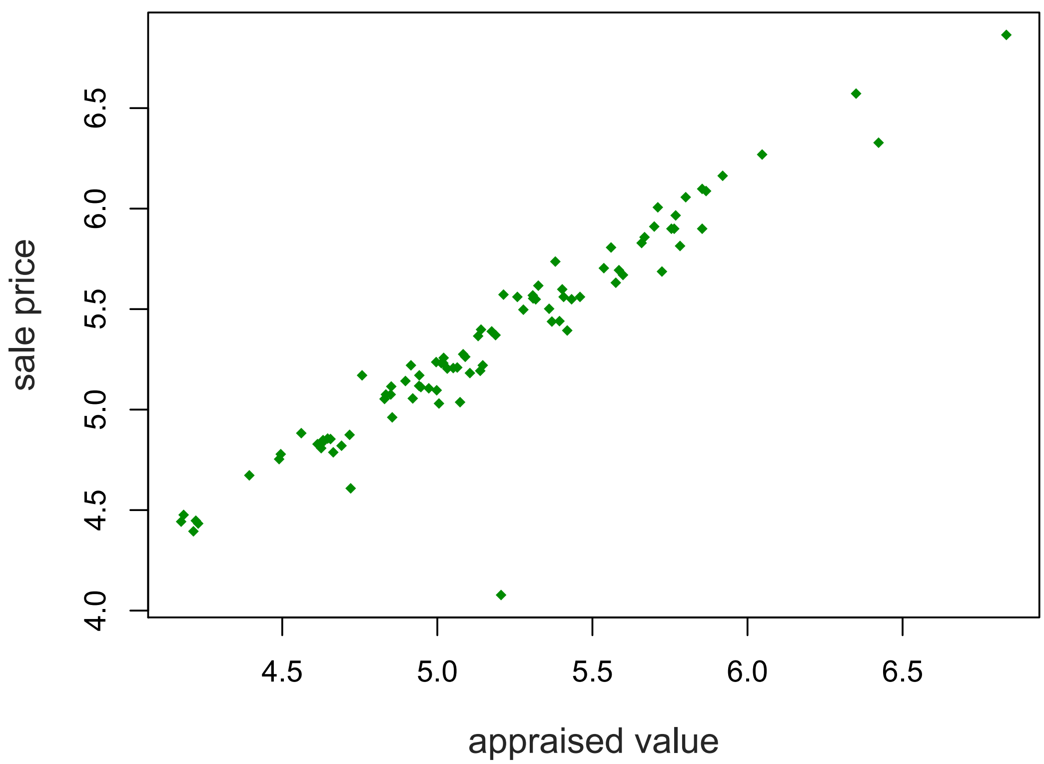
Stat4385

R Project 1 -- Final Report

1. Make a scatterplot of the data. Does it appear that a straight-line model will be an appropriate fit to the data?

R Output -- Scatterplot:



Conclusion: From the Scatterplot, we see a possible positive linear association between sale price and appraised value, thus we conclude that a straight-line-model will be a good fit for the data.

2. Compute the Pearson correlation r, together with a 95% confidence interval for ρ, and interpret.

R-Output:

Pearson's product-moment correlation

data: x and y

t = 27.819, df = 90, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval: (0.9200104   0.9643515)

Cov (x,y) = 0.9464788

Interpretation: this interpret a strong positive linear association between the appraised value and the sale price of a property.

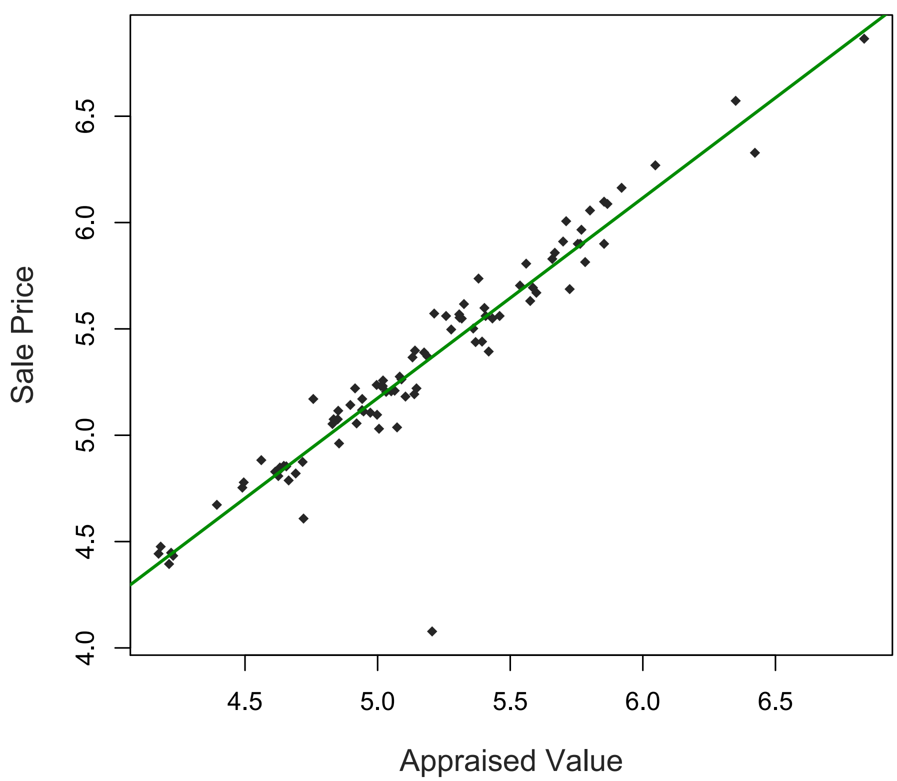
3. Linear regression model is used to relate the appraised property value X to the sale price Y for residential properties in this neighborhood. Compute the LS estimates for the regression parameters and give an unbiased estimate for the constant variance σ^2. Provide the Table of Parameter Estimates and then add the fitted LS line to the scatterplot.

R Output – Table of Parameter Estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.46723	0.17570	2.659	0.00927
x	0.94145	0.03384	27.819	< 2e-16

Where β_0 hat =0.46723; β_1 hat = 0.94145; σ^2 hat =0.0274.

LS Fitted Line:



4. Obtain the ANOVA table. What is the R^2 value of the fitted model?

R Output -- ANOVA Table:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	21.1767	21.1767	773.91	< 2.2e-16 ***
Residuals	90	2.4627	0.0274		
Total	91	23.6394			

R^2 = SSR/SSTO =21.1767 / (21.1767+2.4627) = 89.58%

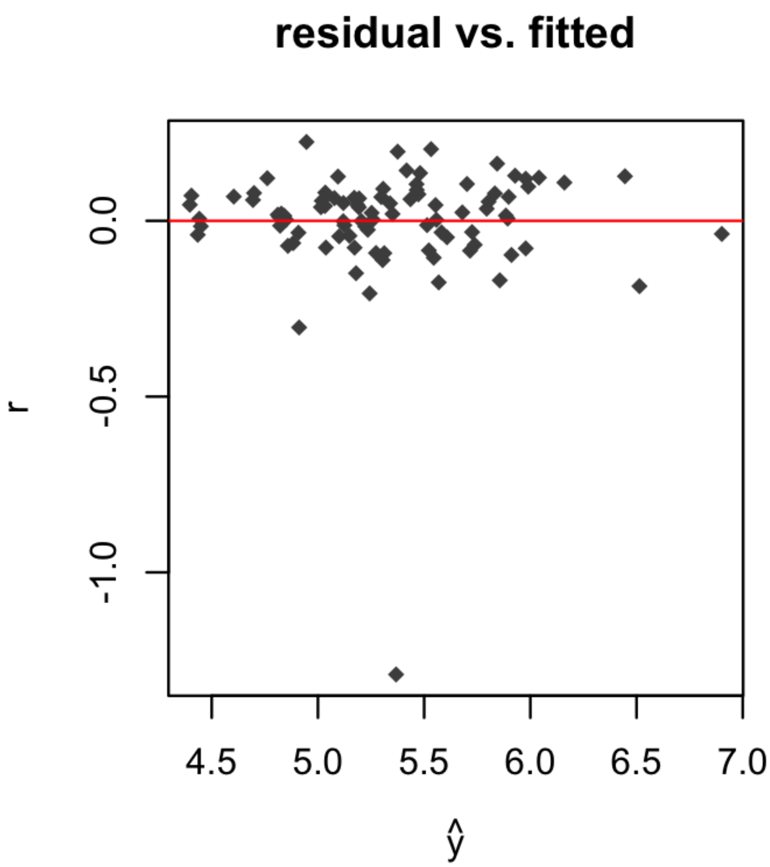
This represents 89.58% of sample variation in sale price of a property that can be explained by using appraised value to predict the sale price of a property in the Simple Linear Regression model.

5. Obtain the fitted values $\hat{y}_i$ and residuals r_i from the fitted model. Plot r_i versus $\hat{y}_i$ and comment.

R Output – fitted Values and residuals

ID	Appraised	Sale	Fitted	Residual
1	5.138336	5.192957	5.304713	-0.1117564479
2	5.360480	5.501666	5.513850	-0.0121835456
3	4.221418	4.447346	4.441481	0.0058650676
4	4.182126	4.476200	4.404490	0.0717094047
…				
91	4.920769	5.055609	5.099885	-0.0442759703
92	5.213326	5.572154	5.375312	0.1968420575

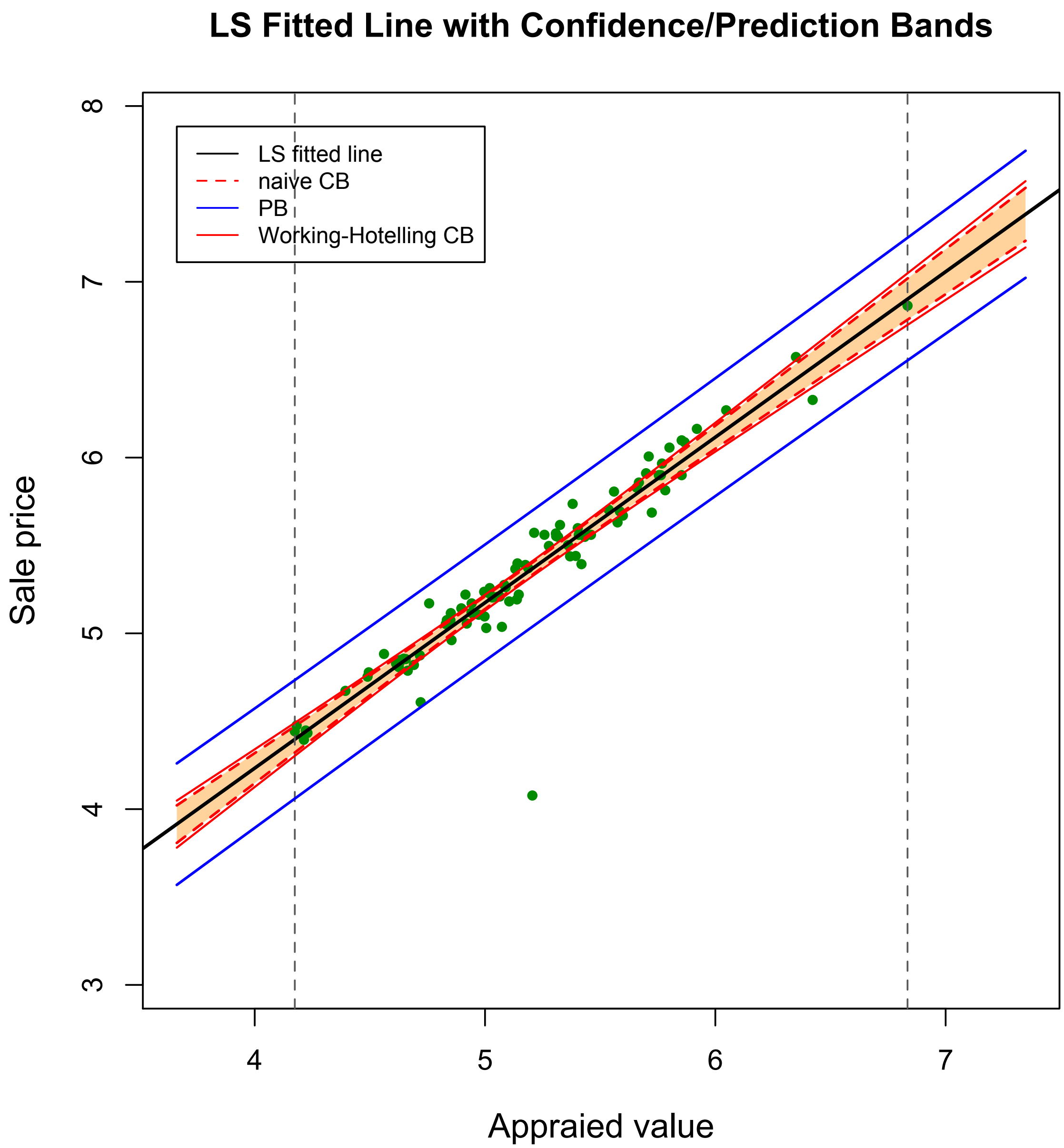
Plot- Residual VS Fitted value



Comment: From the plot we see that most of the fitted values are randomly scattered around 0 residual, this indicates a good fit of the SLR model

6. Plot the naïve 95% confidence band (as well as the 95% Working-Hotelling one) and 95% prediction band and comment on the model fit and potential outliers.

R Output – LS Fitted Line with Confidence/ Prediction Bands



Comment: From the plot, we see most of the data are in the range of the prediction band except one outliner, we can future conclude that we are 95% confident that the true best-fitted linear regression is enclosed in the confidence band and that the SLR model is a good fit to the data.

Appendix –- R Code

```
setwd("/Users/Meng/Documents/school/spring 2017/regression analysis/R
 PROJECT/PRO 1")
# read in data
tampalms <- read.table("tampalms.dat", header=F,
col.names=c("appraised", "sale"))
x <- log(tampalms$appraised)
 y <- log(tampalms$sale)

 #scatterplot (log)
 #.........
 plot(x, y, xlab="appraised value", ylab="sale
 price",col="green4",pch=18, cex=.8, cex.lab=1.2, col.lab="gray15")

 #Correlation
 #...................
 cor(x, y)
cor.test(x, y, alternative = "two.sided",
method = "pearson", conf.level=.95) # ONLY FOR rho0=0, BUT HAVE
 CONFIDENCE INTERVAL


 #SLR
 #......................
 fit <- lm(y~x); summary(fit); anova(fit)

 par(mfrow=c(1,1), mar=c(7, 5, 7, 5))
 plot(x, y, xlab="Appraised Value", ylab="Sale Price", col="gray15",
pch=18, cex=0.9, cex.lab=1.2, col.lab="gray15")

abline(lsfit(x,y), col="green4", lwd=2)
2017-03-25 16:02:08.090 R[22407:870631] kCFURLVolumeIsAutomountedKey
 missing for
 file:///private/var/folders/zz/zyxvpxvq6csfxvn_n0000000000000/T/FPInst
 allMountPoint/: The file “FPInstallMountPoint” couldn’t be opened
 because you don’t have permission to view it.

 #Model Diagnostics
#...................

 y.hat <- fitted(fit)
 r <- resid(fit)
dat.sheet <- data.frame(ID=1:92, appraised=x, sale=y, fitted=y.hat,
```

```r
residual=r)
 dat.sheet
 write.csv(dat.sheet, file="residual.csv", row.names =F)

 # Diagnostic plots
 #....................
 par(mfrow=c(2, 2), mar=rep(4, 6, 4, 6))
 plot(y.hat, r, pch=18, col="grey25", main="residual vs. fitted",
    xlab=expression(hat(y)))
 abline(h=0, col="red")


 #confidence and prediction band
 #..............................
 # AT ONE SINGLE POINT OR SEVERAL
predict(fit, newdata=data.frame(x=20),
 se.fit=TRUE,interval="confidence", level=0.95);
predict(fit, newdata=data.frame(x=20),
 se.fit=TRUE,interval="prediction", level=0.95);

# AT SEVERAL POINTS
predict(fit, newdata=data.frame(x=c(10, 15, 20, 25)),
 se.fit=TRUE,interval="confidence", level=0.95);

# function plot.CB ()
#.....................
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)
}
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")
```