Week 8: Simple Linear Regression

Modeling Linear Associations Between Variables

Scott Schwartz

Oct 31, 2022

Scatter Plot Visualization of Variable Association

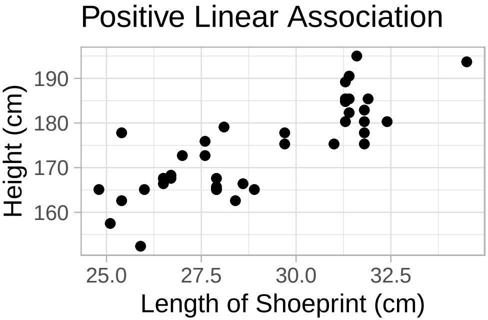
Estimation of stature from foot and shoe length: applications in forensic science

- by B. Rohren

```r
library(tidyverse)
heights <- read_csv("heights.csv")
glimpse(heights)
```

```
## Rows: 40
## Columns: 7
## $ ...1       <dbl> 1, 2, 3, 4, 5, 6, 7, 8,
## $ sex        <chr> "M", "M", "M", "M", "M",
## $ age        <dbl> 67, 47, 41, 42, 48, 34,
## $ footLength <dbl> 27.8, 25.7, 26.7, 25.9,
## $ shoePrint  <dbl> 31.3, 29.7, 31.3, 31.8,
## $ shoeSize   <dbl> 11.0, 9.0, 11.0, 10.0, 1
## $ height     <dbl> 180.3, 175.3, 184.8, 177
```

```r
# {r, fig.width=3, fig.height=2}
heights %>%
  ggplot(aes(x=shoePrint, y=height)) +
  geom_point() + theme_light() +
  labs(title="Positive Linear Association",
       x="Length of Shoeprint (cm)",
       y="Height (cm)")
```



Scatter Plot Visualization of Variable Association

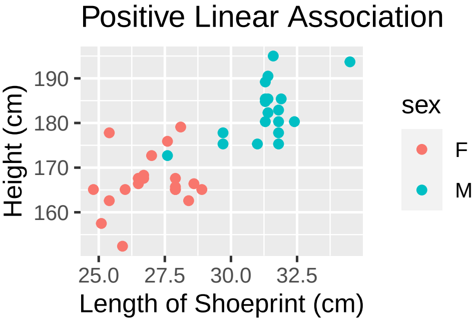
Estimation of stature from foot and shoe length: applications in forensic science

- by B. Rohren

```r
library(tidyverse)
heights <- read_csv("heights.csv")
glimpse(heights)
```

```
## Rows: 40
## Columns: 7
## $ ...1       <dbl> 1, 2, 3, 4, 5, 6, 7, 8,
## $ sex        <chr> "M", "M", "M", "M", "M",
## $ age        <dbl> 67, 47, 41, 42, 48, 34,
## $ footLength <dbl> 27.8, 25.7, 26.7, 25.9,
## $ shoePrint  <dbl> 31.3, 29.7, 31.3, 31.8,
## $ shoeSize   <dbl> 11.0, 9.0, 11.0, 10.0, 1
## $ height     <dbl> 180.3, 175.3, 184.8, 177
```

```r
# {r, fig.width=3, fig.height=2}
heights %>% ggplot(aes(
  x=shoePrint, y=height, color=sex)) +
  geom_point() + theme_gray() +
  labs(title="Positive Linear Association",
       x="Length of Shoeprint (cm)",
       y="Height (cm)")
```



Scatter Plot Visualization of Variable Association

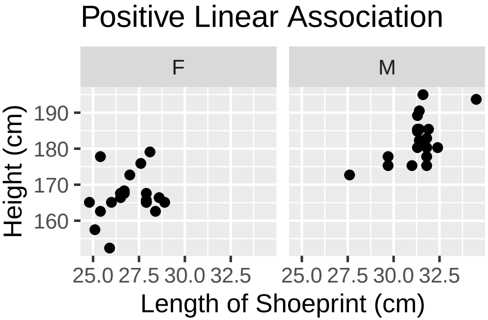
Estimation of stature from foot and shoe length: applications in forensic science

- by B. Rohren

```r
library(tidyverse)
heights <- read_csv("heights.csv")
glimpse(heights)
```

```
## Rows: 40
## Columns: 7
## $ ...1       <dbl> 1, 2, 3, 4, 5, 6, 7, 8,
## $ sex        <chr> "M", "M", "M", "M", "M",
## $ age        <dbl> 67, 47, 41, 42, 48, 34,
## $ footLength <dbl> 27.8, 25.7, 26.7, 25.9,
## $ shoePrint  <dbl> 31.3, 29.7, 31.3, 31.8,
## $ shoeSize   <dbl> 11.0, 9.0, 11.0, 10.0, 1
## $ height     <dbl> 180.3, 175.3, 184.8, 177
```

```r
# {r, fig.width=3, fig.height=2}
heights %>%
  ggplot(aes(x=shoePrint, y=height)) +
  geom_point() + facet_wrap(~sex) +
  labs(title="Positive Linear Association",
       x="Length of Shoeprint (cm)",
       y="Height (cm)")
```



\+ facet_wrap() / facet_grid()

Histogram Visualization of Variable Association

Estimation of stature from foot and shoe length: applications in forensic science

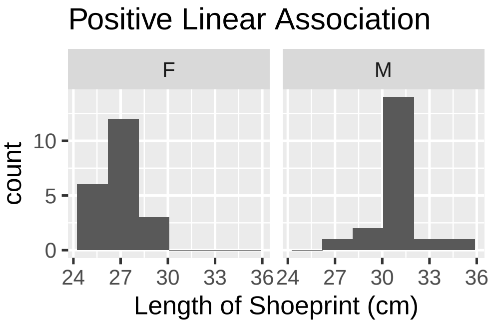
- by B. Rohren

```
library(tidyverse)
heights <- read_csv("heights.csv")
glimpse(heights)
```

```
## Rows: 40
## Columns: 7
## $ ...1       <dbl> 1, 2, 3, 4, 5, 6, 7, 8,
## $ sex        <chr> "M", "M", "M", "M", "M",
## $ age        <dbl> 67, 47, 41, 42, 48, 34,
## $ footLength <dbl> 27.8, 25.7, 26.7, 25.9,
## $ shoePrint  <dbl> 31.3, 29.7, 31.3, 31.8,
## $ shoeSize   <dbl> 11.0, 9.0, 11.0, 10.0, 1
## $ height     <dbl> 180.3, 175.3, 184.8, 177
```

+ facet_wrap() / facet_grid()

```
# {r, fig.width=3, fig.height=2}
heights %>%
  ggplot(aes(x=shoePrint)) +
  geom_histogram(bins=6) +
  facet_wrap(~sex) +
  labs(title="Positive Linear Association",
       x="Length of Shoeprint (cm)")
```



Sample Correlation r: Linear Association in (x_i, y_i)

❶ First, **Correlation is not Causation**

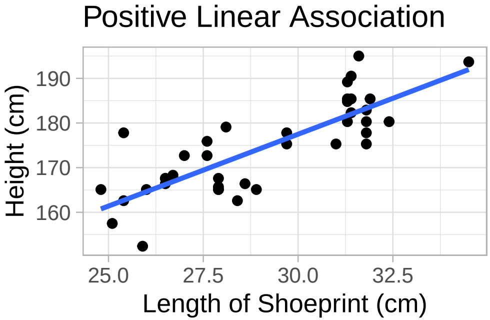
❷ $$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{(n-1)s_{xy}}{\sqrt{(n-1)s_x^2 (n-1)s_y^2}} = \frac{s_{xy}}{s_x s_y} = \frac{\text{Cov}(x,y)}{\text{SD}(x)\text{SD}(y)}$$

```r
cor(heights$shoePrint, heights$height)
```

```
## [1] 0.812948
```

- *Play "guess the correlation" at* http://www.istics.net/Correlations/

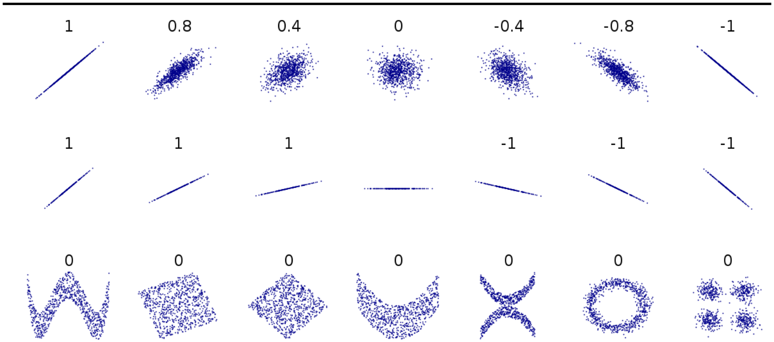
```r
# {r, fig.width=3, fig.height=2}
heights %>% ggplot(aes(
  x=shoePrint, y=height)) + geom_point() +
  geom_smooth(method=lm, se=FALSE) +
  labs(title="Positive Linear Association",
       x="Length of Shoeprint (cm)",
       y="Height (cm)") + theme_light()
```



Sample Correlation r: Linear Association in (x_i, y_i)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{(n-1)s_{xy}}{\sqrt{(n-1)s_x^2(n-1)s_y^2}} = \frac{s_{xy}}{s_x s_y} = \frac{\mathrm{Cov}(x,y)}{\mathrm{SD}(x)\mathrm{SD}(y)}$$

- The denominator scales the numerator so that the total $-1 \leq r \leq 1$ always
- r measures *linear association*, with $r > 0$ *positive* and $r < 0$ means *negative*

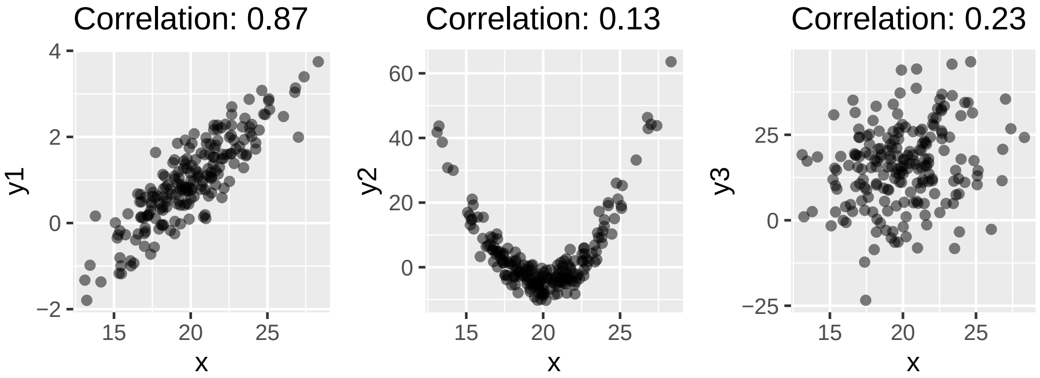


Why no Scale?

Does it matter?

Sample Correlation r: Linear Association in (x_i, y_i)

Is x more strongly associated with $y1$, $y2$, or $y3$?



Correlation is not Causation but it is also only measures **LINEAR Association**

Simple Linear Regression is a Normal Model

$$E[Y_i|x_i] = \beta_0 + \beta_1 x_i$$

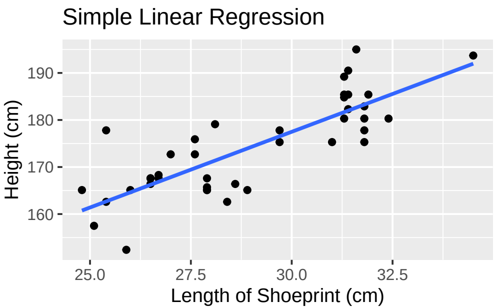
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Parameters of the Normal Model

- β_0: 'Y if x is 0' ***intercept parameter***
- β_1: 'rise over run' ***slope parameter***
- ϵ_i: the 'noise' or 'error' term
- x_i: predictor, feature, covariate, or explanatory or independent variable; *as if a parameter was known without error*
- Y_i: response, outcome or dependent variable

```r
# https://ggplot2.tidyverse.org/reference/
# geom_smooth.html
heights %>% ggplot(aes(
  x=shoePrint,y=height)) + geom_point() +
  geom_smooth(method=lm, se=FALSE) +
  labs(title="Simple Linear Regression",
       x="Length of Shoeprint (cm)",
       y="Height (cm)")
```



Simple Linear Regression is a Normal Model

$$E[Y_i|x_i] = \beta_0 + \beta_1 x_i$$

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

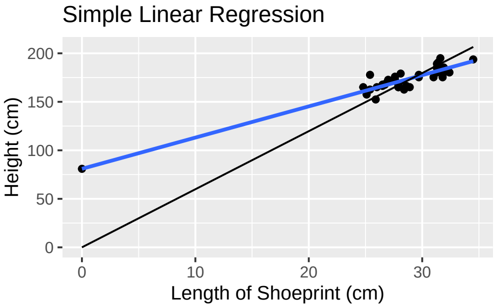
$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Parameters of the Normal Model

- β_0: 'Y if x is 0' ***intercept parameter***
- β_1: 'rise over run' ***slope parameter***
- ϵ_i: the 'noise' or 'error' term
- x_i: predictor, feature, covariate, or explanatory or independent variable; *as if a parameter was known without error*
- Y_i: response, outcome or dependent variable

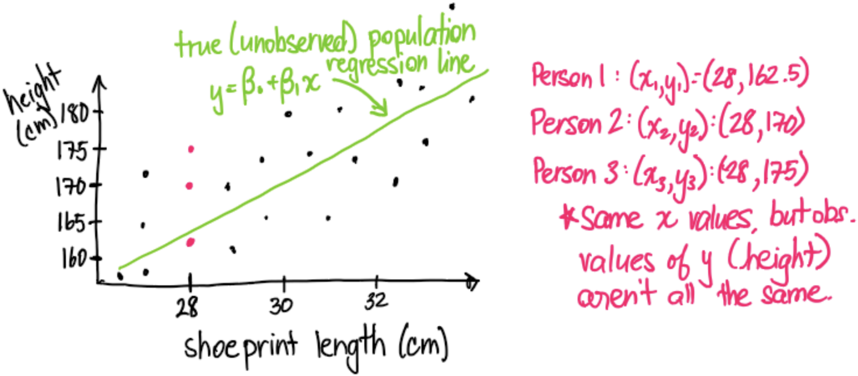
```
intercept <- lm(height~shoePrint, data=heights)$coeff[1]
# "-1" means a "zero intercept" model going through (0,0)
lm(height ~ -1 + shoePrint, data=heights)$coeff[1] ->
  slope_if_intercept_is_0; coef0 <- slope_if_intercept_is_0

heights %>% add_row(tibble(shoePrint=0, height=intercept)) %>%
  ggplot(aes(x=shoePrint,y=height)) + geom_point() +
  geom_smooth(method=lm, se=FALSE) +
  geom_segment(aes(x=x, y=y, xend=xend, yend=yend),
    data=tibble(x=0, y=0, xend=34.5, yend=34.5*coef0)) +
  labs(title="Simple Linear Regression", y="Height (cm)",
       x="Length of Shoeprint (cm)")
```



Simple Linear Regression is a Normal Model

For $x_i = x_j$ we will have $y_i \neq y_j$

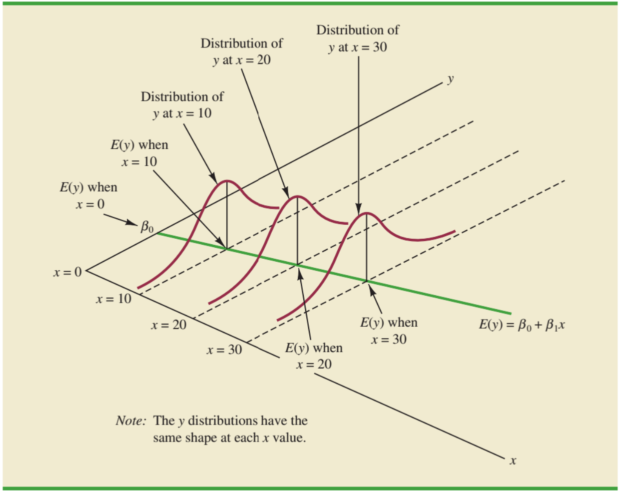


The expected value is $E[Y_i] = \beta_0 + \beta_1 x_i$ but the actual value is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

Simple Linear Regression is a Normal Model

For $x_i = x_j$ we will have $y_i \neq y_j \quad \rightarrow \quad y_i = E[Y_i|x_i] + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$



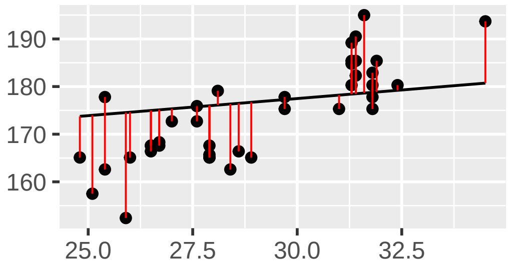
Fitting Simple Linear Regression Models

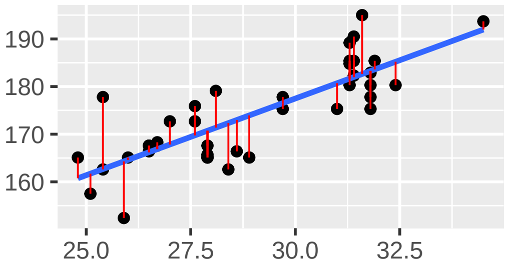
$$\min_{\hat\beta_0,\hat\beta_1} \sum_{i=1}^{n} (y_i - \hat y_i)^2 = \min_{\hat\beta_0,\hat\beta_1} \sum_{i=1}^{n} \hat\epsilon_i^2$$

where $\hat y_i = \hat\beta_0 + \hat\beta_1 x_i$

estimates $E[Y_i|x_i] = \beta_0 + \beta_1 x$

Method of Least Squares: minimize the sum of the squared residuals





- Why square the *residual* $\hat\epsilon_i = y_i - \hat y_i$?
- $\hat\beta_0 = 80.930$ and $\hat\beta_1 = 3.219$
- What are the *three largest residuals* under the fit model on the right?

Fitting Simple Linear Regression Models

$$\hat\beta_1 = r\frac{s_y}{s_x} = \frac{\sum_{i=1}^n (x_i-\bar x)(y_i-\bar y)}{\sqrt{\sum_{i=1}^n (x_i-\bar x)^2 \sum_{i=1}^n (y_i-\bar y)^2}} \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i-\bar y)^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i-\bar x)^2}}$$

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

$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x$$

What is the *predicted* expected value of Y_i $E[Y_i|x_i] = \beta_0 + \beta_1 x_i$ if $x_i = 30$ cm? ***How*** much taller on average are people with 5 cm larger shoe prints?

$\hat\beta_0$ and $\hat\beta_1$ estimates of β_0 and β_1 and the `Response ~ Predictor` Notation

```r
least_squares_fit <- lm( height ~ shoePrint , data=heights) # "response ~ predictor"
# least_squares_fit %>% summary() %>% broom::tidy() # library(broom)
summary(least_squares_fit)$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 80.930409 10.8933945 7.429310 6.504368e-09
## shoePrint    3.218561  0.3740081 8.605591 1.863474e-10
```

Simple Linear Regression can be Interpreted

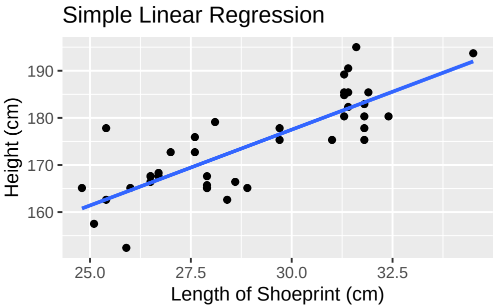
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_i = 80.93 + 3.22 x_i$$

$$\hat{\epsilon}_i = y_i - (80.93 + 3.22 x_i)$$

all based on the idea (model) that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon \;/\; y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$



The **parameters** of a **regression model** are often called ***regression coefficients***

The *slope coefficient* is interpreted as "difference in Y per unit change in x"

"`Height` *increases* by 3.22 cm **on average** per 1 cm *increase* in `shoePrint` length"

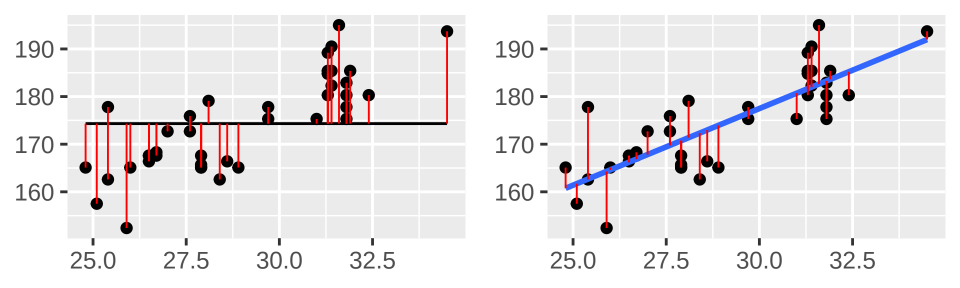
Linear model $\hat{\beta}_1 = r\frac{s_y}{s_x}$ associations are not causal: **"correlation is not causation"**

Should we predict (extrapolate) the expected value of Y_i using this data if $x_i = 10$ cm?

The Coefficient of Determination R^2

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = r^2$$

The "Proportion of Variation Explained" is a *Measure of Model Fit*

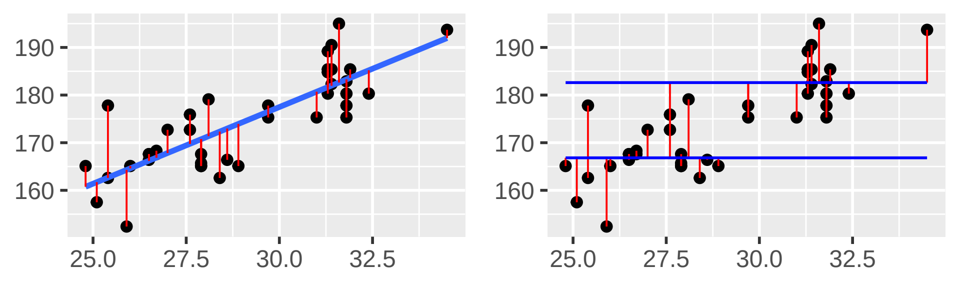


- What is the proportion reduction in the *squared residuals*?
  [R^2 is equal to r^2 only for Simple Linear Regression]

The Coefficient of Determination R^2

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = r^2$$

Which of these looks like a better explanatory model?



- What is the proportion reduction in the *squared residuals* for each?
  [The two lines are "Male" and "Female": is this Simple Linear Regression?]

The Coefficient of Determination R^2

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = r^2$$

The "Proportion of Variation Explained" is a *Measure of Model Fit*

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

```
summary(least_squares_fit)$r.squared
```

```
## [1] 0.6608845
```

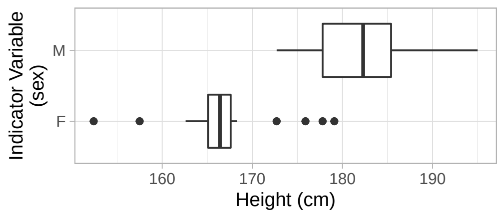
```
#for Simple Linear Regression that equals
cor(heights$shoePrint, heights$height)^2
```

```
## [1] 0.6608845
```

```
least_squares_fit2 <- # different "x"
  lm(height~sex, data=heights)
summary(least_squares_fit2)$r.squared
```

```
## [1] 0.6283874
```

```
heights %>% ggplot(aes(y=sex, x=height))+
  geom_boxplot() + theme_light() +
  labs(y="Indicator Variable\n(sex)",
       x="Height (cm)")
```



How Indicator Variables Work

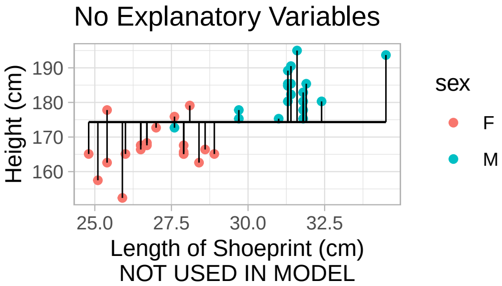
```
##
## Call:
## lm(formula = height ~ sex, data = heights)
##
## Coefficients:
## (Intercept)         sexM
##      166.82        15.79
```

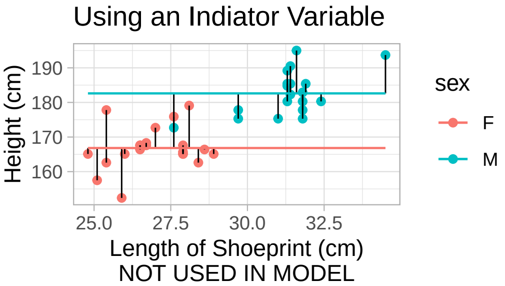
- $\hat{\beta}_1$ is called a ***contrast*** as it captures a difference between groups

$$Y_i = \beta_0 + \beta_1 \mathrm{I}(x_i = \mathrm{M}) + \epsilon_i$$

$$\hat{y}_i = 166.82 + 15.79\mathrm{I}(x_i = \mathrm{M})$$

$$\hat{y}_i = \begin{cases} 166.82[\hat{\beta}_0] + 15.79[\hat{\beta}_1] & \text{if sex} = \mathrm{M} \\ 166.82(\hat{\beta}_0) \quad [\text{baseline}] & \text{otherwise} \end{cases}$$





How Indicator Variables Work

```r
# default lm(height~I(sex=="M")) could specify lm(height~I(sex=="F"))
lm(height~sex, data=heights)$coefficients
```

```
## (Intercept)        sexM
##   166.82381    15.79198
```

$$Y_i = \beta_0 + \beta_1 \mathsf{I}(x_i = \mathsf{M}) + \epsilon_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathsf{I}(x_i = \mathsf{M})$$

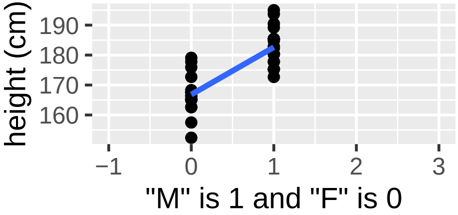
$$\mathsf{I}(x_i = \mathsf{M}) = \begin{cases} 1 & \text{if sex} = \mathsf{M} \\ 0 \quad \text{[baseline]} & \text{otherwise} \end{cases}$$

$$\hat{y}_i = \begin{cases} (\hat{\beta}_0 + \hat{\beta}_1) & \text{if sex} = \mathsf{M} \\ \hat{\beta}_0 \quad \text{[baseline]} & \text{otherwise} \end{cases}$$

How Indicator Variables Work

```
tibble(sexM=as.numeric(heights$sex=="M"),
       y=heights$height) -> version_1
lm(y~sexM, data=version_1) %>%
  tidy() %>% select(-std.error)
```

```
## # A tibble: 2 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)   167.      123.   5.09e-51
## 2 sexM           15.8       8.02 1.09e- 9
```

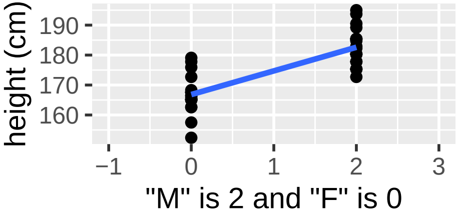



```
cor(heights$sex=="M", heights$height)
```

```
## [1] 0.7927089
```

```
tibble(sexM=2*as.numeric(heights$sex=="M"),
       y=heights$height) -> version_2
lm(y~sexM, data=version_2) %>%
  tidy() %>% select(-std.error)
```

```
## # A tibble: 2 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)   167.      123.   5.09e-51
## 2 sexM            7.90      8.02 1.09e- 9
```



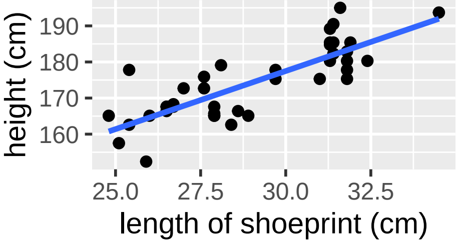

```
cor(2*(heights$sex=="M"), heights$height)
```

```
## [1] 0.7927089
```

How Scaling Variables Works Generally

```r
lm(height~shoePrint, data=heights) %>%
  tidy() %>% select(-std.error)
```

```
## # A tibble: 2 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)    80.9       7.43 6.50e- 9
## 2 shoePrint       3.22      8.61 1.86e-10
```

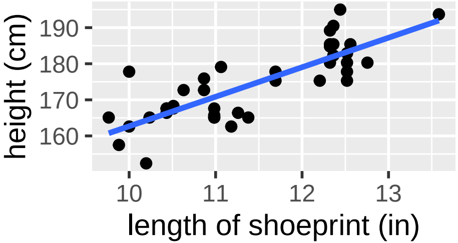



```r
cor(heights$shoePrint, heights$height)
```

```
## [1] 0.812948
```

```r
lm(height~I(shoePrint/2.54), data=heights) %>%
  tidy() %>% select(-std.error)
```

```
## # A tibble: 2 x 4
##   term              estimate statistic  p.value
##   <chr>                <dbl>     <dbl>    <dbl>
## 1 (Intercept)          80.9       7.43 6.50e- 9
## 2 I(shoePrint/2.54)     8.18      8.61 1.86e-10
```




```r
cor(heights$shoePrint/2.54, heights$height)
```

```
## [1] 0.812948
```

Hypothesis Testing for Linear Model Regression

Will $\hat{\beta}_0 = \beta_0$? Will $\hat{\beta}_1 = \beta_1$? What would a NULL hypothesis of $H_0 : \beta_1 = 0$ mean? How about $H_0 : \beta_0 = 0$? What are corresponding the ALTERNATIVE hypotheses here? What's the result and interpretation of $\alpha = 0.05$ significance testing the models below?

The `statistic` and `p.value` outputs are based on $H_0 : \beta = 0$

```r
lm(height~shoePrint, data=heights) %>%
  tidy() %>% select(-std.error)
```

```
## # A tibble: 2 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)    80.9       7.43 6.50e- 9
## 2 shoePrint       3.22      8.61 1.86e-10
```

```r
lm(height~sex, data=heights) %>%
  tidy() %>% select(-std.error)
```

```
## # A tibble: 2 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)    167.     123.   5.09e-51
## 2 sexM            15.8      8.02 1.09e- 9
```

- These tests depend on the $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ model being `TRUE`.
- This entails *normality*, *homoskedasticity*, *independence*, and *linearity* assumptions.

→ Confidence Intervals on Regression Coefficient values are then also possible

Next Class. . .

```r
lm(height~shoePrint, data=heights) %>% tidy() # R^2 0.6608845
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    80.9     10.9        7.43 6.50e- 9
## 2 shoePrint       3.22     0.374      8.61 1.86e-10
```

```r
lm(height~sex, data=heights) %>% tidy() %>% as.matrix() # R^2 0.6283874
```

```
##      term          estimate    std.error  statistic    p.value
## [1,] "(Intercept)" "166.82381" "1.357760" "122.866909" "5.085412e-51"
## [2,] "sexM"        " 15.79198" "1.970046" "  8.016048" "1.085391e-09"
```

```r
lm(height~shoePrint+sex, data=heights) %>% tidy() %>% as.matrix() # R^2 0.6909145
```

```
##      term          estimate     std.error    statistic  p.value
## [1,] "(Intercept)" "112.734096" "19.8103430" "5.690669" "1.647767e-06"
## [2,] "shoePrint"   "  2.007926" " 0.7339256" "2.735872" "9.498622e-03"
## [3,] "sexM"        "  7.001892" " 3.6929712" "1.896005" "6.578969e-02"
```

```r
# summary(lm(height~shoePrint+sex, data=heights))$r.squared
```