ECOLOGICAL DATA ANALYSIS IN R

11.1 Simple (bivariate) Linear Regression

Luis Malpica Cruz
lmalpica@uabc.edu.mx - @luismalpicaC

A case study

PeerJ

View 40 tweets

Related research

✔ PEER-REVIEWED

Sooty tern (*Onychoprion fuscatus*) survival, oil spills, shrimp fisheries, and hurricanes

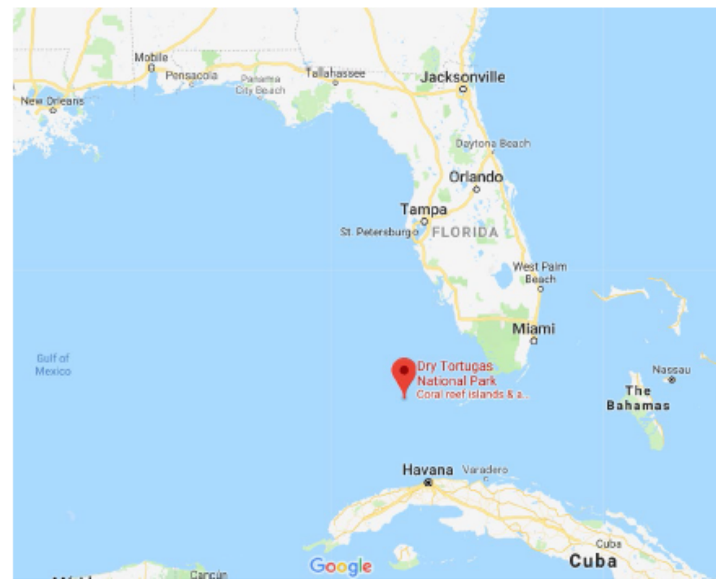
Research article | Conservation Biology | Ecology | Marine Biology | Veterinary Medicine | Zoology

Ryan M. Huang[1], Oron L. Bass Jr[2], Stuart L. Pimm[1]

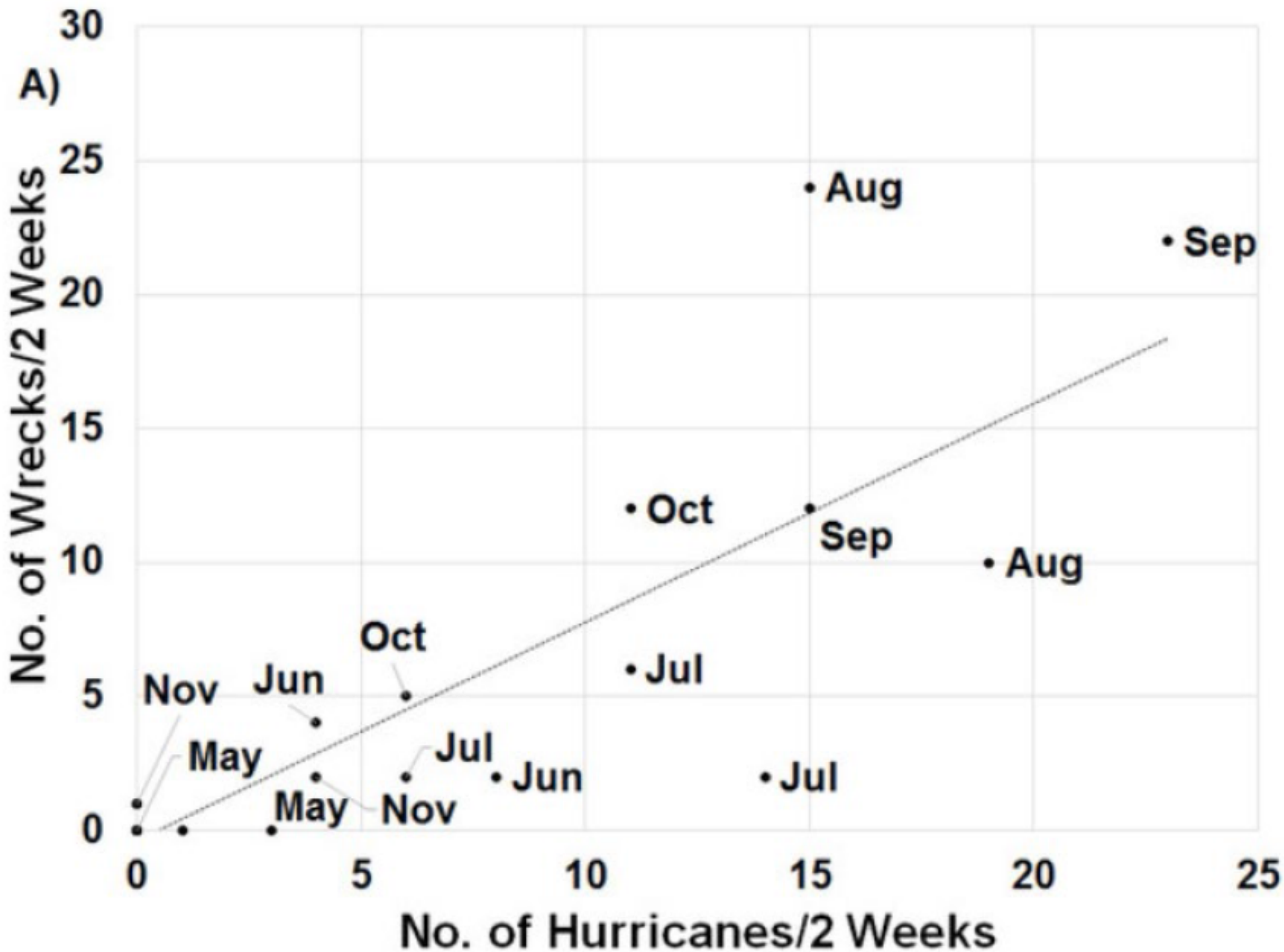
Published May 10, 2017 PubMed 28503374

https://peerj.com/articles/3287/

We combined telemetry data on *Onychoprion fuscatus* (sooty terns) with a long-term capture-mark-recapture dataset from the Dry Tortugas National Park to map the movements at sea for this species, calculate estimates of mortality, and investigate the impact of hurricanes on a migratory seabird. … Indices of hurricane strength and occurrence are positively correlated with annual mortality and indices of numbers of wrecked birds.



Which Q is this graph answering?

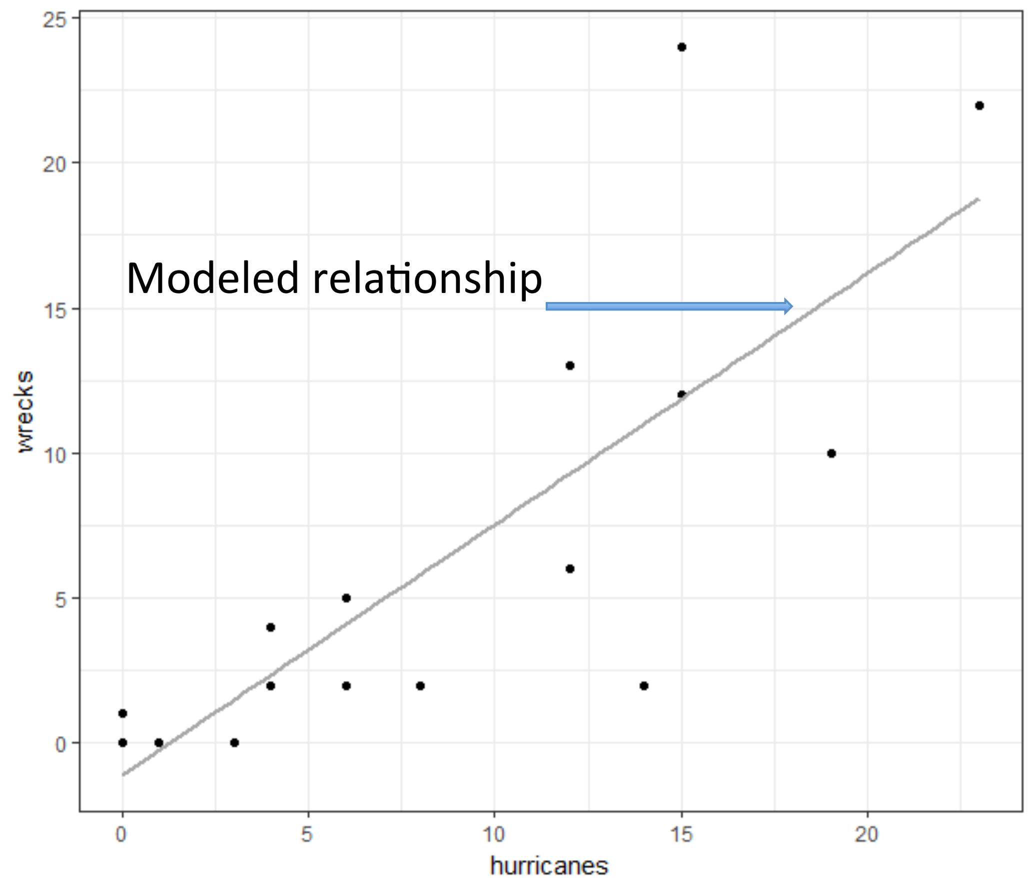


What is the relationship between hurricanes and wrecked birds?

How is this made?

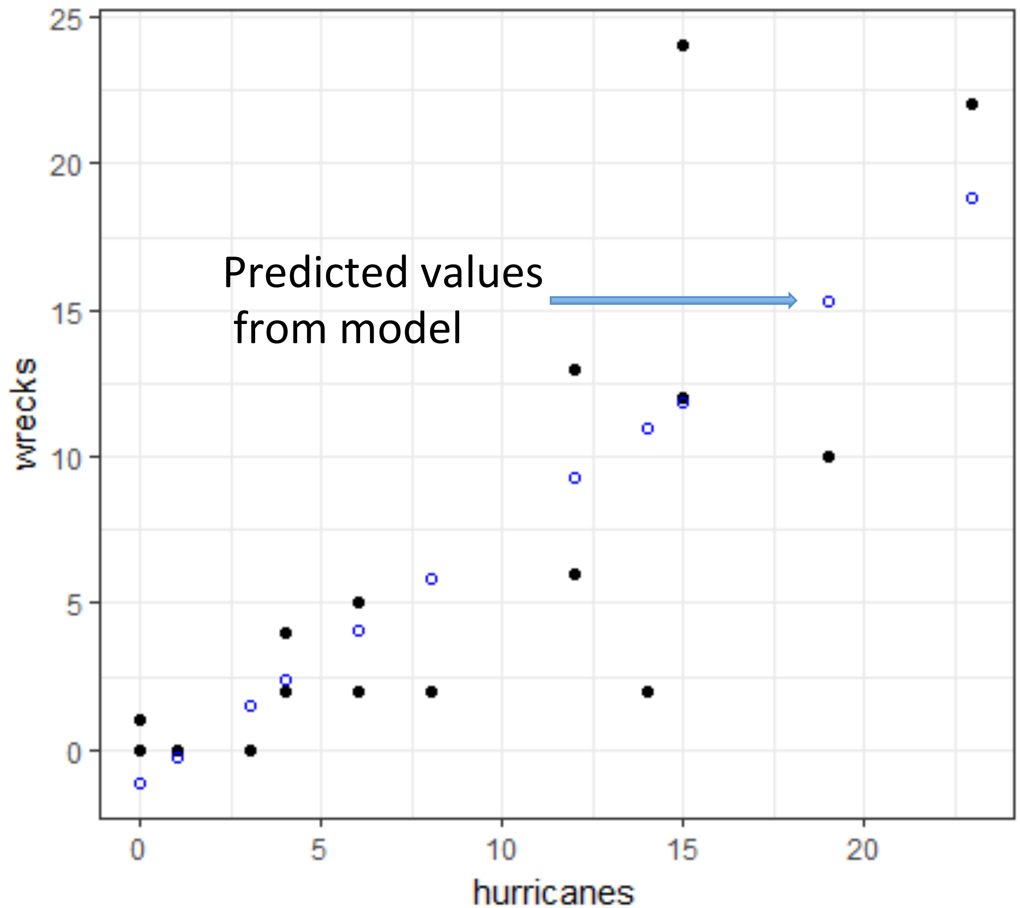
A line ("of best fit") is drawn through the points (data) minimizing the distance between the line and each point

This produces a modelled relationship between X and Y



How is this made?

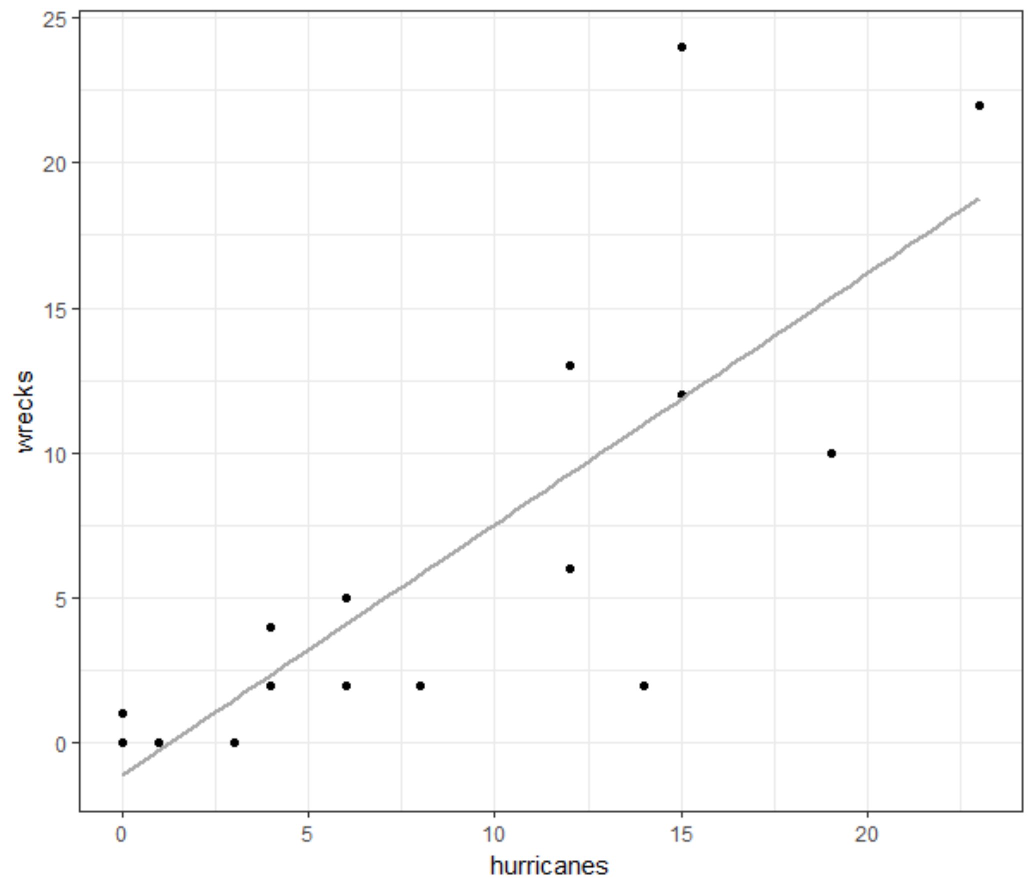
A ‘predicted Y value’ at each X value is calculated and connected with a line



How is this called?

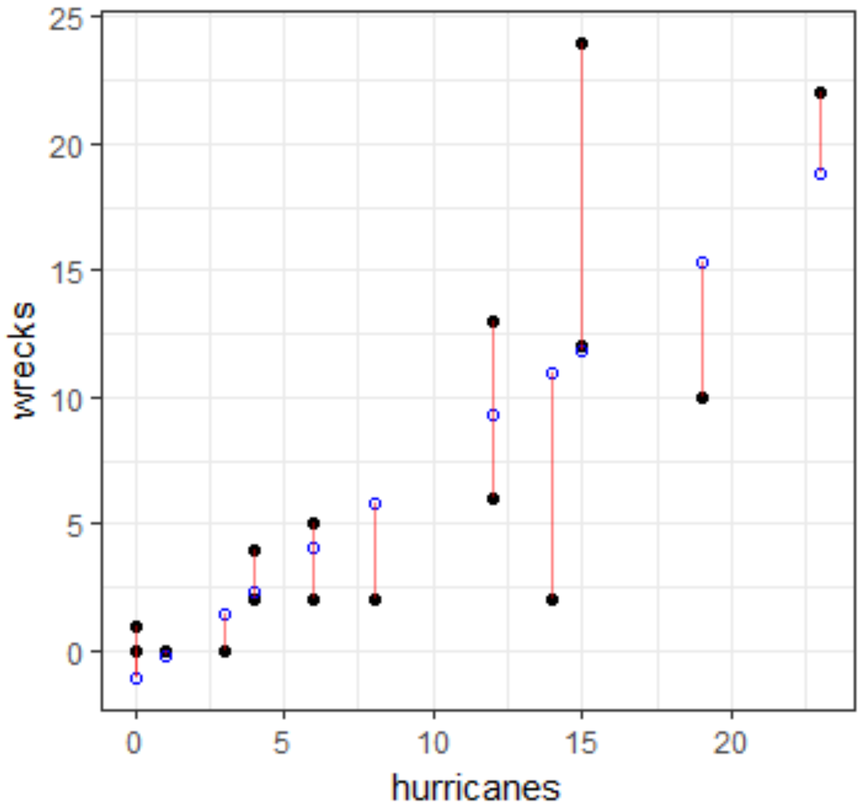
This is a model, a mathematical description of a real-life relationship

“All models are wrong, but some are useful”
- George Box, Statistician



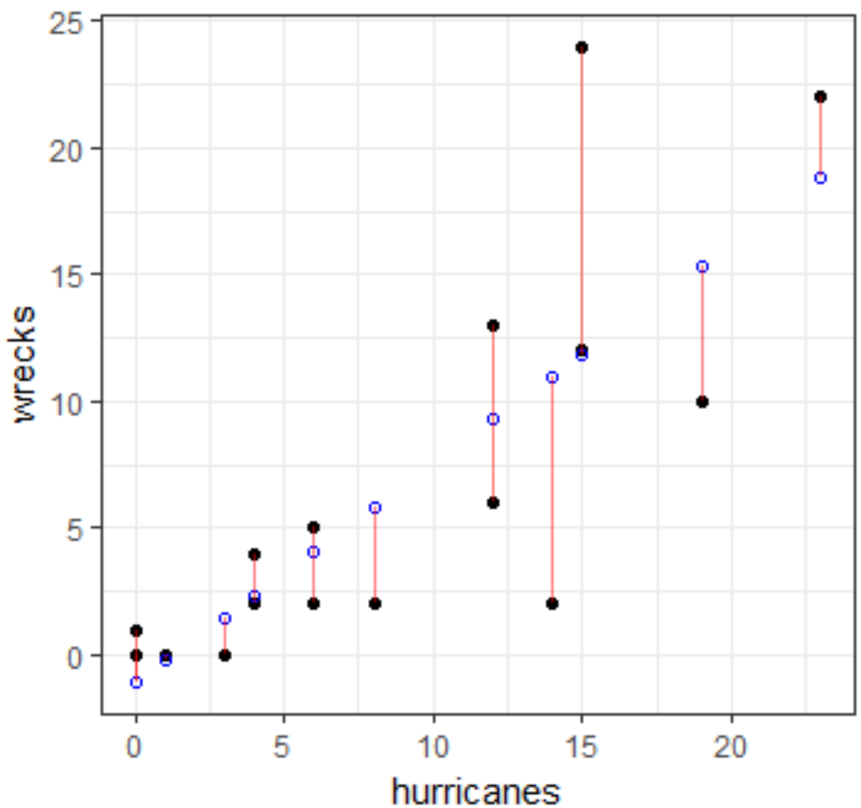
Important terms: Residuals

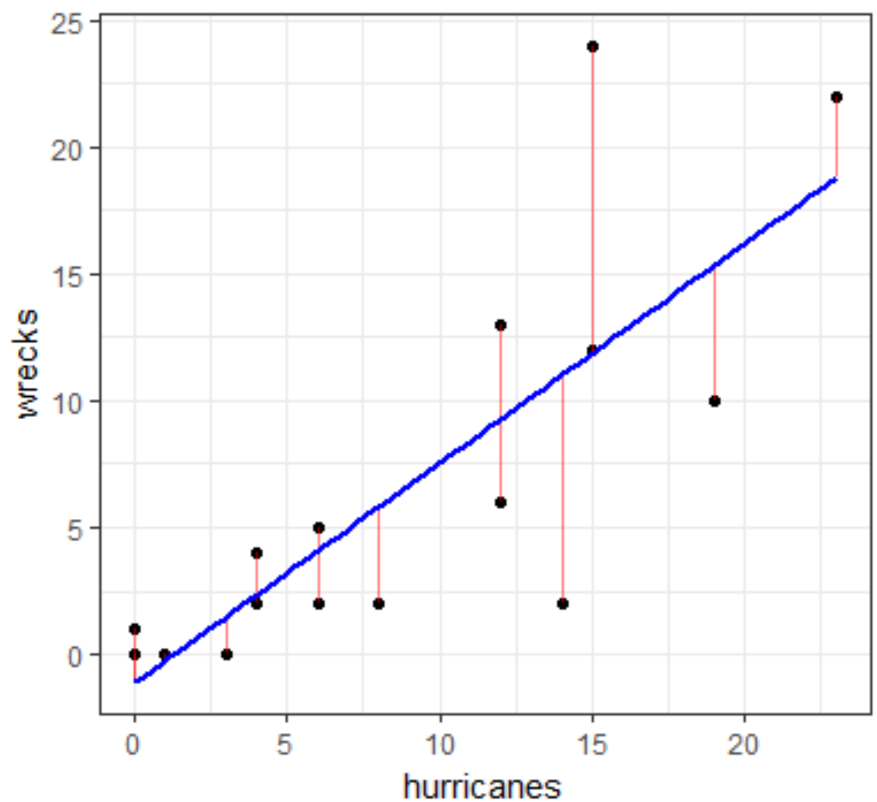
Distance between the observed values and predicted values



Important terms: Residuals

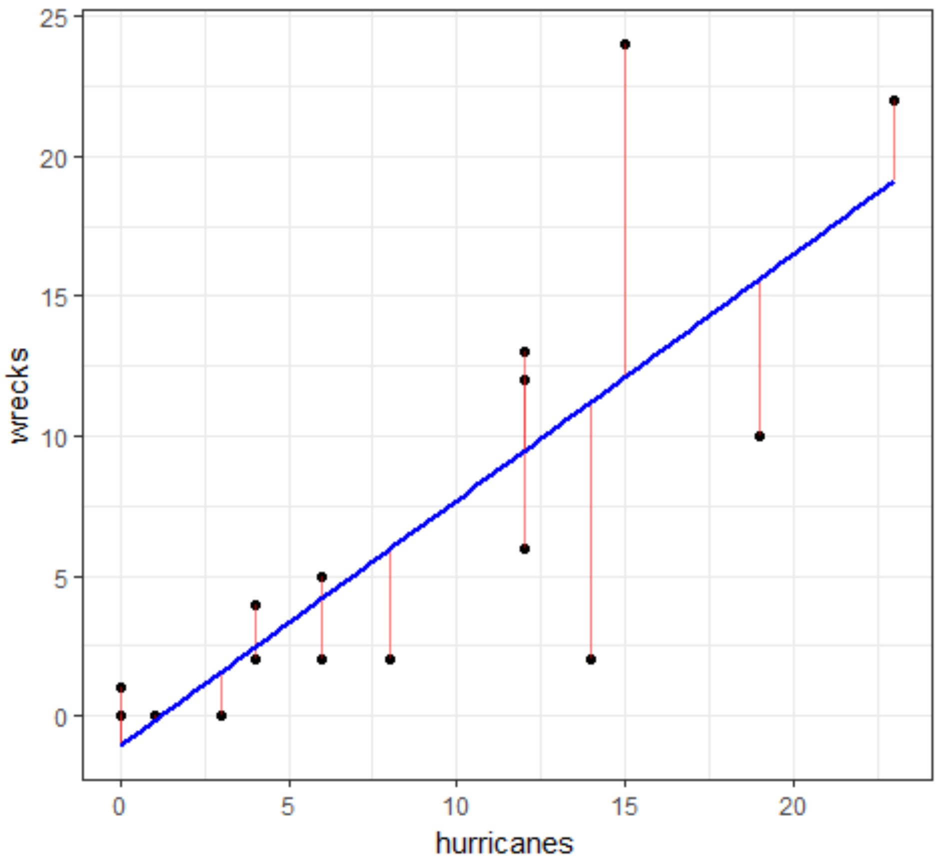
Distance between the observed values and predicted values

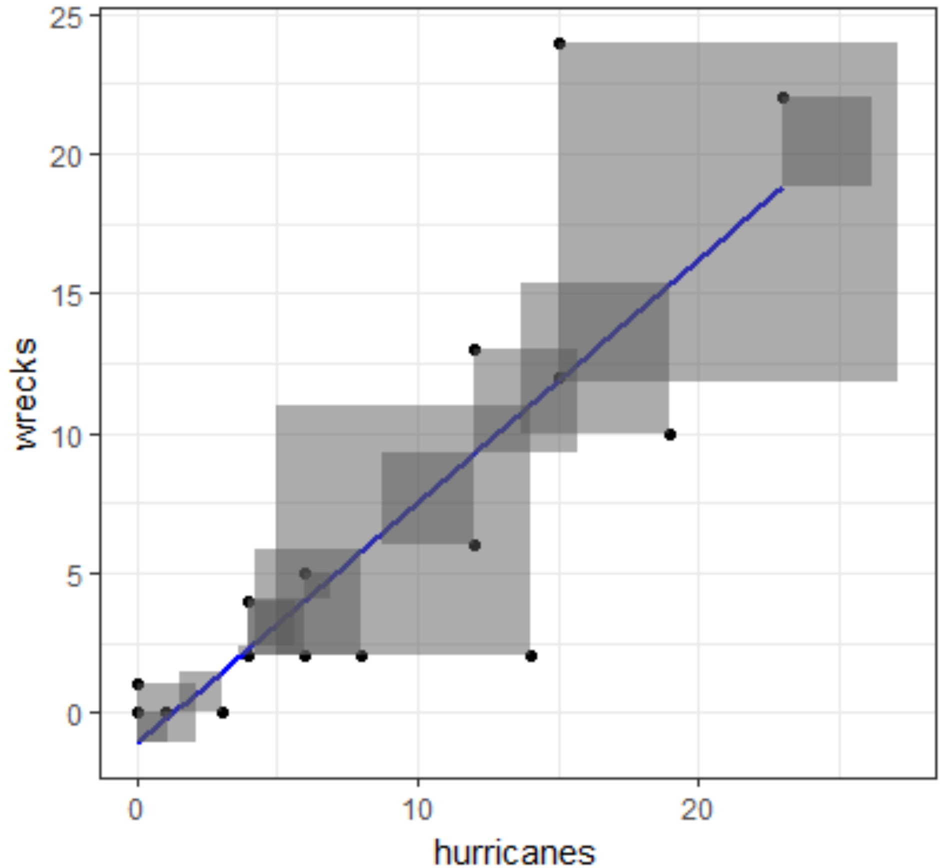




The line is fit so as to minimize residuals

Technically, to minimize the *Residual Sum of Squares* (RSS or SSR)



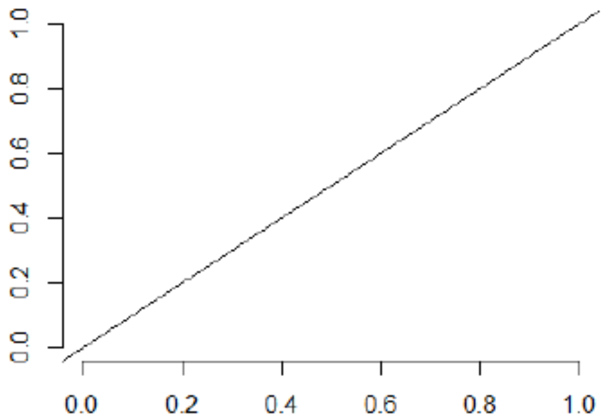


This demonstrates why outliers are so impactful!

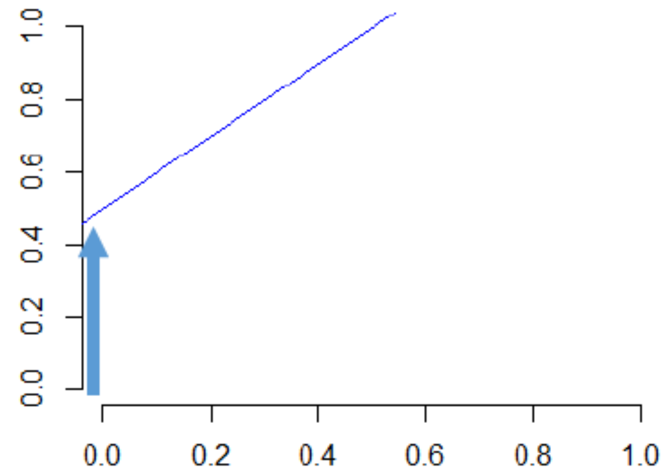
Recall, how to graph a line?

$Y = mX + B$

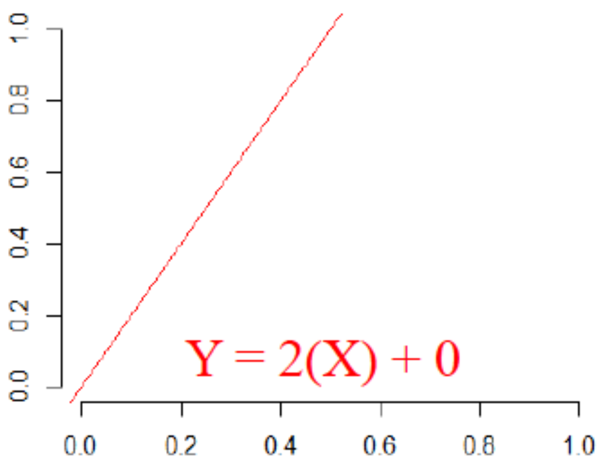
$Y = Y$
m = slope
$X = X$
B = intercept

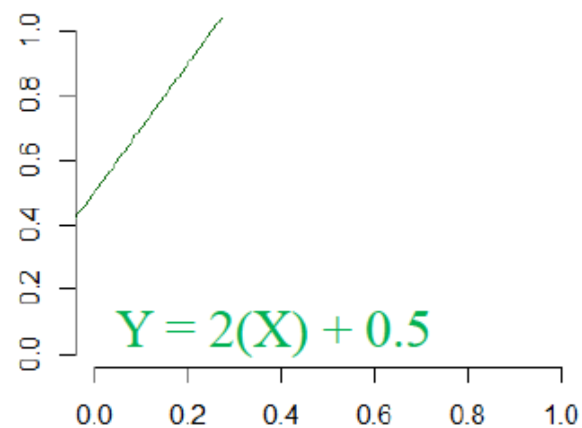


$Y = 1(X) + 0$



$Y = 1(X) + 0.5$





Simple Linear regression model

AKA "Bivariate linear regression"

$Y_i = \beta_o + \beta X_i + \varepsilon_i$

Y_i= Response
X_i= Explanatory variable
β_o= Intercept (sometimes denoted α)
β= Population slope
ε= residual error –information not explained by model

$\varepsilon_i \sim N(0, \sigma_i^2)$

Error is normally distributed with mean of zero, variance σ_i^2

Same form as:
Y = mX + B

Example of a simple linear regression model
(Continuous or predictive (X) variable)

Spoken language:
tern wrecks = intercept + hurricanes + error

Model (math) language:
$wrecks_i = \beta_o + \beta_1 hurricanes_i + error_i$

R language:
```
lm(wrecks ~ hurricanes,
    data = terns)
```

Example of a simple linear regression model
(Continuous or predictive (X) variable)

Spoken language:
tern wrecks = intercept + hurricanes + error

Model (math) language:
$wrecks_i = \beta_o + \beta_1 hurricanes_i + error_i$

R language:
```
lm(wrecks ~ hurricanes,
    data = terns)
```

Let's run a simple linear model in R...

Model output interpretation

```
fit <- lm(wrecks ~ hurricanes, data=terns)
```

```
Call:
lm(formula = wrecks ~ hurricanes, data = terns)

Residuals:
   Min     1Q Median     3Q    Max 
-8.999 -2.372  0.195  1.773 12.135 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.1205     1.9699  -0.569 0.578500    
hurricanes    0.8657     0.1761   4.916 0.000228 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.796 on 14 degrees of freedom
Multiple R-squared:  0.6332,	Adjusted R-squared:  0.607 
F-statistic: 24.16 on 1 and 14 DF,  p-value: 0.0002275
```

Model fit, & coefficient magnitude Vs. significance

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.1205     1.9699  -0.569 0.578500
hurricanes     0.8657     0.1761   4.916 0.000228 ***
```

Assess magnitude of effect on Y

Assess significance

Assess model fit

```
Residual standard error: 4.796 on 14 degrees of freedom
Multiple R-squared:  0.6332,    Adjusted R-squared:  0.607
F-statistic: 24.16 on 1 and 14 DF,  p-value: 0.0002275
```

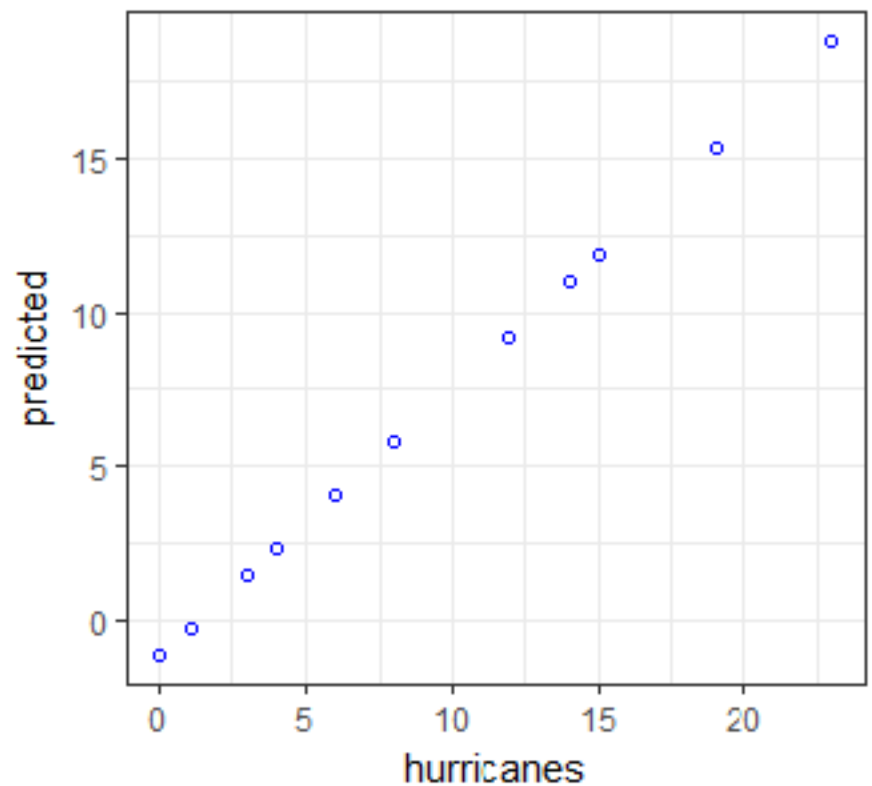
Model coefficient magnitude

```
Coefficients:
             Estimate Std. Error
(Intercept)   -1.1205     1.9699
hurricanes     0.8657     0.1761
```

$$Y = \beta_o + \beta_1 X_1 + \text{error}$$

wrecks = -1.12 + 0.87 * hurricanes

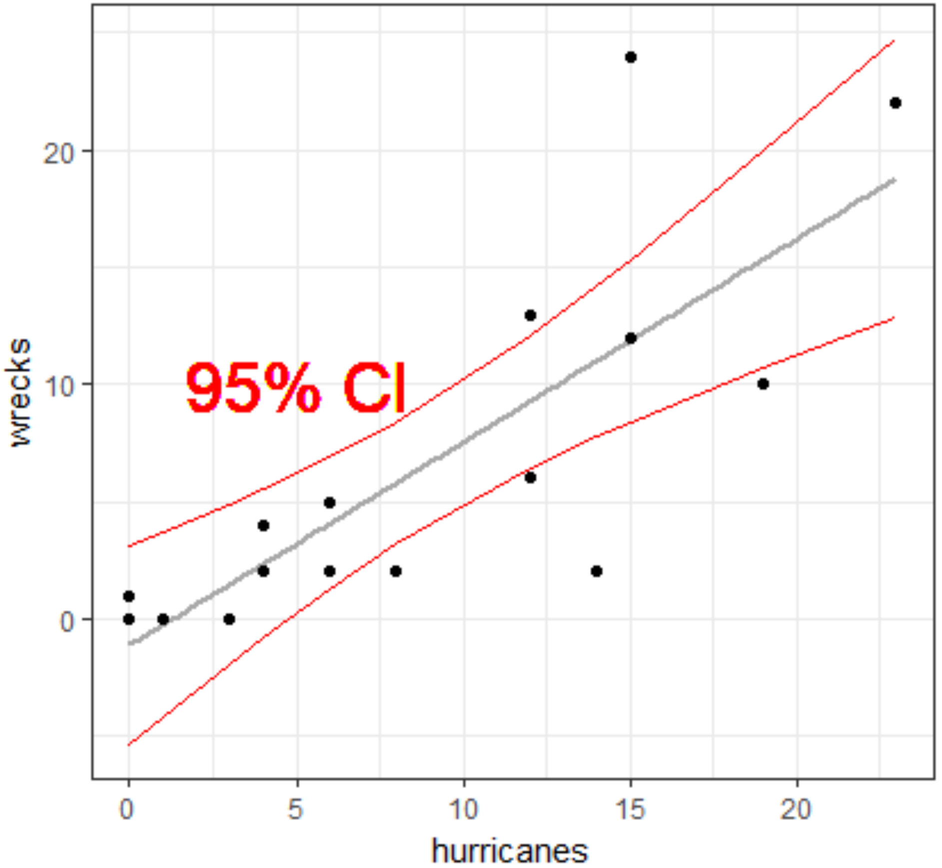
"For every additional hurricane, there were 0.87 more wrecks"



95% Confidence Interval of coefficient

If we repeated this study an infinite number of times, our interval would encapsulate the population mean at that X value 95% of the time

NOT: "95% of values fall within these bands"

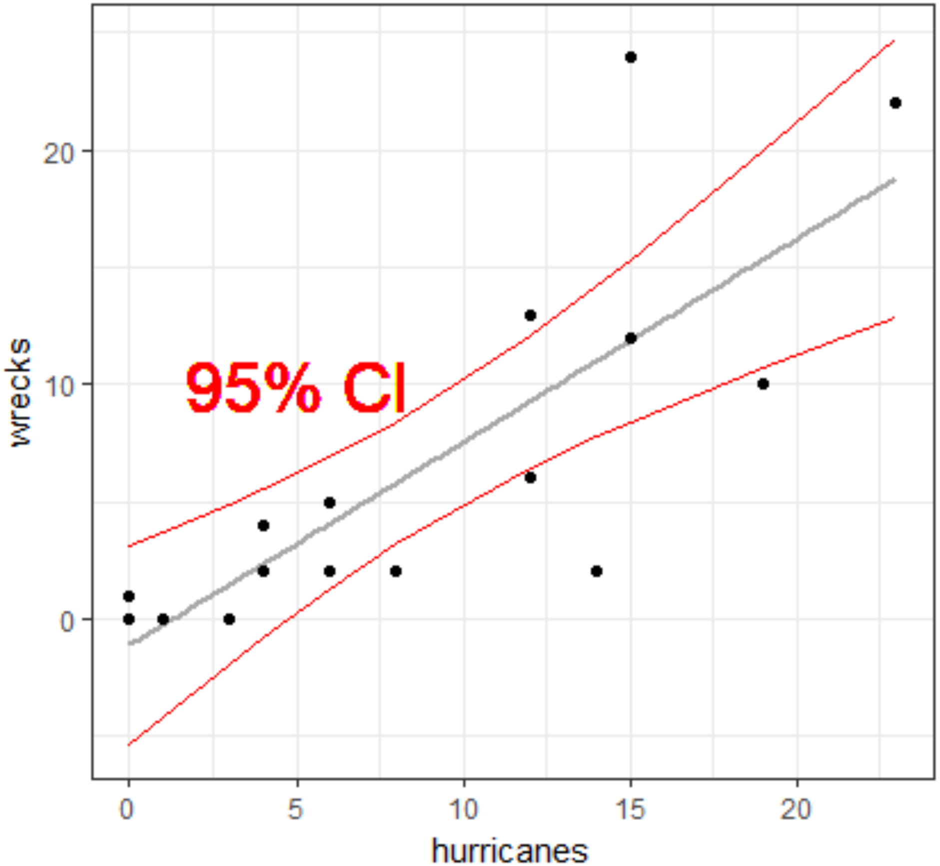


95% CI of coefficient interpretation

```
Coefficients:
             Estimate Std. Error
(Intercept)  -1.1205     1.9699
hurricanes    0.8657     0.1761
```

Here, 95% C.I. of β_1 = 0.8657 $\pm$ 1.96 * 0.1761

= 0.52 to 1.21



"For every additional hurricane, there were between 0.52 and 1.21 additional wrecked birds (95% C.I. = 0.87)"

What if the 95% C.I. of a coefficient (β) spans zero?

$Y = \beta_o + \beta_1 * X + error$
$Y = \beta_o + 0 * X + error$
$Y = \beta_o + error$

Definition of 95% CI: If you did this experiment an infinite number of times, the population Beta would be encapsulated by the interval 95% of the time

In other words, we can't rule out that the 'true' beta value is zero

In OTHER words…
This variable has no statistically significant effect on Y

Model coefficient significance

What's this?

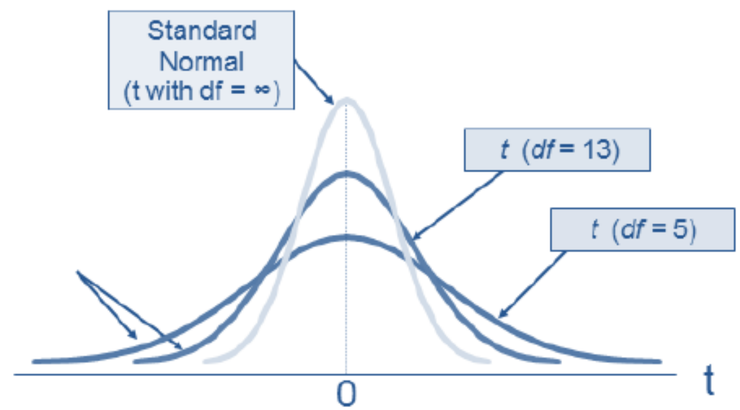
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.1205     1.9699  -0.569 0.578500
hurricanes     0.8657     0.1761   4.916 0.000228 ***
```

"How many standard deviations is our coefficient away from zero?"

This is compared against the T-distribution

Larger absolute value of T →
Further to the tails on the distribution ->
Lower P value

P < 0.05 = "Reject hypothesis that this parameter's value is zero"



https://financetrain.com/students-t-distribution/

Model fit, residual standard error

```
Residual standard error: 4.796 on 14 degrees of freedom
Multiple R-squared:  0.6332,	Adjusted R-squared:  0.607 
F-statistic: 24.16 on 1 and 14 DF,  p-value: 0.0002275
```

Average deviation of observed values from regression line

- When R fits a linear regression model, the sum of all residuals adds to zero

  (intuitively: because there should be as many points “above” line as below, at roughly same overall distance)

- Therefore, calculate ‘spread’ of residuals by the following formula:

```
> sqrt(sum(residuals(fit)^2) / df.residual(fit))
[1] 4.796254
```

14 degrees of freedom (df) because…
16 data points
2 coefficients (intercept and slope)
So… 16 – 2 = 14

Model fit, R-squared (r^2)

```
Residual standard error: 4.796 on 14 degrees of freedom
Multiple R-squared:  0.6332,     Adjusted R-squared:  0.607
F-statistic: 24.16 on 1 and 14 DF,  p-value: 0.0002275
```

% of variance in Y explainable by X

1 = perfect explanatory power (never happens)
0 = no explanatory power

As above, but penalizes model
for having additional parameters

Model fit, F-stat & p-value

```
Residual standard error: 4.796 on 14 degrees of freedom
Multiple R-squared:  0.6332,    Adjusted R-squared:  0.607
F-statistic: 24.16 on 1 and 14 DF,  p-value: 0.0002275
```

$$\frac{\text{Variance explained by model}}{\text{Unexplained variance}}$$

Bigger F-stat means stronger evidence to reject null hypothesis

```
> anova(fit)
Analysis of Variance Table

Response: wrecks
           Df Sum Sq Mean Sq F value    Pr(>F)
hurricanes  1 555.88  555.88  24.165 0.0002275 ***
Residuals  14 322.06   23.00
```

Note: If you have large sample sizes, even an F-ratio of just over 1 may be significant

Simple Linear regression model assumptions

$Y_i = \beta_o + \beta X_i + \varepsilon_i$

Y_i= Response
X_i= Explanatory variable
β_o= Intercept
β= Population slope
ε= residual error –information not explained by model

$\varepsilon_i \sim N(0, \sigma_i^2)$

Assumptions:
Assume error (i.e. residuals) is/are normally distributed with mean of zero, variance σ_i^2

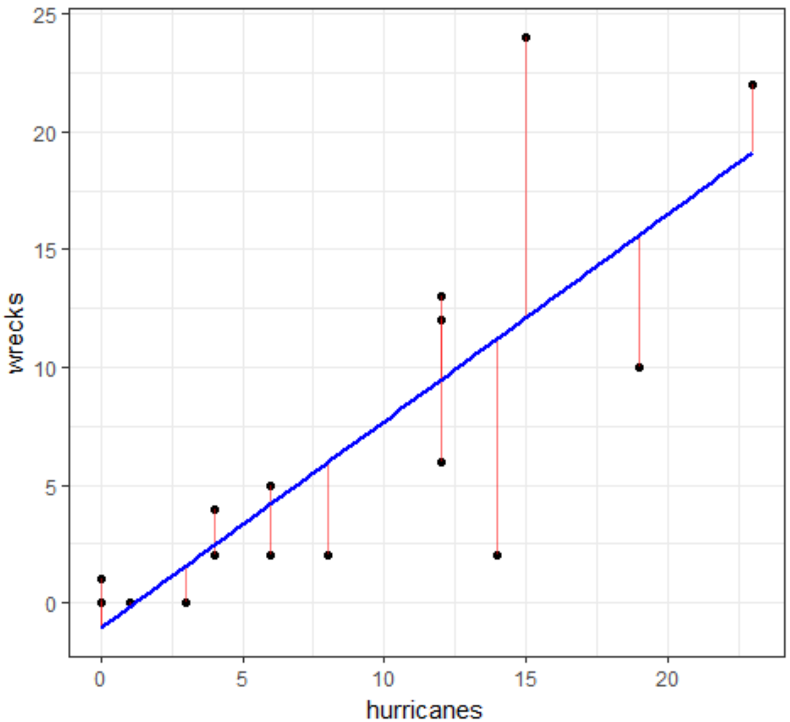
Assume σ_i^2 is equal across the entire range of data

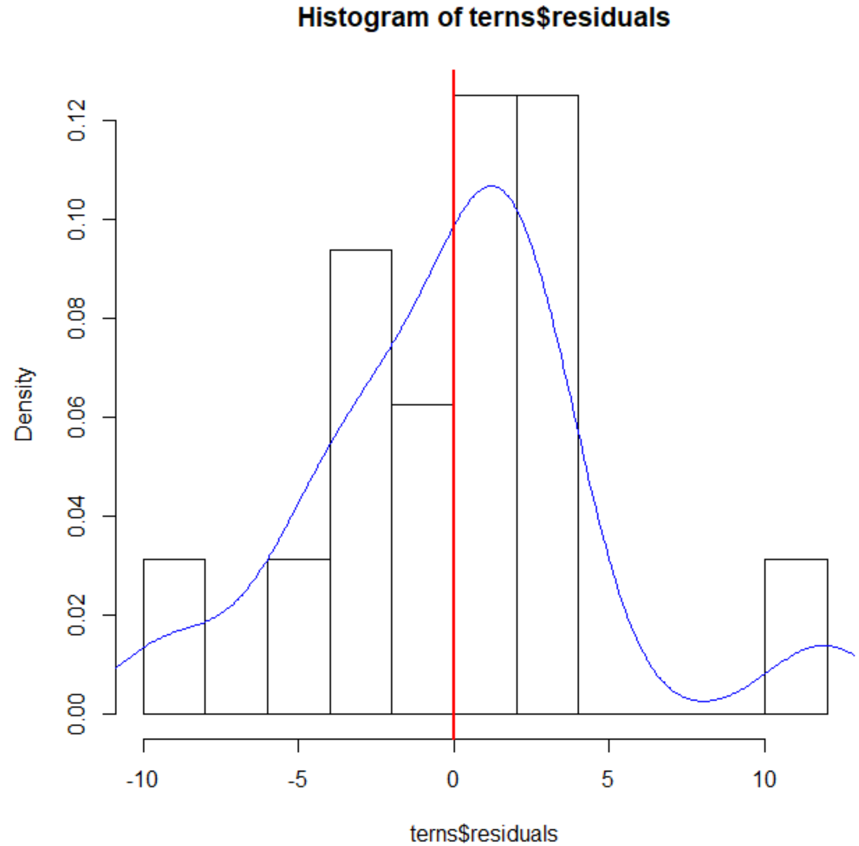
Assume replicates are truly independent.
Y values at a given X should not influence Y values at other X positions

Assume fixed X

Checking model assumptions, normality

$\varepsilon_i \sim N(0, \sigma_i^2)$ ➔ Do our data meet this?





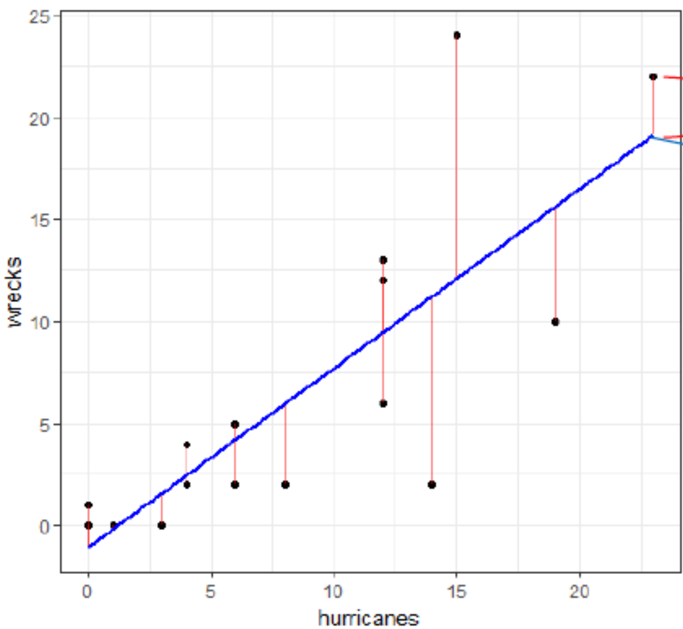
Yes-ish

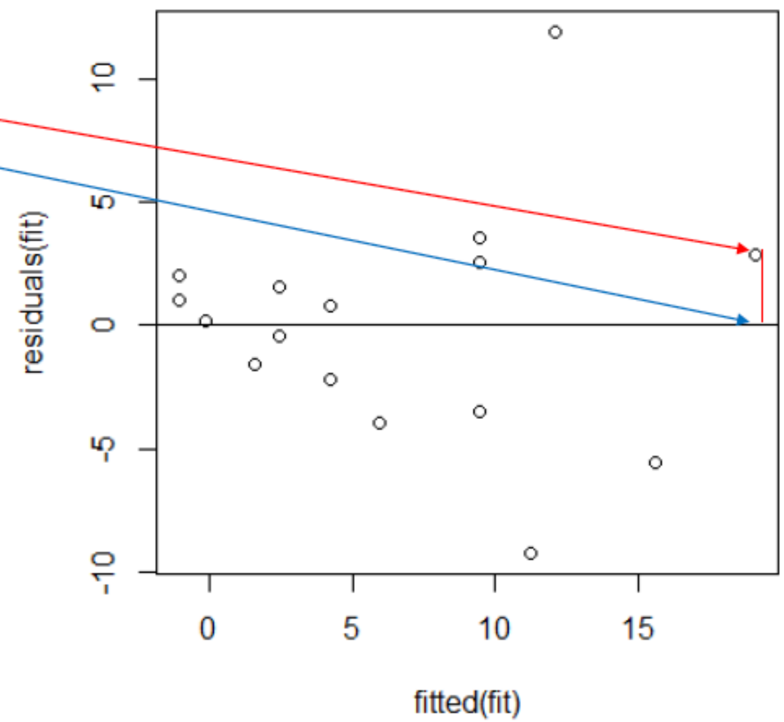
Checking model assumptions, homoscedasticity

Q: Is σ_i^2 equal across the entire range of data?

Plot residuals vs. fitted values

A: No, we see bigger residuals at higher values





Sign of trouble. Need to:

- Transform
- Allow for different variance in Y across X (GLS)
- Allow for different underlying distribution (GLM) ←later

Checking model assumptions, independence

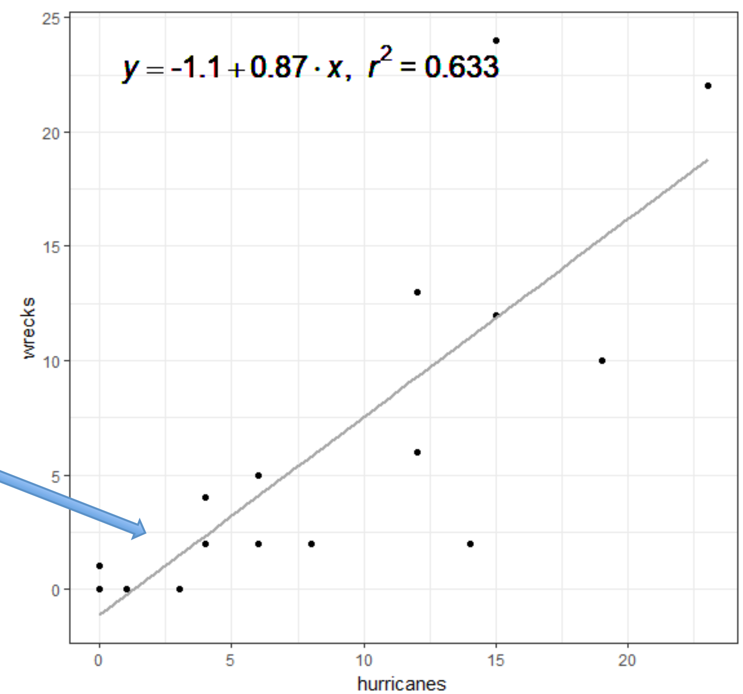
Assume replicates are truly independent.
Y values at a given X should not influence Y values at other X positions

Dependence can be due to study design

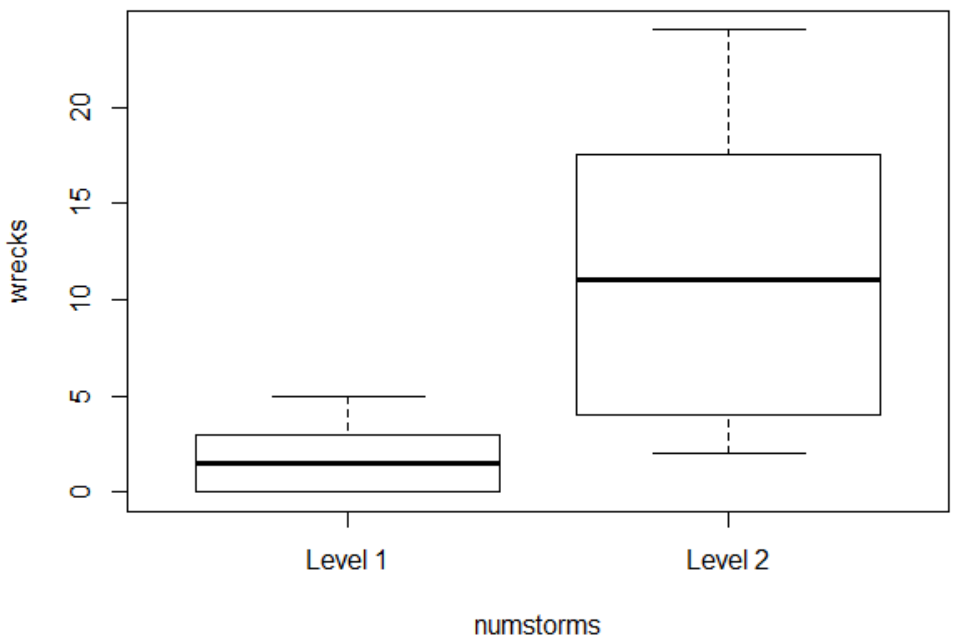
Dependence can also be due to poor model fit

At low X values, Y's are more similar than they are at high X values

Dependence due to model misfit



Categorical predictive (X) variable (2 levels)



wrecks	hurricanes	numstorms
0	0	Level 1
1	0	Level 1
0	1	Level 1
0	3	Level 1
2	4	Level 1
4	4	Level 1
2	6	Level 1
5	6	Level 1
2	8	Level 2
6	12	Level 2
13	12	Level 2
2	14	Level 2
12	15	Level 2
24	15	Level 2
10	19	Level 2
22	23	Level 2

R language:

```
lm(wrecks ~ numstorms, data = terns)
```

If two levels in X, model output is like t-test

```
> categorical_lm1 <- lm(wrecks ~ numstorms, data=terns2)
> summary(categorical_lm1)

Call:
lm(formula = wrecks ~ numstorms, data = terns2)

Residuals:
   Min     1Q Median     3Q    Max 
-9.375 -1.750 -0.250  1.781 12.625 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)   
(Intercept)        1.750      2.128   0.822  0.42474   
numstormsLevel 2   9.625      3.010   3.198  0.00645 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.02 on 14 degrees of freedom
Multiple R-squared:  0.4221,	Adjusted R-squared:  0.3808 
F-statistic: 10.22 on 1 and 14 DF,  p-value: 0.006451
```

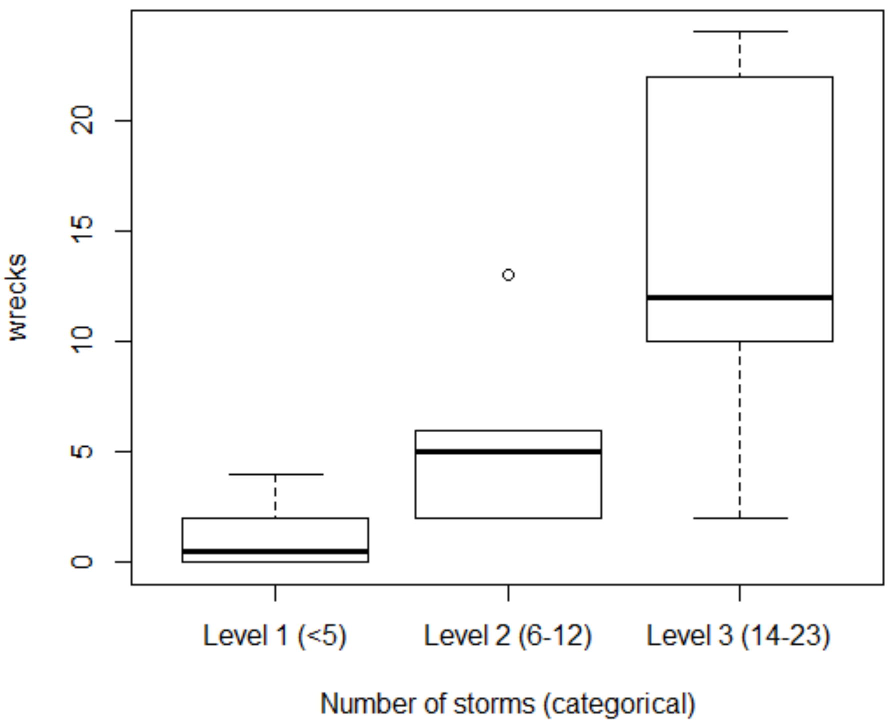
```
> t.test(terns2$wrecks ~ terns2$numstorms)

	Welch Two Sample t-test

data:  terns2$wrecks by terns2$numstorms
t = -3.1976, df = 7.7388, p-value = 0.01322
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.60713  -2.64287
sample estimates:
mean in group Level 1 mean in group Level 2 
                1.750                11.375 
```

$$1.75 + 9.625 = 11.375$$

Categorical predictive (X) variable (>2 levels)



wrecks	hurricanes	numstorms
0	0	Level 1 (<5)
1	0	Level 1 (<5)
0	1	Level 1 (<5)
0	3	Level 1 (<5)
2	4	Level 1 (<5)
4	4	Level 1 (<5)
2	6	Level 2 (6-12)
5	6	Level 2 (6-12)
2	8	Level 2 (6-12)
6	12	Level 2 (6-12)
13	12	Level 2 (6-12)
2	14	Level 3 (14-23)
12	15	Level 3 (14-23)
24	15	Level 3 (14-23)
10	19	Level 3 (14-23)
22	23	Level 3 (14-23)

R language:

lm(wrecks ~ numstorms, data = terns)

If >2 levels in X, model output is like ANOVA

```
> categorical_lm <- lm(wrecks ~ numstorms, data=terns3)
> summary(categorical_lm)

Call:
lm(formula = wrecks ~ numstorms, data = terns3)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0000  -2.4000  -0.8833   1.3333  10.0000 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)   
(Intercept)                 1.167      2.326   0.502  0.62436   
numstormsLevel 2 (6-12)     4.433      3.450   1.285  0.22123   
numstormsLevel 3 (14-23)   12.833      3.450   3.720  0.00257 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.698 on 13 degrees of freedom
Multiple R-squared:  0.5193,	Adjusted R-squared:  0.4453 
F-statistic:  7.022 on 2 and 13 DF,  p-value: 0.008555
```

```
> anova_version <- aov(wrecks ~ numstorms, data=terns3)
> summary(anova_version)
            Df Sum Sq Mean Sq F value  Pr(>F)   
numstorms    2  455.9  227.95   7.022 0.00856 **
Residuals   13  422.0   32.46                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

If >2 levels in X, model output is like ANOVA

```
> categorical_lm <- lm(wrecks ~ numstorms, data=terns3)
> summary(categorical_lm)

Call:
lm(formula = wrecks ~ numstorms, data = terns3)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0000  -2.4000  -0.8833   1.3333  10.0000 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)   
(Intercept)                1.167      2.326   0.502  0.62436   
numstormsLevel 2 (6-12)    4.433      3.450   1.285  0.22123   
numstormsLevel 3 (14-23)  12.833      3.450   3.720  0.00257 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.698 on 13 degrees of freedom
Multiple R-squared: 0.5193,	Adjusted R-squared: 0.4453 
F-statistic: 7.022 on 2 and 13 DF,  p-value: 0.008555
```

```
> anova_version <- aov(wrecks ~ numstorms, data=terns3)
> summary(anova_version)
            Df Sum Sq Mean Sq F value  Pr(>F)   
numstorms    2  455.9  227.95   7.022 0.00856 **
Residuals   13  422.0   32.46                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting β´s between levels:

As you go from Level 1 to Level 2, model predicts 4.43 more wrecks

1.167 + 4.43 = 5.6 wrecks at X = 2

As you go from Level 1 to Level 3, model predicts 12.83 more wrecks

1.167 + 12.83 = ~ 14.0 wrecks at X = 3