

## Midterm I - Practice Exam: Form A

**Instructions** (please read the following carefully before starting the exam):

- **Do not open** the exam unless you are explicitly instructed to do so.
- You may use **one** formula sheet (with writing on front and back), a standard scientific calculator (with **no** internet access), and pencil(s) - all fully functioning (no sharing or borrowing is allowed). **No** cell phones or any other electronic devices are allowed.
- You will need a copy of the Normal and  $t$  distribution tables. They are provided at the back of the exam in that order: Normal table (1 page) and  $t$  table (1 page). (These are Tables A.3 and A.4, respectively, from pages 495–496 of your textbook.)
- **Your exam should have 8 pages + 2 pages of distribution tables** (altogether 10 pages). The exam **duration is 75 minutes** (as in class hours: 2:20 p.m. - 3:35 p.m.).
- There are **16 questions worth 6 points each, for a total of 96 points. Everyone gets a bonus of 4 points!** Your score out of 96 will be increased by 4 and considered out of 100, i.e. if you score “ $s$ ” out of 96, your final score will be “ $s + 4$ ” out of 100.
- There is **NO partial credit. Only ONE answer is correct** for each of the questions. If you mark two or more options in the Scantron for any question, you will get 0 points for that question *irrespective* of whether one of those marked options is correct or not.
- **Mark your answers clearly** (and fully) with a number 2 pencil on your Scantron.
- **Mark which form** you have, **A or B**, on your Scantron.
- Also, **clearly bubble in the following on your Scantron:** *department (STAT)*, the *course number (212)*, the *section number (501)*, your **name and UIN** on the Scantron.
- You may write on your exam (and use the blank side of each page for scratch work). But none of this work will count. Your score will be based solely on the Scantron.
- No cheating or discussing or helping others in any way will be tolerated. If detected, everyone involved will get a grade of 0 on the exam. You **must** work alone.
- **Turn in both your exam and Scantron** when you are done. Bring your TAMU ID.
- Good luck!

**Note:** *Throughout this exam, unless specifically stated otherwise, all notations/symbols and abbreviations used have their usual standard meanings, as used/defined in the lecture notes.*

Read the following information to answer **Questions 1–4**:

Polynomial models of degree 1 to 6 were fitted to a set of  $n = 100$  observations of  $(x, y)$  pairs. Fitted models and a scatterplot of the data are shown on **pg. 8** and information obtained from **R** is given below. Use this information and the plots to answer **Questions 1–4** below.

Polynomial degree	SSR	AIC	BIC
1	149.85	478.36	486.18
2	156.43	479.36	489.78
3	401.20	434.34	447.36
4	401.64	436.23	451.86
5	418.89	433.90	452.14
6	418.92	435.90	456.74

$$\text{SST} = 808.9614$$

1. The value of adjusted  $R^2$  for the fourth degree polynomial model is closest to:
  - (a) 0.4753.
  - (b) 0.7572.
  - (c) 0.4826.
  - (d) 436.23.
  - (e) 0.3165.
  
2. Suppose you were presented with the plots of the third and fifth degree polynomials (as in **pg. 8**) and were asked to choose between the two, *strictly* on the basis of the plots. Then:
  - (a) The fifth degree polynomial is preferable because the two curves look almost the same, but the fifth degree model is more complex.
  - (b) The third degree polynomial is preferable because simpler models are always preferable.
  - (c) The third degree polynomial is preferable because the two curves look almost the same, but the third degree model is simpler.
  - (d) The fifth degree model is preferable because more complex models are always preferable.
  - (e) Not even William of Occam could decide between the two.

3. Based on the plots and all the other given information, the best model choice would be:

- (a) Either the polynomial of degree 1 or degree 2.
- (b) The polynomial of degree 3.
- (c) The polynomial of degree 4.
- (d) The polynomial of degree 5.
- (e) The polynomial of degree 6.

4. The estimate of the error standard deviation,  $\sigma$ , based on the third degree model is:

- (a)  $\sqrt{401.20/96}$ .
- (b)  $\sqrt{808.9614/96}$ .
- (c)  $\sqrt{(808.9614 - 401.20)/96}$ .
- (d)  $401.20/96$ .
- (e)  $(808.9614 - 401.20)/96$ .

5. Suppose that  $x$  is the length and  $Y$  is the weight of smallmouth bass (a type of fish) in a large Texas lake. A sample of smallmouth bass was caught at this lake and a simple linear regression model for  $Y$  versus  $x$  was fitted using the observed data. Using all the results we studied in class regarding inference for simple linear regression, the following two intervals were produced, one of which is a 95% confidence interval for the average weight (in pounds) of smallmouth bass that are 14 inches long, and the other is a 95% prediction interval for the weight (in pounds) of a particular smallmouth bass that is 14 inches long:

$$1.50 \pm 0.15 \quad \text{and} \quad 1.50 \pm 0.50.$$

Which of the following is correct?

- (a) This information implies that the average weight of a 21 inch long smallmouth bass is about 2.25 lbs.
- (b) The interval  $1.50 \pm 0.15$  is the prediction interval and the other is the confidence interval.
- (c) The interval  $1.50 \pm 0.15$  is the confidence interval and the other is the prediction interval.
- (d) Both (a) and (b) are correct.
- (e) Both (a) and (c) are correct.

**6.** Suppose the weights ( $Y$ , in pounds) of female Golden Retriever dogs between the ages of 3 and 12 months follow the simple linear regression model:

$$Y = 11.3 + 4.9x + \epsilon \quad \text{with } \epsilon \sim N(0, \sigma^2),$$

where  $x$  is their age (in months) and  $\epsilon$  has a Normal distribution with mean 0 and standard deviation  $\sigma = 5$  pounds for each  $x$ . Then, the probability that a 6 month old female golden retriever will weigh between 35 and 50 pounds is:

- (a) 0.0456.
- (b) 0.0957.
- (c) 0.1585.
- (d) 0.8415.
- (e) 0.9544.

**7.** In the setting of Question **6**, suppose that a female Golden Retriever is defined to be ‘large’ if she is at or above the 80th percentile of the weights for dogs of her age. Then, what is the smallest weight that qualifies a 9 month old female Golden Retriever as ‘large’?

- (a) 44.9 pounds.
- (b) 50.2 pounds.
- (c) 59.6 pounds.
- (d) 63.6 pounds.
- (e) 70 pounds.

Read the following information to answer **Questions 8–11**:

The following information is from a report on the determination of silver content of galena crystals grown in a closed hydrothermal system over a range of temperatures. Let  $x$  denote the crystallization temperature in degrees centigrade, and  $Y$  the silver content in mol%. A simple linear regression was fitted for  $Y$  versus  $x$ . The observed data are as follows:

$x$	398	292	352	575	568	450	550	408	484	350	503	600
$y$	0.15	0.05	0.23	0.43	0.23	0.4	0.44	0.44	0.45	0.09	0.59	0.63

Some summary statistics from the regression analysis are as follows:

$$\begin{aligned} \sum_{i=1}^{12} x_i &= 5530, & \sum_{i=1}^{12} y_i &= 4.13, & \sum_{i=1}^{12} (x_i - \bar{x})(y_i - \bar{y}) &= 155.4983, \\ \sum_{i=1}^{12} (x_i - \bar{x})^2 &= 113642, & \sum_{i=1}^{12} (y_i - \bar{y})^2 &= 0.39709, & \sum_{i=1}^{12} (y_i - \hat{y}_i)^2 &= 0.18432. \end{aligned}$$

Use this information to answer **Questions 8–11** below.

8. The least-squares estimates of the intercept and slope parameters, respectively, are:

- (a)  $\hat{\beta}_0 = 0.2164$  and  $\hat{\beta}_1 = 0.224$ .
- (b)  $\hat{\beta}_0 = 0.42$  and  $\hat{\beta}_1 = 0.05$ .
- (c)  $\hat{\beta}_0 = 0.2164$  and  $\hat{\beta}_1 = -0.224$ .
- (d)  $\hat{\beta}_0 = -0.2864$  and  $\hat{\beta}_1 = 0.001368$ .
- (e)  $\hat{\beta}_0 = -0.80$  and  $\hat{\beta}_1 = 0.001368$ .

9. To test  $H_0 : \beta_1 = 0$  against  $H_a : \beta_1 \neq 0$ , the test statistic to be used is:

- (a)  $\hat{\beta}_1 / (\hat{\sigma} \sqrt{1/12 + 1/113642})$
- (b)  $(\hat{\beta}_0 + \hat{\beta}_1) / \hat{\sigma}$ .
- (c)  $\hat{\beta}_1 / \hat{\sigma}$ .
- (d)  $\sqrt{12} \hat{\beta}_1 / \hat{\sigma}$ .
- (e)  $(337.108) \hat{\beta}_1 / \hat{\sigma}$ .

10. The proportion of variance in silver content that is explained by temperature is:

- (a) 0.668.
- (b) 0.956.
- (c) 0.810.
- (d) 0.464.
- (e) 0.536.

11. Given that the quantity:  $1/12 + (500 - 5530/12)^2 / 113642$  equals 0.0968, a 95% confidence interval for the *average* silver content of galena crystals grown in a closed hydrothermal system at 500 degrees centigrade is:

- (a)  $(\hat{\beta}_0 + 500\hat{\beta}_1) \pm 1.96\hat{\sigma}\sqrt{0.0968}$ .
- (b)  $(\hat{\beta}_0 + 500\hat{\beta}_1) \pm 1.96\hat{\sigma}\sqrt{1 + 0.0968}$ .
- (c)  $(\hat{\beta}_0 + 500\hat{\beta}_1) \pm 2.228\hat{\sigma}\sqrt{0.0968}$ .
- (d)  $(\hat{\beta}_0 + 500\hat{\beta}_1) \pm 2.228\hat{\sigma}\sqrt{1 + 0.0968}$ .
- (e) Cannot be determined from the information given.

**12.** A simple linear regression model is fitted to  $n = 20$  observations of  $(x, y)$  pairs, and the following summary statistics were computed:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = 202.71 \quad \text{and} \quad \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 157.25.$$

In this case, the estimated error variance  $\hat{\sigma}^2$  is:

- (a)  $157.25/202.71$ .
- (b)  $157.25/18$ .
- (c)  $(202.71 - 157.25)/18$ .
- (d)  $1 - 157.25/202.71$ .
- (e) Cannot be determined from the information given.

**13.** Two random variables  $X$  and  $Y$  have a population correlation coefficient,  $\rho$ , that equals 0. We may thus conclude that:

- (a)  $X$  and  $Y$  are independent.
- (b)  $X$  and  $Y$  are not independent.
- (c)  $X$  and  $Y$  could be independent.
- (d)  $E(X) = E(Y) = 0$ .
- (e) Donkeys can fly.

**14.** A researcher has data  $(x_1, y_1), \dots, (x_n, y_n)$  and wants to fit the following linear model:

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

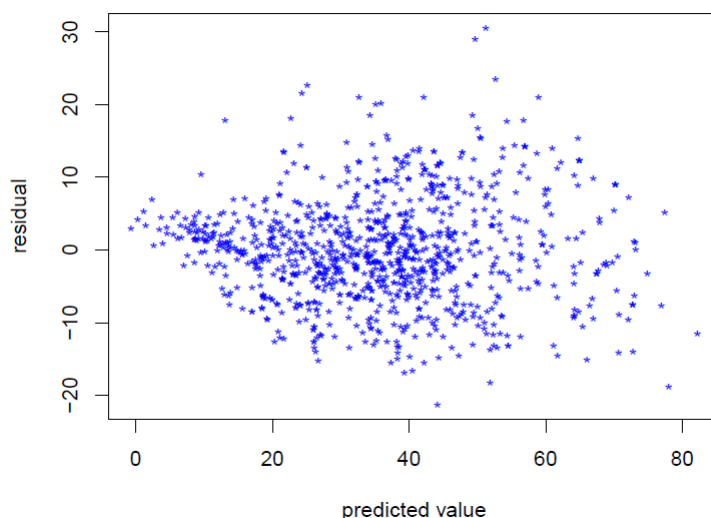
i.e. a linear regression model *without* an intercept parameter. Here,  $\epsilon_1, \dots, \epsilon_n$  are independent and each  $\epsilon_i \sim N(0, \sigma^2)$  as usual. Define  $\bar{y} = \sum_{i=1}^n y_i/n$  and  $\bar{x} = \sum_{i=1}^n x_i/n$ . Then, the correct expression for the least-squares estimate  $\hat{\beta}$  of  $\beta$  for this model is given by:

- (a)  $\hat{\beta} = \bar{x}/\bar{y}$ .
- (b)  $\hat{\beta} = \bar{y}/\bar{x}$ .
- (c)  $\hat{\beta} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2$ .
- (d)  $\hat{\beta} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2$ .
- (e) Cannot be determined from the information given.

**15.** Suppose that a response variable  $Y$  is such that  $Y = \alpha e^{\beta x} \epsilon$ , where  $\epsilon$  is a random variable and  $\alpha$ ,  $\beta$  and  $x$  are constants. Then, the correct expressions for  $\text{Var}(Y)$  and  $\text{Var}(\log Y)$  are:

- (a)  $\text{Var}(Y) = \text{Var}(\epsilon)$  and  $\text{Var}(\log Y) = \text{Var}(\log \epsilon)$ .
- (b)  $\text{Var}(Y) = \alpha e^{\beta x} \text{Var}(\epsilon)$  and  $\text{Var}(\log Y) = \text{Var}(\log \epsilon)$ .
- (c)  $\text{Var}(Y) = \alpha^2 e^{2\beta x} \text{Var}(\epsilon)$  and  $\text{Var}(\log Y) = \log \alpha + \beta x + \text{Var}(\log \epsilon)$ .
- (d)  $\text{Var}(Y) = \alpha^2 e^{2\beta x} \text{Var}(\epsilon)$  and  $\text{Var}(\log Y) = \log \alpha + \log(\beta x) + \text{Var}(\log \epsilon)$ .
- (e)  $\text{Var}(Y) = \alpha^2 e^{2\beta x} \text{Var}(\epsilon)$  and  $\text{Var}(\log Y) = \text{Var}(\log \epsilon)$ .

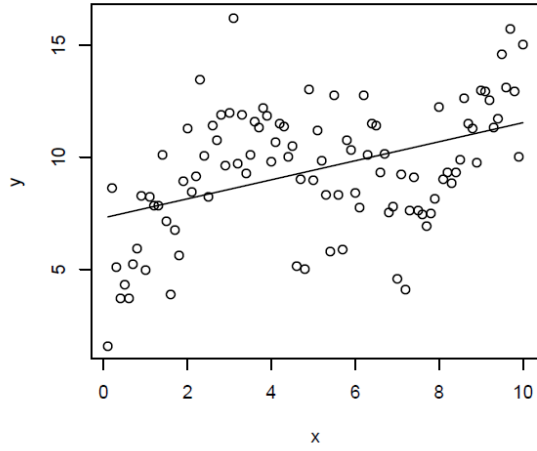
**16.** The residual plot below was obtained from a polynomial regression analysis.



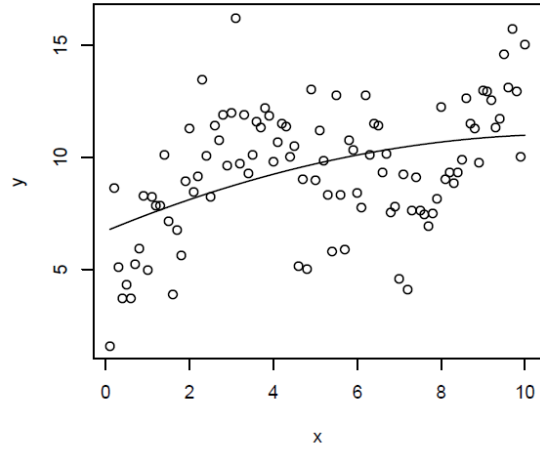
From this plot we can see that:

- (a) The residuals show no pattern.
- (b) The residuals are clearly Normally distributed.
- (c) The residual variance increases somewhat as the predicted value increases.
- (d) Both (b) and (c) are true.
- (e) The residuals were obviously computed incorrectly.

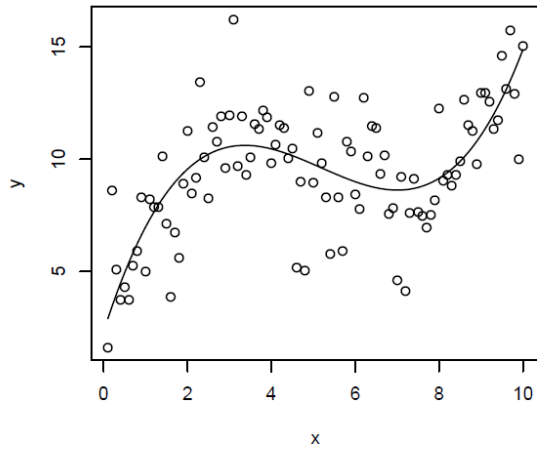
**Straight line**



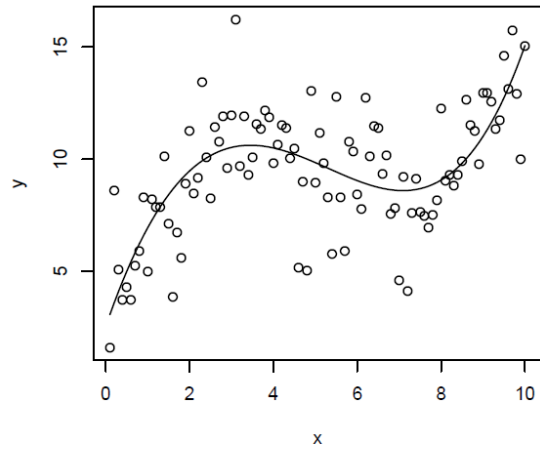
**Second degree polynomial**



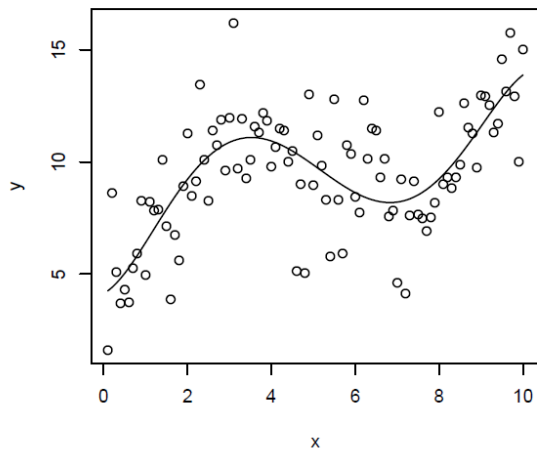
**Third degree polynomial**



**Fourth degree polynomial**



**Fifth degree polynomial**



**Sixth degree polynomial**

