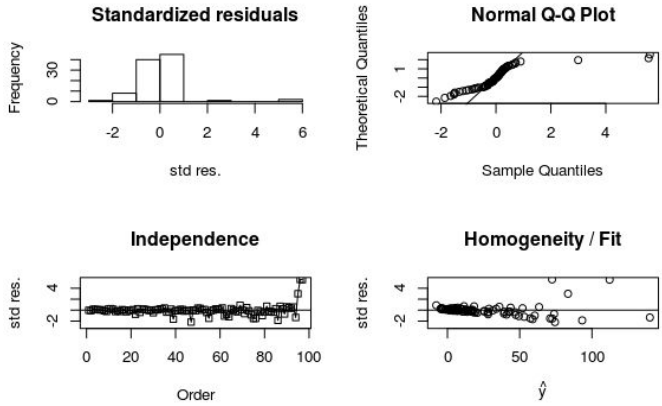
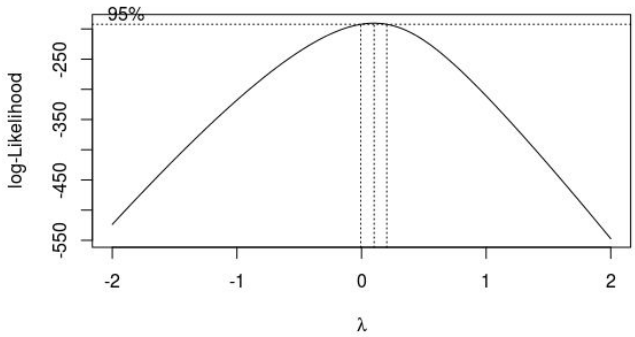
Problem 1

Initial fit of the model:

```
# Variable names(columns) 1-7 below
> psa.reg <- lm(psal ~ cv + weight + age +
          bph + svi + cp + gs, data=psa)
> source("check.R")
> check(psa.reg,tests=TRUE)
```



As we can see from adjacent plots, current fitted model violates condition of normality, equality variance. We need to transform the response.



Box-Cox Transformation:

```
> bc2=boxcox(psa.reg)
> lambda=bc2$x[which.max(bc2$y)];lambda
```

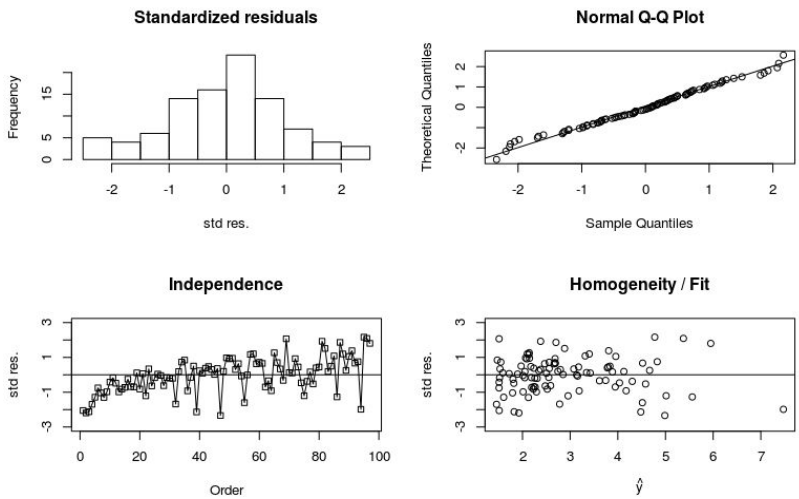
Lambda = 0.10 < 1 ⇒ Need for transformation

```
> psal.T=bcPower(psa$psal,lambda)
> psa=cbind(psa,psal.T)
> psa.model <- lm(psal.T ~ cv + weight +
   age + bph + svi + cp + gs, data=psa)
```

Post-Transformation - Validation Check:

```
> check(psa.model,tests=TRUE)
```

Normality: From Normality Q-Q plot, the points are not deviating much from the straight line, therefore the condition of normality is not violated.



Equal Variance: From std residuals vs fit (4th plot), the variability of residuals is not changing much with fitted values. Therefore condition of equal variance not violated here.

Independence: From Independence plot between std res vs order, we see that there's no specific pattern exists and one cannot conclusively predict where the next observation lies, thus they are truly independent.

Overall fit of the model post-transformation:

$H_0 : \beta_{cv} = \beta_{weight} = \beta_{age} = \beta_{bph} = \beta_{svi} = \beta_{cp} = \beta_{gs} = 0$ vs H_a = At least one $\neq 0$

```
> psa.model <- lm(psal.T ~ cv + weight + age + bph + svi + cp + gs, data=psa) #All Variables
> cor(psa[c("psal.T","cv","weight","age","bph","svi","cp","gs")   > summary(psa.model)
```

```
        psal.T     cv weight    age    bph    svi     cp     gs
psal.T  1.000  0.680  0.111 0.153  0.134  0.587  0.539  0.547
cv             1.000  0.005 0.039 -0.133  0.582  0.693  0.481
weight                1.000 0.164  0.322 -0.002  0.002 -0.024
age                         1.000  0.366  0.118  0.100  0.226
bph                                1.000 -0.120 -0.083  0.027
svi                                       1.000  0.680  0.429
cp                                              1.000  0.462
gs                                                     1.000
```

```
# Variable   p-value
(Intercept)  0.42168
cv           2.6e-06
weight       0.47544
age          0.67319
bph          0.00521
svi          0.00251
cp           0.44681
gs           0.00540
```

```
Multiple R-squared:  0.6127,      Adjusted R-squared:  0.5822  AIC: 278.9192
```

Over fit of the model with all variables has an excellent R2 = 61.7 % ⇒ Large portion of the model variability in response can be accounted by this model.

Simplifying the model:

1. From above table we can see that Seminal Vesicle Invasion(svi) and Capsular Penetration(cp) are strongly correlated. Since p-value for cp > svi, we remove 'cp' from our model.
2. p-value of 'age' is 0.67 >> 0.05, and also 'weight' is 0.475 >> 0.05 ; hence we attempt to remove 'age' & 'weight' from our model.

The reduced model with four variables need to be fitted now.

$$H_0 : \beta_{age} = \beta_{weight} = \beta_{cp} = 0$$

$$H_a : \text{At least one of them} \neq 0$$

Fitting new model:

```
> psa.model.r <- update(psa.model, . ~ . -cp -age -weight) # Removing cp,age,weight
> summary(psa.model.r)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.11432    1.02261  -1.090  0.27870
cv           0.08735    0.01622   5.386  5.5e-07
bph          0.10912    0.03290   3.317  0.00131
svi          0.92999    0.29846   3.116  0.00245
gs           0.42573    0.15567   2.735  0.00749
Multiple R-squared:  0.6072,   Adjusted R-squared:  0.5901    AIC: 274.2819
> anova(psa.model.r) # P-value 0.922 > 0.05
```

⇒ Therefore we fail to reject H_0 null hypothesis ⇒ Capsular Penetration(cp), Age(age), Weight(weight) are not statistically significant predictors.

⇒ Also, in the new model coefficient of determination R^2 hasn't significantly changed compared to old (61.2% → 60.7%) hanged while AIC indeed decreased from 278.9 → 274.28 as expected.

⇒ I tried removing Benign Prostatic Hyperplasia ('bph') from current model but it resulted in significant decrease of R^2 to 56.04%. Hence we could stop our model development and accept this as final model.

Final Model: PSA-transformed = -1.11 + 0.087(CV) + 0.10912(BPH) + 0.9299(SVI) + 0.4251(GS)

```
>newdata=data.frame(cv=4.236,weight=22.783,age=68,bph=1.3500,svi=0,cp=0,gs=
6)
> predict(psa.model.r, newdata, interval="prediction",level=0.90)
       fit       lwr      upr
1 1.957366 0.3403469 3.574385
```

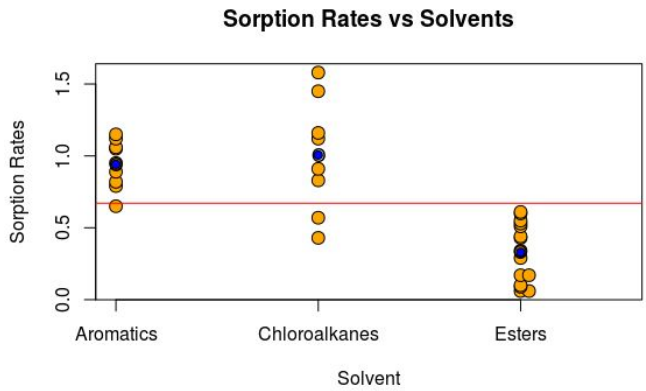
For given data 90% prediction interval for transformed PSA = **(0.34, 3.57)**

Problem 2

We will be using Completely Randomized Design since the problem involves three categories of data and we need to infer whether their means differ. Given are three potential hazardous chemical cleaning solvents and their sorption rates.

Conduct Analysis of Variance:

Using reshape2/melt function we convert given data from wide-large with two columns 'sorption' and 'solvent'. Here solvent takes three given chemicals Aromatics, Chloroalkanes, Esters.




```
> ccs <- read.csv("prob2_data_melt.csv");
attach(ccs)
> m1=aov(sorption~solvent); anova(m1)
Analysis of Variance Table
Response: sorption
          Df Sum Sq Mean Sq F value    Pr(>F)
solvent    2 3.3054 1.65270  24.512 5.855e-07
Residuals 29 1.9553 0.06743
```

Hypothesis Test:

H_0: $\mu_1 = \mu_2 = \mu_3$, all three solvents have have same mean sorption rates

H_a: Not all μ_i's are equal, there is difference in mean sorption rates for at least one solvent.

From above anova output we can see that p-value corresponding to F-statistic is ~ 0.00 i.e. p-value < 0.05. Therefore there is significant evidence to reject null hypothesis and we conclude that mean sorption rates of at least one solvent is different.

Finding the differences:

We employ two available methods discussed in class, we do twice with different methods so as to be confident in results.

Bonferroni procedure:

```
> Bonf(sorption~solvent,data=ccs) # "95 %  Pairwise CIs"
                           Diff.  Lower Upper Differ?
Aromatics - Chloroalkanes -0.064 -0.385 0.257       0
Aromatics - Esters         0.612  0.334 0.890       1
Chloroalkanes - Esters     0.676  0.387 0.965       1
```

Graphical representation → Aromatics Chloroalkanes Esters

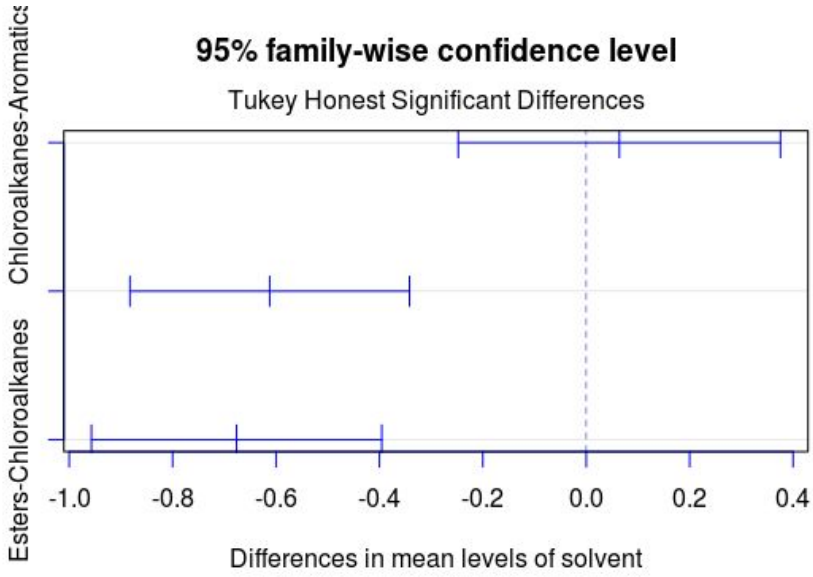
Tukey's Procedure:

```
> ccs.Tukey=TukeyHSD(m1, "solvent")
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

```
                                diff        lwr        upr     p adj
Chloroalkanes-Aromatics  0.06402778 -0.2475781  0.3756337 0.8683140
Esters-Aromatics        -0.61222222 -0.8826095 -0.3418350 0.0000143
Esters-Chloroalkanes    -0.67625000 -0.9570006 -0.3954994 0.0000054
```

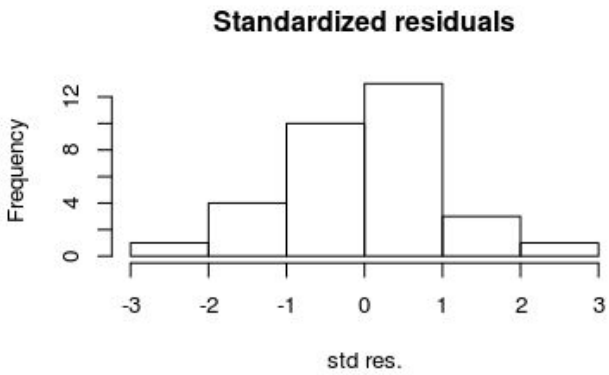
With 95% confidence level, we can say mean sorption rates of Aromatics - Esters & Chloroalkanes -Esters differ.

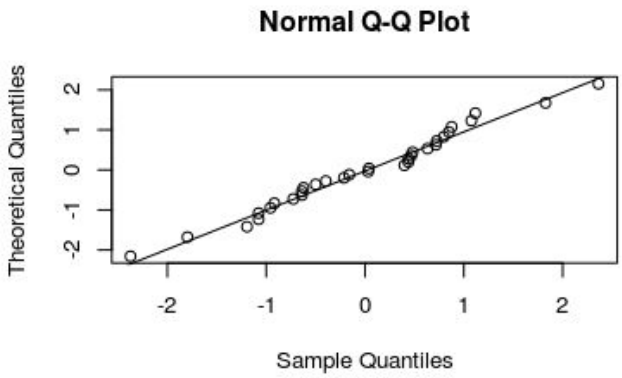
The following graph shows the 95% family-wise intervals. Mean sorption rates of Esters-Aromatics & Esters-Chloroalkanes differ by 61% and 67% (from Tukey's output)

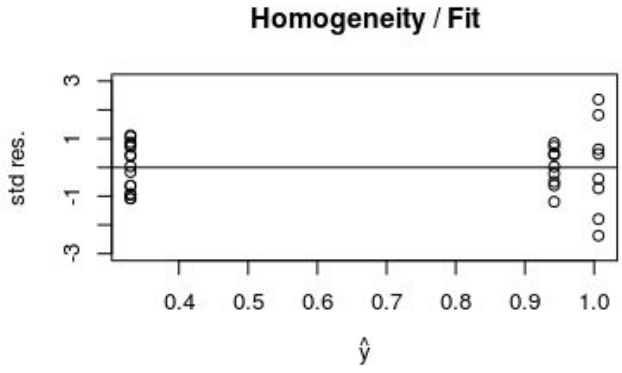


Are the conclusions valid?

In order for the conclusions to be valid, conditions of normality, equal variance and independence must be satisfied. We use residuals from fitted model to assess the three conditions.







Normality: From Normality Q-Q plot, the points are not deviating much from the straight line, therefore the condition of normality is not violated.
Equal Variance: From std residuals vs fit (4th plot), the variability of residuals is not changing much with fitted values. Therefore condition of equal variance not violated here.
Independence: From Independence plot between std res vs order, we see that there's no specific pattern exists and one cannot conclusively predict where the next observation lies, thus they are truly independent.

Since the proposed model satisfies the above conditions, we can confidently conclude that our inferences and results are valid and declare the mean sorption rates differ among three chemical solvents.