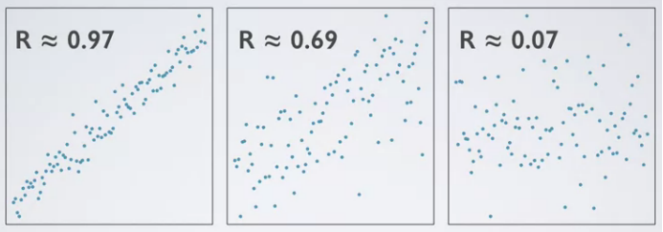
Week1

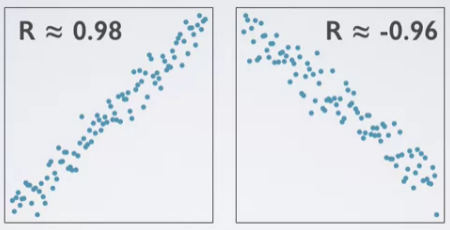
[Relationship between two numerical variables]

1. Correlation

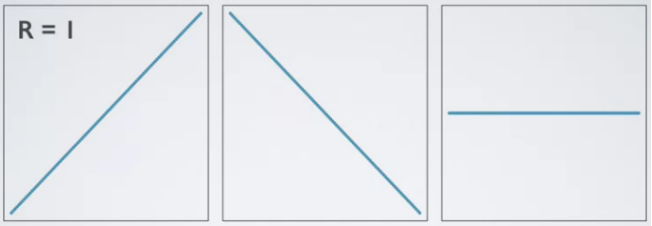
- dealing with the correlation between two numerical variables
- describes the strength of the linear association between two variables
- denoted as R
- properties
    1. the magnitude (absolute value) of the correlation coefficient measures the strength of the linear association between two numerical variables



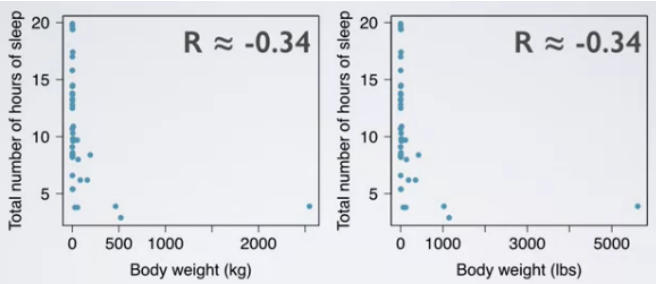
    2. the sign of the correlation coefficient indicates the direction of association



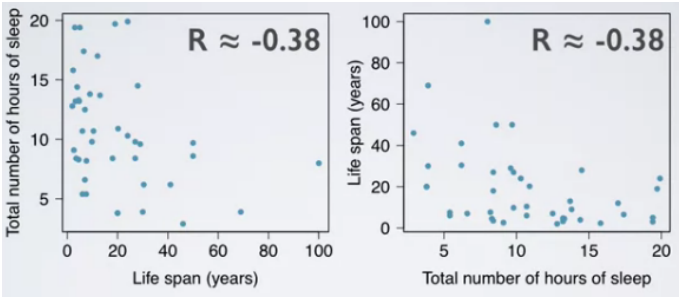
    3. the correlation coefficient is always between -1 and 1 and R = 0 indicates no linear relationship



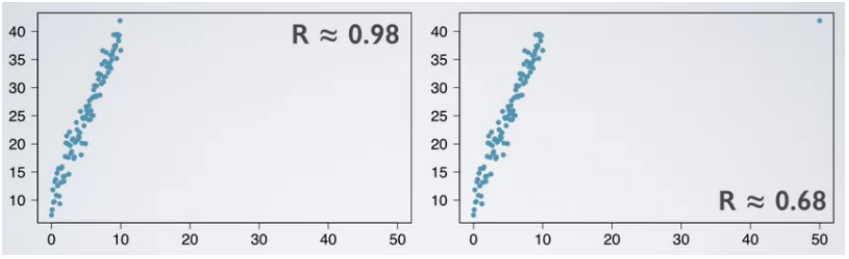
    4. the correlation coefficient is unitless, and is not affected by changes in the center or scale of either variable (such as unit conversions)



    5. the correlation of X with Y is the same as of Y with X



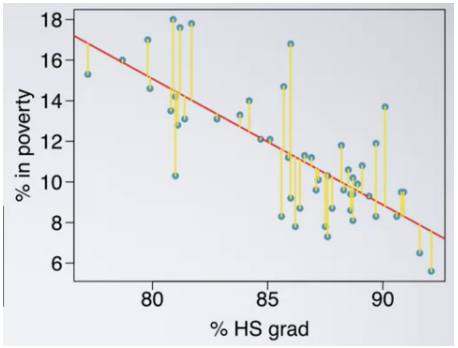
6. the correlation coefficient is sensitive to outliers

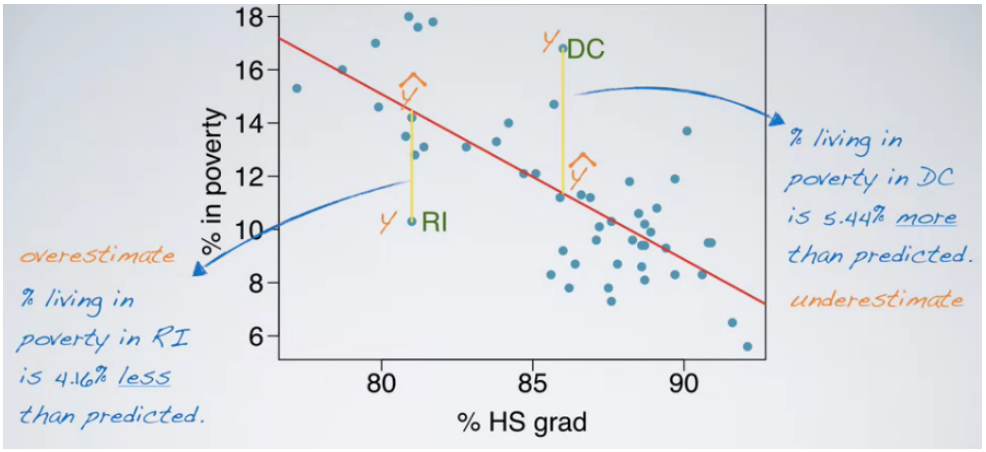


2. Residuals

- leftovers from the model fit
- data = fit + residual
- difference between the observed and predicted y

residual: $e_i = y_i - \hat{y}_i$





2021 October - written by Mihyeon Jeon

3. Least Squares Line

- a measure for the best line

Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$$|e_1| + |e_2| + \cdots + |e_n|$$

✔ **Option 2:** Minimize the sum of squared residuals – least squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Why least squares?
  - most commonly used
  - easier to compute by hand and using software
  - In my applications, a residual twice as large as another is more than twice as bad

- least squares line

$$\hat{y} = \beta_0 + \beta_1 x$$

predicted response, intercept, slope, explanatory

notation

	parameter	point estimate
intercept	β_0	b_0
slope	β_1	b_1

- estimating the regression parameters : slope

slope: $b_1 = \frac{s_y}{s_x} R$

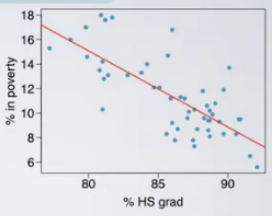
s_x : SD of x
s_y : SD of y
$R = cor(x, y)$

practice)

The standard deviation of % living poverty is 3.1%, and the standard deviation of % HS graduates is 3.73%. Given that the correlation between these variable is -0.75, what is the slope of the regression line for predicting % living poverty from % HS graduates?

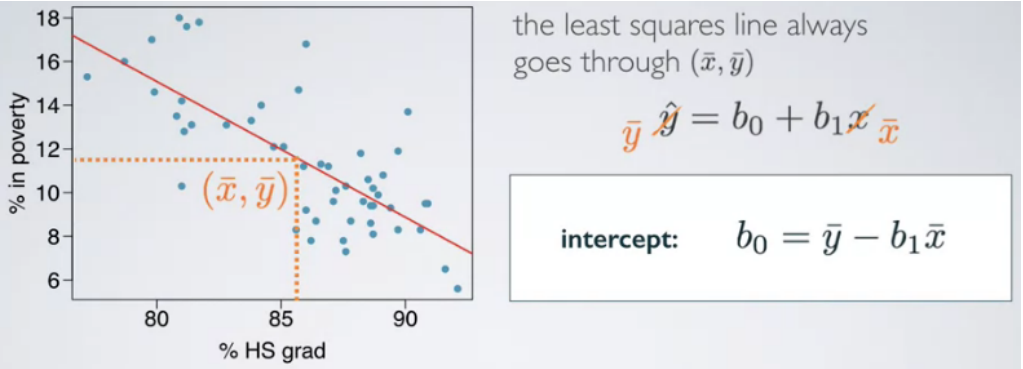
$s_y = 3.1\%$
$s_x = 3.73\%$
$R = -0.75$

$b_1 = \frac{s_y}{s_x} R = \frac{3.1}{3.73} \times -0.75 \approx -0.62$

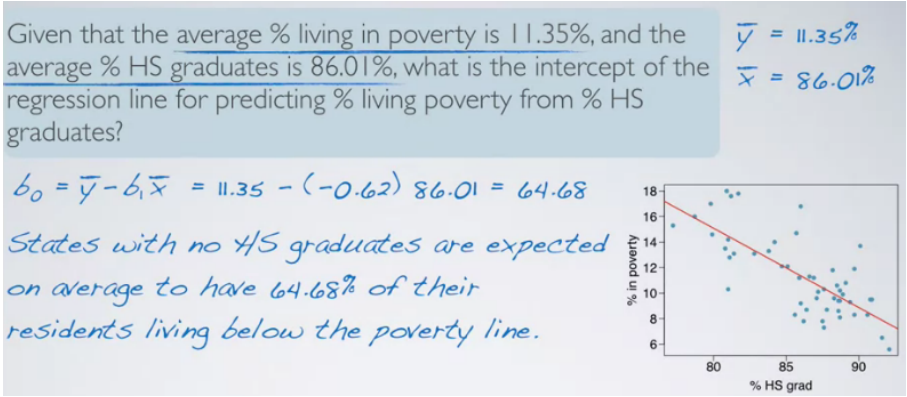


- Mathematically speaking, standard deviation is always positive so the sign of the slope is always going to be the same as the sign of the correlation coefficient.
- For each % point increase in HS graduate rate, we would expect that % living in poverty to be lower on average by 0.62% points.

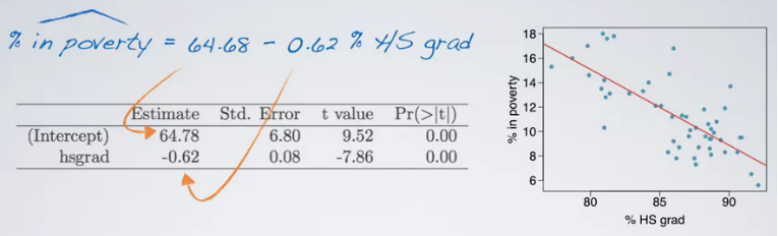
- estimating the regression parameters : intercept



practice)



- States with no HS graduates are expected on average to have 64.68% of their residents living below the poverty line.



- intercept : When x = 0, y is expected to equal the intercept.
  - may be meaningless in context of the data, and only serve to adjust the height of the line.
- slope : For each unit increase in x, y is expected to be higher / lower on average by the slope.

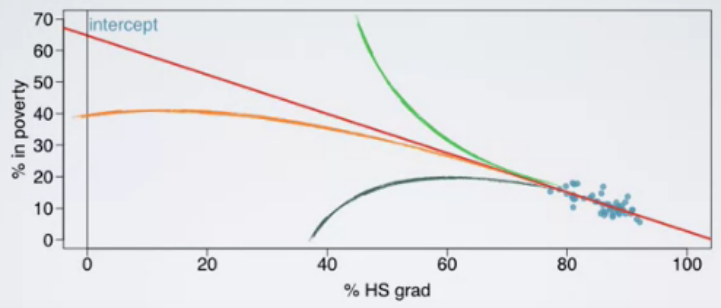
[Linear regression with one predictor]

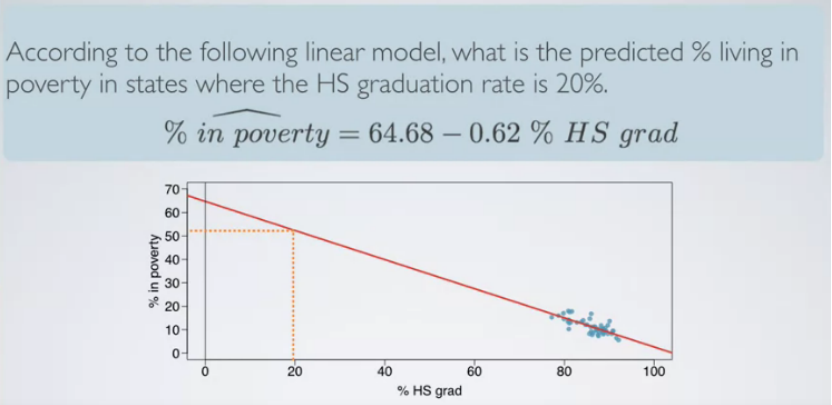
1. Prediction and Extrapolation

- In mathematics, **extrapolation** is a type of estimation, beyond the original observation range, of the value of a variable on the basis of its relationship with another variable.

- prediction
    - using the linear model to predict the value of the response variable for a given value of the explanatory variable is called prediction
    - plug in the value of x in the linear model equation

- extrapolation
    - applying a model estimate to values outside of the realm of the original data is called extrapolation.
    - sometimes the intercept might be an extrapolation



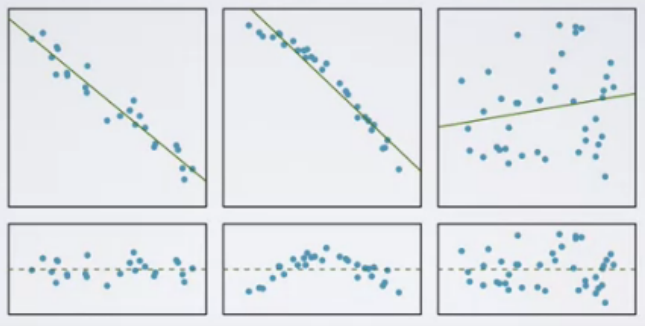


- This is a very simple problem but is it wise to do?
- We always want to look back at our data, this could be in a scatter plot format, or at least looking at the summary statistics.
- We should see whether 20% is within the realm of the data that we observed or not.
- In this case, it is not so we do not want to be doing prediction as it would yield an unreliable estimate.

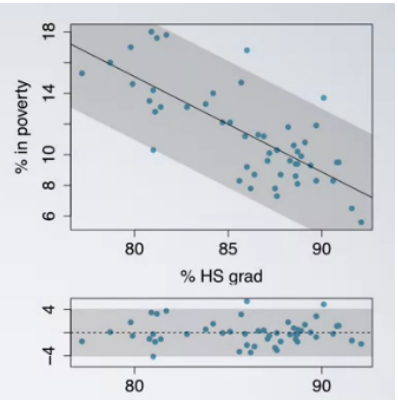
2. Conditions for Linear Regression

1) Linearity
    - relationship between the explanatory and the response variable should be linear.
    - methods for fitting a model to non-linear relationships exist
    - check using a scatterplot of the data, or a residuals plot

2021 October - written by Mihyeon Jeon



2) nearly normal residuals
    - residuals should be nearly normally distributed, centered at 0
    - may not be satisfied if there are unusual observations that don't follow the trend of the rest of the data
3) constant variability
    - variability of points around the least squares line should be roughly constant
    - implies that the variability of residuals around the 0 line should be roughly constant as well
    - also called homoscedasticity
    - check using a residuals plot



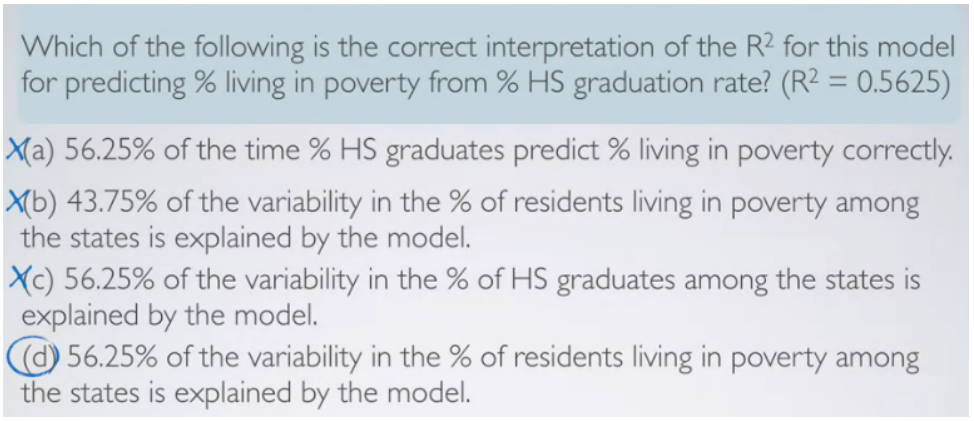
- They seem to be captured around this constantly variable grey band around the regression line. In the residual plot, we can also confirm that the variability of the residuals, that is how far they are from zero, do not vary by the value of the explanatory variable.

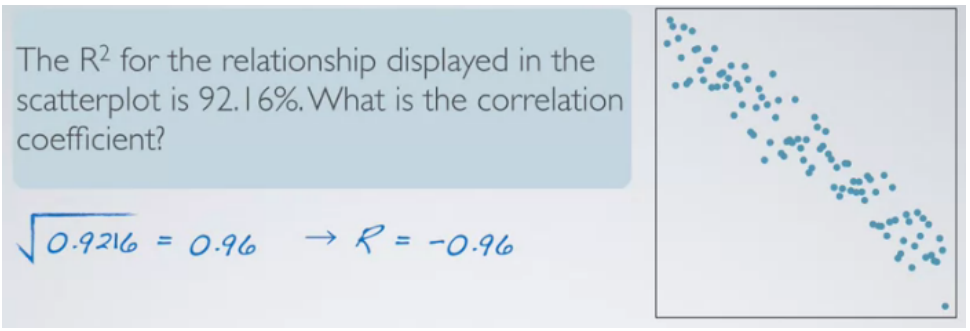
- Diagnostics for simple linear regression

2021 October - written by Mihyeon Jeon

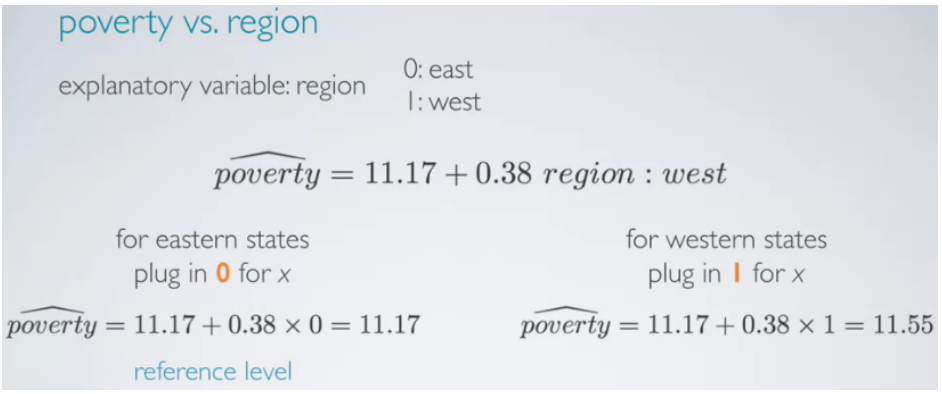
3. R squared

- R squared
  - strength of the fit of a linear model is most commonly evaluated using R squared.
  - calculated as the square of the correlation coefficient
  - tells us what percent of variability in the response variable is explained by the model
  - the remainder of the variability is explained by variables not included in the model
  - always between 0 and 1





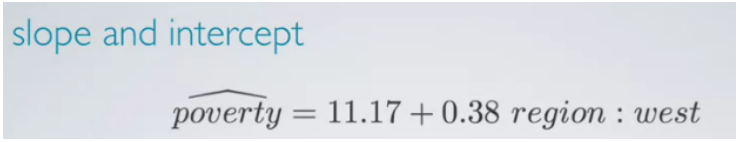
4. Regression with Categorical Explanatory Variables



- In regression models, with explanatory categorical variables, we always code one of the levels of that categorical variable to be what we call the reference level. This is the level that we plug in zero for.

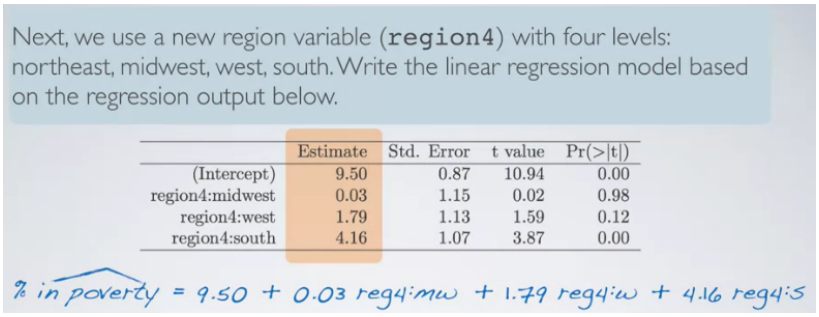
- What do the slope and the intercept mean in this context?

2021 October - written by Mihyeon Jeon

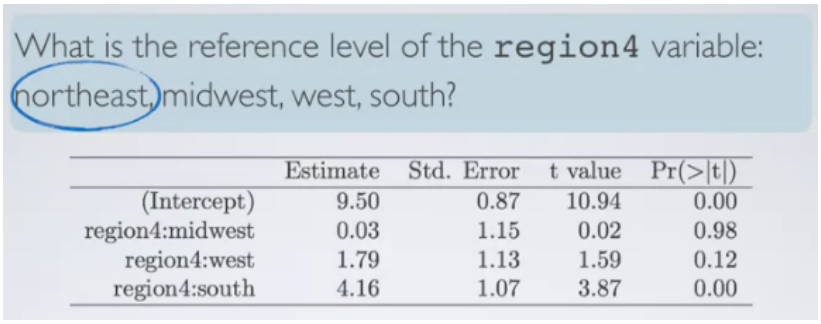


- Intercept
  - It basically tells us that the model predicts an 11.17% average poverty percentage, in eastern states.
  - This is the value we get if we plug in 0 for the explanatory variable
  - labelling some of the levels success and some of the levels failures
- Slope
  - The model predicts that the average poverty percentage in western states is 0.38% higher than in the eastern states.

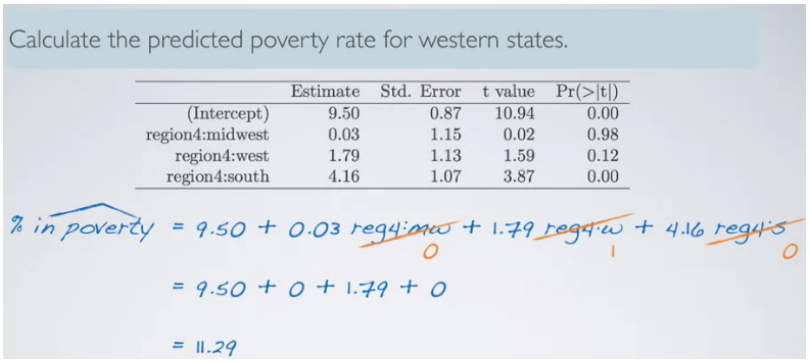
practice)



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00



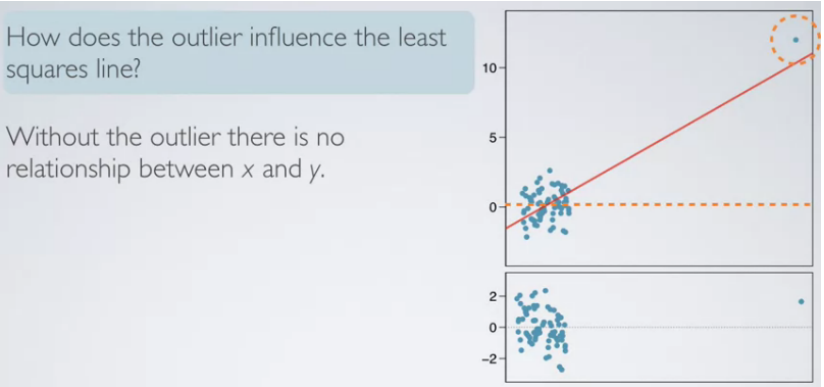
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00

2021 October - written by Mihyeon Jeon

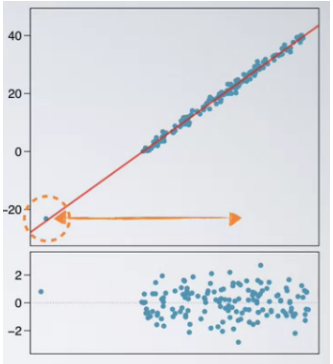
Week2

[Outliers & Inference for regression]

1. Outliers in Regression



- types of outliers
  - outliers are points that fall away from the cloud of points
  - **leverage points** : outliers that fall horizontally away from the center of the cloud but <u>don't influence the slope</u> of the regression line
  - **influential points** : outliers that <u>actually influence the slope</u> of the regression line
    - usually high leverage points
    - to determine if a point is influential, visualize the regression line with and without the point, and ask : Does the slope of the line change considerably?
    -
  - What type of outlier is this?

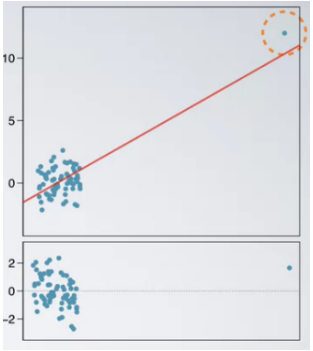


- **Does this point fall away from the rest of the data in the horizontal direction?**

: YES. This makes it a potential leverage point.

- **Is it also influential?**

: It appears that the line would stay in exactly the same place, so the outlier point is actually on the trajectory of the regression line. So it does not influence it. This makes it a leverage point.

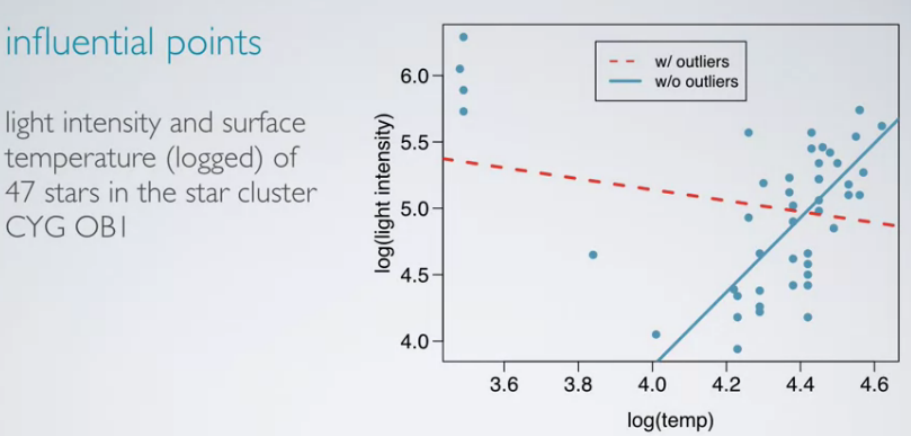


- **Does this point fall away from the rest of the data in the horizontal direction?**

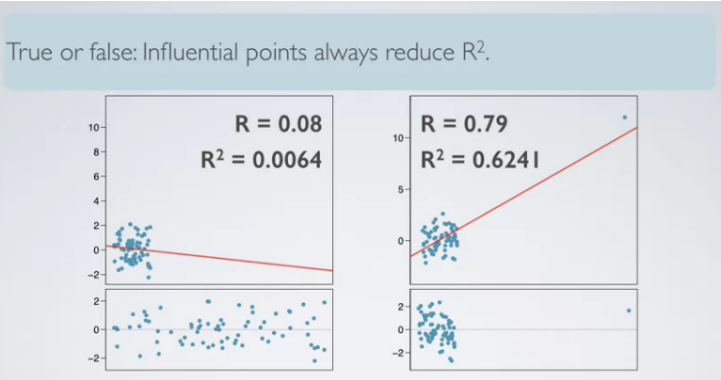
: YES. This makes it a potential leverage point.

- **Is it also influential?**

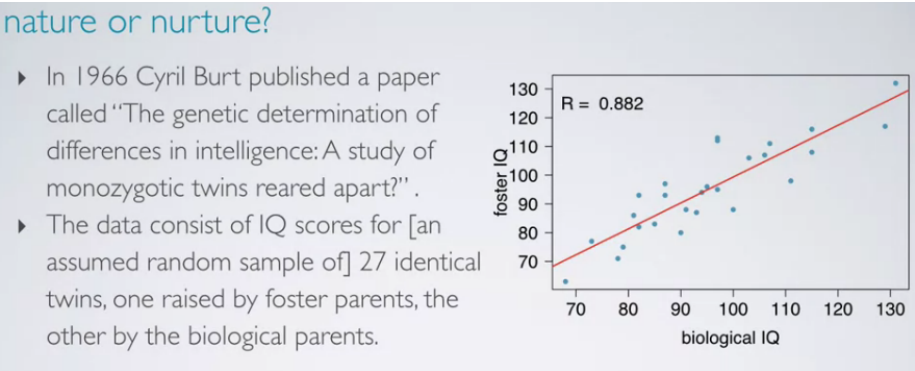
: If we were to remove this point, the line would look considerably different. So we would identify this as an influential point.

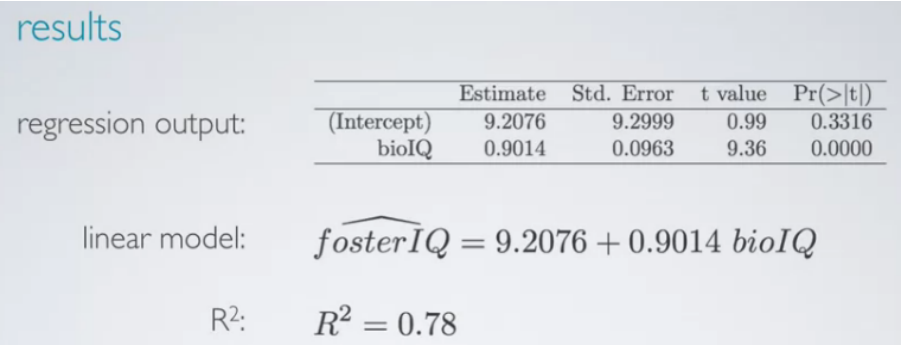


- Obviously, the red-dashed line is not a good fit for this data.
- We might want to split the data into two, those stars with lower temperature and those stars with higher temperature, and model the two groups separately.



2. Inference for Linear Regression





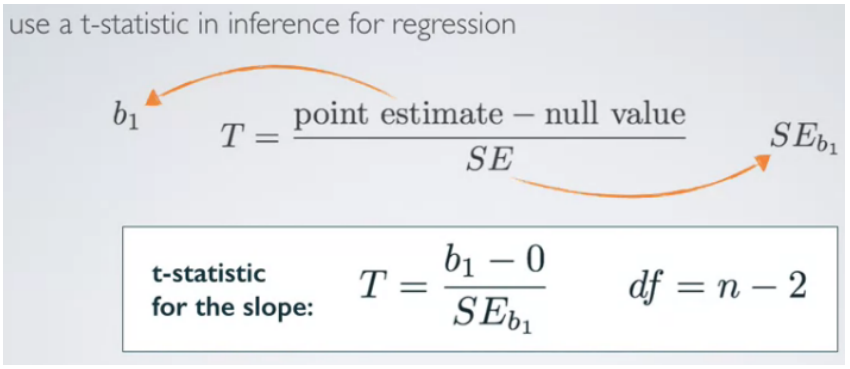
- Testing for the slope - hypothesis

  - Is the explanatory variable a significant predictor of the response variable?
  - H0 (nothing going on) : The explanatory variable is not a significant predictor of the response variable, i.e. no relationship -> slope is 0.

    $H_0: \beta_1 = 0$
  - HA (something going on) : The explanatory variable is a significant predictor of the response variable, i.e. relationship -> slope is different than 0.

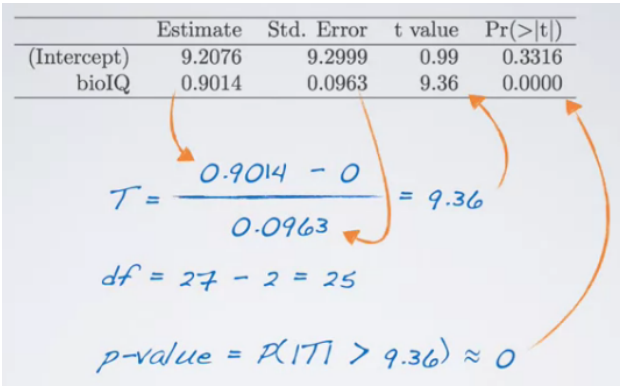
    $H_A: \beta_1 \neq 0$

- Testing for the slope - mechanics

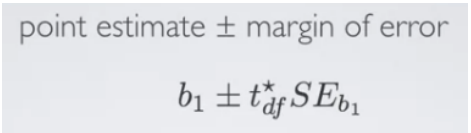


- df = n - 2
  - Lose 1 df for each parameter estimated
  - In linear regression, we estimate 2 parameters : the slope and the intercept

- calculating the test statistic

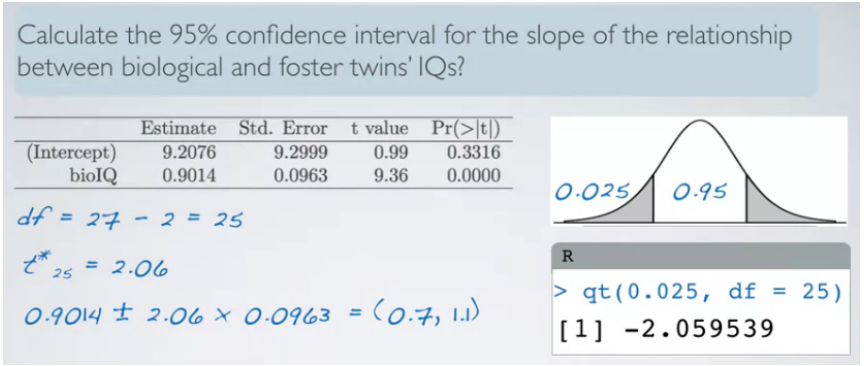


- confidence interval for the slope

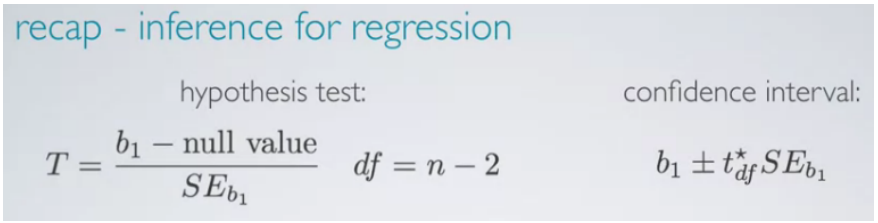


practice)

2021 October - written by Mihyeon Jeon



- **Interpret the 95% confidence interval for the slope of the relationship between biological and foster twins' IQs : (0.7, 1.1)**
  : 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.



- Null value is often 0, since we usually check for any relationship between explanatory and the response variables.
- Regression output gives b1, SEb1, and two-tailed p-value for the t-test for the slope where the null value is 0.
- Inference on the intercept is rarely done as it is not that informative.

- CAUTION!
  - Always be aware of the type of data you're working with : random sample, non-random sample, or population.
  - Statistical inference, and the resulting p-values, are meaningless when you already have population data.
  - If you have a sample that is non-random(biased), the results will be unreliable.
  - The ultimate goal is **to have independent observations** - and you know how to check for those by now.

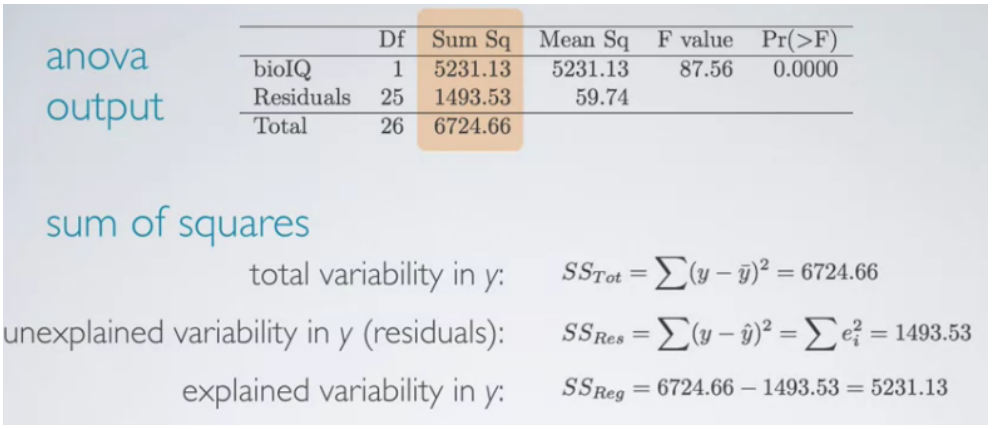
3. Variability Partitioning

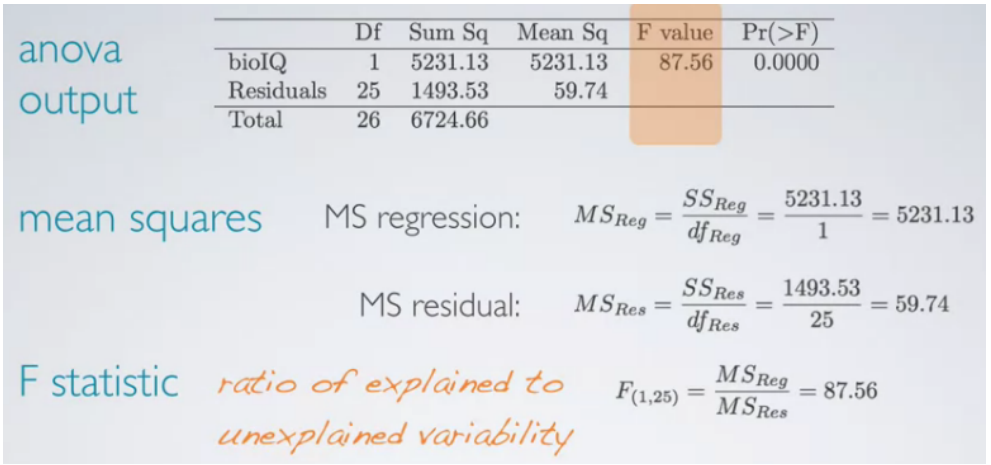
- Variability partitioning
  - So far : t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between x and y.
  - Alternative : consider the variability in y explained by x, compared to the unexplained variability.
  - Partitioning the variability in y to explained and unexplained variability requires **analysis of variance (ANOVA).**

- ANOVA output

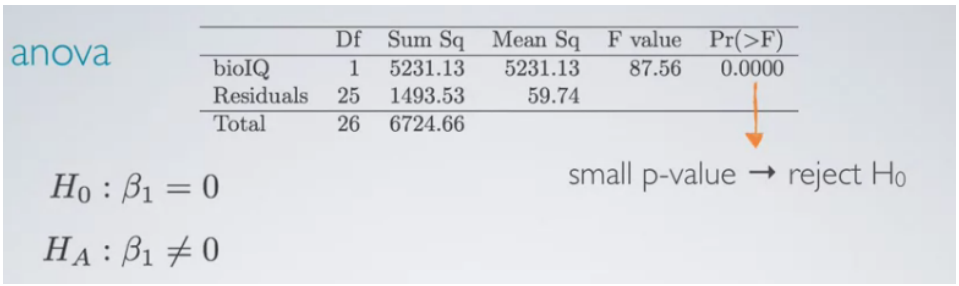
[Study note] Statistics with R specialization - 3. Linear Regression and Modeling

2021 October - written by Mihyeon Jeon



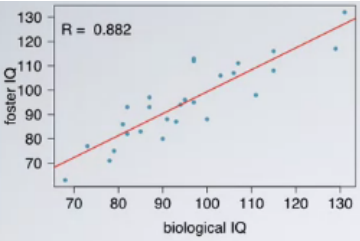


- + Sum of Squares : 제곱의 합, 대표적인 예 - 표준편차. 여러 개의 관측값들이 평균으로부터 얼마나 떨어져있는가?
- + F statistic : 데이터를 모아보니 선형이 잘 되는가? 값이 작을수록 선형이 잘 그려진다는 의미



-> The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

- Revisiting R^2
  - It is the proportion of variability in y explained by the model:
    - large : linear relationship between x and y exists
    - small : evidence provided by the data may not be convincing
  - Two ways to calculate :
    - using correlation : square of the correlation coefficient
    - from the definition : proportion of explained to total variability



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

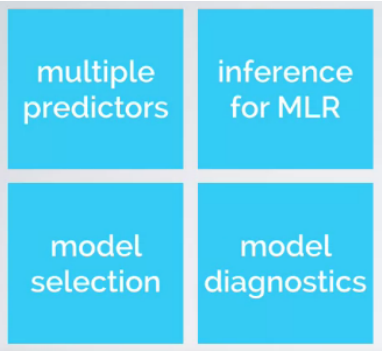
(1) $R^2 = \text{square of correlation coefficient} = 0.882^2 \approx 0.78$

(2) $R^2 = \dfrac{\text{explained variability}}{\text{total variability}} = \dfrac{SS_{Reg}}{SS_{Tot}} = \dfrac{5231.13}{6724.66} \approx 0.78$

Week3

[Regression with multiple predictors]

1. Multiple Predictors



predicting birth weight of babies from a **variety** of variables:

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

$y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6$

2021 October - written by Mihyeon Jeon

R

```
# load data
> library(DAAG)
> data(allbacks)

# fit model
> book_mlr = lm(weight ~ volume + cover, data = allbacks)
> summary(book_mlr)
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 ***
cover:pb    -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
```

- One of the levels of the cover variable is noted on the regression output and it is the non-reference level, which means that the hardcover books must be the reference level.
- This is an observational study!

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

- For hardcover books: plug in **0** for cover:

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 0$$
$$= 197.96 + 0.72\ volume$$

- For paperback books: plug in **1** for cover:

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 1$$
$$= 13.91 + 0.72\ volume$$

- **Slope of volume** : All else held constant, for each 1 cm3 increase in volume the model predicts the books to be heavier on average by 0.72 grams.
- **Slope of cover** : All else held constant, the model predicts that paperback books weigh 185.05 grams lower than hardcover books, on average.
- **Intercept** : Hardcover books with no volume are expected on average to weigh 198 grams.
    - Meaningless in context, serves to adjust the height of the line

prediction

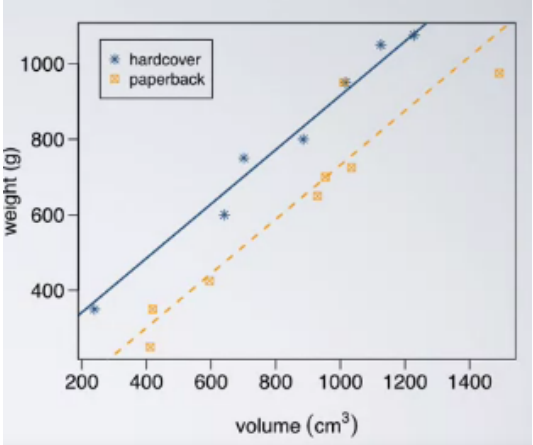
Predict the weight of a paperback book that is 600 cm^3 in volume.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

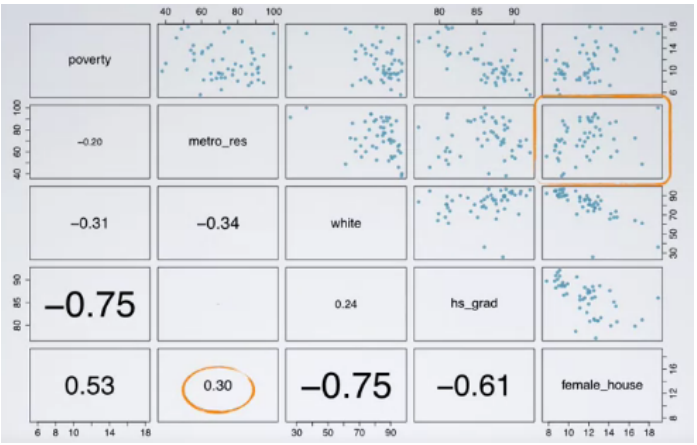
$197.96 + 0.72 \times 600 - 184.05 \times 1 = 445.91$ grams

- Interaction variables



- Model assumes hardcover and paperback books have the same slope for the relationship between their volume and weight.
- If this isn't reasonable, then we would include an interaction variable in the model (beyond the scope of this course).

2. Adjusted R Squared



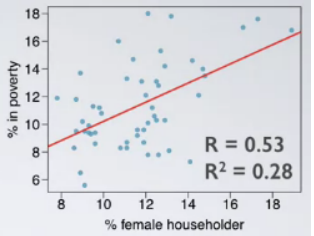
R

```
# load data
> states = read.csv("http://bit.ly/dasi_states")

# fit model
> pov_slr = lm(poverty ~ female_house, data = states)
> summary(pov_slr)

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
```

2021 October - written by Mihyeon Jeon

- dataset : http://d396qusza40orc.cloudfront.net/statistics/lec_resources/states.csv

predicting poverty from % female householder



Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00

another look at R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28$$

predicting poverty from % female householder + % white

R

```
> pov_mlr = lm(poverty ~ female_house + white, data = states)
> summary(pov_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

R

```
> anova(pov_mlr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{132.57 + 8.21}{480.25} = 0.29$$

- **Adjusted R2**

adjusted R^2

adjusted R^2: $R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right)$ k : number of predictors

- The R squared value is going up each time you add a new predictor to your model.
- We need a more honest measure.
- Adjusted R squared applies a penalty to R squared.
- The larger the sample size, the more predictors the model can handle, and therefore the less the penalties are going to be for more additional predictors being added to the model.
- Adjusted R squared is only going up if the added variable is actually of value.
- If the additional percentage of variability in the response variable explained by that new variable can offset the penalty for the additional number of predictors in the model.

2021 October - written by Mihyeon Jeon

Calculate adjusted R^2 for the multiple linear regression model predicting % living in poverty from % female householders and % white. Remember n = 51 (50 states + DC).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right)$$

$$= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1}\right) = 0.26$$

R^2 vs. adjusted R^2

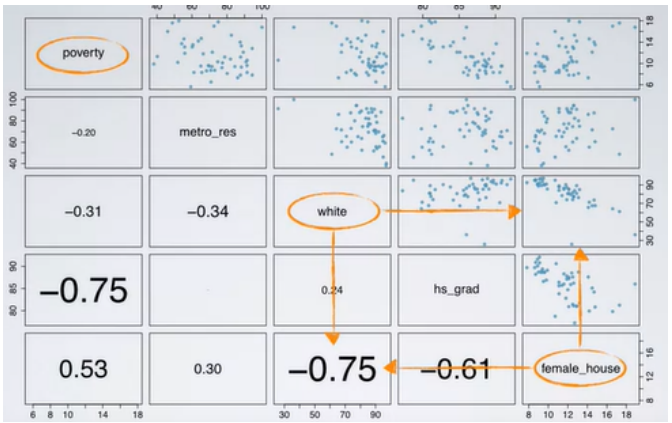
	R^2	adjusted R^2
Model 1 (poverty vs. female_house)	0.28	0.28
Model 1 (poverty vs. female_house + white)	0.29	0.26

- If the added variable doesn't really provide any new information or is completely unrelated, the adjusted R squared does not increase.

- Properties of adjusted R squared
  - k is never negative -> adjusted R^2< R^2
  - adjusted R^2 applies a penalty for the number of predictors included in the model.
  - We choose models with higher adjusted R^2 over others.

3. Collinearity and Parsimony

- Collinearity
  - Two predictor variables are said to be collinear when they are correlated with each other.
  - Remember : Predictors are also called independent variables, so they should be independent of each other.
  - Inclusion of collinear predictors (also called multicollinearity) complicated model estimation.



- The correlation between the variable white and femlae_householder is quite strong.
- They are not independent and we wouldn't want to add the variable white to our existing model as it's going to bring nothing new to the table.
- In addition, using both of these variables in the model is going to result in multicollinearity which we said might also result in unreliable estimates of the coefficients from the model.

- parsimony
    - Avoid adding predictors associated with each other because oftentimes the addition of such variables brings nothing new to the table.
    - Prefer the simplest best model, i.e. the parsimonious model
        - Occam's razor : Among competing hypotheses, the one with the fewest assumptions should be selected
    - Addition of collinear variables can result in biased estimates of the regression parameters
    - While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to control for correlated predictors

[Inference for multiple regression and model selection]

1. Inference for MLR

- Data

modeling cognitive test scores of children

Data: Cognitive test scores of three- and four-year-old children and characteristics of their mothers (from a subsample from the National Longitudinal Survey of Youth).

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
...	...	...	...	...	...
6	98	no	107.90	no	18
...	...	...	...	...	...
434	70	yes	91.25	yes	25

R

```
# load data
> cognitive = read.csv("http://bit.ly/dasi_cognitive")

# full model
> cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive)
> summary(cog_full)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341 *
mom_hs:yes   5.09482    2.31450   2.201   0.0282 *
mom_iq       0.56147    0.06064   9.259   <2e-16 ***
mom_work:yes 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
```

- Inference for the model as a whole

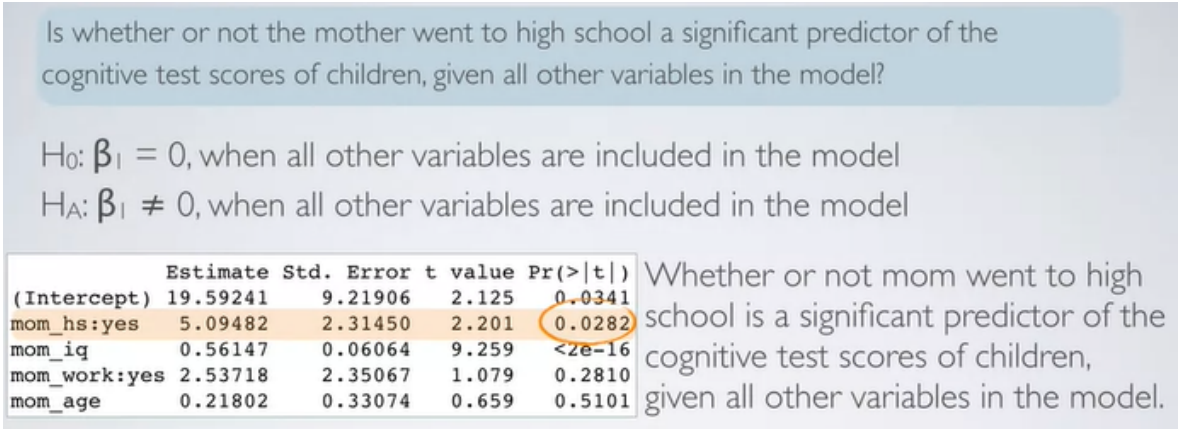
$H_0: \beta_1 = \beta_1 = \ldots = \beta_k = 0$

H_A: At least one β_i is different than 0

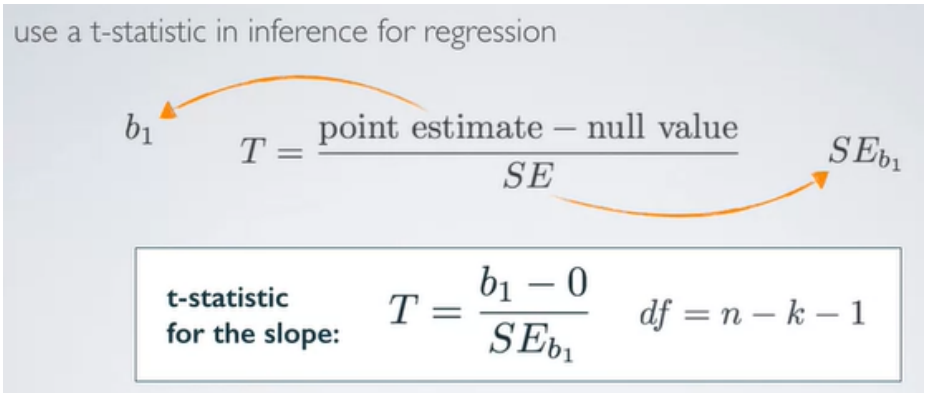
```
F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```

- Since p-value < 0.05, the model as a whole is significant.
    - The F test yielding a significant result doesn't mean the model fits the data well, just means at least one of the betas is non zero.
    - The F test not yielding a significant result doesn't mean individual variables included in the model are not good predictors of y, it just means the combination of these variables doesn't yield a good model.

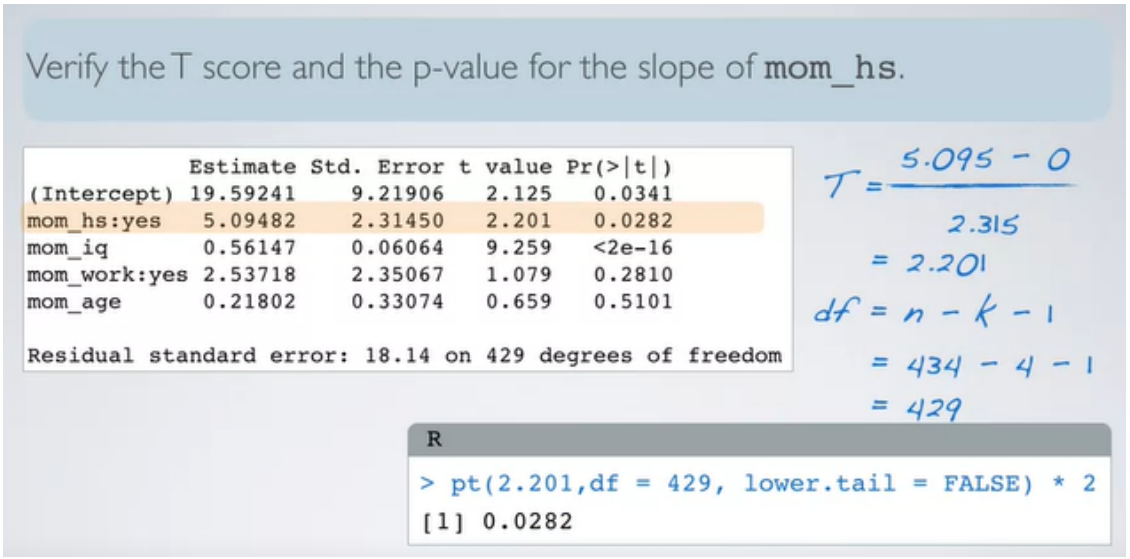
- hypothesis testing for slopes



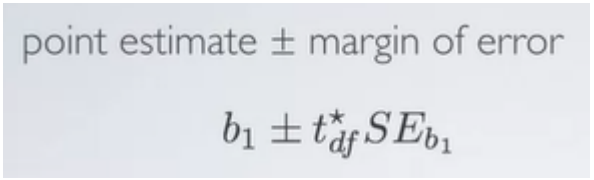
- testing for the slope - mechanics

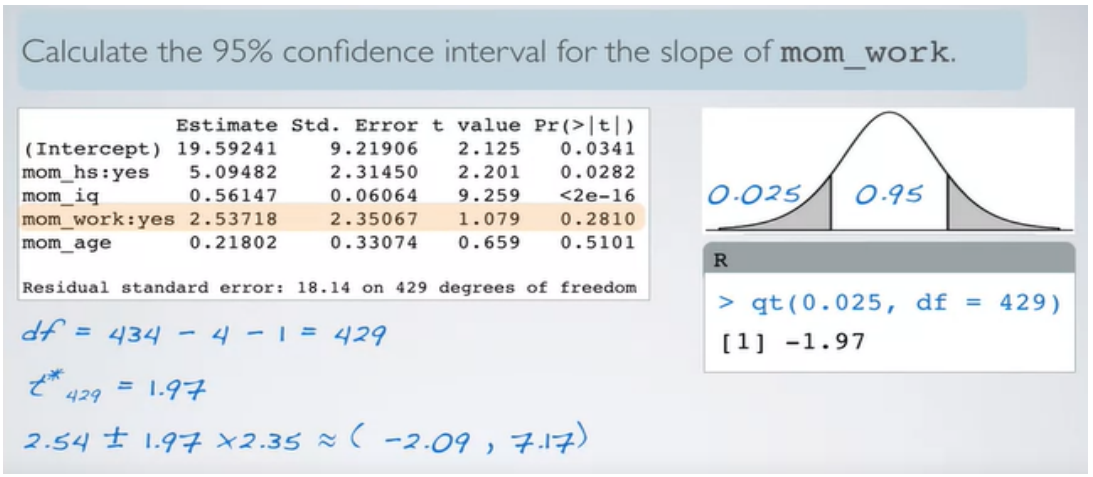


- k is the number of predictors included in the model



- Confidence intervals for slopes





-> We are 95% confident that, all else being equal, the model predicts that children whose moms worked during the first three years of their lives score 2.09 points lower to 7.17 points higher than those whose moms did not work.

2. Model Selection

- Stepwise model selection
  - **backwards elimination** : start with a full model (containing all predictors), drop one predictor at a time until the parsimonious model is reached.
  - **forward selection** : start with an empty model and add one predictor at a time until the parsimonious model is reached.
  - criteria
    - p-value, adjusted R squared
    - AIC, BIC, DIC, Bayes factor, Mallow's Cp (beyond the scope of this course)

- backwards elimination - adjusted R squared
    1. Start with the full model
    2. Drop one variable at a time and record adjusted R squared of each smaller model
    3. Pick the model with the highest increase in adjusted R squared
    4. Repeat until none of the models yield an increase in adjusted R squared

step	variables included	removed	adjusted R^2
FULL	`kid_score ~ mom_hs + mom_iq + mom_work + mom_age`		0.2098
STEP 1	`kid_score ~ mom_iq + mom_work + mom_age`	[-mom_hs]	0.2027
	`kid_score ~ mom_hs + mom_work + mom_age`	[-mom_iq]	0.0541
	`kid_score ~ mom_hs + mom_iq + mom_age`	[-mom_work]	0.2095
	`kid_score ~ mom_hs + mom_iq + mom_work`	[-mom_age]	0.2109
STEP 2	`kid_score ~ mom_iq + mom_work`	[-mom_hs]	0.2024
	`kid_score ~ mom_hs + mom_work`	[-mom_iq]	0.0546
	`kid_score ~ mom_hs + mom_iq`	[-mom_work]	0.2105

- backwards elimination - p-value
    1. Start with the full model
    2. Drop the variable with the highest p-value and refit a smaller model
    3. Repeat until all variables left in the model are significant

FULL	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.5924	9.2191	2.13	0.0341
mom_hs:yes	5.0948	2.3145	2.20	0.0282
mom_iq	0.5615	0.0606	9.26	0.0000
mom_work:yes	2.5372	2.3507	1.08	0.2810
~~mom_age~~	~~0.2180~~	~~0.3307~~	~~0.66~~	~~0.5101~~

STEP 1	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.1794	6.0432	4.00	0.0001
mom_hs:yes	5.3823	2.2716	2.37	0.0183
mom_iq	0.5628	0.0606	9.29	0.0000
~~mom_work:yes~~	~~2.5664~~	~~2.3487~~	~~1.09~~	~~0.2751~~

STEP 2	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.7315	5.8752	4.38	0.0000
mom_hsyes	5.9501	2.2118	2.69	0.0074
mom_iq	0.5639	0.0606	9.31	0.0000

practice)

2021 October - written by Mihyeon Jeon

The following model uses data from the American Community Survey to predict income from hours worked per week, race, and gender. Which variable (if any) should be dropped from the model first when doing backwards elimination using the p-value approach?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2782.5726	6676.5534	0.42	0.6770
hrs_work	1247.2128	146.2013	8.53	0.0000 ✔
race:black	-9565.3090	6393.2168	-1.50	0.1350
race:asian	35816.6156	8690.3484	4.12	0.0000 ✔
race:other	-11112.8617	7213.3220	-1.54	0.1238
gender:female	-16430.0916	3803.4700	-4.32	0.0000 ✔

don't drop any variables

- If you have a categorical variable with multiple levels, you cannot drop part, some of the levels of that variable and keep others.

- Adjusted R squared vs. p-value
  - p-value : significant predictors
  - adjusted R squared : more reliable predictions
  - p-value method depends on the (somewhat arbitrary) 5% significance level cutoff
    - different significance level -> different model
    - used commonly since it requires fitting fewer models(in the more commonly used backwards-selection approach)

- forward selection- adjusted R squared
  1. Start with single predictor regression of response vs. each explanatory variable
  2. Pick the model with the highest adjusted R squared
  3. Add the remaining variables one at a time to the existing model, and pick the model with the highest adjusted R squared
  4. Repeat until the addition of any of the remaining variables does not result in a higher adjusted R squared.

step	variables included	adjusted R^2
STEP 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	0.1991
STEP 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	0.2105
STEP 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	0.2109
STEP 4	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	0.2098

- forward selection - p-value
  1. Start with single predictor regressions of response vs.each explanatory variable
  2. Pick the variable with the lowest significant p-value
  3. Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
  4. Repeat until any of the remaining variables do not have a significant p-value

- expert opinion
    - Variables can be included in (or eliminated from) the model based on expert opinion
    - If you are studying a certain variable, you might choose to leave it in the model regardless of whether it's significant or yield a higher adjusted R squared.

final model

R

```
> cog_final = lm(kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)
> summary(cog_final)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.17944    6.04319   4.001 7.42e-05 ***
mom_hsyes    5.38225    2.27156   2.369   0.0183 *
mom_iq       0.56278    0.06057   9.291  < 2e-16 ***
mom_workyes  2.56640    2.34871   1.093   0.2751

Residual standard error: 18.13 on 430 degrees of freedom
Multiple R-squared:  0.2163, Adjusted R-squared:  0.2109
F-statistic: 39.57 on 3 and 430 DF,  p-value: < 2.2e-16
```

- We selected the model with the variable 'mom_work' although it's not statistically significant.
- We selected this model using the adjusted R squared method, which tells us that including that variable actually gives the model higher predictive power even though the variable may not be statistically significant.

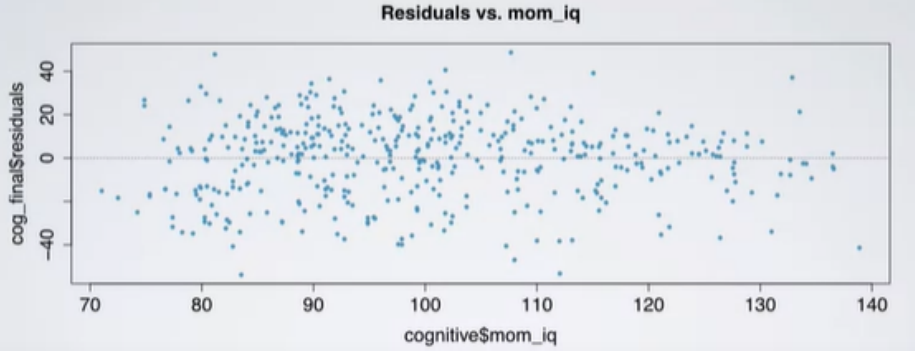
3. Diagnostics for MLR

(1) Linear relationships between (numerical) x and y

- each (numerical) explanatory variable linearly related to the response variable
- check using residuals plots (e vs. x)
    - looking for a random scatter around 0
    - instead of scatterplot of y vs. x: allows for considering the other variables that are also in the model, and not just the bivariate relationship between a given x and y

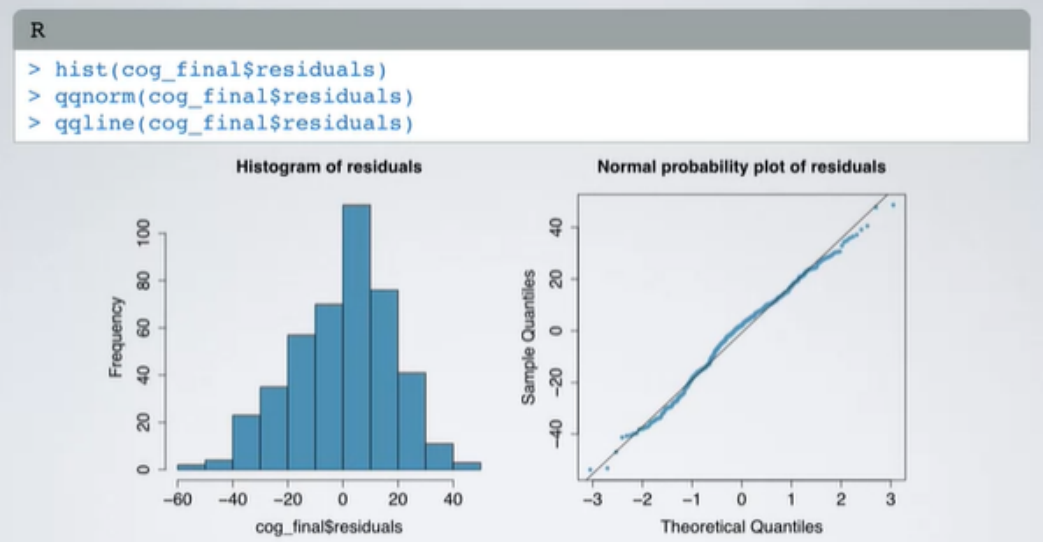
R

```
> cog_final = lm(kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)
> plot(cog_final$residuals ~ cognitive$mom_iq)
```



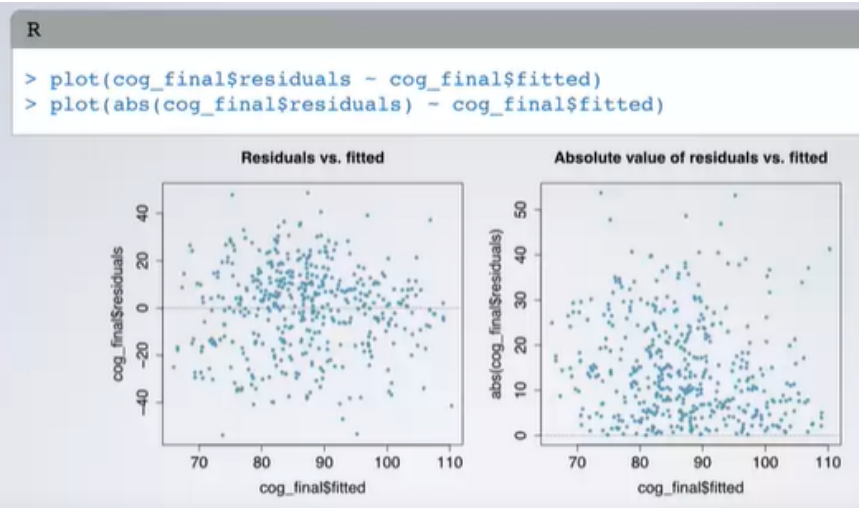
(2) Nearly normal residuals with mean 0

- some residuals will be positive and some negative
- on a residuals plot we look for random scatter of residuals around 0
- this translate to a nearly normal distribution of residuals centered at 0
- check using histogram or normal probability plot



(3) constant variability of residuals

- residuals should be equally variable for low and high values of the predicted response variable
- check using residuals plots of residuals vs. predicted (e vs. $\hat{y}$)
    - residuals vs. predicted instead of residuals vs. x because it allows for considering the entire model (with all explanatory variables) at once
    - residuals randomly scattered in a band with a constant width around 0 (no fan shape)
    - also worthwhile to view the absolute value of residuals vs. predicted to identify unusual observations easily.



(4) independent residuals

- independent residuals -> independent observations

- if time series structure is suspected check using residuals vs. order of data collection
- if not, think about how the data are sampled