QMB Exam 1--Solution Fall 2017

1. New Product Testing: In a random sample of 28 male and 36 female customers, 20 male and 25 female customers prefer the new version of your company's product to the old one. Set up the null and alternative hypotheses most appropriate for each of the following questions and conduct the hypothesis tests. Provide appropriate interpretations and/or list any relevant assumptions.

a. Do over 60% of male customers prefer the new version?

H_0: 60% (or less) of male customers prefer the new version. ρ_m=0.6

H_a: Over 60% of male customers prefer the new version. ρ_m>0.6

$\hat{\rho}_m = 20/28 = 0.714$. Under H_0 $se(\hat{\rho}_m) = \sqrt{0.6 \times 0.4/28} = 0.093$, $z = (0.714 - 0.6)/0.093 = 1.234$

$P(Z > 1.234 \mid \rho = 0.6) = 0.109$. So, if the population proportion were 0.6, we would find a sample proportion as large as 0.714 or larger 109 times per 1000 times we conduct this experiment. If your type 1 error tolerance is above that, reject H_0 in favor of H_a. Otherwise, do not reject H_0. To sum up, there is some evidence more than 60% of males prefer the new version, but it is not as strong as we would want to support a firm conclusion.

b. Do over 60% of female customers prefer the new version?

H_0: 60% (or less) of male customers prefer the new version. ρ_f=0.6

H_a: Over 60% of male customers prefer the new version. ρ_f>0.6

$\hat{\rho}_f = 25/36 = 0.694$. Under H_0 $se(\hat{\rho}_f) = \sqrt{0.6 \times 0.4/36} = 0.082$ $z = (0.694 - 0.6)/0.082 = 1.157$,

$P(Z > 1.157 \mid \rho = 0.6) = 0.124$. So, if the population proportion were 0.6, we would find a sample proportion as large as 0.694 or larger only 124 times per 1000 times we conduct this experiment. If your type 1 error tolerance is above that, reject H_0 in favor of H_a. Otherwise, do not reject H_0. To sum up, there is evidence more than 60% of females prefer the new version, but it is not as strong as we would want like to support a firm conclusion.

c. Do the same proportion of male and female customers prefer the new version?

H_0: The same proportions of male and female customers prefer the new version. $\rho_m = \rho_f$

H_a: The proportions of male and female customers that prefer the new version differ. $\rho_m \neq \rho_f$

$\hat{\rho}_m = 20/28 = 0.714$ and $\hat{\rho}_f = 25/36 = 0.694$. We need the pooled proportion to calculate the standard error of the difference under the null, $\hat{\rho}_c = 45/64 = 0.703$ Under H_0

$se(\hat{\rho}_m - \hat{\rho}_f) = \sqrt{0.703 \times 0.297/36 + 0.703 \times 0.297/28} = 0.115$, $z = (0.714 - 0.694)/0.115 = 0.172$

$P(|Z| > 0.172 \mid \rho_m = \rho_f) = 0.863$. If the population proportions were the same, we would find a difference of sample proportions of 0.02 or larger 863 of 1000 times we conduct this experiment. There is no reason at all to think the proportions differ meaningfully.

d. Assuming male and female customers have the same preferences, is the proportion of all customers that prefer the new version over 60%?

H_0: 60% (or less) of customers prefer the new version. ρ=0.6

H_a: Over 60% of customers prefer the new version. ρ>0.6

$\hat{\rho} = 45/64 = 0.703$. Under H_0 $se(\hat{\rho}) = \sqrt{0.6 \times 0.4/64} = 0.061$, $z = (0.703 - 0.6)/0.061 = 1.684$,

$P(Z > 1.806 \mid \rho = 0.6) = 0.046$. If the population proportion were 0.6, we would find a sample proportion as large as 0703 or larger only 46 times per 1000 times we conduct this experiment. If your type 1 error tolerance is above that, reject H_0 in favor of H_a. To sum up, there is relatively strong evidence more than 60% prefer the new version, if we are willing to maintain the assumption the preferences of males and females are essentially the same.

2. Regression Mechanics. A company collects data on the March totals of number of units sold and minutes of local television advertising for 50 local outlets. Their analyst regressed the natural log of quantity sold (q) on the natural log of advertising (a), using robust standard errors. Various summary statistics are in Table 2.1 to the right. The Stata regression output is in Table 2.2 below.

Table 2.1: Summary Statistics

n	50
$\bar{a}$	5.94
$\bar{q}$	6.99
$\sum_i (a_i - \bar{a})^2$	12.03
$\sum_i (q_i - \bar{q})^2$	5.49
$\sum_i (a_i - \bar{a})(q_i - \bar{q})$	6.39
$\sum_i e_i^2$	2.10
$\sum_i (a_i - \bar{a})^2 e_i^2$	0.4133

Table 2.2 Stata Output

```
Linear regression                               Number of obs     =         50
                                                R-squared         =     0.6178
                                                Root MSE          =     0.20916

------------------------------------------------------------------------------
             |               Robust
           q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           a |   .5312361   .0545477     9.74   0.000     .4215606    .6409116
       _cons |   3.835784   .3203253    11.97   0.000     3.191727    4.479841
------------------------------------------------------------------------------
```

a-e: Using the summary statistics provided, show how each of the following elements of the Stata output table are calculated.

a. Root MSE. $\sqrt{\sum_i e_i^2 / (N-K-1)} = \sqrt{2.1/48} = 0.209$

b. R-squared. $R^2 = SSM/SST$, $SSM = SST - SSR = 5.94 - 2.1 = 3.39$, $R^2 = 3.39/5.49 = 0.618$

c. The advertising coefficient. $\sum_i (a_i - \bar{a})(s_i - \bar{s}) / \sum_i (a_i - \bar{a})^2 = 6.39/12.03 = 0.531$

d. The standard error of the advertising coefficient.

$$\sqrt{\left(N/(N-K-1)\right)\sum_i (a_i - \bar{a})^2 e_i^2 / \left(\sum_i (a_i - \bar{a})^2\right)^2} = \sqrt{(50/48)0.4133/12.03^2} = 0.055$$

e. The 95% confidence interval for the advertising coefficient.

Roughly speaking, the coefficient plus or minus 2 standard errors: $0.531 \pm 2 \times 0.055$. To be a bit more precise, instead of using 2, use the value of the t distribution for 48 degrees of freedom that cuts of 2.5% in each tail, which is 2.011.

f. Carefully explain the interpretation of the p-value associated with the advertising coefficient.

If there is truly no predictive relationship between advertising and quantity in the population, and one were to repeatedly draw random samples of 50 from the population and run this regression, one would get a coefficient as large or larger than in this sample less than one time in 10,000. Thus, there is almost no chance that in the actual population there is no relationship between advertising and quantity.

g. Remembering variables are in natural logs, carefully interpret the advertising coefficient.

Acceptable: When the log of advertising increases by 1, the expected increase in the log of quantity is 0.53.

Better: Noting that log point changes are approximately percentage changes, when advertising increases by 10%, the expected increase in quantity is 5.3%.

h-l: Imagine an outlet with log quantity and log advertising initially equal to the sample average, at which log advertising increases by one standard deviation.

h. What are the initial values of quantity and advertising (not the log values)?

Since *a* and *q* are log values, I'll let *A* and *Q* be values before taking logs.

$A_0 = e^{5.94} = 378.3$, $Q_0 = e^{6.99} = 1084.7$

i. What is the standard deviation of log advertising? $\sqrt{\sum_i (a_i - \bar{a})^2 / (N-1)} = \sqrt{12.03/49} = 0.496$

j. If log advertising increases one standard deviation, what is the new value of advertising?

$a_1 = 5.94 + 0.496 = 6.436$, $A_1 = e^{6.436} = 624$ Rounding may lead to slightly different numbers.

k. If log advertising increases one standard deviation, what is predicted log quantity?

$\hat{q}_1 = 3.836 + 0.531 \times 6.436 = 7.25$

l. If log advertising increases one standard deviation, what is predicted quantity?

$Q_1 = e^{7.25} = 1413$

m-n: The coefficients of the best fitting regression line are found by solving the normal equations, $\sum_i x_{ik} e_i = 0$, for each of the coefficients indexed by k.

m. For *k*=0, the normal equation shows the intercept is determined so the line passes through the sample mean. Explain intuitively why the best fitting line passes through the sample mean.

If the regression is not right at least on average (sum of residuals is 0), the fit would could be improved (sum of squared residuals reduced) by shifting it to be right on average.

n. For any other coefficient, explain intuitively why the sum of the products of the variable and the residual must be 0 for the best fitting line.

If residuals tended to be positive (negative) at high values of X and negative (positive) at low values of x, so that the model is under (0ver) predicting at high values of X and under (over) predicting at low values of X, prediction would be improved by making use of that information by increasing (decreasing) the slope of the line, yielding a better fit.