COVER PAGE

STAT 608 Homework 01, Summer 2017

NAME - RAJAN KAPOOR

EMAIL - r.kapoor@tamu.edu

40/40

Solution 1 (i) $X \sim \mathcal{N}(\mu, \sigma^2)$

probability density function $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

$-\infty < x < \infty$

$Y = \exp(X)$ is monotonically increasing function of X with inverse given by
$X = \log(Y)$ (∵ By inverse function theorem, a continuously differentiable f^n is invertible)

Using change of variable formula for continuous density functions, density function for RV Y,

$$f_Y(y) = f_X(\log y) \left|\frac{d}{dy} \log y\right|$$

$$= \frac{1}{y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\log y - \mu)^2}{2\sigma^2}\right]$$

for $\exp(-\infty) < y < \exp(\infty)$

or $0 < y < \infty$

Another approach is to use derivative of cdf for more general cases.

(ii) $E[\exp(X)] = \int_{-\infty}^{\infty} \exp x \, f_X(x)\, dx$ (by definition of expectation)

$$= \int_{-\infty}^{\infty} e^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(x - \frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

let $X = \mu + \sigma Z$, (Z i.e. is standard normal RV)

$\Rightarrow dx = \sigma dz$ $\qquad x = -\infty \quad z = -\infty$; $x = \infty \quad z = \infty$

$$E[\exp(X)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\mu + \sigma z - \frac{z^2}{2}\right) dz$$

$$= \frac{e^{\mu}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[\frac{-(-2\sigma z + z^2 - \sigma^2 + \sigma^2)}{2}\right] dz$$

$$= \frac{e^{\mu + \sigma^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(z-\sigma)^2}{2}\right) dz$$

$z - \sigma = t$ $\qquad z = -\infty \quad t = -\infty$

$\Rightarrow dz = dt$ $\qquad z = \infty \quad t = \infty$

$$E[\exp(X)] = \frac{e^{\mu + \sigma^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt = \frac{e^{\mu + \sigma^2/2}}{\sqrt{2\pi}} \sqrt{2\pi}$$

$$= e^{\mu + \sigma^2/2}$$

$$\neq \exp[\mu] \text{ or } \exp[E[X]]$$

So

$$E[\exp(X)] \neq \exp[E[X]]$$

(iii) For median, CDF $= 0.5$

cumulative distribution fn for RV Y,

$$F_Y(y) = P(Y \leq y)$$
$$= P(e^X \leq y)$$
$$= P(X \leq \log y)$$
$$= P(\mu + \sigma Z \leq \log y)$$

where Z is standard normal RV

$$F_Y(y) = P\left(Z \le \frac{\log y - \mu}{\sigma}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\log y - \mu}{\sigma}} e^{-z^2/2}\, dz$$

$$= \frac{1}{2}\left[1 + \text{erf}\left(\frac{\log y - \mu}{\sigma\sqrt{2}}\right)\right]$$

$$F_Y(y_m) = 0.5$$

$$\Rightarrow \text{erf}\left(\frac{\log y_m - \mu}{\sigma\sqrt{2}}\right) = 0 \Rightarrow \frac{\log y_m - \mu}{\sigma\sqrt{2}} = 0$$

$$\Rightarrow y_m = e^{\mu} = \exp(m[X])$$

where $y_m = m[Y]$

So $m[\exp(x)] = \exp[\cancel{x}\, m[X]]$

Solution 2

For size < 250 $\quad Y = \beta_0 + \beta_1 X_1 + e$

For size ≥ 250 $\quad Y = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 + e$

$$= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + e$$

For same slope, $\beta_1 = \beta_1 + \beta_3$

$\Rightarrow \beta_3 = 0$ $\qquad$ (d) is correct

Solution 3 From the expression of CI of predicted value of response variable from the least square fit, the variability should increase as we move away from mean. Clearly this is not the case in the figure since variability is increasing on moving right from the mean. $(y_i - \hat{y}_i)^2$

Solution 4

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

For given x_i, $\beta_0 + \beta_1 x_i$ is constant $\Rightarrow Var[\beta_0 + \beta_1 x_i] = 0$

$$Var[y_i|x] = Var[\beta_0 + \beta_1 x_i + e_i | x]$$

$$= \cancel{Var[\beta_0 + \beta_1 x_i | x]}^{0} + Var[e_i|x] + \cancel{2Cov(\beta_0 + \beta_1 x_i, e_i)|x}^{0}$$

$$= Var[e_i].$$

(Assumption of random error)
$Var[e_i|x] = Var[e_i]$

Solution 5

$$Y_i = \beta x_i + e_i$$

(a) Least squares estimate of $\beta = \hat{\beta}$

Residual sum of squares, $RSS = \sum_{i=1}^{n} e_i^2$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

For least squares estimate $\frac{\partial RSS}{\partial \hat{\beta}} = 0$

$$\Rightarrow -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta} x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \hat{\beta} x_i^2 = \hat{\beta} \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum x_i^2}$$

b (i) $$E[\hat{\beta}|X] = E\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \middle| X\right]$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} E\left[\sum_{i=1}^{n} x_i y_i | X\right]$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} \cdot \sum_{i=1}^{n}\left(x_i \; E[y_i|X]\right)$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} \cdot \sum_{i=1}^{n}\left(x_i \; E[\beta x_i|X]\right)$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} \cdot \sum_{i=1}^{n}\left(x_i \cdot x_i \; E[\beta|X]\right)$$

$$= \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} (x_i^2 \cdot \beta) = \frac{\beta}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i^2 = \beta$$

b(ii) $$Var(\hat{\beta}|X) = Var\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \middle| X\right]$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} Var\left[\sum_{i=1}^{n} x_i y_i | X\right]$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \cdot \sum_{i=1}^{n}\left(x_i^2 \; Var[y_i|X]\right)$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} (x_i^2 \sigma^2) = \frac{\sigma^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} x_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

(iii) since the errors $e_i \overset{iid}{\sim} N(0,1)$ i.e. normally distributed,

$\therefore$ $Y_i|X$ is normally distributed

Now $\hat{\beta}|X$ is linear combination of y_i's and therefore normally distributed

From (i) and (ii)

$$\hat{\beta}|X \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$$

Solution 6

Residual Sum of squares, RSS $= \sum_{i=1}^{n} \hat{e}_i^2$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0)^2$$

For least squares fit $\frac{\partial RSS}{\partial \hat{\beta}_0} = 0$

$$\Rightarrow -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\beta}_0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_0$$

$$\Rightarrow \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\Rightarrow \hat{\beta}_0 = \bar{y}$$

Solution 7

The confidence interval calculatior estimates the standard deviation of population distribution using standard deviation of sampling distribution. This estimated value might differ significantly from actual value. Thus 95% of observations can fall outside the 95% CI. if sampling is not random

---

Solution 8

(a) R^2 = coefficient of determination

$$= \frac{SSreg}{SST} = \frac{\text{Variability explained by model}}{\text{Total sample variability}}$$

$$SST = \sum_{i}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSreg = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 1.5026e12$$

$$SST = SSreg + RSS$$

$$= 1.5026e12 + 5.1884e9$$

$$R^2 = \frac{SSreg}{SST} = 0.997$$

(b) F value for testing $H_0: \beta_1 = 0$ against $H_A: \beta_1 \neq 0$

$$F = \frac{SSreg/1}{RSS/(n-2)} = \frac{16\,(SSreg)}{RSS}$$

$$= 4633.721$$

(d) Best estimate of β_1 (slope)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{SXY}{SXX}$$

$$= 0.9821 \approx 0.982$$

(c) Best estimate of β_0 (intercept)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= 6804.886$$

(e) Lower and upper confidence limits for 95% CI for $\hat{\beta}_1$

CI = statistic ± critical value * standard error

95% CI $\alpha = 0.05$

$$t_{\alpha/2, 16} = 2.120$$

$$= 0.9821 \pm 2.120 * \underbrace{1.443e(-2)}_{se(\hat{\beta}_1)}$$

$$= 0.9821 \pm 0.03059$$

Lower limit : 0.952

upper limit : 1.013

(f) Is $\beta_1 = 1$ a plausible value for β_1?

$H_0 : \beta_1 = 1$

$H_a : \beta_1 \neq 1$

$$t = \frac{0.9821 - 1}{1.443e-2} = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)}$$

$se(\hat{\beta}_1) = \frac{S}{\sqrt{Sxx}}$

$S =$ std dev of resid

$= -1.240$

$\alpha = 0.05$

p-value $= P(> |t| \text{ or } < -|t|)$

$= (1 - pt(|t|, 16)) \times 2$

$= (0.116) \times 2 = 0.233 > 0.05$

Cannot reject null hypothesis

$\therefore \beta_1 = 1$ is plausible value for β_1

(g)

$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ $\quad x^* = 369000$

$E[\hat{y}^* | X = x^*] = E[\hat{\beta}_0 | X = x^*] + E[\hat{\beta}_1 | X = x^*] x^*$

$= \beta_0 + \beta_1 x^*$

$= \$369,193$

(h) 95% Confidence interval $\alpha = 0.05$

$$CI = \hat{y}^* \pm t_{\alpha/2, 16}\, S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sxx}}$$

$CI = (\$357322, \$381064)$

(i) 95 % prediction interval

$$PI = \hat{y}^* \pm t_{\alpha/2, n-2} \, s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sxx}}$$

$$= (\$329215.4, \$409170.5)$$

(j) No. \$497,000 lies outside the prediction interval for $x^* = \$369000$.

(k) $$Y_i = \beta x_i + e_i$$

$$\hat{\beta} = 0.99102$$

$H_0 : \hat{\beta} = 1$ $\qquad \alpha = 0.05$

$$t = \frac{0.991 - 1}{0.00607}$$

$$= -1.4827$$

$$\approx -1.483$$

$$p\text{-value} = P(> |t| \text{ or } < -|t|)$$

$$= 0.1575$$

$$\approx 0.158 > 0.05$$

hypothesis $\hat{\beta} = 1$ cannot be rejected

Prediction rule is acceptable

Solution 1

(a) Plot of lognormal