KEY

Stat 328 - Fall 2004 - EXAM 1

The total number of points in this exam is 100. Each definition is worth 5 points and each part of Problems 2 and 3 is worth 8 points. Partial credit will be given so write everything down! I can give you no points if you give me just the final answer and it happens to be wrong. Please read the entire exam before you begin and start with those questions for which you know the answer. No extensive calculations are necessary so make sure that you make use of the enclosed output where appropriate. Good luck!

Problem 1 - Definitions: Define in your own words (no formulas or symbols!) and briefly (no more than 50 words) the following terms:

(a) Regression slope
   The slope in a regression line represents the expected change in the response when the predictor is increased by one unit.

(b) Median
   For some random variable $y$, it is the value such that the probability that $y$ is larger than or smaller than it equals $0.5$.

(c) $(1 - \alpha)\%$ confidence interval for a mean $\mu$:
   The set of values of a variable y that cover the true population mean $\mu$ with probability $1 - \alpha$.

(d) Sample estimator
   A function of sample observations that is used to estimate an unknown population parameter.

(e) Sampling distribution
   Distribution of likely values of the sample statistic. It arises from conceptually repeating the experiment a very large number of times.
**Problem 2:** Economists at Iowa State University wished to investigate whether unemployment rate can be used to predict the number of individuals without health insurance at the State level. To do so, they collected data on the proportion of unemployed persons and the number of individuals (in thousands) without health insurance in each of the 50 states in the country during the year 2003. They then fitted a simple linear regression model to estimate the linear association between the two variables.

(a) Please refer to the scatter plot enclosed. Interpret what you see.

The number of uninsured persons appear to increase with unemployment rate. The association between the two variables appears to be linear, and thus a simple linear regression model might be adequate to summarize the association. If we sketch a regression line through the observations, we observe that none of the states appear to deviate significantly from what might be expected # of uninsured people given a certain unemployment rate.

(b) Write down a simple linear regression model that describes the (potential) association between number of uninsured persons and unemployment rate.

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, \ldots, 50 \]

\[ y_i = \# \text{ uninsured (in 1000s)} \text{ in } i \text{th state} \]

\[ x_i = \text{unemployment rate in } i \text{th state} \]

(c) What are the usual assumptions behind a simple linear regression model?

- \( \epsilon_i \sim N(0, \sigma^2) \)
- Observations are independent.
(d) Please refer to the JMP or to the SAS output that is enclosed. *Notins* stands for number of uninsured persons (in thousands) and *Unemploy* stands for unemployment rate. Using information from the output, please fill in the blanks below:

\[ b_0 = -81 \quad \text{and} \quad \hat{\sigma}_{b_0} = 66.3 \]
\[ b_1 = 163 \quad \text{and} \quad \hat{\sigma}_{b_1} = 10.5 \]

(e) In English, interpret the values of the estimated intercept \( b_0 \) and slope \( b_1 \).

The intercept \( b_0 \) represents the expected number of uninsured persons in a state with 0 unemployment.

The slope \( b_1 \) is the expected change in the number of uninsured persons (in 1000s) if unemployment increases by 1%.

(f) Provide a 90% confidence interval for the true slope \( b_1 \). (\( z \) and \( t \) tables are enclosed). Interpret the result.

90% CI for \( b_1 \):

\[ b_1 \pm t_{0.05, n-2} \frac{\hat{\sigma}_{b_1}}{\sqrt{n-2}} \]
\[ 163 \pm t_{0.05, 48} 10.5 = 163 \pm 1.684 \times 10.5 \]
\[ = (145.3, 180.7) \]

(g) The Governor of a state with unemployment rate equal to 4% in 2003 wishes to predict the number individuals (in thousands) that might have no health insurance in 2004 should unemployment rate decrease by half a percent. Please compute the expected number of uninsured individuals for the Governor.

\[ \hat{y} = b_0 + b_1x \]

For \( x = 3.5\% \):

\[ \hat{y} = -81 + 163 \times 3.5 \]
\[ = 489.5 \]

About half a million persons in the state would be expected to lack health insurance if unemployment decrease to 3.5% in 2004.
(h) Are the predictions you might compute from your model trustworthy? Obtain the coefficient of variation in your analysis and interpret the result.

\[
CV = \left( \frac{\bar{y}}{\bar{y}} \right) \times 100 = \left( \frac{\text{MSE} / \bar{y}}{\bar{y}} \right) \times 100 = \left( \frac{98.94}{921.59} \right) \times 100 = 10.7\%
\]

The standard error of 98.94 is small relative to the magnitude of the response variable. Thus, we trust that our predictions are reasonably accurate.

Problem 3: A marketing firm conducted a study to determine whether the price of a product is associated to demand for the product. An experiment was conducted in which the same product was offered to 10 independent groups of 100 potential buyers at prices increasing by $0.25 increments. Individuals in each group were then asked whether they would purchase the product or not. Therefore, experimental data consisted of 10 pairs of price of the product (x) and number (out of 100 possible) of potential customers who expressed an interest in purchasing the product (y). Data are presented in the table below.

<table>
<thead>
<tr>
<th>Focus group</th>
<th>Price of product ($)</th>
<th>No. who would purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.00</td>
<td>70</td>
</tr>
<tr>
<td>2</td>
<td>2.25</td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>2.50</td>
<td>65</td>
</tr>
<tr>
<td>4</td>
<td>2.75</td>
<td>67</td>
</tr>
<tr>
<td>5</td>
<td>3.00</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>3.25</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>3.50</td>
<td>42</td>
</tr>
<tr>
<td>8</td>
<td>3.75</td>
<td>45</td>
</tr>
<tr>
<td>9</td>
<td>4.00</td>
<td>37</td>
</tr>
<tr>
<td>10</td>
<td>4.25</td>
<td>34</td>
</tr>
</tbody>
</table>

A simple linear regression model was fitted to the data. The fitted regression equation was

\[
\hat{y} = 110.59 - 18.01x.
\]

The standard errors of the intercept and the slope were, respectively, 5.36 and 1.67.
(a) Investigators wish to test the hypothesis that price and demand for the product are linearly associated at the 95% level. Using a two-tailed test of hypothesis, set up the null and alternative hypotheses, compute a test statistic for the test, define the critical region and reach a conclusion. In English, please interpret your results.

\[ H_0: \beta_1 = 0 \]
\[ H_a: \beta_1 \neq 0 \]

\[ t = \frac{b_1 - 0}{\hat{\sigma}_{b_1}} = -10.78 \]

\[ \text{Critical region: } t \leq -t_{\alpha/2, n-2} \text{ and } t \geq t_{\alpha/2, n-2}, \]

\[ \text{for } t_{\alpha/2, n-2} = t_{0.025, 8} = 2.306. \]

Since \( t < -2.306 \), we reject \( H_0 \).

The data strongly suggest that the true association between price and demand for the product is different from zero.

(b) Repeat the hypothesis test above, but now carry out a one-sided test of the hypothesis that the true slope is negative. As in part (a), set up the hypotheses, calculate the test statistic, define the critical region and reach a conclusion. Use the same confidence level as in part (a). In English, please interpret your results.

\[ H_0: \beta_1 = 0 \]
\[ H_a: \beta_1 < 0 \]

\[ t = \frac{b_1}{\hat{\sigma}_{b_1}} = -10.78 \]

\[ \text{Critical region: } t \leq -t_{\alpha, n-2} \]

\[ \text{for } t_{\alpha, n-2} = t_{0.05, 8} = 1.86 \]

Since \( t < -1.86 \) we reject \( H_0 \) and conclude \( H_a \). Data suggest that the true slope \( \beta_1 \) is negative, so price and demand are negatively associated.