

## Stat 150 – Day 27 Multiple Regression

### Example: Pricing Bordeaux wines

(a) Open the `wine.mtw` worksheet from the Stat 150 Data files webpage. Reproduce Figure 1 from the essay, which is a scatterplot of the “price index” variable vs. vintage year. Does this relationship appear linear? *Hint:* You can add a regression fit and/or a smoother to help you see the overall relationship (right click on the graph and choose Add > Regression Fit or Add > Smoother).

(b) In this context, what are the implications for a point to fall above the line vs. below the line?

(c) Recall when a relationship is nonlinear, we can use a transformation to rescale the variable(s) and produce a linear relationship. Create the new variable  $\ln(\text{price})$  and reproduce Figure 2.

```
MTB> let c7=loge(c6)
MTB> plot c7*c1
```

What is the regression equation for these data? What does this line imply about the relationship between price and vintage year? What is the interpretation of the slope coefficient?

(d) Another way to see explore the relationship is to regress  $\ln(\text{price})$  on *age*. What is this regression equation? How do you interpret the slope coefficient?

(e) Another way to obtain this regression equation, as well as additional output, is to choose Stat > Regression > Regression and enter  $\ln(\text{price})$  in the response variable box and *age* in the Predictors box. Scroll up to find the regression equation and “R-Sq.” What percentage of the variation in the log of prices is explained by this regression model?

The key question is then: After accounting for age, can we explain more of the variability in the wine prices? The obvious place to start is adding additional variables into the regression model.

Recall, when using  $\ln(\text{price})$  as the response variable and  $\text{age}$  as the predictor variable, what was the generic representation of the simple functional relationship between the response and predictor variables? (*Let's assume a linear relationship*)

The best fitted line will be one that achieves the least sum of squared errors. Geometrically speaking, what does this imply?

Suppose we are interested in using both  $\text{age}$  and  $\text{Summer temperature}$  as predictor variables. Expanding on the idea from above, how would we write the generic representation of simple functional relationship between the response and predictor variables?

Geometrically speaking, what does this function represent? \_\_\_\_\_

How would we define the best fitting function of the above form? Geometrically speaking, what does this imply?

(f) Return to Stat > Regression > Regression and place both *age* and *Summer temperature* in the Predictors box. Report the resulting regression equation and  $R^2$  value. Does adding this variable to the model appear to be helpful in explaining wine prices? Explain.

(g) What is the sign (positive or negative) of the coefficient of summer temperature? Is this what you would have expected based on Figure 3?

(h) Did the coefficient of *age* change? What does this mean?

(i) Now add *Harvest rain* and *Winter rain* to the model as well, in order to reproduce the equation on p.416. What is the  $R^2$  value? How do you interpret this value?

(j) Minitab automatically flags any observations with unusually large residuals. Which vintage year did it flag? Is the residual positive or negative? What does this indicate about that vintage year?