Linear Regression 4

Iris Kim (6757827) and Ted Tinker (3223468)

May 4, 2017

A Report of Ring Prices by Diamond Weight (Carats)

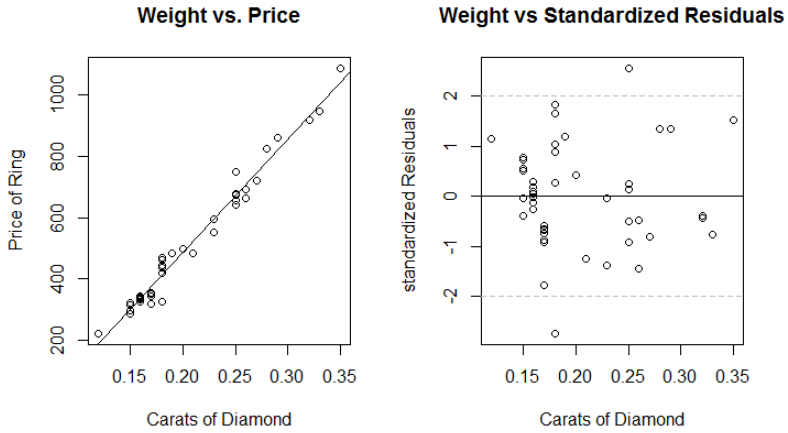
Abstract

This study investigates the relationship between the weight of the diamond on a golden ring (in carats) and the price of that ring. The data consists of 48 diamond weights and the price of the corresponding ring from an ad in the *Straights Times* newspaper by a jewelry retailer in Singapore. The models we create show a strong correlation.

Method

We loaded the dataset .txt into R-Studio and found the least squares regression line for the data. Then we made scatterplots for the data and its standardized residuals to assess whether a linear regression model was appropriate (that is, normality of errors, constant variance of errors across the range, and independence of errors). We used the following code:

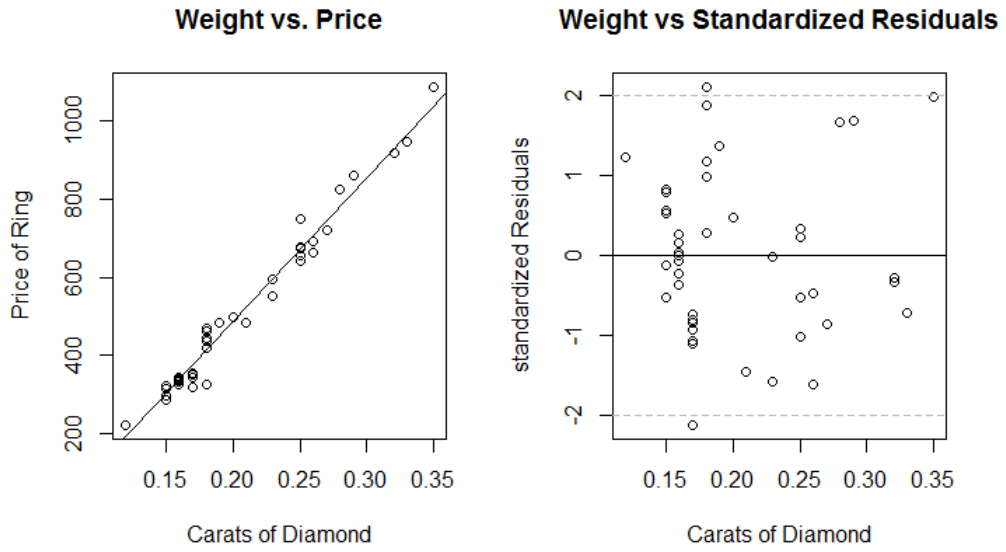
```
diamondLine <- lm(diamonds$Price~diamonds$Size) # Makes regression line for the data
par(mfrow=c(1,2)) # Two side-by-side plots
plot(diamonds,main="Weight vs. Price",xlab="Carats of Diamond",ylab="Price of Ring")
    # First plot: standard
abline(diamondLine$coefficients[1],diamondLine$coefficients[2]) # Add regression line
stanres1 <- rstandard(diamondLine) # Find standardized residuals
plot(diamonds$Size,stanres1,main="Weight vs Standardized Residuals",xlab="Carats of Diamond",
    ylab="standardized Residuals") # Second plot: carats vs standardized residuals
abline(h=2,col="gray",lty=2) # Add dashed lines to residual plot
abline(h=-2,col="gray",lty=2)
abline(h=0)
```



The model appears to capture the data's strong linear correlation, though there are two datapoints more than two standard deviations from the expectation. Errors are fairly well distributed (although perhaps more widely spread for diamonds of lower weight). If we were to use this model, with no alterations, then the regression line would be $Price = -258.05 + 3715.02(carats)$, such that a gold ring with no diamond should cost negative $258.05, and for each carat added to the diamond, the price increases by an expected $3,715.02.

For comparison, we evaluated the same regression line, but omitted the two points with standardized residuals outside two standard deviations using the following code:

```
diamondLine <- lm(diamonds$Price[-c(4,19)]~diamonds$Size[-c(4,19)]) # [-c(4,19)] skips rows 4, 19
par(mfrow=c(1,2)) # Two side-by-side plots
plot(diamonds,main="Weight vs. Price",xlab="Carats of Diamond",ylab="Price of Ring")
    # First plot: standard
abline(diamondLine$coefficients[1],diamondLine$coefficients[2]) # Add regression line
stanres1 <- rstandard(diamondLine) # Find standardized residuals
plot(diamonds$Size[-c(4,19)],stanres1,main="Weight vs Standardized Residuals",
    xlab="Carats of Diamond", ylab="standardized Residuals")
    # Second plot: carats vs standardized residuals
abline(h=2,col="gray",lty=2) # Add dashed lines to residual plot
abline(h=-2,col="gray",lty=2)
abline(h=0)
```



As before, the differences between expected value and actual value are well-distributed except that variance seems to be higher on the left (where there are more data-points anyway, so this is not unexpected). Removing rows 4 and 19 changed the variance such that there are two new outliers, but these are closer to the 2 standard deviation mark than the outliers in the previous graphs. For these reasons, we believe this model more accurately reflects the data. We will discuss the implications of this model in the Analysis section.

The weaknesses of this second model include those two new outliers, which may be pulling the regression line up or down. In addition, omitting two data points could have made the model more accurate, or we could be discarding valuable information. It is difficult to tell without more data.

Analysis

The new regression line, without those two outliers, is $Price = -250.31 + 3677.49(carats)$, such that a gold ring with no diamond should cost negative \$250.31, and for each carat added to the diamond, the price increases by an expected \$3,677.49. The 95% confidence interval for the Intercept is $(-279.8122, -220.8117)$ and that of the slope is $(3537.2549, 3817.7159)$; we are 95% confident that the true intercept and slope are in those ranges. As 0 is not in the confidence interval of the slope, this confirms a relationship between carats of a diamon and the price of the ring in which it is set.

The R^2 value for this regression model is 0.9841, meaning that the regression line accounts for about 98.41% of the variation of the data. Compare this to the R^2 value of the model including the two outliers, 0.9785, to see that our transformed model more accurately fits the data.

Conclusion

Using these models, we establish a clear correlation between the size of a diamond and the price of the ring featuring that diamond. We may use the equation $Price = -250.31 + 3677.49(carats)$ to estimate the price of a ring given the weight of its diamond in carats.

However, the existence of outliers belies the possibility that other factors influence the price of a ring, rather than only the size of the diamond: the more expensive ring whose data we omitted could have had a unique coloration of its diamond, or the gold making the band could have been wider, or more pure. The low-priced outlying ring may have had a defect in the diamond's clarity, or an unattractive band.

Nevertheless, we state confidently that diamond size affects ring price by a linear model.

Problem 2

Using the following code in R-Studio, we find the regression line for the data accounting for the number of datapoints in each case using weights:

```
weightLine <- lm(‘3rdQuartile‘~YearsOfExperience,data=academic,weights=academic$SampleSize_ni)
    # Make a linear regression with weights equal to the number of counts in each discrete case
summary(weightLine) # Asks R for a summary of the line
```

This produces the following output:

```
Call:
lm(formula = ‘3rdQuartile‘ ~ YearsOfExperience, data = academic,
    weights = academic$SampleSize_ni)

Weighted Residuals:
    Min     1Q Median     3Q    Max
    -67614 -40996   5138  51633  87400

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       104757.1     5754.7  18.204 8.52e-08 ***
YearsOfExperience   1173.3      337.1   3.481  0.00831 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57640 on 8 degrees of freedom
Multiple R-squared:  0.6023,Adjusted R-squared:  0.5526
F-statistic: 12.12 on 1 and 8 DF,  p-value: 0.008311
```

Using this, the model predicts that the 3rd Quartile salary of an academic statistician with 6 years of experience in 2005-2006 was $\$104,757.10 + 6 \times \$1,173.30 = \$111,796.90$.