Exploratory Data Analysis and Modeling with R

An Analysis of Access and Quality of Healthcare in India

Karthik Sriram
Indian Institute of Management Ahmedabad
karthiks@iima.ac.in

Outline

1 Introduction

2 Data Visualization and Summaries
- Exploring univariate data
- Exploring association between variables

3 Fitting Linear Equations
- Formulation and Estimation
- Model Assumptions
- Statistical Inference

The Healthcare Sector

- Goods and services to treat patients

  Treatments: curative, preventive, rehablitative and palliative.

- Consists of Hospitals, Medical / Dental pratitioner and other activities: pathology labs, physiotherapy etc.

- Work force as per WHO[1]: 9.2 million physicians, 1.9 million dentists , 2.6 million pharmacists, 1.3 million community health workers.

- Delivery is usually face to face, but nowadays phone, text message, video etc.

The Healthcare Sector in India

- Expected to grow at CAGR of 17% during 2011-2020 to touch US$ 280 billion.
- Drivers for demand: Rising income levels, ageing population, growing health awareness and changing attitude towards preventive healthcare.
- The private sector's share in healthcare delivery is expected to increase from 66 per cent in 2005 to 81 per cent by 2015

  ....accounts for almost 72 per cent of the country's total healthcare expenditure.

  Ref: `http://www.ibef.org/industry/healthcare-india.aspx`

Analytics Broadly

- Data Management: Collect clean and timely data.
- Data Mining and Modeling: Understand data, decipher patterns and associations.
- Drive decision making and practice.

Healthcare Analytics

Real Time Analytics:
Analyze clinical information at the point of care and support health providers as they make prescriptive decisions. Use of patient data to generate case-specific advice.

Batch Analytics:
Retrospectively evaluate population data (i.e. records of patients in a large medical system, or claims data from an insured population) and supplement disease management or population health management efforts.

Learning Objectives

We will learn some useful statistical techniques by using the R software to analyze a real dataset on the Indian health care system.

1. We obtain data from `https://data.gov.in`
2. We will learn some useful data visualizations.
3. We will learn some basic statistical summaries and interpretations.
4. We will learn how to model associations among variables.

Data Analysis Background

The Government of India's National Health Assurance Mission aims to provide all citizens with free drugs and diagnostic treatment, as well as insurance cover to treat serious ailments, by 2020.

While one aspect about the mission will be to provide easy access, the other aspect must be that of quality care.

We wish to analyze the current state of the Indian health system with respect to these aspects.

Our objective

We want to carry out an exploratory analysis to understand some inter-relationshipthat exist between Quality, Access and Utilization of health care. Our interest is in the following questions...

- Is extent of access to healthcare related at all to quality of health ?
- How does Utilization of the facilities plan in ?
- Is it just access or does it matter what kind of facility one has access to ?
- Can we establish an equation relating these variables?

Data

We compiled data from the District Level Household and Facility Survey (DLHS) from the data portal of Government of India.

- Row identificaton is by name of **State** and **District**.
- Variables indicating **Quality**.
- Variables indicating **Access** to healthcare.
- Variables indicating **Type of Access** to healthcare.
- Variables indicating **Utilization** of healthcare.

A Remark on Survey Data

Survey Sampling versus Census

Data obtained from this sample survey is collected from a randomly chosen set of households from various districts. Random sampling is a way to ensure that such a chosen set is representative of the entire population of the country. The number of individuals surveyed is planned with the intention of keeping sampling errors within acceptable limits.

While a Census covers entire population, it is not always fasible to conduct under the given time or monetary resource constraints. Even if conducted, difficult to ensure desirable accuracy of collected data. Inaccuracy and cinonsistency of data collection lead to "Non-Sampling Errors" which when aggregated over a large population can give a completely wrong picture of the population.

R set up

Install Libraries

```
#install.packages(dependencies=TRUE, "colorspace")
#install.packages(dependencies=TRUE, "ggplot2")
#install.packages(dependencies=TRUE,"reshape2")
#install.packages(dependencies=TRUE,"gridExtra")
#install.packages(dependencies=TRUE,"data.table")

require(colorspace)

## Warning:  package 'colorspace' was built under R version 3.3.3

require(ggplot2)

## Warning:  package 'ggplot2' was built under R version 3.3.3

require(reshape2)

## Warning:  package 'reshape2' was built under R version 3.3.3

require(data.table)

## Warning:  package 'data.table' was built under R version 3.3.1

require(gridExtra)
```

Import Data

Set Working Directory

```
setwd('C:/Users/pc1/Google Drive/EPAPB/RSessions')
```

Import State Level Data

```
# Data at State Level
QUAP_state1<- read.csv("QUAP_state1.csv")[,-1]
```

Import District Level Data

```
# Data at District Level
QUAP_dist1<- read.csv("QUAP_dist1.csv")[,-1]
```

What's in the data ? (ctd..)

Rows are identified by **State and District**

```
colnames(QUAP_dist1)[1:2]

## [1] "state"    "district"
```

Variables (i.e. Columns) indicating **Quality** of healthcare

```
colnames(QUAP_dist1)[3:8]

## [1] "Q_prevalence_RespDisease"       "Q_prevalence_cardiovascDisease"
## [3] "Q_prevalence_TB"                "Q_prevalence_AnyInjury"
## [5] "Q_prevalence_AcuteIllness"      "Q_prevalence_ChronicIllness"
```

What's in the data ? (ctd..)

Variables indicating **Access** to healthcare

```
colnames(QUAP_dist1)[18:20]

## [1] "A_accesstoSC3km"         "A_accessPHC10km"
## [3] "A_facilitiesper10000pop"
```

Variables (i.e. Columns) indicating **Type of Access** to healthcare

```
colnames(QUAP_dist1)[12:17]

## [1] "TA_SCinGovBuild"   "TA_LadyMedOfficer" "TA_PHC4beds"
## [4] "TA_Func24hrs"      "TA_newborncare"    "TA_operationTh"
```

What's in the data? (ctd..)

Variables indicating **Utlization** of healthcare

```
colnames(QUAP_dist1)[9:11]

## [1] "U_Anenatalcare"                 "U_pregnancycomplication"
## [3] "U_postdeliverycomplication"
```

Analysis of Quality

There are many variables in the data indirectly representing quality of health. For our purpose we will consider the following indicator:

Quality= "100 minus percentage of people with chronic illness in the region".

In other words, Quality is the "Non-prevalence" rate of chronic illness.

State level Data

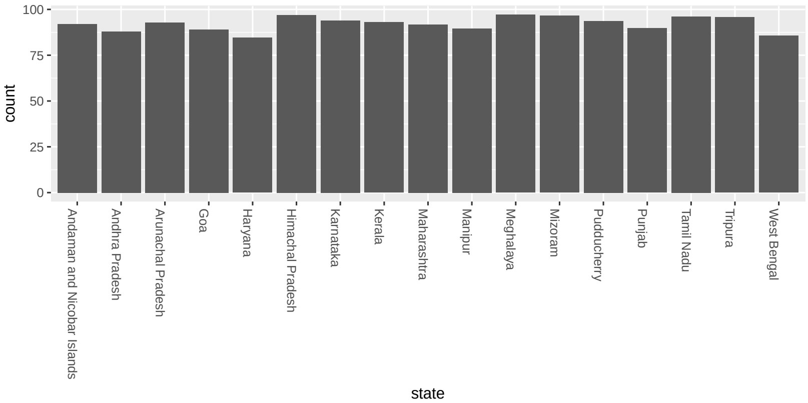
```
Newdata_st<-data.table(state= QUAP_state1$state,
          Quality= (100-QUAP_state1$Q_prevalence_ChronicIllness))
```

District Level Data

```
Newdata_dt<-data.table(state= QUAP_dist1$state,
            district=QUAP_dist1$district,
            Quality= (100-QUAP_dist1$Q_prevalence_ChronicIllness))
```

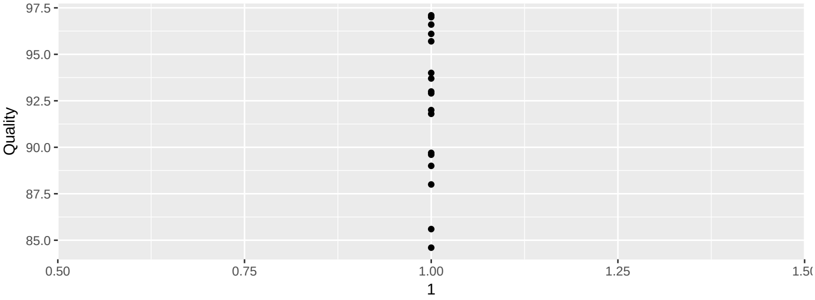
Bar Chart

```
p1<-ggplot(Newdata_st, aes(state,weight=Quality))+ geom_bar()
p1+ theme(axis.text.x=element_text(angle = -90, hjust = 0))
```



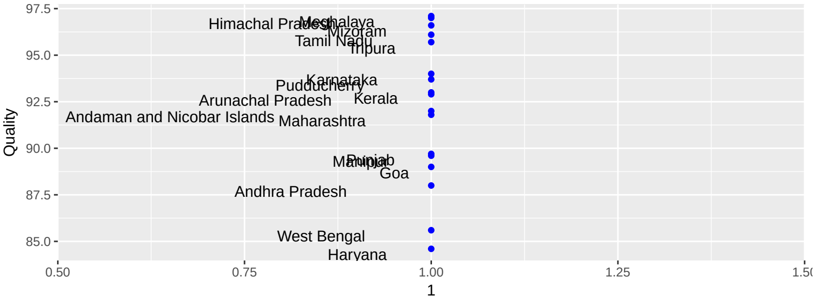
Dot Plot for Quality

```
p1<-ggplot(Newdata_st, aes(1,Quality, label=state))+ geom_point()
p1
```



Dot Plot for Quality with Labels

```
p1<-ggplot(Newdata_st, aes(1,Quality, label=state))+ geom_point(color="blue")
p1+geom_text(size=4, hjust=1.75, vjust=1)
```



Good : Meghalaya, Himachal Pradesh, Mizoram Bad : West Bengal, Haryana

Mean

Mean ($\bar{x}$)

For data points $x_1, x_2, \ldots, x_n$

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

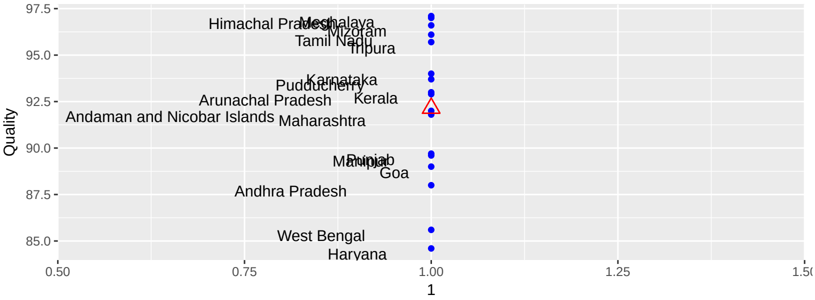
This is a measure of "Central tendency".

```
mm<-mean(Newdata_st$Quality)
mm

## [1] 92.14118
```

Mean Quality across states, i.e. Average non-prevalence of chronic illness (in %) is 92.14

Dot Plot + Mean

```
p1<-ggplot(Newdata_st, aes(1,Quality, label=state))+
  geom_point(color="blue")
p1+geom_text(size=4, hjust=1.75, vjust=1) +
geom_point(aes(x=1, y= mean(Quality)),
           color="red", shape=2, size=4)
```



Avg Quality: Kerala, Maharashtra, Arunachal, Andaman & Nicobar.

Standard Deviation

Clearly, the average value does not speak to all states. There is much variability in quality across states.

Standard Deviation (sd)

For data points $x_1, x_2, \ldots, x_n$

$$sd = s_x = \sqrt{\frac{1}{n-1}\left\{(x_1 - mean)^2 + (x_2 - mean)^2 + \ldots + (x_n - mean)^2\right\}}$$

This is a measure of "Dispersion" or "Spread".

```
ssd<-sd(Newdata_st$Quality)
ssd

## [1] 3.895841
```

SD of Quality in our data across states (in %) is 3.9

Quantiles and Inter Quartile Range

e.g. 75th Quantile = Value below which 75% of the observations lie.

Quantiles (also Percentiles) and IQR

In general, for $0 < p < 1$

Q_{100p}= value below which there are $100p$ observations.

Q_{50} = Median, i.e. Half of the value are below this number.

Q_{25} is called First Quartile

Q_{75} is called Third Quartile

Inter Quartile Range $= Q_{75} - Q_{25}$

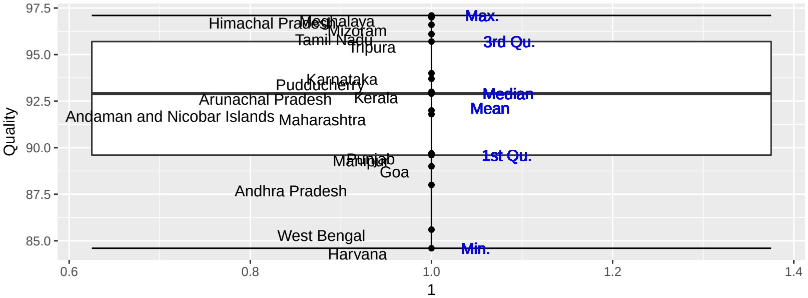
Quantile summary from data

```
summary(Newdata_st$Quality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   84.60   89.60   92.90   92.14   95.70   97.10
```

Box and Whiskers Plot

Pictorically the Quantile Summary can be shown using a Box Plot

```
p1<-ggplot(Newdata_st, aes(1,Quality, label=state))+
  geom_boxplot() +stat_boxplot(geom='errorbar')
p1<-p1+geom_point()+geom_text(size=4, hjust=1.75, vjust=1)
```



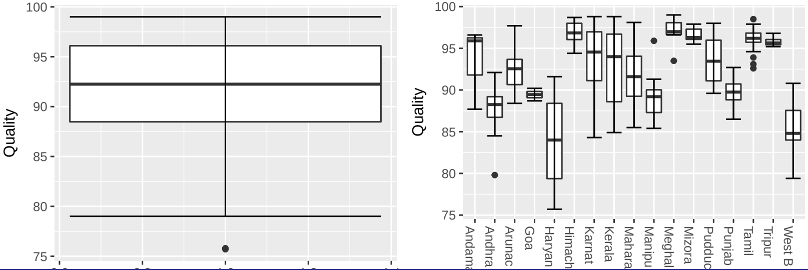
Whiskers

$Q_{25} - 1.5 \times IQR, \; Q_{75} + 1.5 \times IQR.$

Box Plot within states

So far, we looked at state level aggregate data. More insight on variation can be obtained by looking at district level data.

```
p1<-ggplot(Newdata_dt, aes(substr(state,1,6),Quality))+
  geom_boxplot()+stat_boxplot(geom='errorbar')
p1<-p1+theme(axis.text.x=
    element_text(angle = -90, hjust = 0))

p2<-ggplot(Newdata_dt, aes(1,Quality))+
geom_boxplot()+stat_boxplot(geom='errorbar')

grid.arrange(p2,p1, ncol=2)
```

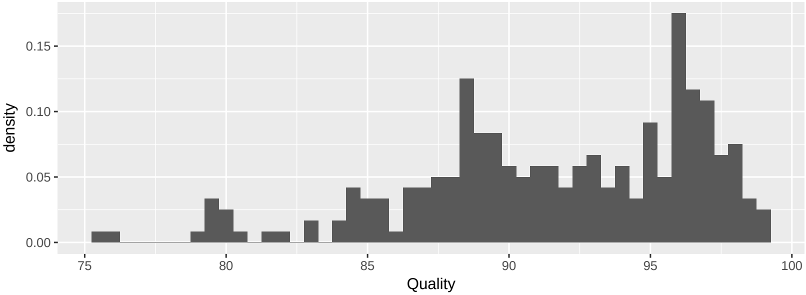


Observations from Box plots

Histogram

Histogram is a (relative) frequency distribution of the data. Higher bars means data in that range occurs more frequently.

```
p1<-ggplot(Newdata_dt, aes(Quality))
p1<-p1+ geom_histogram(aes(y=..density..), binwidth=.5)
p1
```

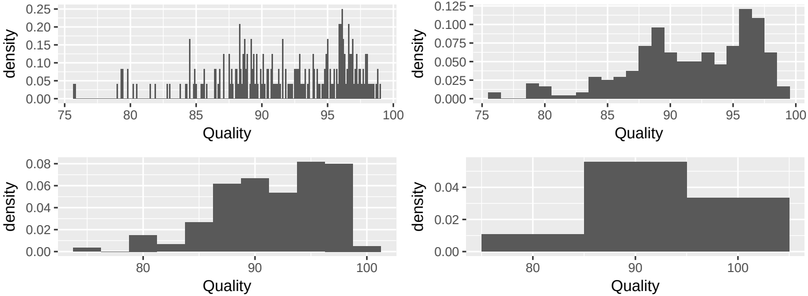


Some observations from histogram

- Bimodal
- Districts are likely to be around 87% or around 97% non-prevalence rate.
- Distribution is Skewed

Histogram (different bin widths)

Histogram depends on choice of bin width. We don't want the bin to be too large or too small.



Freedman-Diaconis Bandwidth

Freedman and Diacons proposed optimal bandwidth for the a histogram for data X

$$bw = \frac{2(Q_3 - Q_1)}{(length(X))^{\frac{1}{3}}}$$

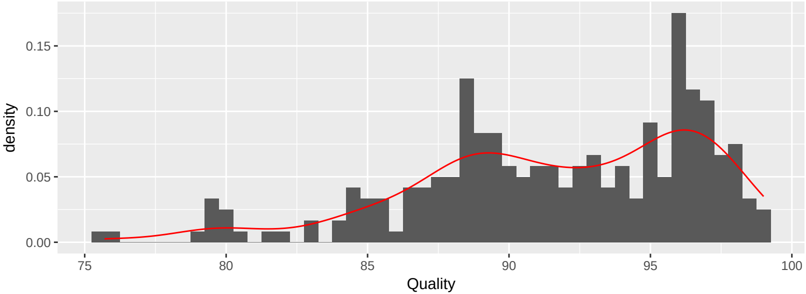
```
x<-Newdata_dt$Quality
bw <- (2 * IQR(x) / length(x)^(1/3))
popt<-ggplot(Newdata_dt, aes(Quality))
popt<-popt+ geom_histogram(aes(y=..density..), binwidth=bw)
```

Kernel Density

Kernel Density Estimate

Smooth version of a histogram. It can help remove some of the noise in the regular histogram and decipher prominent modes in the data.

```
p1<-ggplot(Newdata_dt, aes(Quality))
p1<-p1+ geom_histogram(aes(y=..density..), binwidth=.5)
p1+stat_density(kernel="gaussian", geom="line", color="red")
```



Data Variables

For our discussion, we will explore relationships between Quality, Access, Type of Access and Utlization of healthcare, by defining the variables as:

- Quailty = 100-% reporting chronic illness
- Utilization= % using antenatal care
- Access = % facilities per 10000 population
- Type Access=% health care centers open 24 hours

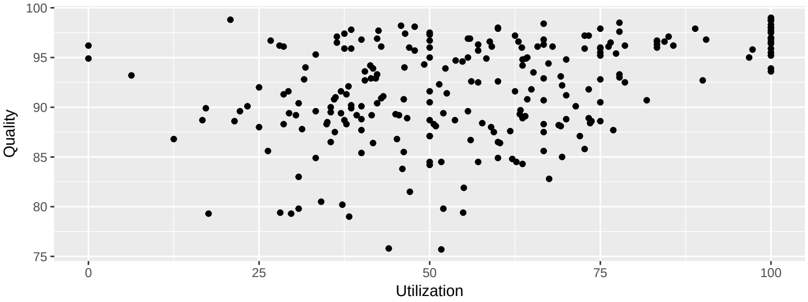
```
Newdata_dt$Utilization=QUAP_dist1$U_postdeliverycomplication
Newdata_dt$Access=QUAP_dist1$A_facilitiesper10000pop
Newdata_dt$Type_Access=QUAP_dist1$TA_Func24hrs
```

Summary of the variables

```
V_Q<-Newdata_dt$Quality*1
V_U<-Newdata_dt$Utilization*1
V_A<-Newdata_dt$Access*1
V_TA<-Newdata_dt$Type_Access*1
summary(cbind(V_Q, V_U, V_A, V_TA))

##       V_Q             V_U              V_A              V_TA
##  Min.   :75.70   Min.   :  0.00   Min.   :0.6273   Min.   :10.30
##  1st Qu.:88.47   1st Qu.: 40.00   1st Qu.:1.2543   1st Qu.:33.55
##  Median :92.25   Median : 54.95   Median :1.4060   Median :47.00
##  Mean   :91.64   Mean   : 56.06   Mean   :1.7353   Mean   :46.51
##  3rd Qu.:96.10   3rd Qu.: 70.00   3rd Qu.:1.8305   3rd Qu.:62.55
##  Max.   :99.00   Max.   :100.00   Max.   :3.9148   Max.   :93.30
```

Scatter Plot: Quality vs. Utilization

```
p2<- ggplot(Newdata_dt, aes(y=Quality, x=Utilization))
p2+geom_point()
```

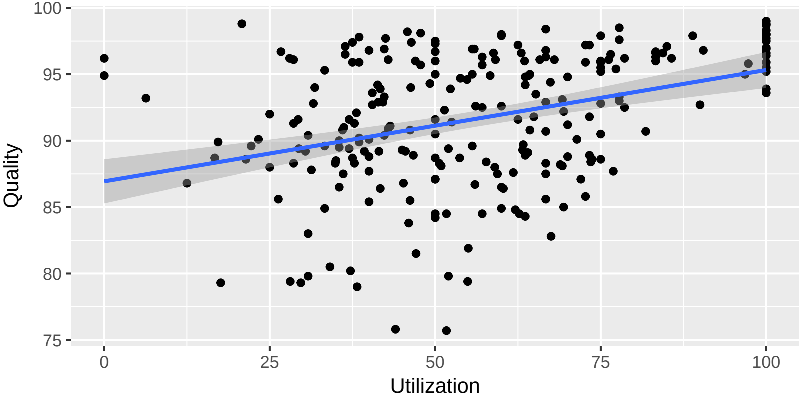


Some observations on scatter plot

- As expected, Quality and Utilization seem to be positively related
- Variation across districts with higher Quality is lesser than variation in quality across districts with lower quality.

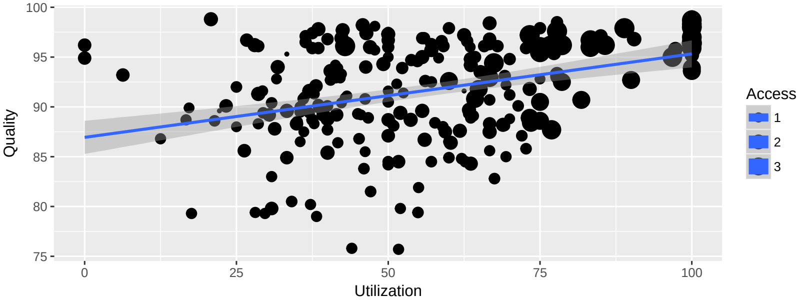
Scatter Plot: Quality vs. Utilization + Line

```
p2<- ggplot(Newdata_dt, aes(y=Quality, x=Utilization))
p2+geom_point()+geom_smooth(method="lm")
```



Scatter Plot: Quality vs. Utilization (size=Access)

```
p2<- ggplot(Newdata_dt, aes(y=Quality,
                   size=Access, x=Utilization))
p2+geom_point()+geom_smooth(method="lm")
```



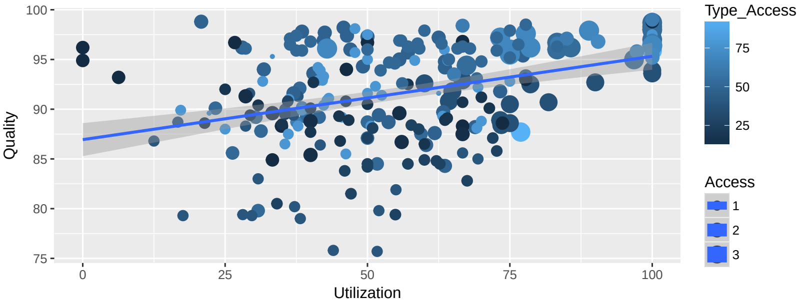
Some observations

- Higher Quality seems to be associated with Better Access
- More number of smaller sized dots below the line than above.

  Some lower quality points at high utilization may be attributable to lower Access levels.

Scatter Plot: Quality vs. Utilization (size=Access, color=Type of Access)

```
p2<- ggplot(Newdata_dt, aes(y=Quality,
  size=Access, x=Utilization, color=Type_Access))
p2+geom_point()+geom_smooth(method="lm")
```



Some observations

- Larger number of light colored smaller dots above the line versus below indicating that type of access matters even if access is limited.

Measuring the strength of association

We want to establish a quantitative relationship between two given variables x and y.

Correlation

Measures the degree of linear association between two variables x and y.

It is a number between -1 and 1. For perfect positive linear relationship it is 1 and for a perfect negative linear relationship, it is -1. It is =0 if there is "no linear relationship"

$$R = Correlation(y, x) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

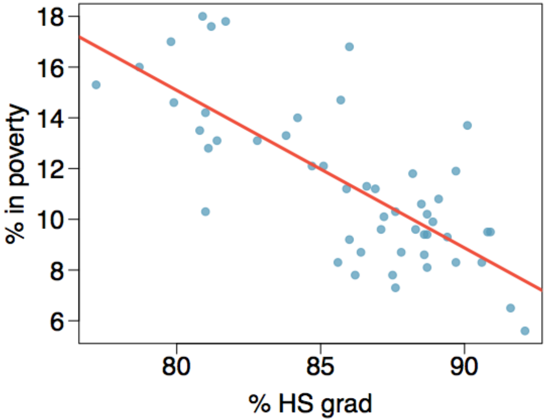
Let us learn a bit about correlation.

Next few slides are borrowed from `www.openintro.org`, which is a free and useful online resource for learning statistics.

Guessing the correlation

Which of the following is the best guess for the correlation between percent in poverty and percent HS grad?

(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between percent in poverty and percent HS grad?

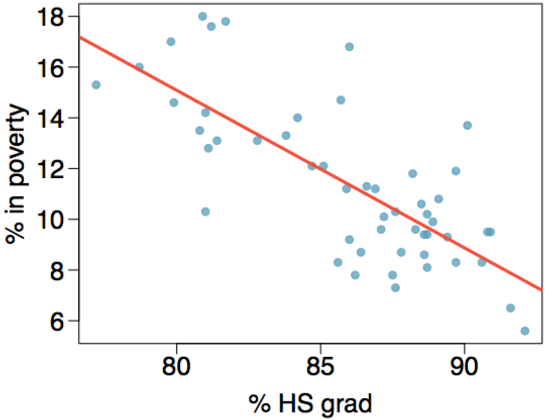
(a) 0.6

(b) -0.75

(c) -0.1

(d) 0.02

(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between percent in poverty and percent female householder?

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5

Guessing the correlation

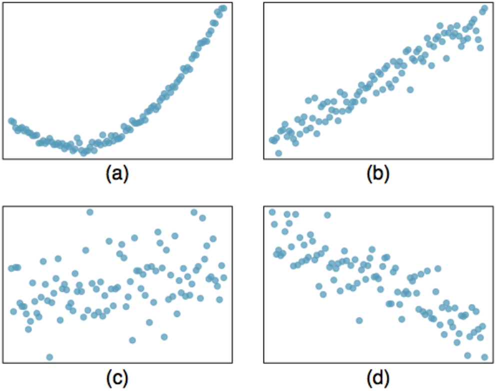
Which of the following is the best guess for the correlation between percent in poverty and percent female householder?

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5

Assessing the correlation

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?

(b) → correlation means linear association



End of general introduction to correlation.

Last few slides were borrowed from `www.openintro.org`, which is a free and useful online resource for learning statistics.

Let us apply this on our data

Correlation between variables in our data

```
V_Q<-Newdata_dt$Quality*1
V_U<-Newdata_dt$Utilization*1
V_A<-Newdata_dt$Access*1
V_TA<-Newdata_dt$Type_Access*1
round(cor(cbind(V_Q, V_U, V_A, V_TA)),2)

##       V_Q  V_U  V_A V_TA
## V_Q  1.00 0.36 0.34 0.30
## V_U  0.36 1.00 0.55 0.18
## V_A  0.34 0.55 1.00 0.07
## V_TA 0.30 0.18 0.07 1.00
```

Some observations

- Quality has similar correlation with Access, Utilization and Type of access
- None of Access, Type of Access and Utilization are perfectly correlated with each other. Given th eprevious observation, this means they will complement each other in explaining quality.
- Access and Utilization have a slightly higher correlation.

We now want to derive a relationship by considering all these variables together..

Simple Linear Regresssion

Here, the objective is to build a linear relationship between

(i) "dependent variable" Y (also called Response)

and

(ii) "independent variable" X (also called explanatory variable or predictor).

Model Formulation

The model formulation is as follows:

$$Y = \beta_0 + \beta_1 X + Error$$

Model Estimation Approach

Method of Least Squares

Obtain data on Y : $y_1, y_2, \ldots, y_n$

Obtain data on X : $x_1, x_2, \ldots, x_n$

Estimate β_0 and β_1 by minimizing error sum of squares

$$\text{Error Sum of Squares} = \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

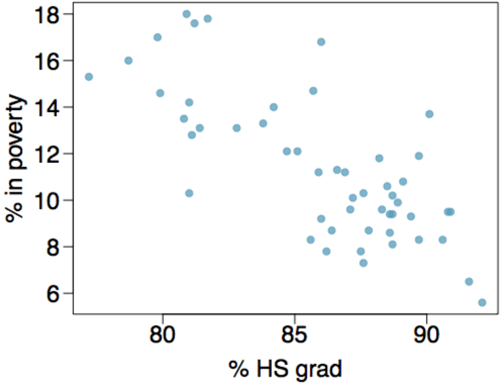
Estimate of $\beta_0 = b_0$
Estimate of $\beta_1 = b_1$

Let us further understand concepts of simple linear regression

Next few slides are borrowed from `www.openintro.org`, which is a free and useful online resource for learning statistics.

Poverty vs. HS graduate rate

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

Explanatory variable?

% HS grad

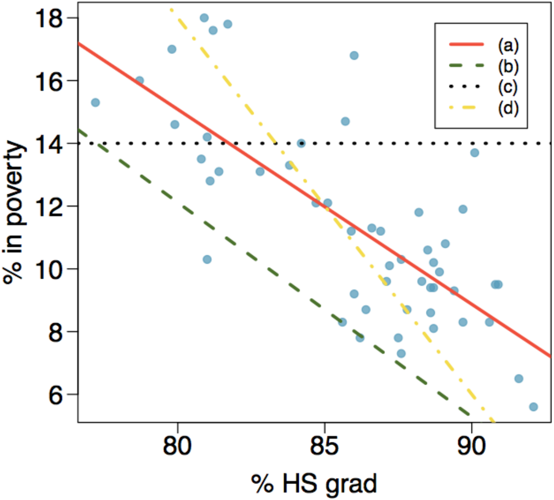
Relationship?

linear, negative,

moderately strong

Eyeballing the line

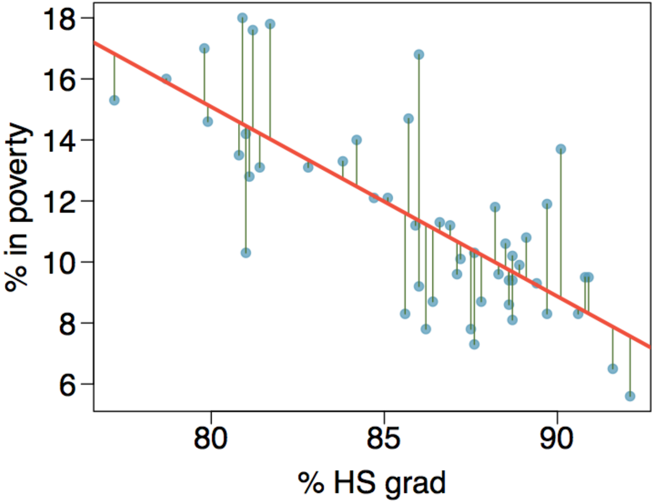
Which of the following appears to be the line that best fits the linear relationship between percent in poverty and percent HS grad? Choose one.



(a)

Residuals

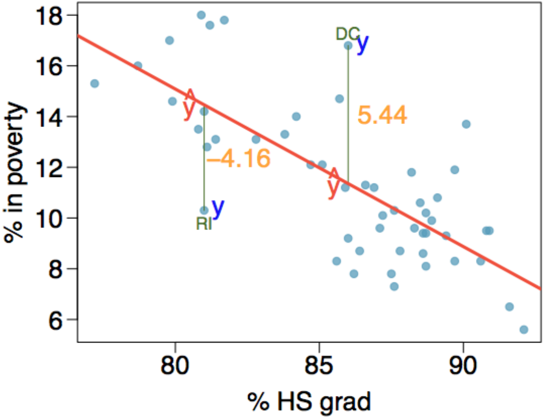
Residuals are the leftovers from the model fit: Data = Fit + Residual



Residuals (cont.)

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

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



Percent living in poverty in DC is 5.44% more than predicted.

Percent living in poverty in RI is 4.16% less than predicted.

A measure for the best line

We want a line that has small residuals

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

$$|e_1| + |e_2| + \dots + |e_n|$$

Option 2: Minimize the sum of squared residuals -- least squares

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

Why least squares?

- Most commonly used
- Easier to compute by hand and using software
- In many applications, a residual twice as large as another is usually more than twice as bad

The least squares line

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

predicted y

intercept

slope

explanatory variable

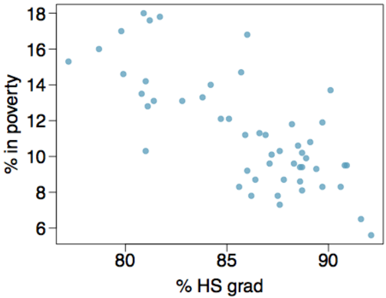
Intercept Notation

- Parameter: β_0
- Point estimate: b_0

Slope Notation

- Parameter: β_1
- Point estimate: b_1

Given...



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

Slope

The slope of the regression can be calculated as

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

In context...

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

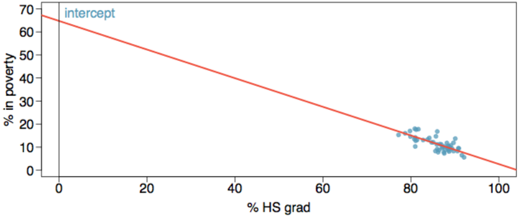
Interpretation

For each additional % point in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.

Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1\,\bar{x}$$



$$b_0 = 11.35 - (-0.62) \times 86.01$$
$$= 64.68$$

Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

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

(e) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

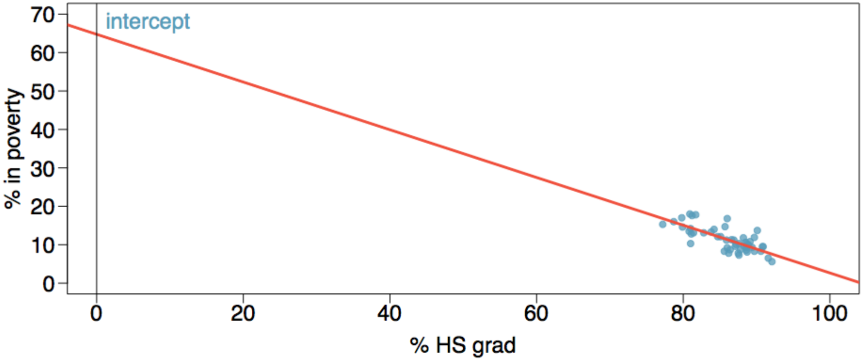
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

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

(e) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

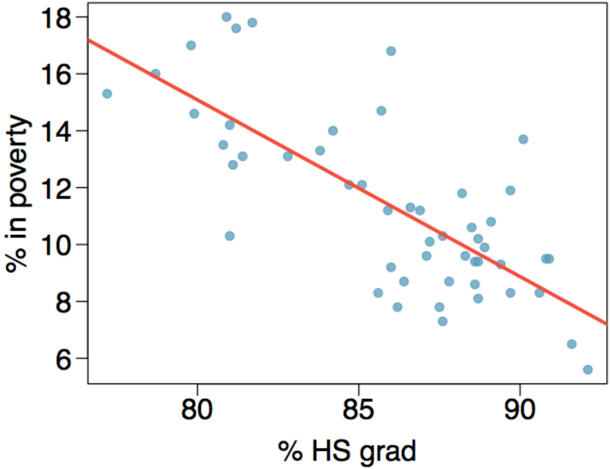
More on the intercept

Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.



Regression line

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



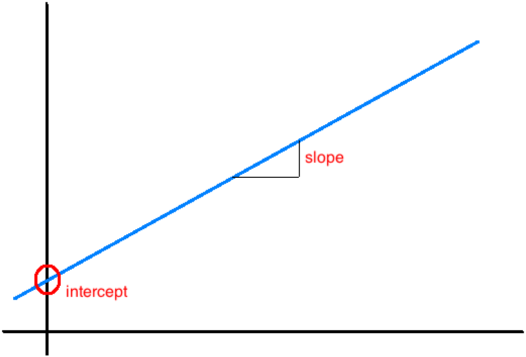
Interpretation of slope and intercept

Slope

For each unit in x, y is expected to increase / decrease on average by the slope.

Intercept

When x = 0, y is expected to equal the intercept.

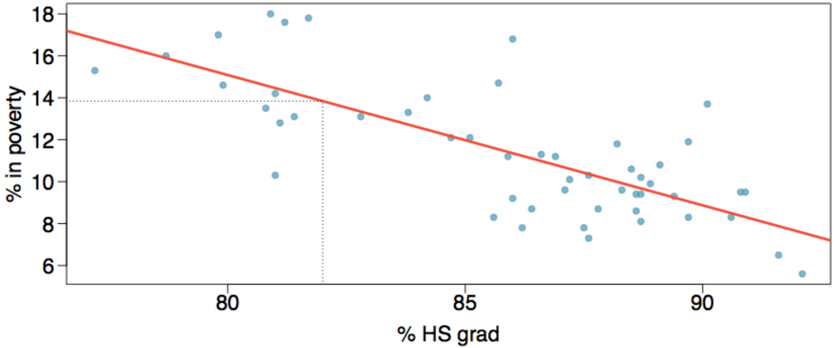


Note: These statements are not causal, unless the study is a randomized controlled experiment.

Prediction

Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called prediction, simply by plugging in the value of x in the linear model equation.

There will be some uncertainty associated with the predicted value.



End of general introduction to Estimation in Simple Linear Regression

Last few slides were borrowed from `www.openintro.org`, which is a free and useful online resource for learning statistics.

Let us apply this on our problem

Model Formulation and Estimation in our problem

We will try to model with

Response=Quality
Explanatory variable = Utilization

Model Formulation

Quality $= \beta_0 + \beta_1$ Utilization + error

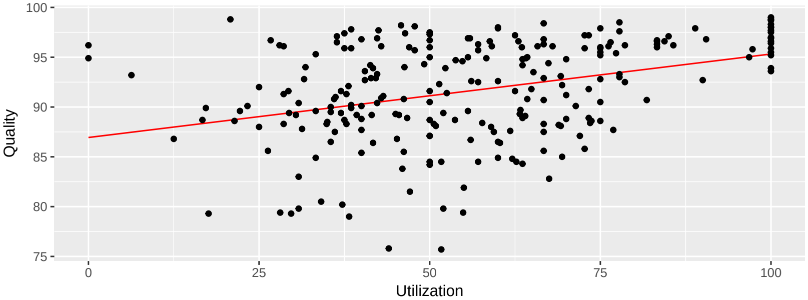
```
mod<-lm(Quality~ Utilization , data=Newdata_dt)
mod

##
## Call:
## lm(formula = Quality ~ Utilization, data = Newdata_dt)
##
## Coefficients:
## (Intercept)  Utilization
##     86.9396       0.0838
```

Estimated Model

$\hat{\text{Quality}} = 86.93 + .0838$ Utilization.

Fitted versus Actual

```
mod<-lm(Quality~ Utilization , data=Newdata_dt)
p1<-ggplot(mod, aes(x= Utilization , y=.fitted) )
p1<-p1+geom_line(color="red")
p1<-p1+geom_point(aes(x=Utilization, y=Quality))
p1+ylab("Quality")
```




```
#%+% fortify(mod, Newdata_dt)
```

Interpretations

- Slope= ? and interpretation = ?
- Intercept= ? and interpretation= ?

Interpretations

- What is the expected quality level at a utilization of 75%?

Interpretations

- Is the following statement True or False ? Why?
  "When utilization is 100% quality is 95%."

Interpretations

- Is the following statement True or False ? Why?

  "When utilization is 100% quality is 95%."

- Take Away : The Model is for predicting "Expected (or Mean) Quality" at a given level of utilization. The actual quality for any district can still be different from the Expected value!

- So far, we have done a mechanical estimation of the simple linear regression model.
- We will now delve more into the statistical side of the model.
- It's usefulness is dependent on key statistical assumptions.
- We will next try to understand these assumptions.

Let us understand the assumptions in linear regression.

Next few slides are borrowed from `www.openintro.org`, which is a free and useful online resource for learning statistics.

Conditions for the least squares line

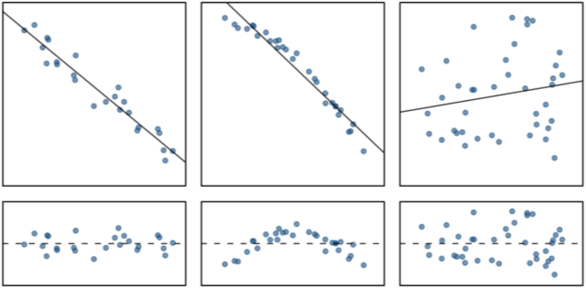
1. Linearity
2. Nearly normal residuals
3. Constant variability

Conditions: (1) Linearity

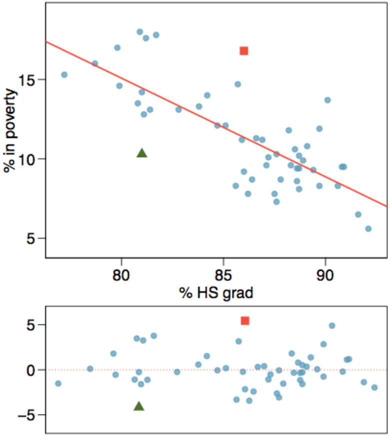
The relationship between the explanatory and the response variable should be linear.

Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.

Check using a scatterplot of the data, or a residuals plot.



Anatomy of a residuals plot



▲ *RI:*

$$\% \text{ HS grad} = 81 \qquad \% \text{ in poverty} = 10.3$$

$$\widehat{\% \text{ in poverty}} = 64.68 - 0.62 * 81 = 14.46$$

$$e = \% \text{ in poverty} - \widehat{\% \text{ in poverty}}$$

$$= 10.3 - 14.46 = -4.16$$

■ *DC:*

$$\% \text{ HS grad} = 86 \qquad \% \text{ in poverty} = 16.8$$

$$\widehat{\% \text{ in poverty}} = 64.68 - 0.62 * 86 = 11.36$$

$$e = \% \text{ in poverty} - \widehat{\% \text{ in poverty}}$$

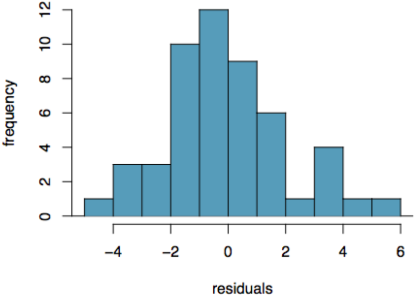
$$= 16.8 - 11.36 = 5.44$$

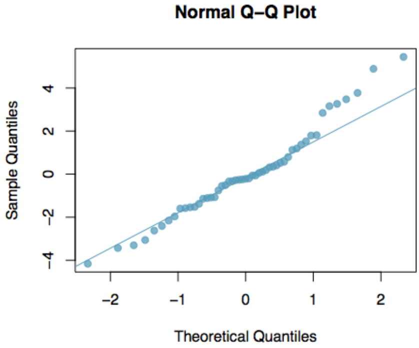
Conditions:
(2) Nearly normal residuals

The residuals should be nearly normal.

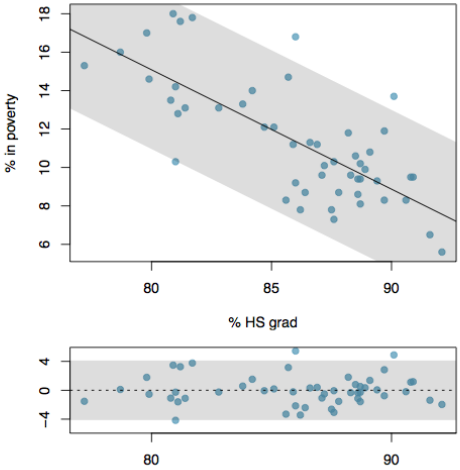
This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.

Check using a histogram or normal probability plot of residuals.





Conditions: (3) Constant variability



The variability of points around the least squares line should be roughly constant.

This implies that the variability of residuals around the 0 line should be roughly constant as well.

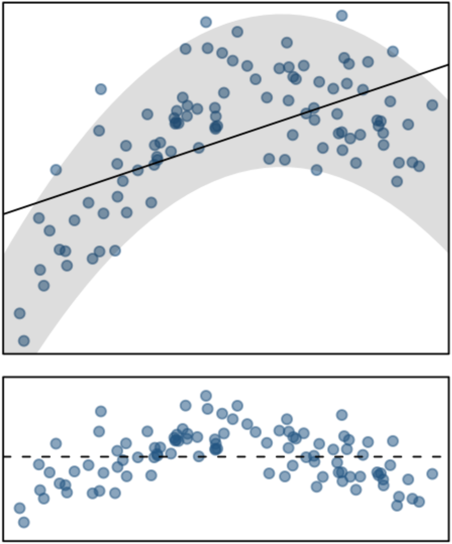
Also called homoscedasticity.

Check using a histogram or normal probability plot of residuals.

Checking conditions

What condition is this linear model obviously violating?

(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Checking conditions

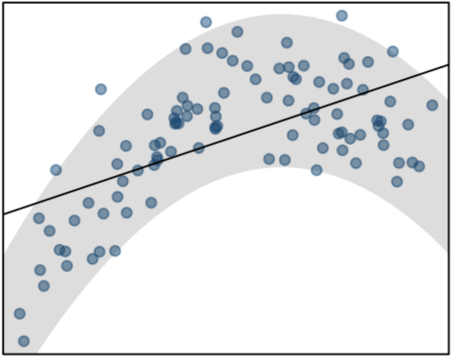
What condition is this linear model obviously violating?

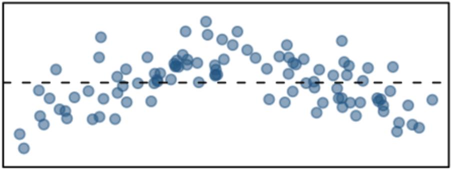
(a) Constant variability

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers

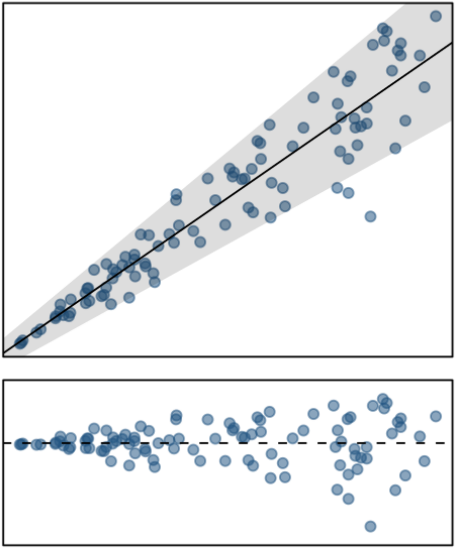




Checking conditions

What condition is this linear model obviously violating?

(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Checking conditions

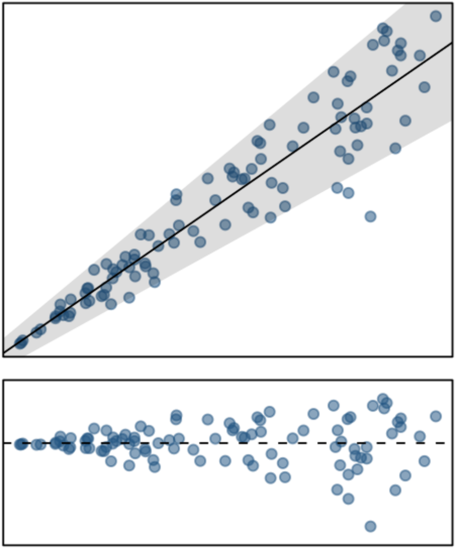
What condition is this linear model obviously violating?

(a) Constant variability

(b) Linear relationship

(c) Normal residuals

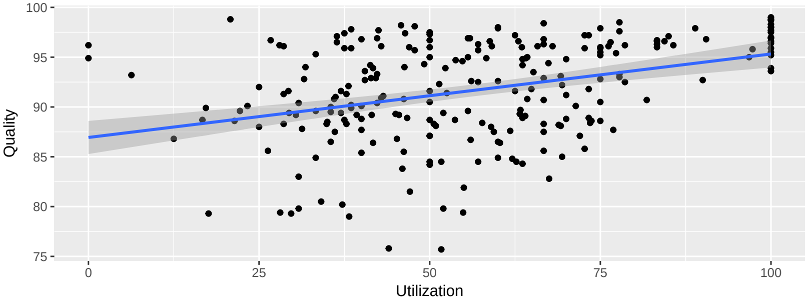
(d) No extreme outliers



End of general discussion on assumptions in linear regression.

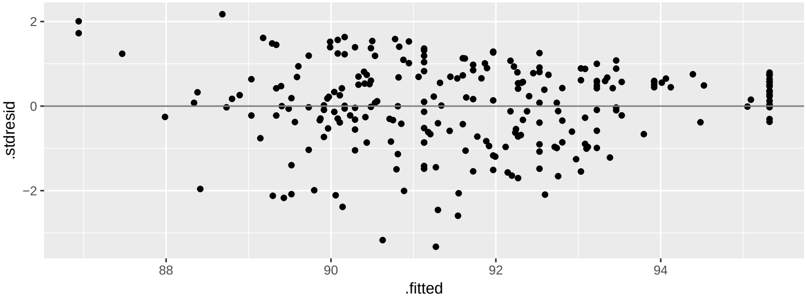
Last few slides were borrowed from `www.openintro.org`, which is a free and useful online resource for learning statistics.

Scatter plot



Residual Plot for our problem

```
mod<-lm(Quality~ Utilization , data=Newdata_dt)
p1<-ggplot(mod, aes(x= .fitted , y=.stdresid) )
p1<-p1+geom_point()
p1+geom_hline(yintercept=0, color="grey50", size=0.5)
```



Interpretations

- Do the points look randomly scattered around zero ?
- Do you see a patteren ?
- Is linearity assumption violated?
- Is constancy of variance assumption violated ?

Remark on standardized residuals

A technical point is that the residuals used for this analysis need to be standardized. The reason being that purely by the statistical property, the estimated residuals have non-constant variance. To look for non-constancy in variance purely due to model inadequacy, it is better to look at standardized resuals, i.e. residuals normalized by diving standard deviation.

How to handle violation of assumptions?

Typically,

- nonlinear relationship is handled by transforming x. e.g. $\log(x)$, x^2, $\sqrt{x}$ etc.
- non-constant variances is handled by transforming y, e.g. $\log(y)$, y^2, $\sqrt{y}$ etc.
- A general class of tranformations is the Box Cox tranformation

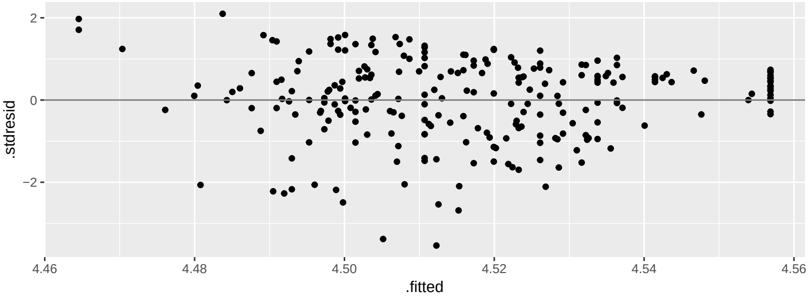
$$y_\lambda = \frac{y^\lambda - 1}{\lambda},\ \lambda \neq 0.$$

Residual Plot for our problem

Let us try to reformulate our model

$$log(Quality) = \beta_0 + \beta_1 \times Utilization + Error.$$

```
mod<-lm(log(Quality)~ Utilization, data=Newdata_dt)
p1<-ggplot(mod, aes(x= .fitted , y=.stdresid) )
p1<-p1+geom_point()
p1+geom_hline(yintercept=0, color="grey50", size=0.5)
```



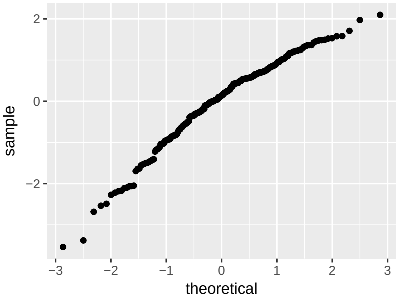
Checking Normality of Residuals

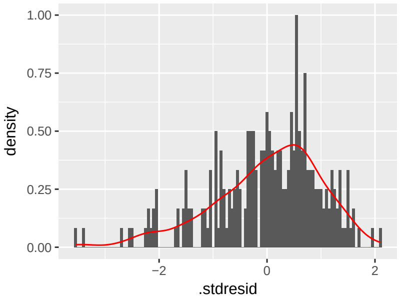
- Check Histogram: Does it resemble the normal distribution ?
- Check QQ plot : A Plot of Quantiles from normal distribution with quantiles of the (standardized) residuals.

....A straight line indicates proximity to normal distribution.

Histogram and QQ Plot Residuals

```
mod<-lm(log(Quality)~ Utilization, data=Newdata_dt)
p1<-ggplot(mod, aes(sample=.stdresid) )
p2<-p1+stat_qq()
p3<-ggplot(mod, aes(.stdresid))
p3<-p3+geom_histogram(aes(y=..density..), binwidth=.05)
p3<-p3+stat_density(kernel="gaussian", geom="line", color="red")
grid.arrange(p2,p3, ncol=2)
```





Some remarks

- The validity of assumptions are important to draw inferences on the model.
- Model building involves much back and forth analysis involving different transformations, variable formulations etc.

Standard Deviations of Estimates

- The data we used is from a survey. It is not the entire population.
- But the model we are building is to understand or infer the population.
- For this, we have built the model by using a representative sample from the population.
- Understandably the b_0 and b_1 will not be exactly equal to the true parameters β_0 and β_1 that corresponds to the entire population.
- There is uncertainty around b_0 and b_1 in inferring about β_0 and β_1
- This uncertainty is measured by standard deviations for b_0 and b_1.

Standard Deviations of Intercept and Slope

```
mod<-lm(log(Quality)~ Utilization, data=Newdata_dt)
summary(mod)$coefficients[,1:2]

##                  Estimate    Std. Error
## (Intercept) 4.4645383245 0.0094480539
## Utilization 0.0009233351 0.0001572969
```

Question:

Can we conclude here that the effect of utilization (slope) is close to 0 ?

The t-test

- The answer to the question depends on how certain we are about our estimate.
- If the sd is very large, then the true value of β_1 can very well be 0.
- If sd is very small, then it indicates that the effect is positive.
- How small is small ?

t Statistic

It is the standardized value of the parameter

$$t = \frac{b_1}{sd(b_1)}$$

If the regression model assumptions hold, then t is approximately distributed as standard normal for large sample sizes.

If the absolute value of t is large, we conclude the parameter is significantly different from 0.
One thumb rule for large enough is $t > 2$ or $t < -2$. Equivalently one can look at $p - value > .05$.

Checking for statistical significance in our model

```
mod<-lm(Quality~ Utilization, data=Newdata_dt)
summary(mod)$coefficients

##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 86.93960843 0.84387657 103.024081 2.019053e-199
## Utilization  0.08380296 0.01404936   5.964894  8.798294e-09
```

We can conclude that the effect of utilization is significantly different from 0 and positive.

Coefficient of Determination R^2

R^2

The percentage of variance in the response that is explained by the model.

Errors : $e_1 = y_1 - b_0 - b_1 x_1,\ e_2 = y_2 - b_0 - b_1 x_2,\ \ldots,\ e_n = y_n - b_0 - b_1 x_n.$

Error Sum of Squares (ESS)$= e_1^2 + e_2^2 + \ldots + e_n^2$

Total Sum of Squares (TSS)$= (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2.$

$R^2 = 1 - \frac{ESS}{TSS}.$

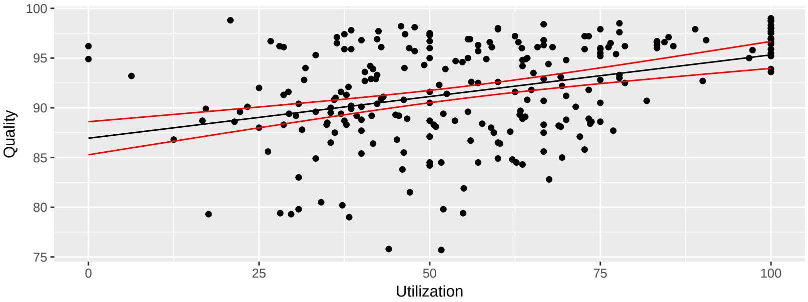
R^2 being close to 1 is one indication of a good model. In our problem, $R^2 = .13$. Hence, model explains 13% of the variation seen in the data.

```
Rsq<- 1-var(mod$resid)/var(Newdata_dt$Quality)
Rsq

## [1] 0.1300532
```

Confidence Interval

95% Confidence Interval for predicted value

This is the interval indicating where the true average value of the response might lie for a given value of the explanatory variable. The 95% indicates the confidence level. If the same survey were to be repeated many times, 95% of the time we would capture the true value.



Interval is shorter for center value on the utilization axis and widens at the extremes.

Confidence Interval (R code)

95% Confidence Interval for predicted value

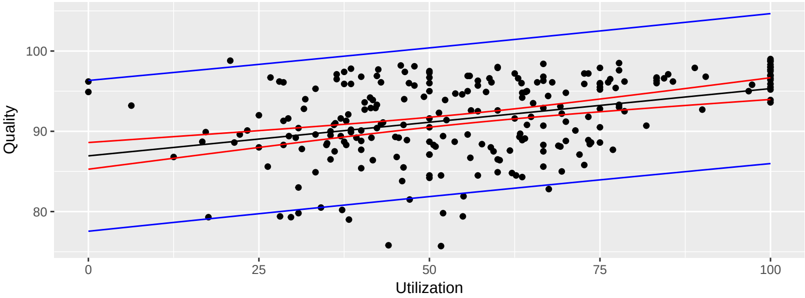
This is the interval indicating where the true average value of the response might lie for a given value of the explanatory variable. The 95% indicates the confidence level. If the same survey were to be repeated many times, 95% of the time we would capture the true value.

```
mod<-lm(Quality~ Utilization, data=Newdata_dt)

CI<-predict(mod, Newdata_dt, interval="confidence")

fitted<-predict(mod, Newdata_dt)

Newdata_dt1<-data.table(Newdata_dt, CI)

p1<-ggplot(Newdata_dt1, aes(x=Utilization, y=Quality))
p1<-p1+ geom_point()
p1<-p1+ geom_line(aes(y=fit))
p1<-p1+geom_line(aes(y=lwr), color="red")
p1<-p1+geom_line(aes(y=upr), color="red")
p1
```

Prediction Interval

95% Prediction Interval for y at a given level of x

This is the uncertainty interval indicating where to expect a new data point to lie for a given value of the explanatory variable. The 95% indicates the degree of confidence. If the survey is repeated many times, we would expect 95% of the data at a given level of the explanatory variable, to lie within this interval.



Prediction Interval (R code)

```
mod<-lm(Quality~ Utilization, data=Newdata_dt)

CI<-predict(mod, Newdata_dt, interval="confidence")
PredI<-predict(mod, Newdata_dt, interval="predict")
fitted<-predict(mod, Newdata_dt)

Newdata_dt1<-data.table(Newdata_dt, CI)

p1<-ggplot(Newdata_dt1, aes(x=Utilization, y=Quality))
p1<-p1+ geom_point()
p1<-p1+ geom_line(aes(y=fit))
p1<-p1+geom_line(aes(y=lwr), color="red")
p1<-p1+geom_line(aes(y=upr), color="red")

Newdata_dt2<-data.table(Newdata_dt, PredI)
p1<-p1+geom_line(data=Newdata_dt2,aes(y=lwr), color="blue")
p1<-p1+geom_line(data=Newdata_dt2,aes(y=upr), color="blue")
p1
```

Regression with multiple explanatory variables

Formulation

$$Quality = \beta_0 + \beta_1 Utilization + \beta_2 Access + \beta_3 TypeAccess + Error$$

```
mod<-lm(Quality~ Utilization+ Access + Type_Access, data=Newdata_dt)
summary(mod)
```

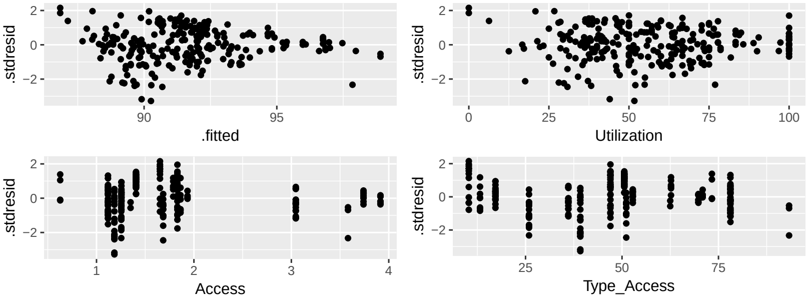
Regression with multiple explanatory variables

Formulation

$$Quality = \beta_0 + \beta_1 Utilization + \beta_2 Access + \beta_3 TypeAccess + Error$$

```
##
## Call:
## lm(formula = Quality ~ Utilization + Access + Type_Access, data = Newdata_dt
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5585  -2.5890   0.1761   3.0349   9.3615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.98511    0.98566  85.207  < 2e-16 ***
## Utilization  0.04562    0.01625   2.806  0.00543 **
## Access       1.36103    0.43208   3.150  0.00184 **
## Type_Access  0.05877    0.01378   4.266 2.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.461 on 236 degrees of freedom
## Multiple R-squared:  0.2203,Adjusted R-squared:  0.2104
## F-statistic: 22.23 on 3 and 236 DF,  p-value: 1.036e-12
```

Regression with multiple explanatory variables - Residual Plots



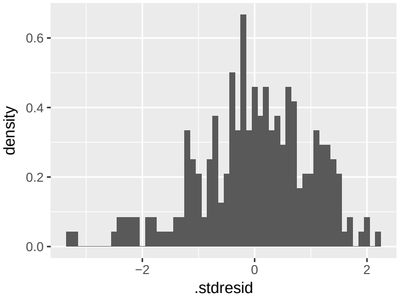
Regression with multiple explanatory variables - Residual Plots

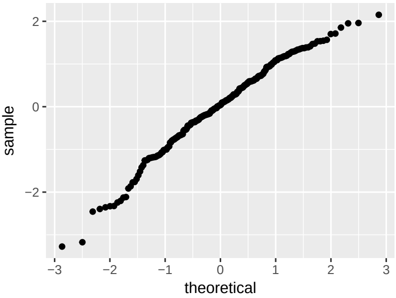
```
mod<-lm(Quality~ Utilization+ Access + Type_Access,
        data=Newdata_dt)
p1<- ggplot(mod, aes(.fitted, .stdresid))+geom_point()
p2<- ggplot(mod, aes(Utilization, .stdresid))+geom_point()
p3<- ggplot(mod, aes(Access, .stdresid))+geom_point()
p4<- ggplot(mod, aes(Type_Access, .stdresid))+geom_point()
grid.arrange(p1,p2,p3,p4)
```

Regression with multiple explanatory variables - QQ Plot

```
mod<-lm(Quality~ Utilization+ Access + Type_Access,
        data=Newdata_dt)
p1<- ggplot(mod, aes(.stdresid))+
  geom_histogram(aes(y=..density..), binwidth=.1)
p2<- ggplot(mod, aes(sample=.stdresid))+ stat_qq()
grid.arrange(p1,p2, ncol=2)
```

Regression with multiple explanatory variables - QQ Plot





Improving fit with more variables

One can introduce more variables

```
X<-as.matrix(QUAP_dist1[,c(9,10,11,12,13,
                    14,15,16,17,18,19,20)])*1
Y<-100-QUAP_dist1[,7]*1
mod<-lm(Y~X)
summary(mod)
```

Variable Selection

Here, the objective is to choose the best subset of variables that optimizes some measure of goodness of the model. R^2 is not a good measure because it always increases with more variables and does not penalize addition of too many variables. A common useful measure is 'Mallow's Cp'.

For a model with K variables,

$$C_p = \frac{ESS/(n-k)}{TSS/(n-1)} - n + 2K$$

Lower Cp value means better model.

Variable Selection

Cp values for the best model of different sizes

```
X<-as.matrix(QUAP_dist1[,c(9,10,11,12,13,14,
                           15,16,17,18,19,20)])*1
Y<-100-QUAP_dist1[,7]*1

#install.packages("leaps")

library(leaps)

## Warning:  package 'leaps' was built under R version 3.3.3

sub<-leaps(X,Y, method="Cp", nbest=1)

sub$Cp

##  [1] 136.230756 112.198462  77.914589  55.664933  26.307706  15.702648
##  [7]  13.222179   9.931041   8.076822   9.247290  11.101965  13.000000
```

Result based on Best model as per Cp

```
sub<-leaps(X,Y, method="Cp", nbest=1)
best_model_index<-which(sub$Cp==min(sub$Cp))
L<-which(sub$which[best_model_index,]==1)
X1<-X[,L]
mod<-lm(Y~X1[,-5])

summary(mod)
```

In conclusion

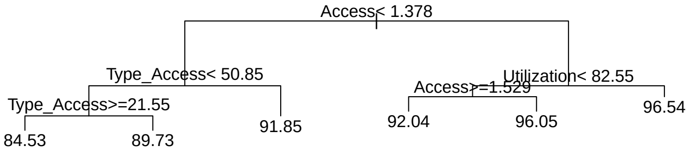
- Acess, Type of Access and Utilization seem to matter for Quality of Health.
- In our particular analysis, we could explain up to 50% of variation.
- More data on utilization could be beneficial. Currently available at state level from DLHS 3. Should be procurable at district level for DLHS 4.
- Model Validation is important. One idea is to see if the model stands test of time. e.g. DLHS 3 versus DLHS 4.

References

1. Data : https://data.gov.in
2. Reference: www.openintro.org
3. Reference Book: Author-Hadley Wickham (2009), Title-"ggplot2, Elegant Graphics for Data Analysis", Publisher-Springer.

A Glimpse of Other Methods

Classification and Regression Trees



Classification and Regression Trees

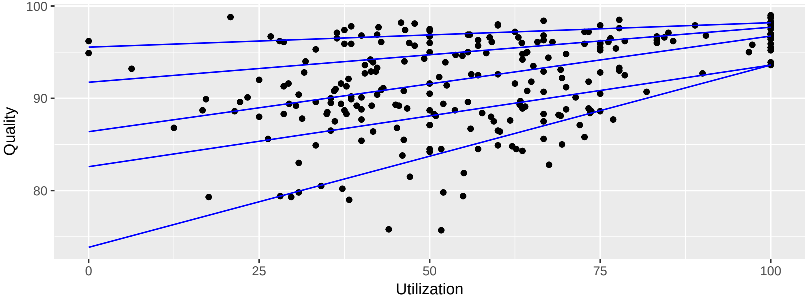
```
#install.packages("rpart")
library(rpart)
cptry<- .01
minn<-.05
fit<-rpart(Quality~ Utilization+Access+Type_Access,
    data=Newdata_dt, method="anova",
    control = rpart.control(cp = 0.05))
par(mfrow = c(1,1), xpd = NA)
plot(fit)
text(fit)

pred<-predict(fit)
R_sq<-sd(predict(fit))^2/sd(Newdata_dt$Quality)^2
R_sq
```

Quantile Regression

Earlier we modeled the "mean-Line". Instead, one could model quantiles.

e.g. Figure below shows the 10th, 25th, 50th (Median), 75th and 90th quantile lines.

```
## Warning:  package 'quantreg' was built under R version 3.3.2
## Warning:  package 'SparseM' was built under R version 3.3.2
```



Quantile Regression

Earlier we modeled the "mean-Line". Instead, one could model quantiles.

e.g. Figure below shows the 10th, 25th, 50th (Median), 75th and 90th quantile lines.

```
library(quantreg)
mod<-rq(Quality~Utilization, data=Newdata_dt, tau=c(.1,.25,.5,.75,.9))
Dat<-data.table(Q10=predict(mod)[,1], Q25=predict(mod)[,2],
                Q50=predict(mod)[,3],Q75=predict(mod)[,4],
                Q90=predict(mod)[,5])

Dat1<-cbind(Newdata_dt,Dat)
p1<-ggplot(Dat1, aes(x=Utilization, y=Quality))
p1<-p1+geom_point()
p1<-p1+geom_line(aes(y=Q10), color="blue")
p1<-p1+geom_line(aes(y=Q25), color="blue")
p1<-p1+geom_line(aes(y=Q50), color="blue")
p1<-p1+geom_line(aes(y=Q75), color="blue")
p1<-p1+geom_line(aes(y=Q90), color="blue")
p1
```