PSTAT220A HW 3

sbsambado

11/6/2021

Full disclaimer: I really botched questions 1 - 3 (theoretical-ish stuff), I think I'm still not grasping what is expected for questions like that (i.e. what kind of work is required, general concepts) but question 4 - 5 (applied stuff) I felt much better about.

Question 1

Part A. I am a bit confused if this was suppose to be in R or by hand.

Model: yi = XB + e, i = 1,. . . ,n

Matrix model: y = $\sum_{j=1}^{p} \beta_j x_j + \epsilon$

```
# expanded
#  Y               XB                   e
# ---  =  ---------------------   +   ----
# [Y1]    [x1,0   x1,1    x1,2]       [e1]
# [Y2]    [x2,0   x1,1    x1,2]       [e2]
# [..]    [....   ....    ....]       [..]
# [Yn]    [xn,0   xn,1    xn,2]       [en]
```

Part B Was confused about this question. Should have asked during office hours!

Attempt: We estimate parameters β by minimizing the least squares.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_i j)^2 = ||y - X\beta||^2$$

Taking derivative with respect to β and setting it equal to zero, we have the normal equation:

$$x^T x\beta = x^T y$$

Assume that X is full column rank. Then the LS estimates of β is

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

Part C Was confused about this question. Should have asked during office hours!

Atempt: Under the hypothesis $H_0 : A\beta = 0$, to test all covariate effects equal to zero, the extra sum of squares should be

$$RSS_{H_0} - RSS = SS_{total} - RSS = SS_{model}$$

where

$$RSS_{H_0} = \sum_{i=1}^{n}(y_i - y_{bar})^2 = SS_{total}$$

Question 2

I am confused about this question.

Question 3

Part A. I am a bit confused if this was suppose to be in R or by hand.

Model: yi = XB + e, i = 1,. . . ,n + 1

Matrix model: y = $\sum_{i=1}^{n+1} \beta_j x_j + \epsilon$

```
# expanded
#  Y               XB                  e
# ---  =  ---------------------   +    ----
# [Y1]    [x1,0   x1,1     x1,2]       [e1]
# [Y2]    [x2,0   x1,1     x1,2]       [e2]
# [..]    [....   ....     ....]       [..]
# [Yn]    [xn+1,0 xn +1,1  xn+1,2]     [en]
```

Part B The derived LS estimates of the two mean parameters:

$$\hat{\beta}_1 = \sum_{i=1}^{n} (X_i - x_{bar})(Y_i - Y_{bar}) / \sum_{i=1}^{n} (X_i - x_{bar})^2$$
$$\hat{\beta}_2 = \sum_{i=2}^{n} (X_i - x_{bar})(Y_i - Y_{bar}) / \sum_{i=2}^{n} (X_i - x_{bar})^2$$

$$\hat{\beta}_0 = Y_{bar} - \hat{\beta}_1 X_{bar} - \hat{\beta}_2 X_{bar}$$

Part C. Would Mallow's Cp be an appropriate test statistic for testing the hypothesis that the (n + 1)st observation has the same population mean as the previous observations.It is an unbiased estimate of MSE/sigmaˆ2.

Question 4

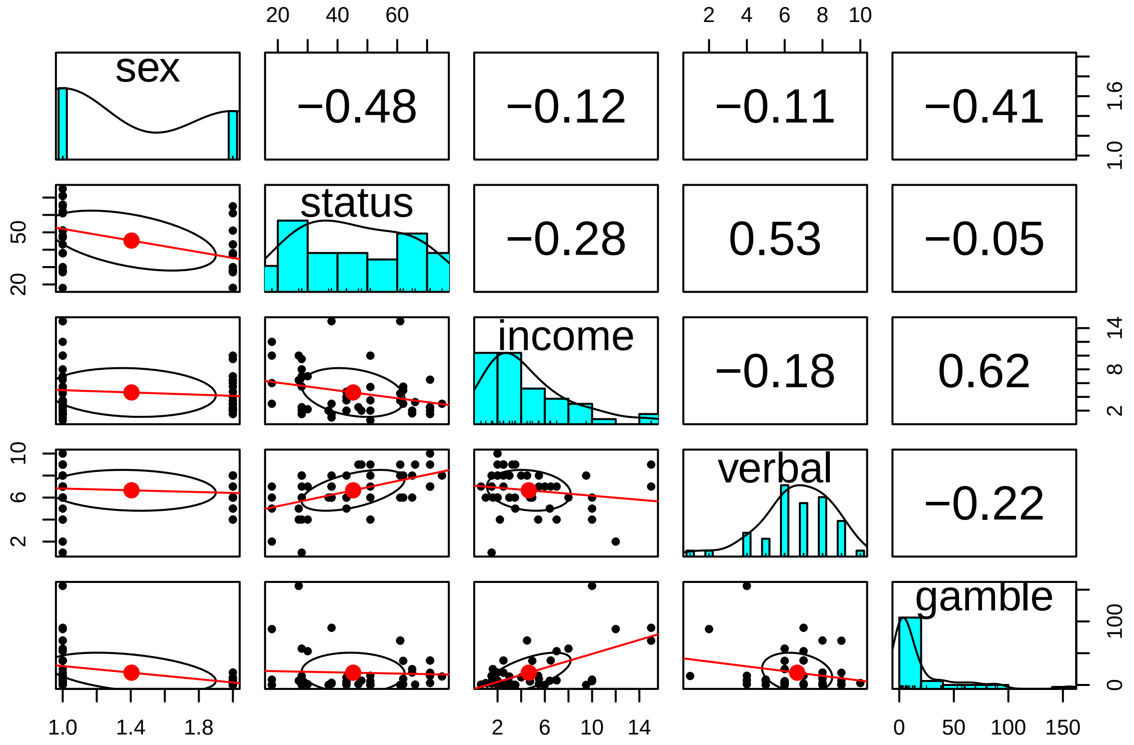
Part A. The teengamb data frame has 47 rows and 5 columns. The survey was conducted to study teenage gambling in Britain. There appears to be a few outliers when evaluating the relationship between gambling expenditure (y) and sex (x) but nothing too concerning. Based on the `pairs.panel()` function from the `car` package, it appears that there is a 1) positive association between gambling expenditure (y) and income in pounds per week (x) (correlation rho = .62) and 2) negative association between gambling expenditure (y) and sex (correlation rho = -.41).

```r
# call in data and omit NAs
teen <- na.omit(teengamb)

# check structure of data
#str(teen)
teen$sex <- factor(teen$sex,
                   levels = c(0,1),
                   labels = c("male","female"))
```

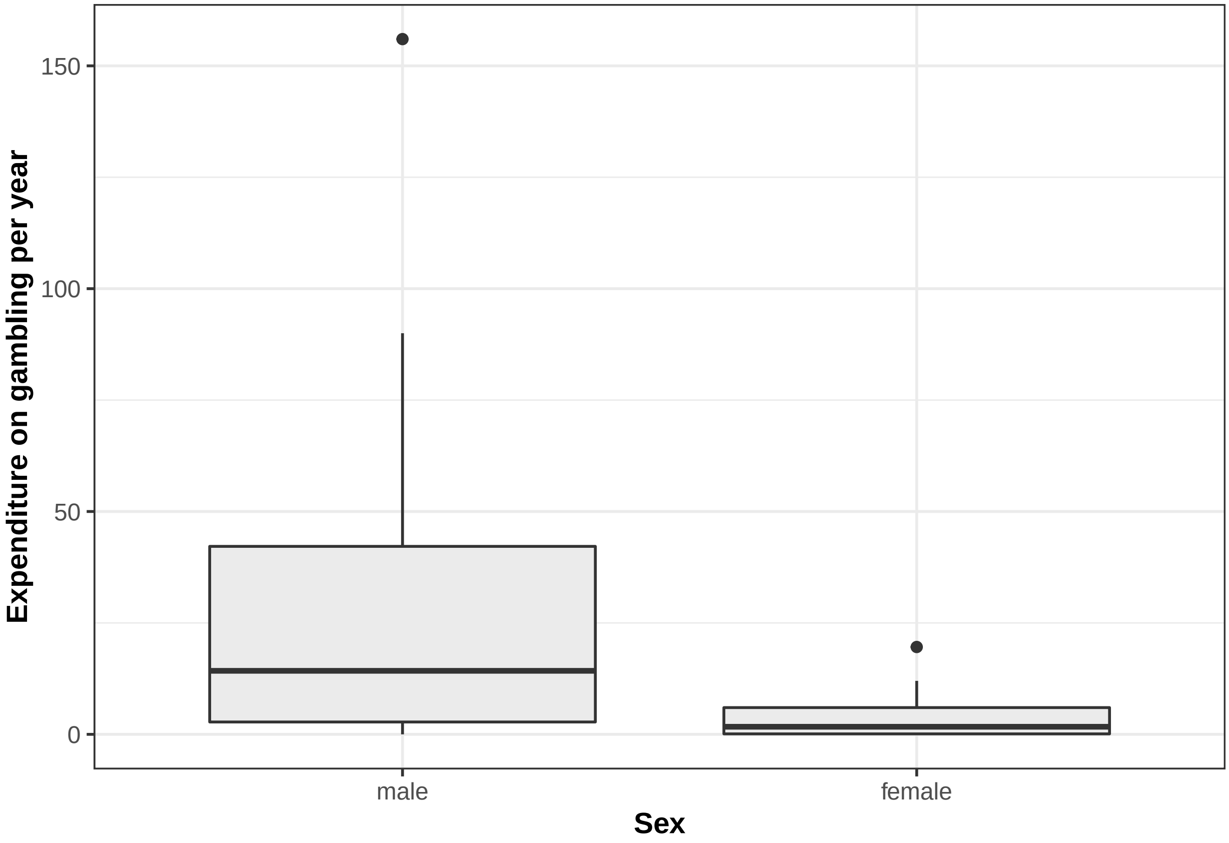
	sex	status	income	verbal	gamble
	male :28	Min. :18.00	Min. : 0.600	Min. : 1.00	Min. : 0.0
	female:19	1st Qu.:28.00	1st Qu.: 2.000	1st Qu.: 6.00	1st Qu.: 1.1
	NA	Median :43.00	Median : 3.250	Median : 7.00	Median : 6.0
	NA	Mean :45.23	Mean : 4.642	Mean : 6.66	Mean : 19.3
	NA	3rd Qu.:61.50	3rd Qu.: 6.210	3rd Qu.: 8.00	3rd Qu.: 19.4
	NA	Max. :75.00	Max. :15.000	Max. :10.00	Max. :156.0

```
# make a numeric summary of data
teen %>%
  group_by(sex) %>%
  summary() %>%
kable() %>%
  kable_styling()
```

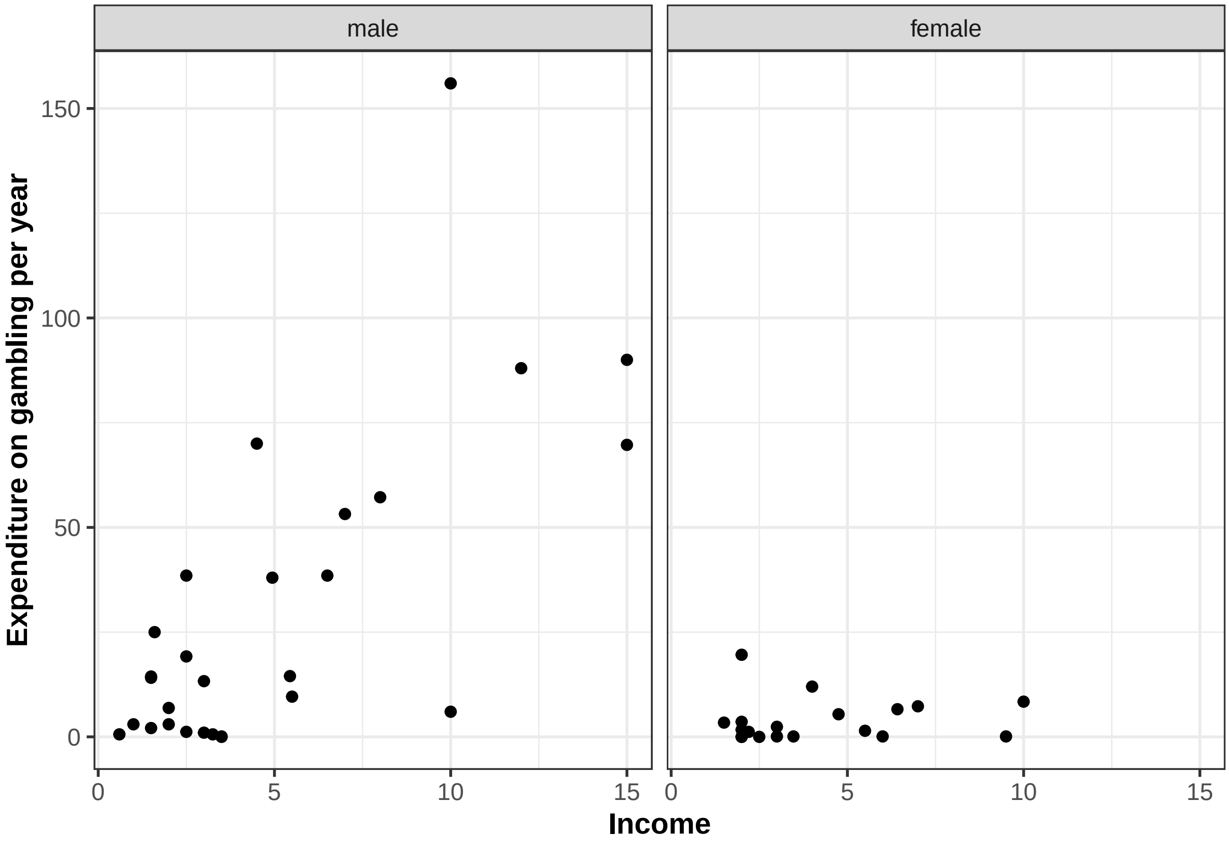
```
# make a graphical summary of data with pairs panel plot

# pairs plot allows me to visualize all variables 1) distribution, 2) linear relationship, and 3) corre

pairs.panels(teen, lm = TRUE)
```



```
# based on the pairs.panels, it seems 1) sex and 2) income have strongest association with gambling exp

# graphical visualization of gambling ~ sex

ggplot(teen, aes(y = gamble, x = sex)) +
  geom_boxplot(fill = "grey92") +
  theme_bw() +
  labs(x = "Sex", y = "Expenditure on gambling per year") +
  theme(axis.title = element_text(face = "bold"))
```



```
# graphical visualization of gambling ~ income and faceted by sex

ggplot(teen, aes(y = gamble, x = income)) +
  geom_point() +
  theme_bw() +
  labs(x = "Income", y = "Expenditure on gambling per year") +
  theme(axis.title = element_text(face = "bold")) +
  facet_wrap(~sex)
```



Part B. I have presented the summary results of the lm model with the function `sum()` in the package `jtools`

```r
# create linear regression model
mod_4b <- lm(gamble ~ income, data = teen)
summary(mod_4b)
```

```
##
## Call:
## lm(formula = gamble ~ income, data = teen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.020 -11.874  -3.757  11.934 107.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.325      6.030  -1.049      0.3
## income         5.520      1.036   5.330 3.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 45 degrees of freedom
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.3734
## F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

```
anova(mod_4b)
```

```
## Analysis of Variance Table
##
## Response: gamble
##           Df Sum Sq Mean Sq F value    Pr(>F)
## income     1  17681 17680.9  28.407 3.045e-06 ***
## Residuals 45  28009   622.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# nicer output of lm model
summ(mod_4b)
```

Observations	47
Dependent variable	gamble
Type	OLS linear regression

F(1,45)	28.41
R^2	0.39
Adj. R^2	0.37

	Est.	S.E.	t val.	p
(Intercept)	-6.32	6.03	-1.05	0.30
income	5.52	1.04	5.33	0.00

Standard errors: OLS

Part C. When I computed the LS estimate with the formula below, I got a different LS estimate than the estimates in part c. For part b my estimates for intercept and income were -6.325 and 5.520, respectively. The LS estimates I calculated for part c are 0.0424 (which is positive) and -0.00455 (which is negative), respectively for intercept and income.

```
X <- model.matrix(mod_4b)
betahat <- solve(t(X)%*%X) %*% t(X) # %*% y
betahat[,1]
```

```
##  (Intercept)       income
##  0.042414699 -0.004553746
```

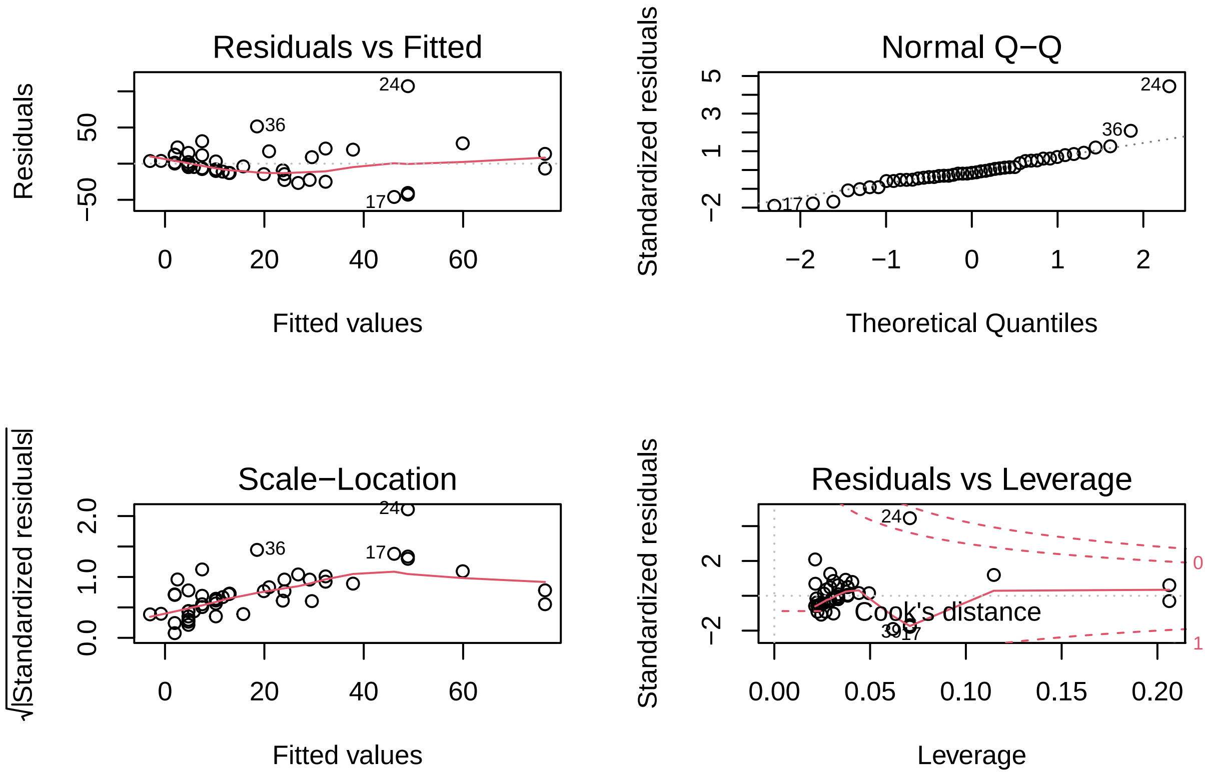
```
mean(betahat)
```

```
## [1] 0.0106383
```

Part D. The percentage of variation in the response is explained by the covariate income is 39%.

Part E. The observation with the largest absolute residual is case number 24, with 36 being another influential residual.

```
# plot model diagnostics
par(mfrow = c(2,2))
plot(mod_4b)
```



Part F. The mean of the residuals from our model is -5.203801e-16 and the median of the residuals is -3.757382. The median is a smaller residual meaning it is a better fit. Because the median is smaller than the mean, I assume there are outliers that are skewing the mean and the outliers are not a good fit.

```
mean(mod_4b$residuals) # -5.203801e-16
```

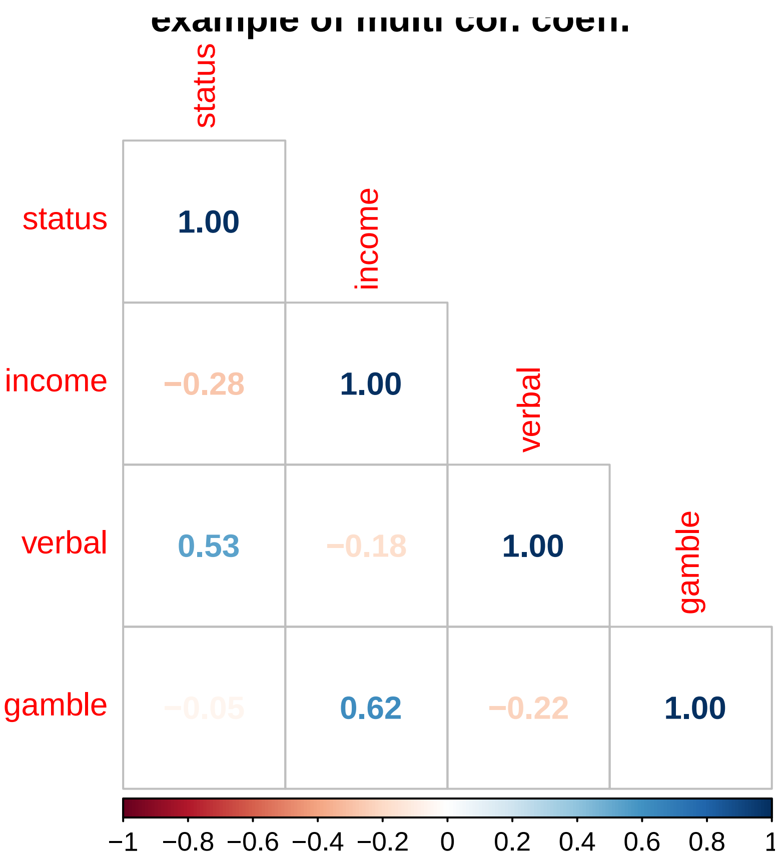
```
## [1] -5.203801e-16
```

```
median(mod_4b$residuals) # -3.757382
```

```
## [1] -3.757382
```

Part G. Rˆ2 is the square of the multiple correlation coefficient which is defined as the sample correlation coefficient betweeny and y_hat. The multiple correlation coefficient is a measure of how well a given individual variable can be predicted using a linear function of a set of other variables. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables.

```
t <- cor(teen[,2:5])
corrplot::corrplot(t, method = "number", type = "lower", main = "example of multi cor. coeff.")
```



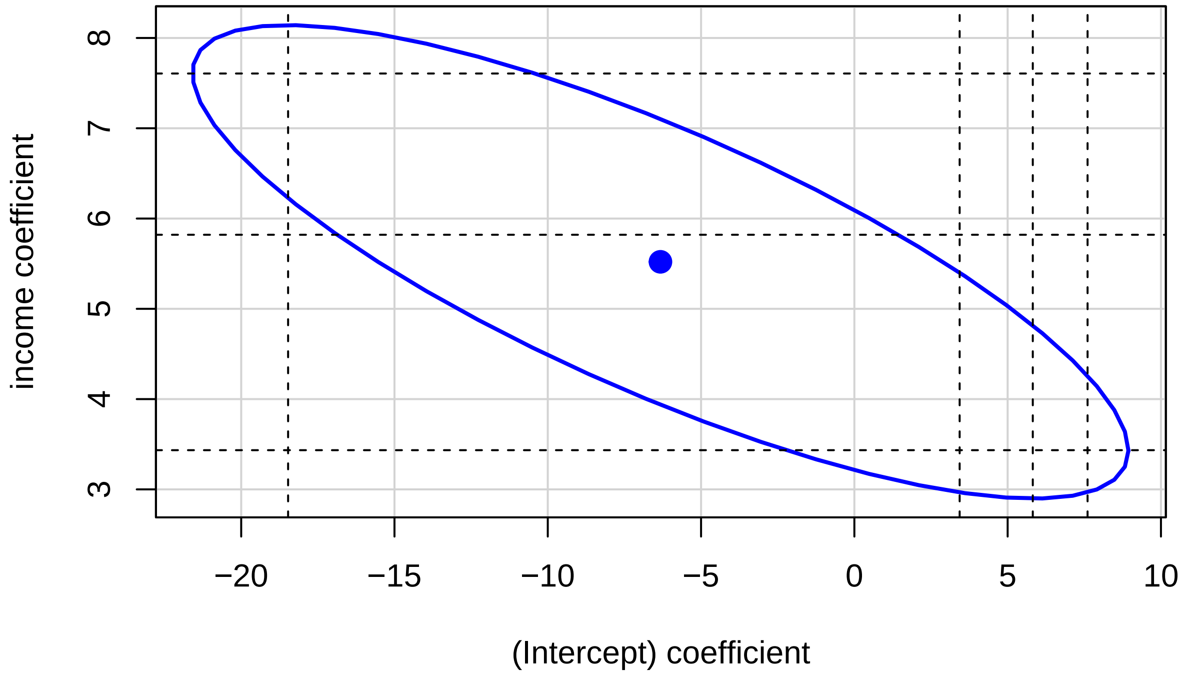
Part H. The 99% confidence intervals (CI) for the intercept are -22.54 - 9.89. The 99% CI for the slope parameter income are 2.73 - 8.31.

```
# 99% CI
confint(mod_4b, level = .99)
```

```
##                   0.5 %   99.5 %
## (Intercept) -22.542419 9.893300
## income        2.734687 8.306283
```

Part I. I have plotted the confidence region for the intercept on x axis and income on the y axis using the library `car`.

```
confidenceEllipse(mod_4b)
abline(v = confint(mod_4b), lty = 2)
abline(h = confint(mod_4b), lty = 2)
```



```
# check correlation between covariates and estimates
#> cor(Wind,Temp)
#> summary(fit3,corr=T)$corr
```

Part J. To compare pointwise and simultaneous bands I would use the code below. Not sure why my code isn't working. I think there is an issue when I make the grid and it doesn't keep the value income which prevents me from using the `predict()` function. Sorry about that! The predict mean of gambling is 19.301 if I calculate it manually.

```
# cheating way to kinda calculate predicted mean
mean(predict(mod_4b))
```

```
## [1] 19.30106
```

```
# intended way to calculate predicted future mean, but it didn't work

# x <- model.matrix(mod_4b)
# grid <- seq(min(x),max(x),len=100)
#
#
# p1 <- predict(mod_4b, newdata=data.frame(x=grid), se=T,
#               interval="confidence")
# p2 <- predict(mod_4b, newdata=data.frame(x=grid), se=T,
#               interval="prediction")
#
# matplot(grid,p1$fit,lty=c(1,2,2),col=c(1,2,2),type="l",
#   xlab="Income",ylab="Gambling",
#   ylim=range(p1$fit,p2$fit,y))
# points(x,y,cex=.5)
# title("Prediction of mean response")
# matplot(grid,p2$fit,lty=c(1,2,2),col=c(1,2,2),type="l",
#   xlab="Income",ylab="Gambling",
#   ylim=range(p1$fit,p2$fit,y))
# points(x,y,cex=.5)
# title("Prediction of future observations")
#
# # compare pointwise and simultaneous bands
# # Sheffe’s method is used
# matplot(grid,p1$fit,lty=c(1,2,2),col=c(1,2,2),type="l",
#   xlab="Body Weight (kg)",ylab="Heart Weight (gm)")
# points(x,y,cex=.5)
# lines(grid,
#       p1$fit[,1]-sqrt(2*qf(.95,2,length(x)-2))*p1$se.fit,
#       lty=3, col="blue")
# lines(grid,
#       p1$fit[,1]+sqrt(2*qf(.95,2,length(x)-2))*p1$se.fit,
#       lty=3, col="blue")


# To compare pointwise and simultaneous bands I would use Sheffe's method

# matplot(grid, p1$fit, lty = c(1,2,2), col = c(1,2,2), type = "l")
# points(x,y, cex = .5)
# lines(grid,
#       p1$fit[,1]-sqrt(2*qf(.95,2,length(x)-2))*p1$se.fit,
#       lty = 3, col = "blue")
# lines(grid,
#       p1$fit[,1]+sqrt(2*qf(.95,2,length(x)-2))*p1$se.fit,
#       lty = 3, col = "blue")
```

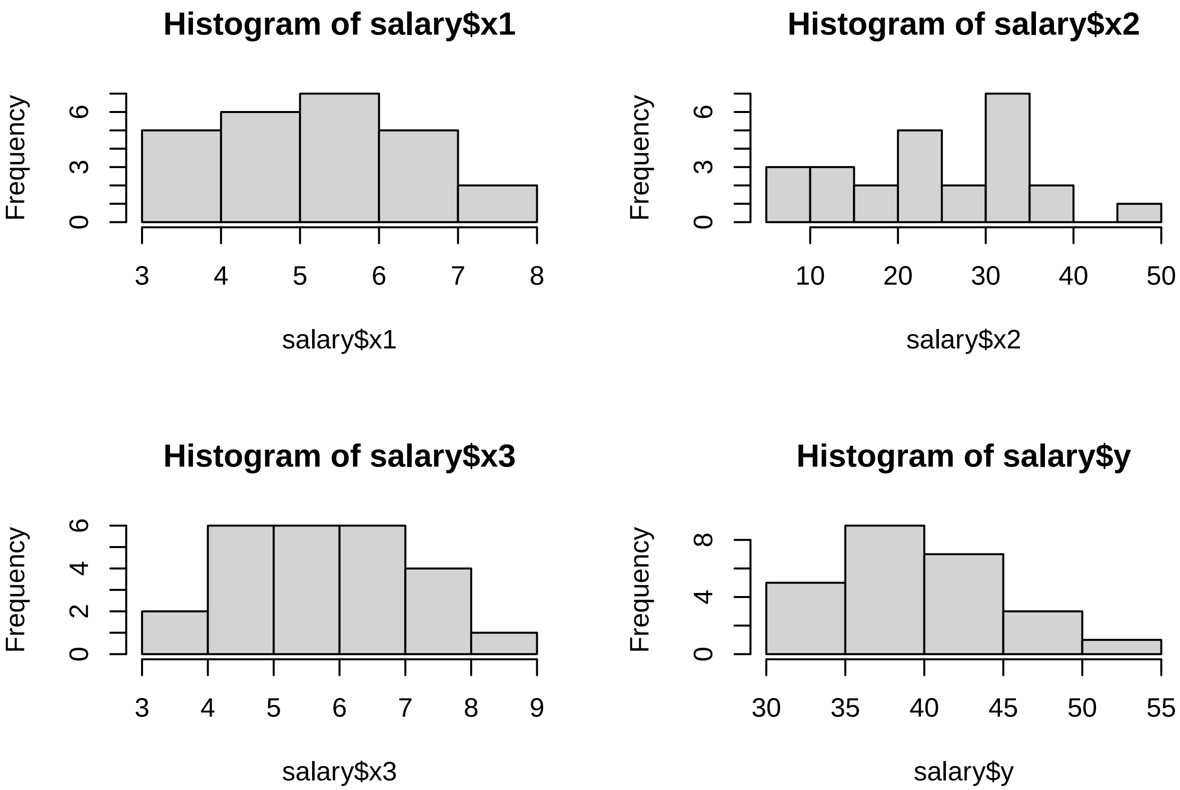
Question 5

Part A. I see that none of the columns/variables are normally distributed before data untransformtion. The outcome variable, y, is right skewed meaning a log transformation may be most appropriate.

```
# call in data and organize
salary <- read.table("salaries.data.csv")
str(salary)
```

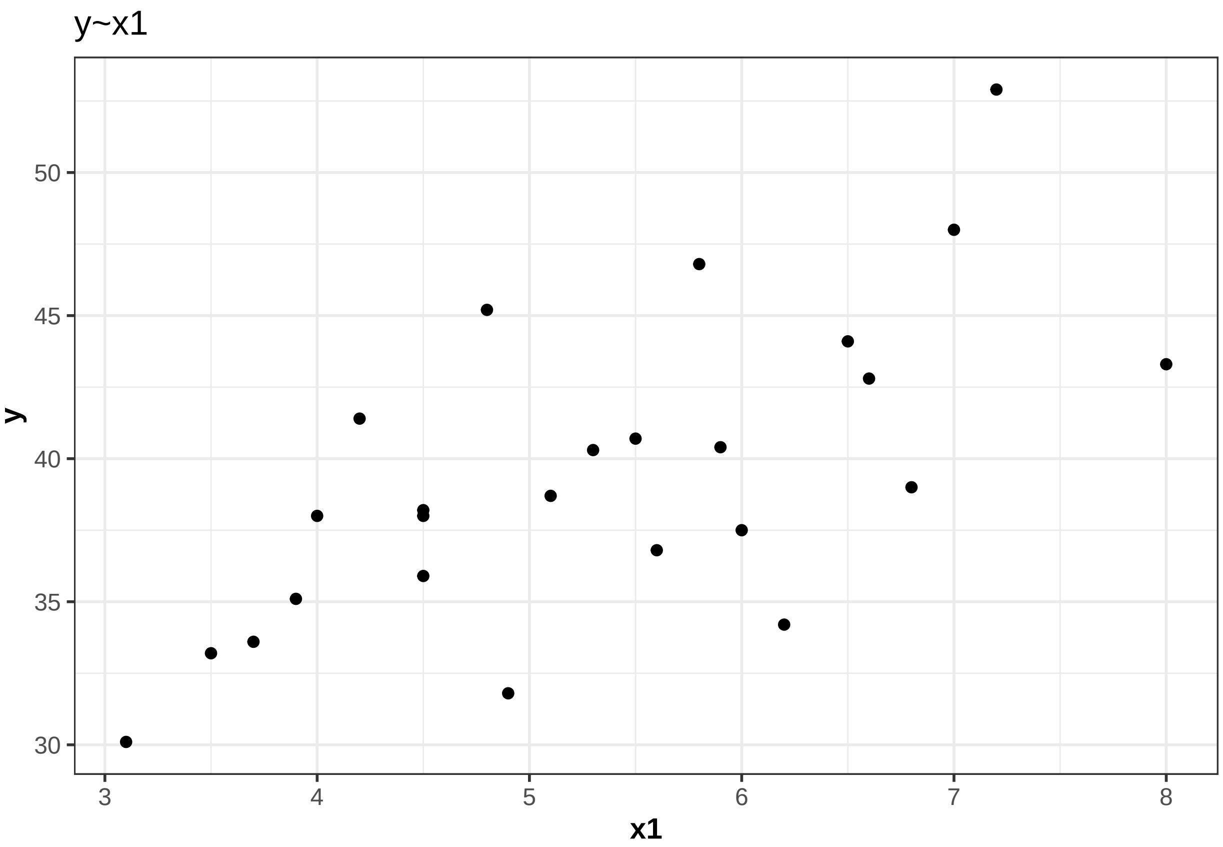
```
## 'data.frame':    25 obs. of  4 variables:
##  $ V1: num  3.5 5.3 5.1 5.8 4.2 6 6.8 5.5 3.1 7.2 ...
##  $ V2: int  9 20 18 33 31 13 25 30 5 47 ...
##  $ V3: num  6.1 6.4 7.4 6.7 7.5 5.9 6 4 5.8 8.3 ...
##  $ V4: num  33.2 40.3 38.7 46.8 41.4 37.5 39 40.7 30.1 52.9 ...
```

```
names(salary) <- c("x1","x2","x3","y")

# create histograms
par(mfrow = c(2,2))
hist(salary$x1)
hist(salary$x2)
hist(salary$x3)
hist(salary$y)
```

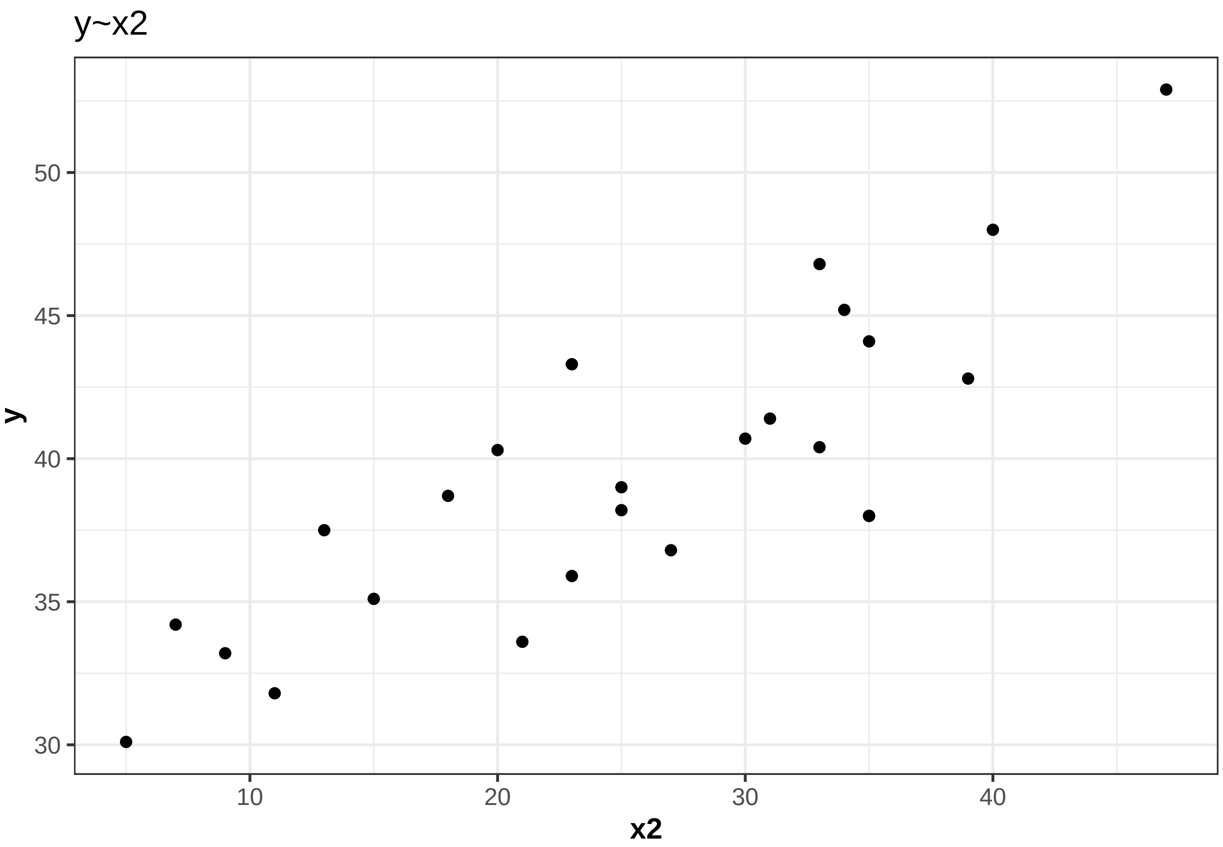


Part B. Based on the scatterplots, it appears the relationship between y ~ x2 may be the most positive linear association. The second most positive linear association is between y ~ x1.

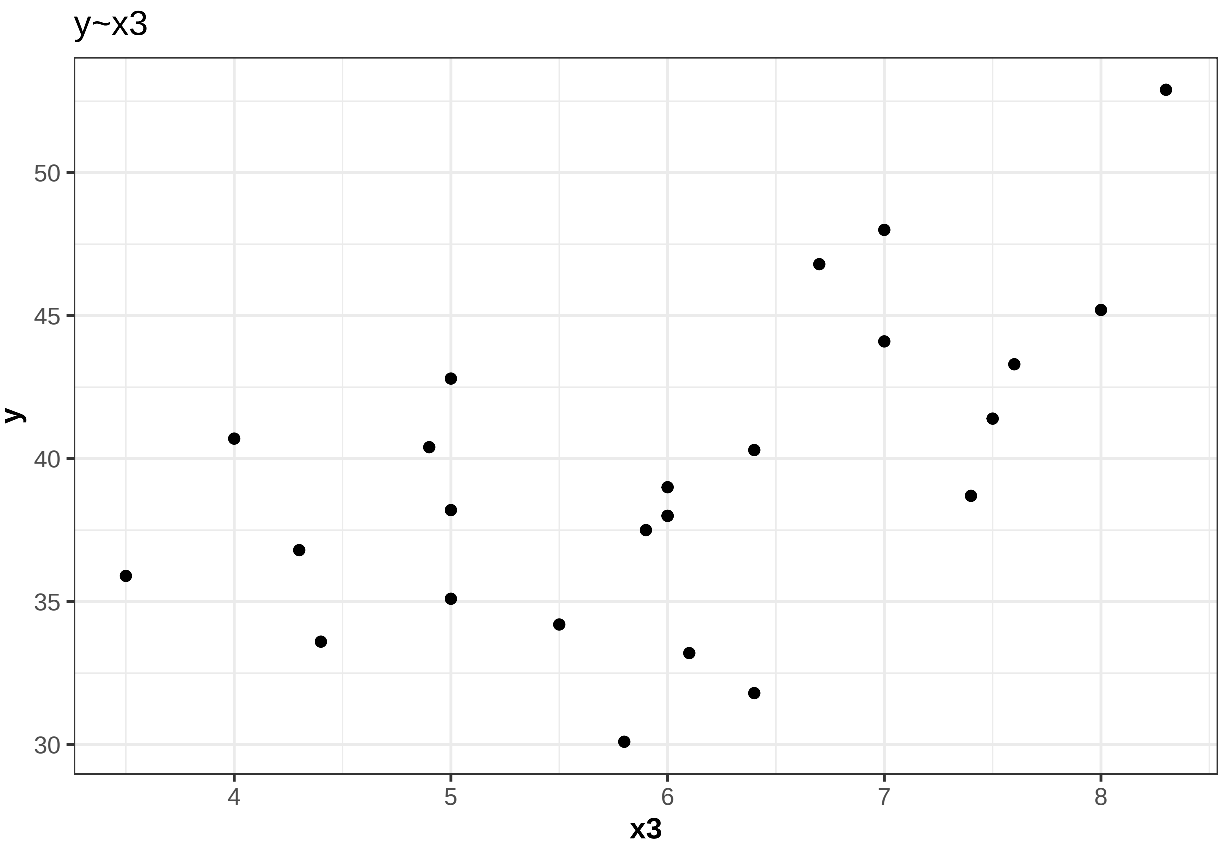
```r
# y ~ x1
ggplot(salary, aes(y = y, x = x1)) +
  geom_point() +
  theme_bw() +
  theme(axis.title = element_text(face = "bold"))+
  ggtitle("y~x1")
```



```
# y ~ x2
ggplot(salary, aes(y = y, x = x2)) +
  geom_point() +
  theme_bw() +
  theme(axis.title = element_text(face = "bold"))+
  ggtitle("y~x2")
```



```
# y ~ x3
ggplot(salary, aes(y = y, x = x3)) +
  geom_point() +
  theme_bw() +
  theme(axis.title = element_text(face = "bold"))+
  ggtitle("y~x3")
```

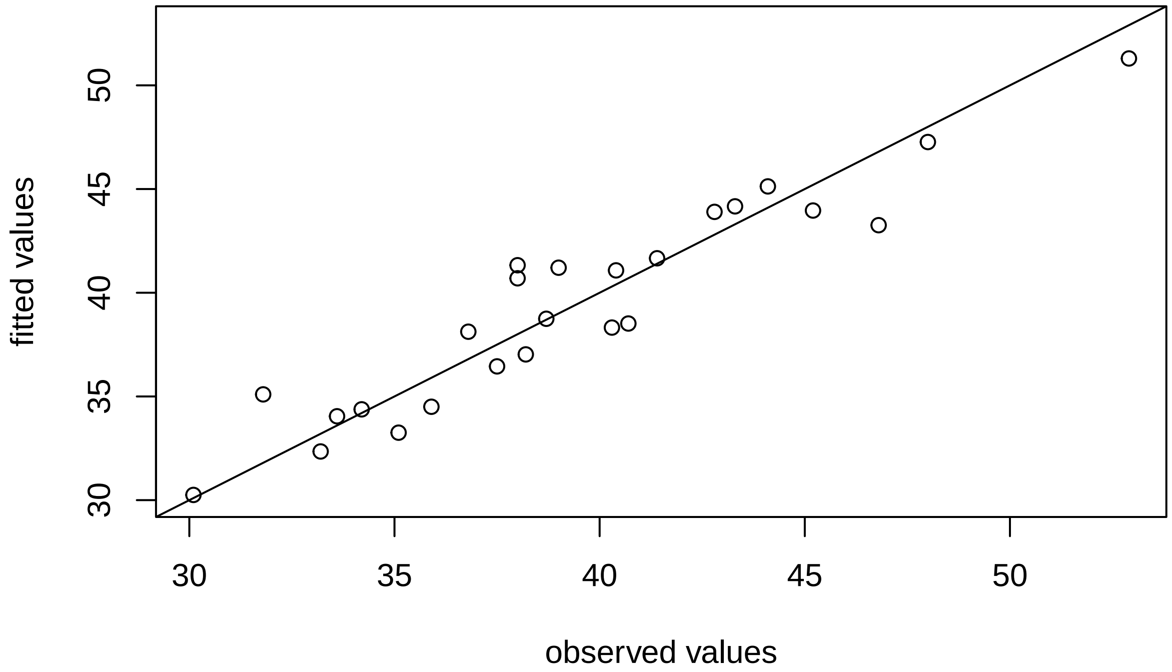


Part C. I fit a linear regression model called mod_5c with 3 explanatory variables (x1,x2,x3) to the response variable y using `lm()` function.

```
mod_5c <- lm(y ~ x1 + x2 + x3, data = salary)
```

Part D. I think the model fits the data well. Looking at the plot the fitted values fall close to the 1:1 line. There are a couple points that are on average farther away than the rest such as when x ~32 and ~ 37.

```
plot(salary$y, mod_5c$fitted.values, data = salary,
     xlab = "observed values", ylab = "fitted values",
     xlim = range(salary$y, mod_5c$fitted.values),
     ylim = range(salary$y, mod_5c$fitted.values))
abline(0,1)
```



Part E. Variables x1, x2, and x3 are all statistically significant. However, x2 is the strongest association with y due to the lowest p-value of 1.08e-07. Followed by x3 (p = 0.000642) and x1 (p = 0.001420).

```
summary(mod_5c)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3261 -1.0274 -0.1519  1.2361  3.5426
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.40780    2.13249   8.163 5.95e-08 ***
## x1           1.26031    0.34324   3.672 0.001420 **
## x2           0.30179    0.03837   7.865 1.08e-07 ***
## x3           1.28073    0.31980   4.005 0.000642 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.878 on 21 degrees of freedom
## Multiple R-squared:  0.8929, Adjusted R-squared:  0.8776
## F-statistic: 58.34 on 3 and 21 DF,  p-value: 2.344e-10
```

Part F. The 90% confidence interval for parameter x2 are 0.24 - 0.37. Based on this interval, I would deduce that x2 is a strong predictor since the confidence interval is a relatively small range. This other two parameters have larger confidence intervals. This makes sense based on the regression summary since x2 had the lowest p-value.

```
confint(mod_5c, parm = "x2", level = .90)
```

```
##           5 %      95 %
## x2 0.2357596 0.3678147
```

```
confint(mod_5c, parm = "x3", level = .90)
```

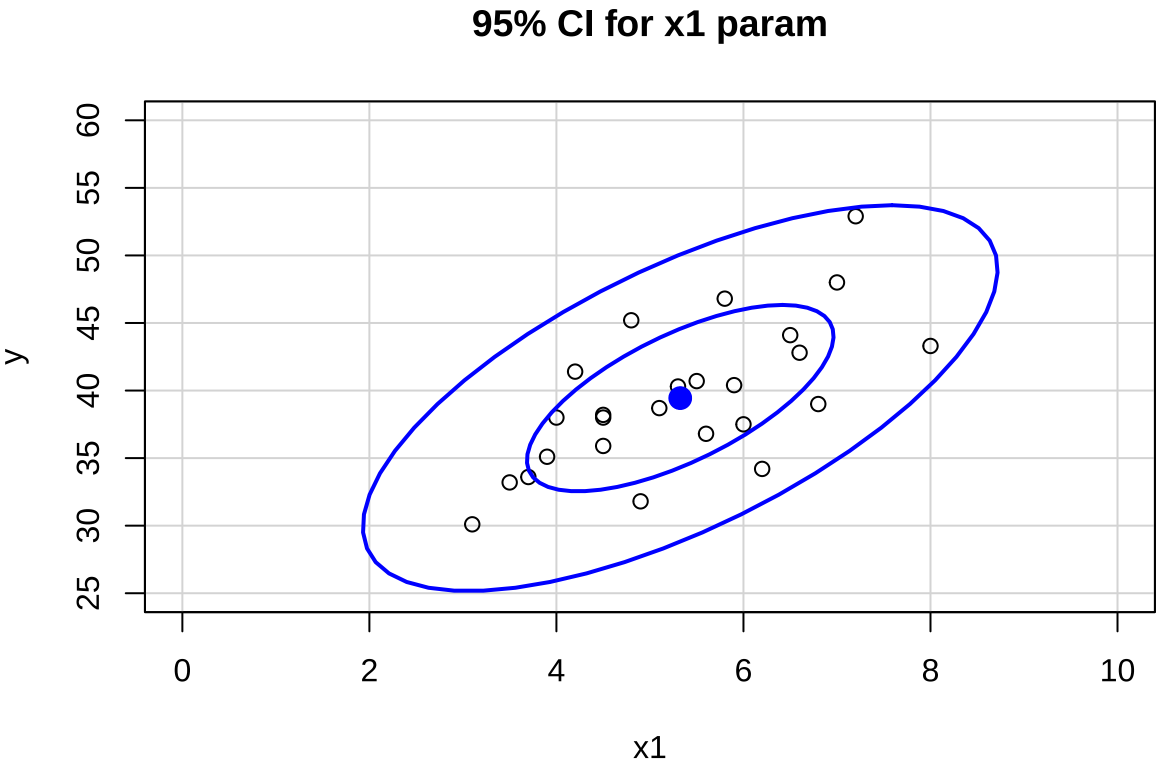
```
##           5 %     95 %
## x3 0.7304341 1.831023
```

```
confint(mod_5c, parm = "x1", level = .90)
```

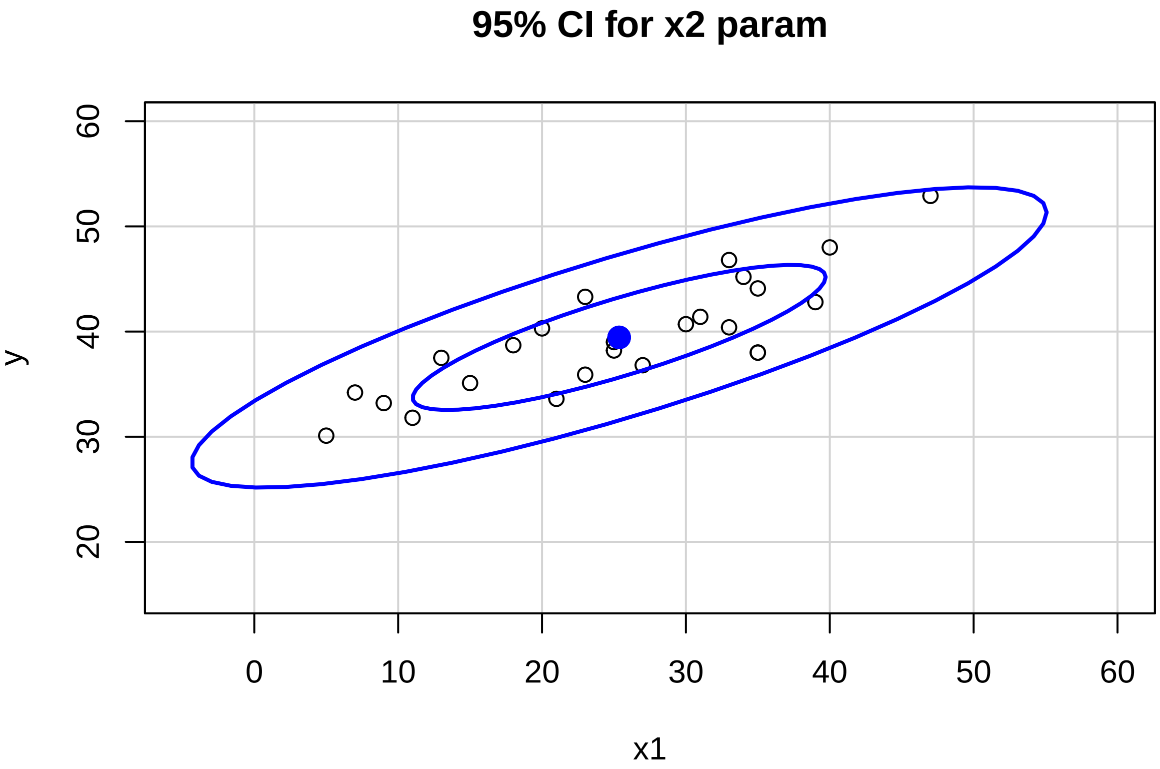
```
##           5 %    95 %
## x1 0.6696767 1.85094
```

Part G. I plotted the 95% joint confidence region with blue bands for parameters associated with x1 and x2 with the blue dot representing the origin. The problem is that there is an infinite number of ellipses that will contain 95% of the data points. However, the origin gives the best fit of the confidence interval as the mean vector and the radius from that origin is the square root of the chisquare value at 0.05 with 2 degrees of freedom.

```
car::dataEllipse(salary$x1 ,salary$y, levels = c(0.55, 0.955),
                 xlim = c(0,10),
                 ylim = c(25,60),
                 xlab = "x1", ylab = "y",
                 main = "95% CI for x1 param")
```

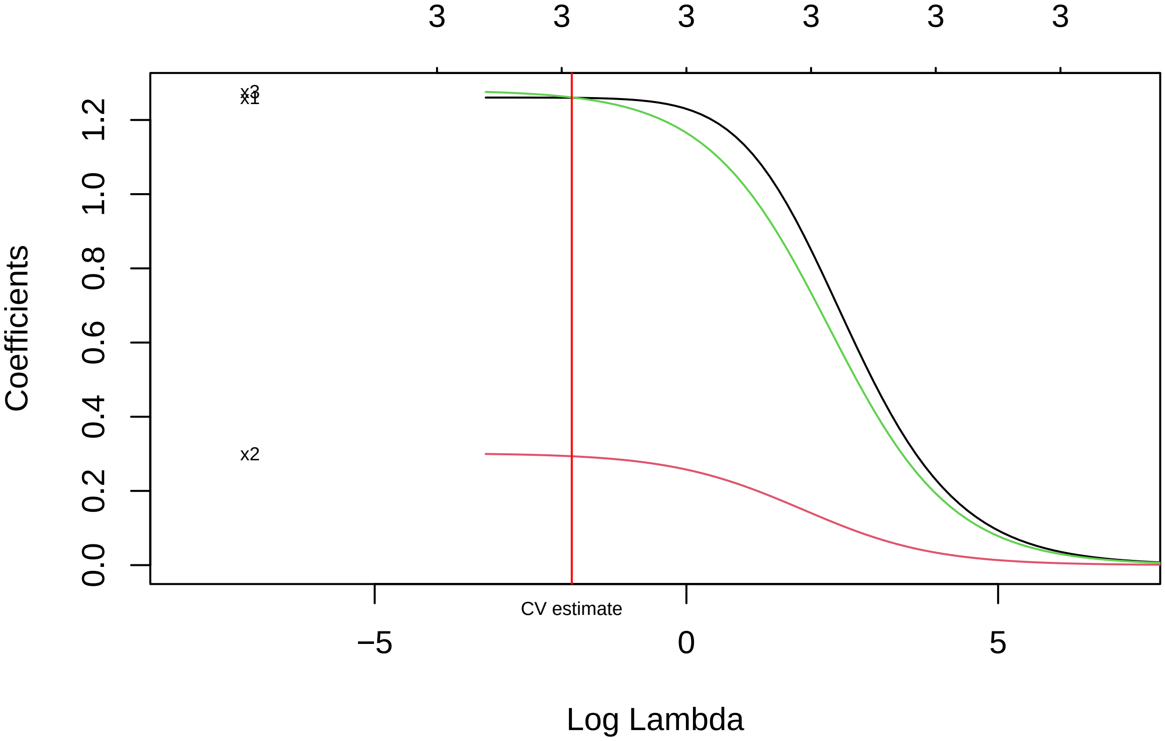


```
car::dataEllipse(salary$x2 ,salary$y, levels = c(0.55, 0.955),
                 xlim = c(-5,60),
                 ylim = c(15,60),
                 xlab = "x1", ylab = "y",
                 main = "95% CI for x2 param")
```

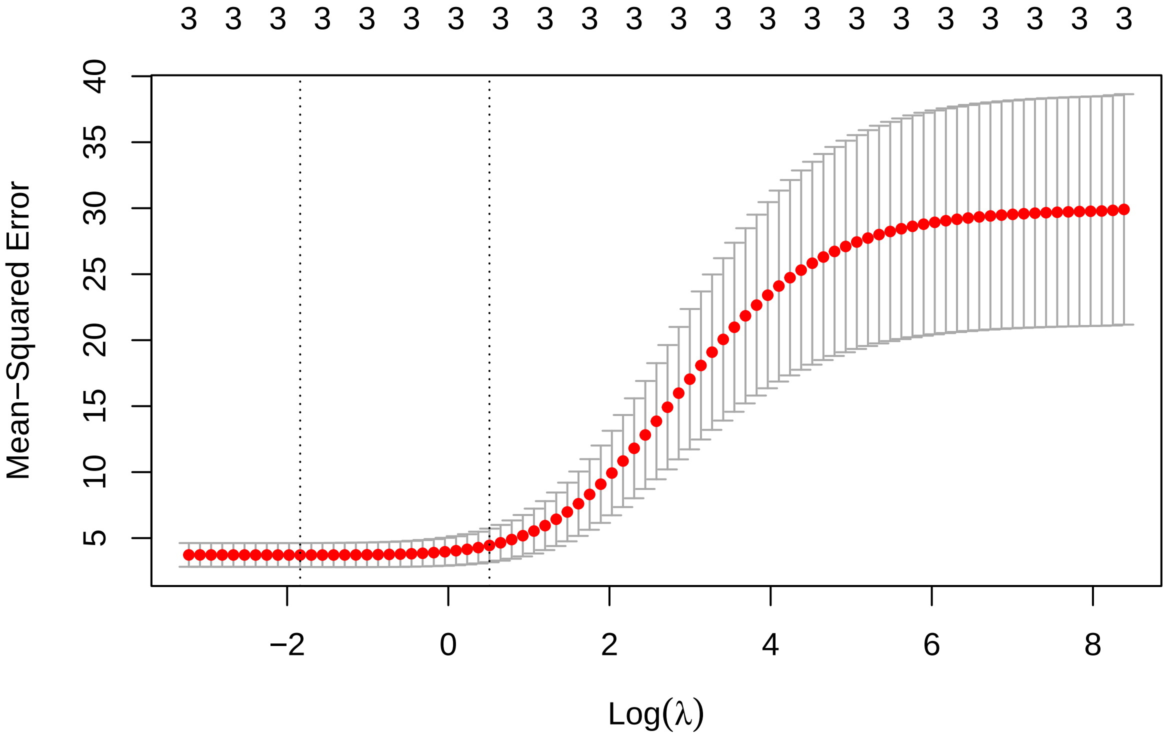


Part H. I would use ridge regression to keep all covariates to obtain simultaneos interval estimates of the mean salary levels assuming that all coefficients should not equal 0 since this is a small dataset and are probably somewhat influential in salary prediction. We can see the CV estimate includes all 3 parameters.

```
library(glmnet)
X <- model.matrix(mod_5c)[,-1]
fit.ridge <- glmnet(X, salary$y, lambda.min=0, nlambda=101, alpha=0)
plot(fit.ridge, xvar="lambda", xlim=c(-8,7))
text(-7,coef(fit.ridge)[-1,length(fit.ridge$lambda)],labels=colnames(X),cex=0.6)


fit.ridge.cv <- cv.glmnet(X, salary$y, lambda.min=0, nlambda=101, alpha=0)
abline(v=log(fit.ridge.cv$lambda.min), col="red")
mtext("CV estimate", side=1, at=log(fit.ridge.cv$lambda.min), cex=.6)
```



```
plot(fit.ridge.cv)
```



Part I. Based on the model estimates and intercept, when given the information provided in the assignment, that research mathematician can claim they are grossly underpaid. Their salary of 35,000 is less than the calculated salary of 39,000 when given the model.

```r
# pulling up B0, B1, B2, B3
summary(mod_5c)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3261 -1.0274 -0.1519  1.2361  3.5426
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.40780    2.13249   8.163 5.95e-08 ***
## x1           1.26031    0.34324   3.672 0.001420 **
## x2           0.30179    0.03837   7.865 1.08e-07 ***
## x3           1.28073    0.31980   4.005 0.000642 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.878 on 21 degrees of freedom
## Multiple R-squared:  0.8929, Adjusted R-squared:  0.8776
## F-statistic: 58.34 on 3 and 21 DF,  p-value: 2.344e-10
```

```r
# given values
x1 = 7
x2 = 10
x3 = 7.9


# linear equation to solve for salary
x = 17.40780 + 1.26031*x1 +  0.30179*x2 + 1.28073*x3
x
```

```
## [1] 39.36564
```