STAT 224 Autumn 2022 HW3

Matthew Zhao

```
food = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/food.txt", h = T)
ee = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/EE.txt", h = T)
te = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/TE.txt", h = T)
```

Q1a — 3 Points

```
lmte = lm(log(V) ~ log(K) + log(L), data = te)
print(lmte$coefficients[2:3])
```

```
## log(K) log(L)
## 0.5057 0.8455
```

```
confint(lmte,'log(L)',level = 0.90)
```

```
##            5 %  95 %
## log(L) 0.09411 1.597
```

Least Squares estimates for β_K and β_L are 0.5057 and 0.8455 respectively, with the 90% CI for β_L: (0.09411,1.597). This means that we are 90% confident that a 1% increase in labor input would lead to an increase in output of 0.09411 to 1.597%.

Q1b — 5 points

```
> summary(lmte)
Residual standard error: 0.07495 on 12 degrees of freedom
Multiple R-squared:  0.8184,    Adjusted R-squared:  0.7881
$F$-statistic: 27.03 on 2 and 12 DF,  p-value: 3.591e-05
```

$SSE = 0.07495$

$R^2 = 1 - \frac{SSE}{SST}, SST = SSR + SSE \Rightarrow R^2 = 1 - \frac{SSE}{SSR+SSE} \Rightarrow SSR = SSE \times \frac{R^2}{1-R^2} \Rightarrow SSR = 0.07495 \times \frac{0.8184}{1-0.8184} = 0.3378$

Since the F-statistic is drawn from F distribution $F_{2,12}$, dfR = 2 and dfE = 12 so $MSR = \frac{SSR}{dfR} = \frac{0.3378}{2} = 0.1689$ and $MSE = \frac{SSE}{dfE} = \frac{0.07495}{12} = 0.006246$

The F-statistic = 27.03

```
pf(27.03,2,12,lower.tail = F)
```

```
## [1] 0.00003593
```

Source	df	SS	MS	F	P-value
Regression	2	SSR = 0.3378	MSR = 0.1689	F = 27.03	0.00003593
Error	12	SSE = 0.07495	MSE = 0.006246		

Q1c — 5 points

Full Model: $\log V = \log \alpha + \beta_K \log K + \beta_L \log L + \epsilon$

Reduced Model ($\beta_K = \beta_L$): $\log V = \log \alpha + \beta_K (\log K + \log L) + \epsilon$

```
full_lmte = lm(log(V) ~ log(K) + log(L), data = te)
reduced_lmte = lm(log(V) ~ I(log(K) + log(L)), data = te)
anova(reduced_lmte,full_lmte)
```

```
## Analysis of Variance Table
##
## Model 1: log(V) ~ I(log(K) + log(L))
## Model 2: log(V) ~ log(K) + log(L)
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     13 0.0682
## 2     12 0.0674  1  0.000787 0.14   0.71
```

$F = \frac{(SSE_{reduced} - SSE_{full})/(dfE_{reduced} - dfE_{full})}{MSE_{full}} = \frac{(0.0682 - 0.0674)/(13-12)}{0.0674/12} = 0.1424$

Degrees of freedom for F-stat is $dfE_r - dfE_f$ and dfE_f so 13-12 = 1 and 12.

```
pf(0.1424,1,12,lower.tail = F)
```

```
## [1] 0.7125
```

Since the F-test is not statistically significant, we say that we fail to reject the null that $\beta_L = \beta_K$.

Q1d — 6 Points

```
full_lm = lm(log(V) ~ log(K) + log(L), data = food)
reduced_lm = lm(log(V) ~ I(log(K)-log(L)), offset=log(L), data = food)
anova(reduced_lm,full_lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(V) ~ I(log(K) - log(L))
## Model 2: log(V) ~ log(K) + log(L)
##   Res.Df    RSS Df Sum of Sq    F  Pr(>F)
## 1     13 0.1210
## 2     12 0.0419  1    0.0791 22.7 0.00046 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For Food: F-stat 22.7 with $dfE_r - dfE_f = 1$ and $dfE_f = 12$, p-value 0.00046.

```
full_lm = lm(log(V) ~ log(K) + log(L), data = te)
reduced_lm = lm(log(V) ~ I(log(K)-log(L)), offset=log(L), data = te)
anova(reduced_lm,full_lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(V) ~ I(log(K) - log(L))
## Model 2: log(V) ~ log(K) + log(L)
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     13 0.0835
## 2     12 0.0674  1    0.0161 2.87   0.12
```

For Transportation Equipment (TE): F-stat 2.87 with $dfE_r - dfE_f = 1$ and $dfE_f = 12$, p-value 0.12

```
full_lm = lm(log(V) ~ log(K) + log(L), data = ee)
reduced_lm = lm(log(V) ~ I(log(K)-log(L)), offset=log(L), data = ee)
anova(reduced_lm,full_lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(V) ~ I(log(K) - log(L))
## Model 2: log(V) ~ log(K) + log(L)
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     13 0.0611
## 2     12 0.0591  1   0.00202 0.41   0.53
```

For Equipment and Supplies (EE): F-stat 0.41 with $dfE_r - dfE_f = 1$ and $dfE_f = 12$, p-value 0.53. At the 0.05 significance level, we fail to reject that $\beta_L + \beta_K = 1$ for TE and EE. We reject for food.

Q1e — 2 Points

```r
model = lm(log(V) ~ log(K) + log(L) + YEAR, data = ee)
summary(model)
```

```
##
## Call:
## lm(formula = log(V) ~ log(K) + log(L) + YEAR, data = ee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06402 -0.01141 -0.00361  0.01420  0.04440
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.41454    2.64406   -5.83  0.00011 ***
## log(K)        0.82098    0.28919    2.84  0.01611 *
## log(L)        0.88249    0.18889    4.67  0.00068 ***
## YEAR          0.02497    0.00346    7.21 0.000017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0306 on 11 degrees of freedom
## Multiple R-squared:  0.914,  Adjusted R-squared:  0.891
## F-statistic:   39 on 3 and 11 DF,  p-value: 0.00000373
```

Since $H_0 : \log \rho = 0$ and $H_a : \log \rho > 0$, we use a 1-sided t-test.

$t = \frac{\hat{\beta}_j - \beta_j^0}{se(\hat{\beta}_j)} = \frac{0.02497 - 0}{0.00346} = 7.217$

```r
pt(7.217,nrow(ee)-3-1,lower.tail = F)
```

```
## [1] 0.000008574
```

Since the p-value is less than 0.05, we reject the null hypothesis, meaning that there was technological progress from 1972 to 1986 for the EE sector.

Q1f — 5 points

```r
full_food = lm(log(V) ~ log(K) + log(L) + YEAR, data = food)
reduced_food = lm(log(V) ~ YEAR, data = food)
anova(reduced_food, full_food)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: log(V) ~ YEAR
## Model 2: log(V) ~ log(K) + log(L) + YEAR
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     13 0.0447
## 2     11 0.0413  2   0.00335 0.45   0.65
```

$F = \frac{(SSE_{reduced} - SSE_{full})/(dfE_{reduced} - dfE_{full})}{MSE_{full}} = \frac{(0.0447 - 0.0413)/(13-11)}{0.0413/11} = 0.4527$

$dfE_{reduced} - dfE_{full} = 13 - 11 = 2$, $dfE_{full} = 11$.

```
pf(0.4527,2,11,lower.tail = F)
```

```
## [1] 0.6472
```

Since the p-value is higher than 0.05, we fail to reject the null and conclude that $\beta_K = \beta_L = 0$

Question 2

Q2a — 3 points

Model 1:

$Price = -6.107 + 0.169 * Horsepower$

Prediction $= -6.107 + 0.169 * 100 = 10.79 =$ \$10,790 USD

Model 2:

$Price = -4.117 + 0.174 * Horsepower + -3.162 * I(USA) + -3.818 * I(Japan) + 0.311 * I(Germany)$

Prediction $= -4.117 + 0.174 * 100 + 0.311 * 1 = 13.59 =$ \$13,590 USD

Model 3:

$Price = -10.882 + 0.237 * Horsepower + 2.076 * I(USA) + 4.755 * I(Japan) + 11.774 * I(Germany) + -0.052 * Horsepower * I(USA) + -0.077 * Horsepower * I(Japan) + -0.095 * Horsepower * I(Germany)$

Prediction $= -10.882 + 0.237 * 100 + 11.774 * 1 - 0.095 * 100 * 1 = 15.09 =$ \$15,090 USD

Q2b — 2 points

Model 2:

$Price = -4.117 + 0.174 * Horsepower + -3.162 * I(USA) + -3.818 * I(Japan) + 0.311 * I(Germany)$

Prediction $= -4.117 + 0.174 * 100 = 13.28 =$ \$13,280 USD

Model 3:

$Price = -10.882 + 0.237 * Horsepower + 2.076 * I(USA) + 4.755 * I(Japan) + 11.774 * I(Germany) + -0.052 * Horsepower * I(USA) + -0.077 * Horsepower * I(Japan) + -0.095 * Horsepower * I(Germany)$

Prediction $= -10.882 + 0.237 * 100 = 12.82 =$ \$12,820 USD

Q2c — 2 points

If a car is from Japan, its expected price is $3,818 lower than if it were not from Japan, given that it is also not from the USA or Germany.

Q2d — 2 points

A unit increase in horsepower (1 additional horsepower) is associated with an increase in expected car price of $174 holding other variables constant.

Q2e — 3 points

A unit increase in horsepower (1 additional horsepower) is associated with an increase in expected car price of $237 given that the car is not from the USA, Japan, or Germany. If it is from one of these countries, then a unit increase in horsepower is associated with either an increase in expected car price of $185 if it is from the USA, $160 if it is from Japan, or $142 if it is from Germany.

The interaction between Country and horsepower means that depending on which country the car is from (if it is a country that the model has an indicator for), additional horsepower changes the expected price by a different amount.

Q2f — 4 points

We would test model 3 against model 2, where model 2 is nested in model 3 i.e. is a reduced version.

$F = \frac{(SSE_{reduced} - SSE_{full})/(dfE_{reduced} - dfE_{full})}{MSE_{full}} = \frac{(1390.31 - 1319.85)/(85-82)}{1319.85/82} = 1.459$

Degrees of Freedom for F-stat are $dfE_{reduced} - dfE_{full} = 85 - 82 = 3$ and $dfE_{full} = 82$

```
pf(1.459,3,82,lower.tail = F)
```

```
## [1] 0.2318
```

The p-value is 0.2318. Since it is higher than 0.05, we fail to reject the null that the interactions are zero. We can conclude that the interactions are insignificantly different from 0 at the 0.05 significance level.

Q2g — 2 points

Model 2:

$Price = -4.117 + 0.174 * Horsepower + -3.162 * I(USA) + -3.818 * I(Japan) + 0.311 * I(Germany)$

Since constant and horsepower are the same for all countries, they do not affect the comparison:

Prediction = $ -3.162 * I(USA) + -3.818*I(Japan) + 0.311*I(Germany)$

If USA: -3.162, if Japan: -3.818, if Germany: 0.311, if Other: 0.

Japan has the least expensive cars based on model 2 since it has a large negative coefficient, meaning that a car from Japan is expected to be cheaper than a car not from Japan, USA, or Germany. Additionally, the p-value for the Japan coefficient is statistically signficant at the 0.05 significance level, meaning that it is significantly different from 0. While the USA also has a negative coefficient that is significant at the 0.05 significance level, it is not as large as Japan's.

Q2h — 4 points

Compare Model 1 (reduced) and Model 2 (full).

$F = \frac{(SSE_{reduced} - SSE_{full})/(dfE_{reduced} - dfE_{full})}{MSE_{full}} = \frac{(1604.44 - 1390.31)/(88-85)}{1390.31/85} = 4.363$

Degrees of Freedom for F-stat are $dfE_{reduced} - dfE_{full} = 88 - 85 = 3$ and $dfE_{full} = 85$

```
pf(4.363,3,85,lower.tail = F)
```

```
## [1] 0.006591
```

The p-value is 0.006591 Since it is lower than 0.05, we reject the null that the country coefficients are zero. We can conclude that the country indicators are significant at the 5% level, meaning that the country a car was made in affects the price.

Q2i — 2 points

Yes. The Germany category is insignificantly different from 0 at the 5% level so it can be moved to other. Additionally, the USA and Japan categories could potentially be merged since their coefficients are both statistically significant and are fairly similar to each other. However, this second recommendation should be tested.