Regression Lectures

Fabio Carvalho Lima

13/11/2019

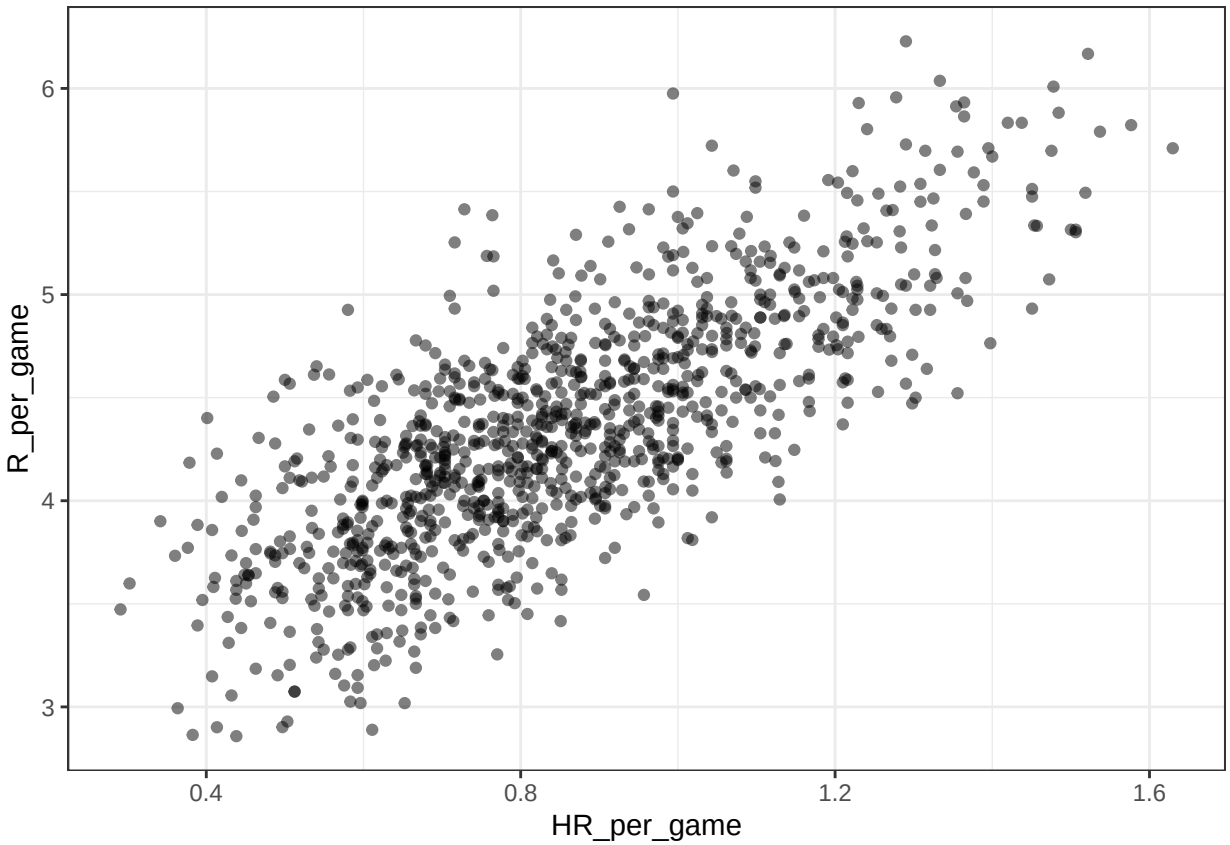
Case study on Moneyball

- Bill James was the originator of the **sabermetrics**, the approach of using data to predict what outcomes best predicted if a team would win.
- The goal of baseball game is to score more runs, than the other team.
- Each team has 9 batters who have an opportunity to hit a ball with a bat in a predetermined order.
- Each time a batter has an opportunity to bat, we call it a plate appearance (PA).
- The PA ends with a binary outcome: the batter either makes an out (failure) and returns to the bench or the batter doesn't (success) and can run around the bases, and potentially score a run (reach all 4 bases).
- There are five ways a batter can succeed (not make an out):

1. Bases on balls (BB): the pitcher fails to throw the ball through a predefined area considered to be hittable (the strike zone), so the batter is permitted to go to first base.
2. Single: the batter hits the ball and gets to first base.
3. Double (2B): the batter hits the ball and gets to second base.
4. Triple (3B): the batter hits the ball and gets to third base.
5. Home Run (HR): the batter hits the ball and goes all the way home and scores a run.

- Historically, the batting average has been considered the most important offensive statistic. To define this average, we define a hit (H) and an at bat (AB). Singles, doubles, triples and home runs are hits. The fifth way to be successful, a walk (BB), is not a hit. An AB is the number of times you either get a hit or make an out; BBs are excluded. **The batting average is simply H/AB and is considered the main measure of a success rate.**

The visualization of choice when exploring the relationship between two variables like home runs and runs is a **scatterplot**.



- Question1

What is the aplication of statistics and data science to baseball called?

Sabermetrics

- Question2

What is the outcome is not included in the batting average?

A base on balls

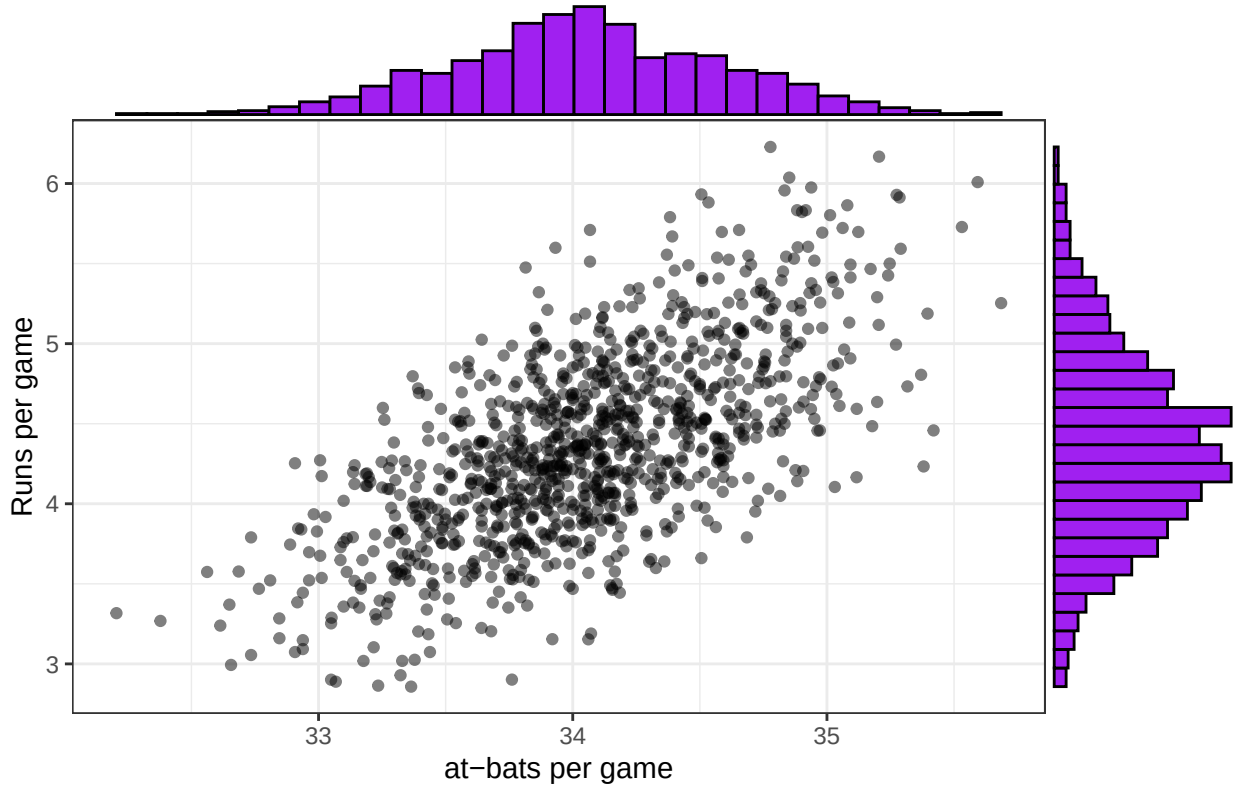
- Question3

Why do we consider team statistics as well as individual player statistcs?

Team statistics are important because the success of individual players depends also on the strength of their team.

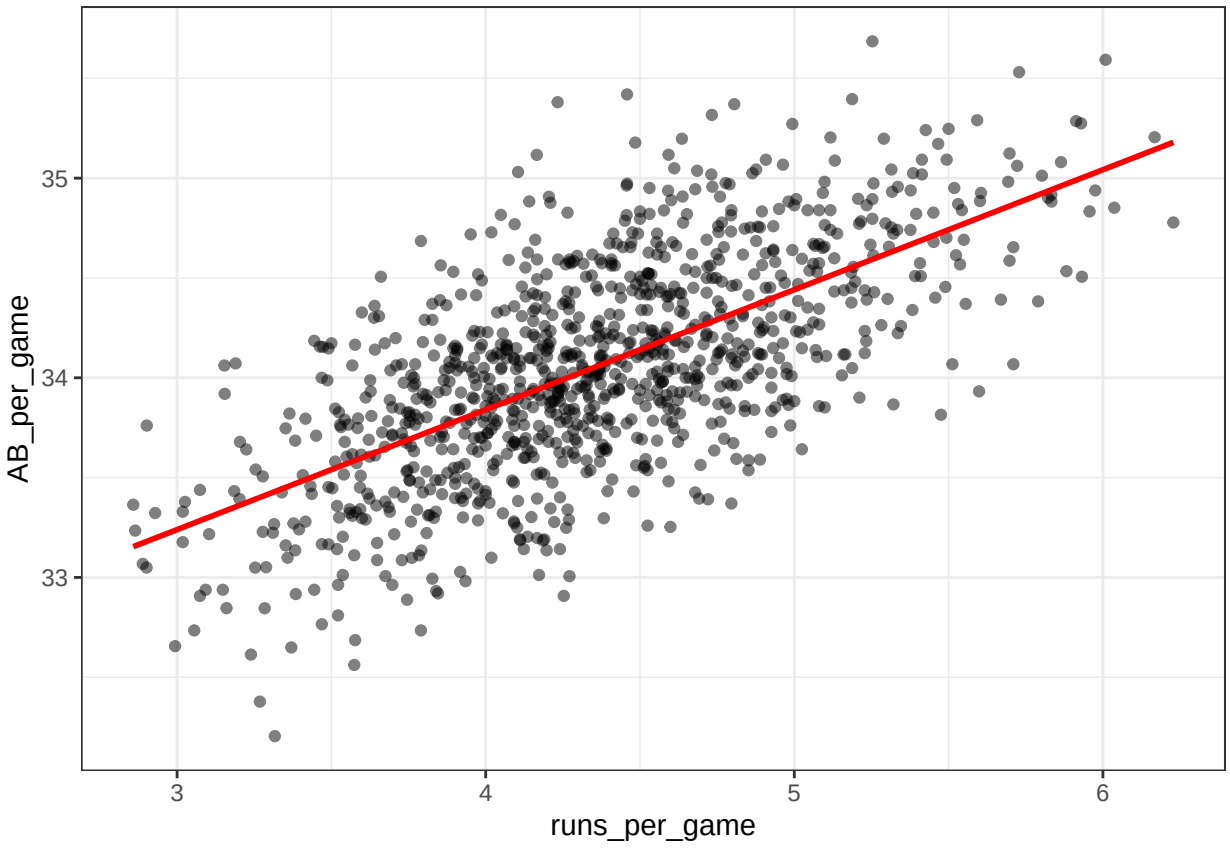
- Question4

You want to know whether teams with more at-bats per game have more runs per game.



- Question6

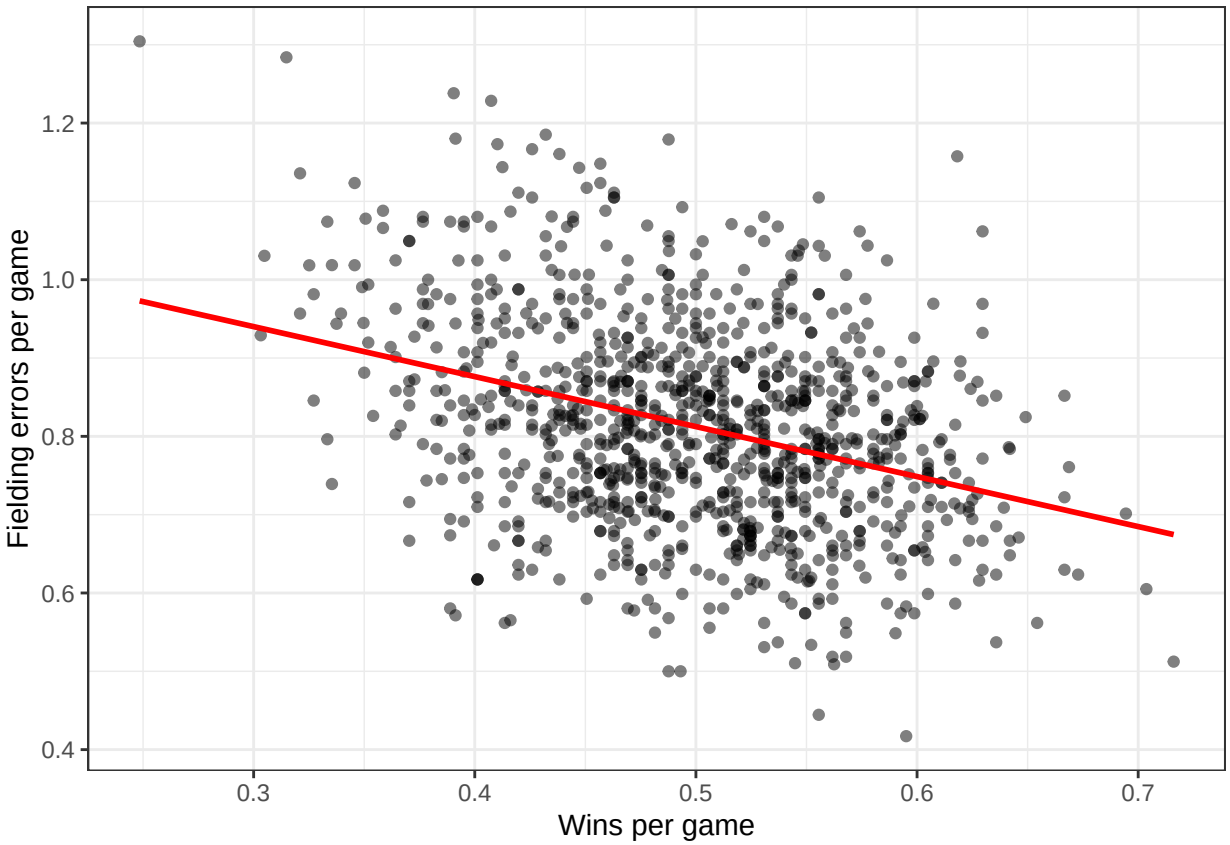
Load the Lahman library. Filter the Teams data frame to include years from 1961 to 2001. Make a scatterplot of runs per game versus at bats (AB) per game.



- Question7

Use the filtered Teams data frame from Question 6. Make a scatterplot of win rate (number of wins per game) versus number of fielding errors (E) per game.

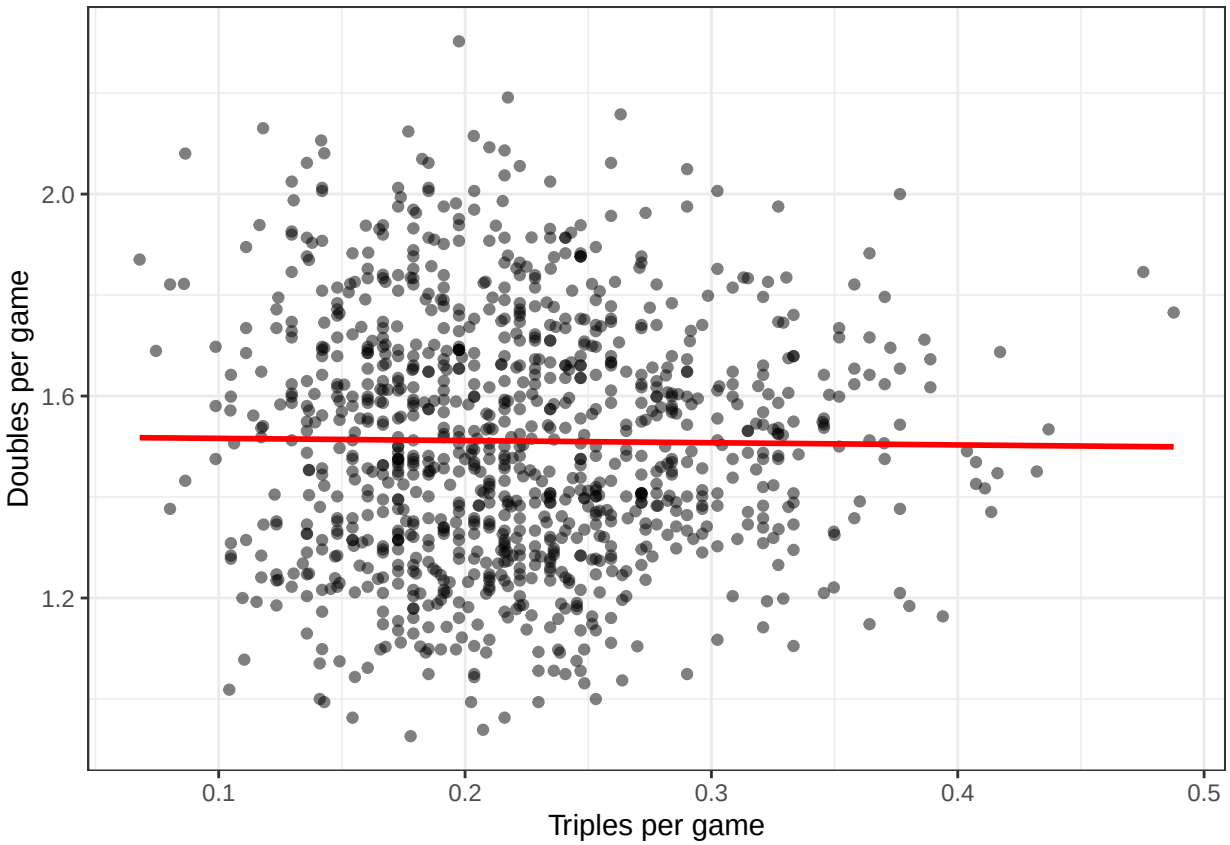
Which of the following is true?



When you examine the scatterplot above, you can see a clear trend towards decreased win rate with increasing number of errors per game.

- Question8

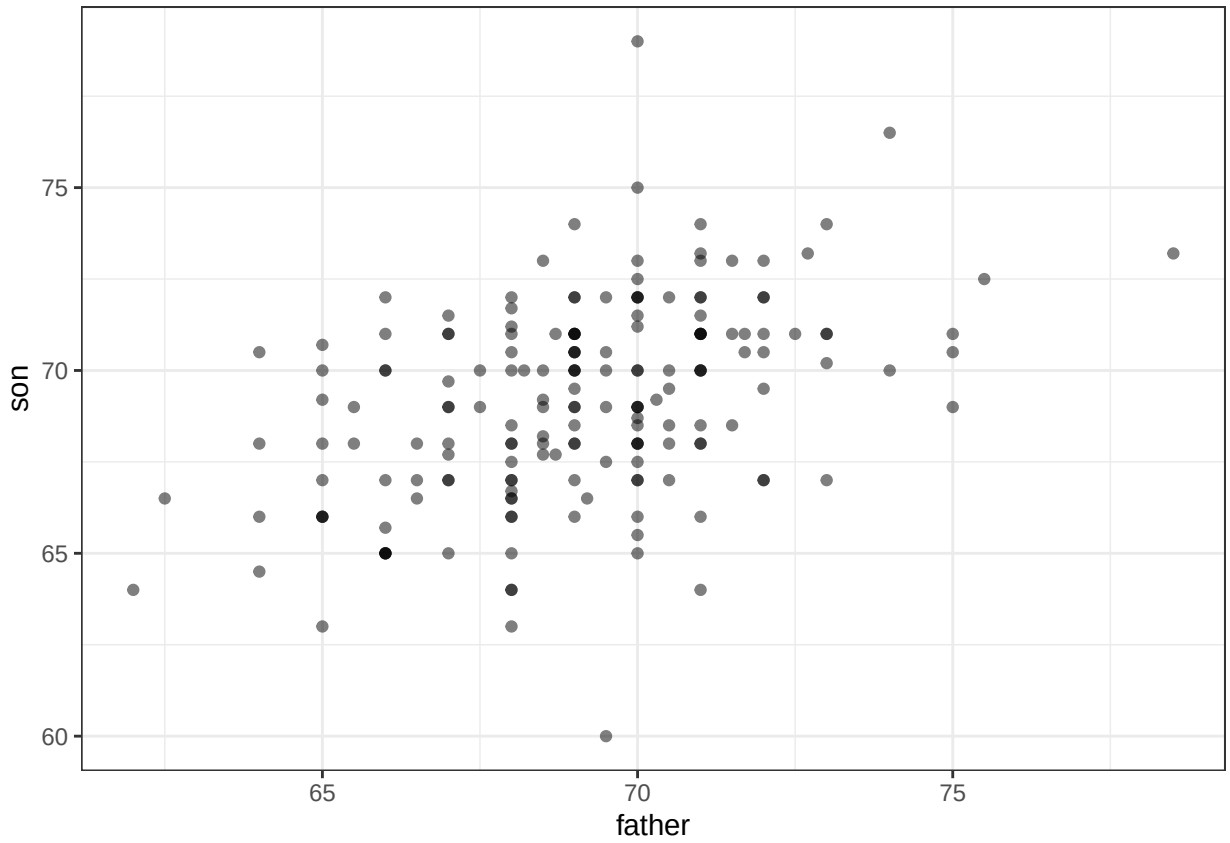
Use the filtered Teams data frame from Question 6. Make a scatterplot of triples (X3B) per game versus doubles (X2B) per game.



Correlation - Case Study: is height hereditary?

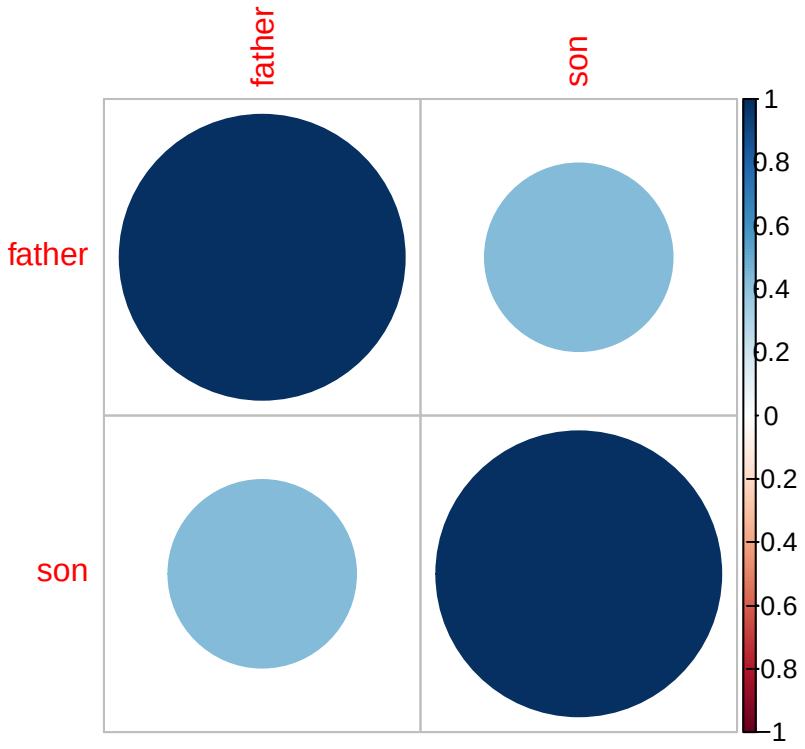
- Galton tried to predict sons' heights based on fathers' heights.
- The mean and standard errors are insufficient for describing an important characteristic of the data: the trend that the taller the father, the taller the son.
- The **correlation coefficient** is an informative summary of how two variables move together that can be used to predict one variable using the other.

```
## # A tibble: 1 x 4
##   `mean(father)` `sd(father)` `mean(son)` `sd(son)`
##            <dbl>        <dbl>       <dbl>     <dbl>
## 1           69.1         2.55        69.2      2.71
```



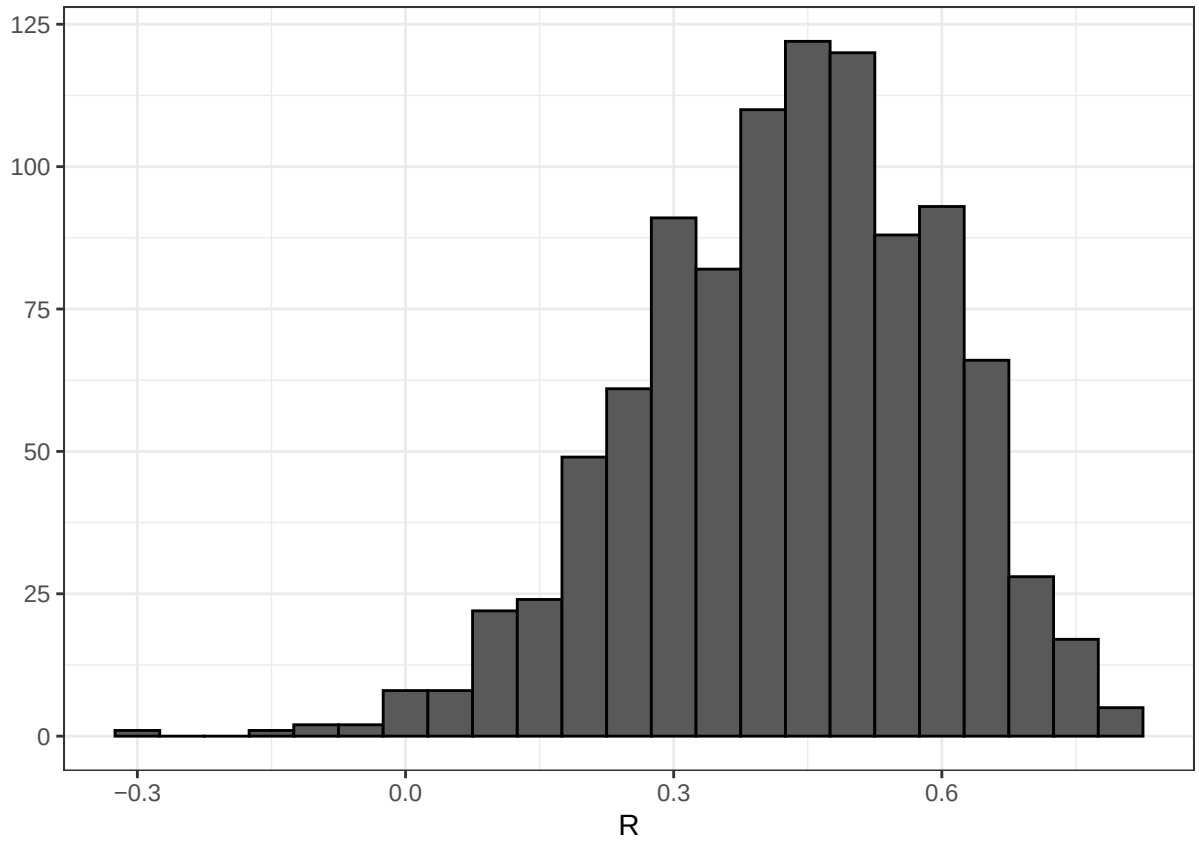

```
##        father    son
## father 1.0000 0.4334
## son    0.4334 1.0000
```

```
## [1] 0.4334
```



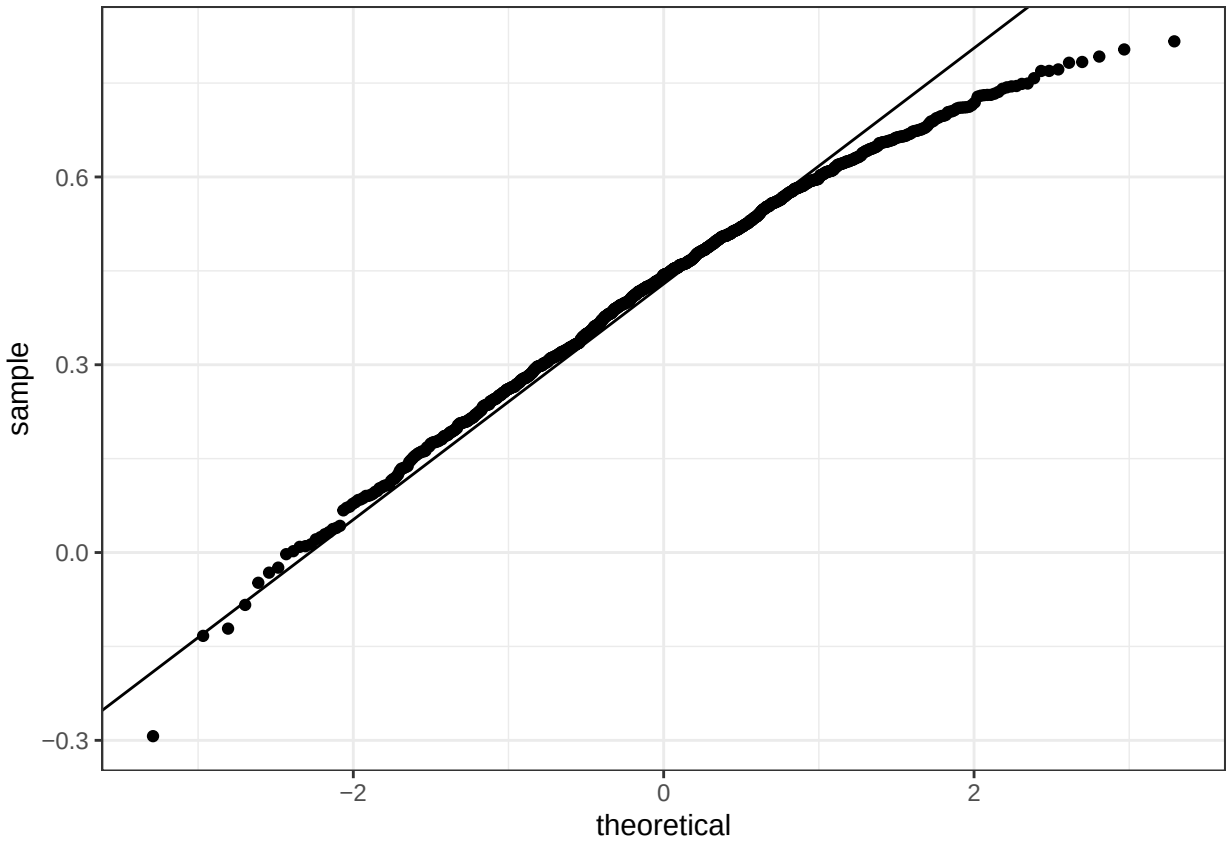
- The correlation that we compute and use as a summary is a random variable.
- When interpreting correlations, it is important to remember that correlations derived from samples are estimates containing uncertainty.
- Because the sample correlation is an average of independent draws, the central limit theorem applies.

```
## # A tibble: 1 x 1
##       r
##   <dbl>
## 1 0.463
```




```
## [1] 0.4292
```

```
## [1] 0.1664
```



- Questions:
- 1.While studying heredity, Francis Galton developed what important statistical concept?

Francis Galton developed the concept of correlation while study heredity.

- 2.The correlation coefficient is a summary of what?

The correlation coefficient is a summary of the trend between two variables.

The standard deviation describes the dispersion of a variable; the mean is a description of a variable's central tendency; the distribution of a variable (e.g., normal, log-normal) describes the possible values of your data and the probability of them occurring.

- 4. Instead of running a Monte Carlo simulation with a sample size of 25 from the 179 father-son pairs described in the videos, we now run our simulation with a sample size of 50. Would you expect the **expected value** of our sample correlation to increase, decrease, or stay approximately the same?

Explanation *Because the expected value of the sample correlation is the population correlation, it should stay approximately the same even if the sample size is increased.*

- 5. Instead of running a Monte Carlo simulation with a sample size of 25 from the 179 father-son pairs described in the videos, we now run our simulation with a sample size of 50. Would you expect the **standard error** of our sample correlation to increase, decrease, or stay approximately the same?

Explanation

As the sample size N increases, the standard deviation of the sample correlation should decrease.

- 6. If X and Y are completely independent, what do you expect the value of the correlation coefficient to be?

Explanation

Variables that are independent of each other have a correlation coefficient of 0.

- 7. Load the Lahman library. Filter the Teams data frame to include years from 1961 to 2001. What is the correlation coefficient between number of runs per game and number of at bats per game?

```
## [1] 0.6581
```

The solution for this correlation is cor(runs per game, at bats per game) = 0.6581

- 8. Use the same filtered dataset from previous question. What is the correlation coeffcient between win rate (number of win per game and number of errors per game)? solution:

```
## [1] -0.3397
```

answer : **correlation = -0.3397**

- 9. What is the correlation coefficient between doubles(X2B) per game and triples (X3B) per game?

```
## [1] -0.01157
```

answer: **correlation = -0.0116**

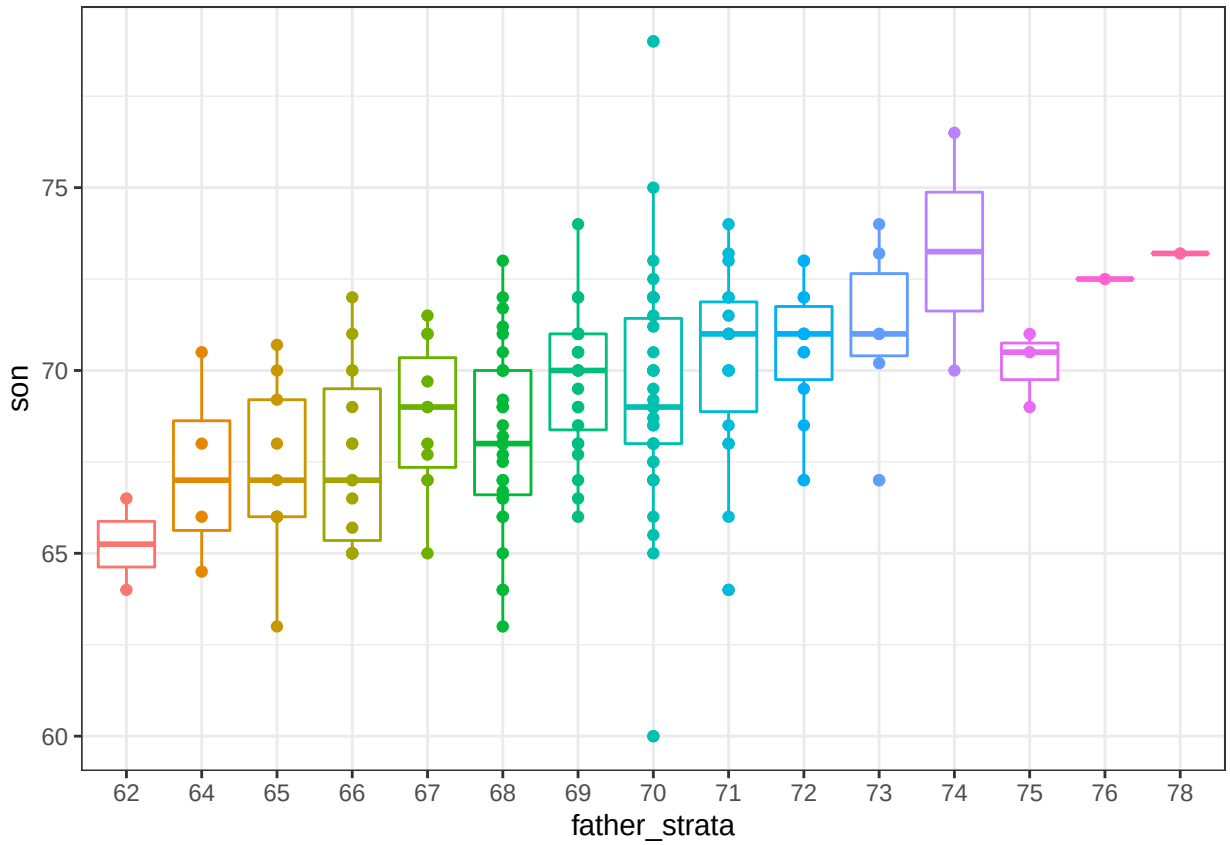
Anscombe's Quartet/Stratification

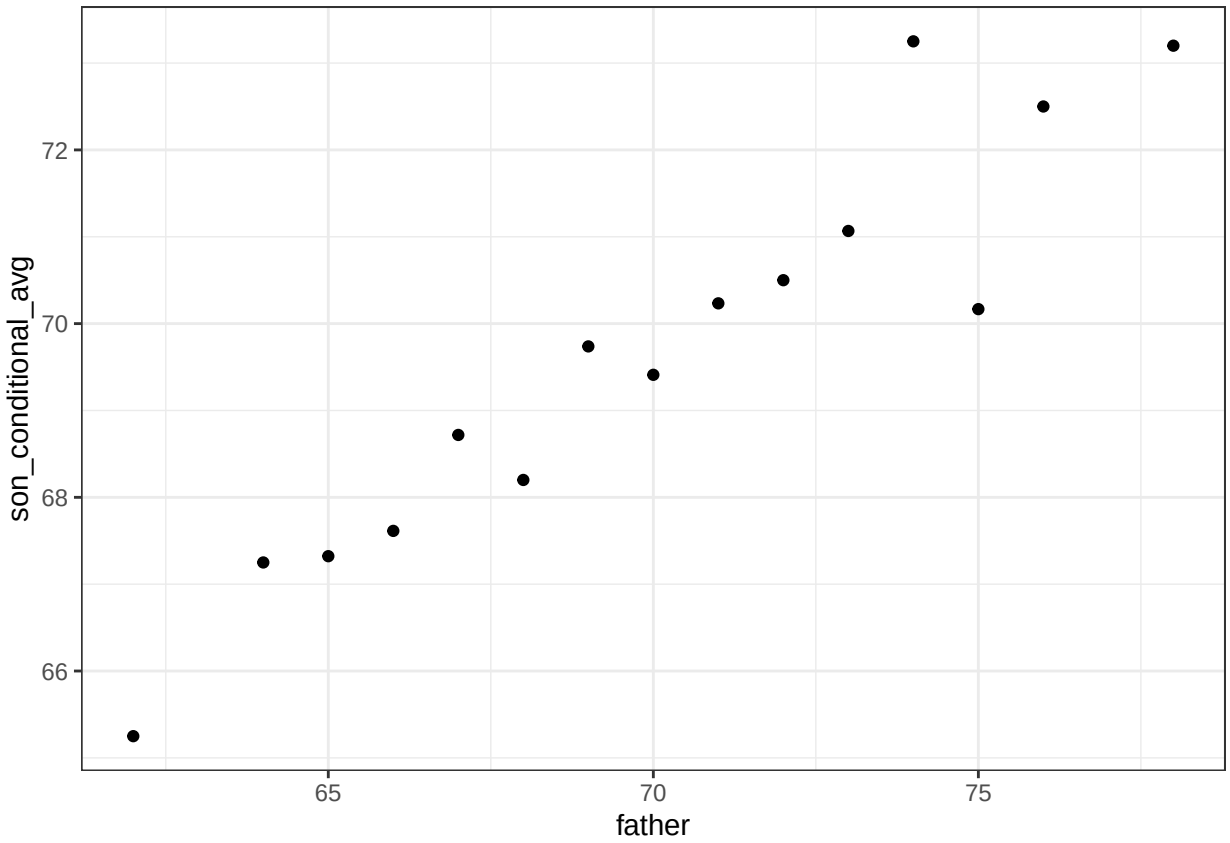
- Correlação não é sempre um bom resumo de relacionamento entre duas variáveis.
- A idéia geral da expectativa condicional é que estratifiquemos uma população em grupos e calculemos resumos em cada grupo.
- Uma maneira prática de melhorar as estimativas das expectativas condicionais é definir estratos com valores semelhantes de x.
- Se houver uma correlação perfeita, a linha de regressão prevê um aumento que é o mesmo número de desvio padrão para ambas as variáveis. Se houver correlação 0, não usamos x para a previsão e simplesmente predizemos a média μ_y . Para valores entre 0 e 1, a previsão está em algum lugar no meio. Se a correlação for negativa, prevemos uma redução em vez de um aumento.

```
## [1] 8
```

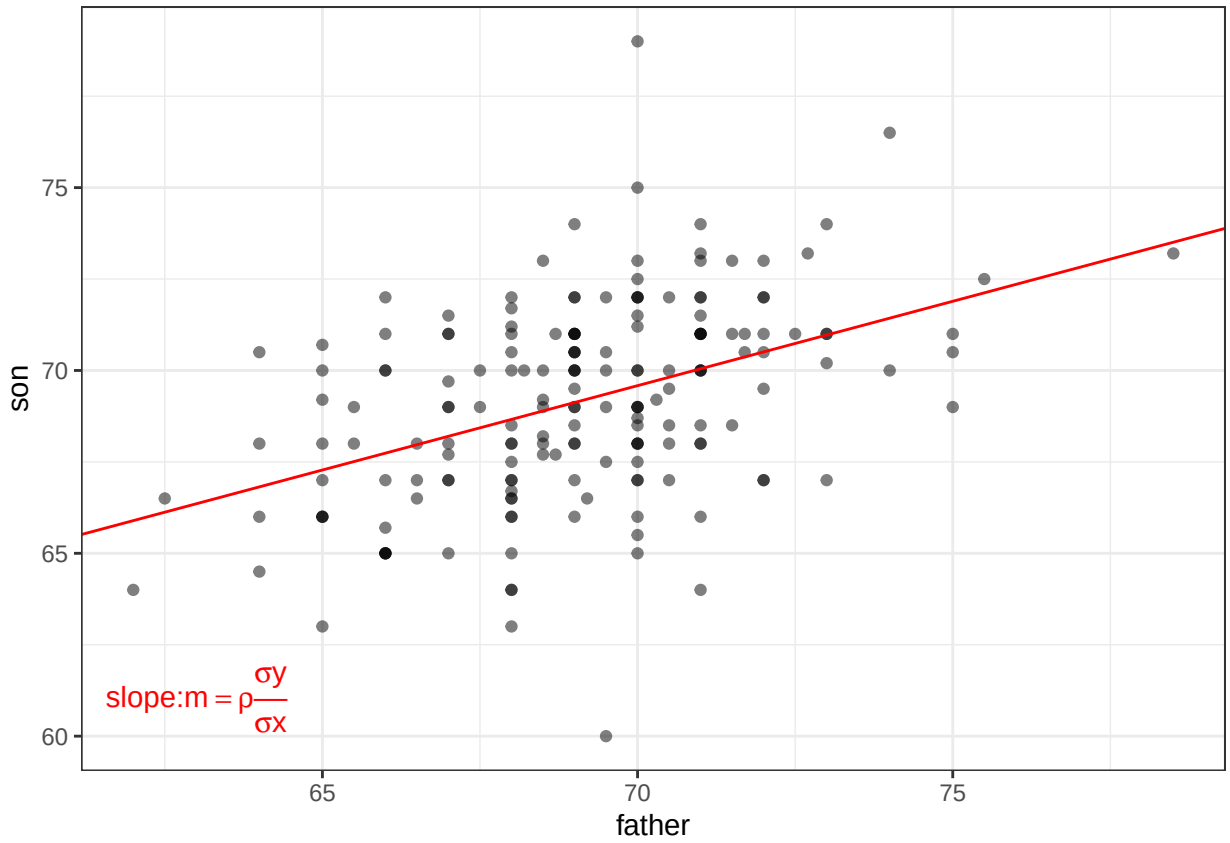
```
## [1] 1
```

```
## [1] 70.5
```

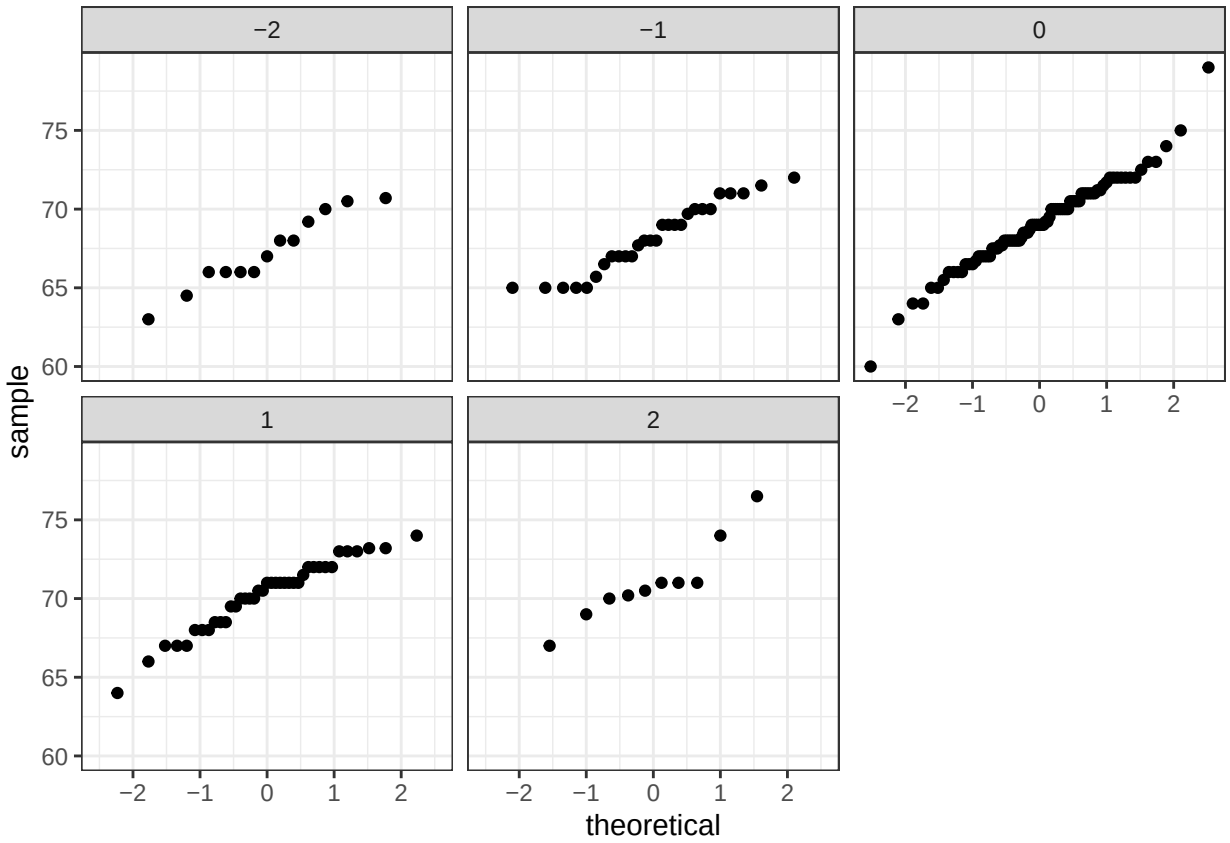

son
75
70
65
60
62
64
65
66
67
68
69
70
71
72
73
74
75
76
78
father_strata




```
## [1] 0.4334 0.4334
```



Sabemos das aulas que a inclinação do gráfico de ajuste é $m = \rho\frac{\sigma_y}{\sigma_x}$, que expresso em palavras é o coeficiente de correlação entre a variável resposta **y** e a variável de controle **x**, vezes o σ_x dividido pelo σ_y.



- Quando um par de variáveis aleatórias é aproximado pela distribuição normal bivariada, os gráficos de dispersão parecem ovais. Eles podem ser finos (alta correlação) ou em forma de círculo (sem correlação).
- Quando duas variáveis seguem uma distribuição normal bivariada, o cálculo da linha de regressão é equivalente ao cálculo das expectativas condicionais.
- Podemos obter uma estimativa muito mais estável da expectativa condicional localizando a linha de regressão e usando-a para fazer previsões.
- O condicionamento em uma variável aleatória X pode ajudar a reduzir a variação da variável de resposta Y.
- O desvio padrão da distribuição condicional é $SD(Y|X = x) = \sigma_y\sqrt{1} - \rho^2$, que é menor do que o desvio padrão sem o condicionamento σ_y.
- Como a variação é o desvio padrão ao quadrado, a variação da distribuição condicional é $\sigma_y^2(1 - \rho^2)$.
- Na afirmação "X explica tal e qual porcentagem da variabilidade", o valor percentual refere-se à variação. A variação diminui em ρ^2 por cento.
- A declaração "variância explicada" só faz sentido quando os dados são aproximados por uma distribuição normal bivariada.
- Existem duas linhas de regressão diferentes, dependendo se estamos assumindo a expectativa de Y dado X ou assumindo a expectativa de X dado Y.

Assessment: Stratification and Variance Explained, Part 1

- 1 - Qual é a inclinação na reta de regressão.

Expressed in words, the slope is the correlation coefficient of the son and father heights times the standard deviation of the sons' heights divided by the standard deviation of the fathers' heights.

- 2 - Por que a linha de regressão simplifica para uma linha com interceptação zero e inclinação ρ quando padronizamos nossas variáveis x e y ?

Quando padronizamos variáveis, elas têm uma média de 0 e um desvio padrão de 1, dando a equação $y_i $= ρ x_i. A equação $ y_i$ = $\rho + x_i$ é para uma linha com intersecção igual ao coeficiente de correlação, mas a linha de regressão é simplificada para 0 qdo padronizamos as variáveis.

- 3. What is a limitation of calculating conditional means?

a. Each stratum we condition on (e.g. a specific father's height) may not have many data points.

b. Because there are limited data points for each stratum, our average values have large standard errors.

c. Conditional means are less stable than a regression line.

- 4. A regression line is the best prediction of Y given we know the value of X when:

In order for the regression line to be the best predictor of Y given a known value of X, X and Y must follow a bivariate normal distribution. It is insufficient for X and Y to each be normally distributed on their own; they must also have a joint bivariate normal distribution

- 5. Which one of the following scatterplots depicts an x and y distribution that is NOT well-approximated by the bivariate normal distribution?

The v-shaped distribution of points from the first plot means that the x and y variables do not follow a bivariate normal distribution.

When a pair of random variables is approximated by a bivariate normal, the scatter plot looks like an oval (as in the 2nd, 3rd, and 4th plots) - it is okay if the oval is very round (as in the 3rd plot) or long and thin (as in the 4th plot).

- 6. We previously calculated that the correlation coefficient ρ between fathers' and sons' heights is 0.5.

Given this, what percent of the variation in sons' heights is explained by fathers' heights?

```
rho <- 0.5
var_rho <- rho^2*100
var_rho
```

```
## [1] 25
```

When two variables follow a bivariate normal distribution, the variation explained can be calculated as $\rho^2 * 100$.

- 7. Suppose the correlation between father and son's height is 0.5, the standard deviation of fathers' heights is 2 inches, and the standard deviation of sons' heights is 3 inches.

Given a one inch increase in a father's height, what is the predicted change in the son's height?

Correct! TThe slope of the regression line is calculated by multiplying the correlation coefficient by the ratio of the standard deviation of son heights and standard deviation of father heights: $\sigma_{son}/\sigma_{father}$

- Associação não é causa !

```
# find regression line for predicting runs from BBs
bb_slope <- Teams %>%
  filter(yearID %in% 1961:2001 ) %>%
  mutate(BB_per_game = BB/G, R_per_game = R/G) %>%
  lm(R_per_game ~ BB_per_game, data = .) %>%
  .$coef %>%
```

```
  .[2]
bb_slope
```

```
## BB_per_game
##      0.7353
```

```
# compute regression line for predicting runs from singles
singles_slope <- Teams %>%
  filter(yearID %in% 1961:2001 ) %>%
  mutate(Singles_per_game = (H - HR - X2B - X3B)/G, R_per_game = R/G) %>%
  lm(R_per_game ~ Singles_per_game, data = .) %>%
  .$coef  %>%
  .[2]
singles_slope
```

```
## Singles_per_game
##           0.4494
```

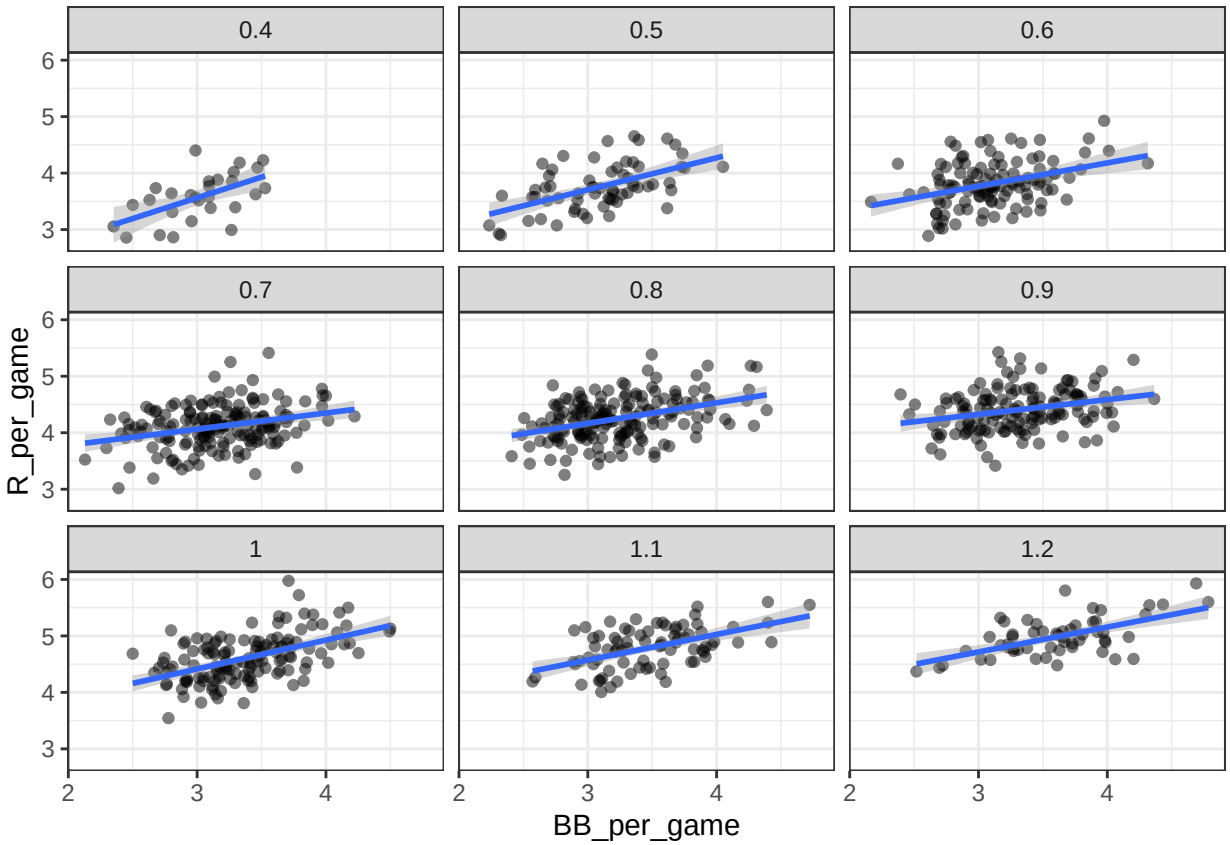
```
# calculate correlation between HR, BB and singles
Teams %>%
  filter(yearID %in% 1961:2001 ) %>%
  mutate(Singles = (H - HR - X2B - X3B)/G, BB = BB/G, HR = HR/G) %>%
  summarize(cor(BB, HR), cor(Singles, HR), cor(BB,Singles))
```

```
##   cor(BB, HR) cor(Singles, HR) cor(BB, Singles)
## 1      0.4039          -0.1737         -0.05604
```

Introduction to Linear Models

- A primeira maneira de checar "counfounding" is manter o HRs fixo até certo valor e então examinar a relação entre BB e runs.
- As inclinações BB depois da estratificação, em HR são reduzidas, mas não até zero, o que indica que BB são úteis para produzir runs, mas nem tanto como se pensava anteriormente.

```
# stratfy HR per game to nearest 10, filter out strata with few points
dat <- Teams %>% filter(yearID %in% 1961:2001) %>%
  mutate(HR_strata = round(HR/G, 1),
         BB_per_game = BB / G,
         R_per_game = R / G) %>%
  filter(HR_strata >= 0.4 & HR_strata <= 1.2)

# scatterplot for each HR stratum
dat %>%
  ggplot(aes(BB_per_game, R_per_game)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  facet_wrap( ~ HR_strata)
```

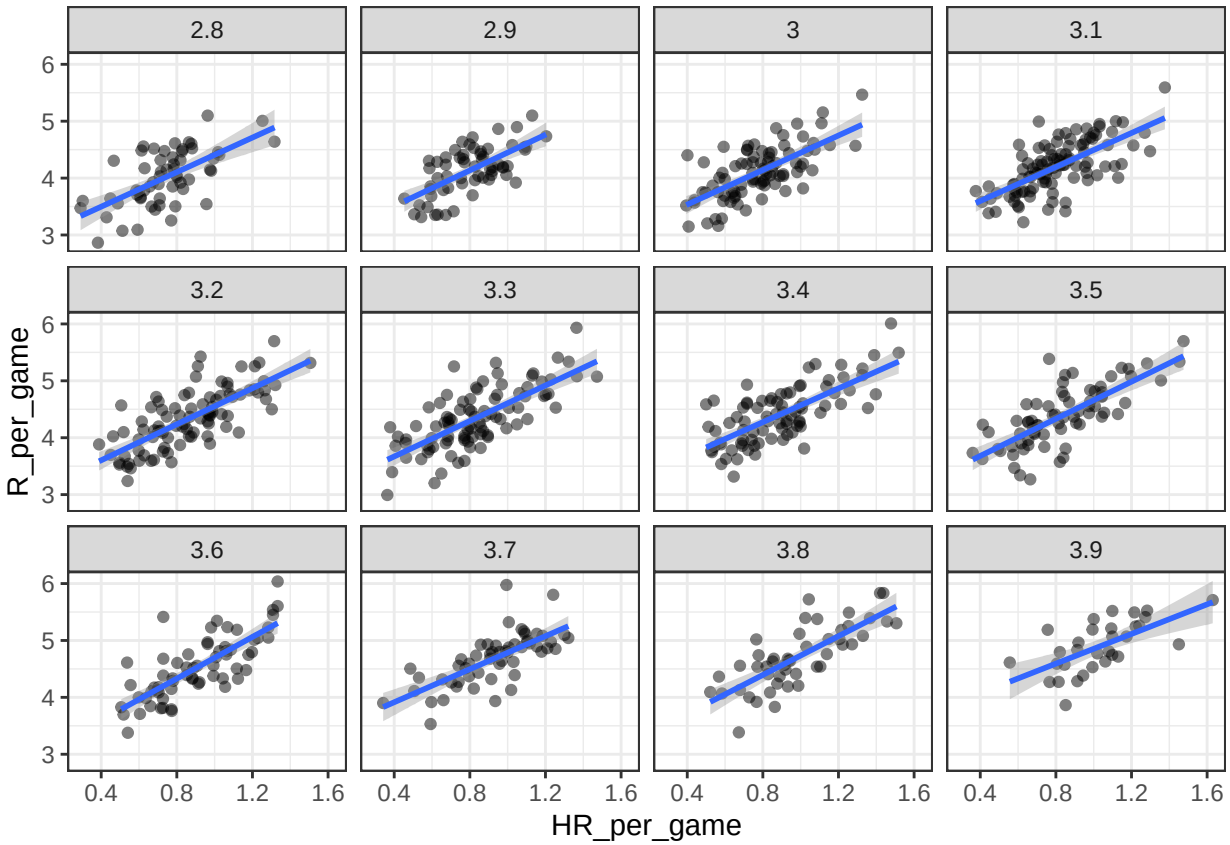



```
# calculate slope of regression line after stratifying by HR
dat %>%
  group_by(HR_strata) %>%
  summarize(slope = cor(BB_per_game, R_per_game)*sd(R_per_game)/sd(BB_per_game))
```

```
## # A tibble: 9 x 2
##   HR_strata slope
##       <dbl> <dbl>
## 1       0.4 0.734
## 2       0.5 0.566
## 3       0.6 0.412
## 4       0.7 0.285
## 5       0.8 0.365
## 6       0.9 0.261
## 7       1   0.511
## 8       1.1 0.454
## 9       1.2 0.440
```

```
# stratify by BB
dat <- Teams %>% filter(yearID %in% 1961:2001) %>%
  mutate(BB_strata = round(BB/G, 1),
         HR_per_game = HR / G,
         R_per_game = R / G) %>%
  filter(BB_strata >= 2.8 & BB_strata <=3.9)

# scatterplot for each BB stratum
dat %>% ggplot(aes(HR_per_game, R_per_game)) +
```

```
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  facet_wrap( ~ BB_strata)
```




```
# slope of regression line after stratifying by BB
dat %>%
  group_by(BB_strata) %>%
  summarize(slope = cor(HR_per_game, R_per_game)*sd(R_per_game)/sd(HR_per_game))
```

```
## # A tibble: 12 x 2
##    BB_strata slope
##        <dbl> <dbl>
##  1       2.8  1.52
##  2       2.9  1.57
##  3       3    1.52
##  4       3.1  1.49
##  5       3.2  1.58
##  6       3.3  1.56
##  7       3.4  1.48
##  8       3.5  1.63
##  9       3.6  1.83
## 10       3.7  1.45
## 11       3.8  1.70
## 12       3.9  1.30
```

Assessment:

- 1. Why is the number of home runs considered a confounder of the relationship between bases on balls and runs per game?

Explanation

Number of home runs is a confounder of the relationship between bases on balls and runs per game because players who get more bases on balls also tend to have more home runs and home runs also increase the points/runs scored per game.

- 2. As described in the videos, when we stratified our regression lines for runs per game vs. bases on balls by the number of home runs, what happened?
- 3. The coefficient for "father" gives the predicted increase in son's height for each increase of 1 unit in the father's height. In this case, it means that for every inch we increase the father's height, the son's predicted height increases by 0.5 inches.
- 4. Because the fathers' heights (the independent variable) have been centered on their mean, the intercept represents the height of the son of a father of average height. In this case, that means that the height of a son of a father of average height is 70.45 inches.

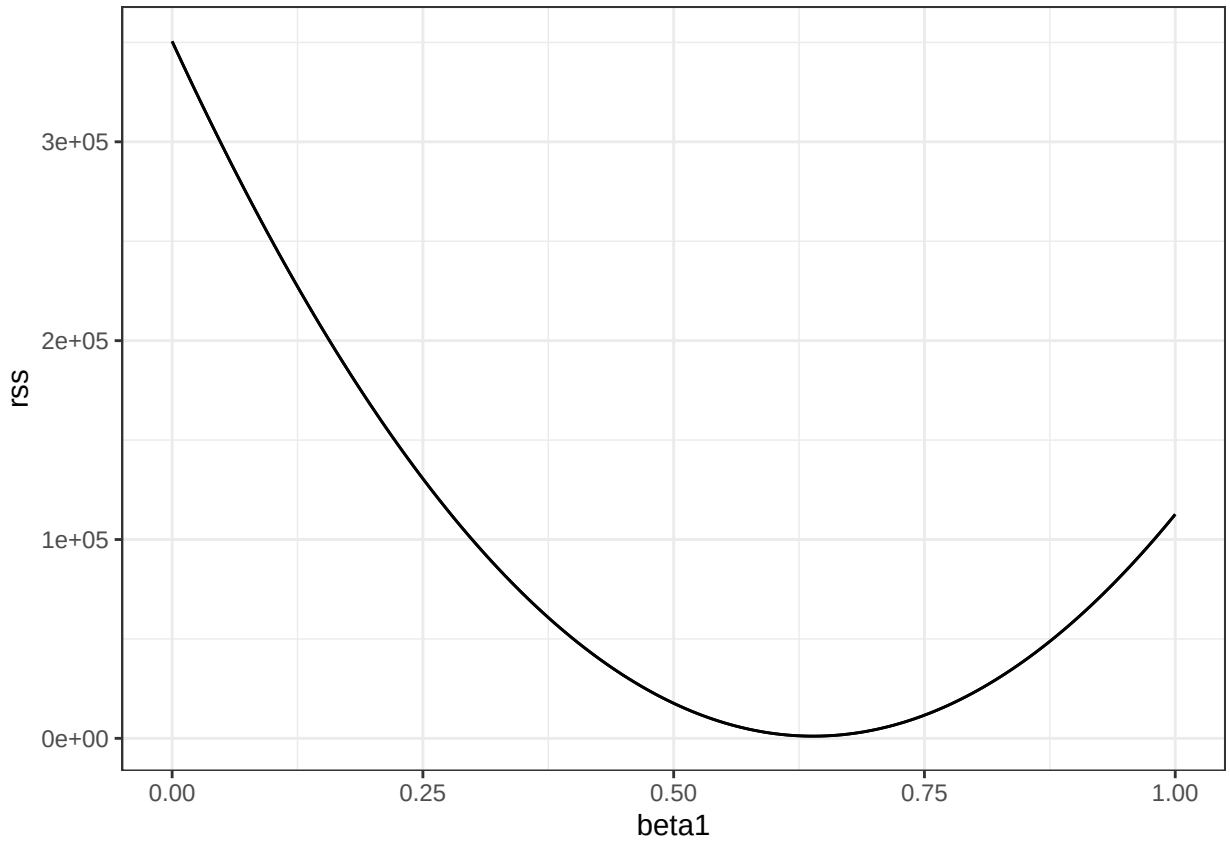
If we had not centered fathers' heights to its mean, then the intercept would represent the height of a son when a father's height is zero.

- 5.If x1 is fixed, then β_1x1 is fixed and acts as the intercept for this regression model. This is the basis of stratificaton.

Least Squares estimates

- Na regressão o objetivo é achar os valores dos coeficientes que minimizam a distância entre o modelo ajustado e os dados.
- Soma quadratica dos residuos **(RSS)** mede a distância entre os valores reais e o valor predito dado pela linha de regressão. Os valores que minimizam a RSS, são chamados de **LSE**.
- Nós podemos usar derivadas parciais para obter os valores para β_0 e β_1 in Galton's data.

```
# compute RSS for any pair of beta0 and beta1 in Galton's data
library(HistData)
data("GaltonFamilies")
set.seed(1983)
galton_heights <- GaltonFamilies %>%
  filter(gender == "male") %>%
  group_by(family) %>%
  sample_n(1) %>%
  ungroup() %>%
  select(father, childHeight) %>%
  rename(son = childHeight)

rss <- function(beta0, beta1, data){
    resid <- galton_heights$son - (beta0 + beta1*galton_heights$father)
    return(sum(resid^2))
}
# plot RSS as a function of beta1 when beta0=25
beta1 = seq(0, 1, len=nrow(galton_heights))
results <- data.frame(beta1 = beta1,
                      rss = sapply(beta1, rss, beta0 = 25))
results %>% ggplot(aes(beta1, rss)) + geom_line() +
  geom_line(aes(beta1, rss))
```




```
# fit regression line to predict son's height from father's height
fit <- lm(son ~ father, data = galton_heights)
fit
```

```
##
## Call:
## lm(formula = son ~ father, data = galton_heights)
##
## Coefficients:
## (Intercept)       father
##      37.288        0.461
```

```
# summary statistics
summary(fit)
```

```
##
## Call:
## lm(formula = son ~ father, data = galton_heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.354 -1.566 -0.008  1.726  9.415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2876     4.9862    7.48  3.4e-12 ***
## father        0.4614     0.0721    6.40  1.4e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.45 on 177 degrees of freedom
## Multiple R-squared:  0.188,  Adjusted R-squared:  0.183
## F-statistic: 40.9 on 1 and 177 DF,  p-value: 1.36e-09
```

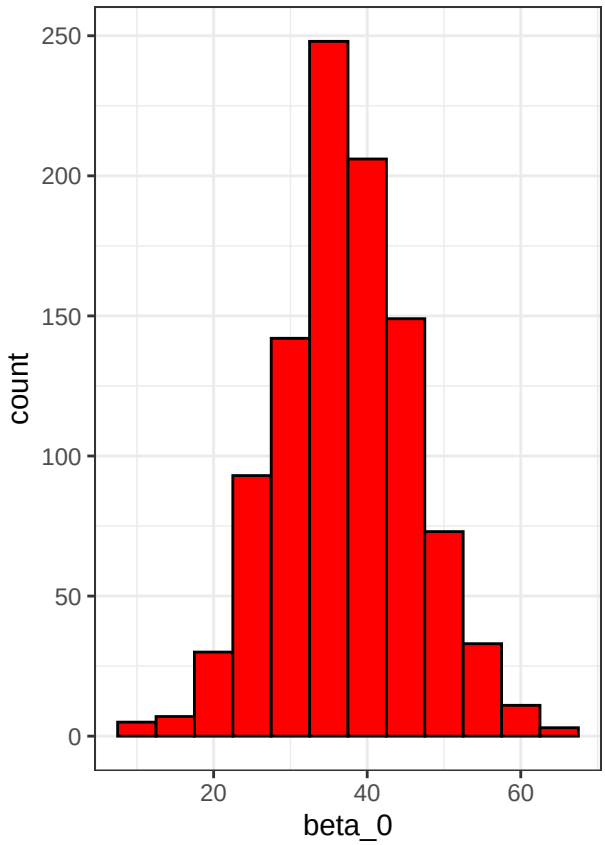
```
# Monte Carlo simulation
B <- 1000
N <- 50
lse <- replicate(B, {
  sample_n(galton_heights, N, replace = TRUE) %>%
    lm(son ~ father, data = .) %>%
    .$coef
})
lse <- data.frame(beta_0 = lse[1,], beta_1 = lse[2,])
```

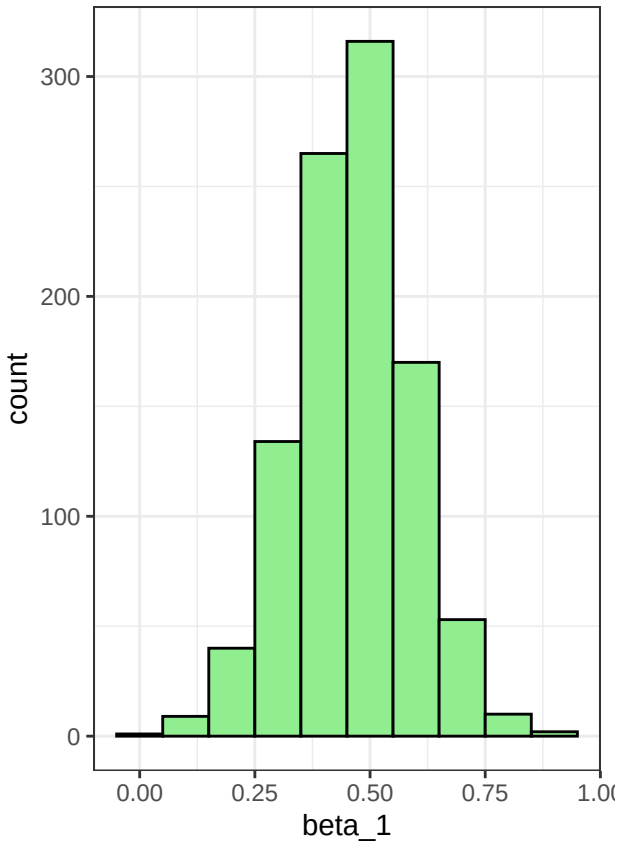
```
# Plot the distribution of beta_0 and beta_1
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
p1 <- lse %>% ggplot(aes(beta_0)) + geom_histogram(binwidth = 5, color = "black", fill = "red" )
p2 <- lse %>% ggplot(aes(beta_1)) + geom_histogram(binwidth = 0.1, color = "black", fill = "lightgreen")
grid.arrange(p1, p2, ncol = 2)
```






```
# summary statistics
sample_n(galton_heights, N, replace = TRUE) %>%
  lm(son ~ father, data = .) %>%
  summary %>%
  .$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  19.2792    11.6565   1.654 1.047e-01
## father        0.7199     0.1694   4.250 9.792e-05
```

```
lse %>% summarize(se_0 = sd(beta_0), se_1 = sd(beta_1))
```

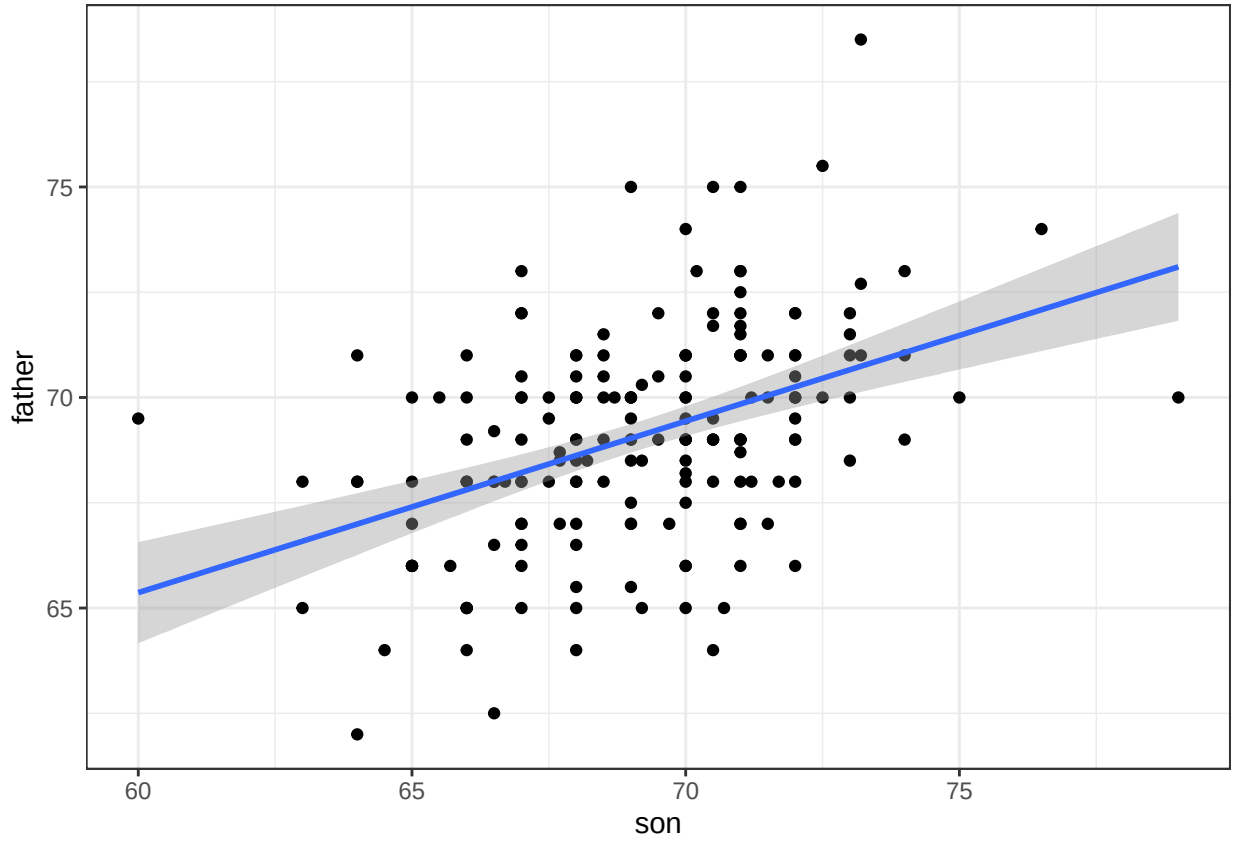
```
##    se_0   se_1
## 1 8.836 0.1279
```

```
# LSE pode ser fortemente correlacionado
#lse %>% summarize(cor(beta_0, beta_1))

# contudo a correlação depende de como a variável controle é definida ou
# transformada
# padronizando a altura dos pais

B <- 1000
N <- 50
lse <- replicate(B, {
      sample_n(galton_heights, N, replace = TRUE) %>%
      mutate(father = father - mean(father)) %>%
      lm(son ~ father, data = .) %>% .$coef
```

```
})
cor(lse[1,], lse[2,])
```

```
## [1] 0.004968
```

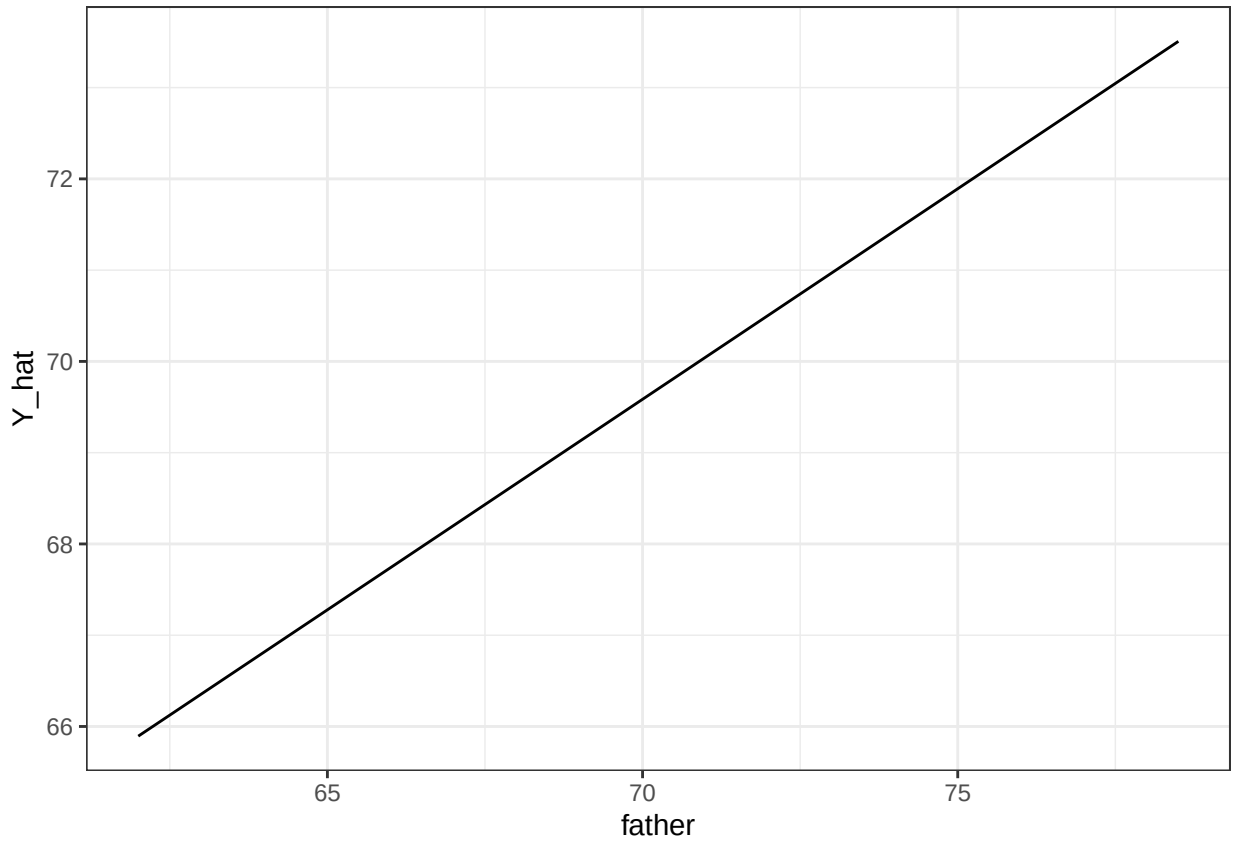
```
# plot predictions and confidence intervals
galton_heights %>% ggplot(aes(son, father)) +
  geom_point() +
  geom_smooth(method = "lm")
```



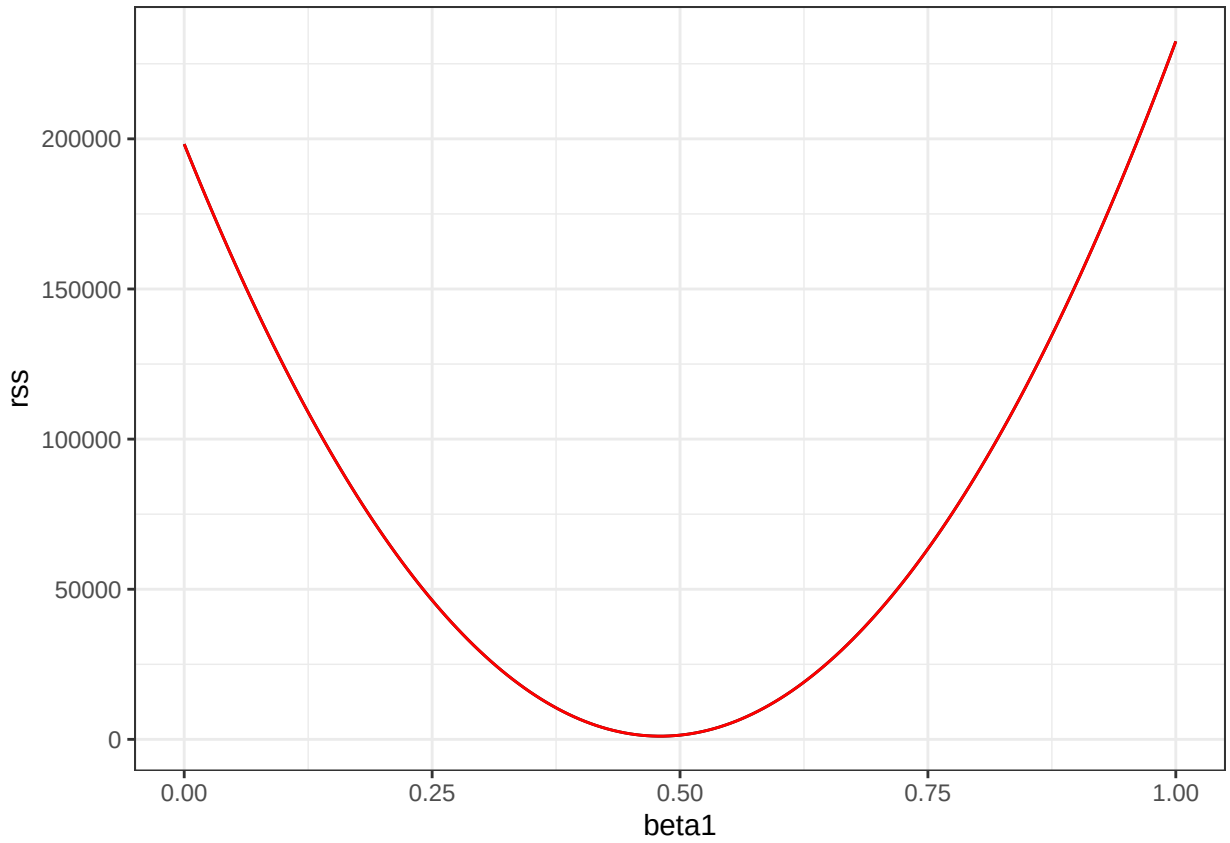

```
# predict Y directly
fit <- galton_heights %>% lm(son ~ father, data = .)
Y_hat <- predict(fit, se.fit = TRUE)
names(Y_hat)
```

```
## [1] "fit"            "se.fit"         "df"             "residual.scale"
```

```
# plot best fit line
galton_heights %>%
  mutate(Y_hat = predict(lm(son ~ father, data =.))) %>%
  ggplot(aes(father, Y_hat)) +
  geom_line()
```




```
beta1 = seq(0, 1, len=nrow(galton_heights))
results <- data.frame(beta1 = beta1,
                      rss = sapply(beta1, rss, beta0 = 36))
results %>% ggplot(aes(beta1, rss)) + geom_line() +
  geom_line(aes(beta1, rss), col=2)
```



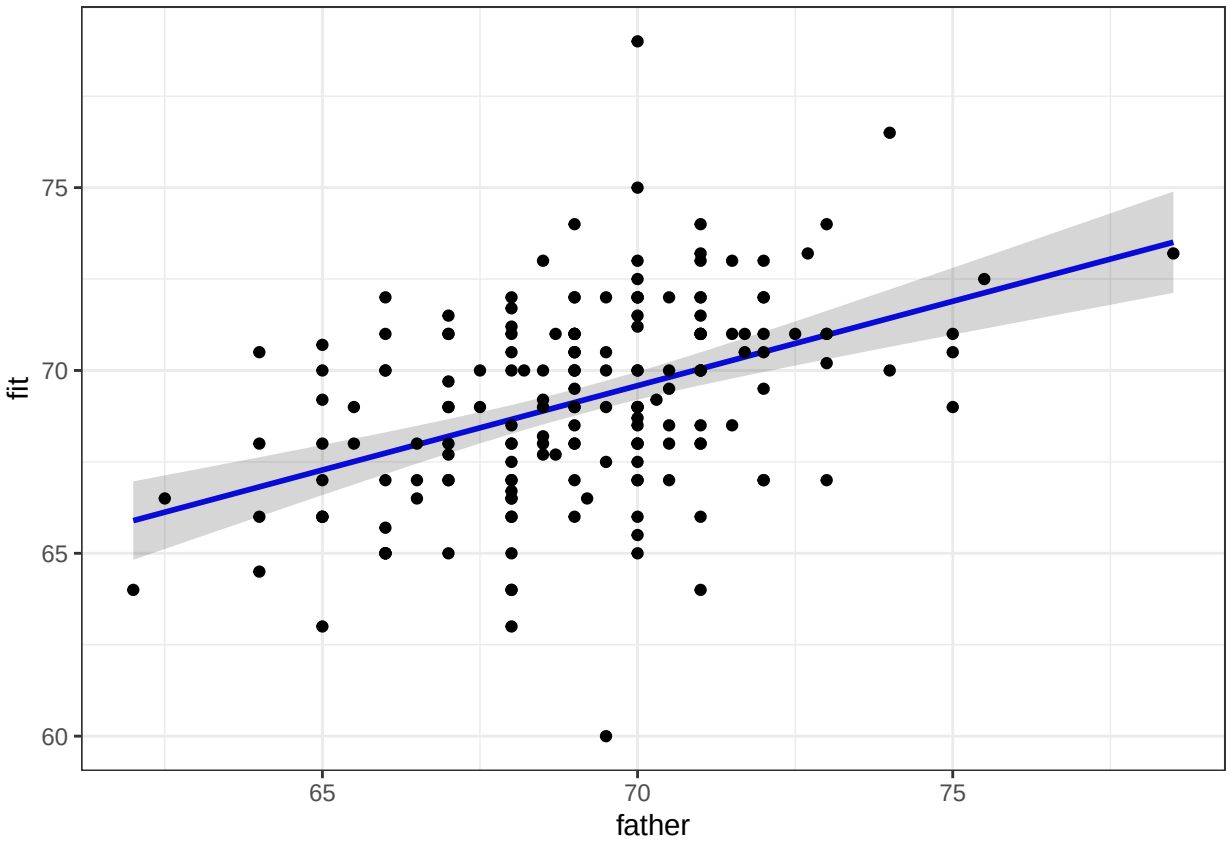

```
fit <- Teams %>% filter(yearID %in% 1961:2001) %>%
    mutate(BB_per_game = BB / G, R_per_game = R / G, HR_per_game = HR / G) %>%
    lm(R_per_game ~ BB_per_game + HR_per_game, data = .) %>%
    .$coef
fit
```

```
## (Intercept) BB_per_game HR_per_game
##      1.7443      0.3874      1.5612
```

```
model <- lm(son ~ father, data = galton_heights)
predictions <- predict(model, interval = c("confidence"), level = 0.95)
data <- as.tibble(predictions) %>% bind_cols(father = galton_heights$father)
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
ggplot(data, aes(x = father, y = fit)) +
    geom_line(color = "blue", size = 1) +
    geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.2) +
    geom_point(data = galton_heights, aes(x = father, y = son))
```




```
set.seed(1989, sample.kind="Rounding") #if you are using R 3.6 or later
```

```
## Warning in set.seed(1989, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
library(HistData)
data("GaltonFamilies")
options(digits = 3)    # report 3 significant digits
female_heights <- GaltonFamilies %>%
    filter(gender == "female") %>%
    group_by(family) %>%
    sample_n(1) %>%
    ungroup() %>%
    select(mother, childHeight) %>%
    rename(daughter = childHeight)
```

```
glimpse(female_heights)
```

```
## Observations: 176
## Variables: 2
## $ mother   <dbl> 67.0, 66.5, 64.0, 64.0, 58.5, 68.0, 68.0, 66.5, 66.0, 65.5...
## $ daughter <dbl> 69.0, 65.5, 68.0, 64.5, 66.5, 69.5, 70.5, 70.5, 66.0, 65.5...
```

```
model <- lm(mother ~ daughter, data = female_heights)
coef(model)
```

```
## (Intercept)    daughter
##       44.18        0.31
```

```r
data2 <- as.tibble(predict(model)) %>%
         bind_cols(mother = female_heights$mother)
head(data2)
```

```
## # A tibble: 6 x 2
##   value mother
##   <dbl>  <dbl>
## 1  65.6   67
## 2  64.5   66.5
## 3  65.3   64
## 4  64.2   64
## 5  64.8   58.5
## 6  65.7   68
```