Stat 504: Statistical Approaches to Interdisciplinary Data Analysis

Linear model concepts

Overview for this session

- Discussion of important concepts in using linear models and assessing model adequacy

Linear model concepts

- In our introductory session we viewed the two sample t test as arising from a particular linear model:
- $Y_{ij} = \mu_i + \varepsilon_{ij}$
- Where the errors ε_{ij} are assumed to be independent and normally distributed with common mean 0 and variance σ^2

Linear model concepts

- More generally many models can be written in the form:
- response = linear model + error,
- including multiple regression and ANOVA models
- Here we focus on how to assess the fitted model, identify potential problems, and possible remedies

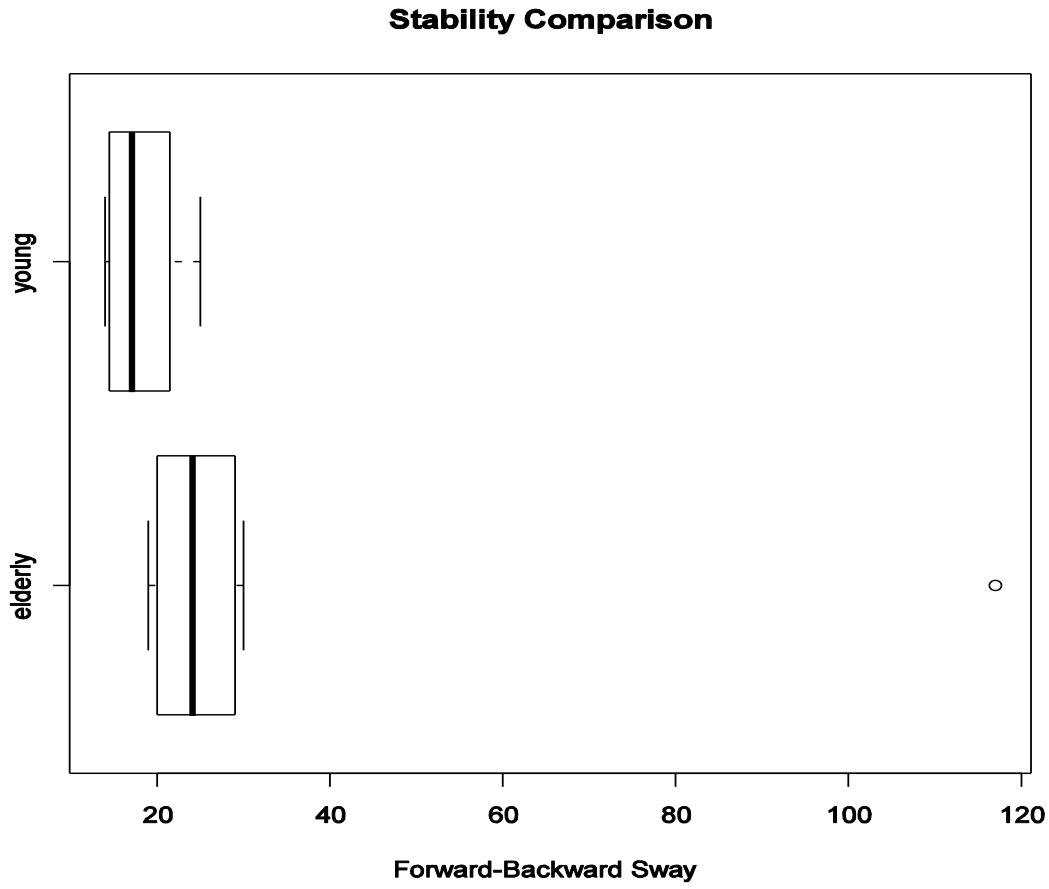
Model assessment

- Assess the assumptions of linearity, normality, independence, and constant variance
- Detect problems caused by outliers, high-leverage points, and highly correlated predictor variables

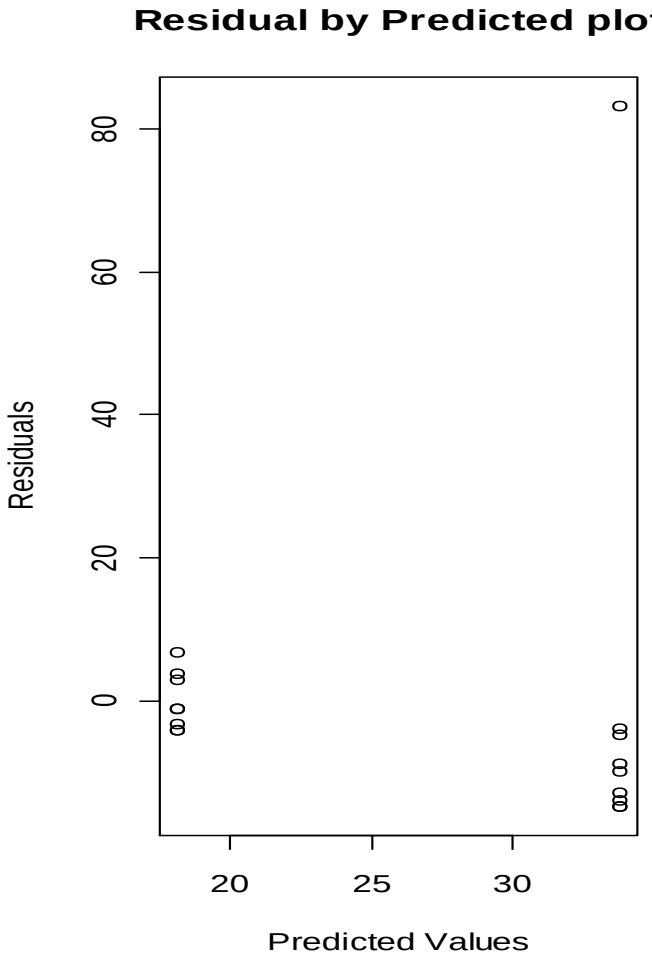
More Model Assessment

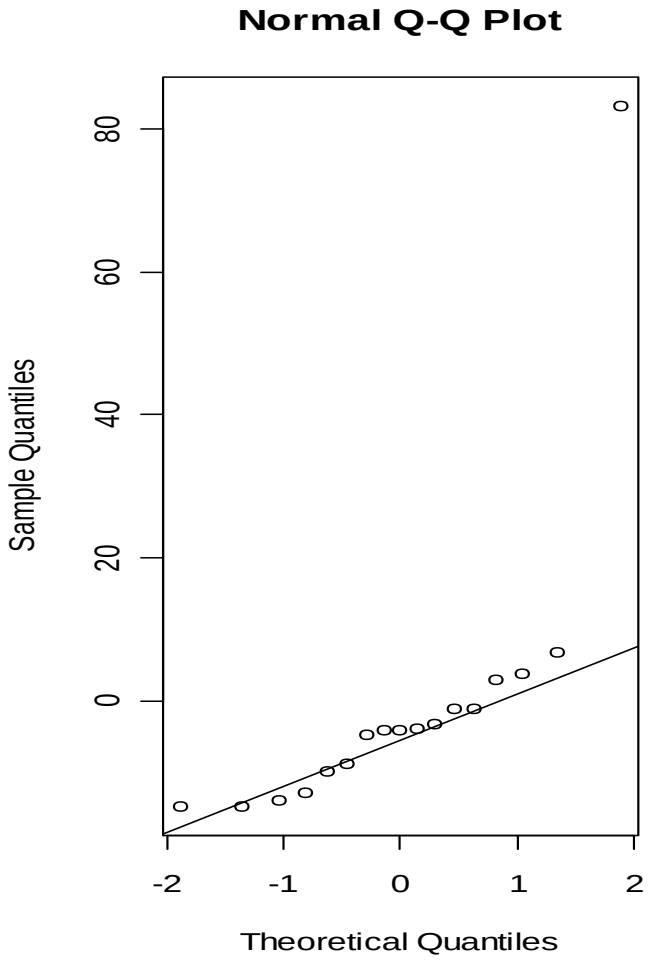
- Our first two tools for assessment are:
- The residuals e_{ij} = actual Y_{ij} – predicted Y_{ij} for each observation, and
- Leverage values (h_i), which measure for each observation how far it is from the mean of the set of predictor variables
- The average value of h_i is (p+1)/n, and a rough guide is that h_i values > 2(p+1)/n are considered high-leverage points

Back to the stability data



Two commonly used plots for any linear model





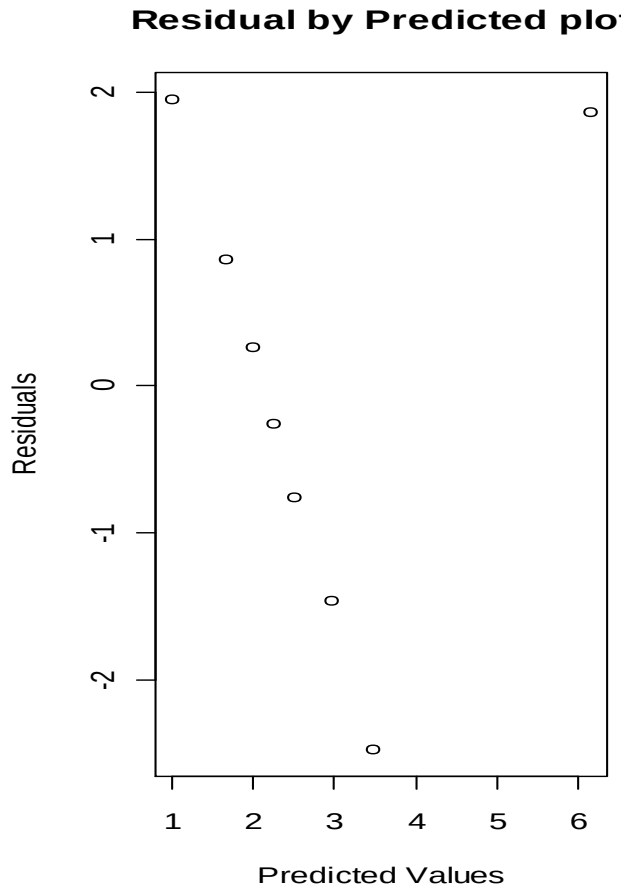
Stability data leverage values

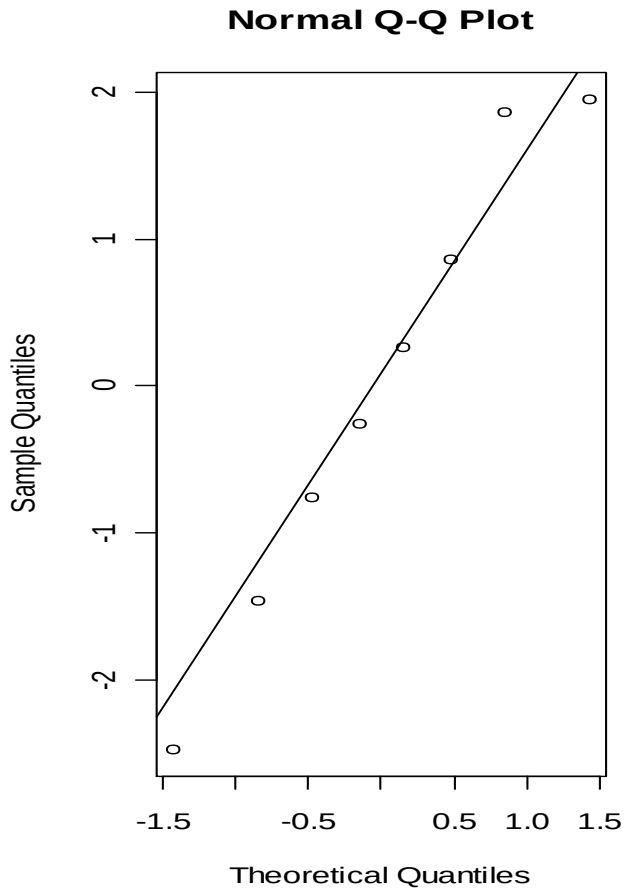
- Since the model has no continuous covariates, leverage is not a problem and is nearly equal for all points (.111 or .125)

Another example

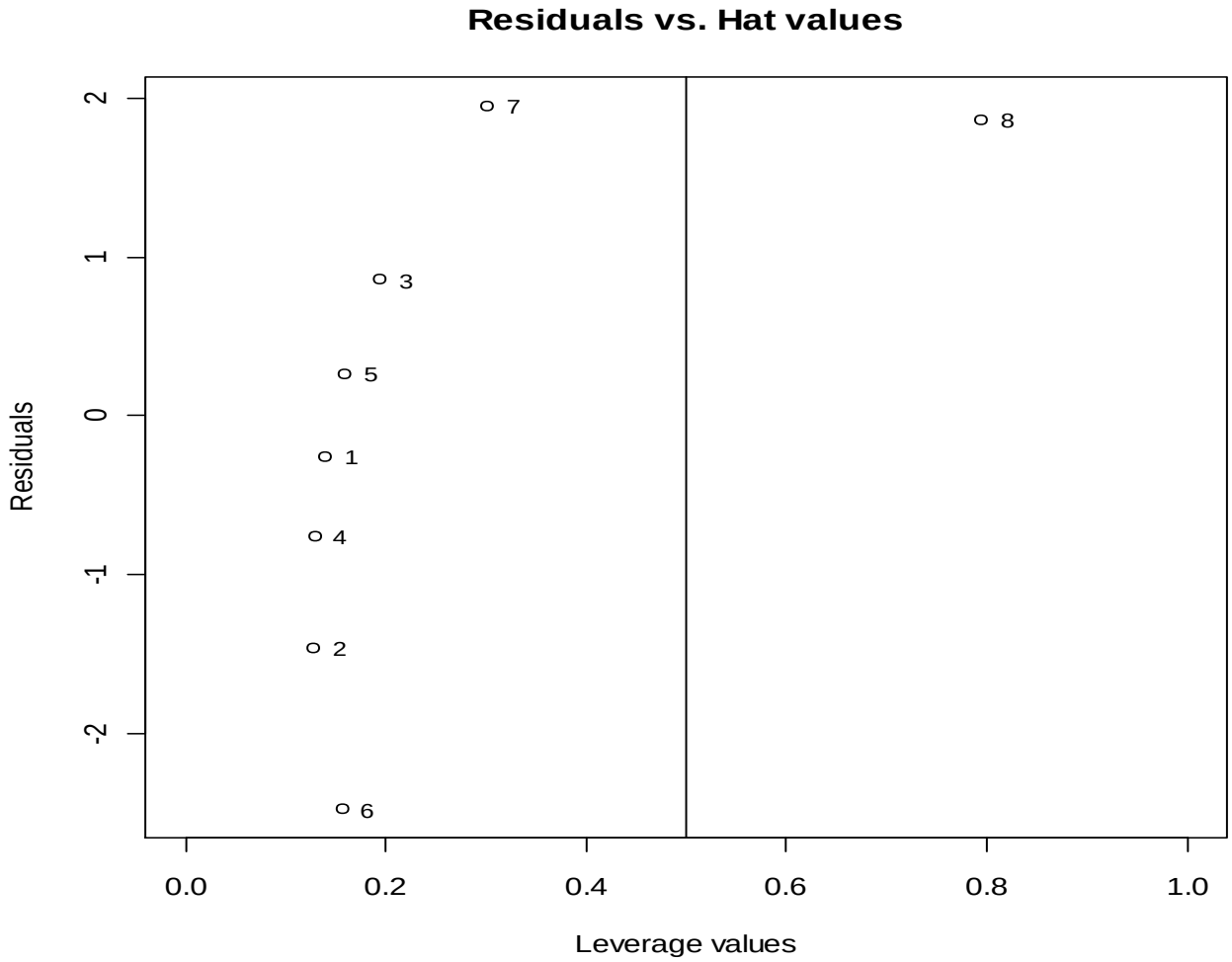
- A simple linear regression model based on 8 observations yields the following least-squares prediction equation:
- Predicted Y = .3 + .65X
- The test statistic for the slope coefficient is t = 2.46 on 6 df, P = .049
- How would you interpret these results?

Residual plots

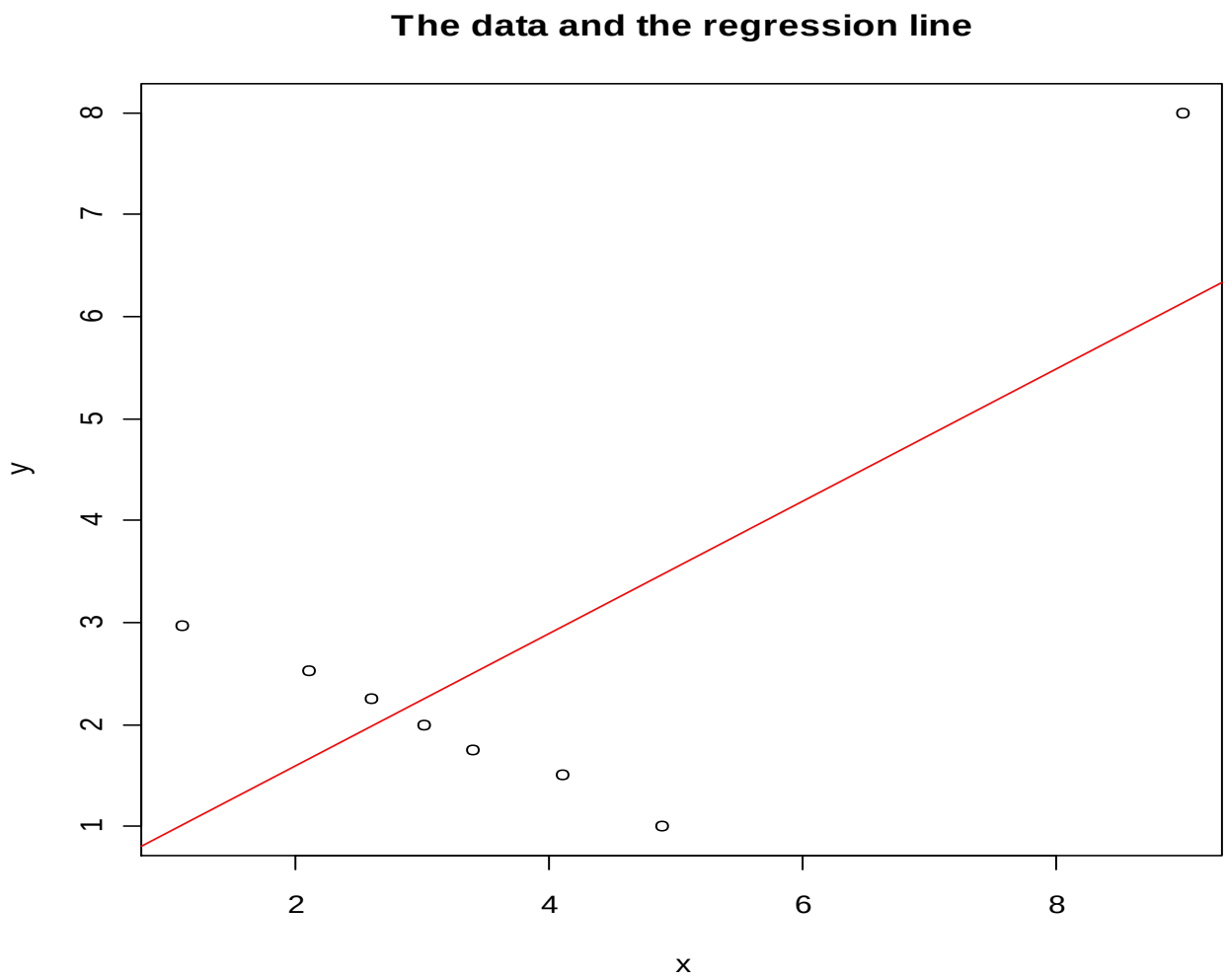




Residuals and leverage values



Finally, the raw data and fitted line



Conclusions for these data

- We can see that one data point has distorted the entire regression line
- There are no outliers (as measured by the raw residuals), the problem is a high-leverage point
- What do you think is an appropriate analysis of these data?

An interesting new problem, for a different data set

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	15253	7626.73151	117.58	<.0001
Error	47	3048.58026	64.86341		
Corrected Total	49	18302			

Root MSE	8.05378	R-Square	0.8334
Dependent Mean	31.34478	Adj R-Sq	0.8263
Coeff Var	25.69417		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.10827	2.45455	-0.45	0.6537
x1	1	-2.14577	3.66422	-0.59	0.5609
x2	1	1.29969	0.90890	1.43	0.1593

- For the previous regression printout, the global model test is highly significant ($F = 117.58$, $P < .0001$), but the test for each individual coefficient is non-significant (P values of .56 and .16)
- This can happen when the predictor variables (x1 and x2 here) are highly related

One more tool for model assessment

- The Variance Inflation Factor (VIF) for a predictor variable X_i is $VIF = 1/(1 - R_i^2)$
- R_i^2 is the R^2 for a regression model to predict X_i from all of the other X variables in the regression model
- A rough guideline is that if $R_i^2 > .9$ ($VIF > 10$) then at least some of the X_i variables are highly linearly related, which leads to many problems

VIF applied to the strange regression model

- For that example, x1 and x2 are highly correlated, and VIF = 328, far above the rough cutoff of 10
- Although this example involved only two predictors, in practice there can be several variables where one variable is a linear function of others. VIF can detect this but individual correlations might not show the problem

How to handle VIF problems

- There are many approaches, here are two:
- Informally, look at variable plots and intercorrelations then choose a reduced set of variables that make sense scientifically and have lower VIF's
- Use some type of model selection procedure (best subset, etc) to pick a smaller set of predictors

Putting it all together

- Use tools like residual and leverage plots and VIF values to examine a model, to see if model assumptions appear tenable and to avoid the effects of high-leverage points and collinearity in regression models
- There are other plots and methods available to examine linear (and other) models, but these tools make a very good start