

# Intro to Social Science Data Analysis

## Week 11 Lecture: Simple Linear Regression

**Christopher Gandrud**

November 12, 2012

- 1 Recap
- 2 Correlation
- 3 Best Fit Lines & Least Squares Regression
- 4 Hypothesis Testing
- 5 Some Special Issues in Simple Linear Regression

## Assignment 3

Like all data assignments in this course, your response to Assignment 3 must be **reproducible**.

## 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 its **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?

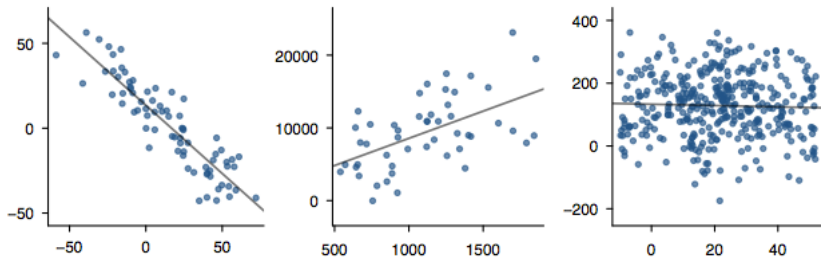
## Quick Quiz 6

Describe the variables in the simple linear regression equation.

$$y = \alpha + \beta x$$

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?

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"
```

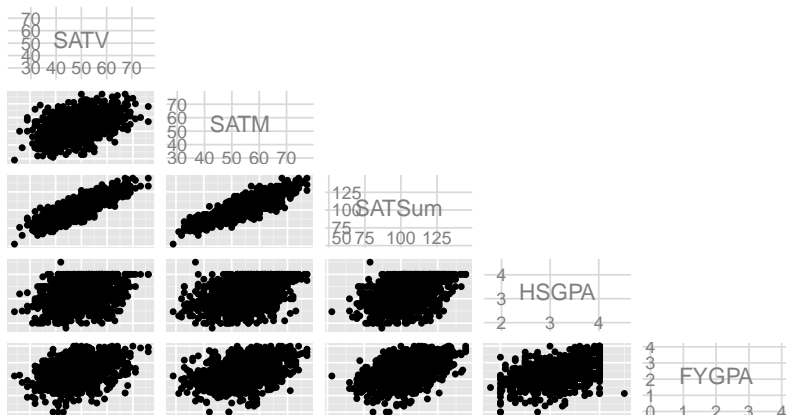
  

```
# Subset to remove the sex variable
satGPASlim <- satGPA[, 2:6]
```

# Plot the SAT Scores & GPAs

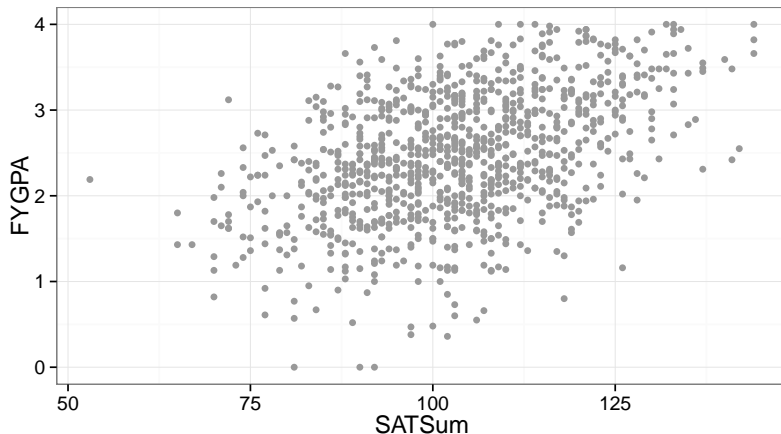
```
library(GGally)
```

```
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).





# Correlation

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**.

# Correlation

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

$$H_0: R = 0$$

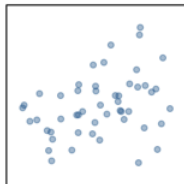
$$H_a: R \neq 0$$

# Hypothesis Testing Correlation Coefficients in R

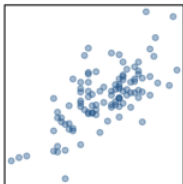
```
cor.test(satGPA$SATSum, satGPA$FYGPA)

##
##  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 0
## 95 percent confidence interval:
##  0.4100 0.5078
## sample estimates:
##      cor
## 0.4603
```

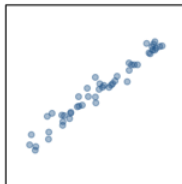
# More Correlation Examples



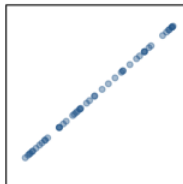
$R = 0.33$



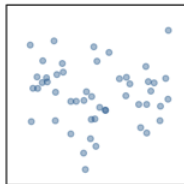
$R = 0.69$



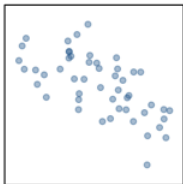
$R = 0.98$



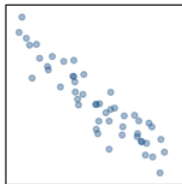
$R = 1.00$



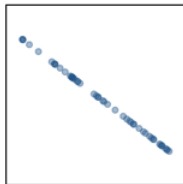
$R = -0.08$



$R = -0.64$



$R = -0.92$



$R = -1.00$

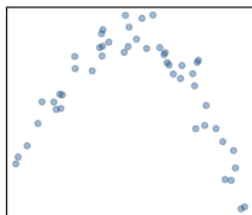
Source: Diaz et al. (2011, 282 )



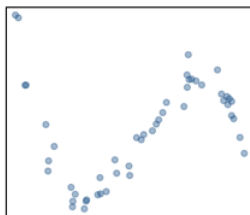
## Caution

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

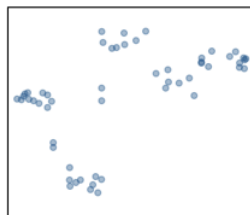
It means a weak **linear** relationship.



$R = -0.23$



$R = 0.31$



$R = 0.50$

Source: Diaz et al. (2011, 282 )

# 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.

# 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

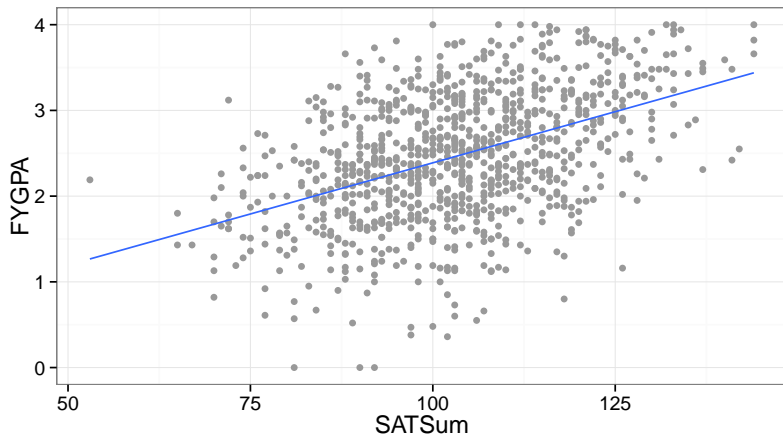
What if we want to be more specific?

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?

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.

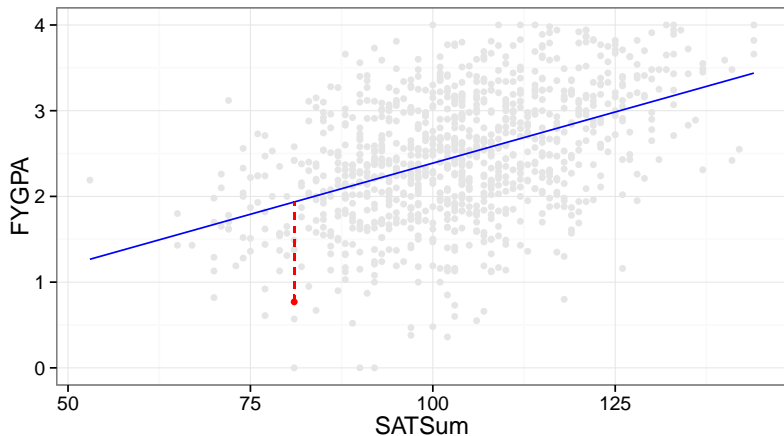
More formally: the residual ( $e_i$ ) of the observation  $(x_i, y_i)$  is the difference between the observed value of  $y_i$  and the expected value  $\hat{y}_i$ :

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



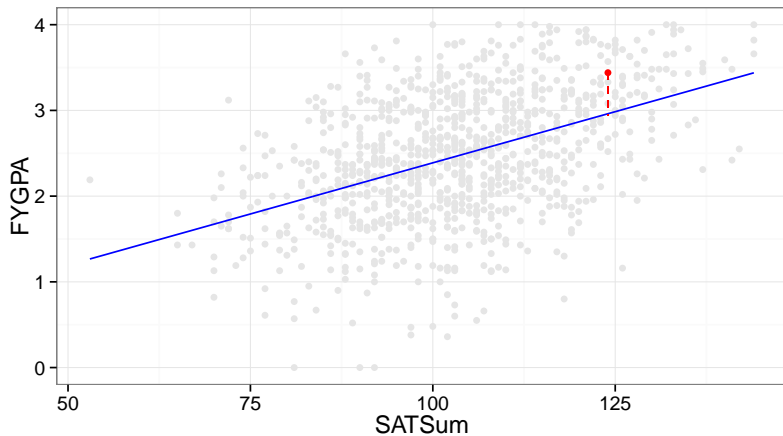
# Residuals

The red point is at (81, 0.77). Given that SATSum is 81, it is expected to be at 1.935. So, its 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, its residual is  $3.44 - 2.94 = 0.5$ .



# Aim

We want to find the line that gets us the smallest **sum of squared residuals** (SSR) where the SSR is:

$$SSR = e_1^2 + e_2^2 + \dots + e_n^2$$

Um...

So where do the expected values ( $\hat{y}$ ) come from?

Remember the simple linear regression equation:

$$y = \alpha + \beta x$$

We want to find the  $\hat{\beta}$  that **minimizes** the SSR.

Formally:

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^p} \text{SSR}$$

This is called **ordinary least squares simple linear regression**.

How?

How do we find the  $\hat{\beta}$  that **minimizes** the SSR?

## By Hand

One way to estimate the correlation coefficient parameter  $\beta$  for  $x$  is with the following equation:

$$\hat{\beta} = \frac{s_y}{s_x} R$$

## Calculate $\hat{\beta}$

```
# Find standard errors
SDy <- sd(satGPA$FYGPA)
SDx <- sd(satGPA$SATSum)

# Find correlation
CorXY <- cor(satGPA$SATSum, satGPA$FYGPA)

# Estimate correlation coefficient
BetaHat <- (SDy/SDx) * CorXY

BetaHat

## [1] 0.02387
```



## Finding the intercept.

How can we find the intercept?

## Finding the intercept ( $\hat{\alpha}$ ).

You might remember from maths that we can find the whole equation for a line if we know:

- ▶ a point on the line.
- ▶ the slope.

We know the slope and we know that the point at the mean of  $x$  and  $y$  ( $\bar{x}, \bar{y}$ ) will be on the line so we can use the equation for the **point-slope** form of a line:

$$y - \bar{y} = \hat{b}(x - \bar{x})$$

## Finding the intercept ( $\hat{\alpha}$ ).

If  $\bar{y} = 2.468$ ,  $\bar{x} = 103.329$ , and  $\hat{\beta} = 0.02387$ , then:

$$y - \bar{y} = \hat{\beta}(x - \bar{x})$$

$$y - 2.468 = 0.02387(x - 103.329)$$

$$y - 2.468 = 0.02387x - 2.466463$$

$$y = 0.2387x + 0.001537$$

$$\widehat{FYGPA} = 0.001537 + 0.2387SATSum$$

## This class

In this class we will let the computer find the  $\hat{\beta}$  and  $\hat{\alpha}$ .

# Linear Model

In R you can use the `lm` (linear model) command. For example,

```
M1 <- lm(FYGPA ~ SATSum, data = satGPA)
```

```
M1
```

```
##
```

```
## Call:
```

```
## lm(formula = FYGPA ~ SATSum, data = satGPA)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)      SATSum
```

```
##      0.00193      0.02387
```

# The Regression Equation

So again, our estimated regression equation is:

$$\widehat{FYGPA} = 0.00193 + 0.02387SATSum$$

## Linear Regression Assumptions:

- ▶ The data follow a **linear trend**,
- ▶ Nearly **normally distributed residuals**,
- ▶ There is **constant variability**.

## Linear Regression Assumptions:

- ▶ The data follow a **linear trend**,
- ▶ Nearly **normally distributed residuals**,
- ▶ There is **constant variability**.

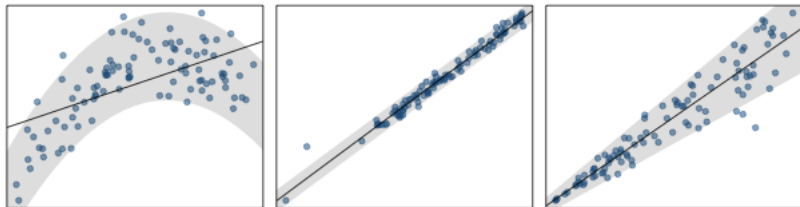


## Linear Regression Assumptions:

- ▶ The data follow a **linear trend**,
- ▶ Nearly **normally distributed residuals**,
- ▶ There is **constant variability**.

## Example Assumption Violations

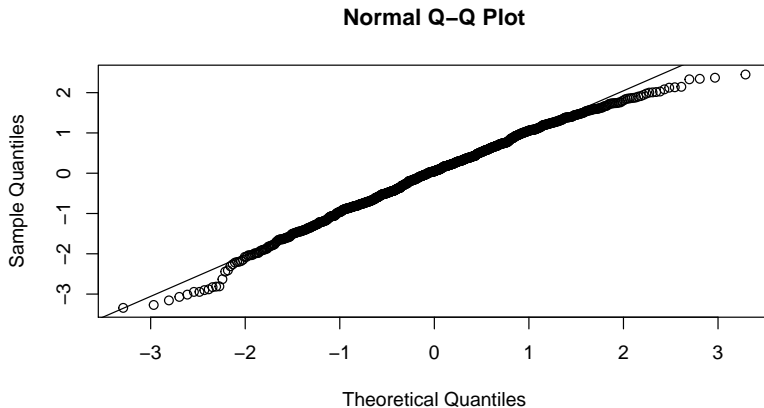
Which assumptions do these data violate?



Source: Diaz et al. (2011, 285)

To determine if the residuals in the model M1 are normally distributed:

```
# Find standardized residuals  
M1.residuals <- rstandard(M1)  
# Create Quantile-Quantile Plot  
qqnorm(M1.residuals)  
qqline(M1.residuals)
```



# Hypothesis Testing

Remember that  $\hat{\beta}$  is a **point estimate** of the **population parameter**  $\beta$ .

How can we make **inferences** about  $\beta$  from  $\hat{\beta}$ ?

Note: some people use  $b$  to refer to  $\hat{\beta}$ .

What would our null and alternative hypotheses be?

# Hypotheses

For our example, with a positive slope of 0.2387:

$$H_0: \beta = 0$$

$$H_a: \beta > 0$$

## P-values for $\hat{\beta}$

We usually assume that the sampling distribution of  $\hat{\beta}$  follows a  $t$  distribution.

Remember the equation for the  $t$  test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{\text{SE}}$$

with  $n - 2$  degrees of freedom.

The procedure is the same as before to find the p-value.

```
# Summarize M1 model output
```

```
summary(M1)
```

```
##
```

```
## Call:
```

```
## lm(formula = FYGPA ~ SATSum, data = satGPA)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

##	-2.1976	-0.4495	0.0315	0.4557	1.6115
----	---------	---------	--------	--------	--------

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
```

## (Intercept)	0.00193	0.15199	0.01	0.99
## SATSum	0.02387	0.00146	16.38	<2e-16

```
##
```

```
## (Intercept)
```

```
## SATSum      ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```



## Inference & p-values

So, we find evidence against the null hypothesis that the slope of the line summarizing the relationship between first year university grades and SAT total scores is 0 in the population.

Note: the p-value given by summary is based on a **two-sided** hypothesis. If we have a one sided hypothesis we can **halve** the p-value.

In our example this would be impractical, since the p-value is already so small.

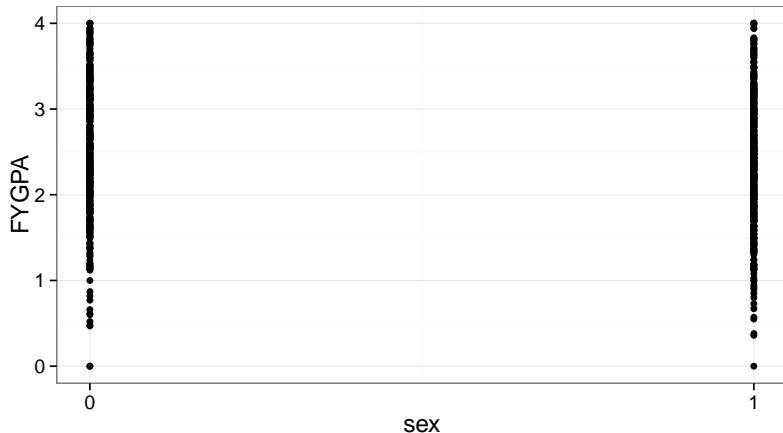
# Dummy Variables

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 (dummy) independent** variable?

# Dichotomous Independent Variables

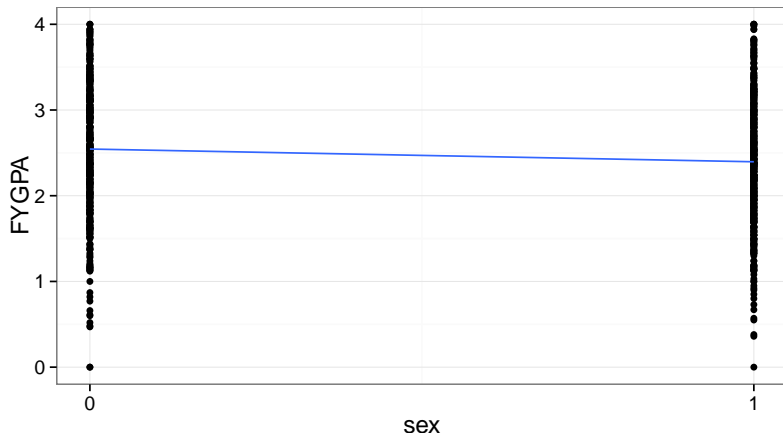
For example:



Note, I recoded the values of the original variable.

## Dichotomous Independent Variables

$\beta$  is pretty similar. It is still the slope of the line for a one unit change in  $x$ . The only difference is that the variable only goes between 0 and 1.



Note, I recoded the values of the original 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, for example **logistic regression**.

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

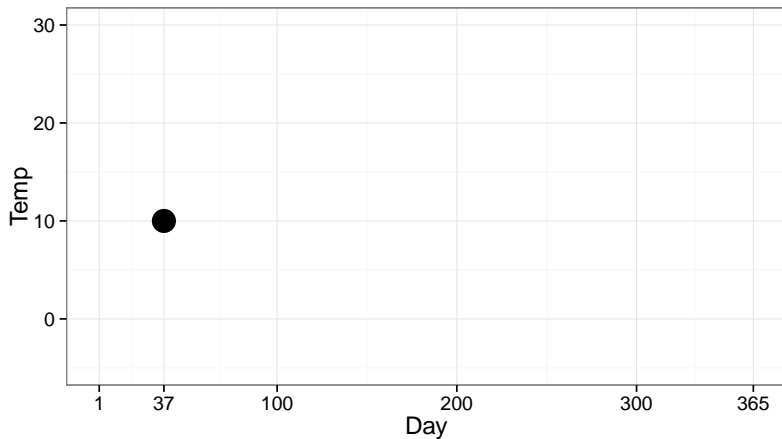
# Extrapolation

Be careful about **extrapolating** beyond your data.

We don't know how the data beyond what we observe will behave.

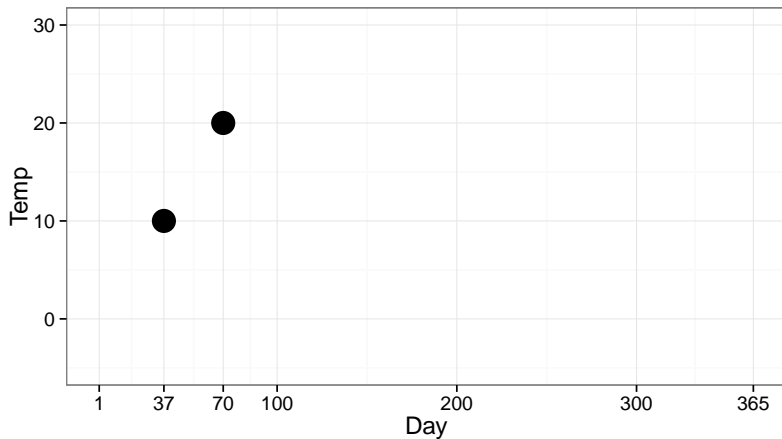
# Extrapolation

On the 37th day of the year it was 10 degrees.



# Extrapolation

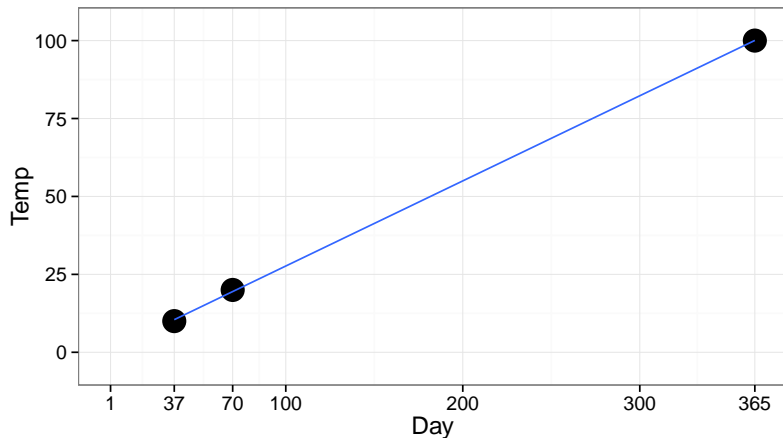
On the 70th day of the year it was 20 degrees.





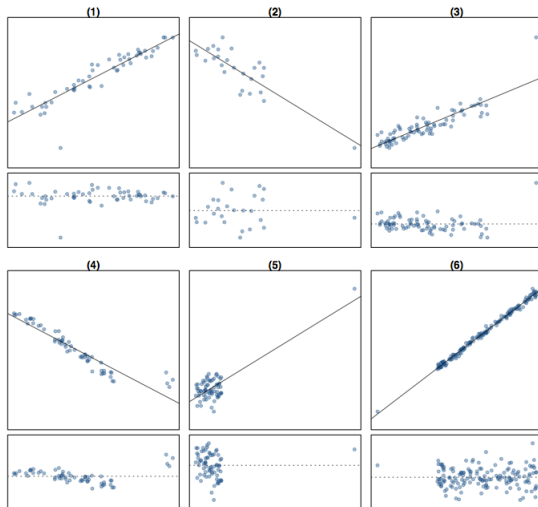
# Extrapolation

So on the last day of the year it will be about 100 degrees?.



# Outliers

Be careful about outliers.



## Removing Outliers

Only remove outliers if you have a **good reason** to.

Try to find out **substantively** why they are outliers.

## 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>.