Fitting a Model to Happiness Using Regression Analysis

By Kevin Ayala and Mohammad Khan
Section: Lizzie/Monday 5:00pm – 5:50pm
June 6, 2018

Abstract

We use a sample size of one hundred to create a best fit model that can accurately predict an individual's happiness level based on three predictors. This report details the methodology that was used in assessing the best possible regression model. First, we describe our data set, then we move on to the statistical methodologies used during this study. We explain appropriate graphs and results. We follow up with a discussion of our results and conclude with the optimal model where the error in variation is minimized.

Introduction

For most people, a goal in life is to find happiness. The question is, what and how does one become happy? We approach this arbitrary question with data from individuals and use statistical analysis to propose a pedestal model. Our dataset contains a sample size of 100 individuals. These Individuals were asked to rate on a 1-10-point scale on how happy they were with their life. A rating of 1 indicated "very unhappy" while a rating of 10 signified they were "extremely happy". Three questions were later then asked; in regard to their relationship status on a 1-10-point scale, the number of hours they worked on a weekly basis, and whether they identified themselves as either female or male. A score of 1 during the relationship status question indicated "loneliness" while a 10 meant "deeply in love". Males were assigned a value of 0 and females were assigned a value of 1.

Social expectations have set marriage as the symbolic representation of love and happiness across a variety of cultures. It is our belief that relationship status will be the biggest contributor of variance explained in an individual's happiness. We expect to see women and men have various levels of relationship satisfaction, we based this on the stereotype that men are on average less happy than females. On work hours, we do not expect this predictor to be indicative on the outcome, we believe it to have great variability amongst individuals as human beings lie within a spectrum of satisfaction catalysts, some people thrive on work while others loathe it.

The Method:

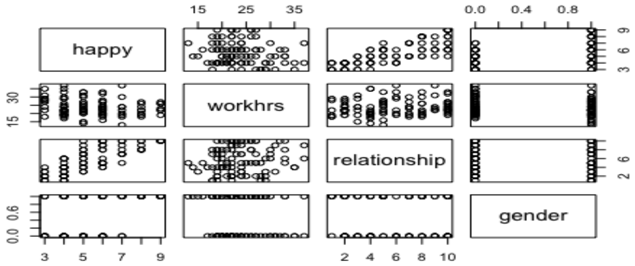
A matrix scatterplot was used to obtain a brief overview of the data and to see which predictors would potentially interact well with each other. Exploratory data analysis was also used regarding evidence of women being happier than men in relationships. After the overview, a first order linear model was fitted to the dataset with happiness as the outcome variable. Workhours, relationship, and gender where then used as the predictor models for the outcome.

After creating this first model, we were able to retrieve an overall p-value as well as corresponding partial p-values for each individual predictor. An F-statistic and coefficient of determination was also retrieved for this model. Following up on potential interactions, a second order model was fitted to see whether the predictors had any significant interaction with one another. We again analyzed this second order model for overall p-value and corresponding p-values for all possible two-way interactions between predictors. Then, we evaluated the extra sum of squares to see if two-way interactions could be recalibrated into the model. We then proceeded to confirm the final model using stepwise regression. We employed the use of forward and backward stepwise regression to conclude a convergence. After having this final model, we again confirmed that we did indeed make an increase in the overall fit, proving the model fits even better than chance alone relative to model 1.

We conducted diagnostic checks to make sure the assumption of normality, identically distributed and constant variance were not violated. Independence is an assumption we will make for the purposes of this study. The normality assumption was tested by a qq-plot and histogram. We checked the constant variance assumption by generating a residuals plot with a line of slope 0. We did this to make comparisons of the expected deviations and sample deviations. Once diagnostic was passed, we again looked at model 3 and analyzed the corresponding partial p-values, F-statistic, and coefficient of determination, we then addressed our hypothesis and compared the results to our expectations.

Results

Overview with a scatterplot matrix revealed a moderate-strong positive relationship between happiness level and relationship status. This was the strongest relationship in the matrix. The weakest was a negative relationship with hours worked and happiness level which indicated a small impact on the outcome. Gender seemed to be insignificant in any scatterplot as the data was either 0 or 1. The only thing that stood out was the Happy and Gender scatter plot where men averaged less on happiness. Men on average answered at a lower mean on the relationship quality question than females did. This is made clearer in plot 1 in the referenced graphs section. Relationship and workhours seemed to have an extremely weak relationship and women seemed to be happier when it came to work hour; there was no interaction with worked hours and happiness for gender differences. (Referenced Graphs, plot 2)



The first fitted model of $\text{Happiness}_i = 3.54 - .0711\text{workhours}_i + .48\text{relationship}_i + 1.55\text{gender}_i$ yielded a strong coefficient of determination of .907, meaning a 90.7% overall fit. With a p-value of $2.2e^{-16}$, there was convincing evidence to reject the null hypothesis that all predictor slopes were equivalent to zero. Thus, we immediately concluded at least one predictor was significant. The partial p-values for each predictor are workhours($2.52e^{-09}$), relationship($2e^{-16}$) and gender($2e^{-16}$). They all are close to zero, thus we concluded not to drop any predictor. The F-statistic for this model was 312.2, a relatively substantial number which is quite indicative that the numerator of variance explained is high compared to the denominator of unexplained variance. Further analysis with an ANOVA test on the first order model yielded sequential p-values close to zero, workhours ($9.218e^{-06}$), relationship ($2.2e^{-16}$) and gender ($2.2e^{-16}$) in that order.

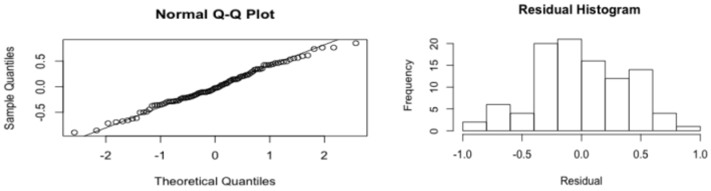
The second order model was model 1 added with all possible interactions, it yielded an F-Statistic of 297.4, with a corresponding p-value of $2.2e^{-16}$. We rejected the null hypothesis that adding interaction terms would contribute nothing to the model, hence we knew the optimal model would contain at least one interaction. All possible interactions had the following partial p-values gender*workhrs(.5810), workhrs*relationship(.5030), and gender*relationship ($3.26e^{-14}$). Gender*relationship was the only significant interaction at an alpha level of .05. The rest were dropped from a final model due to failing to reject the null hypothesis of having a slope equivalent to 0. Gender itself having a p-value of .3447, was not significant. It was not dropped from the final model as it plays a vital interaction role and was kept via convention.

The second order model yielded a R^2 value of 95.05% with all interactions, much better than the original value of 90.7%. Further analysis using extra sums of squares on both model 1 and 2 gave an output in an increase of Residual Sum Squares of 14.384. The overall p-value of $1.047e^{-12}$ during extra sums of squares test confirmed the second model was a better fit than the first model. A ANOVA test was used exclusively for model 2 to obtain the sequential effect of the predictors. We observed sequential p-values of $2.2e^{-16}$, 0.0002352, $2.2e^{-16}$, and $3.01e^{-14}$ for gender, workhours, relationship, and gender*relationship respectively. Each predictor was significant given the previous predictor already in the model. We chose model 3 to have these predictors as it would be the optimal model. This yielded a p-value of 95.05%, retaining model 2 R^2 value, we also noticed that relationship was the key contributor to the model, along with relationship*gender.

We then conducted stepwise regression in both direction to see if there would be any paradox of different final model's due to the direction used. The forward direction ran the model in the order of relationship + gender + workhrs + relationship:gender. The backward direction eliminated predictors in the order of gender + workhrs + relationship + gender: relationship. Both directions yielded the same model solidifying our final model choice of

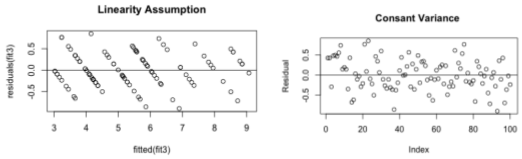
$Happines_i = 4.28774 +.352relationship_i + .178gender_i -.07workhours_i + .242relationship*gender_i$

With the result of a final model, we then ran it through diagnostics, making sure that it did not violate the assumptions of normality, constant variance, and identically distributed. The qq-normal plot checked normality by comparing theoretical quantiles to the sample quantiles. A line was then inputted at a 45-degree angle to check that no plotted residuals were extreme outliers, which would violate our normality assumption.



We also cross validated the normality assumption by creating a histogram of the residuals and concluded that there was no significant violation in the assumption of normality as the qq-plot and residual histogram were passable. There are no extreme outliers in the qq-plot which led us to also conclude that errors were identically distributed.

The constant variance assumption was also checked by employing the use of residuals.



We found no evidence of the variance being non-constant hence it passed the constant variance assumption while the left figure shows compelling evidence of linearity with each predictor following a linear trend. Thus, we concluded model 3 passed all diagnostics.

Significant predictors and interactions in the model was the intercept, work hours, relationship, and a gender*relationship interaction with p-values of 1.01e^{-11}, 0.0112, 1.30e^{-06}, and 3.26e^{-14} respectively. Our final model has an R^2 value of 95.05%, a rather impressive figure. This means that 95.05% of the variance of an individual's happiness level can be explained by knowing their gender, work hours, quality of relationship, and the level of interaction for relationship*gender. (Referenced Graphs, Plot 1)

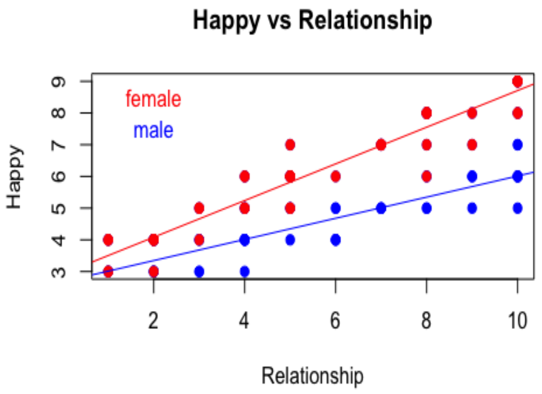
We can infer from our model that a female who does not work and has a 0 level of relationship will have a predicted happiness level of 4.46574 while a male with equivalent circumstances will have a reported happiness level of 4.28774. We can also infer from the model that overall happiness level will decrease by .07 for every hour worked, equivalent across both genders. The final model proves women and men have different happiness levels, women are .178 points above men just by gender alone in relation to happiness.

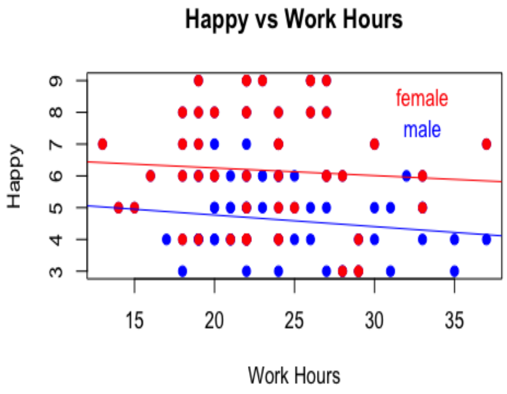
Discussion

. The data proves we were wrong in our belief that working hours would play no significant role on predicted happiness level and instead actually decreased happiness for every hour a person worked. We also correctly assumed that the relationship predictor would have the greatest contribution in the final model as relationship status alone increases happiness for every unit measurement on an individual's relationship quality.

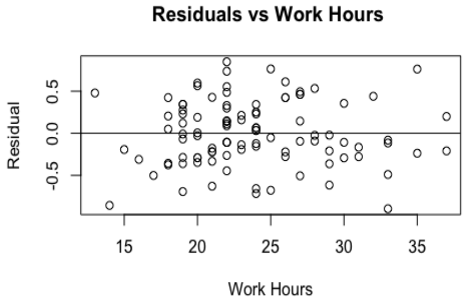
In plot 1 on Referenced Graphs, women on average rate their relationship quality on a much higher frequency than men. Men rate their relationship quality on a much lower scale. Women also are happier than men if they do not work (Referenced Graphs, Plot 2). We find that the only limitations of the model are data collection method used, and being unable to make an inference on the general population as we do not know where the data was collected. Consideration for future study are including predictors such as age, income level, and perhaps even number of past relationships.

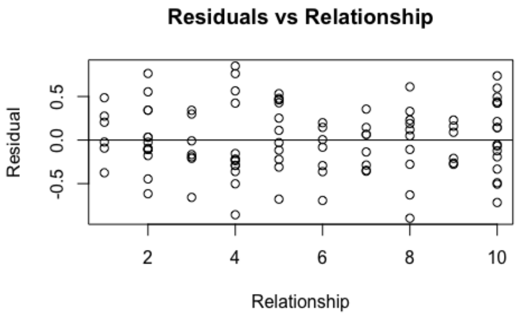
Referenced Graphs

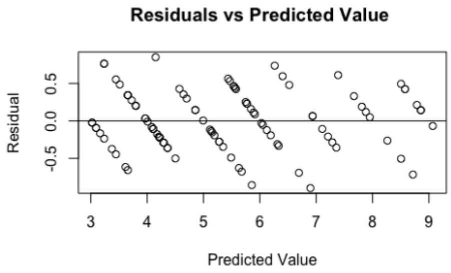




Extra Graphs







Appendix

```
 > #import data file
> projectdata = read.table ("/Users/Kevin/Downloads/projdata126.txt",header=TRUE)
> View(projectdata) #list entire data
> summary(projectdata)
     happy          gender        workhrs       relationship
 Min.   :3.00   Min.   :0.00   Min.   :13.00   Min.   : 1.00
 1st Qu.:4.00   1st Qu.:0.00   1st Qu.:20.00   1st Qu.: 3.00
 Median :5.00   Median :1.00   Median :22.50   Median : 5.00
 Mean   :5.42   Mean   :0.52   Mean   :23.76   Mean   : 5.69
 3rd Qu.:6.00   3rd Qu.:1.00   3rd Qu.:27.00   3rd Qu.: 8.00
 Max.   :9.00   Max.   :1.00   Max.   :37.00   Max.   :10.00
> head(projectdata)
  happy gender workhrs relationship
1     6      1      18            4
2     9      1      26           10
3     4      0      30            6
4     7      1      13            5
5     9      1      27           10
6     6      1      27            5
> str(projectdata)
'data.frame':	100 obs. of  4 variables:
 $ happy       : int  6 9 4 7 9 6 4 7 5 8 ...
 $ gender      : int  1 1 0 1 1 1 0 0 0 1 ...
 $ workhrs     : int  18 26 30 13 27 27 22 22 22 20 ...
 $ relationship: int  4 10 6 5 10 5 2 10 6 8 ...
>
> #Scatterplot matrix and liner model
> pairs(happy~workhrs+relationship+gender, data = projectdata)
> fit1 = lm(happy~workhrs+relationship+gender, data = projectdata)
> summary(fit1)

Call:
lm(formula = happy ~ workhrs + relationship + gender, data = projectdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04590 -0.35802 -0.02218  0.37697  1.26763
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.54123    0.28090  12.607  < 2e-16 ***
workhrs      -0.07118    0.01082  -6.576 2.52e-09 ***
relationship  0.48538    0.01821  26.649  < 2e-16 ***
gender        1.55447    0.10700  14.528  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5302 on 96 degrees of freedom
Multiple R-squared:  0.907,	Adjusted R-squared:  0.9041 
F-statistic: 312.2 on 3 and 96 DF,  p-value: < 2.2e-16

>
>
>
> #Evaluate the Extra SS associated with adding interaction terms to model
> #by comparing a model with 2-way interactions to the model
> fit2 = lm(happy~.^2,data = projectdata)
> summary(fit2)#create linear model (fit3) most significant

Call:
lm(formula = happy ~ .^2, data = projectdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.86671 -0.26448 -0.04598  0.30179  0.86016 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           3.906103   0.502528   7.773 1.01e-11 ***
gender                0.379229   0.399279   0.950   0.3447    
workhrs              -0.053897   0.020836  -2.587   0.0112 *  
relationship          0.401203   0.077494   5.177 1.30e-06 ***
gender:workhrs       -0.008898   0.016067  -0.554   0.5810    
gender:relationship   0.243410   0.027169   8.959 3.26e-14 ***
workhrs:relationship -0.002106   0.003132  -0.672   0.5030    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3933 on 93 degrees of freedom
Multiple R-squared:  0.9505,	Adjusted R-squared:  0.9473 
```

```
F-statistic: 297.4 on 6 and 93 DF,  p-value: < 2.2e-16

> #Use the anova function to compare the two models
> #the Extra SS associated with adding 2-way interactions significant?
> anova(fit2,fit1)
Analysis of Variance Table

Model 1: happy ~ (gender + workhrs + relationship)^2
Model 2: happy ~ workhrs + relationship + gender
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     93 14.384
2     96 26.991 -3   -12.606 27.168 1.047e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(fit2)
Analysis of Variance Table

Response: happy
                     Df  Sum Sq Mean Sq   F value    Pr(>F)
gender                1  61.439  61.439  397.2186 < 2.2e-16 ***
workhrs               1   2.265   2.265   14.6433 0.0002352 ***
relationship          1 199.666 199.666 1290.8985 < 2.2e-16 ***
gender:workhrs        1   0.077   0.077    0.4965 0.4827852
gender:relationship   1  12.460  12.460   80.5558  3.01e-14 ***
workhrs:relationship  1   0.070   0.070    0.4521 0.5029995
Residuals            93  14.384   0.155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> #most significant model before stepwise regression
> fit3 = lm(happy~gender+workhrs+relationship+relationship*gender,
+           data = projectdata)
> summary(fit3)

Call:
lm(formula = happy ~ gender + workhrs + relationship + relationship *
    gender, data = projectdata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89700 -0.26709 -0.02701  0.28099  0.84955
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.287745   0.222865  19.239  < 2e-16 ***
gender              0.178353   0.171396   1.041    0.301
workhrs            -0.070259   0.007978  -8.807 5.85e-14 ***
relationship        0.352098   0.019935  17.662  < 2e-16 ***
gender:relationship  0.241580   0.026716   9.043 1.84e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3908 on 95 degrees of freedom
Multiple R-squared:   0.95,              Adjusted R-squared:  0.9479
F-statistic: 451.7 on 4 and 95 DF,  p-value: < 2.2e-16

>
>  #stepwise regression, confirming if model 3 is the best model
> null=lm(happy~1,data = projectdata) #the null model
> full=lm(happy~.^2,data = projectdata) #full model
> step(null,scope=list(lower=null,upper=full),direction='forward') #stepwise process in the forward direction ect.
Start:  AIC=108.6
happy ~ 1

               Df Sum of Sq    RSS     AIC
+ relationship  1   183.983 106.38  10.182
+ gender        1    61.439 228.92  86.821
+ workhrs       1     6.171 284.19 108.447
<none>                      290.36 108.595

Step:  AIC=10.18
happy ~ relationship

          Df Sum of Sq     RSS     AIC
+ gender   1    67.227  39.150 -87.777
+ workhrs  1    20.043  86.335  -8.694
<none>                 106.377  10.182

Step:  AIC=-87.78
happy ~ relationship + gender

                      Df Sum of Sq    RSS      AIC
+ gender:relationship  1    12.802 26.348 -125.376
+ workhrs              1    12.159 26.991 -122.967
```

```
<none>                        39.150  -87.777

Step:  AIC=-125.38
happy ~ relationship + gender + relationship:gender

          Df Sum of Sq    RSS     AIC
+ workhrs  1    11.843 14.506 -183.06
<none>                 26.348 -125.38

Step:  AIC=-183.06
happy ~ relationship + gender + workhrs + relationship:gender

                       Df Sum of Sq    RSS     AIC
<none>                              14.506 -183.06
+ workhrs:relationship  1  0.073733 14.432 -181.57
+ gender:workhrs        1  0.051244 14.454 -181.42

Call:
lm(formula = happy ~ relationship + gender + workhrs + relationship:gender,
    data = projectdata)

Coefficients:
        (Intercept)         relationship               gender
            4.28774              0.35210              0.17835
            workhrs  relationship:gender
           -0.07026              0.24158

> step(full,direction='backward')
Start:  AIC=-179.9
happy ~ (gender + workhrs + relationship)^2

                       Df Sum of Sq    RSS     AIC
- gender:workhrs        1    0.0474 14.432 -181.57
- workhrs:relationship  1    0.0699 14.454 -181.42
<none>                              14.384 -179.90
- gender:relationship   1   12.4145 26.799 -119.68

Step:  AIC=-181.57
happy ~ gender + workhrs + relationship + gender:relationship +
    workhrs:relationship

                       Df Sum of Sq    RSS     AIC
```

```
- workhrs:relationship  1    0.0737 14.506 -183.06
<none>                              14.432 -181.57
- gender:relationship   1   12.4494 26.881 -121.37

Step:  AIC=-183.06
happy ~ gender + workhrs + relationship + gender:relationship

                      Df Sum of Sq    RSS     AIC
<none>                              14.506 -183.06
- workhrs              1    11.843 26.348 -125.38
- gender:relationship  1    12.485 26.991 -122.97

Call:
lm(formula = happy ~ gender + workhrs + relationship + gender:relationship,
    data = projectdata)

Coefficients:
        (Intercept)               gender              workhrs
            4.28774              0.17835             -0.07026
       relationship  gender:relationship
            0.35210              0.24158

> step(null,scope=list(upper=full),direction='both') #both directions converge to a single model
Start:  AIC=108.6
happy ~ 1

               Df Sum of Sq    RSS     AIC
+ relationship  1   183.983 106.38  10.182
+ gender        1    61.439 228.92  86.821
+ workhrs       1     6.171 284.19 108.447
<none>                      290.36 108.595

Step:  AIC=10.18
happy ~ relationship

               Df Sum of Sq     RSS     AIC
+ gender        1    67.227  39.150 -87.777
+ workhrs       1    20.043  86.335  -8.694
<none>                      106.377  10.182
- relationship  1   183.983 290.360 108.595

Step:  AIC=-87.78
```

```
happy ~ relationship + gender

                        Df Sum of Sq     RSS      AIC
+ gender:relationship  1    12.801  26.348 -125.376
+ workhrs              1    12.159  26.991 -122.967
<none>                              39.150  -87.777
- gender               1    67.227 106.377   10.182
- relationship         1   189.772 228.921   86.821

Step:  AIC=-125.38
happy ~ relationship + gender + relationship:gender

                        Df Sum of Sq    RSS      AIC
+ workhrs              1    11.843 14.506 -183.063
<none>                             26.348 -125.376
- relationship:gender  1    12.802 39.150  -87.777

Step:  AIC=-183.06
happy ~ relationship + gender + workhrs + relationship:gender

                         Df Sum of Sq    RSS     AIC
<none>                              14.506 -183.06
+ workhrs:relationship  1    0.0737 14.432 -181.57
+ gender:workhrs        1    0.0512 14.454 -181.42
- workhrs               1   11.8427 26.348 -125.38
- relationship:gender   1   12.4853 26.991 -122.97

Call:
lm(formula = happy ~ relationship + gender + workhrs + relationship:gender,
    data = projectdata)

Coefficients:
        (Intercept)         relationship               gender
            4.28774              0.35210              0.17835
            workhrs  relationship:gender
           -0.07026              0.24158

>
>
> #checking assumption for fit3 linear model
> resd2<-residuals(fit3)
> pred2<-fitted(fit3)
> qqnorm(resd2)
> qqline(resd2)
> hist(resd2,xlab="Residual",main="Residual Histogram ")
> plot(resd2,ylab="Residual",main="Consant Variance ")
> abline(h=0)
> plot(fitted(fit3),residuals(fit3), main="Linearity Assumption") #added linearity assumption
> abline(h=0)
> #
> plot(resd2~projectdata$workhrs,xlab="Work Hours",ylab="Residual",
+      main="Residuals vs Work Hours ")
> abline(h=0)
> plot(resd2~projectdata$relationship,xlab="Relationship",ylab="Residual",
+      main="Residuals vs Relationship")
> abline(h=0)
> plot(resd2~pred2,xlab="Predicted Value",ylab="Residual",
+      main="Residuals vs Predicted Value ")
> abline(h=0)
>
> # graph of happy and workhours between male(0) and female(1)
> plot(happy~workhrs,xlab="Work Hours",ylab="Happy",pch=19,col="blue",
+      data=projectdata,main="Happy vs Work Hours")
> points(happy[gender==1]~workhrs[gender==1],pch=19,col="red",
+        data=projectdata) #specfying for females
> abline(lm(happy[gender==0]~workhrs[gender==0],data=projectdata),col="blue")
> abline(lm(happy[gender==1]~workhrs[gender==1],data=projectdata),col="red")
> text(33,8.5,"female",col="red")
> text(33,7.5,"male",col="blue")
>
> # graph of happy and relationship between male(0) and female(1)
> plot(happy~relationship,xlab="Relationship",ylab="Happy",pch=19,col="blue",
+      data=projectdata,main="Happy vs Relationship")
> points(happy[gender==1]~relationship[gender==1],pch=19,col="red",
+        data=projectdata) #specfying for females
> abline(lm(happy[gender==0]~relationship[gender==0],data=projectdata),col="blue")
> abline(lm(happy[gender==1]~relationship[gender==1],data=projectdata),col="red")
> text(2,8.5,"female",col="red")
> text(2,7.5,"male",col="blue")
>
```