MSAN 601 - Homework 1

Andre Guimaraes Duarte

September 1, 2016

Question 1

In the normal error regression model, the error terms are assumed to be independent and identically distributed according to a normal distribution with mean 0 and variance σ^2: $e_i \sim N(0, \sigma^2)$ $\forall i$. This assumption greatly simplifies calculations for inference of new data points.

Question 2

We have $(t^*)^2 = (\frac{b_1}{s\{b_1\}})^2 = \frac{b_1^2}{\frac{MSE}{\sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{b_i^2 \sum_{i=1}^n (X_i - \bar{X})^2}{MSE}$.

We also have $F^* = \frac{MSR}{MSE} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{MSE} = \frac{\sum_{i=1}^n (b_0 + b_1 X_i - (b_0 + b_1 \bar{X}))^2}{MSE} = \frac{\sum_{i=1}^n (b_1 X_i - b_1 \bar{X})^2}{MSE} = \frac{b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{MSE}$.

Therefore, we have shown that for SLR, we get $(t^*)^2 = F^*$.

Question 3

Hypotheses are tested concerning the estimated values of the coefficients because the real values are not known. The estimators obtained with OLS are the best linear unbiased estimators that we can get.

Question 4

Normality of the error terms is not a necessary condition to obtain BLUE OLS estimators. Only the Gauss-Markov assumptions are needed:

- the population process is linear in parameters
- the data from the population comes from a random sample
- there is no prefect multicollinearity between independent variables
- zero conditional mean ($E[\epsilon_i | X_{i1} \dots X_{ik}] = 0$)
- homoscedasticity ($V(\epsilon_i | X_{i1} \dots X_{ik}) = \sigma^2$)
- no serial correlation ($cov(\epsilon_i, \epsilon_j) = 0 \forall i \neq j$)

We assume normality of the error terms in order to establish interval estimates and perform tests, i.e. when we want to make inference based on the data.

Question 5

The existence of a statistical relationship between the response variable Y and the explanatory variable X does not imply in any way that Y depends causally on X. No causal relationship can be determined from this result.

Question 6

65.33% of the variation in crime rate per capita is explained by the size of municipal police force.

Question 7

We have $n = 6690$, $SSTO = 36234$, and $SSR = 24332$. Since $SSTO = SSR + SSE$, we get $SSE = SSTO - SSR = 36234 - 24332 = 11902$. In addition, $MSE = \frac{SSE}{n-2}$, so we get $MSE = \frac{11902}{6690-2} \approx 1.78$.

Question 8

Computational part can be found in `HW1.R` file.

1 The OLS estimates are found using the formulas:

$b_0 = \bar{Y} - b_1 \cdot \bar{X}$ and $b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

Using `R`, we obtain:

$b_0 \approx -0.58$ and $b_1 \approx 15.04$.

2 The fitted regression line becomes $\hat{Y} = b_0 + b_1 X = -0.58 + 15.04X$.

3 We have $SSE = \sum_{i=1}^{n} e_i^2$ and $e_i = Y_i - \hat{Y}_i$. Using `R`, we get $SSE \approx 3416.38$.

4 In this context, b_0 does not provide any relevant information. In fact, X represents the number of photocopiers services, and Y the number of minutes pent by the service person. So for $X = 0$, it is only logical that $Y = 0$: if no photocopier is serviced, no minutes were spent by the service person.

5 The point estimate of the mean service time when 5 copiers are serviced is given by $\hat{Y}_5 = b_0 + b_1 \cdot 5$. Using `R`, we get $\hat{Y}_5 \approx 74.60$ minutes.

6 We wish to test the null hypothesis $H_0 : \beta_0 = 0$. In order to do so, we use a t-test. We first need to estimate $s^2\{b_1\} = \frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$, where $MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$. In this case, we have $MSE \approx 79.45$, and $s^2\{b_1\} \approx 0.23$.

We get the test statistic $t^* = \frac{b_1}{s\{b_1\}}$, which we compare to $t_{(1-\alpha/2;n-2)}$. Here, we have $t^* \approx 31.12$ and $t_{(1-\alpha/2;n-2)} \approx 2.02$. Since $31.12 > 2.02$, we reject the null hypothesis that $\beta_0 = 0$.

7 We now wish to test the nyll hypothesis $H_0 : \beta_1 \leq 1$. In order to do so, we use a t-test. We use the same test statistic t^* as the previous question, but this time we compare it to $t_{(1-\alpha;n-2)}$ since it is a two-sided test. Here, we have $t_{(1-\alpha;n-2)} \approx 1.68$. Since $31.12 > 1.68$, we reject the null hypothesis that $\beta_1 \leq 1$.

8 To get a confidence interval of the expected mean service time when 5 copiers are serviced, we first need to estimate $s^2\{\hat{Y}_5\} = MSE[\frac{1}{n} + \frac{(5-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}]$. Using `R`, we get $s^2\{\hat{Y}_5\} \approx 1.77$.

Therefore, the 95% confidence interval is given by $\hat{Y}_5 \pm t_{(1-\alpha/2;n-2)}s\{\hat{Y}_5\} \Leftrightarrow 74.60 \pm 2.68$. We get the confidence interval $[71.91; 77.28]$.

9 We wish to test the hypothesis $H_0 : \beta_1 = 0$ using an F-test. In order to do so, we first need to calculate the MSR. We have $MSR = \frac{SSR}{1} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$. Here, we have $MSR \approx 76960.42$.

We get the test statistic $F^* = \frac{MSR}{MSE}$, which we compare to $F_{(1,n-2)}$. Using `R`, we get $F^* \approx 968.66$ and $F^* \approx 5.39$. Since $968.66 > 5.39$, we reject the null hypothesis that $\beta_1 = 0$.

We can also see that $(t^*)^2 = 31.12^2 = 968.66 = F^*$.

10 Using `R`, we calculate the SSE for several values of (b_0, b_1):

- For $b_0 = 0$ and $b_1 = 15$, we get $SSE = 3424$
- For $b_0 = 0$ and $b_1 = 10$, we get $SSE = 40524$
- For $b_0 = 0$ and $b_1 = 20$, we get $SSE = 42124$
- For $b_0 = 1$ and $b_1 = 15$, we get $SSE = 3505$
- For $b_0 = -1$ and $b_1 = 15$, we get $SSE = 3433$
- For $b_0 = 10$ and $b_1 = 25$, we get $SSE = 207484$

We can see that the SSEs obtained in this fashion are always larger than the one computed in subquestion 3 using the OLS estimates (3416.38). The OLS estimates are indeed the best.

Question 9

Computational part can be found in `HW1.R` file.

1 Using `R`'s `lm` and `summary` functions, we find the OLS estimates for β_0 and β_1. We get $b_0 \approx -0.61$ and $b_1 \approx 0.11$.

2 The fitted regression equation is then given by $\hat{Y} = b_0 + b_1 X = -0.61 + 0.11X$.

3 We can see from the output of `summary` that the p-value for the intercept (b_0) is $0.658 > 0.05$. Therefore we cannot reject the null hypothesis that $b_0 = 0$. In this case, the intercept does not provide any relevant information.

4 We can see from the output of `summary` that the p-value for the slope (b_1) is $8.97e-07 < 0.05$. Therefore, we reject the null hypothesis that $b_1 = 0$: the slope is not null. This means that an increase of 1 in dietary fat consumption will lead to an expected increase of the death rate of 0.11.

5 Using the function `confint` in `R`, we get the following 95% confidence intervals for b_0 and b_1:

$b_0 \in [-3.53; 2.30]$ and $b_1 \in [0.08; 0.14]$.

Note: we can see that 0 is in the confidence interval for the intercept.

Question 10

In the context of SLR, when we regress X on Y, the independent variable is Y, and the dependent variable is X (so X is a function of Y), whereas if we regress Y on X, the independent variable is now X, and the dependent variable is now Y (Y is a function of X). Therefore, the parameter interpretation (and values) for both models is not the same. In fact, β_0 is the intercept of the model (the value of the explained variable when the explanatory variable is 0), and β_1 is the slope of the regression line (the expected variation in the explained variable when the explanatory variable is increased by 1 unit), but the interpretation varies depending on which is the independent variable and which is the dependent variable.

Using the `faithful` data set, we can easily observe this. If we regress `eruptions` on `waiting`, we get $b_0 \approx -1.87$ and $b_1 \approx 0.08$, whereas if we regress `waiting` on `eruptions`, we get $b_0 \approx 33.47$ and $b_1 \approx 10.73$.

Question 11

We can see that the F statistic in the summary of the two models in Question 10 are equal ($F^* = 1162.1$). Since n is the same for both models, we can conclude that the models are significant to the same statistical level. The coefficients obtained by regressing Y on X and X on Y are similar by means of a linear transformation.

Question 12

X = c(1:100) and Y = X ** 2 In this case, the relationship between Y and X is quadratic ($Y = X^2$), so doing a linear model is not the best idea. However, we can still run a SLR model, and get a value for the R^2. In fact, in this case, since we only have positive values of X, R^2 will probably be very high. We can verify using `R` that we get $R^2 \approx 0.94$.

X = c(1:100) and Y = X * 2 In this case, the relationship between Y and X is perfectly linear ($Y = 2X$), so doing a regression model is not useful (i.e. overkill) in this case. The R^2 value will be 1, which is easily verifiable with `R`.

Note: `R` even throws a warning message when calling the `summary` for the SLR model in this case: `essentially perfect fit`.

Question 13

Since the point of SLR is to minimize SSE, the presence of outliers can significantly alter the regression model in order to accomodate bigger errors. Outliers can have a disproportionate impact on the fitting of the regression line. However, one must have a valid proven reason to define an observation of the data set as an outlier and them from the study.