Project for Wine Price

Patricia

I am interested in predicting the price of a bottle of 750ml wine. And I want to know the variables that affect the price. Hence, I collected data of 13 bottles of 750ml wine from LCBO official website.

Observations on 2 variales were collected for 13 bottles of wine.
The exlanatory variables are:
DRY: indicator variable for Sweetness Descriptor : 0 for dry, 1 for extra dry
YEAR: harvest year of the grapes, which the wine was made from

```
# 1)
wines = read.table(file="~/Desktop/wine.txt",header = T)
price = wines[, "price"]
year = wines[, "year"]
dry = wines[, "Dry"]
model = lm(price ~ dry + year, data = wines)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ dry + year, data = wines)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5712  -6.1803   0.9088   7.1197  11.2588
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12387.070   1943.953   6.372 8.12e-05 ***
## dry             8.227      4.911   1.675    0.125
## year           -6.138      0.965  -6.360 8.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.588 on 10 degrees of freedom
## Multiple R-squared:  0.8189, Adjusted R-squared:  0.7827
## F-statistic: 22.61 on 2 and 10 DF,  p-value: 0.0001949
```

```
# 2) check the model is a good fit
criteriaF = qf(.95, df1=3, df2=10)
```

To test H0: B1 = B2 = 0, the F statistics is greater than criteriaF And the p-value < 0.05, so we reject H0 => and conclude the model is overall significant. R^2 indicates it is a good fit.

```
# 3) analyse variables
```

The variable YEAR has the smallest p-value, hence it is the most significant variable among the other variable. The variable DRY has p-value > 0.05, we conclude it does not affect the price

```
# 4) 95% prediction interval
## Predict the price for a new type of DRY wine given YEAR 2015.
## We want to know the 95% prediction interval of the price of this wine.
summary(model)$coef[, "Estimate", drop=F]
```

```
##                 Estimate
```

```
## (Intercept) 12387.070505
## dry             8.227058
## year           -6.137980
```

```
B0 = 12387.0705
B1 = 8.227058
B2 = -6.137980
Bhat = (c(B0, B1, B2))
p = (c(1, 0, 2015))
price_hat = t(p) %*% Bhat
V = vcov(model)
rhat_square = 0.8189^2
se = sqrt(t(p) %*% V %*% p + rhat_square)
t = qt(0.975,10)
c(price_hat - t*se, price_hat + t*se)
```

```
## [1] 10.19611 27.88549
```

we are 95% confident that this interval (e.g.[10,28] dollars) covers the true value of this future response (e.g. the price of this new wine).

```
# 5) Check whether we can find a better model
## Idea: fit another model by excluding the most insignificant predictor
model2 = lm(price ~ year, data = wines)
summary(model2)
```

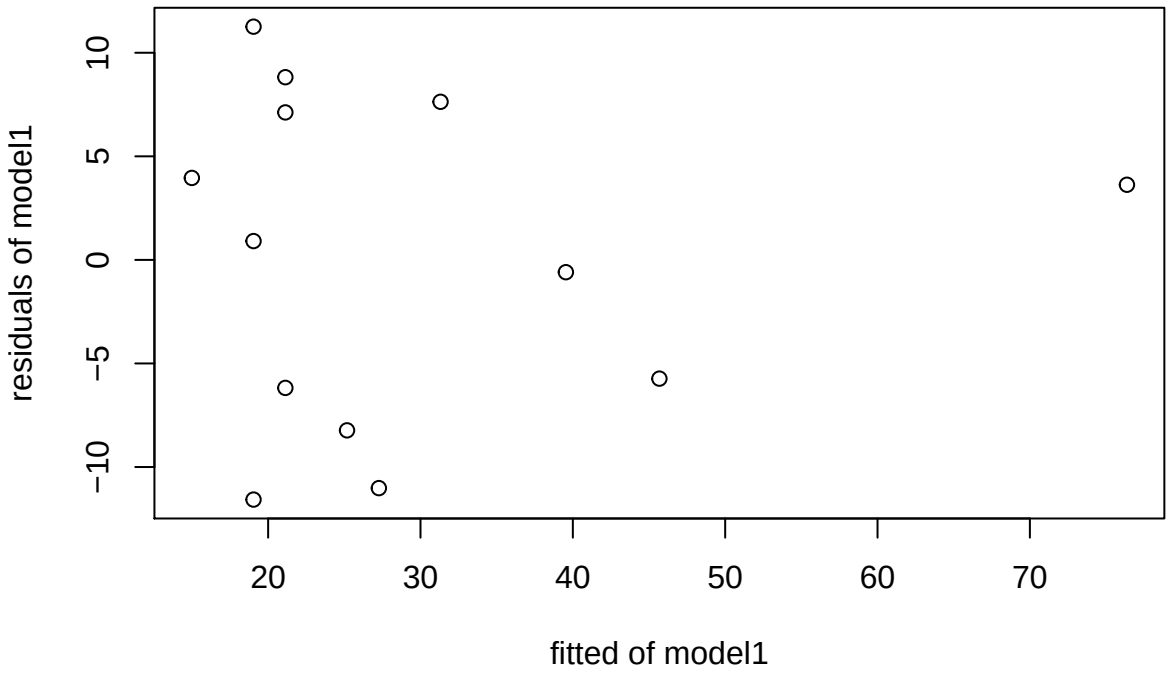
```
##
## Call:
## lm(formula = price ~ year, data = wines)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.526  -4.046   2.423   6.303  12.219
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12648.049   2090.738   6.050 8.31e-05 ***
## year           -6.265      1.038  -6.036 8.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.266 on 11 degrees of freedom
## Multiple R-squared:  0.7681, Adjusted R-squared:  0.747
## F-statistic: 36.43 on 1 and 11 DF,  p-value: 8.483e-05
```

Since the adjusted R squared of the first model is slightly higher, we prefer to use the first model for prediction.

```
# 6) Check final model by plots
```

- Plot residuals vs fitted values (=> model is adequate)

```
plot(fitted(model), residuals(model), xlab = "fitted of model1", ylab = "residuals of model1")
```


residuals of model1
10
5
0
−5
−10
20
30
40
50
60
70
fitted of model1