Lecture 09 - Multivariate Regression

Samuel DeWitt

Outline

I. Random Variables and Expectations

Random Variables and Expectations

The study of statistics is the study of random variables.

A **random variable** is any empirical property which varies stochastically in the population, and whose values cannot be predicted with certainty in advance.

Each possible value has a corresponding probability of occurrence, allowing construction of a probability distribution for their realization.

Expected Values

An expected value is simply a mean of a random variable, but it has the special significance of being a population mean.

A variance is also an expected value (i.e., a mean), but it is the squared expectation centered on another expected value (the population mean).

$$E(Y_i) = \mu \Rightarrow \mu = \frac{1}{N}\sum Y_i$$

$$V(Y_i) = \sigma^2 \Rightarrow \sigma^2 = \frac{1}{N}\sum (Y_i - \mu)^2$$

Expected Values

In these formulas, summation is performed over a (possibly infinite) population of size N. The expected value and variance are also known as the **first** and **second moments**, respectively, of the random variable Y_i. Alternatively, they are known as the parameters of Y_i.

Point and Interval Estimation

Statistics entails point and interval estimation of these unknown parameters, or at least of $E(Y_i)$.

In a finite sample of n subjects randomly selected from a well-defined population, the parameters are estimated by the easily recognizable following formulas:

$$\overline{Y} = \hat{\mu} = \frac{1}{n} \sum Y_i$$

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1} \sum (Y_i - \overline{Y})^2$$

Point and Interval Estimation

Note that the sample mean is itself a random variable, because it is specific to a finite sample of n subjects, and there are infinite samples that can be drawn from the population. Since the sample mean is a random variable, it too has an expected value and a variance:

$$E(\overline{Y}) = \mu$$

$$V(\overline{Y}) = \frac{1}{n}\sigma^2$$

Unbiased and Efficient Estimators

From the first expectation, we regard the sample mean as **unbiased**, meaning that it yields the true parameter on average.

In other words, in a world where we repeatedly (and hypothetically) sample units randomly from the population, the mean of the resulting distribution of sample means (a.k.a., the sampling distribution of means) is the parameter of interest.

Unbiased and Efficient Estimators

Using the second expectation, it is possible to quantify sampling error, or inherent randomness in parameter estimates that arises from the fact that each parameter estimate is sample specific and that there are an infinite number of ways of drawing a sample of n subjects from the population.

The square root of the variance of an estimator yields what is known as a standard error, which represents a quantification of sampling error, and is in fact the standard deviation of this hypothetical distribution of sample means.

Unbiased and Efficient Estimators

Another helpful property of the sample mean is that it is efficient, meaning that it yields the smallest variance around the parameter compared to any other estimator. In other words, if we have two unbiased estimators of μ (the mean and median, say), the efficient estimator is the one that satisfies the following property:

$$\frac{1}{n-1}\sum(Y_i - \hat{\mu}_1)^2 < \frac{1}{n-1}\sum(Y_i - \hat{\mu}_2)^2$$

Central Limit Theorem

An additional property of the sample mean that comes in quite handy for statistical inference is that its distribution is (approximately) normal when the sample size is "large," irrespective of how the random variable Y_i is actually distributed in the population. This is known as the **central limit theorem**, and it allows us to characterize the estimator for the mean in the following way:

$$\overline{Y} \sim N(\mu, \sigma^2/n)$$

The N signifies that the sample mean is normally distributed (this use of N is not to be confused with the size of the population), where the normal distribution is defined by the two terms in parentheses.

II. The Simple Regression Model

The Simple Regression Model

Consider the following two-variable linear model, also known as a simple regression model. Assume for now that both variables are continuous:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

for i = 1,. . . ,n. In this model, Y_i is the dependent variable, X_i is the independent variable, α is the intercept, β is the slope, and ϵ_i is the error.

The subscripts denote the fact that Y_i, X_i, and ϵ_i are unit-specific realizations of random variables, meaning that their values differ across units. Notice that α and β do not have subscripts, representing the fact that they are the same for all units.

The Simple Regression Model

The slope represents the difference in the mean of Y_i for two hypothetical observations which differ by one unit in X_i. The intercept represents the mean of Y_i when X_i equals 0, or else the point at which the regression line crosses the Y-axis.

In sample data, the two-variable linear model is expressed using Latin letters rather than Greek letters:

$$Y_i = a + bX_i + e_i$$

where a is an estimate of α, b is an estimate of β, and e_i is technically known as a residual rather than an error.

Parts of the Simple Regression Model

There are two fundamental parts to the regression model – the prediction (the systematic part) and the error (the idiosyncratic part). This allows the model to be written in the following way:

$$Y_i = \text{prediction}_i + \text{error}_i = \hat{Y}_i + e_i$$

where it is the case that:

$$\hat{Y}_i = a + bX_i$$

Parts of the Simple Regression Model

The prediction refers to the regression or trend line as it would appear in a scatterplot of Y_i against X_i.

The error refers to the fact that the data points do not line up perfectly on the regression line. It absorbs three things: inherent randomness in Y_i, errors in the measurement of X_i and Y_i, and the effects on Y_i of variables that are not included in the model.

Parts of the Simple Regression Model

Importantly, the prediction part of the model can be thought of as a mean, it just happens to be a conditional mean:

$$\hat{Y}_i = \overline{Y} \mid X_i$$

Unbiased and Efficient Estimator - Conditional Mean

By referring to the prediction part of the regression equation as a conditional mean, we can characterize its properties in the same way as we would an unconditional mean.

Specifically, we are interested in estimators of the conditional mean that are unbiased and efficient. We can use the method of ordinary least squares (OLS) to provide unbiased and efficient estimates of α and β.

Unbiased and Efficient Estimator - Conditional Mean

The least squares principle involves choosing values for the regression coefficients that minimize the sum of squared errors:

$$min \sum_{i=1}^{n} \Rightarrow min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \Rightarrow min \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

Unbiased and Efficient Estimator - Conditional Mean

The formal expressions of the OLS estimator for the constant and slope in a two-variable model are:

$$b = \frac{\sum(Y_i - \overline{Y})(X_i - \overline{X})}{\sum(X_i - \overline{X})^2}$$

$$a = \overline{Y} - b\overline{X}$$

Unbiased and Efficient Estimator - Conditional Mean

Because sample means are random variables, and a regression slope is a (conditional) mean, it has expectations:

$$E(b) = \beta$$

$$V(b) = \frac{\sigma_e^2}{\sum(X_i - \overline{X})^2}$$

where the residual variance is estimated from sample data by the formula:

$$\sigma_e^2 = s_e^2 = \frac{1}{n-2}\sum e_i^2 = \frac{1}{n-2}\sum(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\sum(Y_i - a - bX_i)^2$$

Unbiased and Efficient Estimator - Conditional Mean

The expected value confirms that the OLS estimator is unbiased (yielding the parameter, on average, in repeated random samples) and, although the proof is not shown here, the OLS estimator can also be shown to be efficient (yielding the smallest variance compared to other estimators in its class).

A Two-Variable Illustration

Suppose that we regress the violent crime rate per 100,000 population on the percentage of persons that live below the poverty level, using data for all 50 states in the year 2000. The violent crime rate encompasses murder and non-negligent manslaughter, forcible rape, robbery, and aggravated assault. The data are from the 2002 edition of the Statistical Abstracts of the United States.

We are interested in obtaining estimates of the following population model:

$$\text{Violence}_i = \alpha + \beta \text{Poverty}_i + \epsilon_i$$

First, Descriptive Statistics

Let's first get some descriptive statistics of the two variables and inspect a histogram of the dependent variable.

First, Descriptive Statistics

```
summary(state_data00$violrate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    81.0   281.2   377.0   420.3   541.8   812.0
```

```
sd(state_data00$violrate)
```

```
## [1] 189.2019
```

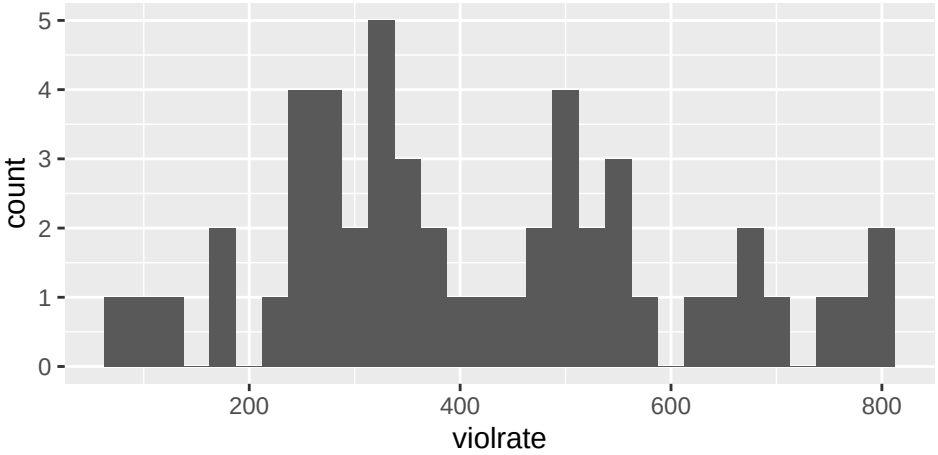
```
summary(state_data00$povrate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.300   9.425  10.550  11.440  13.375  19.300
```

```
sd(state_data00$povrate)
```

```
## [1] 2.937096
```

Scatterplot of Dependent Variable



Intercept-Only Regression Model

So the average state has 420.3 violent crimes per 100,000 ($s_Y = 189.2$) and an 11.4% poverty rate (s_X =2.9). As a convenient starting point, let's estimate an intercept-only regression model:

```
##
## Call:
## lm(formula = violrate ~ 1, data = state_data00)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -339.34 -139.09  -43.34  121.41  391.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   420.34      26.76   15.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 189.2 on 49 degrees of freedom
```

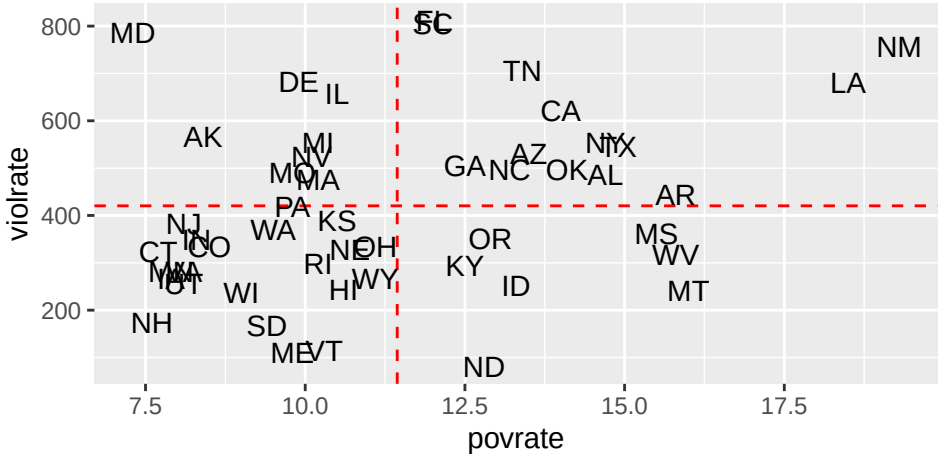
Intercept-Only Regression Model

Notice that a regression model with no regressors yields the sample mean of violrate for the estimate of the intercept ($a = 420.3$). This must be true by definition. Note also that the "Residual standard error" is the standard deviation of violrate ($s_e = 189.2$).

Scatterplot - Poverty & Violence

Before estimating the model of interest, it is always a good idea to have a look at a scatterplot, in which the means of Y_i and X_i can be used to partition the plot into quadrants.

Scatterplot - Poverty & Violence



Scatterplot - Poverty & Violence

Notice that there are 34 states in the diagonal quadrants (B, C), but only 15 states in the off-diagonal quadrants (A, D). (One state, PA, is indeterminate.) This actually provides an important clue about whether the relationship between poverty and violence is positive or negative.

In this instance, the slope will be positive. We just want to know next whether it is significantly positive or not. (As a final aside, note that the regression line will go right through the intersection of the means of Y_i and X_i.)

Bivariate Regression Model

The output from the regression model of interest is provided below:

```
##
## Call:
## lm(formula = violrate ~ povrate, data = state_data00)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -367.04 -125.00  -11.51   97.53  450.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  187.336    104.120   1.799   0.0783 .
## povrate       20.367      8.821   2.309   0.0253 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 181.4 on 48 degrees of freedom
## Multiple R-squared:  0.09997,    Adjusted R-squared:  0.08122
## F-statistic: 5.331 on 1 and 48 DF,  p-value: 0.02529
```

Bivariate Regression Model

The coefficient on the variable povrate is the estimate of the slope, b.

A slope represents the difference in the mean of Y_i for two hypothetical observations which differ by one unit in X_i. In this example, states in which 1 percent more people live in poverty have 20.4 more violent crimes per 100,000, on average.

Bivariate Regression Model

The estimate of the intercept, a, is 187.3, which represents the mean violent crime rate in a state with 0.0% poverty.

Since no such state exists in the data, the actual value of the constant is meaningless in this example.

The linear equation for this model is therefore:

$$\text{Violence}_i = 187.3 + 20.4\text{Poverty}_i + e_i$$

Bivariate Regression Model

In addition to the coefficient for the slope of the relationship between poverty and violence, we are provided an estimate of the standard error as well as the p-value for a test of the null hypothesis H_0: β=0, against the alternative hypothesis, H_1: $\beta \neq 0$. This happens to be a t-test with df $= n - 2$.

$$t = \frac{b}{s_b} = \frac{20.367}{8.821} = 2.309$$

When judged against a 5-percent (two-tailed) critical value of 2.01 (not shown in the output), this t-test is statistically significant. We can thus conclude that there is a statistically significant relationship between poverty and violent crime.

Bivariate Regression Model

Notice that the 95-percent confidence interval for the slope is:

$$b \pm t_{48}^{.05/2} x s_b = 20.367 \pm 2.01 * 8.821 = [2.637, 38.097]$$

The shorthand interpretation of this interval is that we have 95% confidence that the true slope of the poverty-violence relationship is between 2.637 and 38.097.

Bivariate Regression Model - Predicted Values

An advantage of the regression equation shown above is the ability to get predicted values of Y_i for hypothetical values of X_i. Let's have another look at the equation and consider 6 hypothetical states which span the poverty continuum.

$$\text{Violence}_i = 187.3 + 20.4\text{Poverty}_i + e_i$$

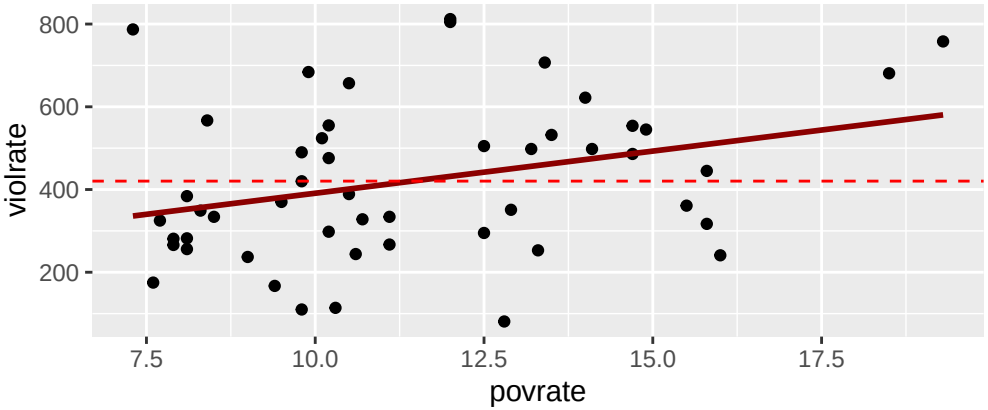
Poverty Rate	Predicted Violent Crime Rate
8.0	187.336 + (20.367*8.0)=350.272
10.0	187.336 + (20.367*10.0)=391.006
12.0	187.336 + (20.367*12.0)=431.74
14.0	187.336 + (20.367*14.0)=472.474
16.0	187.336 + (20.367*16.0)=513.208
18.0	187.336 + (20.367*18.0)=553.942

Bivariate Regression Model - Regression Line

Now we'll have another look at the scatterplot, but this time we'll overlay the regression line.

To motivate additional discussion of the regression model, let's plot both the unconditional mean and the conditional mean in the same graph.

Bivariate Regression Model - Regression Line



Bivariate Regression Model - Regression Line

What we are really interested in knowing when we estimate a regression model is whether the conditional mean, denoted by the solid sloping line, provides a better overall fit to the data points than the unconditional mean, signified by the dashed horizontal line.

In other words, does knowing a state's poverty rate help us to predict, with better than chance accuracy (i.e., better than the mean violent crime rate), that state's violent crime rate?

Bivariate Regression Model - Regression Line

- If the answer is NO:
  - The mean violent crime rate is the same (i.e., a constant) no matter the state's poverty rate, implying that the regression line is equivalent to the sample mean and thus there is no correlation between poverty and violence.
- If the answer is YES:
  - The mean violent crime rate differs depending on a state's poverty rate. In this example, the answer is YES, as indicated by the statistical significance of the poverty-violence relationship.

III. The Multiple (Multivariate) Regression Model

The Multiple Regression Model

Now, when there are multiple independent variables, the notation for the regression model changes slightly, with appropriate substitutions for sample equivalents:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ...B_j X_{ij} + \epsilon_i$$

where i = 1,. . . , n indexes units and $j = 1,\ldots,k$ indexes regressors. The parameter β_0 is used to represent the intercept (in place of α, which was the intercept in the two-variable regression model), and represents $E(Y_i)$ when all of the X_{ij}'s jointly equal 0. The parameter β_j represents the slope corresponding to an arbitrary regressor, which is one of k regressors.

The Multiple Regression Model

Each slope is now technically known as a "partial slope" or "partial coefficient," because it represents the mean difference in $E(Y_i)$ for two hypothetical observations which differ by one unit in X_{ij}, holding all other X_{ij}'s in the model constant. The expected value and variance of the slope in a multiple regression model are:

$$E(b_j) = \beta_j$$

$$V(b_j) = \frac{\sigma_e^2}{\sum(X_{ij} - \overline{X}_j)^2 * (1 - R_j^2)}$$

R_j^2 is the squared multiple correlation from a model in which regressor j is regressed on all of the other $k - 1$ regressors in the model.

The error variance is estimated by:

$$\hat{\sigma}_e^2 = s_e^2 = \frac{1}{n-k-1}\sum e_i^2 = \frac{1}{n-k-1}\sum (Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_j X_{ij})^2$$

The denominator of the error variance estimated from the sample, showing the degrees of freedom that are lost as a penalty for having to estimate $k+1$ unknown parameters (k slopes + 1 intercept).

The Multiple Regression Model - Example

To work through an example of multiple regression, let's supplement our poverty-violence analysis with some additional variables - percent with a college degree (25+ years of age), percent black, and percent of the population that lives in an urban area. The population model of interest is now:

$$\text{Violence}_i = \beta_0 + \beta_1 \text{Poverty}_i + \beta_2 \text{College}_i + \beta_3 \text{Black}_i + \beta_4 \text{Urban}_i + \epsilon_i$$

The Multiple Regression Model - Descriptive Statistics

Let's get the descriptive statistics, followed by the regression output from this model:

```
## # A tibble: 5 x 8
##   var          min    q25 median   q75   max   mean     sd
##   <chr>      <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1 black      0.300   2.35   6.75  15.0  36.3   9.90   9.58
## 2 college   15.3    22.5   24.4   27.4  34.6  24.9    4.31
## 3 povrate    7.30    9.42  10.6   13.4  19.3  11.4    2.94
## 4 urban     27.8    50.2   70.2   84.4 100    67.9   20.6
## 5 violrate  81     281.   377    542.  812   420.   189.
```

The Multiple Regression Model - Model Summary 1

```
##
## Call:
## glm(formula = violrate ~ povrate + college + black + urban)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -205.40   -74.07   -30.21    66.56   374.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -94.285    194.567  -0.485 0.630321
## povrate       18.626      7.819   2.382 0.021497 *
## college       -2.747      5.740  -0.479 0.634599
## black          7.316      2.228   3.284 0.001985 **
## urban          4.382      1.097   3.996 0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 17620.29)
##
##     Null deviance: 1754071  on 49  degrees of freedom
## Residual deviance:  792913  on 45  degrees of freedom
## AIC: 637.47
##
## Number of Fisher Scoring iterations: 2
```

The Multiple Regression Model - Model Summary 2

```
##
## Call:
## lm(formula = violrate ~ povrate + college + black + urban)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -205.40  -74.07  -30.21   66.56  374.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -94.285    194.567  -0.485 0.630321
## povrate       18.626      7.819   2.382 0.021497 *
## college       -2.747      5.740  -0.479 0.634599
## black          7.316      2.228   3.284 0.001985 **
## urban          4.382      1.097   3.996 0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 132.7 on 45 degrees of freedom
## Multiple R-squared:  0.548,  Adjusted R-squared:  0.5078
## F-statistic: 13.64 on 4 and 45 DF,  p-value: 0.0000002324
```

Note - I switched to using the glm() and lm() functions together because each provides only part of the ANOVA output for the errors from the model (the Null and Residual Deviance section) and the F-Ratio.

Regression Output Explanation

Before looking at the coefficient table, let's examine some of the other information provided in the regression output.

Regression Output Explanation - ANOVA

The ANOVA table partitions the variability in state violent crime rates into its constituent sums of squares. Each sum of squares (SS) has degrees of freedom (df), and dividing SS by its corresponding df yields a mean square (MS) or variance. In R output, the **Null*** deviance is equivalent to the **Total** sum of squares. In other words:

$$\text{Total (Null) Mean Square: } MS_T = \frac{SS_T}{df_T} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

$$\text{Model Mean Square: } MS_M = \frac{SS_M}{df_M} = \frac{1}{k}\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

$$\text{Residual Mean Square: } MS_R = \frac{SS_R}{df_R} = \frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Regression Output Explanation - F-Test

The F-statistic is provided in the output from the lm() function and can be computed using the equations above and some algebra - if we know the Total (Null) Sum of Squares and we know the Residual Sum of Squares, then simple subtraction will provide us with the Model Sum of Squares (recall that $SS_T = SS_W + SS_B$).

$$F_{k,n-k-1} = \frac{MS_M}{MS_R} = \frac{SS_M/df_M}{SS_R/df_R} = \frac{(1754071 - 792913)/4}{792913/45} = 13.64$$

where df_M = k, representing the number of regressors, and df_R = n – k – 1.

Regression Output Explanation - F-Test

The F-test and corresponding p-value are for the test of the null hypothesis that the coefficients are jointly zero, that is, $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

The alternative hypothesis for this test is that at least one coefficient differs from zero. In a manner of speaking, this is the test of the significance of the model. The p-value already indicates that the model is highly significant, with $p < 0.001$.

Regression Output Explanation - R^2

The R-square is the coefficient of determination, and quantifies the proportion of the total variability in violent crime rates (technically, the total sum of squares) that is "explained" by the regressors:

$$R^2 = \frac{SS_M}{SS_T} = 1 - \frac{SS_R}{SS_T}$$

The model indicates that 54.8% of the variability in violent crime rates is explained jointly by poverty, education, race, and urbanicity. The adjusted R-square makes a degrees-of-freedom correction to the R^2:

$$\text{Adjusted } R^2 = 1 - \frac{MS_R}{MS_T}$$

Regression Output Explanation - Residual Standard Error

The Residual standard error refers to the square root of MS_R, which is simply the residual standard deviation:

$$s_e = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Regression Output Explanation - Coefficients

The variable povrate remains statistically significant when additional regressors are introduced into the model.

Specifically, two hypothetical states that differ by 1 percentage point in their poverty rate differ by 18.626 violent crimes per 100,000, on average and all else equal.

Regression Output Explanation - Coefficients

We also see that black and urban contribute to state-to-state differences in violent crime rates.

For example, two states which differ by 1% in the percentage black population differ by 7.316 violent crimes per 100,000, on average and all else equal, whereas two states which differ in by 1% in the urban population differ by 4.382 violent crimes per 100,000, on average and all else equal.

The variable college is unrelated to violent crime net of poverty, race, and percent urban.

Regression Output Explanation - Comparing Coefficients

Once we have estimated the model, we can get some sense of the relative influence of each of the regressors on the dependent variable.

It is technically improper to directly compare regression coefficients in a model.

This is because the coefficients are interpreted in the scale of their respective regressors and are not directly comparable.

Regression Output Explanation - Comparing Coefficients

However, we can get standardized coefficients, which provide scale-free coefficients.

Standardized coefficients represent the difference in the mean of Y_i, scaled by the standard deviation of Y_i, for two hypothetical observations that differ by one standard deviation in X_i, holding all other regressors constant:

$$b_j^* = b_j * \frac{s_{X_j}}{s_Y}$$

Regression Output Explanation - Standardized Coefficients

```
##
## Call:
## lm(formula = scale(violrate) ~ scale(povrate) + scale(college) +
##     scale(black) + scale(urban))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0856 -0.3915 -0.1597  0.3518  1.9795
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.165e-16  9.922e-02   0.000 1.000000
## scale(povrate)   2.891e-01  1.214e-01   2.382 0.021497 *
## scale(college) -6.260e-02  1.308e-01  -0.479 0.634599
## scale(black)     3.705e-01  1.128e-01   3.284 0.001985 **
## scale(urban)     4.768e-01  1.193e-01   3.996 0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7016 on 45 degrees of freedom
## Multiple R-squared:  0.548,  Adjusted R-squared:  0.5078
## F-statistic: 13.64 on 4 and 45 DF,  p-value: 0.0000002324
```

Regression Output Explanation - Standardized Coefficients

In this case, the standardized coefficients indicate that urban population is the most strongly related to violent crime rates, followed by the percent black and then poverty rates.

To illustrate, states which differ in their urban population by one standard deviation (20.6, as shown in the descriptive table above) differ in their violent crime rates by 0.477 standard deviations, specifically by 90.8 violent crimes per 100,000 (0.48 * 189.2 = 90.8), on average and all else equal.

NOTE: We can also obtain a variation on the standardized coefficient by using the ppcor package, which will provide partial and semi-partial correlations.

Qualitative Regressors

Now let's consider the case in which we have a regressor that is categorical rather than continuous, such as census region. The distribution of Census region is provided below:

```
##   Midwest Northeast     South      West
##        12         9        16        13
```

Qualitative Regressors

We can inspect regional differences in the mean violent crime rate using the tapply() command:

```
##    Midwest Northeast     South      West
##   344.5000  317.3333  544.2500  409.1538
```

```
##    Midwest Northeast     South      West
##   161.8667  158.5852  182.3025  171.8419
```

Qualitative Regressors

There appear to be obvious regional differences in violence, with southern states leading the pack by a fairly large margin.

Because the numerical categories are nominal rather than ordinal, however, it is impossible to enter it into a regression model and obtain a meaningful slope.

Qualitative Regressors

Instead, it is necessary to tell R that the variable is what is known as a "factor" variable -

I have already done this and you can double check to see if you have by looking at the contents of a data frame.

A factor variable will be listed in the data frame as a "Factor with # levels..." where # is replaced by however many unique categories the variable has (in this case, 4). R will then automatically exclude one category for you as a **reference** category.

Qualitative Regressors - Model Output

```
##
## Call:
## lm(formula = violrate ~ region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -263.50 -142.29  -29.24  139.00  348.85
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       544.25      42.71  12.744  < 2e-16 ***
## regionMidwest    -199.75      65.23  -3.062  0.00367 **
## regionNortheast  -226.92      71.18  -3.188  0.00258 **
## regionWest       -135.10      63.78  -2.118  0.03961 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 170.8 on 46 degrees of freedom
## Multiple R-squared:  0.2348, Adjusted R-squared:  0.1849
## F-statistic: 4.704 on 3 and 46 DF,  p-value: 0.006024
```

Qualitative Regressors - Model Output

Notice that the model excludes the “South” category - this is because I told it to using the relevel() function.

This means that the intercept is now the mean violent crime rate in Southern states, whereas the slopes are contrasts between each region of interest and the South.

Qualitative Regressors - Model Output

This can be confirmed by inspecting the regional means in the table above.

For example, the coefficient for Northeast indicates that Northeastern states have 226.92 fewer violent crimes per 100,000 than Southern states, on average ($a - b_{Northeast}$ = 544.25 - 226.92 = 317.33).

Qualitative Regressors - Model Output

If I want different contrasts I have to use the relevel() function to shift the reference category to another region. Alternatively, I could create a series of dummy variables for each category and enter them individually into the model.

Qualitative Regressors - Model Output

The important thing to remember is that dummy variables provide a convenient way of entering one or more categorical variables into a regression model.

However, the slopes do not have the usual interpretation, in the sense of representing a contrast between two observations which differ by one unit in the regressor.

Instead, it is a contrast between observations which are included in the model and the reference category observations that are absorbed into the intercept value.

Qualitative Regressors - Full Model

Finally, let's estimate a fully specified model of violent crime to ascertain whether regional differences persist once the other relevant regressors are included. The population model is:

$$\text{Violence}_i = \beta_0 + \beta_1 \text{Poverty}_i + \beta_2 \text{College}_i + \beta_3 \text{Black}_i + \beta_4 \text{Urban}_i + \beta_5 \text{Region}_i + \epsilon_i$$

Region is understood to be a categorical variable represented by three dummy variables, meaning that β_5 comprises three coefficients (reference = south) instead of one.

Qualitative Regressors - Full Model

```
##
## Call:
## lm(formula = violrate ~ povrate + college + black + urban + region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -230.01  -77.99  -12.03   60.90  316.99
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -87.69921  205.14994  -0.427  0.67121
## povrate           14.99840    8.27071   1.813  0.07691 .
## college           -0.09191    5.90745  -0.016  0.98766
## black              6.72132    3.15914   2.128  0.03929 *
## urban              4.47399    1.09337   4.092  0.00019 ***
## regionMidwest    -42.70286   66.37047  -0.643  0.52346
## regionNortheast -126.03601   75.50710  -1.669  0.10252
## regionWest         5.10667   73.21686   0.070  0.94473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 129.2 on 42 degrees of freedom
## Multiple R-squared:    0.6,  Adjusted R-squared:  0.5334
## F-statistic: 9.001 on 7 and 42 DF,  p-value: 0.0000009813
```

Qualitative Regressors - Testing Equality

```
## Linear hypothesis test
##
## Hypothesis:
## regionMidwest = 0
## regionNortheast = 0
## regionWest = 0
##
## Model 1: restricted model
## Model 2: violrate ~ povrate + college + black + urban + region
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     45 792913
## 2     42 701601  3     91312 1.8221 0.1578
```

Qualitative Regressors - Testing Equality

I use the linearHypothesis() function from the car package to test the hypothesis that the individual coefficients from the region variable are equal to 0 (the test uses the F distribution).

The alternative hypothesis being tested is that **at least one** of those coefficients does not equal 0 (the null being that all 3 are equal to 0).

Qualitative Regressors - Testing Equality

The important part to focus on in the results from that test is the final column that indicates the probability of observing that F-ratio if the null hypothesis were true - because it is above any conventional level of statistical significance, we have to accept the null hypothesis that all three coefficients jointly equal 0.

Practically, this means that controlling for poverty, college education, percent black, and percent urban accounts for the regional differences in violent crime rate we observed earlier.

IV. Assumptions of the Linear Regression Model

Assumptions of the Linear Model

There are three groups of basic assumptions required by the linear model:(1) assumptions about the response (outcome) variable; (2) assumptions about the regressors (predictor variables); (3) assumptions about the residuals.

Assumptions of the Linear Model - Response

The three assumptions concerning the response variable (Y_i) are:

1. Linearity: the response variable is linearly related to the regressors
2. Additivity: the response variable is additively related to the regressors
3. Reliable measurement: the response variable is measured without error

Assumptions of the Linear Model - Regressors

The three assumptions concerning the regressors (X_i) are:

1. Linear independence: the regressors are not perfectly linearly related
2. Reliable measurement: the regressors are measured without error
3. Exogeneity: the regressors are not correlated with the residual

Assumptions of the Linear Model - Residuals

The four assumptions concerning the residuals (e_i) are:

1. Zero mean: The residuals have an expected value of 0.
2. Homoscedasticity: The residuals have constant variance.
3. Serial independence: The residuals are independent.
4. Normality: The residuals are normally distributed.

Assumptions of the Linear Model

With the exception of the normality assumption, these are known as the Gauss-Markov assumptions.

When these assumptions are satisfied, the least squares estimator is BLUE (the best linear unbiased estimator).

Some assumptions are more central than others, and can be violated with little practical consequence. Other assumptions have straightforward remedies in instances where they are likely to be violated.

Assumptions of the Linear Model

Violation of OLS Assumption	Is OLS Biased?
Assumptions Concerning the Response Variable (Y_i)	
Linearity	Yes
Additivity	Yes
Reliable Measurement	Maybe
Assumptions COncerning the Regressors (X_i)	
Linear Independence	No
Reliable Measurement	Maybe
Exogeneity	Yes
Assumptions Concerning the Residuals (ϵ_i)	
Zero Mean	Yes
Homoscedasticity	No
Serial Independence	No
Normality	No

Assumptions of the Linear Model

For the remainder of this section, the baseline model of interest will be as follows:

$$\text{Violence}_i = \beta_0 + \beta_1 \text{Poverty}_i + \beta_2 \text{College}_i + \beta_3 \text{Black}_i + \beta_4 \text{Urban}_i + \beta_5 \text{Region}_i + \epsilon_i$$

Assumptions Concerning the Response Variable - Linearity

A key functional form assumption of the regression model is that the relationship between Y_i and each of the regressors is well captured by a straight line. This assumption is what allows us to characterize the model in a way that is linear in the variables:

$$Y_i + \beta_0 + \beta_1 X_{i1} + ... + \beta_j X_{ij} + \epsilon_i$$

A much more general assumption, which accommodates a much broader class of models, is that the model can be expressed in a way that is **linear in the parameters**:

$$f(Y_i) = \beta_0 + \beta_1 f(X_{i1}) + ... + \beta_j f(X_{ij}) + \epsilon_i$$

Assumptions Concerning the Response Variable - Linearity

It is therefore possible to estimate non-linear relationships in a completely linear setting, as long as we can transform Y_i and/or X_{ij} in such a way that the basic linear form in the β_j's is preserved.

Assumptions Concerning the Response Variable - Linearity

Some examples of fundamentally non-linear models that can be expressed as linear in the parameters are:

$$\text{Polynomial model: } Y_i + \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i$$

$$\text{Loglinear (exponential) model: } ln(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\text{Semilog (logarithmic) model: } Y_i = \beta_0 + \beta_1 ln(X_{i1}) + \epsilon_i$$

$$\text{Loglog model: } ln(Y_i) = \beta_0 + \beta_1 ln(X_{i1}) + \epsilon_i$$

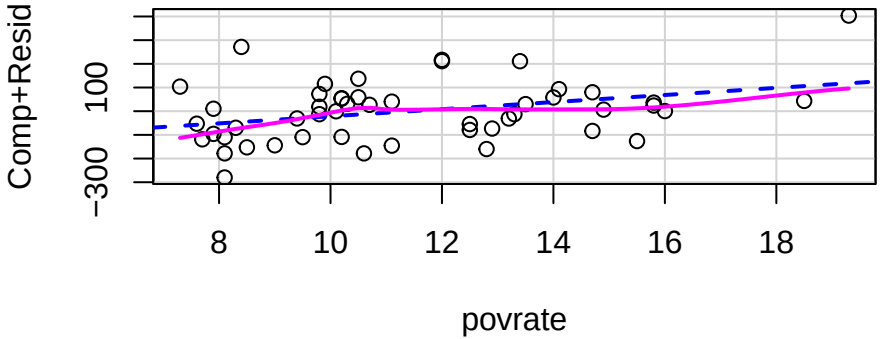
Assumptions Concerning the Response Variable - Linearity Diagnostics

Let's begin with the polynomial model. A good way to evaluate the linearity assumption is to plot the model residuals against each of the regressors.

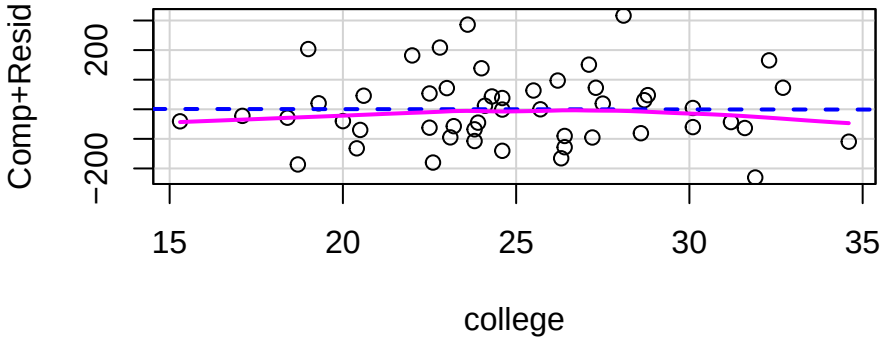
We want to ensure that a straight line provides a reasonably good fit to the data. This can be done using the crPlots() function in the car package, or augmented component-plus-residual plot, which will provide post-estimation plots of the residuals against the regressors.

In addition to including a regression line, we can add a lowess smoother, which is a nonparametric estimator that can accommodate very flexible functional forms:

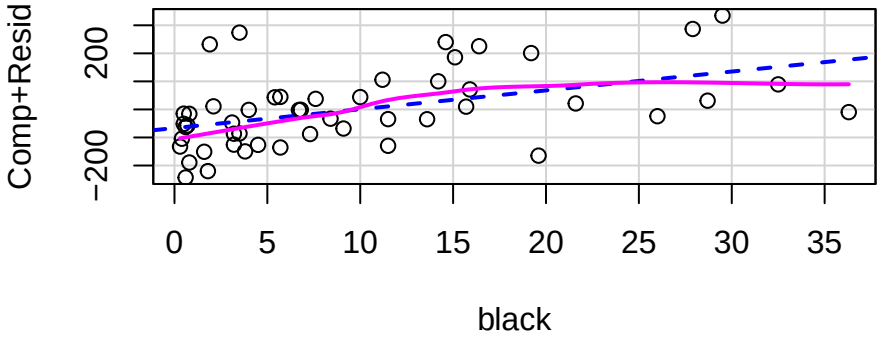
Assumptions Concerning the Response Variable - Linearity Diagnostics



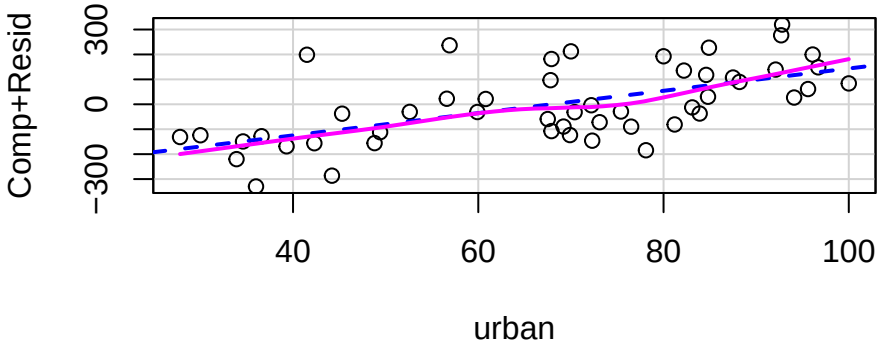
Assumptions Concerning the Response Variable - Linearity Diagnostics



Assumptions Concerning the Response Variable - Linearity Diagnostics



Assumptions Concerning the Response Variable - Linearity Diagnostics



Assumptions Concerning the Response Variable - Polynomial Model

There appears to be some mild non-linearity, mostly in the percent black variable. So let's augment our original regression model with the square of this regressor. The model to be estimated is:

$$\text{Violence}_i = \beta_0 + \beta_1 \text{Poverty}_i + \beta_2 \text{College}_i + \beta_3 \text{Black}_i + \beta_4 \text{Urban}_i + \beta_5 \text{Region}_i + \beta_6 \text{Black}_i^2 + \epsilon_i$$

Assumptions Concerning the Response Variable - Polynomial Model

```
##
## Call:
## lm(formula = violrate ~ povrate + college + black + black_sq +
##     urban + region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -274.405  -88.635    4.148   47.463  275.539
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -136.7779   198.5080  -0.689   0.4947
## povrate           15.7994     7.9571   1.986   0.0538 .
## college            1.8280     5.7491   0.318   0.7521
## black             27.5470    10.2987   2.675   0.0107 *
## black_sq          -0.5761     0.2722  -2.116   0.0404 *
## urban              2.4110     1.4333   1.682   0.1002
## regionMidwest    -20.8969    64.6089  -0.323   0.7480
## regionNortheast  -78.5233    75.9562  -1.034   0.3073
## regionWest        83.7265    79.5671   1.052   0.2988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124.2 on 41 degrees of freedom
## Multiple R-squared:  0.6394, Adjusted R-squared:  0.569
## F-statistic: 9.088 on 8 and 41 DF,  p-value: 0.000000445
```

Assumptions Concerning the Response Variable - Polynomial Model

There is indeed some non-linearity in black, as indicated by the significance of the quadratic term. The fact that the linear term is positive (27.6) and the quadratic term is negative (–0.58) indicates something about the functional form.

The story appears to be that increases in percent black correspond with higher violent crime rates until a threshold is reached, at which point further increases in percent black are uncorrelated with violence.

Assumptions Concerning the Response Variable - Polynomial Model

In order to determine where this threshold occurs, we first need to differentiate the model with respect to percent black:

$$\frac{\partial \text{Violence}_i}{\partial \text{Black}_i} = \frac{\partial[\beta_0 + \beta_1 \text{Poverty}_i + \beta_2 \text{College}_i + \beta_3 \text{Black}_i + \beta_4 \text{Urban}_i + \beta_5 \text{Region}_i + \beta_6 \text{Black}_i^2 + \epsilon_i]}{\partial \text{Black}_i} = \beta_3 + 2 * \beta_6 \text{Black}_i$$

Assumptions Concerning the Response Variable - Polynomial Model

Then, we set the first derivative equal to zero:

$$\beta_3 + 2 * \beta_6 \text{Black}_i = 0$$

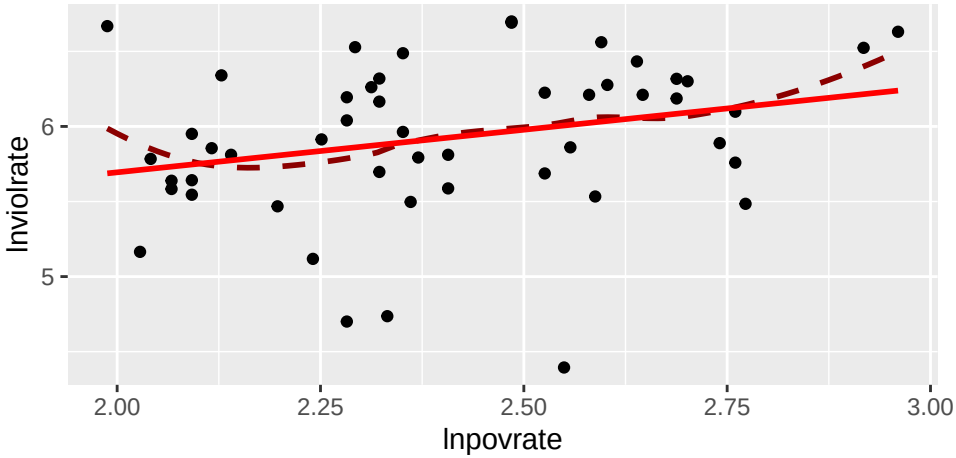
Finally, we solve the equation for black to identify the inflection point, which in this case is a maximum:

$$\text{Max}(\text{Black}_i) = \frac{-\beta_3}{2 * \beta_6} = \frac{-27.547}{2 * (-0.576)} = 23.91$$

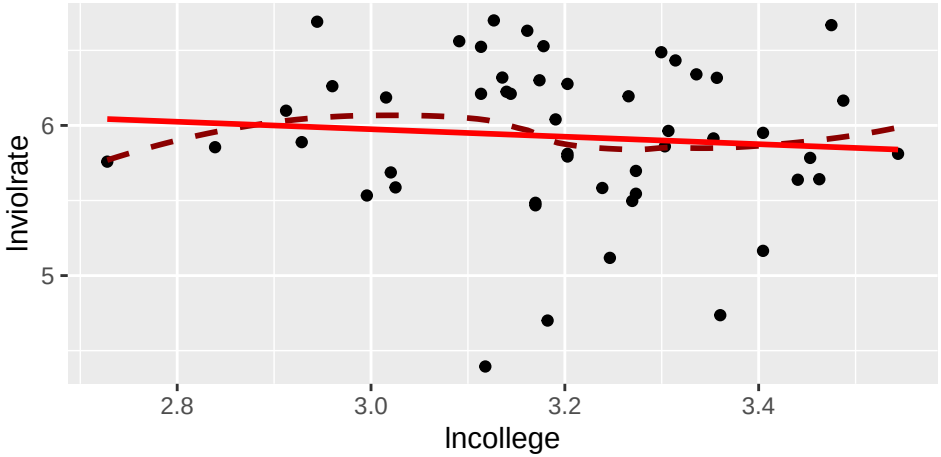
Assumptions Concerning the Response Variable - Logarithmic Model

When we have variables that are continuous and non-zero, it can also be advantageous to consider a logarithmic transformation. Let's first have a look at some scatterplots.

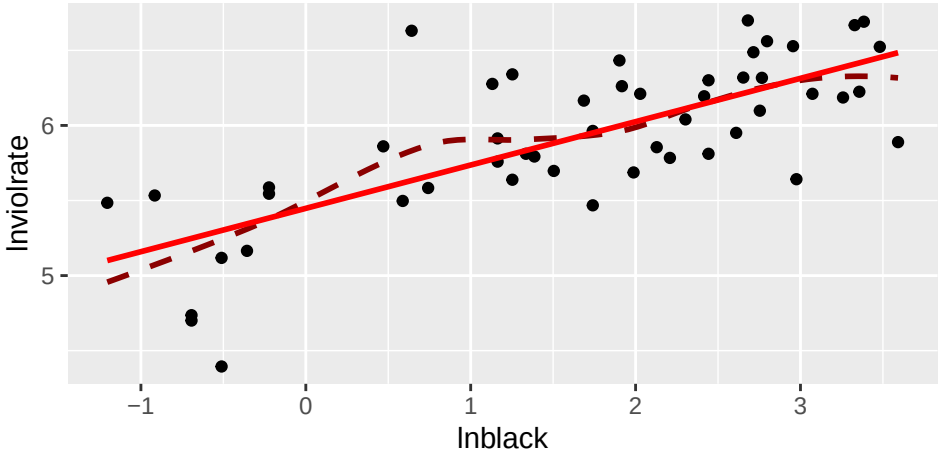
Assumptions Concerning the Response Variable - Logarithmic Model



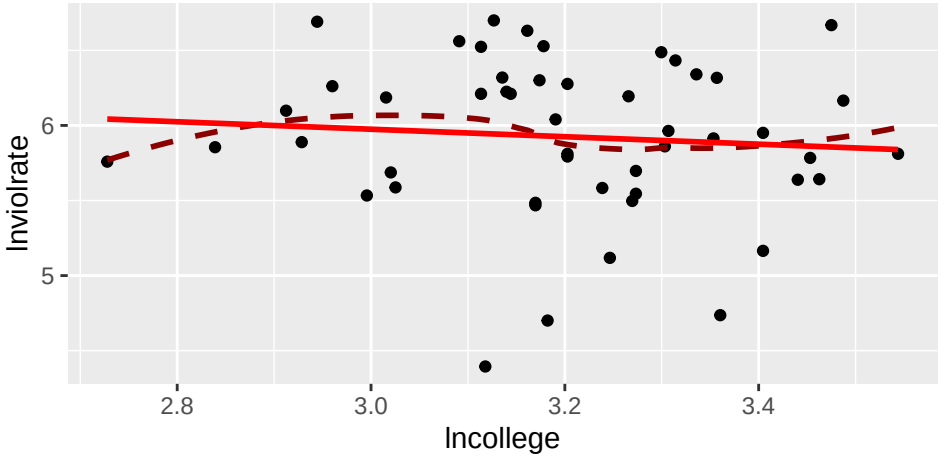
Assumptions Concerning the Response Variable - Logarithmic Model



Assumptions Concerning the Response Variable - Logarithmic Model



Assumptions Concerning the Response Variable - Logarithmic Model



Assumptions Concerning the Response Variable - Logarithmic Model

With the possible exception of lnblack, the functions again all appear to be strictly linear. For the time being, we will assume linearity and retain all of the continuous variables in log metric. We will concern ourselves with a log-log model of the form:

$$\ln(\text{Violence}_i) = \beta_0 + \beta_1 ln(\text{Poverty}_i) + \beta_2 ln(\text{College}_i) + \beta_3 ln(\text{Black}_i) + \beta_4 ln(\text{Urban}_i) + \beta_5 \text{Region}_i + \epsilon_i$$

Assumptions Concerning the Response Variable - Logarithmic Model

In this model, the slopes for the log-transformed regressors (lnpoverty, lncollege, lnblack, lnurban) are all known as "elasticities," and they represent the percentage difference in Yi for two hypothetical observations which differ by one percent in X_i.

The slopes for the dummy regressors (region) represent proportional differences in violent crime between the noted region and the reference region.

Assumptions Concerning the Response Variable - Logarithmic Model

```
##
## Call:
## lm(formula = lnviolrate ~ lnpovrate + lncollege + lnblack + lnurban +
##     region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71902 -0.12846  0.01743  0.16862  0.52029
##
## Coefficients:
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      3.07817    1.34210   2.294    0.0269 *
## lnpovrate        0.36978    0.21257   1.740    0.0893 .
## lncollege        0.08695    0.31832   0.273    0.7861
## lnblack          0.30267    0.06179   4.898 0.0000148 ***
## lnurban          0.26197    0.19016   1.378    0.1756
## regionMidwest   -0.01648    0.13810  -0.119    0.9056
## regionNortheast -0.08426    0.16827  -0.501    0.6192
## regionWest       0.40977    0.16816   2.437    0.0191 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2849 on 42 degrees of freedom
## Multiple R-squared:  0.7431, Adjusted R-squared:  0.7002
## F-statistic: 17.35 on 7 and 42 DF,  p-value: 0.0000000001485
```

Assumptions Concerning the Response Variable - Logarithmic Model

Note that elasticities are more often multiplied by 10 to make their interpretation meaningful.

Let's focus on the coefficient for lnpovrate. It indicates that a state possessing a poverty rate that is 10% higher than a counterpart state (say, 11% poverty versus 10% poverty, or 22% versus 20%) has a violent crime rate that is 3.7% higher, on average and all else equal.

NOTE: Technically, for a state with 10% higher poverty, the percent difference in the violent crime rate is $100 * [110/100]^{0.370} - 1 = 3.59\%$).

Assumptions Concerning the Response Variable - Logarithmic Model

With respect to the regional dummies, the slopes are best multiplied by 100 to give them a **percent difference** interpretation.

For example the coefficient for West indicates that the violent crime rate of Western states is 41% higher than Southern states, on average and all else equal.

Assumptions Concerning the Response Variable - Logarithmic Model

Regressor	Standard Model	Log-Log Model
Poverty	14.998(8.271)+	.370 (.213)+
College	-.092(5.907)	.087(.318)
Black	6.721(3.159)*	.303(.062)***
Urban	4.474(1.093)***	.262(.190)
Midwest	-42.703(66.370)	-.016(.138)
Northeast	-126.036(75.507)	-.084(.168)
West	5.107(73.217)	.410(.168)*
R^2	0.600	0.743

+p<.10; *p<.05; ** p<.01; *** p<.001

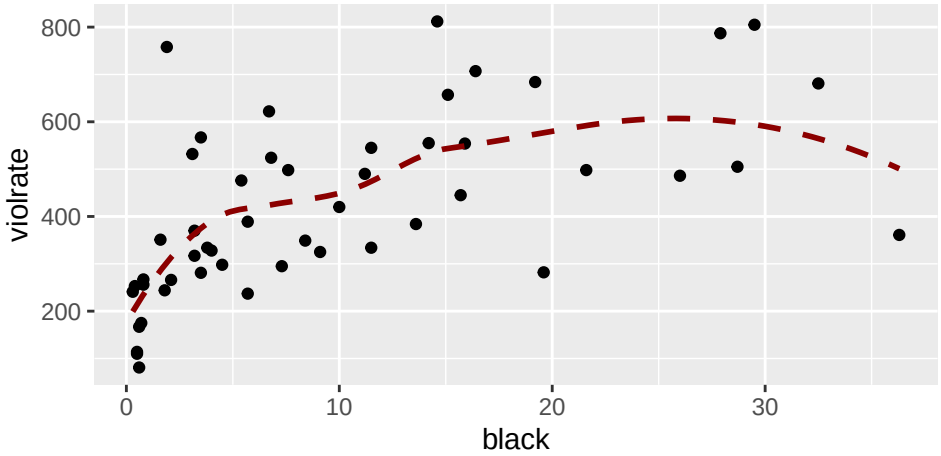
Assumptions Concerning the Response Variable - Logarithmic Model

In addition to altering the metric of the model, the natural logarithm has one characteristic that can be advantageous. Namely, it can accommodate a relationship between X_i and Y_i that is fundamentally non-linear.

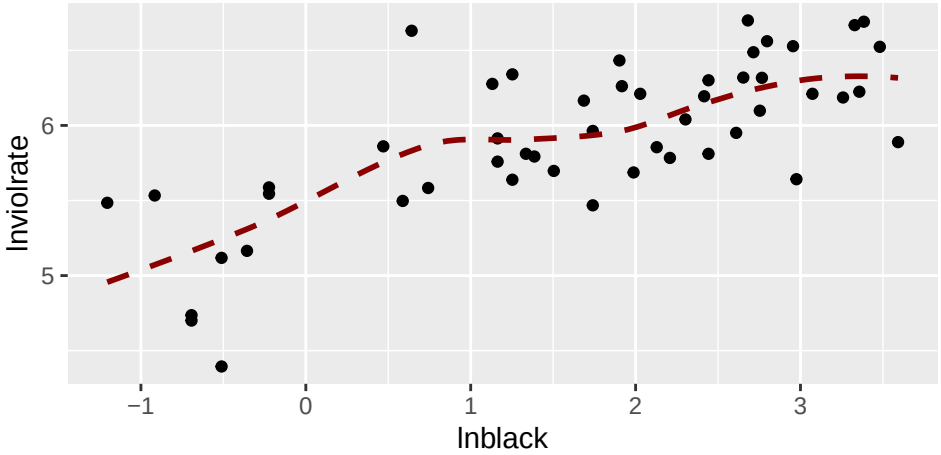
Let's return to the regressor for percent black, for which we have already seen that violent crime exhibits mild non-linearity.

We will inspect two scatterplots - one which keeps **violrate** and **black** in their original metric and one which performs a log-transformation on both.

Assumptions Concerning the Response Variable - Logarithmic Model



Assumptions Concerning the Response Variable - Logarithmic Model



Assumptions Concerning the Response Variable - Logarithmic Model

Interestingly, the log-log scatterplot linearizes the relationship between percent black and violent crime rates.

While there remains very mild non-linearity in the log-log scatterplot, the fact that the regression line is always increasing (i.e., there is no inflection point) indicates that we could preserve degrees of freedom by estimating a log-log model as opposed to a polynomial model.

Assumptions Concerning the Response Variable - Additivity

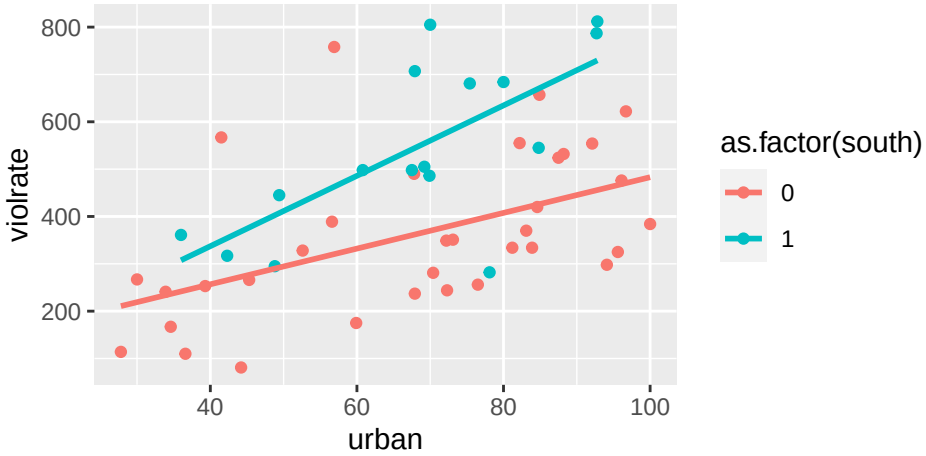
A second functional form assumption is that the relationship between X_i and Y_i does not depend on the value of one or more other regressors.

Assumptions Concerning the Response Variable - Additivity

Suppose we have reason to believe that states which are more urbanized have higher violent crime rates, but this relationship is magnified in the South, where we believe that the percent of the population living in urban areas is even more positively correlated with violence.

We can see this in a scatterplot of violence against percent urban, both of which are retained in their original metric for the time being:

Assumptions Concerning the Response Variable - Additivity



Assumptions Concerning the Response Variable - Additivity

Now consider the fully specified regression model in log-log form:

$$ln(\text{Violence}_i) = \beta_0 + \beta_1 ln(\text{Poverty}_i) + \beta_2 ln(\text{College}_i) + \beta_3(\text{Black}_i) + \beta_4 ln(\text{Urban}_i) + \beta_5(\text{Region}_i) + \beta_6 ln(\text{Urban}_i * (\text{Region}_i = South) + \epsilon_i$$

Assumptions Concerning the Response Variable - Additivity

```
##
## Call:
## lm(formula = lnviolrate ~ lnpovrate + lncollege + lnblack + lnurban +
##     region + south * lnurban, data = state_data00)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72154 -0.14549  0.00226  0.18229  0.46188
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value   Pr(>|t|)
## (Intercept)    4.565378   1.449510   3.150   0.003048 **
## lnpovrate      0.414168   0.200923   2.061   0.045652 *
## lncollege     -0.155150   0.314687  -0.493   0.624621
## lnblack        0.352702   0.061462   5.739 0.00000102 ***
## lnurban        0.044349   0.198718   0.223   0.824509
## regionSouth   -3.496622   1.427197  -2.450   0.018644 *
## regionMidwest -0.009391   0.131272  -0.072   0.943319
## regionWest     0.462306   0.125607   3.681   0.000672 ***
## south                NA         NA      NA         NA
## lnurban:south  0.816235   0.323308   2.525   0.015547 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2682 on 41 degrees of freedom
## Multiple R-squared:  0.7776, Adjusted R-squared:  0.7342
## F-statistic: 17.92 on 8 and 41 DF,  p-value: 0.00000000003823
```

Assumptions Concerning the Response Variable - Additivity

The slope for the relationship between lnurban and lnviolrate is obtained by differentiating the equation with respect to lnurban, and is provided by the formula:

$$\frac{\partial ln(\text{Violence}_i)}{\partial ln(\text{Urban}_i)} = \beta_4 + \beta_6(\text{Region}_i = \text{South}) = 0.044 * 0.816(\text{Region}_i = \text{South})$$

Assumptions Concerning the Response Variable - Additivity

There are two different slopes, one for non-southern states and one for southern states:

$$\frac{\partial ln(\text{Violence}_i)}{\partial ln(\text{Urban}_i)} = \begin{cases} 0.044 \text{ if Region}_i \neq \text{South} \\ 0.860 \text{ if Region}_i = \text{South} \end{cases} \tag{1}$$

Assumptions Concerning the Response Variable - Additivity

Notice that, for non-southern states, the relationship between percent urban and violence is not statistically significant.

In southern states, on the other hand, there is a strong positive relationship, indicating that southern states with 10% higher percent urban (e.g., 33% versus 30%, or 55% versus 50%) have 8.6% higher violent crime rates.

NOTE: Technically, $100 * [110/100^{0.816} - 1] = 8.09\%$.

Assumptions Concerning the Response - Reliable Measurement

It is important that the response variable be perfectly measured. However, measurement error will not introduce any bias into the model as long as that error is random. Instead, it will only make the regression coefficients less precise by inflating the model variance. Assume that the model we want to estimate is as follows:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Assumptions Concerning the Response - Reliable Measurement

However, assume that the response variable is measured with error, so we measure Y_i when we actually want to measure the true Y_i^*:

$$Y_i = Y_i^* + \mu_i$$

Assumptions Concerning the Response - Reliable Measurement

The regression model written in terms of Y_i therefore becomes:

$$Y_i = \alpha + \beta X_i + \epsilon_i + \mu_i \tag{2}$$
$$= \alpha + \beta X_i + \omega_i \tag{3}$$

where

$$\omega_i = \epsilon_i + \mu_i$$

Assumptions Concerning the Response - Reliable Measurement

The residual in this model, ω_i, has more variability than the residual in the theoretical model, ϵ_i. The only consequence of random noise is therefore larger standard errors.

Assumptions Concerning the Response - Reliable Measurement

However, the issue becomes more complicated if the measurement error is non-random, meaning that it is correlated with one or more regressors, In this case, the μ_i represents an omitted variable that is potentially correlated with X_i, which will introduce bias in the coefficient for X_i.

We'll return to this in the next section when we discuss unreliable measurement and endogeneity in the regressors.

Assumptions Concerning the Regressors - Linear Independence

A key assumption for identifiability of the regression model is linear independence of the regressors.

Specifically, a regressor cannot be a perfect linear combination of other regressors, and the regressors cannot be perfectly correlated with each other.

Another way to express this is to say that the matrix of regressors must have "full column rank."

Assumptions Concerning the Regressors - Linear Independence

Perfect linear dependence characterizes a categorical variable which is coded into dummy variables.

This is the reason that, when we perform regression with dummy regressors, we must omit the dummy variable for one of the categories, which is labeled the reference category.

Assumptions Concerning the Regressors - Linear Independence

Consider the census region example used earlier. Perfect linear dependence arises because, once we know that a state is coded 0 on neast, mwest, and south, it must be coded 1 on west.

So the sum of all four dummy variables is 1 for all states, which means there is a perfect linear combination of the dummy regressors unless one of them is excluded from the model.

Assumptions Concerning the Regressors - Linear Independence

It is easy to know if perfect linear combinations of regressors exist because the regression model will break down when it is estimated.

On the other hand, we can still estimate the regression model if we have high correlations among any of the regressors, and it can create problems.

High correlations lead to a problem known as **multicollinearity**, and it can make regression coefficients highly unstable. The first place to start is with an examination of a correlation matrix of the variables included in the model. We are looking for unusually high correlations.

Assumptions Concerning the Regressors - Linear Independence

```
##            lnpovrate lncollege lnblack lnurban neast mwest south  west
## lnpovrate       1.00     -0.55    0.12   -0.19 -0.27 -0.29  0.40  0.10
## lncollege      -0.55      1.00   -0.11    0.36  0.39  0.01 -0.40  0.08
## lnblack         0.12     -0.11    1.00    0.58 -0.14 -0.06  0.61 -0.47
## lnurban        -0.19      0.36    0.58    1.00  0.12 -0.08  0.04 -0.06
## neast          -0.27      0.39   -0.14    0.12  1.00 -0.26 -0.32 -0.28
## mwest          -0.29      0.01   -0.06   -0.08 -0.26  1.00 -0.39 -0.33
## south           0.40     -0.40    0.61    0.04 -0.32 -0.39  1.00 -0.41
## west            0.10      0.08   -0.47   -0.06 -0.28 -0.33 -0.41  1.00
```

Assumptions Concerning the Regressors - Linear Independence

None of these correlations gives cause for alarm.

However, a disadvantage of this diagnostic approach is that it only provides information about collinearity problems which are bivariate.

If there are linear combinations of three or more variables, an alternative test is required. After the model is estimated, we can inspect the variance inflation factors (VIFs). One general rule is to be concerned about a VIF that exceeds 2.0 and especially one that is in proximity to 5.0.

Assumptions Concerning the Regressors - Linear Independence

There might be a reason to be concerned about collinearity problems in this model.

The highest VIF is for lnblack, but this regressor is already statistically significant, so it does not seem that multicollinearity is creating any efficiency problems for this variable.

On the other hand, notice that lnurban and the regional dummies all have VIFs larger than 2.0, and the dummy regressor for southern states has a modestly large VIF.

```
## lnpovrate lncollege   lnblack   lnurban     mwest     south      west
##  1.700140  1.919903  4.081959  2.652077  2.066013  3.796038  2.088499
```

Assumptions Concerning the Regressors - Linear Independence

The following table examines several specifications of the basic regression model (regression output is not shown):

Regressor	Model A All Variables	Model B Black Excluded	Model C Region Excluded	Model D Region Modified
Poverty	.370(.213)+	.436(.263)	.566(.227)*	.368(.199)+
College	.087(.318)	-.070(.392)	-.001(.015)	.038(.300)
Black	.303(.062)***	—	.208(.046)***	.318(.049)***
Urban	.262(.190)	.949(.159)***	.008(.003)*	.232(.176)
Region				
Midwest	.068(.136)	.217(.164)	—	—
South	.084(.168)	.547(.173)**	—	—
West	.484(.133)***	.349(.160)*	—	.457(.112)***
R^2	0.743	0.596	0.642	0.741
Mean VIF	2.61	1.89	1.79	1.99

+ p<.10; * p<.010; ** p<.05; *** p<.001

Assumptions Concerning the Regressors - Linear Independence

The pattern of results suggests that regional differences in violence arise largely because of regional differences in such things as poverty, education, racial composition, and percent urban.

Interestingly, the first and final models indicate that western states have significantly higher violent crime rates than would be predicted based on their profile on these continuous regressors.

When all is said and done, however, the original model (Model A) does not appear to be severely plagued by multicollinearity.

Assumptions Concerning the Regressors - Reliable Measurement

The assumptions concerning the measurement properties of the regressors are more salient than the measurement properties of the response variable.

This is because measurement error in one or more regressors, even if it is random, introduces bias in the coefficients as opposed to just imprecision.

We'll begin with the simplest case – classical measurement error. This refers to measurement error which is independent of both the true variable and the residual.

Assumptions Concerning the Regressors - Reliable Measurement

To formalize this, let's say that our data reveal X_i when we actually want to measure the true X_i^*.

$$X_i = X_i^* + \nu_i$$

Assumptions Concerning the Regressors - Reliable Measurement

So instead of estimating the regression model we want with the true X_i^* on the right hand side, we actually end up estimating:

$$\begin{aligned} Y_i &= \alpha\beta(X_i - \nu_i) + \epsilon_i && (4) \\ &= \alpha + \beta X_i + \epsilon_i - \beta\nu_i && (5) \\ &= \alpha + \beta X_i + \psi_i && (6) \end{aligned}$$

where

$$\psi_i = \epsilon_i - \beta\nu_i$$

Assumptions Concerning the Regressors - Reliable Measurement

In the case of classical measurement error, or measurement error that is random, we assume that:

$$\nu_i \perp X_i^*$$

and

$$\nu_i \perp \epsilon_i$$

Assumptions Concerning the Regressors - Reliable Measurement

It can be shown that the OLS estimator of β is biased downward (and is inconsistent, meaning that the bias does not lessen as the sample size gets infinitely large):

$$b \to \beta * \frac{\sigma^2_{X^*}}{\sigma^2_{X^*} + \sigma^2_{\nu}}$$

Assumptions Concerning the Regressors - Reliable Measurement

The ratio in parentheses represents the **reliability** of X_i, and is always less than (or equal to) one, which leads to an *attenuation bias* in OLS. It means that the noisier the measure, the worse the attenuation.

Additionally, when there are other regressors in the model, bias pervades the other coefficients, even if the other regressors are perfectly measured.

Assumptions Concerning the Regressors - Reliable Measurement

Things get much less predictable when the measurement error is not random, for example, when it is correlated with the true X_i^*.

Because the measurement error is absorbed by the residual, the model now suffers from omitted variables bias, because X_i will be correlated with the residual. We will consider omitted variables bias next.

Assumptions Concerning the Regressors - Exogeneity

The exogeneity assumption is another way of saying that the regressors should not be correlated with the residual, or that there is no omitted variables bias.

If we were able to conduct an experiment in which we randomized the values of X_i, this assumption would be met by definition.

Formally, the regression model requires that:

$$E(X_i, \epsilon_i$$

Assumptions Concerning the Regressors - Exogeneity

Let's consider the following two models:

True model:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \epsilon_i$$

Estimated model:

$$Y_i = \alpha^* + \beta^* X_i + \epsilon_i^*$$

Where the error in the estimated model is:

$$\epsilon_i^* = \gamma Z_i + \epsilon_i$$

Assumptions Concerning the Regressors - Exogeneity

If Z_i is uncorrelated with X_i, there is no omitted variables bias. Bias in the estimator for β arises when Z_i and X_i are correlated.

In fact, the expected value of the OLS estimator when Z_i is excluded is:

$$E(b) = \beta + \gamma \frac{\text{Cov}(X_i, Z_i)}{\text{Var}(X_i)}$$

The second term represents the bias, and it reveals that, assuming $\gamma \neq 0$, OLS remains unbiased when Z_i is excluded if and only if $\text{Cov}(X_i, Z_i) = 0$.

Assumptions Concerning the Regressors - Exogeneity

Unfortunately, there is no foolproof method of diagnosing endogeneity or omitted variables bias. One method that is occasionally encountered is called a RESET test (**regression specification error test**), which involves supplementing the model with a polynomial function of the fitted values from the baseline model:

Baseline Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$

Supplementary Model: $Y_i = \beta_0^* + \beta_1^* X_{i1} + \beta_2^* X_{i2} + \beta_3^* X_{i3} + \beta_4^* \hat{Y}_i^2 + \beta_5^* \hat{Y}_i^3 + \beta_6^* \hat{Y}_i^4 + \epsilon_i$

Assumptions Concerning the Regressors - Exogeneity

The RESET test is the joint significance of the coefficients for the polynomials, i.e., a test of the null hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ against the alternative that at least one coefficient is non-zero.

Rejection of the null hypothesis indicates that the baseline model might be misspecified.

Note that a different order of the polynomial can be easily specified (e.g., a cubic rather than quartic function), and the RESET test is adjusted accordingly.

Assumptions Concerning the Regressors - Exogeneity

```
##
##  RESET test
##
## data:  lnviolrate ~ lnpovrate + lncollege + lnblack + lnurban + region
## RESET = 0.84197, df1 = 4, df2 = 38, p-value = 0.5073
```

Based on this diagnostic, we have little reason to be concerned about model misspecification.

Assumptions Concerning the Regressors - Exogeneity

Note that this test is agnostic about the nature of the misspecification, and provides no guidance about the appropriate alternative model.

For example, had we rejected the null hypothesis, we would not know whether it was necessary to include a polynomial function of one or more regressors, or whether there were omitted variables that were biasing the estimator.

Assumptions Concerning the Residuals - Zero Mean

This assumption is related to the exogeneity assumption and is what allows us to generate the expected of Y_i as:

$$E(Y_i) = \beta_0 + \beta_1 E(X_{i1}) + ... + \beta_j E(X_{ij})$$

This is because the expected value of the residual, conditional on X_{ij}, is assumed to be zero:

$$E(\epsilon_i \mid X_{i1}, ..., X_{ik}) = 0$$

Assumptions Concerning the Residuals - Zero Mean

Violation of the assumption does not have any bearing on any of the coefficients except for the constant. The slope coefficients remain unbiased, as long as the other model assumptions are met.

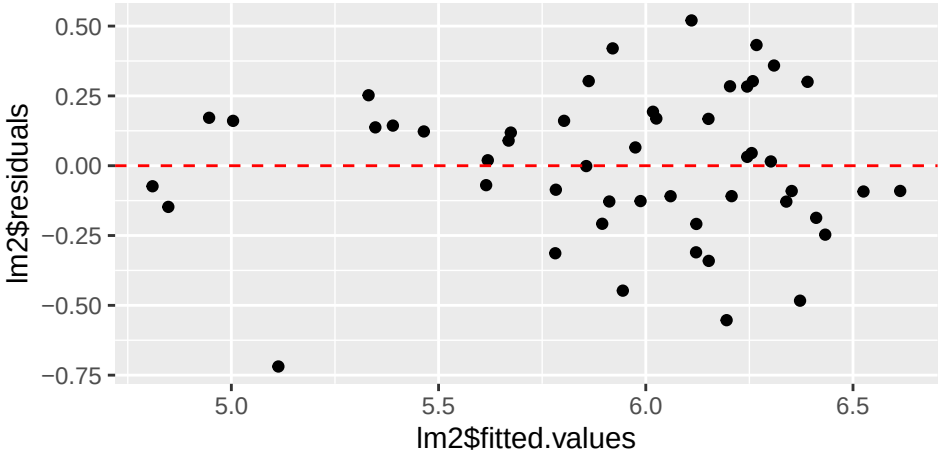
Assumptions Concerning the Residuals - Homoscedasticity

The assumption of **homoscedasticity** is the assumption of constant variance of the residuals, conditional on the regressors. Formally, this assumption is represented as:

$$V(\epsilon_i \mid X_{i1}, ..., X_{ik}) = E(\epsilon_i^2 \mid X_{i1}, ..., X_{ik}) = \sigma_\epsilon^2 \; \forall \; i$$

We can evaluate violation of the constant variance assumption by plotting the residuals against the model-fitted values. We are looking for any evidence that the residuals fan out.

Assumptions Concerning the Residuals - Homoscedasticity



Assumptions Concerning the Residuals - Homoscedasticity

In this figure, heteroscedasticity is difficult to judge, although it does not appear to be a major problem.

There are a couple of numerical tests that we can employ. A caveat, though, is that these tests are sensitive to the normality assumption, which we have not yet evaluated for this model.

Assumptions Concerning the Residuals - Homoscedasticity

```
##
##  studentized Breusch-Pagan test
##
## data:  lm2
## BP = 5.8183, df = 7, p-value = 0.5611
```

```
##
##  Score Test for Heteroskedasticity
##  ---------------------------------
##  Ho: Variance is homogenous
##  Ha: Variance is not homogenous
##
##  Variables: fitted values of lnviolrate
##
##           Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    0.003175691
##  Prob > Chi2   =    0.9550604
```

Assumptions Concerning the Residuals - Homoscedasticity

By these criteria, the residuals appear to be homoscedastic.

Had we concluded otherwise, it would have been wise to use the coeftest() function in the lmtest package to estimate heteroscedasticity-robust standard errors, or what are also known as Huber-White standard errors.

These are appropriate for arbitrary forms of heteroscedasticity. In fact, robust standard errors are generally a good idea in any regression model.

Assumptions Concerning the Residuals - Homoscedasticity

Regular Standard Errors:

```
##
## Call:
## lm(formula = lnviolrate ~ lnpovrate + lncollege + lnblack + lnurban +
##     region, data = state_data00)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71902 -0.12846  0.01743  0.16862  0.52029
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    2.99391    1.39028   2.153  0.037070 *
## lnpovrate      0.36978    0.21257   1.740  0.089263 .
## lncollege      0.08695    0.31832   0.273  0.786063
## lnblack        0.30267    0.06179   4.898 0.0000148 ***
## lnurban        0.26197    0.19016   1.378  0.175605
## regionSouth    0.08426    0.16827   0.501  0.619172
## regionMidwest  0.06778    0.13559   0.500  0.619775
## regionWest     0.49403    0.13273   3.722  0.000582 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2849 on 42 degrees of freedom
## Multiple R-squared:  0.7431, Adjusted R-squared:  0.7002
## F-statistic: 17.35 on 7 and 42 DF,  p-value: 0.0000000001485
```

Assumptions Concerning the Residuals - Homoscedasticity

Robust Standard Errors:

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   2.993911   1.313084  2.2801 0.0277455 *
## lnpovrate     0.369780   0.238888  1.5479 0.1291450
## lncollege     0.086954   0.258589  0.3363 0.7383478
## lnblack       0.302671   0.086472  3.5002 0.0011151 **
## lnurban       0.261973   0.245645  1.0665 0.2923036
## regionSouth   0.084256   0.175483  0.4801 0.6336191
## regionMidwest 0.067775   0.128562  0.5272 0.6008425
## regionWest    0.494029   0.124177  3.9784 0.0002687 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assumptions Concerning the Residuals - Serial Independence

The assumption of serially independent residuals is also known as the assumption of no autocorrelation.

Serial independence is really an assumption about the sampling process. If there is any kind of *clustering* of observations in space or time (e.g., cluster samples, panel data), this will create biased variance estimates.

Assumptions Concerning the Residuals - Serial Independence

Formally, the assumption of serial independence is represented as:

$$E(\epsilon_i \epsilon_j \mid X_{i1}, ..., X_{ik}) = 0 \; \forall \; i \neq j$$

Assumptions Concerning the Residuals - Serial Independence

With state-level data that are cross sectional, there is a distinct possibility of serial (spatial) dependence, because neighboring states might be more like one another than non-contiguous states. Resolving this complexity is beyond the focus of this lecture.

It suffices to say that there is indeed spatial dependence in violent crime rates. It is likely that the inclusion of the regional dummies in the regression model are capable of absorbing much of this spatial dependence, anyhow. We can examine whether this is true by clustering the standard errors by region, as so:

Assumptions Concerning the Residuals - Serial Independence

```
##
## Call:
## lm(formula = lnviolrate ~ lnpovrate + lncollege + lnblack + lnurban +
##     region, data = state_data00)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71902 -0.12846  0.01743  0.16862  0.52029
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    2.99391    1.39028   2.153  0.037070 *
## lnpovrate      0.36978    0.21257   1.740  0.089263 .
## lncollege      0.08695    0.31832   0.273  0.786063
## lnblack        0.30267    0.06179   4.898 0.0000148 ***
## lnurban        0.26197    0.19016   1.378  0.175605
## regionSouth    0.08426    0.16827   0.501  0.619172
## regionMidwest  0.06778    0.13559   0.500  0.619775
## regionWest     0.49403    0.13273   3.722  0.000582 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2849 on 42 degrees of freedom
## Multiple R-squared:  0.7431, Adjusted R-squared:  0.7002
## F-statistic: 17.35 on 7 and 42 DF,  p-value: 0.0000000001485
```

Assumptions Concerning the Residuals - Serial Independence

There does appear to be some spatial dependence here as the errors for different regressors (including the regional dummies) can vary quite a bit if we cluster our errors by region.

Although our primary inferences remain fairly similar, the results suggest a more appropriate model may be one that accounts for spatial dependence in the outcome - the clustering of standard errors is just one way to do this.

There are other options, like spatial autoregression models, that are beyond the scope of this lecture.

Assumptions Concerning the Residuals - Normality

Normality of the residuals is not a required assumption of the linear model.

Even if this assumption is violated, the estimates from the model are still unbiased and efficient.

Rather, approximate normality is only necessary for hypothesis testing because a distributional assumption is made in order to compute p-values for the refression coefficients.

Assumptions Concerning the Residuals - Normality

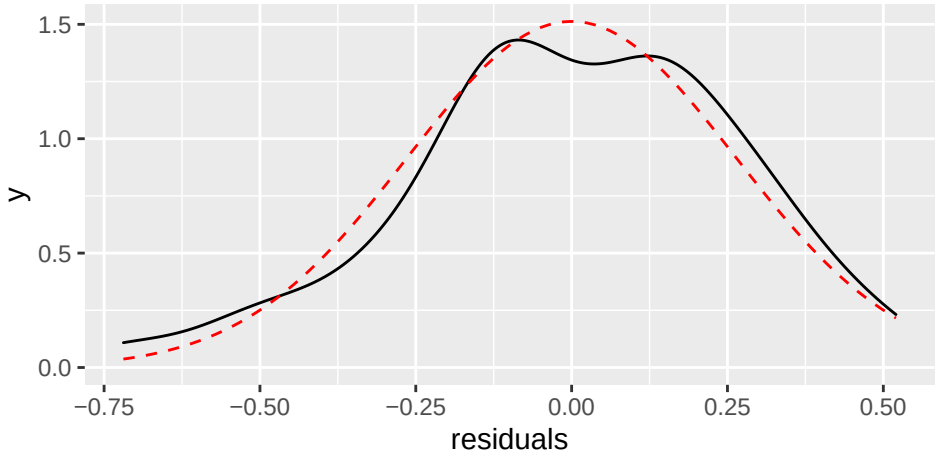
Formally, the assumption is:

$$\epsilon_i \mid X_{i1}, ..., X_{ik} \sim N(0, \sigma_\epsilon^2)$$

In practice, of course, we rely on the t-distribution rather than the z-distribution, with df=k, in order to estimate the p-values.

We can assess normality with a kernel density plot of the residuals, which is like a histogram, but with very narrow bins. The normal density can be overlaid to assess the degree to which the residuals depart from normality:

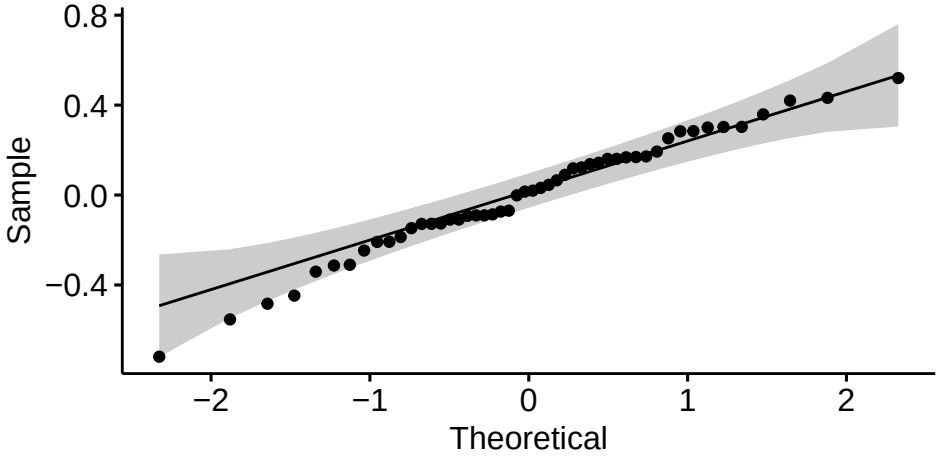
Assumptions Concerning the Residuals - Normality



Assumptions Concerning the Residuals - Normality

As you can see, the log-log model performs quite well with respect to the normality assumption. An alternative to the density plot is a standardized normal probability plot, in which case we would like to see the data points line up along the diagonal:

Assumptions Concerning the Residuals - Normality



Assumptions Concerning the Residuals - Normality

A formal test of the normality of the residuals is provided by the Shapiro-Wilk test (among others):

```
##
##  Shapiro-Wilk normality test
##
## data:  lm2$residuals
## W = 0.9813, p-value = 0.608
```

The test confirms our visual diagnosis of no serious evidence of non-normality.

What are your two questions today?