THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Simple Regression 1

Lecture 5

STA 371G

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** 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:

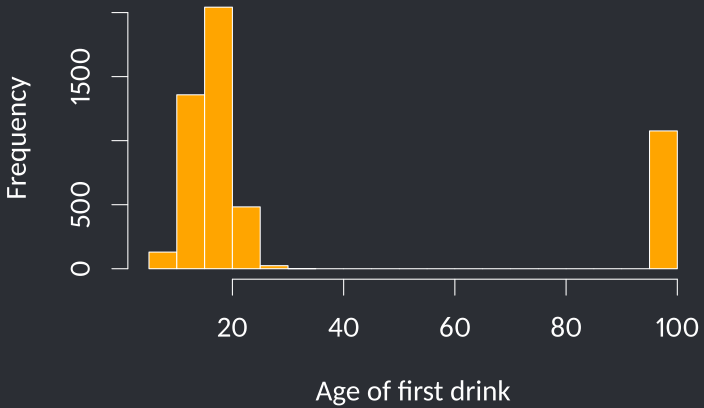
- What is our best **prediction** alcohol consumption if we know at what age had their first drink?
- How good is that prediction?

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** 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**

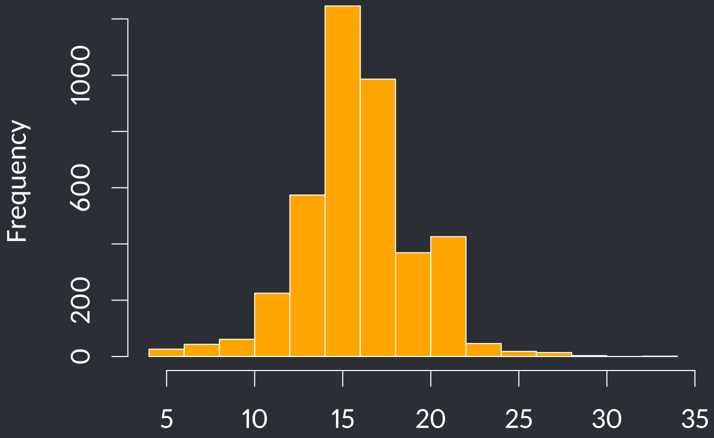
```r
> hist(addhealth_public4$h4to34,
+   main='', xlab='Age of first drink',
+   col='orange')
```



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

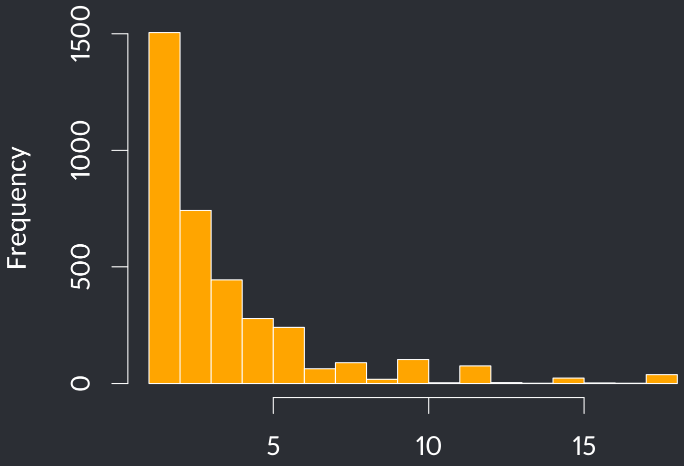
```r
> age <- addhealth_public4$h4to34
> age[age >= 96] <- NA
> hist(age, main='', xlab='', col='orange')
```



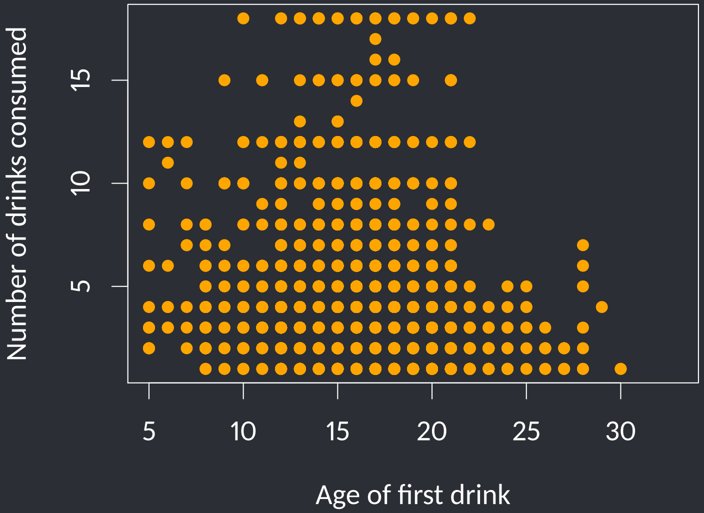
Let's examine our variables

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

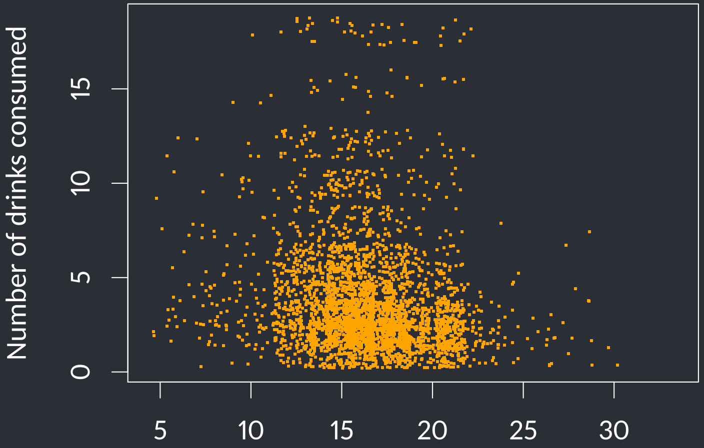
```
> num.drinks <- addhealth_public4$h4to36
> num.drinks[num.drinks >= 96] <- NA
> hist(num.drinks, main='', xlab='How many drinks',
+   col='orange')
```



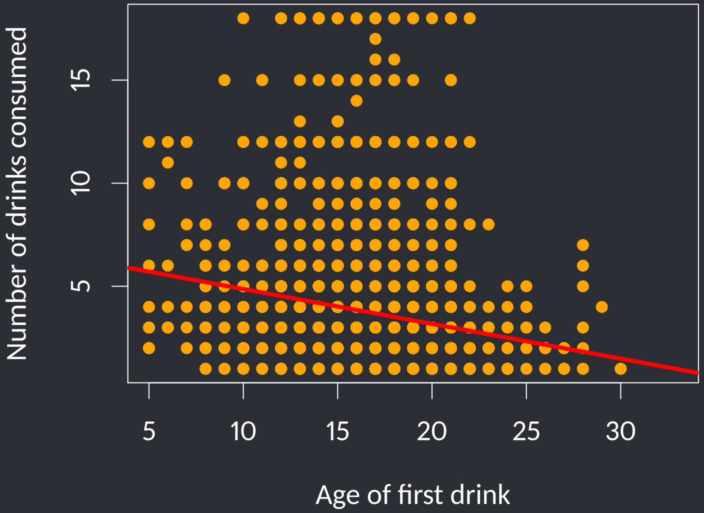
```r
> plot(num.drinks ~ age, pch=16, col='orange',
+   xlab='Age of first drink',
+   ylab='Number of drinks consumed')
```



```
> 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.lm(model, list(age=21))
```

How good are our predictions?

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

- $R^2 = \text{cor}(X, Y)^2$, i.e., the squared correlation between X and Y.

How good are our predictions?

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

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

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

- $R^2 = \text{cor}(X, Y)^2$, 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?"

- **Statistical significance:** Can we reject the null hypothesis that the correlation between X and Y in the *population* is not zero?

In our regression, $R^2 = 0.03$, so $r = \sqrt{0.03} = 0.17$.
Is this "significant?"

- **Statistical significance:** Can we reject the null hypothesis that the correlation between X and Y in the *population* is not zero?
- **Practical significance:** Is the correlation in our sample large enough to be meaningful?

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)
- $\mathrm{cor}(X, Y) = 0$ (in the population)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)
- $\text{cor}(X, Y) = 0$ (in the population)
- $\beta_1 = 0$

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)
- $\text{cor}(X, Y) = 0$ (in the population)
- $\beta_1 = 0$
- The model has no predictive power

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)
- $\text{cor}(X, Y) = 0$ (in the population)
- $\beta_1 = 0$
- The model has no predictive power
- Predictions from this model are no better than predicting $\overline{Y}$ for every case

Two ways to test the overall null hypothesis

- The F-test (tests $H_0 : R^2 = 0$ in the population)
- The t-test for the *slope* (β_1) coefficient (tests $H_0 : \beta_1 = 0$)

Two ways to test the overall null hypothesis

- The F-test (tests $H_0 : R^2 = 0$ in the population)
- The t-test for the *slope* (β_1) coefficient (tests $H_0 : \beta_1 = 0$)

Both of these methods are equivalent; the p-values will be exactly the same!

```
> 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
```

What is our conclusion about β_1?

- There is a **statistically significant** relationship between the age someone starts drinking and how much they drink as an adult.

What is our conclusion about β_1?

- There is a **statistically significant** relationship between the age someone starts drinking and how much they drink as an adult.
- Or: People that start drinking earlier in life consume **significantly more** alcohol when they drink as adults.

What is our conclusion about β_1?

- There is a **statistically significant** relationship between the age someone starts drinking and how much they drink as an adult.
- Or: People that start drinking earlier in life consume **significantly more** alcohol when they drink as adults.
- Each additional year you wait to start drinking is associated with consuming 0.17 fewer drinks as an adult.

What is our conclusion about β_1?

- There is a **statistically significant** relationship between the age someone starts drinking and how much they drink as an adult.
- Or: People that start drinking earlier in life consume **significantly more** alcohol when they drink as adults.
- Each additional year you wait to start drinking is associated with consuming 0.17 fewer drinks as an adult.
- Is this relationship **practically significant**?

Put a confidence interval on it

- Our best estimate for the *effect* of a year's postponement of drinking is 0.17 fewer drinks as an adult

Put a confidence interval on it

- Our best estimate for the *effect* of a year's postponement of drinking is 0.17 fewer drinks as an adult
- We can use a confidence interval to give a range of plausible values for what this effect size is in the population

Put a confidence interval on it

A confidence interval is always of the form

$$\text{estimate} \pm (\text{critical value})(\text{standard error}).$$

Put a confidence interval on it

A confidence interval is always of the form

$$\text{estimate} \pm (\text{critical value})(\text{standard error}).$$

Recall that the critical value for a 95% confidence interval is the cutoff value that cuts off 95% of the area in the middle of the distribution; the sampling distribution of $\hat{\beta}_1$ is a t-distribution.

```
> n <- nobs(model)
> qt(0.975, n-2)

[1] 1.960623
```

Put a confidence interval on it

R will also calculate confidence intervals for us:

```
> confint(model)

                2.5 %     97.5 %
(Intercept)  6.0339847  7.0743549
age         -0.1999713 -0.1376959
```

Put a confidence interval on it

R will also calculate confidence intervals for us:

```
> confint(model)

                 2.5 %     97.5 %
(Intercept)  6.0339847  7.0743549
age         -0.1999713 -0.1376959
```

In other words, we are 95% confident that the effect of each additional year's delay in starting to drink is between 0.14 and 0.2.

Put a confidence interval on it, part 2

We can also put a confidence interval on a prediction!

Two kinds of intervals:

Confidence	Predicting the mean value of Y for a particular X.	Among all people that start drinking at age 21, how many drinks do have on average as adults?
Prediction	Predicting Y for a single new case.	If Bob started drinking at age 21, how many drinks do we think will have as an adult?

Put a confidence interval on it, part 2

```
> predict.lm(model, list(age=21),
+   interval='confidence')

       fit      lwr      upr
1 3.008664 2.83616 3.181167

> predict.lm(model, list(age=21),
+   interval='prediction')

       fit       lwr      upr
1 3.008664 -2.802894 8.820221
```

Put a confidence interval on it, part 2

```
> predict.lm(model, list(age=21),
+   interval='confidence')

       fit      lwr       upr
1 3.008664 2.83616 3.181167

> predict.lm(model, list(age=21),
+   interval='prediction')

       fit       lwr      upr
1 3.008664 -2.802894 8.820221
```

Why is the prediction interval wider?