Propensity Score Matching

Vaishnavi

2024-11-25

### Part a: Preparing necessary variables and packages

# Installing and load necessary packages  
if (!require(haven)) install.packages("haven")

## Loading required package: haven

## Warning: package 'haven' was built under R version 4.2.3

if (!require(MatchIt)) install.packages("MatchIt")

## Loading required package: MatchIt

## Warning: package 'MatchIt' was built under R version 4.2.3

library(haven)  
library(MatchIt)  
  
  
# First I am creating a new indicator variable for exports.   
  
data <- read\_dta("D://Downloads//Myanmardata.dta") ##Importing dta file  
  
data$exports <- ifelse(data$export\_sh > 0, 1, 0) #Create a new variable equal to 1 if the firm exports (export\_sh>0), and 0 otherwise

#### Part b: Propesnity Score Matching

# Convert oage, famowner, ogender, and oeduc to factors  
data$famowner <- as.factor(data$famowner)  
data$ogender <- as.factor(data$ogender)  
data$oeduc <- as.factor(data$oeduc)  
  
# Here, I specify first stage logit regression needed for the matching  
psm\_formula <- glm(exports ~ oage + fshare + famowner + ogender + oeduc + ap\_timeo, family = binomial(), data=data) #whether the firm exports is the treatment  
summary (psm\_formula)

##   
## Call:  
## glm(formula = exports ~ oage + fshare + famowner + ogender +   
## oeduc + ap\_timeo, family = binomial(), data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6843 -0.7837 0.1864 0.9282 1.8607   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.9628305 1.1039366 -0.872 0.3831   
## oage 0.0004268 0.0203275 0.021 0.9832   
## fshare 0.0345964 0.0075783 4.565 4.99e-06 \*\*\*  
## famowner2 1.7375855 0.4103303 4.235 2.29e-05 \*\*\*  
## ogender2 -0.2412194 0.3255372 -0.741 0.4587   
## oeduc3 0.7373563 0.3532108 2.088 0.0368 \*   
## ap\_timeo -1.1081953 0.5140684 -2.156 0.0311 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 352.26 on 259 degrees of freedom  
## Residual deviance: 242.98 on 253 degrees of freedom  
## (140 observations deleted due to missingness)  
## AIC: 256.98  
##   
## Number of Fisher Scoring iterations: 6

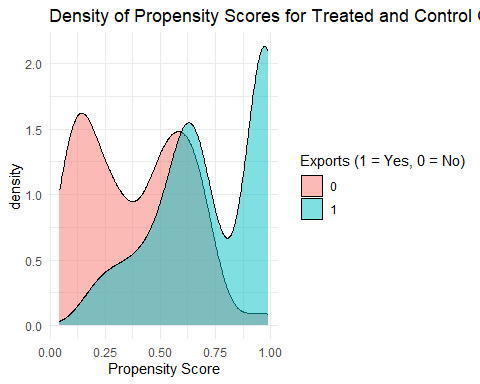
#Now, I predict the probability scores  
prs\_df <- data.frame(pr\_score = predict(psm\_formula, type= "response"), exports=psm\_formula$model$exports)  
head(prs\_df)

## pr\_score exports  
## 1 0.9886374 1  
## 2 0.9886374 1  
## 3 0.9765382 1  
## 4 0.9874752 1  
## 5 0.9874752 1  
## 6 0.9874752 1

#I want to check if the common support assumption holds for the continuous variables  
library(ggplot2) # Loading the library

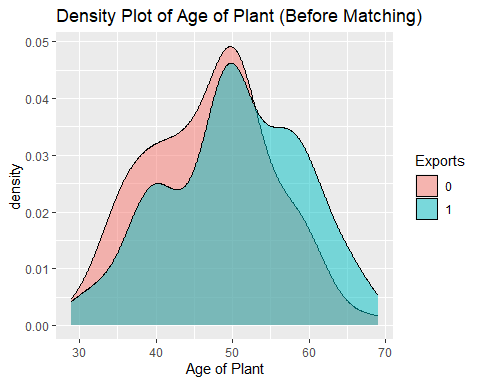
## Warning: package 'ggplot2' was built under R version 4.2.3

# Plotting the propensity scores for treated and control groups  
ggplot(prs\_df, aes(x = pr\_score, fill = as.factor(exports))) + geom\_density(alpha = 0.5) +   
labs(title = "Density of Propensity Scores for Treated and Control Groups", x = "Propensity Score",  
fill = "Exports (1 = Yes, 0 = No)") + theme\_minimal()

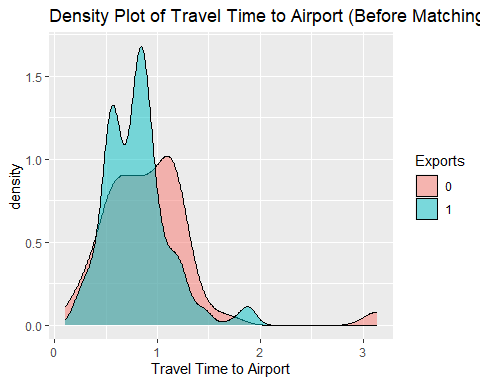


# Plotting the distribution of each covariate by treatment group (before matching)  
  
##Age  
ggplot(data, aes(x = oage, fill = as.factor(exports))) + geom\_density(alpha = 0.5) +   
labs(title = "Density Plot of Age of Plant (Before Matching)", x = "Age of Plant", fill = "Exports")

## Warning: Removed 140 rows containing non-finite values (`stat\_density()`).



##Travel Time to Airport  
ggplot(data, aes(x = ap\_timeo, fill = as.factor(exports))) + geom\_density(alpha = 0.5) +   
labs(title = "Density Plot of Travel Time to Airport (Before Matching)", x = "Travel Time to Airport", fill = "Exports")



### Propensity score matching with replacement  
  
# Remove rows with missing values in the matching variables  
data\_clean <- data[complete.cases(data[, c("oage", "fshare", "famowner", "ogender", "oeduc", "ap\_timeo")]), ]  
  
# PSM with the cleaned data  
psm\_model <- matchit(exports ~ oage + fshare + famowner + ogender + oeduc + ap\_timeo, data = data\_clean, method = "nearest", replace = TRUE)  
summary(psm\_model) ##Model summary/output

##   
## Call:  
## matchit(formula = exports ~ oage + fshare + famowner + ogender +   
## oeduc + ap\_timeo, data = data\_clean, method = "nearest",   
## replace = TRUE)  
##   
## Summary of Balance for All Data:  
## Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean  
## distance 0.7317 0.3836 1.4346 1.1072 0.3290  
## oage 50.2549 47.0280 0.3641 1.1887 0.0863  
## fshare 43.7908 1.8692 0.8422 13.3816 0.2096  
## famowner1 0.0850 0.4673 -1.3712 . 0.3823  
## famowner2 0.9150 0.5327 1.3712 . 0.3823  
## ogender1 0.7451 0.5421 0.4659 . 0.2030  
## ogender2 0.2549 0.4579 -0.4659 . 0.2030  
## oeduc2 0.2745 0.4299 -0.3482 . 0.1554  
## oeduc3 0.7255 0.5701 0.3482 . 0.1554  
## ap\_timeo 0.8248 0.8777 -0.1766 0.8666 0.0647  
## eCDF Max  
## distance 0.5264  
## oage 0.1950  
## fshare 0.4192  
## famowner1 0.3823  
## famowner2 0.3823  
## ogender1 0.2030  
## ogender2 0.2030  
## oeduc2 0.1554  
## oeduc3 0.1554  
## ap\_timeo 0.2226  
##   
## Summary of Balance for Matched Data:  
## Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean  
## distance 0.7317 0.7283 0.0141 0.8639 0.0284  
## oage 50.2549 46.1765 0.4602 0.9536 0.1281  
## fshare 43.7908 43.1373 0.0131 0.8550 0.0033  
## famowner1 0.0850 0.0654 0.0703 . 0.0196  
## famowner2 0.9150 0.9346 -0.0703 . 0.0196  
## ogender1 0.7451 0.3791 0.8399 . 0.3660  
## ogender2 0.2549 0.6209 -0.8399 . 0.3660  
## oeduc2 0.2745 0.2353 0.0879 . 0.0392  
## oeduc3 0.7255 0.7647 -0.0879 . 0.0392  
## ap\_timeo 0.8248 0.8892 -0.2149 0.8957 0.0882  
## eCDF Max Std. Pair Dist.  
## distance 0.2614 0.0309  
## oage 0.3922 1.2214  
## fshare 0.0065 0.0131  
## famowner1 0.0196 0.1641  
## famowner2 0.0196 0.1641  
## ogender1 0.3660 1.4397  
## ogender2 0.3660 1.4397  
## oeduc2 0.0392 0.6151  
## oeduc3 0.0392 0.6151  
## ap\_timeo 0.3660 0.8372  
##   
## Sample Sizes:  
## Control Treated  
## All 107. 153  
## Matched (ESS) 6.52 153  
## Matched 36. 153  
## Unmatched 71. 0  
## Discarded 0. 0

### Part c: Summarizing the matched data

matched\_data <- match.data(psm\_model)  
dim(matched\_data)

## [1] 189 16

# Means for each covariate in the full dataset (before matching)  
summary\_full <- aggregate(cbind(oage, fshare, famowner, ogender, oeduc, ap\_timeo) ~ exports, data = data, FUN = mean)  
print(summary\_full)

## exports oage fshare famowner ogender oeduc ap\_timeo  
## 1 0 47.02804 1.869159 1.532710 1.457944 1.570093 0.8777274  
## 2 1 50.25490 43.790850 1.915033 1.254902 1.725490 0.8248137

# Means for each covariate in the matched dataset (after matching)  
summary\_matched <- aggregate(cbind(oage, fshare, famowner, ogender, oeduc, ap\_timeo) ~ exports, data = matched\_data, FUN = mean)  
  
print(summary\_matched) ##show output

## exports oage fshare famowner ogender oeduc ap\_timeo  
## 1 0 46.88889 5.555556 1.805556 1.444444 1.666667 0.8751870  
## 2 1 50.25490 43.790850 1.915033 1.254902 1.725490 0.8248137

### Part d: Regressions

##To test if exporting firms pay higher salary than non-exporters  
# Run regression on matched data  
reg\_matched <- lm(salaryd ~ exports, data = matched\_data) ##without covars  
reg\_matched\_with\_covars <- lm(salaryd ~ exports + oage + fshare + famowner + ogender + oeduc + ap\_timeo, data = matched\_data) ##with covars  
  
# Run regression on full (unmatched) data  
reg\_full <- lm(salaryd ~ exports, data = data) ##without covars  
reg\_full\_with\_covars <- lm(salaryd ~ exports+ oage + fshare + famowner + ogender + oeduc + ap\_timeo, data = data) ##with covars  
  
# Show regression results for matched and full data  
summary(reg\_matched)

##   
## Call:  
## lm(formula = salaryd ~ exports, data = matched\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -77.565 -12.263 -2.521 8.006 83.869   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 82.798 3.626 22.837 <2e-16 \*\*\*  
## exports 3.933 4.030 0.976 0.33   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.75 on 187 degrees of freedom  
## Multiple R-squared: 0.005069, Adjusted R-squared: -0.000251   
## F-statistic: 0.9528 on 1 and 187 DF, p-value: 0.3303

summary(reg\_matched\_with\_covars)

##   
## Call:  
## lm(formula = salaryd ~ exports + oage + fshare + famowner + ogender +   
## oeduc + ap\_timeo, data = matched\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -75.797 -10.258 -2.493 8.133 84.423   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 94.12788 11.23553 8.378 1.46e-14 \*\*\*  
## exports 5.38870 4.31155 1.250 0.2130   
## oage -0.14413 0.19899 -0.724 0.4698   
## fshare 0.01684 0.03753 0.449 0.6542   
## famowner2 -10.13270 5.76477 -1.758 0.0805 .   
## ogender2 0.08711 3.76842 0.023 0.9816   
## oeduc3 -2.97493 3.57991 -0.831 0.4071   
## ap\_timeo 6.21775 5.37918 1.156 0.2492   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.7 on 181 degrees of freedom  
## Multiple R-squared: 0.04131, Adjusted R-squared: 0.004229   
## F-statistic: 1.114 on 7 and 181 DF, p-value: 0.3561

summary(reg\_full)

##   
## Call:  
## lm(formula = salaryd ~ exports, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -76.567 -10.734 -1.523 8.719 91.081   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 75.586 1.454 51.980 < 2e-16 \*\*\*  
## exports 10.148 1.956 5.187 3.4e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.45 on 398 degrees of freedom  
## Multiple R-squared: 0.06333, Adjusted R-squared: 0.06097   
## F-statistic: 26.91 on 1 and 398 DF, p-value: 3.405e-07

summary(reg\_full\_with\_covars)

##   
## Call:  
## lm(formula = salaryd ~ exports + oage + fshare + famowner + ogender +   
## oeduc + ap\_timeo, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79.263 -10.994 -2.956 8.158 87.834   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 90.49875 8.83122 10.248 < 2e-16 \*\*\*  
## exports 9.27669 3.18882 2.909 0.00395 \*\*   
## oage -0.29341 0.15887 -1.847 0.06594 .   
## fshare 0.02822 0.03452 0.817 0.41455   
## famowner2 -3.31411 3.52006 -0.941 0.34735   
## ogender2 0.89980 2.87888 0.313 0.75488   
## oeduc3 0.81773 2.80761 0.291 0.77110   
## ap\_timeo 3.24357 4.17976 0.776 0.43847   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 20.51 on 252 degrees of freedom  
## (140 observations deleted due to missingness)  
## Multiple R-squared: 0.0606, Adjusted R-squared: 0.03451   
## F-statistic: 2.322 on 7 and 252 DF, p-value: 0.02596

### Part e: Limited dependent variable regressions

library(margins)  
  
# To test if firms that export are more likely to have fire exits.  
  
# Logit regression  
logit\_model <- glm(fexit ~ exports, data = matched\_data, family = binomial) ##without covars  
  
logit\_model\_with\_covars <- glm(fexit ~ exports + oage + fshare + famowner + ogender + oeduc + ap\_timeo, data = matched\_data, family = binomial) ##with covars  
  
# Display logit regression results  
summary(logit\_model)

##   
## Call:  
## glm(formula = fexit ~ exports, family = binomial, data = matched\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.043 0.515 0.515 0.515 0.854   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.8210 0.3618 2.269 0.02326 \*   
## exports 1.1324 0.4370 2.591 0.00957 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 165.40 on 188 degrees of freedom  
## Residual deviance: 159.12 on 187 degrees of freedom  
## AIC: 163.12  
##   
## Number of Fisher Scoring iterations: 4

summary(logit\_model\_with\_covars)

##   
## Call:  
## glm(formula = fexit ~ exports + oage + fshare + famowner + ogender +   
## oeduc + ap\_timeo, family = binomial, data = matched\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3397 0.3630 0.4386 0.5871 1.5789   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.231158 1.500496 -0.821 0.4119   
## exports 0.981327 0.494514 1.984 0.0472 \*  
## oage 0.022980 0.027738 0.828 0.4074   
## fshare -0.001633 0.005503 -0.297 0.7666   
## famowner2 1.071075 0.634151 1.689 0.0912 .  
## ogender2 -0.785319 0.468206 -1.677 0.0935 .  
## oeduc3 -0.268122 0.505939 -0.530 0.5961   
## ap\_timeo 0.837786 0.770302 1.088 0.2768   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 165.40 on 188 degrees of freedom  
## Residual deviance: 147.06 on 181 degrees of freedom  
## AIC: 163.06  
##   
## Number of Fisher Scoring iterations: 5

# Marginal effects for Logit model  
logit\_marginal\_effects <- margins(logit\_model) ##marginal effects (𝜕𝑌/𝜕𝑋)   
logit\_with\_covars\_marginal\_effects <- margins(logit\_model\_with\_covars)  
print("Logit Model Marginal Effects:")

## [1] "Logit Model Marginal Effects:"

summary(logit\_marginal\_effects)

## factor AME SE z p lower upper  
## exports 0.1455 0.0545 2.6682 0.0076 0.0386 0.2523

summary(logit\_with\_covars\_marginal\_effects)

## factor AME SE z p lower upper  
## ap\_timeo 0.0990 0.0905 1.0946 0.2737 -0.0783 0.2764  
## exports 0.1160 0.0572 2.0297 0.0424 0.0040 0.2280  
## famowner2 0.1622 0.1159 1.3995 0.1617 -0.0649 0.3893  
## fshare -0.0002 0.0006 -0.2971 0.7664 -0.0015 0.0011  
## oage 0.0027 0.0033 0.8297 0.4067 -0.0037 0.0091  
## oeduc3 -0.0307 0.0560 -0.5486 0.5833 -0.1405 0.0790  
## ogender2 -0.1019 0.0654 -1.5571 0.1195 -0.2301 0.0264

# Probit regression  
probit\_model <- glm(fexit ~ exports, data = matched\_data, family = binomial(link = "probit")) ##without covars  
probit\_model\_with\_covars <- glm(fexit ~ exports + oage + fshare + famowner + ogender + oeduc + ap\_timeo, data = matched\_data, family = binomial(link = "probit")) ##with covars  
  
# Display probit regression results  
summary(probit\_model)

##   
## Call:  
## glm(formula = fexit ~ exports, family = binomial(link = "probit"),   
## data = matched\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.043 0.515 0.515 0.515 0.854   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.5085 0.2190 2.322 0.0202 \*  
## exports 0.6458 0.2547 2.535 0.0112 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 165.40 on 188 degrees of freedom  
## Residual deviance: 159.12 on 187 degrees of freedom  
## AIC: 163.12  
##   
## Number of Fisher Scoring iterations: 4

summary(probit\_model\_with\_covars)

##   
## Call:  
## glm(formula = fexit ~ exports + oage + fshare + famowner + ogender +   
## oeduc + ap\_timeo, family = binomial(link = "probit"), data = matched\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3480 0.3510 0.4355 0.6019 1.5148   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.6661208 0.8264861 -0.806 0.4203   
## exports 0.5326445 0.2842288 1.874 0.0609 .  
## oage 0.0123795 0.0149748 0.827 0.4084   
## fshare -0.0005926 0.0029235 -0.203 0.8394   
## famowner2 0.6228632 0.3682037 1.692 0.0907 .  
## ogender2 -0.4280618 0.2606396 -1.642 0.1005   
## oeduc3 -0.1453111 0.2734838 -0.531 0.5952   
## ap\_timeo 0.4931091 0.4205376 1.173 0.2410   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 165.40 on 188 degrees of freedom  
## Residual deviance: 147.11 on 181 degrees of freedom  
## AIC: 163.11  
##   
## Number of Fisher Scoring iterations: 5

# Marginal effects for Probit model  
probit\_marginal\_effects <- margins(probit\_model) ##marginal effects (𝜕𝑌/𝜕𝑋)   
probit\_with\_covars\_marginal\_effects <- margins(probit\_model\_with\_covars)  
print("Probit Model Marginal Effects:")

## [1] "Probit Model Marginal Effects:"

summary(probit\_marginal\_effects)

## factor AME SE z p lower upper  
## exports 0.1503 0.0575 2.6118 0.0090 0.0375 0.2630

summary(probit\_with\_covars\_marginal\_effects)

## factor AME SE z p lower upper  
## ap\_timeo 0.1060 0.0899 1.1793 0.2383 -0.0702 0.2821  
## exports 0.1145 0.0601 1.9062 0.0566 -0.0032 0.2322  
## famowner2 0.1672 0.1165 1.4344 0.1515 -0.0613 0.3956  
## fshare -0.0001 0.0006 -0.2028 0.8393 -0.0014 0.0011  
## oage 0.0027 0.0032 0.8281 0.4076 -0.0036 0.0090  
## oeduc3 -0.0304 0.0555 -0.5475 0.5841 -0.1391 0.0784  
## ogender2 -0.1002 0.0655 -1.5311 0.1258 -0.2285 0.0281

# Linear Regression  
lr\_model <- lm(fexit ~ exports, data = matched\_data) ##without covars  
lr\_model\_with\_covars <- lm(fexit ~ exports + oage + fshare + famowner + ogender + oeduc + ap\_timeo, data = matched\_data) ##with covars  
  
# Display LPM regression results  
summary(lr\_model)

##   
## Call:  
## lm(formula = fexit ~ exports, data = matched\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8758 0.1242 0.1242 0.1242 0.3056   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.69444 0.06005 11.564 <2e-16 \*\*\*  
## exports 0.18137 0.06675 2.717 0.0072 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3603 on 187 degrees of freedom  
## Multiple R-squared: 0.03799, Adjusted R-squared: 0.03284   
## F-statistic: 7.384 on 1 and 187 DF, p-value: 0.0072

summary(lr\_model\_with\_covars)

##   
## Call:  
## lm(formula = fexit ~ exports + oage + fshare + famowner + ogender +   
## oeduc + ap\_timeo, data = matched\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.95418 0.04889 0.09368 0.18294 0.59219   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4054800 0.1825091 2.222 0.0275 \*  
## exports 0.1472398 0.0700364 2.102 0.0369 \*  
## oage 0.0022035 0.0032324 0.682 0.4963   
## fshare -0.0002658 0.0006096 -0.436 0.6633   
## famowner2 0.2092057 0.0936424 2.234 0.0267 \*  
## ogender2 -0.1100060 0.0612139 -1.797 0.0740 .  
## oeduc3 -0.0307642 0.0581517 -0.529 0.5974   
## ap\_timeo 0.1005455 0.0873790 1.151 0.2514   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3526 on 181 degrees of freedom  
## Multiple R-squared: 0.1085, Adjusted R-squared: 0.07407   
## F-statistic: 3.148 on 7 and 181 DF, p-value: 0.003653