***COURSERA: STATS W/ R SPECIALIZATION***

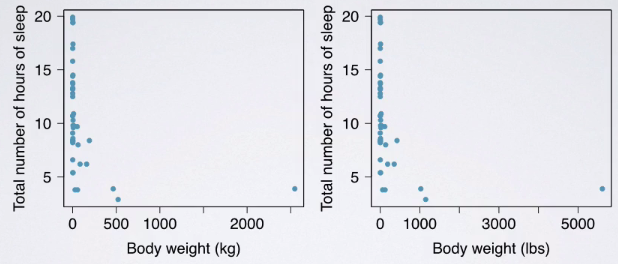
***COURSE 3 - Linear Regression and Modeling***

**WEEK 1 - Linear Regression**

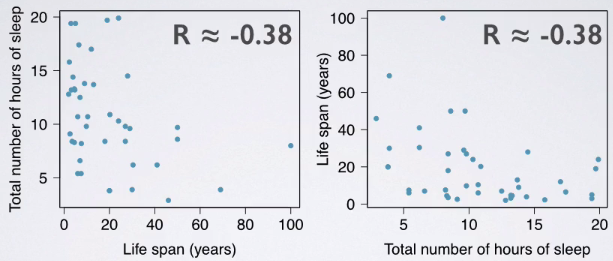
***5.1.1 Relationship between two numerical variables***

**Correlation**

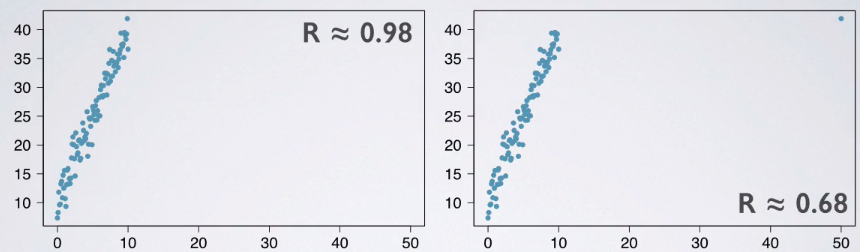
* **Correlation** = a measure of the strength of the linear relationship/association between 2 numerical variables.
* *Only* measure the *linear* association w/ correlation = **R**
* **Magnitude** = absolute value of the **correlation coefficient** **R** = Measures strength of linear association between 2 numerical variables.
* **Sign** of the correlation coefficient = direction of association.
* Correlation coefficient = always between -1 (perfect negative linear association) + 1 (perfect positive linear association), + correlation coefficient = 0 means no linear relationship.
* Correlation coefficient is **unit-less** + is NOT affected by changes in center or scale of either variable, such as unit conversions.



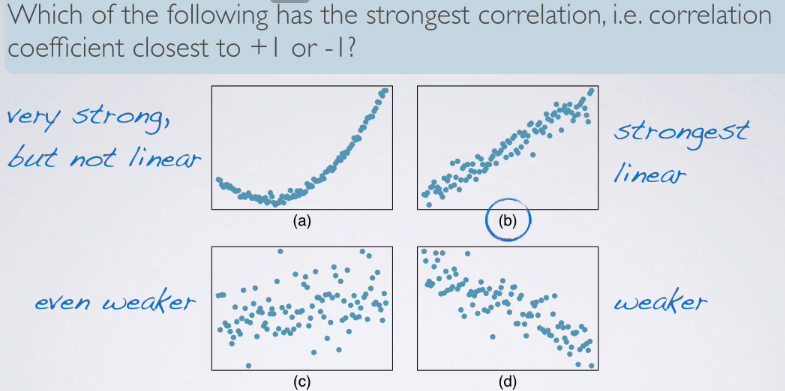
* Shape of relationship of the 2 variables looks very similar + for both plots correlation coefficient R = -0.34, no matter if x in terms of kg or lbs
* Probably wouldn't want to claim a linear relationship between these variables b/c the relationship between these 2 variables do not appear to be linear, but to demonstrate the fact that changing units does not affect the correlation coefficient, these plots still serve a purpose.
* **Correlation of X w/ Y = same as of Y / X**.
* Even if you swap the axis, correlation coefficient R should stay the same



* Correlation coefficient R = **sensitive to outliers**.



* Even w/ 1 outlier, b/c the correlation coefficient is sensitive, it will change greatly.
* So we can see that moving even one data



**Residuals**

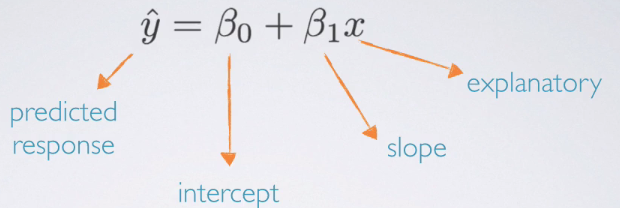
* **Residuals** = leftovers from the model fit = the difference between the observed + the predicted Y
* Observed data = the model fit + the residuals



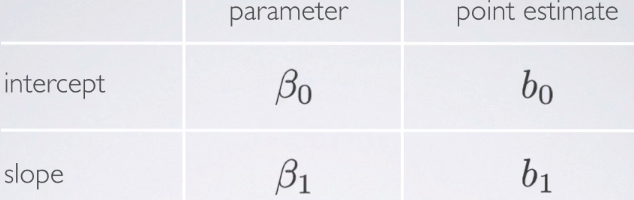
* Residuals tell us how much less/greater an actual DP is than what a model predicts. Similarly in D.C., the observed value is

**Least Squares Line**

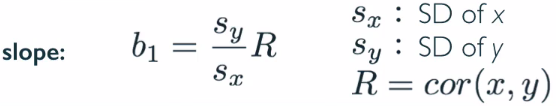
* Need the **least squares line** b/c can't simply add up all residuals b/c some are negative + some positive, depending on whether the model is over or underestimating certain DPs.
* Need to come up w/ a more clever approach:
* Minimize sum of *magnitudes* (absolute values) of their residuals
* Minimize sum of *square residuals* = the least squares
* Least squares = most commonly used approach + is also easier to compute by hand + w/ software.
* But most importantly, a residual twice as large as another = MORE than twice as bad (same idea as SD)



* If we have data from a sample, use that sample to estimate unknown population parameters.



* Greek = population, Latin = sample
* To minimize sum of squared residuals, could actually use calculus + calculate slope + intercept
* Can also use shortcut formula 🡪 **slope b1 = SD\_y / SD\_x (rise / run) times correlation coefficient R**



* Ex: SD of % living in poverty = 3.1% + of % of high school graduates = 3.73% in our data set. Given that the correlation between these variables = -.75, what is the slope of the regression line for predicting % living in poverty from % of high school graduates?

> sd.pov <- 3.1

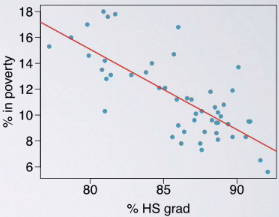
> sd.hs <- 3.73

> R <- -.75

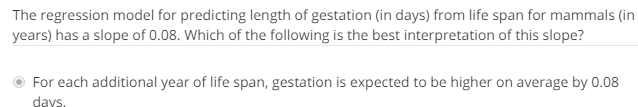
> (slope <- (sd.pov/sd.hs)\*R)

[1] -0.6233244

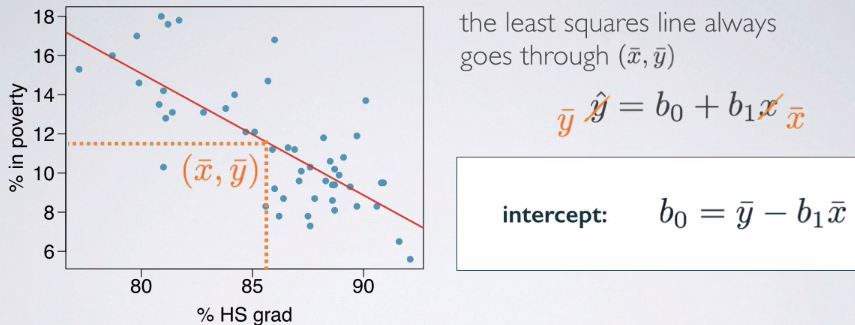
* Note: sign of the correlation coefficient R.
* Conceptually speaking, this is true, b/c we clearly seeing a negative relationship between the 2 variables so it makes sense the slope is negative.



* Mathematically speaking, remember SD = square root of the **variants** + is a measure of variability
* So the SD of y + x are always going to be positive, so we’re always dividing 2 positive #’s + multiplying by a negative or positive value depending on the direction of the relationship between the 2 variables.
* So, mathematically speaking, **sign of the slope always = sign of the correlation coefficient**.
* Calculating the slope by hand is clearly very simple and actually sometimes
* For each % point increase in HS graduate rate, we’d expect the % living in poverty to be lower, *on average*, by 0.62% points.



* Slope = relationship between explanatory + response variable = how we expect response variable to change as we increase explanatory variable by 1 unit.
* When interpreting these, also make sure if dealing w/ an *observational study, avoid causal language*
* We expect this to happen on average
* Next, we want to estimate the intercept (where regression line crosses y axis).
* For this, make use of the property that least squares line always goes through x\_, y\_ = always goes through mean of y + x.
* We know we can write the linear model: y^ = b0 + b1\*x 🡪 plug in x\_ + y\_ in b/c we know the line MUST go through this point
* Then rearrange things a bit to get the formula for the intercept.



* Intercept = average value of response variable – slope times average value of explanatory variable.
* Given the average % living in poverty = 11.35% + average % of HS grads = 86.01%, what is the intercept of the regression line for predicting % living in poverty from % of high school graduates?

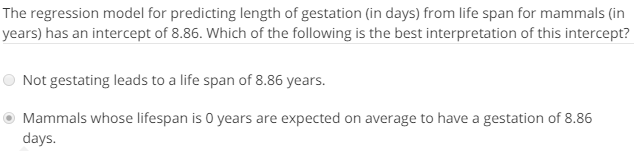
> pov\_bar <- 11.35

> hs\_bar <- 86.01

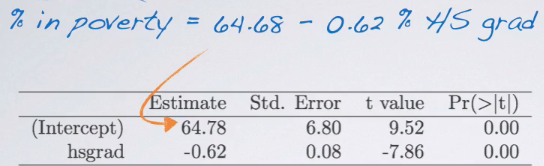
> (int <- pov\_bar - slope\*hs\_bar)

[1] 64.96213

* Understand that the regression line always goes through the *center of the data* + we interpret the intercept as where the regression line crosses the Y axis.
* In other words, it's the expected value of the response variable when the explanatory variable = 0
* So, in context, what we can say states w/ no HS graduates are expected, *on average*, to have 64.68% of residents living below the poverty line.
* Does this seem realistic, that there would be states in the U.S. w/ absolutely no HS graduates?
* Looking at the data we have, it actually seems very unlikely as all states in the US actually have HS graduation rate varying somewhere between 75% to ~95%, maybe.
* So mathematically speaking, this is a construct that is important for putting together a linear model
* However, in context, it is not a very useful number.



* So putting the info together from the previous 2 steps, we can write our linear model as a predicted % living in poverty (64.68%, the intercept) - minus 0.62 (slope) \* % that’re HS graduates.



* Estimate column = we can find our **parameter/coefficient estimates** for slope + intersect.
* So to recap, **intercept**: when x = 0, y is expected to = the intercept (may be a meaningless value in context of the data + in those cases it might only be serving to adjust the height of the line)
* The interpretation of the slope is slightly different = about the relationship between the 2 variables
* For each unit increase in X, Y is expected to be higher/lower on average by the value of the slope.
* Depending on type of study, be careful about interpreting the slope as causal vs. correlational

***5.1.2 Linear Regression w/ 1 Predictor***

**Prediction + Extrapolation**

* Least squares line allows us to evaluate the relationship between 2 numerical variables
* Here, **prediction** = using a liner model to predict the value of the response variable for a given value of the explanatory variable = simple as plugging in value of x in the model + seeing the resulting y

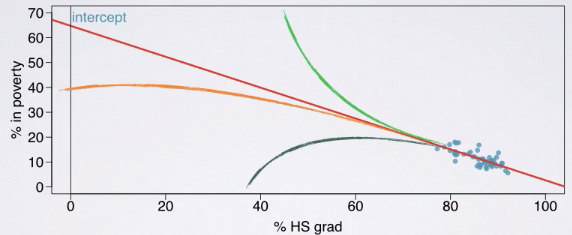


* What is the predicted % living in poverty in states where the HS graduation rate = 82%?

> int + (slope\*82)

[1] 13.84953

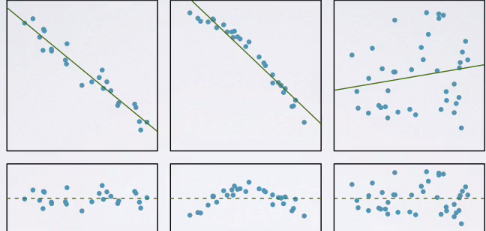
* This model predicts that in states where HS grad rate = 82%, predicted % living in poverty is, on average, 13.84%
* Prediction = useful and powerful tool, but be careful b/c applying a model estimate to values outside the realm of the original data = **extrapolation**.
* **Extrapolation** = simply plugging in a value of x into the model that was *not* in the range of the original observed data.
* Sometimes the intercept might be an extrapolation as well.



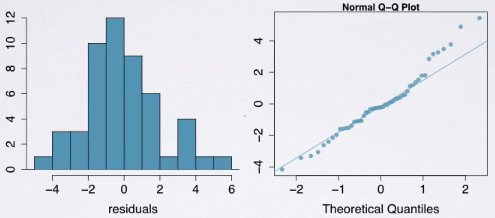
* Plugging in 0 for x into the model gives the intercept, but we have no idea if this line *actually* extends out to infinity as a straight line, or if it curves down, up, down even more (any are possible)
* Since we don't have data from states w/ such HS school grad rates, it's really not wise to believe the value of the intercept is a plausible value of poverty rate, if HS grad rate is 0 for a state.
* **Extrapolation** doesn't have to happen in *extremes* though.
* According to our linear model, the predicted % living in poverty in states where HS grad rate = 20% is a very simple problem, mathematically speaking.
* Plug in 20 for the HS grad rate + formula spits out a predicted value for poverty
* But is it wise to do? Before doing so, always look back at the data, 🡪 in a scatter plot or at least the summary statistics.
* Look for something that will tell us whether 20% is w/in the realm of the data observed or not.
* **If it is not, we do not want to be doing this prediction b/c it would yield an unreliable estimate.**

**Conditions for Linear Regression**

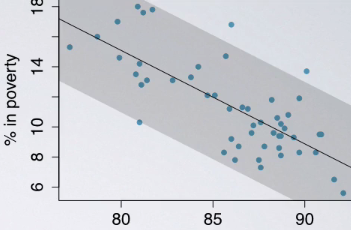
* As w/ any technique, there are conditions associated w/ linear regression = **linearity**, nearly normal **residuals**, + constant **variability/homoscedasticity**.
* **Linearity** = relationship between explanatory + response variables should be linear.
* Makes sense b/c we're using a liner model to predict response from explanatory.
* There are indeed methods for fitting a model to non-linear relationships
* To check if linearity condition has been met, use a scatter plot of the data or a **residuals plot**.



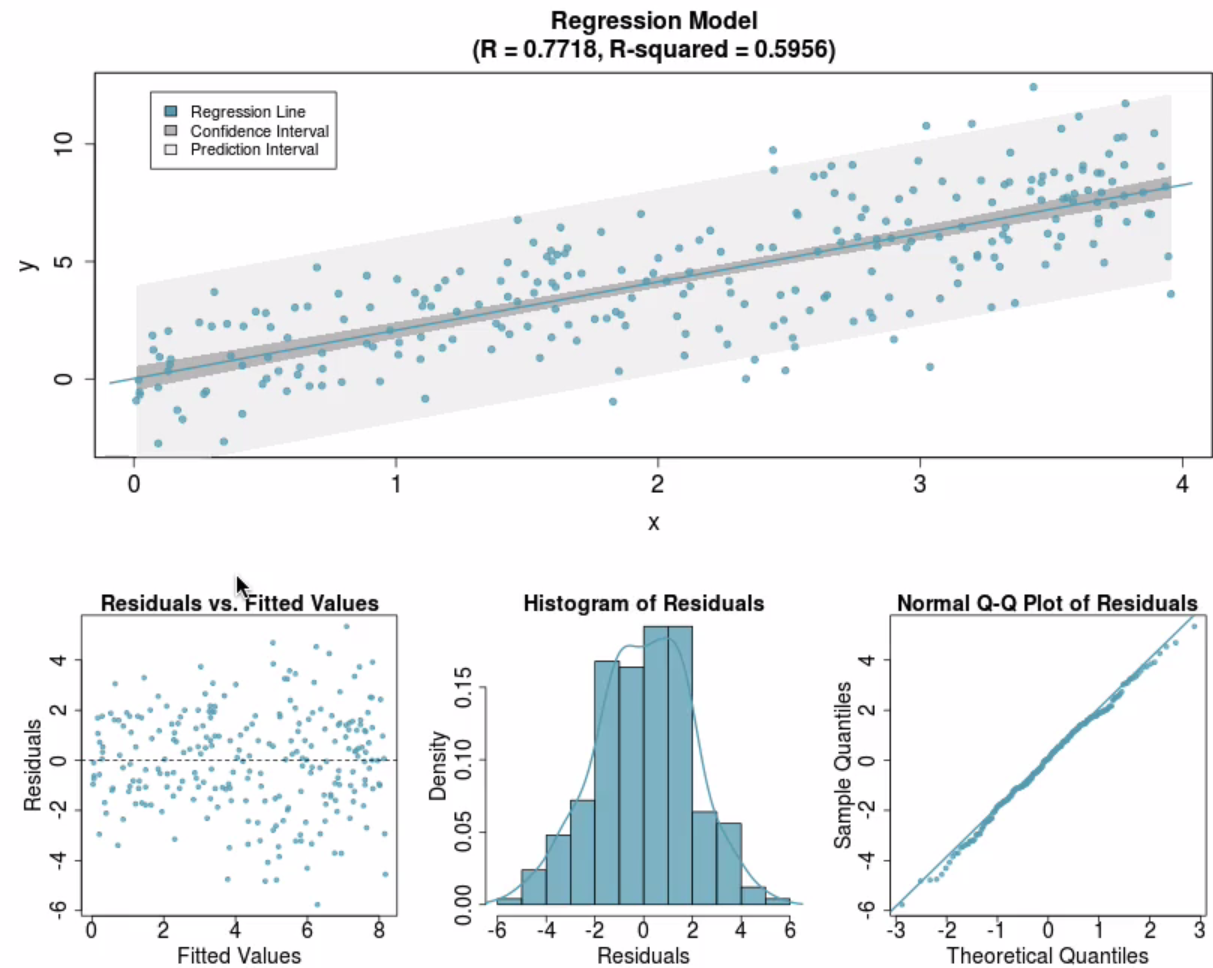
* Based on the scatter plots, plot 1 seems to display a pretty linear relationship and have residuals close to the ideal residual (error) of the horizontal line = 0
* Ideal residual = 0 = means the DP falls exactly on the regression line w/ no difference between predicted + observed values for that particular DP.
* W/ random chance, this is unlikely to happen, but we like **small residuals** + we want residuals in a residuals plot to be RANDOMLY scattered **around zero**.
* Some will be positive + some negative (falling above + below the regression line)
* **We want absolutely no pattern b/c we want the linear model to capture all of the pattern in the data + anything that's left over to be simply random scatter**.
* In the residuals plot, we look for a RANDOM scatter around 0.
* Nearly normal residuals = residuals should be nearly normally distributed, centered at 0.
* May not be satisfied w/ unusual observations that don't follow the trend of the rest of the data
* Check this condition w/ a histogram or a **normal probability plot of residuals (QQ plot).**



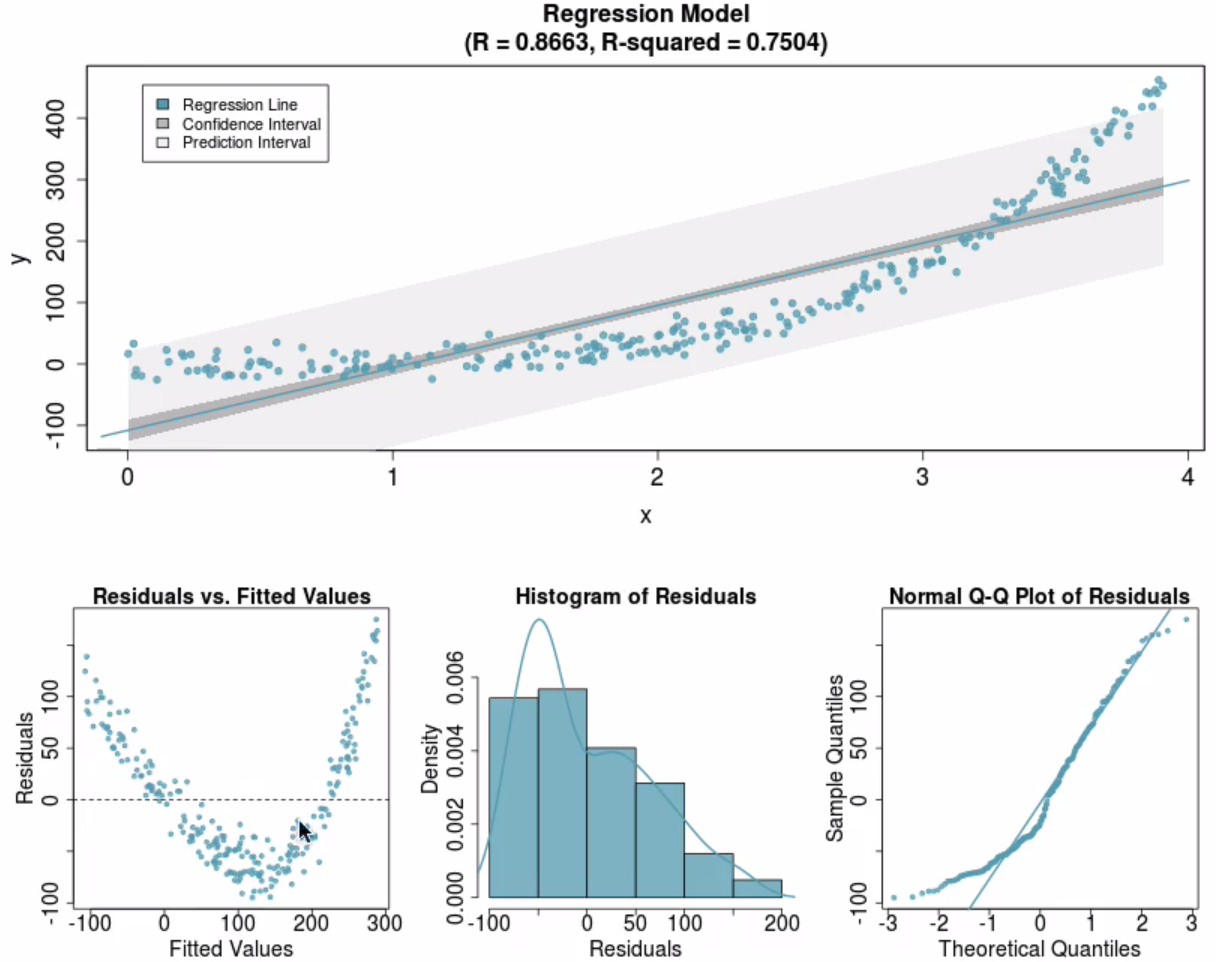
* The histogram shows a somewhat symmetric distribution, indeed centered at 0, + the normal probability/QQ plot shows some values on the higher end of the tail that actually steer away from normality (just a few observations)
* **Homoscedasticity** = variability of points around the least squares line should be roughly constant.
* Implies variability of residuals around the 0-line should be roughly constant as well.
* Can check this using a residuals plot.

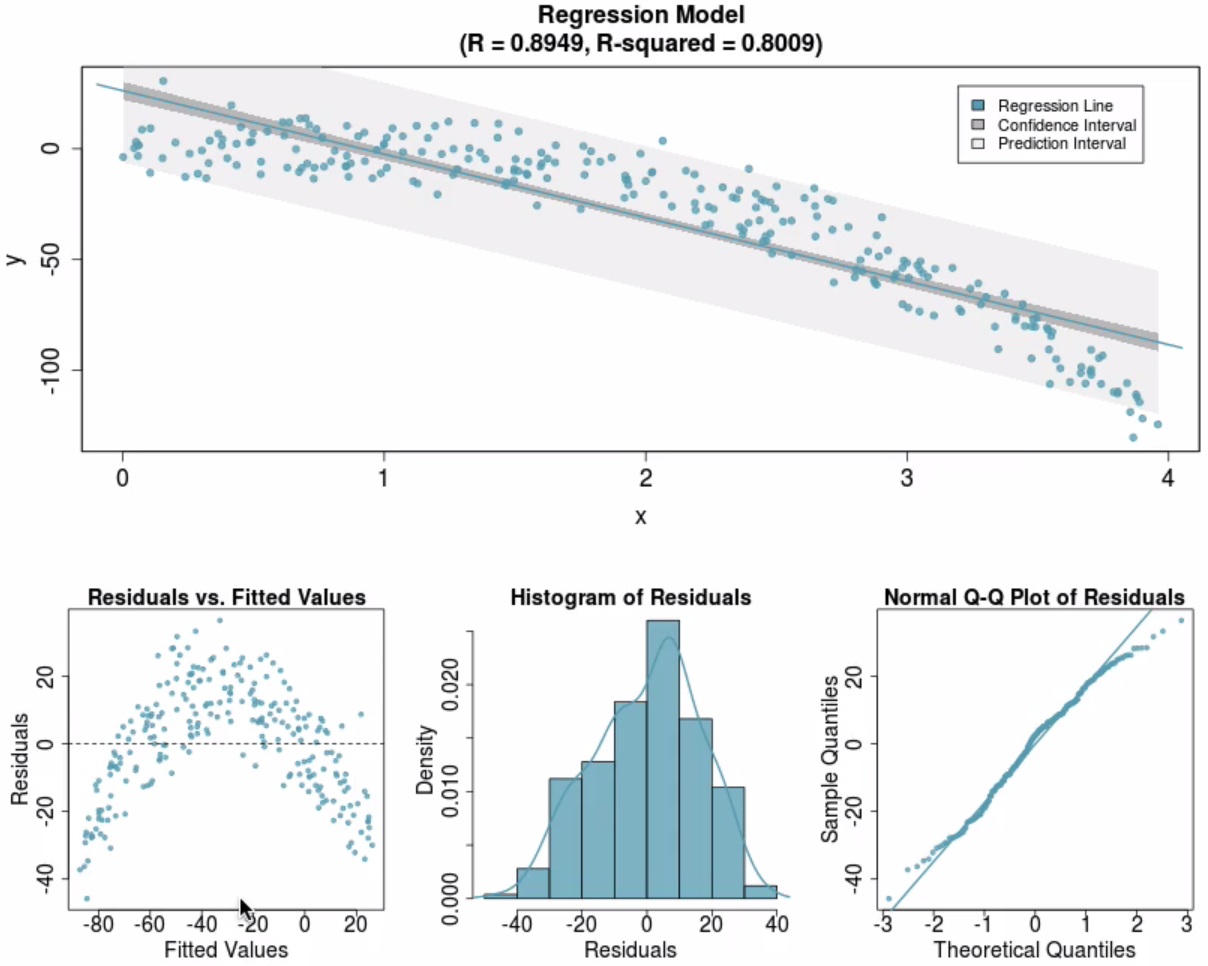
* On the scatter plot, see that as x varies, variability of the data do not vary a whole lot + actually seem to be captured around the constantly variable grey band around the regression line
* In the residuals plot, can confirm the variability of the residuals do not vary by the value of the explanatory variable.
* Checking **regression diagnostics** is somewhat of an art + takes lots of practice to be able to tell when a condition has been met or has not
* Ex: Linear trend between explanatory + response.



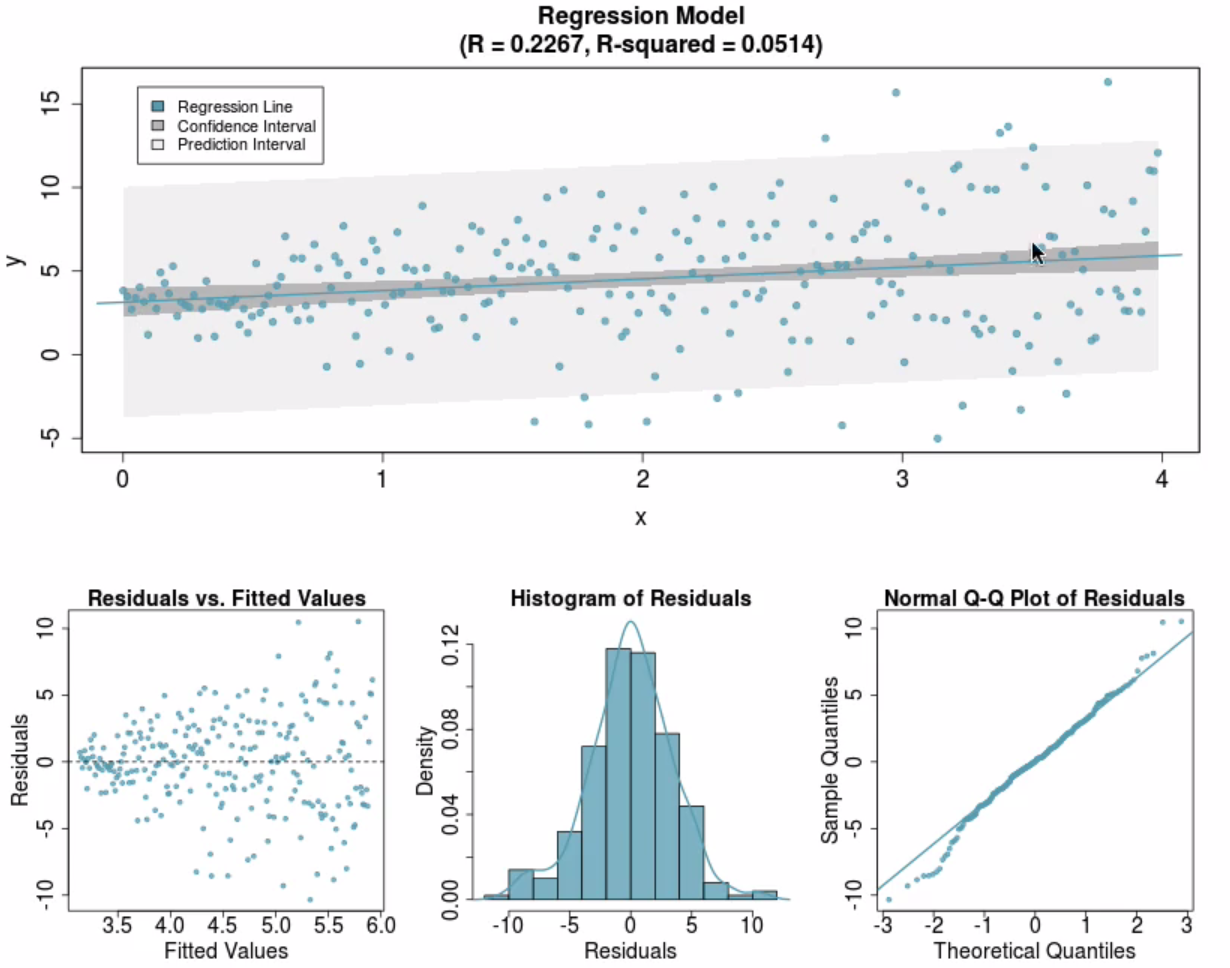
* See a completely random scatter in residuals plot, histogram of residuals is centered at 0 + shape of the distribution looks fairly symmetric, + the normal probability/QQ plot w/ has almost all dots aligned on the straight line (also indicates the distribution of residuals is nearly normal).
* Let's take a look at another example.



* Now see a curved relationship between response + explanatory variable
* Residuals plot is no longer displaying random scatter around 0, histogram of the residuals shows a right skew that is also seen in the normal probability/QQ plot as well.
* In this case, it would NOT be appropriate to fit a linear model to predict y from x



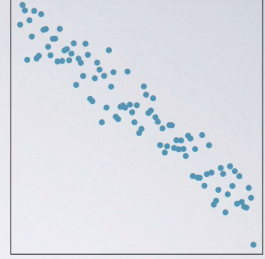
* Here again, we have a curved relationship (not as extreme) + it might actually be somewhat difficult to tell from the scatter plot if we didn't have the grey band around it.
* But, the residuals plot highlights very well for us that the relationship is not linear b/c the we do not see *random* scatter *around 0*.
* Histogram of the residual shows a distribution centered at 0 but is not very normal + the normal probability/QQ plot also shows a lot of the points on the tails steer away from normality.
* These were 2 examples where linearity has not been met.
* What if **homoscedasticity** has not been met?



* See **fan-shaped data** = when explanatory variable is low, variability of the response is low as well, but as x increases, the data are fanning out such that the response variable becomes more + more variable.
* Can clearly see as x increases, variability of the residuals increase as well.
* Histogram of the residuals looks fairly symmetric + is centered at zero, but looking at the normal probability/QQ plot, see steering quite a bit away from normality at tails

**R Squared**

* Once you check your conditions + are convinced a linear model is indeed appropriate to model the relationship between your response + explanatory variables, next step = check the **fit** of your model.
* **R2** = strength of the fit of a linear model = square of the correlation coefficient R.
* R2 = what % of variability in the response is explained by the model (explanatory).
* Remainder of the variability is explained by variables NOT included in the model.
* Since R2 = square of correlation coefficient, it’s always between 0-1
* Interpretation of R2 = .5625 for a model predicting % living in poverty via % of HS grad rate = **56.25% of variability in % of residents living in poverty among the US is explained by the model**
* R2 for a = 92.16% 🡪 correlation coefficient = square root of this = 0.96
* BUT check the scatterplot to see if this is positive or negative .96

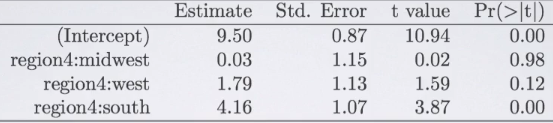


**Regression with Categorical Explanatory Variables**

* Fit a regression model where the response variable = numerical + explanatory variable = categorical
* Explanatory variable = region 🡪 0 if state eastern + 1 if western



* For eastern states, plug in 0 for x = predicted poverty rate for eastern states = 11.17%.
* For western states, plug in 1 (west = success, east = failure) = predicted poverty rate = 11.55%.
* In regression models w/ explanatory categorical variables, always code 1 level of that categorical variable to be the **reference level** = the level that we plug in 0 for.
* In this context, **intercept** tells us the model predicts an 11.17% average poverty % in eastern states
* We have to plug in a numerical variable b/c we can’t simply plug in a level of a categorical variable + solve a *mathematical* equation.
* Making due by labeling some levels “successes” + some levels “failures” + denoting w/ 0‘s + 1’s
* Slope, on the other hand, is, once again, about the actual relationship between explanatory + response variables.
* tells us this model predicts average poverty % in western states = .38% higher than in eastern states
* New variable region4 = 4 levels = Northeast, Midwest, west, south.



* Write the linear regression model based on the regression output.
* To write the regression model all we need are the slope + intercept estimate.
* **y = 9.5 + .03\*mw + 1.79\*w + 4.16\*s**



* Here, reference level = NE = identified as the **0-level** (level not in the regression output)
* Given what we know so far, we can calculate predicted poverty rate for western states w/ our linear model
* Since we're looking for a western state, plug in 0 for MW, 1 for west, + 0 for south
* **= 9.5 + 0 + 1.79 + 0 = 11.29 = 11.29%.**
* The model predicts the poverty rate for western states = 11.29% on average.