THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

# Correlation & Simple Regression 2

## Lecture 10

STA 371G

1. Regression basics

2. Regression assumptions

3. Regression to the mean

Add Health

The National Longitudinal Study of Adolescent to Adult Health

National Longitudinal Study of Adolescent to Adult Health

Nationally representative sample of US students in grades 7-12 were surveyed in the 1994-95 school year (`http://www.cpc.unc.edu/projects/addhealth`)

Students were followed up on with subsequent in-home interviews four times (most recently 2008)

This is an **awesome** data set, with data on:

- family
- relationships
- health
- military service
- religion
- sex and STDs
- economics
- education
- personality
- criminality
- tobacco
- drugs
- alcohol
- pregnancy
- sleep
- daily activities

Do people that start drinking younger tend to drink more (or less) when they become adults?

Do people that start drinking younger tend to drink more (or less) when they become adults?

We want to know:

- What is our best **prediction** of alcohol consumption if we know at what age they had their first drink?

Do people that start drinking younger tend to drink more (or less) when they become adults?

We want to know:

- What is our best **prediction** of alcohol consumption if we know at what age they had their first drink?
- How good is that prediction?

Do people that start drinking younger tend to drink more (or less) when they become adults?

We want to know:

- What is our best **prediction** of alcohol consumption if we know at what age they had their first drink?
- How good is that prediction?
- What is the **relationship** between alcohol consumption and age of first drink?

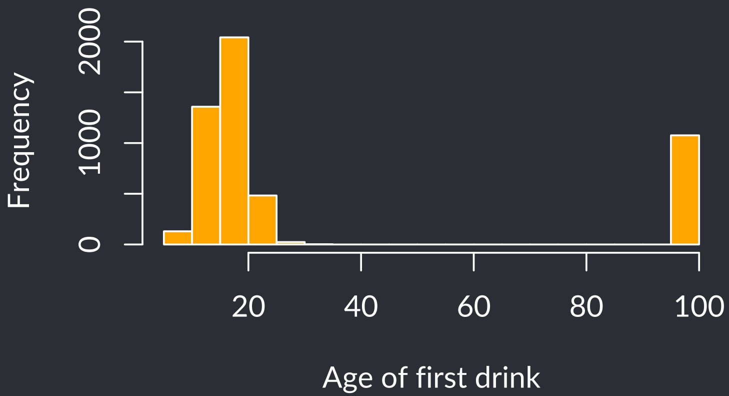
| | |
|---|---|
| Age of first drink | **Explanatory variable** ($X$) |
| Number of drinks consumed as adult | **Response variable** ($Y$) |

# Working with a large data set

Put the variables you will work with into a new data set in R; it's easier to work with and clean up that way.

```
drinking <- data.frame(age=addhealth$h4to34,
                       num.drinks=addhealth$h4to36)
```

```r
hist(drinking$age,
  main='', xlab="Age of first drink",
  col="orange")
```

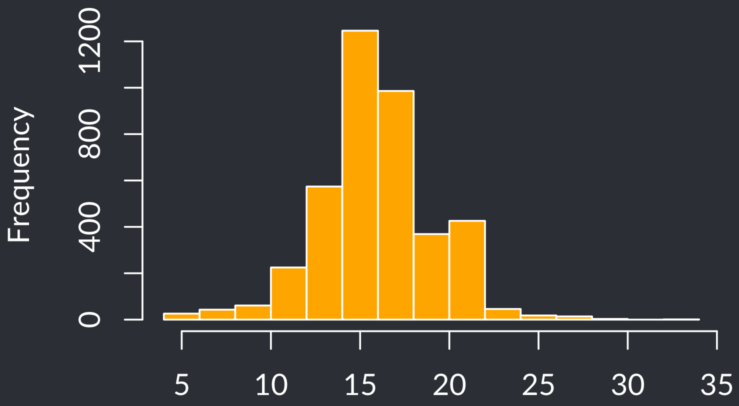


# Let’s examine our variables using the codebook

***If Q.33 = 1, ask Q.34, else skip to Q.63.***

| | | | |
|---|---|---|---|
| **H4TO34** | | Num | 34. How old were you when you first had an alcoholic drink? By drink, we mean a glass of wine, a can or bottle of beer, a wine cooler, a shot glass of liquor, or a mixed drink, not just sips or tastes from someone else's drink.<br>NOTE: Smallest 5 and largest 5 values are displayed. |
| Frequency | Percent | Value | Label |
| 56 | 0.4% | 5 | 5 years |
| 30 | 0.2% | 6 | 6 years |
| 21 | 0.1% | 7 | 7 years |
| 71 | 0.5% | 8 | 8 years |
| 52 | 0.3% | 9 | 9 years |
| 12014 | 76.5% | 10-31 | NOTE: Range of values omitted from display |
| 1 | 0.0% | 32 | 32 years |
| 2 | 0.0% | 33 | 33 years |
| 21 | 0.1% | 96 | refused |
| 3322 | 21.2% | 97 | legitimate skip |
| 111 | 0.7% | 98 | don't know |

```r
drinking$age[drinking$age >= 96] <- NA
hist(drinking$age, main='', xlab='', col="orange")
```

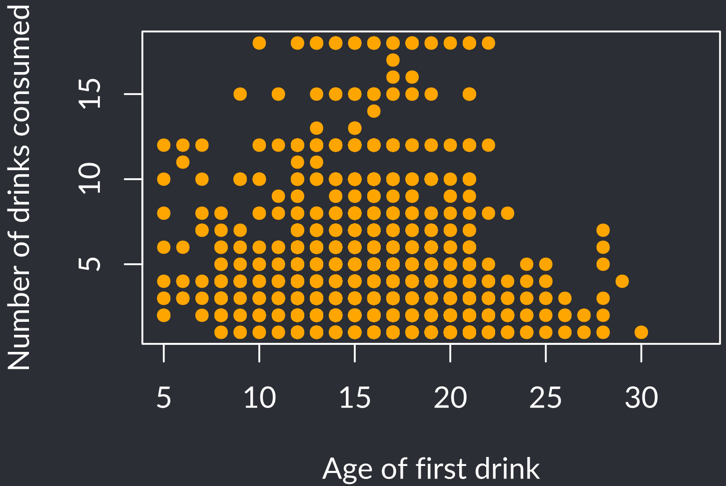


# Let's examine our variables using the codebook

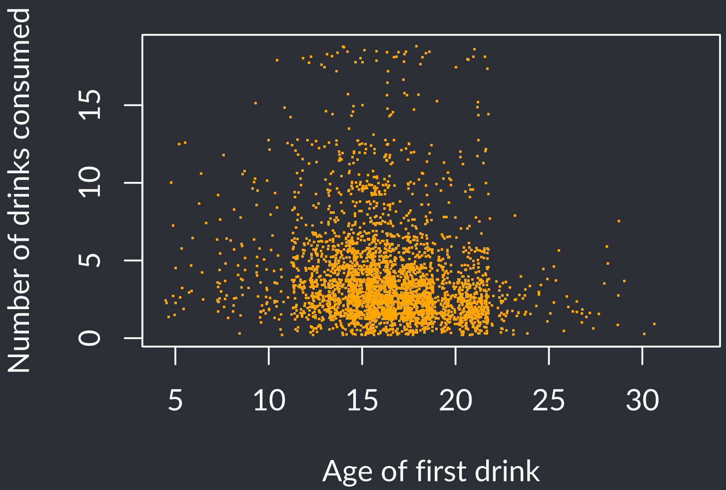
| | | | |
|---|---|---|---|
| ***If Q.35 not equal 0, ask Q.36, else if Q.35 = 0, then skip to Q.43.*** | | | |
| **H4TO36** | | Num | 36. Think of all the times you have had a drink during the past 12 months. How many drinks did you **usually** have each time? A 'drink' is a glass of wine, a can or bottle of beer, a wine cooler, a shot glass of liquor, or a mixed drink.<br>NOTE: Smallest 5 and largest 5 values are displayed. |
| Frequency | Percent | Value | Label |
| 1651 | 10.5% | 1 | 1 drink |
| 3051 | 19.4% | 2 | 2 drinks |
| 2274 | 14.5% | 3 | 3 drinks |
| 1343 | 8.6% | 4 | 4 drinks |
| 891 | 5.7% | 5 | 5 drinks |
| 1815 | 11.6% | 6-16 | NOTE: Range of values omitted from display |
| 4 | 0.0% | 17 | 17 drinks |
| 108 | 0.7% | 18 | 18 drinks |
| 27 | 0.2% | 96 | refused |
| 4427 | 28.2% | 97 | legitimate skip |
| 110 | 0.7% | 98 | don't know |

```r
drinking$num.drinks[drinking$num.drinks >= 96] <- NA
hist(drinking$num.drinks, main='', xlab='How many drinks',
  col="orange")
```

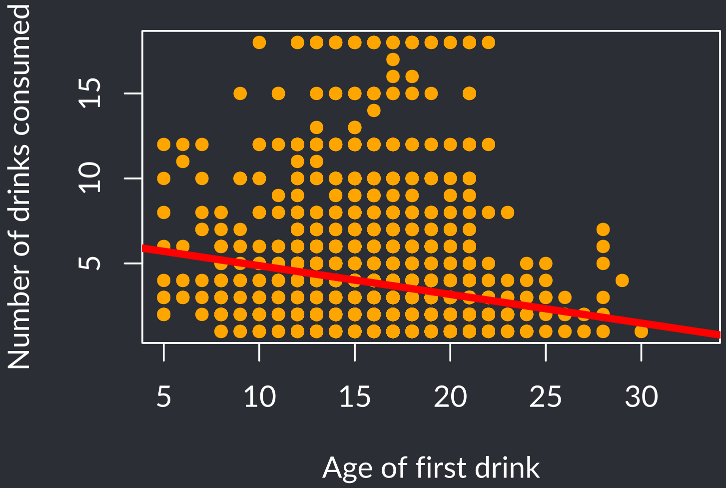
```r
plot(num.drinks ~ age, data=drinking, pch=16, col="orange",
  xlab="Age of first drink",
  ylab="Number of drinks consumed")
```



```r
plot(jitter(num.drinks, 4) ~ jitter(age, 4),
  data=drinking, pch=".", col="orange",
  xlab="Age of first drink",
  ylab="Number of drinks consumed")
```



The regression line is the line of “best fit” through this plot:



# What is linear regression doing?

We model each case ($x_i$ = age for $i$th person, $y_i$ = number of drinks for $i$th person) as a linear relationship plus some error:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$\beta_0$ and $\beta_1$ are the intercept and slope, respectively.

# What is linear regression doing?

We model each case ($x_i$ = age for $i$th person, $y_i$ = number of drinks for $i$th person) as a linear relationship plus some error:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$\beta_0$ and $\beta_1$ are the intercept and slope, respectively.

We find estimates for $\beta_0$ and $\beta_1$ in our sample that *minimize* the errors:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

This is the regression (best fit) line.

# Finding the regression equation

When there is just one $X$ variable, the formulas are straightforward:

1. The slope is $\hat{\beta}_1 = r \cdot \dfrac{\mathrm{SD}(Y)}{\mathrm{SD}(X)}$.
2. Solve for the intercept using the fact that the regression line will always pass through $(\overline{X}, \overline{Y})$: $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$.

```
model <- lm(num.drinks ~ age, data=drinking)
summary(model)


Call:
lm(formula = num.drinks ~ age, data = drinking)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2035 -1.8528 -0.8528  0.8095 15.1602

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.55417    0.26532   24.70   <2e-16 ***
age         -0.16883    0.01588  -10.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.963 on 3600 degrees of freedom
  (2902 observations deleted due to missingness)
Multiple R-squared:  0.03044,Adjusted R-squared:  0.03017
F-statistic:   113 on 1 and 3600 DF,  p-value: < 2.2e-16
```

This translates to a regression line of:

$$\widehat{\text{num drinks}} = 6.55 - 0.17 \cdot \text{age}$$

This translates to a regression line of:

$$\widehat{\text{num drinks}} = 6.55 - 0.17 \cdot \text{age}$$

Predict number of drinks for age $= 21$:

$$\widehat{\text{num drinks}} = 6.55 - 0.17 \cdot 21 = 3.01$$

Or we can use R to do the work for us:

```r
predict(model, list(age=21))
```

## How good are our predictions?

$R^2$ quantifies how closely the model fits the data.

- $R^2$ is the fraction of the variation of $Y$ explained by the model (i.e., $R^2 = \text{Var}(\hat{Y})/\text{Var}(Y)$).

# How good are our predictions?

$R^2$ quantifies how closely the model fits the data.

- $R^2$ is the fraction of the variation of $Y$ explained by the model (i.e., $R^2 = \mathrm{Var}(\hat{Y})/\mathrm{Var}(Y)$).
- $R^2 = \mathrm{cor}(X, Y)^2$, i.e., the squared correlation between $X$ and $Y$.

## How good are our predictions?

$R^2$ quantifies how closely the model fits the data.

- $R^2$ is the fraction of the variation of $Y$ explained by the model (i.e., $R^2 = \text{Var}(\hat{Y})/\text{Var}(Y)$).
- $R^2 = \text{cor}(X, Y)^2$, i.e., the squared correlation between $X$ and $Y$.
- $R^2 = 0$ when the model has no predictive power at all.

# How good are our predictions?

$R^2$ quantifies how closely the model fits the data.

- $R^2$ is the fraction of the variation of $Y$ explained by the model (i.e., $R^2 = \text{Var}(\hat{Y})/\text{Var}(Y)$).
- $R^2 = \text{cor}(X, Y)^2$, i.e., the squared correlation between $X$ and $Y$.
- $R^2 = 0$ when the model has no predictive power at all.
- $R^2 = 1$ when the model yields perfect predictions every time.

# How good are our predictions?

$R^2$ quantifies how closely the model fits the data.

- $R^2$ is the fraction of the variation of $Y$ explained by the model (i.e., $R^2 = \text{Var}(\hat{Y})/\text{Var}(Y)$).
- $R^2 = \text{cor}(X, Y)^2$, i.e., the squared correlation between $X$ and $Y$.
- $R^2 = 0$ when the model has no predictive power at all.
- $R^2 = 1$ when the model yields perfect predictions every time.
- $R^2 = \text{cor}(Y, \hat{Y})^2$, i.e., the squared correlation between the actual and predicted values of $Y$.

```
model <- lm(num.drinks ~ age, data=drinking)
summary(model)


Call:
lm(formula = num.drinks ~ age, data = drinking)

Residuals:
   Min     1Q Median     3Q    Max
-4.204 -1.853 -0.853  0.810 15.160

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5542     0.2653    24.7   <2e-16 ***
age          -0.1688     0.0159   -10.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3 on 3600 degrees of freedom
  (2902 observations deleted due to missingness)
Multiple R-squared:  0.0304,Adjusted R-squared:  0.0302
F-statistic:  113 on 1 and 3600 DF,  p-value: <2e-16
```

In our regression, $R^2 = 0.03$, so $r = \sqrt{0.03} = -0.17$ (negative since the slope is negative).

In our regression, $R^2 = 0.03$, so $r = \sqrt{0.03} = -0.17$ (negative since the slope is negative).

Is this "significant?"

In our regression, $R^2 = 0.03$, so $r = \sqrt{0.03} = -0.17$ (negative since the slope is negative).

Is this "significant?" We'll discuss this next time!

1. Regression basics

2. Regression assumptions

3. Regression to the mean

In finance, the $\beta$ of an asset indicates its volatility relative to the market. An asset with:

In finance, the $\beta$ of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.

In finance, the $\beta$ of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.
- $\beta > 1$ is **more** volatile than the market as a whole.

In finance, the $\beta$ of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.
- $\beta > 1$ is **more** volatile than the market as a whole.
- $\beta < 1$ is **less** volatile than the market as a whole.

In finance, the $\beta$ of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.
- $\beta > 1$ is **more** volatile than the market as a whole.
- $\beta < 1$ is **less** volatile than the market as a whole.

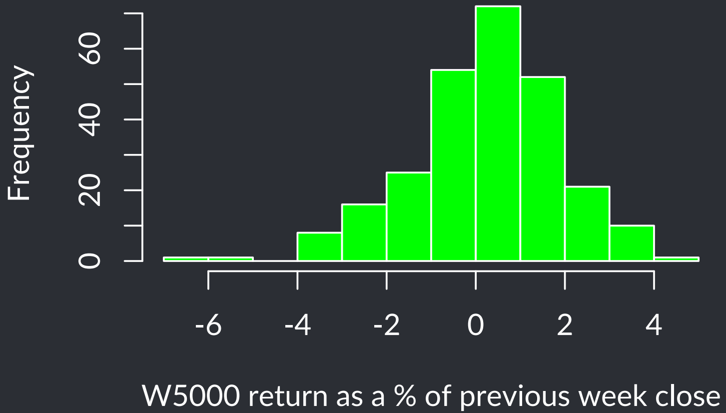
In finance, the $\beta$ of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.
- $\beta > 1$ is **more** volatile than the market as a whole.
- $\beta < 1$ is **less** volatile than the market as a whole.

$\beta$ is just the slope of the regression line (i.e. $\hat{\beta}_1$) when we regress the asset's weekly returns against the weekly returns of a market index.

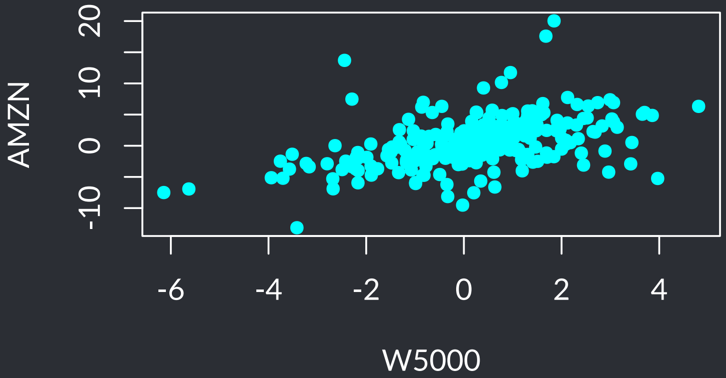
# W5000 (Wilshire 5000, a broad market index)

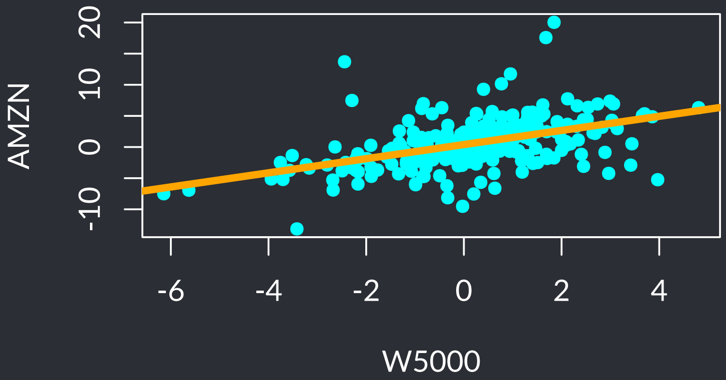
```
hist(stock.market$W5000, col="green", main="",
  xlab="W5000 return as a % of previous week close")
```



# Amazon (AMZN)

```
plot(AMZN ~ W5000, data=stock.market,
  pch=16, col='cyan')
```






The regression line is

$$\widehat{\text{AMZN}} = 0.4 + 1.13 \cdot \text{W5000},$$

with $R^2 = 0.22$.

# Interpreting the regression statistics

- $\hat{\beta}_1 = 1.13$ ("$\beta$") is the predicted increase in returns for AMZN when W5000 returns increase by 1 percentage point—since this is $> 1$ AMZN will swing more than the market as a whole

## Interpreting the regression statistics

- $\hat{\beta}_1 = 1.13$ ("$\beta$") is the predicted increase in returns for AMZN when W5000 returns increase by 1 percentage point—since this is $> 1$ AMZN will swing more than the market as a whole
- $R^2 = 0.22$ indicates how closely AMZN tracks W5000 (the market as a whole)

# Simple regression assumptions

We need four things to be true for a regression model to be a good fit for the data:

1. Both $X$ and $Y$ are quantitative
2. $X$ and $Y$ are approximately linearly related
3. There are no extreme outliers
4. The variance of $Y$ is the same for any value of $X$ ("homoscedasticity")

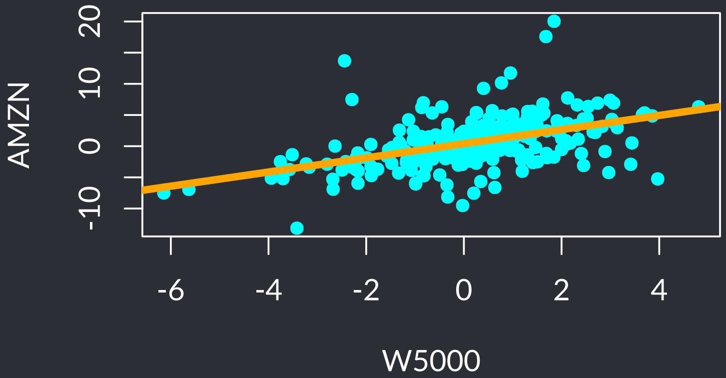
# Simple regression assumptions

We need four things to be true for a regression model to be a good fit for the data:

1. Both $X$ and $Y$ are quantitative $\checkmark$
2. $X$ and $Y$ are approximately linearly related
3. There are no extreme outliers
4. The variance of $Y$ is the same for any value of $X$ ("homoscedasticity")

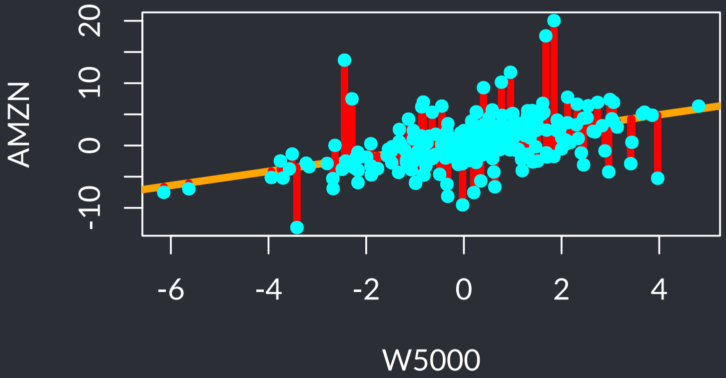
# Assumption 2: Linearity

Step 1: Visually examine to ensure a line is a good fit for the data:



# Assumption 2: Linearity

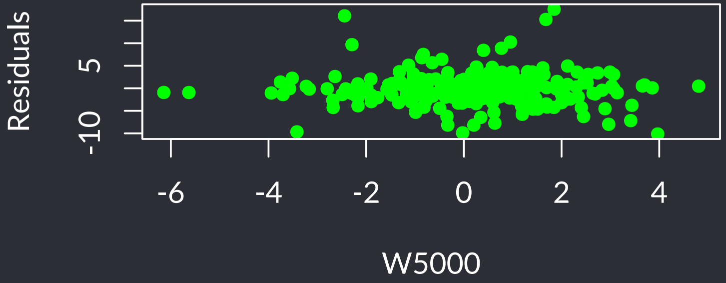
Each point has a **residual** ($Y - \hat{Y}$); this is the over/under-prediction of the model (red lines).



## Assumption 2: Linearity

A **residual plot** (of residuals vs $X$) helps us ensure that there is not subtle nonlinearity. We want to see **no trend** in this plot:

```
model <- lm(AMZN ~ W5000, data=stock.market)
plot(stock.market$W5000, resid(model),
  pch=16, col="green", xlab='W5000', ylab='Residuals')
```



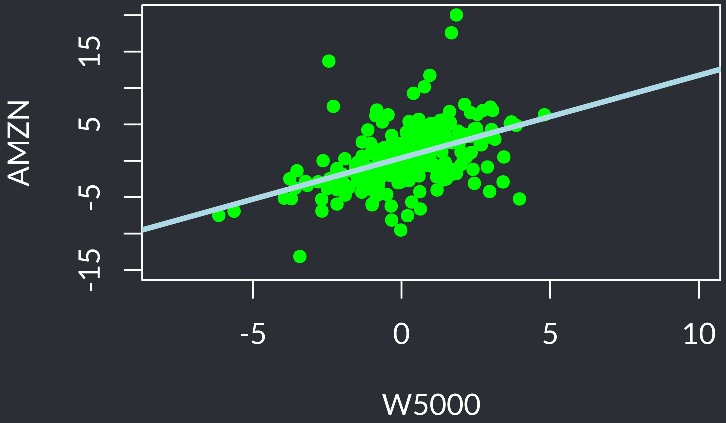
# Simple regression assumptions

We need four things to be true for a regression model to be a good fit for the data:

1. Both $X$ and $Y$ are quantitative $\checkmark$
2. $X$ and $Y$ are approximately linearly related $\checkmark$
3. There are no extreme outliers
4. The variance of $Y$ is the same for any value of $X$ ("homoscedasticity")

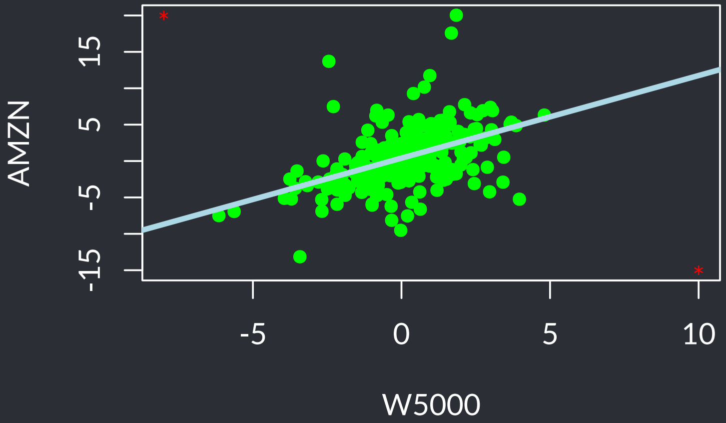
## Assumption 3: There are no extreme outliers

The outliers that would be problematic are those that are *deviate from the existing relationship* between $X$ and $Y$ and are far from the mean on $X$. These would are called influential points:



# Assumption 3: There are no extreme outliers

The outliers that would be problematic are those that are *deviate from the existing relationship* between $X$ and $Y$ and are far from the mean on $X$. These would are called influential points:



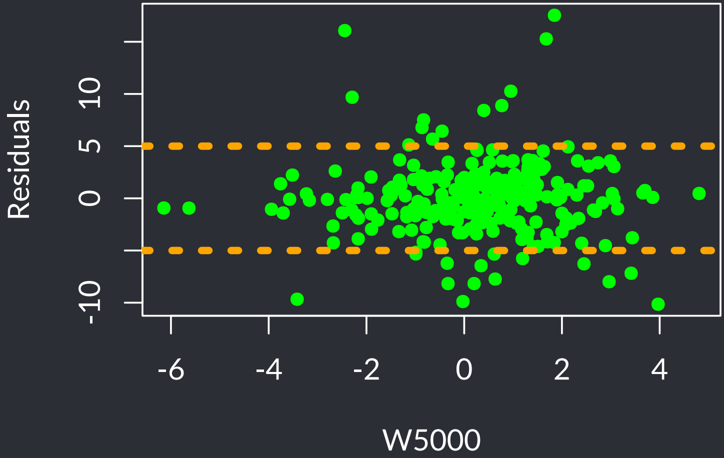
# Simple regression assumptions

We need four things to be true for a regression model to be a good fit for the data:

1. Both $X$ and $Y$ are quantitative ✓
2. $X$ and $Y$ are approximately linearly related ✓
3. There are no extreme outliers ✓
4. The variance of $Y$ is the same for any value of $X$ ("homoscedasticity")

# Assumption 4: Homoscedasticity

Look for the residual plot to have roughly equal vertical spread all the way across:



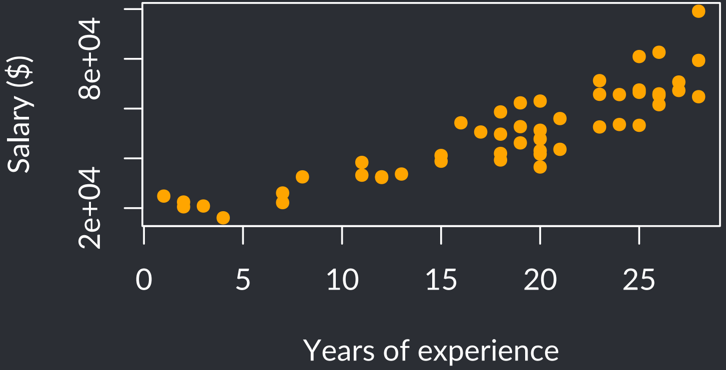
# Simple regression assumptions

We need four things to be true for a regression model to be a good fit for the data:

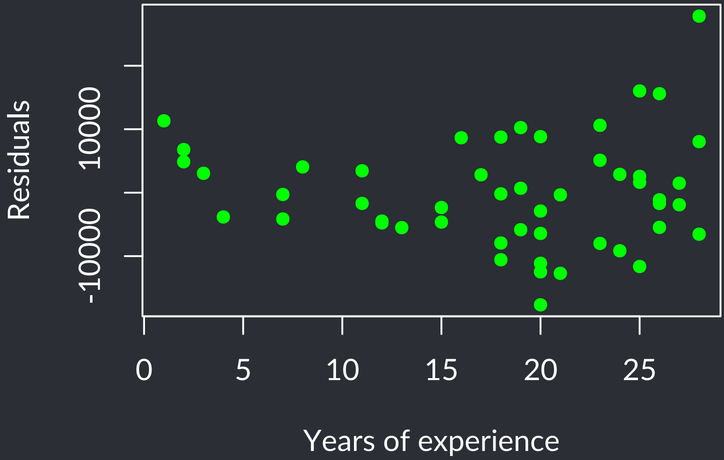
1. Both $X$ and $Y$ are quantitative ✓
2. $X$ and $Y$ are approximately linearly related ✓
3. There are no extreme outliers ✓
4. The variance of $Y$ is the same for any value of $X$ ("homoscedasticity") ✓

# An example where an assumption fails

This is a data set of social worker salaries based on years of experience. Which assumption might be violated here?



# An example where an assumption fails



1. Regression basics
2. Regression assumptions
3. Regression to the mean

EXCLUSIVE | THE ATHLETICISM OF LEBRON + THE DRIVE OF KOBE = RUSSELL WESTBROOK
BY LEE JENKINS P. 38
Sports Illustrated
APRIL 6, 2015
SI.COM
@SINOW
FROM THE BRINK. TO THE BRINK.
KENTUCKY
CLOSES IN ON...
40-0
Karl-Anthony Towns leads Kentucky's quest for college hoops' first undefeated season since 1976.


Techno-Jib Rentals
America's
got Talent


KAPLAN
SAT® Prep

## What is regression to the mean?

- Whenever two variables are imperfectly related (e.g. first SAT attempt vs second SAT attempt), regression to the mean will occur

## What is regression to the mean?

- Whenever two variables are imperfectly related (e.g. first SAT attempt vs second SAT attempt), regression to the mean will occur
- Extreme cases on one variable (very high or very low scores) will likely be not *quite* so extreme on the other variable, since the good (or bad) luck in the extreme case is not likely to be present the next time around:

# What is regression to the mean?

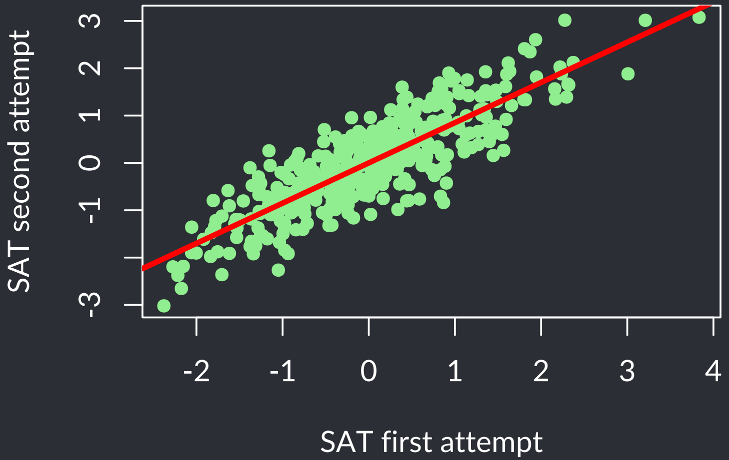
- Whenever two variables are imperfectly related (e.g. first SAT attempt vs second SAT attempt), regression to the mean will occur
- Extreme cases on one variable (very high or very low scores) will likely be not *quite* so extreme on the other variable, since the good (or bad) luck in the extreme case is not likely to be present the next time around:
  - If you got a very high SAT score the first time, you probably will score high again, but you probably won't be as lucky the next time around

## What is regression to the mean?

- Whenever two variables are imperfectly related (e.g. first SAT attempt vs second SAT attempt), regression to the mean will occur
- Extreme cases on one variable (very high or very low scores) will likely be not *quite* so extreme on the other variable, since the good (or bad) luck in the extreme case is not likely to be present the next time around:
  - If you got a very high SAT score the first time, you probably will score high again, but you probably won't be as lucky the next time around
  - If you got a very low SAT score the first time, you probably will score poorly again, but you probably won't be as unlucky the next time around

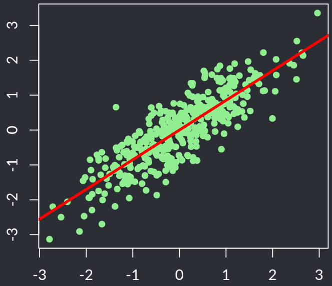
## What does it have to do with regression?

Suppose the correlation between first and second SAT attempts is $r = 0.85$, and scores have been standardized, so $\hat{Y} = 0.85X$:



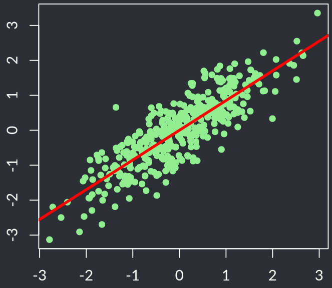
## What does it have to do with regression?

- $X = 2$ (first attempt 2 SD above mean) $\longrightarrow \hat{Y} = 1.7$ (second attempt 1.7 SD above mean)



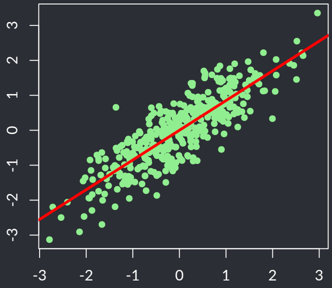
# What does it have to do with regression?

- $X = 2$ (first attempt 2 SD above mean) $\longrightarrow \hat{Y} = 1.7$ (second attempt 1.7 SD above mean)
- $X = 0 \longrightarrow \hat{Y} = 0$



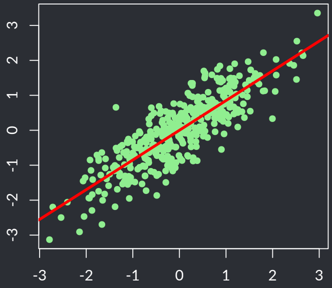
# What does it have to do with regression?

- $X = 2$ (first attempt 2 SD above mean) $\longrightarrow \hat{Y} = 1.7$ (second attempt 1.7 SD above mean)
- $X = 0 \longrightarrow \hat{Y} = 0$
- $X = -2 \longrightarrow \hat{Y} = -1.7$



## What does it have to do with regression?

- $X = 2$ (first attempt 2 SD above mean) $\longrightarrow \hat{Y} = 1.7$ (second attempt 1.7 SD above mean)
- $X = 0 \longrightarrow \hat{Y} = 0$
- $X = -2 \longrightarrow \hat{Y} = -1.7$
- In other words, the second attempts will tend to "regress" towards the mean



## What does it have to do with regression?

- $X = 2$ (first attempt 2 SD above mean) $\longrightarrow \hat{Y} = 1.7$ (second attempt 1.7 SD above mean)
- $X = 0 \longrightarrow \hat{Y} = 0$
- $X = -2 \longrightarrow \hat{Y} = -1.7$
- In other words, the second attempts will tend to "regress" towards the mean
- This has nothing to do with the SAT in particular—it's a statistical phenomenon!