Week 4 Code

Ryan Martin

January 24, 2018

Reminders for Exam

- Take a calculator to calculate p-values
- Bring ID cards
- Be sure to work through past exams, review stata code (all code learned from lab lecture notes should be enough for this class)
- Bring a watch and manage your time carefully
- Double check your work if you have time.
- Familiarize yourself with the formula sheet. There is a lot on there!
- If a question is unclear about the significance level or kind of test, you should ask! But the default is typically two-sided, .95

Chapter 4 Brief Notes

- f for forecast error
- BLUE vs BLUP. Best Linear Unbiased Estimator. Best Linear Unbiased Predictor, yhat is blup, b1 and b2 (these are the betahats, the estimators) are blue.
- $$var(f) = \sigma^2[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}]$$

$var(f)$ is decreasing in N, variation in explanatory variable, and as the point of interest is closer to the mean and as σ^2 (the variance of noise term e) decreases

- estimate of variance of forecast error

$$\widehat{var(f)} = \hat{\sigma}^2[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}]$$

- prediction interval is $\hat{y_0} \pm t_c se(f)$
- compare with CI for a point on the regression line corresponding to $x = x_0$, $b_1 + x_0 b_2 : b_1 + x_0 b_2 \pm t_c se(b_1 + x_0 b_2)$
- Note, in linear regression $\hat{y_0} = b_1 + b_2 x_0$, $\frac{\hat{\sigma^2}}{\sum(x_i - \bar{x})^2} = var\hat{(}b_2)$
- SST = SSR + SSE. SST $= \sum(y_i - \bar{y})^2$. SSR is $\sum(\hat{y_i} - \bar{y})$, SSE is $\sum \hat{e_i}^2$
- $R^2 := SSR/SST = 1 - SSE/SST$ (definition!). So, closer R^2 is to 1, the closer SSE is to 0. Measure of how well your model predicts the data. Interpretation of R^2 is proportion of variation in y about its mean that is explained by regression model.
- sample correlation coefficient, $r_{xy} = \frac{s_{xy}}{s_x s_y}$ measures strength of linear association between x and y. Note that $r_{xy}^2 = R^2 \geq 0$ (because square) in simple (bivariate, 2 term) regression model. Also note that

$R^2 = r^2_{y\hat{y}}$ as well. The second one holds true in more complicated (multiple) regressions. NOte, "y" here is the dependent variable - the variable on the left hand side. For example, for a log-linear regression,

$$\log(price) = \beta_1 + beta_2 quantity$$

, $y = \log(price)$, so your R^2 in this case is $R^2 = r_{\log(price), \hat{log}(price)}$

- How to choose fits:

1. Choose shapes that are consistent with economic theory
2. Choose shapes that are sufficiently flexible to fit the data
3. Choose shapes so SR1-SR6 are verified (can use diagnostic residual plots).

4.5.3 Prediction in Log-Linear Model.

- Consider the estimated regression equation $\widehat{ln(y)} = b_1 + b_2 x$. To predict y from our predictions of $\widehat{ln(y)}$ we use either $\hat{y_n}$ or $\hat{y_c}$.
- y_n is defined by
$$\hat{y_n} = \exp(\widehat{ln(y)}) = exp(b_1 + b_2 x)$$
This one is a natural one and actually probably best for small sample sizes (under 30). This is also the the one we have been using so far in the course.
- Alternatively, y_c is defined by $\hat{y_c} := \exp(b_1 + b_2 x + \hat{\sigma^2}/2) = \hat{y}_n e^{\hat{\sigma}^2/2}$ This is for larger samples. It's derivation is shown in appendix 4c, from properties of log-normal distribution
- Note, y_c is not as good in smaller samples because $\hat{\sigma}^2$ is unlikely to be accurate in small samples.
- Note, since $\hat{\sigma}^2 \geq 0$, this correction increases the value of our prediciotn. The natural predictor tends to systematically underpredict the value of y in a log-linear model.
- If $\hat{y_n}$ denotes the natural predictor, we can define the generalized R^2 for the log-linear model as $R^2_g := r^2_{y\hat{y_n}}$ This is in contrast to the standard R^2, mentioned above, which would be $R^2 = r^2_{logy, \widehat{\log(y)}}$. Further note that
$$corr(aX, Y) = \frac{cov(aX, Y)}{\sqrt{var(aX)var(Y)}} = \frac{acov(X, Y)}{|a|\sqrt{var(X)var(Y)}} = sign(a)corr(X, Y)$$
where
$$sign(a) = \begin{cases} 1 \text{ if } a > 0 \\ 0 \text{ if } a = 0 \\ -1 \text{ if } a < 0 \end{cases}$$
Finally, we can therefore conclude, that since $e^{\hat{\sigma}^2/2} y_n = y_c$ and $e^{\hat{\sigma}^2/2} > 0$, that $corr(\hat{y}_n, y) = corr(\hat{y}_c, y)$ and so $R^2_g = r^2_{y, \hat{y_n}} = r^2_{y, \hat{y_c}}$ for the linear and log-linear models.
- Note R^2 is a.k.a. coefficient of determination.
- Interval prediction for log-linear model does NOT use corrected predictors. It uses the natural predictor, $\hat{y_n}$. recall f is the forecast error; $f = \hat{y_n} - y$ We have $[\exp(\widehat{ln(y)} - t_c se(f)), \exp(\widehat{ln(y)} + t_c se(f))]$

4.6 Log-Log: $log(y) = \beta_1 + \beta_2 log(x)$

- Here, slope is elasticity. %change response in y per 1 % change in x.
- Note, the correction here is similar form as the correction in log-linear. $\hat{y_c} = \hat{y_n} e^{\hat{\sigma}^2/2} = \exp(b_1 + b_2 logx) e^{\hat{\sigma}^2/2}$. The generalized Rˆ2 is still used, $R^2_g = corr(y, \hat{y_c})$

4.15

Does the return to education differ by race and gender? For this exercise, use the file `cps4.dat` (This is a large file with 4,838 observations. If you are using the student version of Stata software, you can use the smaller file `cps4_small.dat`. If you are using R, the size shouldn't be a problem. R runs into problems around the 10 million entries point, whereupon you may need to do some fancier technique and use some packages, but can still get the job done without paying.) In this exercise you will extract subsamples of observations consisting of (i) all males, (ii) all females (iii) all whites, (iv) all blacks, (v) white males, (vii) black males and (vii) black females.

a

For each sample partition, obtain the summary statistics of $WAGE$

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "cps4.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)

dat_male <- dat[dat$female==0,]
dat_female <- dat[dat$female==1,]
dat_white <- dat[dat$white==1,]
dat_black <- dat[dat$black==1,]
dat_wm <- dat[( (dat$white==1)&(dat$female==0) ),]
dat_bm <- dat[( (dat$black==1)&(dat$female==0) ),]
dat_bf <- dat[( (dat$black==1)&(dat$female==1) ),]

#creating a list to hold all datasets
#so can write loops in code
dat_list <- list(dat_male,
                 dat_female,
                 dat_white,
                 dat_black,
                 dat_wm,
                 dat_bm,
                 dat_bf)
my_names = c( "male",
                 "female",
                 "white",
                 "black",
                 "white and male",
                 "black and male",
                 "black and female")
m = length(dat_list)
count = 1
for (mydat in dat_list){
  print( paste("This is the data subset of everyone who is",
               my_names[count]))
  count = 1 + count
  print("Summary Statistics of this subset:")
  print(summary(mydat$wage))
}
```

```
## [1] "This is the data subset of everyone who is male"
```

```
## [1] "Summary Statistics of this subset:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   12.75   19.10   22.26   28.05  173.00
## [1] "This is the data subset of everyone who is female"
## [1] "Summary Statistics of this subset:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.14   10.00   15.00   18.05   22.05   96.17
## [1] "This is the data subset of everyone who is white"
## [1] "Summary Statistics of this subset:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.14   11.64   17.00   20.48   25.61  173.00
## [1] "This is the data subset of everyone who is black"
## [1] "Summary Statistics of this subset:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   13.45   16.44   20.00   72.13
## [1] "This is the data subset of everyone who is white and male"
## [1] "Summary Statistics of this subset:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.50   13.23   19.24   22.83   28.83  173.00
## [1] "This is the data subset of everyone who is black and male"
## [1] "Summary Statistics of this subset:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   13.47   16.21   19.23   72.13
## [1] "This is the data subset of everyone who is black and female"
## [1] "Summary Statistics of this subset:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.75    9.75   13.45   16.62   20.09   72.13
```

b

A variable's *coefficient of variation* is 100 times the ratio of its sample standard deviation to its sample mean. That is, for a variable y it is

$$CV(y) := 100 \times \frac{s_y}{\bar{y}}$$

where s_y is y's standard deviation and $\bar{y}$ is y's average. What is the coefficient of variation for $WAGE$ within each sample partition?

```r
CV = rep(0,m) #creating vector to hold CV estimates
counter = 1
for (mydat in dat_list){
  CV[counter] = 100*sd(mydat$wage) /mean(mydat$wage)
  counter = 1 + counter
}
names(CV) <- paste("CV", my_names) #note, paste can be vectorized
  #which means it will create a vector with each entry
  # CV and the my_names entry at that position, pretty cool
CV
```

```
##             CV male           CV female            CV white
##            60.52906            61.79755            61.69531
##            CV black   CV white and male   CV black and male
##            61.63910            59.86918            58.54793
## CV black and female
##            63.87431
```

Table 4.1 **Some Useful Functions, their Derivatives, Elasticities and Other Interpretation**

Name	Function	Slope $= dy/dx$	Elasticity
Linear	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Quadratic	$y = \beta_1 + \beta_2 x^2$	$2\beta_2 x$	$(2\beta_2 x)\frac{x}{y}$
Cubic	$y = \beta_1 + \beta_2 x^3$	$3\beta_2 x^2$	$(3\beta_2 x^2)\frac{x}{y}$
Log-Log	$\ln(y) = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{y}{x}$	β_2
Log-Linear	$\ln(y) = \beta_1 + \beta_2 x$	$\beta_2 y$	$\beta_2 x$
	or, a 1 unit change in x leads to (approximately) a 100 β_2% change in y		
Linear-Log	$y = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$
	or, a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y		

Figure 1: Figure to help problem solving

c

For each sample partition, estimate the log-linear model

$$\log(WAGE) = \beta_1 + \beta_2 EDUC + e$$

What is the approximate percentage return to another year of education for each group? *Solution*: Note, by the table from the text pictured in this file, the percentage return is $100 \times \beta_2$. This is the percent change in y for a 1 unit change in x.

```
reg_out_vec <- vector(mode = "list", length = m)
returns_vec <- rep(0,m)
counter = 1
for (mydat in dat_list){
  s <- lm(I(log(mydat$wage)) ~ mydat$educ)
  reg_out_vec[[counter]] <- s
  print( paste("This is the data subset of everyone who is",
              my_names[counter]))
  print(summary(s))
  returns_vec[counter] <- as.numeric(coef(s)[2]*100)
  print("The semielasticity for this group is")
  print(as.numeric(coef(s)[2]*100))
  counter = 1 + counter
  print("The R squared for this group is")
  w <- summary(s)
  print(w$r.squared)
}
```

```
## [1] "This is the data subset of everyone who is male"
##
## Call:
## lm(formula = I(log(mydat$wage)) ~ mydat$educ)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79305 -0.31649 -0.00222  0.36355  2.18350
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.732617   0.049857   34.75   <2e-16 ***
## mydat$educ  0.088370   0.003565   24.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.51 on 2393 degrees of freedom
## Multiple R-squared:  0.2043, Adjusted R-squared:  0.204
## F-statistic: 614.5 on 1 and 2393 DF,  p-value: < 2.2e-16
##
## [1] "The semielasticity for this group is"
## [1] 8.836967
## [1] "The R squared for this group is"
## [1] 0.204334
## [1] "This is the data subset of everyone who is female"
##
## Call:
## lm(formula = I(log(mydat$wage)) ~ mydat$educ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81403 -0.31591 -0.00338  0.30661  1.97180
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.242679   0.055943   22.21   <2e-16 ***
## mydat$educ  0.106399   0.003927   27.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4896 on 2441 degrees of freedom
## Multiple R-squared:  0.2312, Adjusted R-squared:  0.2309
## F-statistic:   734 on 1 and 2441 DF,  p-value: < 2.2e-16
##
## [1] "The semielasticity for this group is"
## [1] 10.63988
## [1] "The R squared for this group is"
## [1] 0.231186
## [1] "This is the data subset of everyone who is white"
##
## Call:
## lm(formula = I(log(mydat$wage)) ~ mydat$educ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91828 -0.33543 -0.00356  0.34890  2.28609
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.592440   0.041147    38.7   <2e-16 ***
## mydat$educ  0.091054   0.002909    31.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5152 on 4114 degrees of freedom
## Multiple R-squared:  0.1923, Adjusted R-squared:  0.1921
## F-statistic: 979.8 on 1 and 4114 DF,  p-value: < 2.2e-16
##
## [1] "The semielasticity for this group is"
## [1] 9.105445
## [1] "The R squared for this group is"
## [1] 0.1923438
## [1] "This is the data subset of everyone who is black"
##
## Call:
## lm(formula = I(log(mydat$wage)) ~ mydat$educ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50777 -0.31036 -0.02286  0.26157  1.98107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.245588   0.127762   9.749   <2e-16 ***
## mydat$educ  0.105182   0.009398  11.192   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4659 on 491 degrees of freedom
## Multiple R-squared:  0.2033, Adjusted R-squared:  0.2016
## F-statistic: 125.3 on 1 and 491 DF,  p-value: < 2.2e-16
##
## [1] "The semielasticity for this group is"
## [1] 10.51817
## [1] "The R squared for this group is"
## [1] 0.2032506
## [1] "This is the data subset of everyone who is white and male"
##
## Call:
## lm(formula = I(log(mydat$wage)) ~ mydat$educ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4192 -0.3045 -0.0012  0.3555  2.1564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.790900   0.052184   34.32   <2e-16 ***
## mydat$educ  0.086145   0.003725   23.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.5041 on 2063 degrees of freedom
## Multiple R-squared:  0.2059, Adjusted R-squared:  0.2055
## F-statistic: 534.9 on 1 and 2063 DF,  p-value: < 2.2e-16
##
## [1] "The semielasticity for this group is"
## [1] 8.614524
## [1] "The R squared for this group is"
## [1] 0.2059031
## [1] "This is the data subset of everyone who is black and male"
##
## Call:
## lm(formula = I(log(mydat$wage)) ~ mydat$educ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56621 -0.31152 -0.01395  0.32666  1.40756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.65210    0.21048   7.849 2.04e-13 ***
## mydat$educ   0.07618    0.01584   4.809 2.88e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4908 on 212 degrees of freedom
## Multiple R-squared:  0.09835,    Adjusted R-squared:  0.09409
## F-statistic: 23.12 on 1 and 212 DF,  p-value: 2.878e-06
##
## [1] "The semielasticity for this group is"
## [1] 7.617588
## [1] "The R squared for this group is"
## [1] 0.09834774
## [1] "This is the data subset of everyone who is black and female"
##
## Call:
## lm(formula = I(log(mydat$wage)) ~ mydat$educ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18094 -0.30997 -0.04592  0.23565  2.07730
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.93953    0.15918   5.902 1.04e-08 ***
## mydat$educ   0.12616    0.01151  10.958  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4406 on 277 degrees of freedom
## Multiple R-squared:  0.3024, Adjusted R-squared:  0.2999
## F-statistic: 120.1 on 1 and 277 DF,  p-value: < 2.2e-16
##
## [1] "The semielasticity for this group is"
## [1] 12.6164
```

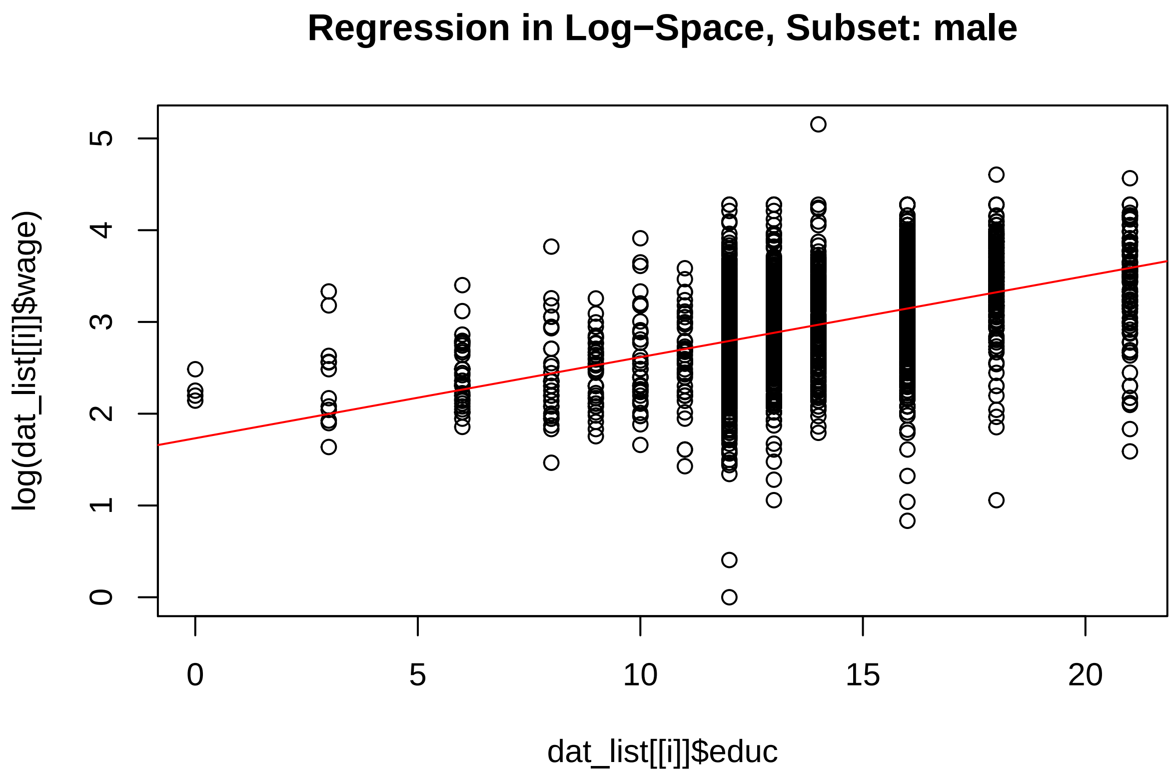
```
## [1] "The R squared for this group is"
## [1] 0.3024115
```

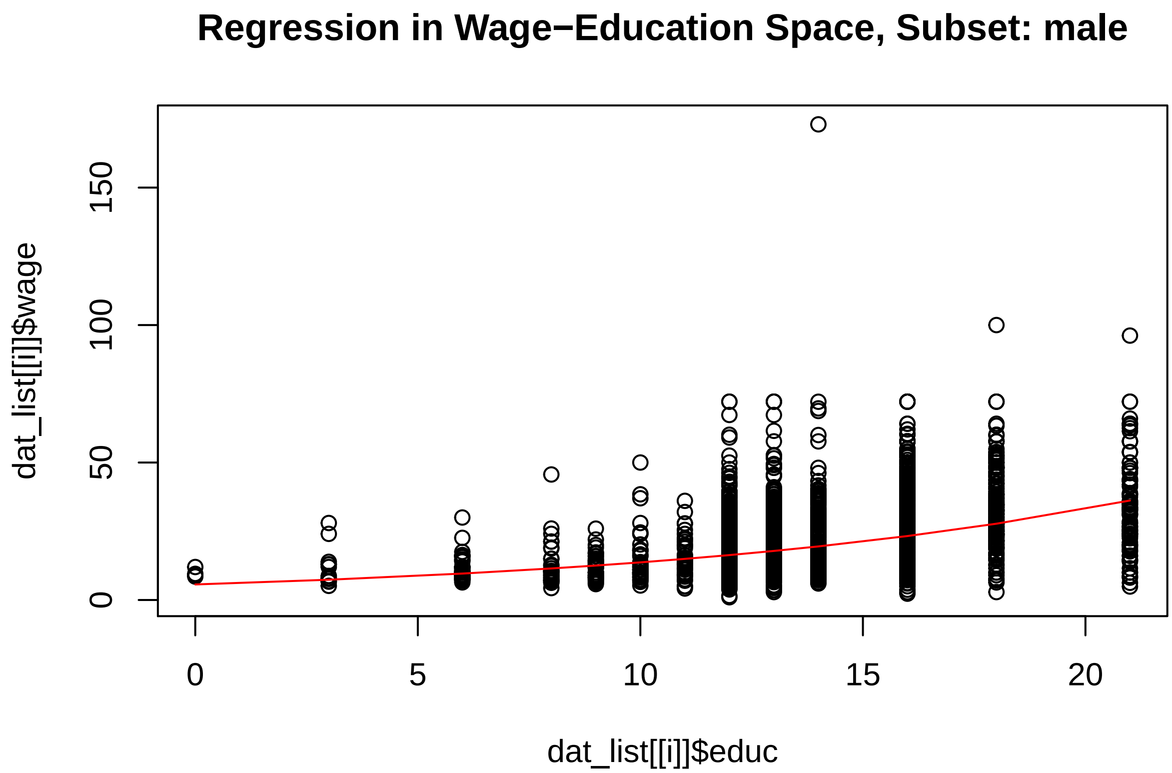
d

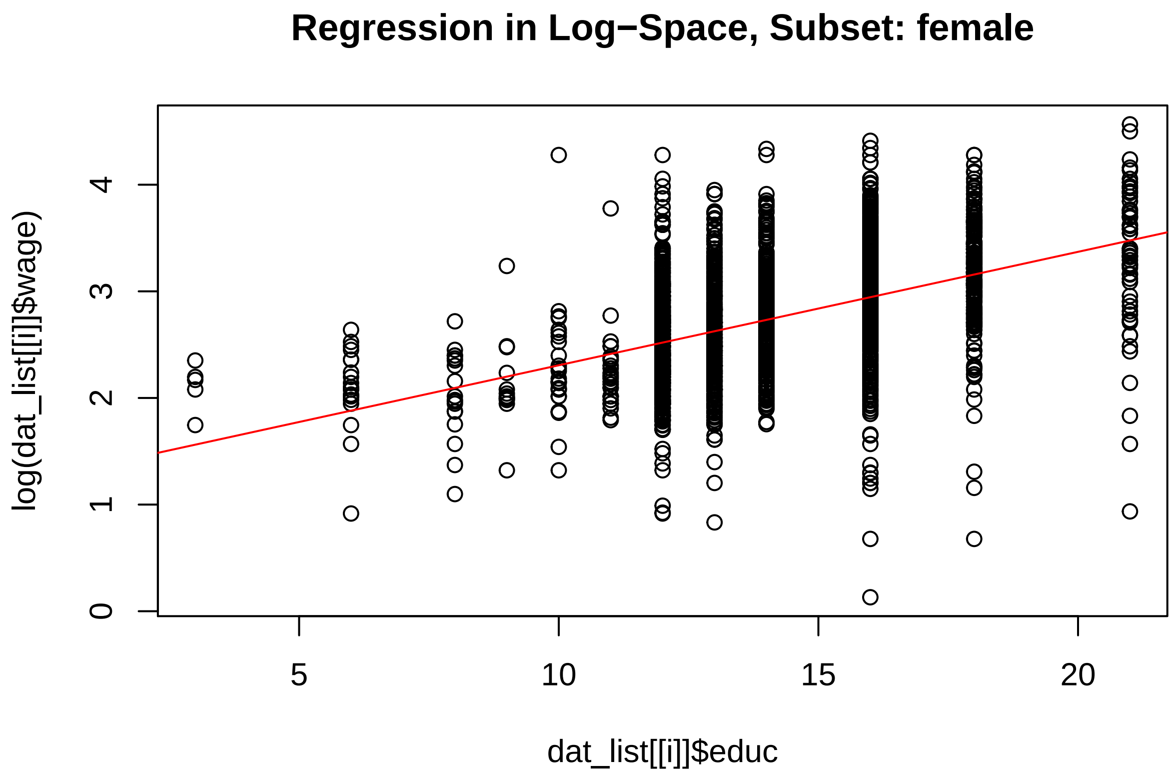
Does the model fit the data equally well for each sample partition?

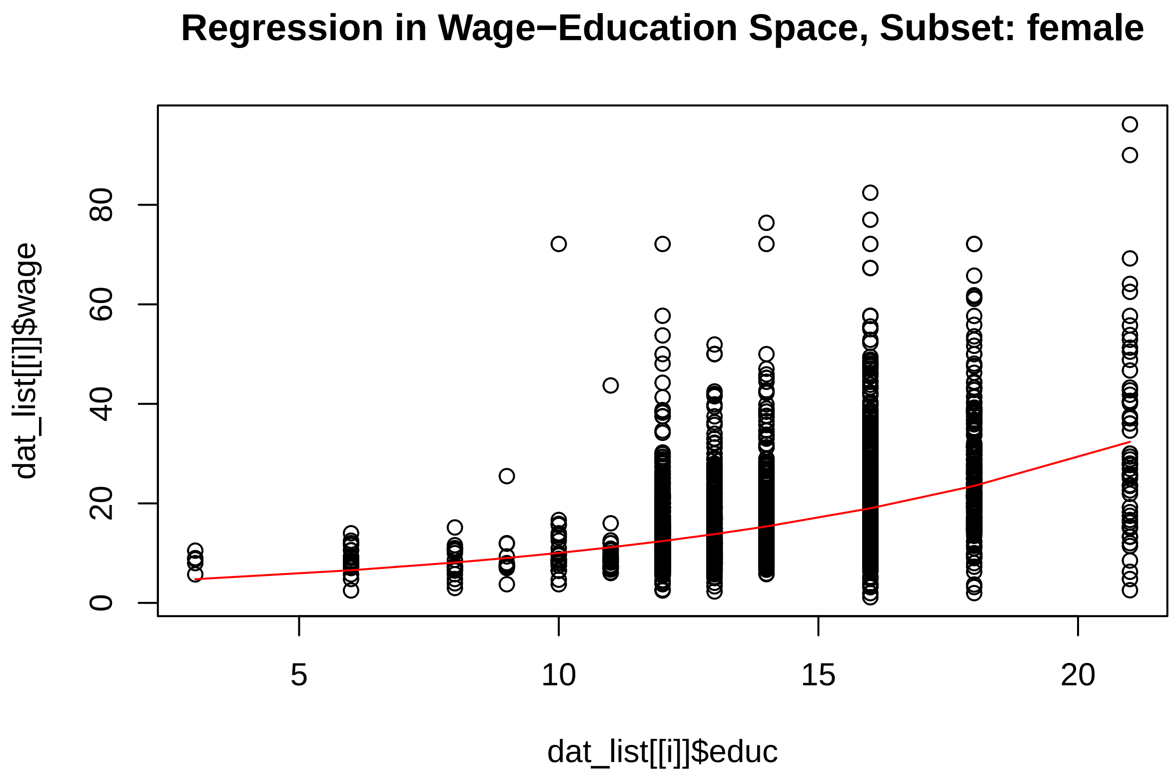
Solution No one data partition stands out as being particularly poorly fit by the model to me. Note, I like to use plots to diagnose fit quality. However, this chapter is emphasizing using the R^2 to diagnose fit quality. So, we could see the variation in R^2 of the model and compare which fit "best" in the sense of explaining variation. From R^2 estimates (above in c) we see none of the R^2 are particularly good, but black and male is much lower than the rest whereas black and female is the highest. The low R^2 in general is evident in the wide spread of points around estimated average outcome.

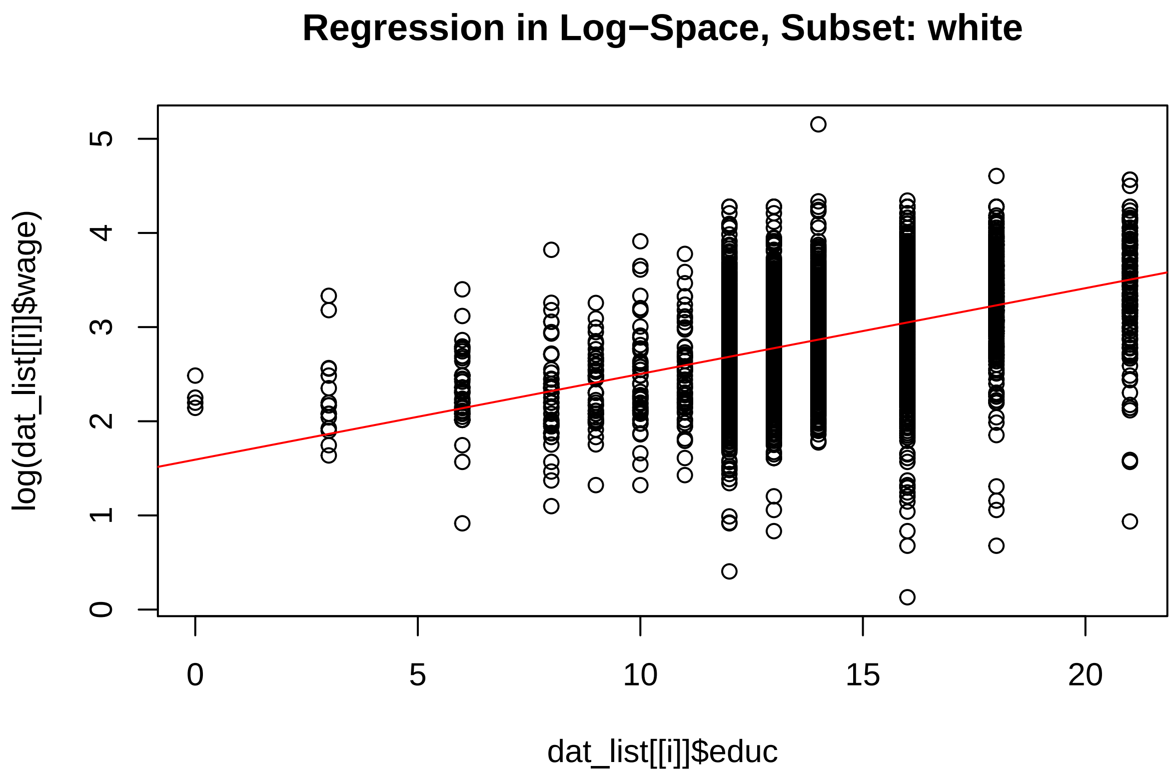
```r
#Looking at fits
for (i in 1:m){

  title1 <- paste("Regression in Log-Space, Subset: ", my_names[i],
                  sep = "")
  #Log transformed space
  plot( dat_list[[i]]$educ, log(dat_list[[i]]$wage ), main = title1)
  abline(reg_out_vec[[i]],col='red')

  #Regular space
  my_dat_frame <- cbind(dat_list[[i]]$educ,
        exp(reg_out_vec[[i]]$fitted.values))

  colnames(my_dat_frame) <- c("educ","log_fit")
  my_dat_frame <- data.frame(my_dat_frame)
  my_dat_frame_sorted <- my_dat_frame[order(my_dat_frame$educ),]
        #orders the data in increasing education,
        #so can connect the lines

  title2 <- paste("Regression in Wage-Education Space, Subset: ",
                  my_names[i],
                  sep = "")
  plot( dat_list[[i]]$educ, dat_list[[i]]$wage , main = title2)
  lines( my_dat_frame_sorted$educ, my_dat_frame_sorted$log_fit,
         col = 'red')


}
```


Regression in Log−Space, Subset: male
log(dat_list[[i]]$wage)
dat_list[[i]]$educ
0
1
2
3
4
5
0
5
10
15
20

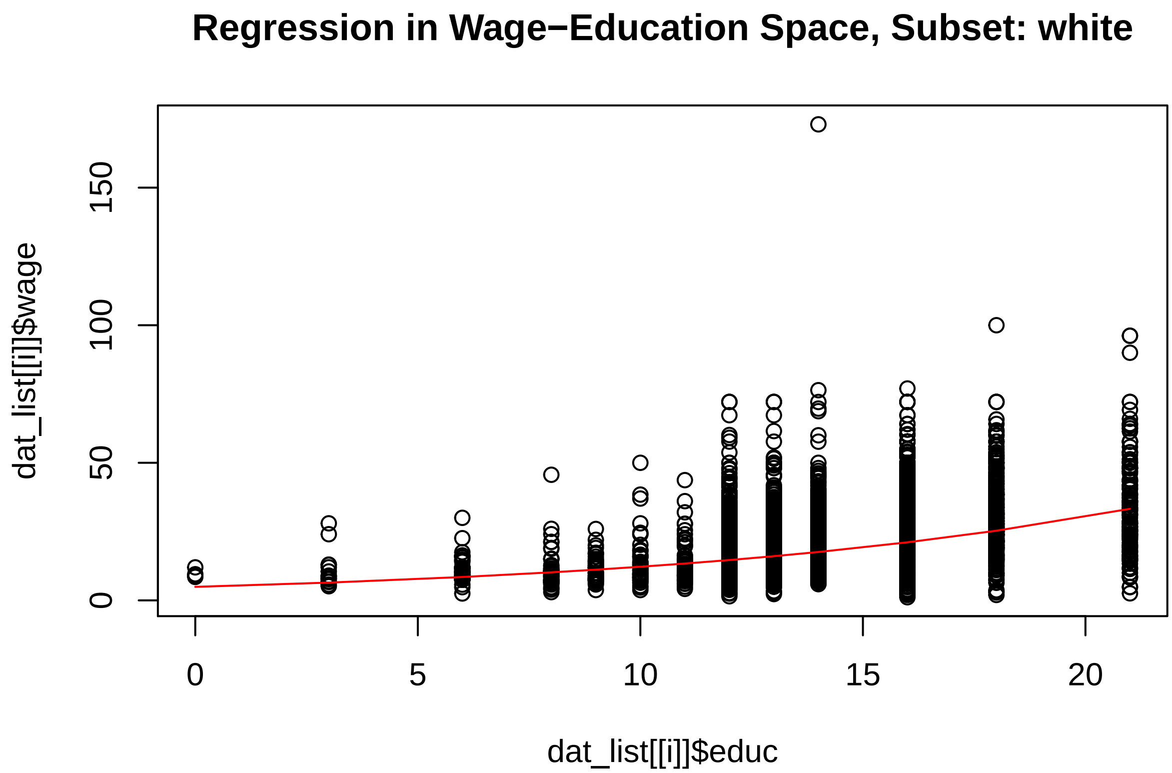

Regression in Wage−Education Space, Subset: male
dat_list[[i]]$wage
0
50
100
150
0
5
10
15
20
dat_list[[i]]$educ

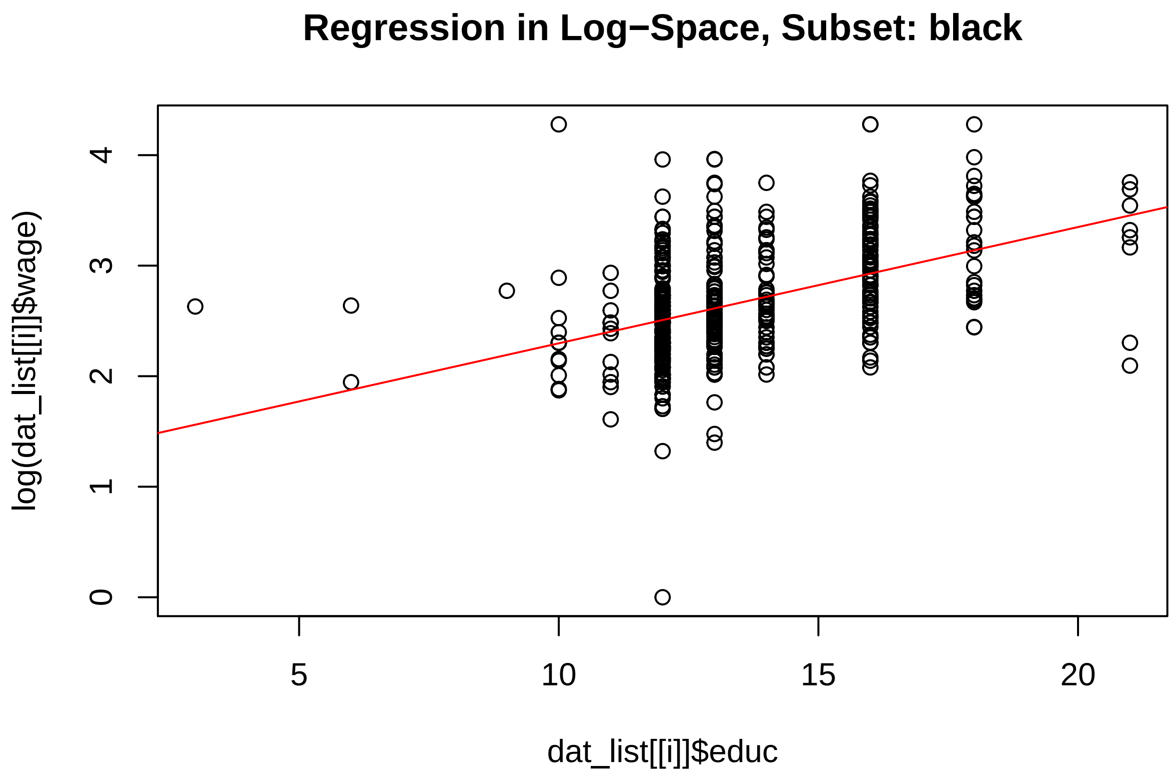

Regression in Log−Space, Subset: female
log(dat_list[[i]]$wage)
dat_list[[i]]$educ
0
1
2
3
4
5
10
15
20


Regression in Wage−Education Space, Subset: female
dat_list[[i]]$wage
0
20
40
60
80
5
10
15
20
dat_list[[i]]$educ

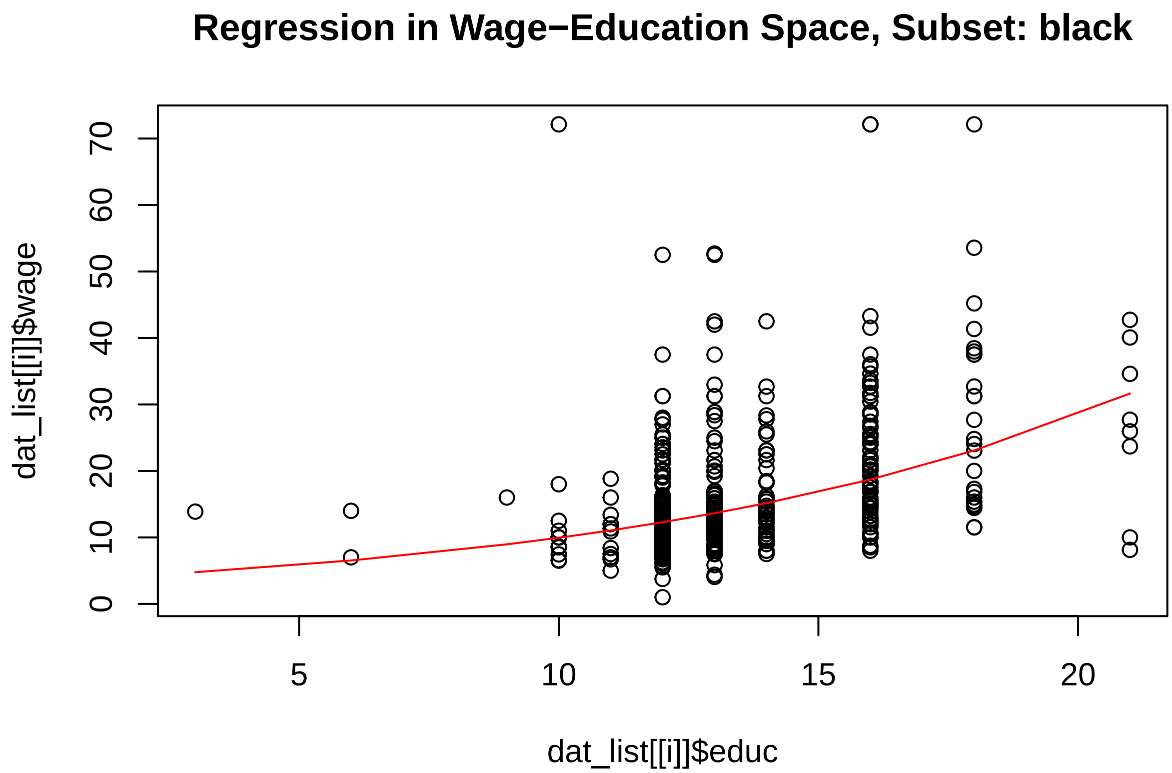


Regression in Log−Space, Subset: white
log(dat_list[[i]]$wage)
dat_list[[i]]$educ
0
1
2
3
4
5
0
5
10
15
20

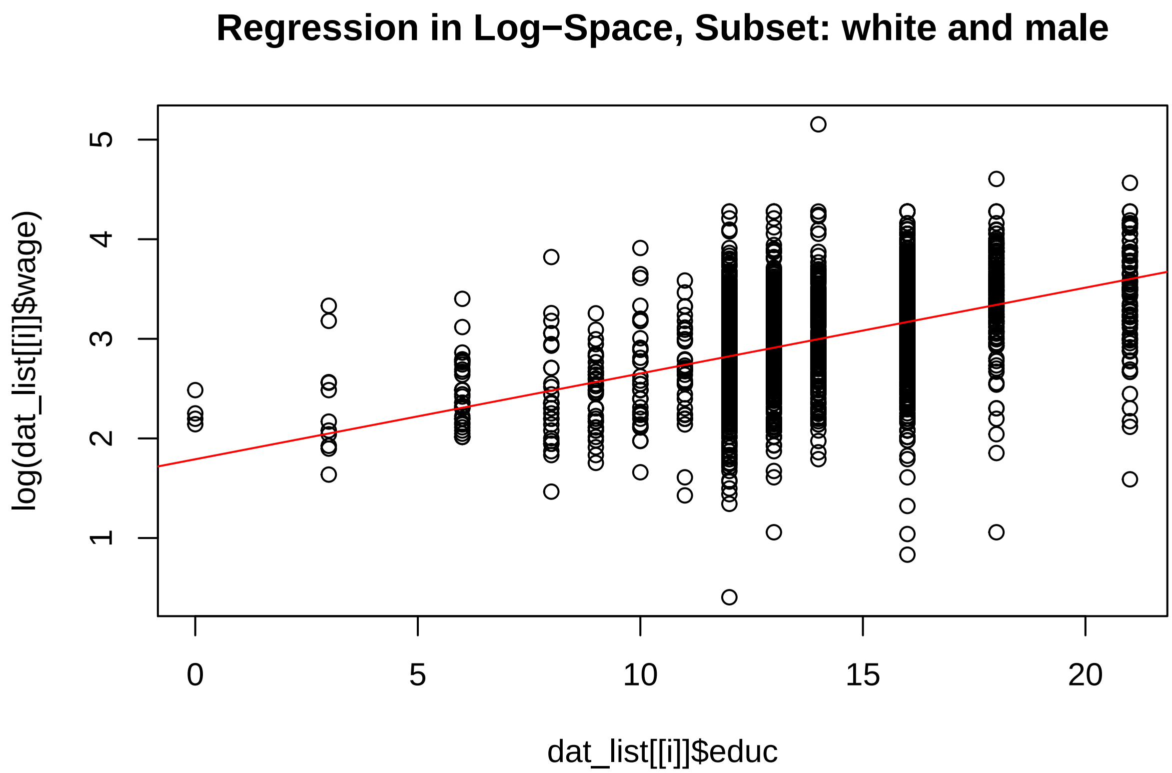

Regression in Wage−Education Space, Subset: white
dat_list[[i]]$wage
dat_list[[i]]$educ
0
50
100
150
0
5
10
15
20

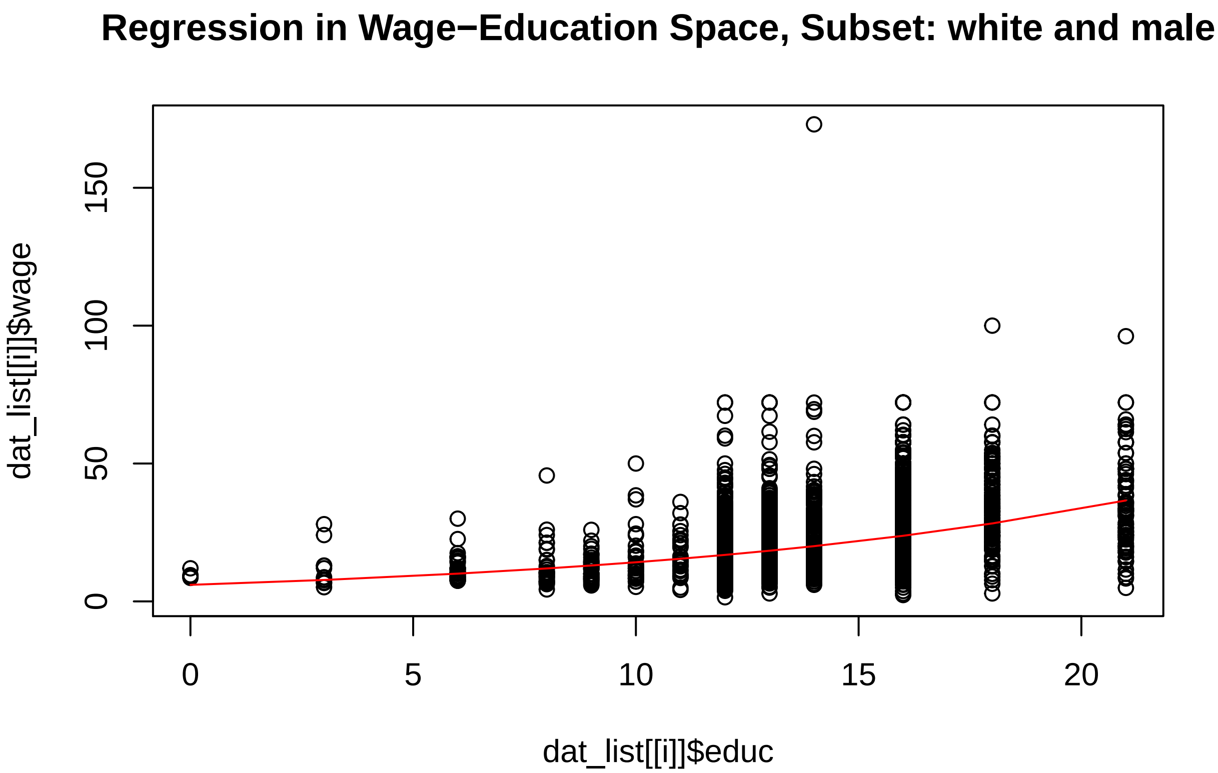


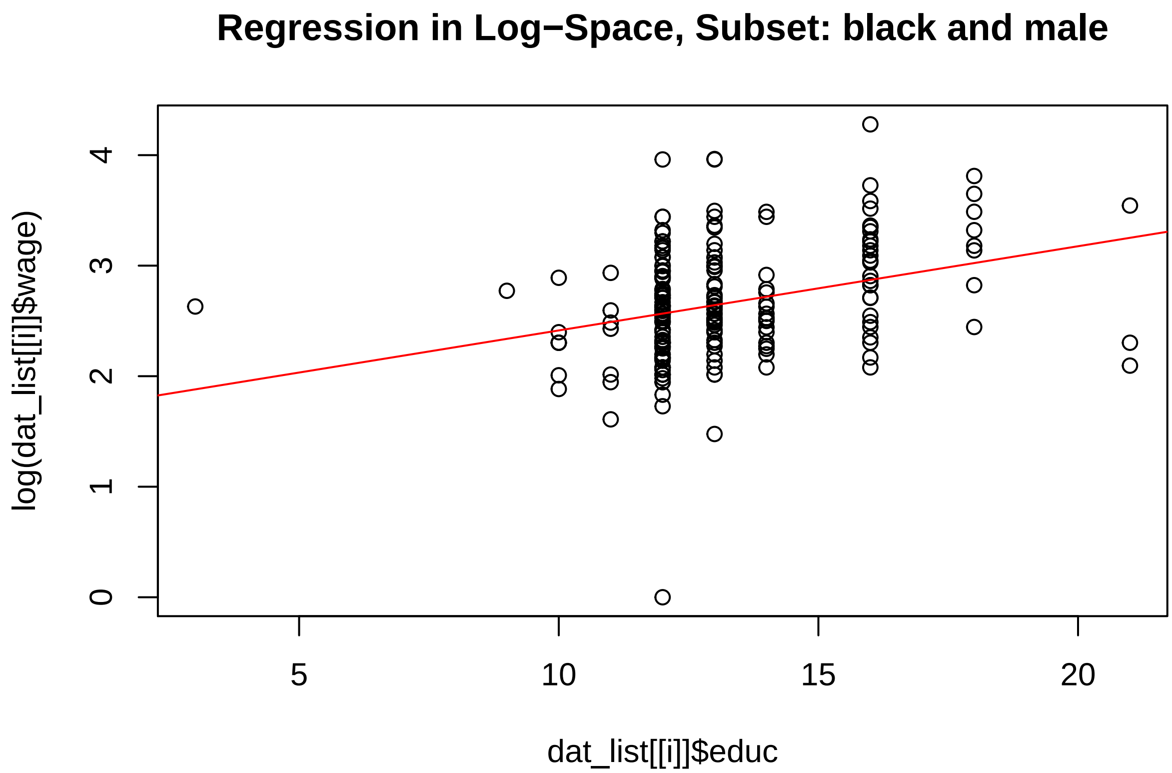
Regression in Log−Space, Subset: black
log(dat_list[[i]]$wage)
0
1
2
3
4
5
10
15
20
dat_list[[i]]$educ

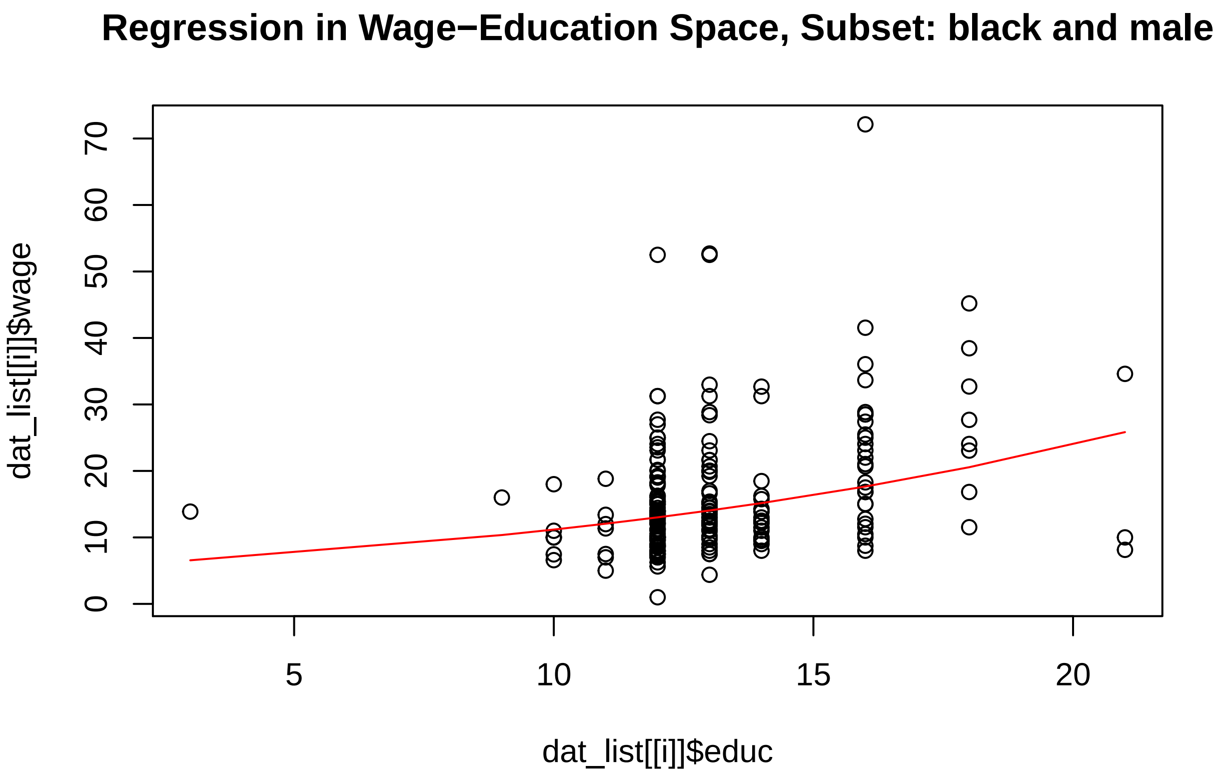


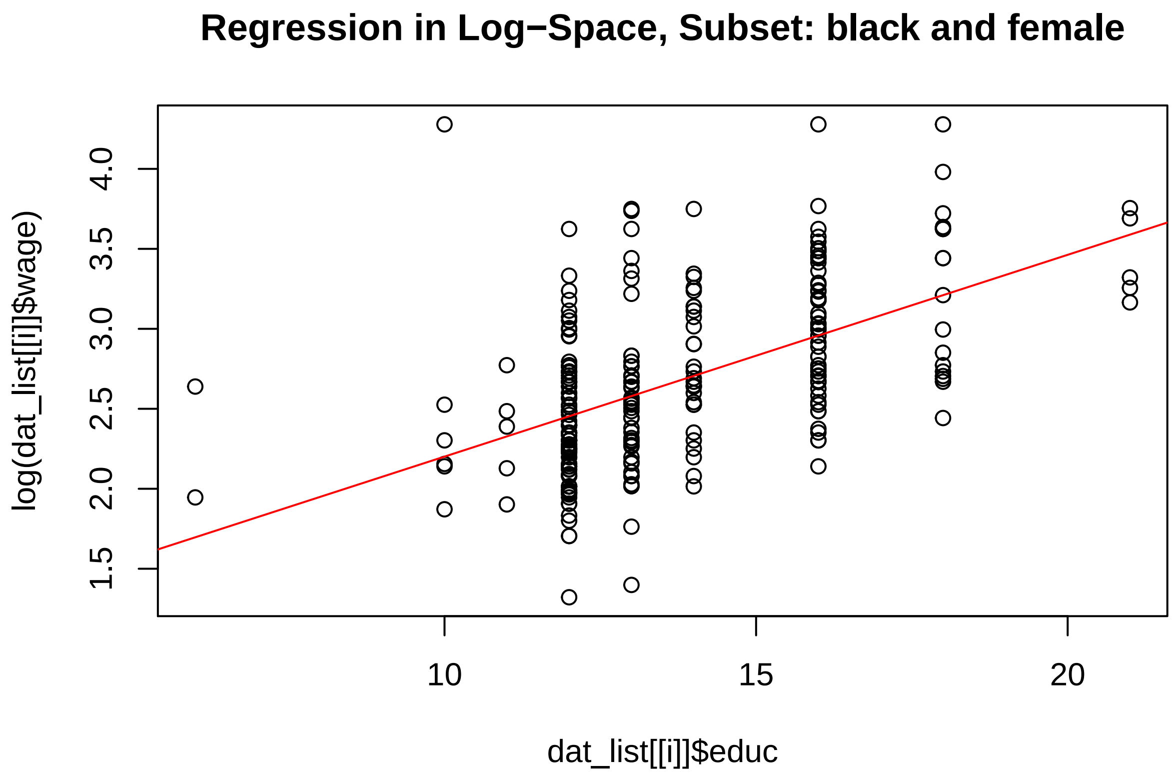
Regression in Wage−Education Space, Subset: black
dat_list[[i]]$wage
0
10
20
30
40
50
60
70
5
10
15
20
dat_list[[i]]$educ


Regression in Log−Space, Subset: white and male
log(dat_list[[i]]$wage)
dat_list[[i]]$educ
0
5
10
15
20
1
2
3
4
5

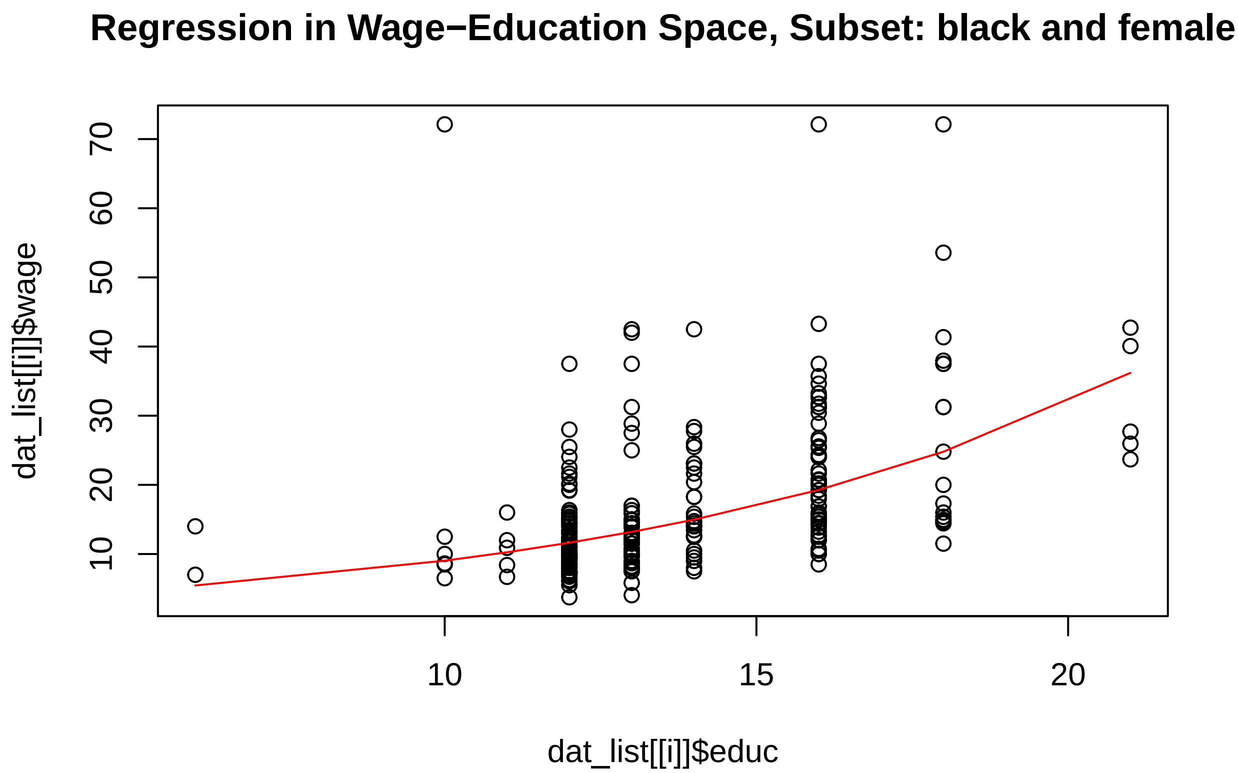

Regression in Wage−Education Space, Subset: white and male
dat_list[[i]]$wage
0
50
100
150
0
5
10
15
20
dat_list[[i]]$educ


Regression in Log−Space, Subset: black and male
log(dat_list[[i]]$wage)
0
1
2
3
4
5
10
15
20
dat_list[[i]]$educ


Regression in Wage−Education Space, Subset: black and male
dat_list[[i]]$wage
0
10
20
30
40
50
60
70
5
10
15
20
dat_list[[i]]$educ



Regression in Log−Space, Subset: black and female
log(dat_list[[i]]$wage)
1.5
2.0
2.5
3.0
3.5
4.0
10
15
20
dat_list[[i]]$educ



e

For each sample partition, test the null hypothesis that the rate of return to education is 10% against the alternative that it is not, using a two-tail test at the 5% level of significance.

Solution For each subset of people, our test is $H_0 : 100 \times \beta_2 = 10\%$ $H_1 : 100 \times \beta_2 \neq 10\%$

Note that our test statistic is therefore $t_{stat} = \frac{\hat{\beta}_2 \times 100 - 10}{se(\hat{\beta}_2 \times 100)} = \frac{\hat{\beta}_2 - .10}{se(\hat{\beta}_2)}$

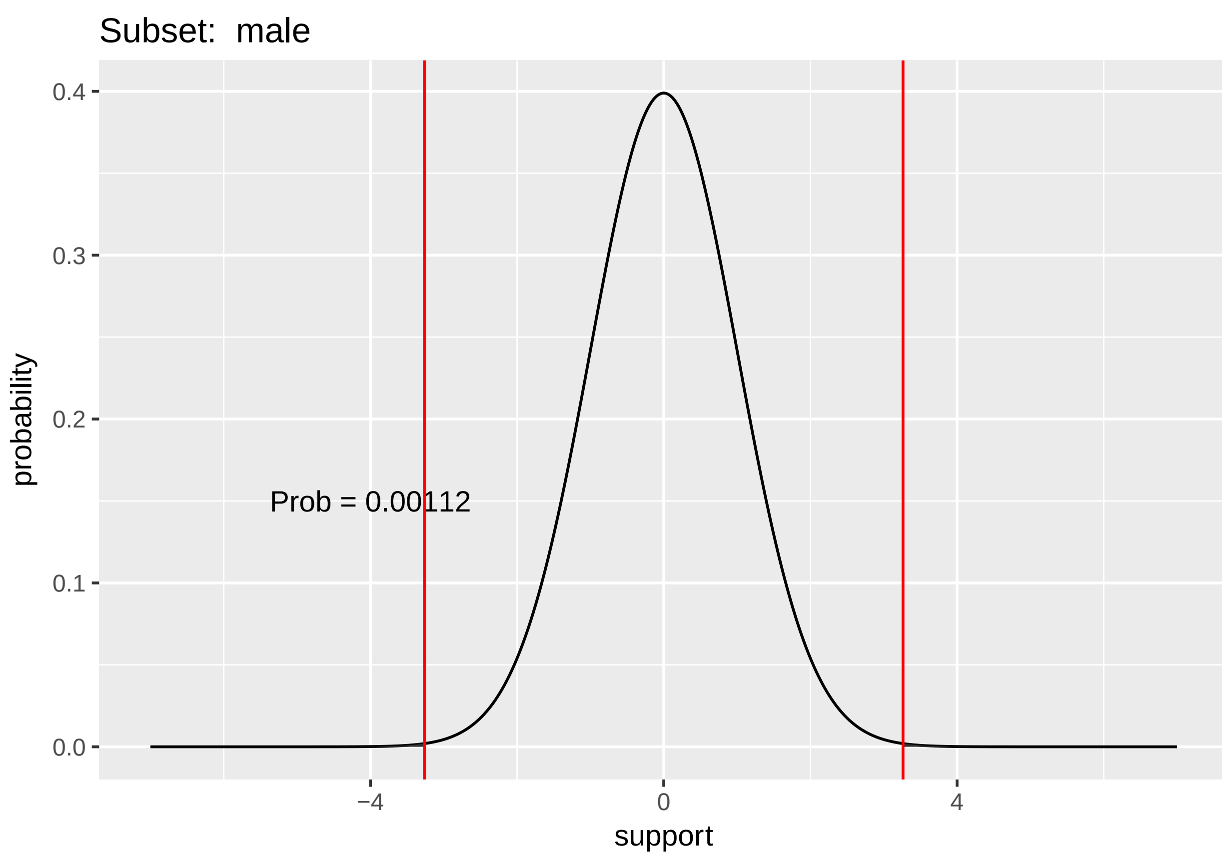
```
#Note the critical value depends on the number of data points
#(through degree of freedom).
t_c_vec <- rep(0,m)
test_stat.vec  = rep(0,m)
p_val_vec <- rep(0,m)
for (i in 1:m){
  s <- summary(reg_out_vec[[i]])
  test_stat = (coef(s)[2,1] -   .10)/( coef(s)[2,2])
  test_stat.vec[i] = test_stat
  t_c <- pt(.975,reg_out_vec[[i]]$df.residual)
  t_c_vec[i] <- t_c
    #note reg_out_vec[[i]]$df.residual = nrow(my_dat[[i]]) - 2
  print( paste("Subset: ", my_names[i]))
  print( paste("Test Statistic:", test_stat))
  print( paste("Critical Value:", t_c))
  print( paste("Reject the Null?", abs(test_stat)>t_c ))
  p_val <- 2*(1 - pt(abs(test_stat), df =
```

```
                           reg_out_vec[[i]]$df.residual))
  p_val_vec[i] = p_val
  print( paste("P-value: "))
}
```

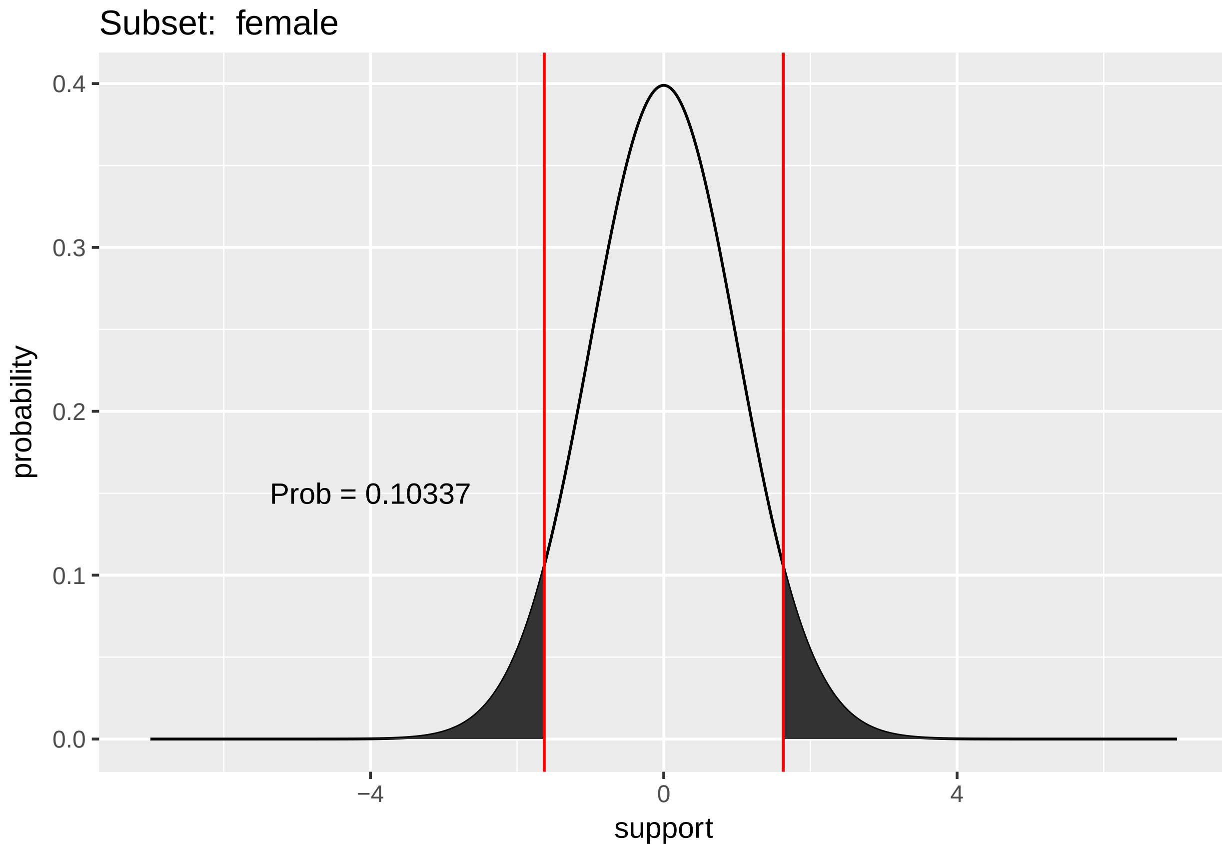
```
## [1] "Subset:  male"
## [1] "Test Statistic: -3.26261040596478"
## [1] "Critical Value: 0.835170596272869"
## [1] "Reject the Null? TRUE"
## [1] "P-value: "
## [1] "Subset:  female"
## [1] "Test Statistic: 1.62934975530218"
## [1] "Critical Value: 0.835171565097832"
## [1] "Reject the Null? TRUE"
## [1] "P-value: "
## [1] "Subset:  white"
## [1] "Test Statistic: -3.07513372364614"
## [1] "Critical Value: 0.835191207612912"
## [1] "Reject the Null? TRUE"
## [1] "P-value: "
## [1] "Subset:  black"
## [1] "Test Statistic: 0.551351183776422"
## [1] "Critical Value: 0.834979818787136"
## [1] "Reject the Null? FALSE"
## [1] "P-value: "
## [1] "Subset:  white and male"
## [1] "Test Statistic: -3.71973654091978"
## [1] "Critical Value: 0.835162715323056"
## [1] "Reject the Null? TRUE"
## [1] "P-value: "
## [1] "Subset:  black and male"
## [1] "Test Statistic: -1.5039380307703"
## [1] "Critical Value: 0.834664269064866"
## [1] "Reject the Null? TRUE"
## [1] "P-value: "
## [1] "Subset:  black and female"
## [1] "Test Statistic: 2.27252293780425"
## [1] "Critical Value: 0.83479452878925"
## [1] "Reject the Null? TRUE"
## [1] "P-value: "
```

```
#Two-sided p-value plot example
#Then use this new data.frame with geom_polygon
my_plots <- vector("list", m)

for( i in 1:m){
p_val <- p_val_vec[i]
test_stat <- test_stat.vec[i]
support = -700:700/100
plot_data <- as.data.frame(cbind(support,
                   probability = dt(support, nrow(dat)-2)))
#note shade order changed
shade <- as.data.frame(rbind(
               subset(plot_data, support < -abs(test_stat)),
```

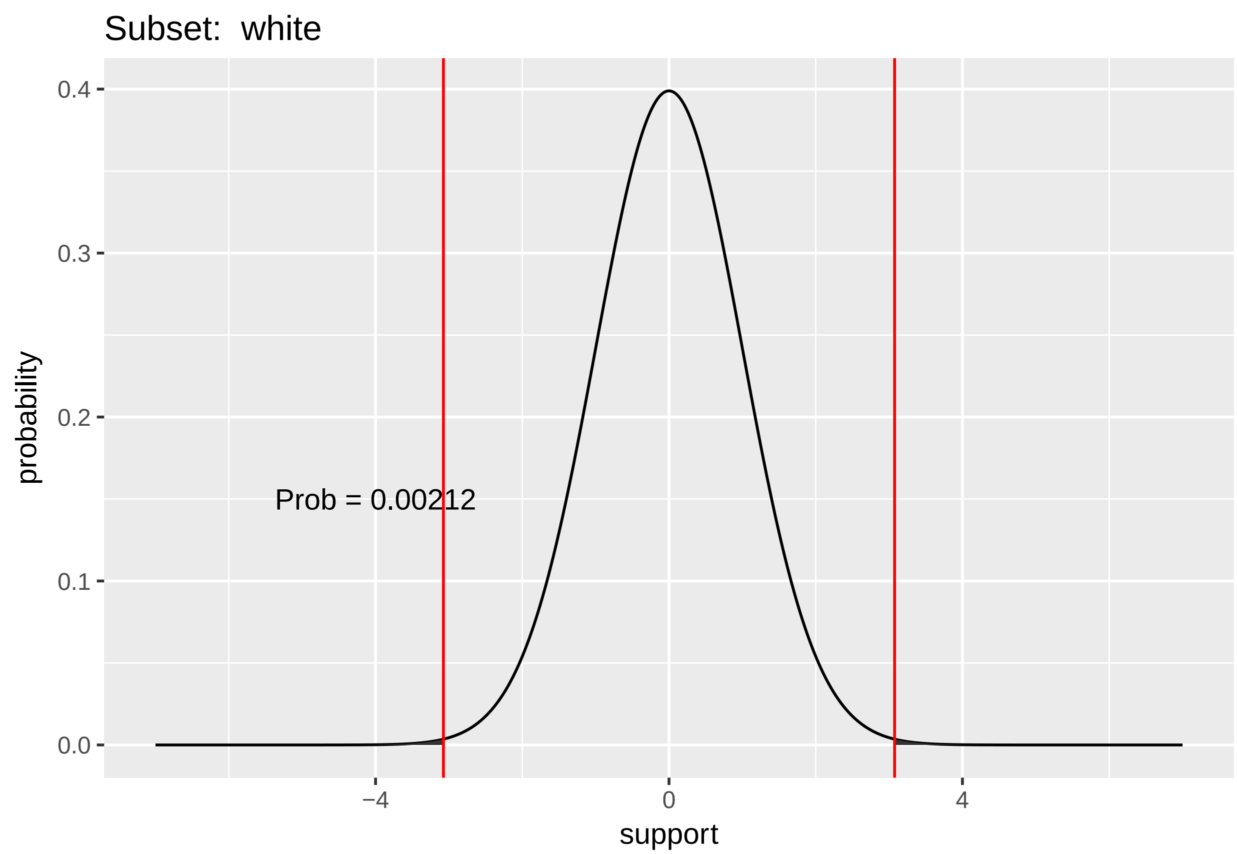
```r
                c( -abs(test_stat),0)))
                #c(0, plot_data[nrow(plot_data), "support"])))
names(shade) <- c("x","y")
shade2 <- as.data.frame(rbind(
               c(abs(test_stat),0),
                subset(plot_data, subset = support > abs(test_stat)),
                c(plot_data[nrow(plot_data), "support"], 0)))
names(shade2) <- c("x2","y2")


library(ggplot2)
my_plots[[i]] <- ggplot(data = plot_data, aes(x = support, y= probability))  +
  geom_line() +
  annotate("text", x =-4, y = .15, label =
             paste("Prob = ", round(p_val,5), sep = "")) +
  geom_polygon(data = shade, aes(x,y )) +
  geom_polygon(data = shade2, aes(x2,y2 )) +
  ggtitle("P-value of Test Statistic")  +
  geom_vline(xintercept = abs(test_stat), col = 'red'  ) +
 geom_vline(xintercept = - abs(test_stat), col = 'red'  ) +
  ggtitle( paste("Subset: ", my_names[i]))
}

my_plots[[1]]
```



```
my_plots[[2]]
```

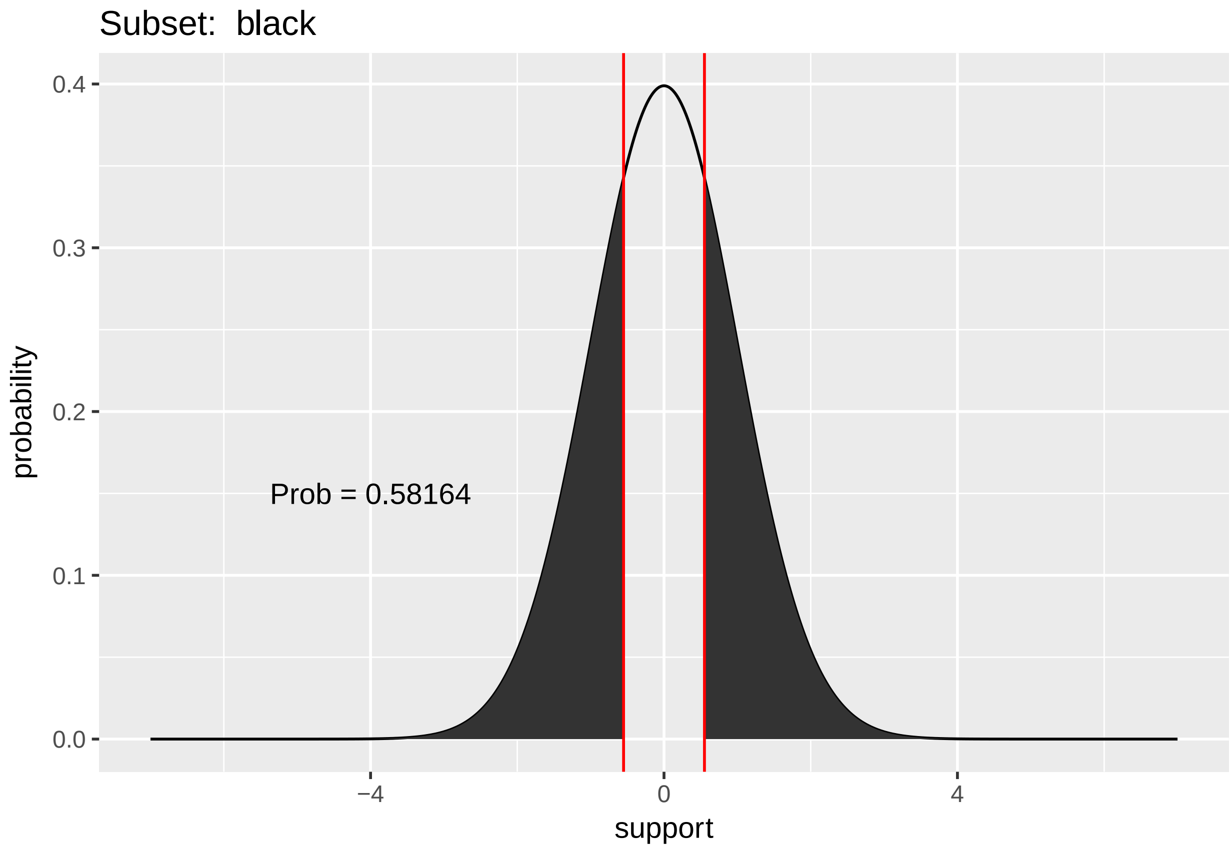



```
my_plots[[3]]
```



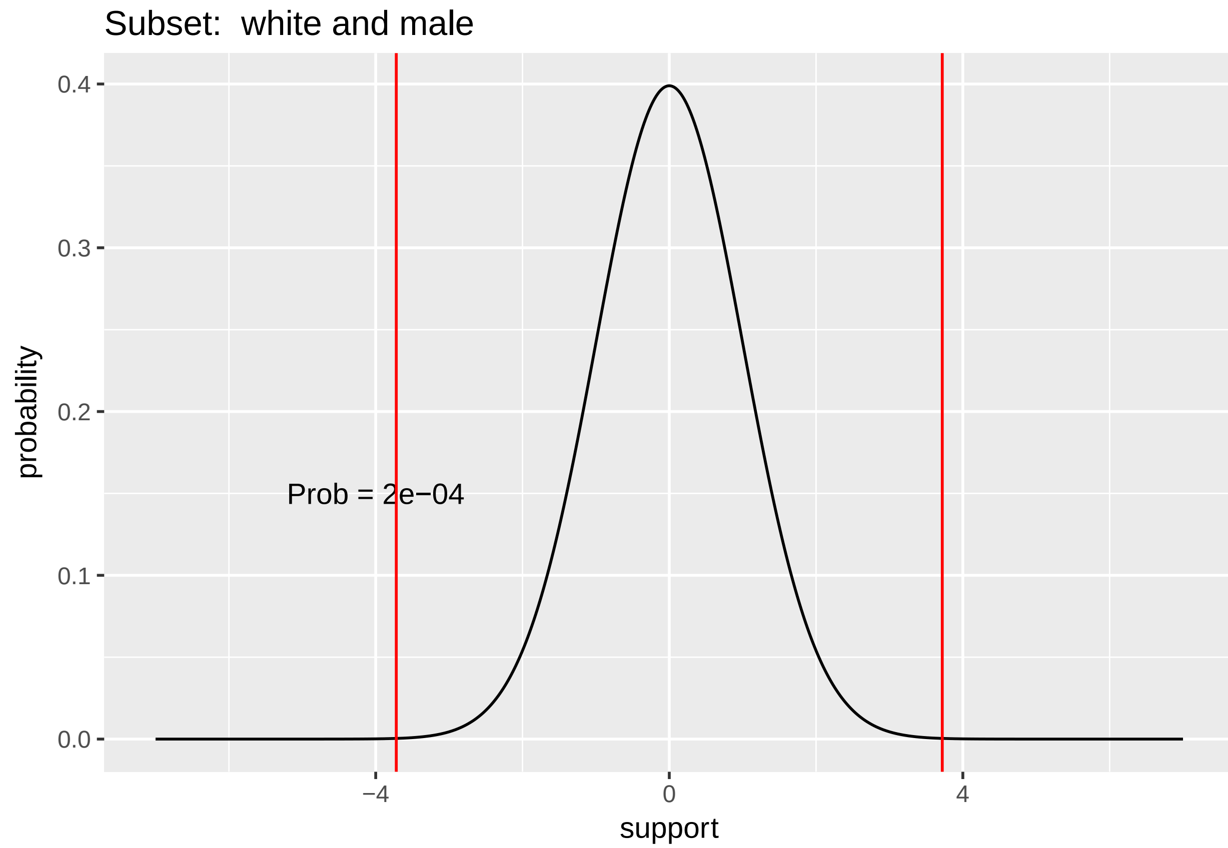
Subset: white
0.4
0.3
0.2
0.1
0.0
probability
Prob = 0.00212
−4
0
4
support


```
my_plots[[4]]
```

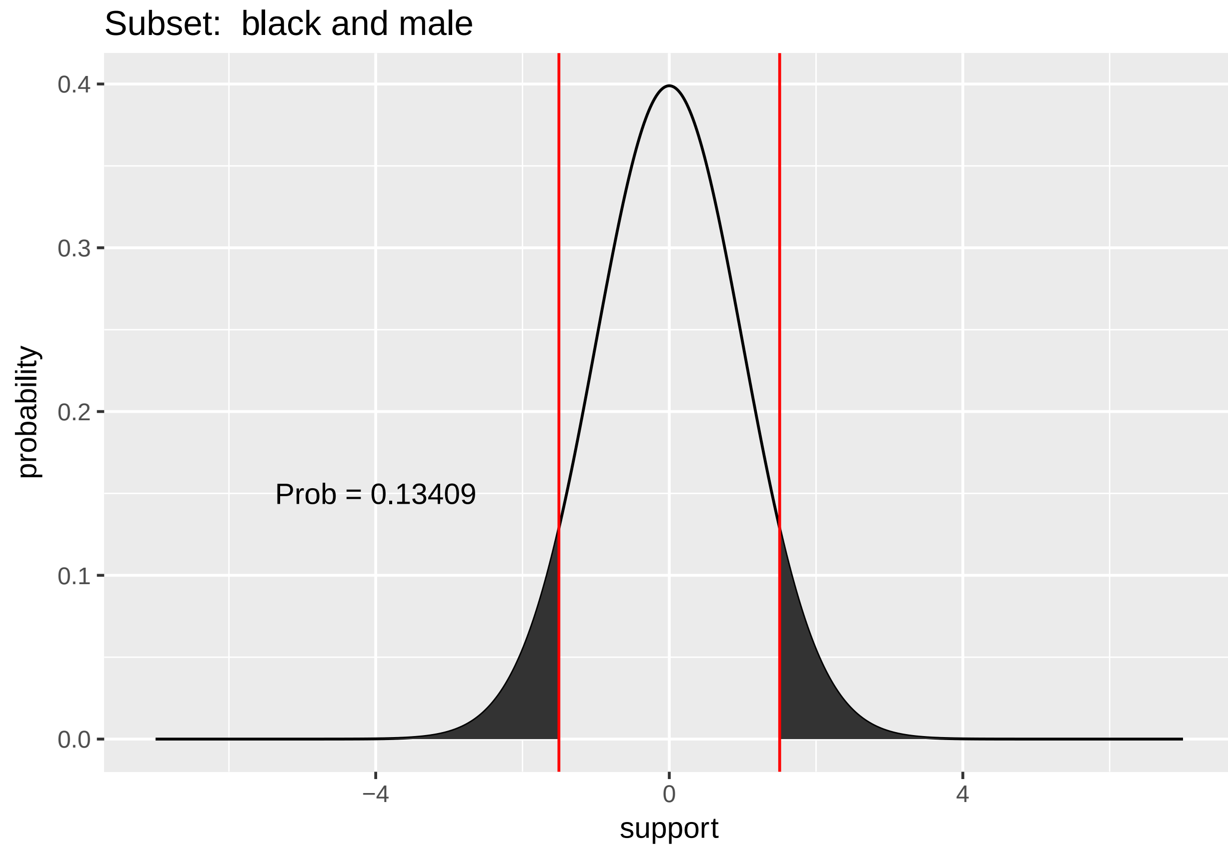


Subset: black
0.4
0.3
0.2
0.1
0.0
probability
Prob = 0.58164
−4
0
4
support

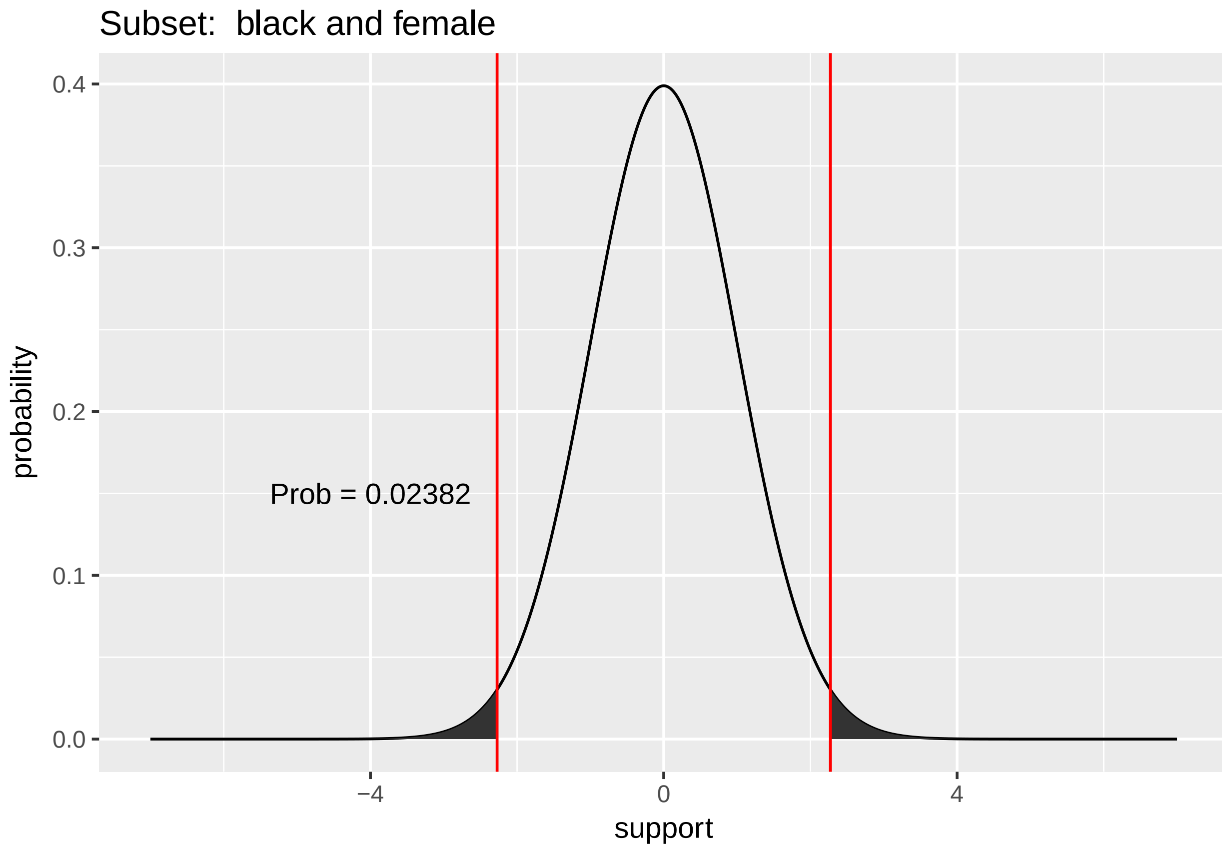

```
my_plots[[5]]
```


Subset: white and male
probability
0.4
0.3
0.2
0.1
0.0
Prob = 2e−04
−4
0
4
support

```
my_plots[[6]]
```


Subset: black and male
probability
0.4
0.3
0.2
0.1
0.0
Prob = 0.13409
−4
0
4
support

```
my_plots[[7]]
```


Subset: black and female
Prob = 0.02382
probability
0.4
0.3
0.2
0.1
0.0
−4
0
4
support