Lab 2: Linear Regression I

ACTL3142 and ACTL5110

Questions

Conceptual Questions

Simple linear regression questions

1. ⋆ Prove that the Least Squared coefficient estimates (LSE) for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are:

   $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$

   $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{S_{xy}}{S_{xx}}$

   Solution

2. Prove that the estimates in Q1 are unbiased.

   Solution

3. Prove that the MLE estimates of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are equal to the ones given by LSE (from Q1).

   Solution

4. Prove SST=SSE+SSM

   Solution

5. Express SSM in terms of a) β_1 and b) β_1^2

   Solution

6. Prove the following variance formulas:

$$\begin{aligned}\mathbb{V}\left(\widehat{\beta}_0|X\right) &= \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\\ \mathbb{V}\left(\widehat{\beta}_1|X\right) &= \frac{\sigma^2}{S_{xx}}\\ \mathrm{Cov}\left(\widehat{\beta}_0,\widehat{\beta}_1|X\right) &= -\frac{\overline{x}\sigma^2}{S_{xx}}\end{aligned}$$

   Solution

7. Prove $\mathbb{V}(\widehat{y}_0|X) = \left(\dfrac{1}{n}+\frac{(\overline{x}-x_0)^2}{S_{xx}}\right)\sigma^2$, where $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$.

   Remember that (x_0, y_0) is a new (but fixed) observation, i.e. not in the training set used to find $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

   Solution

8. Prove:

   - $\mathbb{E}[Y_0 - \widehat{y}_0|X] = 0$
   - $\mathbb{V}(Y_0 - \widehat{y}_0|X) = \sigma^2\left(1+\frac{1}{n}+\frac{(\overline{x}-x_i)^2}{S_{xx}}\right)$

   Solution

9. Forensic scientists use various methods for determining the likely time of death from post-mortem examination of human bodies. A recently suggested objective method uses the concentration of a compound (3-methoxytyramine or 3-MT) in a particular part of the brain. In a study of the relationship between post-mortem interval and the concentration of 3-MT, samples of the approximate part of the brain were taken from coroners cases for which the time of death had been determined form eye-witness accounts. The intervals (x; in hours) and concentrations (y; in parts per million) for 18 individuals who were found to have died from organic heart disease are given in the following table. For the last two individuals (numbered 17 and 18 in the table) there was no eye-witness testimony directly available, and the time of death was established on the basis of other evidence including knowledge if the individuals' activities.

Observation number	Interval (x)	Concentration (y)
1	5.5	3.26
2	6.0	2.67
3	6.5	2.82
4	7.0	2.80
5	8.0	3.29
6	12.0	2.28

7	12.0	2.34
8	14.0	2.18
9	15.0	1.97
10	15.5	2.56
11	17.5	2.09
12	17.5	2.69
13	20.0	2.56
14	21.0	3.17
15	25.5	2.18
16	26.0	1.94
17	48.0	1.57
18	60.0	0.61

$\sum x = 337$, $\sum x^2 = 9854.5$, $\sum y = 42.98$, $\sum y^2 = 109.7936$, $\sum xy = 672.8$

In this investigation you are required to explore the relationship between concentration (regarded the responds/dependent variable) and interval (regard as the explanatory/independent variable).

a. Construct a scatterplot of the data. Comment on any interesting features of the data and discuss briefly whether linear regression is appropriate to model the relationship between concentration of 3-MT and the interval from death.

b. Calculate the correlation coefficient for the data, and use it to test the null hypothesis that the population correlation coefficient is equal to zero.

c. Calculate the equation of the least-squares fitted regression line and use it to estimate the concentration of 3-MT:

   i. after 1 day and

   ii. after 2 days.

   Comment briefly on the reliability of these estimates.

d. Calculate a 99% confidence interval for the slope of the regression line. Using this confidence interval, test the hypothesis that the slope of the regression line is equal to zero. Comment on your answer in relation to the answer given in part (2) above.

Solution

10. $\star$ A university wishes to analyse the performance of its students on a particular degree course. It records the scores obtained by a sample of 12 students at the entry to the course, and the scores obtained in their final examinations by the same students. The results are as follows:

Student	A	B	C	D	E	F	G	H	I	J	K	L
Entrance exam score x (%)	86	53	71	60	62	79	66	84	90	55	58	72
Final paper score y (%)	75	60	74	68	70	75	78	90	85	60	62	70

$\sum x = 836$, $\sum y = 867$, $\sum x^2 = 60,016$, $\sum y^2 = 63,603$, $\sum(x-\overline{x})(y-\overline{y}) = 1,122$.

a. Calculate the fitted linear regression equation of y on x.

b. Assuming the full normal model, calculate an estimate of the error variance σ^2 and obtain a 90% confidence interval for σ^2.

c. By considering the slope parameter, formally test whether the data is positively correlated.

d. Find a 95% confidence interval for the mean finals paper score corresponding to an individual entrance score of 53.

e. Test whether this data come form a population with a correlation coefficient equal to 0.75.

f. Calculate the proportion of variance explained by the model. Hence, comment on the fit of the model.

Solution

11. $\star$ Complete the following ANOVA table for a simple linear regression with 60 observations:

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio
Regression	______	______	______	______
Error	______	______	8.2	
Total	______	639.5		

Solution

12. $\star$ Suppose you are interested in relating the accounting variable EPS (earnings per share) to the market variable STKPRICE (stock price). Then, a regression equation was fitted using STKPRICE as the response variable with EPS as the regressor variable. Following is the computer output from your fitted regression. You are also given that: $\overline{x} = 2.338$, $\overline{y} = 40.21$, $s_x = 2.004$, and $s_y = 21.56$. (Note that: $s_x^2 = \frac{S_{xx}}{n-1}$ and $s_y^2 = \frac{S_{yy}}{n-1}$)

```
Regression Analysis
The regression equation is
STKPRICE = 25.044 + 7.445 EPS

Predictor                       Coef      SE Coef    T       p
Constant                        25.044    3.326      7.53    0.000
EPS                             7.445     1.144      6.51    0.000

Analysis of Variance
SOURCE                          DF        SS         MS      F       p
Regression                      1         10475      10475   42.35   0.000
Error                           46        11377      247
Total                           47        21851
```

a. Calculate the correlation coefficient of EPS and STKPRICE.

b. Estimate the STKPRICE given an EPS of \$2. Provide a 95% confidence interval of your estimate.

c. Provide a 95% confidence interval for the slope coefficient β.

d. Compute s and R^2.

e. Describe how you would check if the errors have constant variance.

f. Perform a test of the significance of EPS in predicting STKPRICE at a level of significance of 5%.

g. Test the hypothesis $H_0 : \beta = 24$ against $H_a : \beta > 24$ at a level of significance of 5%.

Solution

13. (Modified from an Institute of Actuaries exam problem) An insurance company issues house buildings policies for houses of similar size in four different post-code regions A, B, C, and D. An insurance agent takes independent random samples of 10 house buildings policies for houses of similar size in each of the four regions. The annual premiums (in dollars) were as follows:

Region A :	229	241	270	256	241	247	261	243	272	219
	$(\sum x = 2,479, \quad \sum x^2 = 617,163)$									
Region B :	261	269	284	268	249	255	237	270	269	257
	$(\sum x = 2,619, \quad \sum x^2 = 687,467)$									
Region C :	253	247	244	245	221	229	245	256	232	269
	$(\sum x = 2,441, \quad \sum x^2 = 597,607)$									
Region D :	279	268	290	245	281	262	287	257	262	246
	$(\sum x = 2,677, \quad \sum x^2 = 718,973)$									

Perform a one-way analysis of variance at the 5% level to compare the premiums for all four regions. In order words, test whether the mean of each 4 region are significantly different to each other. State briefly the assumptions required to perform this analysis of variance.

Solution

14. (Past Institute Exam) As part of an investigation into health service funding a working party was concerned with the issue of whether mortality could be used to predict sickness rates. Data on standardised mortality rates and standardised sickness rates collected for a sample of 10 regions and are shown in the table below:

Region	Mortality rate m (per 100,000)	Sickness rate s (per 100,000)
1	125.2	206.8
2	119.3	213.8
3	125.3	197.2
4	111.7	200.6
5	117.3	189.1
6	100.7	183.6
7	108.8	181.2
8	102.0	168.2
9	104.7	165.2
10	121.1	228.5

Data summaries: $\sum m = 1136.1$, $\sum m^2 = 129,853.03$, $\sum s = 1934.2$, $\sum s^2 = 377,700.62$, and $\sum ms = 221,022.58$.

a. Calculate the correlation coefficient between the mortality rates and the sickness rates and determine the probability-value for testing whether the underlaying correlation coefficient is zero against the alternative that it is positive.

b. Noting the issue under investigation, draw an appropriate scatterplot for these data and comment on the relationship between the two rates.

c. Determine the fitted linear regression of sickness rate on mortality rate and test whether the underlaying slope coefficient can be considered to be as large as 2.0.

d. For a region with mortality rate 115.0, estimate the expected sickness rate and calculate 95% confidence limits for this expected rate.

Solution

15. (Past institute Exam) Consider the following data, which comprise of four groups sizes (y), each comprising four observations. In scenario I, information is also given on the sum assured under the policy concerned - the sum assured is the same for all four policies in a group. In scenario II, we regard the policies in the different groups as having been issued by four different companies - the policies in a group are all issued the same company.

All monetary amounts are in units of £10,000. Summaries of the claim sizes in each group are given in a second table.

Group	1	2	3	4
Claim sizes y	0.11 0.46	0.52 1.43	1.48 2.05	1.52 2.36
	0.71 1.45	1.84 2.47	2.38 3.31	2.95 4.08
I: sum assured x	1	2	3	4
II: Company	A	B	C	D

Summaries of claim sizes:

Group	1	2	3	4
$\sum y$	2.73	6.26	9.22	10.91
$\sum y^2$	2.8303	11.8018	23.0134	33.2289

a. In scenario I, suppose we adopt the linear regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

where Y_i is the i^{th} claim size and x_i is the corresponding sum assured, $i = 1, \ldots, 16$.

i. Calculate the total sum of squares and its partition into the regression (model) sum of squares and the residual (error) sum of squares.

ii. Fit the model and calculate the fitted values for the first claim size of group 1 (namely 0.11) and the last claim size of group 4 (namely 4.08).

iii. Consider a test of the hypothesis $H_0 : \beta = 0$ against a two-sided alterative. By preforming appropriate calculations, assess the strength of the evidence against this “no linear relationship” hypothesis.

b. In scenario II, suppose we adopt the analysis of variance model

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

where Y_{ij} is the j^{th} claim size for company i and τ_i is the i^{th} company effect, $i = 1, 2, 3, 4$ and $j = A, B, C, D$.

i. Calculate the partition of the total sum of squared into the "between companies" (model) sum of squares and the "within companies" (residual/error) sum of squares.

ii. Fit the model.

iii. Calculate the fitted values for the first claim size of group 1 and the last claim size of group 4.

iv. Consider a test of hypothesis $H_0 : \tau_i = 0$, $i = A, B, C, D$ against a general alternative. By preforming appropriate calculations, assess the strength of the evidence against this "no company effects" hypothesis.

Solution

Multiple linear regression questions

1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

   Solution

2. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, $\beta_5 = -10$.

   a. Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

      ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

      iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

   b. Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

   c. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Solution

3. $\star$ I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

   a. Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   b. Answer (a) using test rather than training RSS.

   c. Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   d. Answer (c) using test rather than training RSS.

   Solution

4. a. Write down the design matrix for the simple linear regression model.

   b. Write out the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ for the simple linear regression model.

   c. Write out the vector $\boldsymbol{X}^\top \boldsymbol{y}$ for the simple linear regression model.

   d. Write out the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ for the simple linear regression model.

   e. Calculate $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ using your results above.

   Where $\boldsymbol{y}$ is the vector of the response variable and $\widehat{\boldsymbol{\beta}}$ is the vector of coefficients.

   Solution

5. $\star$ The following model was fitted to a sample of supermarkets in order to explain their profit levels:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

   where

   - $y =$ profits, in thousands of dollars
   - $x_1 =$ food sales, in tens of thousands of dollars
   - $x_2 =$ nonfood sales, in tens of thousands of dollars, and
   - $x_3 =$ store size, in thousands of square feet.

The estimated regression coefficients are given below:

$$\widehat{\beta}_1 = 0.027 \text{ and } \widehat{\beta}_2 = -0.097 \text{ and } \widehat{\beta}_3 = 0.525.$$

Which of the following is TRUE?

a. A dollar increase in food sales increases profits by 2.7 cents.

b. A 2.7 cent increase in food sales increases profits by a dollar.

c. A 9.7 cent increase in nonfood sales decreases profits by a dollar.

d. A dollar decrease in nonfood sales increases profits by 9.7 cents.

e. An increase in store size by one square foot increases profits by 52.5 cents.

Solution

6. ⋆ In a regression model of three explanatory variables, twenty-five observations were used to calculate the least squares estimates. The total sum of squares and regression sum of squares were found to be 666.98 and 610.48, respectively. Calculate the adjusted coefficient of determination (i.e adjusted R^2).

   a. 89.0%

   b. 89.4%

   c. 89.9%

   d. 90.3%

   e. 90.5%

   Solution

7. ⋆ In a multiple regression model given by:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} + \varepsilon,$$

   which of the following gives a correct expression for the coefficient of determination (i.e R^2)?

   I. $\frac{\text{SSM}}{\text{SST}}$

   II. $\frac{\text{SST}-\text{SSE}}{\text{SST}}$

   III. $\frac{\text{SSM}}{\text{SSE}}$

   Options:

   a. I only

b. II only

c. III only

d. I and II only

e. I and III only

Solution

8. The ANOVA table output from a multiple regression model is given below:

Source	D.F.	SS	MS	F-Ratio	Prob($> F$)
Regression	5	13326.1	2665.2	13.13	0.000
Error	42	8525.3	203.0		
Total	47	21851.4			

Compute the adjusted coefficient of determination (i.e adjusted R^2).

a. 52%

b. 56%

c. 61%

d. 63%

e. 68%

Solution

9. $\star$ You have information on 62 purchases of Ford automobiles. In particular, you have the amount paid for the car y in hundreds of dollars, the annual income of the individuals x_1 in hundreds of dollars, the sex of the purchaser (x_2, 1=male and 0=female) and whether or not the purchaser graduated from college (x_3, 1=yes, 0=no). After examining the data and other information available, you decide to use the regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

You are given that:

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \begin{bmatrix} 0.109564 & -0.000115 & -0.035300 & -0.026804 \\ -0.000115 & 0.000001 & -0.000115 & -0.000091 \\ -0.035300 & -0.000115 & 0.102446 & 0.023971 \\ -0.026804 & -0.000091 & 0.023971 & 0.083184 \end{bmatrix}$$

and the mean square error for the model is $s^2 = 30106$. Calculate SE($\widehat{\beta}_2$).

a. 0.17

b. 17.78

c. 50.04

d. 55.54

e. 57.43

Solution

10. Suppose in addition to the information in question 9., you are given:

$$\boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 9\,558 \\ 4\,880\,937 \\ 7\,396 \\ 6\,552 \end{bmatrix}.$$

Calculate the expected difference in the amount spent to purchase a car between a person who graduated from college and another one who did not.

Possible answers:

a. 233.5
b. 1,604.3
c. 2,195.3
d. 4,920.6
e. 6,472.1

Solution

11. ⋆ A regression model of y on four independent variables x_1, x_2, x_3 and x_4 has been fitted to a data consisting of 212 observations and the computer output from estimating this model is given below:

```
Regression Analysis
The regression equation is
y = 3894 - 50.3 x1 + 0.0826 x2 + 0.893 x3 + 0.137 x4

Predictor     Coef  SE Coef      T
Constant    3893.8    409.0   9.52
x1          -50.32    9.062  -5.55
x2         0.08258  0.02133   3.87
x3         0.89269  0.04744  18.82
x4         0.13677  0.05303   2.58
```

Which of the following statement is NOT true?

a. All the explanatory variables have insignificant influence on $\boldsymbol{y}$.

b. The variable $\boldsymbol{x}_1$ is a significant variable.

c. The variable $\boldsymbol{x}_2$ is a significant variable.

d. The variable $\boldsymbol{x}_3$ is a significant variable.

e. The variable $\boldsymbol{x}_4$ is a significant variable.

Where $\boldsymbol{x}_i$'s are vectors of explanatory variables and $\boldsymbol{y}$ is the vector of response variable.

Solution

12. The estimated regression model of fitting life expectancy from birth (LIFE_EXP) on the country's gross national product (in thousands) per population (GNP) and the percentage of population living in urban areas (URBAN%) is given by:

$$\text{LIFE\_EXP} = \ 48.24 \ + \ 0.79 \text{ GNP} \ + \ 0.154 \text{ URBAN\%}.$$

For a particular country, its URBAN% is 60 and its GNP is 3.0. Calculate the estimated life expectancy at birth for this country.

a. 49

b. 50

c. 57

d. 60

e. 65

Solution

13. What is the use of the scatter plot of the fitted values and the residuals?

a. to examine the normal distribution assumption of the errors

b. to examine the goodness of fit of the regression model

c. to examine the constant variation assumption of the errors

d. to test whether the errors have zero mean

e. to examine the independence of the errors

Solution

KNN question

1. Consider a k-nearest neighbours model where $Y = f(X) + \epsilon$, $\mathbb{E}(\epsilon) = 0, \mathbb{V}(\epsilon) = \sigma^2$, and the estimated model is $\widehat{f}(x)$. The weight function is $\frac{1}{k}$. Show that

$$\text{EPE}_k(x_0) = \sigma^2 + \left[f(x_0) - \frac{1}{k} \sum_{l \in N(x_0)} f(x_{(l)}) \right]^2 + \frac{\sigma^2}{k}$$

Where $N(x_0)$ are x_0's k-nearest neighbours. Note that:

$$\text{EPE}_k(x_0) = \mathbb{E}[(Y - \widehat{f}(x_0))^2 | X = x_0]$$

Solution

Applied Questions

1. ⋆ (ISLR2, Q3.8) This question involves the use of simple linear regression on the Auto data set.

   a. Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output.
   For example:

      i. Is there a relationship between the predictor and the response?

      ii. How strong is the relationship between the predictor and the response?

      iii. Is the relationship between the predictor and the response positive or negative?

      iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

   b. Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

   c. Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

   Solution

2. (ISLR2, Q3.11) In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
set.seed(1)
x <- rnorm(100)
y <- 2 * x + rnorm(100)
```

a. Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate $\widehat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (You can perform regression without an intercept using the command `lm(y ~ x+0)`.)

b. Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.

c. What is the relationship between the results obtained in (a) and (b)?

d. For the regression of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes the form $\widehat{\beta}/\mathrm{SE}(\widehat{\beta})$, where $\widehat{\beta}$ is given by (3.38), and where

$$\mathrm{SE}(\widehat{\beta}) = \sqrt{\frac{\sum_i^n (y_i - x_i\widehat{\beta})^2}{(n-1)\sum_{i'=1}^n x_{i'}^2}}$$

(These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in R, that the t-statistic can be written as

$$\frac{(\sqrt{n-1})\sum_{i=1}^n x_i y_i}{\sqrt{(\sum_{i=1}^n x_i^2)(\sum_{i'=1}^n y_{i'}^2) - (\sum_{i'=1}^n x_{i'} y_{i'})^2}}$$

e. Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

f. In R, show that when regression is performed with an intercept, the t-statistic for $H_0 : \beta_1 = 0$ is the same for the regression of y onto x as it is for the regression of x onto y.

Solution

Solutions

Conceptual Questions

Simple linear regression questions

1. We determine $\widehat{\beta}_0$ and $\widehat{\beta}_1$ by minimizing the error. Hence, we use least squares estimates (LSE) for $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$\min_{\beta_0,\beta_1}\left\{S\left(\widehat{\beta}_0,\widehat{\beta}_1\right)\right\}=\min_{\beta_0,\beta_1}\left\{\sum_{i=1}^{n}\left(y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1x_i\right)\right)^2\right\}.$$

The minimum is obtained by setting the first order condition (FOC) to zero:

$$\begin{aligned}\frac{\partial S\left(\widehat{\beta}_0,\widehat{\beta}_1\right)}{\partial\widehat{\beta}_0}&=-2\sum_{i=1}^{n}y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1x_i\right)\\ \frac{\partial S\left(\widehat{\beta}_0,\widehat{\beta}_1\right)}{\partial\widehat{\beta}_1}&=-2\sum_{i=1}^{n}x_i\left(y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1x_i\right)\right).\end{aligned}$$

The LSE $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by setting the FOC equal to zero:

$$\begin{aligned}\sum_{i=1}^{n}y_i=&n\widehat{\beta}_0+\widehat{\beta}_1\sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}x_iy_i=&\widehat{\beta}_0\sum_{i=1}^{n}x_i+\widehat{\beta}_1\sum_{i=1}^{n}x_i^2.\end{aligned}$$

So we have

$$\begin{aligned}\widehat{\beta}_0=&\frac{\sum_{i=1}^{n}y_i-\widehat{\beta}_1\sum_{i=1}^{n}x_i}{n}=\overline{y}-\widehat{\beta}_1\overline{x},\text{ and}\\ \widehat{\beta}_1=&\frac{\sum_{i=1}^{n}x_iy_i-\widehat{\beta}_0\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2}.\end{aligned}$$

Next step: Rearranging so that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ become functions of $\sum_{i=1}^{n}y_i$, $\sum_{i=1}^{n}x_i$, $\sum_{i=1}^{n}x_i^2$, and $\sum_{i=1}^{n}x_iy_i$

$$\begin{aligned}
\widehat{\beta}_0 =& \frac{\sum_{i=1}^n y_i - \left(\frac{\sum_{i=1}^n x_i y_i - \widehat{\beta}_0 \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2}\right) \sum_{i=1}^n x_i}{n} \\
\left(1 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n \sum_{i=1}^n x_i^2}\right) \widehat{\beta}_0 =& \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \left(\sum_{i=1}^n x_i y_i\right) \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2} \\
\widehat{\beta}_0 \overset{*}{=}& \frac{\sum_{i=1}^n y_i \left(\sum_{i=1}^n x_i^2\right) - \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}.
\end{aligned}$$

$*(1 - a/b)c = d/b \to (bc - ac)/b = d/b \to c = d/(b - a)$.

And $\widehat{\beta}_0$'s in line (6) was subbed into $\widehat{\beta}_1$ in line (7). At this point, $\widehat{\beta}_0$ is done. So we'll continue with $\widehat{\beta}_1$.

From the previous steps we have:

$$\begin{aligned}
\widehat{\beta}_0 =& \frac{\sum_{i=1}^n y_i - \widehat{\beta}_1 \sum_{i=1}^n x_i}{n} \\
\widehat{\beta}_1 =& \frac{\sum_{i=1}^n x_i y_i - \widehat{\beta}_0 \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2}.
\end{aligned}$$

thus:

$$\begin{aligned}
\widehat{\beta}_1 =& \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n y_i - \widehat{\beta}_1 \sum_{i=1}^n x_i\right) \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2} \\
\left(1 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n \sum_{i=1}^n x_i^2}\right) \widehat{\beta}_1 =& \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2} \\
\widehat{\beta}_1 \overset{*}{=}& \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}.
\end{aligned}$$

$*(1 - a/b)c = d/b \to (bc - ac)/b = d/b \to c = d/(b - a)$.

Using the notations, we have an easier way to write $\widehat{\beta}_1$:

$$\begin{aligned}
\widehat{\beta}_1 =& \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
=& \frac{n \left(\sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i \cdot \frac{n}{n^2}\right)}{n \left(\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 \cdot \frac{n}{n^2}\right)} \\
=& \frac{\sum_{i=1}^n x_i y_i - n \overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 - n \overline{x}^2} \\
\overset{*}{=}& \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \overline{y} - \sum_{i=1}^n y_i \overline{x} + n \overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n \overline{x}^2 - 2 \sum_{i=1}^n x_i \overline{x}} \\
=& \frac{\sum_{i=1}^n (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}.
\end{aligned}$$

*$\sum_{i=1}^n x_i\overline{y} = \sum_{i=1}^n x_i \frac{\sum_{i=1}^n y_i}{n} = \sum_{i=1}^n y_i \frac{\sum_{i=1}^n x_i}{n} = \sum_{i=1}^n y_i\overline{x} = n\frac{\sum_{i=1}^n x_i}{n}\frac{\sum_{i=1}^n y_i}{n} = n\overline{xy}$.

2. For $\widehat{\beta}_0$, using the equation in line (10) in Q1:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_0|X\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^n y_i\left(\sum_{i=1}^n x_i^2\right) - \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}\right] \\
&= \frac{\sum_{i=1}^n \mathbb{E}\left[y_i\right]\left(\sum_{i=1}^n x_i^2\right) - \sum_{i=1}^n x_i \mathbb{E}\left[y_i\right] \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\sum_{i=1}^n \left(\beta_0 + \beta_1 x_i\right)\left(\sum_{i=1}^n x_i^2\right) - \sum_{i=1}^n x_i\left(\beta_0 + \beta_1 x_i\right) \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{n\beta_0\left(\sum_{i=1}^n x_i^2\right) - \sum_{i=1}^n \left(\beta_0 x_i\right) \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \beta_0.
\end{aligned}$$

For $\widehat{\beta}_1$, using equation in line (13) from Q1:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_1|X\right] &= \mathbb{E}\left[\frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}\right] \\
&= \frac{n\sum_{i=1}^n x_i \mathbb{E}\left[y_i\right] - \sum_{i=1}^n \mathbb{E}\left[y_i\right] \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{n\sum_{i=1}^n x_i\left(\beta_0 + \beta_1 x_i\right) - \sum_{i=1}^n \left(\beta_0 + \beta_1 x_i\right) \sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\beta_1\sum_{i=1}^n x_i^2 + n\sum_{i=1}^n x_i\beta_0 - \sum_{i=1}^n \beta_0 \sum_{i=1}^n x_i - \beta_1\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
&= \beta_1.
\end{aligned}$$

3. In the regression model there are three parameters to estimate: β_0, β_1, and σ^2.

   Joint density of $Y_1, Y_2, \ldots, Y_n$ — under the (strong) normality assumptions — is the product of their marginals (independent by assumption) so that the likelihood is:

$$\begin{aligned}
L\left(\beta_0, \beta_1, \sigma; \{(x_i, y_i)\}_{i=1}^n\right) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2}{2\sigma^2}\right) \\
&= \frac{1}{\left(2\pi\right)^{n/2}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n\left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2\right) \\
\ell\left(\beta_0, \beta_1, \sigma; \{(x_i, y_i)\}_{i=1}^n\right) &= -n\log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n\left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2.
\end{aligned}$$

Taking partial derivatives of the log-likelihood with respect to β_0:

$$\begin{aligned}\frac{\partial l}{\partial \beta_0} &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ &= \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i\,.\end{aligned}$$

Equate the above to 0 and solve for β_0 should give

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}\,.$$

Similarly, taking partial derivatives of the log-likelihood with respect to β_1:

$$\begin{aligned}\frac{\partial l}{\partial \beta_1} &= \sum_{i=1}^{n} 2x_i(y_i - (\beta_0 + \beta_1 x_i)) \\ &= 2\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \beta_0 - \sum_{i=1}^{n} \beta_1 x_i^2)\right) \\ &= 2\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i(\overline{y} - \beta_1 \overline{x}) - \sum_{i=1}^{n} \beta_1 x_i^2\right) \\ &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \overline{y} - \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_1 \sum_{i=1}^{n} x_i \overline{x} \\ &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\,.\end{aligned}$$

The last line was derived using the fact that

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}\,.$$

Equate the above equation to 0 and solve for β_1, we'll get:

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\,.$$

The rest is the same as the derivation from Q1

4. We have that

$$\begin{aligned}\mathrm{SST} &= \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n} y_i^2+\overline{y}^2-2\overline{y}y_i\,.\\ \mathrm{SSE}+\mathrm{SSM} &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2\\ &= \sum_{i=1}^{n}\Big(y_i^2+\widehat{y}_i^2-2y_i\widehat{y}_i+\widehat{y}_i^2+\overline{y}^2-2\overline{y}\widehat{y}_i\Big)\\ &= \sum_{i=1}^{n}\Big(y_i^2+2\widehat{y}_i^2-2y_i\widehat{y}_i+\overline{y}^2-2\overline{y}\widehat{y}_i\Big)\\ &\overset{*}{=} \sum_{i=1}^{n}\Big(y_i^2+2\widehat{y}_i^2-2(\widehat{y}_i+\widehat{\epsilon}_i)\widehat{y}_i+\overline{y}^2-2\overline{y}(y_i-\widehat{\epsilon}_i)\Big)\end{aligned}$$

* using $\widehat{\epsilon}_i=y_i-\widehat{y}_i$, continue:

$$\begin{aligned}\mathrm{SSE}+\mathrm{SSM} &= \sum_{i=1}^{n}\Big(y_i^2+2\widehat{y}_i^2-2(\widehat{y}_i+\widehat{\epsilon}_i)\widehat{y}_i+\overline{y}^2-2\overline{y}(y_i-\widehat{\epsilon}_i)\Big)\\ &= \sum_{i=1}^{n}\Big(y_i^2-2\widehat{y}_i\widehat{\epsilon}_i+\overline{y}^2-2\overline{y}y_i+2\overline{y}\widehat{\epsilon}_i\Big)\\ &\overset{**}{=} \sum_{i=1}^{n}\Big(y_i^2+\overline{y}^2-2\overline{y}y_i\Big)=\mathrm{SST}\,.\end{aligned}$$

** uses $\sum\widehat{\epsilon}_i=0$ (which is self-explanatory) and $\sum x_i\widehat{\epsilon}_i=0$ (We'll prove this at the end of this question), we have the following results

$$\sum_{i=1}^{n}2\overline{y}\widehat{\epsilon}_i=2\overline{y}\sum_{i=1}^{n}\widehat{\epsilon}_i=0\,,$$

$$\sum_{i=1}^{n}2\widehat{y}_i\widehat{\epsilon}_i=\sum_{i=1}^{n}2(\widehat{\beta}_0+\widehat{\beta}_1x_i)\widehat{\epsilon}_i\,.$$

Proof of $\sum x_i\widehat{\epsilon}_i=0$:

Using the estimates of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we have:

$$\begin{aligned}
\sum_{i=1}^{n} x_i \widehat{e_i} &= \sum_{i=1}^{n} x_i \left(y_i - (\widehat{\beta}_0 - \widehat{\beta}_1 x_1) \right) \\
&= \sum_{i=1}^{n} x_i y_i - \widehat{\beta}_0 \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} \frac{y_i}{n} - \widehat{\beta}_1 \sum_{i=1}^{n} \frac{x_i}{n} \right) \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_1^2 \\
&= \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} + \widehat{\beta}_1 \left(\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} - \sum_{i=1}^{n} x_i^2 \right) \\
&\overset{***}{=} \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} + \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \left(\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} - \sum_{i=1}^{n} x_i^2 \right) \\
&= \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} - \left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \right) \\
&= 0 \,.
\end{aligned}$$

*** uses equation (13) from Q1.

5. The SSM is

$$\begin{aligned}
\text{SSM} &= \sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2 \\
&= \sum_{i=1}^{n} \left(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_i - \overline{y} \right)^2 \\
&= \sum_{i=1}^{n} \left((\overline{y} - \widehat{\beta}_1 \cdot \overline{x}) + \widehat{\beta}_1 \cdot x_i - \overline{y} \right)^2 \\
&= \sum_{i=1}^{n} \widehat{\beta}_1^2 \cdot (x_i - \overline{x})^2 \\
&\overset{b)}{=} \widehat{\beta}_1^2 \cdot S_{xx} \overset{a)}{=} \widehat{\beta}_1 \cdot S_{xy}
\end{aligned}$$

using $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$.

6. We first consider $\mathbb{V}\left(\widehat{\beta}_1 | X\right)$.

Note that we have:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \overset{*}{=} \frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} .$$

*uses:

$$\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\overline{y} = \overline{y}\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\overline{x}.\overline{y}$$
$$= \overline{y}.\overline{x}n-\overline{y}.\overline{x}n=0.$$

Therefore

$$\begin{aligned}\mathbb{V}\left(\widehat{\beta}_1|X\right) &= \mathbb{V}\left(\left.\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)y_i}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}\right|X\right)\\ &= \frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\mathbb{V}\left(y_i|X\right)}{\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)^2}\\ &= \frac{\sigma^2\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)^2}=\frac{\sigma^2}{S_{xx}}\,.\end{aligned}$$

This uses $\mathbb{V}(y_i|X)=\sigma^2$ because $y_i=\beta_0+\beta_1x_i+\epsilon_i$, where the β's are constant and x_i is given, hence $\mathbb{V}(y_i|X)=\mathbb{V}(\epsilon|X)=\sigma^2$.

We next consider $\mathbb{V}\left(\widehat{\beta}_0|X\right)$.

Using that:

$$\widehat{\beta}_0=\overline{y}-\widehat{\beta}_1\overline{x},$$

$$\begin{aligned}\mathbb{V}\left(\widehat{\beta}_0|X\right) =& \mathbb{V}\left(\overline{y}-\widehat{\beta}_1\overline{x}|X\right)\\ =& \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}y_i\mid X\right)+\overline{x}^2\mathbb{V}\left(\widehat{\beta}_1|X\right)\\ =& \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left(y_i|X\right)/n^2+\overline{x}^2\frac{\sigma^2}{S_{xx}}\\ =& \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right).\end{aligned}$$

Finally, we consider $\mathrm{Cov}\left(\widehat{\beta}_0,\widehat{\beta}_1|X\right)$.

$$\widehat{\beta}_0=\overline{y}-\widehat{\beta}_1\overline{x},$$

we have:

$$\begin{aligned}
\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1|X\right) &= \text{Cov}\left(\overline{y} - \widehat{\beta}_1\overline{x}, \widehat{\beta}_1|X\right) \\
&= \text{Cov}\left(-\widehat{\beta}_1\overline{x}, \widehat{\beta}_1|X\right) \\
&= -\overline{x} \cdot \text{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_1|X\right) \\
&= -\overline{x} \cdot \mathbb{V}\left(\widehat{\beta}_1|X\right) \\
&= -\frac{\overline{x}\sigma^2}{S_{xx}}.
\end{aligned}$$

7. We have that

$$\begin{aligned}
\mathbb{V}\left(\widehat{y}_0|X\right) &= \mathbb{V}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_0|X\right) \\
&= \mathbb{V}\left(\widehat{\beta}_0|X\right) + x_0^2\mathbb{V}\left(\widehat{\beta}_1|X\right) + 2x_0\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1|X\right) \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2 + x_0^2\frac{\sigma^2}{S_{xx}} + 2x_0\left(\frac{-\overline{x}\sigma^2}{S_{xx}}\right) \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2 - 2x_0\overline{x} + x_0^2}{S_{xx}}\right)\sigma^2 \\
&= \left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_{xx}}\right)\sigma^2.
\end{aligned}$$

8. Expectation:

$$\begin{aligned}
\mathbb{E}\left[Y_0 - \widehat{y}_0|X\right] &= \mathbb{E}\left[Y_0|X\right] - \mathbb{E}\left[\widehat{y}_0|X\right] \\
&= \mathbb{E}\left[\beta_0 + \beta_1 x_0 + \epsilon_i\right] - \mathbb{E}\left[(\widehat{\beta}_0 + \widehat{\beta}_1 x_0)|X\right] \\
&\overset{*}{=} \beta_0 + \beta_1 x_0 - (\beta_0 + \beta_1 x_0) \\
&= 0.
\end{aligned}$$

*uses the fact that the expected value of the random error ϵ is 0.

Variance:

$$\begin{aligned}
\mathbb{V}\left(Y_0 - \widehat{y}_0|X\right) &= \mathbb{V}\left(Y_0|X\right) + \mathbb{V}\left(\widehat{y}_i|X\right) - 2\,\text{Cov}\left(Y_0, \widehat{y}_0|X\right) \\
&= \mathbb{V}\left(\beta_0 + \beta_1 x_0 + \epsilon_0|X\right) + \sigma^2\left(\frac{1}{n} + \frac{(\overline{x} - x_i)^2}{S_{xx}}\right) - 0 \\
&\overset{**}{=} \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(\overline{x} - x_i)^2}{S_{xx}}\right) \\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{(\overline{x} - x_i)^2}{S_{xx}}\right).
\end{aligned}$$

** uses $\mathbb{V}(\beta_0 + \beta_1 x_0 | X) = 0$ as it contains only constants, and the covariance is 0 because the observed point is not used in making predictions, and should be independent of the predicted point, and the conditional variance of $\widehat{y}_0$ was derived in earlier questions.

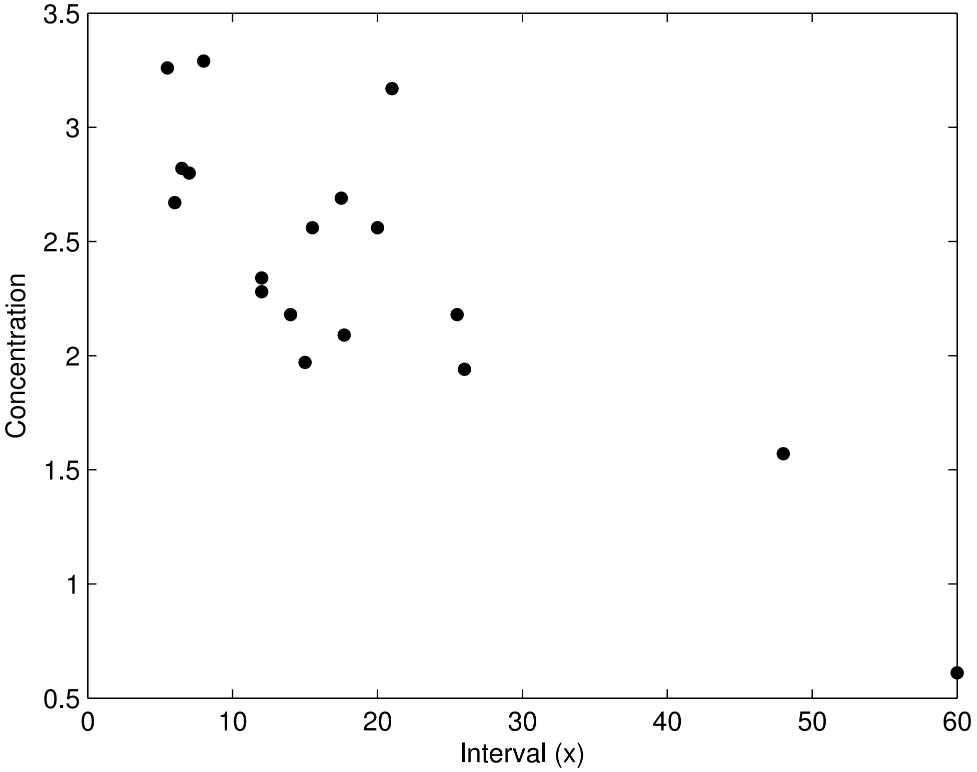


Figure 1: Scatterplot of concentration against interval.

9. a. Interesting features are that, in general, the concentration of 3-MT in the brain seems to decrease as the post mortem interval increases. Another interesting feature is that we observe two observations with a much higher post mortem interval than the other observations.
   The data seems to be appropriate for linear regression. The linear relationship seems to hold,especially for values of interval between 5 and 26 (we have enough observations for that). Care should be taken into account when evaluating y for x lower than 5 and larger than 26 (only two observations) because we do not know whether the linear relationship between x and y still holds then.

   b. We test:

$$H_0 : \rho = 0 \quad \text{v.s.} \quad H_1 : \rho \neq 0$$

The corresponding test statistic is given by:

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{n-2}.$$

We reject the null hypothesis for large and small values of the test statistic. We have $n = 18$ and the correlation coefficient is given by:

$$\begin{aligned} r =& \frac{\sum x_i \cdot y_i - n\overline{xy}}{\sqrt{(\sum x_i^2 - n\overline{x}^2)(\sum y_i^2 - n\overline{y}^2)}} \\ =& \frac{672.8 - 18 \cdot 337/18 \cdot 42.98/18}{\sqrt{(9854.5 - 337^2/18) \cdot (109.7936 - 42.98^2/18)}} = -0.827 \end{aligned}$$

Thus, the value of our test statistic is given by:

$$T = \frac{-0.827\sqrt{16}}{\sqrt{1-(-0.827)^2}} = -5.89.$$

From Formulae and Tables page 163 we observe $\mathbb{P}(t_{16} \leq -4.015) \overset{*}{=} \mathbb{P}(t_{16} \geq 4.015) = 0.05\%$, * using symmetry property of the student-t distribution. We observe that the value of our test statistic (-5.89) is smaller than -4.015, thus our p-value should be smaller than $2 \cdot 0.05\% = 0.1\%$. Thus, we can reject the null hypothesis even at a significance level of 0.1%, hence we can conclude that there is a linear dependency between interval and concentration. Note that the alternative hypothesis is here a linear dependency and not negative linear dependency, so you do accept the alternative by rejecting the null hypothesis. Although, when you would use as alternative hypothesis negative dependency, you would accept this alternative, due to the construction of the test we have to use the phrase "a linear dependency" and not "a negative linear dependency".

c. The linear regression model is given by:

$$y = \alpha + \beta x + \epsilon$$

The estimate of the slope is given by:

$$\begin{aligned} \widehat{\beta} =& \frac{\sum x_i y_i - n \sum x_i/n \sum y_i/n}{\sum x_i^2 - n(\sum x_i/n)^2} \\ =& \frac{672.8 - 337 \cdot 42.98/18}{9854.4 - 334^2/18} = -0.0372008 \end{aligned}$$

The estimate of the intercept is given by:

$$\begin{aligned} \widehat{\alpha} =& \overline{y} - \widehat{\beta}\overline{x} \\ =& 42.98/18 + 0.0372008 \cdot 337/18 = 3.084259 \end{aligned}$$

Thus, the estimate of y given a value of x is given by:

$$\begin{aligned}\widehat{y} =& \widehat{\alpha} + \widehat{\beta}x \\ =& 3.084259 - 0.0372008x\end{aligned}$$

i. One day equals 24 hours, i.e., $x = 24$, thus $\widehat{y} = \widehat{\alpha} + \widehat{\beta}24 = 3.084259 - 0.0372008 \cdot 24 = 2.19$

ii. Two day equals 48 hours, i.e., $x = 48$, thus $\widehat{y} = \widehat{\alpha} + \widehat{\beta}24 = 3.084259 - 0.0372008 \cdot 48 = 1.30$

The data set contains accurate data up to 26 hours, as for observations 17 and 18 (at 48 hour and 60 hours respectively) there was no eye-witness testimony direct available. Predicting 3-MT concentration after 26 hours may not be advisable, even though $x = 48$ is within the range of the x-values (5.5 hours to 60 hours).

d. The pivotal quantity is given by:

$$\frac{\beta - \widehat{\beta}}{\text{SE}(\widehat{\beta})} \sim t_{n-2}.$$

First, we calculate

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n-2}\left(\sum y_i^2 - (\sum y_i)^2/n - \frac{(\sum x_i y_i - \sum x_i \sum y_i/n)^2}{\sum x_i^2 - (\sum x_i)^2/n}\right) \\ &= \frac{1}{16}\left(109.7936 - 42.98^2/18 - \frac{(672.8 - 337 \cdot 42.98/18)^2}{9854.5 - 337^2/18}\right) = 0.1413014,\end{aligned}$$

then the standard error is

$$\text{SE}(\widehat{\beta}) = \sqrt{\frac{\widehat{\sigma}^2}{\sum x_i^2 - n\overline{x}^2}} = \sqrt{\frac{0.1413014}{9854.5 - 337^2/18}} = 0.00631331.$$

From Formulae and Tables page 163 we have $t_{16,1-0.005} = 2.921$. Using the test statistic, the 99% confidence interval of the slope is given by:

$$\begin{aligned}\widehat{\beta} - t_{16,1-\alpha/2}\text{SE}(\widehat{\beta}) <& \beta < \widehat{\beta} + t_{16,1-\alpha/2}\text{SE}(\widehat{\beta}) \\ -0.0372008 - 2.921 \cdot 0.00631331 <& \beta < -0.0372008 + 2.921 \cdot 0.00631331 \\ -0.055641979 <& \beta < -0.0188.\end{aligned}$$

Thus the 99% confidence interval of β is given by: $(-0.055641979, -0.0188)$. Note that $\beta = 0$ in not within the 99% confidence interval, therefore we would reject the null hypothesis that β equals zero and accept the alternative that $\beta \neq 0$ at a 1% level of significance. This confirms the result in (2) where the correlation coefficient was shown to not equal zero at the 1% significance level.

10. a. The linear regression model is given by:

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ i.i.d. distributed for $i = 1, \ldots, n$.
The fitted linear regression equation is given by:

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta} x.$$

The estimated coefficients of the linear regression model are given by (see Formulae and Tables page 25):

$$\begin{aligned}
\widehat{\beta} = \frac{s_{xy}}{s_{xx}} &= \frac{1122}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\
&= \frac{1122}{60016 - 12 \cdot \left(\frac{836}{12}\right)^2} = \frac{1122}{1774.67} = 0.63223 \\
\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} &= \frac{\sum_{i=1}^{n} y_i}{n} - \widehat{\beta}\frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{867}{12} - 0.63223 \cdot \frac{836}{12} = 28.205.
\end{aligned}$$

Thus, the fitted linear regression equation is given by:

$$\widehat{y} = 28.205 + 0.63223 \cdot x.$$

b. The estimate for σ^2 is given by:

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\
&= \frac{1}{n-2} \text{SSE} \\
&= \frac{1}{n-2} (\text{SST} - \text{SSM}) \\
&\overset{*}{=} \frac{1}{n-2} \left(\sum_{i=1}^{n} (y_i - \overline{y})^2 - \widehat{\beta}_1^2 \cdot S_{xx} \right) \\
&= \frac{1}{n-2} \left(\sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2 - \frac{\left(\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \right) \\
&= \frac{1}{10} \cdot \left(63603 - 12 \cdot \left(\frac{867}{12}\right)^2 - \frac{1122^2}{60016 - 836^2/12} \right) = 25.289\,.
\end{aligned}$$

* uses the result from Q5 for SSM.

We know the pivotal quantity:

$$\frac{s^2}{\sigma^2/(n-2)} \sim \chi^2_{n-2}\,.$$

Note: we have $n-2$ degrees of freedom because we have to estimate two parameters form the data ($\widehat{\alpha}$ and $\widehat{\beta}$). We have that $s^2 = \widehat{\sigma}^2$. Thus we have that the 90% confidence interval is given by:

$$\begin{gathered}
\frac{10\widehat{\sigma}^2}{\chi^2_{0.95,10}} < \sigma^2 < \frac{10\widehat{\sigma}^2}{\chi^2_{0.05,10}} \\
\frac{10\cdot 25.289}{18.3} < \sigma^2 < \frac{10\cdot 25.289}{3.94} \\
13.8 < \sigma^2 < 64.2
\end{gathered}$$

Thus the 90% confidence interval of σ^2 is given by $(13.8, 64.2)$.

c. i. We test the following:

$$H_0 : \beta = 0 \quad \text{v.s.} \quad H_1 : \beta > 0,$$

with a level of significance $\alpha = 0.05$.

ii. The test statistic is:

$$T = \frac{\widehat{\beta}-\beta}{\sqrt{\widehat{\sigma^2}/\left(\sum_{i=1}^n (x_i-\overline{x})^2\right)}} \sim t_{n-2}$$

iii. The rejection region of the test is given by:

$$C = \{(X_1,\ldots,X_n) : T \in (t_{10,1-0.05},\infty)\} = \{(X_1,\ldots,X_n) : T \in (1.812,\infty)\}$$

iv. The value of the test statistic is given by:

$$T = \frac{0.63223-0}{\sqrt{25.289/(\sum_{i=1}^n x_i^2 - n\overline{x}^2)}} = \frac{0.63223-0}{\sqrt{25.289/(60016-836^2/12)}} = 5.296.$$

v. The value of the test statistic is in the rejection region, hence we reject the null hypothesis of a zero correlation.

d. We have that $\frac{(y_i|x_i)-(\widehat{y}|x_i)}{\sqrt{\mathbb{V}(y_i|x_i)}}$ has a student-t distribution:

$$\frac{y_i|x_i - \widehat{y}|x_i}{\sqrt{\mathbb{V}(y_i|x_i)}} \sim t_{n-2}$$

The predicted value is given by:

$$\widehat{y}|x_i = \widehat{\alpha} + \widehat{\beta}x_i = 28.205 + 0.63223 \cdot 53 = 61.713.$$

The estimated variance of the observation $x = 53$ is give by:

$$\begin{aligned}\mathbb{V}(y_i|x_i = 53) =& \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\widehat{\sigma^2} \\ =& \left(\frac{1}{12} + \frac{(53 - 836/12)^2}{60016 - 836^2/12}\right) 25.289 = 6.0657.\end{aligned}$$

Thus, the 95% confidence interval for the value of y given that $x = 53$ is given by:

$$\begin{gathered}\widehat{y} - t_{1-0.05/2} \cdot \sqrt{\mathbb{V}(y_i|x_i = 53)} < y|x = 53 < \widehat{y} + t_{1-0.05/2} \cdot \sqrt{\mathbb{V}(y_i|x_i = 53)} \\ 61.713 - 2.228 \cdot \sqrt{6.0657} < y|x = 53 < 61.713 + 2.228 \cdot \sqrt{6.0657} \\ 56.2 < y|x = 53 < 67.2\end{gathered}$$

Thus the 95% confidence interval of y given $x = 53$ is $(56.2, 67.2)$.

e. i. We test the following hypothesis:

$$H_0 : \rho = 0.75 \quad \text{v.s.} \quad H_1 : \rho \neq 0.75$$

ii. The test statistic is given by:

$$T = \frac{Z_r - z_\rho}{\sqrt{\frac{1}{n-3}}} \sim N(0,1)$$

iii. The critical region is given by:

$$C = \{(X_1, \ldots, X_n) : T \in \{(-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)\}\}$$

iv. The value of the test statistic is given by:

$$\frac{Z_r - z_\rho}{\sqrt{\frac{1}{9}}} = 3(z_r - z_\rho) = 3(1.2880 - 0.97296) = 0.94512$$

where

$$\begin{aligned}z_r =& \frac{1}{2}\log\left(\frac{1+0.85860}{1-0.85860}\right) = 1.2880 \\ z_\rho =& \frac{1}{2}\log\left(\frac{1+0.75}{1-0.75}\right) = 0.97296 \\ r =& \frac{\sum_{i=1}^n (x_i - \overline{x})(x_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}} \\ =& \frac{1122}{(\sum_{i=1}^n y_i^2 - n\overline{y}_i^2)(\sum_{i=1}^n x_i^2 - n\overline{x}^2)} \\ =& \frac{1122}{\sqrt{962.25 \cdot 1774.667}} = 0.85860\end{aligned}$$

v. We have that $z_{0.82894} = 0.95$. Thus, the p-value is given by $2 \cdot (1 - 0.82894) = 0.34212$. The value of the test statistic is not in the critical region if the level of significance is lower than 0.34212 (which is normally the case). Hence, for reasonable values of the level of significance we would not reject the null hypothesis.

f. The proportion of the variability explained by the model is given by:

$$\begin{aligned}
R^2 =& \frac{\text{SSM}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \\
=& 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y}_i)^2} \\
=& 1 - \frac{\sum_{i=1}^n y_i^2 - n\overline{y}^2 - \frac{\left(\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}}{\sum_{i=1}^n y_i^2 - n\overline{y}_i^2} \\
=& \frac{\left(\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right)^2}{(\sum_{i=1}^n y_i^2 - n\overline{y}_i^2)(\sum_{i=1}^n x_i^2 - n\overline{x}^2)} \\
=& \frac{1122^2}{962.25 \cdot 1774.667} = 0.737193.
\end{aligned}$$

Hence, a large proportion of the variability of Y is explained by X.

11. The completed ANOVA table is given below:

Source	D.F.	Sum of Squares	Mean Squares	F-Ratio
Regression	1	$639.5 - 475.6 = 163.9$	163.9	$\frac{163.9}{8.2} = 19.99$
Error	58	$8.2 \times 58 = 475.6$	8.2	
Total	59	639.5		

12. A simple linear regression problem:

a. Since we know that $\widehat{\beta} = r\frac{s_y}{s_x}$, then $r = \widehat{\beta}\frac{s_x}{s_y} = 7.445(2.004/21.56) = 69.2\%$. where s_x, s_y are sample standard deviations. Alternatively, you can use the fact that $R^2 = r^2$, so that from 4. below, $r^2 = 0.4794 \Longrightarrow r = +\sqrt{0.4794} = 69.2\%$. You take the positive square root because of the positive sign of the coefficient of EPS.

b. Given $EPS = 2$, we have:

$$\widehat{STKPRICE} = 25.044 + 7.445\,(2) = 39.934.$$

A 95% confidence interval of this estimate is given by:

$$\begin{aligned}
&\left(\widehat{\alpha}+\widehat{\beta}x_0\right) \pm t_{1-\alpha/2,n-2} \times s\sqrt{\left(\frac{1}{n}+\frac{(\overline{x}-x_0)^2}{(n-1)\,s_x^2}\right)} \\
&= (39.934) \pm \underbrace{t_{1-0.025,46}}_{=2.012896} \times \sqrt{247}\sqrt{\left(\frac{1}{48}+\frac{(2.338-2)^2}{(47)\,(2.004^2)}\right)} \\
&= 39.934 \pm 4.636 = (35.298, 44.570)\,.
\end{aligned}$$

where s_x^2 is the sample variance of X.

c. A 95% confidence interval for β is:

$$\begin{aligned}
\widehat{\beta} \pm t_{1-\alpha/2,n-2} \cdot \mathrm{SE}(\widehat{\beta}) &= 7.445 \pm 2.0147 \times \frac{\sqrt{247}}{2.004\sqrt{47}} \\
&= 7.445 \pm 2.305 \\
&= (5.14, 9.75)\,.
\end{aligned}$$

d. $s = \sqrt{247} = 15.716$ and $R^2 = \frac{\mathrm{SSM}}{\mathrm{SST}} = \frac{10475}{21851} = 47.94\%$.

e. A scatter plot or diagram of the fitted values against the residuals (standardised) will provide us an indication of the constancy of the variation in the errors.

f. To test for the significance of the variable EPS, we test $H_0 : \beta = 0$ against $H_a : \beta \neq 0$. The test statistic is:

$$t(\widehat{\beta}) = \frac{\widehat{\beta}}{\mathrm{SE}(\widehat{\beta})} = \frac{7.445}{1.144} = 6.508.$$

This is larger than $t_{1-\alpha/2,n-2} = 2.0147$ and therefore we reject the null. There is evidence to support the fact that the EPS variable is a significant predictor of stock price.

g. To test $H_0 : \beta = 24$ against $H_a : \beta > 24$, the test statistic is given by:

$$t(\widehat{\beta}) = \frac{\widehat{\beta}-\beta_0}{\mathrm{SE}(\widehat{\beta})} = \frac{7.445-24}{1.144} = -14.47.$$

Thus, since this test statistic is smaller than $t_{1-\alpha,n-2} = t_{0.95,46} = 1.676$, do not reject the null hypothesis.

13. The grand total sum is $\sum x = 2479 + 2619 + 2441 + 2677 = 10216$ so that the grand mean is $\overline{\overline{x}} = 10216/40 = 255.4$. Also, $\sum x^2 = 617163 + 687467 + 597607 + 718973 = 2621210$. Therefore the total sum of squares is:

$$\begin{aligned}
\mathrm{SST} &= \sum \left(x-\overline{\overline{x}}\right)^2 = \sum x^2 - N\overline{\overline{x}}^2 \\
&= 2621210 - (40)(255.4)^2 = 12043.6.
\end{aligned}$$

The sum of squares between the regions is:

$$\begin{aligned}\text{SSM} &= \sum n_i \left(\overline{x}_{i.} - \overline{\overline{x}}\right)^2 \\ &= 10\left((247.9-255.4)^2 + (261.9-255.4)^2 + (244.1-255.4)^2 + (267.7-255.4)^2\right) \\ &= 3774.8.\end{aligned}$$

The difference gives the sum of squares within the regions:

$$\text{SSE} = \text{SST} - \text{SSM} = 12043.6 - 3774.8 = 8268.8.$$

The one-way ANOVA table is then summarised below:

ANOVA Table for the One-Way Layout

Source	d.f.	Sum of Squares	Mean Square	F-Statistic
Between	3	3774.8	1258.27	$\frac{1258.27}{229.69} = 5.478$
Within	36	8268.8	229.69	
Total	39	12043.6		

Thus, to test the equality of the mean premiums across the regions, we test:

$$H_0 : \alpha_A = \alpha_B = \alpha_C = \alpha_D = 0 \quad \text{all variances are equal}$$

against the alternative:

$$H_a : \text{at least one } \alpha \text{ is not zero} \quad \text{all variances are equal}$$

using the F-test. Since $F = 5.478 > F_{0.95}(3, 36) = 2.9$ (approximately), we therefore reject H_0. There is evidence to support a difference in the mean premiums across regions. The one-way ANOVA model assumptions are as follows: each random variable x_{ij} is observed according to the model

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \text{ for } i = 1, \ldots, I, \text{ and } j = 1, 2, \ldots, n_i$$

where ε_{ij} refers to the random error in the j^{th} observation of the i^{th} treatment which satisfies:

- $\mathbb{E}[\varepsilon_{ij}] = 0$ and $\mathbb{V}(\varepsilon_{ij}) = \sigma^2$ for all i, j.
- The ε_{ij} are independent and normally distributed (normal errors), and where μ is the overall mean and α_i is the effect of the i^{th} treatment with:

$$\sum_{i=1}^{I} \alpha_i = 0.$$

14. a. We have the estimated correlation coefficient:

$$\begin{aligned}
r =& \frac{s_{ms}}{\sqrt{s_{mm} \cdot s_{ss}}} \\
=& \frac{\sum ms - n\overline{ms}}{\sqrt{(\sum m^2 - n\overline{m}^2) \cdot (\sum s^2 - n\overline{s}^2)}} \\
=& \frac{221,022.58 - 1136.1 \cdot 1934.2/10}{\sqrt{(129,853.03 - 1136.1^2/10) \cdot (377,700.62 - 1934.2^2/10)}} = 0.764.
\end{aligned}$$

i. We have the hypothesis:

$$H_0 : \rho = 0 \quad \text{v.s.} \quad H_1 : \rho > 0$$

ii. The test statistic is:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

iii. The critical region is given by:

$$C = \{(X_1, \ldots, X_n) : T \in (t_{n-2,1-\alpha}, \infty)\}$$

iv. The value of the test is:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.764\sqrt{10-2}}{\sqrt{1-0.764^2}} = 3.35$$

v. We have $t_{8,1-0.005} = 3.35$. Thus the p-value is 0.005 and we reject the null hypothesis of a zero correlation for level of significance less than 0.005 (usually it is larger, thus then we reject the null).

b. Given the issue of whether mortality can be used to predict sickness, we require a plot of sickness against mortality:

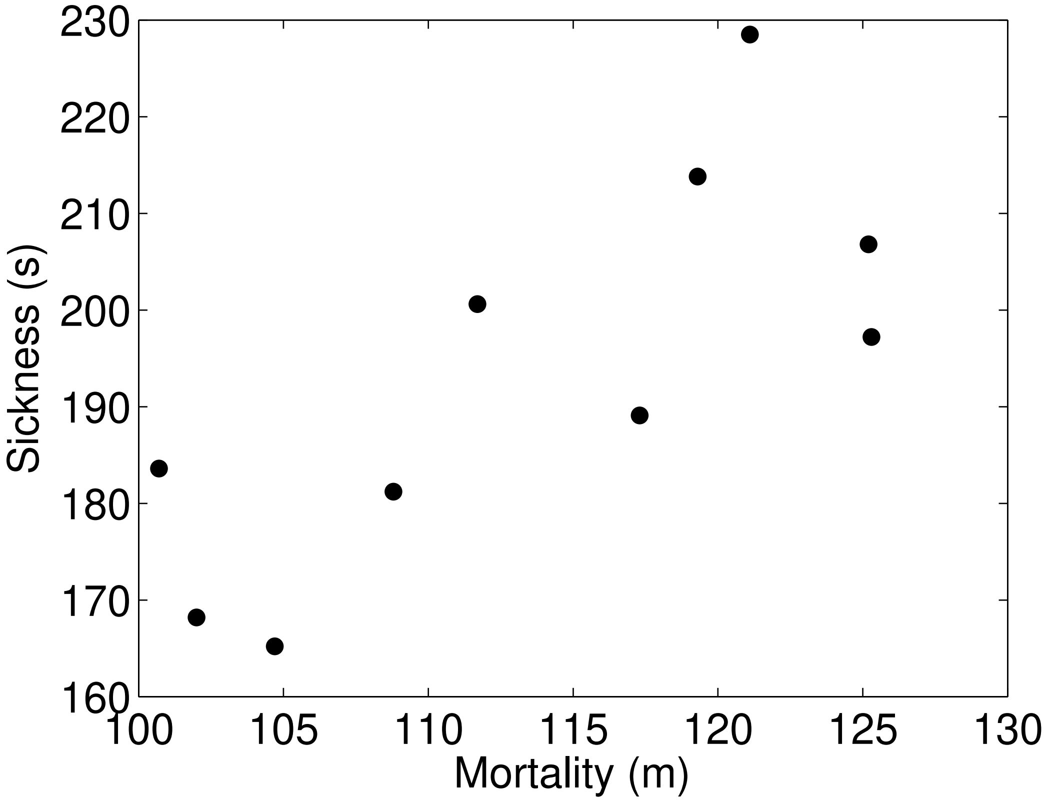


Figure 2: Scatterplot of sickness and mortality.

There seems to be an increase linear relationship such that mortality could be used to predict sickness.

c. We have the estimates:

$$\begin{aligned}
\widehat{\beta} =& \frac{s_{ms}}{s_{mm}} = \frac{\sum ms - n\overline{ms}}{\sum m^2 - n\overline{m}^2} \\
=& \frac{221,022.58 - 1136.1 \cdot 1934.2/10}{129,853.03 - 1136.1^2/10} = 1.6371 \\
\widehat{\alpha} =& \overline{y} - \widehat{\beta}\overline{x} = \frac{1934.2}{10} - 1.6371\frac{1136.1}{10} = 7.426 \\
\widehat{\sigma^2} =& \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n-2}\left(s_{ss} - \frac{s_{ms}^2}{s_{mm}}\right) \\
=& \frac{1}{8}\left((\sum s^2 - n\overline{s}^2) - \frac{(\sum ms - n\overline{ms})^2}{(\sum m^2 - n\overline{m}^2)}\right) \\
=& \frac{1}{8}\left(3587.656 - \frac{(1278.118)^2}{780.709}\right) = 186.902 \\
\mathbb{V}(\widehat{\beta}) =& \widehat{\sigma}^2/s_{mm} = 186.902/780.709 = 0.2394
\end{aligned}$$

i. Hypothesis:

$$H_0 : \beta = 2 \quad \text{v.s.} \quad H_1 : \beta < 2$$

ii. Test statistic:

$$T = \frac{\widehat{\beta} - \beta}{\sqrt{\widehat{\sigma^2}/s_{xx}}} \sim t_{n-2}$$

iii. Critical region:

$$C = \{(X_1, \ldots, X_n) : T \in (-\infty, -t_{n-2,1-\alpha})\}$$

iv. Value of statistic:

$$T = \frac{\widehat{\beta} - \beta}{\sqrt{\widehat{\sigma^2}/s_{xx}}} = \frac{1.6371 - 2}{\sqrt{0.2394}} = -0.74$$

v. We have from Formulae and Tables page 163: $t_{8,1-0.25} = 0.7064$ and $t_{8,1-0.20} = 0.8889$. Thus the p-value (using symmetry) is between 0.2 and 0.25. Thus, we accept the null hypothesis if the level of significance is smaller than the p-value (which is usually the case). Note: exact p-value using computer package is 0.2402.

d. For a region with $m = 115$ we have the estimated value:

$$\widehat{s} = 7.426 + 1.6371 \cdot 115 = 195.69$$

with corresponding variance:

$$\widehat{\sigma^2}\left(\frac{1}{n}+\frac{(x_0-\overline{x})^2}{s_{mm}}\right)=186.902\left(\frac{1}{10}+\frac{(115-113.61)^2}{780.709}\right)=19.1528$$

The corresponding 95% confidence limits are $195.69-t_{8,1-0.025}\cdot \mathrm{SE}(s|m=115)=195.69-2.306\cdot\sqrt{19.1528}=185.60$ and $195.69+t_{8,1-0.025}\cdot\mathrm{SE}(s|m=115)=195.69+2.306\cdot\sqrt{19.1528}=205.78$.

15. a. i. We have:

$$\begin{aligned}
\mathrm{SST} &= \sum y^2-\left(\sum y\right)^2/n = 70.8744-29.12^2/16 = 17.8760\,,\\
\sum x &= 4\cdot(1+2+3+4)=40 \sum x^2 = 4\cdot(1^2+2^2+3^2+4^2)=120\,,\\
\sum xy &= 1\cdot 2.73+2\cdot 6.26+3\cdot 9.22+4\cdot 10.91 = 86.55\,,\\
s_{xy} &= \sum xy-\sum x\sum y/n = 86.55-40\cdot 29.12/16 = 13.75\,,\\
\mathrm{SSM} &= \widehat{\beta}_1^2\cdot s_{xx} = \left(\frac{13.75}{20}\right)^2\cdot 20 = 9.453125\,,\\
\mathrm{SSE} &= \mathrm{SST}-\mathrm{SSM} = 17.8760-9.453125 = 8.422875.
\end{aligned}$$

ii. We have:

$$\begin{aligned}
\widehat{\beta} &= \frac{s_{xy}}{s_{xx}} = \frac{13.75}{20} = 0.6875\\
\widehat{\alpha} &= \overline{y}-\widehat{\beta}\overline{x} = (29.12-0.6875\cdot 40)/16 = 0.1012\,.
\end{aligned}$$

Thus, the fitted model is given by $\widehat{y}=\widehat{\alpha}+\widehat{\beta}x=0.1012+0.6875x$.
For $x=1$ we have: $\widehat{y}=\widehat{\alpha}+\widehat{\beta}x=0.1012+0.6875\cdot 1=0.7887$
For $x=4$ we have: $\widehat{y}=\widehat{\alpha}+\widehat{\beta}x=0.1012+0.6875\cdot 4=2.8512$

iii. We have $\mathrm{SE}(\widehat{\beta})=\sqrt{\frac{8.4229/14}{20}}=0.1734$.

i) Hypothesis:

$$H_0:\beta=0 \quad \text{v.s.} \quad H_1:\beta\neq 0$$

ii) Test statistic:

$$T=\frac{\widehat{\beta}-\beta}{\mathrm{SE}(\widehat{\beta})}\sim t_{n-2}$$

iii) Critical region:

$$C=\{(X_1,\ldots,X_n):T\in\{(-\infty,-t_{n-2,1-\alpha/2})\cup(t_{n-2,1-\alpha/2},\infty)\}\}$$

iv) Value of statistic:

$$T=\frac{\widehat{\beta}-\beta}{\mathrm{SE}(\widehat{\beta})}=\frac{0.6875-0}{0.1734}=3.965$$

v) We have $t_{14,1-0.001} = 3.787$ and $t_{14,1-0.0005} = 4.140$. Thus the p-value is between 0.1% and 0.2%. Accept the null hypothesis if the level of significance is lower than the p-value (which is usually not the case). Hence, we have strong evidence against the "no linear relationship" hypothesis. Note: exact p-value using computer package is 0.00070481.

b. i. Calculating the sums of squares in this question is done similarly to question 13. We have:

$$\begin{aligned}\text{SST} &= 17.8760\,,\\ \text{SSM} &= \sum n_i\left(\overline{y}_{i\cdot} - \overline{\overline{y}}\right)^2 = 4\sum\left(\overline{y}_i - \overline{\overline{y}}\right)^2 = 9.6709\,,\\ \text{SSE} &= \text{SST} - \text{SSM} = 17.8760 - 9.6709 = 8.2051\,.\end{aligned}$$

ii.

$$\begin{aligned}\widehat{\mu} &= 29.12/16 = 1.82\\ \widehat{\tau}_1 &= 2.73/4 - 1.82 = -1.1375\\ \widehat{\tau}_2 &= 6.26/4 - 1.82 = -0.255\\ \widehat{\tau}_3 &= 9.22/4 - 1.82 = 0.485\\ \widehat{\tau}_4 &= 10.91/4 - 1.82 = 0.9075\end{aligned}$$

iii. Company A: fitted value = 2.73/4 = 0.6825
Company D: fitted value = 10.91/4 = 2.7275

iv. Observed F statistic is $(9.6709/3)/(8.2051/12) = 4.715$ on (3,12) d.f..

v. From Formulae and Tables page 173 and 174 we observe that $F_{3,12}(4.474) = 2.5\%$ and $F_{3,12}(5.953) = 1\%$. Thus the p-value is between 0.025 and 0.01, so we have some evidence against the "no company effects" hypothesis. Note: exact p-value using computer package is 0.0213.

Multiple linear regression questions

1. In Table 3.4, the null hypothesis for `TV` is that in the presence of radio ads and newspaper ads, TV ads have no effect on sales. Similarly, the null hypothesis for `radio` is that in the presence of TV and newspaper ads, radio ads have no effect on sales. (And there is a similar null hypothesis for `newspaper`.) The low p-values of TV and radio suggest that the null hypotheses are false for TV and radio. The high p-value of newspaper suggests that the null hypothesis is true for newspaper.

2. The fitted model is given by

$$Y = 50 + 20 \text{ GPA} + 0.07 \text{ IQ} + 35 \text{ Gender} + 0.01\text{GPA} \times \text{IQ} - 10 \text{ GPA} \times \text{Gender}.$$

   For males, Gender = 0, so

$$Y = 50 + 20 \text{ GPA} + 0.07 \text{ IQ} + 0.01\text{GPA} \times \text{IQ}.$$

   For females, Gender = 1, so

$$Y = 85 + 10 \text{ GPA} + 0.07 \text{ IQ} + 0.01\text{GPA} \times \text{IQ}.$$

   a. False. For a fixed value of IQ and GPA, if GPA < 3.5, males earn less on average than females.

   b. False. For a fixed value of IQ and GPA, if GPA > 3.5, females earn less on average than males.

   c. True. For a fixed value of IQ and GPA, if GPA > 3.5, males earn more on average than females.

   d. False. See above.

3. a. I would expect the polynomial regression to have a lower training RSS than the linear regression because it could make a tighter fit against data that matched with a wider irreducible error $\mathbb{V}(\epsilon)$.

   b. I would expect the polynomial regression to have a higher test RSS as the overfit from training would have more error than the linear regression.

   c. Polynomial regression has lower train RSS than the linear fit because of higher flexibility: no matter what the underlying true relationship is the more flexible model will closer follow points and reduce train RSS. An example of this behaviour is shown on Figure 2.9 from Chapter 2.

   d. There is not enough information to tell which test RSS would be lower for either regression given the problem statement is defined as not knowing "how far it is from linear" If it is closer to linear than cubic, the linear regression test RSS could be lower than the cubic regression test RSS. Or, if it is closer to cubic than linear, the cubic regression test RSS could be lower than the linear regression test RSS. It is dues to bias-variance trade-off: it is not clear what level of flexibility will fit data better.

4. a. The design matrix is

$$\boldsymbol{X} = [\mathbf{1}_n \ \boldsymbol{x}] = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

b. The matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n\end{bmatrix}\begin{bmatrix} 1 & x_1\\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n\end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2\end{bmatrix} = n\begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \frac{1}{n}\sum_{i=1}^n x_i^2\end{bmatrix}$$

c. The matrix $\boldsymbol{X}^\top \boldsymbol{y}$ is

$$\boldsymbol{X}^\top\boldsymbol{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n\end{bmatrix}\begin{bmatrix} y_1\\ y_2 \\ \vdots \\ y_n\end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i\end{bmatrix}$$

d. Note: the inverse of a 2 × 2 matrix is given by:

$$M^{-1} = \begin{bmatrix} a & b \\ c & d\end{bmatrix}^{-1} = \frac{1}{\det(M)}\cdot\begin{bmatrix} d & -b \\ -c & a\end{bmatrix} = \frac{1}{ad-bc}\cdot\begin{bmatrix} d & -b \\ -c & a\end{bmatrix}$$

Using this and the result from 2. we have:

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - n^2\overline{x}^2}\cdot\begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n\end{bmatrix} = \frac{1}{s_{xx}}\cdot\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\overline{x} \\ -\overline{x} & 1\end{bmatrix}$$

e. Using the result of 3. and 4. we have:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} =&(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \frac{1}{s_{xx}}\cdot\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\overline{x} \\ -\overline{x} & 1\end{bmatrix}\begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i\end{bmatrix}\\
=&\frac{1}{s_{xx}}\begin{bmatrix} \sum_{i=1}^n y_i\cdot\frac{1}{n}\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i\cdot\overline{x} \\ -\sum_{i=1}^n y_i\cdot\overline{x} + \sum_{i=1}^n x_i y_i\cdot 1\end{bmatrix} = \begin{bmatrix} \overline{y}\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i\cdot\overline{x} \\ \sum_{i=1}^n x_i y_i - n\overline{xy}\end{bmatrix}\\
=&\begin{bmatrix} \overline{y}\left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right) - \left(\sum_{i=1}^n x_i y_i\cdot\overline{x} - n\overline{x}^2\overline{y}\right) \\ \sum_{i=1}^n x_i y_i - n\overline{xy}\end{bmatrix}\\
=&\begin{bmatrix} \overline{y}\left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right) - \overline{x}\left(\sum_{i=1}^n x_i y_i - n\overline{xy}\right) \\ \sum_{i=1}^n x_i y_i - n\overline{xy}\end{bmatrix} = \begin{bmatrix} \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x} \\ \frac{s_{xy}}{s_{xx}}\end{bmatrix}
\end{aligned}$$

5. Statement **(E)** is correct. Note that statement (A) is incorrect because, if food sales increases with one, the expected profit increases with $\widehat{\beta}_1 \cdot 10$ (note the difference in the scale of profit (thousands) and food sales (in **ten** thousands). Similarly, (B), (C) and (D) are incorrect.

6. Statement **(D)** is correct. We have $n = 25$ observations, $p = 3 + 1 = 4$ parameters (three explanatory variables and the constant), SST = 666.98, and SSM = 610.48. Thus we have:

$$\begin{aligned}\text{SSE} &= \text{SST} - \text{SSM} = 666.98 - 610.48 = 56.5\\ R_a^2 &= 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} = 1 - \frac{56.5/(25-4)}{666.98/(25-1)} = 1 - \frac{56.5/21}{666.98/24} = 0.903\,.\end{aligned}$$

7. Statement **(D)** is correct.

$$\begin{aligned}R^2 &\overset{*}{=} \frac{\text{SSM}}{\text{SST}} \overset{**}{=} \frac{\text{SST} - \text{SSE}}{\text{SST}} \quad \text{I and II correct}\\ &\overset{*}{=} \frac{\text{SSM}}{\text{SST}} \overset{**}{=} \frac{\text{SSM}}{\text{SSM} + \text{SSE}} \neq \frac{\text{SSM}}{\text{SSE}} \quad \text{because SSM} > 0\text{, III incorrect}\end{aligned}$$

* using definition of R^2 and ** using SST=SSM+SSE.

8. Statement **(B)** is correct.

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} = 1 - \frac{8525.3/(47-5)}{21851.4/(48-1)} = 1 - \frac{8525.3/42}{21851.4/46} = 0.563$$

9. Statement **(D)** is correct. Let $\boldsymbol{C} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and c_{33} the third diagonal element of the matrix $\boldsymbol{C}$. We have:

$$\text{SE}\left(\widehat{\beta}_2\right) = \sqrt{c_{33} \cdot s^2} = \sqrt{0.102446 \cdot 30106} = 55.535928$$

10. Statement **(C)** is correct. We have:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

In order to find the estimate of the parameter related to x_3 (having graduated from college) we need the fourth (note β_1 corresponds to the constant) row of the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and multiply that with the vector $\boldsymbol{X}^\top \boldsymbol{y}$. We have:

$$\widehat{\beta}_3 = \begin{bmatrix} -0.026804 & -0.000091 & 0.023971 & 0.083184 \end{bmatrix} \begin{bmatrix} 9,558 \\ 4,880,937 \\ 7,396 \\ 6,552 \end{bmatrix} = 21.953$$

Note that y is in hundreds of dollars, so having a graduated from college leads to $21.953 \cdot 100 = 2,195.3$ on the amount paid for a car.

11. Statement **(A)** is correct. We have that the distribution of $\widehat{\beta}_k$ for $k = 1, \ldots, p$ is given by:

$$\frac{\widehat{\beta}_k - \beta_k}{\mathrm{SE}\left(\widehat{\beta}_k\right)} \sim t_{n-p}$$

We have $p = 5$, and $n = 212$. Note, $n-p$ is large, thus the standard normal approximation for the student-t is appropriate (Formulae and Tables page 163 only shows a table for degrees of freedom up to 120 and ∞ = standard normal). We have $z_{1-0.05/2} = 1.96$. This provides the well-known rule of thumb that the absolute value of the T value should be larger than 2 for parameter estimates to be significant ($|T| > 2$). This is the case for all parameters.

12. Statement **(D)** is correct.

$$\begin{aligned}
\text{LIFE\_EXP} &= 48.24\ +\ 0.79\ \text{GNP}\ +\ 0.154\ \text{URBAN\%} \\
&= 48.24\ +\ 0.79\ \cdot 3 +\ 0.154\ \cdot 60 \\
&= 59.85
\end{aligned}$$

13. Statement **(C)** is correct.

    a. Can be done by the scatterplot, but a QQ-plot is better.

    b. Can be done by the scatterplot, but R^2 is better method.

    c. Is the correct one, need both the errors and the corresponding value of the endogenous variable.

    d. Should be by definition by selecting the LS estimator, so does not need to be tested.

    e. Errors should be independent of X not Y.

KNN question

1.

$$\begin{aligned}
\mathrm{EPE}_k(x_0) &= \mathbb{E}[(Y - \widehat{f}(x_0))^2 | X = x_0] \\
&= \mathbb{E}[(\epsilon + f(X) - \mathbb{E}(\widehat{f}(x_0)) + \mathbb{E}(\widehat{f}(x_0)) - \widehat{f}(x_0))^2 | X = x_0] \\
&= \mathbb{E}[\epsilon^2 + (f(x_0) - \mathbb{E}(\widehat{f}(x_0)))^2 + (\mathbb{E}(\widehat{f}(x_0)) - \widehat{f}(x_0))^2 - 2\epsilon(f(x_0) - \widehat{f}(x_0)) + \\
&\quad 2(\mathbb{E}(\widehat{f}(x_0)) - \widehat{f}(x_0))(f(x_0) - \mathbb{E}(\widehat{f}(x_0)))] \\
&= \mathbb{E}[\epsilon^2 + (f(x_0) - \mathbb{E}(\widehat{f}(x_0)))^2 + (\mathbb{E}(\widehat{f}(x_0)) - \widehat{f}(x_0))^2] \\
&= \sigma^2 + \left[f(x_0) - \frac{1}{k} \sum_{l \in N(x_0)} f(x_{(l)}) \right]^2 + \frac{\sigma^2}{k}
\end{aligned}$$

Note that ϵ is independent zero-mean noise, and:

$$\begin{aligned}\mathbb{E}(\mathbb{E}(\widehat{f}(x_0)) - \widehat{f}(x_0))(f(x_0) - \mathbb{E}(\widehat{f}(x_0))) &= \mathbb{E}[\mathbb{E}(\widehat{f}(x_0))f(x_0) - \widehat{f}(x_0)f(x_0) - (\mathbb{E}(\widehat{f}(x_0)))^2 + \\ &\quad \widehat{f}(x_0)\mathbb{E}(\widehat{f}(x_0))] \\ &= 0 \quad \text{since } f(x_0)\text{'s true value constant for fixed } X = x_0\end{aligned}$$

Applied Questions

1. a. Please install the package and load the data by the following command first.

```r
install.packages("ISLR2")
```

```r
library(ISLR2)
```

```r
fit <- lm(mpg ~ horsepower, data = Auto)
summary(fit)
```

```

Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5710  -3.2592  -0.3435   2.7630  16.9240 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049 
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i. Yes

ii. Very significant (p-value of $< 2.10^{-16}$)

iii. Negative

iv. `predict(fit, newdata = data.frame(horsepower = c(98)), interval = "confidence")`

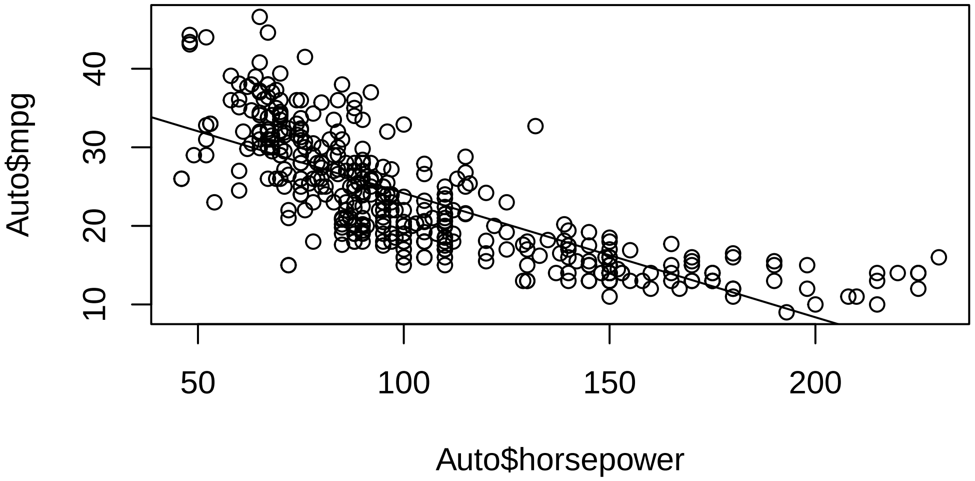
```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

```
predict(fit, newdata = data.frame(horsepower = c(98)), interval = "prediction")
```

```
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

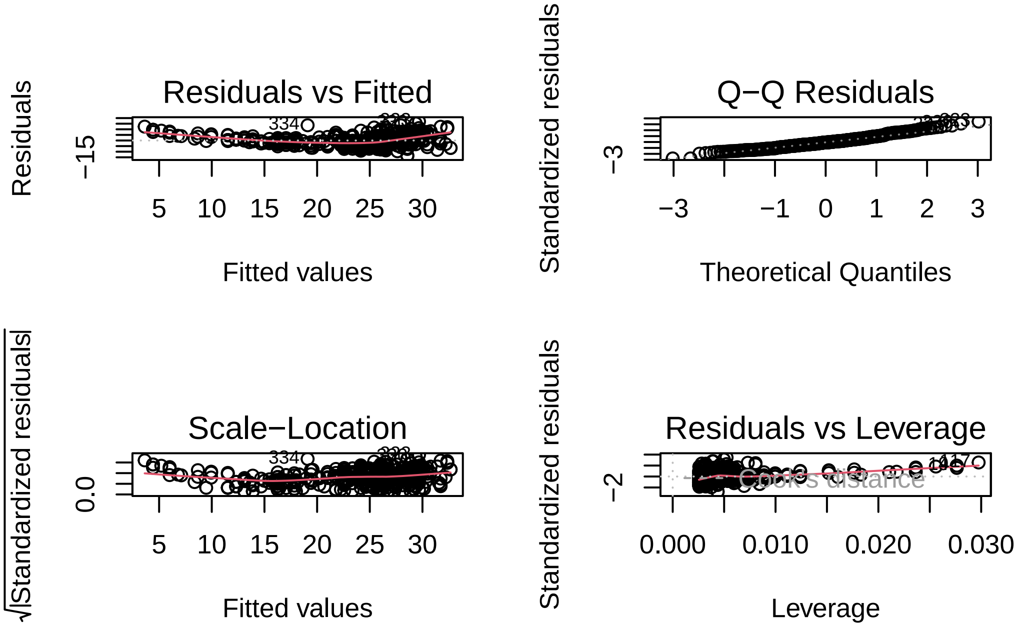
b.

```
plot(Auto$horsepower, Auto$mpg)
abline(a = fit$coefficients[1], b = fit$coefficients[2])
```



c.

```
par(mfrow = c(2, 2))
plot(fit)
```



There appears to be some trend in the residuals, indicating a linear fit is not appropriate.

2.
   ```r
   set.seed(1)
   x <- rnorm(100)
   y <- 2 * x + rnorm(100)
   ```

   a. ```r
      summary(lm(y ~ x + 0))
      ```

      ```
      Call:
      lm(formula = y ~ x + 0)

      Residuals:
          Min      1Q  Median      3Q     Max 
      -1.9154 -0.6472 -0.1771  0.5056  2.3109

      Coefficients:
        Estimate Std. Error t value Pr(>|t|)    
      x   1.9939     0.1065   18.73   <2e-16 ***
      ---
      Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      Residual standard error: 0.9586 on 99 degrees of freedom
      Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776 
      F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
      ```

      Result is fairly close to what's expected (2).

   b. ```r
      summary(lm(x ~ y + 0))
      ```

      ```
      Call:
      lm(formula = x ~ y + 0)

      Residuals:
          Min      1Q  Median      3Q     Max 
      -0.8699 -0.2368  0.1030  0.2858  0.8938 

      Coefficients:
        Estimate Std. Error t value Pr(>|t|)    
      y  0.39111    0.02089   18.73   <2e-16 ***
      ---
      Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      Residual standard error: 0.4246 on 99 degrees of freedom
      Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776 
      ```

```
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

Result is a bit far from what is expected (0.5), and it doesn't land in its 95% confidence interval.

c. The estimate in (a) is about 5 times the estimate in (b). The t-statistics, however, are identical.

d. See:

$$
\begin{aligned}
t &= \frac{\sum_i x_i y_i}{\sum_j x_j^2} \times \sqrt{\frac{(n-1)\sum_j x_j^2}{\sum_i (y_i - x_i\widehat{\beta})^2}} \\
&= \frac{\sqrt{n-1}\sum_i x_i y_i}{\sum_j x_j^2} \times \sqrt{\frac{\sum_j x_j^2}{\sum_i (y_i - x_i\widehat{\beta})^2}} \\
&= \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{\sum_j x_j^2}} \times \sqrt{\frac{1}{\sum_i (y_i - x_i\widehat{\beta})^2}} \\
&= \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{\sum_j x_j^2}} \times \sqrt{\frac{1}{\sum_i y_i^2 - 2y_i x_i\widehat{\beta} + x_i^2\widehat{\beta}^2}} \\
&= \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{\sum_j x_j^2}} \times \sqrt{\frac{1}{\sum_i y_i^2 - 2y_i x_i \frac{\sum_j x_j y_j}{\sum_k x_k^2} + x_i^2(\frac{\sum_j x_j y_j}{\sum_k x_k^2})^2}} \\
&= \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{(\sum_i y_i^2)(\sum_j x_j^2) - 2(\sum_i x_i y_i)^2 + (\sum_i x_i y_i)^2}} \\
&= \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{(\sum_i y_i^2)(\sum_j x_j^2) - (\sum_i x_i y_i)^2}}
\end{aligned}
$$

In R, this is written as

```
(sqrt(100 - 1) * sum(x * y)) / sqrt(sum(x^2) * sum(y^2) - sum(x * y)^2)
```

```
[1] 18.72593
```

This returns the same value as the t-statistic.

e. Due to the symmetry of x and y, we find we have the same formula as above. Hence the t-statistic is the same.

f.
```
fit <- lm(y ~ x)
fit2 <- lm(x ~ y)
summary(fit)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8768 -0.6138 -0.1395  0.5394  2.3462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03769    0.09699  -0.389    0.698
x            1.99894    0.10773  18.556   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9628 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
summary(fit2)
```

```
Call:
lm(formula = x ~ y)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90848 -0.28101  0.06274  0.24570  0.85736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03880    0.04266    0.91    0.365
y            0.38942    0.02099   18.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4249 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```