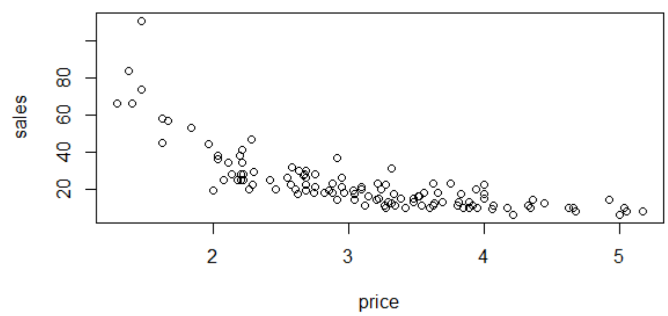
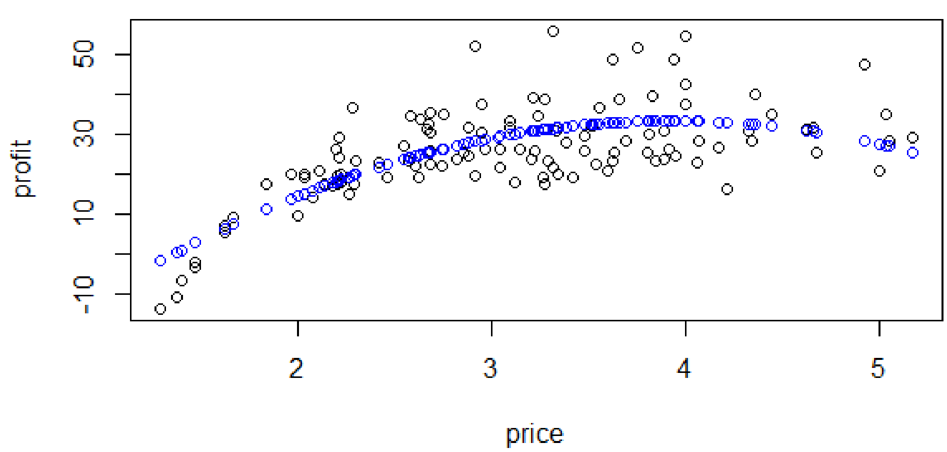
Michelle Lee
9:30 A.M. - 11:00 A.M.

1/29 Class Notes

1. **Milkprice.R**
    a. We started class by importing 'milk.csv' and writing our own script to find the best price to pay for milk (Professor Scott has now uploaded his script 'milkprice.R' on his website)
        i. Needed to find a model that optimized profit given this data

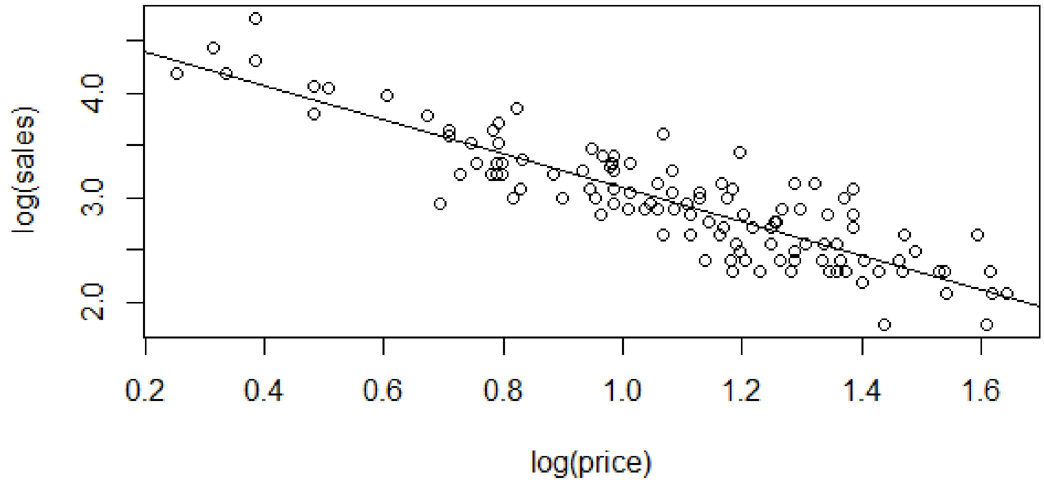


    b. **Method 1:** plot profit vs price and fit a quadratic model
        i. Profit = revenue (# sold) - cost (# sold)
        Code: profit = milk$sales * (milk$price - 1.50)
            1. $1.50 was assigned as the cost per unit
        ii. The data appears quadratic so use a linear model to fit a quadratic equation
        Code: lm2 = lm(profit ~ price + I(price^2), data=milk)
        iii. Plot the model. The blue points should give you an idea of the average profit given a price, showing that the optimal price is around 4.



c. **Method 2:** Demand curve

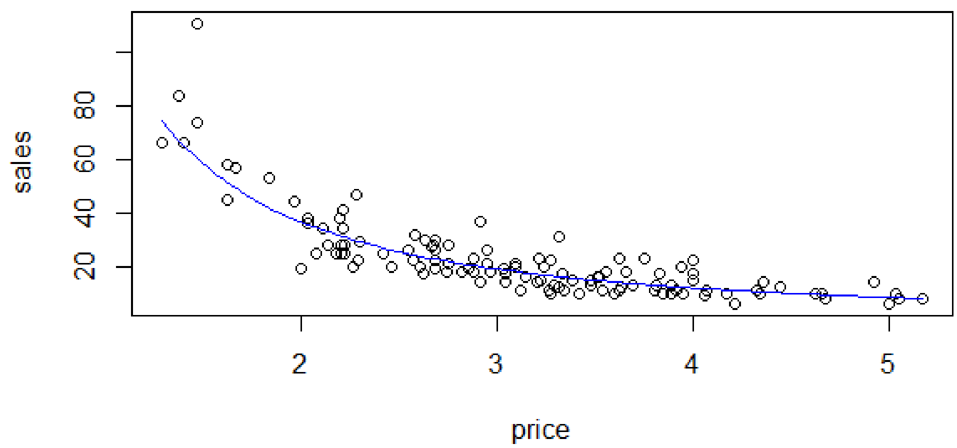
i. Note that when you plot log of sales vs. log of price, the points appear linear. Use a linear model to find a line of best fit.



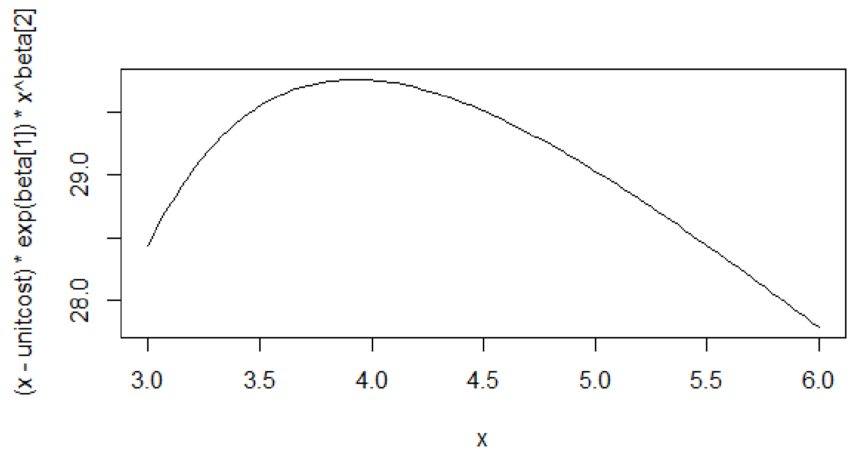
ii. The linear relationship of the logs of both variables implies a Power Law relationship. Based on the Power Law, the intercept and slope (beta[1] and beta[2]) can be plugged into the Power Law equation as K and β to find an equation for profit

1. $y = Kx^{\beta}$
2. Code: curve(exp(beta[1])*x^beta[2], add=TRUE, col='blue')

   This gives us a curve that fits the data



iii. We can use R to plot the profit curve as seen below. Here it shows the profit for a price range of 3 to 6.
Code: curve((x-unitcost)*exp(beta[1])*x^beta[2], from=3, to=6)



2. Predictions

The rest of class focused on ways to predict where a data point would lie given the data.

a. Given the equation for a line: $y = \beta_0 + \beta_1 x_i + e_i$
The equation for a data point (x*) would be: $y^* = \beta_0 + \beta_1 x^* \pm a * s_e$

   i. s_e = standard deviation of residuals
   a = a number, e.g. 1, 2, ... (some multiple)

   ii. $a * s_e$ = measure of variation for unpredictable piece

   iii. Review: The Empirical Rule states that when a=1 around 68% of the data is covered. a=2 covers around 95% of the data.

b. Coverage Intervals

   i. The spread of original data points is greater than the spread of residuals, which is why the residuals can be useful in determining a coverage interval to predict data.

   ii. When using a histogram of data in R, a coverage interval starts at either side and goes in to determine where the interval begins and ends. A central symmetric coverage interval starts at the center and goes out in both directions to cover a certain percentage of data.

   iii. When comparing a histogram of data and a histogram of residuals of groupwise model (mean of each group), the histogram of residuals is centered around 0 and the 50% coverage interval has a much smaller width.

1. The histogram of residuals has reduced uncertainty so looking at a data's residuals can provide a more accurate predictor.

iv. The coverage interval of the residuals provides a prediction interval.

   1. To predict using linear data, plot the linear model.
   2. Then, plot the residuals and find the standard deviation of the residuals.
      a. Code: sd(resid(model1))
   3. We can now plot the zones plus or minus a standard deviation on the linear model to give us a ~70%chance of covering the point.

c. Miscellaneous Terms

   i. s_e = standard deviation of residuals

   ii. s_y = standard deviation of original data

   iii. $R^2 = 1 - (\frac{s_e}{s_y})^2$

      1. The amount of variation in y that can be predicted by x
      2. The higher the R^2, the better the fit.