Statistics

Solutions to Practice Questions

ESD, SUTD

Term 5, 2017

Maximum likelihood estimation for the normal distribution:

Since we are differentiating with respect to one parameter at a time (while treating the other one as fixed), we should have used partial derivatives.

$\hat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2$ is biased.

Q1. The residuals seem to fall on a cubic curve, so $y -$ regression line $\approx$ cubic, so y can also be approximated by a cubic.

Thus a sensible model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

Q2. (a) The regression line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Sub in $x = \bar{x}$, we get $\hat{y} = \bar{y}$.

(b) Appending $(\bar{x}, \bar{y})$ does not affect $\bar{x}$, $\bar{y}$, s_{xy}/s_x^2 or $s_{xy}/(s_x s_y)$ (as you can check using the formulas for variance and covariance), hence the line and r^2 are unchanged.

(Another way to see this is to note that, by part (a), $(\bar{x}, \bar{y})$ lies on the original regression line, so adding it does not change SSE, and also no other line can give a smaller SSE, so the original line is also the least square line for the appended data.)

Q3. (a) $r^2 =$ SSR/SST. This formula also holds in multiple regression.

(b) Note that $k = 1$. $F =$ MSR/MSE $= (n-2)$ SSR/SSE.

(c) SSE + SSR = SST.

(d) Using the result of (b), we eliminate SSE with the help of (c):

$$F = \frac{(n-2)\,\mathsf{SSR}}{\mathsf{SST}-\mathsf{SSR}} = \frac{(n-2)\,\mathsf{SSR}/\mathsf{SST}}{1-\mathsf{SSR}/\mathsf{SST}} = \frac{(n-2)\,r^2}{1-r^2},$$

where we have used (a) for the last step.

(Note that this is the square of the test statistic for $H_0 : \rho = 0$. This is not too surprising, since in simple linear regression, F is testing for $H_1 : \beta_1 = 0$, which is equivalent to the previous null.)

Q4. We first observe that

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}).$$

Then, using this expression for $\hat{y}_i$, we have

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{y}_i) \times (\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} \big[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\big] \times \hat{\beta}_1 (x_i - \bar{x}) \\
&= \hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \hat{\beta}_1 (n-1) s_{xy} - \hat{\beta}_1^2 (n-1) s_x^2 \\
&= \hat{\beta}_1 (n-1) \big(s_{xy} - \hat{\beta}_1 s_x^2\big) = 0,
\end{aligned}$$

where in the last step we have used $\hat{\beta}_1 = s_{xy}/s_x^2$.

Q5. (a) Using properties of variance and covariance, we have $s_{x'y'} = ac\, s_{xy}$, $s_{x'} = a\, s_x$, and $s_{y'} = c\, s_y$.

So $s_{x'y'}/(s_{x'} s_{y'}) = s_{xy}/(s_x s_y)$, thus r will remain the same, while the new slope will be $\hat{\beta}_1' = \dfrac{c}{a}\hat{\beta}_1$.

(b) In this case, $a = 1/s_x$, $c = 1/s_y$, so $\hat{\beta}_1' = \dfrac{c}{a}\dfrac{s_{xy}}{s_x^2} = r$.

Moreover, note that the standardized data has x and y mean 0, so $\hat{\beta}_0' = 0$.

Q6. (a) Under H_0 (that the distribution is discrete uniform), the expected attendance is $(69 + 63 + 55 + 57 + 60 + 44)/6 = 58$ for each week.

Thus we have

$$\chi^2 = \frac{11^2 + 5^2 + 3^2 + 1^2 + 2^2 + 14^2}{58} = 6.138.$$

Since this is less than the critical value, we do not reject H_0.

(b) From (a) we have already calculated $\bar{y} = 58$; also $\bar{x} = 3.5$. The most involved calculation is s_{xy}; it is given by

$$\frac{1}{5}\sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y}) = -13.2.$$

So we have $\hat{\beta}_1 = s_{xy}/s_x^2 = -3.77$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 71.2$.

Q7. Using the formula in the Week 10 slides and that $s^2 = \text{MSE}$, we find that $t_{n-2,\ 0.975}/\sqrt{n-1} = 0.14$. Looking at this value numerically, we deduce that n must be large, so we can approximate $t_{n-2,\ 0.975}$ by $z_{0.975} = 1.96$.

Hence $\sqrt{n-1} \approx 14$, so $n \approx 200$.

Q8. No, r^2 is always between 0 and 1.

On the other hand, if r^2 is close to 0, and if the number of predictors p is close to the number of data points n, then

$$r^2_{adj} = 1 - \frac{n-1}{n-1-p}(1-r^2)$$

can be less than 0.

Q9. If H_0 is true, then the proportion of supporters is (approximately) normally distributed with mean 0.6 and variance $0.6 \times 0.4/100 = 0.0024$. The two-sided p-values is

$$2\,P\left(Z < \frac{0.48 - 0.6}{\sqrt{0.0024}}\right) = 2\,P(Z < -\sqrt{6}) = 0.01431.$$

(b) As $\chi^2 = \sum$ (observed − expected)^2/expected, we have

$$\chi^2 = \frac{(48-60)^2}{60} + \frac{(52-40)^2}{40} = 6.$$

(c) χ^2 is a chi-squared random variable with 1 degree of freedom, which is also the square of a standard normal random variable. Thus the p-value is

$$P(\chi^2 > 6) = P(Z^2 > 6) = 2\,P(Z > \sqrt{6}) = 2\,P(Z < -\sqrt{6}) = 0.01431.$$

The final evaluation comes from (and is the same as) part (a).

Q10. Note that $N = nk$. By carefully applying the formulas from Week 11 lecture 1, we find that

$$c(n,k) = \frac{nk}{k-1}.$$

Q11. (a) No, because F is a unitless ratio (since F is the ratio of MSA and MSE, both of which have the same units).

(b) Actually no. You can use, for instance, the paper airplane data from the Week 12 recitation to provide a counterexample.

(c) Use the definition of the F distribution as a ratio of two distributions! The question is asking for x such that

$$P\left(\frac{\chi_5^2/5}{\chi_2^2/2} < x\right) = 0.05,$$

or equivalently,

$$P\left(\frac{\chi_5^2/5}{\chi_2^2/2} > x\right) = 0.95,$$

or equivalently,

$$P\left(\frac{\chi_2^2/2}{\chi_5^2/5} < \frac{1}{x}\right) = 0.95.$$

But we are actually given $1/x = 5.786$, so $x = 1/5.786 = 0.1728$.