

**Statistics 512: Review Problems for First Midterm Exam**  
Solutions

1. In a simple linear regression problem with  $n = 28$ , the following estimates were obtained:  $b_0 = 3.30$ ,  $b_1 = 0.46$ ,  $s\{b_0\} = 0.014$ ,  $s\{b_1\} = 0.036$ ,  $s = 0.1385$ ,  $\bar{X} = -0.0773$ ,  $SS_X = 14.598$ . Values of  $X$  were in the range  $-1.43$  to  $0.683$ .

- (a) Write the simple linear regression model. Include the distributional assumption.

**Solution:**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \epsilon_i \sim^{iid} N(0, \sigma^2)$$

- (b) Write the estimated regression line and the estimated error variance.

**Solution:**

$$\hat{Y} = 3.30 + 0.46X, s^2 = 0.1385^2 = \mathbf{0.01918}.$$

- (c) If appropriate, estimate the mean value of  $Y$  for the given  $X$ . If not appropriate, say why.

- i.  $X = 0$ .

**Solution:**

$$\hat{Y} = 3.30 + 0.46 \times 0 = \mathbf{3.30}$$

- ii.  $X = 2$ .

**Solution:** Not appropriate, since outside the scope of the model ( $X = -1.43$  to  $0.683$ ).

- (d) Give a 95% confidence interval for the slope of the regression line.

**Solution:**

$$\begin{aligned} CI &= b_1 \pm t_{n-2}(1 - \alpha/2)s\{b_1\}; t_{26}(0.975) = 2.056 \\ &= 0.46 \pm 2.056 \times 0.036 \\ &= 0.46 \pm 0.074 = [\mathbf{0.386}, \mathbf{0.534}] \end{aligned}$$

(e) Give a 99% prediction interval for the next value of  $Y$  when  $X = 0.5$ .

**Solution:**

$$\begin{aligned}PI &= \hat{Y}_h \pm t_{n-2}(1 - \alpha/2) \times s\{pred\} \\ \hat{Y}_h &= 3.30 + 0.46 \times 0.5 = 3.30 + 0.23 = 3.53 \\ s\{pred\} &= s \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{SS_X}} \\ &= 0.1385 \sqrt{1 + \frac{1}{28} + \frac{(0.5 - (-0.0773))^2}{14.598}} \\ &= 0.1385 \sqrt{1 + 0.0357 + 0.0228} \\ &= 0.1385 \times 1.029 = 0.142 \\ t_{26}(0.9995) &= 2.779 \\ PI &= 3.53 \pm 2.779 \times 0.142 = 3.53 \pm 0.53 = [\mathbf{3.13}, \mathbf{3.93}].\end{aligned}$$

2. Short answer questions. Unless stated otherwise, each part is unrelated.

(a) If the design matrix has dimensions  $20 \times 4$ , what is the dimension of the error vector?

**Solution:** The error vector has dimension  $\mathbf{20 \times 1}$ .

(b) If the power for detecting a difference of means is 0.03, what can you say about the test's  $\alpha$  level?

**Solution:** The  $\alpha$  level is given by the probability of being in the rejection region when the difference between hypothesized value and the true value is zero. The power in this case is the smallest it could be (see Simple Linear Regression notes); so, in general, the power must be as big or bigger than the  $\alpha$  level. The  $\alpha$  level of the test is, thus, **less than 0.03**.

(c) A particular linear model has parameter values  $\beta_0 = 20$ ,  $\beta_1 = 0.5$ , and  $\sigma^2 = 4$ . Assuming all the linear model assumptions are true, what's the probability that an observation  $Y$  would be greater than 19.4, given that  $X = 6$ ?

**Solution:** The response values  $Y$  have a normal distribution with mean  $20 + 0.5 \times 6 = 23$  and standard deviation equal to 2. Thus, we have that

$$\begin{aligned}P(Y_{new} > 19.4) &= P\left(Z > \frac{19.4 - 23}{2}\right) \\ &= P(Z > -1.8) \\ &= \mathbf{0.964}.\end{aligned}$$

(d) A quarter of the data fall outside the prediction region, what's the approximate confidence level for the prediction intervals?

**Solution:** The  $\alpha$  level is approximately 0.25, so the confidence level is approximately **75%**.

- (e) For a model with 1 predictor and 40 observations, if  $R^2$  is 0.3 what is the test statistic for the ANOVA  $F$  test?

**Solution:** With 1 predictor,  $p = 2$  including the intercept.

$$\begin{aligned} F &= \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \\ &= (0.3/1)/(0.7/38) \\ &= \mathbf{16.28} \end{aligned}$$

- (f) The  $\alpha$  levels of 2 confidence intervals are 0.05 and 0.01. Find a **lower bound** on the overall coverage rate for the two confidence intervals. (In other words, the probability that the confidence intervals cover **both** their values is at least what?)

**Solution:**

$$\begin{aligned} P(\text{both COR}) &\geq 1 - P(I_1\text{INC}) - P(I_2\text{INC}) \\ &= 1 - 0.05 - 0.01 = \mathbf{0.94} \end{aligned}$$

- (g) For the confidence intervals in the previous problems, what would you need to know to get the **exact** probability of joint coverage?

**Solution:** You need to know the **probability that neither interval covers its true value**. Thus,

$$\begin{aligned} P(\text{both COR}) &= 1 - [P(I_1\text{INC}) + P(I_2\text{INC}) - P(\text{both INC})] \\ &= 0.94 + P(\text{both INC}). \end{aligned}$$

- (h) You run a least squares regression on SAS and get an  $SSE$  value of  $20 m^2$ . Your officemate claims to be able to find a line (using the same data) with an  $SSE$  of  $15 m^2$ . Should you believe your officemate? Why or why not?

**Solution:** SAS runs least squares regression; as such, the SAS value of the  $SSE$  is the least possible value for this data (fit with a regression line). Thus, the officemate is mistaken.

- (i) For a particular  $X$  value, the prediction interval has length 10; the confidence interval has length 5. If  $SS_X$  is 3, what's the length of the confidence interval for the slope?

**Solution:** Let  $\ell(PI)$  and  $\ell(CI)$  be the length of the prediction interval and the length of the confidence interval, respectively. Their values are  $2t_{cs}\{pred\} = 2s\sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{SS_X}}$  and  $t_{cs}\{\hat{Y}_h\} = 2s\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{SS_X}}$ . We have that

$$\begin{aligned} \ell(PI)^2 - \ell(CI)^2 &= 4t_c^2(s^2\{pred\} - s^2\{\hat{Y}\}) \\ &= 4t_c^2s^2 \end{aligned}$$

The length of the confidence interval for the slope is  $2t_{cs}/\sqrt{SS_X}$ . Since  $\ell(PI)^2 - \ell(CI)^2 = 100 - 25 = 75$ ,

$$2t_{cs}/\sqrt{SS_X} = \sqrt{\frac{75}{3}} = \sqrt{25} = 5.$$

- (j) For a given set of data, the optimal Box-Cox transformation (of the response) is at  $\lambda = -0.5$ . You decide to use the “suggested” transformation and take the reciprocal of the response. What’s the optimal Box-Cox transformation for the **transformed** data?

**Solution:** The optimal transformation of  $Y$  is

$$Y_{opt} \leftarrow 1/\sqrt{Y}.$$

The suggested transformation of  $Y$  is

$$Y_{sug} \leftarrow 1/Y.$$

To get from  $Y_{sug}$  to  $Y_{opt}$ , it is necessary to use the square root transformation:  $Y_{opt} = \sqrt{Y_{sug}}$ . The Box-Cox procedure will find that the “optimal” transformation of the transformed data occurs at  $\lambda = \frac{1}{2}$ .

- (k) Using the matrices from least squares estimation, what are the entries of the vector  $\mathbf{X}'\mathbf{e}$ ?

**Solution:**

$$\begin{aligned} \mathbf{X}'\mathbf{e} &= \mathbf{X}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \\ &= \mathbf{X}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\ &= (\mathbf{X}' - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\ &= (\mathbf{X}' - \mathbf{X}')\mathbf{Y} \\ &= \mathbf{0}'\mathbf{Y} \\ &= \mathbf{0}, \end{aligned}$$

where  $\mathbf{0}$  is the  $p \times 1$  vector of 0's.