**Introduction**

For aspiring data scientists, like yourselves, regression analysis the first real "learning" algorithm that they come across with. It is one of the simplest algorithms to master but still requires some statistical understanding of the underlying. This lesson will introduce you to regression process based on the statistical ideas we have discovered so far.

**Objectives**

You will be able to:

* Describe statistical modeling with simple regression
* Explain simple linear regression as solving for equation y=mx+c
* Calculate the slope and y-intercept using the slope value
* Draw a regression line based on calculated slope and intercept
* Predict the target of a previously unseen input feature
* الأهداف
* وصف النمذجة الإحصائية مع الانحدار البسيط
* o شرح الانحدار الخطي البسيط كحل للمعادلة y = mx + c
  + احتساب الميل وتقاطع y باستخدام قيمة الميل
* رسم خط الانحدار على أساس الميل وتقاطع y
* التنبؤ بمخرجات نموذج انحدار خطي لبيانات جديدة

**Linear regression**

So far, we have covered topics like basic hypothesis testing, variable relationships, statistical learning. We shall now build on these ideas to explain the regression process. Regression is a parametric technique used to **predict** the value of a target variable Y based on one or more input feature variables X. Regression is proven to give credible results if the data follows parametric assumptions which will be covered in upcoming lessons. Regression Analysis helps us with analytics in following ways:

* Finding an association, relationship between variables.
* Identifying which variables contribute more towards the outcomes.

and most importantly ..

* **Prediction** of future observations.

حتى الآن ، قمنا بتغطية مواضيع مثل اختبار الفرضيات الأساسية والعلاقات المتغيرة والتعلم الإحصائي. سنبني الآن على هذه الأفكار لشرح عملية الانحدار. الانحدار هو أسلوب حدودي يستخدم للتنبؤ بقيمة المتغير المستهدف Y استنادًا إلى متغير واحد أو أكثر من متغيرات ميزة الإدخال X. ثبت أن الانحدار يعطي نتائج موثوقة إذا كانت البيانات تتبع افتراضات حدودية سيتم تغطيتها في الدروس القادمة. تحليل الانحدار يساعدنا في التحليلات بالطرق التالية:

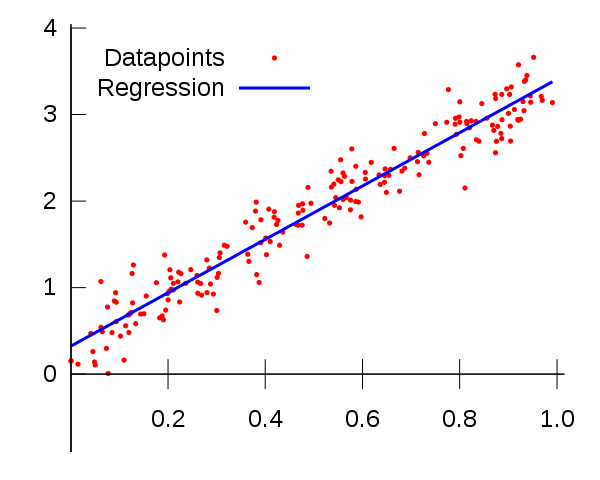
o إيجاد ارتباط ، العلاقة بين المتغيرات.

o تحديد المتغيرات التي تسهم أكثر في تحقيق النتائج.

و الاهم من ذلك ..

o التنبؤ بالملاحظات المستقبلية.

**Why is it called "linear" regression?**

As we saw in pevious lesson, linear implies that the model functions along a straight or nearly straight line. Linear suggests that the relationship between dependent and independent variable can be expressed in a straight line. A **Simple Linear Regression** uses a single feature (independent variable) to predict a target (dependent variable) by fitting a best linear relationship, whereas **Multiple Linear Regression** uses more than one features to predict a target variable by fitting a best linear relationship. In this section, we shall mainly focus on simple regression to build a sound understanding. 

So let's move on and see how to calculate such a line.

**Calculating Slope and Intercepts**

We all know from elementary geometry that equation of a stright line can be written as:

**y = mx + c**

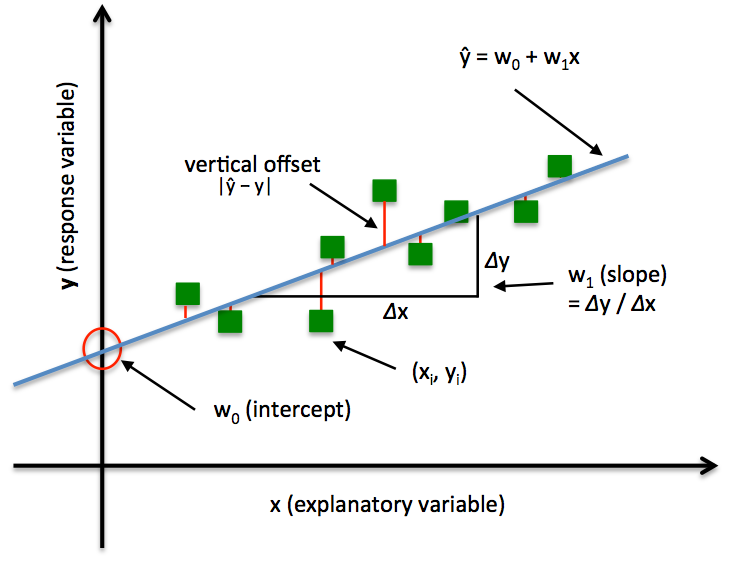
Following what we have covered so far, we can say from the equation that:

* y is the dependent variable i.e. the variable that needs to be estimated and predicted.
* x is the independent variable i.e. the variable that is controllable. It is the input.
* m is the slope. It determines what will be the angle of the line. It is the parameter denoted as β.
* c is the intercept. A constant that determines the value of y when x is 0.

Linear regression is nothing but a manifestation of this simple equation. The formula for the **best-fit** line (or regression line) is still "a line".

A model really cant get any simpler than this. This equation here is the same one used to find a line in algebra, but in statistics the points don’t lie perfectly on a line as shown above

The line is a model around which the data lie if a strong linear pattern exists.

Following image that we saw in previous lesson explains this further. 

The slope (m) of a line is the **change in Y over the change in X** (Δy/Δx shown above). For example, a slope of 4/3 means as the x-value increases by 4 units, the y-value moves up by 3 units on average.

The y-intercept (b) is the value on the y-axis **where the line crosses the axis**. For example, in the equation y= 2x +2, the line crosses the y-axis at the value b = 2. The coordinates of this point would be (0, 2).

When a line crosses the y-axis, the x-value is always 0.

You may be thinking that you have to try lots and lots of different lines to see which one fits best. Fortunately, this task is not as complicated as in may seem . Given some data points, the best-fit line always has a distinct slope and y-intercept that can be calculated using simple linear algebraic approaches.

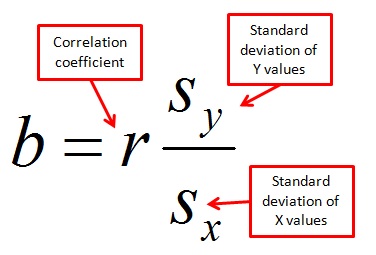
**Best-Fit Line Ingredients**

Before we calculate the best-fit line, we have to make sure that we have calculated following measures for variables x and y:

* The mean of the x (x\_bar)
* The mean of the y (y\_bar)
* The standard deviation of the x values (denoted Sx)
* The standard deviation of the y values (denoted Sy)
* The correlation between X and Y (denoted r - following Pearson Correlation)

**Calculating Slope**

The formula for the slope (shown as b below), of the best-fit line is



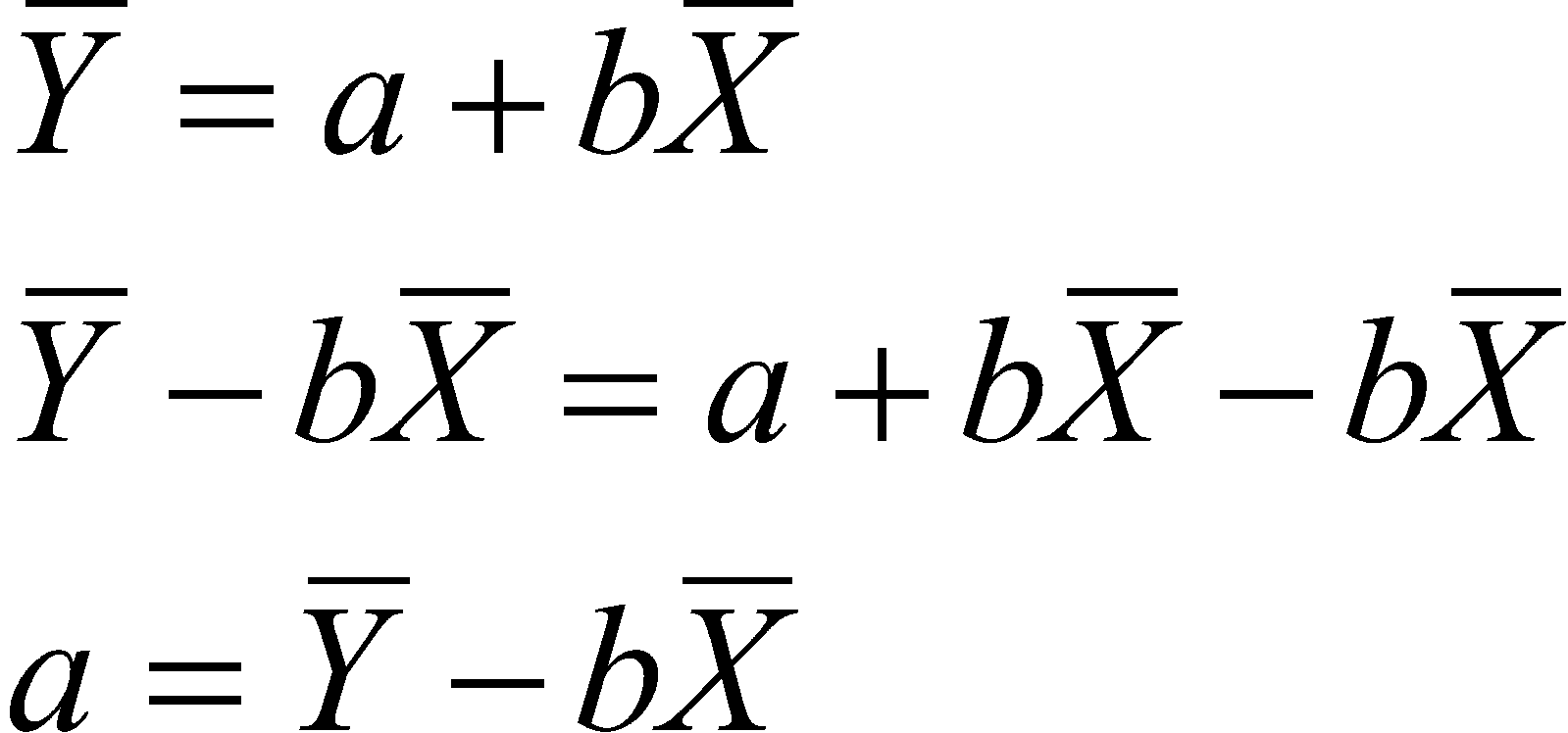
So You simply divide sy by sx and multiply the result by r.

The slope of the best-fit line can be a negative number following a negative correlation. For example, if an increase in police officers is related to a decrease in the number of crimes in a linear fashion, the correlation and hence the slope of the best-fitting line is negative in this case.

Slope is not same as correlation.Slope takes the untiless correlation and attaches units to it. Think of Sy divided by Sx as the variation (resembling change) in Y over the variation in X, **in units of X and Y**.

**Calculating Intercept**

So now that we have slope value (b), we can put it back into our formula to calculate intercept (shown as a below).



x\_bar and y\_bar are the mean values for variables x and y. So to calculate the y-intercept of the best-fit line, you start by finding the slope of the best-fit line using the above steps. Then to find the y-intercept, you multiply slope value mean of x and subtract the result from mean of y.

**Predict from the model**

Once we have a regression line with defined parameters - slope and intercept as shown above, We can easily predict the y\_hat (target) value for a new x (feature) value using the parameter values:

Remember that the different between y and y\_hat is that y\_hat is the y value predicted by the fitted model, whereas y carries actual values of variable (called the truth values) that were used to calculated the best fit.

Next we shall move on and try to code these equations in to draw regression line to a simple dataset to see all of this in action.

**Summary**

In this lesson, we learnt the basics of simple linear regression between two variables as a problem of fitting a straight line to best describe the data associations on a 2-dimensional plane. Remember this is only half the process. Once we have coded these equations as functions, we shall move on to calculating the loss in our model.

**P-value**

P-value is the probability of getting a test statistics that is at least as extreme as the one representing the sample data.

***Low P****=> reject NULL*  ***High P*** *=> fail to reject NULL.*

[α](https://en.wiktionary.org/wiki/%CE%B3%CE%BB%E1%BF%B6%CF%83%CF%83%CE%B1#Ancient_Greek):*Level of significance*

p< [α](https://en.wiktionary.org/wiki/%CE%B3%CE%BB%E1%BF%B6%CF%83%CF%83%CE%B1#Ancient_Greek) =>. *Statistically significant*

قيمة P هي احتمال الحصول على إحصائيات اختبار test statistics تكون على الأقل متطرفة مثل تلك التي تمثل بيانات العينة.

considering a null hypothesis test assess if our sample is extreme enough to reject the null hypothesis

النظر في اختبار فرضية صفرية تقييم ما إذا كانت عينتك المدقع بما فيه الكفاية لرفض الفرضية الصفرية Null hypothesis

considering a null hypothesis test assess if our sample is extreme enough to reject the null hypothesis

قيمة P اذا تقيس مدى تطرف العينة التي لدينا.

**TSS**

amount of variability before the regression : مقدار التباين قبل الانحدار

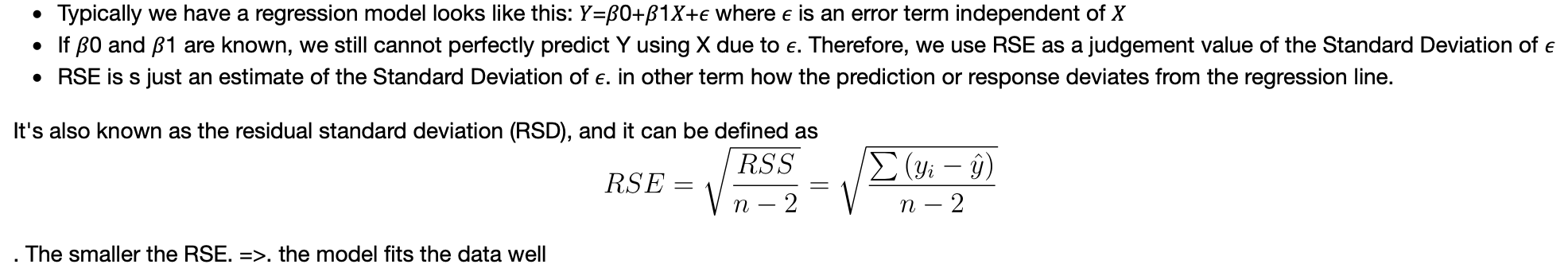
**RSS**

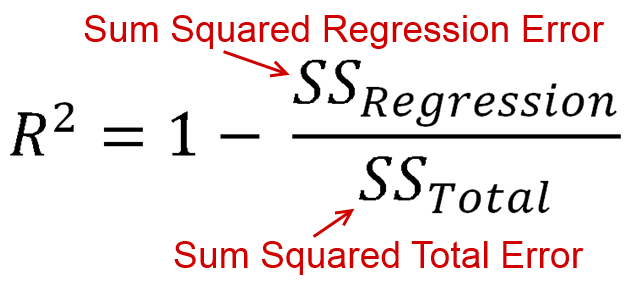
measure of variability left after the regression: قياس التباين المتبقي بعد الانحدار

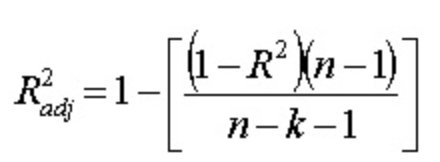
التباين الذي يفسره الانحدار:TSS-RSS

**RSE**:

Is an estimate of the standard deviation e (residual). It’s alose known as RSD







R2 : معامل التحديد

R2- adjusted : معامل التحديد المصحح

Mean Value: القيمة المتوسطة

Factor: مؤشر فئوي

Coef-regression /Slope : معامل الانحدار

Residuals: البواقي

Deviation: الانحراف

Dependent variable/Predicted Value: القيمة المتوقعة

Independent variable: متغير مستقل

IQR: الانحراف الرُبيعي

AIC: Akaiki Information Criteria

**Df**: degree of freedom.

*Answers the Questions:* ***What is the minimum Nbr of Observations required to estimate this regression*?**

n : Nbr of Observations

k : Nbr of explanatory (x) variables

في علم الإحصائيات ، يكون عدد درجات الحرية هو عدد القيم في الحساب النهائي للإحصاء الذي يتمتع بحرية التغيير. لذا فإن " عدد درجات الحرية " degree of freedom هو عدد الأجزاء المستقلة من المعلومات التي تدخل في تقدير المعامل Parameter.

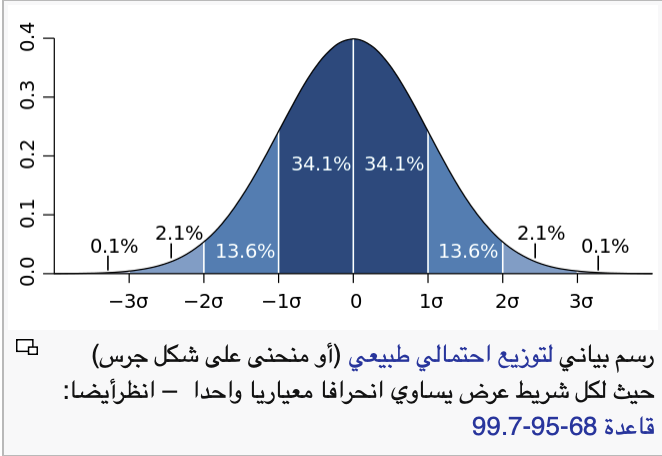
Chisq:

**chisq.test**(iris$**Petal.Length**, iris$**Petal.Width**)

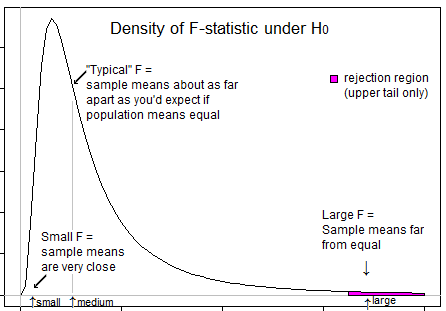
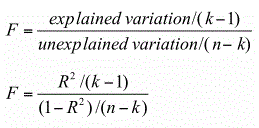
# We have a high Chi-squared value and a P-value of less than. 5% significance level.

# So, we reject the null Hyothesis and conclude that. Petal.Length and Petal.Width have

# a SIGNIFICANT relationship



**F statistic:**

**Linear Regression. Ev:**

**Q:1**

# fit model

mod = lm(tailL ~ totalL, possum)

# define new observations

new\_data = ….. (totalL = c(76.9,63.5))

# predict response

predict(mod, new\_data)

\_\_

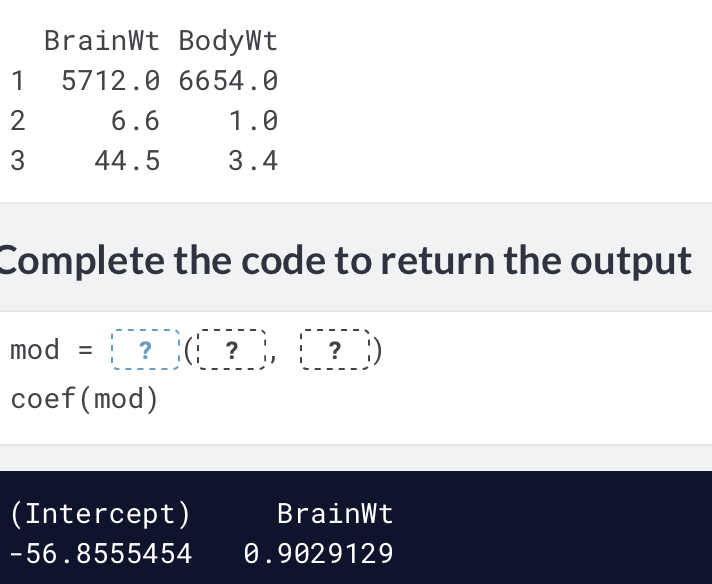
**data.frame** **list** **matrix**

Q:2

You are studying 39 species of mammals distributed over 13 orders. You would

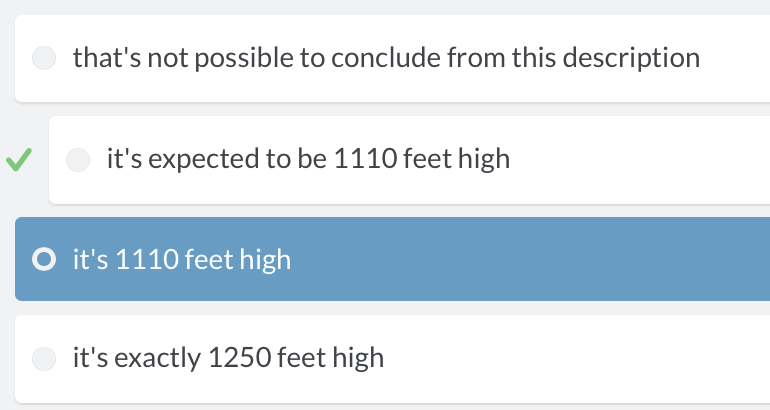
like to fit a linear regression model to understand a mammal’s total body weight

( BodyWt) as a function of its brain’s weight ( BrainWt). Here’s a preview of the mammal dataset:



A fitted model for the height of abuilding ( in feet) as a function of its number of floors is:

height = 90 +10\*floors. What can you conclude about the height of a Bahrain Four Season HOtel that counts 68 floors.



**The residual standard error reported for the regression model for poverty rate of U.S. counties in terms of high school graduation rate is 4.67. What does this mean?**

##### Answer the question

**50 XP**

##### Possible Answers

* *The typical difference between the observed poverty rate and the poverty rate predicted by the model is about 4.67 percentage points.*
* The typical difference between the observed poverty rate and the poverty rate predicted by the model is about 4.67%.
* The model explains about 4.67% of the variability in poverty rate among counties.
* The model correctly predicted the poverty rate of 4.67% of the counties.

# Standard error of residuals

One way to assess strength of fit is to consider how far off the model is for a typical case. That is, for some observations, the fitted value will be very close to the actual value, while for others it will not. The magnitude of a typical residual can give us a sense of generally how close our estimates are.

# View summary of model

summary(mod)

# Compute the mean of the residuals

mean(mod$residuals)

# Compute RMSE

sqrt(sum(residuals(mod)^2) / df.residual(mod))

RMSE(mod)

* View a summary() of mod.
* Compute the mean of the residuals() and verify that it is approximately zero.
* Use residuals() and df.residual() to compute the root mean squared error (RMSE), a.k.a. residual standard error.

Recall that the coefficient of determination (R**2**), can be computed as

R**2**=1−SSE/SST=1−Var(e)/Var(y)

where e is the vector of residuals and y is the response variable. This gives us the interpretation of R**2** as the percentage of the variability in the response that is explained by the model, since the residuals are the part of that variability that remains unexplained by the model.

The bdims\_tidy data frame is the result of augment()-ing the bdims data frame with the mod for wgt as a function of hgt.

* Use the summary() function to view the full results of mod.
* Use the bdims\_tidy data frame to compute the R2 of modmanually using the formula above, by computing the ratio of the variance of the residuals to the variance of the response variable.

# View model summary

summary(mod)

# Compute R-squared

bdims\_tidy %>%

summarize(var\_y = var(wgt), var\_e = var(mod$residuals)) %>%

mutate(R\_squared = 1- var\_e/var\_y)

var(bdims\_tidy$hgt)

# Interpretation of R2

The R2 reported for the regression model for poverty rate of U.S. counties in terms of high school graduation rate is 0.464.

lm(formula = poverty ~ hs\_grad, data = countyComplete) %>%

summary()

How should this result be interpreted?

##### Instructions

**50 XP**

##### Possible Answers

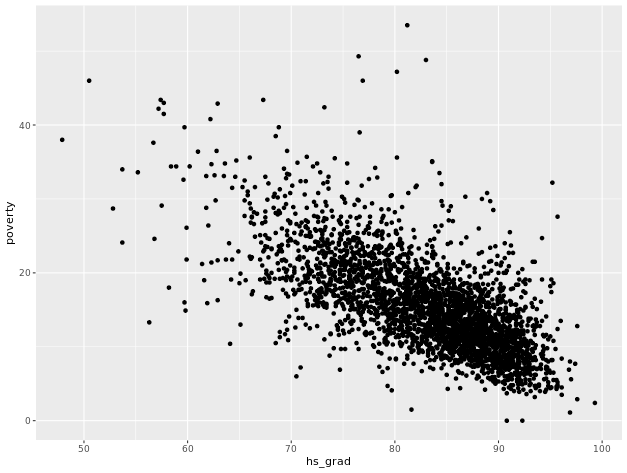
* *46.4% of the variability in high school graduate rate among U.S. counties can be explained by poverty rate.*
* 46.4% of the variability in poverty rate among U.S. counties can be explained by high school graduation rate.

Using the bdims dataset, create a scatterplot illustrating how a person's weight varies as a function of their height. Use color to separate by sex, which you'll need to coerce to a factor with factor().

# Characterizing scatterplots

This scatterplot shows the relationship between the poverty rates and high school graduation rates of counties in the United States.

Describe the form, direction, and strength of this relationship.



##### Possible Answers

* Linear, positive, strong
* Linear, negative, weak
* ***Linear, negative, moderately strong***
* Non-linear, negative, strong

Submit Answer

Use coord\_trans() to create a scatterplot showing how a mammal's brain weight varies as a function of its body weight, where both the x and y axes are on a "log10" scale.

ggplot(data = mammals, aes(x = BodyWt, y = BrainWt)) +

geom\_point() +

coord\_trans(x = "log10", y = "log10")

Same as : # Scatterplot with scale\_x\_log10() and scale\_y\_log10()

ggplot(data = mammals, aes(x = BodyWt, y = BrainWt)) +

geom\_point() +

scale\_x\_log10() +

scale\_y\_log10()

Profil du candidat à la majistrature supreme à qui je vais voter:

* + Il a eu la chance de taper sur un vieux clavier d’un ordinateur Durant son adolescence.
  + Un entrepreneur qui voit une grande opportunité dans cette course aux technologie de l’IA et autres.
  + Un candidat qui a lu le contrat social de J.J.Rousseau
  + Un candidat qui croix que la bonne gouvernance passe par la technologie et que le BlockChain pourra erradiquer la majeur partie de nos maux admonitratifs

J’ai beau cherché ce oisau rare, un beau film de J.Prévert dans les années 30s, mais aucun candidat ne répond aux critères ci-dessus.