Data Analysis in R

Andrew Cobble

January 29, 2024

Data Download and Documentation

Download the data `divusa` from the `faraway` package. Online documentation is available for determining what variable names correspond to.

```
library(faraway)
data(divusa)
```

The relationship between female labor force participation rate and divorce rate has long been a subject of study for social scientists, who have posited that women moving into the labor market grants them the economic freedom and stability to leave troublesome marriages, which would help explain the increase in divorce rate. I will be analyzing and assessing the relationship between female labor force participation rate and divorce rate over time in the United States using the longitudinal dataset from the faraway package in R.

Scatter Plot and model assumptions

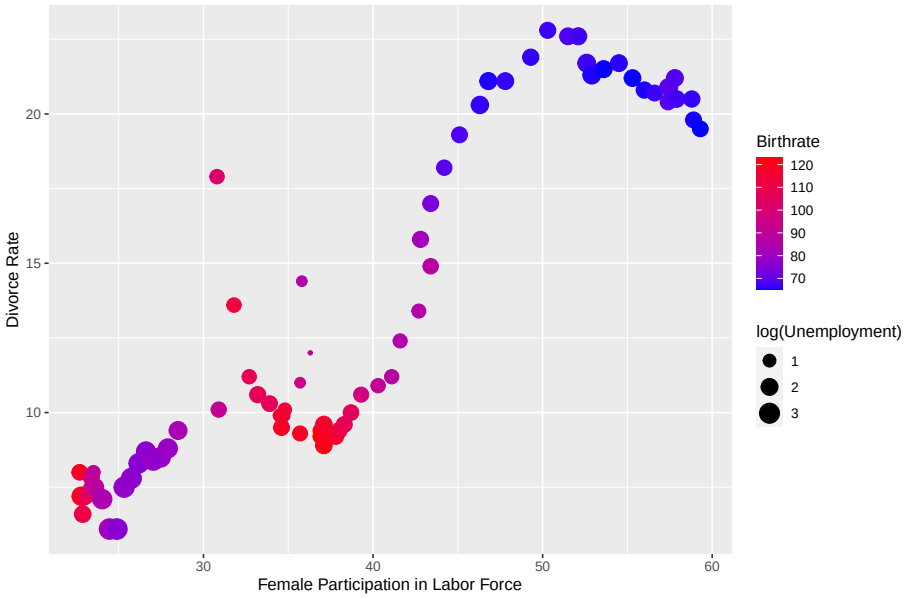
```
summary(divusa)
str(divusa)
summary(divusa$year)
summary(divusa$divorce)
summary(divusa$unemployed)
summary(divusa$femlab)
summary(divusa$marriage)
summary(divusa$birth)
summary(divusa$military)
```

```
                year   divorce unemployed   femlab  marriage     birth military
mean    1958.00000 13.268831   7.172727 38.58052  72.97273  88.88831 12.36477
sd        22.37186  5.669082   5.081070 11.76538  13.12491  19.51975 14.85396
median  1958.00000 10.600000   5.600000 37.10000  74.10000  85.90000  9.10200
min     1920.00000  6.100000   1.200000 22.70000  49.70000  65.30000  1.94010
max     1996.00000 22.800000  24.900000 59.30000 118.10000 122.90000 86.64070
```

The dataset consists of 77 observations and 7 variables. The first is "year", which is an integer variable that includes every year from 1920 to 1996. The next is "divorce", a numeric variable which is the divorce rate per 1000 women over 15. Divorce ranges from 6.10 to 22.80, with a standard deviation of 22.37, a median of 10.60, and a mean of 13.27, which indicates a rightward skew. "Unemployed" is a numeric variable measuring unemployment rate. It ranges from 1.20 to 24.90. The mean of 7.17 and median of 5.60 indicate rightward skew. It has a standard deviation of 5.08. Next is another numeric variable, "femlab", which measures the percent of female participation in labor force over age 16. Femlab has a range of 22.70 to 59.30. The median of 37.10 is fairly close to the mean of 38.58, which indicates a slight rightward skew. It has a standard deviation of 11.77. "Marriage" measures marriages per 1000 unmarried women over 16. It has a range of 49.70 to 118.10 and a standard deviation of 13.12. A mean of 72.97 and median of 74.10 reveals a leftward skew.

Next is "birth", a measure of births per 1000 women aged 15-44. This has a range of 65.30 to 122.90. With a mean of 88.89 and a median of 85.90, birthrate is skewed to the right. Lastly, "military" is a measure of military personnel per 1000 population. It ranges from 1.94 to 86.64. It has a standard deviation of 15.85. The mean of 12.36 and median of 9.10 show a right leaning skew.

```
ggplot(divusa, aes(x = femlab, y = divorce,
  size=log(unemployed), color = birth)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red") +
  xlab("Female Participation in Labor Force") +
  ylab("Divorce Rate") +
  labs(color = "Birthrate", size = "log(Unemployment)")
```



Female participation in the labor market appears to have a generally linear positive relationship with the divorce rate. Despite this, OLS appears to be a poor model for the relationship between divorce rate and female labor force participation. The relationship does not appear to be linear in the scatterplot. There is a linear positive trend up to about female participation rate reaches 50, where the relationship suddenly changes into a consistent negative relationship until the end of the plot. This means that there can be no safe assumption of parametric linearity. Judging from the scatterplot, it is unlikely that the requirements of normally distributed residuals and uniform variance would be met.

Modeling

Fitting models around the full dataset runs the risk of including some of the "noise", or unexplained variability, of the sample data. This challenge of overfitting the data can be partially overcome by training the model on a subset of the data and then testing it on the smaller remainder. This approach allows one to test the model fit on novel data points which it was not originally fit on and evaluate its predictive power on unused data obtained through the same generation process.

I split the data into testing and training data of 80% and 20%. For reproducibility I have set my randomization seed to 1.

```
set.seed(1)
index <- rbinom(nrow(divusa), 1, 0.2)
mean(index)
test <- subset(divusa, index==1)
train <- subset(divusa, index==0)
```

I now model the effect of women's participation in the labor market on the divorce rate using a restricted and unrestricted model. In the restricted model, I run a bivariate regression on women's participation in the labor market on the divorce rate. In the unrestricted model I assess women's participation in the labor market on the divorce rate and control for year, birth rate, and unemployment rate. All models are run on the training data set.

```
Rmodel <- lm(divorce ~ femlab, data = train)
Umodel <- lm(divorce ~ femlab + year + birth +
unemployed, data = train)
```

Table 1: Unrestricted Model

	Dependent variable:
	divorce
femlab	0.485^{***}
	(0.137)
year	-0.088
	(0.067)
birth	-0.111^{***}
	(0.020)
unemployed	-0.174^{***}
	(0.063)
Constant	178.583
	(125.217)
Observations	63
R^2	0.888
Adjusted R^2	0.881
Residual Std. Error	1.950 (df = 58)
F Statistic	115.349^{***} (df = 4; 58)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Restricted Model

	Dependent variable:
	divorce
femlab	0.455^{***}
	(0.028)
Constant	-4.225^{***}
	(1.109)
Observations	63
R^2	0.815
Adjusted R^2	0.812
Residual Std. Error	2.448 (df = 61)
F Statistic	268.443^{***} (df = 1; 61)
Note:	*p<0.1; **p<0.05; ***p<0.01

For both models, I will be using a 95% power level, which is the accepted standard in most fields. This means that any p-values below 0.05 will be considered statistically significant. The restricted model yields the following formula:

$$y = -4.225 + 0.455x$$

The independent variable femlab was found to statistically significant with a p-value of $2 * 10^{-}16$. This model's F statistic, which tests the model's goodness of fit against the intercept-only model, yielded a p-value of $2.2 * 10^{-}16$, which is statistically significant, leading us to reject the intercept only model in favor of this one.

The unrestricted model yields the following formula:

$$y = 178.583 + 0.485x_L - 0.088x_Y - 0.111x_B - 0.174x_U$$

In this model, labor force participation was once again found to be significant with a p-value of 0.00078. Birthrate was also significant with a p-value of $9.26 * 10^{-}7$, and unemployment rate was also significant with a p-value of 0.008. This model has an F statistic p-value of $2.2 * 10^{-}16$, once again verifying the model's usefulness compared to the intercept-only model.

In the unrestricted model, a one-unit increase in the female labor force participation rate, while all other variables are held constant, predicts an increase in the divorce rate of .485. An increase by one year, while all other variables are held constant, predicts a decrease in the divorce rate by 0.088. An increase in the birth rate by one unit, while all other variables are held constant, predicts

a decrease in the divorce rate by 0.111. A one-unit increase in the unemployment rate, while all other variables are held constant, predicts a decrease in the divorce rate by 0.174.

```
confint(Umodel)
coef(Umodel)
table <- cbind(coef(Umodel), confint(Umodel))
colnames(table) <- c("Mean", "Lower Bound", "Upper Bound")
table
stargazer(table)
```

Table 3: Unrestricted Model 95% Confidence Interval

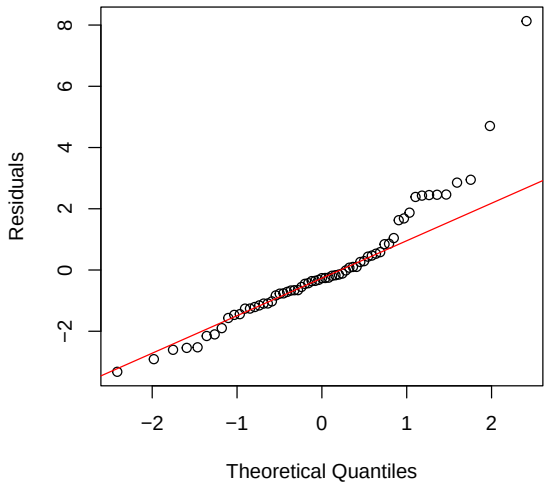
	Mean	Lower Bound	Upper Bound
(Intercept)	178.583	-72.066	429.231
femlab	0.485	0.211	0.759
year	-0.088	-0.222	0.045
birth	-0.111	-0.152	-0.071
unemployed	-0.174	-0.300	-0.047

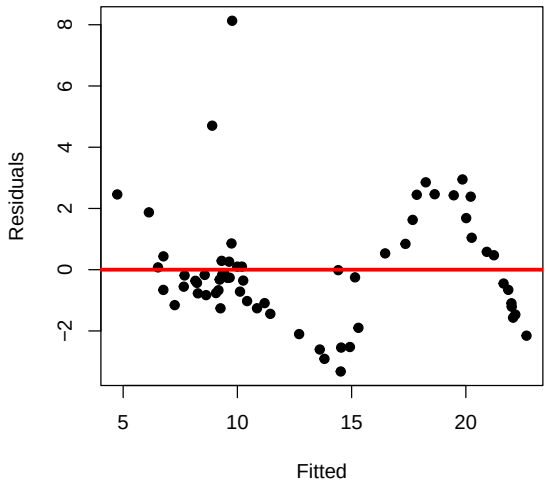
Model Fit

I assess the model fit using a qqnorm plot and residual by fitted plot.

```
qqnorm(residuals(Umodel),ylab="Residuals", main="")
qqline(residuals(Umodel), col="red")

plot(predict(Umodel), residuals(Umodel), xlab="Fitted",
ylab="Residuals", pch=19)
abline(h=0, lwd=3, col="red")
```





The model appears to be a good fit in the qqnorm until the upper range,

which create much larger residuals than the rest of the data, casting doubts on the model's fit. The scatterplot of residuals over fitted values reveals more issues with the model by presenting a clearly heteroscedastic pattern.

```
anova(Umodel, Rmodel)
Rmsum <- summary(Rmodel)
nreps <- 5000
set.seed(1)
fstats <- numeric(nreps)
for(i in 1:nreps){
  temp <- lm(sample(divorce) ~ femlab + year + birth + unemployed, data = train)
  fstats[i] <- summary(temp)$fstat[1]
}
sum(fstats > Rmsum$fstatistic[1])
```

Table 4: F Test

Statistic	N	Mean	St. Dev.	Min	Max
Res.Df	2	59.500	2.121	58	61
RSS	2	293.020	102.601	220.470	365.570
Df	1	−3.000		−3	−3
Sum of Sq	1	−145.099		−145.099	−145.099
F	1	12.724		12.724	12.724
Pr(>F)	1	0.00000		0.00000	0.00000

The F test yields a p-value of $1.699 * 10^-6$, so I reject the null hypothesis that there is no difference in model goodness of fit at the .05 significance level.

The restricted model has a much higher F-statistic (268.4) than the unrestricted model (115.3). I ran a permutation test to check the rarity of this occurrence. 5000 simulations of the sampling process produced no versions of the unrestricted model with an F statistic exceeding the restricted model statistic 268.4. The unrestricted model had only marginally higher predictive power ($R^2 = 0.8883$) compared to the restricted model ($R^2 = 0.8148$). Given this information, I conclude that the restricted model offers a better fit for the data, generating higher F statistics and offering similar predictive power using three fewer explanatory variables. Before ending, I decided to drop the largest residual value in the unrestricted model and rerun both models to see if an outlier is producing spurious results.

```
stud <- rstudent(NewUModel)
```

```
stud[which.max(abs(stud))]
cook <- cooks.distance(model)
halfnorm(cook, 3, ylab="Cook's Distance")
NewUModel <- lm(divorce ~ femlab + year + birth + unemployed, data = train, subset=(cook < m
summary(NewUModel)
summary(Umodel)
summary(Rmodel)
```

Observation 27 generates the largest residuals, which is 5.119 standard deviations from the mean. I dropped 27 from the unrestricted model and reran both models.

Table 5: Old Unrestricted Model

	Dependent variable:
	divorce
femlab	0.485^{***}
	(0.137)
year	-0.088
	(0.067)
birth	-0.111^{***}
	(0.020)
unemployed	-0.174^{***}
	(0.063)
Constant	178.583
	(125.217)
Observations	63
R^2	0.888
Adjusted R^2	0.881
Residual Std. Error	1.950 (df = 58)
F Statistic	115.349^{***} (df = 4; 58)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: New Unrestricted Model

	Dependent variable:
	divorce
femlab	0.561^{***}
	(0.115)
year	-0.119^{**}
	(0.056)
birth	-0.106^{***}
	(0.017)
unemployed	-0.137^{**}
	(0.053)
Constant	234.118^{**}
	(105.106)
Observations	62
R^2	0.923
Adjusted R^2	0.917
Residual Std. Error	1.628 (df = 57)
F Statistic	169.937^{***} (df = 4; 57)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Restricted Model

	Dependent variable:
	divorce
femlab	0.455^{***}
	(0.028)
Constant	-4.225^{***}
	(1.109)
Observations	63
R^2	0.815
Adjusted R^2	0.812
Residual Std. Error	2.448 (df = 61)
F Statistic	268.443^{***} (df = 1; 61)
Note:	*p<0.1; **p<0.05; ***p<0.01

The transformative effects of removing the highest residual observation on the unrestricted model appear to be substantial, increasing the F statistic from 115.349 to 169.937, a small increase in the R^2 value from 0.888 to 0.923, and a new significant effect for the year variable, as well as new coefficients for all parameters. This makes the unrestricted model much more competetive when comparing it to the restricted model.

I plot the effect of a one unit increase in women's labor market participation on the divorce rate using the predicted confidence intervals and simulated data.

```
newU_df <- data.frame(cbind(Upreds[,1],test$femlab))
newR_df <- data.frame(cbind(Rpreds[,1],test$femlab))
Upreds.lm <- lm(X2 ~ X1, data = newU_df)
Rpreds.lm <- lm(X2 ~ X1, data = newR_df)
summary(Upreds.lm)
summary(Rpreds.lm)
```

Table 8: Unrestricted Model Predicted Values Over Testing Data

	Dependent variable:
	X2
X1	2.120***
	(0.136)
Constant	10.382***
	(2.079)
Observations	14
R^2	0.953
Adjusted R^2	0.949
Residual Std. Error	3.282 (df = 12)
F Statistic	242.515*** (df = 1; 12)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9: Restricted Model Predicted Values Over Testing Data

	Dependent variable:
	X2
X1	2.196***
	(0.000)
Constant	9.276***
	(0.000)
Observations	14
R^2	1.000
Adjusted R^2	1.000
Residual Std. Error	0.000 (df = 12)
F Statistic	218,036,180,587,741,707,344,828,604,808,048.000*** (df = 1; 12)
Note:	*p<0.1; **p<0.05; ***p<0.01

When plotted against the unrestricted model's prediction interval data, a one-unit increase in female labor market participation rate predicts an increase of 2.120 in the divorce rate.

When plotted against the restricted model's prediction interval data, a one-unit increase in female labor market participation rate predicts an increase of 2.196 in the divorce rate.

Out-of-Sample Testing

Lastly, I run the models on the out-of-sample testing data set.

```
Rpreds <- predict(Rmodel, new = test, interval = "prediction")
Upreds <- predict(NewUModel, new = test, interval = "prediction")
Rsq.err <- (Rpreds[,1] - test$divorce)^2
mean(Rsq.err)
Usq.err <- (Upreds[,1] - test$divorce)^2
mean(Usq.err)
sort(Rsq.err)
sort(Usq.err)
```

The largest miss occurs in the restricted model, on observation 41, resulting in a squared error of 14.374. The largest miss in the unrestricted model is smaller, occurring in observation observation 77, with a squared error of 12.911.

Going forward I would prefer to use the unrestricted model for predictions. The removal of the most problematic outlier substantially improved this model's predictive capacity and revealed a new significant effect for year. I would prefer the unrestricted model, which controls for some important factors which could affect the relationship between female labor force participation and divorce rate.