Statistical Analysis for a Publishing Company

Shih-Yu Huang

Data Dictionary

The data provided contains information on e-book sales over a period of many months. Each row in the data represents one book. The values of the variables are taken across the entire time period, so daily.sales is the average number of sales (minus refunds) across all days in the period, and sale.price is the average price for which the book sold across all sales in the period.

Variables	Description
sold by	Publishing or E-commerce company that sold the book
publisher.type	Type of publisher
genre	Genre of the book
avg.review	Average rating received for the book
daily.sales	Average number of sales (minus refunds) across all days in the period Total number of reviews received for the book
total.reviews	Average price for which the book sold across all sales in the period

Data Preparation and Manipulation

```
# Read the dataset
publisher_sales <- read_csv("publisher_sales.csv")
```

```
## Rows: 6000 Columns: 7
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): sold by, publisher.type, genre
## dbl (4): avg.review, daily.sales, total.reviews, sale.price
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Check the structure
str(publisher_sales)
```

```
## spc_tbl_ [6,000 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ sold by       : chr [1:6000] "Random House LLC" "Amazon Digital Services,  Inc." "Amazon Digital Services,
Inc." "Amazon Digital Services,  Inc." ...
##  $ publisher.type: chr [1:6000] "big five" "indie" "small/medium" "small/medium" ...
##  $ genre         : chr [1:6000] "childrens" "non_fiction" "non_fiction" "fiction" ...
##  $ avg.review    : num [1:6000] 4.44 4.19 3.71 4.72 4.65 4.81 4.33 4.21 3.95 4.66 ...
##  $ daily.sales   : num [1:6000] 61.5 74.9 66 85.2 37.7 ...
##  $ total.reviews : num [1:6000] 92 130 118 179 111 106 205 86 161 81 ...
##  $ sale.price    : num [1:6000] 8.03 9.08 9.48 12.32 5.78 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `sold by` = col_character(),
##   ..   publisher.type = col_character(),
##   ..   genre = col_character(),
##   ..   avg.review = col_double(),
##   ..   daily.sales = col_double(),
##   ..   total.reviews = col_double(),
##   ..   sale.price = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# Clean data and store in another dataframe
publisher_sales <- na.omit(publisher_sales)

# Change data type
column_f <- c("sold by", "publisher.type", "genre")
publisher_sales[column_f] <- lapply(publisher_sales[column_f], as.factor)

#Check the structure again
str(publisher_sales)
```

```
## tibble [6,000 × 7] (S3: tbl_df/tbl/data.frame)
##  $ sold by       : Factor w/ 13 levels "Amazon Digital Services,  Inc.",..: 11 1 1 1 13 13 1 6 1 1 ...
##  $ publisher.type: Factor w/ 5 levels "amazon","big five",..: 2 3 5 5 2 2 5 2 5 5 ...
##  $ genre         : Factor w/ 3 levels "childrens","fiction",..: 1 3 3 2 1 1 2 1 2 1 ...
##  $ avg.review    : num [1:6000] 4.44 4.19 3.71 4.72 4.65 4.81 4.33 4.21 3.95 4.66 ...
##  $ daily.sales   : num [1:6000] 61.5 74.9 66 85.2 37.7 ...
##  $ total.reviews : num [1:6000] 92 130 118 179 111 106 205 86 161 81 ...
##  $ sale.price    : num [1:6000] 8.03 9.08 9.48 12.32 5.78 ...
```

Do books from different genres have different daily sales on average?

```
#Test whether genre has a significant effect on daily sales
daily.sales.by.genre <- lm(daily.sales ~ genre, data = publisher_sales)
anova(daily.sales.by.genre)
```

```
## Analysis of Variance Table
##
## Response: daily.sales
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## genre        2 2562528 1281264  2590.5 < 2.2e-16 ***
## Residuals 5997 2966133     495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Explain the difference
summary(daily.sales.by.genre)
```

```
##
## Call:
## lm(formula = daily.sales ~ genre, data = publisher_sales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.396  -13.326   -0.076   13.249  102.094
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       55.5773     0.4973  111.76   <2e-16 ***
## genrefiction      50.3087     0.7033   71.53   <2e-16 ***
## genrenon_fiction  20.2886     0.7033   28.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.24 on 5997 degrees of freedom
## Multiple R-squared:  0.4635, Adjusted R-squared:  0.4633
## F-statistic:  2590 on 2 and 5997 DF,  p-value: < 2.2e-16
```

```
#Calculate estimated marginal means to get the means for each genre
daily.sales.by.genre.emm <- emmeans(daily.sales.by.genre, ~genre)
daily.sales.by.genre.emm
```

```
##  genre       emmean    SE   df lower.CL upper.CL
##  childrens     55.6 0.497 5997     54.6     56.6
##  fiction      105.9 0.497 5997    104.9    106.9
##  non_fiction   75.9 0.497 5997     74.9     76.8
##
## Confidence level used: 0.95
```

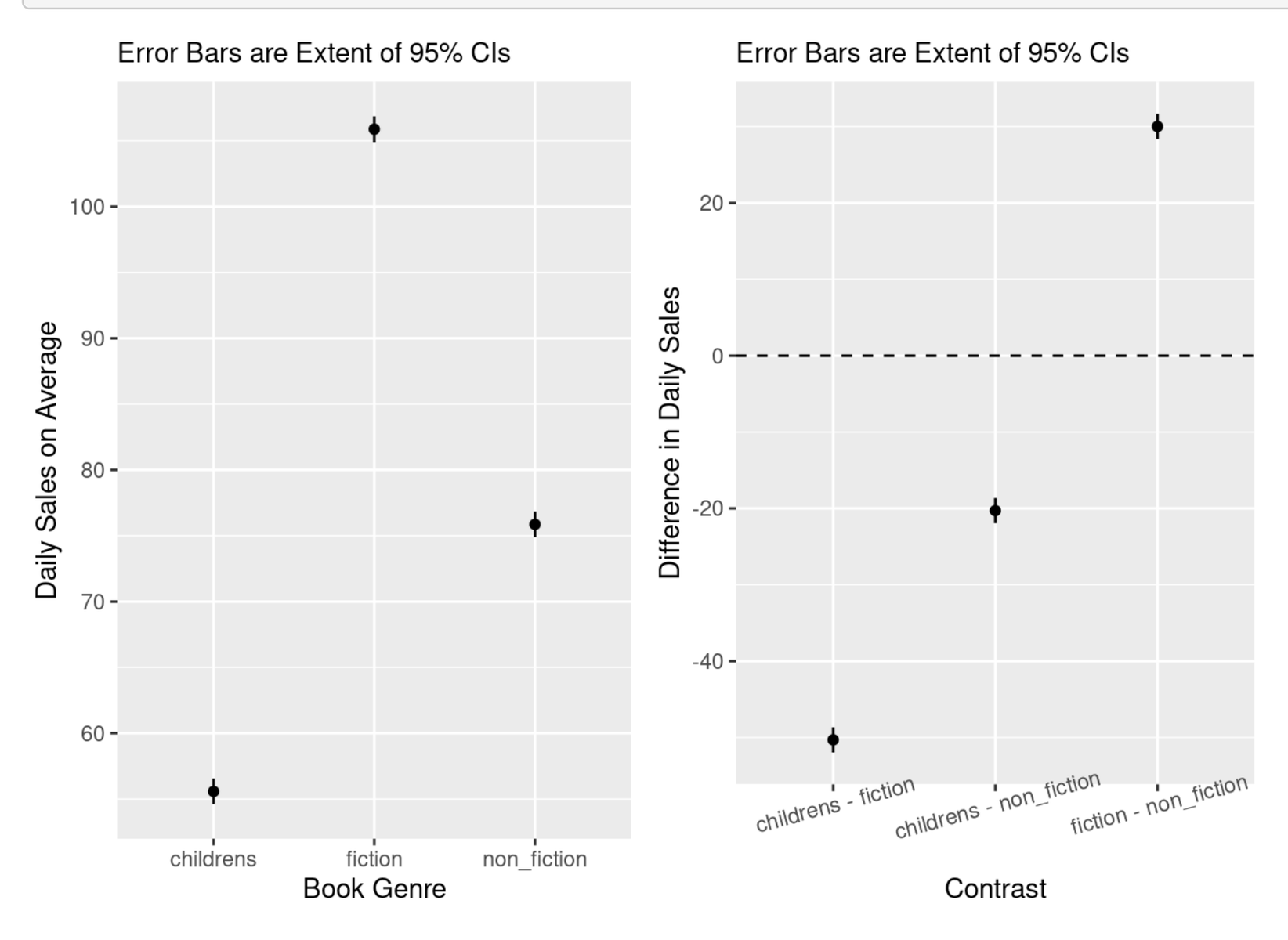
```
#Do pairwise contrasts
daily.sales.by.genre.pairs <- confint(pairs(daily.sales.by.genre.emm))
daily.sales.by.genre.pairs
```

```
##  contrast                 estimate    SE   df lower.CL upper.CL
##  childrens - fiction         -50.3 0.703 5997    -52.0    -48.7
##  childrens - non_fiction     -20.3 0.703 5997    -21.9    -18.6
##  fiction - non_fiction        30.0 0.703 5997     28.4     31.7
##
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
```

```
#Visualisation
p.daily.sales <- ggplot(summary(daily.sales.by.genre.emm), aes(x=genre, y=emmean, ymin=lower.CL, ymax=upper.CL))
+ geom_point() + geom_linerange() + labs(x="Book Genre", y="Daily Sales on Average", subtitle="Error Bars are Ext
ent of 95% CIs")

p.contrasts <- ggplot(daily.sales.by.genre.pairs, aes(x=contrast, y=estimate, ymin=lower.CL, ymax=upper.CL)) + ge
om_point() + geom_linerange() + geom_hline(yintercept=0, lty=2) + labs(x="Contrast", y="Difference in Daily Sale
s", subtitle="Error Bars are Extent of 95% CIs") + theme(axis.text.x = element_text(angle = 15))

grid.arrange(p.daily.sales, p.contrasts, ncol = 2)
```



Do books have more/fewer sales depending upon their average review scores and total number of reviews.

```
#Test whether average review score and review numbers have a significant effect on daily sales
daily.sales.by.review.main <- lm(daily.sales ~ avg.review + total.reviews, data = publisher_sales)
summary(daily.sales.by.review.main)
```

```
##
## Call:
## lm(formula = daily.sales ~ avg.review + total.reviews, data = publisher_sales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.396  -14.645   -1.059   13.690  122.429
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.870506   2.341271  10.196  < 2e-16 ***
## avg.review    -3.943548   0.513120  -7.685 1.77e-14 ***
## total.reviews  0.543329   0.007823  69.451  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.6 on 5997 degrees of freedom
## Multiple R-squared:  0.4458, Adjusted R-squared:  0.4456
## F-statistic:  2412 on 2 and 5997 DF,  p-value: < 2.2e-16
```

```
cbind(coef(daily.sales.by.review.main), confint(daily.sales.by.review.main))
```

```
##                             2.5 %     97.5 %
## (Intercept)   23.870506 19.2807719 28.4602399
## avg.review    -3.943548 -4.9494480 -2.9376473
## total.reviews  0.543329  0.5279926  0.5586653
```

```
#Consider interaction between review score and review numbers
daily.sales.by.review.intr <- lm(daily.sales ~ avg.review * total.reviews, data = publisher_sales)
summary(daily.sales.by.review.intr)
```

```
##
## Call:
## lm(formula = daily.sales ~ avg.review * total.reviews, data = publisher_sales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -104.08  -14.63   -0.92   13.82   92.33
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              63.546900   4.178047  15.210  < 2e-16 ***
## avg.review              -13.683765   0.993159 -13.778  < 2e-16 ***
## total.reviews             0.164754   0.034068   4.836 1.36e-06 ***
## avg.review:total.reviews   0.091688   0.008035  11.411  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.36 on 5996 degrees of freedom
## Multiple R-squared:  0.4576, Adjusted R-squared:  0.4573
## F-statistic:  1686 on 3 and 5996 DF,  p-value: < 2.2e-16
```

```
cbind(coef(daily.sales.by.review.intr), confint(daily.sales.by.review.intr))
```

```
##                                              2.5 %      97.5 %
## (Intercept)               63.54690004  55.35642562  71.7373745
## avg.review               -13.68376484 -15.63071313 -11.7368165
## total.reviews              0.16475390   0.09796872   0.2315391
## avg.review:total.reviews   0.09168842   0.07593650   0.1074403
```

```
anova(daily.sales.by.review.main, daily.sales.by.review.intr)
```

```
## Analysis of Variance Table
##
## Model 1: daily.sales ~ avg.review + total.reviews
## Model 2: daily.sales ~ avg.review * total.reviews
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   5997 3064100
## 2   5996 2998976  1     65125 130.21 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

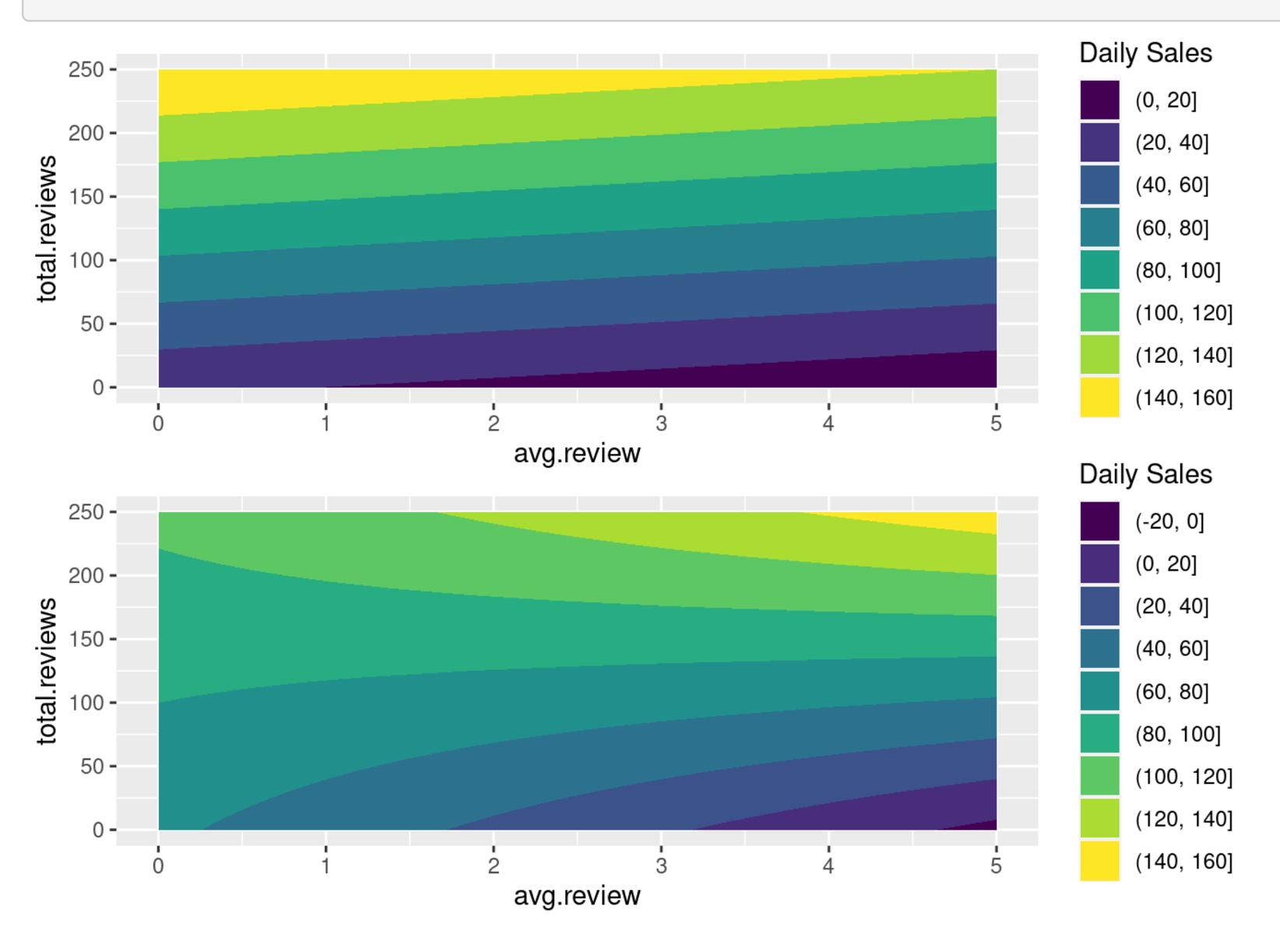
```
#Visualisation
sale.preds <- tibble(avg.review = unlist(expand.grid(seq(0, 5, 1), seq(0, 250, 5))[1]),
                     total.reviews = unlist(expand.grid(seq(0, 5, 1), seq(0, 250, 5))[2]))

sale.preds <- mutate(sale.preds,
                     sale.hat.main = predict(daily.sales.by.review.main, sale.preds),
                     sale.hat.intr = predict(daily.sales.by.review.intr, sale.preds))

p1 <- ggplot(sale.preds, aes(avg.review, total.reviews)) + geom_contour_filled(aes(z = sale.hat.main)) + guides(f
ill=guide_legend(title="Daily Sales"))

p2 <- ggplot(sale.preds, aes(avg.review, total.reviews)) + geom_contour_filled(aes(z = sale.hat.intr)) + guides(f
ill=guide_legend(title="Daily Sales"))

grid.arrange(p1, p2)
```



What is the effect of sale price upon the number of sales, and is this different across genres?

```
#Examine the effect of sale price upon book sales
daily.sales.by.price <- lm(daily.sales ~ sale.price, data = publisher_sales)
summary(daily.sales.by.price)
```

```
##
## Call:
## lm(formula = daily.sales ~ sale.price, data = publisher_sales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.769 -20.650  -4.633  17.099 130.315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 112.0820     1.5207   73.70   <2e-16 ***
## sale.price   -3.8156     0.1705  -22.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.17 on 5998 degrees of freedom
## Multiple R-squared:  0.07706,    Adjusted R-squared:  0.07691
## F-statistic: 500.8 on 1 and 5998 DF,  p-value: < 2.2e-16
```

```
#Examine the effect of sale price upon book sales across genre
daily.sales.by.price.main <- lm(daily.sales ~ sale.price + genre, data = publisher_sales)
summary(daily.sales.by.price.main)
```

```
##
## Call:
## lm(formula = daily.sales ~ sale.price + genre, data = publisher_sales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.357  -13.311    0.031   13.097  102.924
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       63.8931     1.5195   42.05  < 2e-16 ***
## sale.price        -0.8324     0.1438   -5.79  7.4e-09 ***
## genrefiction      48.6713     0.7562   64.36  < 2e-16 ***
## genrenon_fiction  18.5587     0.7624   24.34  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.18 on 5996 degrees of freedom
## Multiple R-squared:  0.4665, Adjusted R-squared:  0.4662
## F-statistic:  1748 on 3 and 5996 DF,  p-value: < 2.2e-16
```

```
#Consider interaction between sale price and genre
daily.sales.by.price.all <- lm(daily.sales ~ sale.price * genre, data = publisher_sales)
summary(daily.sales.by.price.all)
```

```
##
## Call:
## lm(formula = daily.sales ~ sale.price * genre, data = publisher_sales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -102.38  -13.37    0.03   13.08  102.37
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  72.8781     2.5025  29.122  < 2e-16 ***
## sale.price                   -1.7319     0.2456  -7.053 1.95e-12 ***
## genrefiction                 35.1993     3.2740  10.751  < 2e-16 ***
## genrenon_fiction              6.5492     3.2040   2.044 0.040989 *
## sale.price:genrefiction       1.4587     0.3546   4.114 3.94e-05 ***
## sale.price:genrenon_fiction   1.2817     0.3469   3.695 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.15 on 5994 degrees of freedom
## Multiple R-squared:  0.4683, Adjusted R-squared:  0.4679
## F-statistic:  1056 on 5 and 5994 DF,  p-value: < 2.2e-16
```

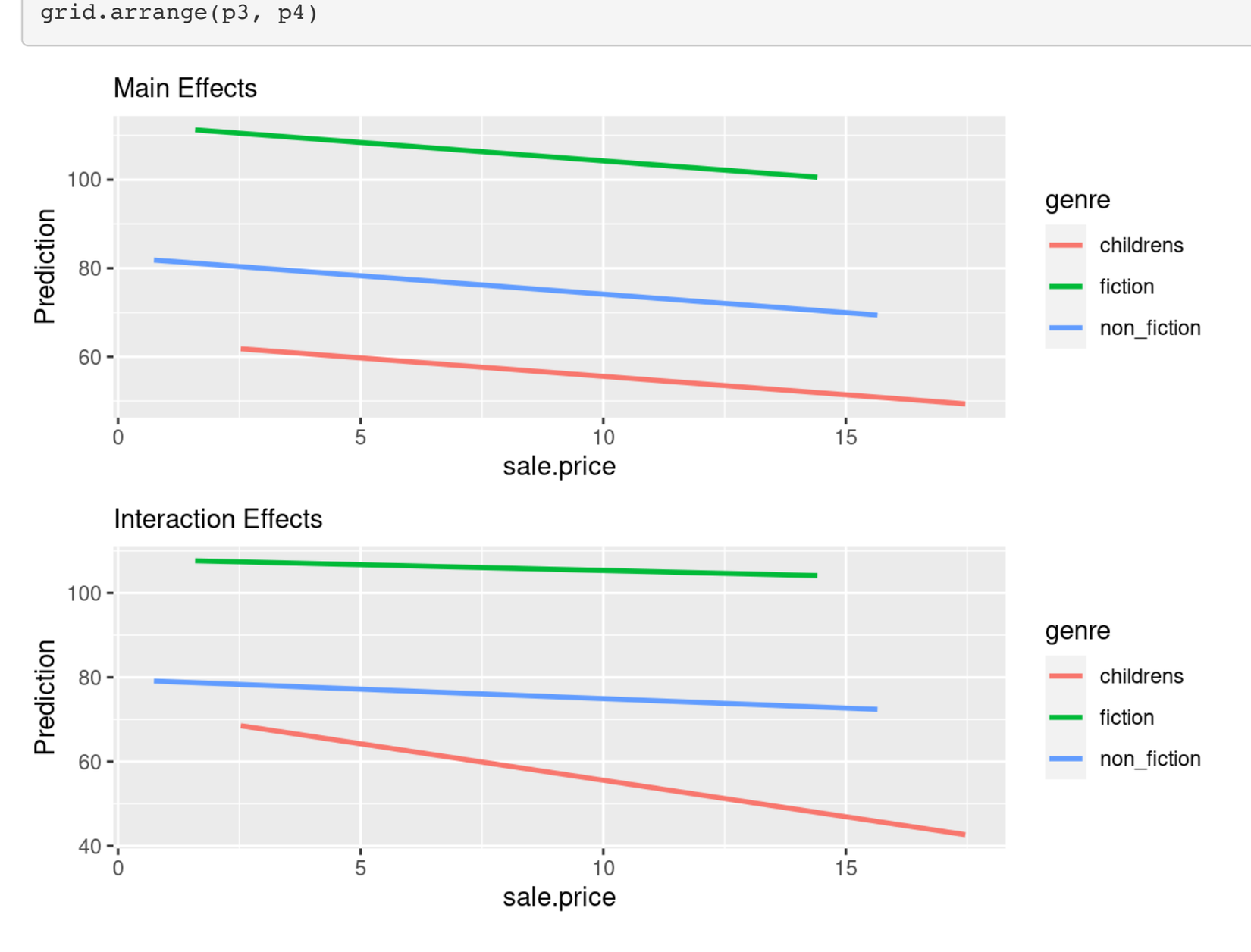
```
#Use vif scores to check multicollinearity
vif(daily.sales.by.price.main)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## sale.price 1.229697  1        1.108917
## genre      1.229697  2        1.053051
```

```
#Visualisation
p3 <- mutate(publisher_sales,
       main.hat = predict(daily.sales.by.price.main, publisher_sales),
       intr.hat = predict(daily.sales.by.price.all, publisher_sales)) %>%
ggplot() +
    geom_line(aes(sale.price, main.hat, colour = genre), linewidth = 1) +
  labs(y = "Prediction", subtitle = "Main Effects")

p4 <- mutate(publisher_sales,
       main.hat = predict(daily.sales.by.price.main, publisher_sales),
       intr.hat = predict(daily.sales.by.price.all, publisher_sales)) %>%
ggplot() +
    geom_line(aes(sale.price, intr.hat, colour = genre), linewidth = 1) +
  labs(y = "Prediction", subtitle = "Interaction Effects")

grid.arrange(p3, p4)
```



Analysis and Conclusion

This report displays the analysis results requested by the management team of publishing company. In this report, we especially focus on the factors (genre, review numbers, review score and sale price) and discuss whether they are more likely to affect the book sales. The data used in this analysis is collected by the publishing company's sale records over a period of many months and expected to offer some insights to understand the key to growth in sales.

First of all, we take a look at the sales in different genres. We can see there are different average sales in different genres. From the table 2, the average sale of children genre is approximate 55.6. The fiction genre has a difference of 50.3 from children genre and reaches 105.9 in its average sale, while the non-fiction genre possesses a difference of around 20.3 and an average sale of 75.9.

Table 3. The Average Sale and Difference in Each Genre.

Genre	Difference	Estimated Marginal Mean
Children	55.6	55.6
Fiction	50.3	105.9
Non Fiction	20.3	75.9

Digging deeper, we research on the difference between each genre, not only limited to compare with the children genre. The left panel of figure 5 states the mean daily book sale for each genre. The fiction genre has the highest average sales, followed by the non-fiction genre. Genre for children has the lowest average sales. The right panel shows the estimates of the difference in book sales for each pair of genres.

1. The estimate for the Childrens-Fiction comparison shows a point estimate of 50.2 greater sales for the fiction genre, but the 95% CI spans 48.7 to 52 greater for the same genre.
2. The estimate for the Childrens-NonFiction comparison sees a point estimate of 20.3 greater sales for the non-fiction genre, but the 95% CI spans 18.6 to 21.9 greater for the same genre.
3. The estimate for the Fiction-NonFiction comparison exhibits a point estimate of 30 greater sales for the fiction genre, but the 95% CI spans 28.4 to 31.7 greater for the same genre.

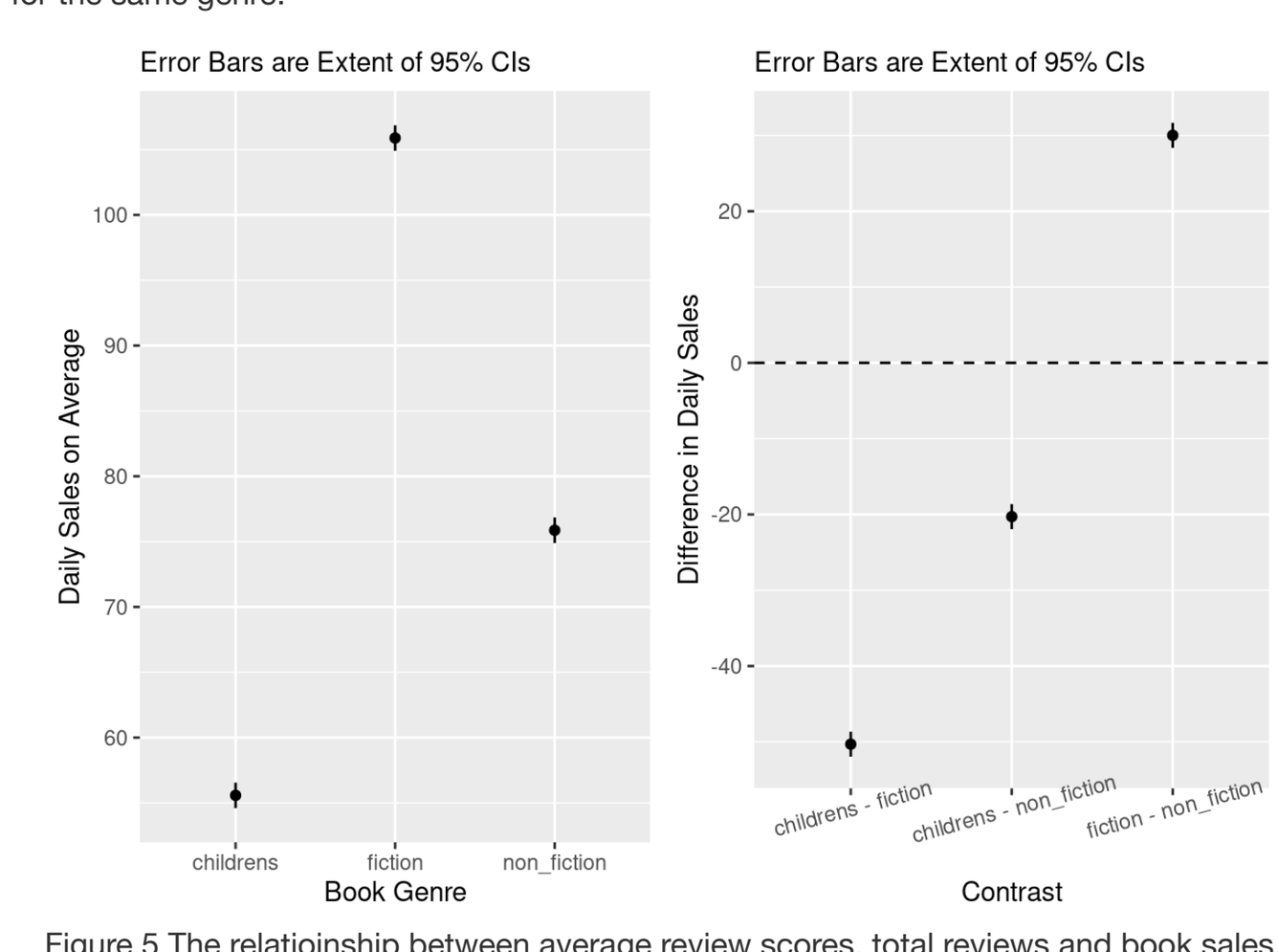


Figure.5 The relatioinship between average review scores, total reviews and book sales.

Next, we choose other two variables (average review score and total reviews) to evaluate their impacts on book sales. From the table 4, we can see there are a negative and a positive correlations in review scores and total reviews respectively. The book sales decrease by 3.94 for every extra average review score, while the sales increase by 0.54 for every extra review. When considering the interaction between average scores and reviews, we can spot changes in both two variables. The interaction terms tell that each additional review increases the effectiveness of average review scores on book sales by 0.09.

Table 4. The Average Review Scores, Total Reviews and Interaction.

Variables	Coefficient	Coefficient with Interaction
Average Review Scores	-3.94	-13.68
Total Reviews	0.54	0.16
Interaction	N/A	0.09

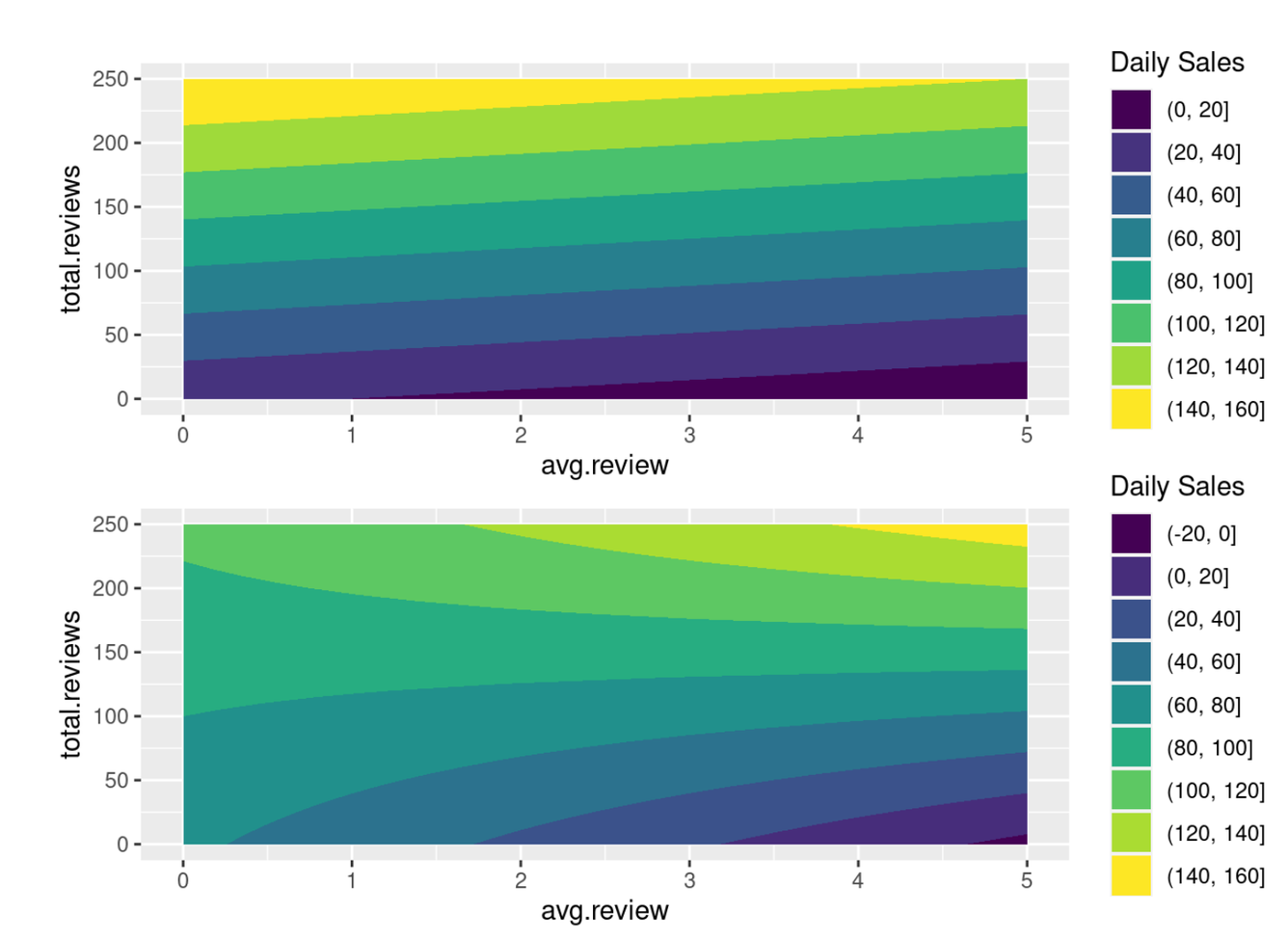


Figure.6 Main Effect vs. Interaction Effect.

Finally, we evaluate the effect of sale price upon book sales. From table 5, the book sales decrease by 3.81 for every extra sale price. When considering book genres and their interactions, the book sales slightly decrease by 0.81 and 1.73 for every extra sale price, respectively. The figure 7 shows that there is a significant main effect of sale price upon book sales and of genres upon book sales. There is also a significant interaction effect between sale price and genres, with the positive effect of sale price being significantly larger when genres are present. For example, we can see the sale grows by 48.67 for every extra fiction book. Furthermore, the effect of sale price is different across genres. Each additional fiction book increases the effectiveness of sale price on book sales by 1.46, while every extra non-fiction book enhances the effectiveness of sale price on book sales by 1.28.

Table 5. The Sale price, Genre and Interaction.

Variables	Coefficient	Coefficient with Genre	Coefficient with Interaction
Sale Price	-3.81	-0.83	-1.73
Genre: Children	N/A	63.89	72.88
Genre: Fiction	N/A	48.67	35.2
Genre: Non-Fiction	N/A	15.56	6.55
Interaction 1 (Fiction)	N/A	N/A	1.46
Interaction 2 (Non-Fiction)	N/A	N/A	1.28

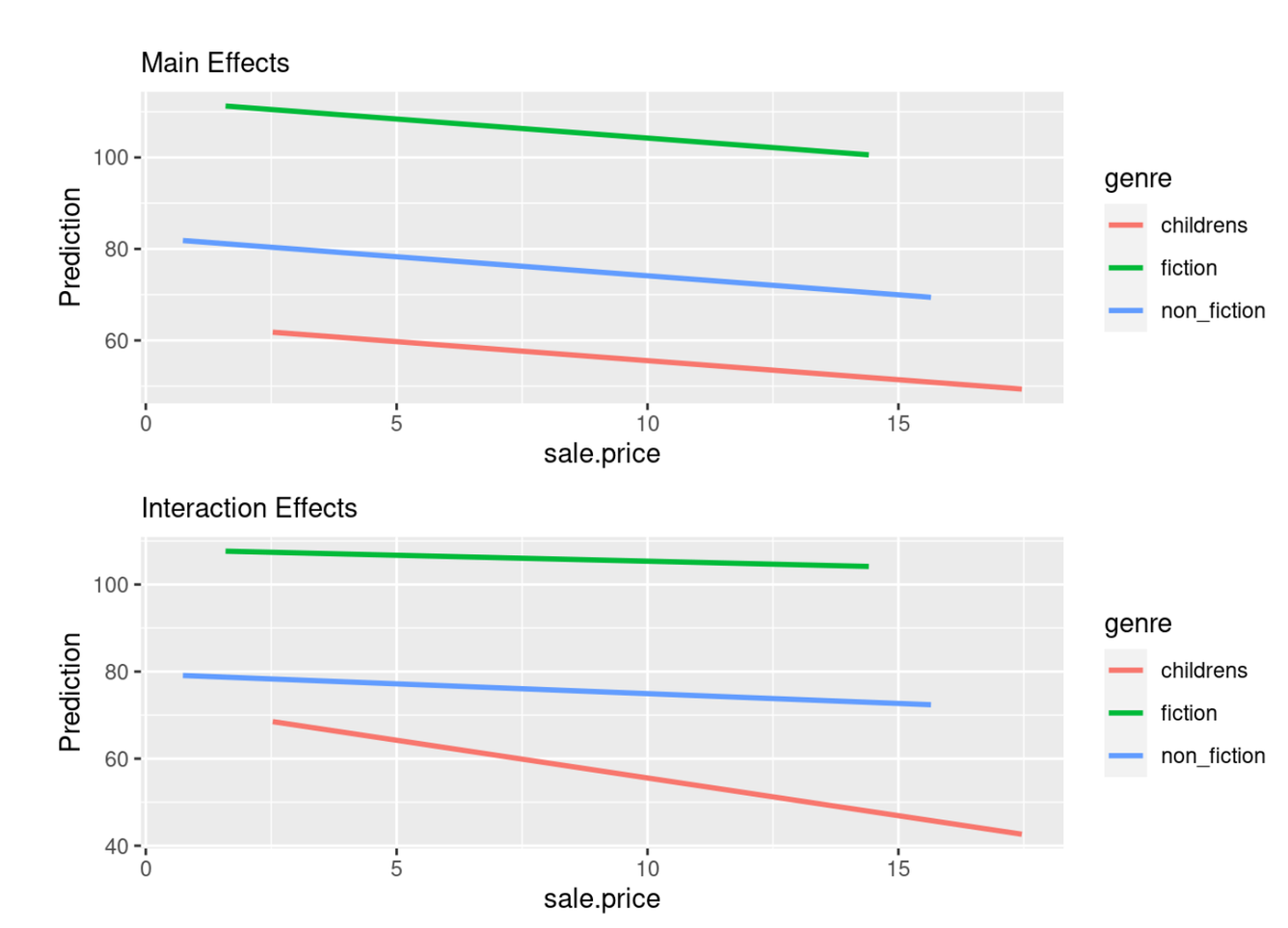


Figure.7 Main Effect vs. Interaction Effect.