HW2_r

Tianqi Wu

2/16/2020

```
library(faraway)
library(ellipse)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
attach(sat)
attach(prostate)
attach(punting)
```

Problem 1

1(a)

Since the p-value of t-statistic for $\beta_{salary} = 0$ is $0.0667 > 0.05$, we do not have enough evidence to reject the null hypothesis that $\beta_{salary} = 0$ at 95% level of significance. Since the p-value for F-statistic is $0.01209 < 0.05$, we reject the null hypothesis that $\beta_{salary} = \beta_{expend} = \beta_{ratio} = 0$ at 95% level of significance. At least one of these predictors have an effect on the response.

```
model.1a = lm(total~expend+ratio+salary,data=sat)
summary(model.1a)
```

```
##
## 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
```

1(b)

Since the p-value of t-statistic for $\beta_{takers} = 0$ is 2.61e-16 < 0.05, we reject the null hypothesis that $\beta_{takers} = 0$ at 95% level of significance. Since F-stat is 157.74 with p-value 2.607e-16 < 0.05, we reject the null hypothesis and conclude that the variation missed by the reduced model, when being compared with the error variance, is significantly large at 95% level of significance. It means that adding takers improves the model. Since t-stat for $\beta_{takers} = 0$ is -12.559, $(t - stat)^2 = (-12.559)^2 = 157.7285 \approx 157.74 = F - stat$. F-test is equivalent to the t-test.

```
model.1b = lm(total~expend+ratio+salary+takers,data=sat)
summary(model.1b)
```

```
##
## 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
```

```
anova(model.1a,model.1b)
```

```
## 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
```

Problem 2

2(a)

The 95% CI for age is (-0.0418, 0.0025) and 90% CI for age is (-0.0382, -0.0010). The 95% CI is wider.

```
model.2a = lm(lpsa~.,data=prostate)
summary(model.2a)
```

```
##
## 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
```

```
confint(model.2a,'age',level=0.95)
```

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

```
confint(model.2a,'age',level=0.9)
```

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

2(b)

Since the p-value for F-stat is 0.2167. We do not have enough evidence to reject the null hypothesis that the reduced model suffices. Despite the full model has higher adjusted R-squared, the reduced model is preferred according to the result of anova.

```
model.2b = lm(lpsa~lcavol+lweight+svi,data=prostate)
summary(model.2b)
```

```
##
## 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
```

```
anova(model.2b,model.2a)
```

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

2(c)

It tests whether the LS coefficients for *age* and *lbph* are zero or not. Since origin is included in the ellipsoid, we do not have enough evidence to reject the null hypothesis that the LS coefficients for *age* and *lbph* are zero at 95% level of significance.

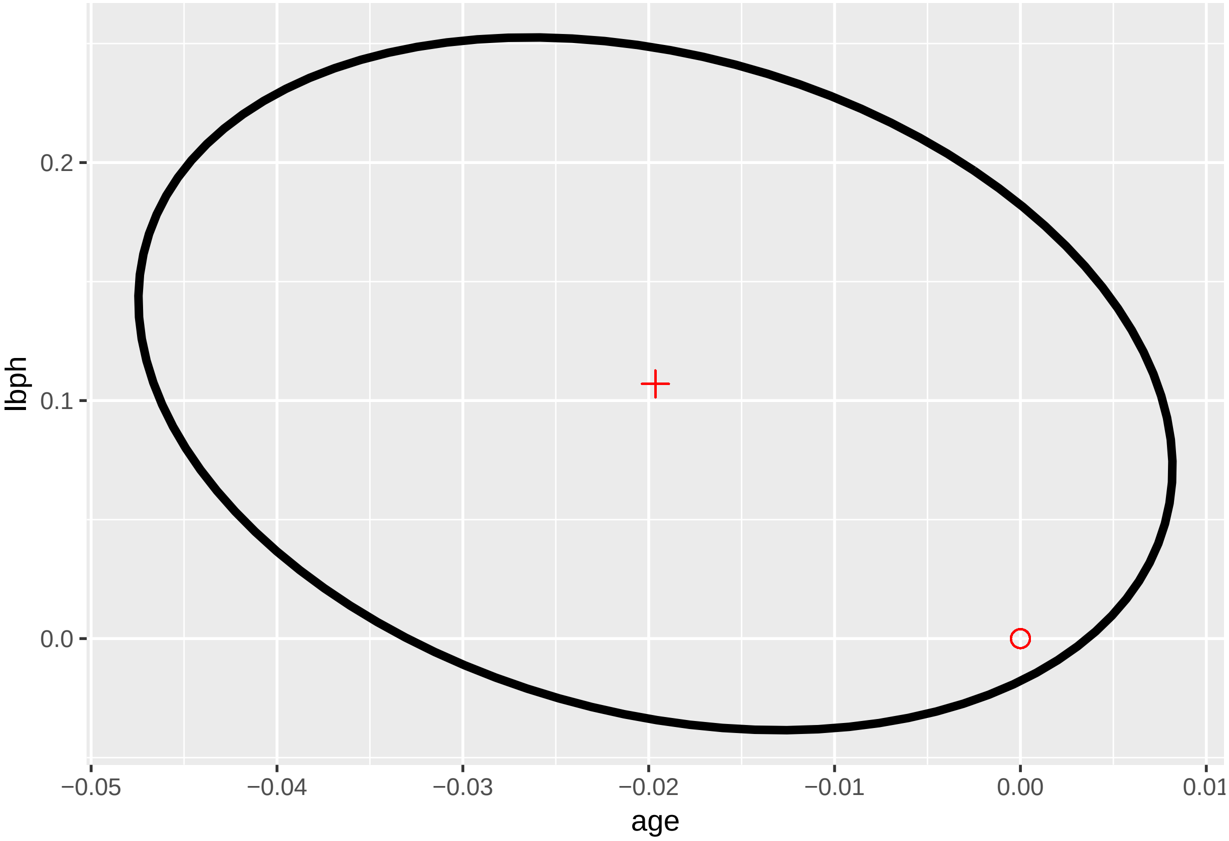
```
CR95 = ellipse(model.2a, c(4,5))
CR95 = data.frame(CR95)
dim(CR95)
```

```
## [1] 100   2
```

```
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
```

```
ggplot(data=CR95, aes(x=age, y=lbph)) +
  geom_path(size=1.5) +
  geom_point(x=coef(model.2a)[4], y=coef(model.2a)[5], shape=3, size=3, colour='red') +
  geom_point(x=0, y=0, shape=1, size=3, colour='red')
```



2(d)

The permutation test indicates that the p-value of t-test for *age* is 0.0852 when setting seed as 123. The p-value of t-test for *age* in the original full model is 0.08229. They are very close.

```
set.seed(123)
n.iter = 5000;
tstats = numeric(n.iter);
for(i in 1:n.iter){
  newprostate=prostate;
  newprostate[,3]=prostate[sample(97),3];
  ge = lm(lpsa ~., data=newprostate);
  tstats[i] = summary(ge)$coef[4,3]
}
p.value= length(tstats[tstats > abs(summary(model.2a)$coef[4,3]) |
                tstats < -abs(summary(model.2a)$coef[4,3])])/n.iter
p.value
```

```
## [1] 0.0852
```

Problem 3

3(a)

Since none of p-value of the predictors are less than 0.05, none of the predictors are significant at the 5% level.

```
model.3a = lm(Distance~RStr+LStr+RFlex+LFlex,data=punting)
summary(model.3a)
```

```
##
## 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
```

3(b)

Since the F-stat is 5.59 with p-value $0.01902 < 0.05$, we can reject the null hypothesis that $\beta_{RStr} = \beta_{LStr} = \beta_{RFlex} = \beta_{LFlex} = 0$. These four predictors collectively are significant at the 5% level

3(c)

$H_0 : \beta_{RStr} = \beta_{LStr}$ *vs.* $H_1 : \beta_{RStr} \neq \beta_{LStr}$ Since p-value is $0.468 > 0.05$, we do not have enough evidence to reject the null hypothesis that right and left strength have the same effect at 95% level of significance.

```
model.3c = lm(Distance~I(RStr+LStr)+RFlex+LFlex, data=punting)
anova(model.3c,model.3a)
```

```
## 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
```

3(d)

The test in 3(c) tests test whether the right and left strength have the same effect. Since the line *RStr*=*LStr* has intersection with the confidence region, we do not have enough evidence to reject the null hypothesis that right and left strength have the same effect.

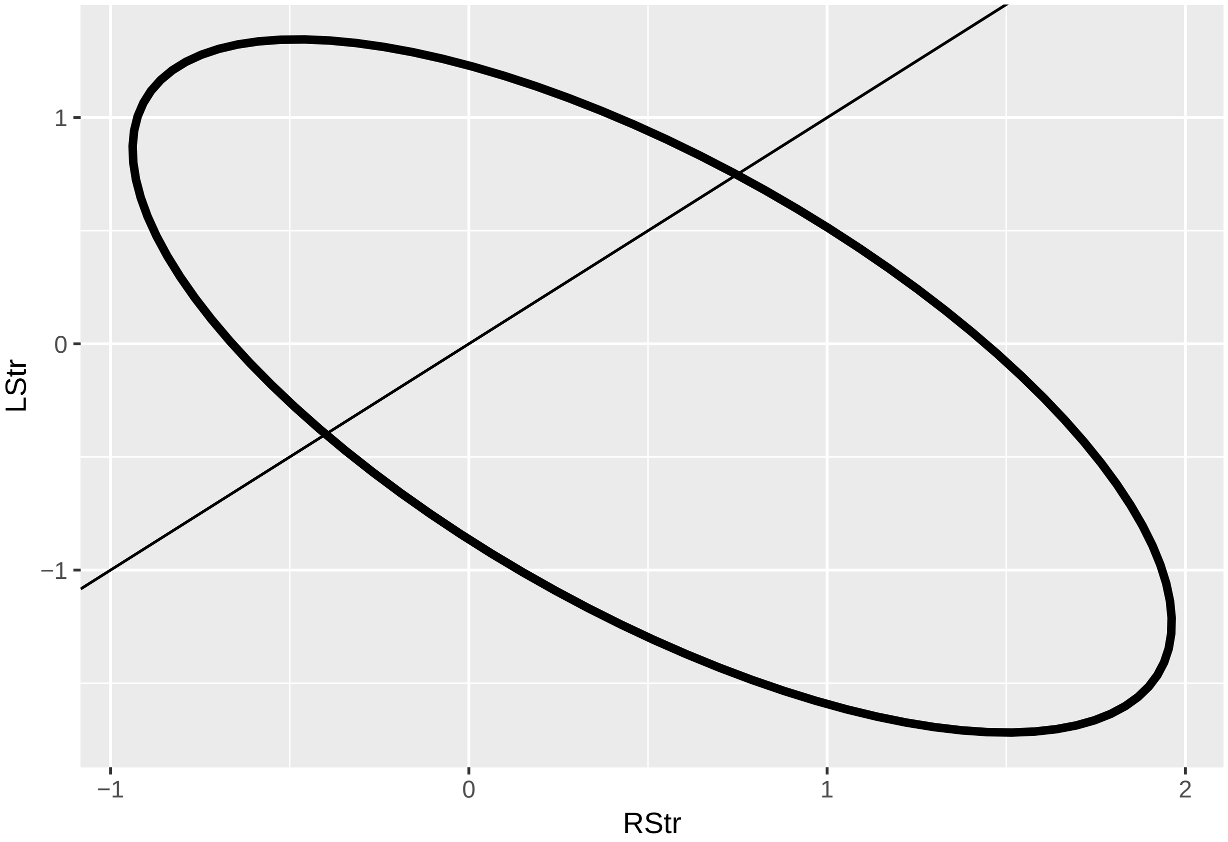
```
CR95.3d = ellipse(model.3a, c(2,3))
CR95.3d = data.frame(CR95.3d)
dim(CR95.3d)
```

```
## [1] 100   2
```

```
head(CR95.3d)
```

```
##        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
```

```
ggplot(data=CR95.3d, aes(x=RStr, y=LStr)) +
  geom_path(size=1.5) +
  geom_abline(intercept = 0, slope = 1)
```



3(e)

The test result of 3(c) indicates that *RStr* and *LStr* have the same effect with present of *RFlex* and *LFlex*. Now we test whether they are same without present of *RFlex* and *LFlex*. F-stat has p-value $0.5978 > 0.05$ indicates that we do not have enough evidence to reject the null hypothesis that *RStr* and *LStr* have the same effect without present of *RFlex* and *LFlex*. Then, we test whether total leg strength alone is sufficient. F-stat has p-value $0.2694 > 0.05$ indicates that we do not have enough evidence to reject the null hypothesis and we conclude that total leg strength alone is sufficient to predict the response at 95% level of significance, in comparison to using individual left and right strengths.

```
model.3e.total = lm(Distance~I(RStr+LStr),data=punting)
model.3e.small = lm(Distance~RStr+LStr,data=punting)
model.3e.large = lm(Distance~I(RStr+LStr)+RFlex+LFlex,data=punting)
anova(model.3e.total,model.3e.small)
```

```
## 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
```

```
anova(model.3e.total,model.3e.large)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: Distance ~ I(RStr + LStr)
## Model 2: Distance ~ I(RStr + LStr) + RFlex + LFlex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     11 3061.3
## 2      9 2287.4  2       774 1.5227 0.2694
```

3(f)

$H_0 : \beta_{RFlex} = \beta_{LFlex}$ *vs.* $H_1 : \beta_{RFlex} \neq \beta_{LFLex}$ Since p-value is $0.2017 > 0.05$, we do not have enough evidence to reject the null hypothesis that right and left leg flexibilities have the same effect at 95% level of significance.

```
model.3f = lm(Distance~RStr+LStr+I(RFlex+LFlex), data=punting)
anova(model.3f,model.3a)
```

```
## 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
```

3(g)

$H_0 : \beta_{RFlex} = \beta_{LFlex}$ *and* $\beta_{RStr} = \beta_{LStr}$

Since p-value is $0.337 > 0.05$, we do not have enough evidence to reject the null hypothesis that left and right is symmetric at 95% level of significance.

```
model.3g = lm(Distance~I(RStr+LStr)+I(RFlex+LFlex), data=punting)
anova(model.3g,model.3a)
```

```
## 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
```

3(h)

Since none of p-value of the predictors are less than 0.05, none of the predictors are significant at the 5% level. We cannot compare this model with model in 3(a) since this model is not nested in 3(a) and the model in 3(a) is not nested in this model.

```
model.3h = lm(Hang~RStr+LStr+RFlex+LFlex,data=punting)
summary(model.3h)
```

```
##
## 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
```

Problem 5

The predicted *lpsa* for this patient is 2.389053 with 95% CI (2.172437, 2.605669).

5(a)

```
model.5a = lm(lpsa~.,data=prostate)
new_patient=data.frame(lcavol=1.44692,lweight=3.62301,age=65,lbph=0.3001,
                       svi=0,lcp=-0.79851,gleason=7,pgg45=15)
predict.lm(model.5a,newdata=new_patient,interval='confidence',level=0.95)
```

```
##        fit      lwr      upr
## 1 2.389053 2.172437 2.605669
```

5(b)

The predicted *lpsa* for this patient is 3.272726 with 95% CI (2.260444, 4.285007). Since *age* has min(41), mean(63.87), sd(7.445) for this dataset, *age*=20 deviates a lot from the mean. Hence, it is more difficult for the model to predict and it leads to wider confidence interval.

```
new_patient.5b=data.frame(lcavol=1.44692,lweight=3.62301,age=20,lbph=0.3001,
                       svi=0,lcp=-0.79851,gleason=7,pgg45=15)
predict.lm(model.5a,newdata=new_patient.5b,interval='confidence',level=0.95)
```

```
##        fit      lwr      upr
## 1 3.272726 2.260444 4.285007
```

```
summary(prostate$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   41.00   60.00   65.00   63.87   68.00   79.00
```

5(c)

At 5% level of significance, we only keep *lcavol*, *lweight* and *svi*. The predicted *lpsa* for this patient is 2.372534 with 95% CI (2.197274, 2.547794). Since *age* is not included as predictor for this reduced model, change of *age* would not affect the prediction outcome. The new CI is almost the same as 5(a) and much narrower than 5(b). I would prefer the prediction from this new reduced model since it eliminates the predictors that are not significant at the 5% level. It makes the model more robust to the changes of relatively unimportant variables.

```
summary(model.5a)
```

```
##
## 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
```

```
model.5c = lm(lpsa~lcavol+lweight+svi,data=prostate)
predict.lm(model.5c,newdata=new_patient.5b,interval='confidence',level=0.95)
```

```
##        fit      lwr      upr
## 1 2.372534 2.197274 2.547794
```

STAT 425 HW#2

TIANQI WU

Problem 4:

$$F = \frac{Ms(reg)}{Ms(err)} = \frac{FSS/(P-1)}{RSS/(n-P)} = \frac{(TSS-RSS)}{RSS} \cdot \frac{n-P}{P-1} \quad \textcircled{1}$$

$$R^2 = \frac{FSS}{TSS} = \frac{TSS-RSS}{TSS}$$

$$\frac{R^2}{1-R^2} = \frac{(TSS-RSS)/TSS}{RSS/TSS} = \frac{TSS-RSS}{RSS}$$

Plug in $\textcircled{1}$:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-P}{P-1}$$

where n is # of observations,
P is # of predictors.

Problem 6:

Since $\hat{Y} = HY$.

$$\begin{aligned} Var(\hat{Y} \mid X) &= var(HY \mid X) \\ &= H^T H \, Var(Y \mid X) && \text{since } H \text{ only depends on } X. \\ &= H \, Var(Y \mid X) && \text{since } H^T H = HH = H \\ &= H\sigma^2 && \text{since } var(Y \mid X) = \sigma^2 I \end{aligned}$$