BUILDING A REGRESSION MODEL

Ruth Anne Sullivan
Saint Mary's College
Department of Mathematics

September 6, 2016

Contents

Introduction

Statistics offers a way to interpret large amounts of data and turn it into useful information. The goal of this project is to show the process of building a regression model with the purpose of identifying city characteristics related to violent crime in 2010. We want to find the most efficient model with the variables collected.

The data used in the model comes from city information on violent crime from the FBI and information from the United States Census Bureau for 2010. Each city has several variables that were collected such as population, percentage of ethnicity, income, and education. [7] The FBI reports the total number of crimes, but the model's response variable is the violent crimes against persons rate per thousand people.[8] The formula used to find the crime rate is

$$\text{City Violent Crime Rate} = \frac{\text{Number of Violent Crimes Committed in 2010}}{\text{Population in 2010}} * 1000.$$

There were two hundred cities picked from all of the United States, and the collection of cities can be seen in **Table A1**.

1 Tools

When we start building our model, we need to look at different considerations involved in building our model. We will be using multiple regression and the tools associated with it, including F-statistic, degrees of freedom, sample size, R^2, t-test, and adjusted R^2. We will also look at multicollinearity, interaction, and failure of variance.

1.1 Multiple Regression

We will be dealing with multiple regression as the basis when we are building the model. There will be different regression models, but they will all follow the basic format of

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon.$$

The Y is the response variable, which is violent crime in our case, and the terms $X_1, ..., X_n$ should be independent variables. The $\beta_1, ..., \beta_n$ are fixed parameters that indicate how to interpret the response variable. The ϵ stands for the error, which is the unexplained variation in the model.

1.2 Degrees of Freedom and Sample Size

When deciding how large of a sample to use in the model, we need to take into account how many variables we might consider since this helps decide on an appropriate sample size. The more variables we use, the larger we want the sample size in order to have more accurate results. The higher the degrees of freedom, the better the prediction of the model. Degrees of Freedom measures the number of independent observations in the sample minus the number of population parameters that must be estimated with the sample data. The formula for degrees of freedom error is,

$$n - k - 1 > 0,$$

where the n stands for the sample size and the k stands for the number of variables. The sample size needs to be at least five to ten times the number of variables in order to be large enough to produce significant results. The higher the degrees of freedom, the larger our sample size is compared to number of predictors, making our model more accurate. [1, p.318]

1.3 R^2

When we want to measure the strength of our model, we can look at the R^2 value. One way of thinking of the formula is that it is the explained variation in the model divided by the total variation. The formula for R^2 is

$$R^2 = \frac{\text{SSR}}{\text{SST}}. \tag{1}$$

The Sum of Squares Regression, SSR is the sum of squared differences between prediction from our model for each observation and population mean. The SST stands for the total sum of squares, which is the sum of the squared differences of each observation from the overall mean. Thus, R^2 is the percentage of the response variable variation that is explained by a linear model.

A high R^2 means that the linear model explains most of the variability in the data. The value R^2 is a fraction between zero and one. When R^2 equals one, all points lie exactly on a straight line with no scatter. This means that we want the R^2 value to be as close to one as we can get. [9]

1.4 F-statistic

While R^2 measures how well the estimate explains the dependent variable in the sample, the F-statistic looks at the population. One of the ways it is able to do this is by taking into account the degrees of freedom. The *F-statistic* is measured using the F-distribution that looks at the Mean Square Regression compared to the Sum of Squares Regression where the higher the value on the F-distribution the more significant the value. It is the partition of the unexplained variance by the standard error. The equation is

$$F = \frac{\text{MSR}}{\text{MSE}}. \tag{2}$$

The MSR in the numerator stands for Mean Square Regression, which is the SSR divided by degrees of freedom; while the MSE, Mean Square Error, is an estimator that measures the average of the squares of errors.

The F-statistic checks to see whether the variables together are useful, but does not check individual variables. The higher the F-statistic, the more reliable the model. As the F-statistic increases, the more significant the value on the F-distribution. [4]

1.5 Partial F-Statistic

When we want to see what happens if variables are added or removed from the model in order to determine the best variables for the model, we will use the *partial F-statistic*. The partial F-statistic compares the expanded model against the reduced model to see if the variables in the expanded model are significant enough to keep. The formula given is

$$F = \frac{\frac{SSE(reduced) - SSE(expanded)}{df(reduced) - df(expanded)}}{\frac{SSE(expanded)}{df(expanded)}}. \tag{3}$$

The SSE is the is the sum of the squares of residuals, which is the deviation of the predictions from the values of the observed data. The null hypothesis is that the expanded model does not significantly improve the reduced model. The alternative hypothesis is that the new variables are significant enough to be added to create a new model. We will use this test to help eliminate terms and find the best model. [11]

1.6 T-test

When we are looking at the significance of individual variables in our model, we will evaluate the *T-test*. The formula for the T-test is

$$T = \frac{\widehat{\beta^*}}{S_{\widehat{\beta^*}}} \tag{4}$$

and we test

$$\beta^* = 0 \text{ vs } \beta^* \neq 0. \tag{5}$$

The $\widehat{\beta^*}$ stands for the corresponding estimated coefficient and $S_{\widehat{\beta^*}}$ is the estimate of the standard error of the $\widehat{\beta^*}$. It tests whether the difference between a variable and zero is unlikely to have occurred because of random chance in sample selection. The difference is more significant if the difference from zero is large enough. We have a large enough sample size, and responses are relatively consistent with each other. The greater the magnitude of T, either positive or negative, the greater the evidence against the null hypothesis that there is no significant difference. If T is closer to zero, the chance that there is not a significant difference is greater. [3]

1.7 Adjusted R^2

While R^2 helps test how well the terms in the model fit, adding more variables automatically makes it closer to one even if the variance has no relation to the sample. We can use adjusted R^2 to help resolve this issue. *Adjusted* R^2 is an unbiased estimate of the fraction of variance explained, taking into account the sample size and number of variables. The formula for adjusted R^2 is

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}. \tag{6}$$

The n stands for sample size and p stands for the number of predictors. Usually adjusted R^2 is only slightly smaller than R^2, but it is possible for adjusted R^2 to be zero or negative if a model with insufficiently informative variables is fitted to too small a sample of data. [5]

2 Developing a Model and Potential Problems

After considering some of the analysis that goes into multiple regression, we will look at the procedure for developing a model and the variables in it. We will first look at the issues of multicollinearity, interaction, and the failure of the variance condition. Then we will go through the forward elimination procedure to help prune the model to the most important variables.

2.1 Multicollinearity

One of the key issues is *multicollinearity* because it means that the predictor variables in the model are not all independent. The estimated regression coefficient of any one variable depends on which of the other predictors are included in the model. The model is unable to separate out the effects of the correlated variables. The precision of the estimated regression coefficients decreases as more correlated variables are added to the model. The dependent variables hide the effect that they have on the model. The best way to deal with multicollinearity is to remove all except one of the correlated variables since that variable will now become independent. This should improve the significance of the selected variable. [6]

For an example, consider the following model. We will let our response variable be reading ability with our predictor variables being shoe size and age. The model would look something like:

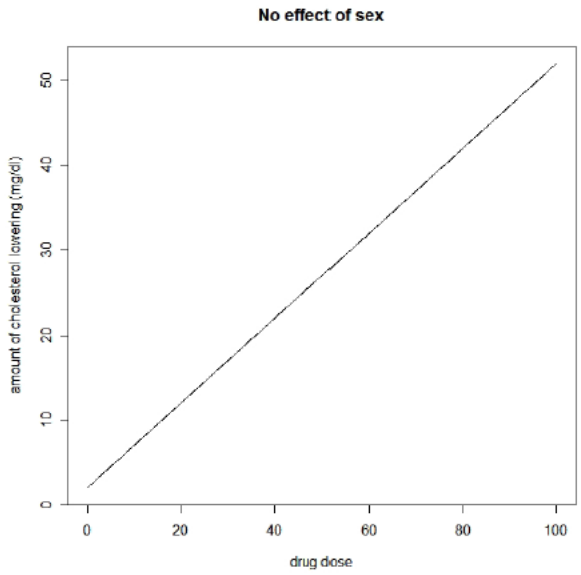
$$\text{Reading ability} \sim \text{shoe size} + \text{age}.$$

Shoe size and age would be correlated because as a child gets older their feet get larger. The older the child or the larger their feet, the more likely the child is to have better reading ability. This means that since shoe size and age are correlated, the shoe size hides

some of the effect of age, which is what we do not want. This means that we would want to pick either age or shoe size to be the only term in the regression model.

2.2 Interaction

Before we define interaction, we will look at an example of interaction to help us understand it. Let's suppose that we are conducting a clinical trial of a cholesterol lowering drug and that we are expecting a linear dose-response over a given range of the drug dose. This observation is seen in **Graph 1**.



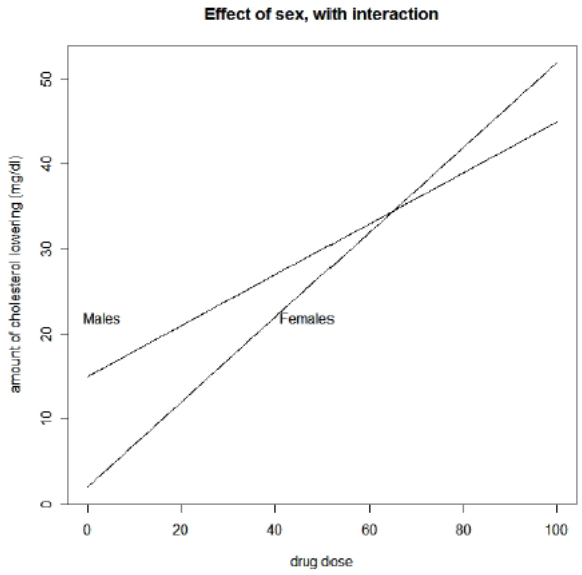
Graph 1

This figure depicts the linear response that we were expecting. We can see that the model ignores possible gender effects. The equation would be

$$Y = \beta_0 + \beta_1 * dose + \epsilon.$$

We can see that Y is the response to the cholesterol lowering drug where β_1 represents the effect of the drug.

Now let's suppose that our clinical trial does include a gender effect. Assume that men have a less steep dose-response curve compared to women. This can be seen in **Graph 2**.



Graph 2

Here, we see interaction occurring between gender and the effect that the drug has on lowering cholesterol. This is seen based on how men respond to change in dosage in **Graph 2** as compared to women's response to dosage. The standard model that we would use to describe a change in reaction to the drug based on gender would be

$$Y = \beta_0 + \beta_1 * dose + \beta_2 * gender + \epsilon.$$

In this case, we see that men's response to the dose in **Graph 2** is not constant as the standard model assumes. Since the response is not constant for men, we will model this by replacing β_1 with a function of gender so that it can vary based on gender. We will replace β_1 with

$$\beta_1 = \alpha_0 + \alpha_1 * gender.$$

Now that we have created a new β_1 to accurately describe the changes in the model based on gender, we create our new model of

$$Y = \beta_0 + (\alpha_0 + \alpha_1 * gender) * dose + \beta_2 * gender + \epsilon.$$

We still have the constant coefficients on dosage and gender, but have now created an interaction term of "$\alpha_1 * gender * dose$" with its own constant coefficient. This is an example of first-order interaction when one of the variables is binary. [10]

We will now look at the general formula for interaction, which is

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 X Z + \epsilon.$$

The coefficient β_3 is an interaction contrast that indicates by how many units the slope of Y on X is predicted to change given a one-unit change in the moderator variable, Z. An interaction term makes the graph more complicated since the terms are no longer distinct, and this requires us to test if the interaction significantly effects the model before adding it to the model.

Interaction occurs when one independent variable has a different effect on the outcome depending on the values of another independent variable. In interaction, the effect of the *focal* independent variable on the outcome variable must differ depending on the level of the *moderator*. There are other interactions such as second-order interactions, squaring of one of the variables, and other higher-order interactions that we will not be exploring. We will be looking for first-order interaction of continuous variables in our model; however, the same rules apply. [2, p.7,18]

2.3 Failure of the Variance Condition

When using multiple regression, there is an assumption that variance of error is constant across the coefficients, known as homoscedasticity. The failure of the variance condition is when the size of the error term differs across values of an independent variable. The solution to this problem is to scale variables. [12]

2.4 Maximum Model

We will start with all the variables that look interesting and refer to this as our first maximum model. It is a model that covers many different options for variables. The maximum model contains all conceivable basic predictors, high-order powers of basic predictors, other transformations of predictors, interactions among predictors, covariance, and all possible "control" variables.

With our model, we will start off with many possible variables and will eliminate variables. In a scenario where this data is being tested for a company or some other research, we would go to an expert on the issue to help choose variables. We will have different steps in creating the model. Each time we have eliminated all the variables that we plan on for that model, we have a new maximum model. We will continue to reduce the maximum model until we have found our final model. [1, p.315]

2.5 How to Choose a Model: Forward Elimination Procedure

We will use the forward elimination procedure to find the most significant terms to keep in our model. The first step is to find the variable most highly correlated with the response variable and create the associated straight-line regression model.

Step two is to calculate the partial F-Statistic associated with each of the remaining variables of the regression equation that contains the previously selected variables using an ANOVA test and to find the variables with the largest possible F-Statistic.

Step three is to choose the the largest and most significant partial F-statistic. If the selected variable is statistically significant, the variable is added to the regression equation. If not significant, the variable is not added to the regression equation.

At each step, we determine the partial F-Statistic for variables not currently in the model and add variables with largest F-Statistic that is significant. We stop adding variables to the model when there are no longer any significant F-Statistics. [1, p.325]

3 Analysis

Now we will start building our model from the information we collected to find city information relevant to the crime rate. The response variable is the violent crimes against persons rate per thousand people. We are building the model looking for the most significant coefficients from the Census Bureau that are related to violent crime. We will start with all the variables and use different elimination tactics until we have our significant model. During this process, we will look at multicollinearity, interaction, and the forward elimination procedure.

3.1 Picking the First Set of Terms

The Census Bureau had over forty variables that were included in their city analysis from 2010. Any information that was not based on 2010, was incomplete, or irrelevant was removed. This left a total of twenty-seven variables seen in the first column of the table below, which is still a high number of variables. The next step was to check for

high correlation between the response variable and the other variables. Higher correlation means a higher chance that the variable will be significant in the model. The correlation to violent crime is also seen in the table below.

Table 1: Correlation to Violent Crime

Categories	**Variable**	**Correlation**
Population	Population Change	-0.39
	Population	0.04
Age & Sex	Persons Under 5	0.17
	Persons Under 18	0.14
	Persons 65 over	-0.09
Race	White	-0.54
	African American	0.60
	Asian	-0.23
	Hispanic	0.00
	Other Race	-0.11
Population Characteristics	Veterans	0.04
	Foreign Born	-0.14
Housing	Housing Units	0.06
	Owner Occupied Housing	-0.36
	Median Rent	-0.26
	Households	0.05
Families & Living Arrangements	Persons Per Household	-0.10
	Language Other Than English	-0.06
Education	High School Graduate	-0.37
	Bachelor Degree	-0.35
Health	Disability Under 65	0.53
	No Health Insurance	0.11
Income & Poverty	Labor 16 and Over	-0.31
	Median Household Income	-0.47
	Per Capital Income	-0.34
	Poverty	0.56
Business	Population Density	0.22

A variable will be picked from each category that we are working with for our model. We only want one or two variables from each category since the variables tend to be

related. If we pick a second variable from a category, it is because of how highly they were correlated to violent crime. Under the category "Population," we are using Population Change. From the category "Age and Sex," the variable we will be using is Persons Under Five. Under the "Race" category, we will be using the variables White and African American. Foreign Born will be from the category, "Population Characteristics." Owner Occupied Housing is a variable from "Housing." From "Families and Living Arrangements," we will be using Persons Per Household. We will be using both "Education" variables, High School Graduate and Bachelors Degree. Under "Health," we will be using Disability Under Sixty-Five. The variables from "Income and Poverty" that we will be using are Median Household Income and Poverty. Population Density will be the last variable, from the category "Business." Most of these variables were picked because they had some of the highest correlation with our response variable.

3.2 First Model

After the list was down to fourteen variables, it was time to run the first regression model. This is our maximum model since it contains all the terms we are originally testing. Any model after this one will be a reduced model since we will start removing variables. The variables in the test are in **Table 2** along with their reference name.

Table 2: Model Variables

Variable	**Reference**
Population Change	Change
Persons Under 5	U5
White	W
African American	AA
Foreign Born	FB
Owner Occupied Housing	OH
Persons Per Household	PPH
High School Graduate	HG
Bachelors Degree	BD
Disability Under 65	Dis
Median Household Income	MHI
Poverty	Pvty
Population Density	Den

The model is

$$\text{VC} = 23.10 - 0.02\text{Change} + 0.26\text{U5} - 0.07\text{W} + 0.03\text{AA} - 0.02\text{FB} + 0.001\text{OH}$$

$$- 4.11\text{PPH} - 0.10\text{HG} - 0.04\text{BD} + 0.13\text{Dis} + 0.00003\text{MHI} + 0.15\text{Pvty} + 0.001\text{Den}.$$

The coefficients and general model information are shown below in **Table 3**.

Table 3: Model 1

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	2.310e+01	9.159e+00	2.522	0.01250	*
Change	-1.571e-02	7.709e-02	-0.204	0.83876	
U5	2.633e-01	2.473e-01	1.065	0.28842	
W	-6.641e-02	2.986e-02	-2.224	0.02736	*
AA	2.510e-02	2.735e-02	0.918	0.35994	
FB	-2.045e-02	4.410e-02	-0.464	0.64346	
OH	1.136e-03	3.829e-02	0.030	0.97636	
PPH	-4.114e+00	1.408e+00	-2.921	0.00392	**
HG	-1.027e-01	6.832e-02	-1.504	0.13440	
BD	-4.403e-02	4.627e-02	-0.952	0.34257	
Dis	1.348e-01	1.375e-01	0.980	0.32820	
MHI	3.141e-05	4.891e-05	0.642	0.52146	
Pvty	1.486e-01	7.844e-02	1.895	0.05967	.
Den	1.113e-04	8.771e-05	1.268	0.20622	

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.551 on 186 degrees of freedom

Multiple R**-squared:** 0.5793, **Adjusted** R**-squared:** 0.5499

F**-statistic:** 19.7 on 13 and 186 DF, p**-value:** $<$ 2.2e-16

The intercept, the percent of white people, and persons per household are the only variables that were recorded as significant according to the p-value.

We will now look at the effectiveness of the total model. The R^2 values are high since there are several variables, making it easier to find a hyperplane that fits the model well. The F-statistic was 19.7 and has a significant p-value, which registers that the model could be considered as effective. The problem is that we are looking for the most efficient model, which means that we want all the variables to be significant and there to be a minimal number of variables.

3.3 Elimination Due to Correlation Between Variables

The first maximum model shows that many of the variables did not register as significant suggesting that many of variables might be correlated. Non-significance can occur in other ways, but we will be looking at correlation to remove several variables. We want to remove variables since we are trying to get the best model with the least amount of variables and **Table 4** shows the correlation. This is the multicollinearity that was previously mentioned in Section 2.1.

The first variable that will be removed is Foreign Born. We are removing it since it is highly correlated to five other variables, it has a low correlation to our response variable, and a high p-value.

Bachelor Degree is the next variable that will be removed since it is highly correlated with High School Graduate, Disability Under 65, and Persons Under 5.

The next variable to be removed will be Median Household Income since it is highly correlated with Poverty, Disability Under 65, and four other variables.

Table 4: Correlation Between Variables

	Persons Under 5	White	African American	Persons Per Household	High School Graduate	Bachelor Degree	Disability Under 65	Poverty	Population Density	Population Change	Foreign Born	Owner Housing	Median Income
Persons Under 5	1												
White	-0.24	1											
African American	0.03	-0.73	1										
Persons Per Household	0.58	-0.24	-0.22	1									
High School Graduate	-0.57	0.4	-0.13	-0.64	1								
Bachelor Degree	-0.63	0.19	-0.19	-0.4	0.65	1							
Disability Under 65	0.25	-0.17	0.39	-0.17	-0.31	-0.65	1						
Poverty	0.07	-0.28	0.44	-0.08	-0.47	-0.33	0.56	1					
Population Density	-0.11	-0.32	0.06	0.09	-0.3	0.1	-0.03	0.14	1				
Population Change	-0.06	0.18	-0.25	0.13	0.23	0.47	-0.58	-0.38	-0.06	1			
Foreign Born	0.11	-0.24	-0.31	0.59	-0.43	0.14	-0.46	-0.18	0.53	0.25	1		
Owner Housing	0.16	0.41	-0.27	0.12	0.27	-0.14	-0.04	-0.51	-0.57	0	-0.29	1	
Median Income	-0.12	0.03	-0.4	0.23	0.33	0.55	-0.67	-0.81	0.12	0.41	0.5	0.25	1

The set of variables that are most highly correlated are White and African American. Since White registered significant in the first test model, we will remove the variable African American.

We will continue removing more correlated variables, and the next set of variables that are most highly correlated are High School Graduate and Persons Per Household. We will remove High School Graduates since it is correlated with more variables and Persons Per Household was more significant in the first model.

Persons Per Household and Persons Under 5 are also highly correlated. Persons Per Household is more significant and will be kept.

The next variable removed is Owner Occupied Housing because of its correlation with Population Density, Poverty, and White.

Disability will be the final variable removed since it is highly correlated with Poverty, but not as highly correlated with violent crime as Poverty.

3.4 Second Model

After we have removed all the variables that are correlated, we are left with Population Change, White, Persons Per Household, Poverty, and Population Density. Using these variables, the second model is

$$\text{VC} = 13.96 - .17\text{Change} - 0.095\text{W} - 2.01\text{PPH} + 0.20\text{Pvty} + .00004\text{Den}.$$

The results of this model are seen below.

Table 5: Model 2

```
              Estimate  Std. Error t-value Pr(>|t|)
(Intercept)  1.396e+01  2.237e+00   6.240 2.69e-09 ***
Change      -1.652e-01  6.179e-02  -2.674  0.00814 **
W           -9.522e-02  1.232e-02  -7.726 5.81e-13 ***
PPH         -2.013e+00  6.363e-01  -3.164  0.00181 **
Pvty         2.030e-01  3.135e-02   6.475 7.58e-10 ***
Den          4.104e-05  6.436e-05   0.638  0.52444
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Residual standard error: 2.676 on 194 degrees of freedom

Multiple R-squared: 0.5171, **Adjusted R-squared:** 0.5046

F-statistic: 41.54 on 5 and 194 DF, **p-value:** $<$ 2.2e-16

Here we see from the t-values and p-values that all of our terms are significant except Population Change. This is probably because it was not very highly correlated with our response variable, but we will leave the variable in to see how it reacts in the forward elimination process. We can see the model is more efficient, since it has less variables creating a higher F-statistic than the first model. The F-statistic value from our second model is over twice as high meaning that the overall model more effectively predicts violent crime.

We will also note that the R^2 and adjusted R^2 values are higher in the first model. This is because with more variables it is easier to estimate violent crime, while the F-statistic

takes into account that a predictor based on fewer variables is more efficient. This goes back to the point that there are less variables in the second model, and this diminishes the value of R^2. But since the R^2 is only about 6% less than the first model and the F-statistic is higher, we conclude that our second model is better than our first.

3.5 Looking For Interaction

Now that our model has been pruned down, we will check some of our variables for interaction as seen in Section 2.2. To check for interaction, we multiply the two variables we are testing to create the interaction variable. The interaction variable is then tested by creating a model to test for significance with the response variable, the terms used to create the interaction variable, and the interaction variable. After checking the variables in the second model for interaction, we find two significant interactions that we will explore. Other tests were run, but nothing of significance was found in these tests.

We will first look at the interaction between White and Poverty. The model is

$$\text{VC} = \text{-2.14} + \text{0.07W} + \text{0.64Pvty} - \text{0.007W*Pvty}.$$

The results of the model are seen below.

Table 6: Interaction Model 1

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-2.136759	1.981810	-1.078	0.282	
W	0.066070	0.030591	2.160	0.032	*
Pvty	0.645567	0.078173	8.258	2.16e-14	***
W:Pvty	-0.006982	0.001279	-5.459	1.44e-07	***

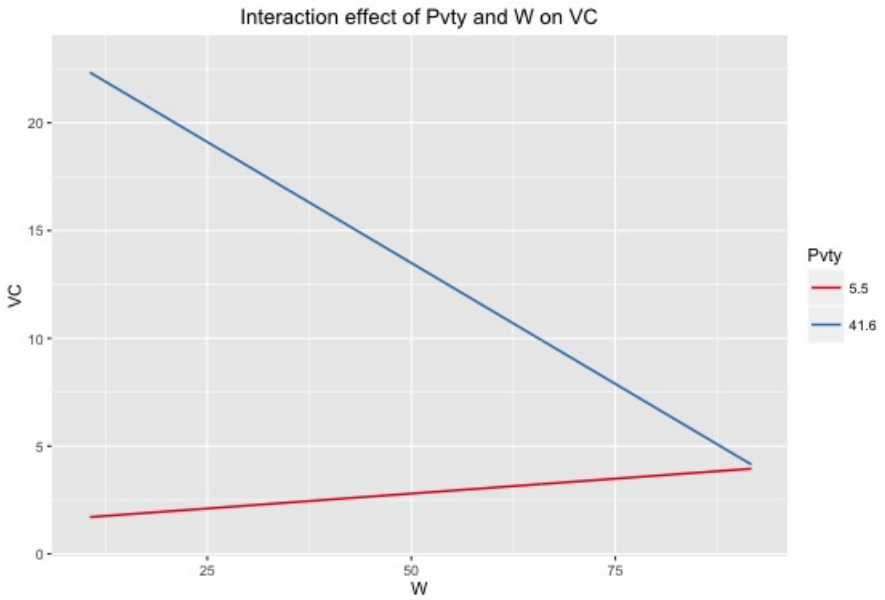
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.605 on 196 degrees of freedom

Multiple R-squared: 0.5376, **Adjusted R-squared:** 0.5305

F**-statistic:** 75.97 on 3 and 196 DF, p**-value:** $<$ 2.2e-16

For the interaction between White and Poverty, we have a negative, but high T-value with a low p-value demonstrating significance of the interaction term. Below in Graph 3, we have the corresponding graph to the interaction term.



Graph 3

Each line in the graph represents one factor level of our moderator variable, Poverty. The y-axis is Violent Crime and the x-axis is percent White. The graph reads that when Poverty is high (around 41.6%), the lower the percentage of White people, the higher the crime. If poverty is low (around 5.5%), the higher the percentage of White, the higher the crime. There is a negative interaction seen in the decreasing slope of the White variable.

We will now look at the interaction between Population Density and Population Change. The model is

$$\text{VC} = 6.57 - 0.21\text{Change} + 0.0005\text{Den} - 0.00007\text{Change*Den}.$$

The results of the model are seen below.

Table 7: Interaction Model 2

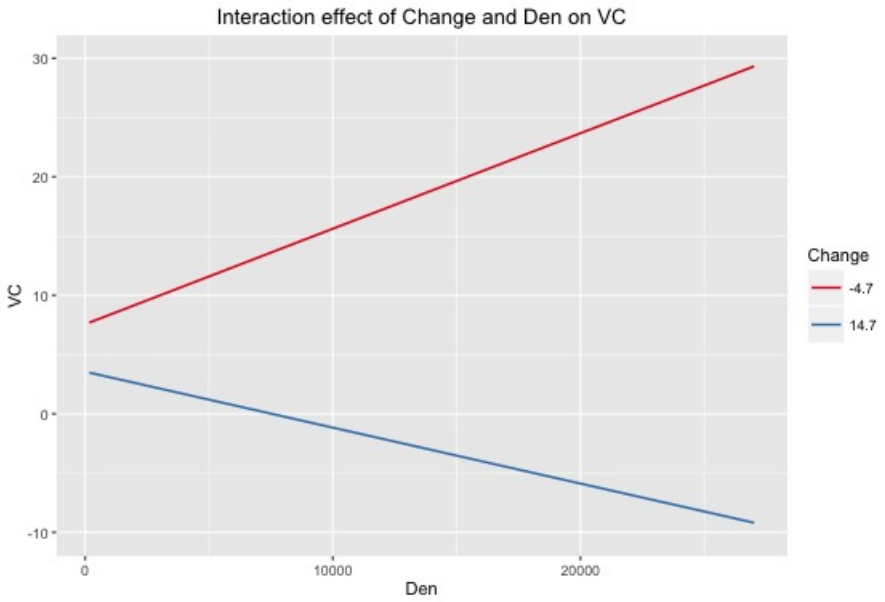
```
              Estimate Std. Error  t-value Pr(>|t|)
(Intercept)  6.586e+00  6.405e-01  10.283  < 2e-16 ***
Change      -2.065e-01  1.279e-01  -1.614  0.10812
Den          4.961e-04  1.493e-04   3.322  0.00107 **
Change:Den  -6.584e-05  3.214e-05  -2.048  0.04187 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Residual standard error: 3.422 on 196 degrees of freedom

Multiple R-squared: 0.2024, **Adjusted R-squared:** 0.1902

F-statistic: 16.58 on 3 and 196 DF, **p-value:** 1.219e-09

For the interaction between Density and Population, we have a negative, but slightly significant p-value, demonstrating significance of the interaction term. Below in Graph 4, we have the corresponding graph to the interaction term.



Graph 4

The moderator variable is Population Change and the focal variable is Population Density. The y-axis is Violent Crime and the x-axis is Population Density. Note that Violent Crime is projected to be negative in one case, which is not possible. The slope shows as Density increases that Violent Crime declines; and, if it were possible, Violent Crime would decline into the negatives. The graph reads that if Change is negative or in this case is around -4.7%, the higher the Population Density, the higher the crime. If Population Change is positive or in this case around 14.7%, the higher the Density, the lower the crime; and as we see Violent Crime gets very low in high density areas.

3.6 Third Model

After we have created our interaction variables, we will create a new maximum model with the terms from the second model and the interaction terms. The third model is

VC = 3.32 - 0.26Change + 0.05W - 1.4PPH + 0.6Pvty - 0.0002Den - 0.007W*Pvty + 0.00005Change*Den.

The results are seen below.

Table 8: Model 3

```
             Estimate  Std. Error  t-value Pr(>|t|)
(Intercept)  3.323e+00  3.015e+00   1.102   0.2717
Change      -2.560e-01  9.787e-02  -2.616   0.0096 **
W            5.425e-02  3.238e-02   1.675   0.0955 .
PPH         -1.364e+00  6.224e-01  -2.191   0.0297 *
Pvty         6.142e-01  8.830e-02   6.956 5.35e-11 ***
Den         -1.594e-04  1.266e-04  -1.259   0.2096
W:Pvty      -6.688e-03  1.350e-03  -4.954 1.59e-06 ***
Change:Den   4.722e-05  2.680e-05   1.762   0.0797 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Residual standard error: 2.533 on 192 degrees of freedom

Multiple R-squared: 0.5719, **Adjusted R-squared:** 0.5563

F**-statistic:** 36.64 on 7 and 192 DF, p**-value:** $<$ 2.2e-16

We can see from the model that our F-statistic is smaller because of the two added terms. The R^2 and adjusted R^2 are up by over five percent. The significance of the variables has been affected by the interaction terms. The interaction of Density and Change is not significant in our model probably because it was not as significant of an interaction of White and Poverty. Population Density is still not significant, meaning it probably will not be kept in the model. White is not significant, and this is probably because of the effect of the interaction variable, White and Poverty. The intercept is also not significant in this model.

3.7 Forward Elimination Procedure

Now that we have our new maximum model of seven variables, we will use the Forward Elimination Procedure. Here, we will start with one variable and test to see if we need to add variables until we have our new model. Depending on the variables left in the new model, we will check correlating variables to see if they fit better within the model. This will help us come to our final model.

The first step is to find the variable most highly correlated with violent crime. Poverty has the highest correlation to Violent Crime so the beginning model is

$$\text{VC} = \text{-0.04} + 0.31\text{Pvty}.$$

Step two involves individually adding the other variables to our first model and testing for the partial F-statistic in each new model. We will have models VC $\sim$ Pvty + W, VC $\sim$ Pvty + Change, and so on. We are looking for the most significant p-value from the partial F-statistic in equation 3. The table below shows our first time through the Forward Elimination Procedure.

Table 9: Partial F-statistic I

Variable	F**-Statistic**	p**-value**
White	57.731	1.189e-12
Change	10.49	0.001409
Persons Per Household	1.113	0.2927
Density	5.7588	0.01734
White*Poverty	47.986	<2.2e-16
Density*Change	5.5234	0.001164

Step three is to choose the variable with the largest F-statistic and highest significance and add it to our model. From this table, we can see that our White*Poverty interaction term has the most significant p-value. White*Poverty has the most significant p-value, but

its F-statistic is slightly lower than White since it has an extra variable. White*Poverty automatically adds White since the interaction needs it to work. The next step is to redo step two with the new model that we have just created. We will keep testing until there is no more significant partial F-statistic values.

We now have a new model of

$$\text{VC} = -2.137 + 0.0646\text{Pvty} + 0.067\text{W} - 0.007\text{W*Pvty}$$

and our new table of partial F-statistics is seen below. From this table, we can see that

Table 10: Partial F-statistic II

Variable	F**-Statistic**	p**-value**
Density	0.2201	0.6395
Change	4.9924	0.02659
Persons Per Household	7.7522	0.005893
Change*Density	3.4537	0.01759

we will add Persons Per Household to our model since it has the most significant value.

Our new model is

$$\text{VC} = 3.833 + 0.599\text{Pvty} + 0.044\text{W} - 0.006\text{W*Pvty} - 1.705\text{PPH}.$$

We now check if Change, Density, and Change*Density will fit in our model. The new table of partial F-statistics is seen below. Population Change is significant enough to be

Table 11: Partial F-statistic III

Variable	F**-Statistic**	p**-value**
Density	0.3089	0.579
Change	3.9499	0.04828
Change*Density	2.479	0.0625

included.

Our new model is

$$\text{VC} = 4.810 + 0.554\text{Pvty} + 0.037\text{W} - 0.006\text{W*Pvty} - 1.580\text{PPH} - 0.118\text{Change}.$$

The table below will check to see if the last two variables will make it into the model. We

Table 12: Partial F-statistic IV

Variable	F**-Statistic**	p**-value**
Density	0.3487	0.5555
Change*Density	1.7287	0.1803

can see that neither Density or Change*Density are significant enough to make the cut. We have now found the best model from the seven variables that we selected in our third model.

The next step is to go back to the twelve variables that we had in the second model to see if we can replace any variables to find a better model. We will leave poverty alone since it was the starting term with one of the highest correlations to violent crime and is required for our interaction term. We will leave White and White*Poverty alone since they were both extremely significant in our tests. This leaves us to test Persons Per Household and Population Change. The variables that are correlated to Persons Per Household include are High School Graduate, Persons Under 5, Bachelor Degree, and Foreign Born. The table below contains the partial F-statistics of the model,

$$\text{VC} = -0.530 + 0.591\text{Pvty} + 0.056\text{W} - 0.006\text{W*Pvty} - 0.134\text{Change},$$

with the new test variables. From this table, we can see that Person Per Household is still the most significant variable, therefore we will leave it in our model.

Next, we will look at the variables correlated to Population Change and find that Bachelor's Degree and Disability Under 65 are the variables correlated to Change in our model.

Table 13: Partial F-statistic V

Variable	F**-Statistic**	p**-value**
Persons Per Household	6.6815	0.01048
High School Graduate	0.0065	0.9356
Persons Under 5	0.2601	0.6106
Bachelor Degree	0.8934	0.3457
Foreign Born	1.6533	0.2000

The table below shows the partial F-statistics of the difference between Change and Bachelor's Degree. This table shows the Population Change is not the best variable since Bachelors Degree is the most significant. This concludes the Forward Elimination Procedure, and we have our final model.

Table 14: Partial F-statistic VI

Variable	F**-Statistic**	p**-value**
Change	3.9499	0.04828
Bachelor Degree	13.428	0.0003198
Disability	12.133	0.0006121

3.8 Final Model

Since we concluded the Forward Elimination Procedure, we now have all the variables for our final model. The final model is

$$\text{VC} = 12.22 + 0.46\text{Pvty} + 0.006\text{W} - 0.005\text{W*Pvty} - 2.95\text{PPH} - 0.07\text{BD}.$$

The summary of the model is seen below.

Table 15: Final Model

```
             Estimate  Std. Error  t value Pr(>|t|)
(Intercept) 12.221547   3.623702   3.373   0.000899 ***
Pvty         0.461584   0.084975   5.432   1.66e-07 ***
W            0.006352   0.031859   0.199   0.842173
PPH         -2.949837   0.684104  -4.312   2.57e-05 ***
BD          -0.066752   0.018216  -3.664   0.000320 ***
Pvty:W      -0.004763   0.001313  -3.627   0.000366 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Residual standard error: 2.484 on 194 degrees of freedom

Multiple R-squared: 0.5841, **Adjusted R-squared:** 0.5734

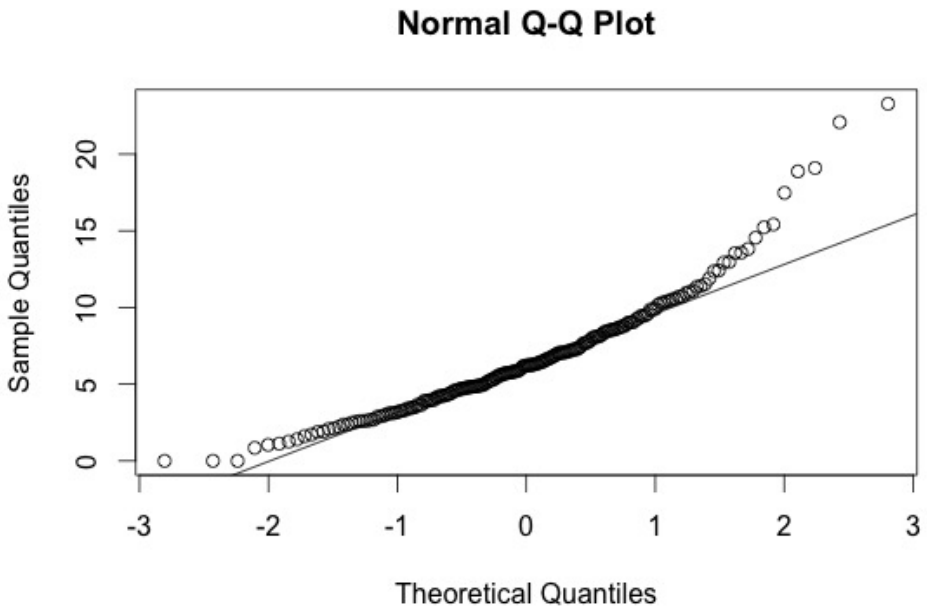
F**-statistic:** 54.49 on 5 and 194 DF, p**-value:** $<$ 2.2e-16

Persons Per Household is the only value that was not measured in percents, which is why it has a much larger coefficient. The rest of the coefficients represent percents of the population.

When we look at overall effectiveness, we can see that our R^2 value is 0.58 and our adjusted R^2 is slightly lower at 0.57. This is good, and better than the corresponding values for the third model despite the fact that we lost two variables. The F-statistic is the highest, at 54.49, that we have seen and is much better compared to the third model F-statistic of 36.64. The T-value and p-value shows the variable White to be insignificant in the model. This is because of the interaction term, since without the interaction term White would

still be significant. We need the interaction term since it gives us more accurate results. We now have a significant model for predicting violent crimes per thousand people in a city.

Lastly, we will check our model to see if the model fails the variance condition. We will do this through a Q-Q plot to see if the distribution is skewed, We want the plot to be as close to the line as possible. We will look at the graph below to see if we need to make adjustments.



Graph 5

We can see from the graph that the Q-Q plot does not perfectly fit the line, but it does fit most of the line with the exception of the beginning and the end. Because it fits most of the line, we will not make any adjustments to the model for the variance condition.

4 Conclusion

We started out with over thirteen terms and have significantly reduced our model. The table below demonstrates the progress we made with each model. The p-values are unlisted

Table 16: Model Comparison

	Model 1	**Model 2**	**Model 3**	**Final Model**
R^2	0.5793	0.5171	0.5719	0.5841
Adjusted R^2	0.5499	0.5046	0.5563	0.5734
F**-statistic**	19.70	41.54	36.64	54.49
Number of Variables	13	5	7	5

since they were all significant and had values of less than 2.2e-16. We can tell that our final model is the most significant with only five variables. This is the best model we found for 2010's violent crime rate, but there are different ways of approaching the problem. A professional statistician probably would have approached a professional about the different characteristics that might contribute to Violent Crime. They probably have the time to collect a larger sample size that would try to include most cities in the United States. This would make the results more significant. Other things to consider might be whether the city is partly suburban or the size of the suburbs outside the city and the size of the police force. There are different ways to approach the problem, but we worked with different methods to try to find one of the most significant models. Our results are defined a model with five variables, but it might be interesting to compare information from different years.

References

[1] David G. Kleinbaum, Lawrence L. Kupper, Keith E. Muller,*Applied Regression Analysis and Other Multivariable Methods*, PWS-KENT Publishing Company, 1988.

[2] James Jaccard, Robert Turrisi,*Interaction Effects In Multiple Regression*, Sage Publications, 2003.

[3] Statwing, "T-Test (Independent Samples)", <http://docs.statwing.com/examples-and-definitions/t-test>.

[4] Jim Frost, "What Is the F-test of Overall Significance in Regression Analysis?", <http://blog.minitab.com/blog/adventures-in-statistics/what-is-the-f-test-of-overall-significance-in-regression-analysis>.

[5] Laura Simon, Dr. Derek Young, "Whats a good value for R-squared?", <http://people.duke.edu/~rnau/rsquared.htm>.

[6] Robert Nau, "Multicollinearity and other Regression Pitfalls", <https://onlinecourses.science.psu.edu/stat501/node/343>.

[7] United States Census Bureau, "Quick Facts", <http://www.census.gov/quickfacts/table/PST045215/0662000>.

[8] Federal Bureau of Investigation, "Crime in the United States by Metropolitan Statistical Area, 2010", <https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/tables/table-6>.

[9] Matthew S. Chambers, "Multiple Regression Analysis", http://pages.towson.edu/mchamber/chapter4eco306.pdf

[10] Lawrence Joseph, "Interactions In Multiple Regression", http://www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-621/interaction.pdf

[11] Jack Weiss, "Lecture 35: Partial F-statistics", https://www.unc.edu/courses/2007spring/biol/145/001/docs/lectures/Nov19.html

[12] Statistics Solution "Homoscedasticity", https://www.statisticssolutions.com/homoscedasticity/

Table A1: Cities Used In Analysis

New York City, NY	Honolulu, HI	Reading, PA	Albany, NY
Los Angeles, CA	Fresno, CA	Salinas, CA	North Charleston, SC
Philadelphia, PA	Bridgeport, CT	Fairfield, CA	Council Bluffs, IA
Houston, TX	Albuquerque, NM	Santa Barbara, CA	Las Cruces, NM
Miami, FL	Omaha, NE	Santa Maria, CA	Fargo, ND
Washington D.C.	Dayton, OH	Brownsville, TX	Longview, TX
Dallas, TX	Kansas City, KS	Salem, OR	Tyler, TX
Detroit, MI	Bakersfield, CA	Anchorage, AK	Springfield, IL
Atlanta, GA	Oxnard, CA	Shreveport, LA	Bellingham, WA
Boston, MA	Tacoma, WA	Killeen, TX	Racine, WI
San Francisco, CA	Worcester, MA	Davenport, IA	Medford, OR
Phoenix, AZ	Grand Rapids, MI	Beaumont, TX	Lafayette, IN
Seattle, WA	Greensboro, NC	McAllen, TX	Johnson City, TN
Minneapolis, MN	Syracuse, NY	Thousand Oaks, CA	Lake Charles, LA
San Diego, CA	Knoxville, TN	Allentown, PA	St. Cloud, MN
Anaheim, CA	Akron, OH	Round Rock, TX	Yuma, AZ
St. Louis, MO	Little Rock, AR	Peoria, IL	Rochester, MN
Tampa, FL	Springfield, MA	Montgomery, AL	Bloomington, IN
Sacramento, CA	Stockton, CA	Tallahassee, FL	Redding, CA
Fort Worth, TX	Charleston, SC	Wilmington, NC	Loveland, CO
Newark, NJ	Toledo, OH	Fayetteville, NC	Bend, OR
Orlando, FL	Colorado Springs, CO	Evansville, IN	Columbia, MO
San Antonio, TX	Wichita, KS	Rockford, IL	Wilmington, DE
Kansas City, MO	Boise City, ID	Ann Arbor, MI	Waterloo, IA
Cleveland, OH	Fremont, CA	Savannah, GA	Eau Claire, WI
Las Vegas, NV	Hayward, CA	South Bend, IN	Janesville, WI
San Jose, CA	Cape Coral, FL	Green Bay, WI	Lynchburg, VA
Santa Clara, CA	Madison, WI	Lincoln, NE	Merced, CA
Columbus, OH	Des Moines, IA	Sunnyvale, CA	Sioux City, IA
Fort Lauderdale, FL	Jackson, MS	Roanoke, VA	Duluth, MN
Austin, TX	Palm Bay, FL	Boulder, CO	Fort Smith, AR
Virginia Beach, VA	Chattanooga, TN	Fort Collins, CO	High Point, NC
Norfolk, VA	Modesto, CA	Columbus, GA	Newport News, VA
Nashville, TN	Lancaster, PA	Erie, PA	Nampa, ID
Providence, RI	Durham, NC	Lubbock, TX	Bossier City, LA
Milwaukee, WI	Winston-Salem, NC	Clarksville, TN	Abilene, TX
Cambridge, MA	Lexington-Fayette, KY	Lafayette, LA	Billings, MT
Jacksonville, FL	Santa Rosa, CA	Gainesville, FL	Pueblo, CO
Memphis, TN	Spokane, WA	Cedar Rapids, IA	Iowa City, IA
Camden, NJ	Lansing, MI	Greeley, CO	Santa Fe, NM
Louisville Metro, KY	Springfield, MO	Amarillo, TX	Grand Junction, CO
Richmond, VA	Visalia, CA	Laredo, TX	Rocky Mount, NC
Oklahoma City, OK	Riverside, CA	Yakima, WA	Wichita Falls, TX
New Orleans, LA	Reno, NV	Waco, TX	Dothan, AL
Raleigh, NC	Flint, MI	Topeka, KS	Napa, CA
Salt Lake City, UT	Fort Wayne, IN	Chico, CA	St. Paul, MN
Buffalo, NY	Huntsville, AL	Tuscaloosa, AL	Edison Township, NJ
Hartford, CT	Mobile, AL	College Station, TX	Cary, NC
Rochester, NY	Corpus Christi, TX	Bryan, TX	Warren, MI
Tucson, AZ	Port St. Lucie, FL	Asheville, NC	Sparks, NV

Table A2: Census Bureau Information

Census Bureau City Information	**Variable Name**	**Correlation to Crime**	$\mathbb{R}$
Population, percent change - April 1, 2010 (estimates base) to July 1, 2014, (V2014)	PopulationChange	-0.39	Change
Population, Census, April 1, 2010	Population	0.04	
Persons under 5 years, percent, April 1, 2010	PersonsUnder5	0.17	U5
Persons under 18 years, percent, April 1, 2010	PersonsUnder18	0.14	
Persons 65 years and over, percent, April 1, 2010	Persons65over	-0.09	
White alone, percent, April 1, 2010 (a)	White	-0.54	W
Black or African American alone, percent, April 1, 2010 (a)	AfricanAmerican	0.60	AA
Asian alone, percent, April 1, 2010 (a)	Asian	-0.23	
Hispanic or Latino, percent, April 1, 2010 (b)	Hispanic	0.00	
Other Race Ethnicity	Other	-0.11	
Veterans, 2010-2014	Veterans	0.04	
Foreign born persons, percent, 2010-2014	ForeignBorn	-0.14	FB
Housing units, April 1, 2010	HousingUnits	0.06	
Owner-occupied housing unit rate, 2010-2014	OwnerOccupiedHousing	-0.36	OH
Median gross rent, 2010-2014	MedianRent	-0.26	
Households, 2010-2014	Households	0.05	
Persons per household, 2010-2014	PersonsPerHousehold	-0.10	PPH
Language other than English spoken at home, percent persons 5 years+, 2010-2014	LanguageOtherEnglishHome	-0.06	
High school graduate or higher, percent of persons age 25 years+, 2010-2014	HighSchoolGraduate	-0.37	HG
Bachelor's degree or higher, percent of persons age 25 years+, 2010-2014	BachelorDegree	-0.35	BD
With a disability, under age 65 years, percent, 2010-2014	DisabilityUnder65	0.53	Dis
Persons without health insurance, under age 65 years, percent	NoHealthInsurance	0.11	
In civilian labor force, total, percent of population age 16 years+, 2010-2014	Labor16over	-0.31	
Median household income (in 2014 dollars), 2010-2014	MedianHouseholdIncome	-0.47	MH
Per capita income in past 12 months (in 2014 dollars), 2010-2014	PerCapitaIncome	-0.34	
Persons in poverty, percent	Poverty	0.56	Pvty
Population per square mile, 2010	PopulationDensity	0.22	Den

This geographic level of poverty and health estimates are not comparable to other geographic levels of these estimates.
Some estimates presented here come from sample data, and thus have sampling errors that may render some apparent differences between geographies statistically indistinguishable.
(a) Includes persons reporting only one race.
(b) Hispanics may be of any race, so also are included in applicable race categories.
QuickFacts data are derived from: Population Estimates, American Community Survey, Census of Population and Housing, Current Population Survey, Small Area Health Insurance Estimates, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits.

Appendix 1: `R` code for Correlation to Violent Crime

VC <- ViolentCrime
cor(VC,Population)
0.04145199
cor(VC,PersonsUnder5)
0.1703413
cor(VC,PersonsUnder18)
0.1372125
cor(VC,Persons65over)
-0.09111918
cor(VC,White)
-0.5361357
cor(VC,AfricanAmerican)
0.6030502
cor(VC,Asian)
-0.2305441
cor(VC,Hispanic)
-0.001557841
cor(VC,Other)
-0.1113062
cor(VC,Veterans)
0.03506498
cor(VC,ForeignBorn)

-0.1381813
cor(VC,HousingUnits)
0.06292743
cor(VC,OwnerOccupiedHousing)
-0.3587487
cor(VC,MedianRent)
-0.2552356
cor(VC,Households)
0.04902502
cor(VC,PersonsPerHousehold)
-0.104659
cor(VC,LanguageOtherEnglishHome)
-0.0574769
cor(VC,HighSchoolGraduate)
-0.3659584
cor(VC,BachelorDegree)
-0.3534227
cor(VC,DisabilityUnder65)
0.5272406
(VC,NoHealthInsurance)
0.1064405
cor(VC,Labor16over)
-0.307077
cor(VC,MedianHouseholdIncome)
-0.4730326
cor(VC,PerCapitaIncome)
-0.3429419
cor(VC,Poverty)
0.5578717
cor(VC,PopulationDensity)
0.215073

Appendix 2: R code for First Model

ModelA <− lm(formula = VC ∼ Change + U5 + W + AA + FB + OH + PPH + HG + BD + Dis + MHI + Pvty + Den)
summary(ModelA)
Call:
lm(formula = VC ∼ Change + U5 + W + AA + FB + OH + PPH + HG + BD + Dis + MHI + Pvty + Den, data = City.List.test.1A)

Residuals:
Min 1Q Median 3Q Max
-6.343 -1.644 -0.312 1.078 8.194
Coefficients:
Estimate Std. Error t value Pr(> |t|)
(Intercept) 2.310e+01 9.159e+00 2.522 0.01250 *
Change -1.571e-02 7.709e-02 -0.204 0.83876
U5 2.633e-01 2.473e-01 1.065 0.28842
W -6.641e-02 2.986e-02 -2.224 0.02736 *
AA 2.510e-02 2.735e-02 0.918 0.35994
FB -2.045e-02 4.410e-02 -0.464 0.64346
OH 1.136e-03 3.829e-02 0.030 0.97636
PPH -4.114e+00 1.408e+00 -2.921 0.00392 **
HG -1.027e-01 6.832e-02 -1.504 0.13440
BD -4.403e-02 4.627e-02 -0.952 0.34257
Dis 1.348e-01 1.375e-01 0.980 0.32820
MHI 3.141e-05 4.891e-05 0.642 0.52146
Pvty 1.486e-01 7.844e-02 1.895 0.05967 .
Den 1.113e-04 8.771e-05 1.268 0.20622
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.551 on 186 degrees of freedom
Multiple R-squared: 0.5793, Adjusted R-squared: 0.5499
F-statistic: 19.7 on 13 and 186 DF, p-value: $<$ 2.2e-16

Appendix 3: `R` code Correlation Between Variables

cor(U5,W)
-0.244937
cor(U5,AA)
0.03439263
cor(U5,PPH)
0.5837726
cor(U5,HG)
-0.5701214
cor(U5,BD)
-0.6272998
cor(U5,Dis)

0.2494157
cor(U5,Pvty)
0.06874991
cor(U5,Den)
-0.1063759
cor(W,AA)
-0.7327755
cor(W,PPH)
-0.2385232
cor(W,HG)
0.4015484
cor(W,BD)
0.1872944
cor(W,Dis)
-0.1737167
cor(W,Pvty)
-0.2814416
cor(W,Den)
-0.316337
cor(AA,PPH)
-0.218435
cor(AA,HG)
-0.1291732
cor(AA,BD)
-0.1887807
cor(AA,Dis)
0.3948774
cor(AA,Pvty)
0.4405419
cor(AA,Den)
0.05849759
cor(PPH,HG)
-0.6447135
cor(PPH,BD)
-0.3964677
cor(PPH,Dis)
-0.1674323
cor(PPH,Pvty)
-0.07642432
cor(PPH,Den)
0.09159653

cor(HG,BD)
0.6486332
cor(HG,Dis)
-0.3083551
cor(HG,Pvty)
-0.4657271
cor(HG,Den)
-0.3046702
cor(BD,Dis)
-0.6545776
cor(BD,Pvty)
-0.3320978
cor(BD,Den)
0.09838038
cor(Dis,Pvty)
0.5603093
cor(Dis,Den)
-0.03313815
cor(Pvty,Den)
0.1371773
cor(U5,Change)
-0.0592979
cor(W,Change)
0.1791781
cor(Change,AA)
-0.2538461
cor(Change,PPH)
0.125447
cor(Change,HG)
0.2267743
cor(Change,Dis)
-0.5848972
cor(Change,BD)
0.4696285
cor(Change,Pvty)
-0.3811883
cor(Change,Den)
-0.060058
cor(Change,FB)
0.2488695
cor(Change,OH)

-0.004642784
cor(Change,MHI)
0.4114252
cor(FB,U5)
0.1115908
cor(FB,W)
-0.2444886
cor(FB,AA)
-0.3058707
cor(FB,PPH)
0.5890402
cor(FB,HG)
-0.4315825
cor(FB,BD)
0.1404387
cor(FB,Dis)
-0.461603
cor(FB,Pvty)
-0.1783083
cor(FB,Den)
0.5266085
cor(FB,OH)
-0.2858362
cor(FB,MHI)
0.49783
cor(U5,OH)
0.1570003
cor(W,OH)
0.4099799
cor(OH,AA)
-0.27145
cor(OH,PPH)
0.1190998
cor(OH,HG)
0.2656568
cor(OH,BD)
-0.1387183
cor(OH,Dis)
-0.0405264
cor(OH,Pvty)
-0.5052747

cor(OH,Den)
-0.5694494
cor(OH,MHI)
0.2458565
cor(MHI,U5)
-0.1231721
cor(MHI,W)
0.02938865
cor(MHI,AA)
-0.4025104
cor(MHI,PPH)
0.2296891
cor(MHI,HG)
0.3343419
cor(MHI,BD)
0.5482958
cor(MHI,Dis)
-0.6685899
cor(MHI,Pvty)
-0.8102228
cor(MHI,Den)
0.1184676

Appendix 4: `R` code for Second Model

ModelB $<-$ lm(formula = VC $\sim$ Change + W + PPH + Pvty + Den)
summary(ModelB)

Call:
lm(formula = VC $\sim$ Change + W + PPH + Pvty + Den)

Residuals:
Min 1Q Median 3Q Max
-6.747 -1.764 -0.256 1.275 8.598

Coefficients:
Estimate Std. Error t value Pr($>|t|$)
(Intercept) 1.396e+01 2.237e+00 6.240 2.69e-09 ***
Change -1.652e-01 6.179e-02 -2.674 0.00814 **
W -9.522e-02 1.232e-02 -7.726 5.81e-13 ***

PPH -2.013e+00 6.363e-01 -3.164 0.00181 **
Pvty 2.030e-01 3.135e-02 6.475 7.58e-10 ***
Den 4.104e-05 6.436e-05 0.638 0.52444
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.676 on 194 degrees of freedom
Multiple R-squared: 0.5171, Adjusted R-squared: 0.5046
F-statistic: 41.54 on 5 and 194 DF, p-value: $<$ 2.2e-16

Appendix 5: R code for Interactions

DeCh $<-$ Den * Change
ModelC $<-$ lm(formula = VC $\sim$ Change + Den + DeCh)
summary(ModelC)

Call:
lm(formula = VC $\sim$ Change + Den + DeCh)

Residuals:
Min 1Q Median 3Q Max
-7.2173 -2.1168 -0.3393 1.7506 12.6248

Coefficients:
Estimate Std. Error t value Pr($>|t|$)
(Intercept) 6.586e+00 6.405e-01 10.283 ¡ 2e-16 ***
Change -2.065e-01 1.279e-01 -1.614 0.10812
Den 4.961e-04 1.493e-04 3.322 0.00107 **
DeCh -6.584e-05 3.214e-05 -2.048 0.04187 *
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 3.422 on 196 degrees of freedom
Multiple R-squared: 0.2024, Adjusted R-squared: 0.1902
F-statistic: 16.58 on 3 and 196 DF, p-value: 1.219e-09

PW $<-$ W * Pvty
ModelD $<-$ lm(formula = VC $\sim$ W + Pvty + PW)
summary(ModelD)

Call:
lm(formula = VC ∼ W + Pvty + PW)

Residuals:
Min 1Q Median 3Q Max
-6.2635 -1.6623 -0.5073 1.1301 8.6603

Coefficients:
Estimate Std. Error t value Pr(> |t|)
(Intercept) -2.136759 1.981810 -1.078 0.282
W 0.066070 0.030591 2.160 0.032 *
Pvty 0.645567 0.078173 8.258 2.16e-14 ***
PW -0.006982 0.001279 -5.459 1.44e-07 ***
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.605 on 196 degrees of freedom
Multiple R-squared: 0.5376, Adjusted R-squared: 0.5305
F-statistic: 75.97 on 3 and 196 DF, p-value: < 2.2e-16

Appendix 6: R code for Failed Interactions

PPHP < − PPH * Pvty
ModelE < − lm(formula = VC ∼ PPH + Pvty + PPHP)
summary(ModelE)

Call:
lm(formula = VC ∼ PPH + Pvty + PPHP)

Residuals:
Min 1Q Median 3Q Max
-8.7229 -1.7082 -0.3474 1.2507 11.7941

Coefficients:
Estimate Std. Error t value Pr(> |t|)
(Intercept) -2.55649 6.23984 -0.410 0.6825
PPH 0.94817 2.34366 0.405 0.6862
Pvty 0.51499 0.26991 1.908 0.0579 .
PPHP -0.07758 0.10153 -0.764 0.4457
—

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 3.166 on 196 degrees of freedom
Multiple R-squared: 0.3171, Adjusted R-squared: 0.3067
F-statistic: 30.34 on 3 and 196 DF, p-value: 3.719e-16

PPHW < − PPH * W
ModelF < − lm(formula = VC ∼ PPH + W + PPHW)
summary(ModelF)

Call:
lm(formula = VC ∼ PPH + W + PPHW)

Residuals:
Min 1Q Median 3Q Max
-7.6654 -1.6593 -0.2544 1.3030 12.9680

Coefficients:
Estimate Std. Error t value Pr(> |t|)
(Intercept) 34.35718 6.54926 5.246 4.02e-07 ***
PPH -7.54016 2.47469 -3.047 0.00263 **
W -0.32215 0.10107 -3.187 0.00167 **
PPHW 0.07450 0.03879 1.921 0.05623 .
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 3.073 on 196 degrees of freedom
Multiple R-squared: 0.3569, Adjusted R-squared: 0.347
F-statistic: 36.26 on 3 and 196 DF, p-value: < 2.2e-16

Appendix 7: `R` code for Interaction Graphs

library(sjPlot)
library(ggplot2)
library(sjmisc)
data(efc)
library(effects)

Graph3 < − lm(VC ∼ Pvty * W)
sjp.int(Graph3, type = ”eff”)

Graph 4 < − lm(VC ∼ Change * Den)
sjp.int(Graph4, type = ”eff”)

Appendix 8: R code for Third Model

¿ ModelG < − lm(formula = VC ∼ Change + W + PPH + Pvty + Den + W*Pvty + Change*Den)
¿ summary(ModelG)
Call:
lm(formula = VC ∼ Change + W + PPH + Pvty + Den + W * Pvty + Change * Den)

Residuals:
Min 1Q Median 3Q Max
-6.9111 -1.5890 -0.3591 1.2038 8.0786
Coefficients:
Estimate Std. Error t value Pr(> $|t|$)
(Intercept) 3.323e+00 3.015e+00 1.102 0.2717
Change -2.560e-01 9.787e-02 -2.616 0.0096 **
W 5.425e-02 3.238e-02 1.675 0.0955 .
PPH -1.364e+00 6.224e-01 -2.191 0.0297 *
Pvty 6.142e-01 8.830e-02 6.956 5.35e-11 ***
Den -1.594e-04 1.266e-04 -1.259 0.2096
W:Pvty -6.688e-03 1.350e-03 -4.954 1.59e-06 ***
Change:Den 4.722e-05 2.680e-05 1.762 0.0797 .
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Appendix 9: R code for Partial F-Statistic

FEP1 < − lm(formula = VC ∼ Pvty)
FEP1B < − lm(formula = VC ∼ Pvty + W)
FEP1C < − lm(formula = VC ∼ Pvty + Change)
FEP1D < − lm(formula = VC ∼ Pvty + PPH)
FEP1E < − lm(formula = VC ∼ Pvty + Den)
FEP1F < − lm(formula = VC ∼ Pvty + W*Pvty)
FEP1G < − lm(formula = VC ∼ Pvty + Den*Change)
anova(FEP1,FEP1B)
Analysis of Variance Table

Model 1: VC ~ Pvty
Model 2: VC ~ Pvty + W
Res.Df RSS Df Sum of Sq F Pr(>F)
1 198 1981.8
2 197 1532.6 1 449.14 57.731 1.189e-12 ***
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
anova(FEP1,FEP1C)
Analysis of Variance Table

Model 1: VC ~ Pvty
Model 2: VC ~ Pvty + Change
Res.Df RSS Df Sum of Sq F Pr(>F)
1 198 1981.8
2 197 1881.6 1 100.19 10.49 0.001409 **
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

anova(FEP1,FEP1D)
Analysis of Variance Table

Model 1: VC ~ Pvty
Model 2: VC ~ Pvty + PPH
Res.Df RSS Df Sum of Sq F Pr(>F)
1 198 1981.8
2 197 1970.6 1 11.133 1.113 0.2927
anova(FEP1,FEP1E)
Analysis of Variance Table

Model 1: VC ~ Pvty
Model 2: VC ~ Pvty + Den
Res.Df RSS Df Sum of Sq F Pr(>F)
1 198 1981.8
2 197 1925.5 1 56.286 5.7588 0.01734 *
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
anova(FEP1,FEP1F)
Analysis of Variance Table

Model 1: VC ~ Pvty
Model 2: VC ~ Pvty + W * Pvty

Res.Df RSS Df Sum of Sq F Pr(>F)
1 198 1981.8
2 196 1330.3 2 651.4 47.986 ¡ 2.2e-16 ***
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
anova(FEP1,FEP1G)
Analysis of Variance Table

Model 1: VC ∼ Pvty
Model 2: VC ∼ Pvty + Den * Change
Res.Df RSS Df Sum of Sq F Pr(>F)
1 198 1981.8
2 195 1826.5 3 155.21 5.5234 0.001164 **
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
FEP2 < − lm(formula = VC Pvty + W + W*Pvty)
FEP2B < − lm(formula = VC Pvty + W + Den + W*Pvty)
FEP2C < − lm(formula = VC Pvty + W + Change + W*Pvty)
FEP2D < − lm(formula = VC Pvty + W + PPH + W*Pvty)
FEP2E < − lm(formula = VC Pvty + W + Change*Den + W*Pvty)
anova(FEP2,FEP2B)
Analysis of Variance Table

Model 1: VC ∼ Pvty + W + W * Pvty
Model 2: VC ∼ Pvty + W + Den + W * Pvty
Res.Df RSS Df Sum of Sq F Pr(>F)
1 196 1330.3
2 195 1328.8 1 1.4996 0.2201 0.6395
anova(FEP2,FEP2C)
Analysis of Variance Table
Model 1: VC ∼ Pvty + W + W * Pvty
Model 2: VC ∼ Pvty + W + Change + W * Pvty
Res.Df RSS Df Sum of Sq F Pr(>F)
1 196 1330.3
2 195 1297.1 1 33.209 4.9924 0.02659 *
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

anova(FEP2,FEP2D)
Analysis of Variance Table
Model 1: VC ∼ Pvty + W + W * Pvty

Model 2: VC ~ Pvty + W + PPH + W * Pvty
Res.Df RSS Df Sum of Sq F Pr(>F)
1 196 1330.3
2 195 1279.5 1 50.866 7.7522 0.005893 **
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

anova(FEP2,FEP2E)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + W * Pvty
Model 2: VC ~ Pvty + W + Change * Den + W * Pvty
Res.Df RSS Df Sum of Sq F Pr(>F)
1 196 1330.3
2 193 1262.6 3 67.78 3.4537 0.01759 *
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

FEP3 < − lm(formula = VC ~ Pvty + W + PPH + W*Pvty)
FEP3Z < − lm(formula = VC ~ Pvty + W + PPH)
anova(FEP3,FEP3Z)
Analysis of Variance Table

Model 1: VC ~ Pvty + W + PPH + W * Pvty
Model 2: VC ~ Pvty + W + PPH
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1279.5
2 196 1443.3 -1 -163.79 24.962 1.296e-06 ***
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
FEP3A < − lm(formula = VC ~ Pvty + W + PPH + W*Pvty + Change)
FEP3B < − lm(formula = VC ~ Pvty + W + PPH + W*Pvty + Den)
FEP3C < − lm(formula = VC ~ Pvty + W + PPH + W*Pvty + Change*Den)

anova(FEP3,FEP3A)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + PPH + W * Pvty
Model 2: VC ~ Pvty + W + PPH + W * Pvty + Change
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1279.5
2 194 1254.0 1 25.531 3.9499 0.04828 *
—

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

anova(FEP3,FEP3B)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + PPH + W * Pvty
Model 2: VC ~ Pvty + W + PPH + W * Pvty + Den
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1279.5
2 194 1277.4 1 2.0337 0.3089 0.579

anova(FEP3,FEP3C)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + PPH + W * Pvty
Model 2: VC ~ Pvty + W + PPH + W * Pvty + Change * Den
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1279.5
2 192 1231.8 3 47.711 2.479 0.0625 .
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
FEP4 < − lm(formula = VC Pvty + W + PPH + Change + W*Pvty)
FEP4A < − lm(formula = VC Pvty + W + PPH + Change + Den + W*Pvty)
FEP4B < − lm(formula = VC Pvty + W + PPH + Change + W*Pvty + Change*Den)
anova(FEP4,FEP4A)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + PPH + Change + W * Pvty
Model 2: VC ~ Pvty + W + PPH + Change + Den + W * Pvty
Res.Df RSS Df Sum of Sq F Pr(>F)
1 194 1254.0
2 193 1251.7 1 2.2615 0.3487 0.5555

anova(FEP4,FEP4B)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + PPH + Change + W * Pvty
Model 2: VC ~ Pvty + W + PPH + Change + W * Pvty + Change * Den
Res.Df RSS Df Sum of Sq F Pr(>F)
1 194 1254.0
2 192 1231.8 2 22.18 1.7287 0.1803
FEP5 < − lm(formula = VC ~ Pvty + W + W*Pvty + Change)
FEP5A < − lm(formula = VC ~ Pvty + W + W*Pvty + Change + PPH)
FEP5B < − lm(formula = VC ~ Pvty + W + W*Pvty + Change + HG)
FEP5C < − lm(formula = VC ~ Pvty + W + W*Pvty + Change + U5)

FEP5D < − lm(formula = VC ∼ Pvty + W + W*Pvty + Change + BD)
FEP5E < − lm(formula = VC ∼ Pvty + W + W*Pvty + Change + FB)

anova(FEP5, FEP5A)
Analysis of Variance Table
Model 1: VC ∼ Pvty + W + W * Pvty + Change
Model 2: VC ∼ Pvty + W + W * Pvty + Change + PPH
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1297.1
2 194 1254.0 1 43.187 6.6815 0.01048 *
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

anova(FEP5, FEP5B)
Analysis of Variance Table
Model 1: VC ∼ Pvty + W + W * Pvty + Change
Model 2: VC ∼ Pvty + W + W * Pvty + Change + HG
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1297.1
2 194 1297.1 1 0.04376 0.0065 0.9356
FEP5C < − lm(formula = VC ∼ Pvty + W + W*Pvty + Change + U5)

anova(FEP5, FEP5C)
Analysis of Variance Table
Model 1: VC ∼ Pvty + W + W * Pvty + Change
Model 2: VC ∼ Pvty + W + W * Pvty + Change + U5
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1297.1
2 194 1295.4 1 1.7367 0.2601 0.6106
FEP6 < − lm(formula = VC ∼ Pvty + W + W*Pvty + PPH)
FEP6A < − lm(formula = VC ∼ Pvty + W + W*Pvty + PPH + Change)
FEP6B < − lm(formula = VC ∼ Pvty + W + W*Pvty + PPH + BD)
FEP6C < − lm(formula = VC ∼ Pvty + W + W*Pvty + PPH + Dis)

anova(FEP5, FEP5D)
Analysis of Variance Table
Model 1: VC ∼ Pvty + W + W * Pvty + Change
Model 2: VC ∼ Pvty + W + W * Pvty + Change + BD
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1297.1
2 194 1291.2 1 5.9461 0.8934 0.3457

anova(FEP5, FEP5E)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + W * Pvty + Change
Model 2: VC ~ Pvty + W + W * Pvty + Change + FB
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1297.1
2 194 1286.2 1 10.961 1.6533 0.2

anova(FEP6, FEP6A)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + W * Pvty + PPH
Model 2: VC ~ Pvty + W + W * Pvty + PPH + Change
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1279.5
2 194 1254.0 1 25.531 3.9499 0.04828 *
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

anova(FEP6, FEP6B)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + W * Pvty + PPH
Model 2: VC ~ Pvty + W + W * Pvty + PPH + BD
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1279.5
2 194 1196.7 1 82.83 13.428 0.0003198 ***
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

anova(FEP6, FEP6C)
Analysis of Variance Table
Model 1: VC ~ Pvty + W + W * Pvty + PPH
Model 2: VC ~ Pvty + W + W * Pvty + PPH + Dis
Res.Df RSS Df Sum of Sq F Pr(>F)
1 195 1279.5
2 194 1204.2 1 75.312 12.133 0.0006121 ***

Appendix 10: R for Final Model

summary(FEP6B)

Call:
lm(formula = VC ∼ Pvty + W + W * Pvty + PPH + BD)

Residuals:
Min 1Q Median 3Q Max
-6.4916 -1.6215 -0.3525 1.0023 8.3781

Coefficients:
Estimate Std. Error t value Pr(> |t|)
(Intercept) 12.221547 3.623702 3.373 0.000899 ***
Pvty 0.461584 0.084975 5.432 1.66e-07 ***
W 0.006352 0.031859 0.199 0.842173
PPH -2.949837 0.684104 -4.312 2.57e-05 ***
BD -0.066752 0.018216 -3.664 0.000320 ***
Pvty:W -0.004763 0.001313 -3.627 0.000366 ***
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.484 on 194 degrees of freedom
Multiple R-squared: 0.5841, Adjusted R-squared: 0.5734
F-statistic: 54.49 on 5 and 194 DF, p-value: < 2.2e-16