THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Simple regression

Lecture 2

STA 371G

About the course staff

- Instructor: **Brian Lukoff, Ph.D.**
  - Office hours: M/W 11 AM-12 PM in CBA 3.440
  - Contact: `brian.lukoff@utexas.edu` or 415-652-8853
- TAs:
  - Office hours: M 11:30 AM-1:30 PM, 2-4 PM, W 12-2 PM, Th 4-6 PM in CBA 4.304
  - Help session: T 6-7 PM in the ModLab

Vasko Lalkov

Nicole Chia

Zameer Vaswani

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 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 had their first drink?
- How good is that prediction?

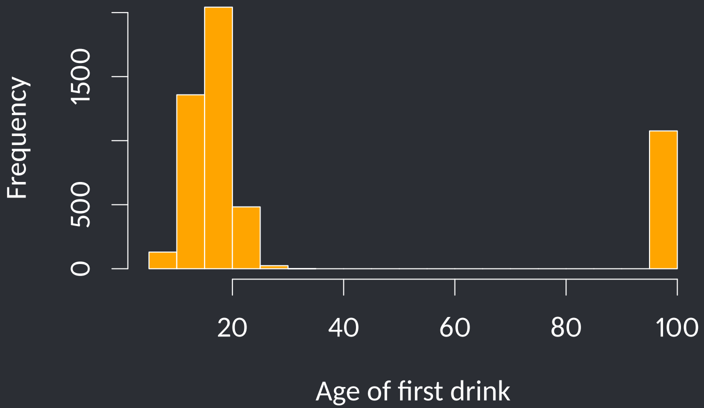
Do people that start drinking younger tend to drink more (or less) when they become adults?

We want to know:

- What is our best **prediction** of alcohol consumption if we know at what age 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	**Predictor variable**
Number of drinks consumed as adult	**Response variable**

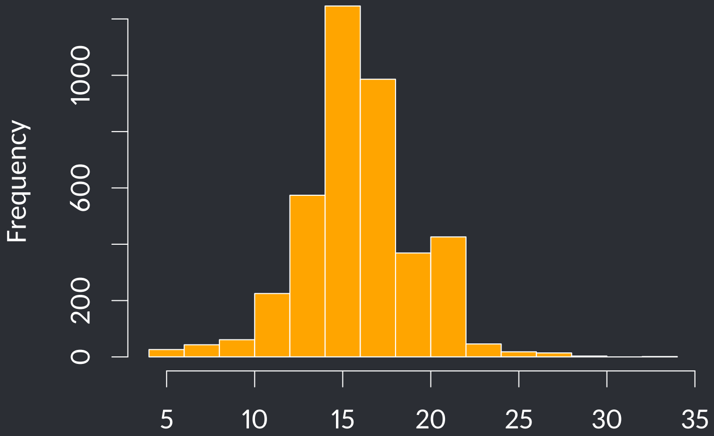
```
> hist(addhealth$h4to34,
+   main='', xlab='Age of first drink',
+   col='orange')
```



Let’s examine our variables

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. 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

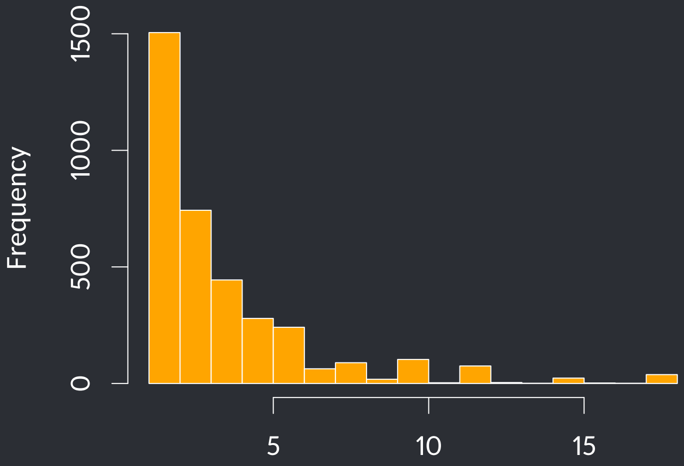
```
> age <- addhealth$h4to34
> age[age >= 96] <- NA
> hist(age, main='', xlab='', col='orange')
```



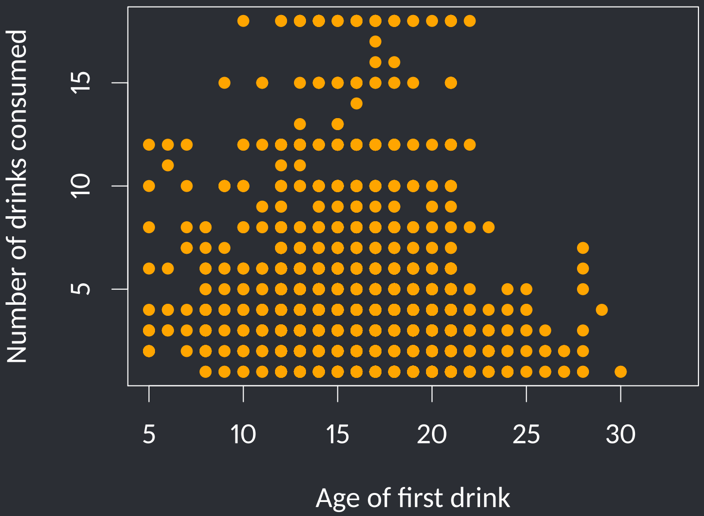
Let's examine our variables

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. 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

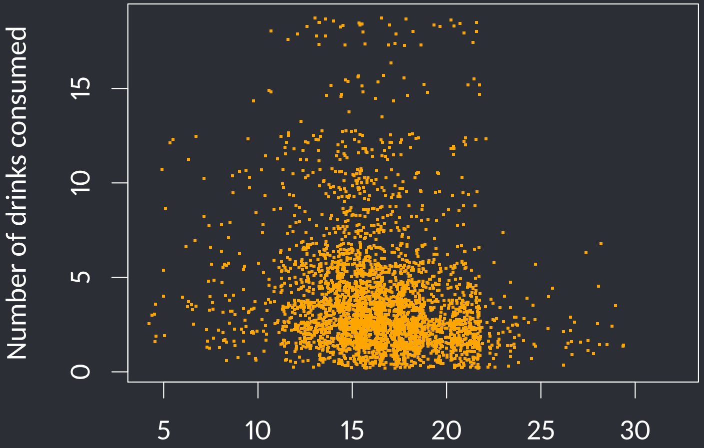
```
> num.drinks <- addhealth$h4to36
> num.drinks[num.drinks >= 96] <- NA
> hist(num.drinks, main='', xlab='How many drinks',
+   col='orange')
```



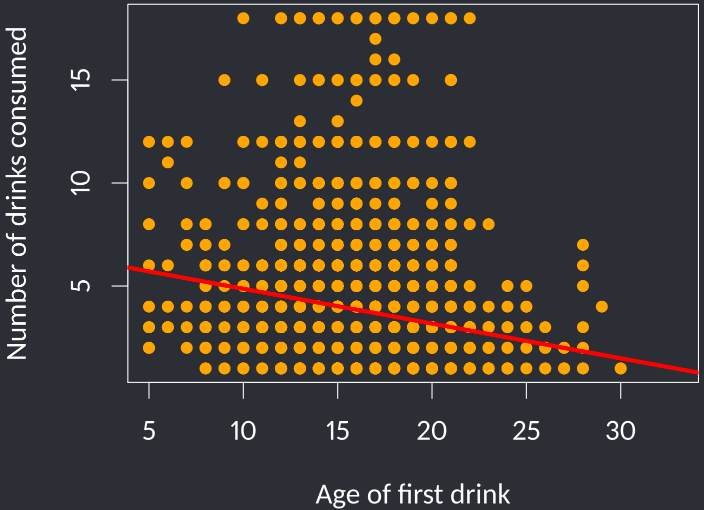
```
> plot(num.drinks ~ age, pch=16, col='orange',
+   xlab='Age of first drink',
+   ylab='Number of drinks consumed')
```



```
> plot(jitter(num.drinks, 4) ~ jitter(age, 4),
+   pch=46, 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 ith person, y_i = number of drinks for ith person) as a linear relationship plus some error:

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

β_0 and β_1 are the intercept and slope, respectively.

What is linear regression doing?

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

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

β_0 and β_1 are the intercept and slope, respectively.

We find estimates for β_0 and β_1 in our sample that *minimize* the errors:

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

This is the regression (best fit) line.

```
> model <- lm(num.drinks ~ age)
> summary(model)


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

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:

```
> 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 X.

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 X.
- $R^2 = \mathrm{cor}(X, Y)^2$, 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 X.
- $R^2 = \text{cor}(X, Y)^2$, 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 X.
- $R^2 = \text{cor}(X, Y)^2$, 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 X.
- $R^2 = \text{cor}(X, Y)^2$, 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)
> summary(model)


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

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 $r = \sqrt{0.03} = 0.17$.

Is this "significant?"

In our regression, $R^2 = 0.03$, so $r = \sqrt{0.03} = 0.17$.

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