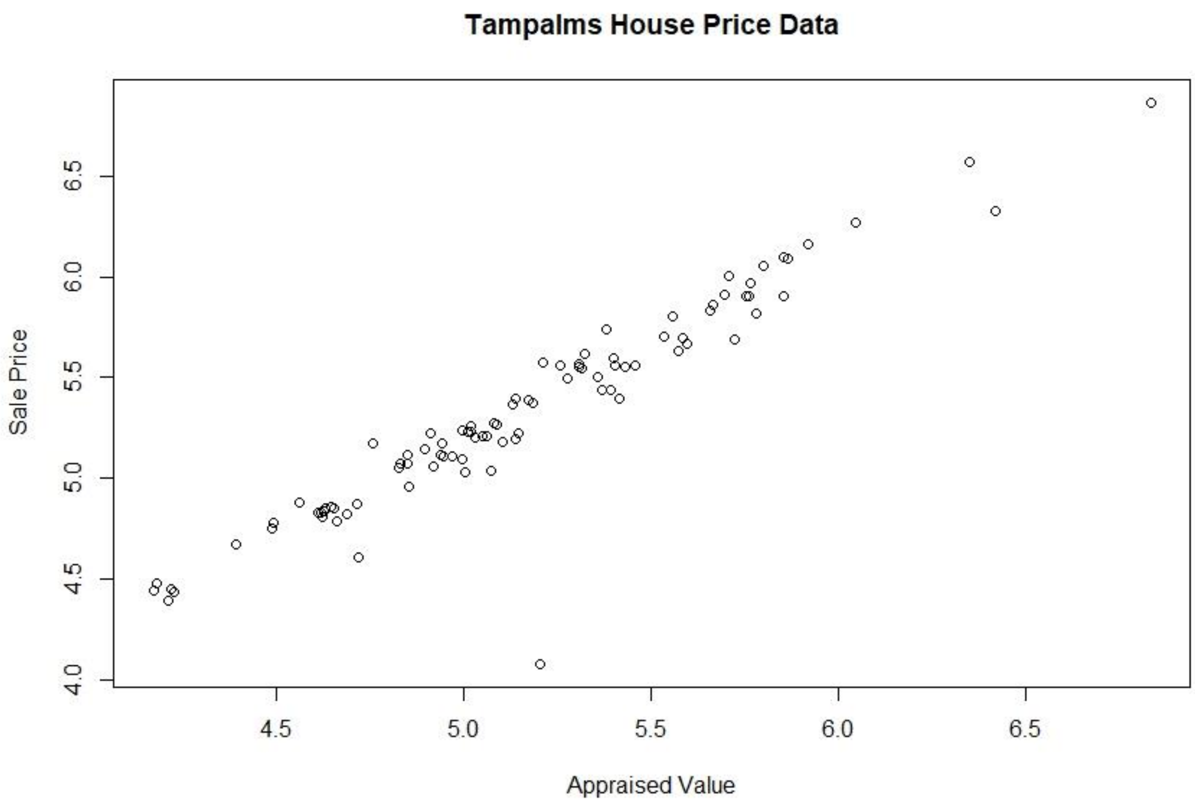
Edna Diaz
80396667
Stat 4380
Project I

Tampa Palms House Price Data:
Project I



1. Based on the strong positive correlation of the points in the scatterplot, a straight line model is an appropriate fit to the data.
2. t-value represent the t-test statistic value which is 39.452
   DF is the degrees of freedom which is 90
   P-value < 2.2e-16
   The confidence interval = [.9583052, .9816214]
   The sample estimate stands for the correlation coefficient which is: .9722849

   Our P-value indicates that it is smaller than 2.2e-16 which means that it is less than our significance level α = .05. Based on this result we can conclude that appraised values and sale values are strongly correlated, with a coefficient of .9722849 and P-value < 2.2e-16.

Edna Diaz
80396667
Stat 4380
Project I

```
        Pearson's product-moment correlation

data:  tampalms$appraised and tampalms$sale
t = 39.452, df = 90, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9583052 0.9816214
sample estimates:
      cor
0.9722849
```

3. Unbiased Constant Variance:

```
var(fit$residuals)
[1] 1063.141
```

LS estimates for regression parameters:

```
Coefficients:
       (Intercept)  tampalms$appraised
            20.942               1.069
```

Table of parameter estimates:

p-value is less than .05 therefore we reject the null hypothesis, which shows that there is a strong correlation between appraised values and sales prices. If appraised values increase, sales prices will likely increase and vice versa.

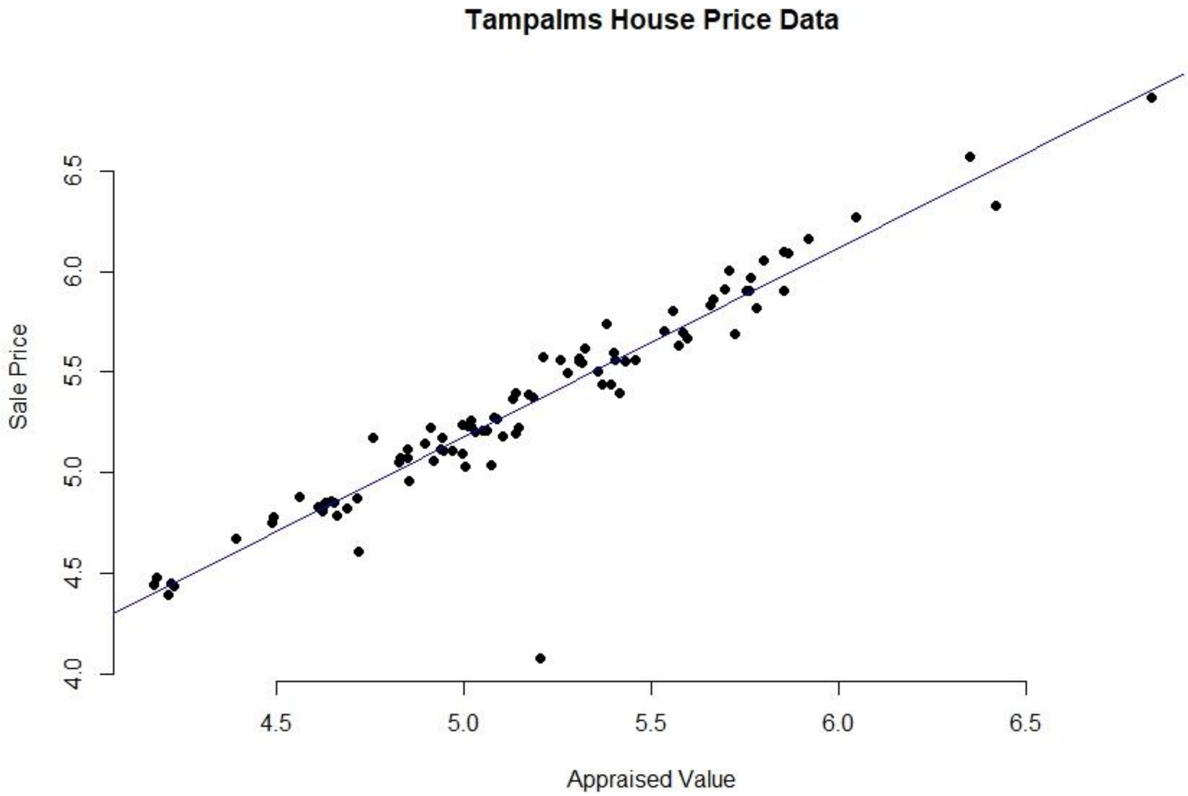
The slope coefficient 1.06873 shows us that the slope will be positive and line will be positively increasing.

```
Residuals:
     Min       1Q   Median       3Q      Max
-156.741  -10.976   -0.979   12.258   82.188

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        20.94193    6.44617   3.249  0.00163 **
tampalms$appraised  1.06873    0.02709  39.452  < 2e-16 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 32.79 on 90 degrees of freedom
Multiple R-squared:  0.9453,         Adjusted R-squared:  0.9447
F-statistic:  1556 on 1 and 90 DF,  p-value: < 2.2e-16
```

Edna Diaz
80396667
Stat 4380
Project I



This represents the LS line on the scatterplot for Tampalms house price data. By the looks of this representation, we can say that there is a high linear correlation. This indicates that appraised value strongly affects the sales price, in this scenario.

4. Anova Table:

```
            Df Sum Sq Mean Sq F value Pr(>F)    
y            1 21.404  21.404   773.9 <2e-16 ***
Residuals   90  2.489   0.028                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value is less than 2e-16, which means that we reject the null hypothesis in this scenario, the population means are not all equal, and there exists a strong correlation between appraised values and Sale Price.

F-value is greater than p-value, so it is assumed that the regression model is a better fit in this case.
MS – contains that average number of squares divided by degrees of freedom or DF, the variation in sample mean is 21.404 and for residual is 2.489.

Edna Diaz
80396667
Stat 4380
Project I

SS – shows the variation in the observed data.
DF- shows the degrees of freedom

R^2 = .9453 which means, that 94.53% of variation in appraisal values is explained by the model. Most of the data is represented by the fitted model in this situation.

5.List of y-hat, and residuals

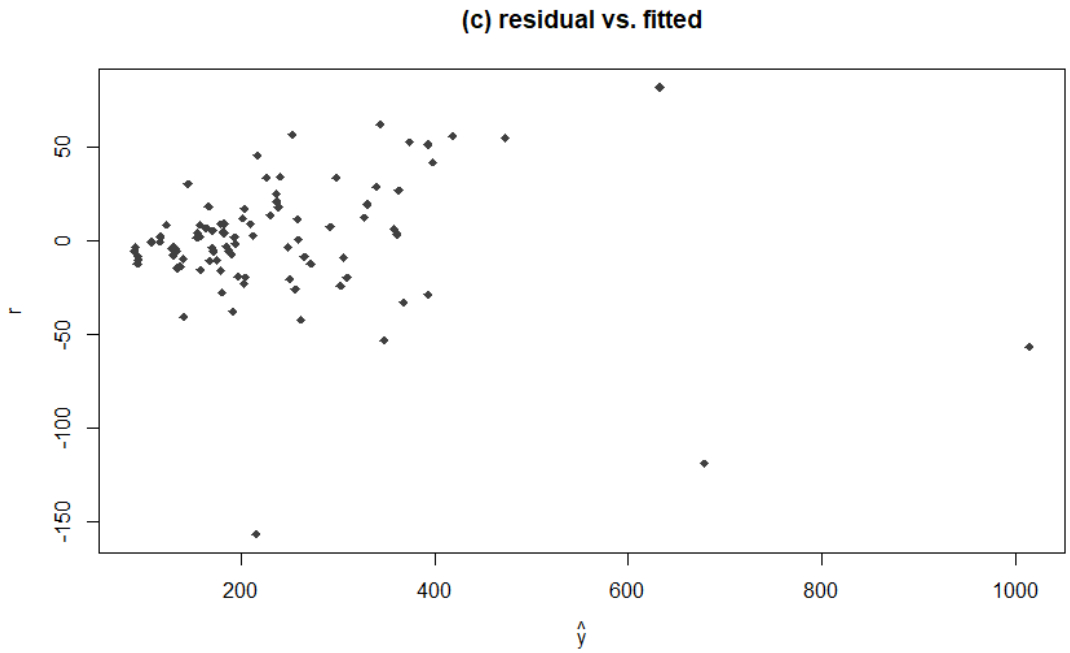
	fitted	residual
1	203.08689	-23.0868923
2	248.39549	-3.2954934
3	93.75418	-8.3541760
4	90.94877	-3.0487726
5	94.31526	-10.1152567
6	90.38769	-5.3876919
7	93.19310	-12.1930953
8	128.73462	-3.7346176
9	137.41267	-13.4126655
10	130.51084	-4.5108387
11	132.30630	-3.8062969
12	130.67970	-3.1796973
13	133.34510	-5.1450977
14	107.45951	-0.4595068
15	129.43356	-4.4335638
16	116.18565	-0.1856473
17	130.02243	-7.5224314
18	116.68688	2.2131206
19	134.35291	-14.3529055
20	178.96897	9.0310273
21	190.07837	-7.0783703
22	193.33691	2.1630869
23	194.49755	-1.4975486
24	182.82707	9.1729296
25	238.77590	18.1241013
26	258.21173	11.7882664
27	250.42714	-20.4271398
28	298.64802	33.8519831
29	252.98887	57.0111261
30	256.10528	-25.6052764
31	265.47265	-8.4726520
32	292.26559	7.7344095
33	240.59594	34.4040625
34	361.33943	3.6605681
35	236.96013	21.0398653
36	302.92185	-23.9218486
37	327.39245	12.6075525
38	367.82659	-32.8265933
39	305.80206	-8.8020628
40	171.35751	-5.3575125
41	182.64539	4.3546129
42	179.23615	-15.8361540
43	215.74060	-156.7405976
44	203.54751	17.4524872
45	309.54046	-19.5404632
46	272.18531	-12.1853144
47	393.47172	51.5282789
48	343.83906	62.1609418
49	166.62520	18.3748023
50	145.39070	30.6093015

Edna Diaz
80396667
Stat 4380
Project I

51	164.05705	7.0429487
52	184.72406	-2.7240575
53	157.66394	8.8360624
54	348.15564	-53.1556389
55	340.14875	28.8512496
56	330.32610	19.6739021
57	358.29891	6.7010911
58	362.95214	27.0478619
59	393.47172	-28.4717211
60	398.18587	42.1141324
61	140.89778	-40.5977781
62	262.06021	-42.0602127
63	181.62262	5.3773828
64	201.85786	12.1421416
65	204.58311	-19.5831075
66	187.90245	-5.4024459
67	175.24874	-10.2487406
68	170.49933	-3.4993262
69	157.42454	2.5754568
70	140.48525	-9.5852502
71	155.31381	4.6861889
72	158.09036	-15.2903590
73	678.83415	-118.8341505
74	632.81164	82.1883590
75	170.60620	5.3938013
76	197.12020	-19.1202000
77	154.73242	1.7675753
78	180.39786	-27.3978582
79	472.91968	55.0803229
80	418.91058	56.0894151
81	374.21329	52.7867054
82	1014.21078	-56.7107777
83	226.24937	33.7506283
84	236.70257	25.2974280
85	191.62268	-37.6226780
86	259.26870	0.7312973
87	212.30358	2.6964223
88	230.26457	13.7354281
89	209.94597	9.0540299
90	123.23816	8.7618356
91	167.47270	-10.5726967
92	217.27101	45.7289880

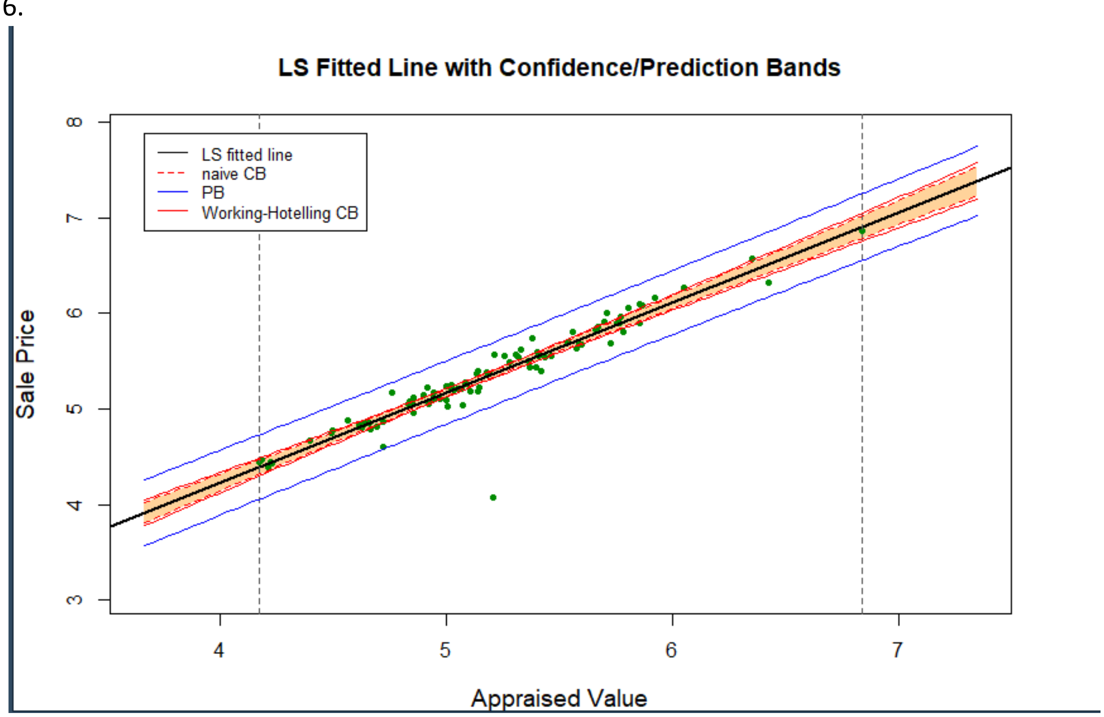
Edna Diaz
80396667
Stat 4380
Project I

Residual vs. Fitted



Based on this image, one can assume that because these points are do not have a general pattern, that the relationship is linear. One can also see that there are three points which could be outliers in this scenario.

6.



Edna Diaz
80396667
Stat 4380
Project I

-What this graph shows, is that within the PB, prediction interval, 95% of the Sale Price values will be found for a certain appraised value within the interval range that is around the linear regression line. And as we can see, most if not all points are within this range. There is one point that is a clear outlier when looking at PB.
-For CB, there is a 95% probability that the best-fit line for the population will be within the confidence interval, there are a few outliers, outside of the CB interval.
-Since working hoteling band contains all mean responses, there are a few possible mean outliers within this band.

Appendix

#1

```
##read data, and create plot
tampalms <- read.table("tampalms.dat", header=F,
                col.names=c("appraised", "sale"))
x <- log(tampalms$appraised)
y <- log(tampalms$sale)
plot(x,y, main="Tampalms House Price Data",xlab="Appraised Value", ylab="Sale Price")
```

#2

```
##compute correlation coefficient 95%
cor.test(Appraised Value, Sale Price)
cor.test(tampalms$appraised,tampalms$sale, method="pearson")
```

#3

```
##Linear Regression Model
fit <- lm(tampalms$sale ~ tampalms$appraised)
fit
attributes(fit)
var(fit$residuals)
summary(fit)
```

#4

```
plot(x, y, main = "Tampalms House Price Data",
   xlab = "Appraised Value", ylab = "Sale Price",
   pch = 19, frame = FALSE)
abline(lm(y ~ x, data = tampalms), col = "blue")

aov.out = aov(x ~ y, data = tampalms)
summary(aov.out)
```

Edna Diaz
80396667
Stat 4380
Project I

#5

```
y.hat <- fitted(fit)
r <- resid(fit)
dat.sheet <- data.frame( fitted=y.hat, residual=r)
dat.sheet
```

#6

```
plot(y.hat, r, pch=18, col="grey25", main="(c) residual vs. fitted",
    xlab=expression(hat(y)))

predict(fit, newdata=data.frame(x), se.fit=TRUE,interval="confidence", level=0.95)

plot.CB <- function(x, y, prediction.band=TRUE, working.hotelling=TRUE,
                confidence.level=0.95, xlab="x", ylab="y", legend=TRUE){
  # COULD HAVE ADDED SOME ERROR CHECKING STEPS
  fit <- lm(y~x)
  x0 <- min(x)-sd(x); x1 <- max(x) + sd(x);
  y0 <- min(y)-2*sd(y); y1 <- max(y) + 2*sd(y)
  new <- data.frame(x= seq(x0, x1, length=100))
  CI95 <- predict(fit, newdata=new, se.fit=TRUE,interval="confidence", level=confidence.level);

  par(mar=rep(4,4), mfrow=c(1, 1))
  plot(c(x0, x1), c(y0, y1), type="n", ylab=ylab, xlab=xlab,
     main="LS Fitted Line with Confidence/Prediction Bands", cex.lab=1.2)
  polygon(c(new$x, rev(new$x)), c(CI95$fit[,2], rev(CI95$fit[,3])),
        col = "burlywood1", border = NA)
  points(x, y, pch=20, col="green4")
  abline(lsfit(x,y), lwd=2)
  abline(v=min(x), col="gray35", lty=2)
  abline(v=max(x), col="gray35", lty=2)
  lines(new$x, CI95$fit[,2], lty=2, col="red", lwd=1.5)
  lines(new$x, CI95$fit[,3], lty=2, col="red", lwd=1.5)

  # PREDICTION BAND
  if (prediction.band) {
    PI95 <- predict(fit, newdata=new, se.fit=TRUE,interval="prediction",
                level=confidence.level)
    lines(new$x, PI95$fit[,2],lty=1, col="blue", lwd=1.5)
    lines(new$x, PI95$fit[,3],lty=1, col="blue", lwd=1.5)
  }

  # WORKING-HOTELLING JOINT CONFIDENCE BAND
  if (working.hotelling) {
    n <- length(x)
    W.Hoteling <-  sqrt(2 * qf(confidence.level, 2, n-2))
    LB <- CI95$fit[, 1] - W.Hoteling*CI95$se.fit
    UB <- CI95$fit[, 1] + W.Hoteling*CI95$se.fit
    lines(new$x, LB,lty=1, col="red", lwd=1.2)
    lines(new$x, UB,lty=1, col="red", lwd=1.2)
  }
```

Edna Diaz
80396667
Stat 4380
Project I

```
  if (prediction.band && working.hotelling && legend){
    legend(x0, y1, c("LS fitted line", "naive CB", "PB", "Working-Hotelling CB"),
           lty=c(1, 2, 1, 1), col=c("black", "red", "blue", "red"), lwd=1, cex=0.8)
  }
}


plot.CB(x, y, prediction.band=TRUE, working.hotelling=TRUE,
        confidence.level=0.95, ylab="Sale Price", xlab="Appraised Value")
```