Statistical Methods for Data Science
Mini Project 6 **Charu Arora**
Cxa150730

1. Using adjusted R-square, we choose another variable to add to the final model

```
> sysBP <- read.csv(file="/Users/charuarora/Downloads/prostate_cancer.csv",
header=TRUE,sep=",");

> fit1 <- lm(log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$weight)
> summary(fit1)
```

```
Call:
lm(formula = log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$weight)

Residuals:
   Min      1Q  Median      3Q     Max 
-1.6425 -0.4095  0.0769  0.5119  1.9086 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.369545   0.144288   9.492 2.21e-15 ***
log(sysBP$cancervol) 0.718260   0.067477  10.644  < 2e-16 ***
sysBP$weight         0.003071   0.001740   1.765   0.0808 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7792 on 94 degrees of freedom
Multiple R-squared:  0.5533,  Adjusted R-squared:  0.5438 
F-statistic: 58.21 on 2 and 94 DF,  p-value: < 2.2e-16
```

ADJUSTED R-SQUARED: 0.5438

```
> fit2 <- lm(log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$age)
> summary(fit2)
```

```
Call:
lm(formula = log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$age)

Residuals:
    Min       1Q   Median       3Q      Max 
-1.67696 -0.42084  0.09667  0.50971  1.90063 

Coefficients:
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.4609508  0.7010997   2.084   0.0399 *
log(sysBP$cancervol) 0.7171620  0.0703897  10.188   <2e-16 ***
sysBP$age           0.0007794  0.0111432   0.070   0.9444
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.792 on 94 degrees of freedom
Multiple R-squared:  0.5385,  Adjusted R-squared:  0.5287
F-statistic: 54.84 on 2 and 94 DF,  p-value: < 2.2e-16
```

ADJUSTED R-SQUARED: 0.5287

```
> fit3 <- lm(log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$benpros)
> summary(fit3)
```

```
Call:
lm(formula = log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$benpros)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54494 -0.48609  0.06774  0.52572  1.90286

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.37891    0.13599  10.140   <2e-16 ***
log(sysBP$cancervol)  0.71496    0.06714  10.649   <2e-16 ***
sysBP$benpros        0.05318    0.02611   2.037   0.0445 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7751 on 94 degrees of freedom
Multiple R-squared:  0.558,    Adjusted R-squared:  0.5486
F-statistic: 59.33 on 2 and 94 DF,  p-value: < 2.2e-16
```

ADJUSTED R-SQUARED: 0.5486

```
> fit4 <- lm(log(sysBP$psa) ~ log(sysBP$cancervol) + factor(sysBP$vesinv))
> summary(fit4)
```

```
Call:
lm(formula = log(sysBP$psa) ~ log(sysBP$cancervol) + factor(sysBP$vesinv))

Residuals:
    Min      1Q  Median      3Q     Max
-1.6217 -0.5281  0.1209  0.4840  1.6907
```

```
        Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
        (Intercept)           1.53531    0.11733  13.086  < 2e-16 ***
        log(sysBP$cancervol)   0.59118    0.07767   7.611 2.07e-11 ***
        factor(sysBP$vesinv)1  0.67187    0.22113   3.038  0.00308 **
        ---
        Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Residual standard error: 0.7558 on 94 degrees of freedom
        Multiple R-squared:  0.5797,  Adjusted R-squared:  0.5708
        F-statistic: 64.84 on 2 and 94 DF,  p-value: < 2.2e-16
```

ADJUSTED R-SQUARED: 0.5708

```
> fit5 <- lm(log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$capspen)
> summary(fit5)
```

```
        Call:
        lm(formula = log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$capspen)

        Residuals:
             Min       1Q   Median       3Q      Max
        -1.64429 -0.42310  0.06919  0.49755  1.91878

        Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
        (Intercept)          1.52299    0.12229  12.454  < 2e-16 ***
        log(sysBP$cancervol)  0.65531    0.08664   7.564  2.6e-11 ***
        sysBP$capspen        0.03172    0.02699   1.175    0.243
        ---
        Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Residual standard error: 0.7863 on 94 degrees of freedom
        Multiple R-squared:  0.5452,  Adjusted R-squared:  0.5355
        F-statistic: 56.33 on 2 and 94 DF,  p-value: < 2.2e-16
```

ADJUSTED R-SQUARED: 0.5355

```
> fit6 <- lm(log(sysBP$psa) ~ log(sysBP$cancervol) + factor(sysBP$gleason))
> summary(fit6)
```

```
Call:
lm(formula = log(sysBP$psa) ~ log(sysBP$cancervol) + factor(sysBP$gleason))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.51308 -0.47216  0.06966  0.48801  1.79892 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             1.42858    0.14420   9.907 3.22e-16 ***
log(sysBP$cancervol)    0.59438    0.07749   7.670 1.65e-11 ***
factor(sysBP$gleason)7  0.19471    0.18070   1.078  0.28403    
factor(sysBP$gleason)8  0.74617    0.24948   2.991  0.00356 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7595 on 93 degrees of freedom
Multiple R-squared:  0.5801,	Adjusted R-squared:  0.5665 
F-statistic: 42.83 on 3 and 93 DF,  p-value: < 2.2e-16
```

ADJUSTED R-SQUARED: 0.5665

The model with vesinv has the highest adjusted R squared value. So, we chose the fit4 model

```
        > fit4 <- lm(log(sysBP$psa) ~ log(sysBP$cancervol) + sysBP$vesinv)
```

```
> z1<-log(sysBP$psa)
> z2<-log(sysBP$cancervol)
> z3<-(sysBP$vesinv)
> model1<-lm(z1 ~z2 + factor(z3))
> model2<-lm(z1 ~z2)
> summary(model1)
        Call:
        lm(formula = z1 ~ z2 + factor(z3))

        Residuals:
            Min      1Q  Median      3Q     Max 
        -1.6217 -0.5281  0.1209  0.4840  1.6907 

        Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
        (Intercept)  1.53531    0.11733  13.086  < 2e-16 ***
        z2           0.59118    0.07767   7.611 2.07e-11 ***
        factor(z3)1  0.67187    0.22113   3.038  0.00308 ** 
        ---
        Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Residual standard error: 0.7558 on 94 degrees of freedom
        Multiple R-squared:  0.5797,	Adjusted R-squared:  0.5708 
```

```
F-statistic: 64.84 on 2 and 94 DF,  p-value: < 2.2e-16
```

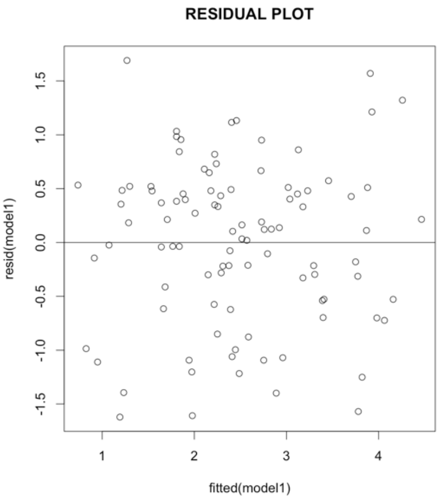
```
> anova(model1)
        Analysis of Variance Table

        Response: z1
                  Df Sum Sq Mean Sq  F value  Pr(>F)
        z2         1 68.801  68.801 120.4445 < 2e-16 ***
        factor(z3)  1  5.273   5.273   9.2313 0.00308 **
        Residuals  94 53.695   0.571
        ---
        Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The estimated intercept, b_0=11.53531 and slope is b_1=0.59118 and b_2=0.67187.
From the, adjusted R-squared value, cancervol and vesinv together explain 57.08 % of the total variability of PSA level.

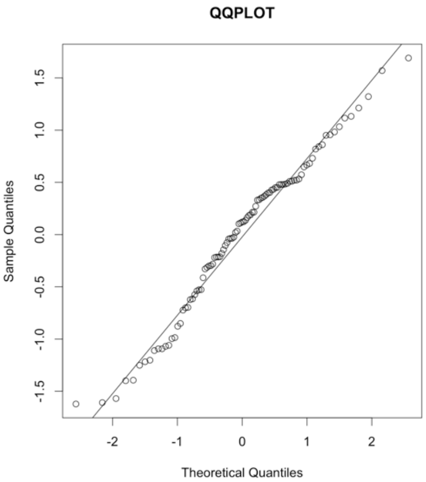
Key Assumptions:

- Errors are constant
- Errors are independent
- Errors follow normal distribution

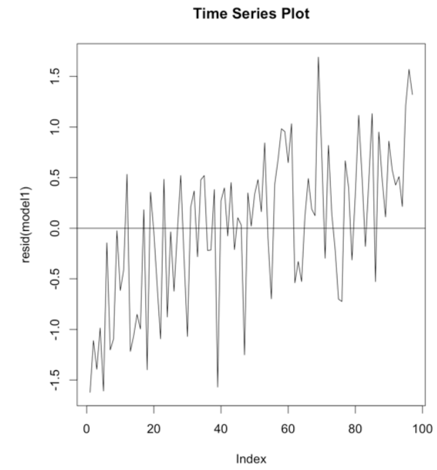
#Check for errors whether they have mean zero constant variance, normality and independence of residual error



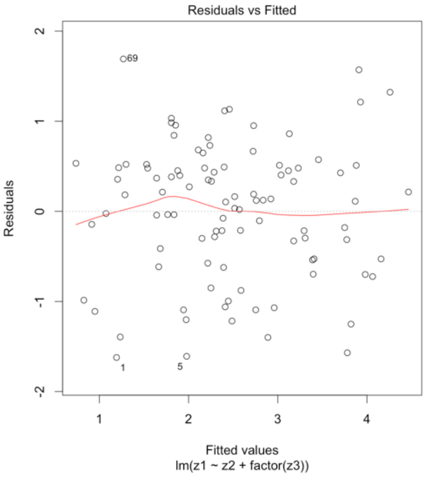
```
> plot(fitted(model1),resid(model1),
main="RESIDUAL PLOT")
abline(h=0)
```



```
> qqnorm(resid(model1),
main="QQPLOT")
> qqline(resid(model1))
```



```
> plot(resid(model1), type="l",main="Time
Series Plot")
> abline(h=0)
```



```
> plot(model1)
```

It is clear from the plots, that all the key assumptions are satisfied.

2. PSA value for a patient whose predictor variables are at the sample medians of the variable.

```
> cancervol<-sysBP$cancervol
> psa<-sysBP$psa
> vesinv<-sysBP$vesinv
> #prediction
> PredictedValue<-
exp(1)^predict(model1,data.frame(z2=log(median(cancervol)),z3=((vesinv=0))))
>
>
> PredictedValue
       1
10.94097
```