HW 0324

Question 1.

Let $K=2$, show that (b1, b2) in p. 29 of slides in Ch 5 reduces to the formula of (b1,b2) in (2.7) - (2.8):

$$b_2 = \frac{\sum_1 (x_i - \bar{x})(y_i - \bar{y})}{\sum_1 (x_i - \bar{x}_2)} \ (2.7) \qquad b_1 = \bar{y} - b_2\bar{x} \ (2.8) \qquad \beta = (X'X)^{-1}X'y \quad \text{where:}$$

$$\beta = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \Rightarrow X'X = \begin{bmatrix} \sum_1 1 & \sum_1 x_i \\ \sum_1 x_i & \sum_1 x_i^2 \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum_1 x_i & \sum_1 x_i^2 \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \Rightarrow X'y = \begin{bmatrix} \sum_1 y_i \\ \sum_1 x_i y_i \end{bmatrix}$$

$$S_x = \sum_1 x_i$$
$$S_{xx} = \sum_1 x_i^2$$
$$\Rightarrow (X'X)^{-1} = \frac{1}{nS_{xx} - S_x^2} \begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix}$$

$$(X'X)^{-1}X'y = \hat{\beta} = \frac{1}{nS_{xx} - S_x^2} \begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix} \begin{bmatrix} \sum_1 y_i \\ \sum_1 x_i y_i \end{bmatrix}$$

$$b_1 = \frac{1}{D}\left(S_{xx}\sum_1 y_i - S_x \sum_1 x_i y_i\right) \quad b_2 = \frac{1}{D}\left(-n\sum_1 x_i y_i + S_x \sum_1 y_i\right) \quad \text{where } D = nS_{xx} - S_x^2$$

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum_1 y_i$$

$$\sum_1 (x_i - \bar{x})^2 = \sum_1 x_i^2 - n\bar{x}^2 = S_{xx} - \frac{S_x^2}{n} \Rightarrow b_2 = \frac{\sum_1 x_i y_i - n\bar{x}\bar{y}}{\sum_1 x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_1 = \bar{y} - b_2\bar{x}$$

Question 2.

Let $K=2$, show that cov(b1, b2) in p. 30 of slides in Ch 5 reduces to the formula of in (2.14) - (2.16).

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1} \text{ where } \hat{\beta} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \text{ need to show that : } Var(b_2) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right] =$$

$$= - \frac{\sigma^2 \bar{x}}{\sum_i (x_i - \bar{x})^2}$$

$$y_i = b_1 + b_2 x_i + u_i$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \Rightarrow X'X = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} = \begin{bmatrix} n & S_x \\ S_x & S_{xx} \end{bmatrix} \text{ so we find } (X'X)^{-1} = \frac{1}{nS_{xx} - S_x^2} \begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix}$$

$$\text{hence } var(\hat{\beta}) = \sigma^2 * (X'X)^{-1} = \frac{\sigma^2}{nS_{xx} - S_x^2} \begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix}$$

let:

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

$$\sum_i (x_i - \bar{x})^2 = S_{xx} - \frac{S_x^2}{n} \Rightarrow \sum_i (x_i - \bar{x})^2 = TSS_x$$

$$nS_{xx} - S_x^2 = n\sum_i (x_i - \bar{x})^2 \quad \text{from } Var(\hat{\beta}) = \frac{\sigma^2}{n * \sum (x_i - \bar{x})^2} \begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix} \quad \text{we get: } Var(b_2) = \frac{\sigma^2 * n}{n * \sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

$$Var(b_1) = \frac{\sigma^2 * \sum x_i^2}{n \sum_i (x_i - \bar{x})^2}$$

$$\sum_i x_i^2 = \sum (x_i - \bar{x})^2 + n\bar{x}^2 \text{ so } Var(b_1) = \frac{\sigma^2}{n * \sum (x_i - \bar{x})^2} \left[\sum_i (x_i - \bar{x})^2 + n\bar{x}^2 \right] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right]$$

$$Cov(b_1, b_2) = - \frac{\sigma^2 \sum_i x_i}{n \sum_i (x_i - \bar{x}^2)} = - \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Question 3.

TABLE 5.6				
Output for Exercise 5.3				
Dependent Variable: *WALC*				
Included observations: 1200				
Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	1.4515	2.2019	0.659	0.5099
ln(*TOTEXP*)	2.7648	0.484	5.7103	0.0000
NK	− 1.459	0.3695	−3.9376	0.0001
AGE	−0.1503	0.0235	−6.4019	0.0000
R-squared	0.058	Mean dependent var		6.19434
S.E. of regression	6.27	S.D. dependent var		6.39547
Sum squared resid	46221.62			

Part a & b.

A 1% increase in total household expenditure (TOTEXP) leads to an approximate 0.0276 percentage point increase in the budget share spent on alcohol, suggesting that wealthier households spend relatively more on alcohol. Each additional child in the household reduces the alcohol budget share by about 1.454 percentage points, indicating that families with more children allocate less to alcohol. Additionally, for every extra year in the age of the household head, the budget share spent on alcohol decreases by 0.15 percentage points. This implies that older household heads tend to spend less on alcohol compared to younger ones. All these effects are statistically significant and highlight how income, family size, and age influence alcohol spending.

Part c.

Given:

$\beta_4 = -0.1503$

$SE(\beta_4) = 0.0235$

$df \approx 1200 - 4 = 1196 \rightarrow t_{0.025} \approx 1.96$

$CI = \hat{\beta}4 \pm t0.025, df \times SE(\hat{\beta}_4) = -0.1503 \pm 1.96 \times 0.0235 =$

$-0.1503 \pm 0.0461 = [-0.1964, -0.1042]$

This 95% confidence interval means we are 95% confident that the true effect of age on alcohol budget share lies between -0.1964 and -0.1042 percentage points per year. Since the interval is entirely negative, it confirms that as age increases, the budget share spent on alcohol decreases. The fact that zero is not in the interval supports that the effect is statistically significant.

Part d.

At the 5% significance level, the coefficients for ln(TOTEXP), NK, and AGE are all statistically significant because their p-values are less than 0.05. This means these variables have a real and measurable effect on the budget share spent on alcohol (WALC). The coefficient for ln(TOTEXP) indicates that higher total expenditure increases alcohol spending, while NK and AGE both reduce it. The intercept, however, is not significant since its p-value is greater than 0.05, meaning we cannot confidently interpret its value. Overall, only the variables ln(TOTEXP), NK, and AGE are important predictors in the model at the 5% level.

Part e.

Hypothesis:

$H_0: \beta_3 = -2$

$H_1: \beta_3 \neq -2$

Given:

β_3 estimate $= -1.454$

Std. error $= 0.3695$

Significance level $= 5\%$ $(\alpha = 0.05)$

$$t = \frac{\hat{\beta}_3 - (-2)}{SE(\hat{\beta}_3)} = \frac{-1.454 - (-2)}{0.3695} = 1.477$$

$df \approx 1200$ so $t_{0.025} \approx 1.96$

$|t| = 1.477 < 1.96 \Rightarrow$ we fail to reject the null hypothesis

Question 23.

Part a

I would expect β_2 (QUANT) to be negative because larger cocaine sales likely come with quantity discounts, reducing the price per gram. β_3 (QUAL) is expected to be positive, as higher purity cocaine should command a higher price due to its better quality. β_4 (TREND) may be positive if prices increased over time due to inflation or changes in enforcement, though it could also be negative if market supply grew. Overall, quantity should lower price, quality should raise it, and the time trend's effect depends on broader market factors.

Part b.

```
Call:
lm(formula = price ~ quant + qual + trend, data = cocaine)

Residuals:
    Min      1Q  Median      3Q     Max 
-43.479 -12.014  -3.743  13.969  43.753 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 90.84669    8.58025  10.588 1.39e-14 ***
quant       -0.05997    0.01018  -5.892 2.85e-07 ***
qual         0.11621    0.20326   0.572   0.5700    
trend       -2.35458    1.38612  -1.699   0.0954 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.06 on 52 degrees of freedom
Multiple R-squared:  0.5097,    Adjusted R-squared:  0.4814 
F-statistic: 18.02 on 3 and 52 DF,  p-value: 3.806e-08
```

The regression results show that quantity has a significant negative effect on price, with a coefficient of -0.05997. This confirms that larger sales come with quantity discounts, which aligns with the initial assumption.. Quality has a positive coefficient of 0.11621, which aligns with the assumption that higher purity would increase price, but this effect is not statistically significant, suggesting purity does not strongly influence price in this sample. Trend has a negative coefficient of -2.35, indicating that prices decreased over time, which contrasts with the expectation of a possible increase, but could be explained by increased market supply or competitive pressures.

Part c

The R-squared value of 0.5097 indicates that about 51% of the variation in cocaine price is explained by differences in quantity, quality, and time. This suggests that these three factors jointly have a strong influence on the price per gram. The remaining 49% of the variation is due to other factors not included in the model or random fluctuations. While not a perfect fit, the model captures more than half of the price variation, which is substantial. Overall, quantity, quality, and time are important determinants of cocaine pricing in this dataset.

Part d

```
> cat("b2 (quant):", round(b2, 5), "\n")
b2 (quant): -0.05997
> cat("Standard Error:", round(se_b2, 5), "\n")
Standard Error: 0.01018
> cat("t-value:", round(t_value, 3), "\n")
t-value: -5.892
> cat("p-value:", round(p_value, 5), "\n")
p-value: 0
```

The coefficient for quantity (β_2) is estimated at -0.05997, with a standard error of 0.01018 and a highly significant t-value of -5.892. The corresponding p-value is effectively zero, indicating very strong evidence against the null hypothesis. At the 5% significance level, we reject the null hypothesis that quantity has no effect or increases price. This confirms that larger quantities are associated with lower prices per gram, supporting the idea that sellers accept lower prices to reduce the risk of making many small sales. Overall, the results provide strong evidence for a quantity discount effect in the cocaine market.

Part e

```
> cat("b3 (qual):", round(b3, 5), "\n")
b3 (qual): 0.11621
> cat("Standard Error:", round(se_b3, 5), "\n")
Standard Error: 0.20326
> cat("t-value:", round(t_value, 3), "\n")
t-value: 0.572
> cat("p-value:", round(p_value, 5), "\n")
p-value: 0.285
```

The estimated coefficient for quality (β_3) is 0.11621, with a standard error of 0.20326 and a t-value of 0.572. The resulting p-value is 0.285, which is much greater than 0.05. At the 5% significance level, we fail to reject the null hypothesis that quality has no influence on price. This means there is no statistically significant evidence that a premium is paid for higher-quality cocaine in this dataset. Although the coefficient is positive, suggesting a possible premium, the effect is not strong enough to be considered meaningful based on the data.

Part f

Based on the trend coefficient, cocaine's price has experienced an average annual decreased of about 2.35 dollars per annum. This indicates that prices steadily declined from 1984 to 1991. One possible reason for this decline could be an increase in supply, leading to more competition and lower prices. Additionally, shifts in market dynamics, such as more efficient distribution or reduced demand, may have contributed to falling prices. Other factors like changes in purity levels or enforcement pressures could also influence this downward trend