School of Mathematical Sciences

MTH5120 Statistical Modelling I — Mid-term Sample Test

This is a multiple choice test. There are 20 small problems. Choose only one statement for each problem, which you think is true, and mark it on the answer sheet by crossing a box. Each problem carries 5 marks. Total time for the test is 40 minutes. Calculators are not permitted.

Part 1

We will write the Simple Linear Regression Model (SLRM) for the the response variable Y and the explanatory variable X as

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

where Y_i denotes $Y|X = x_i$, β_0 and β_1 are unknown constant parameters. We will refer to this model throughout the test as the SLRM.

1. In the SLRM, the assumptions about the error, ε_i, mean that Y_i, $i = 1, \ldots, n$, are normally distributed with

   (a) $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\mathrm{var}(Y_i) = \sigma^2$, $\mathrm{cov}(Y_i, Y_j) = 0$ for $i \neq j$.
   (b) $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$, $\mathrm{var}(Y_i) = 0$, $\mathrm{cov}(Y_i, Y_j) = 0$ for $i \neq j$.
   (c) $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$, $\mathrm{var}(Y_i) = \sigma^2$, $\mathrm{cov}(Y_i, Y_j) = 0$ for $i \neq j$.
   (d) $\mathrm{E}(Y_i) = \varepsilon_i$, $\mathrm{var}(Y_i) = \sigma^2$, $\mathrm{cov}(Y_i, Y_j) = 0$ for $i \neq j$.

2. In the SLRM, the Least Squares Estimator of the parameter β_1 is given by

   (a) $\dfrac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$,

   (b) $\dfrac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}$,

   (c) $\overline{Y} - \beta_0 \overline{x}$,

   (d) $\overline{Y}$,

   where $\overline{Y}$ and $\overline{x}$ respectively denote the average of Y_i and of x_i, $i = 1, \ldots, n$.

3. The 95% confidence interval $[A, B]$ for $\mu_0 = \mathrm{E}(Y|X = x_0)$ means that

   (a) the probability that a true mean response at x_0 is between A and B is 0.95.
   (b) the probability that a true response at x_0 is between A and B is 0.95.
   (c) the probability that an estimate of a true mean response at x_0 is between A and B is 0.95.
   (d) the probability that an estimate of a true response at x_0 is between A and B is 0.95.

4. The Error Sum of Squares, SS_E, is defined as

   (a) $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$,

   (b) $\sum_{i=1}^{n}(\overline{Y} - \widehat{Y}_i)^2$,

   (c) $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$,

   (d) $\sum_{i=1}^{n}(Y_i - Y_j)^2$,

   where $\overline{Y}$ is the average of Y_i and $\widehat{Y}_i$ is the model fit at x_i, $i = 1, \ldots, n$.

5. The meaning of the error degrees of freedom in the regression ANOVA is

   (a) the number of independent pieces of information used to estimate MS_E.
   (b) the number of independent pieces of information used to estimate MS_R.
   (c) the number of dependent pieces of information used to estimate MS_E.
   (d) the number of dependent pieces of information used to estimate MS_R.

6. In the ANOVA table of a SLRM, the test function for the null hypothesis of non-significance of regression is

   (a) $F = \dfrac{MS_E}{MS_R}$ and it is distributed as $\mathcal{F}_{1,n-2}$.

   (b) $F = \dfrac{MS_R}{MS_E}$ and it is distributed as $\mathcal{F}_{1,n-2}$

   (c) $F = \dfrac{MS_R}{MS_E}$ and it is distributed as $\mathcal{F}_{2,n-1}$

   (d) $F = \dfrac{MS_E}{MS_R}$ and it is distributed as $\mathcal{F}_{2,n-1}$

   where MS_R denotes the mean regression sum of squares and MS_E denotes the mean error sum of squares.

7. Coefficient of Determination $R^2 = 0\%$ means that

   (a) all of the variability in the observations is due to the random error.
   (b) a quadratic model would fit the data better.
   (c) all observations fall exactly on the fitted line.
   (d) none of the variability in the observations is explained by the model fit.

8. In the SLRM, if there is no evidence, at a significance level $\alpha = 0.05$, to reject the null hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, then we can say that

   (a) the slope parameter is zero and the model is $Y_i = \beta_0 + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
   (b) the slope parameter is nonsignificant and the model is $Y_i = \beta_0 + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
   (c) the data do not contradict the null hypothesis when it is tested at the significance level $\alpha = 0.05$ and so a possible model is $Y_i = \beta_0 + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
   (d) the test is not valid.

9. In the SLRM, the test statistic for $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$ is

   (a) $T = \dfrac{\beta_0}{S/\sqrt{S_{xx}}}$,

   (b) $T = \dfrac{\widehat{\beta}_0}{S/\sqrt{S_{xx}}}$,

   (c) $T = \dfrac{\widehat{\beta}_0}{S/\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}}$,

   (d) $T = \dfrac{\beta_0}{S/\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}}$,

   where $\widehat{\beta}_0$ is the Least Squares Estimator of β_0, $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$ and $S = \sqrt{MS_E}$.

10. In the SLRM, the Lack of Fit sum of squares identity is

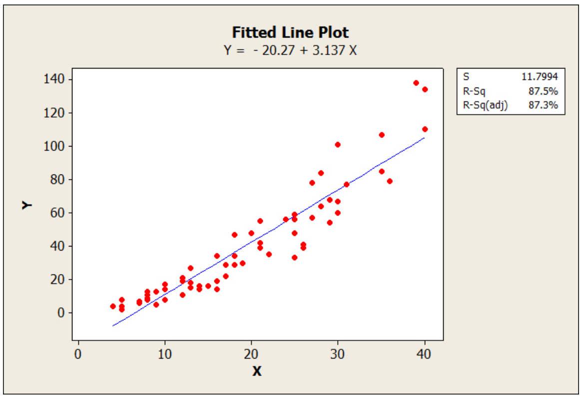
    (a) $SS_{LoF} = SS_{PE} + SS_E$,
    (b) $SS_{LoF} = SS_{PE} - SS_E$,
    (c) $SS_E = SS_{PE} - SS_{LoF}$,
    (d) $SS_E = SS_{PE} + SS_{LoF}$,

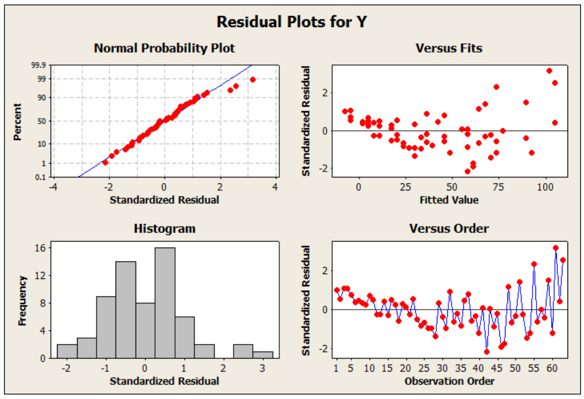
    where SS_{LoF} denotes the sum of squares for lack of fit, SS_{PE} denotes the sum of squares for pure error and SS_E denotes the error sum of squares.

Part 2

The rest of the problems refer to the MINITAB output for the following example. A company producing cars was testing one of their car models with respect to the stopping distance Y [feet] as a function of speed X [miles per hour]. Although several cars were used, there was only one driver and the data were collected in order of nondecreasing speed.

A SLRM was fitted. The data, the fitted line plot and the residual plots are shown below.





1. The Fitted Line Plot above shows that
   (a) the fitted line adequately represents the increasing stopping time as a function of speed.
   (b) the stopping time does not seem to increase linearly and its variability seems to increase as the speed increases.
   (c) there is a random scatter of the response values about the fitted line.
   (d) the only problem with the model fit are a few outliers.

2. The Normal Probability Plot shown above suggests that
   (a) the residuals are a sample from a distribution with light tails.
   (b) the residuals are definitely not a sample from a normal distribution.
   (c) apart from a few outliers all the observations lie close to the fitted distribution line.
   (d) the residuals are a sample from a log-normal distribution.

3. The Residuals Versus Fitted Values Plot shown above suggests that
   (a) the residuals are a sample from a normal distribution.
   (b) there is no problem apparent regarding the constant variance assumption.
   (c) the constant variance assumption may be violated.
   (d) the residuals are independently, identically distributed.

Below there is a part of the MINITAB output.

```
The regression equation is
Y = - 20.3 + 3.14 X

Predictor      Coef  SE Coef      T      P
Constant    -20.273    3.238  -6.26  0.000
X            3.1366   0.1517  20.68  0.000

S = 11.7994   R-Sq = 87.5%   R-Sq(adj) = 87.3%

Analysis of Variance
Source           DF     SS     MS       F      P
Regression        1  59540  59540  427.65  0.000
Residual Error   61   8493    139
  Lack of Fit    27   5253    195    2.04  0.025
  Pure Error     34   3240     95
Total            62  68033
```

4. The numerical output above

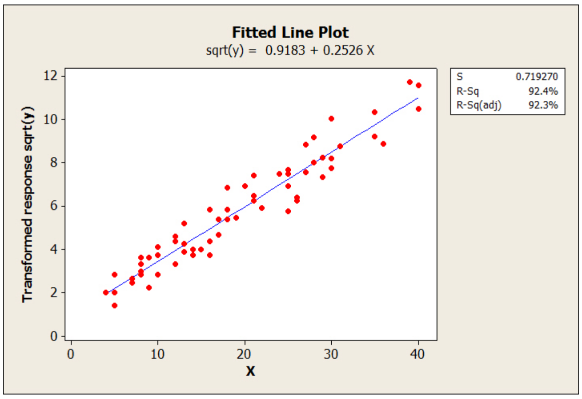
   (a) shows that the regression is highly significant, hence we can conclude that the stopping time is increasing linearly with speed.

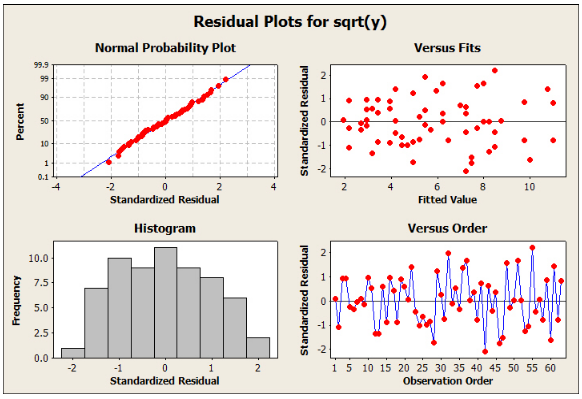
   (b) shows the reasonably high value of R^2 (87.5%) which allows us to ignore any violated model assumptions and to test the model parameters.

   (c) should not be strictly interpreted as the assumption of constant variance is unlikely to be met.

   (d) shows that there is no evidence to doubt that the SLRM is a true model.

The response was transformed to square root of y and a new SLRM was fitted. The plots are shown below.





5. The plots of the standardized residuals shown above suggest that

   (a) the transformation did not help to meet the model assumptions.

   (b) there is still clear curvature in the transformed stopping distance as the speed increases.

   (c) there is no contradiction to the model assumptions of normality and constant variance of the residuals.

   (d) there are apparent problems with outliers.

A part of the MINITAB output for the transformed response is given below.

```
The regression equation is
sqrt(y) = 0.918 + 0.253 X

Predictor       Coef    SE Coef      T      P
Constant      0.9183     0.1974   4.65  0.000
X           0.252568   0.009246  27.32  0.000

S = 0.719270   R-Sq = 92.4%   R-Sq(adj) = 92.3%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        1  386.06  386.06  746.22  0.000
Residual Error   61   31.56    0.52
  Lack of Fit    27   12.89    0.48    0.87  0.643
  Pure Error     34   18.67    0.55
Total            62  417.62
```

6. The model fit for the transformed data tells you that the increase in speed by one mile per hour would, on average,
   (a) increase the stopping distance by about 0.25.
   (b) increase the square root of the stopping distance by about 0.25.
   (c) decrease the square root of the stopping distance by about 0.25.
   (d) would not make any difference for the stopping distance.

7. The result of Lack of Fit test allows you to say that
   (a) there is no evidence in the data against the null hypothesis that the model is true.
   (b) there is no evidence in the data against the null hypothesis that the model is not true.
   (c) we can reject the null hypothesis that the model is true at the significance level $\alpha = 0.05$.
   (d) we can reject the null hypothesis that the model is not true at the significance level $\alpha = 0.05$.

8. The test of the hypothesis $H_0 : \beta_1 = 0$ indicates that the slope β_1 is
   (a) highly non-significant at a significance level $\alpha = 0.002$.
   (b) highly significant at a significance level $\alpha = 0.002$.
   (c) non-significant at a level significance $\alpha = 0.002$.
   (d) significant at any significance level α.

9. The value of the test statistic for testing non-significance of the intercept β_0 is
   (a) 0.87.
   (b) 4.65.
   (c) 27.32.
   (d) 746.22.

10. The coefficient of determination R^2 indicates that
    (a) about 92% of total variability in the observations is explained by the fitted model.
    (b) about 92% of total variability in the observations is explained by the residuals.
    (c) about 92% of total random error variability is explained by the lack of fit.
    (d) about 92% of total random error variability is explained by the pure error.