MA678 homework 01

yourname

Septemeber 6, 2018

Introduction

For homework 1 you will fit linear regression models and interpret them. You are welcome to transform the variables as needed. How to use `lm` should have been covered in your discussion session. Some of the code are written for you. Please remove `eval=FALSE` inside the knitr chunk options for the code to run.

This is not intended to be easy so please come see us to get help.

Data analysis

Pyth!

```
gelman_example_dir<-"http://www.stat.columbia.edu/~gelman/arm/examples/"
pyth <- read.table (paste0(gelman_example_dir,"pyth/exercise2.1.dat"),
                    header=T, sep=" ")
```

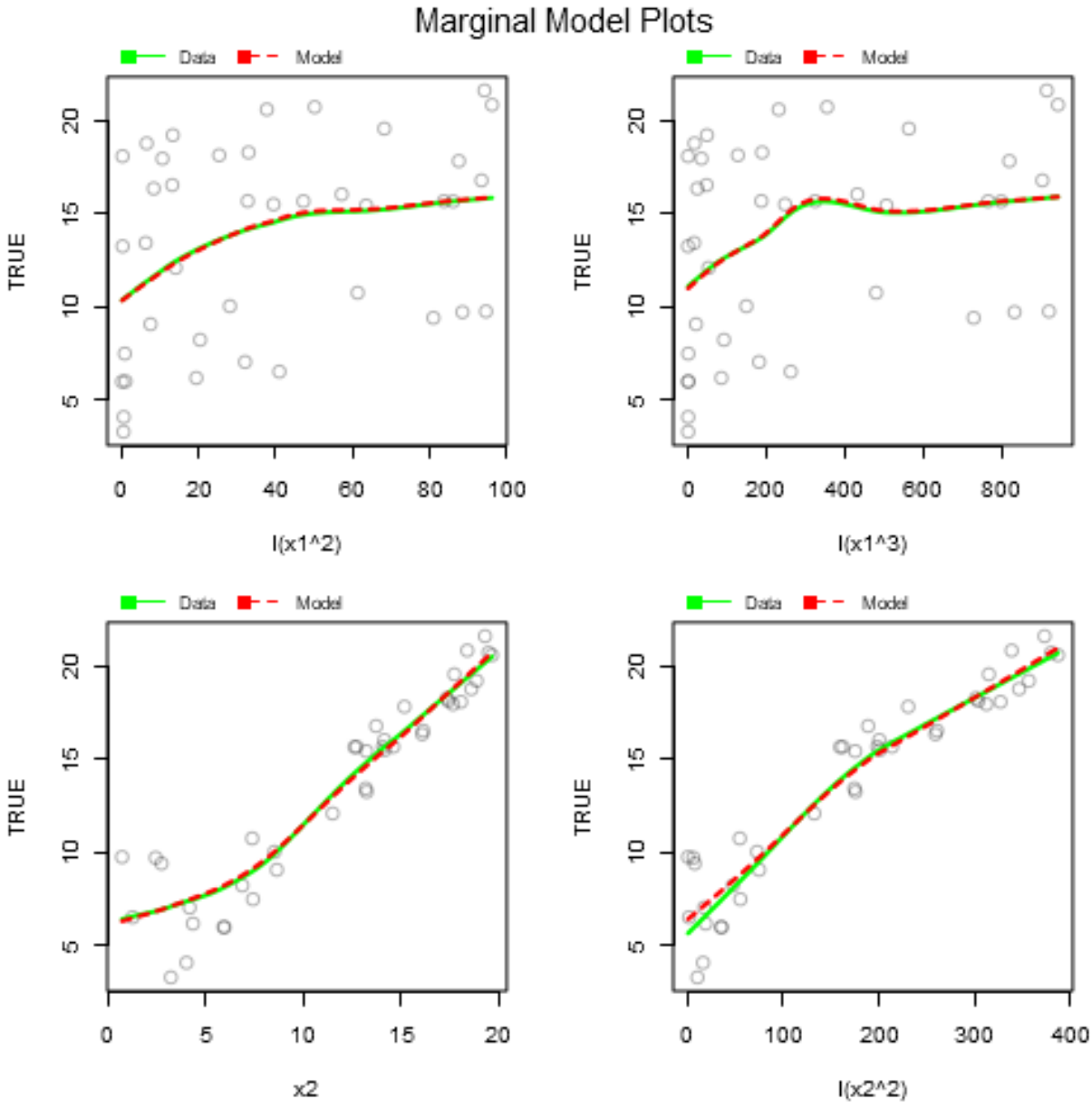
The folder pyth contains outcome `y` and inputs `x1`, `x2` for 40 data points, with a further 20 points with the inputs but no observed outcome. Save the file to your working directory and read it into R using the `read.table()` function.

1. Use R to fit a linear regression model predicting `y` from `x1`,`x2`, using the first 40 data points in the file. Summarize the inferences and check the fit of your model.

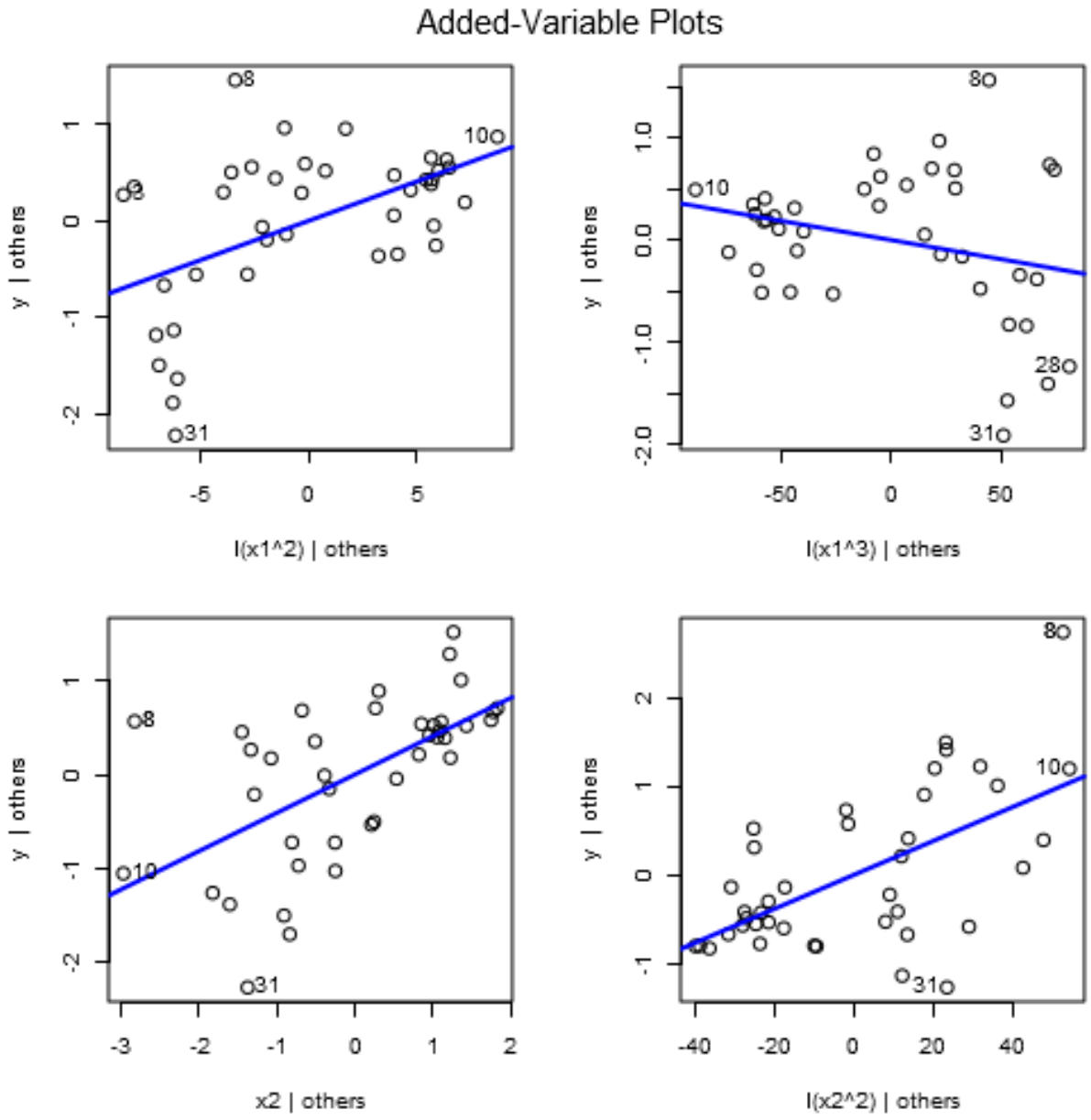
```
fir_40=pyth[1:40,]

model=lm(y ~ I(x1^2)+I(x1^3)+x2+I(x2^2),data = fir_40)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ I(x1^2) + I(x1^3) + x2 + I(x2^2), data = fir_40)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71471 -0.42596 -0.02504  0.48144  1.72788
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.454925   0.472539   7.311 1.52e-08 ***
## I(x1^2)      0.081619   0.022900   3.564 0.001078 **
## I(x1^3)     -0.003750   0.002318  -1.618 0.114713
## x2           0.409301   0.093869   4.360 0.000109 ***
## I(x2^2)      0.019133   0.004300   4.450 8.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7395 on 35 degrees of freedom
## Multiple R-squared:  0.9824, Adjusted R-squared:  0.9804
## F-statistic: 488.2 on 4 and 35 DF,  p-value: < 2.2e-16
```

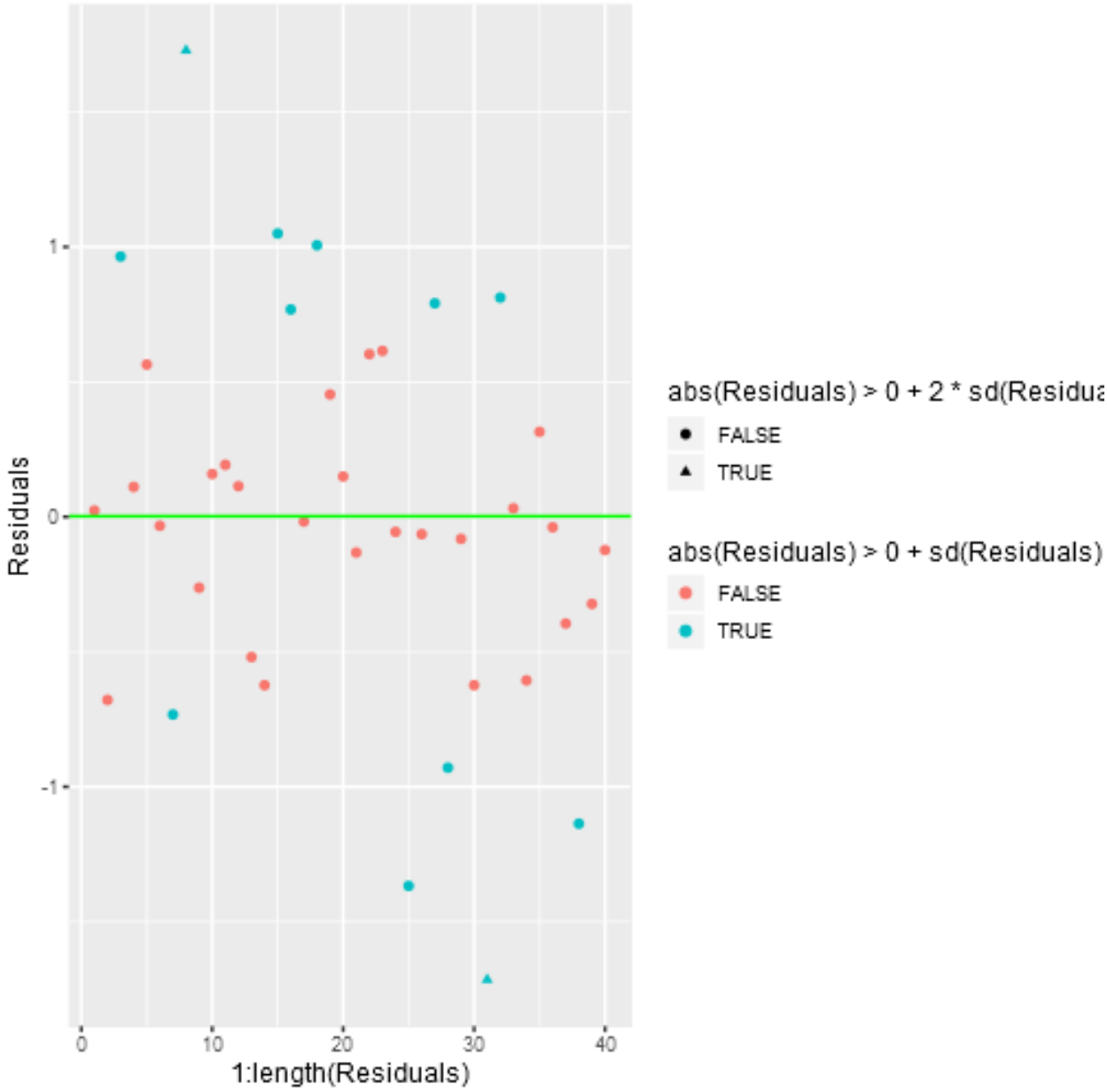
```
car::marginalModelPlots(model,col=rgb(0,0,0,alpha=0.3),col.line = c("green","red"),
                        fitted = FALSE,layout = c(2,2),grid = FALSE)
```




```
car::avPlots(model, id.n=0, id.cex=0.6,layout = c(2,2),grid = FALSE)
```




```
x_re=cbind(fir_40$x1,fir_40$x2,model$residuals)
colnames(x_re)=c('X1','X2','Residuals')
x_re=data.frame(x_re)
ggplot(data=x_re) +
  geom_point(mapping = aes(y=Residuals,x=1:length(Residuals),
                           color=abs(Residuals) > 0+sd(Residuals),
                           shape = abs(Residuals) > 0+2*sd(Residuals))) +
  geom_hline(yintercept = mean(x_re$Residuals), color="green")
```

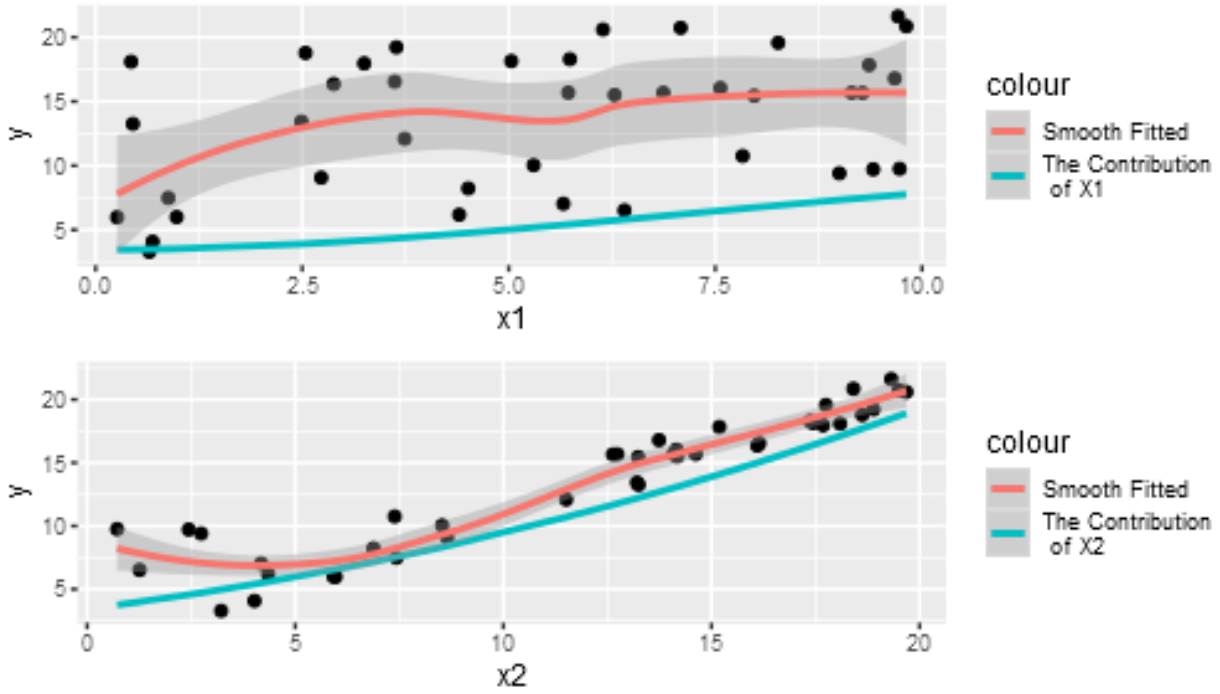


Simply judge from R^2, which indicates the model nicely fit the data and explains 97% of the variance of y, and all the coefficients are significant at confidence level of 99%, and F Test indicates that the model is effective. and the plot of residuals indicates that residuals do comply with normal distibution which obey with the assumption of linear regression model. So we may use simple linear regression model to fit the data.

2. Display the estimated model graphically as in (GH) Figure 3.2.

```
grid.arrange(
  ggplot(data = fir_40) + geom_point(mapping = aes(x = x1,y = y))
  + geom_line(mapping =
    aes(x = x1,y=model$coefficients[1]+
          model$coefficients[2]*x1^2+model$coefficients[3]*x1^3,
                    colour = "The Contribution \n of X1"),size = 1)
  + geom_smooth(mapping = aes(x = x1,y = y,colour = "Smooth Fitted"))
  ,ggplot(data = fir_40) + geom_point(mapping = aes(x = x2,y = y))
```

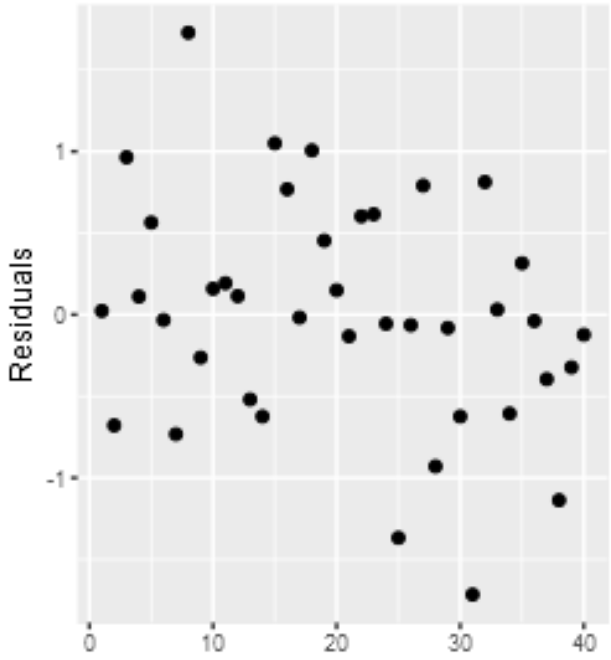
```
  + geom_line(mapping =
    aes(x = x2,y = model$coefficients[1]+
          model$coefficients[4]*x2+model$coefficients[5]*x2^2,
                    colour = "The Contribution \n of X2"),size = 1)
  + geom_smooth(mapping = aes(x = x2,y = y, colour = "Smooth Fitted"))
  ,ncol = 1
)
```

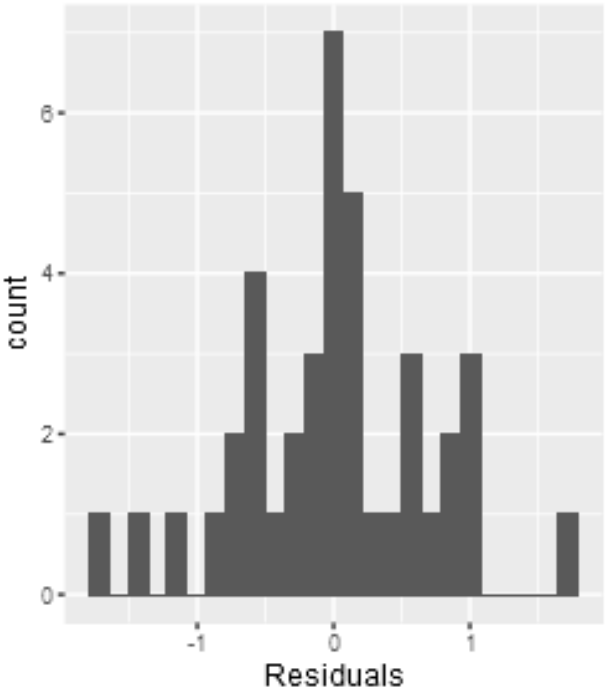


Different with (GH) Figure 3.2, which has only one predictor, here we have several predictors, it makes no sense to plot y against any single indecator.

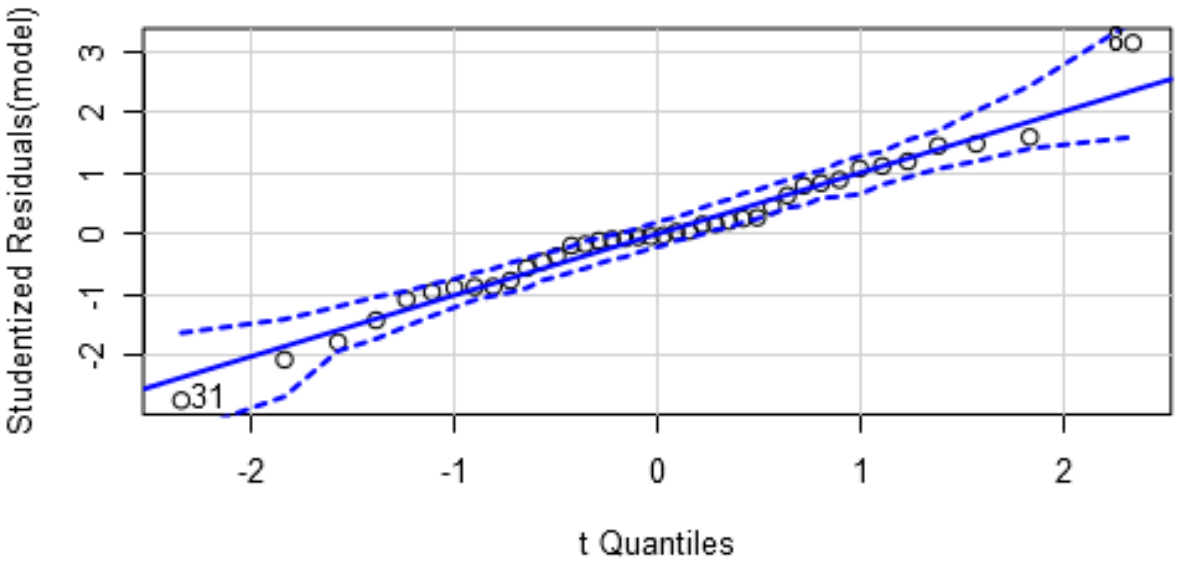
3. Make a residual plot for this model. Do the assumptions appear to be met?

```
a1 = ggplot(data = x_re) +
      geom_point(mapping = aes(x = 1: length(Residuals), y = Residuals)) + xlab("")
a2 = ggplot(data = x_re) + geom_histogram(mapping = aes(x = Residuals), bins = 25)
grid.arrange(a1,a2,ncol = 2)
```






```
car::qqPlot(model,envelope = 0.95)
```

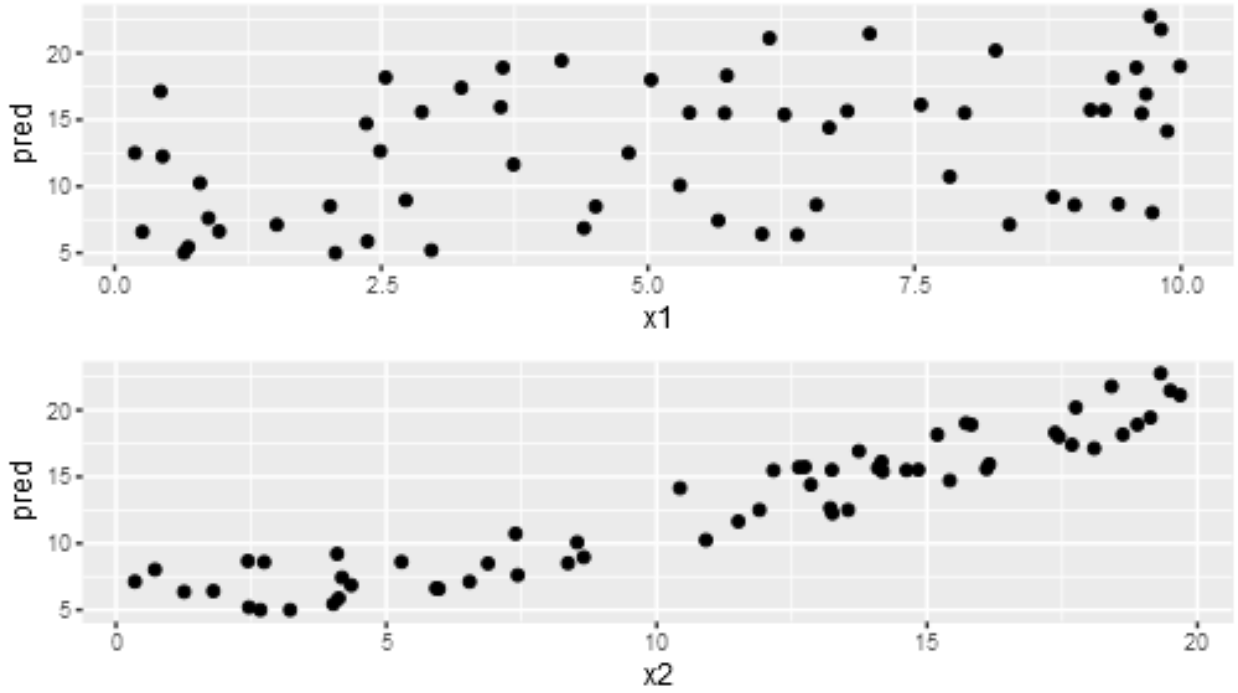



```
## [1]  8 31
```

From the Scatter and histogram plot, we can guess that residuals comply with normal distribution, also according to QQ plot, we can tell that all the point of residuals fall in 95% confidence interval, so we can conclude that residuals comply with normal distribution that comply with assumption of linear regression model.

4. Make predictions for the remaining 20 data points in the file. How confident do you feel about these predictions?

```
pred = model$coefficients[1]+model$coefficients[2]*pyth$x1^2+model$coefficients[3]*pyth$x1^3+
  model$coefficients[4]*pyth$x2+model$coefficients[5]*pyth$x2^2
pred = cbind(pred,pyth$x1,pyth$x2)
pred = data.frame(pred)
colnames(pred) = c('pred','x1','x2')
pe1 = ggplot(pred) + geom_point(aes(x = x1,y = pred))
pe2 = ggplot(pred) + geom_point(aes(x = x2,y = pred))
grid.arrange(pe1,pe2,ncol = 1)
```



Because the residuals of model complies with normal distribution and R-square of model is 0.9811, so I am pretty confident about the prediction of our model. After doing this exercise, take a look at Gelman and Nolan (2002, section 9.4) to see where these data came from. (or ask Masanao)

Earning and height

Suppose that, for a certain population, we can predict log earnings from log height as follows:

- A person who is 66 inches tall is predicted to have earnings of $30,000.
- Every increase of 1% in height corresponds to a predicted increase of 0.8% in earnings.
- The earnings of approximately 95% of people fall within a factor of 1.1 of predicted values.

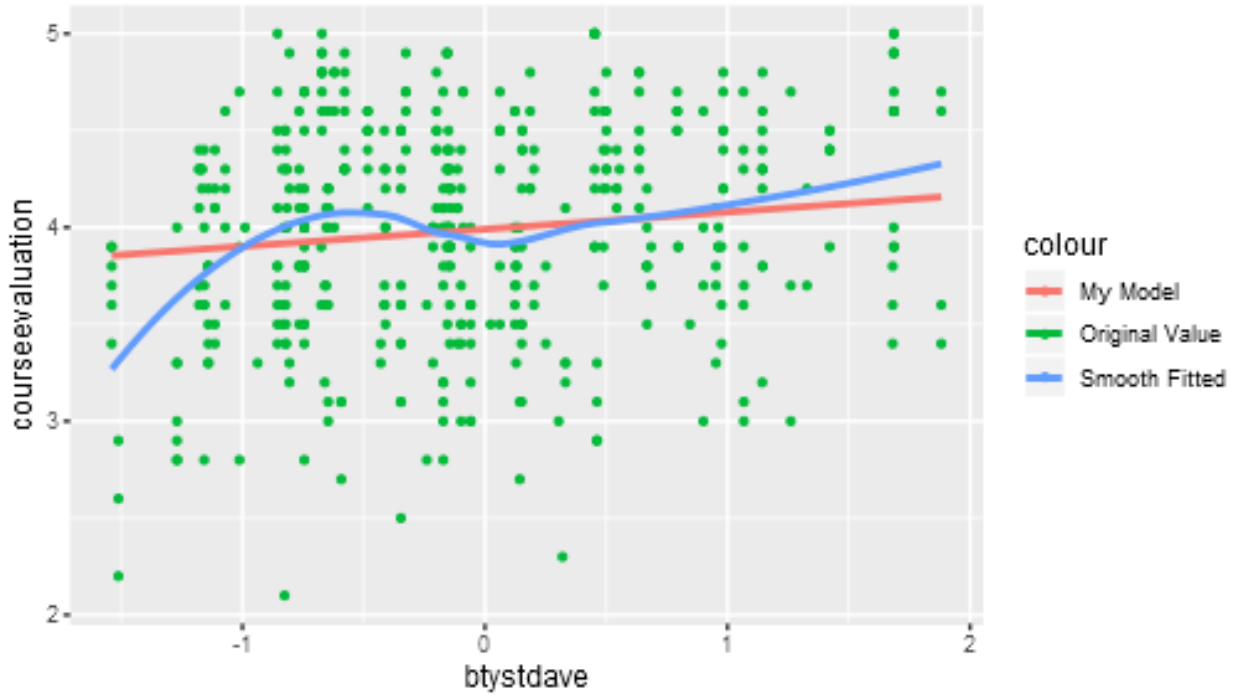
1. Give the equation of the regression line and the residual standard deviation of the regression.
   Ans: $\ln(Earning) = \beta_0 + \beta_1 * \ln(height) + \varepsilon$, we can rearrange this equation and we can get $Earning = e^{\beta_0} * Hight^{\beta_1} * \varepsilon$,increase height by 1% means that multiply height by 1.01, which equals to multiply the right-hand side by 1.01^{β_1}, so we can deduce that $1.01^{\beta_1} = 1.008$, so $\beta_1 = \log_{1.01} 1.008 = 0.8007944$,and we also know that a person who is 66 inches is predicted to have earnings of 30000, which means$ln(30000) = \beta_0 + \beta_1 * ln(66)$, so we can plug $\beta_1 = 0.8007944$, we obtain $\beta_0 = \ln(30000) - \ln(66) * \beta_1 = 6.9539004$.
2. Suppose the standard deviation of log heights is 5% in this population. What, then, is the R^2 of the regression model described here? ### Beauty and student evaluation

The folder beauty contains data from Hamermesh and Parker (2005) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations.

```
beauty.data <- read.table (paste0(gelman_example_dir,
                                  "beauty/ProfEvaltnsBeautyPublic.csv"), header=T, sep=",")
```

1. Run a regression using beauty (the variable btystdave) to predict course evaluations (courseevaluation), controlling for various other inputs. Display the fitted model graphically, and explaining the meaning of each of the coefficients, along with the residual standard deviation. Plot the residuals versus fitted values.

```
mo_be = lm(courseevaluation ~ btystdave, data = beauty.data)
ggplot(data = beauty.data) +
  geom_point(aes(x = btystdave,y = courseevaluation,
                 colour = "Original Value"), alpha = 1,size = 0.7) +
  geom_line(aes(x = btystdave,y = 3.99066+ 0.08827 * btystdave, colour = "My Model"),size = 1) +
  geom_smooth(aes(x = btystdave,y = courseevaluation,colour = "Smooth Fitted"), se = FALSE)
```




```
summary(mo_be)
```

```
##
## Call:
## lm(formula = courseevaluation ~ btystdave, data = beauty.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80015 -0.36304  0.07254  0.40207  1.10373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

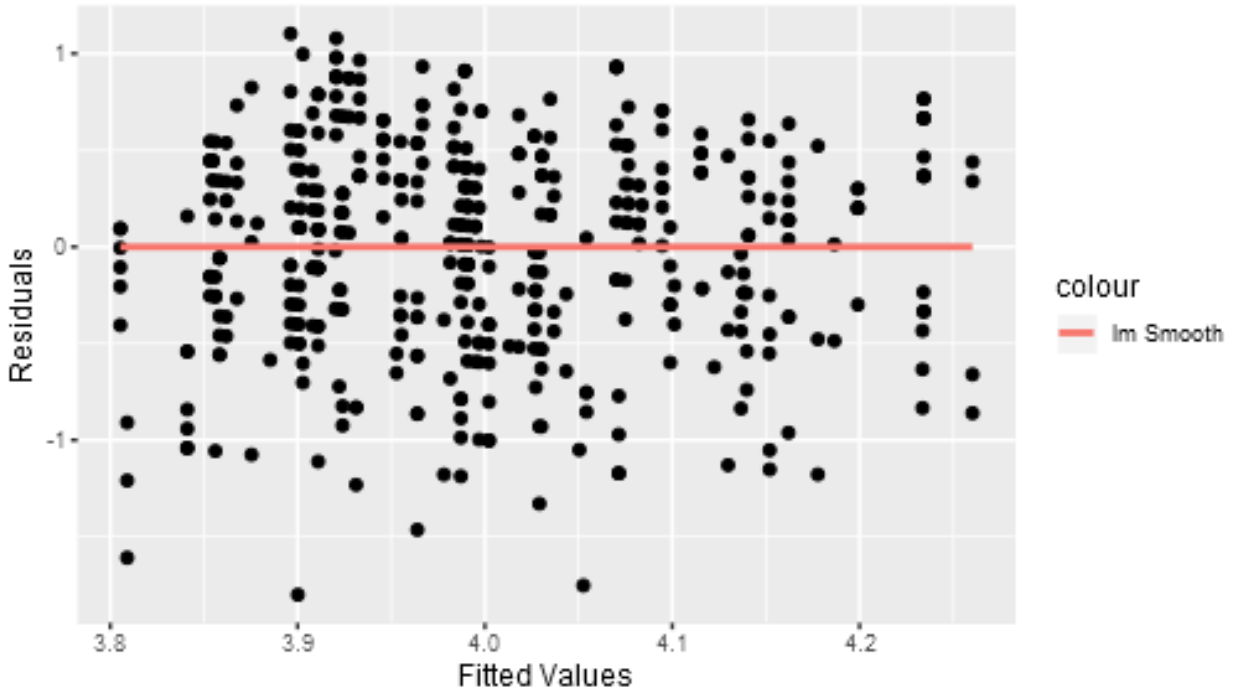
```
## (Intercept)  4.01002    0.02551 157.205  < 2e-16 ***
## btystdave    0.13300    0.03218   4.133 4.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5455 on 461 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.03364
## F-statistic: 17.08 on 1 and 461 DF,  p-value: 4.247e-05
```

The performance of this model is poor in prediction, although all of its coefficients are significantly differ than 0, but its R^2 is less than 4%, which means this model can barely explain the variance of courseevaluation. The coefficients explaination:

Intercept: When the value of average of Standardized beauty rating equals to 0, the average of course evaluation is 4.01002

Coefficient of btystdave: When keep others unchanged, the average of Standardized beauty rating increase one unit, the course evaluation will increase by 0.13300.

```
ggplot() + geom_point(aes(x = mo_be$fitted.values,y = mo_be$residuals)) +
  geom_smooth(aes(x = mo_be$fitted.values,y = mo_be$residuals,colour = "lm Smooth"),
              method = 'lm',se = FALSE) + ylab("Residuals") + xlab("Fitted Values")
```



2. Fit some other models, including beauty and also other input variables. Consider at least one model with interactions. For each model, state what the predictors are, and what the inputs are, and explain the meaning of each of its coefficients.

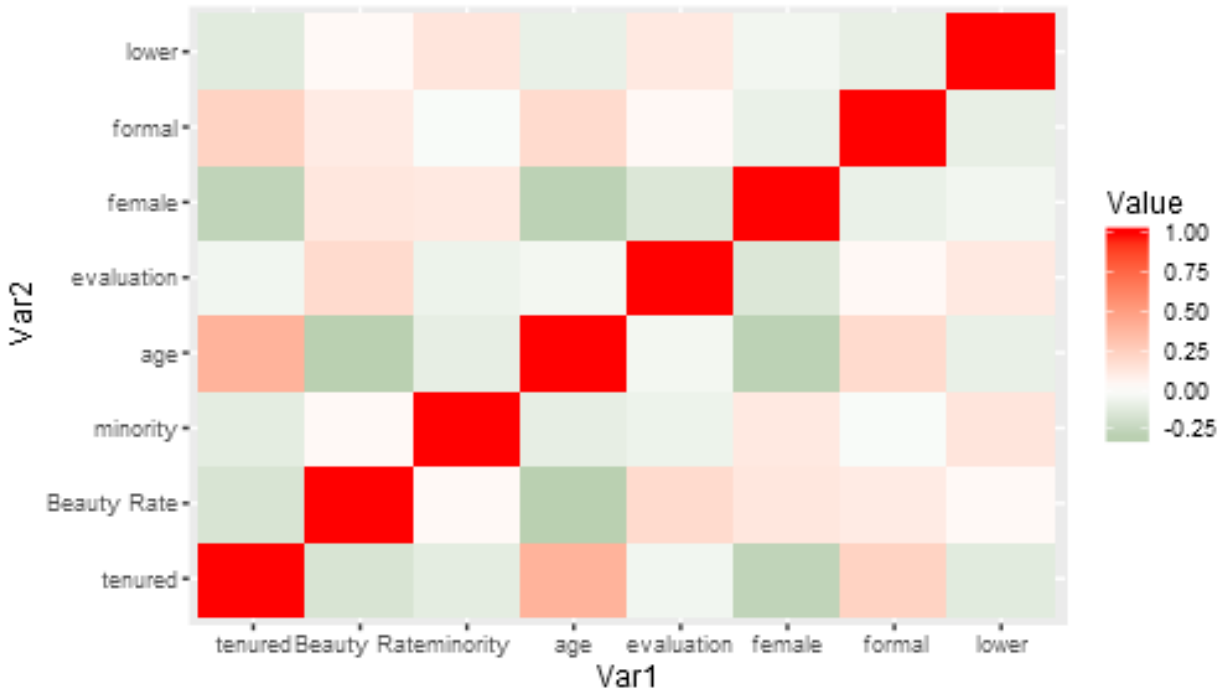
```
da = cbind(beauty.data$tenured,beauty.data$btystdave,beauty.data$minority,beauty.data$age,
           beauty.data$courseevaluation,beauty.data$female,beauty.data$formal,beauty.data$lower)
colnames(da) = c("tenured","Beauty Rate","minority","age","evaluation","female","formal","lower")
kable(cor(da),digits = 3, align = "c")
```

	tenured	Beauty Rate	minority	age	evaluation	female	formal	lower

	tenured	Beauty Rate	minority	age	evaluation	female	formal	lower
tenured	1.000	-0.166	-0.113	0.395	-0.058	-0.268	0.232	-0.126
Beauty Rate	-0.166	1.000	0.033	-0.298	0.189	0.126	0.103	0.033
minority	-0.113	0.033	1.000	-0.103	-0.077	0.115	-0.028	0.136
age	0.395	-0.298	-0.103	1.000	-0.052	-0.285	0.188	-0.091
evaluation	-0.058	0.189	-0.077	-0.052	1.000	-0.150	0.037	0.115
female	-0.268	0.126	0.115	-0.285	-0.150	1.000	-0.087	-0.057
formal	0.232	0.103	-0.028	0.188	0.037	-0.087	1.000	-0.099
lower	-0.126	0.033	0.136	-0.091	0.115	-0.057	-0.099	1.000

This is the Correlation of chosen variables

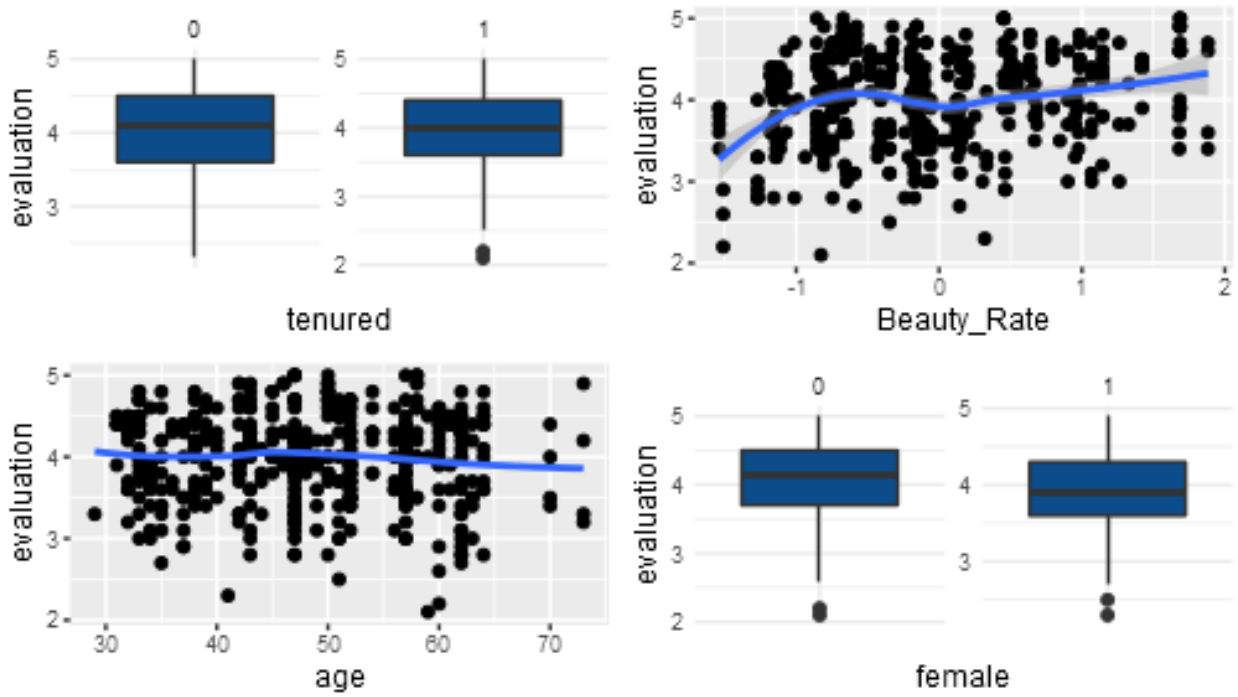
```
cor = reshape2::melt(cor(da))
ggplot(data = cor, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() + scale_fill_gradient2(name="Value",
                                     low = "darkgreen",high = "red", mid = "white")
```



This is the Correlation Heatmap, We choose those variables that is highly correlated as predictors.

```
dad=cbind(beauty.data$tenured,beauty.data$btystdave,beauty.data$age,
          beauty.data$courseevaluation,beauty.data$female)
colnames(dad) = c("tenured","Beauty_Rate","age","evaluation","female")
dad=data.frame(dad)
grid.arrange(
  ggplot(dad) +aes(x = "", y = evaluation) +geom_boxplot(fill = "#0c4c8a") +
    theme_minimal() +facet_wrap(vars(tenured), scales = "free") + xlab("tenured"),
  ggplot(dad,aes(y = evaluation,x = Beauty_Rate))
```

```
  + geom_point()
  + geom_smooth(),
  ggplot(dad,aes(y = evaluation,x = age))
  + geom_point(se = FALSE)
  + geom_smooth(se = FALSE),
  ggplot(dad) +aes(x = "", y = evaluation) +geom_boxplot(fill = "#0c4c8a") +
    theme_minimal() +facet_wrap(vars(female), scales = "free")+ xlab("female"),
  ncol = 2
)
```



We can tell that course evaluation does not significantly differ when we consider tenure, so we can exclude these two variables.
Also we can tell that the effect that beauty rate has on course evaluation may not be linear, so we may add higher order of beauty rate.

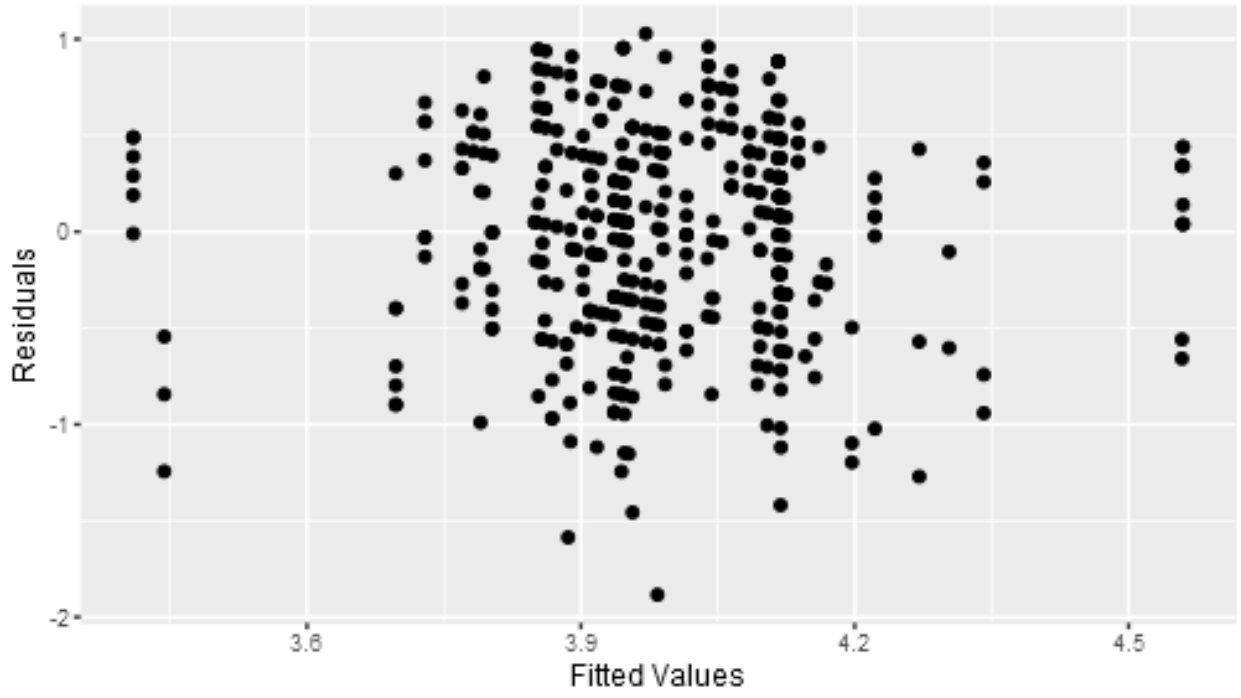
```
m = lm(evaluation ~ I(Beauty_Rate^2)+I(Beauty_Rate^3)+
         I(female*Beauty_Rate)+female,data = data.frame(dad))
summary(m)
```

```
##
## Call:
## lm(formula = evaluation ~ I(Beauty_Rate^2) + I(Beauty_Rate^3) +
##     I(female * Beauty_Rate) + female, data = data.frame(dad))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88387 -0.35636  0.05068  0.40523  1.02892
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.12007    0.03915 105.231  < 2e-16 ***
```

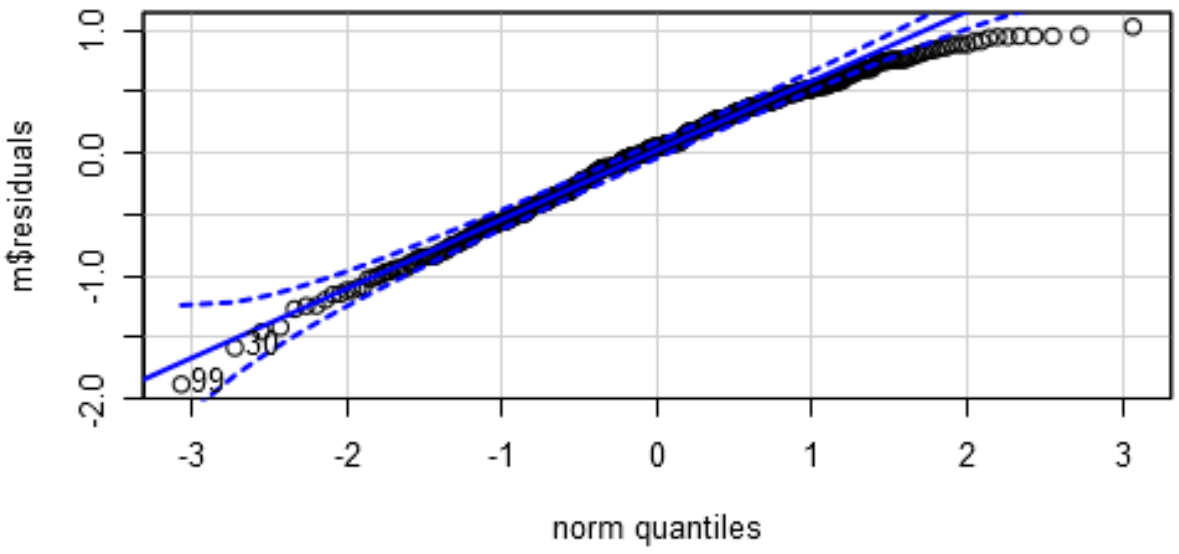
```
## I(Beauty_Rate^2)        -0.08312    0.03653  -2.275  0.02336 *
## I(Beauty_Rate^3)         0.14099    0.02434   5.793 1.29e-08 ***
## I(female * Beauty_Rate) -0.12388    0.06018  -2.058  0.04011 *
## female                  -0.19058    0.05006  -3.807  0.00016 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5299 on 458 degrees of freedom
## Multiple R-squared:  0.09571,    Adjusted R-squared:  0.08782
## F-statistic: 12.12 on 4 and 458 DF,  p-value: 2.26e-09
```

Compare to simply regress course evaluation on beauty rates, new model with more predictors are slightly better in explaining the variance of course evaluation. Coefficient explaination: **intercept:** The average course evaluation when the beauty rate is 0, and female is 0 (The description in original data set didn't give detailed information about what female refers to)
Coefficient of Beauty_Rate^2: When keeps others unchanged, one unit increase in the $BeautyRate^2$, the average of course evaluation of male will decrease by 0.08312.
Coefficient of Beauty_Rate^3: When keeps others unchanged, one unit increase in the $BeautyRate^3$, the average of course evaluation of male will decrease by 0.14099.
Coefficient of female: average course evaluation of female is 0.19058 lower than that of male.

```
ggplot() + geom_point(aes(x = m$fitted.values,y = m$residuals)) +
           ylab("Residuals") + xlab("Fitted Values")
```




```
car::qqPlot(m$residuals,envelope = 0.95)
```




```
## [1] 99 30
```

See also Felton, Mitchell, and Stinson (2003) for more on this topic link

Conceptula excercises

On statistical significance.

Note: This is more like a demo to show you that you can get statistically significant result just by random chance. We haven't talked about the significance of the coefficient so we will follow Gelman and use the approximate definition, which is if the estimate is more than 2 sd away from 0 or equivalently, if the z score is bigger than 2 as being "significant".

(From Gelman 3.3) In this exercise you will simulate two variables that are statistically independent of each other to see what happens when we run a regression of one on the other.

1. First generate 1000 data points from a normal distribution with mean 0 and standard deviation 1 by typing in R. Generate another variable in the same way (call it var2).

```
var1 <- rnorm(1000,0,1)
var2 <- rnorm(1000,0,1)
```

Run a regression of one variable on the other. Is the slope coefficient statistically significant? [absolute value of the z-score(the estimated coefficient of var1 divided by its standard error) exceeds 2]

```
fit  <- lm (var2 ~ var1)
z.scores <- coef(fit)[2]/se.coef(fit)[2]
z.scores
```

```
##      var1
## 0.6246654
```

2. Now run a simulation repeating this process 100 times. This can be done using a loop. From each simulation, save the z-score (the estimated coefficient of var1 divided by its standard error). If the

absolute value of the z-score exceeds 2, the estimate is statistically significant. Here is code to perform the simulation:

```
z.scores <- rep (NA, 100)
for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit  <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
}
sum(z.scores>2)
```

```
## [1] 4
```

How many of these 100 z-scores are statistically significant?
Ans: most time, 2 or 3 out 100. What can you say about statistical significance of regression coefficient? Ans: it is useful but we cannot rely on them to much. ### Fit regression removing the effect of other variables

Consider the general multiple-regression equation

$$Y = A + B_1X_1 + B_2X_2 + \cdots + B_kX_k + E$$

An alternative procedure for calculating the least-squares coefficient B_1 is as follows:

1. Regress Y on X_2 through X_k, obtaining residuals $E_{Y|2,\ldots,k}$.
2. Regress X_1 on X_2 through X_k, obtaining residuals $E_{1|2,\ldots,k}$.
3. Regress the residuals $E_{Y|2,\ldots,k}$ on the residuals $E_{1|2,\ldots,k}$. The slope for this simple regression is the multiple-regression slope for X_1 that is, B_1.

(a) Apply this procedure to the multiple regression of prestige on education, income, and percentage of women in the Canadian occupational prestige data (http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/datasets/Prestige.pdf), confirming that the coefficient for education is properly recovered.

```
fox_data_dir<-"http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/datasets/"
Prestige<-read.table(paste0(fox_data_dir,"Prestige.txt"))
kable(head(Prestige))
```

	education	income	women	prestige	census	type
GOV.ADMINISTRATORS	13.11	12351	11.16	68.8	1113	prof
GENERAL.MANAGERS	12.26	25879	4.02	69.1	1130	prof
ACCOUNTANTS	12.77	9271	15.70	63.4	1171	prof
PURCHASING.OFFICERS	11.42	8865	9.11	56.8	1175	prof
CHEMISTS	14.62	8403	11.68	73.5	2111	prof
PHYSICISTS	15.64	11030	5.13	77.6	2113	prof

```
st1 = lm(prestige~women+income,data = Prestige)
st2 = lm(education~income+women,data = Prestige)
st3 = lm(st1$residual~st2$residual)
rl = lm(prestige~education+women+income,data = Prestige)
st3$coefficient[2]
```

```
## st2$residual
##     4.186637
```

```
rl$coefficient[2]
```

```
## education
##  4.186637
```

(b) The intercept for the simple regression in step 3 is 0. Why is this the case?
Because the intercept represents the average of the out-come values when we hold the predictors equal to zero, for the third step regression, the left-hand side of regression equation is residual of the second step regression, the average should of it should be zero.

(c) In light of this procedure, is it reasonable to describe B_1 as the "effect of X_1 on Y when the influence of $X_2, \cdots, X_k$ is removed from both X_1 and Y"?
It is reasonable to explain the B_1 as the effect of X_1 on Y, because in the first two regressions, we have removed the effects of X_2 and X_3.

(d) The procedure in this problem reduces the multiple regression to a series of simple regressions (in Step 3). Can you see any practical application for this procedure?
When the data set is too large to be processed in one time, we can break this regression model into several pieces so that we can get the final answer.

Partial correlation

The partial correlation between X_1 and Y "controlling for" $X_2, \cdots, X_k$ is defined as the simple correlation between the residuals $E_{Y|2,\ldots,k}$ and $E_{1|2,\ldots,k}$, given in the previous exercise. The partial correlation is denoted $r_{y1|2,\ldots,k}$.

1. Using the Canadian occupational prestige data, calculate the partial correlation between prestige and education, controlling for income and percentage women.

```
cor(st1$residual,st2$residual)
```

```
## [1] 0.7362604
```

2. In light of the interpretation of a partial regression coefficient developed in the previous exercise, why is $r_{y1|2,\ldots,k} = 0$ if and only if B_1 is 0?

Because when X_1 in independent to the other predictors X_2, X_3, the coefficient of X_1 equals to the coefficient in a simple linear regression, while in simple linear regression, the coefficient of X_1 equals to $\frac{cov(Y,X_1)}{Var(X_1)}$, and $cov(Y, X_1)$ equals to $\rho_{X_1,Y} * sd(X_1) * sd(Y)$, as we know, variance of X_1 and Y can not be zero, so when the covariance of X_1 and Y equals to 0, the coefficients of X_1 and the correlation bwtween X_1 and Y will be zero.

Mathematical exercises.

Prove that the least-squares fit in simple-regression analysis has the following properties:

1. $\sum \hat{y}_i \hat{e}_i = 0$
$\sum \hat{y}_i \hat{e}_i = \hat{Y} \cdot \hat{e}$, according the model and assumptions, $\hat{Y}$ is independent to e, which means $\hat{Y}$ is orthorgonal to y, so the inner product should be zero. Also $\hat{e} = (I - P)Y$,where I is identity matrix and P is projection matrix, so $\hat{y} \cdot \hat{e} = YP(I - P)Y = Y(P - PP)Y = Y(P - P)Y = 0$

2. $\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum \hat{e}_i(\hat{y}_i - \bar{y}) = 0$
$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum \hat{e}_i(\hat{y}_i - \bar{y}) = \sum (\hat{e_i}\hat{y_i} - \hat{e_i}\bar{y})$, for$\sum \hat{e_i}\hat{y_i} = 0$, for$\sum \hat{e_i}\bar{y_i} = \bar{y_i}\sum \hat{e_i} = 0$, so $\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

Suppose that the means and standard deviations of $\boldsymbol{y}$ and $\boldsymbol{x}$ are the same: $\bar{y} = \bar{x}$ and $sd(y) = sd(x)$.

1. Show that, under these circumstances

$$\beta_{y|x} = \beta_{x|y} = r_{xy}$$

where $\beta_{y|x}$ is the least-squares slope for the simple regression of $\boldsymbol{y}$ on $\boldsymbol{x}$, $\beta_{x|y}$ is the least-squares slope

for the simple regression of $\boldsymbol{x}$ on $\boldsymbol{y}$, and r_{xy} is the correlation between the two variables. Show that the intercepts are also the same, $\alpha_{y|x} = \alpha_{x|y}$.
$\beta_{y|x} = \frac{cov(y,x)}{sd(x)sd(x)}$,because $sd(y) = sd(x) = \sigma$, so $\beta_{y|x} = \frac{cov(y,x)}{sd(x)sd(x)} = \frac{cov(x,y)}{\sigma^2} = \beta_{x|y} = \rho_{xy}$
because in simple linear regression, the intercept is the average of the left-hand side of the regression equation when hole the predectors equals to 0, so for here the intercepts equal to the average of X and Y, and $\bar{X} = \bar{Y}$, so the intercepts equal.

2. Why, if $\alpha_{y|x} = \alpha_{x|y}$ and $\beta_{y|x} = \beta_{x|y}$, is the least squares line for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ different from the line for the regression of $\boldsymbol{x}$ on $\boldsymbol{y}$ (when $r_{xy} < 1$)?
When regress Y on x, we have $Y = \alpha_{Y|X} + \beta_{Y|X} \cdot X$
When regress X on Y, we have $X = \alpha_{X|Y} + \beta_{X|Y} \cdot Y$, than we rearrange this equation let Y be represented as a function of X, we get $Y = \frac{-\alpha_{X|Y}}{\beta_{X|Y}} + \frac{1}{\beta_{X|Y}} \cdot X$
because $\beta_{X|Y} = \beta_{Y|X} < 1$ and $\alpha_{X|Y} = \alpha_{Y|X}$, so $\alpha_{Y|X} \neq \frac{-\alpha_{X|Y}}{\beta_{X|Y}}, \beta_{Y|X} \neq \frac{1}{\beta_{X|Y}}$, so the two regression line are different

3. Imagine that educational researchers wish to assess the efficacy of a new program to improve the reading performance of children. To test the program, they recruit a group of children who are reading substantially vbelow grade level; after a year in the program, the researchers observe that the children, on average, have imporved their reading performance. Why is this a weak research design? How could it be improved?
This research is designed because it can obtain a good result when assessing the new program. It can be improved by randomly obtain identities in the sample and then compare their performance before and after the program.

Feedback comments etc.

If you have any comments about the homework, or the class, please write your feedback here. We love to hear your opnions.