# Intro to Social Science Data Analysis

# Week 11: Simple Linear Regression

**Christopher Gandrud**

November 5, 2012

# Quick Quiz 1

Find the sample proportions of the following party's supporters:

| Saenuri | DUP | Other | Total |
|---|---|---|---|
| 1064 | 891 | 520 | 2475 |

# Quick Quiz 1

| Saenuri | DUP | Other | Total |
|---|---|---|---|
| 1064 | 891 | 520 | 2475 |
| (0.43) | (0.36) | (0.21) | (1) |

# Quick Quiz 2

If we wanted to make inferences about **population proportions** from sampling proportions, what **distribution** do we often assume the sampling proportions follow?

What are it's **parameters**?

# Quick Quix 3

Imagine we have a two-way contingency table.

| | Attend University | No University |
|---|---|---|
| Married | | |
| Not Married | | |

If we conducted a $\chi^2$ test with this data and found a p-value of $< 0.001$ what would we conclude?

# Quick Quiz 4

Write the simple linear regression equation for how a person's height is related to their income.

# Quick Quiz 5

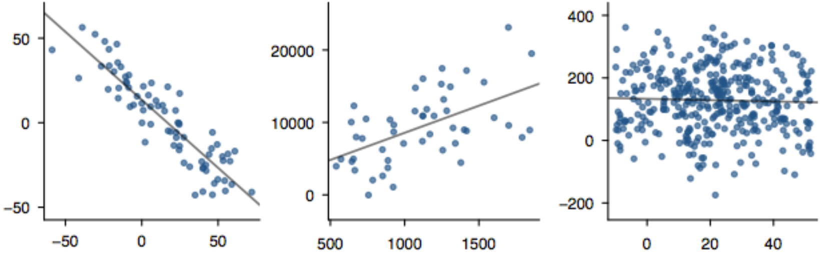
Describe how a linear regression line would look if the relationship between two variables was negative.

How would it look if the relationship was positive?

What about no relationship?

Motivation Since almost no interesting relationship is perfectly linear, how do we find the **best fit line** that describes the relationship between some $x$ and some $y$?

# How?

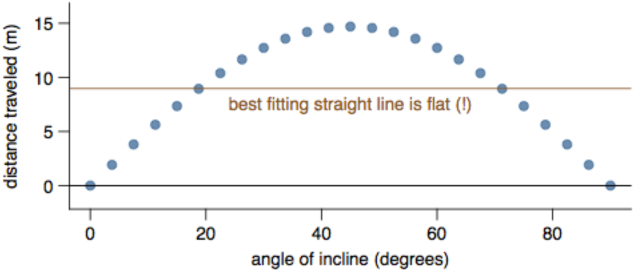



Source: Diaz et. al. (2011, 216)

In **simple linear regression** we are trying to find the straight line that is **as close to all of the data points as possible**.

How do we find this line?

# How?



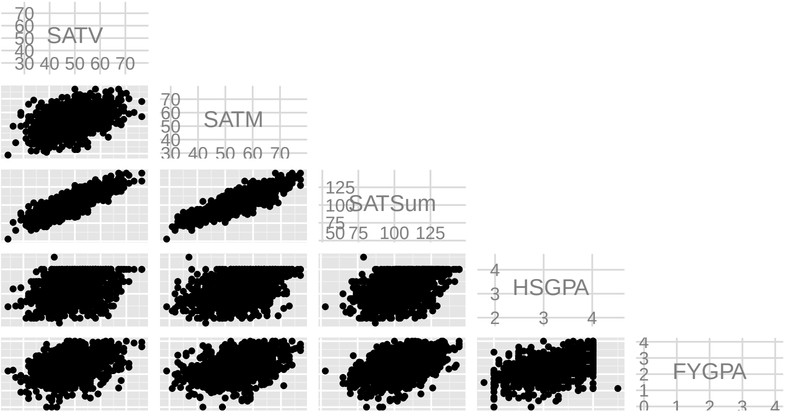

Source: Diaz et. al. (2011, 216)

Let's use the SAT/GPA data from the `openintro` package:

```
# Load library
library(openintro)

# Load data
data(satGPA)

# Show variables
names(satGPA)

## [1] "sex"    "SATV"   "SATM"   "SATSum" "HSGPA"
## [6] "FYGPA"


# Subset to remove the sex variable
satGPASlim <- satGPA[, 2:6]
```

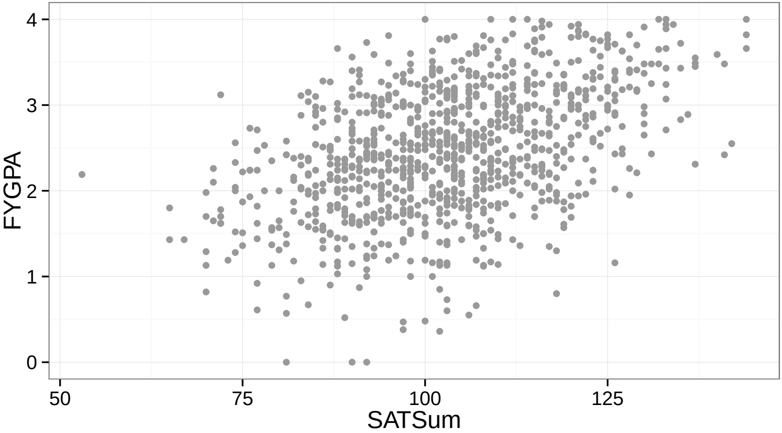
# Plot the SAT Scores & GPAs

```r
library(GGally)

ggpairs(satGPASlim, upper = "blank")
```



# First Year GPA

Universities want to know how well student's total SAT scores (`SATSum`) relate to their academic performance in the first year of university (`FYGPA`.



One way to describe the overall relationship between `SATSum` and `FYGPA` is to find the **correlation** between the two variables.

# Correlation ($R$):

Describes the **strength** of a linear relationship.

It ranges from -1 to 1.

-1 indicates a **perfect negative relationship**.

1 indicates a **perfect positive relationship**.

0 indicates **no correlation/relationship**.

To find the correlation for observations
$(x_1, y_1),\ (x_2, y_2) \dots (x_n, y_n)$

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

# Or. . .

Or we can have R do the maths for us.

```
cor(satGPA$SATSum, satGPA$FYGPA)

## [1] 0.4603
```

# Statistical Significance & Correlation

If we wanted to test to see if the correlation is statistically significant, what would the null hypothesis be?

# Statistical Significance & Correlation

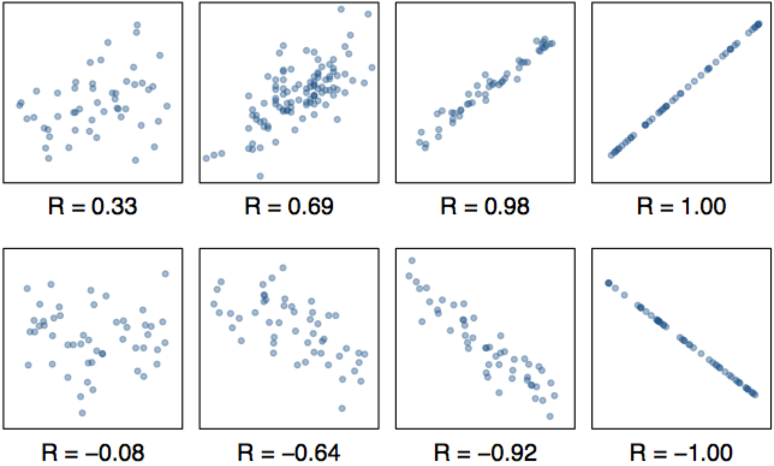
$H_0$: $R = 0$

$H_a$: R $\neq 0$

# Hypothesis Testing Correlation Coefficients in R

```
cor.test(satGPA$SATSum, satGPA$FYGPA)

##
##  Pearson's product-moment correlation
##
## data:  satGPA$SATSum and satGPA$FYGPA
## t = 16.38, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.4100 0.5078
## sample estimates:
##    cor
## 0.4603
```

# More Correlation Examples

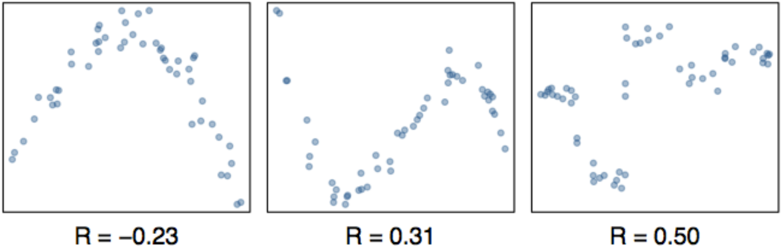



Source: Diaz et al. (2011, 282 )

# Caution

A low linear correlation **does not necessarily** mean a weak relationship.

It means a weak **linear** relationship.




Source: Diaz et al. (2011, 282 )

Ok, linear correlations are useful for finding:

- the **direction** of a linear relationship,
- the **strength** of a linear relationship.

# Best Fit Lines & Least Squares Regression

Ok, linear correlations are useful for finding:

- the **direction** of a linear relationship,
- the **strength** of a linear relationship.

# More specific

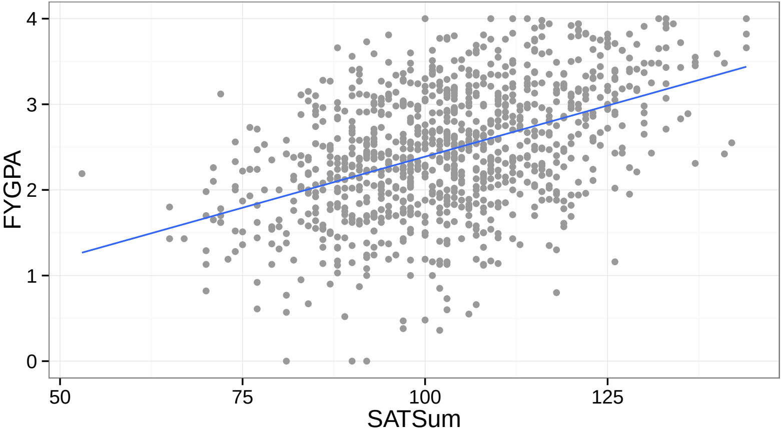
What if we want to be more specific?

For example, using a student's total SAT score to predict their first year university GPA.

**Note:** the estimated value of the dependent variable ($y$) is often written $\hat{y}$ ( *"y hat"*).

# The Linear Best Fit Line

The blue line is the closest straight line ("best fit") to all of the data points.



How do we find the best fit line?

# Residuals

Well, the best fit line would do something like have the smallest **residuals** possible.

What is a residual?

# Residual:

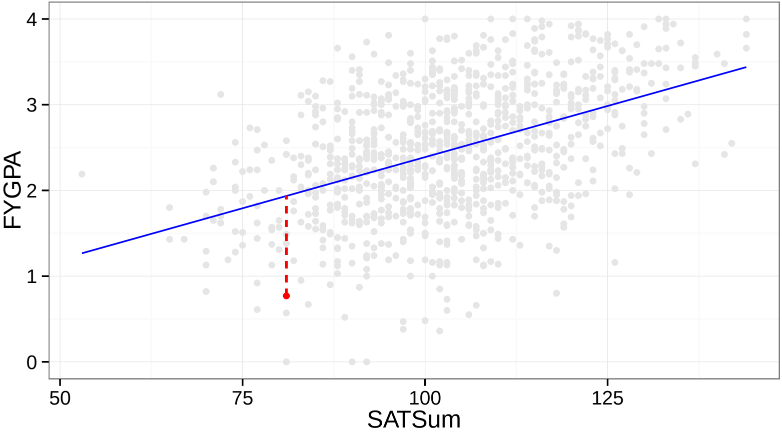
the difference between the observed and expected values based on the best fit model.

More formally: the residual $(e_i)$ of the observation $(x_i, y_i)$ is the difference between the observed value of $y_i$ and the expected value $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i$$

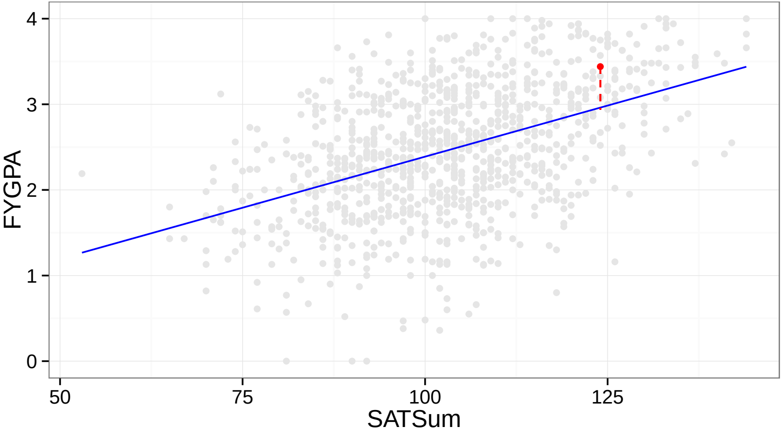
# Residuals

The red point is at (81, 0.77). Given that `SATSum` is 81, it is expected to be at 1.935. So, it's residual is $0.77 - 1.935 = -1.65$.



# Residuals

The red point is at (124, 3.44). Given that `SATSum` is 124, it is expected to be at 2.94. So, it's residual is $3.44 - 2.94 = 0.5$.



# Outliers

# Dummy Variables

So far we have only looked at creating simple linear regression models with **continuous numeric** dependent and independent variables.

What if we have a continuous dependent variable and a dichotomous independent variable?

# Categorical Dependent

What if our **dependent variable** is categorical, for example, the party someone voted for?

For these situations you need to use a different type of regression, usually **logistic regression**.

We do not cover this type of regression in this course.

# Caution: Non-Linear Relationships

It is always a good idea to check for **non-linear relationships**.

One way to address non-linear relationships is to **transform** the data using, for example:

- logs
- squares, cubes.

One way to address non-linear relationships is to **transform** the data using, for example:

- logs
- squares, cubes.

# References I


Crawley, Michael J. 2005. Statistics: An Introduction Using R. Chichester: John Wiley Sons. Ltd.

Diaz, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2011. OpenIntro Statistics. 1st ed. http://www.openintro.org/stat/downloads.php.