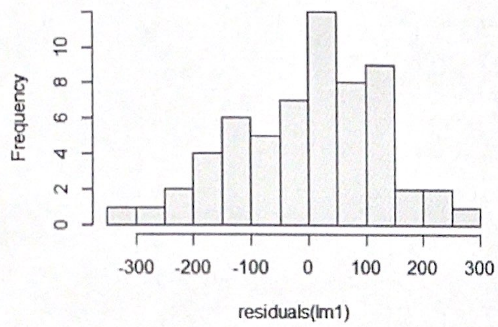
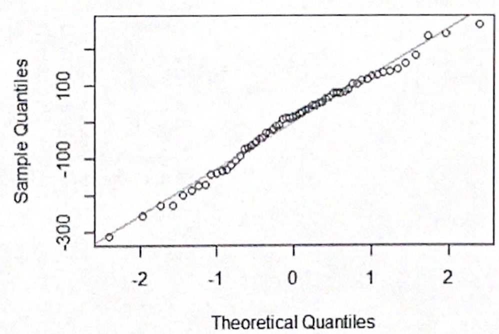
You may also wish to check that residuals are 'nearly Normal', depending on modeling applications:

```
# make histogram of residuals to check Normality
# also identify outliers among residuals
hist(residuals(lm1), breaks =15, col = 'oldlace')
```




```
# or make a Q-Q plot to assess Normality
qqnorm(residuals(lm1))
abline(mean(residuals(lm1)), sd(residuals(lm1)), col = 'tomato')
```

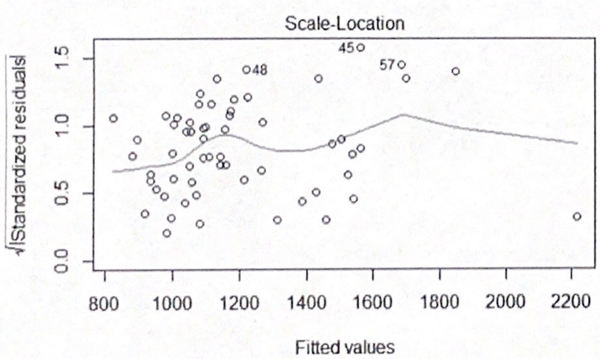


Q-Q plots :

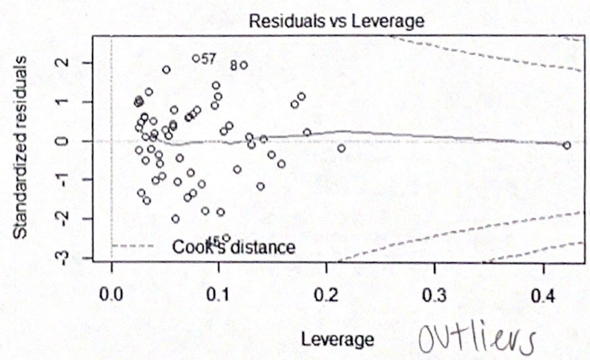
"How similar are the quantiles in my data compared to what the quantiles would be if my data followed a theoretical probability distribution?"

```
## R shortcut for quick assumptions check! Use plot() function
## with name of the model >> get four key model diagnostic plots
plot(lm1)
```

In addition to (1) Residuals Plot and (2) QQ-plot, the `plot()` *shortcut produces a (3) Scale-location plot and (4) Leverage plot*



This plot shows if residuals are spread equally along the ranges of predictors; use it to check for equal variance assumption. A horizontal line with equally spread points is indicative of homoscedasticity. The three observations with largest |residuals| are labeled by default.



outliers

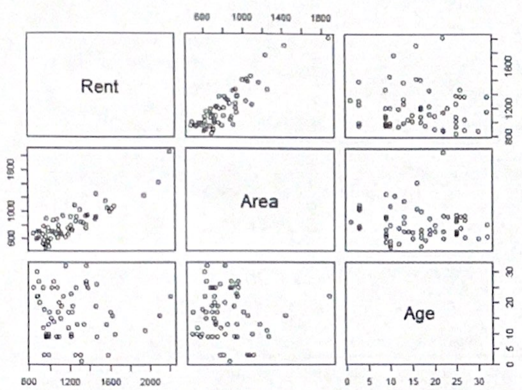
This plot helps us find influential observations (if any) – i.e., data points outside of the dashed line representing Cook's distance. Note that high-leverage or influential points are not necessarily 'problems'. Same with outliers, although all such points should be investigated in context.

Austin Apartment Rents — Checking regression assumptions with base R functions

Before model fitting, use scatterplots to visualize relationships between Y response and X predictors; check for linearity and outliers.

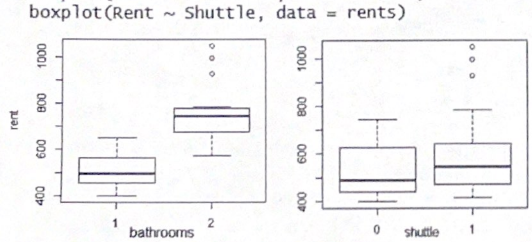
Use a **scatterplot matrix** for multiple plots:

```
plot(rents[c(1,2,9)])
```



Side-by-side boxplots may be preferable for visualizing Y at different levels of a categorical predictor (and useful for identifying outliers)

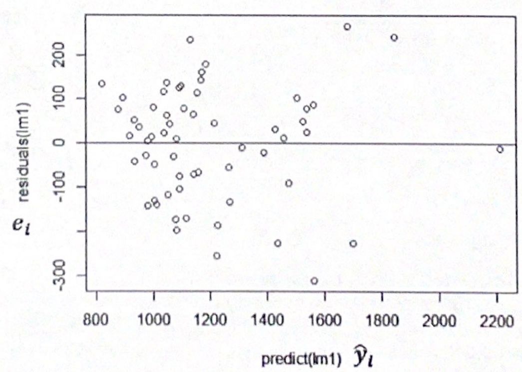
```
boxplot(Rent ~ Bathrooms, data = rents)
boxplot(Rent ~ Shuttle, data = rents)
```




```
# fitting multiple regression model
lm1 <- lm(Rent ~ Area + Age + Bathrooms + Shuttle, data = rents)
summary(lm1)
             Estimate Std. Error t value        Pr(>|t|)
(Intercept) 197.4565    88.3532   2.235        0.02951 *
Area          0.8087     0.1036   7.809 0.000000000179 ***
Age           1.5816     2.2102   0.716        0.47726
Bathrooms   193.5512    57.7597   3.351        0.00146 **
Shuttle      88.4124    50.8820   1.738        0.08788 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 132.6 on 55 degrees of freedom
Multiple R-squared:  0.8118,    Adjusted R-squared:  0.7981
F-statistic: 59.31 on 4 and 55 DF,  p-value: < 0.000000000000000022
```

After model fitting, check assumptions with a series of plots based on model residuals and fitted values.

```
# adding FITTED VALUES to data frame with mutate() and predict()
rents = rents %>%
   mutate(fitted = predict(model))

# adding RESIDUALS to data frame with mutate() and residuals()
rents = rents %>%
   mutate(residuals = residuals(model))

# check linearity (and get a sense of equal spread) by plotting
# residuals against fitted Y values
plot(predict(lm1), residuals(lm1))
abline(0, 0, col = 'blue')          # adds reference line at 0
```



Relatively equal spread above and below zero reference line as $\hat{y}$ increases is evidence of homoscedasticity.

If the residual plot shows a pattern, make separate plots for each x_j vs residuals e_i to identify which x_j variable(s) is/are the source of a violation of regression assumptions.