LM Homework 3

Joshua Ingram

2/25/2020

Problem 1

1.

a.

$\hat{\beta} = \frac{n\Sigma_i x_i y_i - \Sigma_i x_i \Sigma_i y_i}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2}$

$= \frac{n(x_1 y_1 + ... + x_n y_n) - \Sigma_i x_i (y_1 + ... + y_n)}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2} = \frac{(nx_1 y_1 + ... + nx_n y_n) - \Sigma_i x_i (y_1 + ... + y_n)}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2}$

$= \frac{y_1(nx_1 - \Sigma_i x_i) + ... + y_n(nx_n - \Sigma_i x_i)}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2} = y_1(\frac{nx_1 - \Sigma_i x_i}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2}) + ... + y_n(\frac{nx_n - \Sigma_i x_i}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2})$

$= \Sigma_i y_i m_i, m_i = (\frac{nx_i - \Sigma_i x_i}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2})$

b.

Given $m_i = \frac{x_i - \bar{x}}{\Sigma_i (x_i - \bar{x})^2}$ and $\Sigma_i x_i (x_i - \bar{x}) = \Sigma_i (x_i - \bar{x})^2$

$E[\hat{\beta}] = E[\Sigma_i m_i y_i] = \Sigma_i [m_i y_i] = \Sigma_i m_i E[y_i] = \Sigma_i m_i (\alpha + \beta x_i)$

$= \Sigma_i m_i \alpha + \Sigma_i m_i \beta x_i = \alpha \Sigma_i m_i + \beta \Sigma_i m_i x_i$

$\alpha(\frac{\Sigma_i x_i - \bar{x}}{\Sigma_i (x_i - \bar{x})^2}) + \beta(\frac{\Sigma_i x_i (x_i - \bar{x})}{\Sigma_i (x_i - \bar{x})^2})$

$= \alpha(\frac{0}{\Sigma_i (x_i - \bar{x})^2}) + \beta(\frac{\Sigma_i (x_i - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2}) = 0 + \beta(1)$

$=> E[\hat{\beta}] = \beta$

c.

$V(\hat{\beta}) = V(\Sigma_i m_i y_i) = \Sigma_i V(m_i y_i) = \Sigma_i m_i^2 V(y_i) = \Sigma_i m_i^2 \sigma^2$

$= \sigma^2 \Sigma_i m_i^2 = \sigma^2 \frac{\Sigma_i (x_i - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2 \Sigma_i (x_i - \bar{x})^2} = \sigma^2 \frac{1}{\Sigma_i (x_i - \bar{x})^2}$

$=> V(\hat{\beta}) = \frac{\sigma^2}{\Sigma_i (x_i - \bar{x})^2}$

d.

i. $\hat{\alpha} = \bar{y} - \bar{x}\Sigma m_i y_i = \frac{1}{n}\Sigma y_i - \frac{1}{n}\Sigma x_i \Sigma m_i y_i = \frac{1}{n}(\Sigma y_i - \Sigma x_i \Sigma m_i y_i)$

$= \frac{1}{n}((y_1 + ... + y_n) - \Sigma x_i(m_1 y_1 + ... + ...m_n y_n)) = \frac{1}{n}(y_1 + ... + y_n) - (\frac{1}{n}\Sigma x_i m_1 y_1 + ... + \frac{1}{n}\Sigma x_i m_n y_n) = y_1(\frac{1}{n} - (\frac{m_1}{n}\Sigma x_i)) + ... + y_n(\frac{1}{n} - (\frac{m_n}{n}\Sigma x_i))$

$m_i^* = (\frac{1}{n} - (\frac{m_i}{n}\Sigma x_i)$

$=> \hat{\alpha} = \Sigma m_i^* y_i$

ii. $E[\hat{\alpha}] = E[\bar{y} - \hat{\beta}\bar{x}] = E[\bar{y}] - E[\hat{\beta}\bar{x}] = E[\frac{\Sigma y_i}{n}] - \bar{x}E[\hat{\beta}]$

$= \frac{1}{n}\Sigma\alpha + \beta x_i - \bar{x}\beta = \frac{1}{n}n\alpha + \frac{1}{n}\Sigma\beta x_i - \bar{x}\beta$

$= \alpha + \beta\bar{x} - \bar{x}\beta$

$=> E[\hat{\alpha}] = \alpha$

iii. $V[\hat{\alpha}] = V[\bar{y} - \hat{\beta}\bar{x}] = V[\bar{y}] + V[\hat{\beta}\bar{x}]$

Finding $V[\bar{y}]$ first:

$v[\bar{y}] = V[\frac{\Sigma y_i}{n}] = \frac{1}{n^2}\Sigma V[y_i] = \frac{1}{n^2}\Sigma\sigma^2 = \frac{1}{n}n\sigma^2 = \frac{\sigma^2}{n}$

Finding $V[\hat{\beta}\bar{x}]$:

$V[\hat{\beta}\bar{x}] = \bar{x}^2 V[\hat{\beta}] = \frac{\sigma^2\bar{x}^2}{\Sigma(x_i - \bar{x})^2}$

Putting the two togeter:

$v[\hat{\alpha}] = V[\bar{y} - \hat{\beta}\bar{x}] = \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}^2}{\Sigma(x_i-\bar{x})^2} = \frac{\sigma^2\Sigma(x_i-\bar{x})^2}{n\Sigma(x_i-\bar{x})^2} + \frac{\sigma^2\bar{x}^2 n}{n\Sigma(x_i-\bar{x})^2}$

$= \frac{\sigma^2\Sigma(x_i-\bar{x})^2 + \sigma^2\bar{x}^2 n}{n\Sigma(x_i-\bar{x})^2} = \frac{\sigma^2(\Sigma(x_i-\bar{x})^2 + \bar{x}^2 n)}{n\Sigma(x_i-\bar{x})^2} = \frac{\sigma^2(\Sigma x_i^2 - \bar{x}\Sigma x_i + n\frac{(\Sigma x_i)^2}{n^2})}{n\Sigma(x_i-\bar{x})^2}$

$= \frac{\sigma^2(\Sigma x_i^2 - \bar{x}\Sigma x_i + \bar{x}\Sigma x_i)}{n\Sigma(x_i-\bar{x})^2}$

$=> V[\hat{\alpha}] = \frac{\sigma^2(\Sigma x_i^2 - \bar{x}\Sigma x_i + \bar{x}\Sigma x_i)}{n\Sigma(x_i-\bar{x})^2}$

(b) usees the assumptionof linearity

(c) uses the assumptions of linearity, constant variance, and independence

2.

a.

Efficiency refers to the functions variance, so the smaller the variance the more "efficient" the estimator. When all the assumptions of simple linear regression are satisfied, Least squares estimate is the MOST efficient estimator.

b.

When the normality assumption is broken, the LS estimate is the most efficient LINEAR estimator, but is not the most efficient estimator as there may be other non-linear estimators that are more efficicent (smaller variance).

Problem 2

1.

Given the stated formula $\frac{\hat{\beta}-\beta}{SE(\hat{\beta})}$ t_{n-2}

$P(t_{\frac{\alpha}{2}} < T < t_{1-\frac{\alpha}{2}}) = 1 - \alpha$

$P(t_{\frac{\alpha}{2}} < \frac{\hat{\beta}-\beta}{SE(\hat{\beta})} < t_{1-\frac{\alpha}{2}}) = 1 - \alpha$

$P(t_{\frac{\alpha}{2}} SE(\hat{\beta}) < \hat{\beta} - \beta < t_{1-\frac{\alpha}{2}} SE(\hat{\beta})) = 1 - \alpha$

$P(\hat{\beta} - t_{1-\frac{\alpha}{2}} SE(\hat{\beta}) < \beta < \hat{\beta} - t_{\frac{\alpha}{2}} SE(\hat{\beta})) = 1 - \alpha$

$=> \beta \in (\hat{\beta} - t_{1-\frac{\alpha}{2}} SE(\hat{\beta}), \hat{\beta} - t_{\frac{\alpha}{2}} SE(\hat{\beta}))$

2.

```
x <- c(0, 0, 2, 2)
y <- c(0, 0, 2, 2)
x_bar <- mean(x)
y_bar <- mean(y)

beta <- sum((x - x_bar)*(y - y_bar))/sum((x - x_bar)^2)
alpha <- y_bar - beta*x_bar
```

```
my.simple.lm <- function(x, y){

  x_bar <- mean(x)
  y_bar <- mean(y)

  # estimates for slope and beta

  beta_estimate <- sum((x - x_bar)*(y - y_bar))/sum((x - x_bar)^2)

  alpha_estimate <- y_bar - (beta_estimate * x_bar)

  # fitted values from estimates

  fitted <- alpha_estimate + (beta_estimate * x)

  rss <- sum((y - fitted)^2)

  # rse

  rse <- sqrt(rss / (length(x) - 2))

  # something is wrong here: the RSE is correct
  #but when I square it (to get sigma^2 estimate) it gives
  #completely wrong results for the confidence interval and
  #standard errors of the estimates
  rse_2 <- rss / (length(x) - 2)
```

```
  # standard errors of estimates

  se_alpha <- (rse * sum(x^2))/(length(x)*sum((x - x_bar)^2))
  se_beta <- rse/sum((x - x_bar)^2)

  # 95%  confidence intervals for estimates

  lower_alpha <- alpha_estimate - 1.96 * se_alpha

  upper_alpha <- alpha_estimate + 1.96 * se_alpha

  lower_beta <- beta_estimate - 1.96 * se_beta

  upper_beta <- beta_estimate + 1.96 * se_beta

  # output as a list

  output <- list("alpha_est" = alpha_estimate, "beta_est" = beta_estimate, "RSE" = rse, "se_alpha" = se

  return(output)
}
```

3.

```
my.output <- my.simple.lm(anscombe$income, anscombe$education)
my.output
```

```
## $alpha_est
## [1] 17.71003
##
## $beta_est
## [1] 0.05537594
##
## $RSE
## [1] 34.9384
##
## $se_alpha
## [1] 23.86196
##
## $se_beta
## [1] 2.228007e-06
##
## $conf_int_alpha
## [1] -29.05941  64.47948
##
## $conf_int_beta
## [1] 0.05537157 0.05538031
```

```
lm.output <- lm(education ~ income, data = anscombe)
summary(lm.output)
```

```
##
## Call:
## lm(formula = education ~ income, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.077 -21.868  -4.617  17.523 124.701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.710031  28.873840   0.613    0.542
## income       0.055376   0.008823   6.276 8.76e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.94 on 49 degrees of freedom
## Multiple R-squared:  0.4457, Adjusted R-squared:  0.4343
## F-statistic: 39.39 on 1 and 49 DF,  p-value: 8.762e-08
```

```
confint(lm.output)
```

```
##                    2.5 %      97.5 %
## (Intercept) -40.31412380 75.73418534
## income        0.03764572  0.07310616
```

Note: So for my "sanity check", I found that my standard errors and confidence intervals were off for my estimates. I narrowed down the problem to being the Residual standard error estimate. I found that my RSE was the same as the lm() function, but when I square it (rse_2 in my function) the standard errors and confidence intervals of the estimates become ridiculously large... I went ahead and used rse (instead of rse_2) because they gave more "accurate" estimates for the se and conf int, however, I understand I should be using rse^2. I was following the formulas for standard error of beta and alpha, as well as the confidence intervals... so there shouldn't be a problem... but there is... so something is wrong with the change from the rse to rse^2 (which is the variance estimate to be plugged in for the se formulas) and I couldn't find the issue. If you know what might be going on, I'd appreciate a comment in the grading on how to fix this issue. Thank you!