Stat 331 Applied Linear Models – Assignment 2
Solution

1(a) 2 points

Notice that $SSE = (Y - \hat{Y})^T(Y - \hat{Y})$. Because $\hat{Y} = HY$, $Y - \hat{Y} = (I - H)Y$, hence $SSE = Y^T(I - H)^T(I - H)Y$.

1(b) 3 points

Consider $r = (I - H)Y$. Clearly SSE is a function of r ($SSE = r^T r$). Hence we are to prove r and $\hat{Y}$ are independent.

Notice that both r and $\hat{Y}$ are linear combinations of normally distributed random variables Y. Thus, it suffices to prove $Cov(r, \hat{Y}) = 0$.

Notice that $Cov(r, \hat{Y}) = (I - H)Var(Y)H = \sigma^2(I - H)H = 0$. Hence the claim is valid.

1(c) 3 points

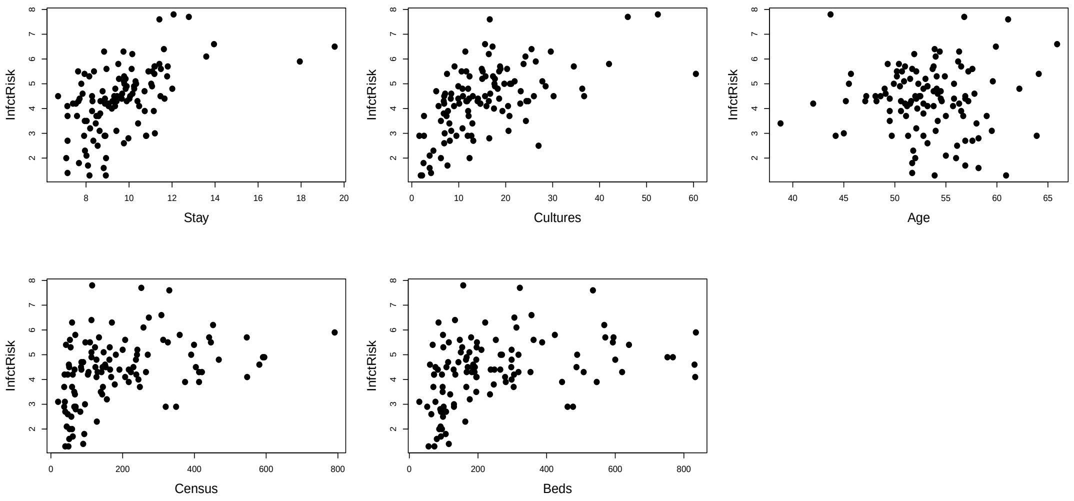
$$\begin{aligned}
SSR &= \sum(\hat{y}_i - \bar{y})^2 \\
&= \sum(\hat{y}_i^2 - 2\hat{y}_i\bar{y} + \bar{y})^2 \\
&= \sum \hat{y}_i^2 - 2\bar{y}\sum \hat{y}_i + n\bar{y}^2 \\
&= \hat{Y}^T\hat{Y} - 2n\bar{y}^2 + n\bar{y}^2 \\
&= Y^T H^T H Y - n\bar{y}^2 \\
&= Y^T H Y - n\bar{y}^2.
\end{aligned}$$

In the third step above, prove $\sum \hat{y}_i = n\bar{y}$:

$$\begin{aligned}
\sum \hat{y}_i &= \sum \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= \sum(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i) \\
&= n\bar{y} + \hat{\beta}_1 \sum(x_i - \bar{x}) \\
&= n\bar{y} + \hat{\beta}_1 0 \\
&= n\bar{y}
\end{aligned}$$

2(a) 2 points

Based on the plots, it seems Stay and Cultures are strongly related to InfctRisk.



2(b) 3 points

$\hat{\beta}_2 = 0.059$.

Its interpretation is: The estimated infection risk increases by 0.059 unit (0.059%) if x_2 (Cultures) increases by 1 unit, while the other predictors are fixed.

2(c) 3 points

$\hat{y} = 3.466$

The 95% prediction interval is $(1.449, 5.482)$.

2(d) 5 points

The null hypothesis is $H_0:\ \beta_3 = \beta_4 = \beta_5 = 0$.

The alternative hypothesis is H_α : At least one of $\beta_3,\ \beta_4,\ \beta_5$ is not 0.

The calculated F statistic is 1.781, and the critical value is $F_{0.05}(3, 107) = 2.689$. Thus we fail to reject H_0.

2(e) 3 points

The R^2 value is 0.450.

Interpretation: 45% of total variation in the response variable InfctRisk can be explained by the regression model (with Stay and Cultures as the predictors).

The R^2 value for the five-predictor model is 0.477. Therefore, the R^2 value for the full model is larger.

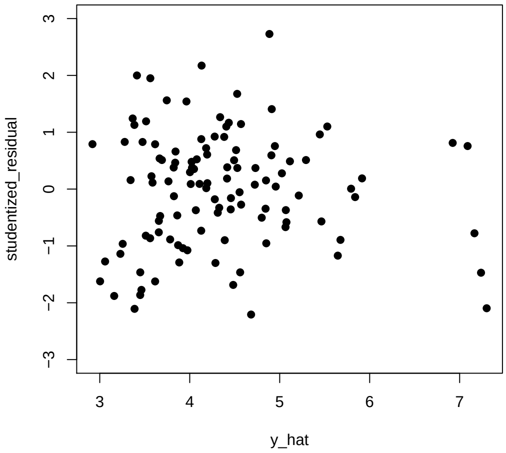
Comment: Although Age, Census, and Beds are not significantly related to the response variable, the R^2 value for the full model is always no smaller than that of the reduced model. Thus, R^2 value is not always a good criterion to assess whether one model is good or not. Further analysis is needed to determine the best model.

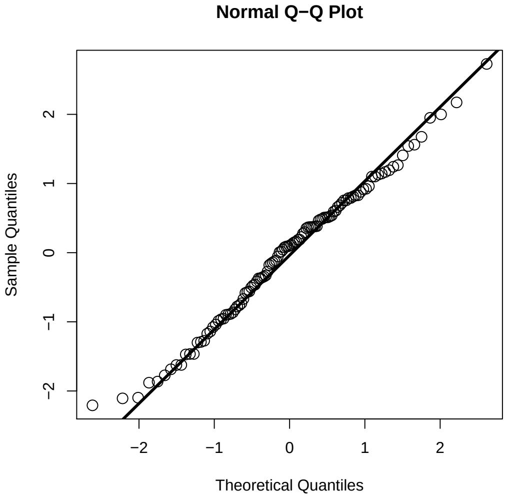
On the other hand, the increase in R^2 is very small (2.7%), which also indicates that Age, Census, and Beds are not significant predictors when the other two are in the model.

2(f) 4 points

For the studentized residual vs. fitted y plot, all of the residuals are in $[-3, 3]$, and most are in $[-2, 2]$. There is no special trend in the plot. The variance seems to be constant with respect to $\hat{y}$. Therefore, there is no significant evidence that the assumption $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$ are violated.

For the QQ plot, the dots are close to a straight line. There is no strong evidence that the normality assumption is violated.



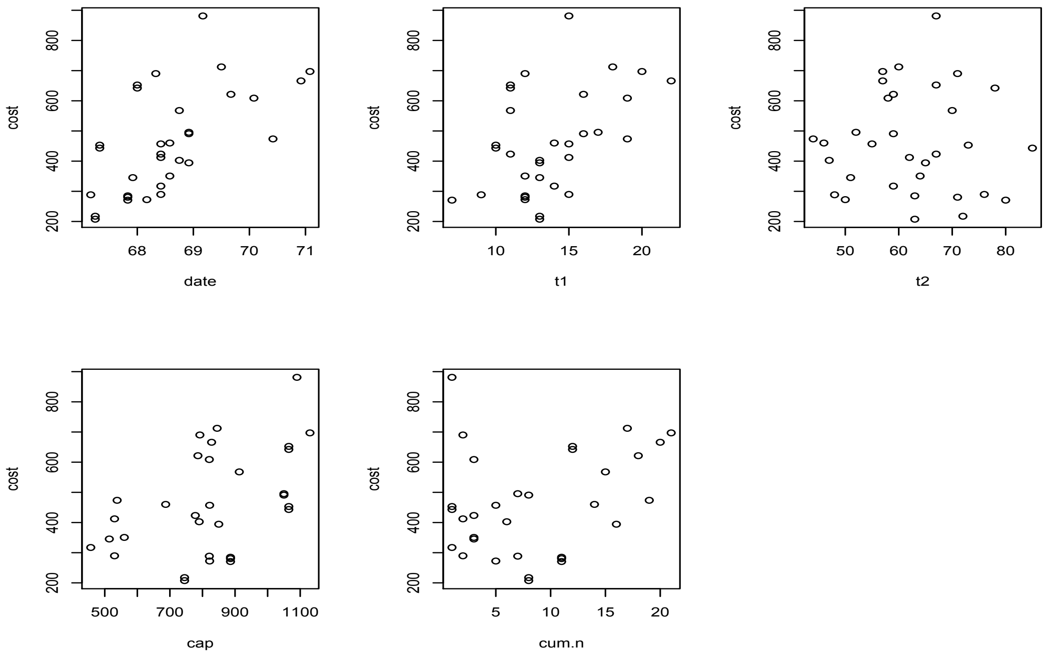


```
setwd("") ## set your own work directory

senic = read.table("Senic.txt",header=T)

x = senic[,c(2,5,3,10,7)]

y = senic$InfctRsk

### (a)

par(mfrow=c(2,3))

for (i in 1:5)
{
plot(x[,i],y, main="", xlab=colnames(x)[i], ylab="InfctRisk",
pch=19,cex=1.4,cex.lab=1.5)
}

### (b)

senic.data = data.frame(x,y)

fm=lm(y~Stay+Cultures+Age+Census+Beds,data=senic.data)

summary(fm)

### (c)

new.x = data.frame( t(as.matrix(c(7,9,56,200,250))))

colnames(new.x)=colnames(x)

y_hat_p = predict(fm,new.x)

y_hat_p

sigma_square = summary(fm)$sigma^2

standard_error = sqrt(sigma_square * (1 + c(1,7,9,56,200,250) %*%
summary(fm)$cov.unscaled %*% c(1,7,9,56,200,250)))

t_star = qt(0.975,df=summary(fm)$df[2])

lower = round(y_hat_p-t_star*standard_error,digits=3)
upper = round(y_hat_p+t_star*standard_error,digits=3)

print(paste("(",lower,",",upper,")",sep=""))

### (d)
```

```
SSE_full = anova(fm)[6,2]

SSE_reduced=SSE_full+anova(fm)[3,2]+anova(fm)[4,2]+anova(fm)[5,2]

F = ( (SSE_reduced - SSE_full)/3 ) / (SSE_full / summary(fm)$df[2])

F0 = qf(0.95,3,summary(fm)$df[2])

F

F0

### (e)

x2=senic[,c(2,5)]

senic.data2 = data.frame(x2,y)

fm2=lm(y~Stay+Cultures,data=senic.data2)

summary(fm2)$r.squared

summary(fm)$r.squared

### (f)

x3 = cbind(1,as.matrix(x2))

H = x3 %*% solve(t(x3) %*% x3) %*% t(x3)

y_hat = H%*%y

residual = y - y_hat

s2 = sum(residual^2)/(113-2-1)

studentized_residual = residual / sqrt(s2*(1-diag(H)))

plot(y_hat, studentized_residual,ylim=c(-3,3),pch=19)


qqnorm(studentized_residual,cex=1.5)

qqline(studentized_residual,lwd=3)
```

3(a) 2 points

All predictors seem to be related to the cost except t2. The plot of cost vs. t2 looks like completely random.



3(b) 2 points

We conduct Box-Cox analysis and decide to choose $\lambda = 0$ (the log transformation) since value 0 is in the 95% CI for λ.

3(c) 3 points

The fitted equation is: $\log(\hat{y}) = -1.063 + 0.228date + 0.0053t1 + 0.0056t2 + 0.00088cap - 0.108pr + 0.260ne + 0.116ct + 0.037bw - 0.012cum.n - 0.22pt$

$\exp(0.0008837) - 1 = 0.09\%$.

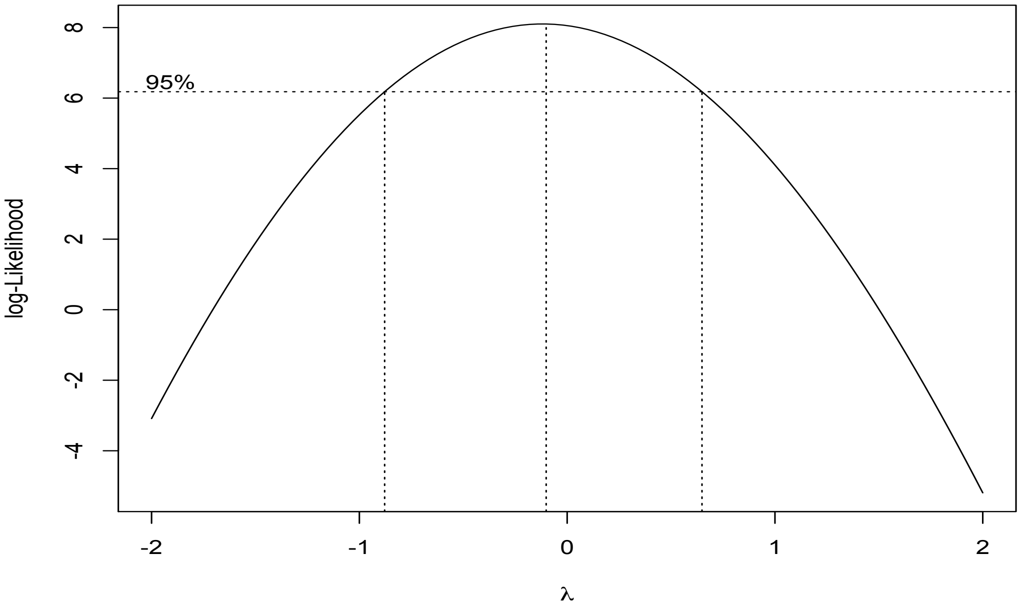
For one unit increase in the net capacity, the cost increases by 0.09% when other predictors are held constant.

$\exp(0.2595) - 1 = 29.63\%$.

The cost of construction of an LWR plant in the north-east region of USA is 29.63% higher than the cost of construction of an LWR plant in the other region of USA, when other predictors are held constant.

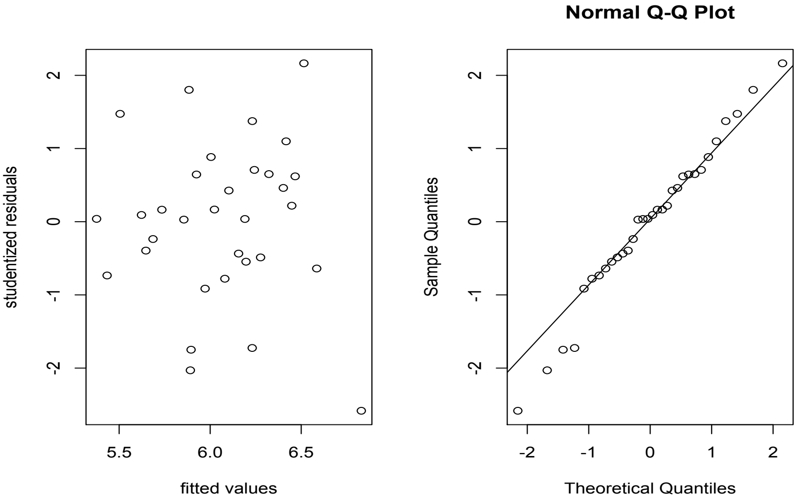
3(d) 2 points

The signs of estimates are reasonable. For example, date, t1 and t2 are all positively correlated with cost since the longer it takes to get the permit and license, the construction cost increases. The cost is also increased if net capacity increases, a cooling tower is used or if constructed in the north-east region. Also the cost decreases as experience of architect-engineer increases, and there exists a prior LWR on same site.



3(e) 3 points

The plot based on transformed data gives no strong evidence of any systematic departure from the assumed model. The residuals are randomly scattered. There is no obvious trend. The variance of the residuals seems constant. The QQ plot shows the residuals are normally distributed.



```
library(boot)
data(nuclear)

### (a)
par(mfrow=c(2,3))
attach(nuclear)
plot(date,cost,xlab="date",ylab="cost")
plot(t1,cost,xlab="t1",ylab="cost")
plot(t2,cost,xlab="t2",ylab="cost")
plot(cap,cost,xlab="cap",ylab="cost")
plot(cum.n,cost,xlab="cum.n",ylab="cost")

### (b)
boxcox(cost~.,data=nuclear)

### (c)
fit=lm(log(cost)~.,data=nuclear)
summary(fit)

### (e)
par(mfrow=c(1,2))
plot(fitted.values(fit),rstudent(fit),xlab="fitted values",
ylab="studentized residuals")
qqnorm(rstudent(fit))
qqline(rstudent(fit))
```