17.871 - PS4 Solutions

Mike Sances

Due: April 13, 2012

1. We use the formulas:

   (a) The model is

   $$proud_i = \alpha + \beta_1 dem_i + \beta_2 black_i + \epsilon_i$$

   We can plug in values from the variance-covariance matrix into our formula for regression coefficients. Let $\hat{\beta}_k^b$ be the bivariate coefficient for the k'th variable and let $\hat{\beta}_k^m$ be the multivariate coefficient. Then we have,

$$\begin{aligned}
\hat{\beta}_1^b &= Cov(dem, proud)/Var(dem) \\
&= 0.46/4.01 \\
&= 0.11 \\
\hat{\beta}_2^b &= Cov(black, proud)/Var(black) \\
&= 0.08/0.19 \\
&= 0.42 \\
\hat{\beta}_1^m &= \frac{Cov(dem, proud)}{Var(dem)} - \hat{\beta}_2^m \frac{Cov(dem, black)}{Var(dem)} \\
&= \hat{\beta}_1^b - \hat{\beta}_2^m \frac{0.33}{4.01} \\
&= 0.11 - \hat{\beta}_2^m * 0.08 \\
\hat{\beta}_2^m &= 0.42 - \hat{\beta}_1^m \frac{Cov(dem, black)}{Var(black)} \\
&= 0.42 - \hat{\beta}_1^m 0.33/0.19 \\
&= 0.42 - \hat{\beta}_1^m * 1.73 \\
&= 0.42 - \Big(0.11 - \hat{\beta}_2^m * 0.08\Big) * 1.73 \\
&= 0.42 - 0.19 + \hat{\beta}_2^m 0.14 \\
\Leftrightarrow 0.86\hat{\beta}_2^m &= 0.23 \\
\Leftrightarrow \hat{\beta}_2^m &= 0.27 \\
\Rightarrow \hat{\beta}_1^m &= 0.11 - 0.27 * 0.08 \\
&= 0.09
\end{aligned}$$

(b) The bivariate and multivariate coefficients for any X_1 in a regression with two independent variables X_1 and X_2 will differ if and only if

$$\beta_2^m Cov(X_1, X_2)/Var(X_1) \neq 0$$

In this problem we determined that both $\hat{\beta}_2^m$ – the "effect" of being black on being proud – and the covariance between being black and being Democratic, are both non-zero. Further we can determine the sign of the difference – that is, if it is positive or negative – by analyzing the signs on these two terms. Here, β_2^m is positive and so is $Cov(X_1, X_2)$. Thus the bivariate coefficient on X_1 will overstate the true relationship.

2. Since the coefficients don't change across the models–that is, the coefficient on newspaper is roughly the same in model (1) and (3), and similarly with the coefficient on wheel in (2) and (3)–we infer that wheel and newspaper are basically uncorrelated. But we can go further and determine the sign of this small correlation. Recall from the previous question that

$$\hat{\beta}_1^b - \hat{\beta}_1^m = \hat{\beta}_2^m Cov(X_1, X_2)/Var(X_1)$$

Plugging in from the table, we have,

$$\begin{aligned}
-0.036 - -0.037 &= 0.053 * Cov(news, wheel)/Var(wheel) \\
\Leftrightarrow 0.001 &= 0.053 * \frac{Cov(news, wheel)}{Var(wheel)} \\
\Leftrightarrow \frac{0.001}{0.053} &= \frac{Cov(news, wheel)}{Var(wheel)}
\end{aligned}$$

On the left hand side we have something greater than zero. On the right hand side we have a fraction where the denominator is non-zero by definition (i.e., variances are never negative). The numerator is the covariance, which always has the same sign as the correlation. Therefore we conclude there is a slight, *positive* correlation between reading newspapers and watching Wheel of Fortune.

3. The answer here depends on how you interpreted "variable." To answer this question, recall the Central Limit Theorem, which tells us that for (basically) any random variable X that has population mean μ and population variance σ^2, a random sample of size n from the population distribution of X will be normally distributed with mean μ and variance σ^2/n. Then we know that the first estimator has variance $\sigma^2/100$ and the second estimator has variance $\sigma^2/1000$. So it is in fact 100 times more variable, if by varable we are referring to variances. What about if we are talking standard errors? Well, in that case the standard error of the first estimator is $\sigma/10$ and that of the second estimator is $\sigma/100$. So if you phrased your answer such that you made clear you interpreted the question in terms of standard errors, I'll give credit.

4. There is both an easy and a hard way to go about this question. The harder way is to use z-scores; the easier way is to use some useful facts about normal distributions.

(a) First the "easy" way. The slides on inference tell us that on a normal curve, $p\%$ of the curve lies within k standard deviations for $k = 1, 2, 3$. For a sampling distribution we substitute the standard error for the standard deviation, since we are dealing with a normal sampling distribution. In our case we have a population standard deviation of 16 and a sample of 100, so the standard error is $\sigma/\sqrt{n} = 16/10 = 1.6$. We have a mean of 100, and we want to know how much of the curve is above 105, or above about $5/16 = 3.125$ standard errors. Thus we know – returning to the picture in the slides – that more than 99% of the curve lies within $[-3.125\sigma, 3.125\sigma]$. To be greater than 105 would mean to be outside this interval, an event which occurs with probability $< 1\%$.

Now let's use z-scores (the "hard way"). We subtract the (known) population mean and divide by the standard error,

$$\begin{aligned} z &= \frac{105 - 100}{16/\sqrt{100}} \\ &= 5/1.6 \\ &= 3.125 \end{aligned}$$

If we plug this into a table for the standard normal, we get$\Pr(z \leq 3.125) = 0.999$ which means that $\Pr(z \geq 3.125) = 1 - 0.999 < 1\%$.

(b) Calculate the interval $[90, 110]$ in terms of standard errors. To be at 90 is to be 10 away from the mean or $10/1.6 = 6.25$ standard deviations. We know from looking at the figure that basically 0% of the sample lies within -6σ and 6σ away from the mean. I won't cover the "hard way", but the logic is similar; if you tried using z-scores, I will be lenient in grading.

(c) To be above 140 is to be more than 40 above the mean or more than $40/1.6 = 25$ standard deviations above the mean. Again the chances of this happening are about 1 in infinity. So basically none of the sample will be above 140.

5. Convert the categories to "real" numbers:

Table 1:

Amount	n
0	150
50	50
150	40
350	30
750	20
1500	5

(a) We have a sample of (totaling the second column) $n = 295$. We then compute the average using the formula for expectation:

$$E[X] = \sum_{i=1}^{6} x_i p_i$$

$$
\begin{aligned}
&= \frac{1}{295}(150*0+50*50+40*150+30*350+20*750+5*1500)\\
&= 41,500/295\\
&= 140.68
\end{aligned}
$$

So the average spending was a little more than 100 dollars. The standard error is the sample standard deviation s divided by $\sqrt{n}$. The formula for s is:

$$
\begin{aligned}
s &= \sqrt{\frac{\sum_{i=1}^{6}(x_i-\bar{x})^2}{n-1}}\\
&= \sqrt{\frac{150*(0-140.68)^2+50*(50-140.68)^2+...+5*(1500-140.68)^2}{295-1}}\\
&= \sqrt{\frac{40,655.92}{294}}\\
&= 269.55
\end{aligned}
$$

Thus the standard error of the mean is $269.55/\sqrt{295} = 15.69$.

6. Since this is a binary variable we know that the mean is $p = .46$ and the sample standard error is

$$
\begin{aligned}
\sqrt{\frac{p*(1-p)}{n}} &= \sqrt{.25/1200}\\
&= 0.014
\end{aligned}
$$

Then relying on the central limit theorem we can construct a confidence interval of the standard error multiplied by the 95% critical value of 1.96:

$$
\begin{aligned}
CI &= p \pm 1.96*0.014\\
&= p \pm 0.027\\
&= [0.432,\ 0.487]
\end{aligned}
$$

7. The CLT tells us that any two random samples from a population with the same mean μ and standard deviation σ will both have sample means μ and standard deviations σ/n. Thus the averages and standard errors should not differ: they should be the same. The pictures are a "red herring." They show us slightly different population distributions, but in fact the CLT doesn't depend on the population distribution. Rather, the CLT tells us what the sampling distribution will look like *for any* distribution.

8. What is the sample probability?

$$23,209/36,000 = .644$$

Then what's the probability we get a value as high as this, given the true population mean is 56%? We can do this with a z-score:

$$\frac{.644-.56}{\sqrt{\frac{.644(1-.644)}{36,000}}} = 33.20$$

So very unlikely we'd get a probability this high. Thus the sample number is "too high."

We can also figure this out the following way: we know the sample standard error is $\sqrt{\frac{.644(1-.644)}{36{,}000}} = 0.0025 = 0.25\%$. We have obtained a sample mean that is $8.4\%/0.2\% \approx 33$ standard errors away from the mean. Examining the nromal curve, we see that being 33σ away from the mean occurs with a probability that is much smaller than 1%, given that the true turnout is .56. So again, "too high."

9. While we did not cover t-tests in lecture, they are mentioned in the readings by Kellstedt and Whitten. The general formula for the t-statistic is,

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{SE(\bar{Y}_1 - \bar{Y}_2)}$$

and in our case, we replace $\bar{Y}_k$ with p_k, since the sample mean for proportions is p. In this case $p_1 = 11,109/16,776 = 0.662$ and $p_2 = 13,164/19,224 = 0.684$. For the denominator, we use the formula for the standard error for the difference in proportions from the lecture slides on inference:

$$\begin{aligned} SE(\bar{Y}_1 - \bar{Y}_2) &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\ &= \sqrt{\frac{0.662 * (1-0.662)}{16,776} + \frac{0.684 * (1-0.684)}{19,224}} \\ &= 0.00495 \\ \Rightarrow t &= (0.662 - 0.684)/0.00495 \\ &= -4.44 \end{aligned}$$

Some of you asked about degrees of freedom. It turns out our sample is so large here we don't need to worry about that. It also turns out that asymptotically, the t distribution is the normal distribution. Thus we know from the previous problem on the normal curve that we reject the null hypothesis that the two proportions (turnout levels) are equal at the less than 1% level. That is, under the null hypothesis that the proportions are equal, there is a less than 1% chance that we would get a t value as large as -4.44.