Math 342 Workshop 2 Winter 2022

Jonathan Lee, Phelix Tang, Jeffery Smith

1/25/2022

R Markdown

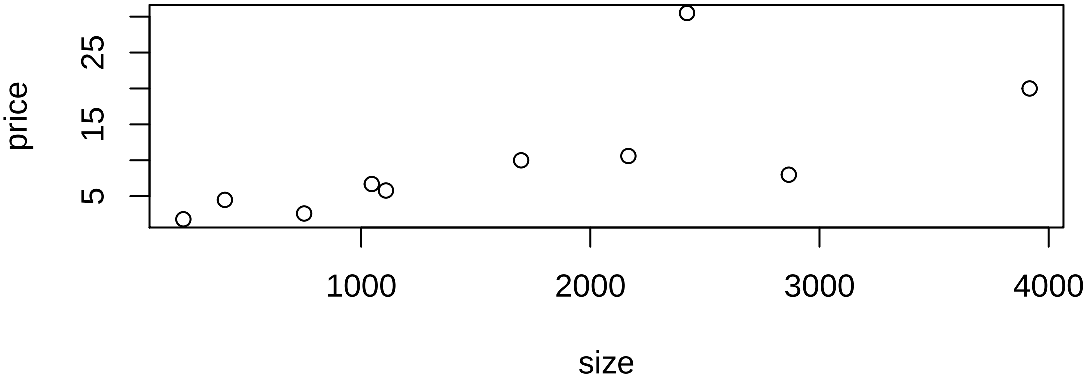
The following data on sale price, size, and land-to-building ratio for 10 large industrial properties appeared in the paper "Using Multiple Regression Analysis in Real Estate Appraisal" (*Appraisal Journal* [2002]:424-430):

Sale Price (in millions)	10.6	2.6	30.5	1.8	20.0	8.0	10.0	6.7	5.8	4.5
Size (thousand sq. ft.)	2166	751	2422	224	3917	2866	1698	1046	1108	405

```
price = c(10.6, 2.6, 30.5, 1.8, 20.0, 8.0, 10.0, 6.7, 5.8, 4.5)
size = c(2166, 751, 2422, 224, 3917, 2866, 1698, 1046, 1108, 405)
```

(a) Make a scatterplot of the data using size as x and price as y.

```
plot(size, price)
```



## (b) Estimate the correlation coefficient, r.

```
cor(size, price)
```

```
## [1] 0.7001976
```

(c) Calculate the LS estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.

```
sum_x = sum(size)
sum_y = sum(price)
sum_x2 = sum(size^2)
sum_y2 = sum(price^2)
sum_xy = sum(size*price)
n = length(size)
```

```
Sxy = sum_xy - (1/n)*sum_x*sum_y
Sxx = sum_x2 - (1/n)*sum_x^2
beta1 = Sxy/Sxx
ybar = sum_y/n
xbar = sum_x/n
beta0 = ybar - beta1*xbar
```

(d) What proportion of variation in sale price is accounted for by the regression on size?

```
SST = sum_y2 - (1/n)*sum_y^2
yhat = beta0 + beta1*size
ei = price - yhat
SSE = sum(ei^2)
r2 = 1 - SSE/SST
r2
```

```
## [1] 0.4902767
```

(e) Use F-test to determine if size is a significant predictor of sale price. Show complete steps in hypothesis testing.

h_0 h_1

```
fit = lm(price~size)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: price
##           Df Sum Sq Mean Sq F value  Pr(>F)
## size       1 345.82  345.82  7.6948 0.02415 *
## Residuals  8 359.54   44.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(f) Construct a 95% prediction interval for the price when size = 1000.

```
new = data.frame(size=c(1000))
predict(fit, new)
```

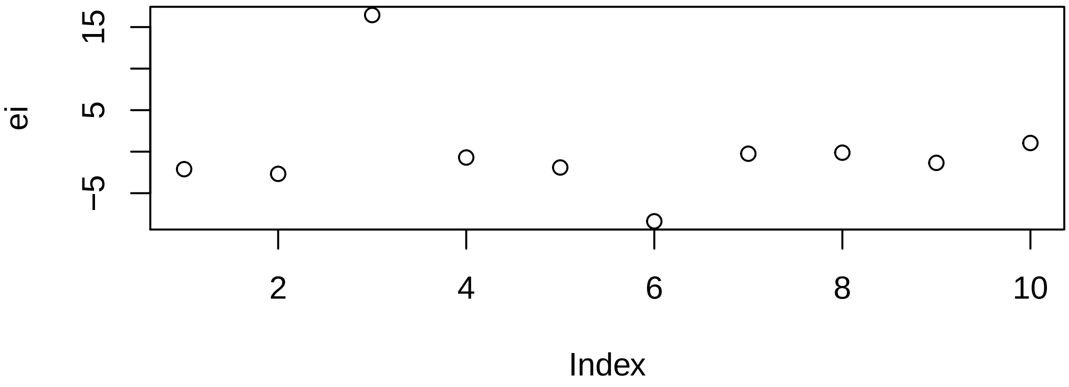
```
##        1
## 6.581324
```

```
predict(fit, new, interval = "confidence")
```

```
##        fit       lwr      upr
## 1 6.581324 0.9056113 12.25704
```

(g) Make a scatterplot of the residuals. Based on the plot, is the assumption of linearity and constant variance satisfied? Justify.

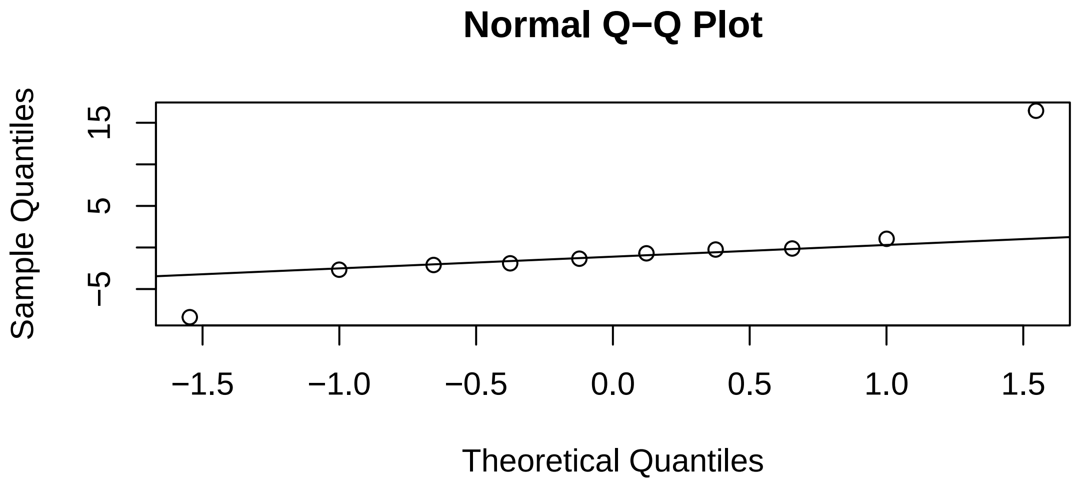
```r
plot(ei)
```



### The residuals seem to be scattered around 0 and no obvious pattern is visible, hence the assumption is satisfied.

(h) Obtain a normal probability plot of the residuals. Based on the plot, is the normality assumption satisfied?

```r
qqnorm(ei)
qqline(ei)
```



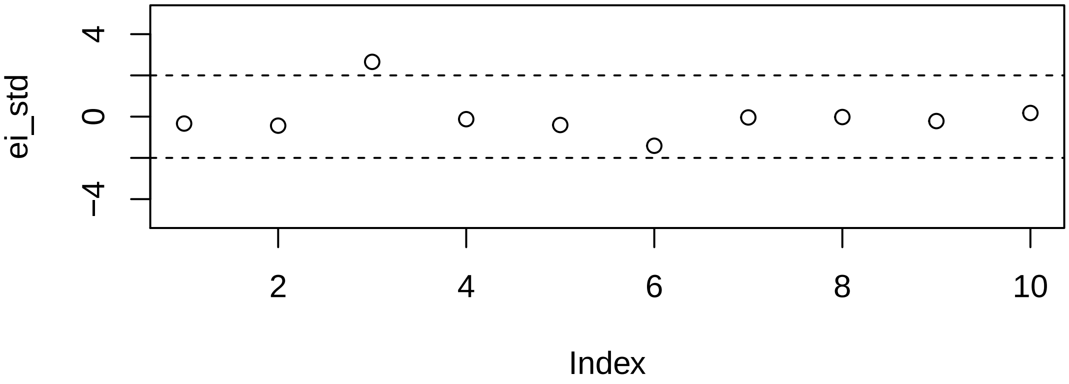
```r
shapiro.test(ei)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.72311, p-value = 0.00168
```

Based on the plot: yes, the assumption of normality assumption is satisfied disregarding the outlier at Theoretical Quantiles = 1.5.

(i) Make a graph of the standardized residuals, including a boundary for identifying outliers. Are there any possible outliers based on the plot?

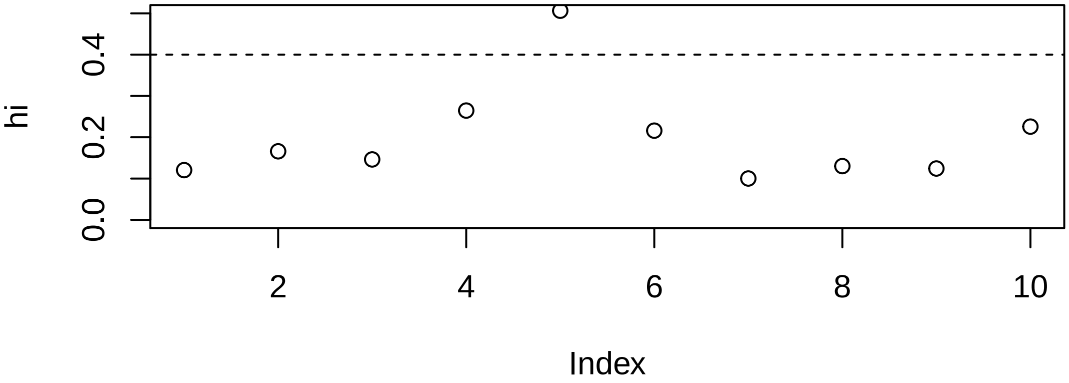
```
ei_std = rstandard(fit)
plot(ei_std, ylim = c(-5,5))
abline(h=2, lty=2)
abline(h=-2, lty=2)
```



### Yes, there seems to be an outlier at around Index=3

(j) Make a plot of the leverage values, including the boundary for identifying influential observations. Are there any influential observations in the data?

```
hi = hatvalues(fit)
b=4/10
plot(hi, ylim=c(0,0.5))
abline(h=b, lty=2)
```



## (k) Summarize the results of the analysis.

Based on the leverage plot, there is a clear influential point at index=5 that would influence the regression line.