JE Problem Set 4

1 $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i$

$u_i \perp\!\!\!\perp (X_{1i}, \ldots, X_{ki})$

$E(u_i) = 0$

$E(u_i \mid X_{1i}, \ldots, X_{ki})$

$= E(u_i)$

since $u_i \perp\!\!\!\perp (X_{1i}, \ldots, X_{ki})$

$= 0$

$E(u_i^2 \mid X_{1i}, \ldots, X_{ki})$

$= E(u_i^2)$

since $u_i \perp\!\!\!\perp (X_{1i}, \ldots, X_{ki})$

Hence u_i is homoskedastic

2a $Y = \beta_0 + \beta X + u$

where $E(u) = 0$ and $E(X_i u) = 0$

$H_0: \beta_1 = 0$, $H_1: \beta_1 > 0$

under H_0,

~~$n^{1/2}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, \omega^2 \beta_1)$~~

~~$n^{1/2}(\hat{\beta}_1 - \beta_1)/\hat{\omega}_{\beta_1}$~~

$se(\hat{\beta}_1) = n^{-1/2} \hat{\omega}_{\beta_1} = n^{-1/2} \hat{sd}(\hat{u}) / \hat{sd}(\hat{X}_1)$

t-statistic

$t = (\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1) = n^{1/2}(\hat{\beta}_1 - \beta_1)/\hat{\omega}_{\beta_1} \xrightarrow{d} N(0,1)$ where $\hat{\omega}_{\beta_1} = \hat{sd}(\hat{X}_1 \hat{u}) / \hat{var}(\hat{X}_1)$

For one-sided test at level of significance $\alpha = 10\%$,

Reject H_0 if $t > c_\alpha$

$\alpha = 0.10 = \Phi(-c_\alpha)$, $c_\alpha = 1.2816$

2bi when $\beta_1 = 0.01$, sampling distribution of $\hat{\beta}_1$ is closer to the sampling distribution under the null than when $\beta_1 = 100$. observed value of the t-statistic is likely to be less positive. Given the decision rule from (a), we are less likely to reject the false null. The test is less powerful when $\beta_1 = 0.01$ than when $\beta_1 = 100$.

ii when $\beta_1 = -1$, sampling distribution of $\hat{\beta}_1$ has mean -1. observed value of the t-statistic is likely to be less positive than when $\beta_1 = 0.01$, 1, or 100. Given the decision rule from (a), we are less likely to reject the false null. The test is less powerful when $\beta_1 = -1$ than when $\beta_1 \in [0, \sout{100}\ +\infty)$

2ci $\hat{\beta}_1$ is approximately normally distributed with mean β_1. The closer the mean of the sampling distribution to the hypothesised value of β_1 under the null, the greater the probability mass of the sampling distribution ~~of~~ around this hypothesised value, hence the lower the probability of observing a t-statistic with a large absolute value, and of rejecting the null, under a two-sided test. For all $\beta_1 \neq 0$,

as $|\beta_1|$ increases, so does the power of the two sided test.

3a ~~Restricted model: $Y = \beta_0 + u$~~

~~where $E(u) = 0$~~

~~$Y_i = \beta_0 + u_i$~~

~~$E(Y) = E(\beta_0) + E(u) = \beta_0$~~

~~$Y_i = \beta_0 + u_i$~~

~~where $\bar{u} = 0$~~

~~$\bar{Y} = \beta_0 + \bar{u} = \beta_0$~~

~~$SSR_{rs} = \sum_{i=1}^{n} u_i^2$~~

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

~~$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$~~

~~$= TSS_{RS}$~~

Restricted model:

~~$Y = \beta_0 + u_{rs}$~~

$Y = \beta_{0,rs} + u_{rs}$ where $E(u_{rs}) = 0$

$Y_i = \beta_{0,rs} + u_{i,rs}$ where $\bar{u}_{rs} = 0$

$\bar{Y} = \beta_{0,rs} + \bar{u}_{rs} = \beta_{0,rs}$

$SSR_{rs} = \sum_{i=1}^{n} (Y_i - \beta_{0,rs})^2$

$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

$= TSS$

b ~~Unrestricted model:~~

~~$Y = \beta_0 + \beta_1 X + u_{un}$ where $E(u_{un}) = 0$ and $E(X u_{un}) = 0$~~

~~$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + u_{un,i}$ where $\bar{u}_{un} = 0$ and $\widehat{cov}(X, u_{un}) = 0$~~

~~From $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, $\hat{\beta}_1 = \widehat{cov}(Y,X)/\widehat{var}(X)$~~

Unrestricted model

~~$Y = \beta_0 + \beta_1 X$~~

$Y = \beta_{0,un} + \beta_1 X + u_{un}$

where $E(u_{un}) = 0$ and $E(X u_{un}) = 0$

$Y_i = \hat{\beta}_{0,un} + \hat{\beta}_1 X_i + \hat{u}_{un,i}$

where $\bar{u}_{un} = 0$ and $\widehat{cov}(X, \hat{u}_{un}) = 0$

$\hat{\beta}_1 = \widehat{cov}(Y,X)/\widehat{var}(X)$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

~~$SSR_{un} = \sum_{i=1}^{n} \hat{u}_{un,i}^2$~~

~~$= \sum_{i=1}^{n} (Y_i - \hat{\beta}_{0,un} - \hat{\beta}_1 X_i)^2$~~

~~$= \sum_{i=1}^{n} (Y_i - \bar{Y} - \hat{\beta}_1 (X_i - \bar{X}))^2$~~

~~$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$~~

~~$- 2\hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$~~

$SSR_{un} = \sum_{i=1}^{n} \hat{u}_{un,i}^2$

$= \sum_{i=1}^{n} (Y_i - \hat{\beta}_{0,un} - \hat{\beta}_1 X_i)^2$

$= \sum_{i=1}^{n} (Y_i - \bar{Y} - \hat{\beta}_1 (X_i - \bar{X}))^2$

$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$

$- 2\hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$

$= SSR_{rs} + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$

only if $\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 0$ or $\hat{\beta}_1 = 0$

↓ implies

$\widehat{cov}(Y,X) = 0$

↓ implies

$\hat{\beta}_1 = 0$

SSR_{rs}

$= TSS$

$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

$= \sum_{i=1}^{n} (\hat{Y}_i + \hat{u}_i - \bar{Y})^2$

$= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u}_i^2 - 2\sum_{i=1}^{n} \hat{u}_i (\hat{Y}_i - \bar{Y})$

$= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u}_i^2$

$= \sum_{i=1}^{n} (\hat{\beta}_{0,un} + \hat{\beta}_1 X_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u}_i^2$

$= \sum_{i=1}^{n} (\hat{\beta}_0 \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u}_i^2$

$= \sum_{i=1}^{n} (\hat{\beta}_1 (X_i - \bar{X}))^2 + \sum_{i=1}^{n} \hat{u}_i^2$

$= \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} \hat{u}_i^2$ SSR_{un}

$SSR_{rs} - SSR_{un} = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$

c $F = LR/q = LR$

$= [(SSR_{rs} - SSR_{un})/q] / [SSR_{un}/(n-k-1)]$

$= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 / (\sum_{i=1}^{n} \hat{u}_{un,i}^2 / (n-k-1))$

$t = (\hat{\beta}_1 - 0^{\#}) / se(\hat{\beta}_1)$

~~$= \hat{\beta}_1 / [s_{\hat{u}} / (\sum_{i=1}^{n} (x_i - \bar{x})^2)^{1/2}]^{1/2}$~~

~~$= \hat{\beta}_1 / [(1/n-k-1 \sum_{i=1}^{n} \hat{u}_i^2)^{1/2} / \sum_{i=1}^{n} (x_i - \bar{x})^2]^{1/2}$~~

$= \hat{\beta}_1 / [s_{\hat{u}} / [\sum_{i=1}^{n} (x_i - \bar{x})^2]^{1/2}$

$= \hat{\beta}_1 [\sum_{i=1}^{n} (x_i - \bar{x})^2]^{1/2} / [1/n-k-1 \sum_{i=1}^{n} \hat{u}_i^2]^{1/2}$

$t^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 / [1/n-k-1 \sum_{i=1}^{n} \hat{u}_i^2]$

$= F$

d ~~$R^2 = ESS/TSS$~~

4a $R^2 = ESS/TSS = 12851/95948 = 0.13394$

$\bar{R}^2 = 1 - (SSR/n-k-1)/(TSS/n-1)$

$= 1 - (TSS - ESS/n-k-1)/(TSS/n-1)$

$= 1 - (95948 - 12851/6028-9-1)/(95948/6028-1)$

$= 0.13264$

$s_{\hat{u}} = \sqrt{s_{\hat{u}}^2} = \sqrt{1/n-k-1 \sum_{i=1}^{n} \hat{u}_i^2}$

$= \sqrt{1/n-k-1 (TSS - ESS)}$

$= \sqrt{1/6028-9-1 (95948 - 12851)}$

$= 3.7159$

~~H_0:~~

Let the ~~regre~~ as regression model be

$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_q X_{qi} + \hat{u}_i$

where $\bar{\hat{u}}_i = 0$ and $\hat{cov}(x_{ki}, \hat{u}) = 0$ for $k \in \{1, \ldots, q\}$

$H_0: \beta_1 = \beta_2 \ldots = \beta_q = 0$

$H_1: \beta_k \neq 0$ for some $k \in \{1, \ldots, q\}$

F-statistic

$F = LR/q = [SSR_{rs} - SSR_{un}]$

~~$= [(SSR_{rs} - ESS$~~

$= [(SSR_{rs} - SSR_{un})/q] / [SSR_{un}/(n-k-1)]$

$= [(TSS - SSR)/q] / [SSR/(n-k-1)]$

~~$= [(95948 - 12851/9] / [(95948 - 12851)/(6028-9-1)$~~

$= [12851/9] / [(95948 - 12851)/(6028-9-1)]$

$= 103.41$

~~F-statistic high~~

~~F-stat~~
~~Sampling distr.~~
F-statistic converges in distribution to $F_{9,\infty}$

Reject H_0 if $F > c_\alpha$
$\alpha = 0.05 = P(F_{9,\infty} > c_\alpha)$, $c_\alpha = 1.8799$

Reject H_0

b $H_0: \beta_{teacher\_exp} = 0$
$H_1: \beta_{teacher\_exp} \neq 0$

Under the null,
t-statistic
$t = (\hat{\beta}_t - \beta_{t,0}) / se(\hat{\beta}_t)$
$= 0.04 / 0.02$
$= 2$
~~t-statistic~~ $t \xrightarrow{d} N(0,1)$

For two-sided test at level of significance $\alpha = 5\%$
Reject H_0 if $|t| > c_\alpha$
$\alpha = 0.05$, $c_\alpha = -\Phi^{-1}(0.05/2) = 1.9600$

Reject H_0

c $H_0: \beta_{small\_class} = 0$
$H_1: \beta_{small\_class} \neq 0$

Under the null,
t-statistic
$t = (\hat{\beta}_s - \beta_{s,0}) / se(\hat{\beta}_s)$
$= 0.66 / 0.30$
$= 2.20$
$t \xrightarrow{d} N(0,1)$

p-value
$p = 2\Phi(-|t|)$
$= 2\Phi(-2.20)$
$= 0.0139$

Under the null, the probability of observing a t statistic at least as unfavourable to the null as that actually observed is 0.0139. Reject the null at all levels of significance $\alpha > 0.0139$.

d ~~$H_0: \beta_1 = b$~~ $H_0: \beta_{summer\_bday} = b$
~~$H_1: \beta_1 \neq$~~ $H_1: \beta_{summer\_bday} \neq b$
$\alpha = 0.01$, $c_\alpha = |\Phi^{-1}(\alpha/2)| = 2.58$

Confidence interval
$C = \{b \in \mathbb{R} \mid H_0: \beta_{s-b} = b \text{ is accepted}\}$
$= \{b \in \mathbb{R} \mid |t(b)| < c_\alpha\}$
$= [\hat{\beta}_{sb} - c_\alpha se(\hat{\beta}_1), \hat{\beta}_{sb} + c_\alpha se(\hat{\beta}_1)]$
$= [-0.8538, -0.2862]$

There is a 0.9 probability that the interval $[-0.8538, -0.2862]$ contains the population regression parameter β_{summer_baby}.

e Elasticity $= (\partial maths\_test / maths\_test) / (\partial teacher\_exp / teacher\_exp)$
$= \hat{\beta}_1 (\overline{maths\_test} / \overline{teacher\_exp})$
$= 0.04 (61.81 / 13.93)$
$= 0.17749$

$H_0: E = e$
$H_1: E \neq e$
$\alpha = 0.10$, $c_\alpha = -|\Phi^{-1}(\alpha/2)| = 1.645$

confidence interval
$C = \{e \in \mathbb{R} \mid H_0: E = e \text{ is accepted}\}$
$= \{e \in \mathbb{R} \mid H_0: \hat{\beta}_1 (\overline{maths\_test} / \overline{teacher\_exp}) = e \text{ is accepted}\}$
$= \{e \in \mathbb{R} \mid H_0: \hat{\beta}_1 = e \overline{teacher\_exp} / \overline{maths\_test} \text{ is accepted}\}$
$= \{e \in \mathbb{R} \mid |t(e \overline{teacher\_exp} / \overline{maths\_test})| \leq c_\alpha\}$
$= [(\hat{\beta}_1 - c_\alpha se(\hat{\beta}_1))(\overline{maths\_test} / \overline{teacher\_exp}), (\hat{\beta}_1 + c_\alpha se(\hat{\beta}_1))(\overline{maths\_test} / \overline{teacher\_exp})]$
$= [0.0315, 0.323]$

Ignoring the sampling variability of sample means, there is a 0.90 probability that the interval $[0.0315, 0.323]$ contains the elasticity of math test score with respect to teacher experience.

f ~~$H_0: \beta_{black} = 0$~~
~~(there is no difference in math test scores of Black students and non-Black students)~~
~~$H_1: \beta_{black} \neq 0$~~
~~(there is some differe~~

$H_0: \beta_{black} = 0$
$H_1: \beta_{black} \neq 0$

Under the null,
t-statistic
$t = (\hat{\beta}_{black} - \beta_{black}) / se(\hat{\beta}_{black})$
$= -1.531 / 0.16$
$= -9.5625$

p-value
$p = 2\Phi(-|t|)$
$= 0$

Under the null, the probability of observing a t statistic as unfavourable to the null as that actually observed is effectively zero. Reject the null at all non-zero levels of significance α, there is strong evidence for differences in math test scores between Black and non-Black students.

$H_0: \beta_{other\_non\_wh} = 0$
$H_1: \beta_{other\_non\_wh} \neq 0$

under the null,
t-statistic
$t = (\hat{\beta}_{other} - \beta_{other}) / se(\hat{\beta}_{other})$
$= 0.90 / 0.598$
$= 1.5254$
p-value
$p = 2\Phi(-|t|)$
$= 0.12716$

Under the null, the probability of observing a t-statistic at least as unfavourable to the null as that actually observed is 0.12716, reject H_0 at all levels of significance greater than 0.12716. There is some evidence for difference in test scores between other non white students and Black and white students.

g The coefficient of free_lunch in the regression model gives a reliable estimate of the effect of free_lunch on math_test only if the other determinants of math_test excluded from the model are uncorrelated with free_lunch. This is unlikely.

math_test is likely to be affected by parents' income and hence access to educational resources. Free school lunches are (presumably) part of a programme to support underprivileged children, whose parents' income is likely to be lower. free_lunch is likely correlated with parents' income, which is a determinant of math_test excluded from the model.

Abolishing the provision of free school lunches could have no effect or even negative effect on math test scores if the relationship between free_lunch and math_test is entirely accounted for by parents' income, since parents' income remains unchanged ~~etc~~ when free school lunches are abolished.

5a $C_i + T_i + R_i = 1$
C_i, T_i, and R_i are perfectly multicollinear

bi On average, a one-year increase in years of experience is associated with a $\beta_x = r_x$ increase in hourly wages, all other factors being equal.

ii On average, an individual who lives in the city has an hourly wage $ higher than an

individual who lives in a rural area, and γ_C higher than an individual who lives in a town.

c $W_i = \beta_0 + \beta_X X_i + \beta_C C_i + \beta_T T_i + u_i$
where $E(u)=0$, $E(Xu)=0$, $E(Cu)=0$, $E(Tu)=0$
$W_i = \gamma_0 + \gamma_X X_i + \gamma_C C_i + \gamma_R R_i + v_i$
where $E(v)=0$, $E(Xv)=0$, $E(Cv)=0$, $E(Rv)=0$

$W_i = \beta_0 + \beta_X X_i + \beta_C C_i + \beta_T T_i + u_i$
$= \beta_0 + \beta_X X_i + \beta_C C_i + \beta_T (1 - C_i - R_i) + u_i$
$= \beta_0 + \beta_X X_i + (\beta_C - \beta_T) C_i - \beta_T R_i + u_i$
Since R_i is a linear function of C_i and T_i, and
$E(Cu)=0$, $E(Tu)=0$, $E(Ru)=0$
$W_i = \beta_0 + \beta_X X_i + (\beta_C - \beta_T) C_i - \beta_T R_i + u_i$
where $E(u)=0$, $E(Xu)=0$, $E(Cu)=0$, $E(Ru)=0$
$\beta_0 = \gamma_0$, $\beta_X = \gamma_X$, $\beta_C - \beta_T = \gamma_C$, $-\beta_T = \gamma_R$, $u_i = v_i$ since the solution to the regression model is unique

$-\beta_T = \gamma_R$
~~γ_R gives the~~ An individual living in a rural area, on average, has an hourly wage γ_R higher than an individual living in a ~~town~~, all other factors ~~beg~~ being equal. ~~th~~ γ_R is equal to $-\beta_T$, where ~~β_T is the diff~~ on average, an individual living in a town has an hourly wage β_T higher than an individual living in a rural area, all other factors being equal.

d $H_0: \beta_C = \beta_T = 0$
$H_1: \beta_C \neq 0$ or $\beta_T \neq 0$

Restricted model: $W_i = \beta_0 + \beta_X X_i + u_i$
Unrestricted model: $W_i = \beta_0 + \beta_X X_i + \beta_C C_i + \beta_T T_i + u_i$

F-statistic
$F = [(SSR_{RS} - SSR_{UR})/q] / [SSR_{UR}/(n-k-1)]$
$F \sim F_{2,\infty}$

Reject H_0 if $F > c_\alpha$
where $\alpha = P(F_{2,\infty} > c_\alpha)$

Under H_0, on average, ~~individual~~ an individual living in a city and an individual living in a town both have hourly wage no higher or lower than an individual living in a rural area, holding years of experience equal.

e Fit
$W_i = \beta_0 + \beta_X X_i + \beta_C C_i + \beta_T T_i + \beta_{CX} C_i X_i + \beta_{TX} T_i X_i + u_i$

Not sure how to explain (formally) why this works

F-test
$H_0: \beta_{CX} = \beta_{TX} = 0$
$H_1: \beta_{CX} \neq 0$ or $\beta_{TX} \neq 0$

6a ~~$Y_i = A_i K_i^{\alpha} C_i^{\beta} e^{\varepsilon_i}$~~

~~$\rightarrow A K_i^{\alpha} C_i^{\beta} e^{\varepsilon_i}$~~

~~$\ln Y_i = \ln A + \alpha \ln K_i + \beta \ln C_i$~~

$Y_i = A_i C_i^{\alpha} K_i^{\beta} e^{\varepsilon_i}$

$= A C_i^{\alpha} K_i^{\beta} e^{\varepsilon_i}$

$\ln Y_i = \ln A + \alpha \ln C_i + \beta \ln K_i + \varepsilon_i$

Let k be such that $E(\varepsilon_i + k) = 0$

$\ln Y_i = (\ln A - k) + \alpha \ln C_i + \beta \ln K_i + (\varepsilon_i + k)$

Given that $\varepsilon_i \perp\!\!\!\perp C_i, K_i$,

$(\varepsilon_i + k) \perp\!\!\!\perp \ln C_i, \ln K_i$

Since $\exp$ and $\ln$ are one-to-one

Let $Y_i^*, \gamma_0^*, C_i^*, K_i^*, \varepsilon_i^*$ be

$\ln Y_i, (\ln A - k), \ln C_i, \ln K_i, (\varepsilon_i + k)$ respectively

$Y_i^* = \gamma_0^* + \alpha C_i^* + \beta K_i^* + \varepsilon_i^*$

where $E(\varepsilon_i^*) = 0$ and $\varepsilon_i^* \perp\!\!\!\perp C_i^*, K_i^*$

$Y_i^* = \gamma_0^* + \alpha C_i^* + \beta K_i^* + \varepsilon_i^*$

is a population linear regression model of Y_i^* on C_i^* and K_i^*

$\hat{\alpha} = \widehat{cov}(Y_i^*, C_i^*) / \widehat{var}(C_i^*)$

$\hat{\beta} = \widehat{cov}(Y_i^*, K_i^*) / \widehat{var}(K_i^*)$

consistently estimate α and β

Is this right?

b The production function exhibits constant returns to scale iff $\alpha + \beta = 1$

H_0: ~~#~~ $\alpha + \beta = 1$

H_1: $\alpha + \beta \neq 1$

Under H_0:

$Y_i^* = \gamma_0^* + \alpha C_i^* + (1-\alpha) K_i^* + \varepsilon_i^*$

$= \gamma_0^* + \alpha (C_i^* - K_i^*) + K_i^* + \varepsilon_i^*$

~~$Y_i^* - K_i^*$~~

$Y_i^* - K_i^* = \gamma_0^* + \alpha (C_i^* - K_i^*) + \varepsilon_i^*$

Let Y_i^{**}, ~~γ_0^{**}~~ X_i^{**} be $Y_i^* - K_i^*$ and $C_i^* - K_i^*$ respectively

~~$Y_i^{**} = \gamma_0^* + \alpha X_i^{**} + \varepsilon_i^*$~~

~~Restricted model~~

Restricted model

$Y_i^{**} = \gamma_0^* + \alpha X_i^{**} + \varepsilon_i^*$

Compute

SSR_{un}, ~~SSR~~

SSR_R

$F = [(SSR_R - SSR_{un})/q] / [SSR_{un}/(n-k-1)]$

where $q = 1, k = 2$

At level of significance α

Reject H_0 if $F > c_\alpha$

where $\alpha = P(F_{1,\infty} > c_\alpha)$

I want to write the formulae for these but the same symbols γ_0^* and $\hat{\alpha}$ appear in both, but refer to different estimators, is there any way to avoid this notational difficulty?

c Yes.

$\hat{\alpha}$ and $\hat{\beta}$ would then estimate both the direct effect of K_i and C_i on Y_i and the respective indirect effects through A_i.

This seems unsatisfactory but I don't know how to make it more precise

d $Y_{i,2015}/Y_{i,2014} = A_{i,2015} C_{i,2015}^{\alpha} K_{i,2015}^{\beta}$

$Y_{i,2015}/Y_{i,2014} = L_{i,2015}^{\alpha} K_{i,2015}^{\beta} e^{\varepsilon_{i,2015}}$
$L_{i,2014}^{-\alpha} K_{i,2014}^{-\beta} e^{-\varepsilon_{i,2014}}$
$= (L_{i,2015}/L_{i,2014})^{\alpha} (K_{i,2015}/K_{i,2014})^{\beta}$
$e^{\varepsilon_{i,2015} - \varepsilon_{i,2014}}$

~~$\ln(Y_{i,2015}/Y_{i,2014})$~~
~~$= \alpha(\ln Y_{i,2015} - \ln Y_{i,2014}) + \beta(K_{i,2015}/K_{i,2014})$~~
~~$+\beta(\ln K_{i,2015}$~~

$\ln(Y_{i,2015}) - \ln(Y_{i,2014})$
$= \alpha(\ln L_{i,2015} - \ln L_{i,2014})$
$+\beta(\ln K_{i,2015} - \ln K_{i,2014})$
$+(\varepsilon_{i,2015} - \varepsilon_{i,2014})$
let k be such that $E(\varepsilon_{i,2015} - \varepsilon_{i,2014} - k)$
$= 0$
$\ln Y_{i,2015} - \ln Y_{i,2014}$
$= \alpha(\ln L_{i,2015} - \ln L_{i,2014})$
$+\beta(\ln K_{i,2015} - \ln K_{i,2014})$
$+(\varepsilon_{i,2015} - \varepsilon_{i,2014} - k)$
$+k$
let $Y_i^*, L_i^*, K_i^*, \varepsilon_i^*$ be
$\ln Y_{i,2015} - \ln Y_{i,2014}$,
$\ln L_{i,2015} - \ln L_{i,2014}$,
$\ln K_{i,2015} - \ln K_{i,2014}$,
$\varepsilon_{i,2015} - \varepsilon_{i,2014} - k$
respectively

$Y_i^* = k + \alpha L_i^* + \beta K_i^* + \varepsilon_i^*$
where $E(\varepsilon_i^*) = 0$, $\varepsilon_i^* \perp\!\!\!\perp L_i^*, K_i^*$

How can I justify this?

$Y_i^* = k + \alpha L_i^* + \beta K_i^* + \varepsilon_i^*$
is a population linear regression model of Y_i^* on L_i^* and K_i^*
$\hat{\alpha} = \widehat{cov}(Y_i^*, L_i^*) / \widehat{var}(L^*)$
$\hat{\beta} = \widehat{cov}(Y^*, K^*) / \widehat{var}(K^*)$
consistently estimate α and β

7a $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_1 X_2 + u$ (1)

$Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 X_2 + \beta_3 X_2 \ln X_1 + u$ (2)

(1)

Let $X_1^*, X_2^*, X_3^*, X_4^*$

be $X_1, X_1^2, X_2, X_1 X_2$ respectively

$Y = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \beta_3 X_3^* + \beta_4 X_4^* + u$

(1) is linear in the parameters

(2)

Let X_1^*, X_2^*, X_3^*

be ~~X_1~~ $\ln X_1, X_2, X_2 \ln X_1$ respectively

$Y = \beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \beta_3 X_3^* + u$

(2) is linear in the parameters

5i (1)

~~$E(u) = 0, E(X_i^* u) = 0$ for $i \in \{1, \dots, 4\}$~~

(2)

~~$E(u) = 0, E(X_i^* u) = 0$ for $i \in \{1, \dots, 3\}$~~

~~$E(u | X_1^*, \dots, X_4^*) = 0$~~

~~$E(u | X_1^*, \dots, X_3^*) = 0$~~

(1) $E(u) = E(X_1 u) = E(X_1^2 u) = E(X_2 u) = E(X_1 X_2 u) = 0$

(2) $E(u) = E(u \ln X_1) = E(X_2 u) = E(X_2 u \ln X_1) = 0$

ii (1) $E(u | X_1, X_2) = 0$

(2) $E(u | X_1, X_2) = 0$

Problem-Set-4-Question-8.R

r1454158

2022-05-15

```
# Clear the environment
rm(list = ls())

# Install packages
install.packages("estimatr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("car")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("margins")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
# Load packages
library(estimatr)
library(car)
```

```
## Loading required package: carData
```

```
library(margins)

# Load data
height = read.csv("height.csv")
dim(height)
```

```
## [1] 17870    11
```

```
head(height)
```

```
##   sex age mrd educ cworker region race earnings height weight occupation
## 1   0  48   1   13       1      3    1 84054.75     65    133          1
## 2   0  41   6   12       1      2    1 14021.39     65    155          1
## 3   0  26   1   16       1      1    1 84054.75     60    108          1
## 4   0  37   1   16       1      2    1 84054.75     67    150          1
## 5   0  35   6   16       1      1    1 28560.39     68    180          1
## 6   0  25   6   15       1      4    1 23362.87     63    101          1
```

```
summary(height)
```

```
##       sex              age             mrd             educ
##  Min.   :0.0000   Min.   :25.00   Min.   :1.000   Min.   : 0.00
```

```
##  1st Qu.:0.0000   1st Qu.:33.00   1st Qu.:1.000   1st Qu.:12.00
##  Median :0.0000   Median :40.00   Median :1.000   Median :13.00
##  Mean   :0.4419   Mean   :40.92   Mean   :2.362   Mean   :13.54
##  3rd Qu.:1.0000   3rd Qu.:48.00   3rd Qu.:4.000   3rd Qu.:16.00
##  Max.   :1.0000   Max.   :65.00   Max.   :6.000   Max.   :19.00
##     cworker          region          race          earnings
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 4726
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:23363
##  Median :1.000   Median :3.000   Median :1.000   Median :38925
##  Mean   :1.964   Mean   :2.551   Mean   :1.386   Mean   :46875
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:84055
##  Max.   :6.000   Max.   :4.000   Max.   :4.000   Max.   :84055
##      height          weight        occupation
##  Min.   :48.00   Min.   : 80.0   Min.   : 1.000
##  1st Qu.:64.00   1st Qu.:140.0   1st Qu.: 2.000
##  Median :67.00   Median :163.0   Median : 5.000
##  Mean   :66.96   Mean   :170.4   Mean   : 6.011
##  3rd Qu.:70.00   3rd Qu.:190.0   3rd Qu.: 8.000
##  Max.   :84.00   Max.   :501.0   Max.   :15.000
```

```
# Regression of earnings on height
reg1 = lm_robust(earnings ~ height, data = height)
summary(reg1)
```

```
##
## Call:
## lm_robust(formula = earnings ~ height, data = height)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   -512.7     3380.0 -0.1517 8.794e-01  -7137.9   6112.4 17868
## height         707.7       50.4 14.0419 1.490e-44    608.9    806.5 17868
##
## Multiple R-squared:  0.01088 ,   Adjusted R-squared:  0.01082
## F-statistic: 197.2 on 1 and 17868 DF,  p-value: < 2.2e-16
```

```
# Estimated slope coefficient is statistically significant. Under the null
# hypothesis that slope coefficient is zero, t-statistic = 14.0419, p-value
# = 1.490e-44. Under the null, the probability of observing a t-statistic as
# unfavourable to the null as that actually observed is 1.490e-44. Reject the
# null at all conventional levels of significance (10%, 5%, 1%).

# Regression of earnings on height, assuming homoskedasticity
reg2 = lm(earnings ~ height, data = height)
summary(reg2)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -512.73    3386.86  -0.151     0.88
## height        707.67      50.49  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

```
# Standard error of the estimated slope coefficient increases from 50.4 to 50.49
# when homoskedacity is assumed.

# Regression of earnings on height for men
reg3 = lm_robust(earnings ~ height, data = height, subset = sex == 1)
summary(reg3)
```

```
##
## Call:
## lm_robust(formula = earnings ~ height, data = height, subset = sex ==
##     1)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)   -43130    6926.25  -6.227 4.995e-10   -56708   -29553 7894
## height          1307      98.87  13.217 1.826e-39     1113     1501 7894
##
## Multiple R-squared:  0.02086 ,   Adjusted R-squared:  0.02074
## F-statistic: 174.7 on 1 and 7894 DF,  p-value: < 2.2e-16
```

```
# Regression of earnings on height for women
reg4 = lm_robust(earnings ~ height, data = height, subset = sex == 0)
summary(reg4)
```

```
##
## Call:
## lm_robust(formula = earnings ~ height, data = height, subset = sex ==
##     0)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)  12650.9     6299.9   2.008 4.466e-02    301.7  25000.0 9972
## height         511.2       97.6   5.238 1.655e-07    319.9    702.5 9972
##
## Multiple R-squared:  0.002672 ,  Adjusted R-squared:  0.002572
## F-statistic: 27.44 on 1 and 9972 DF,  p-value: 1.655e-07
```

```
# Hypothesis test
# H0: slope coefficient of regression of earnings on height for men = slope
# coefficient of regression of earnings on height for women.
# H1: slope coefficient of regression of earnings on height for men != slope
```

```r
# coefficient of regression of earnings on height for women.
# Under the null
# t-statistic
t = (reg3$coefficients[["height"]] - reg4$coefficients[["height"]])/
  (reg3$std.error[["height"]]^2 + reg4$std.error[["height"]]^2)^(1/2)
t
```

```
## [1] 5.726938
```

```r
# p-value
p = 2 * pnorm(-abs(t))
p
```

```
## [1] 1.022592e-08
```

```r
# Under the null, the probability of observing a t-statistic as unfavourable to
# the null as that actually observed is 1.0226e-08. Reject the null at all
# levels of significance greater than 1.0226e-08. Reject the null at all
# conventional levels of significance (10%, 5%, 1%).

# Polynomial regression of earnings on height
reg5 = lm_robust(earnings ~ height + I(height^2) + I(height^3), data = height)
summary(reg5)
```

```
##
## Call:
## lm_robust(formula = earnings ~ height + I(height^2) + I(height^3),
##     data = height)
##
## Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)   CI Lower  CI Upper    DF
## (Intercept)  1.962e+05  4.413e+05  0.4445   0.6567 -6.688e+05 1.061e+06 17866
## height      -8.023e+03  1.978e+04 -0.4056   0.6850 -4.680e+04 3.075e+04 17866
## I(height^2)  1.288e+02  2.951e+02  0.4365   0.6625 -4.496e+02 7.073e+02 17866
## I(height^3) -6.316e-01  1.465e+00 -0.4311   0.6664 -3.503e+00 2.240e+00 17866
##
## Multiple R-squared:  0.01089 ,   Adjusted R-squared:  0.01072
## F-statistic: 65.76 on 3 and 17866 DF,  p-value: < 2.2e-16
```

```r
# Hypothesis test
# H0: coefficient of I(height^2) and coefficient of I(height^3) in regression
# of earnings on height, I(height^2), and I(height^3) are both equal to zero.
# H1: at least one of coefficient of I(height^2) and coefficient of I(height^3)
# in regression of earnings on height, I(height^2), and I(height^3) is non-zero.
# F-test
linearHypothesis(reg5, c("I(height^2)", "I(height^3)"), test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## I(height^2) = 0
## I(height^3) = 0
##
## Model 1: restricted model
```

```
## Model 2: earnings ~ height + I(height^2) + I(height^3)
##
##   Res.Df Df      F Pr(>F)
## 1  17868
## 2  17866  2 0.1058 0.8996
```

```
# Under the null, the probability of observing an F-statistic as unfavourable to
# the null as that actually observed is 0.8996. Fail to reject the null at all
# levels of significance lower than 0.8996. Fail to reject the null at all
# conventional levels of significance (10%, 5%, 1%).

# Construct dummy variables for education categories
height$lths = as.numeric(height$educ <= 11)
height$hs = as.numeric(height$educ == 12)
height$scoll = as.numeric(height$educ >= 13 & height$educ <= 15)
height$coll = as.numeric(height$educ >= 16)

# Construct subset of female workers
height_women = subset(height, sex == "0")

# Regression of earnings on height
reg6 = lm_robust(earnings ~ height, data = height_women)
summary(reg6)
```

```
##
## Call:
## lm_robust(formula = earnings ~ height, data = height_women)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)  12650.9     6299.9   2.008 4.466e-02    301.7  25000.0 9972
## height         511.2       97.6   5.238 1.655e-07    319.9    702.5 9972
##
## Multiple R-squared:  0.002672 ,  Adjusted R-squared:  0.002572
## F-statistic: 27.44 on 1 and 9972 DF,  p-value: 1.655e-07
```

```
# Regression of earnings on height, lths, hs, scoll
reg7 = lm_robust(earnings ~ height + lths + hs + scoll, data = height_women)
summary(reg7)
```

```
##
## Call:
## lm_robust(formula = earnings ~ height + lths + hs + scoll, data = height_women)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)  CI Lower CI Upper   DF
## (Intercept)  50749.5    6004.58   8.452  3.270e-17  38979.33  62519.7 9969
## height         135.1      92.33   1.464  1.433e-01    -45.84    316.1 9969
## lths        -31857.8     835.15 -38.146 2.558e-297 -33494.87 -30220.7 9969
## hs          -20417.9     637.78 -32.014 3.776e-214 -21668.07 -19167.7 9969
## scoll       -12649.1     716.58 -17.652  1.082e-68 -14053.70 -11244.4 9969
##
```

```
## Multiple R-squared:  0.1382 ,    Adjusted R-squared:  0.1378
## F-statistic: 447.4 on 4 and 9969 DF,  p-value: < 2.2e-16
```

```
# Coefficient on height is smaller in the regression model of earnings on
# height, lths, hs, and scoll than in the regression model of earnings on
# height alone. This is consistent with the proposition that the apparently
# positive relationship between earnings and height is partially explained by
# the correlation between height and cognitive ability, which is relevant for
# earnings. The coefficient on height in reg7 estimates the relationship between
# earnings and the component of height uncorrelated with lths, hs, and scoll,
# which are proxies for cognitive ability. Holding lths, hs, and scoll fixed, on
# average, a one inch increase in height is associated with a $135.1 increase in
# annual labour earnings. This increase is significantly lower than the average
# $511.2 increase in annual labour earnings associated witha  one inch increase
# in height, where lths, hs, and scoll are not fixed.

# lths, hs, scoll, and coll are perfectly multicollinear. If all of lths, hs,
# scoll, and coll are included in the regression model, there is no unique
# solution for the estimates of the regression parameters.

# Hypothesis test
# H0: coefficients of lths, hs and scoll in regression of earnings on height,
# lths, hs, and scoll are all equal to zero.
# H1: at least one of the coefficients of lths, hs, and scoll in regression of
# earnings on height, lths, hs, scoll, is non-zero.
# F-test
linearHypothesis(reg7, c("lths", "hs", "scoll"), test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## lths = 0
## hs = 0
## scoll = 0
##
## Model 1: restricted model
## Model 2: earnings ~ height + lths + hs + scoll
##
##   Res.Df Df     F    Pr(>F)
## 1   9972
## 2   9969  3 577.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Under the null, the probability of observing an F-statistic as unfavourable to
# the null as that actually observed is 2.22e-16. Reject the null at any level
# of significance greater than 2.2e-16. Reject the null at all conventional
# levels of significance (10%, 5%, 1%).

# The coefficients of lths, hs, and scoll give the average change in earnings
# associated with having less than high school education, high school education,
# and some college education as opposed to having completed college education
# respectively, holding height fixed. It is unsurprising that the three
# coefficients are negative since college graduates are expected to have higher
# earnings than non-college graduates. It is also unsurprising that the
```

```
# coefficients increase in magnitude from scoll to hs to lths since we expect
# that individuals with some college education have higher earnings than
# individuals with only high school education, and individuals with high school
# education have higher earnings than individuals with less than high school
# education. The magnitude of the coefficients are large relative to the mean
# earnings.
```