Quantitative Sociological Analysis

Inferential Statistics
Hypothesis Testing and Bivariate Statistics

Part 7

April 17, 2025

Simple Linear Regression Model (LRM)

- can be used to estimate an association between a continuous Y (DV) and any type of X (IV)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y is the DV (outcome trying to predict)

X is the IV (predictor of the outcome)

β_0 is the y-intercept (value of Y when X = 0)

β_1 is the slope coefficient (change in Y for a one-unit change in X)

ε is the error term (random variation or omitted variables (theoretical "true error"))

- goal is to estimate β_0 and β_1 using data, given assumptions are met
  - more on this later, [maybe]

Let's break this down…

Simple LRM continued

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

slope

$$\beta_1 = \frac{cov(X,Y)}{var(X)} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

how much Y changes for a unit change in X, standardized by the spread of X

intercept

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

what remains from the mean of Y after accounting for the linear effect of X

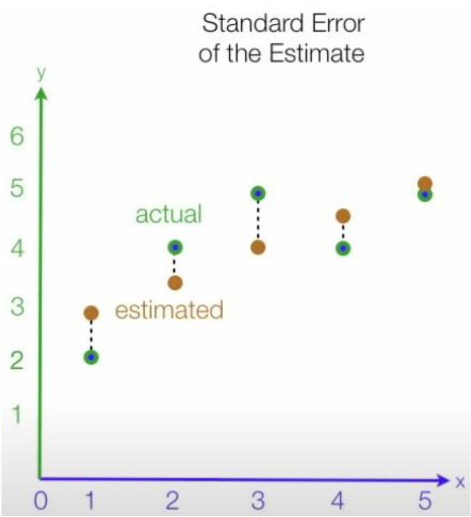
What about ε, the error term (random variation)…

Simple LRM continued

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- The error term (ε) represents unobserved variability in Y not explained by X
  - often denoted as σ^2

- This is directly connected to the precision of the estimated coefficient β_1
  - how far off, on average, predicted values are from the actual Y values
    - the residual standard error, commonly called the [standard error of the estimate](#)(SEE)



$$SEE = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{N-2}}$$

Y_i actual value

$\hat{Y}_i$ predicted (estimated) value

$N - 2$ degrees of freedom (df)

Simple LRM continued

Begin to recognize how this connects back to the Central Limit Theorem (CLT) and probability theory

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Standard error of the coefficient measures uncertainty due to random sampling variability
  - $SE_{\beta 1}$: how much a change in Y for a one-unit change in X would vary across random samples
  - $SE_{\beta 0}$: how much the value of Y when X = 0 would vary across random samples

- See how the standard error of the slope ($SE_{\beta 1}$) is equal to the standard error of the estimate (SEE) divided by the square root of the total variability in X

$$SE_{\beta 1} = \frac{SEE}{\sqrt{\sum (X_i - \bar{X})^2}}$$

The standard error of the coefficient (SE_{β}) is used to compute its margin of error (MoE) …

Simple LRM continued

Begin to recognize how this connects back to the Central Limit Theorem (CLT) and probability theory

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Margin of error (MoE) indicates the range within which the true parameter is expected
  - with a certain level of confidence (e.g., 95% when set alpha at 0.05)

$$MoE = Z^* \times SE_\beta$$

- Given that the standard error of the estimate (SEE) is used to compute the standard error of regression coefficients (SE_β)
  - this directly affects the width of the confidence interval (CI)

- The CLT holds that the distribution of coefficient estimates is approximately normal, which
  - validates the use of critical values from Z or t probability distributions to find the MoE

Let's further consider why the error term (ε) is essential for assessing uncertainty due to sampling variability…

Simple LRM continued

Begin to recognize how this connects back to the Central Limit Theorem (CLT) and probability theory

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- The error term (ε) is a theoretical concept that is never actually observed
  - It reflects everything not explained by the linear relationship between X and Y, such as
    - measurement error, omitted variables, model misspecification, and random variation

- If no ε (i.e., $\varepsilon_i = 0$ for all i), then for each X_i the value of Y_i would be perfectly predictable
  - every sample from a population would return the same exact estimates (i.e., β_0 and β_1)

- In reality, each individual observation has some random variability (noise), which means
  - each time a population is sampled the observed Y_i values bounce around the true regression line

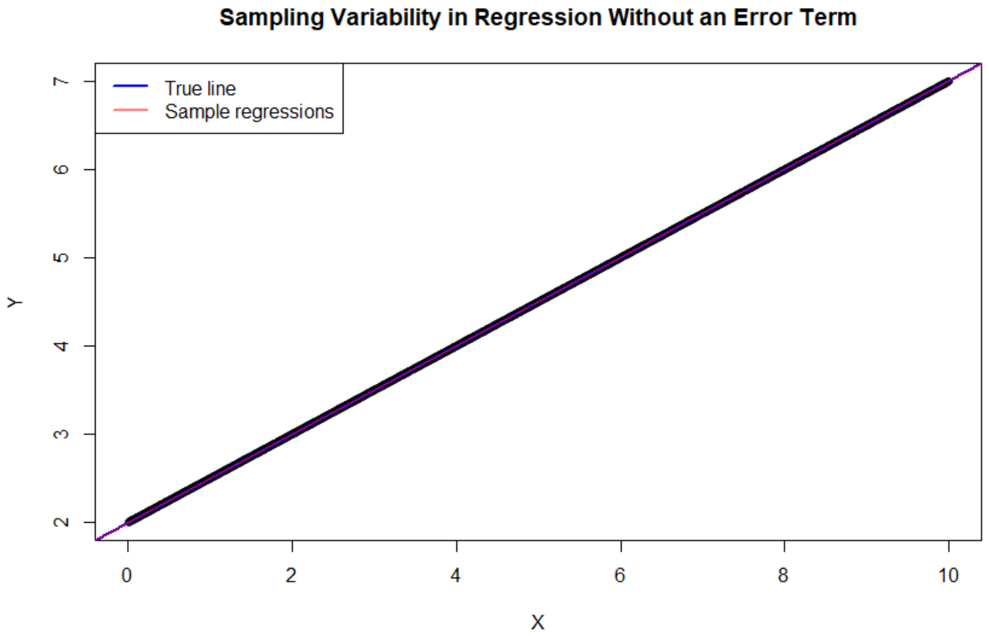
Let's consider a hypothetical example with a visualization…

Simple LRM continued

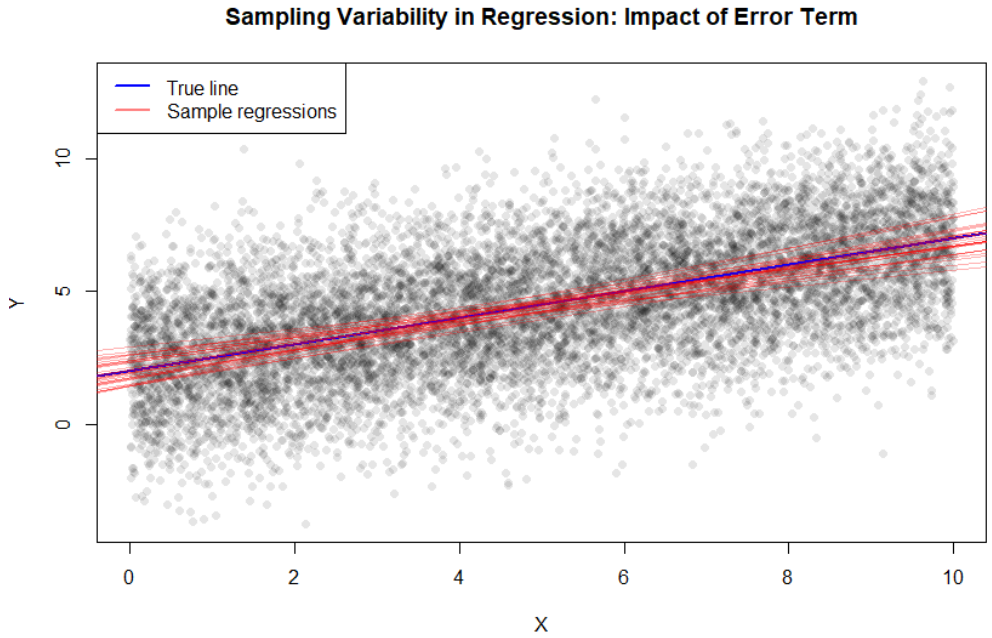
Extra Examples

LRM_ImpactOfErrorTerm_Example.R

$$Y = \beta_0 + \beta_1 X$$



$$Y = \beta_0 + \beta_1 X + \varepsilon$$



Simulated data from which 30 random samples (n = 100) were drawn. A linear regression model (LRM) was estimated for each sample and a regression line was fit to show the results, respectively.

Recall our coin toss exercise that demonstrated how random sampling variability underlies uncertainty in estimates.

Let’s extend the present example based on these simulated data where the population parameter for the true slope (β_1) = 0.5, and consider how estimated slopes (b_1) across random samples will approximate a normal distribution

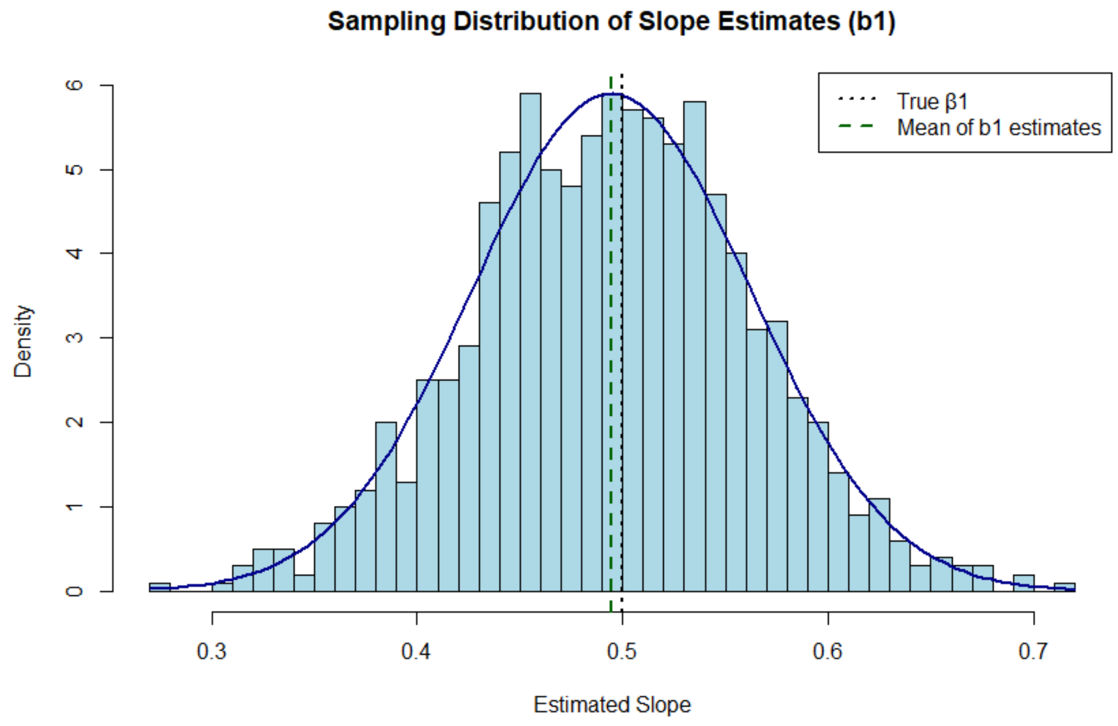
Simple LRM continued

- Each bar denotes a slope coefficient (b_1) estimated by a linear regression model (LRM) from one of 1,000 independent random samples
  - drawn from the same underlying population
- Sampling variability is reflected in how the slope estimates (b_1) differ even though they came from the same underlying population
- The spread of the histogram ≈ average of the standard errors (SE_{b1}) from the samples
  - like what a single LRM estimates, how much the coefficient is expected to vary across samples
- Empirical example showing why the CLT and probability theory are foundational for estimating confidence intervals, p-values and hypothesis tests
  - accounting for uncertainty in estimates

▾ Extra Examples

LRM_SamplingVariabilitySlopes_Example.R

$$\hat{Y} = b_0 + b_1 X$$



Since in practice we are often limited to working with only one random sample from the broader population, the standard errors of the regression coefficients ($SE_{\beta 0}$ and $SE_{\beta 1}$), derived from the residual variability (SEE), allow us to assess how estimated coefficients would theoretically vary across random samples.

Simple LRM continued

$$\hat{Y} = b_0 + b_1 X$$

Note how the linear regression equation for a sample (estimated) model does not include an error term (ε).

ε is never known, but estimated using residuals (e_i), what's left over after fitting linear association between X and Y

- Residuals (e_i) are the best guess at what the error terms are, based on sample data
  - $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$
    - how far each actual Y value is from its predicted value

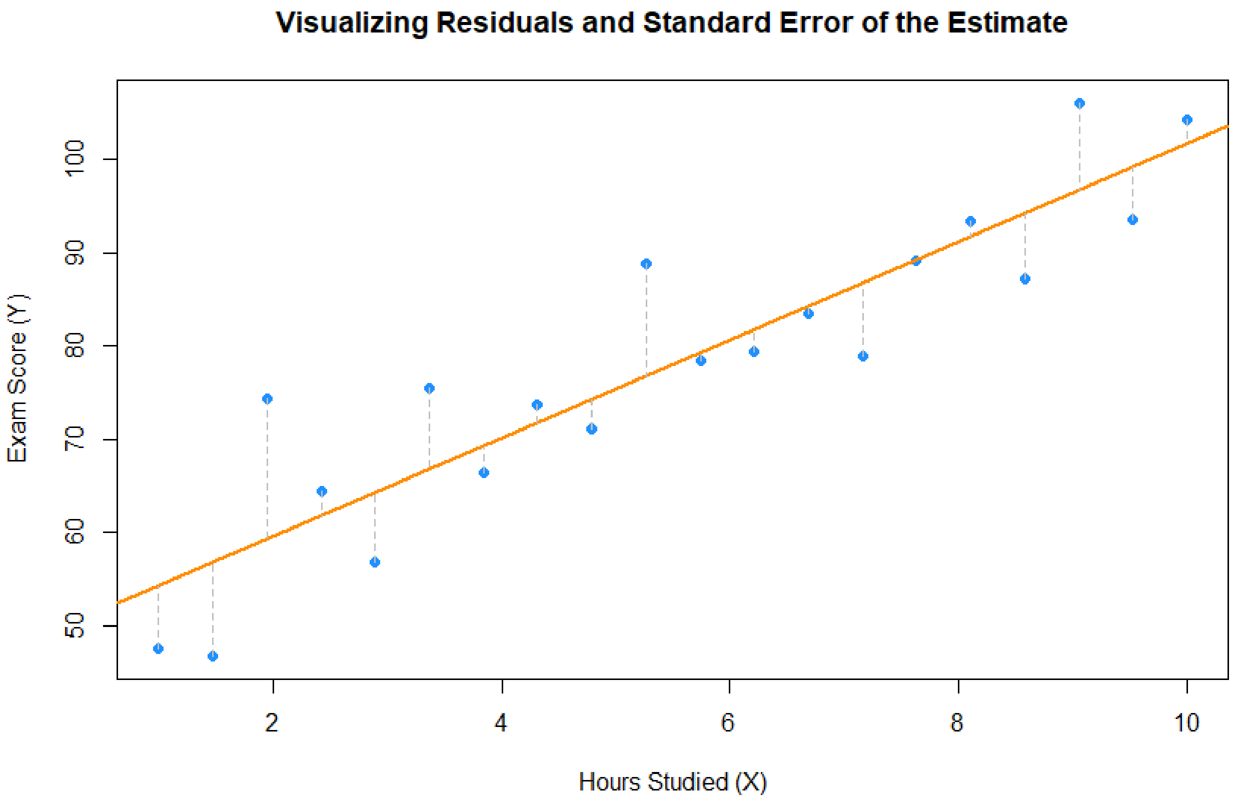
Blue dots: observed data points (actual Y value)

Orange line: regression line (predicted $\hat{Y}$ value)

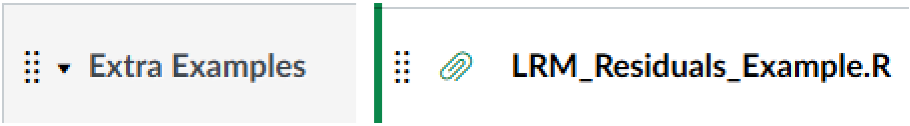
Gray dashes: vertical distance between each observed point and the regression line ($Y_i - \hat{Y}_i$)

The standard error of the estimate (SEE) is the standard deviation of the residuals

$$SEE = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{N - 2}}$$



Simple LRM continued

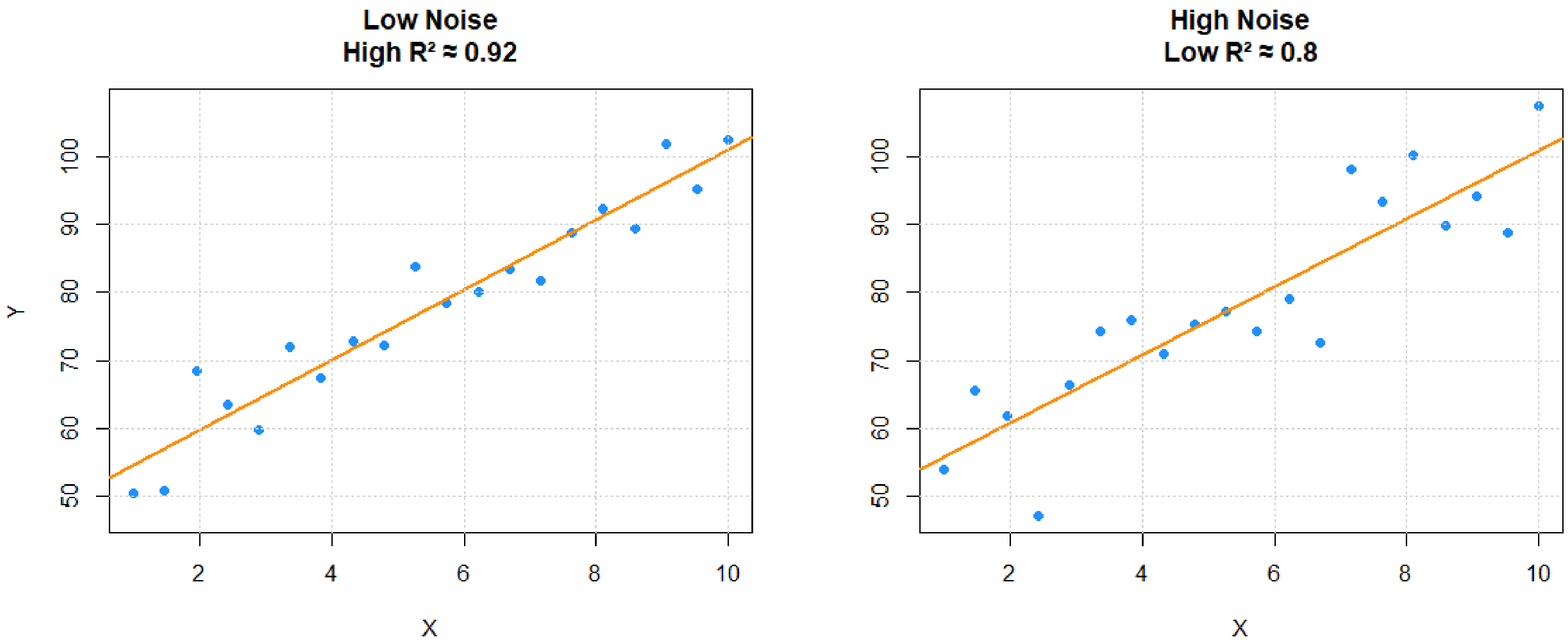


$$\hat{Y} = b_0 + b_1 X$$

Let's revisit R-squared (R^2), sometimes called the coefficient of determination

- R^2 is an estimate of the proportion of variance in Y explained by X
  - relatively larger residuals (e_i) result in a smaller R^2
    - less variance in Y explained by X

Consider the simulated example below with the same X but less vs more noisy Y



Simple LRM continued

$$\hat{Y} = b_0 + b_1 X$$

- R^2 is computed by decomposing the variance
  - like with ANOVA and Pearson's r

- Total Sum of Squares (TSS) = Explained Sum of Square (ESS) + Residual Sum of Squares (RSS)

$$TSS: \sum_{i=1}^{n}(y_i - \bar{y})^2 = \text{ESS}: \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + RSS: \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- TSS is the total variability in Y
- ESS is variance in Y explained by X
- RSS is variance in Y not explained by X

Let's return to our education and cohort example to explore why $R^2 = \frac{ESS}{TSS}$ …

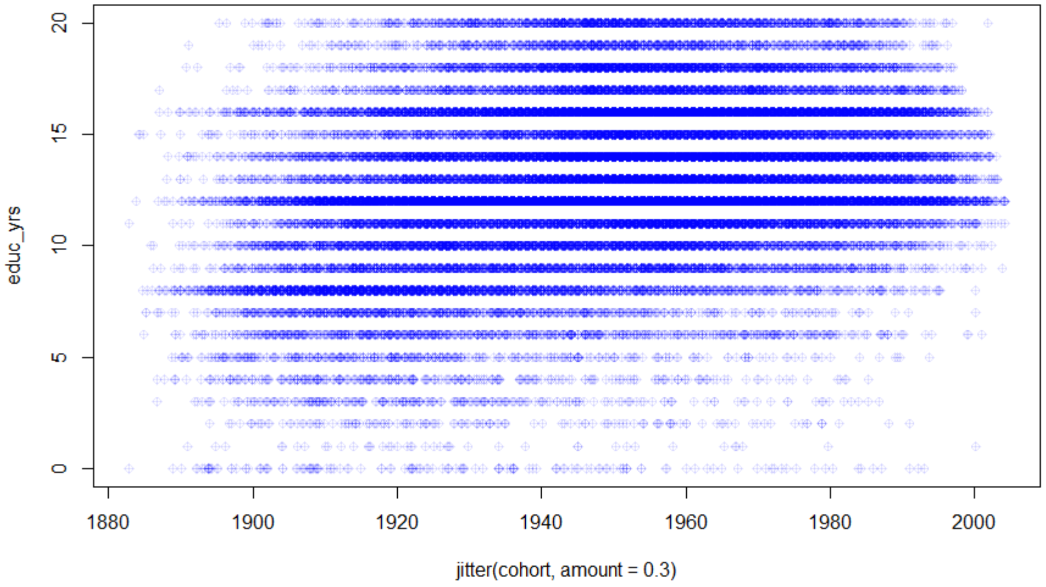
Simple LRM: example

R + RStudio instructions and tutorial

RScript_5.R

$$\hat{Y} = b_0 + b_1 X$$

Pearson's r = 0.297




```
cor(cohort,educ_yrs,method="pearson")
```

```
> cor(cohort,educ_yrs,method="pearson")
[1] 0.2970501
```

- Let's model a linear association using regression
- First, lets calculate the slope and intercept
    - and compute predicted values ($\hat{y}_i$)

slope

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

```
b1<-sum((cohort-mean_cohort)*(educ_yrs-mean_educ))/sum((cohort-mean_cohort)^2)
```

b_1 = 0.043164

intercept

$$b_0 = \bar{y} - b_1\bar{x}$$

```
b0<-mean_educ-b1*mean_cohort
```

b_0 = -71.19411

predicted values

$$\hat{y}_i = b_0 + b_1 x_i$$

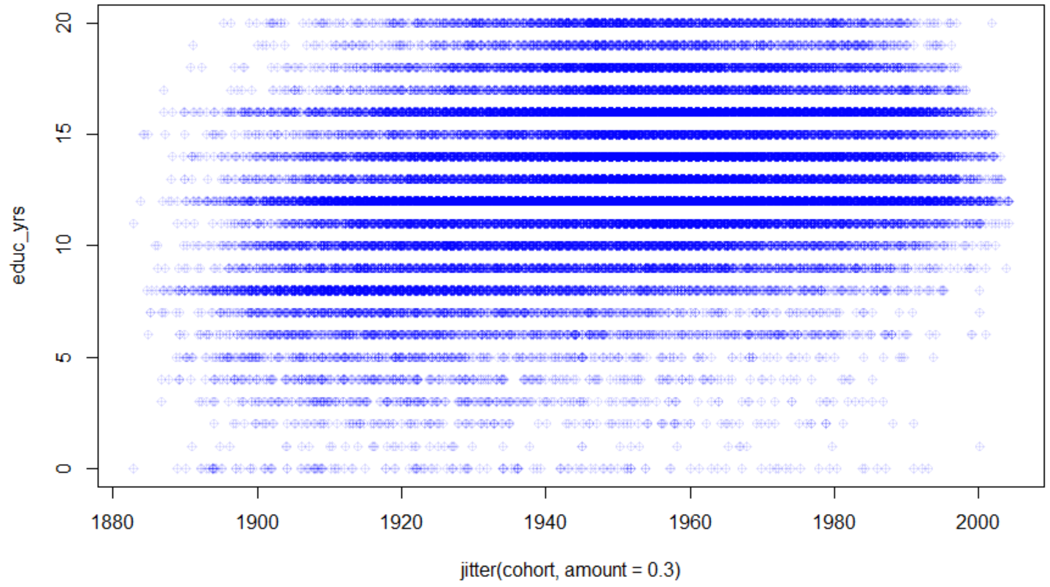
```
predicted_educ<-b0+b1*cohort
```

$\hat{y}_i$ = the estimated value of education for the *i*-th observation, based on its corresponding value of cohort

Simple LRM: example

$$\hat{Y} = b_0 + b_1 X$$

Pearson's r = 0.297




```
cor(cohort,educ_yrs,method="pearson")
```

```
> cor(cohort,educ_yrs,method="pearson")
[1] 0.2970501
```

- Next, let's calculate the total sum of squares (TSS), explained sum of square (ESS), and residual sum of squares (RSS)

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

TSS = 655,387.3

```
TSS<- sum((educ_yrs - mean_educ)^2)
```

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

ESS = 57,830.6

```
ESS<-sum((predicted_educ-mean_educ)^2)
```

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

RSS = 597,556.8

```
RSS<-sum((educ_yrs-predicted_educ)^2)
```

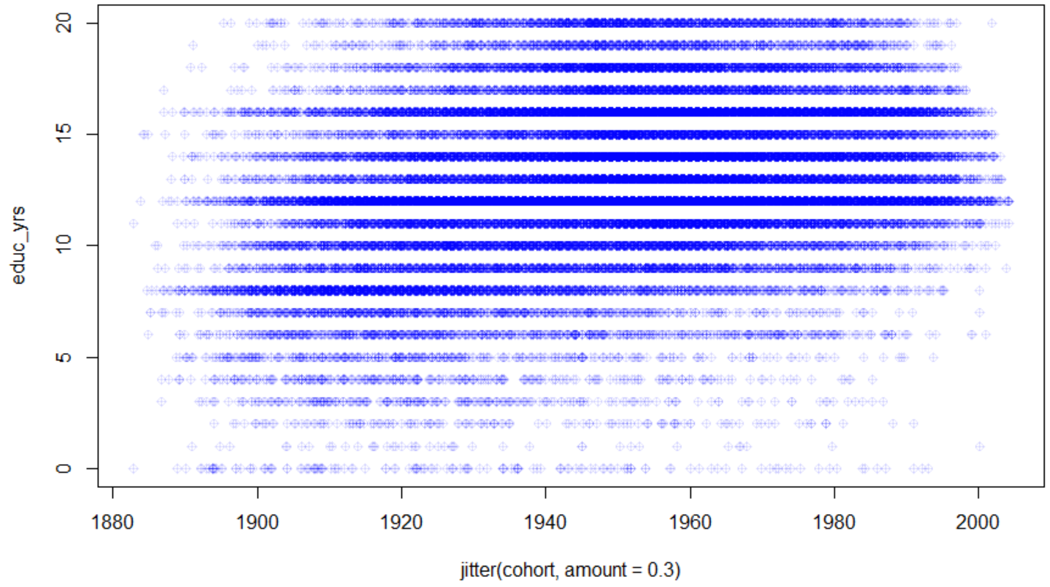
$$TSS = ESS + RSS$$

TSS = 57,830.6 + 597,556.8 = 655,387.3

Simple LRM: example

$$\hat{Y} = b_0 + b_1 X$$

Pearson's r = 0.297




```
790   cor(cohort,educ_yrs,method="pearson")
```

```
> cor(cohort,educ_yrs,method="pearson")
[1] 0.2970501
```

- Now we can decompose the sums of squares to determine how much of the total variance in Y was explained by X in the linear regression model
  - R-squared (R^2)

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$R^2 = \frac{57{,}830.6}{655{,}387.3} = 0.088\text{, alternatively}$$

$$R^2 = 1 - \frac{597{,}556.8}{655{,}387.3} = 0.088\text{, also}$$

$$R^2 = Pearson's\ r^2$$

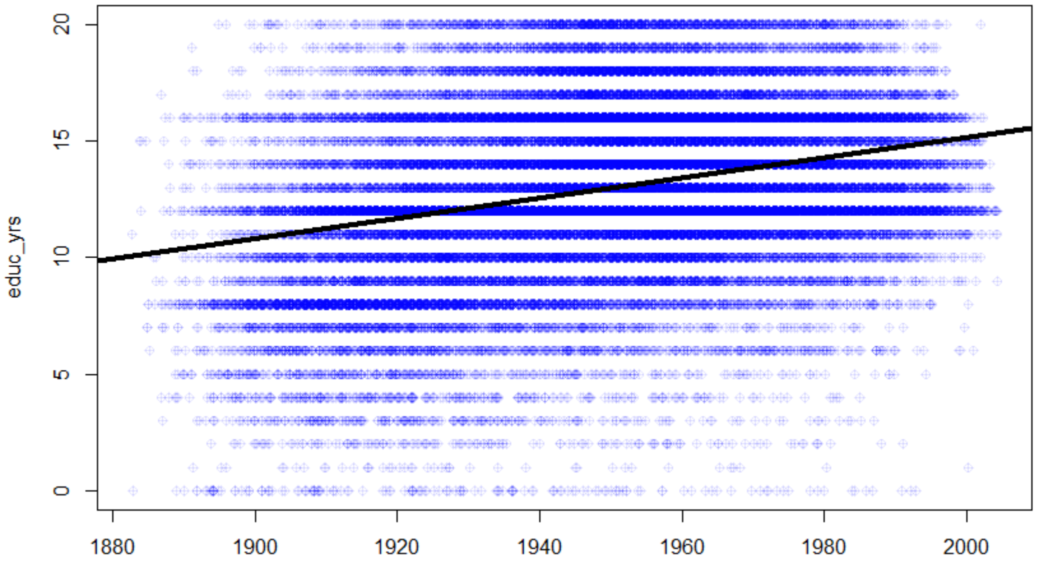
$$R^2 = r^2 = 0.297^2 = 0.088$$

If a simple linear regression model provides the same R^2 estimate, then why not just use Pearson's r^2?

Simple LRM: example

$$\hat{Y} = b_0 + b_1 X$$

R^2 = 0.088




```
898  # add the regression line to the X Y plot
908  abline(model,lwd=4)
```

- intercept coefficient (b_0) is the value of Y when X = 0
- slope coefficient (b_1) is the change in Y for a one-unit change in X

```
892  model<-lm(educ_yrs~cohort)
896  print(summary(model))

              Estimate  Std. Error t value             Pr(>|t|)
(Intercept) -71.1941083   1.0649655  -66.85 <0.0000000000000002 ***
cohort        0.0431640   0.0005461   79.04 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.043 on 64553 degrees of freedom
Multiple R-squared:  0.08824,   Adjusted R-squared:  0.08822
F-statistic:  6247 on 1 and 64553 DF,  p-value: < 0.00000000000000022
```

- So, the predicted value of edu for the first cohort is

$$\hat{Y} = -71.19411 + 0.043164 \times 1880 = 9.95421$$

- and the predicted value of edu for the last cohort is

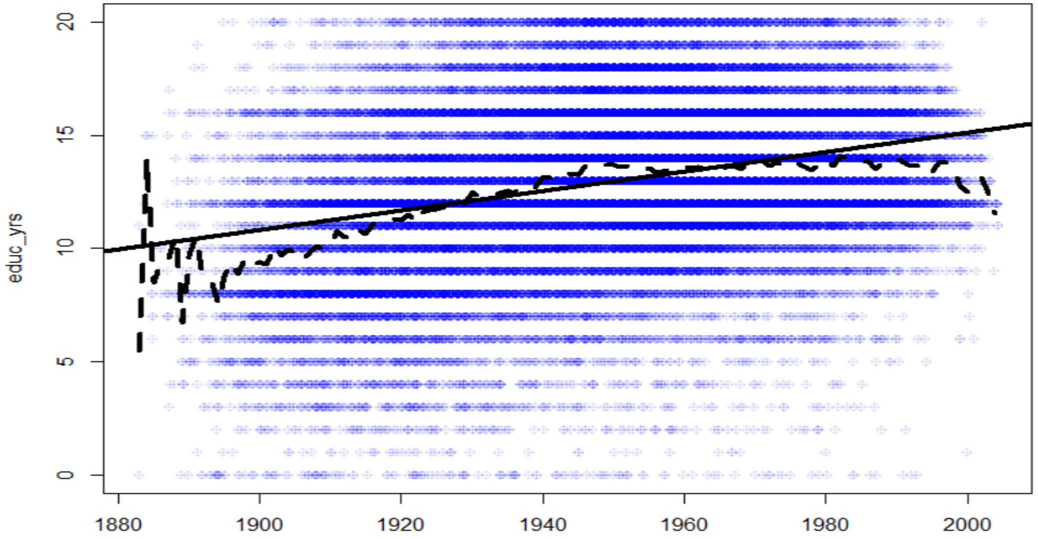
$$\hat{Y} = -71.19411 + 0.043164 \times 2004 = 15.30655$$

- The LRM minimized the sum of squared residuals
  - difference between the observed and predicted values
- Best guess for value of Y given X

Simple LRM: example

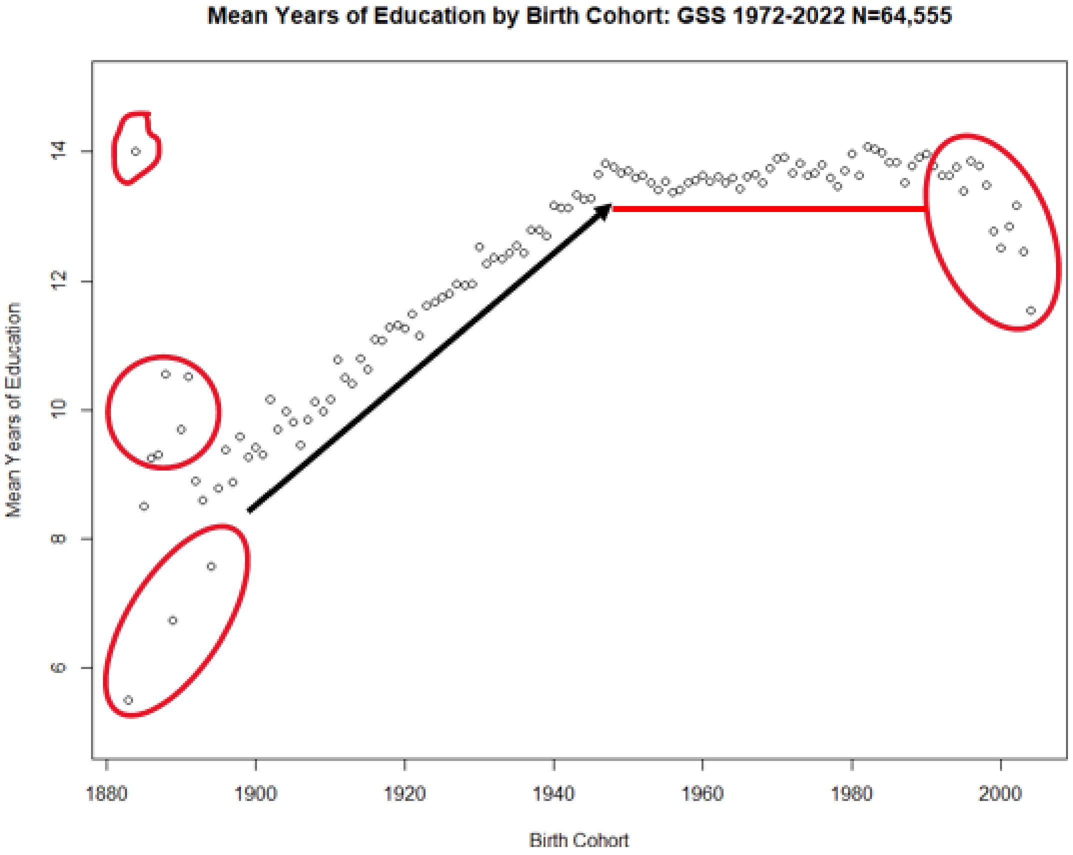
$$\hat{Y} = b_0 + b_1 X$$

$R^2 = 0.088$




```
910 lines(educ_mean_by_cohort$cohort,educ_mean_by_cohort$educ_yrs,lty=2,lwd=4)
```

Recall from PPT 11



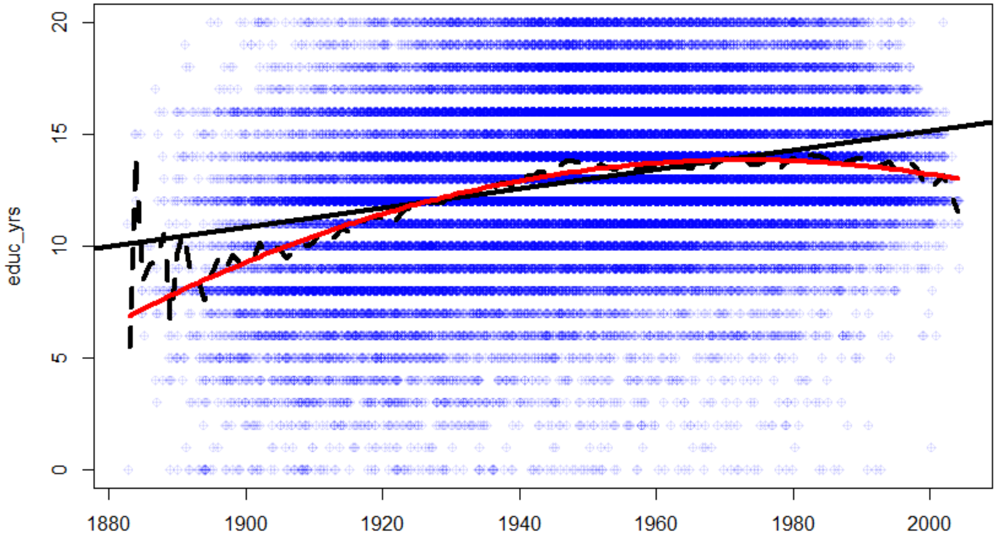
- Although difficult to see in scatter plot, we know from previous analyses that this relationship is not linear
  - let's overlay the cohort-specific mean pattern in education

Can a linear regression model estimate non-linear patterns?

~~Simple~~ LRM: example

$$\hat{Y} = b_0 + b_1 X \boxed{+ b_2 X}$$

$R^2 = 0.1152$




```
# overlay the regression line from the polynomial model
lines(cohort_seq, predicted_vals, col = "red", lwd = 4)
```

```
model_2b<-lm(educ_yrs~cohort + cohort_sqr)
print(summary(model_2b))
```

```
               Estimate     Std. Error t value             Pr(>|t|)
(Intercept) -3340.14951755   73.71866467  -45.31 <0.0000000000000002 ***
cohort          3.40020576    0.07569936   44.92 <0.0000000000000002 ***
cohort_sqr     -0.00086177    0.00001943  -44.35 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.997 on 64552 degrees of freedom
Multiple R-squared:  0.1152,	Adjusted R-squared:  0.1152
F-statistic:  4202 on 2 and 64552 DF,  p-value: < 0.00000000000000022
```

- So, the predicted value of edu for the first cohort is

$$\hat{Y} = -3340.15 + 3.40 \times 1880 + (-0.00086 \times 1880^2) = 6.40$$

and the predicted value of edu for the last cohort is

$$\hat{Y} = -3340.15 + 3.40 \times 2004 + (-0.00086 \times 2004^2) = 12.98$$

- One way to address non-linear patterns in a LRM is to add polynomial transformation(s) of X
  - notice that the amount of explained variance (R^2) increased

With more than one X this is now a multiple linear regression model…

LRM continued

$$Y = \beta_0 + \beta_1 X + \beta_2 X \ldots + \beta_k X + \varepsilon$$

Let’s consider hypothesis testing before moving further toward multiple regression

- F-test: is the overall model statistically significant?
  - at least one of the βX’s is not equal to 0

H_0: $\beta_1 = \beta_2 \ldots = \beta_k = 0$

H_a: At least one $\beta_j \neq 0$

$$F = \frac{Mean\ Saure\ Estimate}{Mean\ Square\ Residual} = \frac{ESS/df\_regression}{RSS/df\_residual}$$

$df\_regression$: number of predictors (k) in the model

$df\_residual$: $n - k - 1$

- If F is large and corresponding p-value is small ($\leq \alpha$) then reject null hypothesis (H_0)

What about particular parameters (β)?

LRM continued

$$Y = \beta_0 + \beta_1 X + \beta_2 X \ldots + \beta_k X + \varepsilon$$

- Hypothesis test for a regression coefficient (β_j)

  $H_0\!: \ \beta_j = 0$

  $H_a\!: \beta_j \neq 0$

$$t = \frac{\beta_j}{SE(\beta_j)}$$

Compare the t-statistic to the t-distribution with $df = n - k - 1$

Or use the associated p-value

$$SE(\hat{\beta}_1) = \frac{SEE}{\sqrt{\sum(x_i - \bar{x}_i)^2}}$$

Recall, standard error of the estimate (SEE): how far off, on average, predicted values are from actual Y values.

The $SE(\hat{\beta}_1)$ is also used to construct confidence intervals around the estimate…

Simple LRM example

$$\hat{Y} = b_0 + b_1 X$$

Let's check this out CIs with our example using education (Y) and cohort (X)

```
               Estimate  Std. Error t value            Pr(>|t|)    
(Intercept) -71.1941083   1.0649655  -66.85 <0.0000000000000002 ***
cohort        0.0431640   0.0005461   79.04 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.043 on 64553 degrees of freedom
Multiple R-squared:  0.08824,	Adjusted R-squared:  0.08822 
F-statistic:  6247 on 1 and 64553 DF,  p-value: < 0.00000000000000022
```

$$CI = \hat{\beta}_j \pm Z^* or\ t^* \times SE(\hat{\beta}_j)$$

- 95% CI = 0.0431640 $\pm$ 1.96 $\times$ 0.0005461 = 0.0431640 $\pm$ 0.00107
  - 95% CI = (0.042,0.044)

```
951  confint(model, level = 0.95)

                   2.5 %      97.5 %
(Intercept) -73.28144137 -69.10677516
cohort        0.04209363   0.04423436
```

LRM interpretation

Differs based on level of measurement of X

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Continuous: change in Y for a one-unit change in X
- Binary (0,1): difference in Y between group =1 and group = 0
  - sometimes called "dummy variable"
- Categorical: difference in Y between each category and the reference group

Let's practice some examples…

Simple LRM interpretation: example

Let’s try another continuous X variable

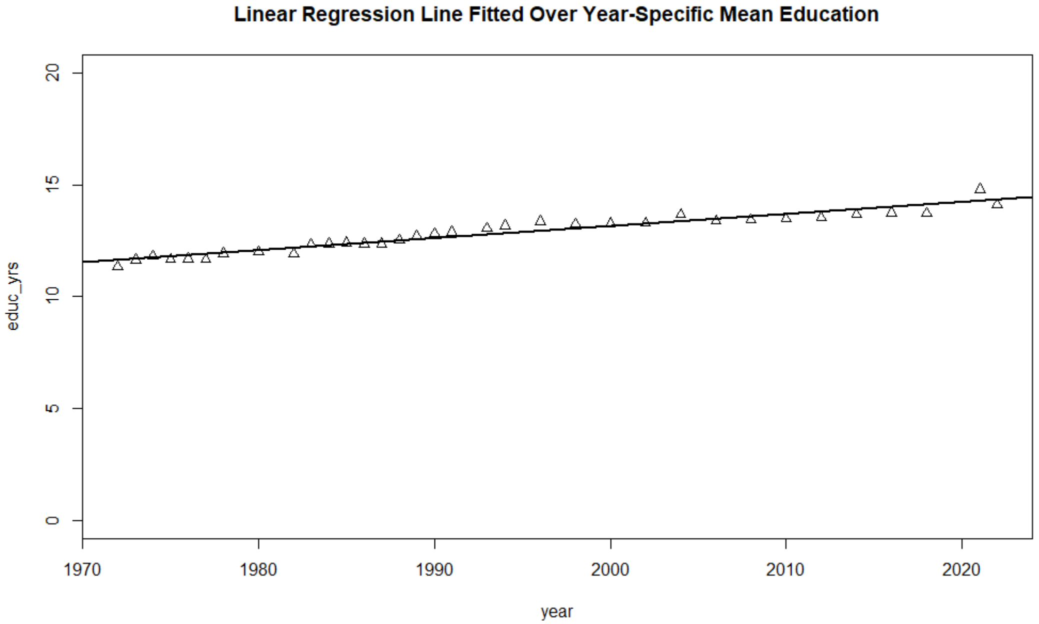
$$\hat{Y} = b_0 + b_1 X$$

- X: year (range 1972 – 2022)
  - year that survey was administered

```
957  model_eduyear<-lm(educ_yrs~year)

              Estimate  Std. Error t value            Pr(>|t|)
(Intercept) -94.9585105  1.6075532  -59.07 <0.0000000000000002 ***
year          0.0540629  0.0008052   67.14 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.081 on 64553 degrees of freedom
Multiple R-squared:  0.06528,	Adjusted R-squared:  0.06526
F-statistic:  4508 on 1 and 64553 DF,  p-value: < 0.00000000000000022
```



- What is predicted value of education ($\hat{Y}$) when year (X) = 0?
- Can you interpret the slope coefficient (β_1)?
- What is $\hat{Y}$ when year = 1972?
  - What is $\hat{Y}$ when year = 2022?

Simple LRM interpretation: example

Let's try a binary X variable

$$\hat{Y} = b_0 + b_1 X$$

- X: race (0=non-white, 1=white)

```
975  model_edurace<-lm(educ_yrs~white)

             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 12.40271    0.02823  439.39 <0.0000000000000002 ***
white        0.71231    0.03148   22.63 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.174 on 64553 degrees of freedom
Multiple R-squared:  0.007871,  Adjusted R-squared:  0.007855
F-statistic: 512.1 on 1 and 64553 DF,  p-value: < 0.00000000000000022

> confint(model_edurace, level = 0.95)
                 2.5 %     97.5 %
(Intercept) 12.3473800 12.4580305
white        0.6506191  0.7740091
```

- What is predicted value of education ($\hat{Y}$) when white (X) = 0?
- Can you interpret the slope coefficient (β_1)?
- What is $\hat{Y}$ when white = 1?
    - Can you interpret the 95% CI for β_1?

LRM interpretation: example

Let's try a categorical X variable with three or more categories

$$\hat{Y} = b_0 + b_1 X + b_2 X$$

- X: political party affiliation (1=democrat, 2=independent, 3=republican)
  - one category must be excluded from the model to be the reference group

```
987  model_edupolit<-lm(educ_yrs~relevel(factor(partyafil),ref="Rep"))

                                    Estimate Std. Error t value            Pr(>|t|)
(Intercept)                         13.35902    0.02145  622.92 <0.0000000000000002 ***
relevel(factor(partyafil), ref = "Rep")Dem -0.50826    0.02790  -18.22 <0.0000000000000002 ***
relevel(factor(partyafil), ref = "Rep")Ind -0.78626    0.03703  -21.23 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.173 on 64552 degrees of freedom
Multiple R-squared:  0.008398,	Adjusted R-squared:  0.008367 
F-statistic: 273.4 on 2 and 64552 DF,  p-value: < 0.00000000000000022
```

- What is predicted value of education ($\hat{Y}$) for the reference category?
- Can you interpret the slope coefficients (β_1, β_2)?
- What is $\hat{Y}$ when partyafil = 1 "Dem"?
  - What is $\hat{Y}$ when partyafil = 2 "Ind"?

Let's change the reference group and see what happens…

LRM interpretation: example

Let's try a categorical X variable with three or more categories

$$\hat{Y} = b_0 + b_1X + b_2X$$

- X: political party affiliation (1=democrat, 2=independent, 3=republican)
  - one category must be excluded from the model to be the reference group

```
993  model_edupolit<-lm(educ_yrs~relevel(factor(partyafil),ref="Dem"))
```

```
                                      Estimate Std. Error t value              Pr(>|t|)
(Intercept)                           12.85077    0.01784 720.133 < 0.0000000000000002 ***
relevel(factor(partyafil), ref = "Dem")Ind -0.27801    0.03506  -7.928  0.00000000000000226 ***
relevel(factor(partyafil), ref = "Dem")Rep  0.50826    0.02790  18.218 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.173 on 64552 degrees of freedom
Multiple R-squared:  0.008398,  Adjusted R-squared:  0.008367
F-statistic: 273.4 on 2 and 64552 DF,  p-value: < 0.00000000000000022
```

- What is predicted value of education ($\hat{Y}$) for the reference category?
- Can you interpret the slope coefficients (β_1, β_2)?
- What is $\hat{Y}$ when partyafil = 2 “Ind”?
  - What is $\hat{Y}$ when partyafil = 3 “Rep”?

Let's see what this looks like if we don't treat X as a categorical variable…

LRM interpretation: example

Why a reference group is needed for categorical X with 3+ categories

$$\hat{Y} = b_0 + b_1 X + b_2 X$$

- X: political party affiliation (1=democrat, 2=independent, 3=republican)
  - one category must be excluded from the model to be the reference group

```
999  model_edupolit<-lm(educ_yrs~polit_party)
```

```
             Estimate Std. Error t value             Pr(>|t|)    
(Intercept) 12.53703    0.02865  437.59 <0.0000000000000002 ***
polit_party  0.23711    0.01394   17.01 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.179 on 64553 degrees of freedom
Multiple R-squared:  0.004464,	Adjusted R-squared:  0.004449 
F-statistic: 289.5 on 1 and 64553 DF,  p-value: < 0.00000000000000022
```

- Can you interpret the intercept coefficient (β_0)?
- Can you interpret the slope coefficient (β_1)?