Statistical Models and Regression

EN 625.661.82

Prepared by:

Sean Mahoney
sean.mahoney743@gmail.com
843-901-0071

Professor: Dr. Kelly Rooker

Title: Test #1
Date Submitted: 9/9/2018

Enclosed pages: 5

Sean Mahoney
Test # 1

(1) $\hat{y}_i = B_0 + B_1(x-\bar{x}) + \varepsilon$, $\Sigma(y_i - \hat{y})^2$ is what we minimize

$y = B_0 + B_1 x_i - B_1\bar{x} \Rightarrow$ minimize $\Sigma(y_i - B_0 - B_1 x_i + B_1\bar{x})^2$

$\frac{\partial S}{\partial B_0} = -2\Sigma(y_i - B_0 - B_1 x_i + B_1\bar{x}) = 0 \Rightarrow \Sigma y_i - \Sigma B_0 - \Sigma B_1 x_i + \Sigma B_1\bar{x} = 0$

$n\bar{y} - nB_0 - n\bar{x}B_1 + n\bar{x}B_1 = 0 \Rightarrow B_0 = \bar{y}$

$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{(\Sigma(x_i-\bar{x})^2 \cdot \Sigma(y_i-\bar{y})^2)^{1/2}} = \frac{S_{xy}}{(S_{xx}\cdot SS_T)^{1/2}}$, Prove $r^2 = R^2 = \Sigma\frac{(\hat{y}_i-\bar{y})^2}{(y_i-\bar{y})^2}$

$r = \frac{\Sigma y_i(x_i-\bar{x}) - \bar{y}\overset{0}{\cancel{(x_i-\bar{x})}}}{(S_{xx}\cdot SS_T)^{1/2}}$; $\Sigma\bar{y}(x_i-\bar{x}) = \Sigma\bar{y}x_i - \Sigma\bar{y}\bar{x} = n\bar{x}\bar{y} - n\bar{x}\bar{y} = 0$

numerator $r = \Sigma y_i(x_i-\bar{x}) = x_i(y_i-\bar{y})$

$\Sigma(\hat{y}_i-\bar{y})^2 = \Sigma(B_0 + B_1x_i - B_1\bar{x} - \bar{y})^2 = \Sigma(\overset{B_0=\bar{y}}{\bar{y}} + B_1x_i - B_1\bar{x} - \bar{y})^2$

$= \Sigma(B_1x_i - B_1\bar{x})^2 = \boxed{B_1^2\Sigma(x_i-\bar{x})^2}$

Solve for B_1

$\frac{\partial S}{\partial B_1} = -2\Sigma(-x_i+\bar{x})(y_i - B_0 - B_1(x-\bar{x})) = 0 \Rightarrow \Sigma\big[(-x_i+\bar{x})y_i - B_0(-x_i+\bar{x}) - B_1x_i(-x_i+\bar{x})$
$+ B_1\bar{x}(-x_i+\bar{x})\big] = 0$

$\Sigma(-x_iy_i + \bar{x}y_i + B_0x_i - B_0\bar{x} + B_1x_i^2 - B_1x_i\bar{x} - B_1\bar{x}x_i + B_1\bar{x}^2) = 0$

$\Sigma x_iy_i - n\bar{x}\bar{y} = nB_0\cancel{\bar{x}} - n\cancel{B_0\bar{x}} + B_1\Sigma x_i^2 - B_1n\bar{x}^2 - B_1\cancel{n\bar{x}^2} + \cancel{nB_1\bar{x}^2}$

$B_1 = \Sigma x_iy_i - \frac{\Sigma x_i\Sigma y_i}{n} \Big/ \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = \frac{S_{xy}}{S_{xx}}$ (same as normal ols)

$= \left(\frac{S_{xy}}{S_{xx}}\right)^2\Sigma(x_i-\bar{x})^2 = \frac{(S_{xy})^2}{S_{xx}} \Rightarrow R^2 = \frac{(S_{xy})^2}{(S_{xx})(\Sigma y_i-\bar{y})}$

$R^2 = \frac{[\Sigma(x_i-\bar{x})(y_i-\bar{y})]^2}{\Sigma(x_i-\bar{x})^2\cdot\Sigma(y_i-\bar{y})^2} = r^2 = \left[\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{[\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2]^{1/2}}\right]^2$ ✓

$\Rightarrow \boxed{R^2 = r^2 \text{ for } \hat{y}_i = B_0 + B_1(x-\bar{x}) + \varepsilon}$

2. $R^2 = 0.82$, $S_{yy} = 50$, $n = 25$, 95% PI. of y for $x = \bar{x}_0$

$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$, $SS_{Res} = \Sigma(y_i - \hat{y})^2$, $SS_R = \Sigma(\hat{y}_i - \bar{y})^2$

$S_{yy} = \Sigma(y_i - \bar{y})^2$, $SS_T = SS_R + SS_{Res}$, $SS_{TOT} = \Sigma(y_i - \bar{y})^2 = S_{yy} = 50$

$R^2 = 1 - \frac{SS_{Res}}{SS_T} \Rightarrow .82 = 1 - \frac{SS_{Res}}{SS_T} \Rightarrow (SS_T)(.82) = SS_T - SS_{Res}$

$SS_{Res} = SS_T - (SST)(.82) = 50 - (50)(.82) = 9$

$SS_R = SS_T - SS_{Res} = 50 - 9 = 41$

$MS_{Res} = \hat{\sigma}^2 = SS_{Res}/n-2 = 9/23 = \boxed{.3913}$

$B_1 = \frac{S_{xy}}{S_{xx}}$; $R^2 = \hat{B}_1^2 \frac{S_{xx}}{SS_T} \Rightarrow \hat{B}_1 = \sqrt{\frac{R^2 SS_T}{S_{xx}}} = \sqrt{\frac{.82 \cdot 50}{S_{xx}}} \to$ still don't know S_{xx}, oops

$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \Rightarrow r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} \Rightarrow$

$\psi = y_0 - \hat{y}_0$

$E(y|x=\bar{x}) = \hat{B}_0 + \hat{B}_1 \bar{x}$, $Var(\psi) = var(y_0 - \hat{y}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$

Standard error for $\psi = \sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$.

for $x_0 = \bar{x}$ this simplifies to $\sqrt{MS_{Res}(1 + \frac{1}{n})} = \sqrt{MS_{Res}(1 + \frac{1}{25})} = \sqrt{1.04 \cdot MS_{Res}}$

$\Rightarrow SE = .6379$

$t_{.05/2, 23} = 2.069$

Prediction Interval: $\hat{y}_0 - (t_{.975, 23})(se) \leq y_0 \leq \hat{y}_0 + (t_{.975, 23})\, se$

$\boxed{\hat{y}_0 - 1.32 \leq y_0 \leq \hat{y}_0 + 1.32}$

I tried everything I could think of to find B_1/B_0, but no S_{xx} or S_{xy} term

I do know that OLS passes through $(\bar{x}, \bar{y}) \Rightarrow \hat{y}_0$ for $\bar{x} = \bar{y}$

$\Rightarrow \bar{y} - 1.32 \leq \bar{y} \leq \bar{y} + 1.32$

Scanned by CamScanner

(3) (8,10), (8,10), (7,9), (6,10), (3,6), (4,8), (4,8), (5,9), (3,7), (6,9)

$\bar{x} = \sum x_i/n = (8+8+7+6+3+4+4+5+3+6)/10 = 5.4$

$\bar{y} = \sum y_i/n = (10+10+9+10+6+8+8+9+7+9)/10 = 8.6$

$S_{xx} = \sum(x_i - \bar{x})^2 = (8-5.4)^2 + (8-5.4)^2 + (7-5.4)^2 + (6-5.4)^2 + (3-5.4)^2 + (4-5.4)^2 + (4-5.4)^2$
$+ (5-5.4)^2 + (3-5.4)^2 + (6-5.4)^2 = 32.4$

$S_{xy} = \sum y_i(x_i - \bar{x}) = (10)(8-5.4)^2 + 10(8-5.4)^2 + 9(7-5.4)^2 + (10)(6-5.4)^2 + (6)(3-5.4)^2 + (8)(4-5.4)^2$
$+ (8)(4-5.4)^2 + 9(5-5.4)^2 + (7)(3-5.4)^2 + (9)(6-5.4)^2 = 20.6$

$\hat{B}_1 = \frac{S_{xy}}{S_{xx}} = \frac{20.6}{32.4} = \underbrace{\boxed{.6358}}_{\text{slope}}$, $\hat{B}_0 = \bar{y} - \hat{B}_1\bar{x} = 8.6 - (.6358)(5.4) = \underbrace{\boxed{5.167}}_{\text{y-int}}$

(b) $\hat{var}(y) = \frac{SS_{Res}}{n-2}$; $\sum y_i^2 - n\bar{y}^2 - \hat{B}_1 S_{xy}$

$\sum(y_i^2) = 10^2 + 10^2 + 9^2 + 10^2 + 6^2 + 8^2 + 8^2 + 9^2 + 7^2 + 9^2 = 756$

$SS_{Res} = 756 - (10)(8.6)^2 - (.6358)(20.6) = \boxed{3.303}$

$\hat{\sigma} = SS_{Res}/n-2 = 3.303/8 = \boxed{.412875} = MS_{Res}$

(c) Test for statistical confidence of slope

$H_0: \hat{B}_1 = 0$, $H_a: B_1 \neq 0$ (2 way test)

Standard error: $se(\hat{B}_1) = \sqrt{MS_{Res}/S_{xx}} = \sqrt{.412875/32.4} = .1129$

$t_0 = \hat{B}_1/se = .6358/.1129 = 5.632$

Testing at 95% significance $\Rightarrow t_{\frac{.05}{2},8} = 1.860$

Since $t_0 >> t_{\frac{.05}{2},8} \Rightarrow$ Reject Null hypothesis. I conclude that there is a statistically significant linear relationship between x and y

(d) Construct 95% CI on slope

$100(1-\alpha)$ CI given by: $\hat{\beta}_1 - t_{\alpha/2,n-2} \cdot se(\hat{\beta}_1) \le \hat{\beta}_1 \le \hat{\beta}_1 + (t_{\alpha/2,n-2}) se(\hat{\beta}_1)$

$t_{.05/2,8} = 1.860$, $se(\hat{\beta}_1) = .1129$ (prev problem), $\hat{\beta}_1 = .6358$

CI: $(.6358) - (1.860)(.1129) \le \hat{\beta}_1 \le (.6358) + (1.860)(.1129)$

$\boxed{.42581 \le \hat{\beta}_1 \le .8458}$

(e)

Source of Variation	Sum of Squares	Degrees Freedom	Mean Square	F_0	p-value
Regression	13.1	1	13.1	31.76	4.9×10^{-4}
Residual	3.3	8	.4125		
Total	16.4	9			

$SS_R = (\hat{\beta}_1) S_{xy} = (.6358)(20.6) = 13.1$; $SST = SS_R + SS_{Res} = 13.1 + 3.303 = 16.4$

$MS_R = SS_R/dof = 13.1$, $MS_{Res} = SS_{Res}/dof = 3.3/8$

$F_0 = MS_R / MS_{Res} = 13.1/.4125 = 31.76$; F_0

If we test at 95% significance, $F_{.05,1,8} = 5.32$

Since $F_0 >> F_{.05,1,8}$ we reject null hypothesis and say significant

(f) ε is random variable w/ $\sim N(0, \sigma^2)$ → Applicable to all cases

Y is random variable, X is fixed → Applicable to all cases

Because $E(\varepsilon)=0$ and $Var(\varepsilon)=\sigma^2$, OLS estimates are unbiased (a,b,c,d)

LSE are best linear unbiased estimators (minimum variance)

uncorrelated → Choice of X does not affect random error of Y and since mean of $\varepsilon = 0$, it does not matter what Ind Var we choose (a,b,c)

Regression model is linear in coefficient and error - model must fit linear pattern (a-e)

Error term has constant variance (homoscedacity) - (a,b)

Since homoscedacity + no autocorrelation → error term is Independent and identically distributed

Scanned by CamScanner

Error term is normally distributed - not required to be Normal for unbiased but allows for more accurate testing → (c,d,e)

(4) $Var(e_i) = Var(y_i - \hat{y}) = Var(y_i - (\bar{y} + B_1(x-\bar{x}))) = Var((y_i-\bar{y}) - B_1(x-\bar{x}))$

$Var(y_i - \bar{y}) = Var(y_i) + Var(\bar{y}) = \sigma^2 + \sigma^2/n \quad (Cov(y_i, \bar{y}) = 0)$

$Var(B_1(x-\bar{x})) = (x-\bar{x})^2 \cdot Var B_1 = (x-\bar{x})^2 \cdot \sigma^2/S_{xx} \quad (Var(B_1) \to HW)$

$$Var((y_i-\bar{y}) - B_1(x-\bar{x})) = \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} \cdot (x-\bar{x})^2 + 2\,cov((y_i-\bar{y}), (B_1(x-\bar{x})))$$

$$2\,cov(y_i - \hat{y}, B_1(x-\bar{x})) = \left(\frac{1}{n-1}\right) \sum (x_i-\bar{x})(y_i-\bar{y})$$

$$Var(e_i) = \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} \cdot (x_i-\bar{x})^2 + \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}\right] + \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

$Var(\epsilon) = \sigma^2 \Rightarrow Var(e_i) > Var(\epsilon)$ since it gets multiplied by a number bigger than 1, and the covariance term is in there being added

I think this is due to the fact that you have variance contributions for both predicted ($\hat{y}$) and observed (y_i) where as the ϵ_i term is by itself and contributes to y_i ($\hat{B}_0 + B_1 x_i + \epsilon$). So y_i includes variance from ϵ and variance from $\hat{y}$.

At first I thought y_i variance would be less since we subtract $\hat{y}$ from it in $Var(y_i - \bar{y})$ but

$Var(a-b) = Var(a) + Var(b)$ not $Var(a) - Var(b)$