R_Project4_S24_

Zheng Yi Lai

2024-02-26

Question 1. In the punting data from the faraway library, we find the average distance punted 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 leg strengths and flexibilities as predictors. Which predictors are significant at the 5% level?

```
#Import the library and the dataset
library(faraway)
data(punting)
#Fit in the model
lm_1<- lm(Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
summary(lm_1)
```

```
##
## 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 them are at the 5% significant level meaning that all the predictor value are not statistically siginificant and has a p-value that are larger than 0.05.

b) Use an F -test to determine whether collectively these four predictors have a rela- tionship to the response.

```r
#Look at the summary of the linear model
summary(lm_1)
```

```
##
## 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
```

Null hypothesis: All the four predictors does not have a relationship to the response. Alternative hypothesis: All the four predictors have a relationship to the response. Test-Statistics: 5.59 p-value: 0.01902 Conclusion: We reject the null hypothesis since the p-value is 0.0192 and it is smalller than 0.05, thus we favor the alternative hypothesis which saying that the four predictors has a relationship to the response.

(c) Relative to the model in (a), which is the full model, test whether the right and left leg strengths have the same effect using the partial F -test.

```r
#Fit the model
lm_1<- lm(Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
lm_2 <- lm(Distance ~ I(RStr + LStr) + RFlex + LFlex, data = punting)
anova(lm_2,lm_1)
```

```
## 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
```

Null hypothesis: Alternative model with right and left leg strengths have the same effect. Alternative hypothesis: Full model with right and left leg strengths doesn't have the same effect. Test-Statistics: 0.5804 p-value: 0.468 Conclusion: We fail to reject the null hypothesis since the p-value is 0.468 and it is larger than 0.05, thus we favor the null hypothesis which saying that right and left leg strengths have the same effect.

d) Construct a 95% confidence region for (beta_RStr, betaLStr). Explain how the test in (c) relates to this region

```r
#Annova to compare two models
#Import library
library(ellipse)
```

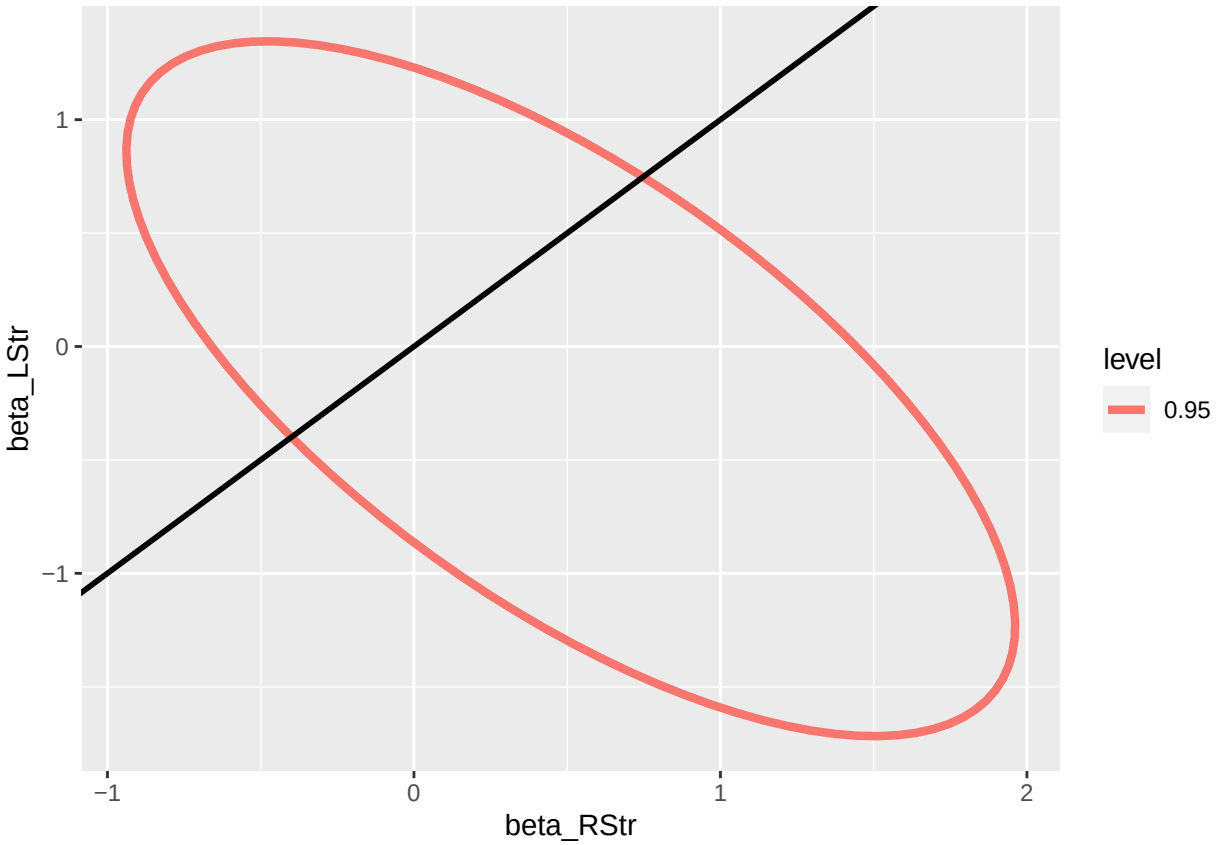
```
##
## Attaching package: 'ellipse'
```

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

```r
library(ggplot2)

#Draw the graph
CR95 = ellipse(lm_1, c(2,3))

myCR = rbind(CR95);
myCR = data.frame(myCR);
names(myCR) = c("beta_RStr","beta_LStr");
myCR[, 'level']=as.factor(c(rep(0.95, dim(CR95)[1])));


ggplot(data=myCR, aes(x=beta_RStr, y=beta_LStr, colour=level)) +
  geom_path(aes(linetype=level), size=1.5) +
  geom_abline(x=coef(lm_1)[2], y=coef(lm_1)[3], size=1, colour='black')
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning in geom_abline(x = coef(lm_1)[2], y = coef(lm_1)[3], size = 1, colour =
## "black"): Ignoring unknown parameters: `x` and `y`
```



Null hypothesis: Alternative model with right and left leg strengths have the same effect. Alternative hypothesis: Full model with right and left leg strengths doesn't have the same effect. Conclusion: We fail to reject the null hypothesis since the line is crossing the confident interval, thus we favor the null hypothesis which saying that right and left leg strengths have the same effect.

(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 leg strengths.

```
#Fit into the model for two situations
lm_3<- lm(Distance ~ RStr + LStr, data = punting)
lm_4 <- lm(Distance ~ I(RStr + LStr), data = punting)
anova(lm_4,lm_3)
```

```
## 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: Alternative model with total leg strength defined by adding the right and left leg strengths. Alternative hypothesis: Full model with individual left and right leg strengths. Test-Statistics: 0.2969 p-value: 0.5978 Conclusion: We fail to reject the null hypothesis since the p-value is 0.5978 and it is larger than 0.05, thus we favor the null hypothesis which saying that total leg strength defined by adding the right and left leg strengths is more sufficient to predict the response.

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

```
#Fit into the model for two situations
lm_1<- lm(Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
lm_5 <- lm(Distance ~ RStr + LStr + I(RFlex + LFlex), data = punting)
anova(lm_5,lm_1)
```

```
## 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
```

Null hypothesis: Alternative model with total leg strength defined by adding the right and left leg flexibilities. Alternative hypothesis: Full model with individual left and right leg flexibilities. Test-Statistics: 1.9346 p-value: 0.2017 Conclusion: We fail to reject the null hypothesis since the p-value is 0.2017 and it is larger than 0.05, thus we favor the null hypothesis which saying that total leg flexibilities defined by adding the right and left leg flexibilities are more sufficient to predict the response.

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

```
#Fit into the model for two situations
lm_1<- lm(Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
lm_6 <- lm(Distance ~ I(RStr + LStr) + I(RFlex + LFlex), data = punting)
anova(lm_6,lm_1)
```

```
## 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: Alternative model with total leg strength defined by adding the right and left leg flexibilities, strength meaning that they are left-right symmetry. Alternative hypothesis: Full model with individual left and right leg flexibilities, strength meaning they are not left-right symmetry. Test-Statistics: 1.25 p-value: 0.337 Conclusion: We fail to reject the null hypothesis since the p-value is 0.337 and it is larger than 0.05, thus we favor the null hypothesis which saying that they are left-right symmetry.

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

```
lm_1<- lm(Distance ~ RStr + LStr + RFlex + LFlex, data = punting)
lm_7 <- lm(Hang ~ RStr + LStr + RFlex + LFlex, data = punting)
anova(lm_7,lm_1)
```

```
## Warning in anova.lmlist(object, ...): models with response '"Distance"' removed
## because response differs from model 1
```

```
## Analysis of Variance Table
##
## Response: Hang
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## RStr       1 1.98540 1.98540 30.0416 0.0005867 ***
## LStr       1 0.19827 0.19827  3.0001 0.1214978
## RFlex      1 0.14699 0.14699  2.2241 0.1742114
## LFlex      1 0.00833 0.00833  0.1260 0.7317905
## Residuals  8 0.52871 0.06609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the two model are not nested relationship thus we can't compare two model using a F-test to compare this model to that used in (a), and while we try to put the model into anova function will pops up a Warning that our response are differ from the model.

Question 2. For the prostate data in faraway library, fit a model with lpsa as the response and the other variables as predictors.

```
library(faraway)
data(prostate)
fit <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate )
summary(fit)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     gleason + pgg45, 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) For patients with the following values: lcavol lweight age lbph svi lcp gleason pgg45 1.44692 3.62301 65 0.30010 0 -0.79851 7 15 Estimate the mean lpsa for the patients along with a 95% CI

```
#Fit the model
fit <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate )

df <- data.frame(
  lcavol=1.44692,
  lweight=3.62301,
  age=65,
  lbph=0.30010,
  svi=0,
  lcp=-0.79851,
  gleason=7,
  pgg45=15
)

# mahalanobis distance -- only relevant to age
```

```
prostate.X = prostate[, 1:8]
mahalanobis(df, colMeans(prostate.X), cov(prostate.X))
```

```
## [1] 1.283065
```

```
mean_lpsa <-predict(fit, df, interval="confidence")
mean_lpsa
```

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

The estimated mean is 2.389053 and the mean will be in the 95% interval between 2.172437 and 2.605669.

(b) Repeat (a) for patients with the same values except that the age is 20. Explain why the CI is wider.

```
#Fit the model
fit <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate )

df <- data.frame(
  lcavol=1.44692,
  lweight=3.62301,
  age=20,
  lbph=0.30010,
  svi=0,
  lcp=-0.79851,
  gleason=7,
  pgg45=15
)

# mahalanobis distance -- only relevant to age
prostate.X = prostate[, 1:8]
mahalanobis(df, colMeans(prostate.X), cov(prostate.X))
```

```
## [1] 48.64388
```

```
mean_lpsa <-predict(fit, df, interval="confidence")
mean_lpsa
```

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

The confident interval is wider because the Mahalobis distance for 2b is larger than 2a, which is 48.64388 larger than 1.283065 in 2a. Points with larger Mahalanobis distances in 2b contribute to the widening of the confidence intervals. The estimated mean is 3.27272 and the mean will be in the 95% interval between 2.260444 and 4.285007.

Question 3. Using the sat dataset from faraway library, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say. Suggest possible improvements or corrections to the model where appropriate.

a) Check for large leverage points.

```
library(faraway)
data(sat)

n = 50; p = 5;
g = lm(total ~ expend + salary + ratio + takers, data = sat)
lev = influence(g) $hat
lev[lev > 2 * p / n]
```

```
##  California Connecticut  New Jersey        Utah
##   0.2821179   0.2254519   0.2220978   0.2921128
```

There are four high leverage points in the data which are Utah, California, New Jersey and Connecticut. Meaning that these four points are the points that they are four most far points from the mean.

(b) Check for outliers

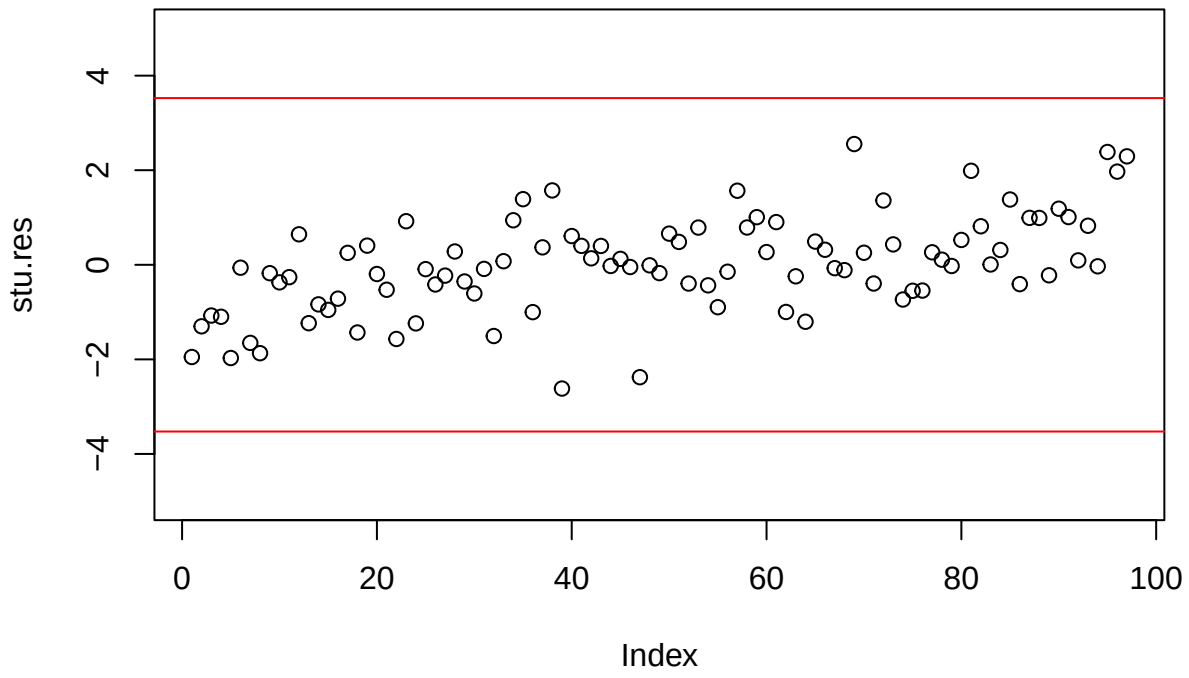
```
library(faraway)
data(sat)

l = lm(sr ~.,data = savings);
n = 50; p = 5;
g = lm(total ~ expend + salary + ratio + takers, data = sat)

jack = rstudent(g);
sort(abs(jack), decreasing = TRUE)[1:5]
```

```
## West Virginia          Utah  North Dakota New Hampshire        Nevada
##      3.124428      2.529587      2.213686      2.190006      1.732004
```

```
stu.res <- rstudent(fit) # studentized residual
cutoff <- qt(0.05/(2*n), (n-1)-p) # Bonferroni adjustment is used, n-1 due to only (n-1) obs used to fi

cutoff
```

```
## [1] -3.525801
```

```
plot(stu.res, ylim=c(-5, 5))
abline(h = c(cutoff, -cutoff), col="red")
```



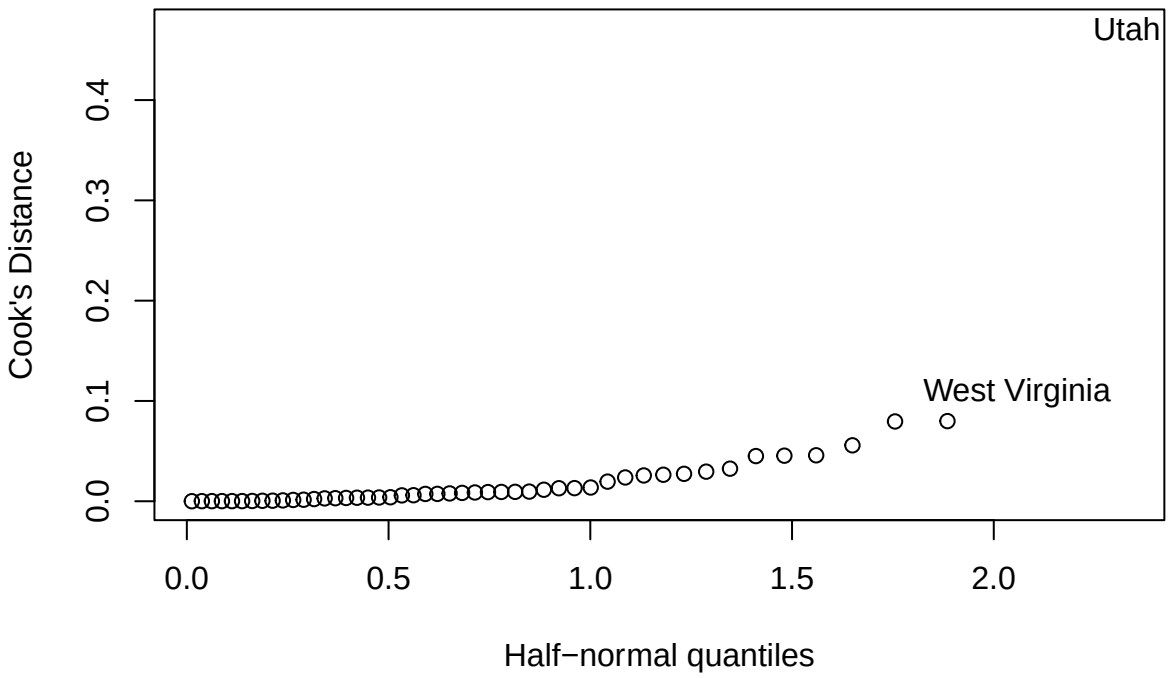
The cutoff point is 3.525801, all the outliers are inside the cutoff region(red line) which is showing in the graph, that means we don't have any outliers.

(c) Check for influential points

```
cook = cooks.distance(g)
max(cook)
```

```
## [1] 0.4715287
```

```
halfnorm(cook, labs = row.names(sat), ylab = "Cook's Distance")
```



Although there are no high influential points based on the rule-of-thumb, the Cook's distance for Utah is much larger than the other samples. The graph also shows that Utah data is far from other data points too.

Possible improvements or corrections to the model where appropriate: Remove Utah, refit the model, and check the changes.