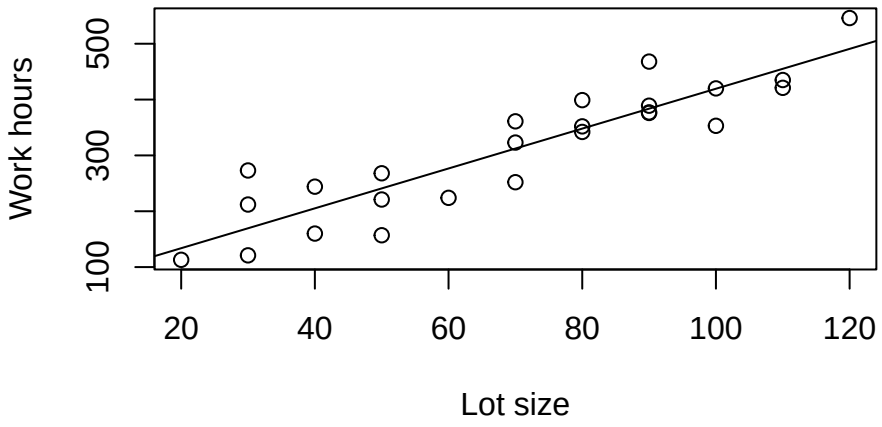
Rexample5

Lingxiao Zhou

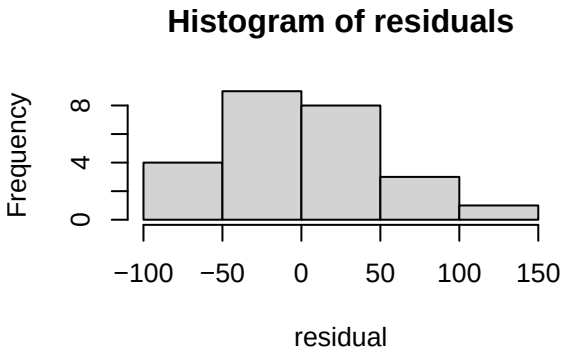
```r
# find relationship between lot size (X) and work hours (Y)
toluca <- read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData
    col.names = c("lotsize", "workhrs"))
plot(toluca$lotsize, toluca$workhrs, xlab = "Lot size", ylab = "Work hours")
toluca.reg <- lm(workhrs ~ lotsize, data = toluca)
summary(toluca.reg)
```

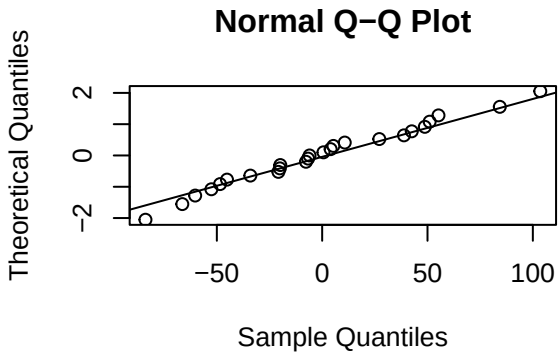
```
##
## Call:
## lm(formula = workhrs ~ lotsize, data = toluca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## lotsize        3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

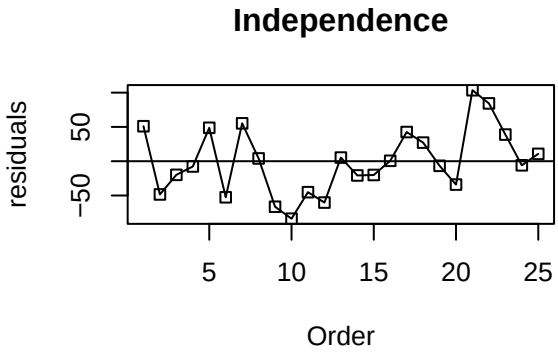
```r
abline(toluca.reg)
```

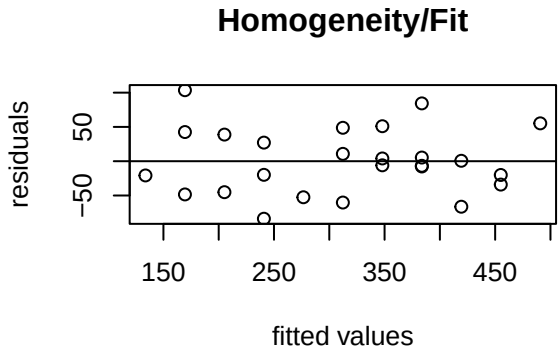



```
e <- toluca.reg$residuals  # get the raw residuals
```









Normality

Shapiro test

- H_0 : The residuals are drawn from normal distribution
- Since P-value > 0.05, we fail to reject the normal assumption at the significance level 0.05

```
shapiro.test(e)
```

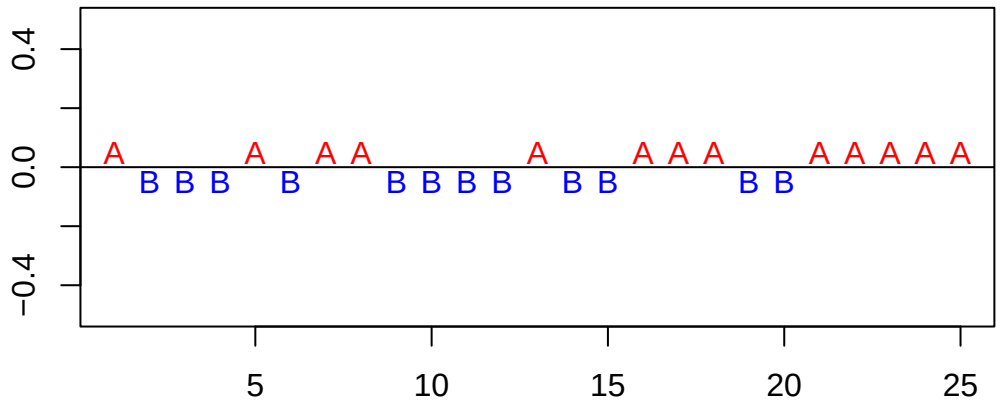
```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.9789, p-value = 0.8626
```

Independence

runs test

- H_0 : Independence
- We fail to reject the null hypothesis since p-value is large.
- According to the plot, there are 11 runs out of maximum 25

```
par(mfrow = c(1, 1))
library(lawstat)  # may need to install package
runs.test(e, plot.it = TRUE)
```




```
##
```

```
##  Runs Test - Two sided
##
## data:  e
## Standardized Runs Statistic = -1.015, p-value = 0.3101
```

Durbin-Watson

- Assume $\epsilon_i = \rho\epsilon_{i-1} + u_i$ where $u_i \overset{iid}{\sim} N(0, \sigma^2)$.
- $H_0 : \rho = 0$
- For two-sided test, $H_a : \rho \neq 0$
- Notice that `durbinWatsonTest()` and `dwtest()` use different method to calculate p-value, so you may observe that they did not give consistent results.
- Here we fail to reject the H_0.

```
set.seed(1234)

library(car)  # for durbinWatsonTest
durbinWatsonTest(toluca.reg, alternative = "two.sided")
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2593193       1.43179   0.138
##  Alternative hypothesis: rho != 0
```

```
library(lmtest)  # for dwtest
dwtest(toluca.reg, alternative = "two.sided")
```

```
##
##  Durbin-Watson test
##
## data:  toluca.reg
## DW = 1.4318, p-value = 0.1616
## alternative hypothesis: true autocorrelation is not 0
```

Homogeneity of Variance

Levene's test

- Need categorical x
- $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$

- Split the data into two groups depending on whether lot size is greater than 75 or not (half-way point).
- With a p-value greater than 0.05 we fail to reject the null.

```
ind <- I(toluca$lotsize > 75)  # convert lotsize to categorical variable
temp <- cbind(toluca$lotsize, e, ind)
temp  # print X, residual and converted X
```

```
##                   e ind
## 1   80  51.0179798   1
```

```
## 2   30 -48.4719192   0
## 3   50 -19.8759596   0
## 4   90  -7.6840404   1
## 5   70  48.7200000   0
## 6   60 -52.5779798   0
## 7  120  55.2098990   1
## 8   80   4.0179798   1
## 9  100 -66.3860606   1
## 10  50 -83.8759596   0
## 11  40 -45.1739394   0
## 12  70 -60.2800000   0
## 13  90   5.3159596   1
## 14  20 -20.7698990   0
## 15 110 -20.0880808   1
## 16 100   0.6139394   1
## 17  30  42.5280808   0
## 18  50  27.1240404   0
## 19  90  -6.6840404   1
## 20 110 -34.0880808   1
## 21  30 103.5280808   0
## 22  90  84.3159596   1
## 23  40  38.8260606   0
## 24  80  -5.9820202   1
## 25  70  10.7200000   0
```

```
leveneTest(temp[, 2], ind)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  1.7331  0.201
##       23
```

Breusch-Pagan/Cook-Weisberg test

- H_0 : The residuals are distributed with equal variance
- Fail to reject the null since the p-value $= 0.36491 > 0.05$

```
ncvTest(toluca.reg)  # car library
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8209192, Df = 1, p = 0.36491
```

Linearity of regression

Lack of fit

- $H_0 : E(Y_i) = \beta_0 + \beta_1 X_i$
- $H_a : E(Y_i) \neq \beta_0 + \beta_1 X_i$

- P-value = 0.6893. We fail to reject the null hypothesis and there is no violation for the linearity assumption based on this test

```
# check t<<n If t is approximately equal to n, then the
# test is not applicable
(t <- length(unique(toluca$lotsize)))
```

```
## [1] 11
```

```
(n <- nrow(toluca))
```

```
## [1] 25
```

```
Reduced <- toluca.reg  # reduced model: SLR model
Full <- lm(workhrs ~ 0 + as.factor(lotsize), data = toluca)  # full model: use group mean
summary(Full)
```

```
##
## Call:
## lm(formula = workhrs ~ 0 + as.factor(lotsize), data = toluca)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -81.0  -25.5    0.0   33.5   71.0
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## as.factor(lotsize)20    113.00      51.81   2.181 0.046732 *
## as.factor(lotsize)30    202.00      29.91   6.753 9.28e-06 ***
## as.factor(lotsize)40    202.00      36.64   5.514 7.63e-05 ***
## as.factor(lotsize)50    215.33      29.91   7.199 4.57e-06 ***
## as.factor(lotsize)60    224.00      51.81   4.323 0.000701 ***
## as.factor(lotsize)70    312.00      29.91  10.430 5.53e-08 ***
## as.factor(lotsize)80    364.33      29.91  12.180 7.73e-09 ***
## as.factor(lotsize)90    402.50      25.91  15.537 3.19e-10 ***
## as.factor(lotsize)100   386.50      36.64  10.550 4.79e-08 ***
## as.factor(lotsize)110   428.00      36.64  11.683 1.32e-08 ***
## as.factor(lotsize)120   546.00      51.81  10.538 4.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.81 on 14 degrees of freedom
## Multiple R-squared:  0.9863, Adjusted R-squared:  0.9756
## F-statistic:  91.7 on 11 and 14 DF,  p-value: 4.428e-11
```

```
anova(Reduced, Full)  # get lack-of-fit test
```

```
## Analysis of Variance Table
##
## Model 1: workhrs ~ lotsize
## Model 2: workhrs ~ 0 + as.factor(lotsize)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     23 54825
## 2     14 37581  9     17245 0.7138 0.6893
```