Stat 425 Homework 2

Sharvi Tomar (stomar2)

30/08/21

Contents

Problem 1: Using the sat data from the faraway library

(a) Fit a model with total sat score as the response and expend, ratio and salary as predictors. Test the hypothesis that βsalary=0. Test the hypothesis that βsalary = βratio = βexpend = 0. Do any of these predictors have an effect on the response?

```
# Fitting a model with total sat score as the response, expend, ratio and salary as predictors.
library(faraway)
model1 = lm(total ~ expend + ratio + salary, data = sat)
summary(model1)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

Test:

H0 : βsalary $= 0$

Ha : βsalary $\neq 0$

p-value for salary(0.0667) is greater than 0.05. Thus at 5% level, we fail to reject the null hypothesis, i.e., we do not have enough evidence to conclude that salary is statistically significant in predicting the response.

Test

H0 : βsalary= βratio= βexpend $= 0$

Ha : Atleast one of the βsalary,βratio, βexpend is non-zero

Overall p-value for the model is 0.0129 which is less than 0.05. Thus at 5% level, we reject the null hypothesis in favor of alternate hypothesis that atleast one of the coefficients is non-zero.

Yes, when taken together atleast one variable has effect on the response.

(b) Now add takers to the model. Test the hypothesis that βtakers $= 0$. Compare this model to the previous one using an F-test. Demonstrate that the F-test is equivalent to the t-test.

```
model2 = lm(total ~ expend + ratio + salary + takers, data = sat)
summary(model2)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

Test:

H0 : βtakers= 0

H1 : βtakers $\neq$ 0

Since the p-value for takers(2.61e-16) is less than 0.05, we reject the null hypothesis and the accept the alternate. Thus, at 5% level, takers is statistically significant in predicting the response.

```
# Compare this model to the previous one using an F-test.
a=anova(model1, model2)
a
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio + salary
## Model 2: total ~ expend + ratio + salary + takers
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     46 216812
## 2     45  48124  1    168688 157.74 2.607e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Showing how f-test is same as t-test for this
f_value = a[2,5]
t_sqrd = summary(model2)$coef[5,3]^2
round(f_value,4)
```

```
## [1] 157.7379
```

```
round(t_sqrd,4)
```

```
## [1] 157.7379
```

F = t^2. Thus, F_test is same as t_test.

Problem 2: For the prostate data from the faraway library, fit a model with lpsa as the response and the other variables as predictors:

```
# fit a model with lpsa as the response and the other variables as predictors:
model3 = lm(lpsa ~ ., data = prostate)
summary(model3)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

(a) Compare 90% and 95% CIs for the parameter associated with age.

```
# 95% CIs for the parameter associated with age
confint(model3, 'age', level = 0.95)
```

```
##           2.5 %      97.5 %
## age -0.04184062 0.002566267
```

```
# 90% CIs for the parameter associated with age
confint(model3, 'age', level = 0.90)
```

```
##            5 %         95 %
## age -0.0382102 -0.001064151
```

b) Remove all predictors that are not significant at the 5% level. Test this model against the original model. Which model is preferred?

```
# Remove all predictors that are not significant at the 5% level.
model4=lm(lpsa ~lcavol+lweight+svi, data = prostate)
summary(model4)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
```

```
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
# Test this model against the original model.
anova(model3, model4)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
## Model 2: lpsa ~ lcavol + lweight + svi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     88 44.163
## 2     93 47.785 -5   -3.6218 1.4434 0.2167
```

Null Hypothesis : Reduced Model

Alternate hypothesis: Full Model

P-value for the anova F-test is greater than 0.05, thus we fail to reject the null hypothesis. Thus, our reduced model is useful in explaining the variance in response variable.Thus, we will go ahead with reduced model

c) Compute and display a 95% joint confidence region for the parameters associated with age and lph. Plot the origin (0; 0) on this display. The location of the origin on the display tell us the outcome of a certain hypothesis test. State that test and its outcome.

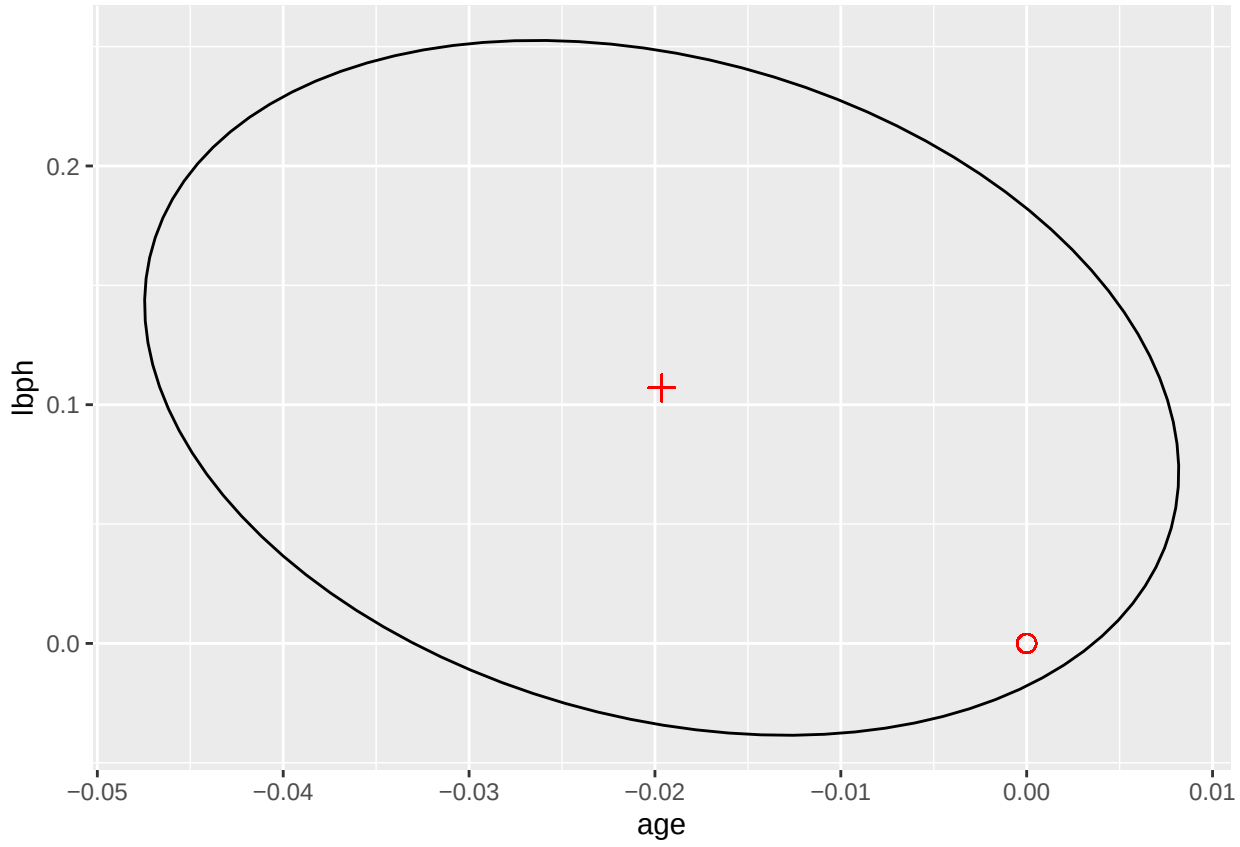
```
# Compute and display a 95%joint confidence regionfor the parameters associatedwithageandlph.
require(ellipse)
```

```
## Loading required package: ellipse

##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```

```
CR95=ellipse(model3,c(4,5))
head(CR95)
```

```
##                age      lbph
## [1,] -0.002508614 0.1966607
## [2,] -0.003933479 0.2037540
## [3,] -0.005421577 0.2104578
## [4,] -0.006966916 0.2167454
## [5,] -0.008563274 0.2225912
## [6,] -0.010204222 0.2279718
```

```
library(ggplot2)
ggplot(data = data.frame(CR95), aes(x = age, y = lbph)) +
  geom_path() +
  geom_point(
    x = coef(model3)[4],
    y = coef(model3)[5],
    shape = 3,
    size = 3,
    colour = 'red'
  ) +
  geom_point(
    x = 0,
    y = 0,
    shape = 1,
    size = 3,
    colour = 'red'
  )
```



Test:

H0 : βage= βlbph= 0

H1 : Atleast one of βage, βlbph is non-zero.

Thus, we fail to reject the null hypothesis, i.e. we do not have enough evidence to conclude that atleast one of age or lbph is significant in predicting lpsa.

d) In class we discussed a permutation test corresponding to the F-test for the significance of a set of predictors. Execute the permutation test corresponding to the t-test for age in this model.

```
# permutation test
fullmodel_ps = lm(lpsa~., data = prostate)
n.iter=2000
fstats=numeric(n.iter)
for (i in 1:n.iter) {
newprostate=prostate
newprostate[,3]=prostate[sample(97),3]
lm.fit=lm(lpsa~.,data=newprostate)
fstats[i]=summary(lm.fit)$fstat[1]
}
length(fstats[fstats > summary(fullmodel_ps)$fstat[1]])/n.iter
```

```
## [1] 0.088
```

P-value for age is greater than 0.05. Thus at 5% level, we fail to reject the null hypothesis,i.e, we do not have enough evidence to conclude that age is significant in predicting lpsa.

Problem 3: In the punting data from the faraway library we find average distance and hang times of 10 punts of an American football as related to various measures of leg strength for 13 volunteers.

a) Fit a regression model with Distance as the response, and the right and left strengths, and the right and left exibilities as predictors. Which predictors are significant at the 5% level?

```
fullmodel_pt=lm(Distance~RStr+LStr+RFlex+LFlex,data=punting)
summary(fullmodel_pt)
```

```
##
## Call:
## lm(formula = Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.941  -8.958  -4.441  13.523  17.016
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -79.6236    65.5935  -1.214    0.259
## RStr          0.5116     0.4856   1.054    0.323
## LStr         -0.1862     0.5130  -0.363    0.726
## RFlex         2.3745     1.4374   1.652    0.137
## LFlex        -0.5277     0.8255  -0.639    0.541
##
## Residual standard error: 16.33 on 8 degrees of freedom
## Multiple R-squared:  0.7365, Adjusted R-squared:  0.6047
## F-statistic:  5.59 on 4 and 8 DF,  p-value: 0.01902
```

None of the variables is significant individually at 5% level.Overall p-value of the F-test is 0.01902 which is less than 0.05.Thus, all 4 predictors taken together explains the variance better than an intercept model. Thus, all 4 variables are significant jointly but not individually

(c) Relative to the model in (a) (full model), test whether the right and left strength have the same effect.

```
reducedmodel_1_pt = lm(Distance ~ I(RStr+LStr) + RFlex+LFlex , data = punting)
anova(reducedmodel_1_pt,fullmodel_pt)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr) + RFlex + LFlex
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2287.4
## 2      8 2132.6  1    154.72 0.5804  0.468
```

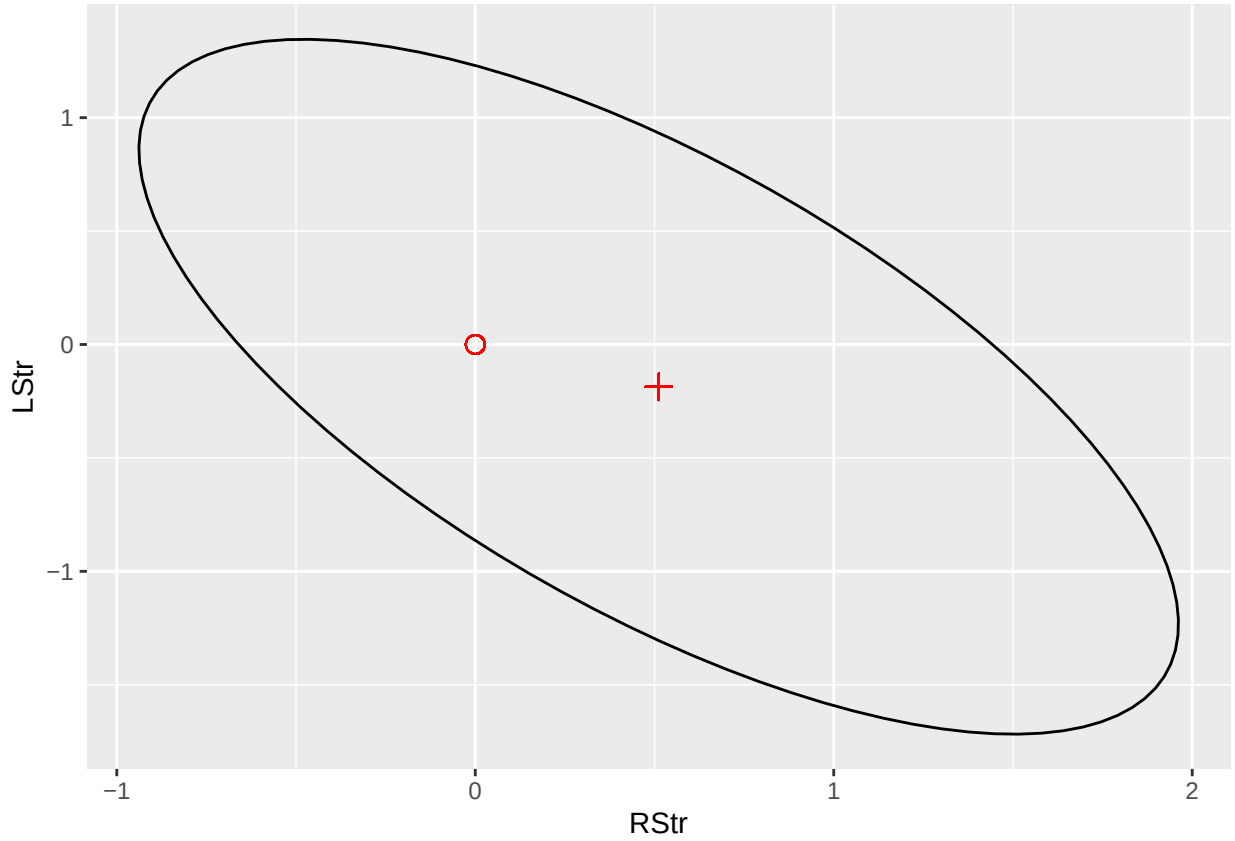
Conclusion:We fail to reject the null hypothesis. Thus, both coefficients are not statistically different, i.e, they can have same effect.

(d) Construct a 95% confidence region for (βRStr; βLStr). Explain how the test in (c) relates to this region.

```
library(ellipse)
CR95=ellipse(fullmodel_pt,c(2,3))
head(CR95)
```

```
##           RStr      LStr
## [1,] 1.0890670 0.4238078
## [2,] 1.0035429 0.5117008
## [3,] 0.9160381 0.5967835
## [4,] 0.8269049 0.6787135
## [5,] 0.7365023 0.7571607
## [6,] 0.6451942 0.8318094
```

```
library(ggplot2)
ggplot(data = data.frame(CR95), aes(x = RStr, y = LStr)) +
  geom_path() +
  geom_point(
    x = coef(fullmodel_pt)[2],
    y = coef(fullmodel_pt)[3],
    shape = 3,
    size = 3,
    colour = 'red'
  ) +
  geom_point(
    x = 0,
    y = 0,
    shape = 1,
    size = 3,
    colour = 'red'
  )
```



Since (0,0) lies inside the 95% confidence region, we fail to reject the null hypothesis,i.e, we fail to conclude that atleast one of LStr and RStr is significant in predicting the Distance.

e) Fit a model to test the hypothesis that it is total leg strength, defined by adding the right and left leg strengths, that is sufficient to predict the response, in comparison to using individual left and right strengths.

```
fullmodel_2_pt = lm(Distance ~ RStr+LStr, data = punting)
reducedmodel_2_pt = lm(Distance ~ I(RStr+LStr), data = punting)
anova(reducedmodel_2_pt,fullmodel_2_pt)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr)
## Model 2: Distance ~ RStr + LStr
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     11 3061.3
## 2     10 2973.1  1    88.281 0.2969 0.5978
```

Null Hypothesis : Reduced Model

Alternate hypothesis: Full Model

Since p-value of the f-test is greater than 0.05, we fail to reject null. Thus, proposed new variable is useful in explaining the variance.

(f) (4CR) Relative to the model in (a), test whether the right and left leg flexibilities have the same effect.

```
reducedmodel_3_pt = lm(Distance ~ RStr+LStr + I(RFlex+LFlex) , data = punting)
anova(reducedmodel_3_pt,fullmodel_pt)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ RStr + LStr + I(RFlex + LFlex)
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2648.4
## 2      8 2132.6  1    515.72 1.9346 0.2017
```

Test:

H0 : βRFlex = βLFlex

Ha : βRFlex $\neq$ βLFlex

Conclusion:We fail to reject the null hypothesis. Thus, both coefficients are not statistically different, i.e, they can have same effect.

(g) (4CR) Test for the left-right symmetry by performing the tests in (c) and (f) simultaneously.

```
reducedmodel_4_pt = lm(Distance ~ I(RStr+LStr) + I(RFlex+LFlex) , data = punting)
anova(reducedmodel_4_pt,fullmodel_pt)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr) + I(RFlex + LFlex)
## Model 2: Distance ~ RStr + LStr + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     10 2799.1
## 2      8 2132.6  2    666.43 1.25  0.337
```

Null Hypothesis : Reduced Model

Alternate hypothesis: Full Model

We fail to reject the null hypothesis.Thus, coefficients for left and right strengths can be same & coefficients for left and right flexibilities can be same. Thus, left and right maybe symmetric; however, we do not have enough evidence to confirm this.

(h) (4CR) Fit a model with Hang as the response, and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.

```
fullmodel_pd = lm(Hang ~ RStr+LStr+RFlex+LFlex, data = punting)
summary(fullmodel_pd)
```

```
##
## Call:
## lm(formula = Hang ~ RStr + LStr + RFlex + LFlex, data = punting)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.36297 -0.13528 -0.07849  0.09938  0.35893
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.225239   1.032784  -0.218    0.833
## RStr         0.005153   0.007645   0.674    0.519
## LStr         0.007697   0.008077   0.953    0.369
## RFlex        0.019404   0.022631   0.857    0.416
## LFlex        0.004614   0.012998   0.355    0.732
##
## Residual standard error: 0.2571 on 8 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7235
## F-statistic: 8.848 on 4 and 8 DF,  p-value: 0.004925
```

No, we can not make a test to compare this model to that used in a) as the response variable is different in both cases

Problem 4: Find a formula relating R2 and the F-test for the regression.

$$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)}$$

TSS = Total Sum of Squares
RSS = Residual Sum of Squares
p = # of predictors
n = # of observations

Simplifying above,

$$F = \frac{\left(1 - \frac{RSS}{TSS}\right)/(p-1)}{\left(\frac{RSS}{TSS}\right)/(n-p)}$$

Taking $\frac{RSS}{TSS} = K$

$$F = \frac{(1-K)/(p-1)}{K/(n-p)} \Rightarrow \left(\frac{1-K}{p-1}\right) \times \left(\frac{n-p}{K}\right)$$

$$\Rightarrow \frac{F(p-1)}{n-p} = \frac{1-K}{K}$$

$$\Rightarrow \frac{F(p-1)}{n-p} + 1 = \frac{1}{K} \quad --- \quad (1)$$

We know that $R^2 = 1 - \frac{RSS}{TSS}$ --- (2)

Substituting (1) in (2) we get,

$$\boxed{R^2 = 1 - \left(1 + \frac{F.(p-1)}{n-p}\right)^{-1}}$$

#

Problem 5: For the prostate data, fit a model with lpsa as the response and the other variables as predictors.

```
fullmodel_pd = lm(lpsa~., data = prostate)
summary(fullmodel_pd)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

(a) Suppose a new patient with the following values arrives. Predict lpsa for this patient along with an appropriate 95% CI.

```
# New patient with the following values arrives.
test = data.frame(
  "lcavol" = 1.44692 ,
  "lweight" = 3.62301,
  "age" = 65,
  "lbph" = 0.3001,
  "svi" = 0,
  "lcp" = -0.79851,
  "gleason" = 7,
  "pgg45" = 15
)

# Predict lpsa for this patient along with an appropriate 95% CI.
predict(fullmodel_pd, test, interval = "prediction")
```

```
##        fit       lwr      upr
## 1 2.389053 0.9646584 3.813447
```

(b) Predict the last question for a patient with the same values except that he is age 20.Explain why the CI is wider.

```
# Another patient with the same values except that he is age 20.
test2 = data.frame(
  "lcavol" = 1.44692 ,
  "lweight" = 3.62301,
  "age" = 20,
  "lbph" = 0.3001,
  "svi" = 0,
  "lcp" = -0.79851,
  "gleason" = 7,
  "pgg45" = 15
)
# Predict lpsa for this patient along with an appropriate 95% CI.
predict(fullmodel_pd, test2, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.272726 1.538744 5.006707
```

```
# Checking range of age predictor
range(prostate$age)
```

```
## [1] 41 79
```

```
# Checking mean of age predictor
mean(prostate$age)
```

```
## [1] 63.86598
```

Confidence interval is wider as age = 20 is outside the range of age (41-79) and much farther from the mean of age(63.86) as compared to age = 65

(c) For the model of the previous question, remove all predictors that are not significant at the 5% level. Now recompute the predictions for the new patient of the previous question and its 95% CI. Are the CIs wider or narrower? Which predictions would you prefer?

```
reducedmodel_pd = lm(lpsa~lcavol+lweight+svi, data = prostate)
predict(reducedmodel_pd,test,interval='prediction')
```

```
##        fit       lwr      upr
## 1 2.372534 0.9383436 3.806724
```

Confidence Interval is wider as compared to the previous question. I will prefer predictions which have narrower intervals at the same confidence level. Thus, in this case I will prefer predictions from the previous question

Problem 6

Given:

$y = X\beta + e$ $\quad \cdot Var(e) = \sigma^2 I$ $\quad Var(Y/X) = \sigma^2 I$

$\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy$

1) Taking variance of $\hat{y}$ given X

$\hat{y} = X\hat{\beta}$

$Var(\hat{y}) = Var(X\hat{\beta})$

$= X\, Var(\hat{\beta})\, X^T$

$= X\sigma^2(X^TX)^{-1}X^T$ $\quad [Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}]$

$= \sigma^2 X(X^TX)^{-1}X^T$

$= \sigma^2 H$

$\boxed{Var(\hat{y}) = \sigma^2 H}$

2) $r = y - \hat{y} = y - Hy = (I-H)y$

Taking variance of $r = (I-H)y$, we get

$Var(r) = Var((I-H)y)$

$= (I-H)\, Var(y)\, (I-H)^T$

$= \sigma^2 (I-H)(I-H)^T$

$= \sigma^2 (I - IH^T - HI^T + HH^T)$

$\Rightarrow \sigma^2 (I - \{X(X^TX)^{-1}X^T\}^T - \{X(X^TX)^{-1}X^T\} + \{X(X^TX)^{-1}X^T\})$ $\quad [$using $HH^T = H]$

$\Rightarrow \sigma^2 (I - 2H + H) = \sigma^2 (I - H)$ $\quad [$using $H^T = H]$

Thus, $\boxed{Var(r) = \sigma^2 (I-H)}$

#