, the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points.

True errors have constant variance. The variability of points around the least squares line remains roughly constant.

Paired observations (x, y) are independent. Be cautious about applying regression to data collected sequentially in what is called a time series. Such data may have an underlying structure that should be considered in a model and analysis.

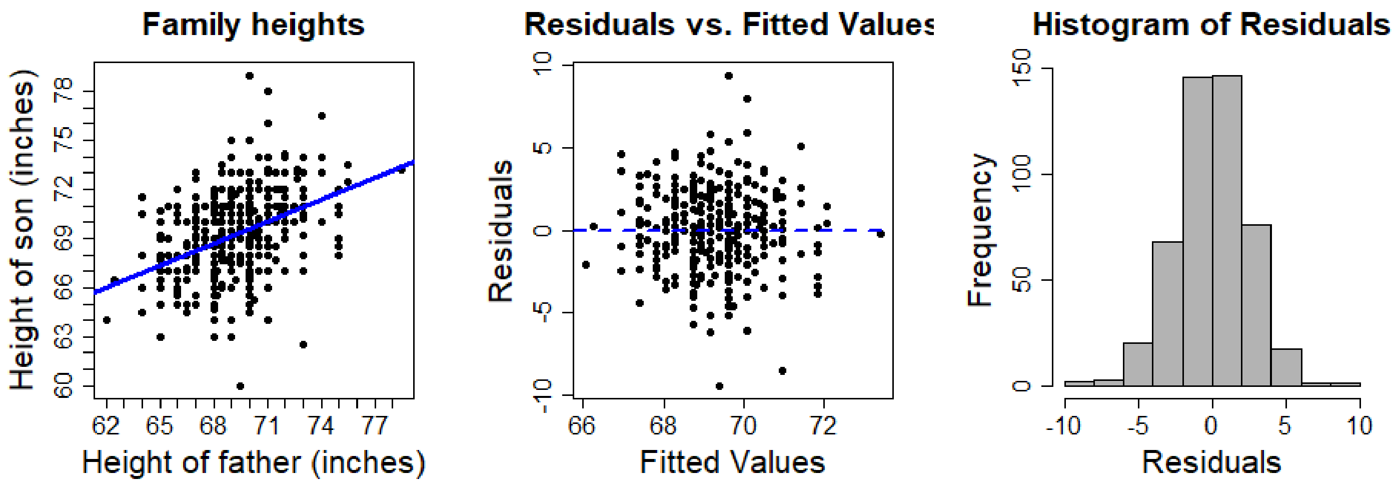
Diagnostics

In general, among the best ways to check whether it is safe to assume these conditions are met in each research scenario is by checking

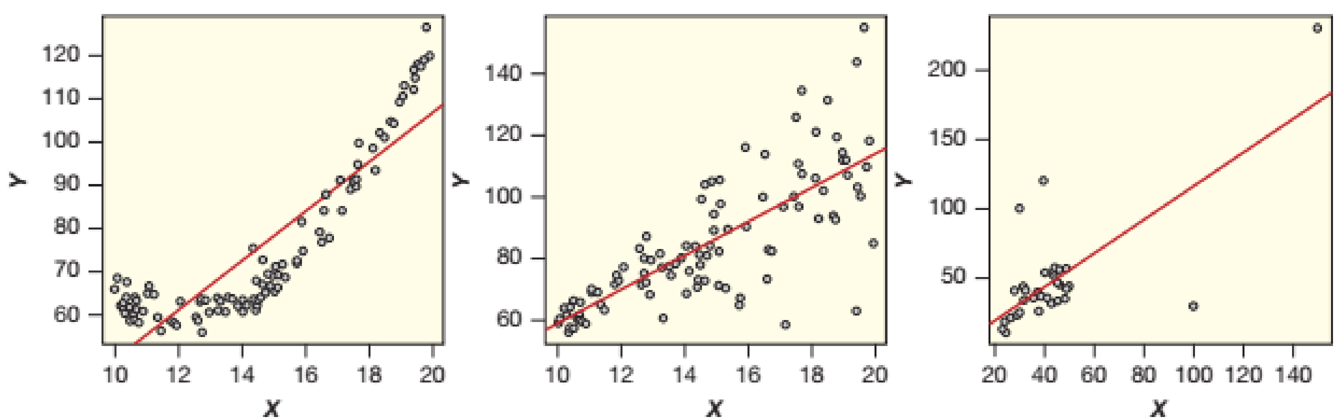
a residuals vs. fitted plot.

Diagnostics

Ideally...



Examples of violations...



Lecture 5-3: Inference for OLS

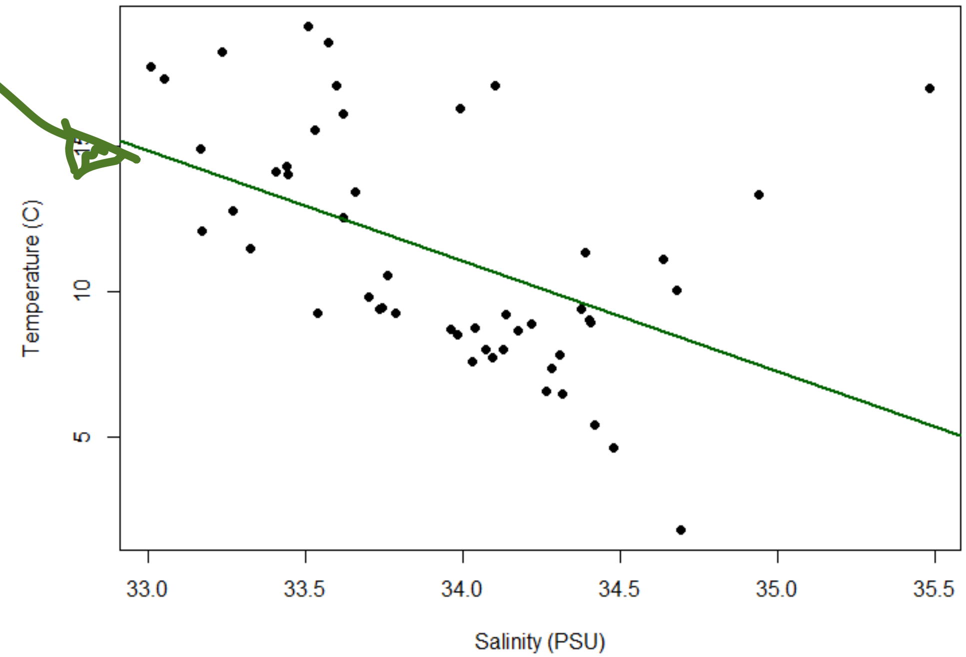
The California Cooperative Oceanic Fisheries Investigation (CalCOFI) data set represents the longest (1949-present) and most complete (more than 50,000 sampling stations) data set of oceanographic and larval fish data in the world.

It includes abundance data on the larvae of over 250 species of fish; larval length frequency data and egg abundance data on key commercial species; and oceanographic and plankton data.

The physical, chemical, and biological data collected at regular time and space intervals quickly became valuable for documenting climatic cycles in the California Current and a range of biological responses to them.

Lecture 5-3: Inference for OLS

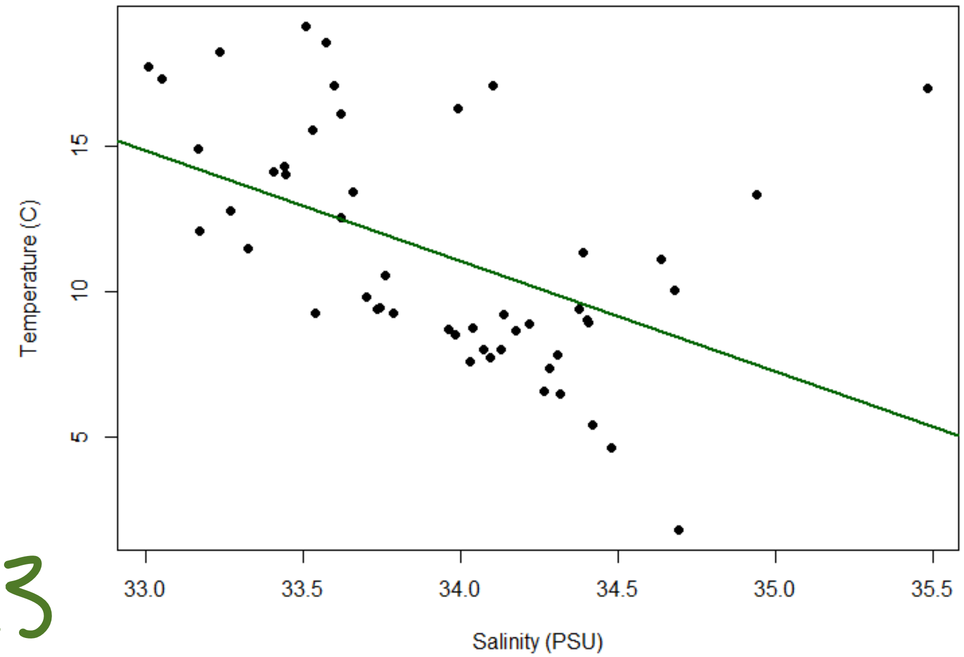
$\hat{y} = 140.4988 - 3.807x$




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 140.4988    33.6989   4.169 0.000127
df$Salnty    -3.8070     0.9928  -3.834 0.000366

Residual standard error: 3.62 on 48 degrees of
freedom
Multiple R-squared:  0.2345
```

Lecture 5-3: Inference for OLS

a. What is the correlation between ocean water salinity and ocean water temperature?



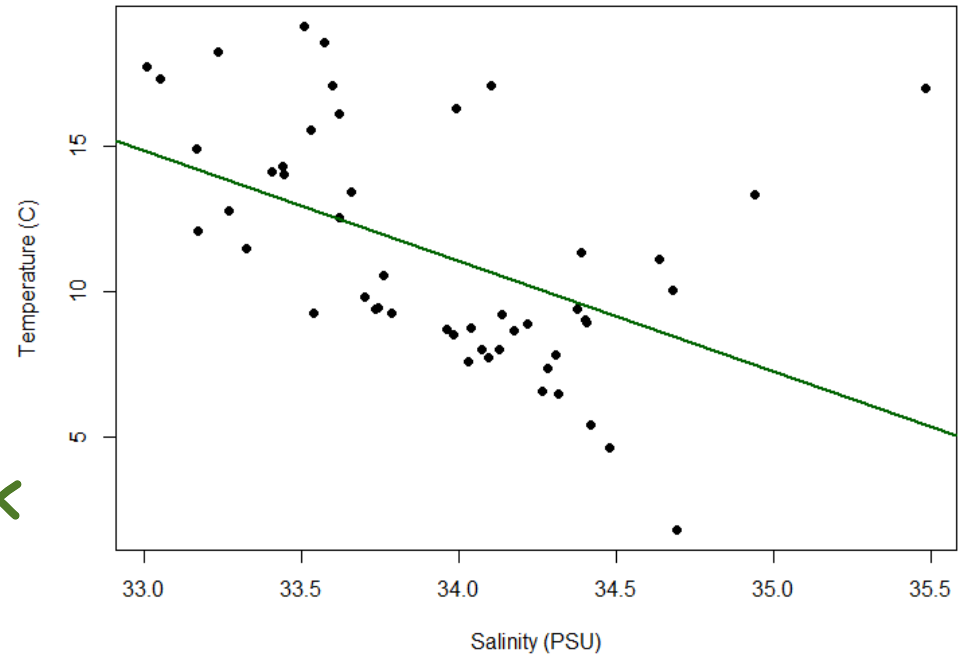
$R^2 = 0.2345$, so

$r = \sqrt{.2345} = 0.4843$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 140.4988    33.6989   4.169 0.000127
df$Salnty    -3.8070     0.9928  -3.834 0.000366

Residual standard error: 3.62 on 48 degrees of
freedom
Multiple R-squared:  0.2345
```

Lecture 5-3: Inference for OLS

b. What is the OLS equation to predict the water temperature based on its salinity?



$\hat{y} = 140.4988 - 3.807x$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 140.4988    33.6989   4.169 0.000127
df$Salnty    -3.8070     0.9928  -3.834 0.000366

Residual standard error: 3.62 on 48 degrees of
freedom
Multiple R-squared:  0.2345
```

Think about it: *What if …*

we had all possible salinity and temperature measurements for all station measurements since 1949,

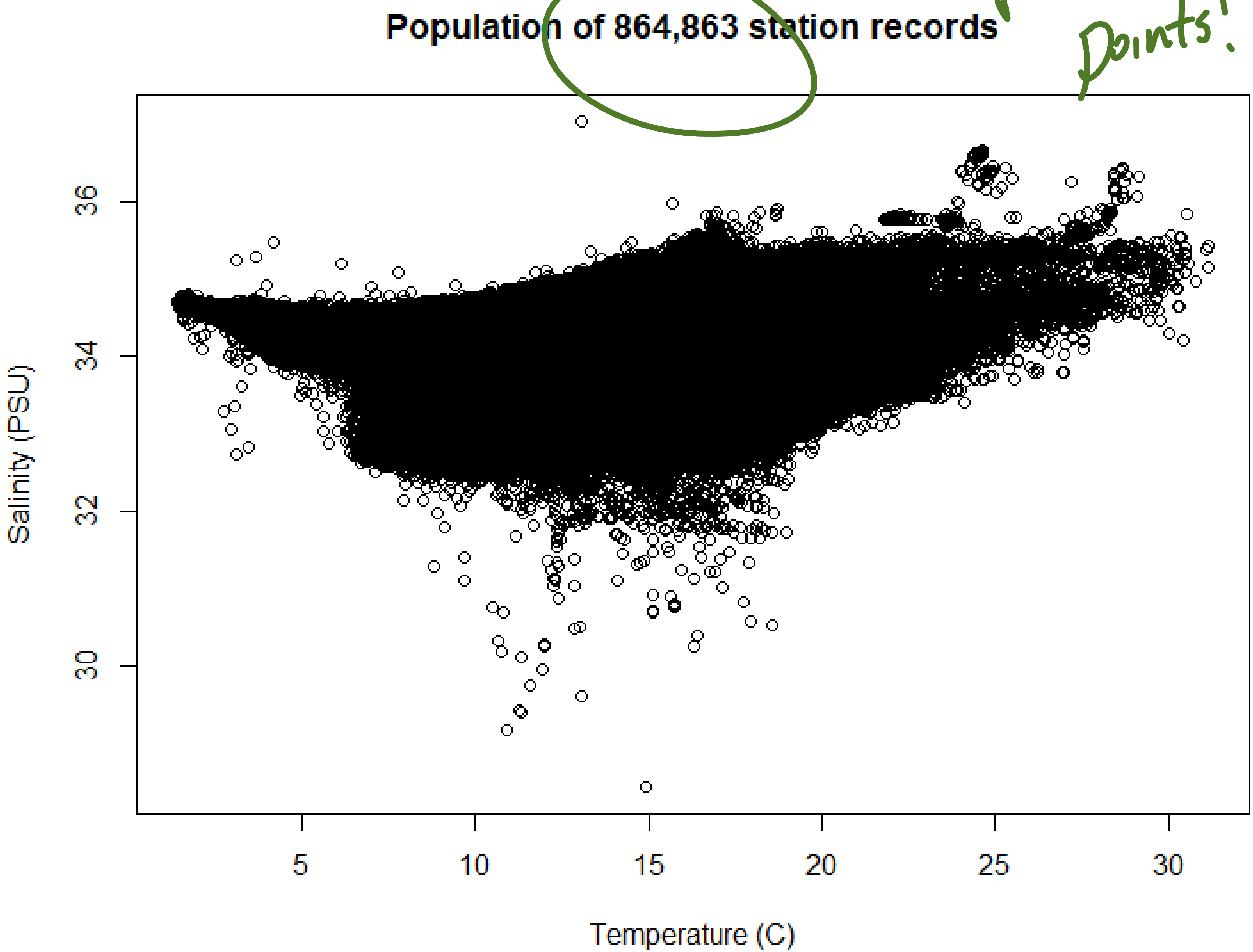
… and we added all of these points to the scatterplot,

… and then found the best fitting line for this population of all points;

… then we could think of that line as the 'true' regression line,

the regression line for the population; and we can start thinking more about inference.

So many data Points!



Inference for regression

The material covered so far focused on using the data from a **sample** to graph and describe a relationship.

The slope and y-intercept values we computed from the sample are statistics (and thus, R.V.'s); they are estimates of the corresponding true slope and true y-intercept for the underlying true relationship for the larger population.

Inference for regression

Sample level

Relationship for an individual response:

$$\hat{y} = b_0 + b_1 x + e$$

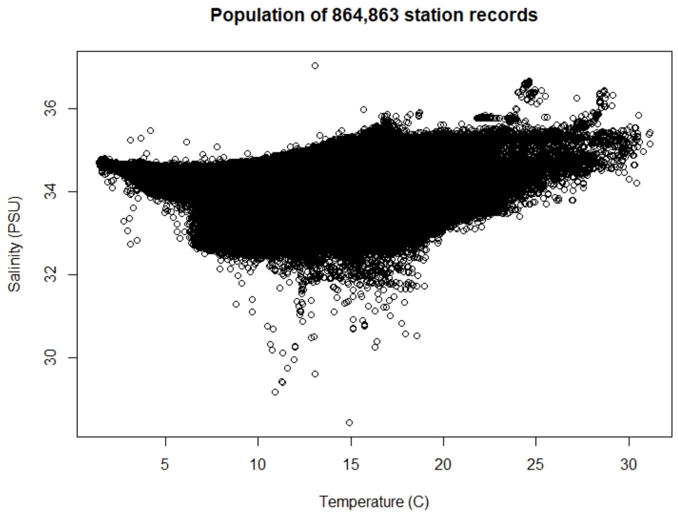
Population level

Relationship for an individual response:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Inference for regression

Consider the population of all station records. When we run a linear regression on all 846,863 observations, we get the following regression line:




```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  167.350630   0.296226   564.9   <2e-16 ***
bottle$Salnty  -4.624236   0.008753  -528.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.646 on 814245 degrees of freedom
  (50616 observations deleted due to missingness)
Multiple R-squared:  0.2553,  Adjusted R-squared:  0.2553
F-statistic: 2.791e+05 on 1 and 814245 DF,  p-value: < 2.2e-16
```

Inference for regression

d. How does our sample regression line from (b) above compare to the true population line?

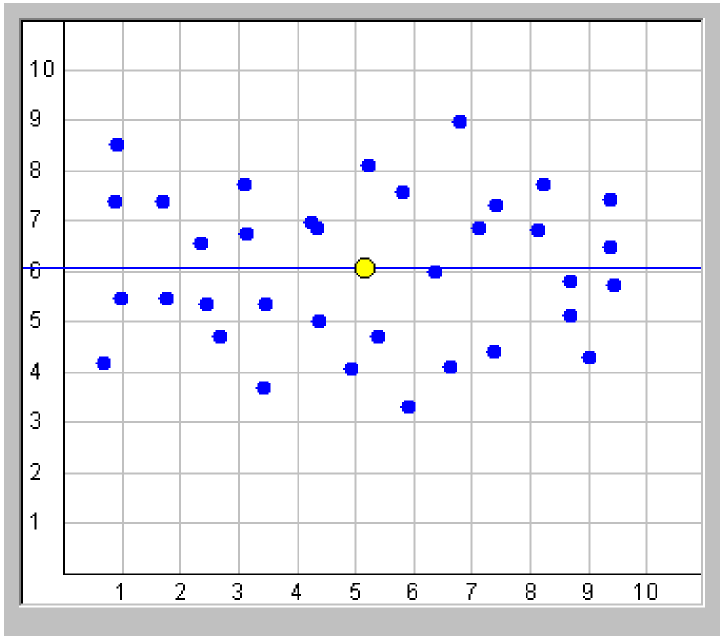
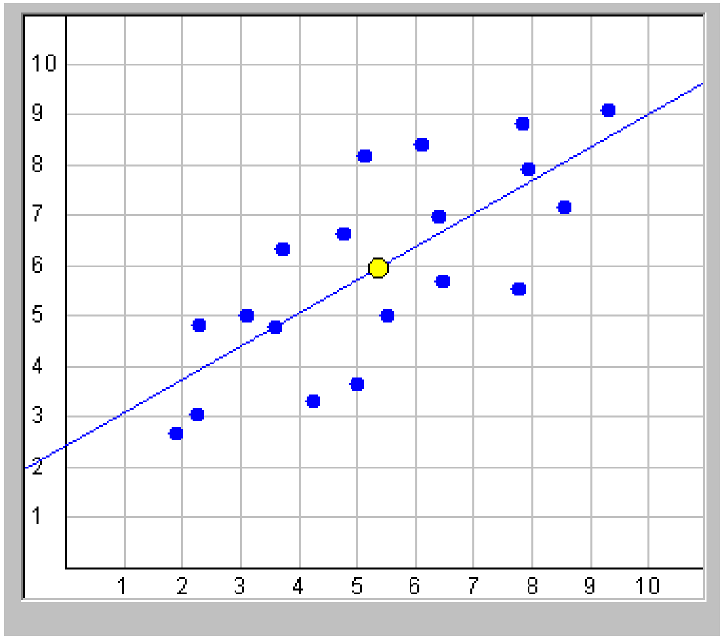
Sample Coefficients:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 140.4988    33.6989   4.169 0.000127
df$Salnty    -3.8070     0.9928  -3.834 0.000366

Residual standard error: 3.62 on 48 degrees of
Freedom Multiple R-squared:  0.2345
```

Population Coefficients:

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    167.350630   0.296226   564.9   <2e-16 ***
bottle$Salnty   -4.624236   0.008753  -528.3   <2e-16 ***

Residual standard error: 3.646 on 814245 degrees of freedom
Multiple R-squared:  0.2553
```

Is there a true relationship?

Null hypothesis $H_0: \beta_1 = 0$	Alternative hypothesis $H_A: \beta_1 \neq 0$
Meaning: The linear model has slope zero; i.e., there is NO linear relationship between x and y	Meaning: The linear model has a non-zero slope; i.e., some linear relationship exists between x and y
	

Is there a true relationship?

There are a number of ways to test this hypothesis. One way is through a t-test statistic (think about why it is a t and not a z test).

The general form for a t test statistic is:

$$t = \frac{sample\ statistic - null\ value}{standard\ error\ of\ the\ sample\ statistic}$$

Is there a true relationship?

t-test for the population slope

To test $H_0 : \beta_1 = 0$ we would use

$$t = \frac{b_1 - 0}{s.e(b_1)} \quad , \quad \text{where } s.e.(b_1) = \frac{s}{\sqrt{\sum(x-\bar{x})^2}}$$

and the degrees of freedom for the *t*-distribution are *n* – *2.*

Is there a true relationship?

Consider the regression output from our earlier sample of 50 station records.

Use it to conduct a hypothesis test of whether there is a linear relationship between the salinity and temperature of ocean water.

In other words, test $H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 140.4988    33.6989   4.169 0.000127
df$Salnty    -3.8070     0.9928  -3.834 0.000366

Residual standard error: 3.62 on 48 degrees of
freedom
Multiple R-squared:  0.2345
```

What does it mean if we have evidence against H_0?

In the t-test for the slope, evidence that the null hypothesis is not consistent with our sample result means that the idea that there is no linear relationship b/t x & y is called into doubt. That is, we have reason to believe there IS a linear relationship between the explanatory and response variables.

What does it mean if we have evidence against H_0?

Confidence Interval for the population slope β_1

$$b_1 \pm t^*[s.e.(b_1)]$$

where df = n – 2 for the t^* value

b. Compute the 95% confidence interval for the slope β_1 for the water salinity & temperature example.

invT(.025, 48)

$= -2.0106$

$b_1 \pm t^* se(b_1)$

$-3.807 \pm 2.0106(0.9928) = (-5.8031, -1.8109)$

Predicting local species diversity

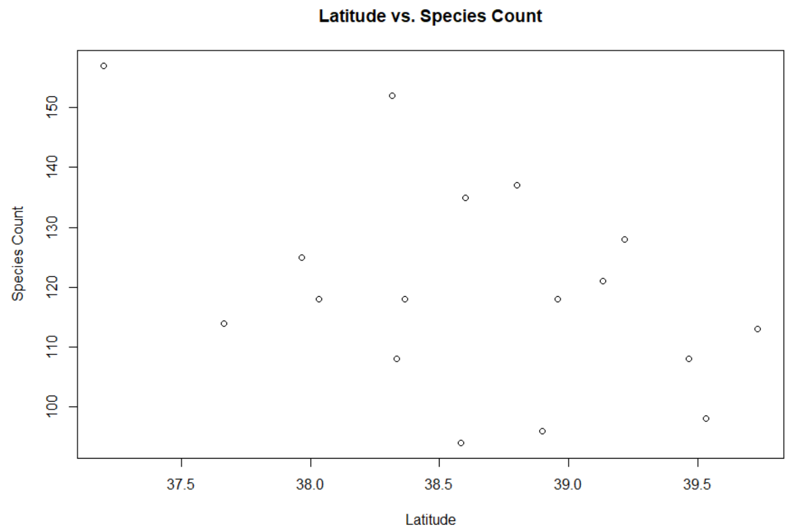
A common observation in ecology is that species diversity is higher in warmer climates than in colder ones.

To examine this association, data was sampled from random locations participating in the Audubon Society's Christmas Bird Count. During the annual Christmas Bird Count, participants attempt to count all birds in a 15-mile diameter area.

Assuming participants' records have errors at random, we can use the latitude of their Bird Count submissions as an explanatory (or predictor) variable of the number of unique species observed that day.

Predicting local species diversity

Location	Lat	Count
Bombay Hook, DE	39.22	128
Cape Henlopen, DE	38.8	137
Middletown, DE	39.47	108
Milford, DE	38.96	118
Rehoboth, DE	38.6	135
Seaford-Nanticoke, DE	38.58	94
Wilmington, DE	39.73	113
Crisfield, MD	38.03	118
Denton, MD	38.9	96
Elkton, MD	39.53	98
Lower Kent County, MD	39.13	121
Ocean City, MD	38.32	152
Salisbury, MD	38.33	108
S. Dorchester County, MD	38.37	118
Cape Charles, VA	37.2	157
Chincoteague, VA	37.97	125
Wachapreague, VA	37.67	114



Statistic	Mean	SD	Cor
x = latitude	38.6358	0.6877	-0.4623
y = No. of species observed	120	17.8851	

Predicting local species diversity

Page 132-133

```
Call:
lm(formula = SpeciesDiversity$Count ~ SpeciesDiversity$Lat

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           XXXXX     230.024   2.544   0.0225
SpeciesDiversity$Lat  XXXXX       5.953  -2.022   0.0613

Residual standard error: 16.37 on 15 degrees of freedom
Multiple R-squared:  XXXX,    Adjusted R-squared:  0.1619
F-statistic:  4.09 on 1 and 15 DF,  p-value: 0.06134
```

a. Notice that the OLS estimates for the population slope and intercept are missing from the regression output, as well as the coefficient of determination. Use the provided sample statistics on the previous page to fill in these missing terms.

Predicting local species diversity

b. The researchers who collected this study are interested in assessing whether there is a significant linear relationship between the temperature during the month of birth and the age of locomotor onset. Use the regression output to conduct the appropriate hypothesis test for this researcher question and draw a conclusion based on your findings.

Hypotheses: H_0:______________ H_a :______________

Test Statistic Value: _________ *p*-value: _____________

Conclusion:

Predicting local species diversity

c. Explain why this model might not reliably predict the number of birds one could expect to see in East Lansing, MI, which has a latitude of 42.74 degrees.