Homework 2

Sarah Adilijiang

Problem 1

(a)

```
# generate simulated data set, there are n = 100 data points

# choose the values of correlation, coefficients, and simga
cor = 0.8
beta0 = 5 ; beta1 = -1 ; beta2 = 2
sigma = 1

# suppose x1, x2 ~ N(0,0,1,1,cor)
library(MASS)
mu = c(0,0)
cov_matrix = matrix(c(1,cor,cor,1),2,2)
bvn = mvrnorm(n=100, mu, cov_matrix)
x1 = bvn[,1]
x2 = bvn[,2]

# linear model for Y
error = rnorm(n=100, mean = 0, sd = sigma)
y = beta0 + beta1 * x1 + beta2 * x2 + error

# fit model Y ~ X1
model1 = lm(y ~ x1)
model1$coefficients[2]
```

```
##        x1
## 0.3676747
```

```
# fit model Y ~ X1 + X2
model2 = lm(y~ x1 +x2)
model2$coefficients[2]
```

```
##        x1
## -1.151557
```

Answer:

Under the above parameters setting, if fit a linear model of Y on covariate X_1 only , the fiited slope is generally positive, and if fit a linear model of Y on both covariates X_1 and X_2, the coefficient on X_1 is generally negative.

(b)

Answer:

For example, Y is the exam score of a student, X_1 is the time the student is present in office hour in hours per week, X_2 is the time the student spend on study in hours per week. In this case, if fit a linear model of Y on X_1 only , the fiited slope is plausibly positive, since the more time in office hour may help a student better understand the course materials thereby get a higher exam score. However, if fit a linear model of Y on both X_1 and X_2, the coefficient on X_1 is plausibly negative, because when including the time the student spend

on study per week, a student who study more time a week may have already had a better understanding of course materials so he/she is less likely to go to the office hour for further help.

Problem 2

```
library(faraway)
data(prostate)
model = lm(lpsa ~ ., data = prostate)
summary(model)
```

```
##
## 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)

```
# 90% Confidence Interval for parameter associated with "age"
confint(model, "age", level = 0.90)
```

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

```
# 95% Confidence Interval for parameter associated with "age"
confint(model, "age", level = 0.95)
```

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

Answer:

In regression summary, H_0: $\beta_{age} = 0$

90% CI does not contain 0, Reject H_0 at $\alpha = 0.1$ level, so p-value < 0.1

95% CI contains 0, Do Not Reject H_0 at $\alpha = 0.05$ level, so p-value > 0.05

Therefore, we can deduce that $0.05 <$ p-value < 0.1

Indeed, the regression summary shows that the parameter associated with age has p-value $= 0.08229$

(b)

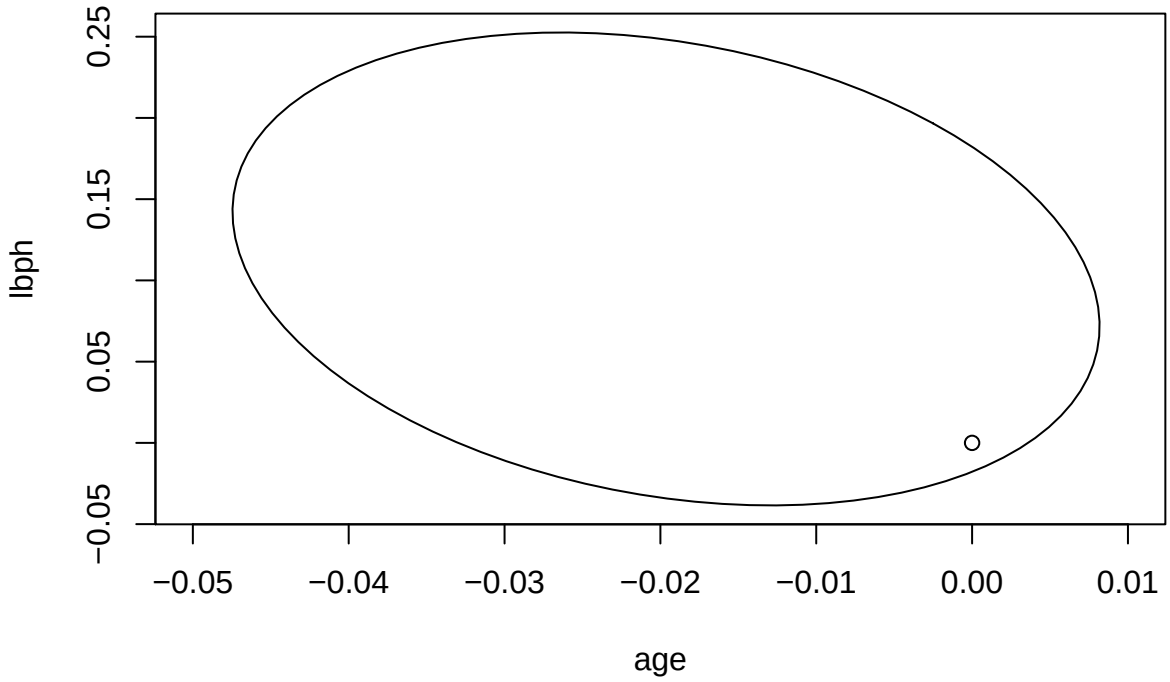
```
# plot the 95% joint confidence region
library(ellipse)
```

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

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

```
plot( ellipse(model, c(4,5)), type = "l", xlim = c(-0.05,0.01)) # level = 0.95

# plot the origin
points(0,0)
```



Answer:

The test is that H_0: $\beta_{age} = \beta_{lbph} = 0$

According to the plot, the origin lies inside the ellipse, so we Do Not Reject H_0 at $\alpha = 0.05$ level

(c)

```
new_x = data.frame(lcavol=1.44692,lweight=3.62301,age=65.00000,
                   lbph=0.30010,svi=0.00000,lcp=-0.79851,
                   gleason=7.00000,pgg45=15.00000)
predict(model, new_x, interval = "prediction", level = 0.95)
```

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

Answer:

The predicted value of lpsa is 2.389053, and the 95% prediction interval is (0.9646584, 3.813447)

(d)

```
new_x = data.frame(lcavol=1.44692,lweight=3.62301,age=20.00000,
                   lbph=0.30010,svi=0.00000,lcp=-0.79851,
                   gleason=7.00000,pgg45=15.00000)
predict(model, new_x, interval = "prediction", level = 0.95)
```

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

Answer:

The predicted value of lpsa is 3.272726, and the 95% prediction interval is (1.538744, 5.006707)

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

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

Answer:

Because age=65 is within the observation range of "age" data but age=20 is already out of the observation range of "age" data, so the prediction interval for age=20 is wider.

On the other hand, the mean of age is 63.87, so age=20 is much farther away from the mean than age=65, therefore the prediction interval for age=20 is wider.

Problem 3

(a)

```
library(faraway)
data(teengamb)
model = lm(gamble ~ ., data = teengamb)
summary(model)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
```

```
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

Answer:

Variables "sex" and "income" are statistically significant.

(b)

Answer:

```
help(teengamb)
```

```
## starting httpd help server ... done
```

Answer:

According to the introduction of dataset "teengamb", we see that for variable "sex": 0=male and 1=female. And the variable "gamble" represents the expenditure on gambling in pounds per year.

So the coefficient of "sex", which equals to -22.11833, means that when the other covariates are not changed, the average expenditure on gambling in pounds per year for a female is 22.11833 pounds lower than that of a male.

(c)

```
# a male with average status, income and verbal score
x_new_ave = data.frame(sex=0,status=mean(teengamb$status),
                       income=mean(teengamb$income),verbal=mean(teengamb$verbal))
predict(model, x_new_ave, interval = "prediction", level = 0.95)
```

```
##        fit       lwr      upr
## 1 28.24252 -18.51536 75.00039
```

```
# a male with maximal values of status, income and verbal score
x_new_max = data.frame(sex=0,status=max(teengamb$status),
                       income=max(teengamb$income),verbal=max(teengamb$verbal))
predict(model, x_new_max, interval = "prediction", level = 0.95)
```

```
##        fit      lwr    upr
## 1 71.30794 17.06588 125.55
```

Answer:

The predicted value of gamble for a male with average status, income and verbal score is 28.24252 pounds per year. And the 95% prediction interval is (-18.51536, 75.00039) pounds per year.

The predicted value of gamble for a male with maximal values of status, income and verbal score is 71.30794 pounds per year. And the 95% prediction interval is (17.06588, 125.55) pounds per year.

The prediction interval for a male with maximal values of status, income and verbal score is wider because the maximal data values are farther away from the observation range of data than the average data values.

(d)

```
model2 = lm(gamble~income, data = teengamb)
anova(model2, model)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ income
## Model 2: gamble ~ sex + status + income + verbal
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 28009
## 2     42 21624  3    6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The F-test p-value is $0.01177 < 0.05$, so we Reject the null model (only with income as a predictor) at $\alpha = 0.05$ significance level comparing with the full model.

Problem 4

Shown in the next page.

STAT 34300 HW2

Problem 4.

(a) $\hat{y}^{(0)} - \hat{y}^{(1)} = x^{(0)T}\hat{\beta} - x^{(1)T}\hat{\beta} = (x^{(0)} - x^{(1)})^T \hat{\beta}$

(b) set fixed vetor $v = x^{(0)} - x^{(1)} \in \mathbb{R}^p$ $\therefore$ $v^T\hat{\beta} \sim N(v^T\beta, \sigma^2 (x^Tx)^T v)$

$\Rightarrow$ a $(1-\alpha)$ level confidence interval for $\hat{y}^{(0)} - \hat{y}^{(1)} = v^T\hat{\beta}$ is:

$$v^T\hat{\beta} \pm t^*_{1-\frac{\alpha}{2}}(n-p) \cdot \hat{\sigma} \cdot \sqrt{v^T (x^Tx)^{-1} v}$$

$$= (x^{(0)} - x^{(1)})^T \hat{\beta} \pm t^*_{1-\frac{\alpha}{2}}(n-p) \cdot \hat{\sigma} \cdot \sqrt{(x^{(0)} - x^{(1)})^T (x^Tx)^{-1} (x^{(0)} - x^{(1)})}$$

(c) a $(1-\alpha)$ level prediction interval for $\hat{y}^{(0)} - \hat{y}^{(1)} = v^T\hat{\beta}$ is:

$$v^T\hat{\beta} \pm t^*_{1-\frac{\alpha}{2}}(n-p) \cdot \hat{\sigma} \cdot \sqrt{1 + v^T (x^Tx)^{-1} v}$$

$$= (x^{(0)} - x^{(1)})^T \hat{\beta} \pm t^*_{1-\frac{\alpha}{2}}(n-p) \cdot \hat{\sigma} \cdot \sqrt{1 + (x^{(0)} - x^{(1)})^T (x^Tx)^{-1} (x^{(0)} - x^{(1)})}$$