Jessie Wong
STA4155 - EMWA
Professor Yue
September 28, 2022

R HW #2

1. Let's take Sales as the response variable and Price as a predictor variable. Fit a linear regression model to each of the four cities. Write down the four fitted models. In which city do the pizza sales seem to be more sensitive to price than in others? Explain.

Answer:
Chicago is more sensitive to price than the 3 other cities because, for every unit of increase in x, Sales volume decreased by $331,152. Compared to Denver, Dallas, and Baltimore, their Sales volumes range from -$33,527 to -$52,796, so their slopes are much smaller than Chicago Sales volume's slope. Therefore, Chicago is more sensitive to price than the other 3 cities.

R code:
```
pizza <- read.table('Frozen_Pizza.txt', sep = '\t', header = TRUE)
Chicago <-lm(Chicago.Volume ~ Chicago.Price, data=pizza)
Chicago
Denver <- lm(Denver.Volume ~ Denver.Price, data=pizza)
Denver
Dallas <- lm(Dallas.Volume ~ Dallas.Price, data=pizza)
Dallas
Baltimore <- lm(Baltimore.Volume ~ Baltimore.Price, data=pizza)
Baltimore
```

R output:
```
Call:
lm(formula = Chicago.Volume ~ Chicago.Price, data = pizza)

Coefficients:
  (Intercept)  Chicago.Price
      1094047        -331152
```

```
lm(formula = Denver.Volume ~ Denver.Price, data = pizza)

Coefficients:
 (Intercept)  Denver.Price
      181218        -52796
```

```
Call:
lm(formula = Dallas.Volume ~ Dallas.Price, data = pizza)

Coefficients:
 (Intercept)  Dallas.Price
      139547        -33527
```

```
Call:
lm(formula = Baltimore.Volume ~ Baltimore.Price, data = pizza)

Coefficients:
    (Intercept)  Baltimore.Price
         126625           -34956
```

2. For each of the models fitted above produce a residual plot in the time order, a residual plot against the fitted values, and a Q-Q plot. Is there any regression assumption violated in each model? Explain.

Answer:

Chicago

Independence Assumption: There seems to be a pattern in the scatterplot, which violates the Independence Assumption. The points are close to the y = 0 line, forming an almost straight line. Since the Independence Assumption requires that points are random in the graph, this assumption is violated.
Constant variance: Constant variance assumption isn't met because there's clumping in the lower range of x-values. Therefore, because the points aren't spread across the graph in a randomized fashion, this assumption is violated.
Q-Q plot: Q-Q plot has 2 outliers at the top right of the graph, and there's evidence of slight curves in the plot as well. Although this issue isn't severe, the regression analysis should proceed with caution.

Denver

Independence Assumption: The graph appears to have no pattern, but rather, it's random and points are spread across the graph without a pattern. Therefore, this assumption is met.
Constant Variance: The points look random, and the spread is relatively widespread, although there is evidence of some clumping. Therefore, this assumption is met.
Q-Q plot: The QQ plot shows curvature towards the top of the graph, although the majority of the points are relatively linear. Therefore, the regression analysis should be careful of this potential issue.

Dallas

Independence Assumption: The points are scattered and spread across the graph with no pattern, so this assumption is met.
Constant Variance: The graph shows no pattern and is spread across the graph, although there is some minor clumping in the bottom half of the plot. Therefore, this assumption is also met.
Q-Q plot: The plot has a curvature at the bottom and top of the linear line, although it's relatively linear in the middle part of the graph. Therefore, this assumption isn't met.

Baltimore

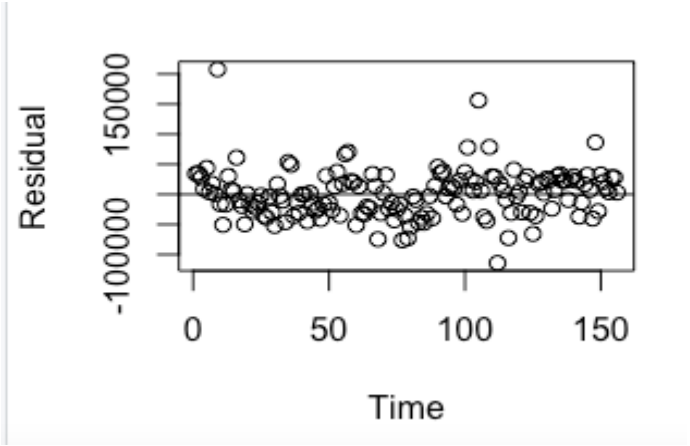
Independence Assumption: This assumption is violated because the points are accumulated close to the line y=0, so there is a linear pattern.
Constant Variance: This assumption is also violated because there is a weak negative correlation, even though for this assumption to be met, the points should be randomly spread across the graph.
Q-Q Plot: There are outliers in the graph that creates an upward curve, even though it should be linear. Therefore, this assumption is violated as well.

R code:

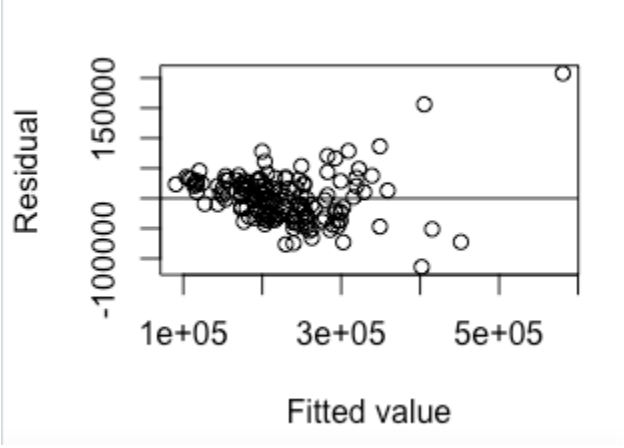
```
plot(Chicago$residuals, xlab = 'Time', ylab = 'Residual')
abline(a = 0, b = 0)
plot(Chicago$fitted.values, Chicago$residuals, xlab = 'Fitted value', ylab = 'Residual')
abline(a = 0, b = 0)
qqnorm(Chicago$residuals)
qqline(Chicago$residuals)

plot(Denver$residuals, xlab = 'Time', ylab = 'Residual')
abline(a = 0, b = 0)
plot(Denver$fitted.values, Denver$residuals, xlab = 'Fitted value', ylab = 'Residual')
abline(a = 0, b = 0)
qqnorm(Denver$residuals)
qqline(Denver$residuals)

plot(Dallas$residuals, xlab = 'Time', ylab = 'Residual')
abline(a = 0, b = 0)
plot(Dallas$fitted.values, Dallas$residuals, xlab = 'Fitted value', ylab = 'Residual')
abline(a = 0, b = 0)
qqnorm(Dallas$residuals)
qqline(Dallas$residuals)

plot(Baltimore$residuals, xlab = 'Time', ylab = 'Residual')
abline(a = 0, b = 0)
plot(Baltimore$fitted.values, Baltimore$residuals, xlab = 'Fitted value', ylab = 'Residual')
abline(a = 0, b = 0)
qqnorm(Baltimore$residuals)
```

qqline(Baltimore$residuals)

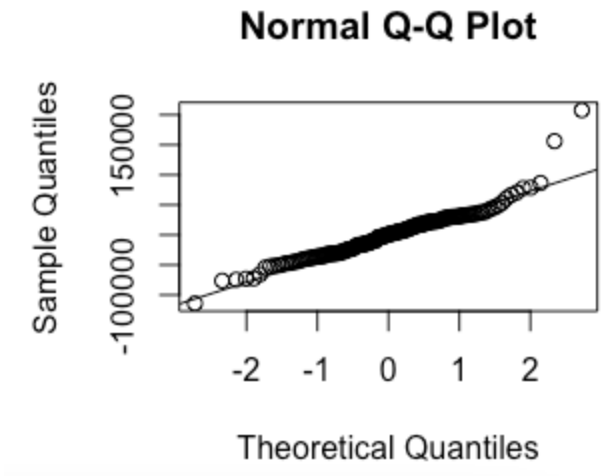
R output:
Chicago: Independence



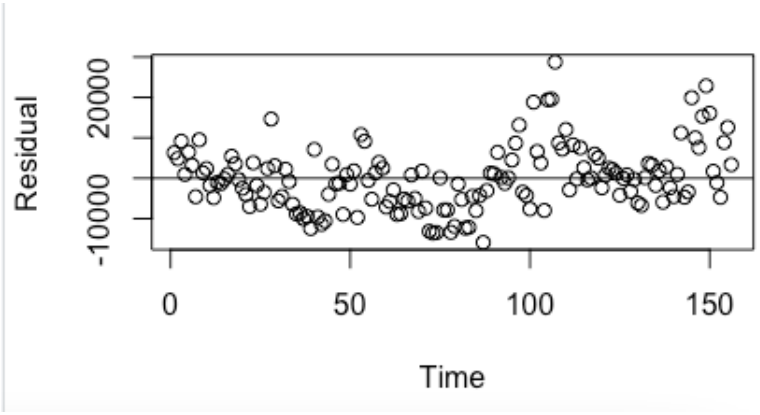
Chicago: Constant Variance



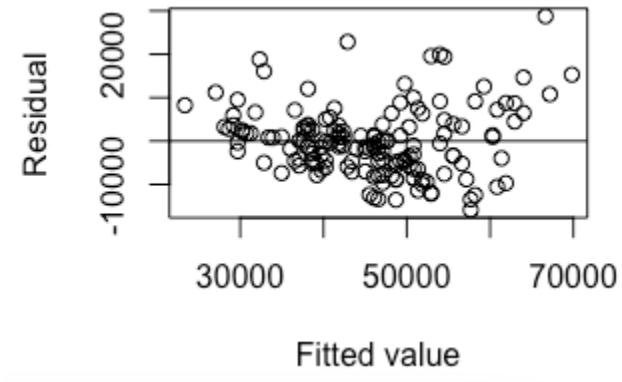
Chicago: Q-Q plot



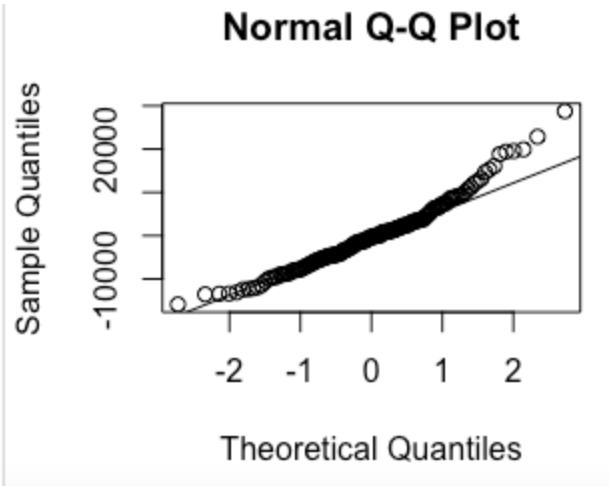
Denver: Independence Assumption



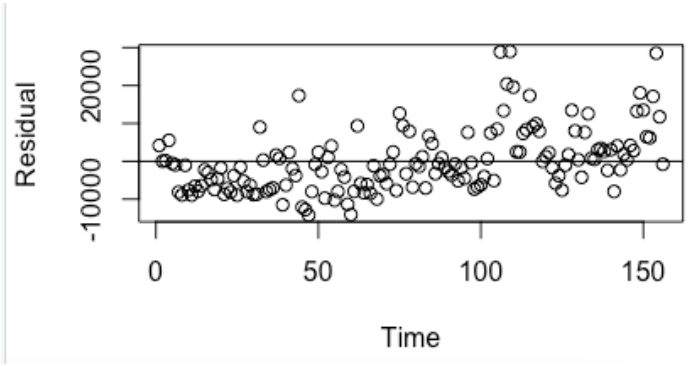
Denver: Constant Variance



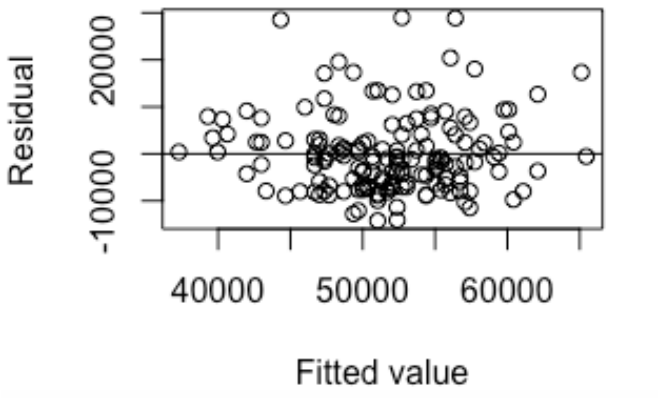
Denver: Q-Q plot:



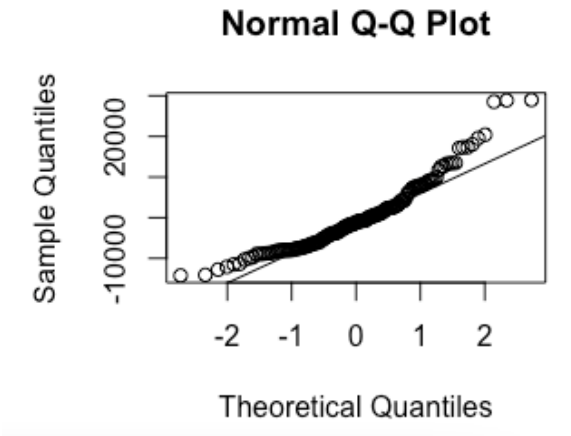
Dallas: Independence Assumption



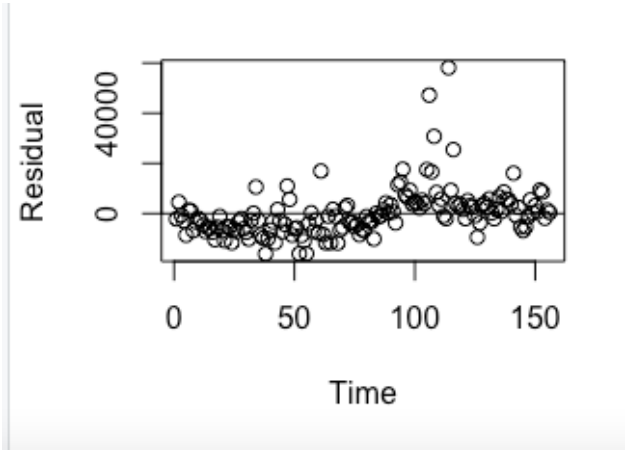
Dallas: Constant Variance



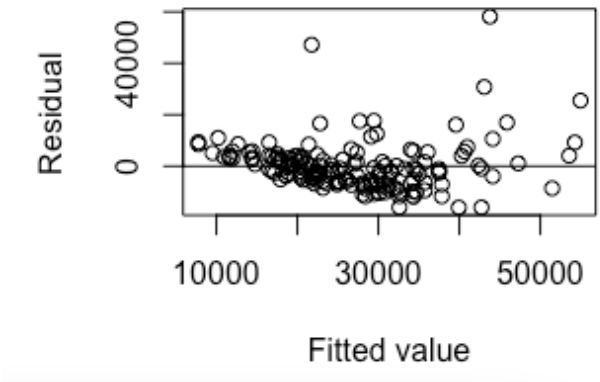
Dallas: Q-Q plot



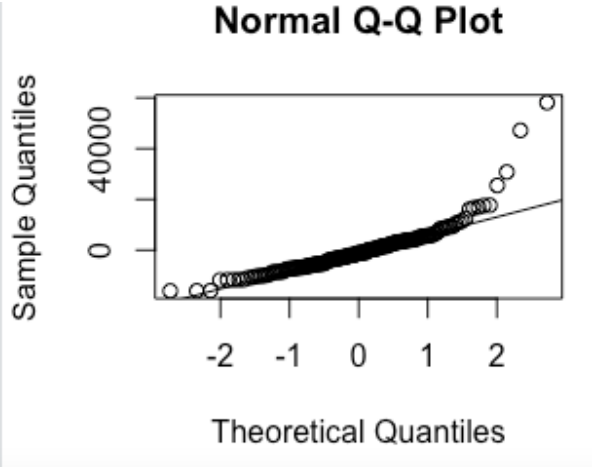
Baltimore: Independence Assumption



Baltimore: Constant Variance



Baltimore: Q-Q plot



3. For the remaining questions let's focus on the model for the city of Dallas. Show a 90% confidence interval for the slope of Price and interpret it. Based on the interval can we say there is a statistically significant linear relationship between Price and Sales volume? Explain.

Answer:
Yes, there's a statistically significant relationship between Price and Sales volume.
We are 95% confident that as the Price increases by 1 unit, Sales will increase by the amount between -$40655.79 and -$26398.58

R code:
confint(Dallas, level=0.90)

R output:

```
> confint(Dallas, level=0.90)
                    5 %      95 %
(Intercept)  120844.48 158250.38
Dallas.Price -40655.79 -26398.58
```

4. Conduct a hypothesis test to see if there is a significant negative correlation between Price and Sales volume in the city of Dallas, i.e., test H0 : β1 = 0 vs Ha : β1 < 0. State your test conclusion.

Answer:
Yes, there is a negative correlation between Price and Sales volume in Dallas, because the slope (b1) is -33527. The t-test statistics values are t=b0/SE(b0) = 12.347, and t = b1/SE(b1) = -7.783.The p-values are < 2e-16 and 9.623-13, which are very small values. Therefore, we can reject the null hypothesis and summarize that there's a negative correlation between Price and Sales volume in Dallas.

R code:
summary(Dallas)

R output:

```
Residuals:
   Min     1Q Median     3Q    Max
-14235  -6345  -1116   3553  28988

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    139547      11302  12.347  < 2e-16 ***
Dallas.Price   -33527       4308  -7.783 9.62e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8365 on 154 degrees of freedom
Multiple R-squared:  0.2823,    Adjusted R-squared:  0.2776
F-statistic: 60.57 on 1 and 154 DF,  p-value: 9.618e-13
```

5. For the city of Dallas, estimate the mean Sales if the Price is $2.50 and $3.00 using 95% confidence intervals. Interpret both intervals. Can we also estimate the mean Sales if the Price is $3.50? Explain.

Answer:
We are 95% confident that the mean sales if the Price is $2.50 are between $54063.14 and $57395.79. The predicted mean Sales is $55729.47. We are 95% confident that the mean sales if the Price is $3.00 are between $35464.31 and $42467.44. The predicted mean Sales is $38965.87. We can't predict sales if the Price is $3.50 because the range is only from $2.21 and $3.05.

R code:
range(pizza$Dallas.Price)
Dallas1 <- data.frame(Dallas.Price = c(2.5))
predict(Dallas, newdata = Dallas1, interval = 'confidence', level = 0.95)
Dallas2 <- data.frame(Dallas.Price = c(3))
predict(Dallas, newdata = Dallas2, interval = 'confidence', level = 0.95)

R output:

```
> range(pizza$Dallas.Price) #It's within the range.
[1] 2.21 3.05
```

```
predict(Dallas, newdata = Dallas1, interval = 'confidence', level = 0.95)
     fit      lwr      upr
55729.47 54063.14 57395.79
```

```
predict(Dallas, newdata = Dallas2, interval = 'confidence', level = 0.95)
     fit      lwr      upr
38965.87 35464.31 42467.44
```

6. For the city Dallas we know the pizza price was $2.77 in the last week of 1996. Suppose the price would increase to $2.99 in the following week. Can you predict the sales for that week and account for the uncertainty of your prediction? Do you think the resulting prediction is useful? Explain.

Answer:
I'm 95% confident that the predicted sales are $39,301.15 for next week given the price increase to $2.99. I can account for uncertainty in my prediction by using the 95% confidence interval. I believe the resulting prediction interval isn't very useful because the range of the lower and upper bounds of the confidence interval is too wide, since the interval is from $22425.65 to $56176.65. However, I believe the resulting prediction is useful because when I computed the confidence interval for the Price of $2.77, the prediction of sales ($46,677.13) is more sales than when the price is $2.99, which makes sense because people will pay less for expensive food.

R code:

```
xx <- data.frame(Dallas.Price = c(2.99))
predict(Dallas, newdata = xx, interval = 'prediction', level = 0.95)

yy <- data.frame(Dallas.Price = c(2.77))
predict(Dallas, newdata = yy, interval = 'prediction', level = 0.95)
```

R output:

When Dallas.Price is $2.99

```
     fit      lwr      upr
39301.15 22425.65 56176.65
```

When Dallas.Price is $2.77

```
     fit      lwr      upr
46677.13 30049.83 63304.42
```

Appendix

```
Chicago <-lm(Chicago.Volume ~ Chicago.Price, data=pizza)
Chicago
Denver <- lm(Denver.Volume ~ Denver.Price, data=pizza)
Denver
Dallas <- lm(Dallas.Volume ~ Dallas.Price, data=pizza)
Dallas
Baltimore <- lm(Baltimore.Volume ~ Baltimore.Price, data=pizza)
Baltimore
plot(Chicago$residuals, xlab = 'Time', ylab = 'Residual')
abline(a = 0, b = 0)
plot(Chicago$fitted.values, Chicago$residuals, xlab = 'Fitted value', ylab = 'Residual')
abline(a = 0, b = 0)
qqnorm(Chicago$residuals)
qqline(Chicago$residuals)
plot(Denver$residuals, xlab = 'Time', ylab = 'Residual')
abline(a = 0, b = 0)
plot(Denver$fitted.values, Denver$residuals, xlab = 'Fitted value', ylab = 'Residual')
abline(a = 0, b = 0)
qqnorm(Denver$residuals)
qqline(Denver$residuals)
plot(Dallas$residuals, xlab = 'Time', ylab = 'Residual')
abline(a = 0, b = 0)
plot(Dallas$fitted.values, Dallas$residuals, xlab = 'Fitted value', ylab = 'Residual')
abline(a = 0, b = 0)
qqnorm(Dallas$residuals)
qqline(Dallas$residuals)
plot(Baltimore$residuals, xlab = 'Time', ylab = 'Residual')
abline(a = 0, b = 0)
plot(Baltimore$fitted.values, Baltimore$residuals, xlab = 'Fitted value', ylab = 'Residual')
abline(a = 0, b = 0)
qqnorm(Baltimore$residuals)
qqline(Baltimore$residuals)
confint(Dallas, level=0.90)
summary(Dallas)
range(pizza$Dallas.Price)
Dallas1 <- data.frame(Dallas.Price = c(2.5))
predict(Dallas, newdata = Dallas1, interval = 'confidence', level = 0.95)
Dallas2 <- data.frame(Dallas.Price = c(3))
```

```
predict(Dallas, newdata = Dallas2, interval = 'confidence', level = 0.95)
xx <- data.frame(Dallas.Price = c(2.99))
predict(Dallas, newdata = xx, interval = 'prediction', level = 0.95)
yy <- data.frame(Dallas.Price = c(2.77))
predict(Dallas, newdata = yy, interval = 'prediction', level = 0.95)
```