Linear Regression

Today:

- Understand linear regression as a statistical model;
- Introduce an example dataset;
- Build the ability to interpret linear regression results.

What is linear regression?

- **Linear regression** is a method that answers three questions simultaneously:
  1. What is the effect of one variable x upon another variable y?
  2. How sure are we that x affects y?
  3. Do the answers to the first two questions depend on other independent variables?

- Typically used to model a dependent variable that is **continuous** (so interval or ratio data) and **unbounded** (so has no theoretical max or min);

- The essence of the linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

What is linear regression?

- **Linear regression** is a method that answers three questions simultaneously:
  1. What is the effect of one variable x upon another variable y?
  2. How sure are we that x affects y?
  3. Do the answers to the first two questions depend on other independent variables?

- Typically used to model a dependent variable that is **continuous** (so interval or ratio data) and **unbounded** (so has no theoretical max or min);

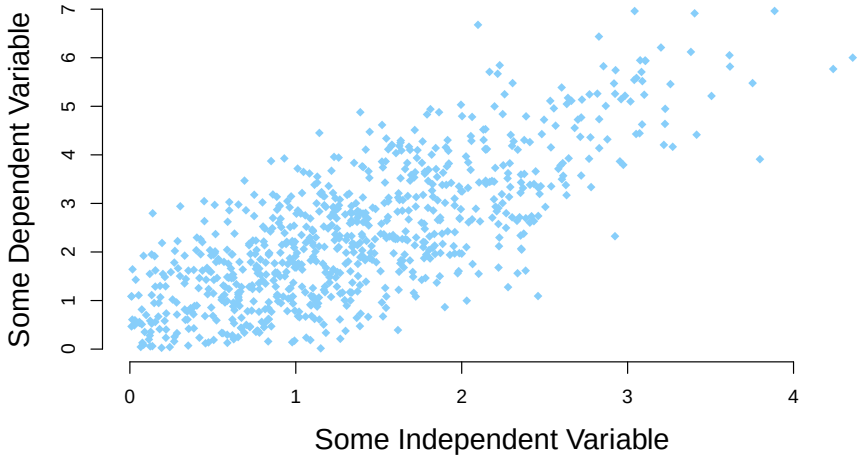
- The essence of the linear regression model is:

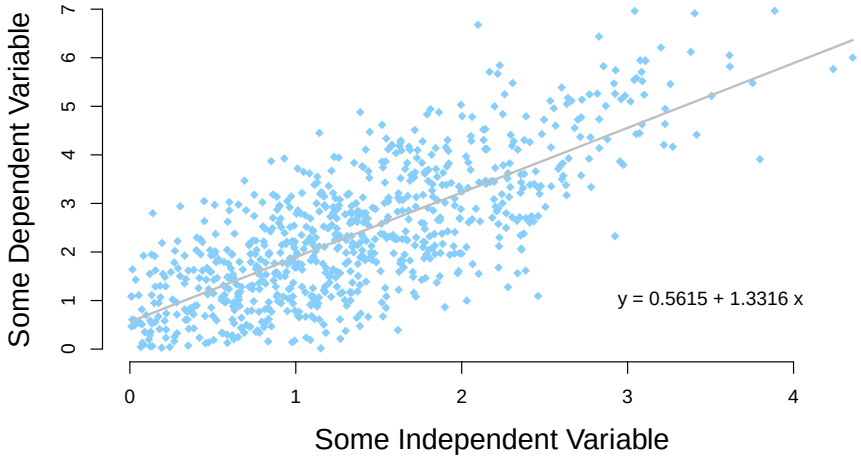
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

What is linear regression?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

- y is the **dependent variable** – this is part of the data set;
- The x's are **independent variables** that explain y – also part of the data set;
- The β's are called **coefficient effects** that control how the x's affect y – they are **learned** from the data;
- Finally the ε is a **noise** term that models the fact that y may also depend on random stuff (we don't observe this but use it to learn the β's – that'll be for next time).


Some Dependent Variable
0
1
2
3
4
5
6
7
0
1
2
3
4
Some Independent Variable


Some Dependent Variable
0
1
2
3
4
5
6
7
y = 0.5615 + 1.3316 x
0
1
2
3
4
Some Independent Variable

An Example: the US Census American Community Survey, 2012.

income	hrs	race	age	gender	cmte	lang	married	edu	disability
1700	40	other	35	female	15	other	yes	hs or lower	yes
45000	84	white	27	male	40	english	yes	hs or lower	no
8600	23	white	69	female	5	english	no	hs or lower	no
33500	55	white	52	male	20	english	yes	hs or lower	no
4000	8	white	67	female	10	english	yes	hs or lower	no
19000	35	white	36	female	15	english	yes	college	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

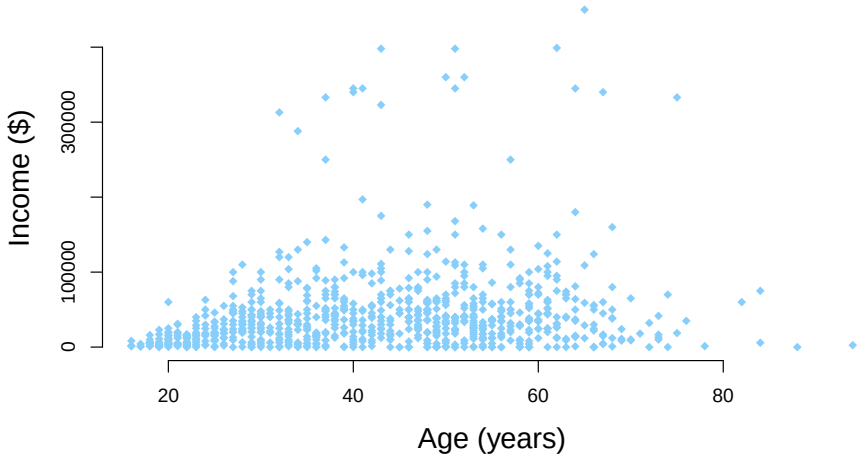
Question: do people earn more money as they get older?

An Example: the US Census American Community Survey, 2012.

income	hrs	race	**age**	gender	cmte	lang	married	edu	disability
1700	40	other	35	female	15	other	yes	hs or lower	yes
45000	84	white	27	male	40	english	yes	hs or lower	no
8600	23	white	69	female	5	english	no	hs or lower	no
33500	55	white	52	male	20	english	yes	hs or lower	no
4000	8	white	67	female	10	english	yes	hs or lower	no
19000	35	white	36	female	15	english	yes	college	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Question: do people earn more money as they get older?

Do people earn more money as they get older?



Do people earn more money as they get older?

- Let's answer this question by building a simple regression model;
  - The dependent variable y will be **income** from the 2012 US Census American Community Survey;
  - The independent variable x will be **age** from the 2012 US Census American Community Survey;
  - We will add an intercept or constant;

- This leads to the regression equation:

$$income = \beta_0 + \beta_{age} * age + \varepsilon;$$

- When we run the linear regression we will learn β_0 and β_{age}.

Do people earn more money as they get older – results!

```
Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)  13141.0     6050.7   2.172       0.0301 *
age            718.8      133.1   5.400 0.0000000868 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Do people earn more money as they get older – results!

```
Call:
lm(formula = income ~ age, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-78205 -27342 -12154  10514 390139

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)  13141.0     6050.7   2.172       0.0301 *
age            718.8      133.1   5.400 0.0000000868 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56190 on 841 degrees of freedom
Multiple R-squared:  0.03351,   Adjusted R-squared:  0.03236
F-statistic: 29.16 on 1 and 841 DF,  p-value: 0.00000008678
```

An Example: the US Census American Community Survey, 2012.

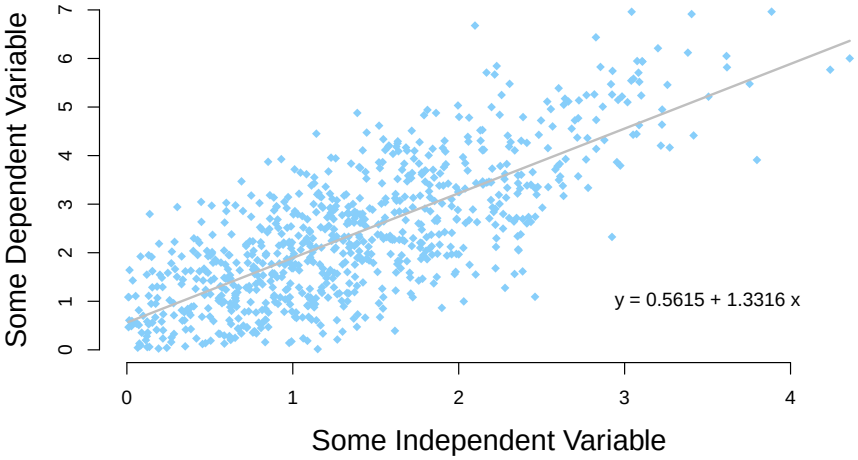
income	**hrs**	race	age	**gender**	**cmte**	lang	married	**edu**	disability
1700	40	other	35	female	15	other	yes	hs or lower	yes
45000	84	white	27	male	40	english	yes	hs or lower	no
8600	23	white	69	female	5	english	no	hs or lower	no
33500	55	white	52	male	20	english	yes	hs or lower	no
4000	8	white	67	female	10	english	yes	hs or lower	no
19000	35	white	36	female	15	english	yes	college	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Question: do people earn more money as they get older?

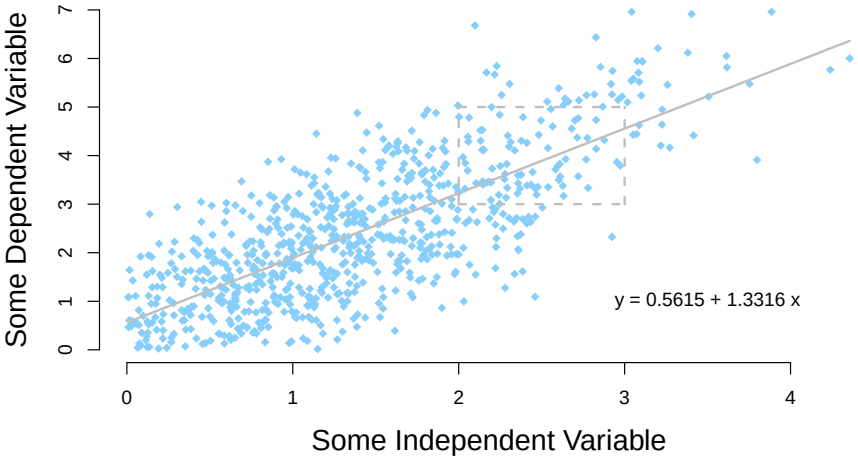
Do people earn more money as they get older – results!

```
Coefficients:
                Estimate Std. Error t value          Pr(>|t|)
(Intercept)    -26945.95    8232.57  -3.273           0.00111 **
age              540.39     122.27   4.420 0.000011286396345 ***
hrs_work        1061.82     149.48   7.103 0.000000000002758 ***
gendermale     19484.80    3688.59   5.282 0.000000165718705 ***
time_to_work      93.06      80.10   1.162           0.24567
edugrad        44734.13    6140.20   7.285 0.000000000000789 ***
eduhs or lower -18519.50    4077.02  -4.542 0.000006446559114 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

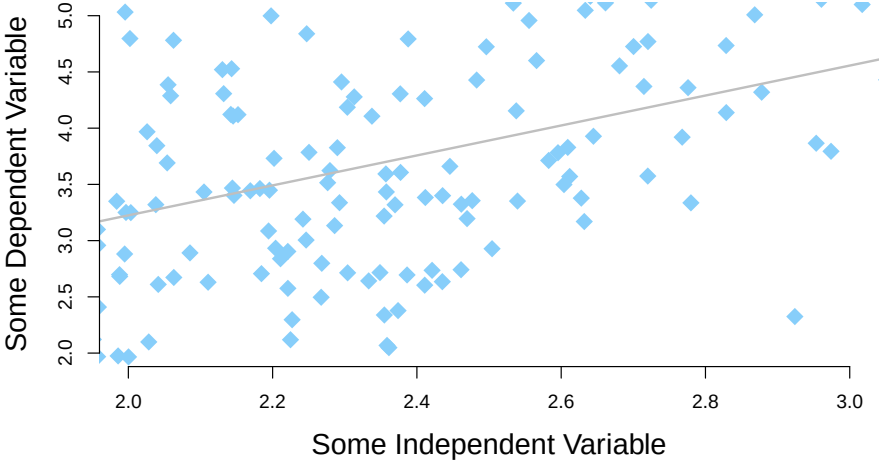
How does linear regression work?



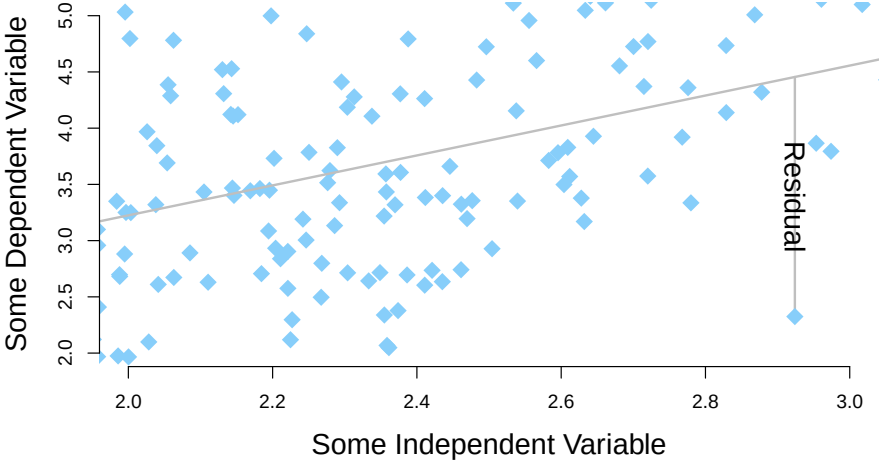
How does linear regression work?



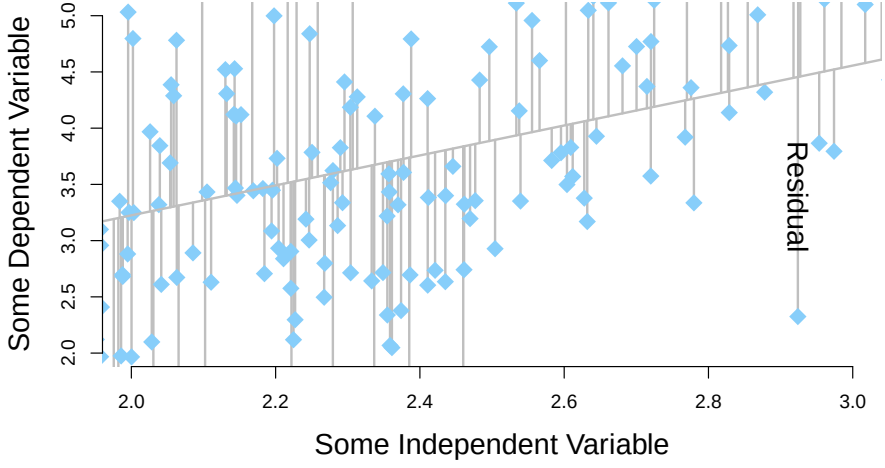
How does linear regression work?



How does linear regression work?



How does linear regression work?



How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...

$$y_i$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...

$$y_i \qquad \beta_0 + \beta_1 x_i$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...
  3. Take the difference between the two (this is the residual)...

$$y_i - (\beta_0 + \beta_1 x_i)$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...
  3. Take the difference between the two (this is the residual)...
  4. Square the difference...

$$(y_i - (\beta_0 + \beta_1 x_i))^2$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...
  3. Take the difference between the two (this is the residual)...
  4. Square the difference...
  5. Add it up for all the observations in the data...

$$\sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$$

How does linear regression work?

- Linear regression finds the 'line of best fit.' What does this mean? It means finding the 'best' βs;
- Linear regression finds the best βs by minimizing the sum of residuals;
- How?
  1. Take the observed dependent variable for an observation...
  2. Given some βs and the independent variables create a prediction...
  3. Take the difference between the two (this is the residual)...
  4. Square the difference...
  5. Add it up for all the observations in the data...
  6. Choose the βs that make this as small as possible...

$$\min_{\beta_0,\beta_1} \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Assumptions of Linear Regression

1. Dependent variable is a **linear** function of independent variables plus noise;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

2. Independent variables have **no measurement error**;

3. Independent variables are not related to each other – **no multicollinearity**;

4. Noise term is a random variable following the **normal distribution**.

Why do we use Linear Regression?

- Theorem (Gauss-Markov) When the assumptions of linear regression are met then it is the Best Linear Unbiased Estimator of the relationship between the dependent and independent variables;
- 'Unbiased' means that it will give you the correct βs on average;
- 'Best' means that it will give you the most precise estimates of those βs possible;
- This is usually called BLUE for short.

Why should we care?

Linear regression combines hypothesis testing and machine learning, and should be a first modeling stop for interval or ratio data.