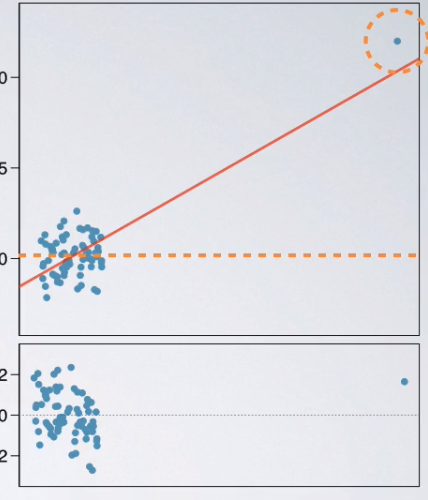
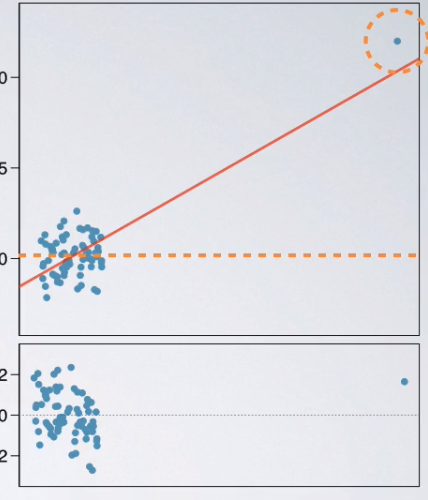
***COURSERA: STATS W/ R SPECIALIZATION***

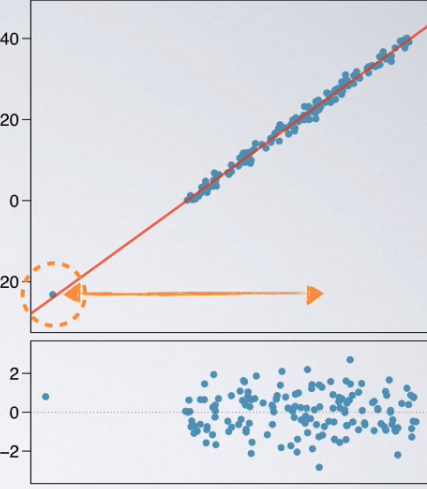
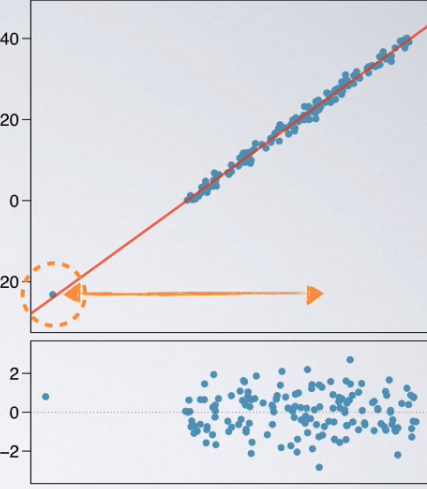
***COURSE 3 - Linear Regression and Modeling***

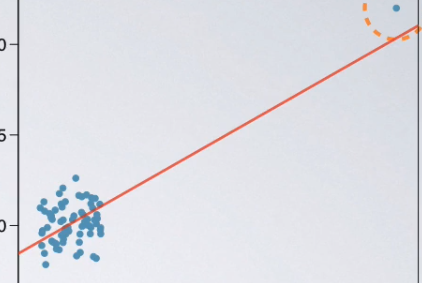
**WEEK 2 - More about Linear Regression**

**Outliers in Regression**

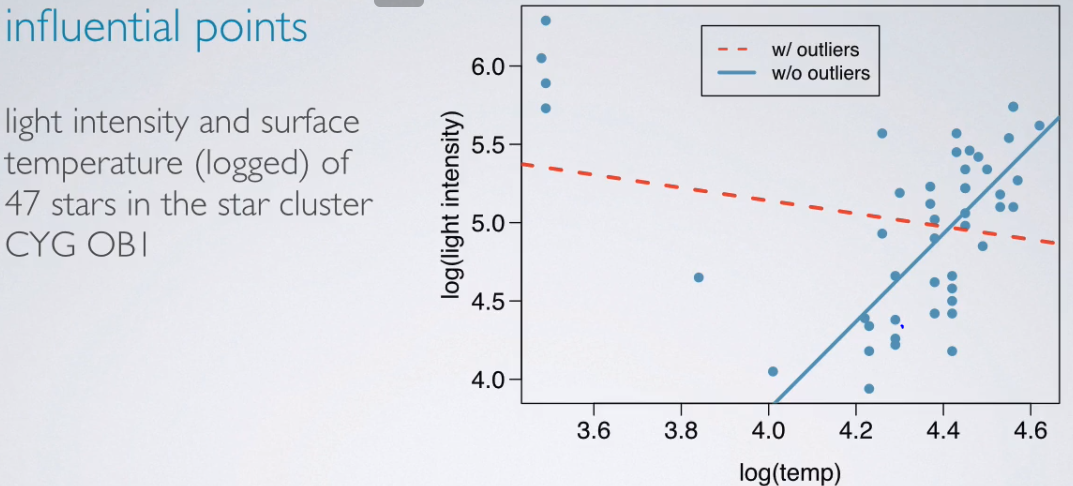
* How do outliers influence the least squares line?
* Think about where the line would go if a particular outlier was not there



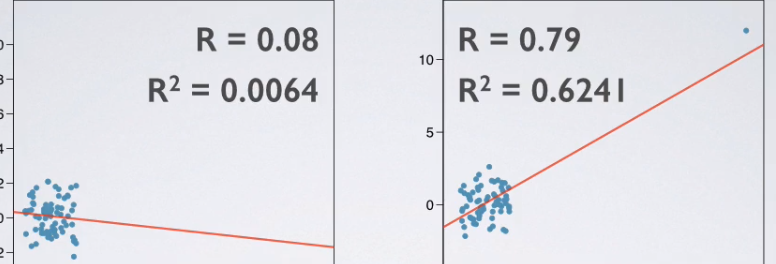
* 1st plot there would be absolutely no relationship between the 2 variables, b/c they’re completely randomly scattered 🡪 line would look horizontal.
* W/out the outlier = no relationship between x + y, + 1 single outlier makes it appear as if there is.
* Various types of outliers + depending on the type is how we decide how to handle them.
* In general, **outliers** = points that fall away from the cloud of points.
* **Leverage points =** outliers that fall horizontally away from the center of the cloud but don't influence the slope of the regression line are called
*  
* **Influential points** = actually influence slope of the regression line
* Usually, these are also high leverage
* To determine if a point is influential = visualize the regression line w/out the point see if slope of the line changes considerably



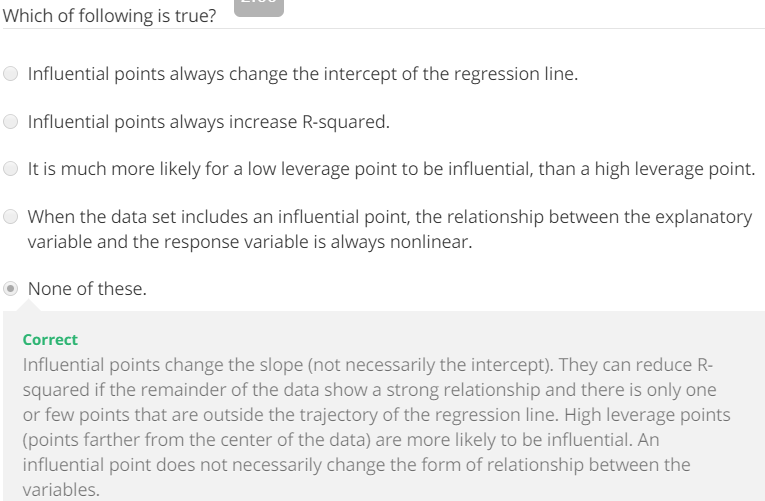
* When we ID an influential point 🡪 decide whether to leave this DP in the analysis or take it out
* If influential, be very careful about leaving it in b/c it's *definitely* going to affect estimates + all decisions made based on the results of the analysis.



* See 2 different types of stars, lower + higher temperatures.
* Solid blue = how regression model looks if we ignore outliers
* Red dash = how regression model looks if we include outliers.
* See 4 stars w/ the lower temps + obviously, the red-dashed line = not a good fit for these data.
* In this case, might want to split data into 2 = stars w/ lower temps + stars w/ higher temps + model the 2 groups separately.
* Don't want to blindly get rid of outliers b/c they actually might be the most interesting cases.
* Perhaps stars that’re much colder than the others are indeed more interesting to look at.
* But we don’t want to lump them along w/ the stars w/ a higher temp + try to model all of them together
* Influential points tend to make life more difficult, but do NOT always reduce R2

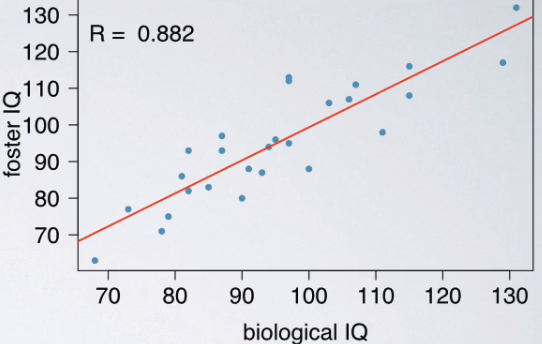


* 1st plot = correlation coefficient is very low = R2 is pretty low as well, at 0.0064.
* 2nd plot = see increase in correlation coefficient as well as in R2
* Even though we’d never want to fit a linear model in the 2nd plot, we actually see a much higher correlation + a much higher R2.
* Always viewing a scatter plot before fitting a 2D model.
* If simply deciding on whether or not a model is a good fit by looking at the correlation coefficient + R2, we’d never catch the anomalies in the data that's driving the entire relationship.

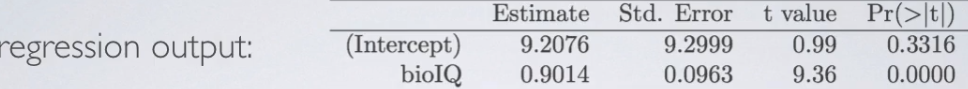


**Inference for Linear Regression**

* Can also use linear regression models to do **inference** = hypothesis testing for significance of a predictor + CI for slope estimate.
* 1966 🡪 Cyril Burt published a paper, The Genetic Determination of Differences in Intelligence: A Study of Monozygotic Twins Reared Apart.
* Data consists of IQ scores for an assumed random sample of 27 identical twins, 1 raised by foster parents, the other by the biological parents.
* This study got a lot of criticism saying the data may have been either non-random, non-representative, or entirely falsified.
* Regardless, we're going to be working w/ the original data set from the paper

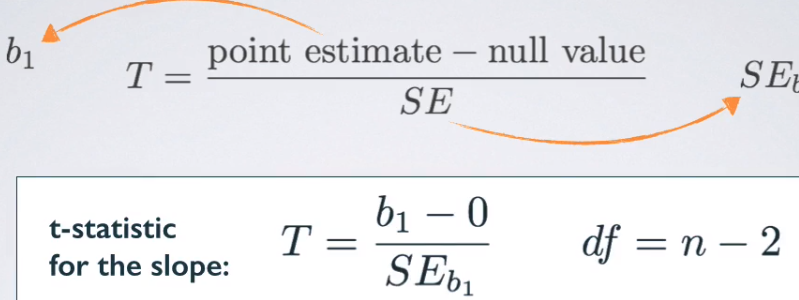


* See the relationship between foster twin + biological twin IQ = as 1 goes up, so does the other
* Positive + relatively strong relationship w/ a correlation coefficient R = 0.882.
* Results of this study can be summarized w/ a regression output + based on it, write our linear model





* To assess fit of the model, look at R2 = 0.78 = **78% of the variability in foster twin IQs can be explained by biological twins IQ's.**
* W/in the framework of inference for regression, we're going to be doing a hypothesis test on slope
* Overall question: Is the explanatory variable a significant predictor of the response variable?
* The null, as usual, says there's nothing going on = explanatory variable is NOT a significant predictor of the response variable = there's no relationship/slope of the relationship = 0
* The alternative says there is something going on = explanatory variable IS a significant predictor of the response = there IS a relationship between these 2 variables/slope of the relationship is != 0
* H0: B1 = 0 (B1 = population parameter for the slope)
* H1: B1 != 0
* In linear regression, always use a **t-statistic** for inference = a point estimate - a null value / a SE
* In this case, our point estimate = slope estimate, b1 + SE = SE of this estimate.



* In the null, we set the b1 = 0 🡪 no relationship/horizontal line
* Whenever we have a t-score, we also have a dF associated w/ it + in this case it’s n – 2
* We always lose a dF for each parameter we estimate
* When we fit a linear regression, even if only interested in slope, you always also end up estimating an intercept as well
* Since we're estimating both an intercept + a slope, we're losing 2 dF
* For calculating the test statistic, make use of the regression output

> b0 <- 9.2076

> b1 <- 0.9014

> se <- .0963

> null <- 0

> (t <- (b1 - null) / se)

[1] 9.360332

* P-value = AUC (for the t curve) that's > 9.36 or < -9.36 (2-sided alternative)

> n <- 27

> dF <- n - 2

> pt(t, dF, lower.tail = F)\*2

[1] 1.196496e-09

* As you can imagine, 9.36 SE’s from the null value is a really unusual outcome, + therefore the p-value is approximately 0.



* Just like we can do HT’s for the slope, we can also do a CI = point estimate +/- a margin of error.
* In this case, point estimate = b1 + margin of error can be calculated as usual, as a critical value \* a SE
* Remember, in linear regression, we always use a t-score, so we to use a **t\*** for our critical value, + the SE of the slope comes from the regression output.

> pe <- b1

> (t.crit <- abs(qt(p = .025, df = dF)))

[1] 2.059539

> mOe <- t.crit\*se

> (lower <- pe - mOe)

[1] 0.7030664

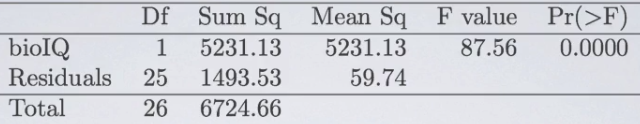
> (upper <- pe + mOe)

[1] 1.099734

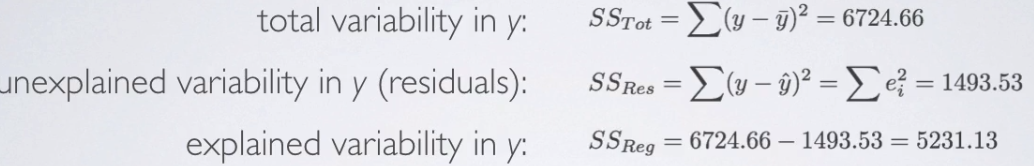
* **We are 95% confident that, for each additional point on the biological twins' IQs, the foster twins' IQs are expected, *on average*, to be higher by 0.7 to 1.1 points.**
* Inference on the *intercept* is rarely done b/c in some cases, the intercept is not very informative
* Usually when we fit a model, we want to evaluate relationships between variables involved in the model, + the parameter that tells us about these relationships = the slope, not the intercept.
* Always be aware of the type of data you're working w/.
* Is it a random sample, a non-random sample, or a population data?
* Statistical inference + the resulting p-value are completely meaningless if you already have population data 🡪 use statistical inference when we have a sample + want to say something about the unknown population.
* If you have a sample that is non-random (biased in some way), results that arise from that sample are going to be unreliable as well
* Ultimate goal = to have *independent observations* to be able to do statistical inference.
* Remember, we like random samples + large samples, but don't want them to be *too* large (check 10% rule if sampling w/out replacement)

**Variability Partitioning**

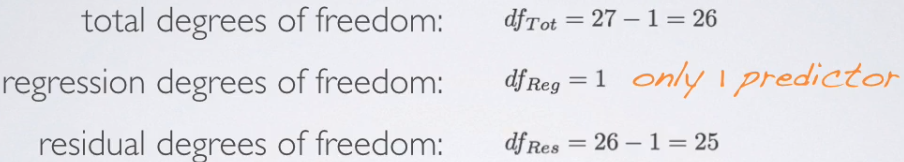
* So far w/in the framework of regression we've used a t-test to evaluate strength of evidence for hypothesis tests for the slope of relationship between x + y.
* Alternatively, can also consider *variability in y explained by x*, compared to **unexplained variability**
* Remember % of variability in y explained by x = **R2**.
* We like large R2’s so, could we use that notion to also do this hypothesis test from another POV?
* **Partitioning variability** in y to explained + unexplained variability measures requires **ANOVA**
* Can get an ANOVA-type output for a regression model



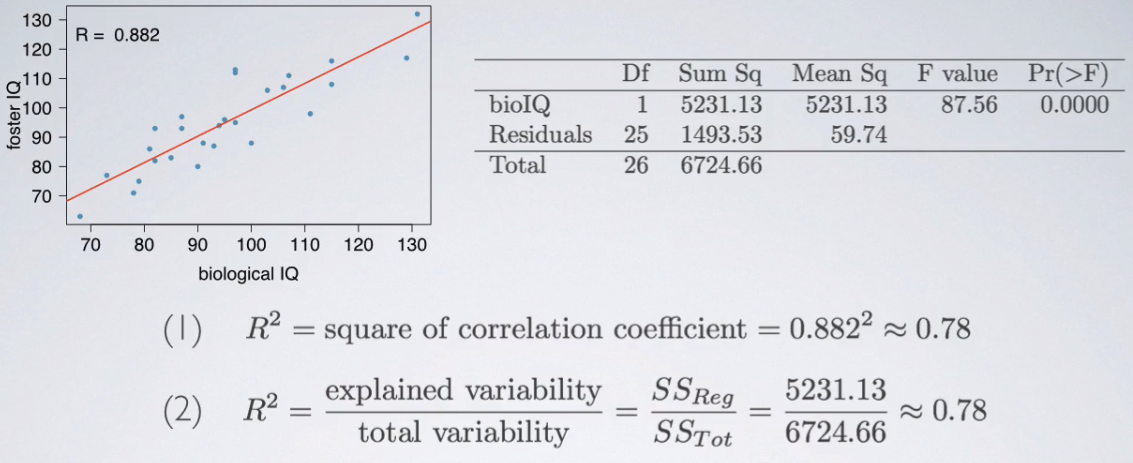
* **Sum of squares** = data on total variability 🡪 total variability in y = **total sum of squares** = looks very much like the variance of y if we didn't scale by sample size.
* **Unexplained variability in y** w/in context of regression = **residual sum of squares**
* Imagine you have the residual for every single DP in a data set, square those + add them up.
* Then, explained variability simply becomes the *balance of these 2* #’s
* Remember the explained + unexplained variability = total variability.



* The total dF = Sample size – 1
* Consider dF associated w/ a regression = only 1 predictor + this Df = 1
* Residual dF = balance of these 2 🡪 26 – 1 = 25.



* Next, get **mean square** = a measure of *average variability* = sum of squares / associated dF
* **Regression mean square =** regression sum of squares / regression dF
* **Residuals mean square** = residual sum of squares / residuals dF
* Finally, **F statistic** = ratio of explained to unexplained variability = Regression MS / Residuals MS
* Hypothesis test goal = **Is the explanatory variable a significant predictor of the response variable?**
* H0: slope = 0 H1: slope != 0
* We have a pretty small p-value, meaning we’d reject the null + in this case it means the data provided convincing evidence the slope is significantly < 0.
* In other words, the explanatory variable is a significant predictor of the response variable.
* Remember R2 = the proportion of variability in y explained by the model.
* If large, there's likely a linear relationship between x + y.
* If small, the evidence provided by the data may not be very convincing.
* 2 ways to calculate R2.
* using the correlation coefficient 🡪 square of correlation coefficient.
* Formal definition of R2 = directly calculated as a **proportion of explained to total variability**.



* R2 = .78 = IQ's can be explained by the model/the biological twins IQ's.