QMB 3200 Dewey Fall 2018

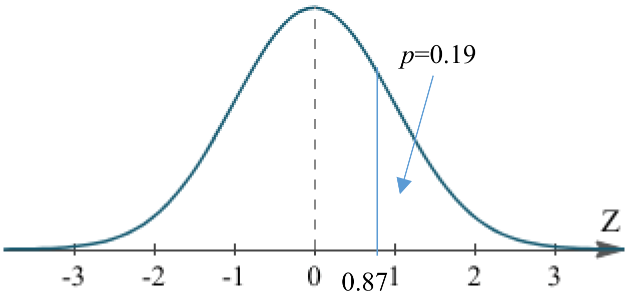
Solution Midterm Exam

I. Evaluating a New Version of Windows

In a random sample of 30 students and 40 workers, 18 students and 28 workers preferred a new software version. Test the hypothesis that the proportion of workers that prefer the new version is higher than the proportion of students that do so by reporting the appropriate *p*-value. Draw a figure to illustrate. (10 points)

$\hat{\rho}_s = 18/30$, $\hat{\rho}_w = 28/40$, $\hat{\rho}_c = 46/70$, under H_0 $\hat{\rho}_w = \hat{\rho}_s$,

$s_{(\hat{\rho}_w - \hat{\rho}_s)} = \sqrt{\frac{(46/70)(24/70)}{30} + \frac{(46/70)(24/70)}{40}} \approx 0.115$, so $z = \frac{28/40 - 18/30}{0.115} \approx 0.87$ =0.5. From the z table, the area to the left of 0.87 is 0. 81. The p-value is 1-0.81=0.19. There is about a one in five chance of getting a different this large in a sample of this size by sheer chance in a world where there is truly no underlying difference. This, there some evidence that workers prefer the new version. So, the null would be rejected only if the acceptable type 1 error rate is 0.2 or larger.



II. Defective Grates

One half of one percent (0.5%) of grates are defective if the stamping machine that presses them is working properly. Each grate is worth $20. (10 points)

1. What is the probability of observing more than one defective grate in a randomly selected production run of 200 units if the machine is working properly?

P(k>1)=1-P(k=0)-P(k=1)= $1 - 0.995^{200} - \frac{200!}{199!1!}0.995^{199}0.005^{1}$ ≈1-0.367-0.369=0.264.

2. What are the mean and standard deviation of the number of defective units for a 200 unit production run?

$\mu = 200 \times 0.005 = 1$ $\sigma^2 = 200 \times 0.005 \times 0.995 = 0.995$, $\sigma = \sqrt{0.995} \approx 0.997$.

3. A company operates four independent and identical stamping machines on 200 unit production runs, producing 800 units in total per cycle. What are the mean and standard deviation of total dollars lost to defective units per cycle if the machines are working correctly?

The expected loss is simply $\$20 \times 1 \times 4 = \$80$.

The variance is $4 \times 20^2 \times 0.995 = 1592$ so the standard deviation is $39.90.

Solution

III. Passenger Enplanements in Medium U.S. MSAs

In this question you will interpret and explain statistical output on the relationship between enplanements per capita (*enppc*) in 152 medium sized U.S. metropolitan statistical areas (MSAs) in 2012 and the distance (in miles) to the nearest MSA with a hub (major) airport (*dist*). The table to the right gives summary statistics. The table below presents output from a regression of the natural log of enppc, *lnenppc*, on the natural log of distance to the nearest MSA with a hub airport, *lndist*, with robust standard error estimates.

Summary Statistics

Variable	Mean	Std Dev
enppc	1.09	1.76
dist	189.23	198.96
lnenppc	-0.53	1.19
lndist	5.05	0.56

Regression Output (Robust Standard Errors)

Source	SS	df	MS	
Model	36.96	1	36.96	N = 152
Residual	177.23	150	1.18	$R^2 = 0.173$
Total	214.19	151	1.42	RMSE = 1.087

lnenppc	Coef	Std Err	t	P>t	95% Conf. Interval	
lndist	0.88	0.16	5.59	<0.001	0.57	1.19
_cons	-4.99	0.80	-6.21	<0.001	-6.58	-3.40

1. Why does using the log of enplanements per capita and the log of distance produce a better model than using the untransformed values? (10 points)

i. Variables that are strictly positive are often modeled more naturally in terms of growth or relative changes, rather than in terms of absolute changes.

ii. As a result of their tendency to reflect growth or decay processes, such non negative variables tend to be right skewed, whereas the log of such variables is much more symmetric, and as you saw in the homework that was definitely the case here—and for small samples many statistical models work better for approximately normal data.

iii. It does not make sense to have a model that could predict negative enplanements, so using the log of enplanements per capita gives a more reasonable model in some sense.

iv. Comparing the two models where the dependent variable is the natural log of enplanements per capita, the R-Squared is slightly higher when we use the log of distance rather than the untransformed value.

v. The scatter plots are truncated in strange ways for all other three models (as a result of the skewed distributions) and as a result the liner fit line does not quite make sense in those figures.

2. Assume both *dist* and *enppc* are at their sample averages. (10 points)

2.1. If *dist* increases by 1% what is the predicted percentage change in *enppc*?

0.88%

2.2. If *dist* increases by 1% what is the predicted change in *enppc*?

Δenppc=0.0088×1.09≈0.01. One more enplanement per 100 residents.

2.3. If *dist* increases by 10 miles, what is the predicted change in *enppc*?

%Δdist=100×10/189.23≈5.29%, so %Δenp≈0.88×5.29%≈4.65%, so Δenp≈0.0465×1.09≈0.05. Five more enplanements per 100 residents.

3. What is the RMSE? Show how it is calculated from the sum of squares information. (10 pts)

The root mean square error is literally the square root of the average squared sample residual, where the sum of squares is divided by degrees of freedom, not N. It is a measure of spread around the regression line—the standard deviation of the unpredictable part of the response variable. In this case, RMSE=(177.23/150)^0.5≈1.087

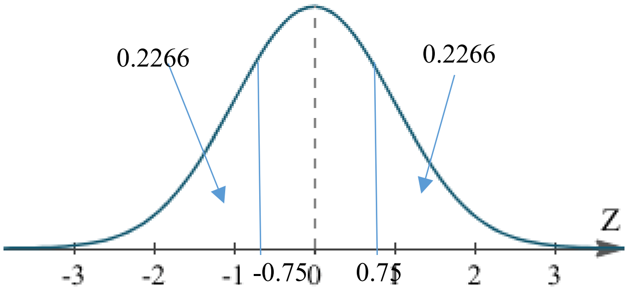
4. What is the R^2? Show how it is calculated from the sum of squares information. (10 points)

It is the percentage of variability in the response variable, measured by the total sum of squared deviation around the mean, predicted by the model. In this case, R^2=36.96/214.19≈0.173

5. Show how the 95% confidence interval for the slope coefficient was calculated. Given the large sample size, just use the critical value from the standard normal table provided. (10 pts)

The value of z that puts 2.5% in the tail is 1.96. The confidence interval is 0.88±1.96×0.16, or 0.57 to 1.19.

6. Test the null hypothesis that the slope coefficient is 1 against the alternative it is either more or less than 1 by reporting the appropriate two-tailed p-value. Again just use the standard normal distribution for your test. Draw a figure to illustrate. (10 points)



The Z score is (0.88-1)/0.16=-0.75. From the table the area to the right of 0.75 is 0.2266 and the area left of -0.75 is 0.2266. Therefore, the p-value is 0.4532. This means that, under the null hypothesis that the coefficient is 1, one would get a coefficient less than 0.88 or larger than 1.12 just over 4,532 times in 10,000. Thus, there is no little evidence that the coefficient is not 1, the null should not be rejected.

7. Why does it make sense to use robust standard errors and how does their calculation differ from the default standard error (speaking broadly and intuitively)? (10 points)

There is typically no reason to assume the spread of the data is the same for all observations (that is that the data is homoscedastic), as is done by the default calculation. The robust standard errors do not make this assumption. Instead, they simply use the pattern of the relationship between the predictor variables and the squared residuals apparent in the data to account for the pattern of heteroscedasticity in estimating the standard errors of the coefficients.