MLDS 401 (Fall 2023): Lab 1 Solution

2.3 (Weighted least squares): Minimize Q w.r.t. β by setting

$$\frac{dQ}{d\beta} = \sum_{i=1}^{n} \frac{d}{d\beta} w_i(y_i - \beta x_i)^2 = -2\sum_{i=1}^{n} w_i x_i (y_i - \beta x_i) = -2\left[\sum_{i=1}^{n} w_i x_i y_i - \beta \sum_{i=1}^{n} w_i x_i^2\right] = 0,$$

whose solution is

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} w_i x_i y_i}{\sum_{i=1}^{n} w_i x_i^2}.$$

2.10 (Price elasticities of steaks):

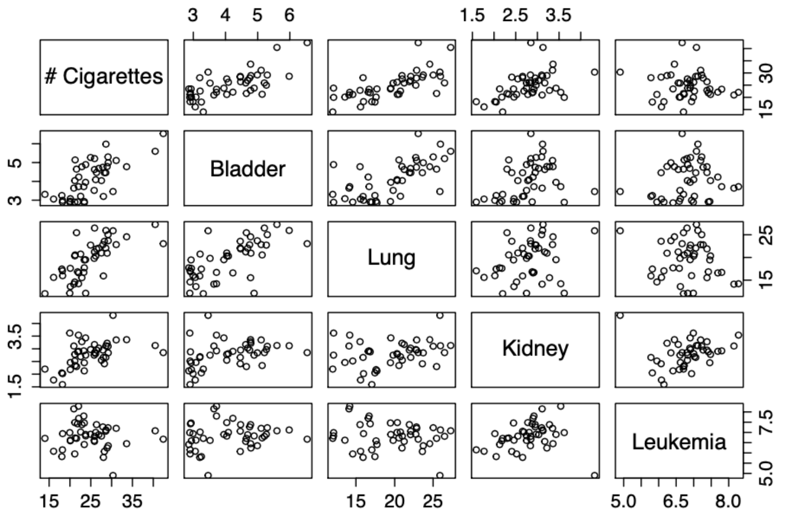
a. Price elasticities are given by the slope coefficients in the regression of log(Quantity) on log(Price). They are as follows:

   Chuck: -1.3687, Porter House: -2.6565, Rib Eye: -1.4460.

   All three coefficients are negative indicating that demand drops as price increases. The negative price effect is most pronounced for porter house steak (most expensive) and least pronounced for chuck steak (least expensive).

b. If price is increased by 10% then chuck demand will drop by 13.69%, the porter house demand by 26.57% and the rib eye demand by 14.46%.

2.11 (Smoking versus cancer): The matrix plot of cigarettes smoked and the four types of cancer deaths is shown below.



The correlation matrix is as shown below. We see that bladder and lung cancer deaths are most highly correlated with cigarettes smoked. Kidney cancer deaths are modestly correlated while leukemia deaths have a very small (nonsignificant) negative correlation.

```
            CIG.SMOKED   BLADDER       LUNG    KIDNEY    LEUKEMIA
CIG.SMOKED  1.00000000 0.7036219  0.6974025 0.4873896 -0.06848123
BLADDER     0.70362186 1.0000000  0.6585011 0.3588140  0.16215663
LUNG        0.69740250 0.6585011  1.0000000 0.2827431 -0.15158448
KIDNEY      0.48738962 0.3588140  0.2827431 1.0000000  0.18871294
LEUKEMIA   -0.06848123 0.1621566 -0.1515845 0.1887129  1.00000000
```

The t-statistics for testing the significance of the correlation between each type of cancer deaths and smoking using the formula $t = r\sqrt{n-2}/\sqrt{1-r^2}$ where $n = 44$ are as follows.
Bladder: $t = 6.417$, Lung: $t = 5.505$, Kidney: $t = 3.617$, Leukemia: $t = -0.445$.
Clearly, the first three are highly significant, but the correlation with leukemia deaths is non-significant.

Question 3

(a) $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = -0.215715$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4.06153 + 5.0394515 = 9.101$. Then our line is $\hat{y} = 9.101 - 0.215715 \times x$.

(b) $9.101 - 0.215715 \times 25 = 3.708125$

(c) $9.101 - 0.215715 \times 45 = -0.606175$. Our maximum x value in our sample is 40.4, so 45 is outside our calculation and because of that our solution is not trustworthy.

(d) $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \dfrac{-341.959231}{(8.790379)(39.81495659)}$ and then $R^2 = r^2 = 0.9546$.

We can explain 95% of the variation in the data with the known values of x.

(e) Our confidence interval for the slope has the form $\hat{\beta}_1 \pm t_{\alpha/2,n-2} s_{\hat{\beta}_1} = -0.2157157 \pm 3.105807 * \frac{0.5644611}{\sqrt{1585.231}} = -0.2157157 \pm 0.044031$ Therefore our CI is $[-0.259747, -0.171684]$

(f) Our confidence interval has the form $\hat{y} \pm t_{\alpha/2,n-2} . s_{\hat{Y}}$. In this question, $\hat{y} = 9.101 - 0.215715 \times 25 = 3.708125$ and $s_{\hat{Y}} = 0.5644611 \sqrt{1/13 + \dfrac{(25 - 23.36154)^2}{1585.231}} = 0.158267$ which gives us the CI $3.708125 \pm 2.200985 * 0.158267 = [3.359784.056468]$.

(g) Our prediction interval has the form $\hat{y} \pm t_{\alpha/2,n-2} . \sqrt{s^2 + s_{\hat{Y}}^2} = 3.708125 \pm 2.200985 * 0.5862292 = [2.41784, 4.9984]$.