Data Analysis Class Test 2

2700298

1 Relationship between house price and the number of bathrooms by type of parking

1.1 Introduction

As we all know,economic status could be reflected by housing prices.The data was collected by a realtor which included the house price (in pounds) of 300 houses at the time of sale within the last six months.Besides,the number of bathrooms and the type of parking for each house were collected as well.The main purpose of this report is to examine the relationship between house price and the number of bathrooms and try to determine whether it changes by type of parking or not.

1.2 Exploratory Analysis

There is no missing value in the data after checking.The summary statistics of the house prices by type of parking are given by Table 1 and the counts for different number of bathrooms by type of parking are given by Table 2.

From Table 1,we see that the mean price of houses with covered parking (510711.6 pounds) is greater than the mean price of houses with open parking (487128.4 pounds).What's more, the maximum price as well as the minimum price of houses with covered parking are respectively higher than the maximum price and minimum price of houses with open parking.In addition, the standard deviation of the price for houses with covered parking is greater than the standard deviation of the price for houses with open parking,which indicates that the difference in price among houses with covered parking is larger than the difference in price among houses with open parking.

Table 1: Summary statistics of house prices by type of parking

Variable	Parking	Mean	SD	Minimum	1st Q.	Median	3rd Q.	Maximum
Price	Covered	510711.6	143488.8	203417	402062.5	488625.0	623916.5	857667
Price	Open	487128.4	138443.6	124333	380854.2	488333.5	580708.2	811333

Table 2: Counts for different number of bathrooms by type of parking

Bathrooms	Coverd.number	Open.number
1	6	14
2	49	88
3	38	84
4	11	10

From Table 2, we notice that for houses with covered parking, 2 bathrooms are the most common type and 1 bathroom is the most uncommon type.For houses with open parking, 2 bathrooms are also the most common type while 4 bathrooms are the most uncommon type.

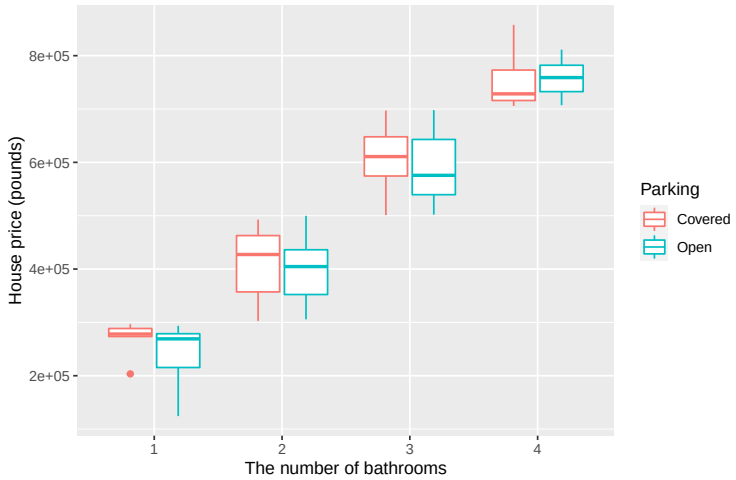


Figure 1: Relationship between house price and the number of bathrooms by type of parking.

Figure 1 shows the relationship between house price and the number of bathrooms by type of parking. From the plot,we see that the price increases with the number of bathrooms both for houses with covered parking and houses with open parking.In addition,the mean price of houses with covered parking is higher than the mean price of houses with open parking when the number of bathrooms equals to one,two,three respectively.However,the mean price of houses with covered parking is less than the mean price of houses with open parking when the number of bathrooms equals to four.

1.3 Formal Analysis

The multiple linear regression model with interaction term between the number of bathrooms and the type of parking (full model) will be fitted as follows:

$$\widehat{\text{Price}} = \hat{\alpha} + \hat{\beta}_{\text{bathroom=2}} \cdot \mathbb{I}_{\text{bathroom=2}}(i) + \hat{\beta}_{\text{bathroom=3}} \cdot \mathbb{I}_{\text{bathroom=3}}(i) +$$

$$\hat{\beta}_{\text{bathroom=4}} \cdot \mathbb{I}_{\text{bathroom=4}}(i) + \hat{\beta}_{\text{open}} \cdot \mathbb{I}_{\text{open}}(i) + \textit{interaction terms}$$

where

- the intercept $\hat{\alpha}$ is the mean house price for the baseline category (the number of bathrooms equals to one);
- $\hat{\beta}_{\text{bathroom=x}}$ is the difference in the mean price of houses with x bathrooms (x=2,3,4) relative to the baseline category (the number of bathrooms equals to one);
- $\mathbb{I}_{\text{bathroom=x}}(i)$ is an indicator function such that

$$\mathbb{I}_{\text{bathroom=x}}(i) = \begin{cases} 1 & \textit{if the } i\text{th observation has x bathrooms} (x = 2, 3, 4) \\ 0 & \text{Otherwise.} \end{cases}$$

Backward stepwise regression selection based on AIC is used for reducing the full model terms and obtaining the final model.The final model is given as follows:

$$\widehat{\text{Price}} = \hat{\alpha} + \hat{\beta}_{\text{bathroom=2}} \cdot \mathbb{I}_{\text{bathroom=2}}(i) + \hat{\beta}_{\text{bathroom=3}} \cdot \mathbb{I}_{\text{bathroom=3}}(i) + \hat{\beta}_{\text{bathroom=4}} \cdot \mathbb{I}_{\text{bathroom=4}}(i) + \hat{\beta}_{\text{open}} \cdot \mathbb{I}_{\text{open}}(i)$$

Table 3 displays the estimated intercept and slope parameters from the final model.

Table 3: Estimates of the regression coefficients from the final model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	260409.9	13211.689	19.711	0.000	234408.77	286410.987
Bathrooms: 2	151857.4	13216.732	11.490	0.000	125846.33	177868.394
Bathrooms: 3	344228.3	13314.300	25.854	0.000	318025.23	370431.330
Bathrooms: 4	500880.9	17309.473	28.937	0.000	466815.23	534946.632
Parking: Open	-14543.9	6738.707	-2.158	0.032	-27805.93	-1281.865

Hence, from Table 3 we obtain the regression models for houses with covered parking and houses with open parking:

$$\widehat{\text{Price}}_{\text{Covered}} = 2.6040988 \times 10^5 + 1.5185736 \times 10^5 \cdot \mathbb{I}_{\text{bathroom=2}}(i) + 3.4422828 \times 10^5 \cdot \mathbb{I}_{\text{bathroom=3}}(i) + 5.0088093 \times 10^5 \cdot \mathbb{I}_{\text{bathroom=4}}(i)$$

$$\widehat{\text{Price}}_{\text{Open}} = 2.4586598 \times 10^5 + 1.5185736 \times 10^5 \cdot \mathbb{I}_{\text{bathroom=2}}(i) + 3.4422828 \times 10^5 \cdot \mathbb{I}_{\text{bathroom=3}}(i) + 5.0088093 \times 10^5 \cdot \mathbb{I}_{\text{bathroom=4}}(i)$$

Firstly, for the price of houses with covered parking: If the number of bathrooms is one,the price would be 2.6040988×10^5 pounds; If the number of bathrooms is two,the price would be 4.1226724×10^5 pounds; If the number of bathrooms is three,the price would be 6.0463816×10^5 pounds; If the number of bathrooms is four,then the price would be 7.6129081×10^5 pounds.

Then, for the price of houses with open parking: If the number of bathrooms is one,the price would be 2.4586598×10^5 pounds; If the number of bathrooms is two,the price would be 3.9772334×10^5 pounds; If the number of bathrooms is three,the price would be 5.9009426×10^5 pounds; If the number of bathrooms is four,then the price would be 7.4674691×10^5 pounds.

After interpreting the model regression results, assumptions checking are shown in Figure 2 and Figure 3.

Figure 2 displays the scatterplots of the residuals against bathrooms(left) and the fitted values (right) by type of parking. From the scatterplots, we see that there is approximately an even spread of the residuals above and below the zero line for all numbers of bathrooms and fitted values after ignoring the outliers,hence our assumption that the residuals have mean zero appears valid. However,the assumption that the residuals have constant variance across all levels of the fitted values seems to be not valid.

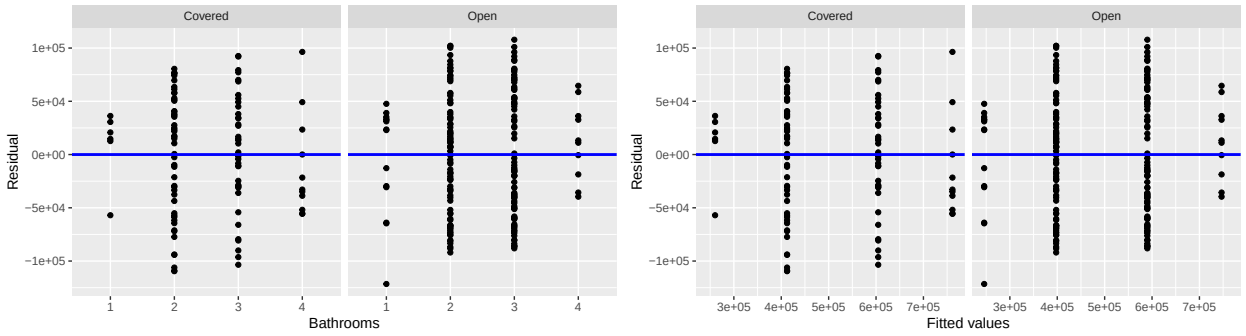


Figure 2: Scatterplots of the residuals against Bathrooms(left) and the fitted values (right) by parking.

Figure 3 shows the histogram of the residuals by type of parking.The histogram is relatively not bell-shaped and doesn't centre around zero. Hence, the assumption of normally distributed errors doesn't appear to hold for the fitted regression model.

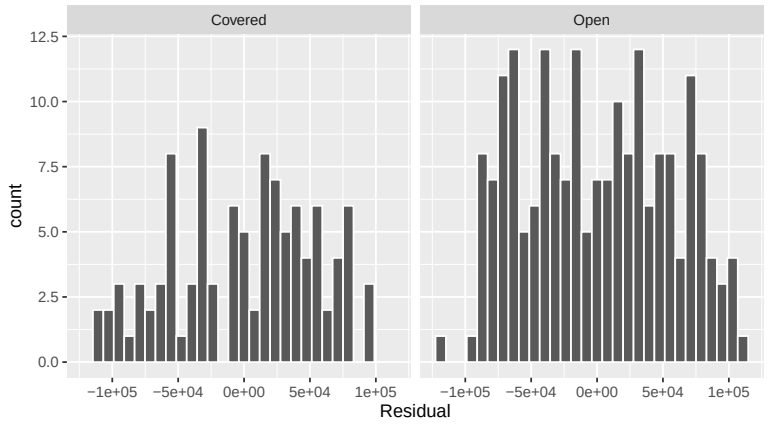


Figure 3: Histogram of the residuals by type of parking.

1.4 Conclusions

The relationship between house price and the number of bathrooms is positive which indicates that the house price will increase with the number of bathrooms.Further more, we find that the relationship between house price and the number of the bathrooms does not differ by type of parking since the interaction terms in the full model is not significant and will be removed.

In conclusion, the house price has a positive trend with the number of bathrooms and the type of parking would not make a change to this relationship. However, the price of houses with covered parking is greater than the price of houses with open parking commonly speaking.

2 Further Question 1

2.1 (a)

```
log.model <- glm(Promotion ~ Sales, data = grocery, family = binomial(link = "logit"))
log.model %>%
  summary()
```

```
Call:
glm(formula = Promotion ~ Sales, family = binomial(link = "logit"),
    data = grocery)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3300  -0.7539  -0.4162   0.7537   2.1987

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.542e+00  6.477e-02  -39.25   <2e-16 ***
Sales        1.267e-03  3.125e-05   40.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9783.1  on 7059  degrees of freedom
Residual deviance: 6761.8  on 7058  degrees of freedom
AIC: 6765.8

Number of Fisher Scoring iterations: 5
```

2.2 (b)

```
mod.coefs <- log.model %>%
              summary() %>%
              coef()

odds.lower <- exp(mod.coefs["Sales", "Estimate"]
                  - 1.96 * mod.coefs["Sales", "Std. Error"])
odds.lower
```

```
[1] 1.001206
```

```
odds.upper <- exp(mod.coefs["Sales", "Estimate"]
                  + 1.96 * mod.coefs["Sales", "Std. Error"])
odds.upper
```

```
[1] 1.001329
```

The odds of an item being on promotion increase by between 0.12% and 0.13% for every 1 number increase in Sales.

2.3 (c)

```
exp(mod.coefs["(Intercept)", "Estimate"] + mod.coefs["Sales", "Estimate"] * 2100)
```

```
[1] 1.124781
```

The odds of an item being on promotion given the sales of 2100 are 12% greater than not being on promotion.

3 Further Question 2

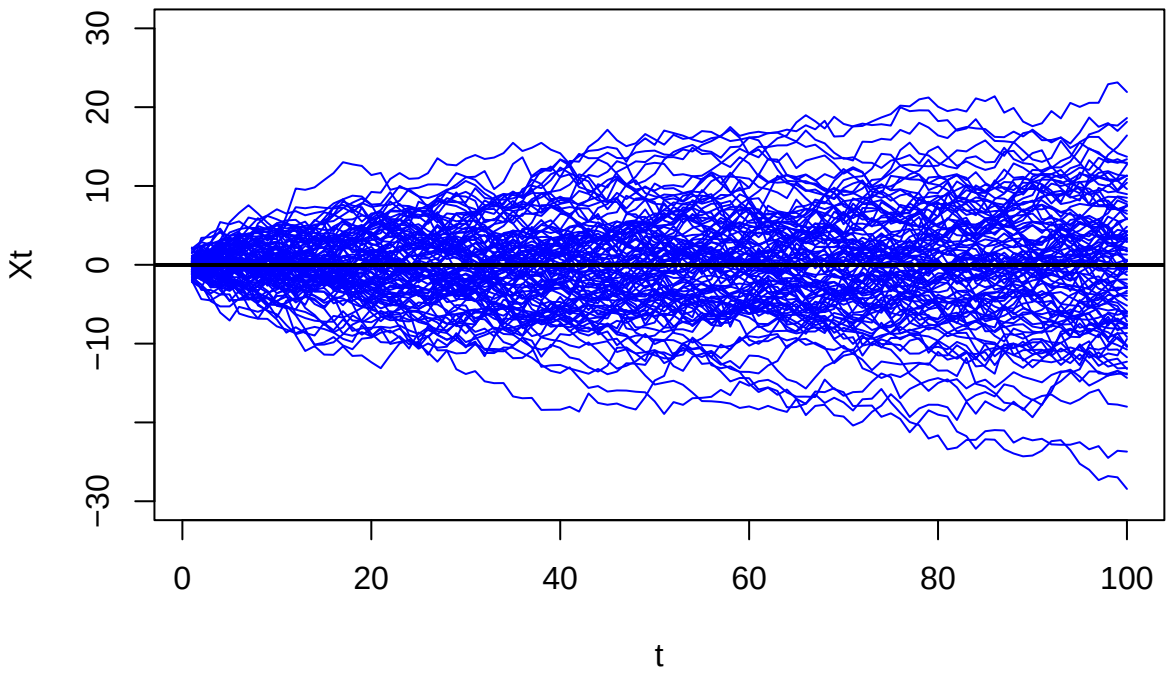
3.1 (a)

```
Xt<-matrix(0,nrow=100,ncol=100)
for(i in 1:100){
  zt<-rnorm(100, mean = 0, sd = 1)
  Xt[i,]<-cumsum(zt)
}
```

3.2 (b)

```
t<-1:100
plot(t,Xt[1,],type="l",ylab="Xt",ylim=c(-30,30),col="blue")

for(i in 2:100){
  lines(t,Xt[i,],col="blue")
}
abline(h=sum(Xt)/10000,lwd=2,col="black")
```



From the plot, we are able to find that the variance of each random walk process would increase as length t becomes larger.However,the overall mean of the 100 random walk processes keeps around 0.