**Linear Regression and Modeling**

Kunyu HE

**Week 1 About Linear Regression and Modeling**

* Correlation and Association
* Correlation: the linear association between two numerical variables

Association: nonlinear relationship

* Describe an association

Direction: positive or negative;

Form: linear or nonlinear;

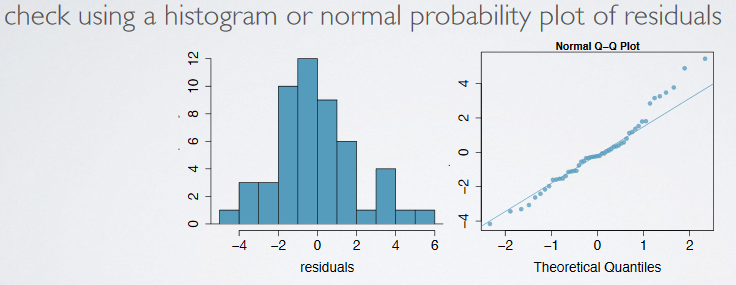
Strength

* Correlation coefficient: R; sensitive to outliers
* Least Square Line
* Residuals: ;
* Conditions

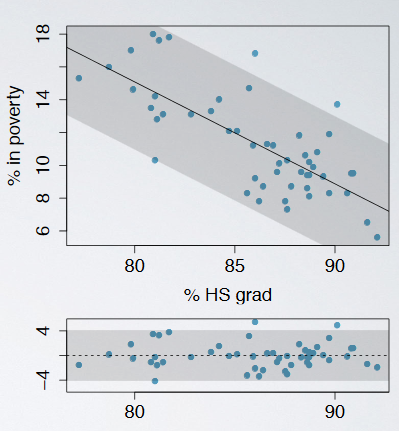
Linearity:



Nearly normal residuals:



Constant variability:’



* Formula:
* Slope and interpretation:

;

Interpretation: ;

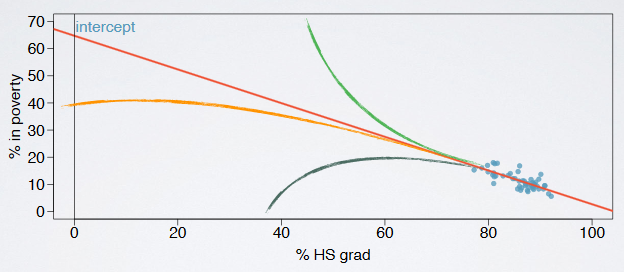
* intercept and interpretation

;

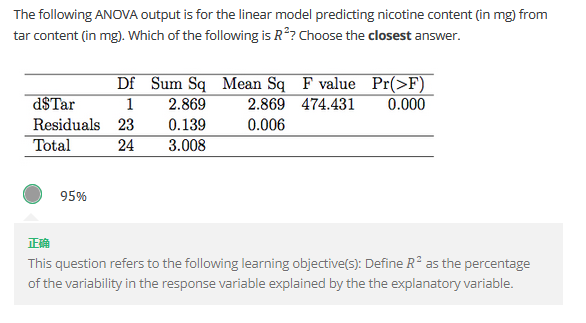
Interpretation:

* Prediction and Extrapolation

Extrapolation: applying a model estimate to values outside of the realm of the original data, sometimes the regression line may change beyond the observed scale



* Notice: ANOVA, with 95% confidence



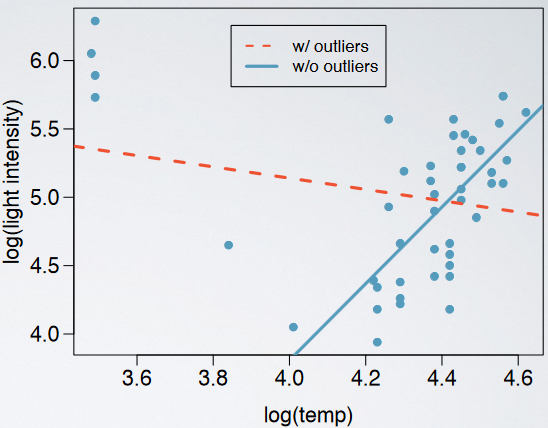
**Week 2 Outliers, Inference and Variability Partitioning**

* Outliers
* Classification:

Leverage points: outliers that fall horizontally away from the center of the cloud but don’t influence the slope of the regression line;

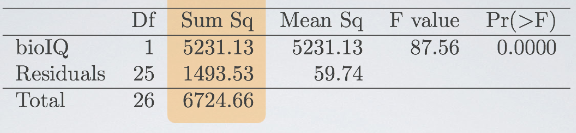
Inferential points: outliers that actually influence the slope of the regression line.

* To remove outliers:

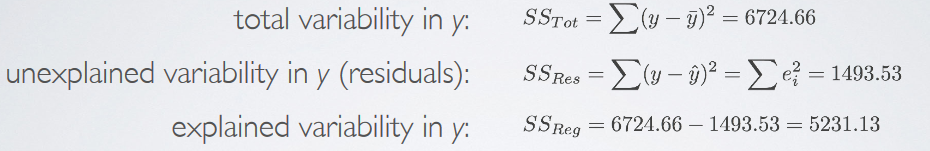


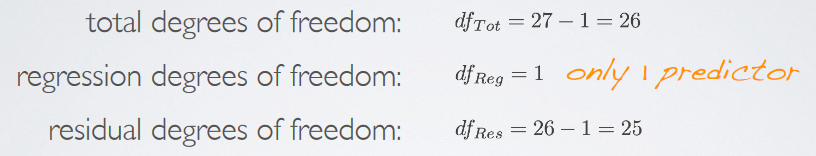
The influential points above can be interesting cases to pay extra attention to.

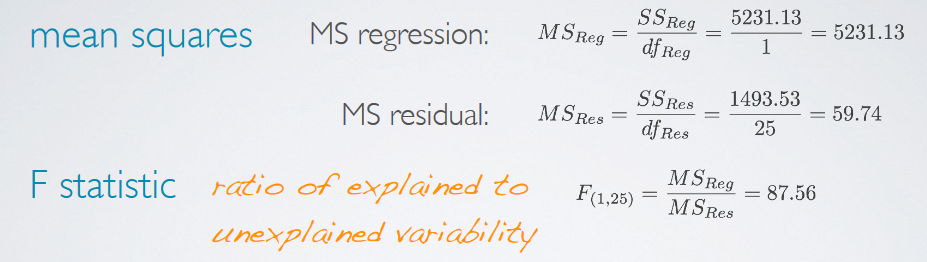
* Hypothesis test for slope
* Conditions: CLT conditions
* Distribution: t-distribution
* Parameters:
* Confidence interval of slope
* Conditions: CLT conditions
* Distribution: t-distribution
* Interval:
* Variability Partitioning and R2
* Variability partitioning: partitioning the variability in y to explained (model) and unexplained variability (residuals) with an ANOVA, and test whether the slope is a significant predictor
* ANOVA output



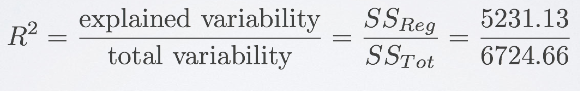
* Parameters



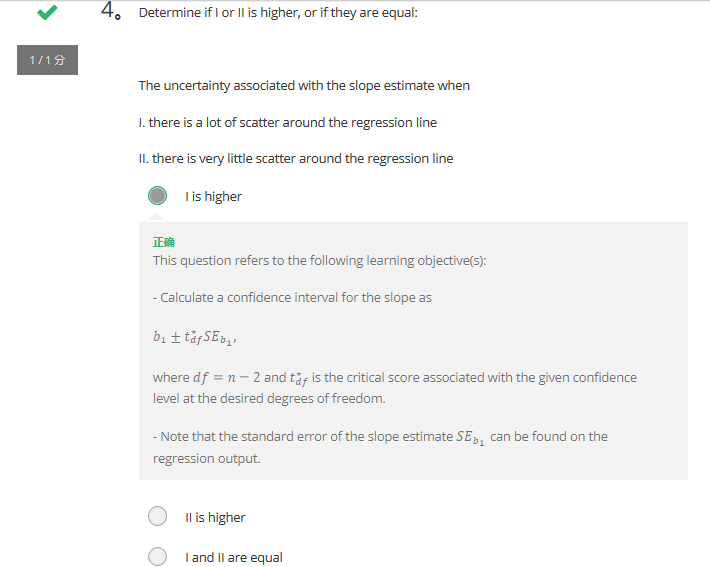








* Notice: the number of scatters refer to n in



**Week 3 Multiple Linear Regression (MLR)**

* Model Selection
* Forward Selection or Backwards Elimination

Criteria: p-value (significant predictors), adjusted R2(reliable predictors)

Other criteria: AIC, BIC, DIC, Mallows Cp

* Adjusted R2

.

* Inference for MLR and Diagnostics
* Inference for the model as a whole

H0: all

H1: at least one of the

* Hypothesis test for slopes

T-statistic;

;

p-value: >pt(t-statistic, df, lower.tail=TRUE/FALSE)

* Confidence intervals for slopes

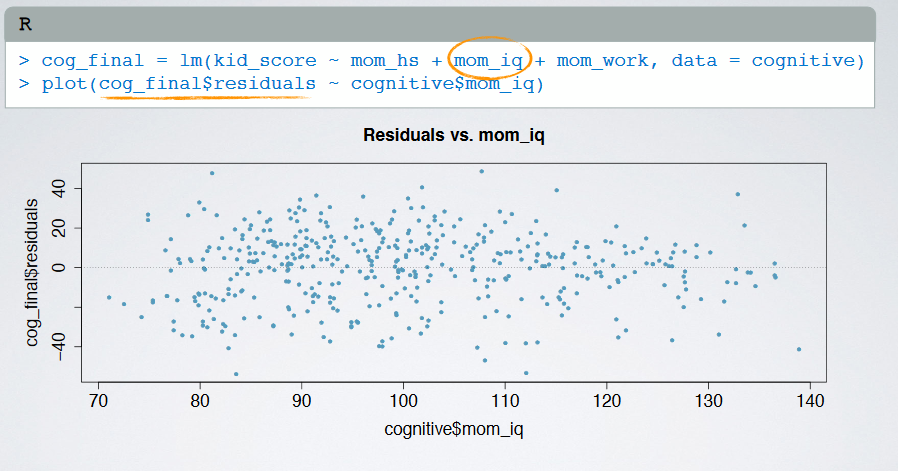
;

;

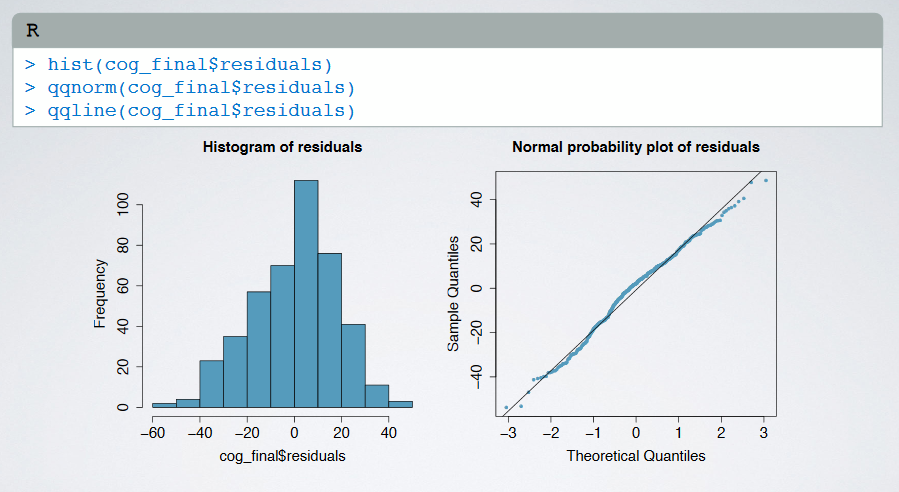
CI:

* Diagnostics

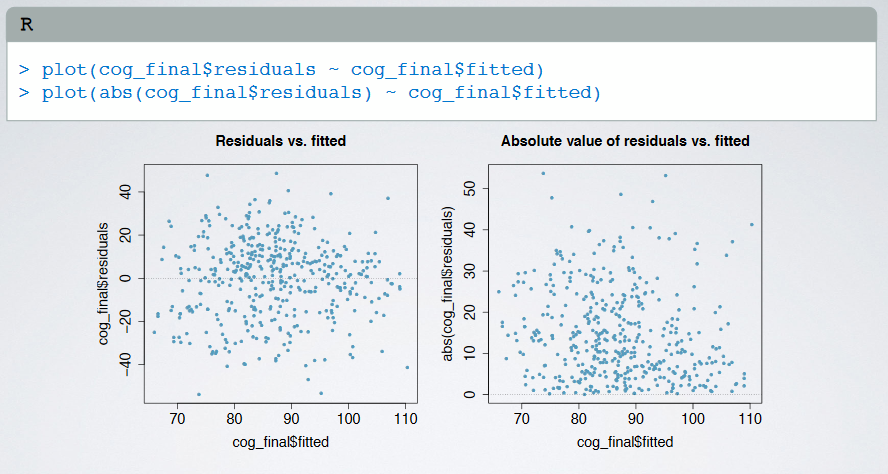
Linearity: scatter plots of residuals against x1



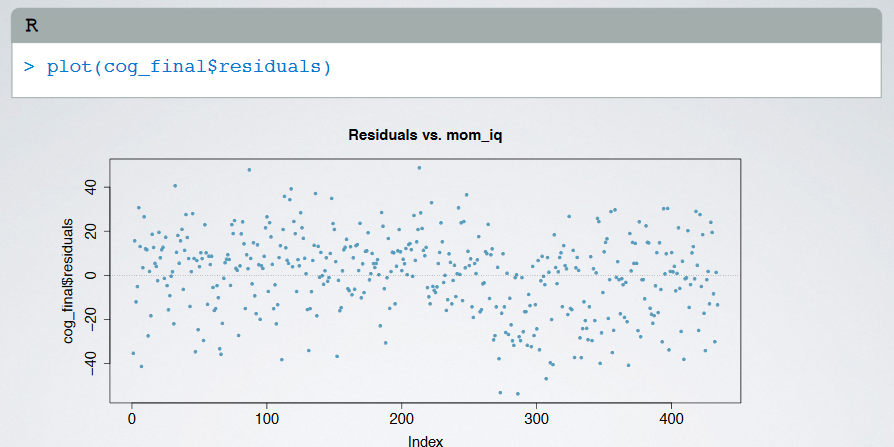
Nearly normal residuals: histograms and normal probability distribution plots



Constant variability of residuals: absolute value of residuals agaist predicted y



Independent residuals: no time series structure



* Notice: in backwards elimination, eliminate the variable with highest p-value first; when a categorical variable has multiple levels, use the lowest for comparison