# Final exam

Kevin Tchouate Mouofo

12/7/2020

## Problem 1 (10 pts=2+2+2+4)

Accurately measuring a person's body fat percentage is difficult. An indirect method is to estimate the body fat percentage based on various body circumference measurements. Data were collected on the body fat percentage and several body measurements. The regression output of the data is given below (some entries in the output are deleted). Answer the following questions.

(a) Is it possible for you to find out the number of observations in the data set? If so, do it :

Answer :

It is possible to find the number of observation with the degrees of freedom. From the anova table, we see that we have regression df = 4, Error df = 247. So, we have n = 4 + 247 + 1 = 252 observations.

(b) Compute the Adjusted $R^2$ .

Answer :

$R_a^2 = 1 - (\frac{n-1}{n-p})\frac{SSE}{SSTO}$

$= 1 - (\frac{251}{247})\frac{4658.236}{17578.990}$

$= 0.7307199$

(c) Which body circumference measurement seems to be the most important in determining the body fat percentage?

Answer :

The wrist circumference measurement seems to be the most important in determining the body fat percentage, because it has the coefficient with the largest absolute value.

(d) Compute a 95% confidence interval for the coefficient of Forearm (t0.025,247 = 1.97).

Answer :

For forearm we have :

$stdError = \frac{Estimate}{tRatio} = \frac{0.4729284}{2.60} = 0.1818955$

Hence, a 95% confidence interval for the coefficient of Forearm is :

$0.4729284 - 0.1818955 * 1.97 \leq \beta_{forearm} \leq 0.4729284 + 0.1818955 * 1.97$

$\Leftrightarrow 0.1145943 \leq \beta_{forearm} \leq 0.8312625$

## Problem 2 (10 pts=2+2+4+2)

Data for 51 U.S. "states" (50 states, plus the District of Columbia) was used to examine the relationship between violent crime rate (violent crimes per 100,000 persons per year) and the predictor variables of urbanization (percentage of the population living in urban areas) and poverty rate. A predictor variable indicating whether or not a state is classified as a Southern state (1 = Southern, 0 = not) was also included.

(a) For the Southern states, what is the fitted regression model for Crime with respect to urbanization and poverty.

Answer :

The fitted regression model for southern states for crime with respect to urbanization and poverty is :

$\hat{crime} = -321.9 + 4.69Urban + 39.3Poverty - 649.3 + 12.1Urban - 5.84Poverty$

$\Leftrightarrow \hat{crime} = -971.2 + 16.79Urban + 33.46Poverty$

(b) Predict the violent crime rate for a Southern state with an urbanization of 55.4 and a poverty rate of 13.7.

Answer :

$\hat{crime} = -971.2 + 16.79 * 55.4 + 33.46 * 13.7 = 417.368$

(c) Calculate the ANOVA F test statistic value, the DF corresponds to Residual Error, and the MSE. What should be the degree of freedom for the F statistic?

Answer :

$df_{Residual} = df_{total} - df_{Regression} = 50 - 5 = 45$

$MSE = \frac{SSE}{DF_{Error}} = \frac{882169}{45} = 19603.76$

$F* = \frac{MSR}{MSE} = \frac{412091}{19603.76} = 21.02102$

The degrees of freedom for F statistic shound be 5 for the numerator and 50 for the denominator.

(d) Which predictors should probably be removed from the model to improve it? Why?

Answer :

We should probably remove Poverty*South to improve the model since it is the only predictor associated to a p-value (0.728) greater than the usual level of significance 0.05.

## Problem 3 (10 pts: 3, 2, 5)

faults were studied by changing the temperature from 40 $\circ$C to 80 $\circ$C in the spinning process.

(a) Perform the two-sided hypothesis test on the coefficient corresponding to x, given $\alpha = 0.05$. (The normal quantile is z0.975 = 1.96.)

Answer :

Alternatives :

H0 : $\beta_x = 0$

Ha: $\beta_x \neq 0$

Test statistic : $t* = \frac{b_x}{std(b_x)}$

Decision rule : if $|t*| \leq 1.96$ conclude H0, if $|t*| > 1.96$ conclude Ha

We have $t* = \frac{0.2067636}{0.0068186} = 30.32347$, we conclude Ha, that x contribute to the model.

(b) Predict the percentage of faults when temperature is at 60 $\circ$C.

Answer :

We have the response equation : $\hat{\pi} = [1 + exp(13.43198 - 0.2067636x - 0.0101906(x - 60)^2)]^{-1}$

For x=60, we have $\hat{\pi} = [1 + exp(13.43198 - 0.2067636 * 60)]^{-1} = 0.2638285$

(c) Find the optimum value of of temperature to minimize the faults. (Hint: ez/(1 + ez) is a monotonic function in z).

Answer :

Let's have $z = 13.43198 - 0.2067636 * x - 0.0101906 * (x - 60)^2$

We have :

$\hat{\pi} = \frac{e^z}{1+e^z}$

The optimum value that minimize fault is found when

$\frac{d\hat{\pi}}{dx} = 0$

$\Leftrightarrow \frac{d(\frac{e^z}{1+e^z})}{dx} = 0$

$\Leftrightarrow \frac{\frac{d(e^z)}{dx}(1+e^z) - e^z\frac{d(1+e^z)}{dx}}{(1+e^z)^2} = 0$

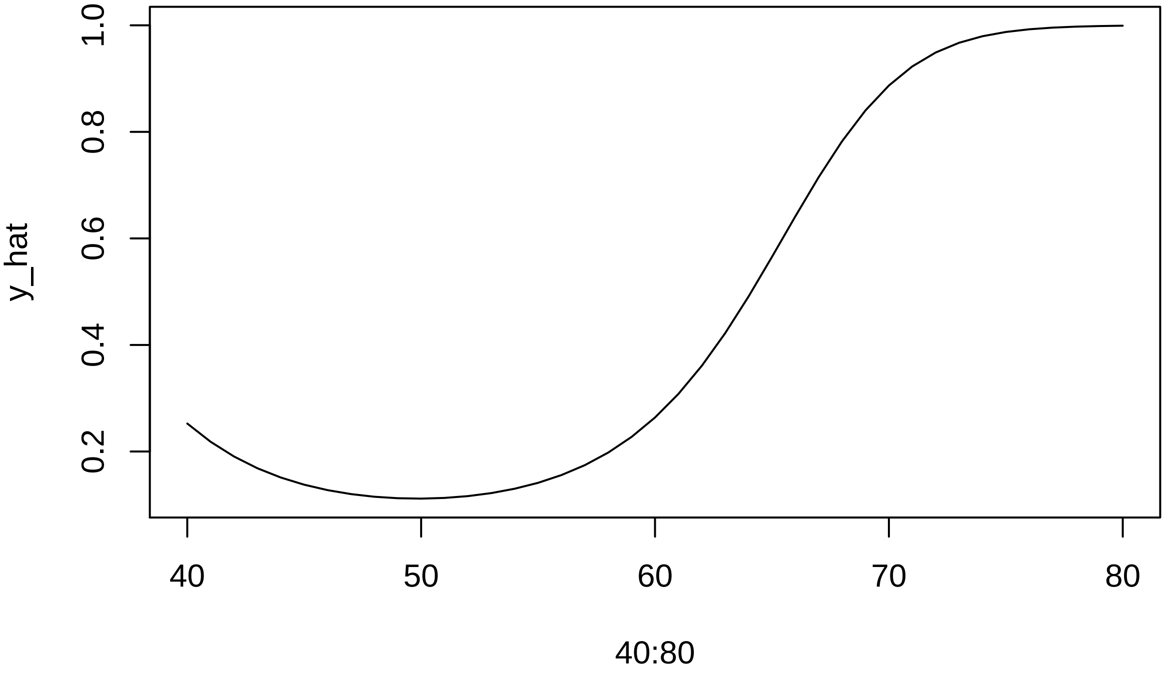
$\Leftrightarrow \frac{[\beta_1 + 2\beta_2(x-60)][(1+e^z)e^z - e^{2z}]}{(1+e^z)^2} = 0$

$\Leftrightarrow \frac{[\beta_1 + 2\beta_2(x-60)]e^z}{(1+e^z)^2} = 0$

$\Leftrightarrow \beta_1 + 2\beta_2(x - 60) = 0$

$\Leftrightarrow x = 60 - \frac{\beta 1}{2\beta_2} = 49.85518$

So the optimum value of temperature to minimize the faults is 50°.



As we can see from the plot, the faults reaches its minimum at 50°

## Problem 4 (10 pts=3+2+2+3) Flu Shots.

A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded Y = 1, and a client who did not receive a flu shot was coded Y = 0. In addition, data were collected on their age (X1) and their health awareness. The latter data were combined into a health awareness index (X2), for which higher values indicate great awareness. Also included in the data were client gender, when males were coded X3 = 1 and females were coded X3 = 0.

(a) Obtain exp($\hat{\beta 2}$) (using the first R output), and interpret this number.

Answer :

$exp(\hat{\beta 2})$ = 0.9057518.

The odds ratio $\hat{OR}_2$ = exp(b2) = 0.9057518 means that the odds of a client receiving a flu shot are estimated to decrease by about 9.4 percent with each unit increase in the health awareness index of the client, with other variables fixed.

(b) What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot? (Use the first R output.)

Answer :

The estimated probability is

$\hat{\pi}(55, 60, 1) = [1 + exp(1.17716 - 0.07279 * (55) + 0.09899 * (60) - 0.43397)]^{-1} = 0.06422$

(c) Based on the residual deviance, perform the Goodness-of-Fit test for $\alpha$ = 0.05. Write down the H0 and Ha for the Goodness-of-Fit test and make conclusion based on your test. (The chi-square quantile is $X^2$(0.95, 155) = 185.0523. Use the first R output.)

Answer :

Alternatives :

H0: $E(Y) = [1 + exp(\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3)]^{-1}$

Ha: $E(Y) \neq [1 + exp(\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3)]^{-1}$

Where $\beta 0$ = 1.17716, $\beta 1$ = -0.07279, $\beta 2$ = 0.09899, $\beta 3$ = -0.43397.

Test statistic :

DEV(1, X1, X2, X3) = -2[log L(R) - log L(F)], where L(F) and L(R) are the maximized likelihood of the satured model and the reduced model respectively(Saturated model E(Yij) = $\pi j = \frac{Y.j}{n_j}$, reduced model $E(Yij) = [1 + exp(X_j'\beta)]^{-1}$)

Decision rules :

if DEV(1, X1, X2, X3) $\leq X^2$(0.95, 155) conclude H0, if DEV(1, X1, X2, X3) $> X^2$(0.95, 155) conclude Ha.

We have DEV(1, X1, X2, X3) = 105.09 $\leq$ 185.0523, so we conlude H0, that the logistic model is a satisfactory model.

(d) Comment on the statistical significance of gender to the probability of getting a flu shot. Use the likelihood ratio test (The chi-square quantile is $X^2$(0.95,1) = 3.84.) to see if X3 should be dropped from the first logistic regression model.

Answer :

Since the p-value of gender (0.40558) is greater than the usual level of significance 0.05, we can see that it is not a significant variable.

Lets carry the likelihood ratio test :

Full model : $log(\frac{\pi(x)}{1-\pi(x)}) = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3$

Reduced model : $log(\frac{\pi(x)}{1-\pi(x)}) = \beta 0 + \beta 1X1 + \beta 2X2$

Alternatives :

H0: $\beta 3 = 0$

Ha: $\beta 3 \neq 0$

Test statistc :

$G^2 = -2log\frac{L(R)}{L(F)} = 105.80 - 105.09 = 0.71$

The decision rule :

if($G^2 \leq X2(0.95, 1) = 3.84$), conclude H0

if($G^2 > X2(0.95, 1) = 3.84$), conclude Ha

Since $G^2 = 0.71 \leq 3.84$, we conclude H0, that $\beta 3$ can be dropped from the regression model.

## Problem 5 (10 pts=4+3+3)

Researchers studied 41 male African elephants over a period of 8 years. The age (X) of the elephant at the beginning of the study and the number of successful mating (Y ) during the 8 years were recorded. We assume the number of matings follows a Poisson distribution, where the mean depends on the age of the elephant in question.

(a) Show that the Poisson distribution is a member of the exponential family.

Answer :

To be sure that poisson distribution is a member of natural exponential family distribution, we need to rewrite its density function in this form :

$f(y, \theta, \phi) = exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi))$

Where $\phi$ is the dispersion paramter and $\theta$ the canonical parameter.

The probability mass function for a Poisson random variable is :

$f(y) = \frac{\mu^y e^{-\mu}}{y!}$, $y = 0, 1, 2, ...$

$E(Y) = \mu$ and $var(Y) = \mu$, where $\mu > 0$

We have :

$logf(y) = ylog(\mu) - \mu - log(y!)$

$exp(logf(y)) = f(y) = exp(ylog(\mu) - \mu - log(y!))$

For $\theta = log(\mu)$, $\mu = e^\theta = b(\theta)$, $c(y, \phi) = -log(y!)$ and $a(\phi) = 1$, we have:

$f(y) = exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi))$ as required.

We now have :

$E(y) = \frac{d}{d\theta}b(\theta) = \frac{d}{d\theta}e^\theta = \mu$

$var(y) = a(\phi)\frac{d^2}{d\theta^2}b(\theta)$

$= 1 * \frac{d^2}{d\theta^2}e^\theta$

$= \mu$

Thus, E(y) = var(y) = $\mu$

Lets have the canonical link function $g(\mu) = \theta$ :

We have : $\mu = e^{\theta} \Rightarrow log(\mu) = \theta \Rightarrow$ log is the canonical link

Hence, the Poisson distribution is a member of the exponential family.

(b) Return the 95% confidence interval on the estimation of the coefficient for Age X. (z0.975 = 1.96).

Answer :

$0.06869 - 1.96 * 0.01375 \leq \beta 1 \leq 0.06869 + 1.96 * 0.01375$

$\Leftrightarrow 0.04174 \leq \beta 1 \leq 0.09564$

(c) How much does the mean number of matings increases if age is increased by one year? Obtain a 95% confidence interval for the estimated increase.

Answer :

If the age increase by 1 year, the mean number of matings increases by 0.06869. We have 95% confidence interval :

0.06869 - 1.96 $\sqrt{\frac{0.06869}{41}} \leq \beta 1 \leq$ 0.06869 + 1.96 $\sqrt{\frac{0.06869}{41}}$

$\Leftrightarrow -0.01153522 \leq E(\beta 1) \leq 0.1489152$

# Problem 6 (10 pts)

Explain how you will use linear regression methods to approximately estimate the parameter in the following nonlinear model:

$y = 1 - e^{-(\beta 0 + \beta 1 x)}$

Given the data (xi,yi) for i = 1,. . . ,n, write down the formula for estimating $\beta 0$ and $\beta 1$. (Hint: you need to use the parameter estimation formula for simple linear regression.)

Answer :

We have :

$y = 1 - e^{-(\beta 0 + \beta 1 x)}$

$\Leftrightarrow 1 - y = e^{-(\beta 0 + \beta 1 x)}$

$\Leftrightarrow log(1 - y) = -(\beta 0 + \beta 1 x)$

$\Leftrightarrow log(\frac{1}{1-y}) = \beta 0 + \beta 1 x$

So, by estimating the paramters for the simple linear model $y' = \beta 0 + \beta 1 x$, with $y' = log(\frac{1}{1-y})$, we will obtain the estimated coefficients for our non linear model.

Or we have the following estimation of the parameters for a simple linear model:

$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}))}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

$b_0 = \bar{y} - b_1 \bar{x}$

Thus for our non linear model we will have :

$\hat{\beta_1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y'_i - \bar{y'}))}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

$\hat{\beta}_0 = \bar{y'} - \hat{\beta}_1\bar{x}$

With $y'_i = log(\frac{1}{1-y_i})$

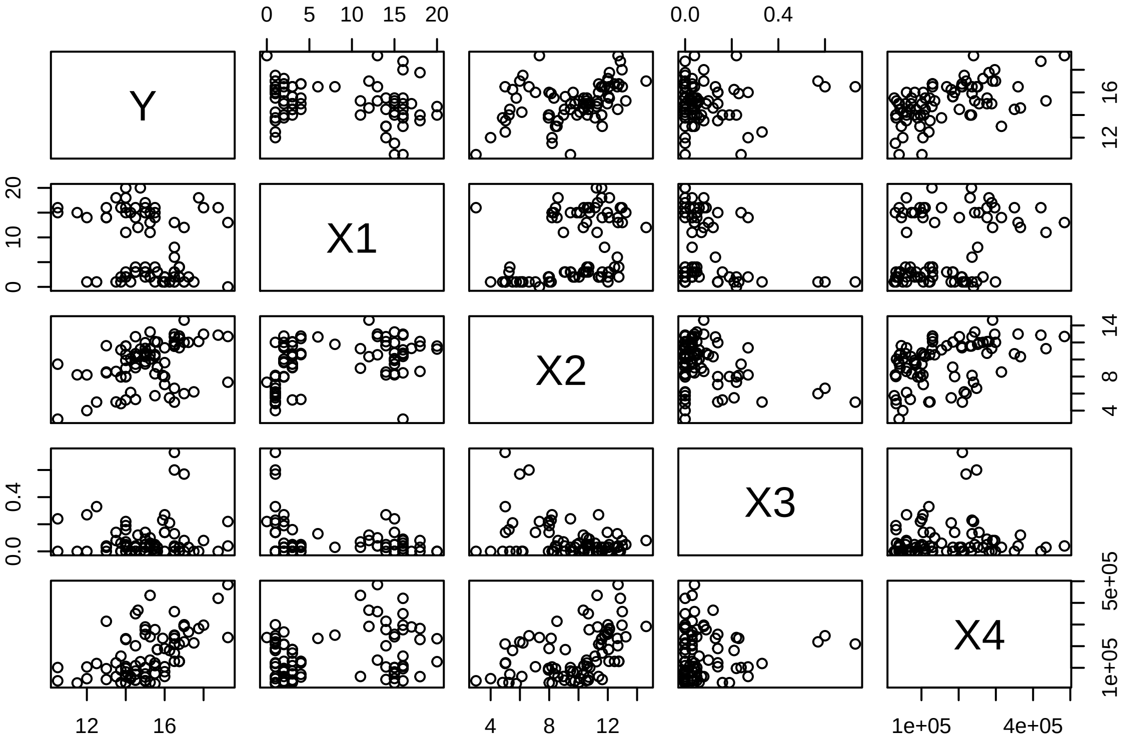
## Problem 7. Commercial Properties.

A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, more attractive, and expensive for five specific geographic areas. The data contain the columns the age (X1), operating expense and taxes (X2), vacancy rates (X3), total square footage (X4) and rental rates (Y ).

1. (2 pts) Obtain the scatter plot matrix of all the variables Y and X1 X4. State and interpret your findings.

Answer :

```r
commercial_data <- read.csv("/Users/kevinmouofo/Desktop/MATH 564/Final/Commercial_Property.txt",
                 sep="", header = TRUE )
plot(commercial_data)
```



```r
cor(commercial_data)
```

```
##              Y         X1         X2          X3         X4
## Y   1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
## X1 -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
## X2  0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
## X3  0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
## X4  0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```

From the scatter plot, we can see that there is no strong linear correlation among the variables, however the correlation between Y and X2, and also the correlation between Y and X4, are relatively high as confirmed by the correlation matrix.

2. (1 pt) Fit regression model $Y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \beta 3 X3 + \beta 4 X4 + \epsilon$ (call this Model 1). State the estimated regression function.

Answer :

```r
model1 <- lm(Y~X1+X2+X3+X4, data=commercial_data)
summary(model1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = commercial_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

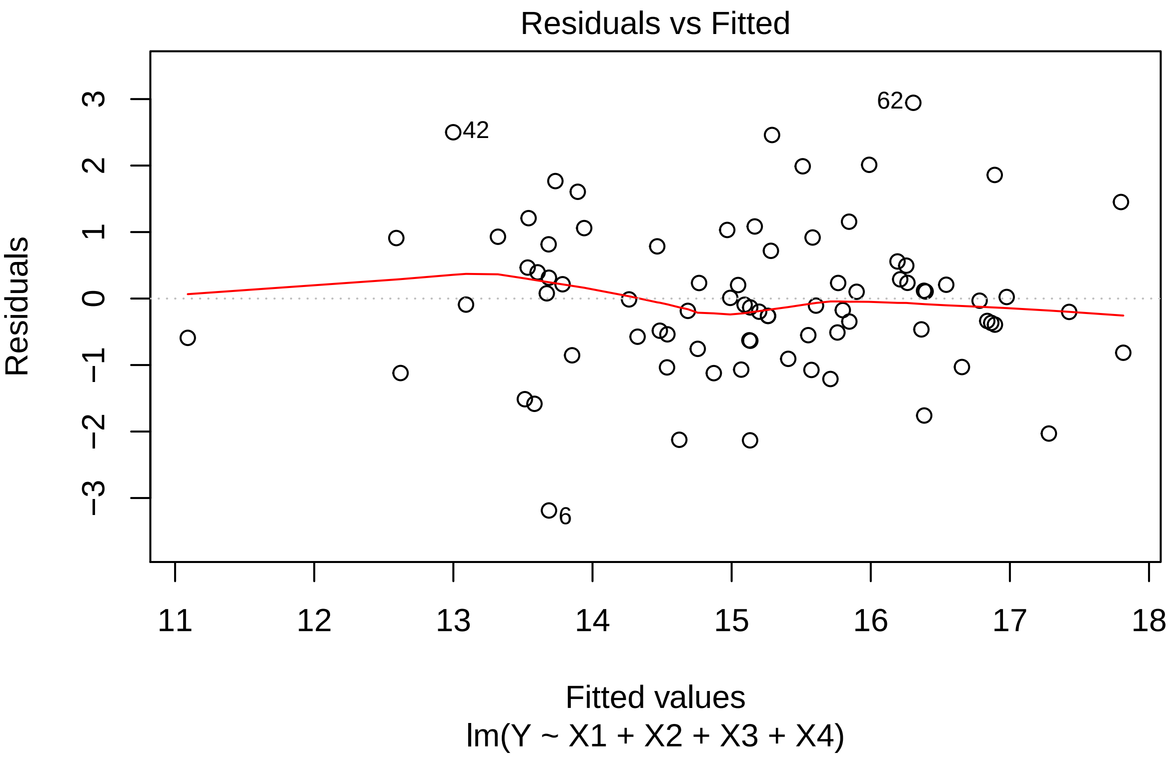
We have the following regression function :

$\hat{y} = (1.220e+01) - (1.420e-01)X1 + (2.820e-01)X2 + (6.193e-01)X3 + (7.924e-06)X4$

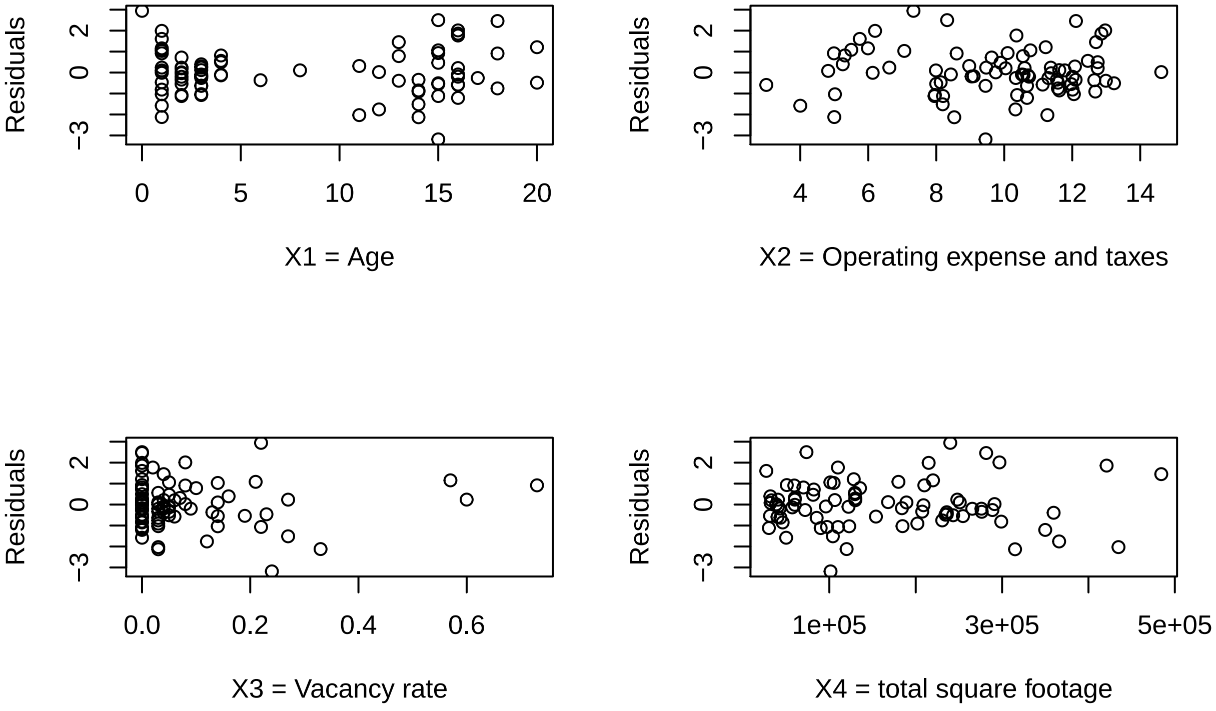
3. (4 pts) Plot the residuals against Y (one plot in one picture), against 4 predictor variables (4 plots in one picture), against each two-factor interaction (6 plots in in one picture). Are the residuals look like i.i.d. normally distributed, i.e., the pattern is unsystematic random around zero?

Answer :

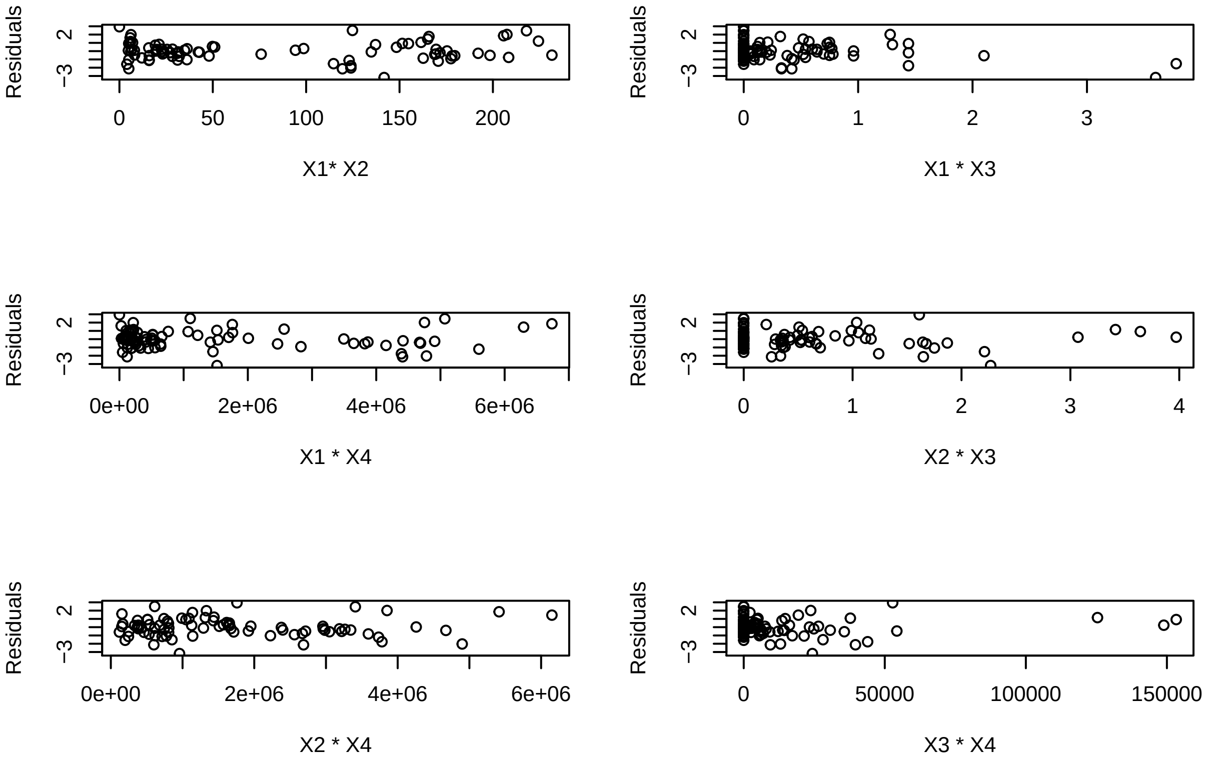
```r
#Residuals against fitted values
plot(model1, which = 1)
```



```
#Residuals against each predictor variables
par(mfrow=c(2,2))
plot(model1$residuals~commercial_data$X1, xlab="X1 = Age", ylab="Residuals")
plot(model1$residuals~commercial_data$X2, xlab="X2 = Operating expense and taxes", ylab="Residuals")
plot(model1$residuals~commercial_data$X3, xlab="X3 = Vacancy rate", ylab="Residuals")
plot(model1$residuals~commercial_data$X4, xlab="X4 = total square footage", ylab="Residuals")
```



```
#Residuals against each interaction term
par(mfrow=c(3,2))
plot(commercial_data$X1*commercial_data$X2, model1$residuals, xlab="X1* X2", ylab="Residuals")
plot(commercial_data$X1*commercial_data$X3, model1$residuals,  xlab="X1 * X3", ylab="Residuals")
plot(commercial_data$X1*commercial_data$X4, model1$residuals, xlab="X1 * X4", ylab="Residuals")
plot(commercial_data$X2*commercial_data$X3, model1$residuals, xlab="X2 * X3", ylab="Residuals")
plot(commercial_data$X2*commercial_data$X4, model1$residuals, xlab="X2 * X4", ylab="Residuals")
plot(commercial_data$X3*commercial_data$X4, model1$residuals, xlab="X3 * X4", ylab="Residuals")
```



From the residuals against $\hat{Y}$ plot, it seems that the residuals are symmetrically distributed around 0. In addition, we observe a slightly increasing trend in the size of the residuals against $\hat{Y}$. From residuals against the variables X1 - X4 plot, it seems that there is a curvature in the residuals v.s. X1. The residuals v.s. X2 - X4 are not conclusive. From the residuals v.s. interaction plots, It seems that the residuals has an increasing linear trend with interaction between X1 and X2, and decreasing linear trend with interaction between X1 and X3.

4. (2 pts) Show the ANOVA table of the regression model. Is the F-ratio significant? ($\alpha = 0.05$).

Answer :

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 14.819  14.819 11.4649  0.001125 **
## X2         1 72.802  72.802 56.3262 9.699e-11 ***
## X3         1  8.381   8.381  6.4846  0.012904 *
## X4         1 42.325  42.325 32.7464 1.976e-07 ***
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MSR <- sum(14.819 + 72.802 + 8.381 + 42.325)/4
MSE <- 1.293
F_ratio <- MSR/MSE
cat("F* =", F_ratio, "F(1-alpha, 4, n-5) = ", qf(.95, 4, 76))
```

```
## F* = 26.74536 F(1-alpha, 4, n-5) =  2.492049
```

We have F* = 26.74536 > F(1-alpha, 4, n-5) = 2.492049, We conclude that F is statiscally significant.

5. (2pts)What are the $R^2$ and $R^2_{adj}$ ?

Answer :

```
SSE <- 98.231
SSR <- sum(14.819+72.802+8.381+42.325)
SSTO <- sum(14.819+72.802+8.381+42.325+98.231)
MSE <- 1.293
R_squared <- SSR/SSTO
R_squared_adj <- 1 - (80/76)*(SSE/SSTO)
cat("R^2 = ", R_squared, ", R^2_adj = ", R_squared_adj)
```

```
## R^2 =  0.5847488 , R^2_adj =  0.5628934
```

6. (3 pts) Provide the point estimate of the mean values of the response at the following three new settings of X1 - X4. Provide the 95% confidence intervals and prediction intervals.

Answer :

```
X <- cbind(rep(1, nrow(commercial_data)), commercial_data$X1, commercial_data$X2, commercial_data$X3, co
Xt <- t(X)
Xp <- Xt%*%X
Xp.inv <- solve(Xp)

#Load the data
Xnew <- read.csv("/Users/kevinmouofo/Desktop/MATH 564/Midterm/new.txt",
                 sep="", header = TRUE )
alpha <- .05
Ynew <- predict(model1, Xnew)

t <- qt(1-(alpha/2),76)

ans <- sapply(1:nrow(Xnew), FUN=function(x){
  xh <- cbind(1, Xnew$X1[x], Xnew$X2[x], Xnew$X3[x],Xnew$X4[x])
  xht <- t(xh)
  s_yh <- sqrt(MSE*(xh%*%Xp.inv%*%xht))
  s_pred <- sqrt(MSE*(1 + xh%*%Xp.inv%*%xht))
  cat("For i = ", x, ", yh_hat =", Ynew[x], ", CI : {", Ynew[x]- t*s_yh, ",",
      Ynew[x] + t*s_yh, "}", ", PI : {", Ynew[x]- t*s_pred, ",",
      Ynew[x] + t*s_pred, "}\n")
})
```

```
## For i =  1 , yh_hat = 15.1485 , CI : { 14.76822 , 15.52877 } , PI : { 12.85206 , 17.44493 }
## For i =  2 , yh_hat = 15.54249 , CI : { 15.15358 , 15.9314 } , PI : { 13.24461 , 17.84037 }
## For i =  3 , yh_hat = 16.91384 , CI : { 16.18344 , 17.64424 } , PI : { 14.53424 , 19.29344 }
```

7. (3 pts) Fit the regression model $Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X4 + \epsilon$ (call this Model 2). State the

estimated regression function. Use partial F test to compare Model 1 and Model 2. What is your conclusion.

Answer :

```
model2 <- lm(Y~X1+X2+X4, data = commercial_data)
summary(model2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4, data = commercial_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0620 -0.6437 -0.1013  0.5672  2.9583
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.237e+01  4.928e-01  25.100  < 2e-16 ***
## X1          -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
## X2           2.672e-01  5.729e-02   4.663 1.29e-05 ***
## X4           8.178e-06  1.305e-06   6.265 1.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.132 on 77 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.5667
## F-statistic: 35.88 on 3 and 77 DF,  p-value: 1.295e-14
```

```
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2 + X4
## Model 2: Y ~ X1 + X2 + X3 + X4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     77 98.650
## 2     76 98.231  1   0.41975 0.3248 0.5704
```

We have the model 2 $\hat{Y} = (1.237e+01) - (1.442e-01)X1 + (2.672e-01)X2 + (8.178e-06)X4$

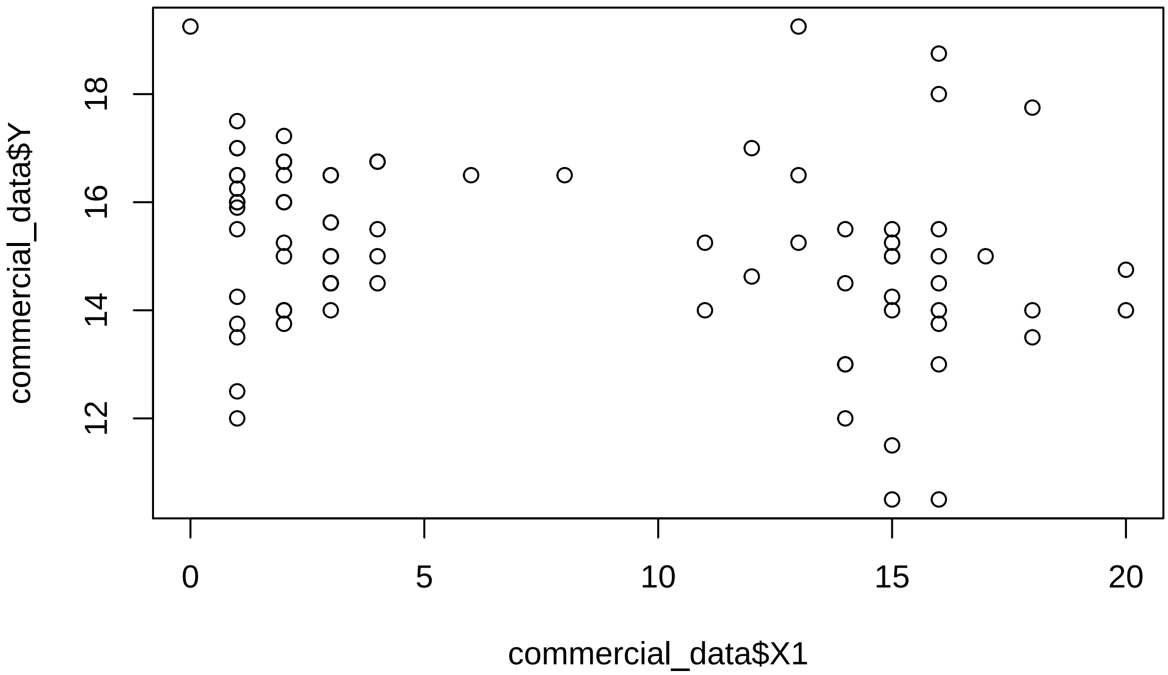
From anova table above, we have a p-value is 0.5704, indicating that dropping X3 does not significantly change the model. Thus it is reasonable to drop X3 from Model 1.

```
SSE_model2 <- 98.650
F_star <- (SSE_model2 - SSE) / (SSE/76)
if (F_star <= qf(.95, 1, 76)){
  cat("F* = ", F_star, "F(1-alpha, 1, n-5) = ", qf(.95, 1, 76),
      "\n Conclude X3 can be dropped from regression model that already have X1, X2, X4")
}else {
  cat("F* = ", F_star, "F(1-alpha, 1, n-5) = ", qf(.95, 1, 76), ", conclude can not be dropped")
}
```

```
## F* =  0.3241746 F(1-alpha, 1, n-5) =  3.96676
##  Conclude X3 can be dropped from regression model that already have X1, X2, X4
```

8. (2 pts) Plot Y against X1, do you observe any curvature?

```
plot(commercial_data$Y~commercial_data$X1)
```

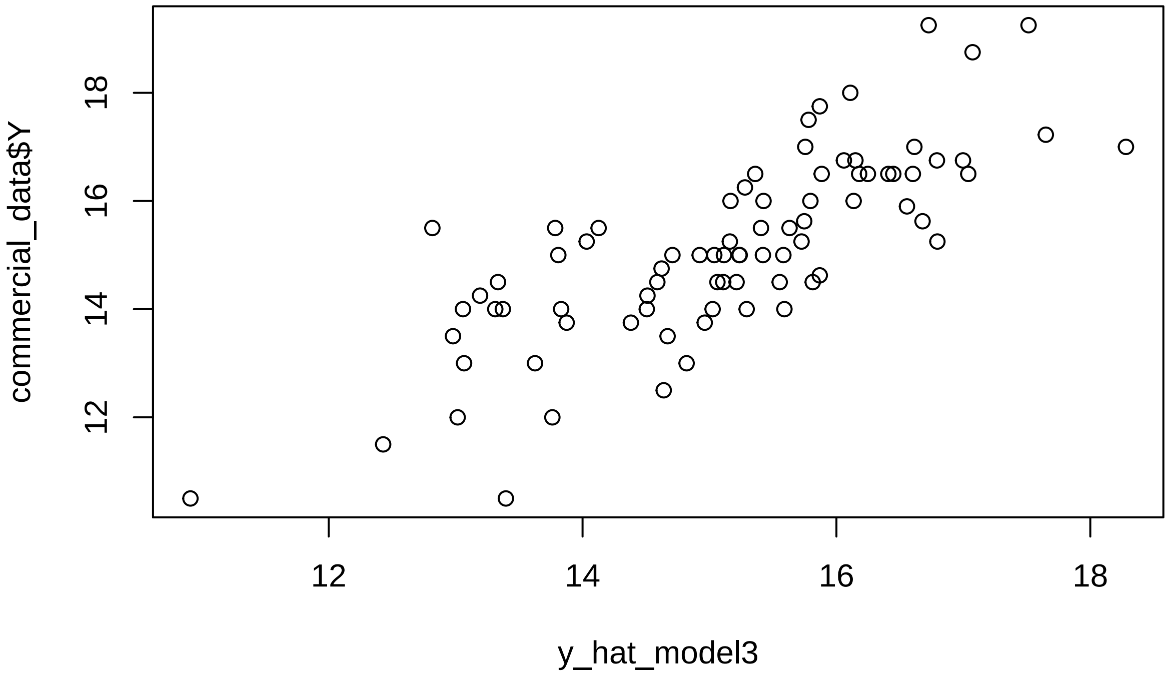


There seems to be a not obvious curvature between Y and X1.

9. (4 pts) Fit the regression model $Y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \beta 3 X4 + \beta 4 X1^2 + \epsilon$ (call this Model 3). State the estimated regression function. Plot Y against the fitted Y. Does Model 3 seem to be a good fit?

Answer :

```
X1_squared <- (commercial_data$X1)^2
commercial_data$X1_squared <- X1_squared
model3 <- lm(Y~X1+X2+X4+X1_squared, data = commercial_data)
dat <- data.frame(X1 = commercial_data$X1, X2 = commercial_data$X2, X4 = commercial_data$X4, X1_squared
y_hat_model3 = predict(model3, dat)
plot(commercial_data$Y~y_hat_model3)
```



```r
summary(model3)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X1_squared, data = commercial_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89596 -0.62547 -0.08907  0.62793  2.68309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.249e+01  4.805e-01  26.000  < 2e-16 ***
## X1          -4.043e-01  1.089e-01  -3.712  0.00039 ***
## X2           3.140e-01  5.880e-02   5.340 9.33e-07 ***
## X4           8.046e-06  1.267e-06   6.351 1.42e-08 ***
## X1_squared   1.415e-02  5.821e-03   2.431  0.01743 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 76 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.5927
## F-statistic:  30.1 on 4 and 76 DF,  p-value: 5.203e-15
```

```r
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X1          1 14.819  14.819 12.3036 0.0007627 ***
## X2          1 72.802  72.802 60.4463 2.968e-11 ***
## X4          1 50.287  50.287 41.7522 8.907e-09 ***
## X1_squared  1  7.115   7.115  5.9078 0.0174321 *
## Residuals  76 91.535   1.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have the regression equation :

$\hat{Y} = (1.249e + 01) - (4.043e - 01)X1 + (3.140e - 01)X2 + (8.046e - 06)X4 + (1.415e - 02)X1^2$

The plot Y against $\hat{Y}$ suggest that this model is a good enough fit.

From the Y v.s. Y plot, we can see that the Y and Y falls on a 45 degree diagonal lines, so it is a good fit.

10. (2 pts) Use partial F test to compare Model 2 and Model 3. Can you conclude $X^2$ is a significant term?

Answer :

```r
SSE_model3 <- 91.535
F_star_2 <- (SSE_model2-SSE_model3) / (SSE_model3/76)
if (F_star_2 <= qf(.95, 1, 76)){
  cat("F* = ", F_star_2, "F(1-alpha, 1, n-5) = ", qf(.95, 1, 76), " Conclude beta4 = 0")
}else {
  cat("F* = ", F_star_2, "F(1-alpha, 1, n-5) = ", qf(.95, 1, 76), ", conclude beta4 is not 0")
}
```

```
## F* =  5.907467 F(1-alpha, 1, n-5) =  3.96676 , conclude beta4 is not 0
```

```r
anova(model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2 + X4
## Model 2: Y ~ X1 + X2 + X4 + X1_squared
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     77 98.650
## 2     76 91.535  1    7.1154 5.9078 0.01743 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have a statistical evidence that $X1^2$ is significant.

From the partial F test above, the p-value is small enough thus we reject the null hypothesis, thus X12 is a significant term.

## Problem 8. Biopsy.

A breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumors for 699 patients up to 15 July 1992; each of nine attributes has been scored on a scale of 1 to 10, and the outcome is also known: benign (Y = 0) or malignant (Y = 1).

1. (5 pts) Build a generalized linear model on the outcome (called class in the data) with respect to the variables V1, . . . , V9. Write down the fitted model. Note that the data contain some NA values. Remove the rows containing NA to do all the questions.

Answer :

```r
library(MASS)
biopsy_data <- biopsy
#count missing values
cat("Total NA values before cleaning : ", sum(is.na(biopsy_data)),"\n" )
```

```
## Total NA values before cleaning :  16
```

```
#get rid of missing values
biopsy_data <- na.omit(biopsy_data)
#verify no missing values
cat("Total NA values after cleaning : ", sum(is.na(biopsy_data)), "\n" )
```

```
## Total NA values after cleaning :  0
```

```
# Get rid of ID column
biopsy_data$ID <- NULL
#fit the model
full.fit <- glm(class ~ ., family = binomial, data = biopsy_data)
summary(full.fit)
```

```
##
## Call:
## glm(formula = class ~ ., family = binomial, data = biopsy_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4841  -0.1153  -0.0619   0.0222   2.4698
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.10394    1.17488  -8.600  < 2e-16 ***
## V1            0.53501    0.14202   3.767 0.000165 ***
## V2           -0.00628    0.20908  -0.030 0.976039
## V3            0.32271    0.23060   1.399 0.161688
## V4            0.33064    0.12345   2.678 0.007400 **
## V5            0.09663    0.15659   0.617 0.537159
## V6            0.38303    0.09384   4.082 4.47e-05 ***
## V7            0.44719    0.17138   2.609 0.009073 **
## V8            0.21303    0.11287   1.887 0.059115 .
## V9            0.53484    0.32877   1.627 0.103788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 102.89  on 673  degrees of freedom
## AIC: 122.89
##
## Number of Fisher Scoring iterations: 8
```

We have the fitted equation : $\hat{\pi} = [1 + exp(10.10394 - 0.53501V1 + 0.00628V2 - 0.32271V3 - 0.33064V4 - 0.09663V5 - 0.38303V6 - 0.44719V7 - 0.21303V8 - 0.53484V9)]^{-1}$

2. (4 pts) From variables V1,. . . , V9, perform the forward stepwise regression with AIC criterion. What is the reduced model?

Answer :

```
fit0 <- glm(class~1, family = binomial, data=biopsy_data)
fit.forward <- step(fit0, scope=list(lower=~1, upper=~V1+V2+V3+V4+V5+V6+V7+V8+V9), direction = "forward
```

```
## Start:  AIC=886.35
## class ~ 1
##
##        Df Deviance    AIC
## + V2    1   254.76 258.76
## + V3    1   267.59 271.59
## + V6    1   340.63 344.63
## + V7    1   388.22 392.22
## + V5    1   452.88 456.88
## + V1    1   458.48 462.48
## + V4    1   463.34 467.34
## + V8    1   464.30 468.30
## + V9    1   717.52 721.52
## <none>      884.35 886.35
##
## Step:  AIC=258.76
## class ~ V2
##
##        Df Deviance    AIC
## + V6    1   166.31 172.31
## + V1    1   196.57 202.57
## + V7    1   207.32 213.32
## + V3    1   221.11 227.11
## + V8    1   223.01 229.01
## + V4    1   229.02 235.02
## + V5    1   239.54 245.54
## + V9    1   241.44 247.44
## <none>      254.76 258.76
##
## Step:  AIC=172.31
## class ~ V2 + V6
##
##        Df Deviance    AIC
## + V1    1   135.56 143.56
## + V8    1   150.94 158.94
## + V7    1   151.77 159.77
## + V3    1   153.08 161.08
## + V9    1   158.47 166.47
## + V4    1   161.03 169.03
## + V5    1   161.56 169.56
## <none>      166.31 172.31
##
## Step:  AIC=143.56
## class ~ V2 + V6 + V1
##
##        Df Deviance    AIC
## + V7    1   123.28 133.28
## + V8    1   123.84 133.84
## + V4    1   126.29 136.29
## + V3    1   129.54 139.54
## + V5    1   131.51 141.51
## + V9    1   133.01 143.01
## <none>      135.56 143.56
##
## Step:  AIC=133.28
## class ~ V2 + V6 + V1 + V7
##
##        Df Deviance    AIC
## + V4    1   114.60 126.60
## + V8    1   116.68 128.68
## + V3    1   119.34 131.34
## + V9    1   120.25 132.25
## + V5    1   120.77 132.77
## <none>      123.28 133.28
##
## Step:  AIC=126.6
## class ~ V2 + V6 + V1 + V7 + V4
##
##        Df Deviance    AIC
## + V8    1   108.97 122.97
## + V9    1   110.83 124.83
## + V3    1   111.52 125.52
## <none>      114.60 126.60
## + V5    1   113.10 127.10
##
## Step:  AIC=122.97
## class ~ V2 + V6 + V1 + V7 + V4 + V8
##
##        Df Deviance    AIC
## + V9    1   105.35 121.35
## <none>      108.97 122.97
## + V3    1   107.02 123.02
## + V5    1   108.38 124.38
##
## Step:  AIC=121.35
## class ~ V2 + V6 + V1 + V7 + V4 + V8 + V9
##
##        Df Deviance    AIC
## + V3    1   103.27 121.27
## <none>      105.35 121.35
## + V5    1   104.74 122.74
##
## Step:  AIC=121.27
## class ~ V2 + V6 + V1 + V7 + V4 + V8 + V9 + V3
##
##        Df Deviance    AIC
## <none>      103.27 121.27
## + V5    1   102.89 122.89
```

```
summary(fit.forward)
```

```
##
## Call:
## glm(formula = class ~ V2 + V6 + V1 + V7 + V4 + V8 + V9 + V3,
##     family = binomial, data = biopsy_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5208  -0.1150  -0.0628   0.0219   2.4113
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.975413   1.142150  -8.734  < 2e-16 ***
## V2           0.007847   0.207449   0.038 0.969827
## V6           0.388292   0.093581   4.149 3.34e-05 ***
## V1           0.533412   0.141690   3.765 0.000167 ***
## V7           0.460963   0.170307   2.707 0.006796 **
## V4           0.341475   0.122235   2.794 0.005213 **
## V8           0.225207   0.113255   1.988 0.046757 *
## V9           0.530181   0.325450   1.629 0.103298
## V3           0.339599   0.228428   1.487 0.137100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 103.27  on 674  degrees of freedom
## AIC: 121.27
##
## Number of Fisher Scoring iterations: 8
```

The reduced model is $\hat{\pi} = [1 + exp(9.975413 - 0.533412V1 - 0.007847V2 - 0.339599V3 - 0.341475V4 - 0.388292V6 - 0.460963V7 - 0.225207V8 - 0.530181V9)]^{-1}$

3. (3 pts) Use likelihood ratio test to compare the model in Question 1 and 2. Is there significant difference between the two?

Answer :

$H0 : \beta 5 = 0$, Ha : $\beta 5 \neq 0$

Test statistic : $G^2 = -2log_e(\frac{L(R)}{L(F)})$.

if ($G^2 \leq X^2(.95;3)$) Conclude H0, Otherwise conlude Ha.

```r
reduced.fit <- glm(class ~V1+V2+V3+V4+V6+V7+V8+V9, family = binomial, data = biopsy_data)
#summary(reduced.fit)
anova(reduced.fit, full.fit)
```

```
## Analysis of Deviance Table
##
## Model 1: class ~ V1 + V2 + V3 + V4 + V6 + V7 + V8 + V9
## Model 2: class ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9
##   Resid. Df Resid. Dev Df Deviance
## 1       674     103.27
## 2       673     102.89  1  0.37714
```

```r
alpha <- .05
x_square <- qchisq(1- alpha,1)
g_square <- 0.37714

if(g_square <= x_square){
  cat("G_square =", g_square, ", chi-squared(0.95;1) =", x_square, ", Conclude H0, X5 can be dropped")
}else
  cat("G_square =", g_square, ", chi-squared(0.95;1) =", x_square,", Conclude Ha")
```

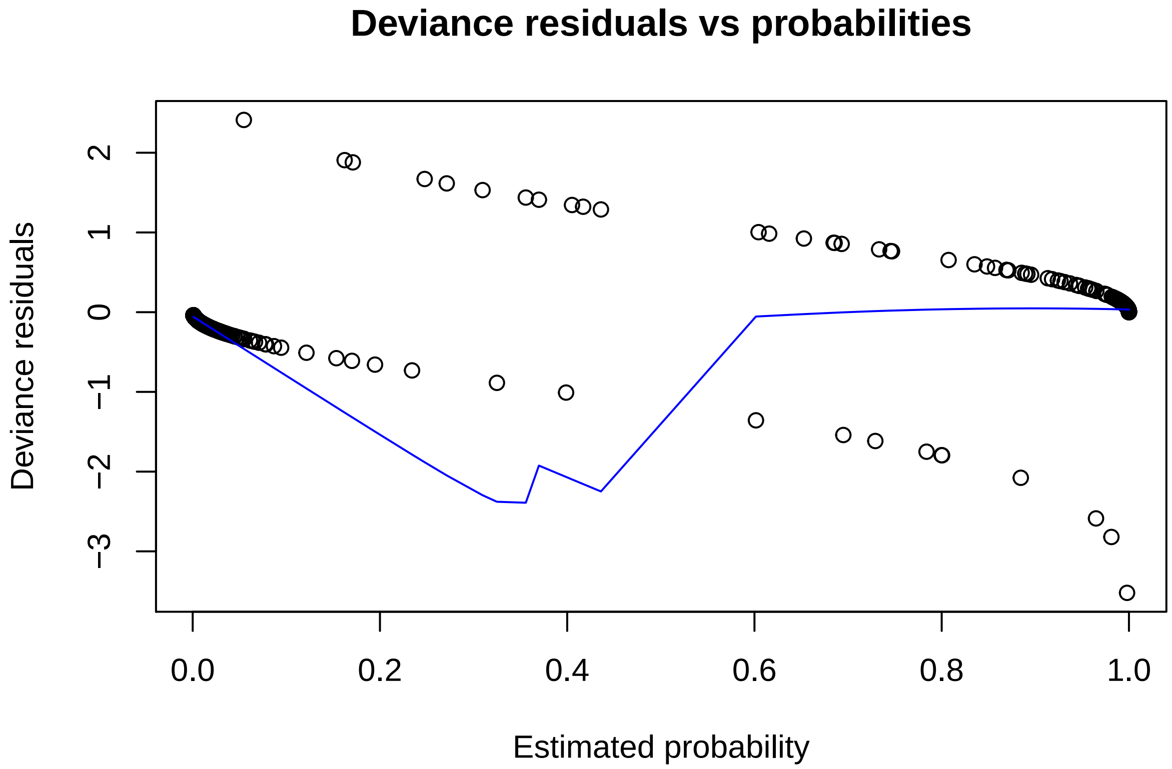
```
## G_square = 0.37714 , chi-squared(0.95;1) = 3.841459 , Conclude H0, X5 can be dropped
```

We can see that there is no significant difference between the 2 models.

4. (3 pts) Using the reduced model from Question 2, plot the deviance residual against the fitted value and perform the residual diagnostics.

Answer :

```r
r_D <- residuals(reduced.fit,type="deviance")
plot(reduced.fit$fitted.values, r_D, main="Deviance residuals vs probabilities",
     xlab = "Estimated probability", ylab = "Deviance residuals")
lines(lowess(reduced.fit$fitted.values, r_D), col="blue")
```



We can see that the smooth lowess line is not horizontal, we can conclude that we have an inadequate model.