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

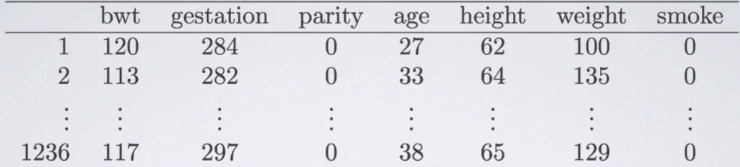
***COURSE 3 - Linear Regression and Modeling***

**WEEK 3 – Multiple Linear Regression**

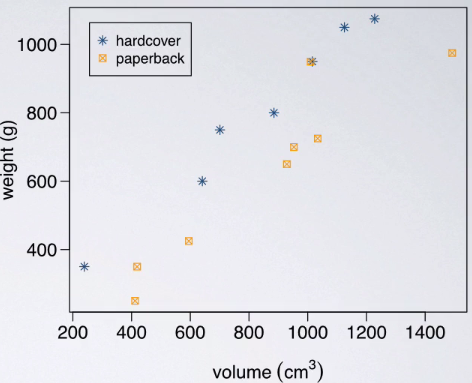
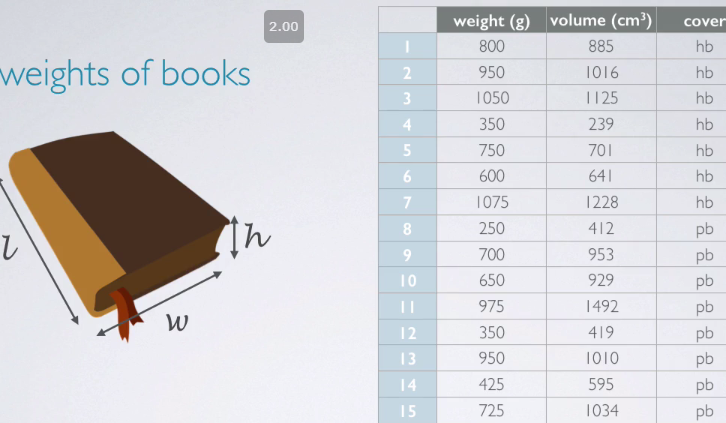
***5.3.1 Regression with multiple predictors***

**Multiple Predictors**

* Set of 1236 observations w/ data from birth weight of babies + a variety of variables on the baby, the birth, or the mother.



* Want to predict birth weight of babies from a **set** of these variables = multiple linear regression
* Explanatory variables in set can be numerical or categorical.
* Build a model predicting weight of a book using its volume + cover type



* It appears that paperbacks orange squares (PB) generally weigh less than HC books.
* Relationship = similar for 2 types of books (this is expected) = As volume increases, so does weight
* Also noting that PB generally weigh less than HC books.
* Next, fit the model w/ data from DAAG library 🡪 data set = allbacks,

> model1 <- lm(weight ~ volume + cover, data = allbacks)

> summary(model1)

Call:

lm(formula = weight ~ volume + cover, data = allbacks)

Residuals:

Min 1Q Median 3Q Max

-110.10 -32.32 -16.10 28.93 210.95

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 197.96284 59.19274 3.344 0.005841 \*\*

volume 0.71795 0.06153 11.669 6.6e-08 \*\*\*

coverpb -184.04727 40.49420 -4.545 0.000672 \*\*\*

Residual standard error: 78.2 on 12 degrees of freedom

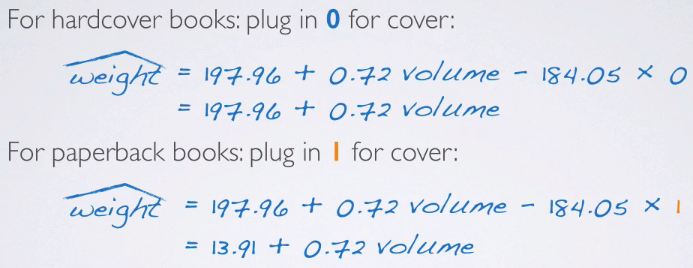
Multiple R-squared: 0.9275, Adjusted R-squared: 0.9154

F-statistic: 76.73 on 2 and 12 DF, p-value: 1.455e-07

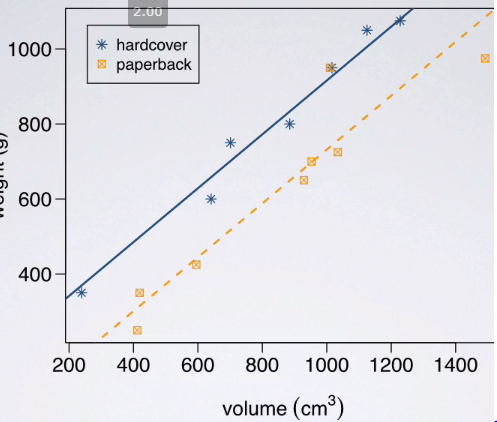
* See estimate for the intercept = 197.96 + estimates for slope of volume + of cover variables.
* Also note only 1 level of **cover** is in the output = the **NON-reference level** 🡪 hardcover books must be the **reference level**.
* Multiple R2 = 92.75% 🡺 92.75% of variability of book weights is explained by volume + cover type
* Pretty high R2 but we’d expect that b/c *what else could the weight of a book depend on*?
* Perhaps paper type + something like that makes up the remaining 7% of unexplained variability
* Using these estimates, we can easily write the linear model.



* Remember PB = non-reference level, meaning HC books = reference level
* **For reference levels, always plug in 0 in our linear model**.
* To simplify this linear model to see what it’d look like only for HC books, plug in a 0 for cover
* For a PB, plug in 1 for cover



* So same slope, different intercepts.



* Multiple linear regression has allowed us to fit these separate parallel lines for the 2 types of books, as opposed to imposing only 1 single line describing all books
* That line would be somewhere in-between these 2 in order to be able minimize the residuals, but it really wouldn't do a good job explaining either the HC or PB books.
* Slope estimate for volume = 0.72 🡺 **All else held constant**, for each 1 cm^3 increase in volume, the model predicts the books to be heavier **on average** by 0.72 grams
* Slope for cover variable = -184.05 + remember since HC = reference level + PB = non-reference level.
* In this case, we can think of this value as when going from HC to PB, there is an expected decrease in weight of 184.05 g
* In context of the data (keeping in mind this is an *observational* study), interpret this as *all else held constant*, the model predicts PB books weigh 184.05 grams lower than HC books, on average.
* The intercept = 197.96 = predicted value of the response variable when x = 0.
* In this case, we have 2 x's, so when volume = 0 + when cover = 0 (i.e. reference level for cover is being considered), weight = 198 g.
* In other words, HC books w/ no volume are expected on average to weigh 198 grams.
* Obviously, this is completely meaningless in context b/c what is a book cannot NOT have a volume
* Intercept is still useful b/c it serves to adjust the height of the regression line.
* Predict weight of a PB book 600 cm^3 in volume.

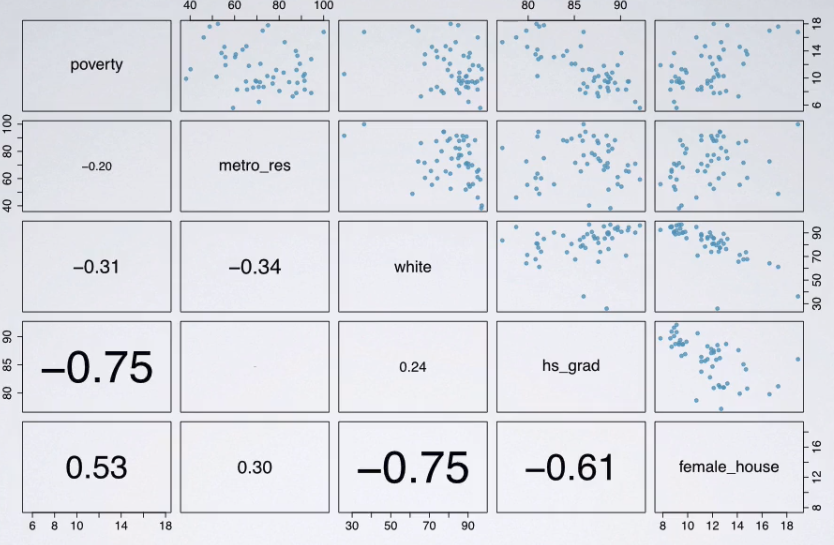
> non\_reference\_model(600)

[1] 444.6856

* This model predicts a PB book 600 cm^3 to weigh 444.68 grams.
* Note this model assumes that HC + PB books have the *same slope* for the relationship between volume + weight.
* This is probably not an unreasonable assumption for books, but this may not be the case.
* If trying to predict calories burned from # of minutes of exercise + a categorical variable sex, the relationship between these may not be the same for males + females
* In that case, it wouldn't really make sense to model this using 2 parallel lines.
* If this assumption of parallel lines/idea of the same slope for the 2 levels of the categorical variable is NOT reasonable, introduce an **interaction variable** in the model.

**Adjusted R2**

* Data from the US states + DC on poverty = % living in poverty, % of residents living in a metropolitan area, % white, % of high school graduates, + % of female head of householders.



* For % living in poverty + % of metropolitan residents, = -.20, + % of metropolitan residents vs. female house heads correlation = 0.30.
* Sizes of correlation coefficients vary by magnitude + highly correlated (negative *or* positive) = larger font sizes + those not highly correlated = smaller font sizes.
* These are **pairwise** plots = very useful for initial EDA, especially, if you have all numerical variables
* We're going to start w/ a simple linear regression for this dataset w/ only 1 predictor.

Residuals:

Min 1Q Median 3Q Max

-5.7537 -1.8252 -0.0375 1.5565 6.3285

Coefficients:

Estimate Std. Error t value Pr(>|t|)

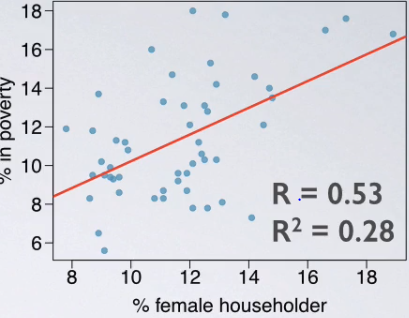
(Intercept) 3.3094 1.8970 1.745 0.0873 .

female\_house 0.6911 0.1599 4.322 7.53e-05 \*\*\*

Residual standard error: 2.664 on 49 degrees of freedom

Multiple R-squared: 0.276, Adjusted R-squared: 0.2613

F-statistic: 18.68 on 1 and 49 DF, p-value: 7.534e-05



* R2 (% of variability in poverty explained by % of female householders) = 28%.
* See w/ a small p value = female householder is a significant predictor of % living in poverty.
* Can take a look at an ANOVA output = allows us to **partition** the variability in our response variable.

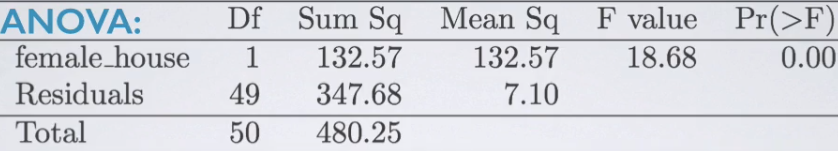
Analysis of Variance Table

Response: poverty

Df Sum Sq Mean Sq F value Pr(>F)

female\_house 1 132.57 132.568 18.683 7.534e-05 \*\*\*

Residuals 49 347.68 7.095



* **Total Sum of Squares =** total variability in the response variable = 480.25.
* this is very similar to the variance of that variable except not scaled by the sample size,
* Can also see how much variability can be attributed to our explanatory variable (% of female head of householders) vs. how much of it is *unexplained* by the model = **sum of square error** = variability left over still in the residuals.
* R2 = ratio of explained variability (regression sum of squares) / to total variability (total sum of squares) = 132.57 / 480.25 = 28%.
* 72% of variability is unexplained (347.68 / 480.25)
* Now that we have our baseline model, we can add another variable to it, % white.

Residuals:

Min 1Q Median 3Q Max

-5.5245 -1.8526 -0.0381 1.3770 6.2689

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.57894 5.78491 -0.446 0.657743

female\_house 0.88689 0.24191 3.666 0.000615 \*\*\*

white 0.04418 0.04101 1.077 0.286755

Residual standard error: 2.659 on 48 degrees of freedom

Multiple R-squared: 0.2931, Adjusted R-squared: 0.2637

F-statistic: 9.953 on 2 and 48 DF, p-value: 0.0002422

Analysis of Variance Table

Response: poverty

Df Sum Sq Mean Sq F value Pr(>F)

female\_house 1 132.57 132.568 18.7447 7.562e-05 \*\*\*

white 1 8.21 8.207 1.1605 0.2868

Residuals 48 339.47 7.072



* Note that the total variability/total sum of squares has not changed b/c this is the **inherent variability** in our response variable, % living in poverty.
* *Regardless of how many variables you're using in a model, total variability should not change.*
* However, what *has* changed = how this variability is partitioned.
* Part of it can be attributed to female householder + a much smaller part can be attributed to white
* If we wanted to calculate R2 based on this output (keeping in mind R2 = % of variability in response variable explained by the model) is comprised of *TWO* explanatory variables.
* We have (132.57 + 8.21) / 480.25 = roughly 29.3%.
* Adding another variable to the model now explains 1 more % of the variability in response variable
* R2 goes to up each time you add a new predictor to your model.
* We need a more *honest* measure of whether an added variable is *actually useful =* **Adjusted R2**
* This measure applies a penalty to R2 for the # of predictors included in the model, + the magnitude of this penalty depends on how k (# of predictors) compared to n (sample size).
* Larger sample size = more predictors a model can handle, + therefore less penalties for additional predictors being added to the model.



* While R2 always increases w/ each additional variable to the model regardless of whether it’s useful or not, adjusted R2 is *only* going to increase if the added variable is *actually of value*
* In other words, if the additional % of variability in the response variable explained by the new variable can **offset the penalty** for the additional # of predictors in the model.
* Calculate adjusted R2.

> sse <- 339.74

> sst <- sum(132.57,8.21,339.47)

> n <- 51

> k <- 2

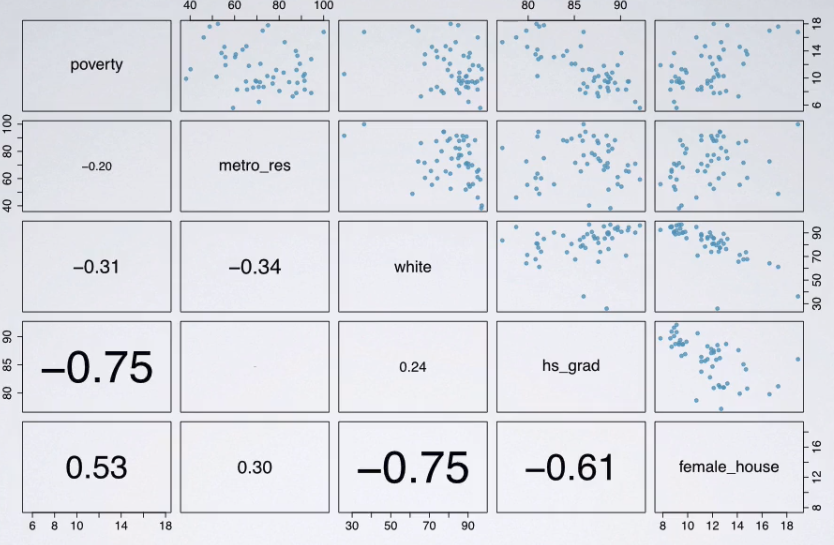
> (adjR2 <- 1 - ((sse/sst)\*((n-1)/(n-k-1))))

[1] 0.2631008

* Adjusted R2 has penalty for additional predictor, so its only 26% compared to 29% for R2
* 1st model w/ female house holder = only predictor, R2 = 28%.
* 2nd model w/ additional variable white, R2 increased to 29% while Adj. R2 w/ penalty stayed at 26%.
* When any variable is added to the model, R2 increases.
* However, if the added variable doesn't really provide any new info or is completely unrelated, adjusted R2 does not increase.
* Properties of Adjusted R2
* k (# of predictors) can never be negative, therefore, adjusted R2 is *always* going to be < R2.
* Adjusted R2 applies a penalty for the # of predictors included in the model
* Choose models w/ higher adjusted R2 over others.
* The decision criteria is based on Adj. R2 as opposed to R2 b/c R2 is always higher for models w/ a higher # of predictors, but those may not always be the favorable ones.

**Collinearity and Parsimony**

* 2 predictor variables are **collinear** when they are correlated w/ each other.
* Predictors = independent variables + *should be independent of each other* + should NOT be collinear w/ each other.
* **Multicollinearity** (inclusion of collinear predictors) complicates model estimation = estimates coming out of the model may no longer be reliable.



* In a model predicting poverty from female householder + white, we saw very little gains from adding the 2nd explanatory variable (R2 went up by just a tiny bit) + Adj. R2 did not go up at all.
* Look at how white + female householder are related to each other 🡪 see little scatter in their relationship scatter plot + correlation coefficient between them is quite high, indicating a strong negative relationship between the 2 variables.
* This means white is highly correlated w/ female householder, + therefore they're not independent of each other.
* If that is the case, we wouldn't want to add white to a model that already has female householder, b/c it's going to bring nothing new to the table.
* Any info gleaned from inclusion of white is probably already being captured by the female householder b/c these 2 variables are highly associated w/ each other.
* Using both of these variables in the model results in **multicollinearity** which might also result in unreliable estimates of the coefficients from the model
* **Parsimony** = avoid adding predictors that are associated w/ each other b/c often times the addition of such variable brings nothing new to the table.
* Prefer the simplest best model = the **parsimonious model** = model w/ the highest **predictive power** but lowest # of variables.
* Idea comes from Occam's razor = Among competing hypotheses, the 1 w/ the fewest assumptions should be selected.
* In other words, among models that’re equally good, select the one w/ fewer variables.
* The addition of collinear variables can result in biased estimates of regression parameters, so not only do we prefer simple parsimonious models, but we also want to be very careful about adding a bunch of explanatory variables to a model.
* If those are co-linear w/ each other, the model estimates may no longer be reliable.
* While it is impossible to avoid collinearity from arising in *observational* data, experiments are usually designed to control for correlated predictors.

***5.3.2 Inference for Multiple Regression + Model Selection***

**Inference for MLR**

* Using inferential techniques, we can determine *which* variables in a model are significant predictors of a response variable.
* Data comes from the National Longitudinal Survey of Youth = cognitive test scores of 3 + 4-year-old children + characteristics of their mothers w/ data on kid's score, whether mom went to high school or not, IQ score of the mother, whether the mom worked during the 1st 3 years of the kid's life, + age of the mother at birth of child.



* Fit the full model (include all explanatory variables given to us in the data set)

Residuals:

Min 1Q Median 3Q Max

-54.045 -12.918 1.992 11.563 49.267

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 19.59241 9.21906 2.125 0.0341 \*

mom\_hsyes 5.09482 2.31450 2.201 0.0282 \*

mom\_iq 0.56147 0.06064 9.259 <2e-16 \*\*\*

mom\_workyes 2.53718 2.35067 1.079 0.2810

mom\_age 0.21802 0.33074 0.659 0.5101

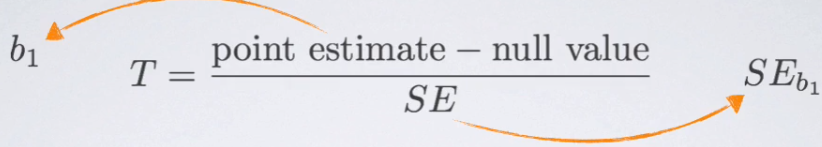
---

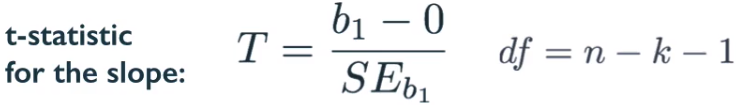
Residual standard error: 18.14 on 429 degrees of freedom

Multiple R-squared: 0.2171, Adjusted R-squared: 0.2098

F-statistic: 29.74 on 4 and 429 DF, p-value: < 2.2e-16

* 1st, do inference for the model as a whole.
* Here, our null H0: each variable’s slope = 0 🡪 none of the explanatory variables is a significant predictor of the response variable
* The alternative H1: At least 1 of the slopes is different than 0.
* Test statistic used = F-statistic (from bottom of the regression output) = 4 + 429 dF 🡪 4 = # of predictors, + 429 = n - k - 1.
* We had 434 observations – 4 predictors - 1, which gives **residual dF**
* We’re also given the p-value, so we really don’t need to do any calculations by hand here.
* Need to focus on the *interpretation* of what this means.
* p-value < 0.05 = we say the model, *as a whole*, is significant = reject the null
* The alternative suggests there is at least *something* interesting to look for here.
* **F test yielding a significant result doesn't mean the model fits the data well** **🡪 just means at least 1 of the betas is non-zero**.
* **The F test, on the other hand, NOT yielding a significant result DOESN'T mean individual variables included in the model aren’t good predictors of y 🡪 just means the combination of these variables doesn't yield a good model**
* Now that we know there’s something worthwhile to look for in this model (b/c we found out at least 1 beta is different than 0), we can do individual tests on the slopes.
* Is whether or not the mother went to HS a significant predictor of cognitive test scores of children, given all other variables in the model?
* H0= Beta associated w/ mother HS status = 0 when all other variables are included in the model
* H1: It's different than 0 when all other variables are included in the model.
* Regression output 🡪 look at row for mother's HS status 🡪 look at p-value for it.
* Since it’s a very, very small p-value, can determine if whether a mom went high school is a significant predictor of the cognitive test scores of children, given all other variables in the model.
* Even though we don't need to do any calculations by hand, it's a good idea to try to understand how calculations included in a regression output are actually done so you understand what they mean.
* Mechanics of testing for slope w/in the framework of a multiple linear regression.
* As usual w/ a regression, use a t-statistic in inference = point estimate - the null value, divided by SE





* Our point estimate = our slope estimate, + the SE = SE of this estimate from the regression output.
* So, **the t-statistic for slope = (b1 – 0) / SE of b1**.
* This is different from the single predictor regression case b/c now we calculate the dF = n – k - 1, where k = # of predictors included in the model + n = sample size
* This is not a new measure at all.
* For a regression w/ a single predictor, the dF = n – 2, b/c in a single predictor regression, # of predictors = 1.
* So, to calculate dF as n - k - 1 for this case, we have n – 1 – 1 = n - 2
* Remember, the additional minus 1 is b/c along w/ every single predictor for which we calculate a slope estimate, we also calculate an intercept, + that's where we're losing that 1 additional dF
* So, while we've introduced these 2 formulas slightly differently, they mean the same thing
* *Start w/ sample size (total dF you have to play w/) + then lose # of dF for however many predictors you have, + then lose 1 more for the intercept.*
* Verify the T-score + the p-value for the slope of mom\_hs

mom\_hsyes 5.09482 2.31450 2.201 0.0282 \*

> pt.estimate <- 5.09482 # slope of var

> null <- 0

> se <- 2.31450 # SE of var

> n <- nrow(cognitive)

> k <- ncol(cognitive) - 1

> dF <- n - k - 1

> (t <- (pt.estimate - null)/se)

[1] 2.201262

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

[1] 0.02824936

* The t-score + p-value are spot on w/ the table = reject the null in favor of the alternative + determine that mom's high school status is indeed a significant predictor for kid's cognitive score.
* Construction of a CI’s always follows the same structure, regardless of the estimate for which you're constructing it 🡪 a point estimate +/- a margin of error
* In this case, our point estimate = slope estimate + margin of error = t-statistic \* SE of the slope.
* Calculate a 95% CI for the slope of mom\_work.



> alpha = .95

> (t.crit <- abs(qt(p = (1 - alpha)/2, df = dF)))

[1] 1.965509

> pt.estimate.work <- 2.53718

> se.work <- 2.35067

> (mOe <- t.crit\*se.work)

[1] 4.620263

> (lower <- pt.estimate.work - mOe)

[1] -2.083083

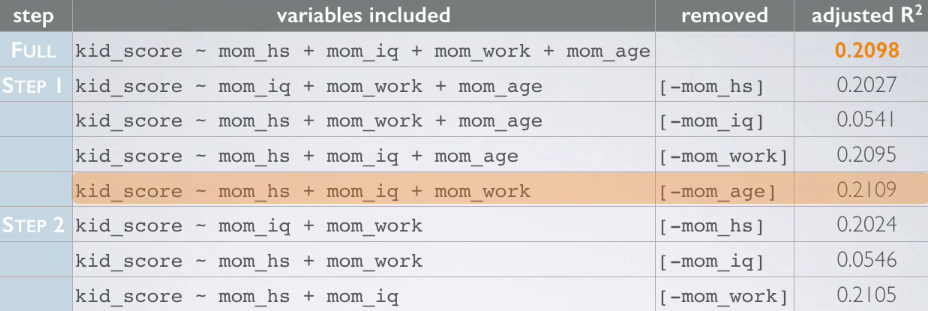
> (upper <- pt.estimate.work + mOe)

[1] 7.157443

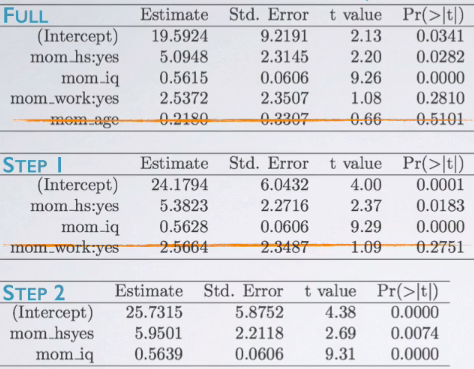
* This CI = the interpretation of the slope for this variable, except we're now also adding a statement to the beginning of that about how confident we are of that estimate.
* So, **we are 95% confident that, all else being equal, the model predicts children whose moms worked during the 1st 3 years of their lives scored 2.09 points lower to 7.17 points higher than those whose moms did not work.**

**Model Selection**

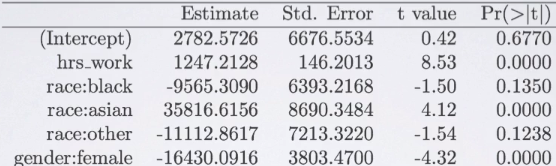
* 1 **stepwise model selection** method = **backwards elimination** = start w/ full model (has all possible co-variants/predictors included) + drop variables 1 at a time until a **parsimonious model** is reached.
* Could also do **forward selection** = start w/ an empty model + add variables 1 at a time until a parsimonious model is reached.
* There are many criteria for model selection 🡪 we focus on p values + Adj. R2s.
* Other model selection criteria = **Aikake Information Criterion (AIC), Bayesian Information Criterion (BIC), Deviance Information Criterion (DIC), Bayes factor, Mallow's Cp, etc.**
* Backwards elimination using the Adj. R2 🡺 start w/ the full model, drop 1 variable at a time, record Adj. R2 of each smaller model, pick the model w/ **highest increase** in Adj. R2, we repeat until no of models yield an increase in Adj. R2.



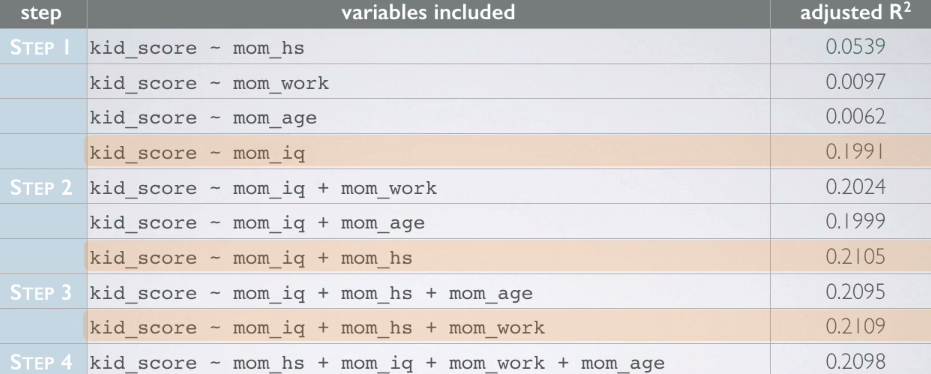
* Removing mom's age at birth of the child gives highest increase for step 1
* None of the options in Step 2 actually yield an increased Adj. R2.
* Therefore, final model predicts kids’ cognitive test score from Mom's HS status, IQ + work status
* Backwards elimination using the p-value 🡪 start w/ full model, drop variable w/ highest p-value, refit the smaller model, repeat until all variables left in the model are significant.



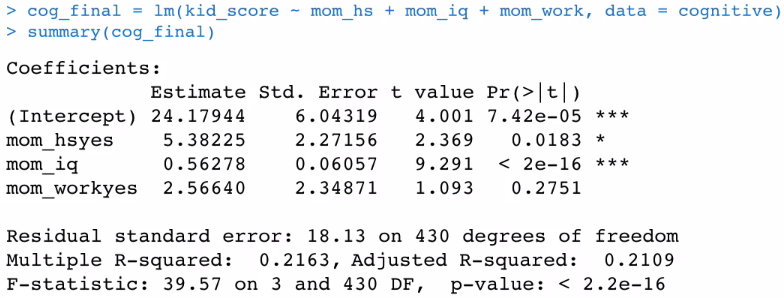
* After step 1, we got in a slightly different model using the p-value approach vs. Adj. R2 approach
* This is not unexpected = would expect very similar models but not necessarily *exactly* the same model b/c decision criteria is different.
* Ex: Data from the American Community Survey to predict income from hours worked per week, race, + gender.



* Need to consider **race** *all at once* b/c we can't simply drop 1 level of an existing variable + b/*c at least 1 level of this variable has a significant p-value*
* We’d actually keep this variable in the model as well, therefore, we don't drop any variables here.
* This is an important point 🡺 **If you have a categorical variable w/ multiple levels, you cannot drop some levels of that variable + keep others.**
* Either need to decide that to *keep* the entire variable *as a whole* or *drop* it *as a whole*
* In this case, b/c at least 1 level has a small p value/has some significance there, we’d keep the entire variable.
* If all levels of a variable had high p-values such that there wouldn't be any significant levels we’d drop the entire variable as a whole.
* **Use p-value approach if interested in *finding out which predictors are statistically significant.***
* **Use the Adj. R2 method if interested in *more reliable predictions* from the model**
* The p-value method depends on the somewhat arbitrary 5% or whatever other % you use for your significance level cutoff + if you use a different significance level, you're going to end up w/ a different model.
* *It's used more commonly though, since it requires fitting fewer models.*
* Remember, at each stage of the Adj. R2 method, we dropped 1 variable at a time + refit a bunch of models to determine which one to go w/
* VS. the p-value approach, simply drop the variable w/ the highest p-value + proceed = more common b/c easier to implement.
* However, b/c it relies on this arbitrary significance level cut-off, might be more favorable to use the Adj. R2 method for model selection.
* **Forward selection (Adj. R2) 🡪** start w/ single-predictor regressions of response vs. each explanatory variable, pick model w/ highest Adj. R2, add remaining variables 1 at a time to the existing model, + pick model w/ highest Adj. R2, repeat until addition of any other remaining variables does not result in a higher Adj. R2.



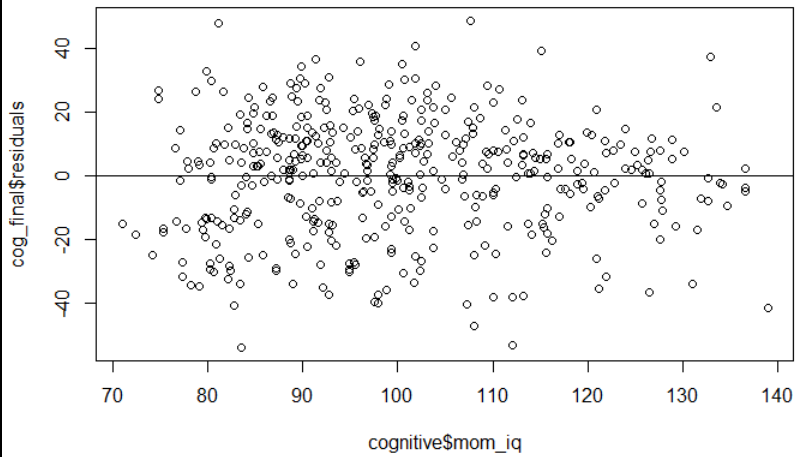
* See when we try the full model, Adj. R2 does not go up, therefore stick w/ the model in step 3
* Note we arrived at the same model w/ this Adj. R2 criteria whether we went backwards or forwards
* To do **forward selection using p-values**, start w/ single-predictor regressions of response vs. each explanatory variable, pick variable w/ lowest significant p-value, add remaining variables 1 at a time to existing model, pick variable w/ lowest significant p-value again, repeat until any remaining variable does not have a significant p-value.
* We talked about *algorithmic ways of doing model selection*, however, sometimes variables can be included in/eliminated from the model based on expert opinion as well.
* Ex: if setting a certain variable, you might choose to leave it in the model regardless of whether its significant or whether it would yield a higher Adj. R2 or not.
* Finally, fit our final model 🡪 Remember we selected variables = mom's high school status, IQ, + work status



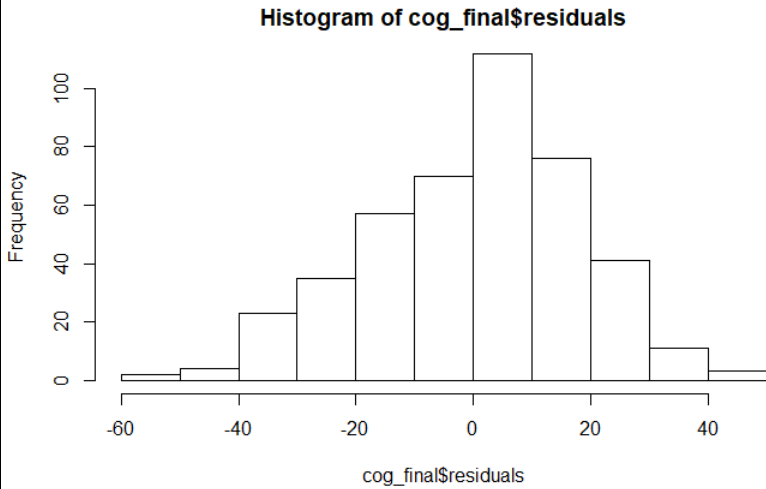
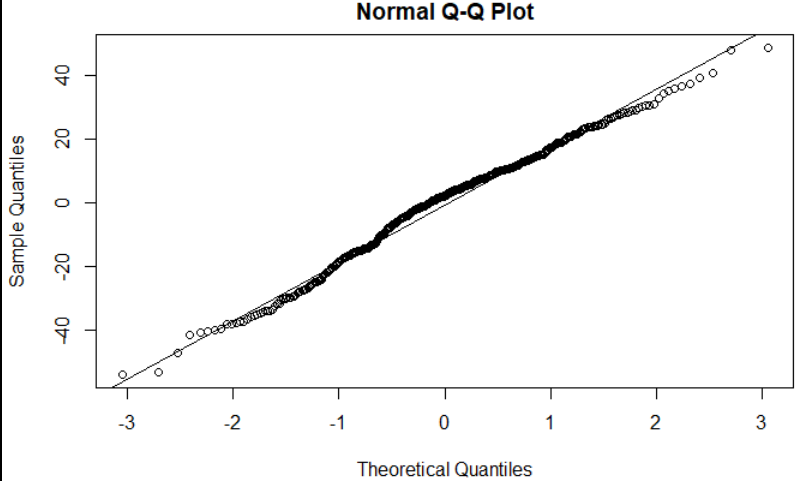
* See that mom's high school status + IQ are statistically significant at the 5% level but work status is not
* But, remember, **we selected this model using the Adj. R2 method,** which tells us including work status actually gives the model higher predictive power even though the variable may not be statistically significant.
* If we had used the p-value approach to do model selection, we would not end up w/ any variables that are not statistically significant in our model.

**Diagnostics for MLR**

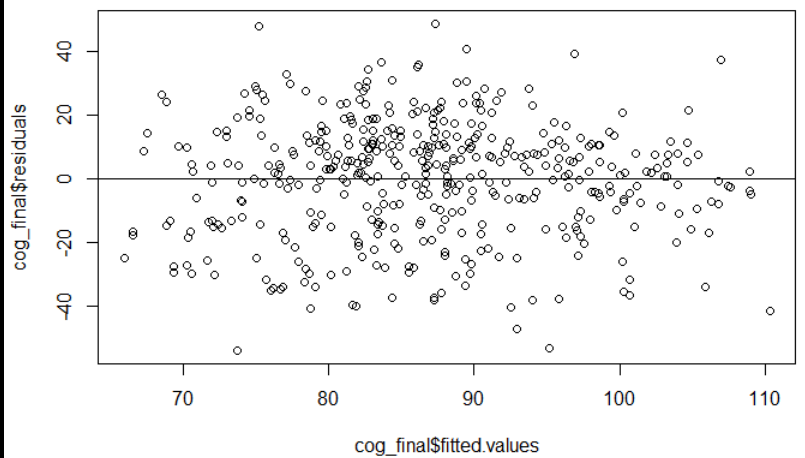
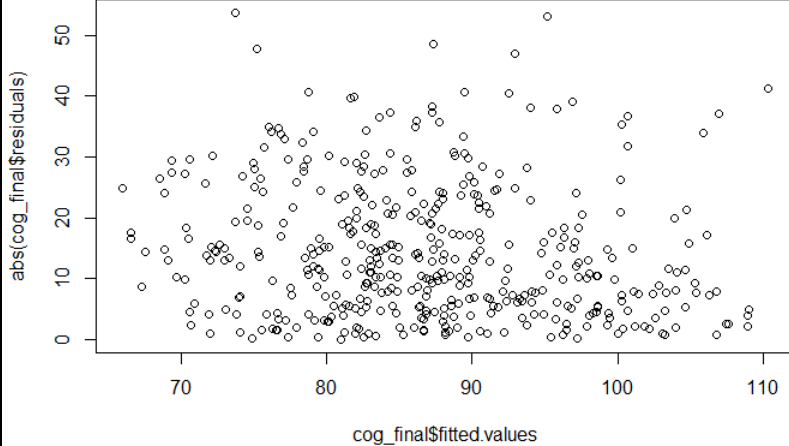
* There are conditions required for the multiple linear regression model to be valid.
* 1) Linear relationships between our numerical explanatory variables + our response variable
* doesn't make sense to ask for linear relationship between a categorical + numerical variable, so each numerical explanatory variable needs to be linearly related to the response variable.
* Check this condition using residuals plots (residuals vs. explanatory variable), looking for a random scatter around y = 0
* Using the residuals plot instead of a scatter plot of the response vs. the explanatory b/c the residuals plot allows for considering the *other variables* also in the model, not *just* the bivariate relationship between a given x + our y.
* Final cognitive model = kid score by mom's HS status, IQ, + work status as explanatory variables
* The only numerical variable = mom's IQ = the variable to focus on for the linearity condition.



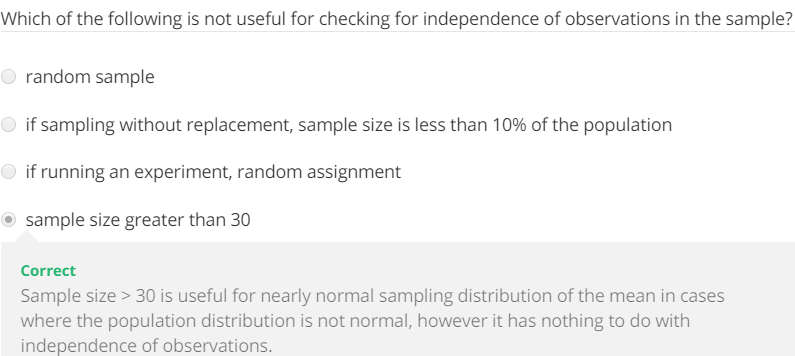
* Want our residuals to be randomly scattered around 0 + it seems we meet this condition
* 2) Residuals need to be nearly normally distributed
* Remember some residuals will be positive + some negative.
* A residuals plot w/ random scatter of residuals around 0 translates to a nearly normal distribution of residuals centered at 0, checked w/ a histogram or normal probability/QQ plot

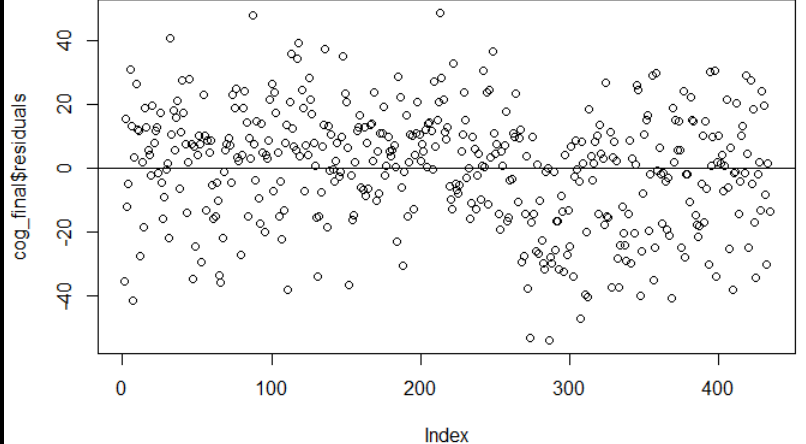
* In histogram, we see a little bit of a skew in the residuals, but not too bad.
* In normal probability/QQ plot, except for at the tail areas, we're not seeing huge deviations from the mean.
* This condition seems to be fairly satisfied.
* 3) Constant variability of residuals
* Want residuals to be equally variable for low + high values of the predicted response variable.
* Check the plot of residuals vs. predicted values (e vs. y\_hat)
* We're using residuals vs. *predicted* instead of residuals versus x b/c it allows for considering the entire model w/ all explanatory variables at once.
* Want residuals to be randomly scattered in a band w/ a constant width around 0.
* Looking to see a fan shape.
* Also worthwhile to view absolute value of residuals vs. predicted values to easily ID any unusual observations

* Don't see a fan shape = appears the variability of the residual stays constant as the value of the fitted/predicted values change, so, the constant variability condition appears to be met.
* The absolute value of residuals plot can be thought of simply the 1st plot folded in half.
* So, if we were to see a fan shape in the 1st plot, we’d see a triangle in the absolute value of residuals versus fitted plot.
* Doesn't exactly seem to be the case, so it seems like this condition is met as well.
* 4) Independence of residuals (speaks to independence of observations in a sample)
* Independent residuals basically means independent observations.
* W/ any time series structure, or if suspecting any time series structure in our data set, we can check for independent residuals using the residuals vs. the order of data collection plot.
* If, on the other hand, that is not a consideration, we don't really have another diagnostic approach we can use.
* Instead, go back to first principles = think about how the data are sampled.



* R plots residuals in the order that they appear in our data set



* Order of data collection = x-axis 🡺 does not show any patterns.
* If there was some non-independent structure we’d see residuals increasing or decreasing, but we don't see any such pattern, so it appears that any sort of time series structure is not a consideration for this dataset.