Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Ranking States by SAT Performance: Model Fit

About This Lesson

Residual Analysis

To evaluate assumptions:

- ***Constant variance & uncorrelated errors***
  - Response variable or fitted values vs residuals
- ***Linearity***
  - Predicting variables vs residuals
- ***Normality***
  - Histogram and QQ normal plot

To evaluate outliers:

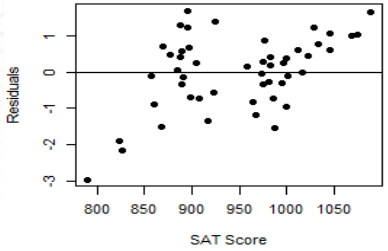
- Cook’s distance plots

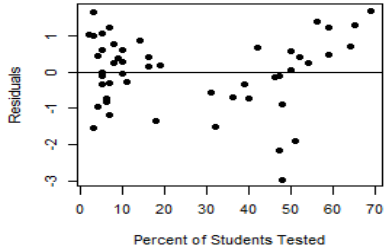
Residual Analysis

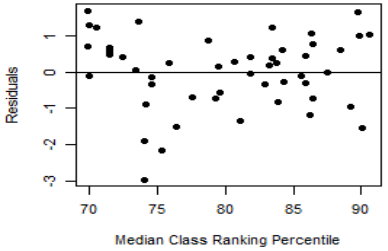
```
## Residual analysis for the reduced model
res = stdres(reduced.line)
cook = cooks.distance(reduced.line)
par(mfrow = c(2,3))
plot(sat, res, xlab = "SAT Score", ylab = "Residuals", pch = 19)
abline(h = 0)
plot(takers, res, xlab = "Percent of Students Tested", ylab = "Residuals", pch = 19)
abline(h = 0)
plot(rank, res, xlab = "Median Class Ranking Percentile", ylab = "Residuals", pch = 19)
abline(h = 0)
hist(res, xlab="Residuals", main= "Histogram of Residuals")
qqnrom(res)
qqline(res)
plot(cook,type="h",lwd=3, ylab = "Cook's Distance")
```

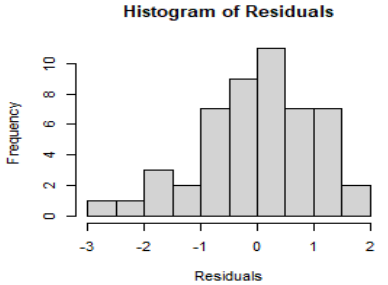
Georgia Tech

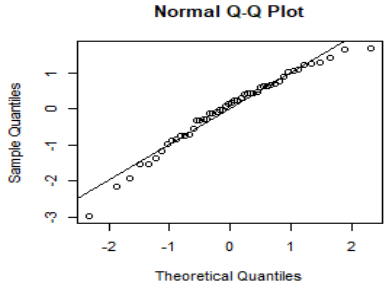
Residual Analysis

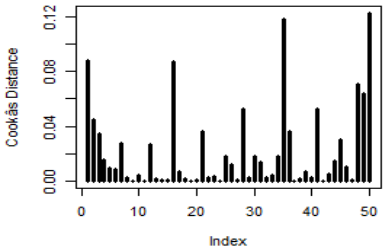




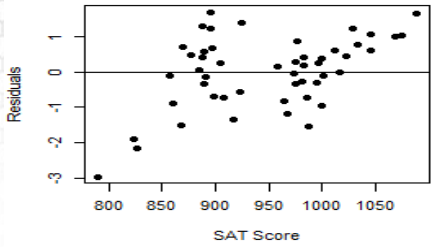


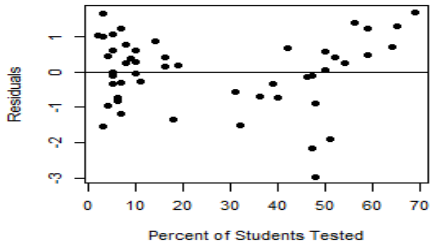


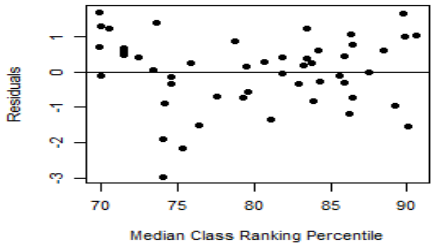


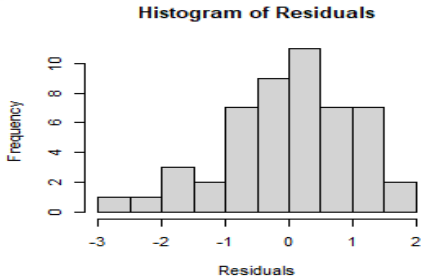


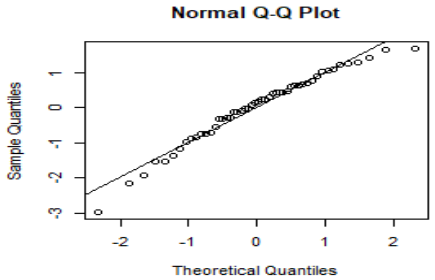
Residual Analysis

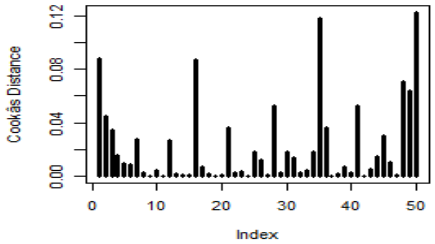












- Transform the predicting variable Percent of Students Tested (*takers*)
- Reanalyze heavy tailed residuals and outliers after transformation

Georgia Tech

Linear Regression Analysis in R

regression.line = lm(sat ~ log(takers)+rank+income+years+public+expend)
summary(regression.line)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	407.53990	282.76325	1.441	0.15675	
log(takers)	-38.43758	15.95214	-2.410	0.02032	*
rank	4.11427	2.50166	1.645	0.10734	
income	-0.03588	0.13011	-0.276	0.78407	
years	17.21811	6.32007	2.724	0.00928	**
public	-0.11301	0.56239	-0.201	0.84168	
expend	2.56691	0.80641	3.183	0.00271	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.86 on 43 degrees of freedom
Multiple R-squared: 0.8919, Adjusted R-squared: 0.8769
F-statistic: 59.15 on 6 and 43 DF, p-value: < 2.2e-16

Georgia Tech

Linear Regression Analysis in R

regression.line <- lm(sat ~ log(takers)+rank+income+years+public+expend)
summary(regression.line)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	407.53990	282.76325	1.441	0.15675	
log(takers)	-38.43758	15.95214	-2.410	0.02032	*
rank	4.11427	2.50166	1.645	0.10734	
income	-0.03588	0.13011	-0.276	0.78407	
years	17.21811	6.32007	2.724	0.00928	**
public	-0.11301	0.56239	-0.201	0.84168	
expend	2.56691	0.80641	3.183	0.00271	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.86 on 43 degrees of freedom
Multiple R-squared: 0.8919, Adjusted R-squared: 0.8769
F-statistic: 59.15 on 6 and 43 DF, p-value: < 2.2e-16

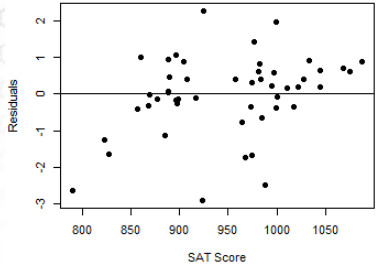
Test for statistical significance:

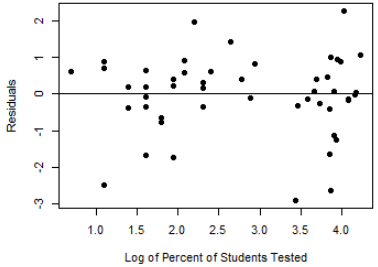
$\hat{\beta}_{\log(takers)}$	Pr(>|t|) ≈ 0.0203 < 0.1
$\hat{\beta}_{rank}$	Pr(>|t|) ≈ 0.1073 > 0.1
$\hat{\beta}_{income}$	Pr(>|t|) ≈ 0.7840 > 0.1
$\hat{\beta}_{years}$	Pr(>|t|) ≈ 0.0093 < 0.1
$\hat{\beta}_{public}$	Pr(>|t|) ≈ 0.8417 > 0.1
$\hat{\beta}_{expend}$	Pr(>|t|) ≈ 0.0027 < 0.1

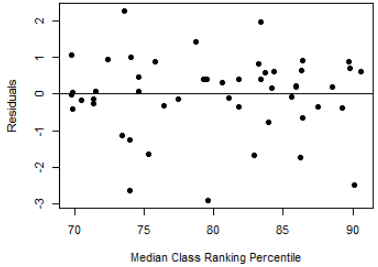
$\hat{\sigma}$ = 24.86, df = *n*-*p*-1 = 43
R^2 ≈ 0.892 ⇒ 89.2% of variability explained

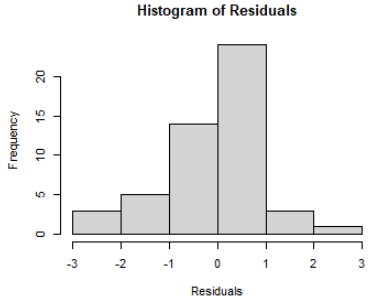
Georgia Tech

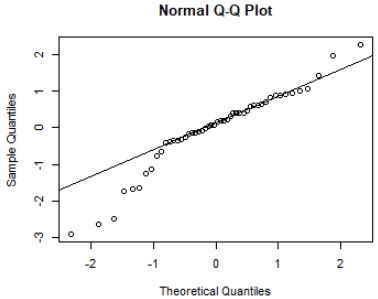
Residual Analysis

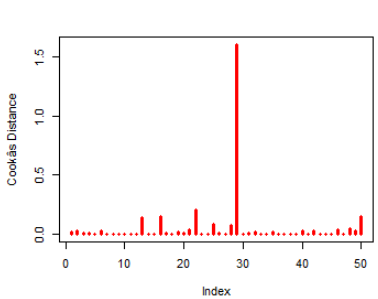




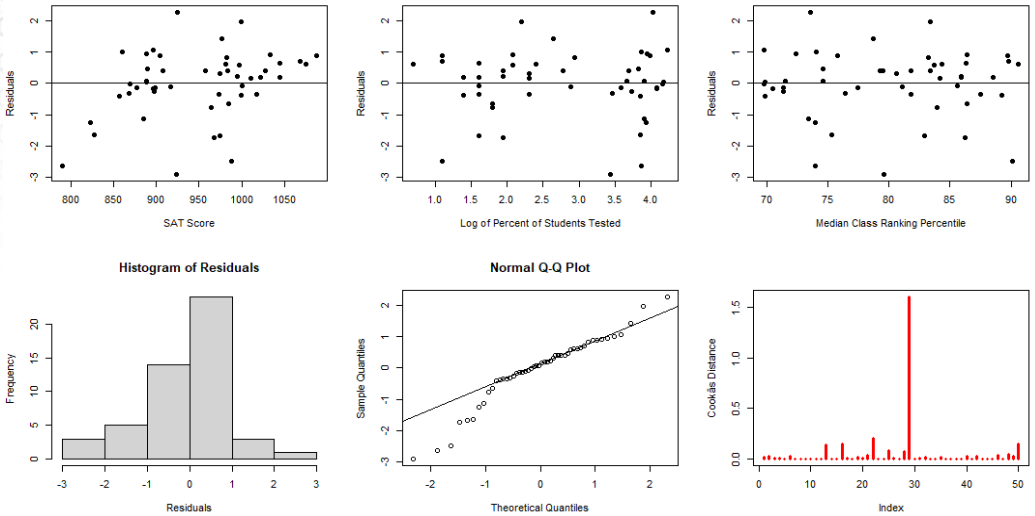








Residual Analysis



- Transformation has improved on the linearity assumption
- Heavy tailed residuals remain
- Cook's distance
  - Alaska is an outlier/influential point for the model

State SAT Performance: Findings

- Given all other predictors in the model:
  - Percent of students taking SAT from a public school and family income of test takers are not statistically significantly associated to SAT score
  - A $100,000 increase in the expenditure on secondary schools results in a 2.56-point increase in the SAT score
  - One additional year that test takers had in social sciences, natural sciences, and humanities leads to a 17.2-point increase in the SAT score
- The predictors in the model explain close to 90% of the variability in SAT score
- We find that the relationship between state average SAT score and the percent of students taking SAT to be nonlinear
- Ranking changes after controlling for the bias selection factors
  - For example, Connecticut moves from 35th to 1st, Massachusetts from 41st to 4th, and New York from 36th to 5th

Georgia Tech

Summary

Georgia Tech