CS6220 — Fall 2015

Midterm

Thursday October 29, 2015

Time: 1 hour 40min

Name (please print): _____

Show all your work and calculations. Partial credit will be given for work that is partially correct. Points will be deducted for false statements, even if the final answer is correct. Please circle your final answer where appropriate.

This exam is closed-book. You may consult one page with your personal notes. Calculators are permitted.

Honor code: I promise not to cheat on this exam. I will neither give nor receive any unauthorized assistance. I will not to share information about the exam with anyone who may be taking it at a different time. I have not been told anything about the exam by someone who has taken it earlier.

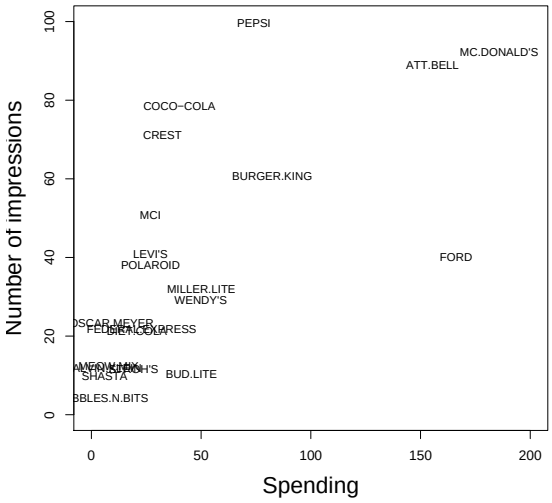
Signature: _____ Date: _____

The 1-0.05/2 quantile of the Normal distribution is 1.959964
The 1-0.05/2 quantile of the Student distribution with 19 df is 2.093024

Question	Possible Points	Actual Points
1	30	
2	12	
3	24	
4	24	
5	6	
6	6	

1. A New York ad-testing company Video Board Tests, Inc., conducted an annual survey to study the success of TV commercials as function of investment. It selected **21** most outstanding recent commercials shown on TV. Then it asked regular product users to cite which commercials they had seen for that product category in the past week.

   The resulting dataset appeared in the Wall Street Journal. It contains TV advertising budget (in millions of dollars), the number of retained impressions of the advertisement, and the name of the company that sponsored the advertisement. The data are plotted below.



   (a) **(6 pts)** A simple linear regression model is considered for use with the data. Describe the model and the assumptions. Comment whether the assumptions are plausible for this dataset. (*Note:* the plausibility of an assumption may or may not be diagnosed from the scatterplot.)

   Answer:

$$\texttt{Impress}_i = \beta_0 + \beta_1 * \texttt{Spend}_i \ + \varepsilon_i \ \textit{where}$$

   ε_i are independent, and distributed according to a Normal distribution with mean 0 and equal variance.

   Assumptions:

   i. *Linear relationship between* `Spend` *and* `Impress`. *A non-linear term may be needed.*
   ii. *The variance is constant across values of* `Spend`. *It is not verified for this dataset.*
   iii. *Non-systematic errors. Cannot tell from this chart.*
   iv. *The error terms are independent and normally distributed. We can not conclude from the plot, however the assumption may be plausible according to the domain knowledge.*

(b) **(6 pts)** The analysts decided to work with the model that uses `log(Spend)` and `log(Impress)`. A partial R output of the model fit is given below. Test for association between `log(Impress)` and `log(Spend)`. State the null and the alternative hypotheses, and your conclusion at the confidence level of 95%.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2999     0.4236
logSpend      0.6135     0.1191

Residual standard error: 0.581
Multiple R-squared:  0.5829
```

Answer:

$$H_0: \ \beta_1 = 0 \ \textit{vs} \ H_a: \ \beta_1 \neq 0$$

$t = \frac{0.6135}{0.1191} = 5.1511 > t_{19}^{1-0.05/2} = 2.093024$ *(see first page)*
Conclusion: we reject H_0 of no association at the confidence level of 95%.

(c) **(6 pts)** Report and interpret the 95% confidence interval for β_1.
Answer:

$$\beta_1 \in (0.6135 \pm 2.093024 \cdot 0.1191) = (0.3642, 0.8627)$$

The probability that the interval contains β_1 is 95%. The confidence interval for β_1 does not contain 0, and this is an additional evidence against H_0 in the tests above.

(d) **(6 pts)** Report the Pearson correlation coefficient between `log(Spend)` and `log(Impress)`.

Answer:

$r = \sqrt{R^2} = \sqrt{0.5829} = 0.7634$, *indicating moderate assosiation.*

(e) **(6 pts)** What are the advantages and disadvantages of characterizing the association between `log(Spend)` and `log(Impress)` in terms of hypothesis testing, confidence intervals and Pearson correlation?

Answer:

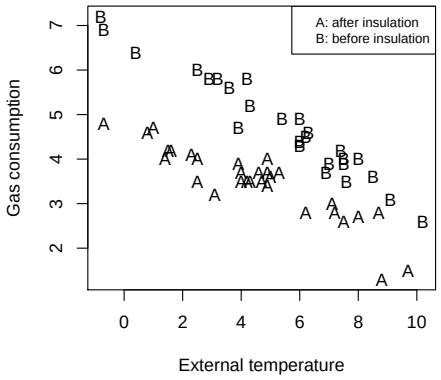
Hypothesis testing *leads to a binary decision of whether there is evidence against the null hypothesis of no association. Advantage: it takes into account the standard error of the parameter, and quantifies the associated uncertainty in terms of a p-value.*

Confidence interval *provides a plausible range of values that contains the parameter. It has the same advantage as the hypothesis testing, i.e. it takes into account the standard error of the parameter and quantifies the associated uncertainty. It is equivalent to hypothesis testing: when the confidence interval does not cover 0 it equivalent to rejecting the null hypothesis in hypothesis testing.*

Pearson correlation *has the advantage of quantifying the uncertainty on a simple scale (between -1 and 1). Disadvantage: it does not express the associated uncertainty, does not have a decision cutoff, and cannot be extended beyond pairwise associations.*

2. Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption (in 1000s of cubic feet) and average external temperature (in Fahrenheit) at his house in South-East England for 56 consecutive weeks, before and after installing a house insulation.

   The figure below shows the plot of the observed data. The goal of the study is to assess the effect of the insulation on gas consumption.



   (a) **(6 pts)** The researcher considers to model gas `Consumption` as a linear regression with `Temperature` as the predictor. In addition, he considers that a separate linear relationship may exist before and after installing the insulation. Therefore, for each data point he creates an additional variable `Insulation`, that takes the value of 1 for "after installing the installation", and 0 otherwise.

   State the linear model which will allow us to express the two linear relationships in one model. State the assumptions, and interpret the parameters.

   Answer: *Linear model:*

   $$\texttt{Consumption}_i = \beta_0 + \beta_1 \texttt{Temperature}_i + \beta_2 \texttt{Insulation}_i + \beta_3 \texttt{Temperature}_i \cdot \texttt{Insulation}_i + \varepsilon_i$$

   $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

   Parameter interpretation:

   - β_0*: intercept when insulation* $= 0$
   - $\beta_0 + \beta_2$*: intercept when insulation* $= 1$
   - β_1*: change in the consumption mean when the temperature increases 1 unit and insulation* $= 0$*.*
   - $\beta_1 + \beta_3$*: change in the consumption mean when the temperature increases 1 unit and insulation* $= 1$*.*

   (b) **(6 pts)** Explain why it is advantageous to consider this model, as opposed to to separate regression models (one for before the insulation, and the other after).

   Answer:

   The model expresses all the available data. If we can assume that the variance of the error is the same for both before and after the insulation, then we use all the data to estimate this variation. This increases the sensitivity of detecting the associations between temperature and consumption.

3. A marketing firm is investigating the patterns of new car purchase in a particular geographic region. A random sample of 33 families from this region was selected. A follow-up interview 12 months later was conducted to record the annual family income (X_1, in thousand dollars), and whether the family purchased a new car (Y=1) or not (Y=0). The output of a statistical model used to analyze the data is given below.

```
glm(formula = Y ~ income, family = binomial(link = "logit"), data = insurance)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.98079    0.85720  -2.311   0.0208 *
income       0.04342    0.02011   2.159   0.0308 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)
```

(a) **(6 pts)** State the model, and interpret the parameters.

Answer:

$$Y_i \overset{ind}{\sim} Bernouilli(\pi_i),\ log\left\{\frac{P(Y=1)}{P(Y=0)}\right\} = \beta_0 + \beta_1 X_1$$

where β_1 is the slope of the relationship between the log odds of purchasing a car, and β_0 is the intercept.

β_1 is also interpreted as the log of the odds ratio for a unit change in `income`.

(b) **(6 pts)** Estimate the odds ratio for the annual income and its 95% confidence interval. Interpret the result.

Answer:

The odds ratio is $\exp\{0.0434\} = 1.044$. *A 95% CI for the odds ratio is*

$$\exp\{0.0434 \pm z_{1-0.05/2}0.0201\} = (\exp\{0.0040\}, \exp\{0.0827\}) = (1.004, 1.08)$$

(c) **(6 pts)** As an alternative approach, the analyst grouped the individuals into 6 levels of income, calculated the number of car purchases per income group, and conducted the Pearson χ^2 test for association between the income and the car purchase. How many terms will be added up to the Pearson χ^2 test statistic, and what are the associated degrees of freedom? (For the degrees of freedom, provide the exact number).

Answer:

Test statistics is $\chi^2 = \sum_{i=1}^{6} \sum_{j=0}^{1} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$. *Since there are 6 income groups, and each group will have two response, either* $'yes'$ *or* $'no'$. *12 items are added to calculate the statistics.*

The degree of freedom will be $(6-1)\cdot(2-1) = 5$.

(d) **(6 pts)** Describe the advantages and disadvantages of the approach in (a-b) and of the approach in (c), for studies of associations between the income and the car purchase.

Answer:

The logistic regression is advantageous because it expresses the relationship between income and purchase with only two parameters. It has an easy interpretation, and can be extended to other predictors. Grouping the data in a table and analyzing it with the χ^2 *test adds more parameters, leads to overfitting and is not easily extended. However, this approach makes fewer assumptions.*

4. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)+\lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ. For parts (a) through (d) below, indicate which of (i) through (v) is correct. Justify your answer.

(a) **(6 pts)** As we increase λ from 0, the training residual sum of squares will:

i Increase initially, and then eventually start decreasing in an inverted U shape

ii Decrease initially, and then eventually start increasing in a U shape

iii Steadily increase

iv Steadily decrease

v Remain constant

Answer:

iii The Residual Sum of Squares is minimized at the ordinary logistic regression model. By adding a penalty term to the model, there will be a tradeoff between RSS and the penalty term. As optimization point of view, the penalty term constraints βs. Increasing λ , force βs twards zero more and more which cause theterm of RSS to increase steadily.

(b) **(6 pts)** Repeat (a) for test residual sum of squares.

Answer:

ii The λ is usually selected through the cross validation to minimize the optimization term on the training set. For the test set, the RSS is decreased initially as λ increased but the penalty term starts to become dominant and then eventually the RSS increases.

(c) **(6 pts)** Repeat (a) for variance.

Answer:

iv As λ increase, at the cost of bias, the variance decreases.

(d) **(6 pts)** Repeat (a) for (squared) bias.

Answer:

Increasing λ caused the β s shift more and more towards zero, and hence leads to greater bias.

5. **(6 pts)** What is the difference between a standard deviation and a standard error?

Answer:

The standard deviation quantifies the variability in the population, and the standard error quantifies the uncertainty in a parameter of a population. The former will not decrease with the increased sample size, while the latter will decrease. For example, if one were to take multiple samples from this population, there will be variations among the means. Then by estimating the variation in the means among samples which is called the standard error of the estimate of the means. There is a relation between the standard error of the sample mean ($\sigma_{\bar{x}}$), which depends on both the standard deviation (σ) and the sample size (n) as follows:

$$\textit{Standard Error of the Mean } (\sigma_{\bar{x}}) = \frac{\textit{Standard Deviation } (\sigma)}{\sqrt{n}}$$

6. **(6 pts)** Consider a simple linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \overset{\textit{independent, identical}}{\sim} (0, \sigma^2)$$

where the assumption of Normally distributed residuals is not appropriate (but the assumption of constant variance is appropriate). Describe a bootstrap-based algorithm that you would use to obtain the standard error of the predicted value of Y given a value X_h of interest.

Answer:

Repeat B times:

(a) Select randomly with replacement n pairs (X_i, Y_i). Denote them (X_i^*, Y_i^*)

(b) Fit linear regression to these pairs, and obtain $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$

(c) Calculate the predicted value for X_h: $\hat{Y}_j^* = \hat{\beta}_0^* + \hat{\beta}_1^* X_h$

Report the standard deviation of the B values $\hat{Y}_j^*$