Simple Linear Regression

Sumanta Basu

Logistics and Reading

Prelim 2 scores posted, HW9 due Wednesday 11/27, Exploration due Monday 12/2

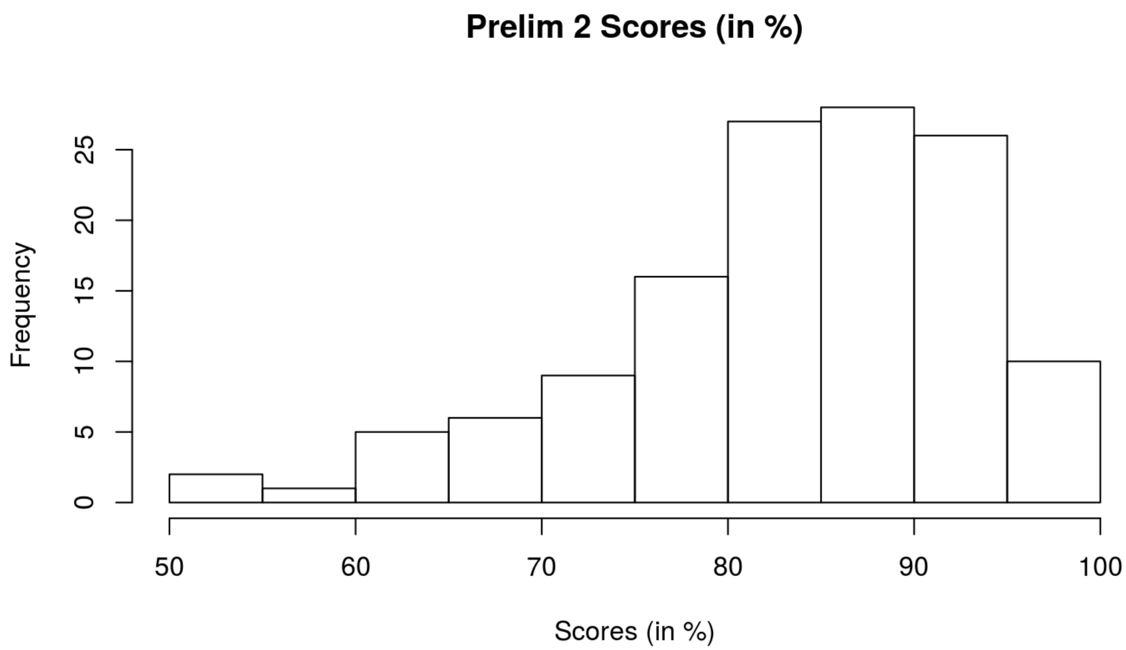
For questions/concerns about scores, email lab TA with a screenshot of the problem

Next week:

- Tuesday: Guest lecture by Francoise Vermeylen (CSCU)
- Thursday: No lecture
- No labs

> Reading: Textbook Sections 7.1, 7.2, 7.4

Prelim 2 Scores



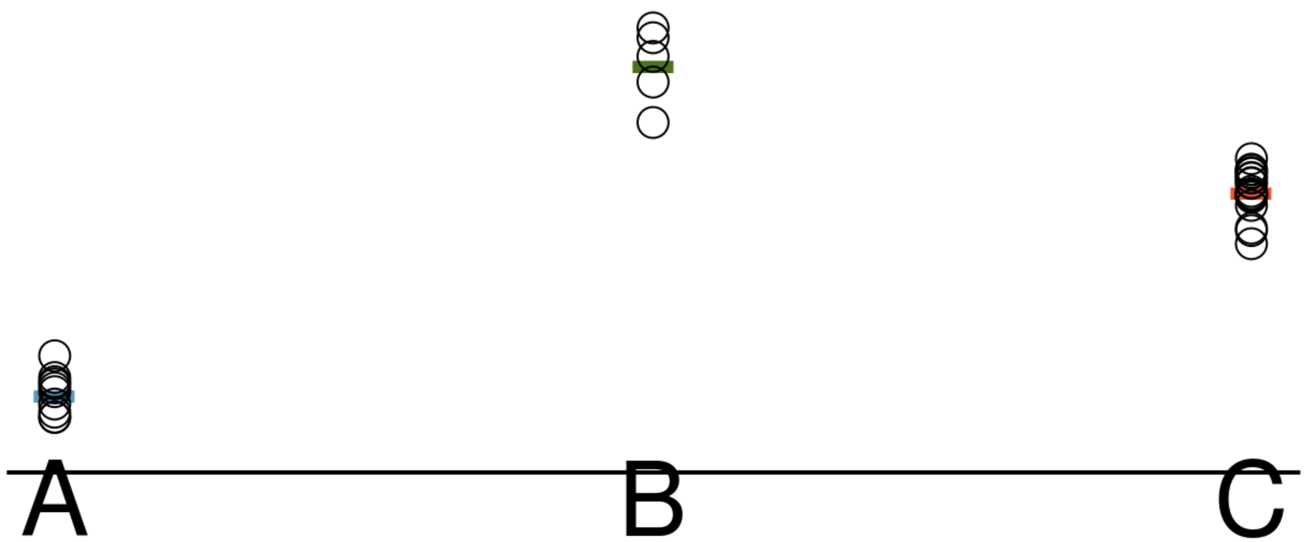

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.67   77.92   85.00   84.01   91.67  100.00
```

Simple Linear Regression & Correlation

Taking a step back…

1. Inference for **one mean**: $\mu = 0$?
2. Inference for **two means**: $\mu_1 = \mu_2$?
3. Inference for **multiple means**: $\mu_1 = \mu_2 = \cdots = \mu_k$?
4. What should come next?

Continuous response across groups



Taking a step back…

1. Inference for **one mean**: $E[Y] = 0$?
2. Inference for **2 means**: *Is continuous Y associated with binary X?*

$$E[Y \mid X = Treatment] = E[Y \mid X = Control]?$$

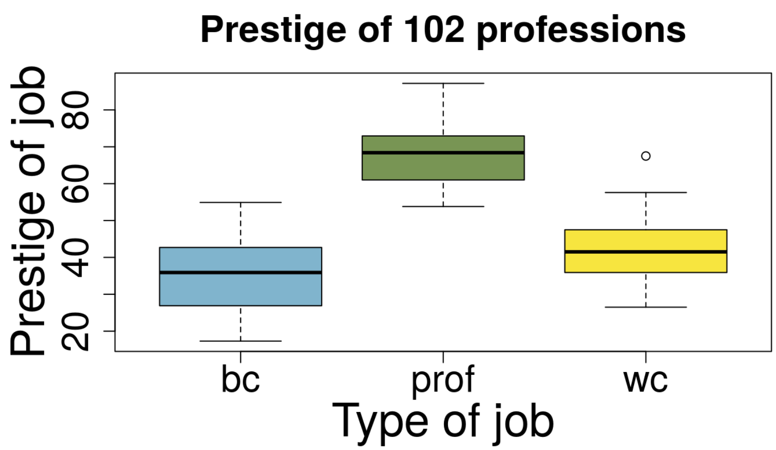
3. Inference for **multiple means**: *Is continuous Y associated with categorical X?*

$$E[Y \mid X = Red] = E[Y \mid X = Blue] = E[Y \mid X = Green]?$$

4. Next: *Is continuous Y associated with* **numerical** *X?*

$$\text{Does } E[Y \mid X = x] \text{ depend on } x?$$

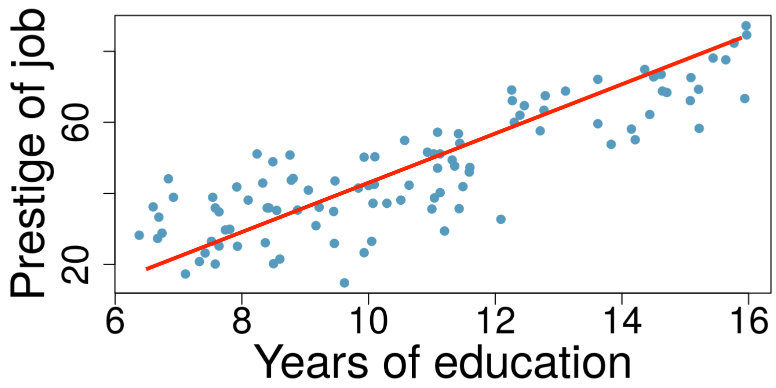
Example: Continuous vs. categorical



- Given a certain job type, what is expected prestige?
- Is job type at all related to prestige?

$$E[Prestige \mid Type = wc]$$

Example: Continuous vs. continous



- Given a certain education level, what is expected prestige?
- Is education at all related to prestige?

$$E[Prestige \mid Educ = 10]$$

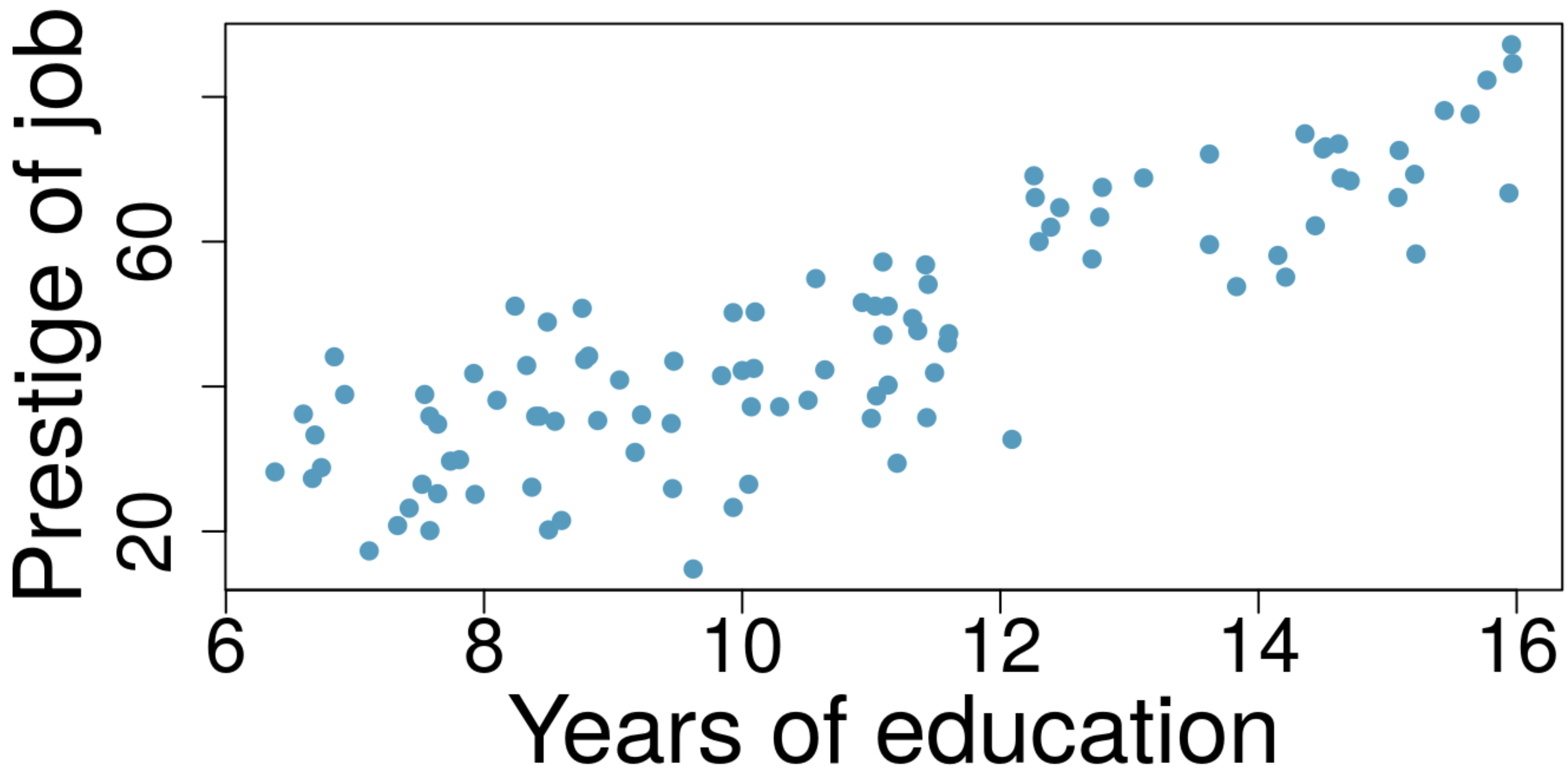
When YoE goes up, prestige goes up.
Approximately linear
Slope 4 to 1
with 1 year of increase YoE average scores goes up by 4 +/-

Average prestiges score associated with 6 YoE is 20 and with every for years of increase the score goes up by 20 units

Baseline: E(Prestige | YoE = 6) =20
With every extra 1 year of education, Average prestige score goes up by 5 units.

Estimation of parameter:
1. Slope parameter - relationship between x and y
2. Prediction - If YoE = 10 prestige = ??
3. Goodness of fit - how sure are you about 1 and 2

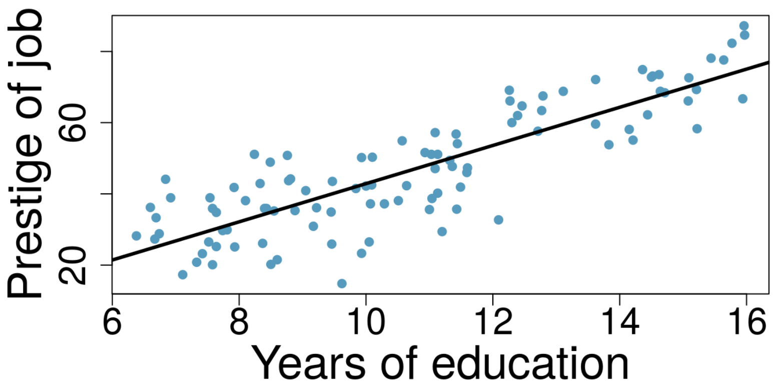
Simplest approach: linear relationship?



Outline1: Modeling Assumptions

- Model conditional mean of Y as a **linear function** of X
- A straight line can be described by two numbers: intercept (β_0) and slope (β_1)
- Every observation Y_i is conditional mean + independent Gaussian noise with equal variance σ^2

Simple model: linear relationship



$$E[Y \mid X = x] = \beta_0 + \beta_1 \cdot x$$

(equation for a straight line)

- intercept β_0
- slope β_1
- here $\beta_0 \approx -10, \beta_1 \approx 5$

Simple linear regression

Assumes mean of Y increases linearly in x:

$$E[Y \mid X = x] = \beta_0 + \beta_1 x.$$

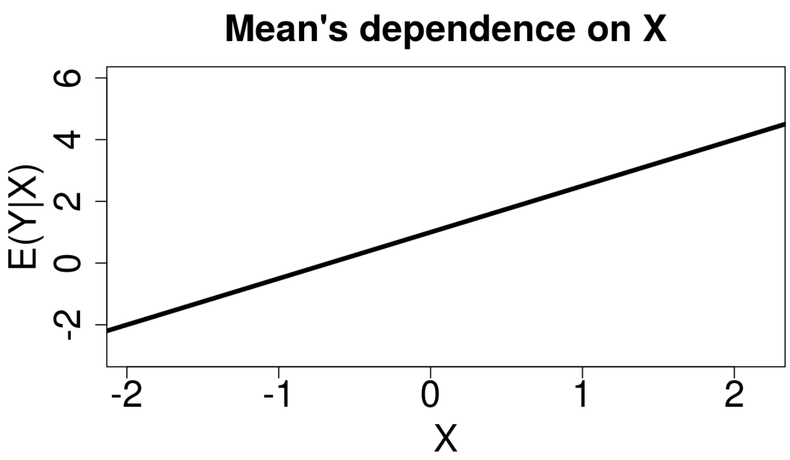
Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma)$ are independent.

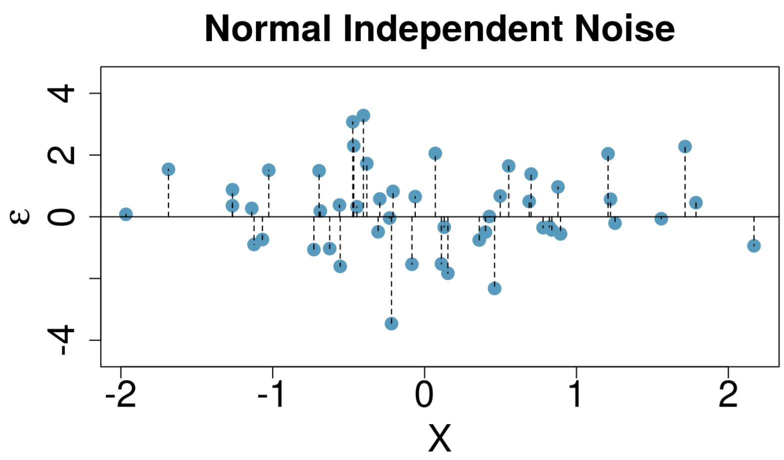
- just like in ANOVA, we assume independent normal noise with fixed variance

In pictures



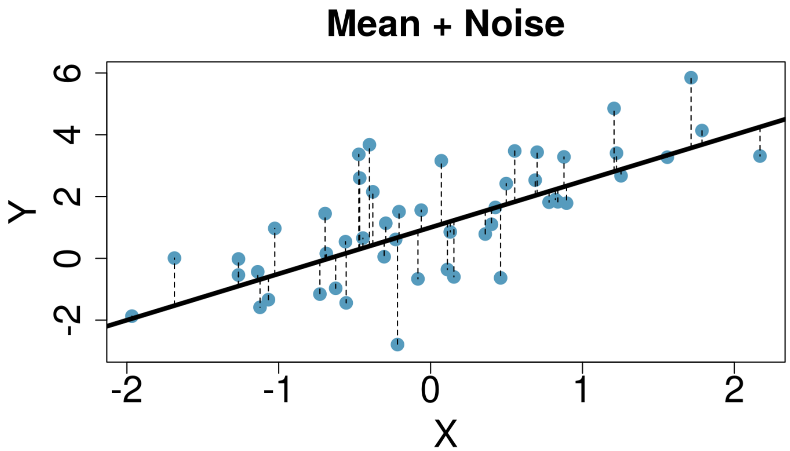
- linear function for Y's mean: $E[Y \mid X = x] = \beta_0 + \beta_1 x$

In pictures



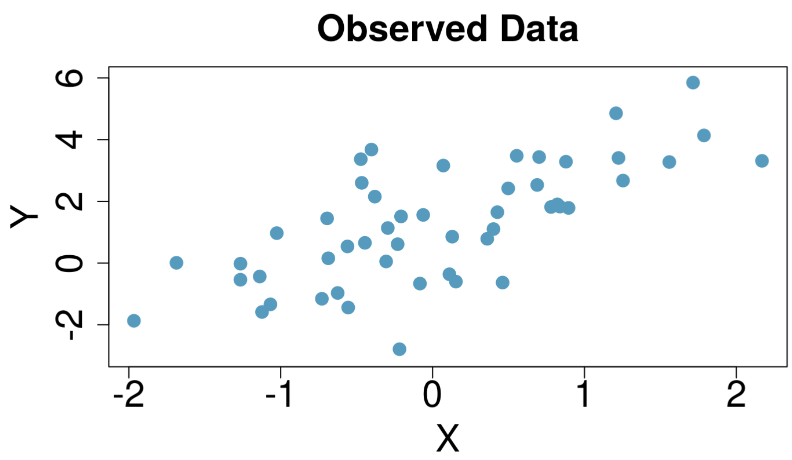
- Noise is independent $\epsilon_i \sim N(0, \sigma)$.
- Variance σ^2 does not depend on X.

In pictures



$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

In pictures



- We observe (X_i, Y_i) pairs.
- How can we estimate β_0 and β_1 from this data?

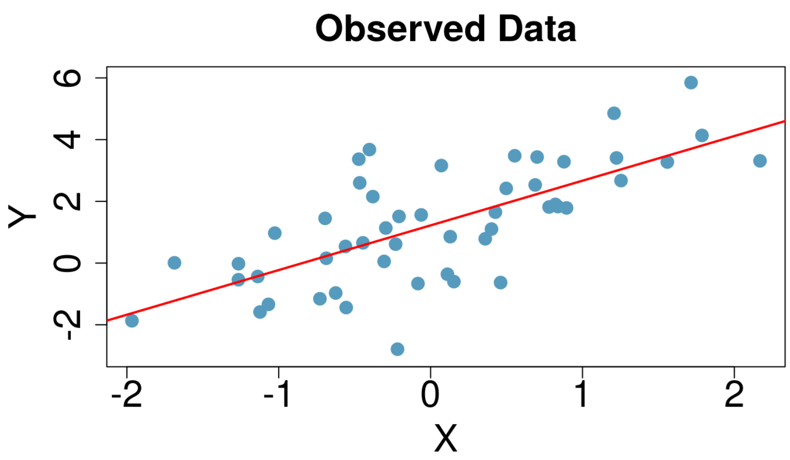
Outline 2: Estimate Parameters

- Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we need to find estimates of β_0, β_1 and σ^2
- Use the **principle of least squares**
- Estimates of slope and intercept are from sample, they have uncertainty associated with them

Remember

minimize (x_1 - c)^2+ (x_2 - c)^2 + ... + (x_n - c)^2 c = mean
minimize |x_1 - c| + |x_2 - c| + ... + |x_n -c| c = median(x)

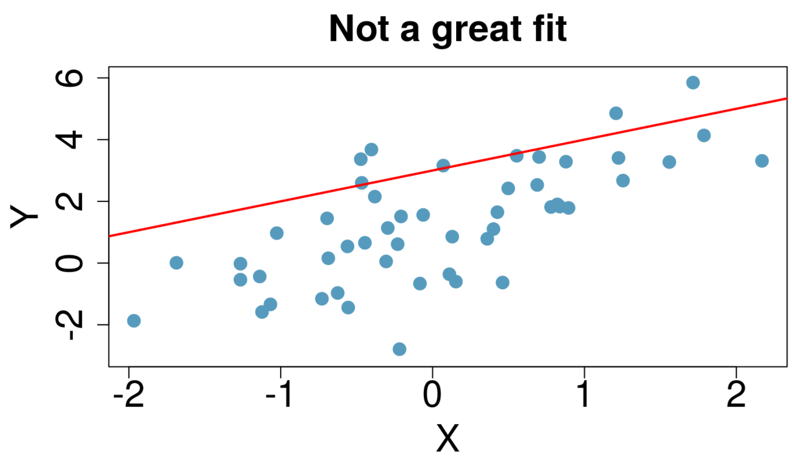
Natural Idea: Best fitting line



Choose a line that “fits the data well”

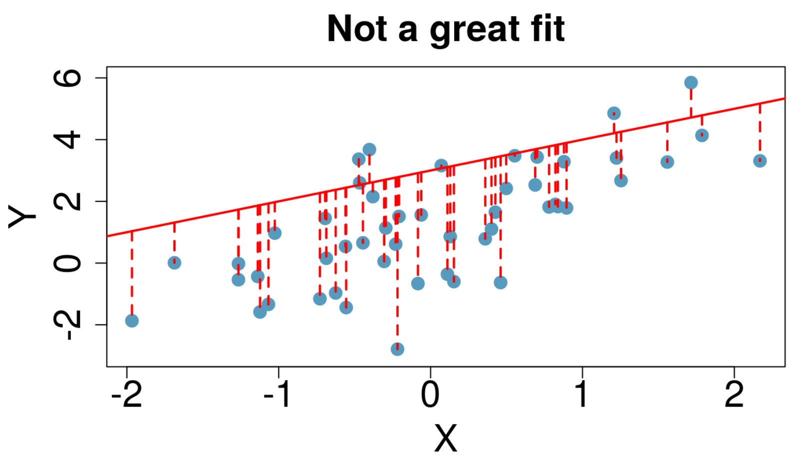
- what do we mean by that?

Natural Idea: Best fitting line



This line doesn't fit well.

Natural Idea: Best fitting line

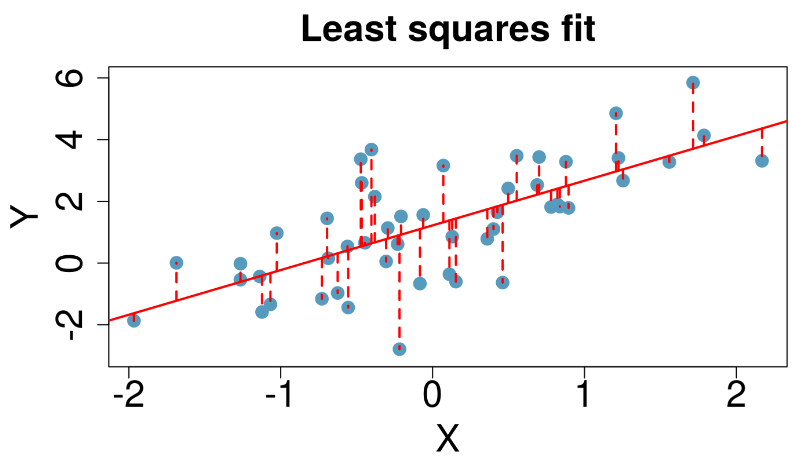


- The "**residuals**" between the line and the data are unnecessarily large.
- Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that makes "*sum of squared errors*"

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

as small as possible.

Natural Idea: Best fitting line



- Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that makes "sum of squared errors"

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

as small as possible.

- This is called the **least squares** line

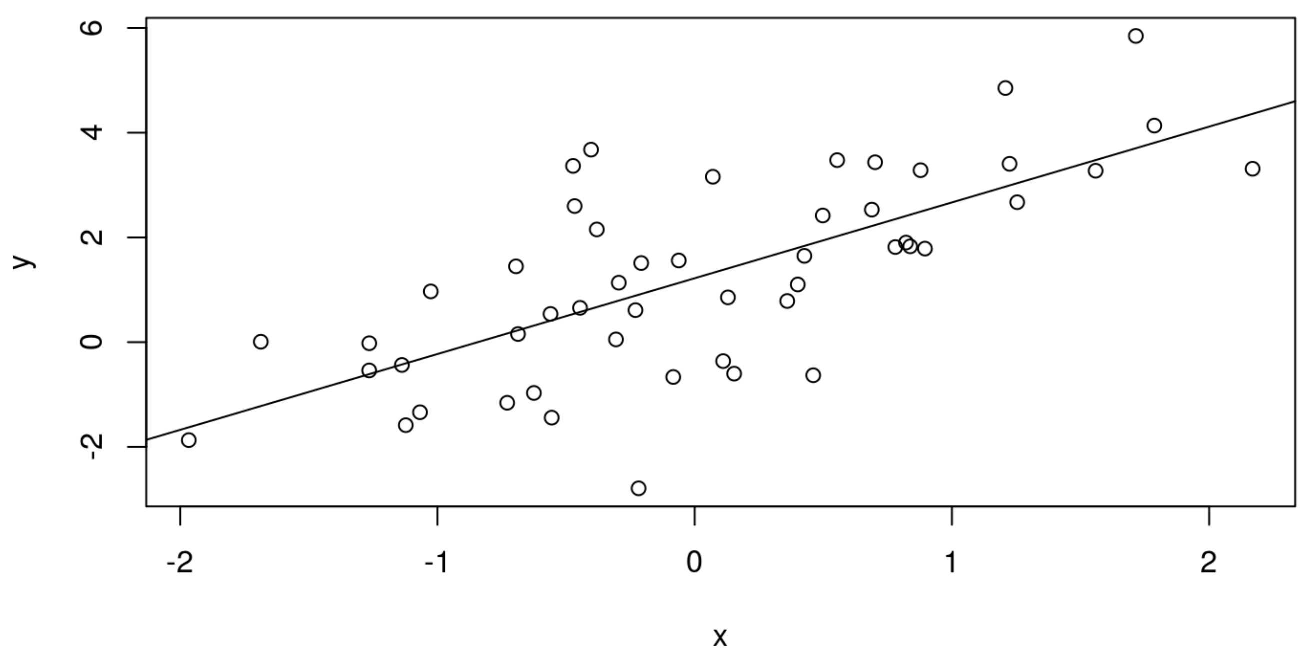
In R...

```r
lm(y ~ x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       1.221        1.447
```

In R...

```
fit = lm(y ~ x)
plot(x, y)
abline(fit)
```



Question

Suppose we get the following result:

```
lm(yy ~ xx)
```

```
##
## Call:
## lm(formula = yy ~ xx)
##
## Coefficients:
## (Intercept)           xx
##   1.174e+00   -4.947e-17
```

- That is, $\hat{\beta}_1 \approx 0$

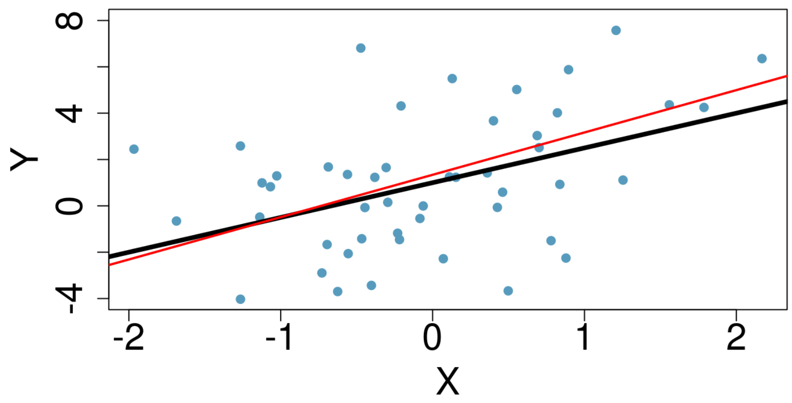
What does this tell us?

> $E[Y \mid X = x]$ has no relationship to x?

No. $E[Y \mid X = x]$ is the population mean of Y when $X = x$.

That'd be analogous to saying that $\mu = 0$ just because $\bar{x}_n = 0$

Least squares line is based on sample



Black line: (population)

$$\mu(x) = \beta_0 + \beta_1 x$$

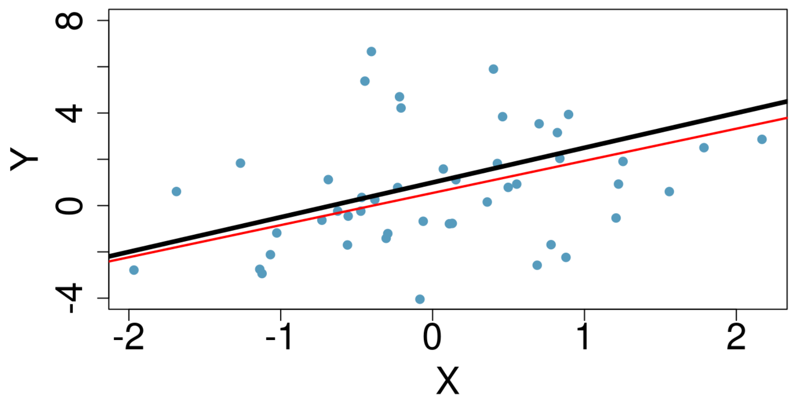
Data points:

$$Y_i = \mu(X_i) + \epsilon_i$$

Red line: (based on data)

$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least squares line is based on sample



Black line: (population)

$$\mu(x) = \beta_0 + \beta_1 x$$

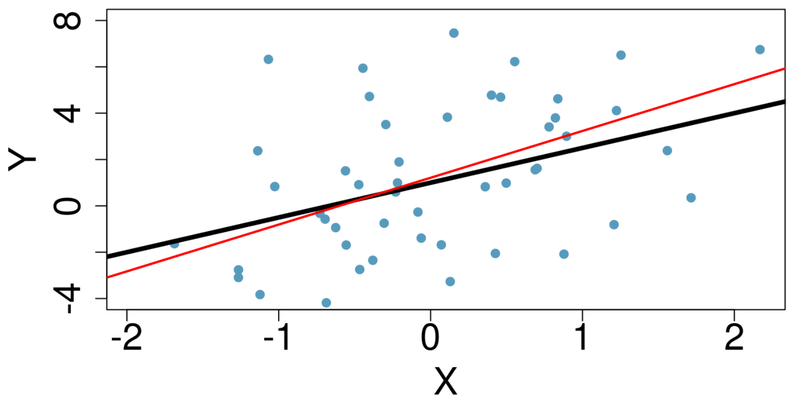
Data points:

$$Y_i = \mu(X_i) + \epsilon_i$$

Red line: (based on data)

$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least squares line is based on sample



Black line: (population)

$$\mu(x) = \beta_0 + \beta_1 x$$

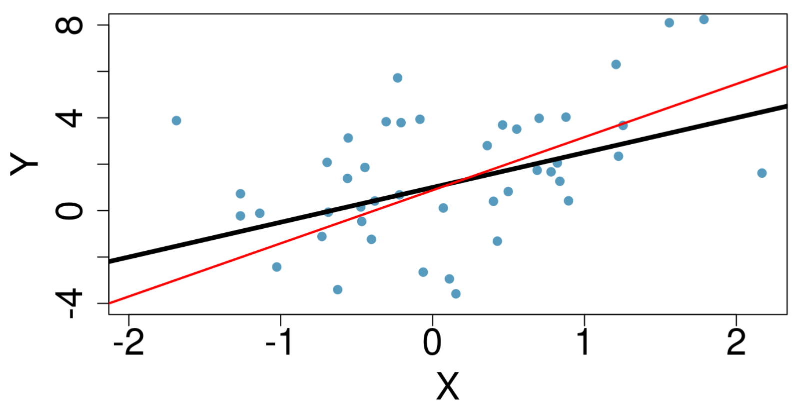
Data points:

$$Y_i = \mu(X_i) + \epsilon_i$$

Red line: (based on data)

$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least squares line is based on sample



Black line: (population)

$$\mu(x) = \beta_0 + \beta_1 x$$

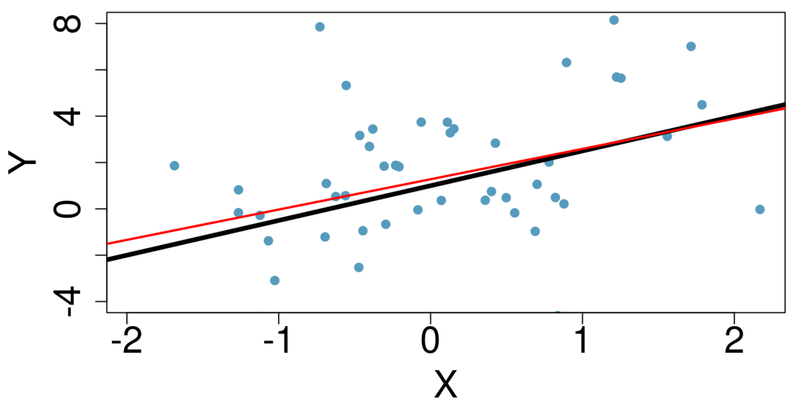
Data points:

$$Y_i = \mu(X_i) + \epsilon_i$$

Red line: (based on data)

$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least squares line is based on sample



Black line: (population)

$$\mu(x) = \beta_0 + \beta_1 x$$

Data points:

$$Y_i = \mu(X_i) + \epsilon_i$$

Red line: (based on data)

$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Outline 3: Linear Relationship, Correlation

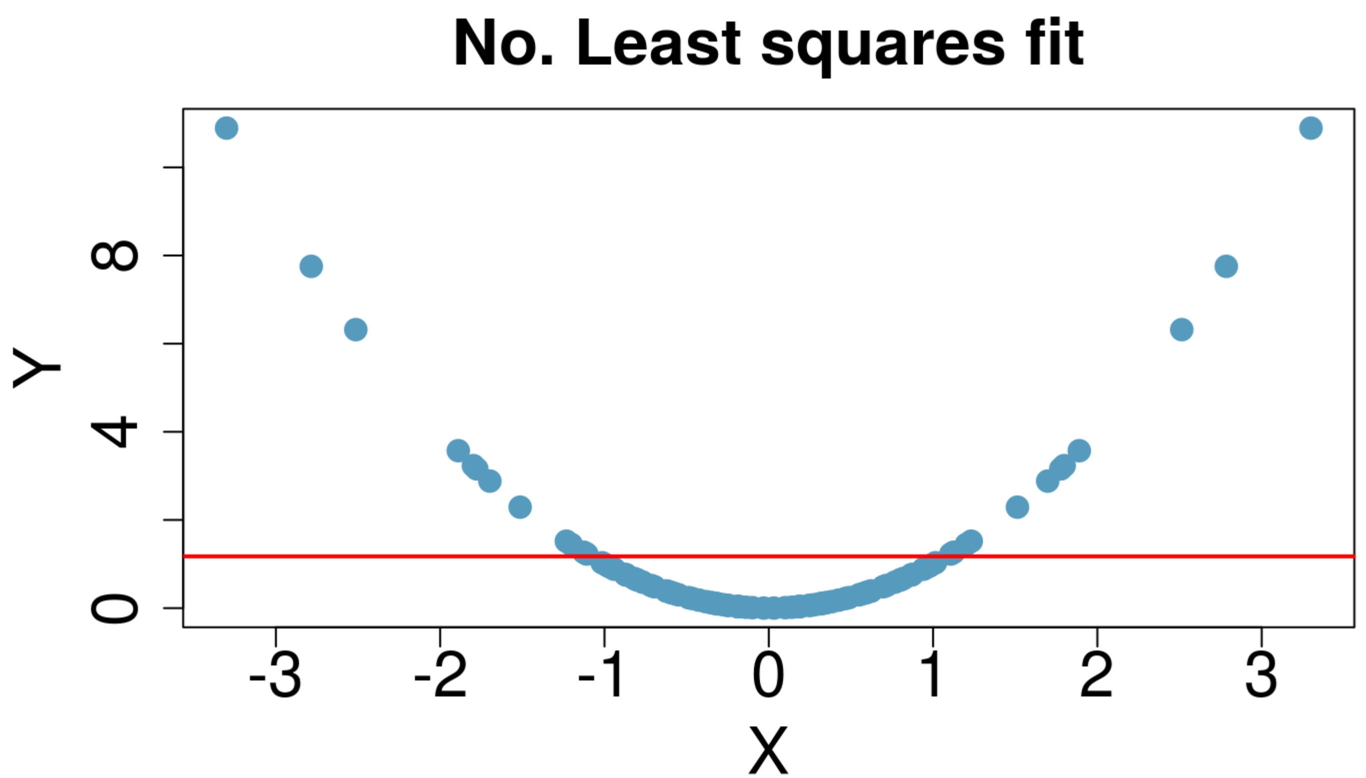
- Linear regression only captures linear relationship between X and Y
- Pearson Correlation (r) is a popular measure of linear association, $-1 \leq r \leq 1$
- Estimated slope $\hat{\beta}_1 = r s_y / s_x$, estimated intercept $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$

Average Years of Edu -> Average Prestige Score

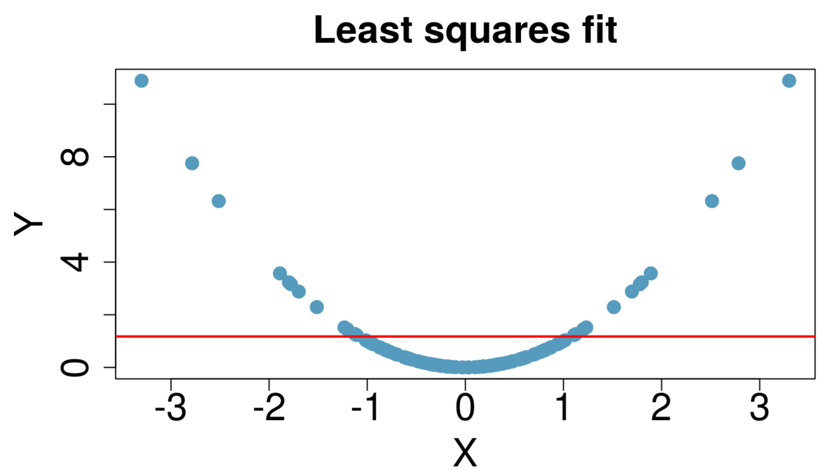
Correlation

What does this tell us?

> We find that the realized y_i's in our sample have no relationship with the x_i's?



Remember it's simple LINEAR regression



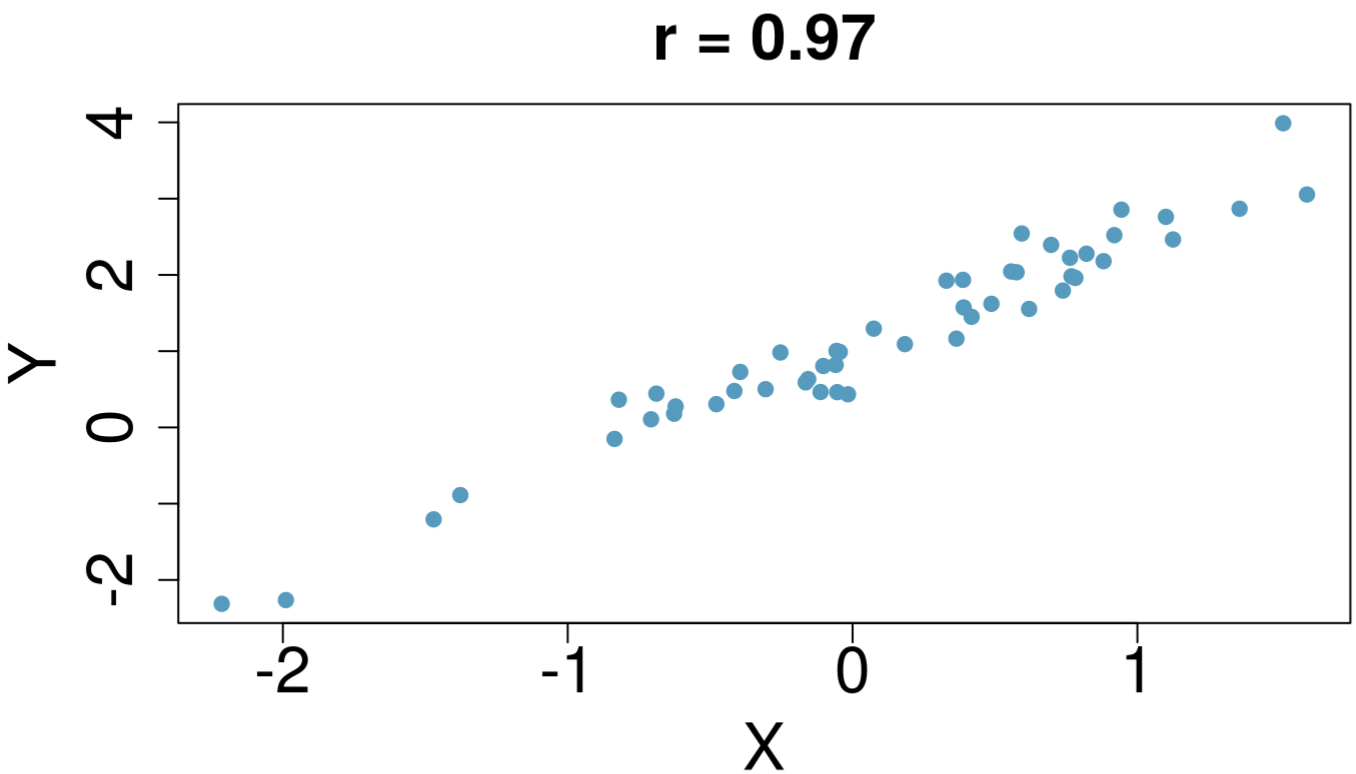
- Finding that $\hat{\beta}_1 \approx 0$ simply says that in fitting a line to the data, the best fitting line is horizontal.
- There could still be a nonlinear relationship present.
- **Moral:** Don't forget to plot the data!

Correlation, r

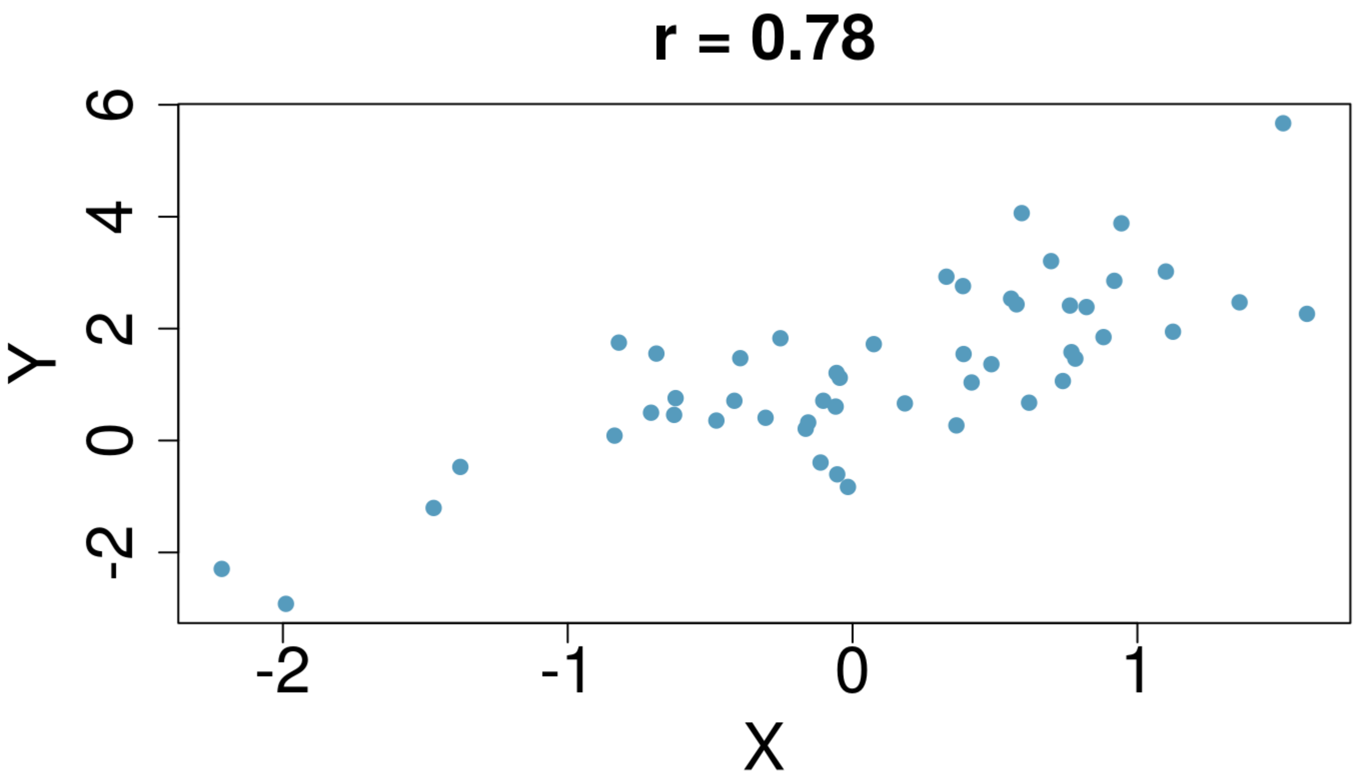
A well-known measure of the **strength of a linear relationship** between two quantitative variables

- unitless, unaffected by shifts and scalings
- between -1 and 1
- $r > 0$: positive linear association between x and y
- $r < 0$: negative linear association
- $r = 0$: no linear association ("**uncorrelated**")
- $r = \pm 1$: perfect linear association

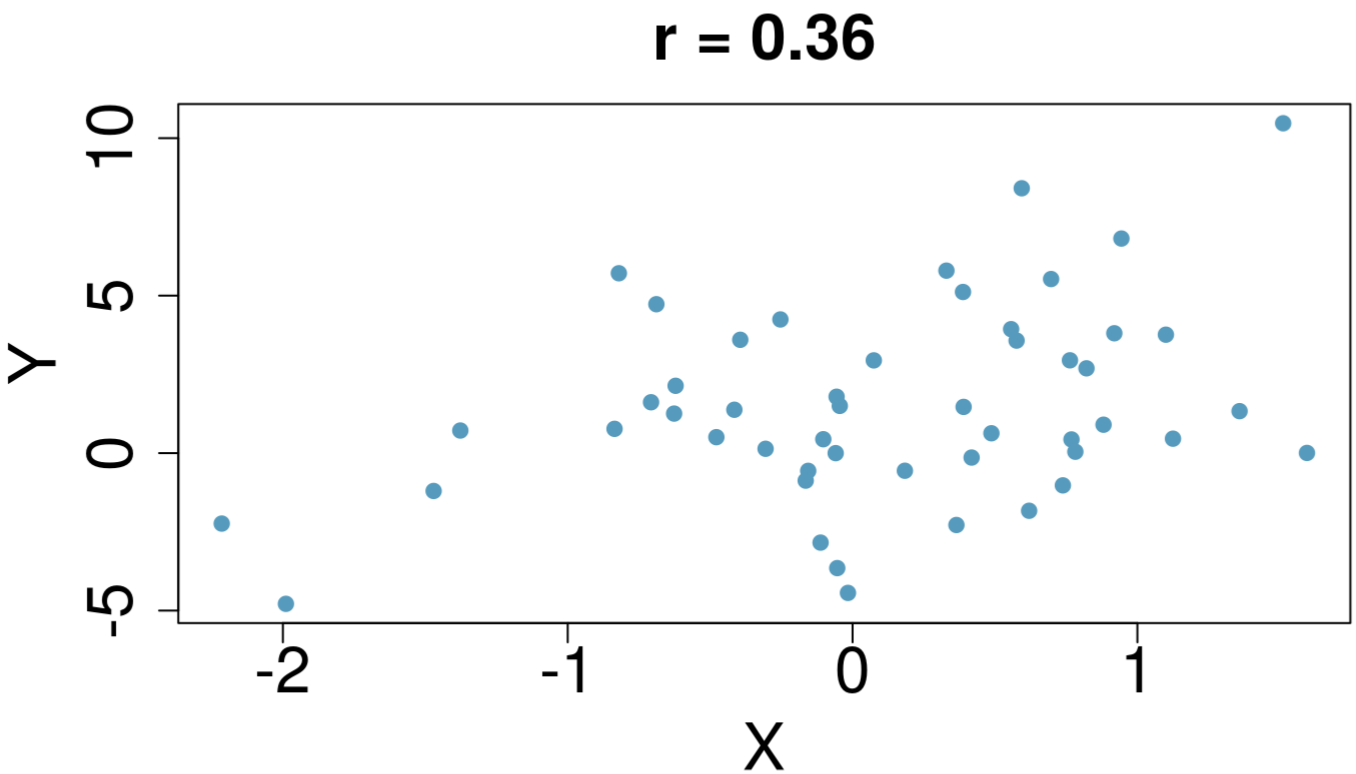
Examples



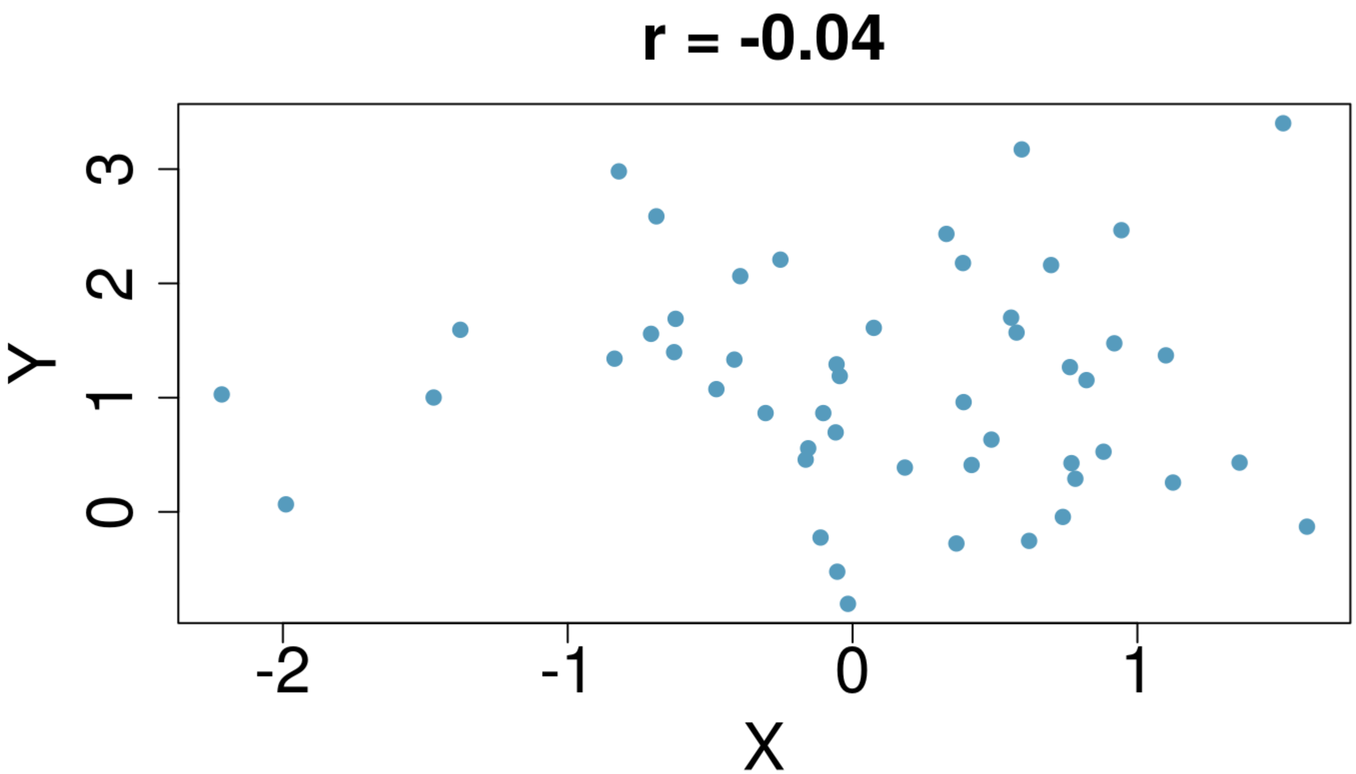
Examples



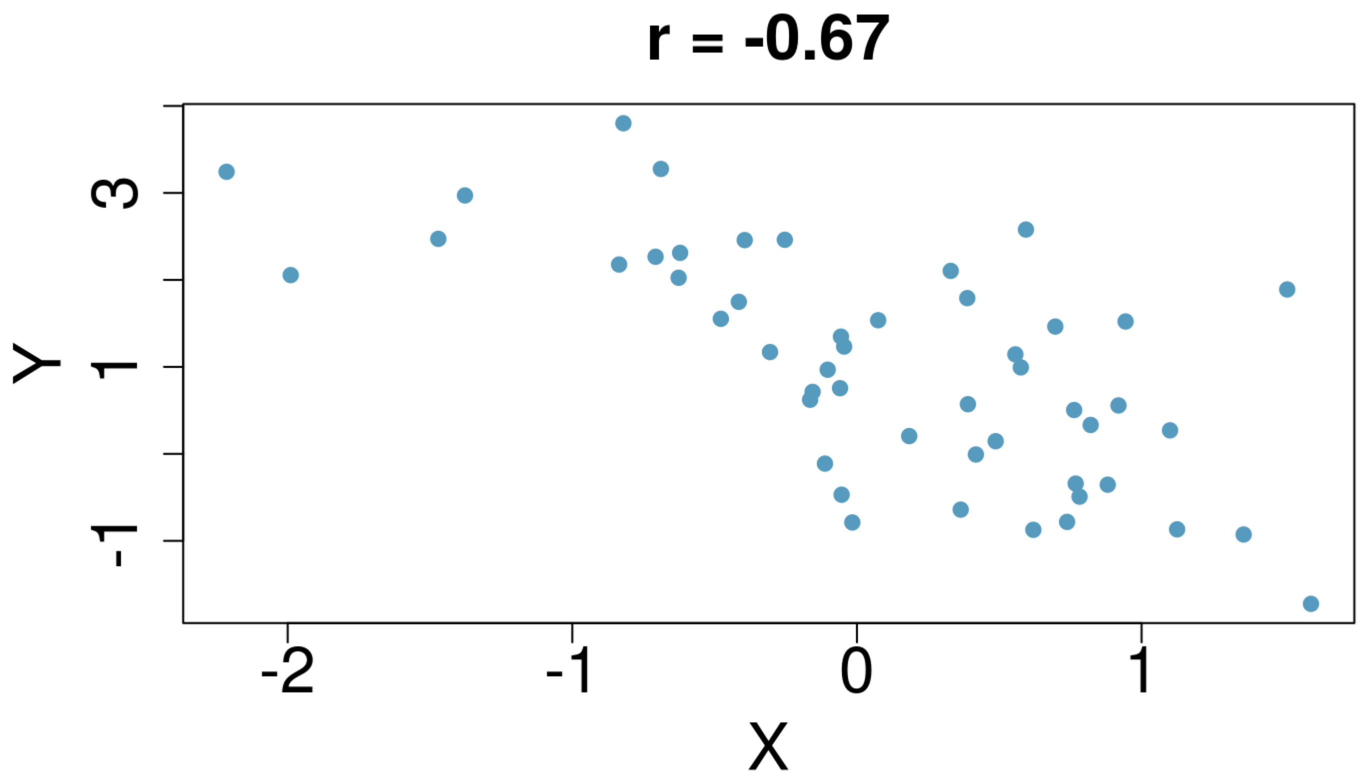
Examples



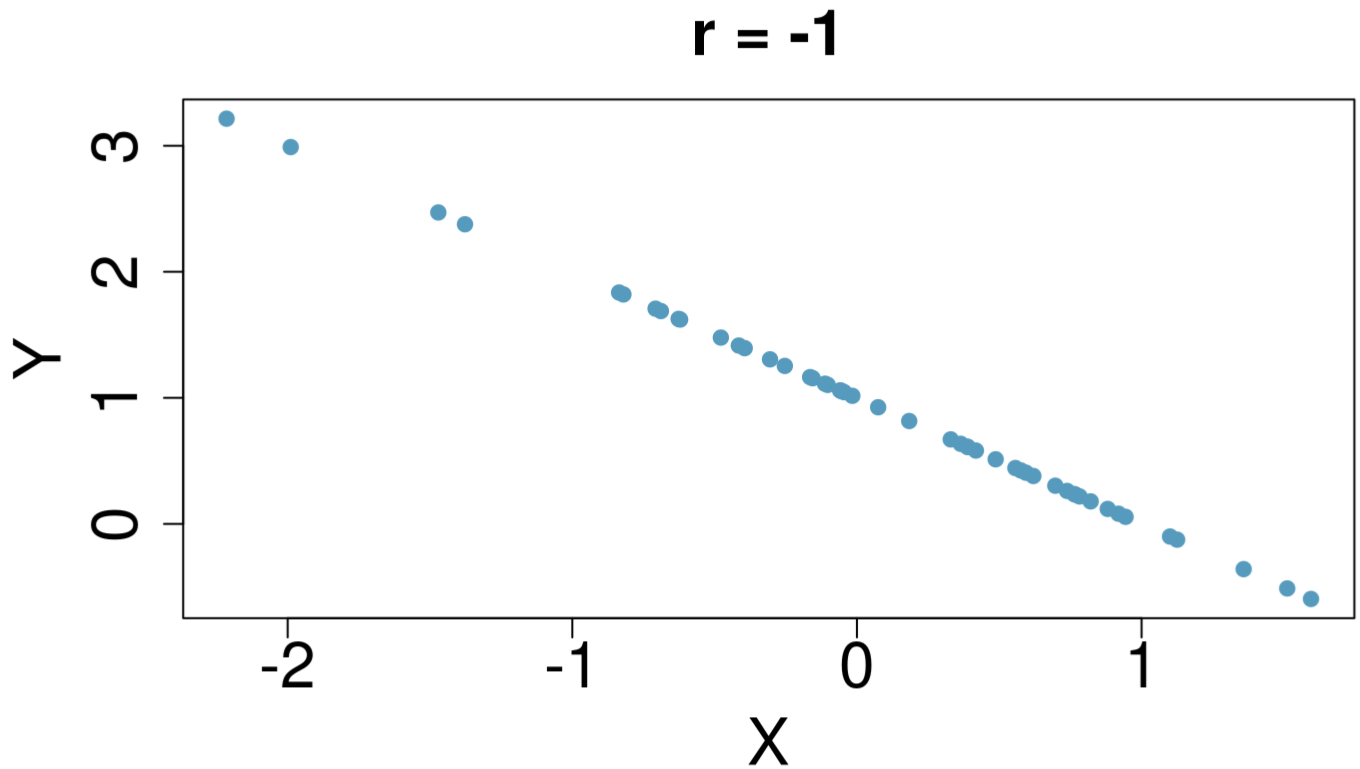
Examples



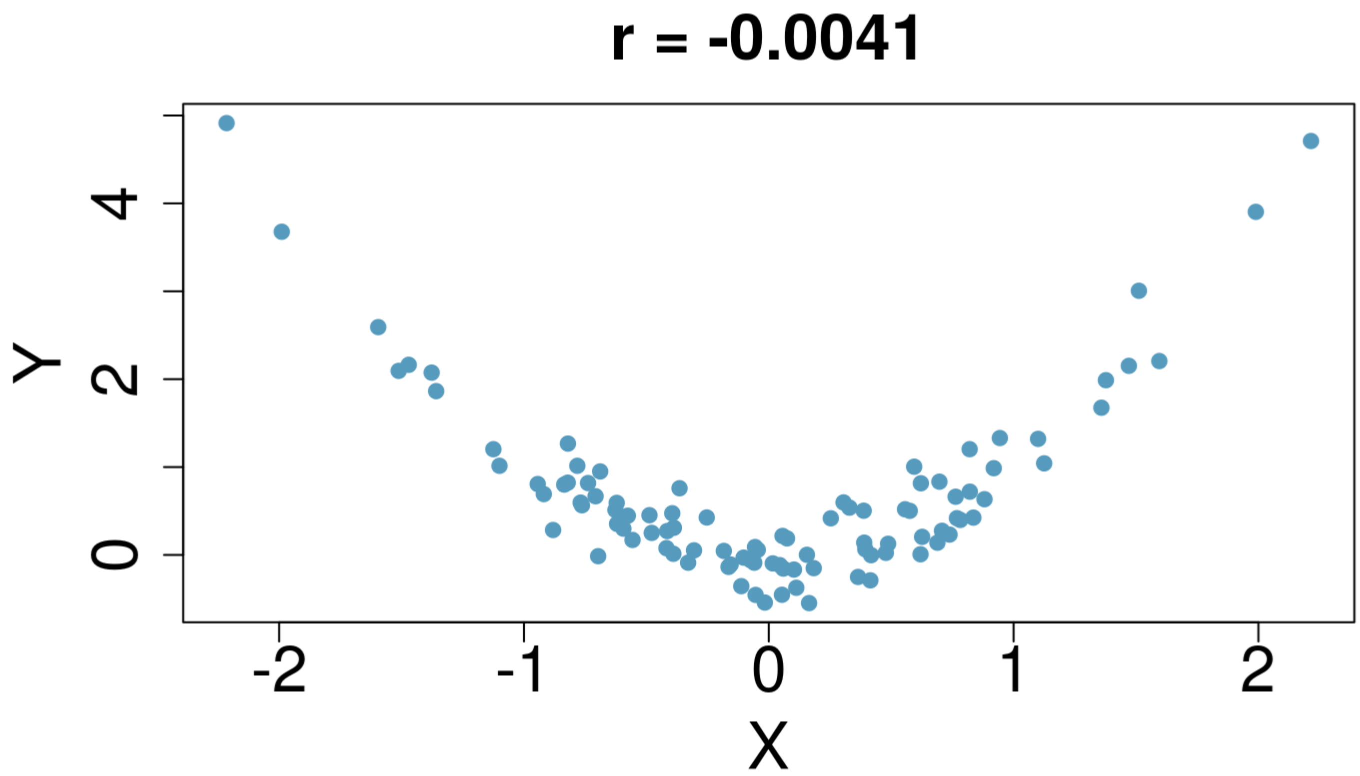
Examples



Examples



Examples



In R…

```r
cor(x, y)
```

```
## [1] 0.3635742
```

Sample correlation defined

Z-score_x Z-score_y

Correlation between pairs $(x_1, y_1), \ldots, (x_n, y_n)$:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}_n}{s_x} \right) \left(\frac{y_i - \bar{y}_n}{s_y} \right)$$

- standardized versions of x_i and y_i are multiplied for each pair and then these are added up.
- if $x_i > \bar{x}_n$ when $y_i > \bar{y}_n$, then term i is positive.
- if $x_i < \bar{x}_n$ when $y_i < \bar{y}_n$, then term i is positive.
- otherwise the term is negative.
- high correlation says that big x's co-occur big y's (and likewise for small ones)

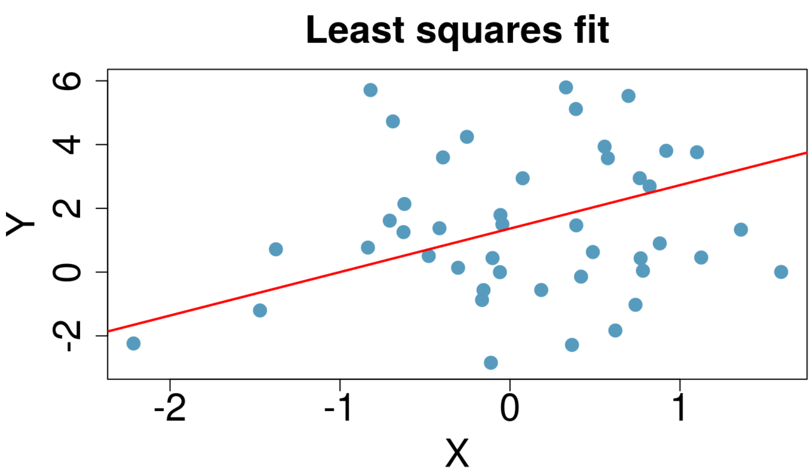
If r is positive, then on average, when x is larger than x_bar accounting for the natural scale of variation of x, y is larger than y_bar when accounting for the natural scale of variation of y

On average, z-score of x is + when z-score of y is +

Additional properties

- also, note that `cor(x,y)=cor(y,x)`
- from pictures:
  - sign of r gives direction of association
  - the closer r is to 1 or -1, the stronger the association and the more tightly the observations are clustered about the straight line

Correlation & simple linear regression



In simple linear regression, best fit line has

- slope $\hat{\beta}_1 = \frac{r s_y}{s_x}$, *Correlation after adjusting for variablity*
- intercept $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$.
- Putting this together:

$$\hat{\mu}(x) = \bar{y}_n + \frac{r s_y}{s_x}(x - \bar{x}_n)$$

Correlation & simple linear regression

$$\hat{\mu}(x) = \bar{y}_n + \frac{r s_y}{s_x}(x - \bar{x}_n)$$

Rewrite as

$$\frac{\hat{\mu}(x) - \bar{y}_n}{s_y} = r \cdot \left(\frac{x - \bar{x}_n}{s_x} \right)$$

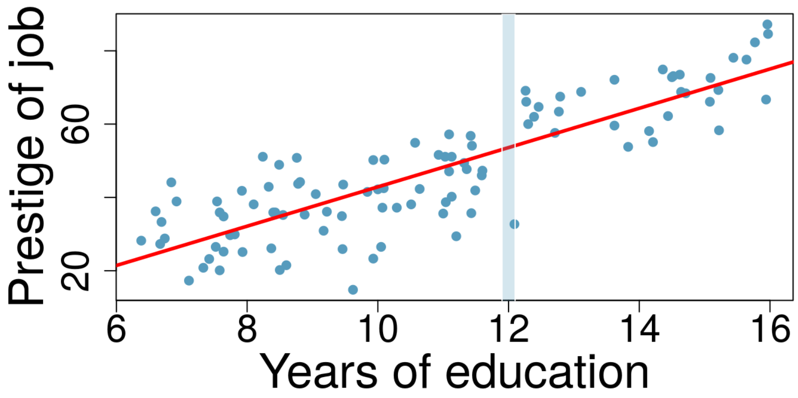
so if x's and y's are each standardized first, simple linear regression is simply the line through the origin of slope r.

Making predictions with SLR

Regression is about prediction

- given that $X = x$, what value of Y do we expect? $\mu(X) = E[Y \mid X = x]$
- if a person has ten years of education, what do we expect the prestige of his/her job to be?
- if today's temperature is 20F, what do we expect tomorrow's temperature to be?
- terminology
- Y is called the **response** ("dependent variable")
- X is called the **predictor** ("independent variable" or "covariate" or "explanatory variable" or "feature")

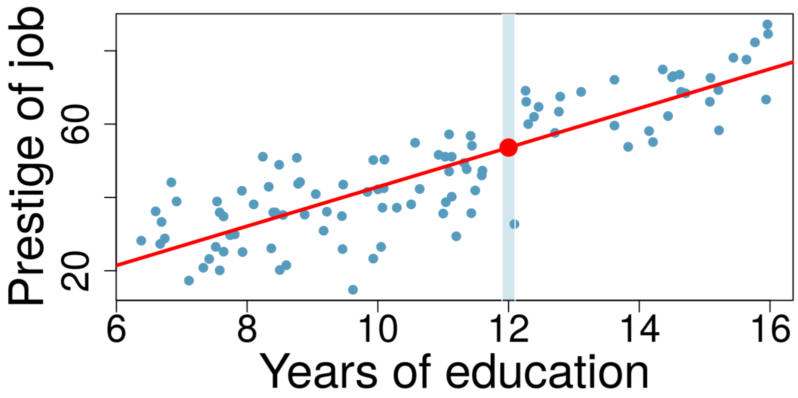
Predicted values



- Given Education = 12, what is expected value of Prestige?
- if we knew population parameters:

$$\mu(12) = E[Y|X = 12] = \beta_0 + \beta_1 \cdot 12$$

Predicted values



- Given Education = 12, what is expected value of Prestige?
- if we knew population parameters:

$$\mu(12) = E[Y|X = 12] = \beta_0 + \beta_1 \cdot 12$$

- instead, we use estimates

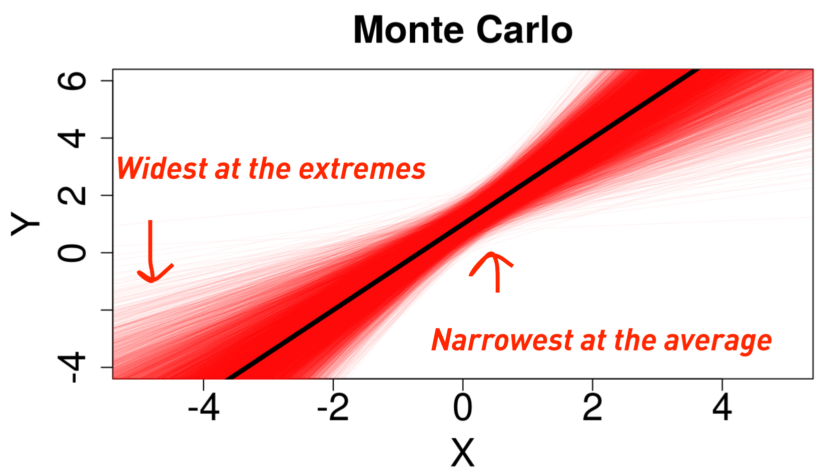
$$\hat{\mu}(12) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 12$$

In R…

```
fit <- lm(prestige ~ education, data = Prestige)
predict(fit, newdata = list(education = 12))
```

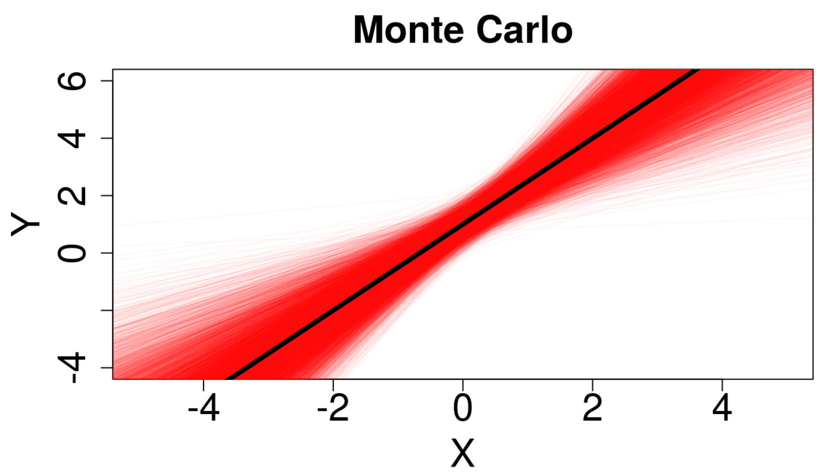
```
##        1
## 53.59855
```

Estimated line is random



- 10,000 Monte Carlo simulations shown
- each time we simulate $n = 20$ points from true (population) model shown in black
- each time we get an estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ of intercept and slope... resulting line shown in red

Danger of extrapolating far from data



- predictions are fairly reliable near $(\bar{x}_n, \bar{y}_n)$
- far from data, predictions are very high variance

Variance of estimate

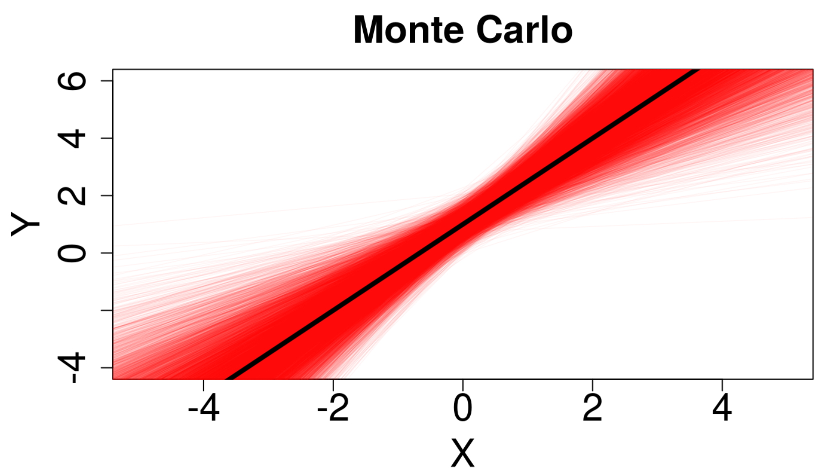
- black line is $\mu(x) = \beta_0 + \beta_1 x$
- each red line is a realization of $\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
- what is the variance of our estimate of $\mu(x)$ at a point x_0?

$$\mathrm{Var}[\hat{\mu}(x_0)] = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{(n-1)s_x^2} \right)$$

Variance increases as you move away from the mean.

- recall σ came from $\epsilon \sim N(0, \sigma)$

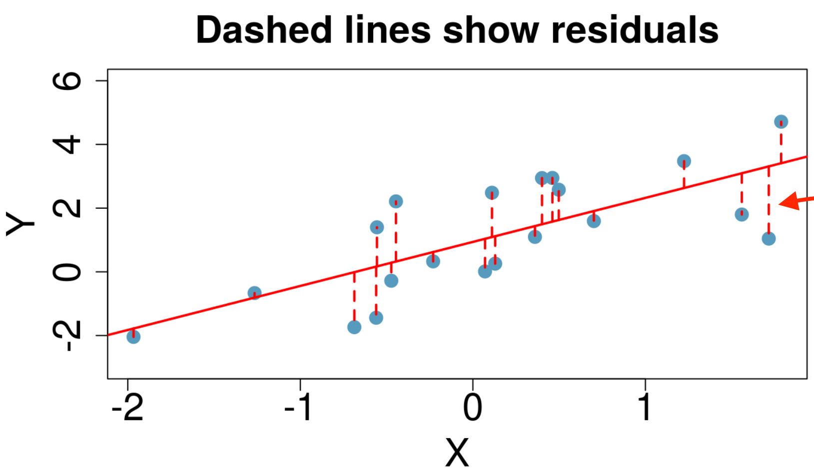
Variance of estimate



$$\mathrm{Var}[\hat{\mu}(x_0)] = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{(n-1)s_x^2} \right)$$

- at $x_0 = \bar{x}_n$ it's just σ^2/n
- gets really large far from $\bar{x}_n$

Estimating sigma



- Recall that $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma)$
- $\sigma^2 = E[\epsilon_i^2]$
- In red, we show *residuals*:

$$r_i = Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i]$$

- To estimate σ^2 we use

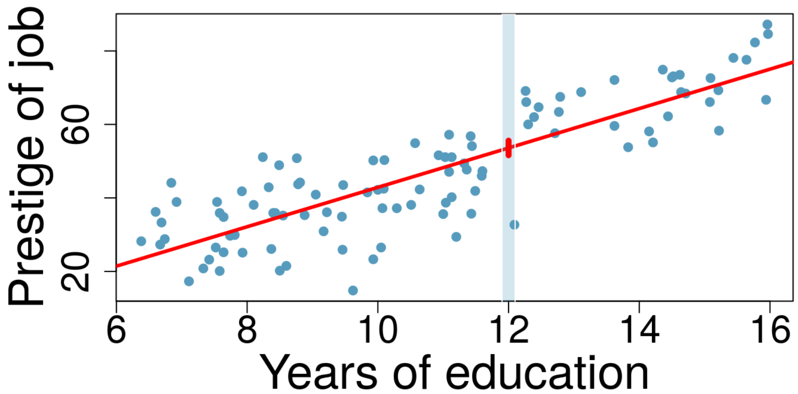
$$\boxed{\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2}$$

Confidence interval

- $100(1-\alpha)\%$ confidence interval for $\mu(x_0)$:

$$\hat{\mu}(x_0) \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{(n-1)s_x^2}}$$

Confidence Interval



We are 95% confident that the mean prestige of a person with 12 years of education is between 51.6 and 55.6.

In R...

```
predict(fit, newdata = list(education = 12),
        interval = "confidence")
```

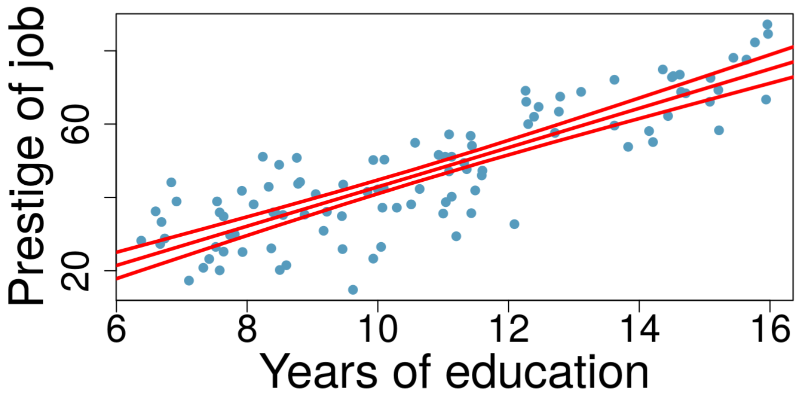
```
##        fit      lwr      upr
## 1 53.59855 51.62654 55.57056
```

- or for multiple education levels:

```
predict(fit, newdata = list(education = c(10, 11, 12)),
        interval = "confidence")
```

```
##        fit      lwr      upr
## 1 42.87680 41.02363 44.72996
## 2 48.23767 46.44110 50.03425
## 3 53.59855 51.62654 55.57056
```

Pointwise confidence intervals



For any x_0, it gives the confidence interval for $\mu(x_0)$.

Prediction interval

- so far, we've seen how to get a confidence interval for $\mu(x_0)$, the mean of Y when $X = x_0$.
- but sometimes we want to have an interval that instead gives **forecast range** for an *individual* Y (not mean) at $X = x_0$.
- imagine a future value Y_{new} at $X = x_0$.
- $Y_{new} = \mu(x_0) + \epsilon_{new}$
- our confidence interval for $\mu(x_0)$ doesn't include uncertainty due to ϵ_{new}.
- a **prediction interval** is a random interval that captures the random Y_{new} with a certain probability.

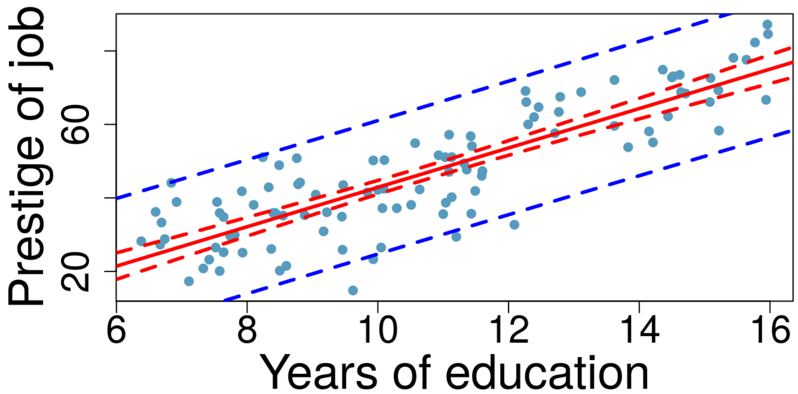
Prediction interval

- $100(1-\alpha)\%$ prediction interval for a future observation Y_{new} at $X = x_0$:

$$\hat{\mu}(x_0) \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{(n-1)s_x^2}}$$

- interval is expanded to account for extra variance of σ^2 from ϵ_{new}.
- when n is enormous, we know $\mu(x_0)$ really well and yet we are still uncertain about Y_{new}.

Pointwise prediction intervals



- For any x_0, red gives the confidence interval for $\mu(x_0)$.
- For any x_0, blue gives the prediction interval for a future observation Y_{new} at $X = x_0$.

In R...

```r
predict(fit, newdata = list(education=12), interval = "confidence")
```

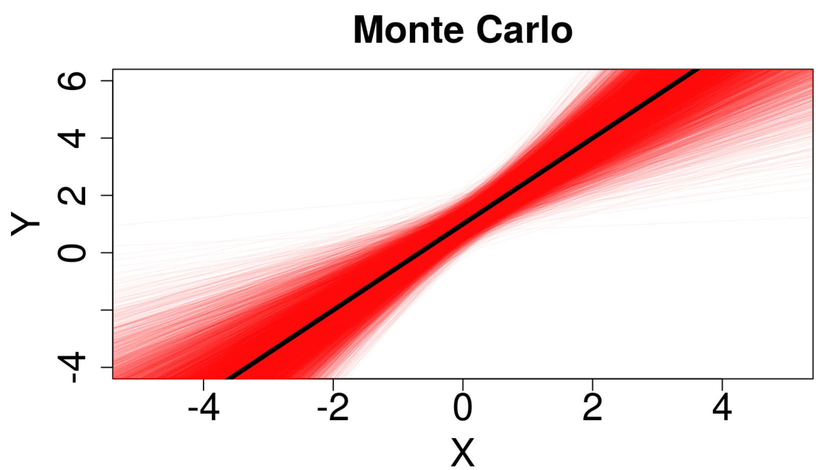
```
##        fit      lwr      upr
## 1 53.59855 51.62654 55.57056
```

```r
predict(fit, newdata = list(education=12), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 53.59855 35.43054 71.76656
```

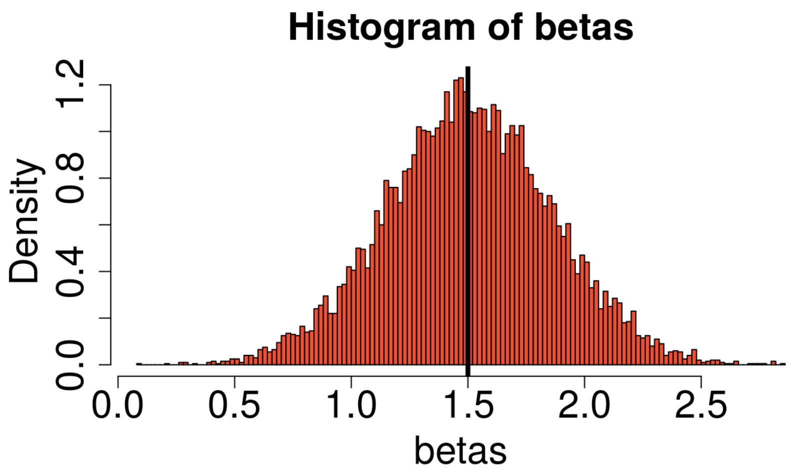
Inference about slope and intercept

Estimated line is random



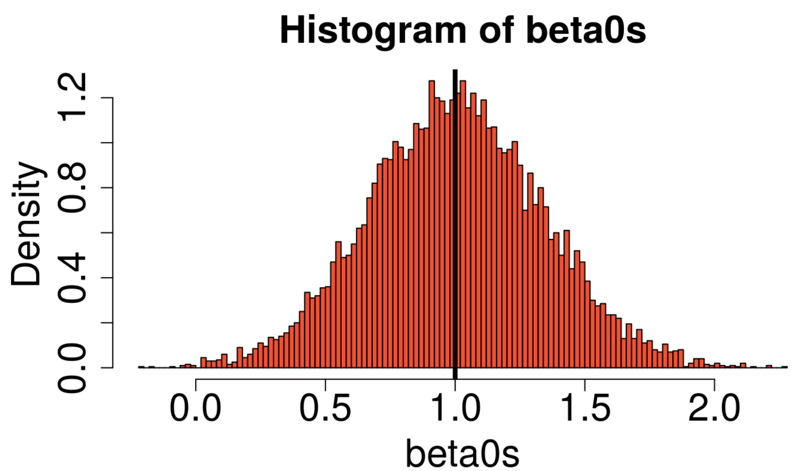
- each time we get an estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ of intercept and slope… resulting line shown in red

Looking at slope from Monte Carlo



- These are the 10,000 slopes (realizations of $\hat{\beta}_1$) we get from the SLR Monte Carlo.
- **Good news** it's normal!
- centered at correct place: population parameter β_1 (black line)

Looking at slope from Monte Carlo



- These are the 10,000 intercepts (realizations of $\hat{\beta}_0$) we get from the SLR Monte Carlo.
- **Good news** it's normal!
- centered at correct place: population parameter β_0 (black line)

Sampling distribution for slope

The SLR estimate $\hat{\beta}_1$ of the slope β_1 has a **normal distribution** with mean β_1 and variance

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{(n-1)s_x^2},$$

- $\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1})$ is equivalent to $\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0, 1)$
- however, in practice we don't know σ so for testing and intervals, we use that

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2} \text{ where } \hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{(n-1)s_x^2}$$

Confidence interval for slope

A $100(1-\alpha)\%$ confidence interval for the population parameter β_1 is given by

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \frac{\hat{\sigma}}{s_x\sqrt{(n-1)}}.$$

Why do we care about this?

- often interested in whether $H_0 : \beta_1 = 0$ or $H_A : \beta_1 \neq 0$
- under H_0, SLR model is $Y = \beta_0 + \epsilon$
- tests whether expected Y doesn't depend on X (assuming SLR model holds)

Likewise for intercept

A $100(1-\alpha)\%$ confidence interval for the population parameter β_0 is given by

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}_n^2}{(n-1)s_x^2}}.$$

- this is less commonly of interest

Example: Prestige

```
fit = lm(prestige~education, data=Prestige)
summary(fit)
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0397  -6.5228   0.6611   6.7430  18.1636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.732      3.677  -2.919  0.00434 **
## education      5.361      0.332  16.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.103 on 100 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:   0.72
## F-statistic: 260.8 on 1 and 100 DF,  p-value: < 2.2e-16
```

Goodness of fit

The above shows that $\hat{\beta}_1 = 5.36$ and $\hat{\sigma}_{\hat{\beta}_1} \approx 0.332$.

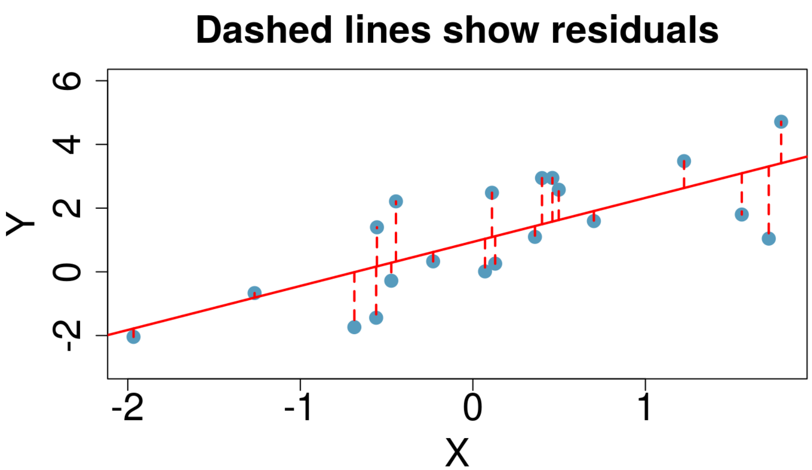
Example: Prestige

$\hat{\beta}_1 = 5.36$ and $\hat{\sigma}_{\hat{\beta}_1} \approx 0.332.$

We are roughly 95% confident that for each additional year of education the mean prestige of job increases by between 5.36 - 2(0.332) and 5.36 + 2(0.332) units.

Explained variance

Estimating sigma



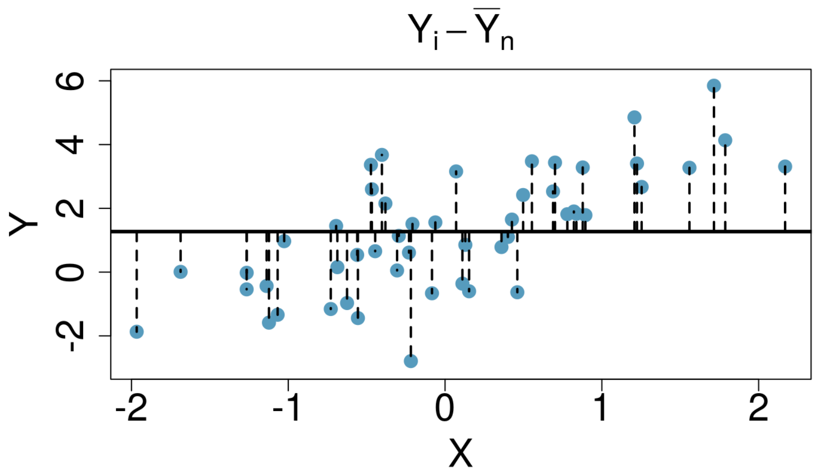
- Recall that $Y_i = \beta_0 + \beta X_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma)$
- $\sigma^2 = E[\epsilon_i^2]$
- In red, we show *residuals*:

$$r_i = Y_i - [\hat{\beta}_0 + \hat{\beta} X_i]$$

- To estimate σ^2 we use

$$\boxed{\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2}$$

Decomposing the variance

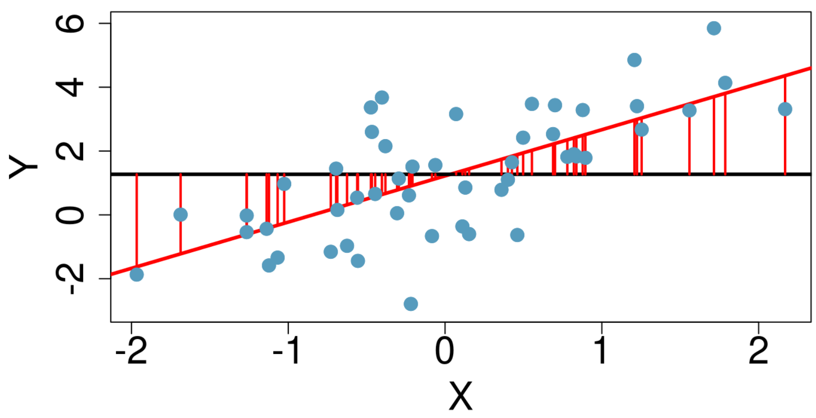


Just as in ANOVA, we can decompose the observed variation

$$tss = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

into the part that's "explained" by X and the part that is not.

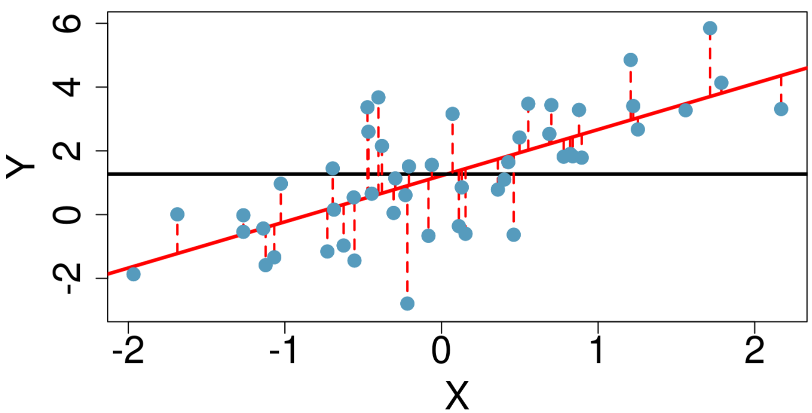
SSR



how much of the variation in y_i's is explained by regression line?

$$ssr = \sum_{i=1}^{n} (\hat{\mu}(x_i) - \bar{y}_n)^2$$

sse



variation not explained by regression line…

$$sse = \sum_{i=1}^{n} (y_i - \hat{\mu}(x_i))^2$$

$$= \sum_{i=1}^{n} r_i^2 = (n-2)\hat{\sigma}^2$$

Amount of variability NOT explained by regression

Decomposing the variance

Just like with ANOVA

$$tss = ssr + sse$$

Definition: Fraction of total variation explained by Linear ~~regression~~ relationship with X:

$$R^2 = \frac{ssr}{tss}$$

Exercise: prove that for SLR, $R^2 = r^2$ (correlation!)

Prestige Example

```
fit = lm(prestige~education, data=Prestige)
summary(fit)
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0397  -6.5228   0.6611   6.7430  18.1636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.732      3.677  -2.919  0.00434 **
## education      5.361      0.332  16.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.103 on 100 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:   0.72
## F-statistic: 260.8 on 1 and 100 DF,  p-value: < 2.2e-16
```

shows that R-squared is about 0.72

Prestige Example

- `summary(fit)` shows that R-squared is about 0.72
- *72% of the variation in job prestige is explained by its relationship with years of education.*

Recall: ANOVA Table (k groups)

Source of variability	Degrees of freedom	Sum of squares	Mean square	F statistic
Between	$k - 1$	ssb	$msb = \frac{ssb}{k-1}$	$\frac{msb}{mse}$
Error (within)	$n_{tot} - k$	sse	$mse = \frac{sse}{n_{tot}-k}$	
Total	$n_{tot} - 1$	tss		

Regression ANOVA Table

Source of variability	Degrees of freedom	Sum of squares	Mean square	F statistic
Regression	1	ssr	$msr = ssr$	$\frac{msr}{mse}$
Error (within)	$n - 2$	sse	$mse = \frac{sse}{n-2}$	
Total	$n - 1$	tss		

- **same idea:** what proportion of variation is explained by X?
- only 1 degree of freedom due to X since we only estimate 1 parameter on account of X (i.e., the slope β)

F test

Which of two competing models is “better”?

Simple mean model: $Y = \beta_0 + \epsilon$ (aka, intercept only model)

SLR model: $Y = \beta_0 + \beta_1 X + \epsilon$

$H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

Idea: compare msr to mse. If msr/mse is much larger than 1, we reject H_0.

p-value: $P(F_{1,n-2} > msr/mse)$

In R…

```
fit = lm(prestige~education, data=Prestige)
anova(fit)
```

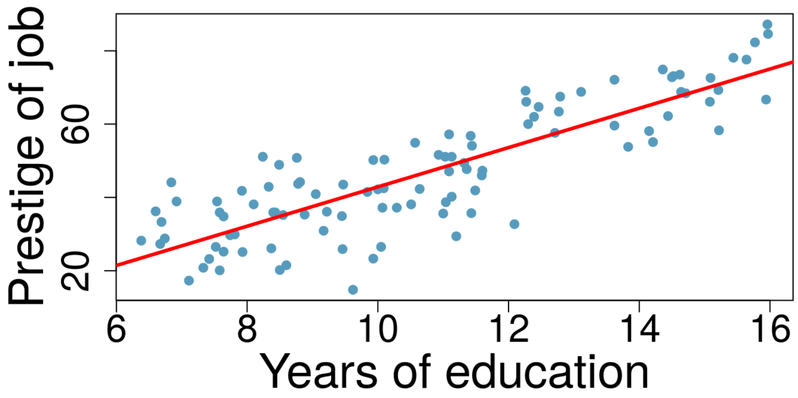
```
## Analysis of Variance Table
##
## Response: prestige
##            Df Sum Sq Mean Sq F value    Pr(>F)
## education   1  21608 21608.4  260.75 < 2.2e-16 ***
## Residuals 100   8287    82.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

But what about our t test for slope?

- recall we know how to construct a confidence interval for the slope β_1 based on the t-distribution
- in SLR, the F test is redundant (and in fact equivalent) to the t test for testing $\beta = 0$.
- in more general regression settings, F allows for more complex tests

Review Slides

Simple model: linear relationship



$$E[Y \mid X = x] = \beta_0 + \beta_1 \cdot x$$

(equation for a straight line)

- intercept β_0
- slope β_1
- here $\beta_0 \approx -10, \beta_1 \approx 5$

Simple linear regression

Assumes mean of Y increases linearly in x:

$$E[Y \mid X = x] = \beta_0 + \beta_1 x.$$

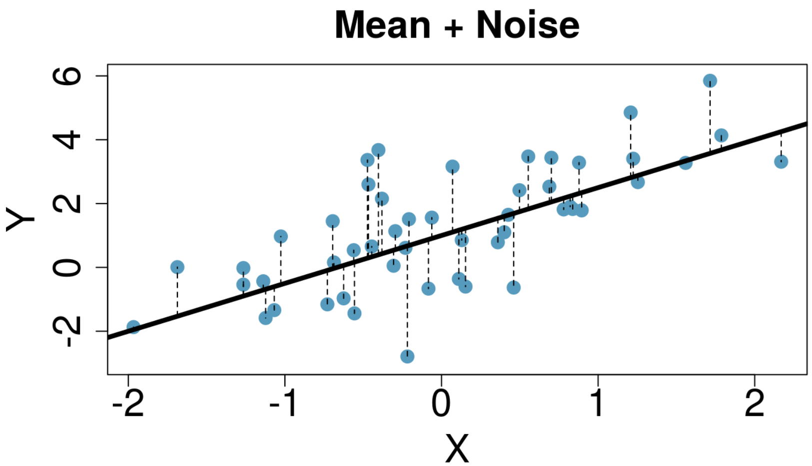
Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma)$ are independent.

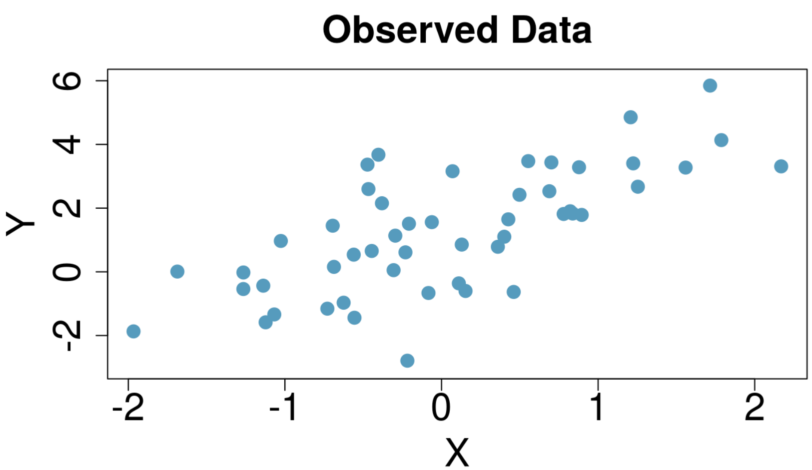
- just like in ANOVA, we assume independent normal noise with fixed variance

In pictures



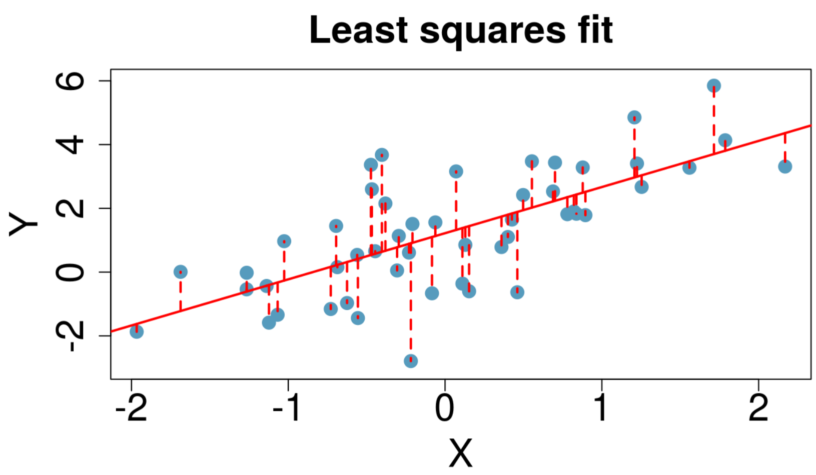
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

In pictures



- We observe (X_i, Y_i) pairs.
- How can we estimate β_0 and β_1 from this data?

Natural Idea: Best fitting line



Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that makes "sum of squared errors"

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

as small as possible.

- This is called the **least squares** line

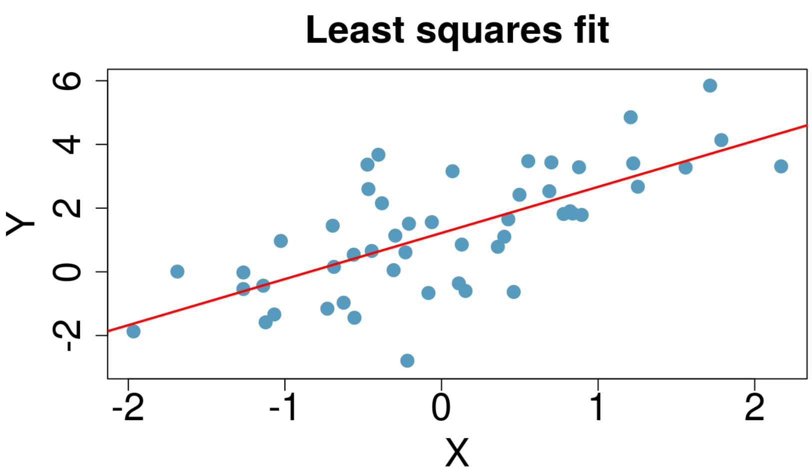
Sample correlation defined

Correlation between pairs $(x_1, y_1), \ldots, (x_n, y_n)$:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}_n}{s_x} \right) \left(\frac{y_i - \bar{y}_n}{s_y} \right)$$

- standardized versions of x_i and y_i are multiplied for each pair and then these are added up.
- if $x_i > \bar{x}_n$ when $y_i > \bar{y}_n$, then term i is positive.
- if $x_i < \bar{x}_n$ when $y_i < \bar{y}_n$, then term i is positive.
- otherwise the term is negative.
- high correlation says that big x's co-occur big y's (and likewise for small ones)
- also, note that `cor(x,y)=cor(y,x)`

Correlation and simple linear regression



Best fit line has

- slope: $\hat{\beta}_1 = \frac{r s_y}{s_x}$
- intercept: $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$
- putting this together:

$$\hat{\mu}(x) = \bar{y}_n + \frac{r s_y}{s_x}(x - \bar{x}_n)$$

Correlation and simple linear regression

$$\hat{\mu}(x) = \bar{y}_n + \frac{rs_y}{s_x}(x - \bar{x}_n)$$

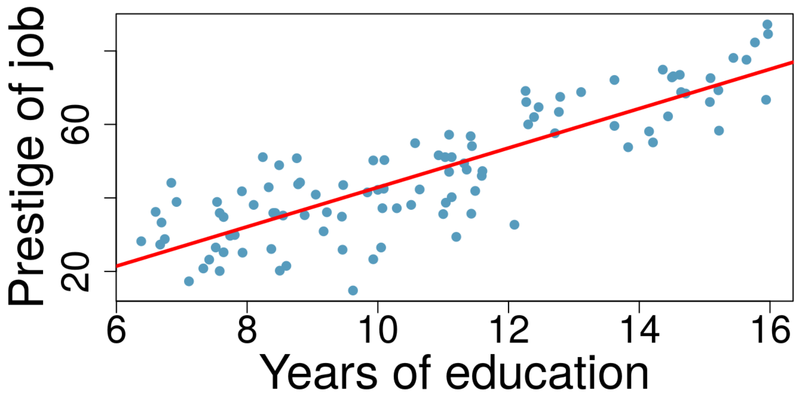
Rewrite as

$$\frac{\hat{\mu}(x) - \bar{y}_n}{s_y} = r \cdot \left(\frac{x - \bar{x}_n}{s_x} \right)$$

That is, if z_y is standardized y and z_x is standardized x, then regression line given by

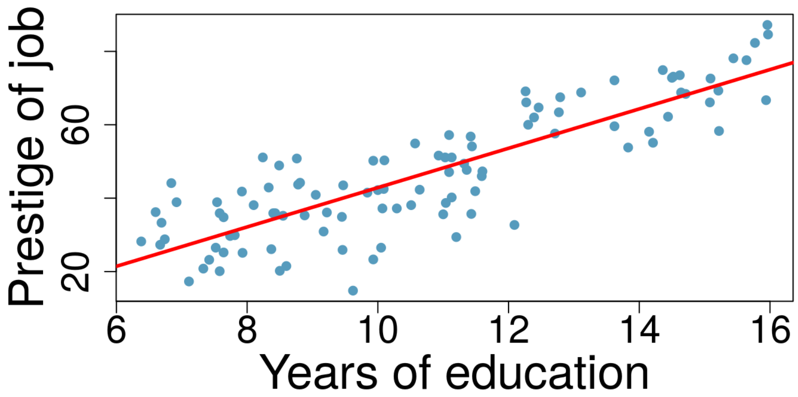
$$z_y = r \cdot z_x$$

Example: Prestige



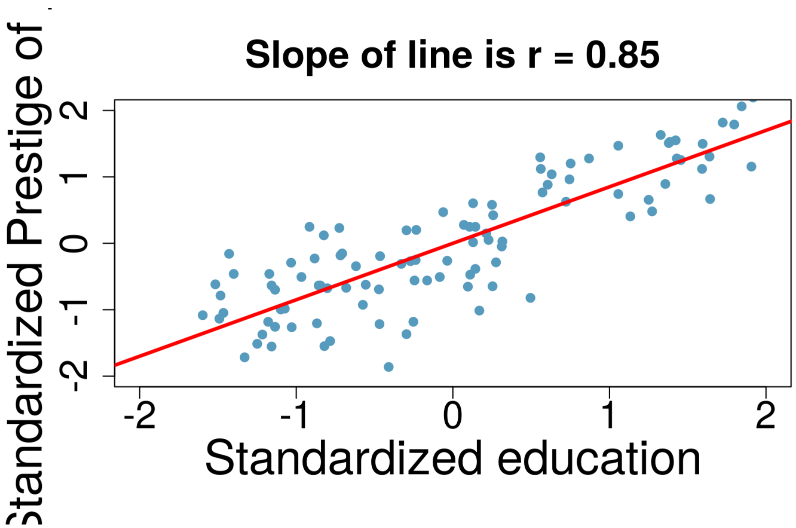
- $s_x = 2.7284442$
- $s_y = 17.2044856$
- $r = 0.8501769$
- $rs_y/s_x = 5.3608777$
- `lm(prestige~education)` gives $\hat{\beta}_1$ = 5.3608777

Example: Prestige



$$Prestige \approx -10 + 5 \cdot Education$$

Example: Prestige



- Standardized education: $z_x = \frac{x - \bar{x}}{s_x}$
- Standardized prestige: $z_y = \frac{y - \bar{y}}{s_y}$
- Equation of line:

$$z_y = r z_x$$