交通大数据

线性回归

- 李豪杰、郭延永、徐铖铖
- guoyanyong@seu.edu.cn

Linear Regression model

- If a regression model has only two unknown parameters, then it is a binary regression model
- If there are more than two parameters, then it is a multiple regression model

BINARY REGRESSION MODEL

A binary regression model takes the form:

$$y_i = \beta_o + \beta_1 x_{1i} + \varepsilon_i$$

where y = dependent variable (response)

x = independent variable (explanatory variable)

β_o and β_1 = parameters to be estimated

ε = random error or disturbance

In order to estimate parameters, specific assumptions regarding the probability distribution of ε must be made. These assumptions are very basic to any statistical regression analysis.

Assumptions of Linear Regression models

- **Assumptions 1:**

  Continuous Dependent Variable Y

  Y -- Count variables: Poisson and negative binomial regression

  Y -- Nominal scale variables: Discrete outcome models

  Y -- Ordered scale variables: Ordered regression models

- **Assumptions 2:**

  Linear-in-Parameters Relationship between Y and X

  $$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

  $$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

Assumptions of Linear Regression models

- **Assumptions 3:**

Observations Independently and Randomly Sampled

Independence requires that the probability that an observation is selected is unaffected by other observations selected into the sample.

Other sampling schemes such as stratified and cluster samples can be accommodated in the regression modeling framework with **corrective measures**

Assumptions of Linear Regression models

- **Assumptions 4:**

Uncertain Relationship between Variables

$$Y_i = \beta_0 + \beta_1 X_{1i} \boxed{+ \varepsilon_i}$$

Variables that were **omitted** from the model

Measurement errors in the dependent variable, or the imprecision in measuring Y.

Random variation inherent in the underlying data-generating process.

Assumptions of Linear Regression models

- **Assumptions 5:**

Disturbance Independent across observations and Expected Value Zero

$$E[\varepsilon_i] = 0$$

$$VAR[\varepsilon_i] = \sigma^2$$

Homoscedasticity -- Variance of ε_i is **independent** across observations

- **Assumptions 6:**

Disturbance Terms Not Autocorrelated

$$COV[\varepsilon_i, \varepsilon_j] = 0 \text{ if } i \neq j$$

Disturbances are **independent** across observations

Violations – Repeated observations, Observations across time

Assumptions of Linear Regression models

- **Assumptions 7:**

Independent Variables and Disturbances **Uncorrelated**

$$COV\left[X_i, \varepsilon_j\right] = 0 \text{ for all } i \text{ and } j$$

Exogeneity -- the values of X_i are determined by influences outside of the model. Y does not directly influence the value of an exogenous independent variables.

Endogeneity -- the values of X_i are determined by influences inside of the model. Y can be consider an **endogenous** variable.

Assumptions of Linear Regression models

- **Assumptions 8:**

Disturbances Approximately **Normally** Distributed

$$\varepsilon \sim N\left(0, \sigma^2 I_n\right)$$

Combined with the assumption 3, disturbances are **independently** and **identically** distributed as **normal** (i.i.d. normal)

$$\mathbf{Y}_i \approx N\left(\mathbf{X}'\boldsymbol{\beta}, \sigma^2\right)$$

Assumptions of Linear Regression models

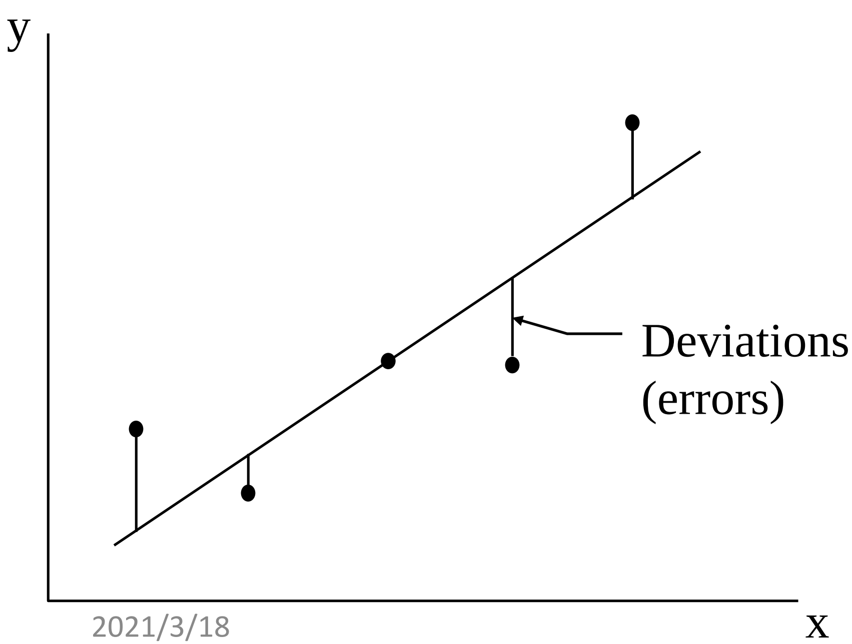
Summary of Ordinary Least Squares Linear Regression Model Assumptions

	Statistical Assumption	Mathematical Expression
1.	Functional form	$Y_i = \beta_0 + \beta_1 X_{1i} + e_i$
2.	Zero mean of disturbances	$E[\varepsilon_i] = 0$
3.	Homoscedasticity of disturbances	$VAR[\varepsilon_i] = \sigma^2$
4.	Nonautocorrelation of disturbances	$COV[\varepsilon_i, \varepsilon_j] = 0$ if $i \neq j$
5.	Uncorrelatedness of regressor and disturbances	$COV\ [X_i, \varepsilon_j] = 0$ for all i and j
6.	Normality of disturbances	$\varepsilon_i \approx N(0, \sigma^2)$
	Continuous Dependent Variable *Y*	

There are techniques to check the validity of the assumptions and remedies available that can be applied if the assumptions are violated in a model estimation effort. These will be discussed later

Least Squares Estimation

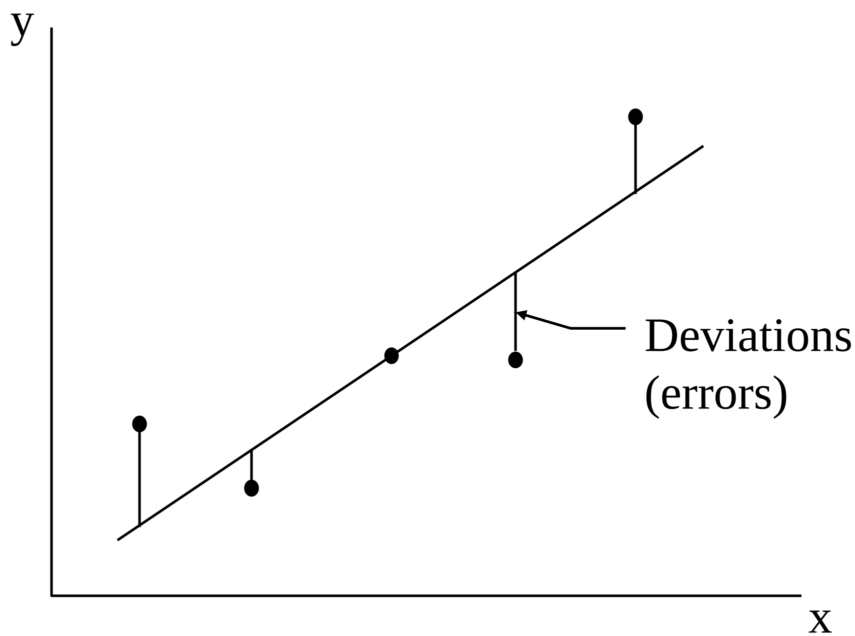
In the figure, the vertical line segments represent deviations of the observation points from the line. One can find many lines for the sum of deviations (errors) is equal to zero. But it can be shown that there is one and only one line for which the SUM of SQUARED DEVIATIONS is a minimum. This sum is called the sum of squared errors (SSE).



$$y = \beta_o + \beta_1 x + \varepsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least Squares Estimation



$$y = \beta_o + \beta_1 x + \varepsilon$$

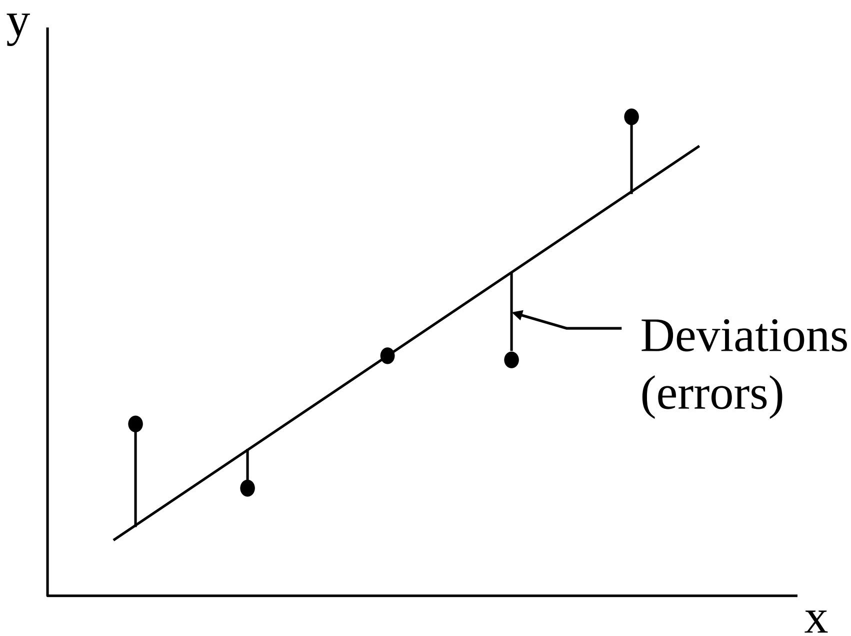
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Thus, $\hat{y}$ is an estimator of the mean value of y, E(y), and a predictor of some future value of y. $\hat{\beta}_o$ and $\hat{\beta}_1$ are least squares estimators of β_o and β_1 respectively.

For a given data point, (x_i, y_i), the observed value of y is y_i and the predicted value of y would be obtained by substituting x_i into the prediction equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least Squares Estimation



$$y = \beta_o + \beta_1 x + \varepsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The deviation of the ith value of y from its predicted value is:

$$(y_i - \hat{y}_i) = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

Then, the sum of squares of the deviations of the y values about their predicted values for all of the observations (n data points):

$$SSE = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Least Squares Estimation

OLS seeks a solution that minimizes the function *Q:*

$$Q_{\min} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2_{\min} = \sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2_{\min} = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_i\right)^2_{\min}$$

By setting the partial derivatives of *Q* with respect to β_o and β_1 equal to zero, the least squares estimated parameters β_o and β_1 are obtained:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_i\right) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i\left(Y_i - \beta_0 - \beta_1 X_i\right) = 0$$

Least Squares Estimation

Solving these equations using B_0 and B_1 to denote the estimates of β_o and β_1 , respectively, and rearranging terms yields:

$$\sum_{i=1}^{n} Y_i = nB_0 + B_1 \sum_{i=1}^{n} X_i \,, \quad \sum_{i=1}^{n} X_i Y_i = B_0 \sum_{i=1}^{n} X_i + B_1 \sum_{i=1}^{n} X_i^2$$

Solving simultaneously for the betas yields:

$$B_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \,, \quad B_0 = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - B_1 \sum_{i=1}^{n} X_i\right) = \overline{Y} - B_1 \overline{X}$$

Least Squares Estimation

A database consisting of 121 observations is available to study annual average daily traffic (*AADT*) in Minnesota. Variables available in this database, and their abbreviations, are provided in Table 3.2. A simple regression model is estimated using least squares estimated parameters as a starting point. The starter specification is based on a model that has *AADT* as a function of *CNTYPOP*, *NUMLANES*, and *FUNCTIONALCLASS*: $AADT = \beta_0 + \beta_1(CNTYPOP) + \beta_2(NUMLANES) + \beta_3(FUNCTIONALCLASS)$ + disturbance.

Variables Collected on Minnesota Roadways

Variable No.	*Abbreviation*: Variable Description
1	*AADT*: Average annual daily traffic in vehicles per day
2	*CNTYPOP*: Population of county in which road section is located (proxy for nearby population density)
3	*NUMLANES*: Number of lanes in road section
4	*WIDTHLANES*: Width of road section in feet
5	*ACCESSCONTROL*: 1 for access controlled facility; 2 for no access control
6	*FUNCTIONALCLASS*: Road sectional functional classifications; 1 = rural interstate, 2 = rural non-interstate, 3 = urban interstate, 4 = urban non-interstate
7	*TRUCKROUTE*: Truck restriction conditions: 1 = no truck restrictions, 2 = tonnage restrictions, 3 = time of day restrictions, 4 = tonnage and time of day restrictions, 5 = no trucks
8	*LOCALE*: Land-use designation: 1 = rural, 2 = urban with population ≤ 50,000, 3 = urban with population > 50,000

Least Squares Estimation

Least Squares Estimated Parameters (Example 3.1)

Parameter	Parameter Estimate
Intercept	–26234.490
CNTYPOP	0.029
NUMLANES	9953.676
FUNCLASS1	885.384
FUNCLASS2	4127.560
FUNCLASS3	35453.679

- For each additional 1000 people in the local county population, there is an estimated 29 additional AADT.
- For each lane there is an estimated 9954 AADT.
- Urban interstates are associated with an estimated 35,454 AADT more than non-interstates

Maximum Likelihood Estimation

- Another popular and sometimes useful statistical estimation method is called maximum likelihood estimation, which results in the maximum likelihood estimates, or **MLEs**. The probability of each observation is given as:

$$\mathbf{Y}_i \approx N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\right) \quad f\left(x_i, \boldsymbol{\theta}\right) = P\left(y_i\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(y_i - \boldsymbol{X}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right)$$

- Likelihood function is the joint density of observing the sample data from a statistical distribution with parameter vector

$$f\left(x_1, x_2, \ldots, x_n, \boldsymbol{\theta}\right) = \prod_{i=1}^{n} f\left(x_i, \boldsymbol{\theta}\right) = L(\boldsymbol{\theta} \mid \mathbf{X})$$

Maximum Likelihood Estimation

- For the regression model, the likelihood function for a sample of n independent, identically, and normally distributed disturbances is given by:

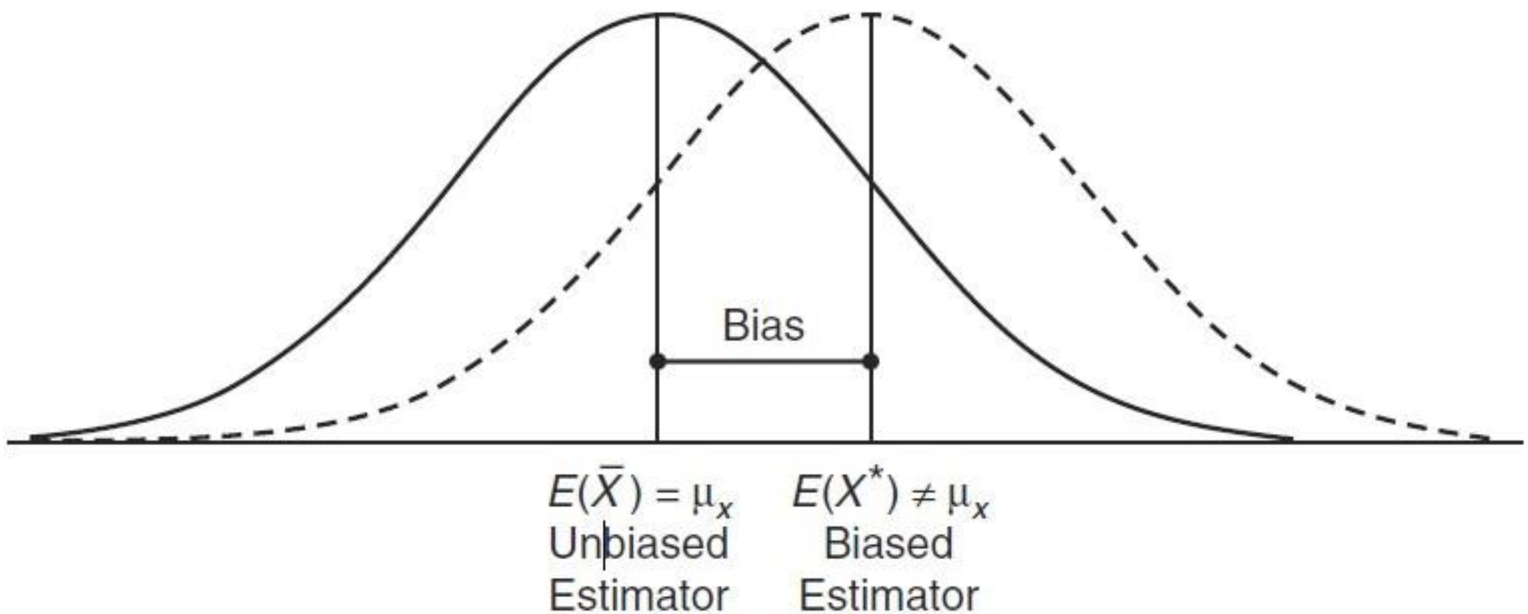
$$L=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}EXP\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-X_i^T\beta\right)^2\right]=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}EXP\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right]$$

- As is usually the case, the logarithm of Equation, or the log likelihood, is simpler to solve than the likelihood function itself, so taking the log of L yields:

$$LN(L)=LL=-\frac{n}{2}LN(2\pi)-\frac{n}{2}LN\left(\sigma^2\right)-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

Properties of OLS and MLE Estimators

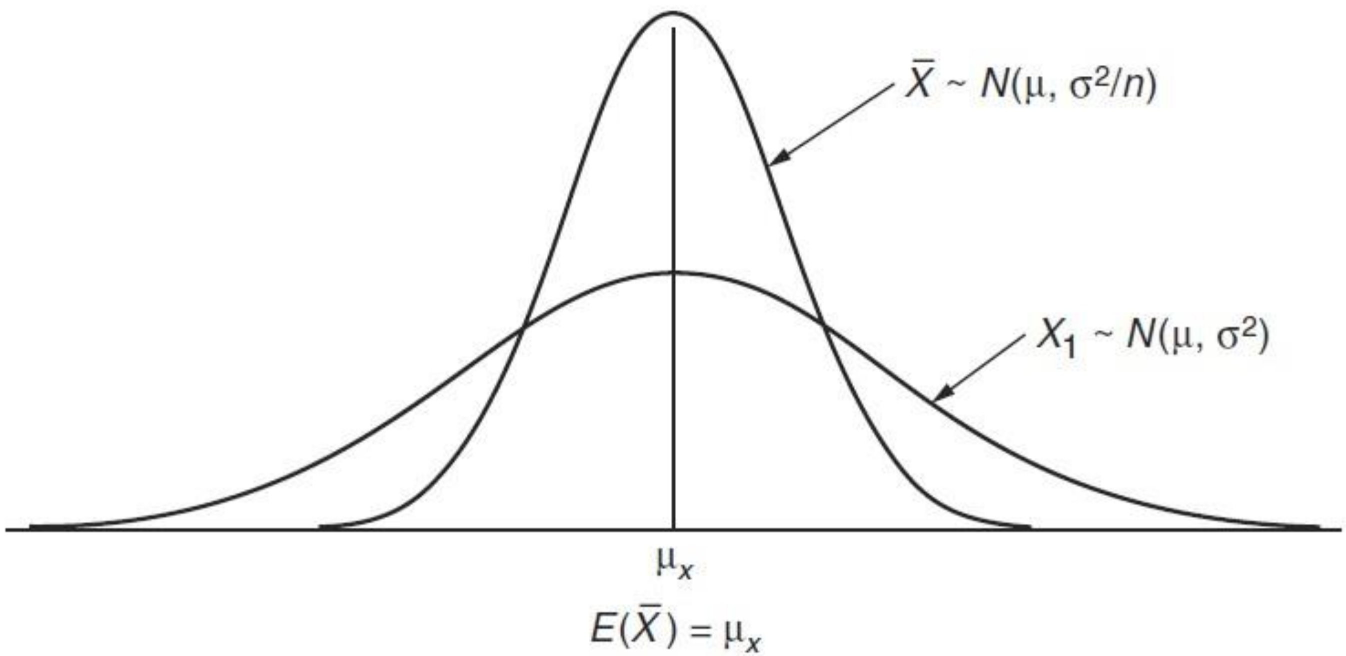
- The MLE and OLS estimators are **unbiased** estimators of the betas



If there are several estimators of a population parameter, and if one of these estimators coincides with the true value of the unknown parameter, then this estimator is called an unbiased estimator.

Properties of OLS and MLE Estimators

- The MLE and OLS estimators are **Efficiency** estimators of the betas



Efficiency is a relative property in that an estimator is efficient relative to another, which means that an estimator has a smaller variance than an alternative estimator.

Inference in Regression Analysis

- The sampling distribution of B_1 is approximately normal such that

$$B_1 \approx N\left(\beta_1, \frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}\right)$$

- Population variance is typically unknown, **MSE** is an estimate of the variance in the regression model and is given as

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-p}$$

Inference in Regression Analysis

- The following t distribution

$$t^* = \frac{B_k - \beta_k}{\sqrt{\frac{MSE}{\sum(X_i - \overline{X})^2}}} = \frac{B_k - \beta_k}{s\{B_k\}} \approx t(\alpha; n-p), \qquad (3.30)$$

where α is the level of significance and $n - p$ is the associated degrees of freedom. This is an important result; it enables a statistical test of the probabilistic evidence in favor of specific values of β_k.

- A confidence interval for the parameter β_1 is given by

$$B_k \pm t\left(1 - \frac{\alpha}{2}; n-p\right)s\{B_k\}$$

Inference in Regression Analysis

- Consider again the study of AADT in Minnesota. Interest centers on the development of confidence intervals and hypothesis tests on some of the parameters estimated in Example 3.1

Least Squares Estimated Parameters (Example 3.2)

Parameter	Parameter Estimate	Standard Error of Estimate	t-value	$P(>\|t\|)$
Intercept	–26234.490	4935.656	–5.314	<0.0001
CNTYPOP	0.029	0.005	5.994	<0.0001
NUMLANES	9953.676	1375.433	7.231	<0.0001
FUNCLASS1	885.384	5829.987	0.152	0.879
FUNCLASS2	4127.560	3345.418	1.233	0.220
FUNCLASS3	35453.679	4530.652	7.825	<0.0001

Manipulating Variables in Regression

- **Standardized** regression models allow for direct comparison of the relative importance of independent variables
- Often interest is focused on the **relative impacts** of independent variables on the response variable Y
- Two independent variables in a model describing the *expected number of daily trip-chains* were *number of children* and *household income*
- Households may have children ranging from 0 to 8, while income may range from $5000 to $500,000, making it difficult to make a useful comparison between the relative impact of these variables

Manipulating Variables in Regression

- The standardized regression model is obtained by **standardizing** all **independent variables**.

$$X_1' = \frac{X_1 - \overline{X}}{s\{X_1\}}$$

- The standardized variables are created with expected values equal to 0 and variances equal to 1.

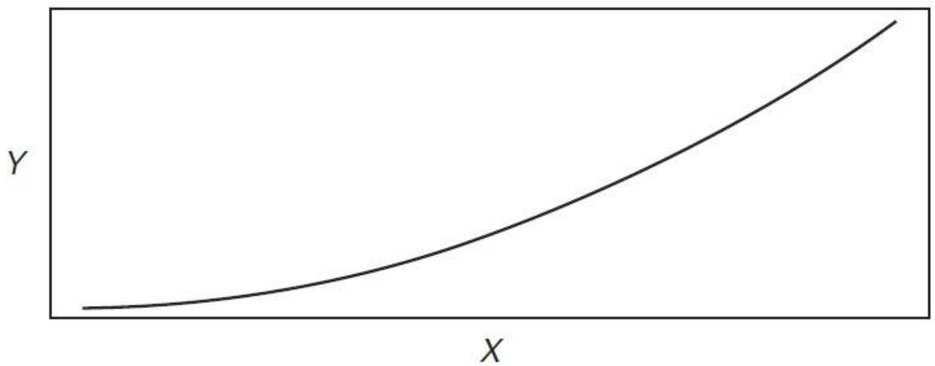
Manipulating Variables in Regression

- **Nonlinear relationships** can be accommodated within the linear regression framework by **Transformations**

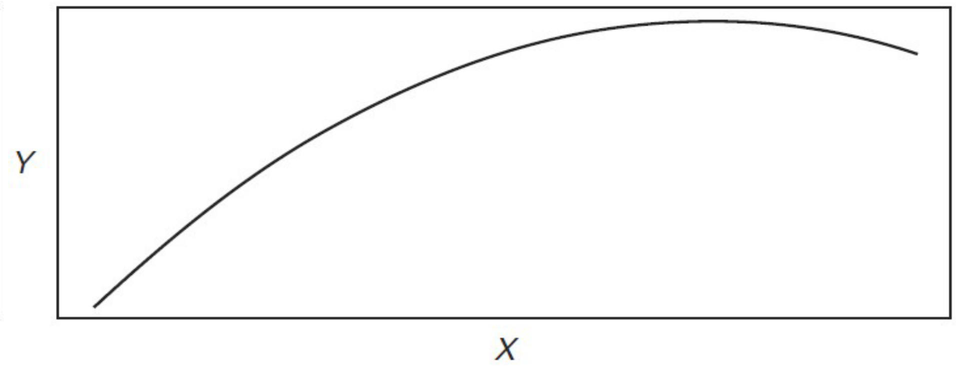
$$\hat{Y} = \frac{X}{B_0 X + B_1 + eX} = \frac{X}{\alpha X + \lambda + Xe} \quad \Rightarrow E\left[\frac{1}{\hat{Y}}\right] = B_0 + B_1 \frac{1}{X} + e$$

$$\hat{Y} = EXP^{B_0} EXP^{\frac{B_1}{X}} EXP^{e} \quad \Rightarrow \quad LN\left(\hat{Y}\right) = B_0 + \frac{B_1}{X} + e$$

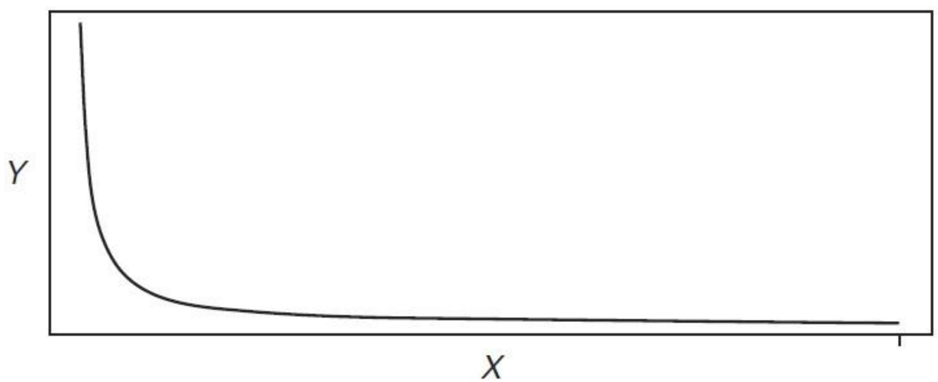
Manipulating Variables in Regression



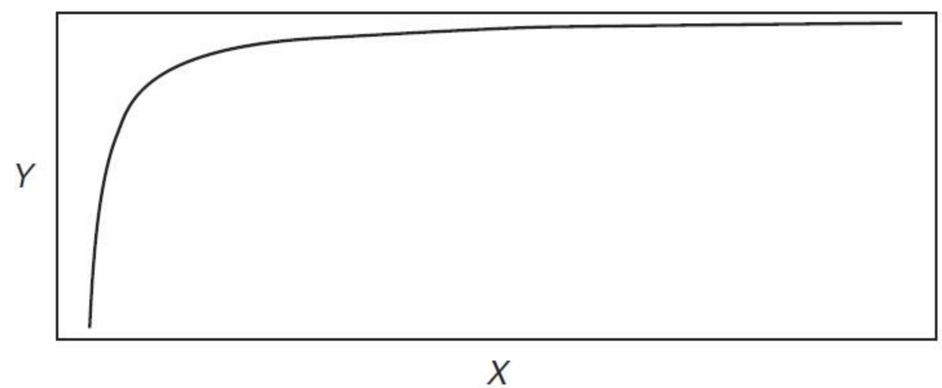
$Y = \alpha + \beta X + \gamma X^2$, where $\alpha > 0, \beta > 0, \gamma > 0$



$Y = \alpha + \beta X + \gamma X^2$, where $\alpha > 0, \beta > 0, \gamma < 0$

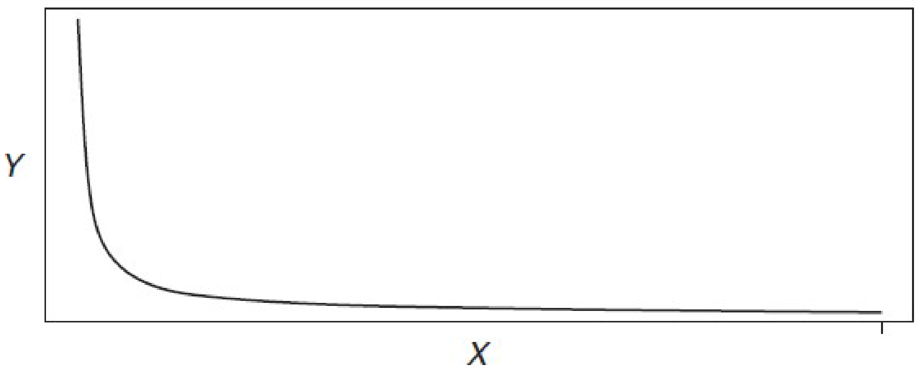


$Y = X/(\alpha + \beta X)$, where $\alpha > 0$

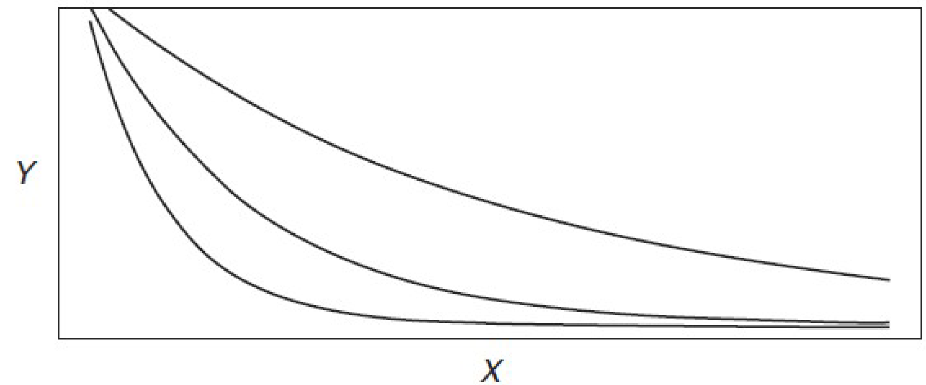


$Y = X/(\alpha + \beta X)$, where $\alpha < 0$

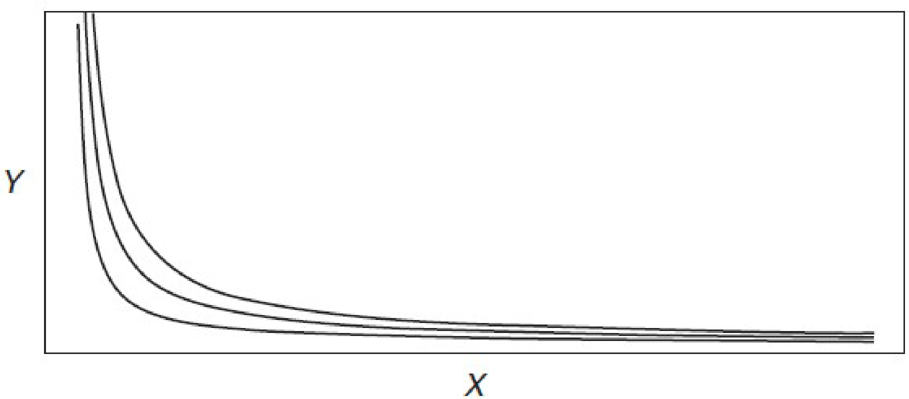
Manipulating Variables in Regression



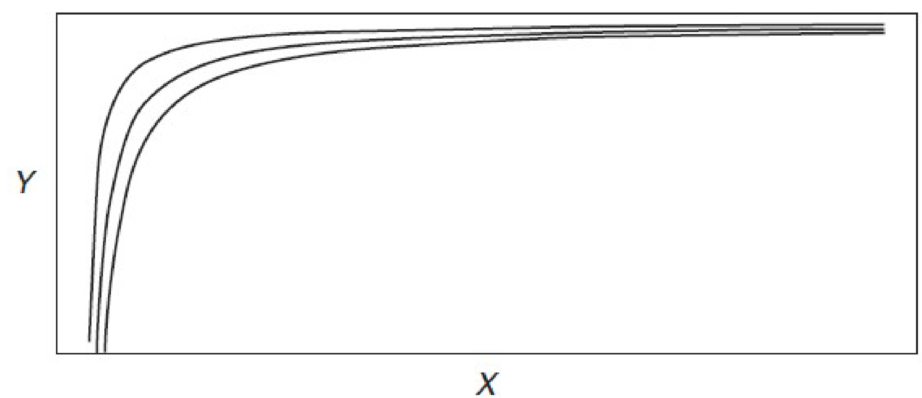
$Y = \alpha EXP(\beta X)$, where $\beta > 0$



$Y = \alpha EXP(\beta X)$, where $\beta < 0$

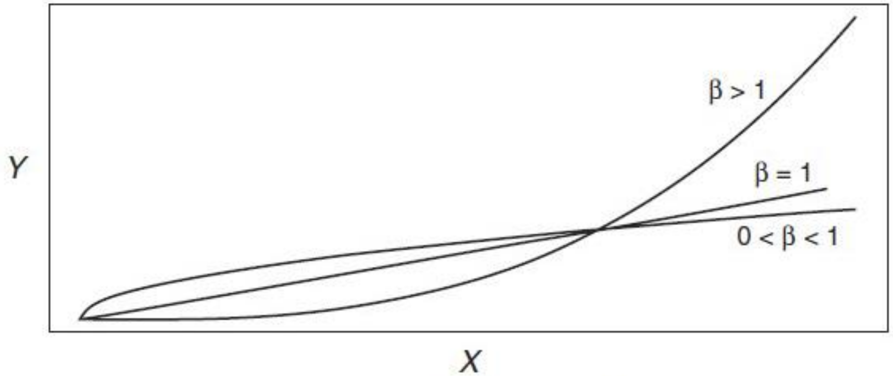


$Y = \alpha EXP(\beta / X)$, where $\beta > 0$

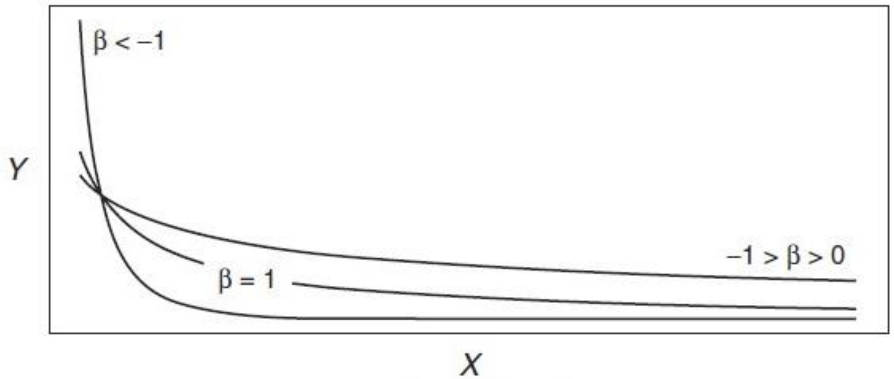


$Y = \alpha EXP(\beta / X)$, where $\beta < 0$

Manipulating Variables in Regression



Y = αXβ, where β > 0



Y = αXβ, where β < 0

Manipulating Variables in Regression

- Ordinal and nominal scale variables need transformed into indicator Variables
- Qualitative variables such as *roadway functional class*, *gender*, *attitude toward transit*, and *trip purpose*
- For **nominal scale variables**, **m-1 indicator variables** must be created to represent all **m levels** of the variable in the regression model.
- **Ordinal scale variables**, unlike nominal scale variables, are ranked, and can complicate matters in the regression. Several methods for dealing with ordinal scale variables in the regression model

Manipulating Variables in Regression

- Estimate a Single Beta Parameter: The assumption in this approach is that the marginal effect is equivalent across increasing levels of the variable
- Estimate Beta Parameter for Ranges of the Variable: Consider the variable NUMLANES used in previous chapter examples. Although the intervals between levels of this variable are equivalent, it may be believed that a fundamentally different effect on AADT exists for different levels of NUMLANES.

$$Ind_1 = \begin{cases} NUMLANES \text{ if } 1 \le NUMLANES \le 2 \\ 0 \text{ otherwise} \end{cases} \qquad Ind_2 = \begin{cases} NUMLANES \text{ if } NUMLANES > 2 \\ 0 \text{ otherwise} \end{cases}$$

- These two indicator variables would allow the estimation of two parameters, one for the lower range of the variable NUMLANES and one for the upper range

Manipulating Variables in Regression

- Estimate a Single Beta Parameter for **m-1** of the **m Levels** of the Variable: The justification for this approach is that each level of the variable has a unique marginal effect on the response

$$Ind_1 = \begin{cases} 1 \text{ if } NUMLANES = 1 \\ 0 \text{ otherwise} \end{cases}$$

$$Ind_2 = \begin{cases} 1 \text{ if } NUMLANES = 2 \\ 0 \text{ otherwise} \end{cases}$$

$$Ind_3 = \begin{cases} 1 \text{ if } NUMLANES = 3 \\ 0 \text{ otherwise} \end{cases}$$

Manipulating Variables in Regression

- **Interactions** in regression models represent a **combined** or **synergistic** effect of two or more variables. That is, the response variable depends on the joint values of two or more variables.

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 .$$

$$\text{level 1:} \quad X_2 = X_3 = X_4 = 0 \qquad \hat{Y} = B_0 + B_1X_1$$

$$\text{level 2:} \quad X_2 = 1 \qquad \hat{Y} = B_0 + B_1X_1 + B_2X_2 = (B_0 + B_2) + B_1X_1$$

$$\text{level 3:} \quad X_3 = 1 \qquad \hat{Y} = B_0 + B_1X_1 + B_3X_3 = (B_0 + B_3) + B_1X_1$$

$$\text{level 4:} \quad X_4 = 1 \qquad \hat{Y} = B_0 + B_1X_1 + B_4X_4 = (B_0 + B_4) + B_1X_1 .$$

Manipulating Variables in Regression

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 .$$

$$\text{level 1:} \quad X_2 = X_3 = X_4 = 0 \qquad \hat{Y} = B_0 + B_1X_1$$

$$\text{level 2:} \quad X_2 = 1 \qquad \hat{Y} = B_0 + B_1X_1 + B_2X_2 = (B_0 + B_2) + B_1X_1$$

$$\text{level 3:} \quad X_3 = 1 \qquad \hat{Y} = B_0 + B_1X_1 + B_3X_3 = (B_0 + B_3) + B_1X_1$$

$$\text{level 4:} \quad X_4 = 1 \qquad \hat{Y} = B_0 + B_1X_1 + B_4X_4 = (B_0 + B_4) + B_1X_1 .$$

- Depending on which of the indicator variables is coded as 1, the slope of the regression line with respect to **X1 remains fixed**, while the **Y-intercept** parameter changes by the amount of the parameter of the indicator variable.

Manipulating Variables in Regression

- Suppose that each indicator variable is thought to interact with the variable X1. That is, each level of the indicator variable has a unique effect on Y when interacted with X1.

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_2X_1 + B_6X_3X_1 + B_7X_4X_1$$

level 1: $X_2 = X_3 = X_4 = 0$ $\quad \hat{Y} = B_0 + B_1X_1$

level 2: $X_2 = 1$ $\quad \hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_5X_2X_1 = (B_0 + B_2) + (B_1 + B_5)X_1$

level 3: $X_3 = 1$ $\quad \hat{Y} = B_0 + B_1X_1 + B_3X_3 + B_6X_3X_1 = (B_0 + B_3) + (B_1 + B_6)X_1$

level 4: $X_4 = 1$ $\quad \hat{Y} = B_0 + B_1X_1 + B_4X_4 + B_7X_4X_1 = (B_0 + B_4) + (B_1 + B_7)X_1$

Manipulating Variables in Regression

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_2X_1 + B_6X_3X_1 + B_7X_4X_1$$

level 1: $X_2 = X_3 = X_4 = 0$ $\quad \hat{Y} = B_0 + B_1X_1$

level 2: $X_2 = 1 \quad \hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_5X_2X_1 = (B_0 + B_2) + (B_1 + B_5)X_1$

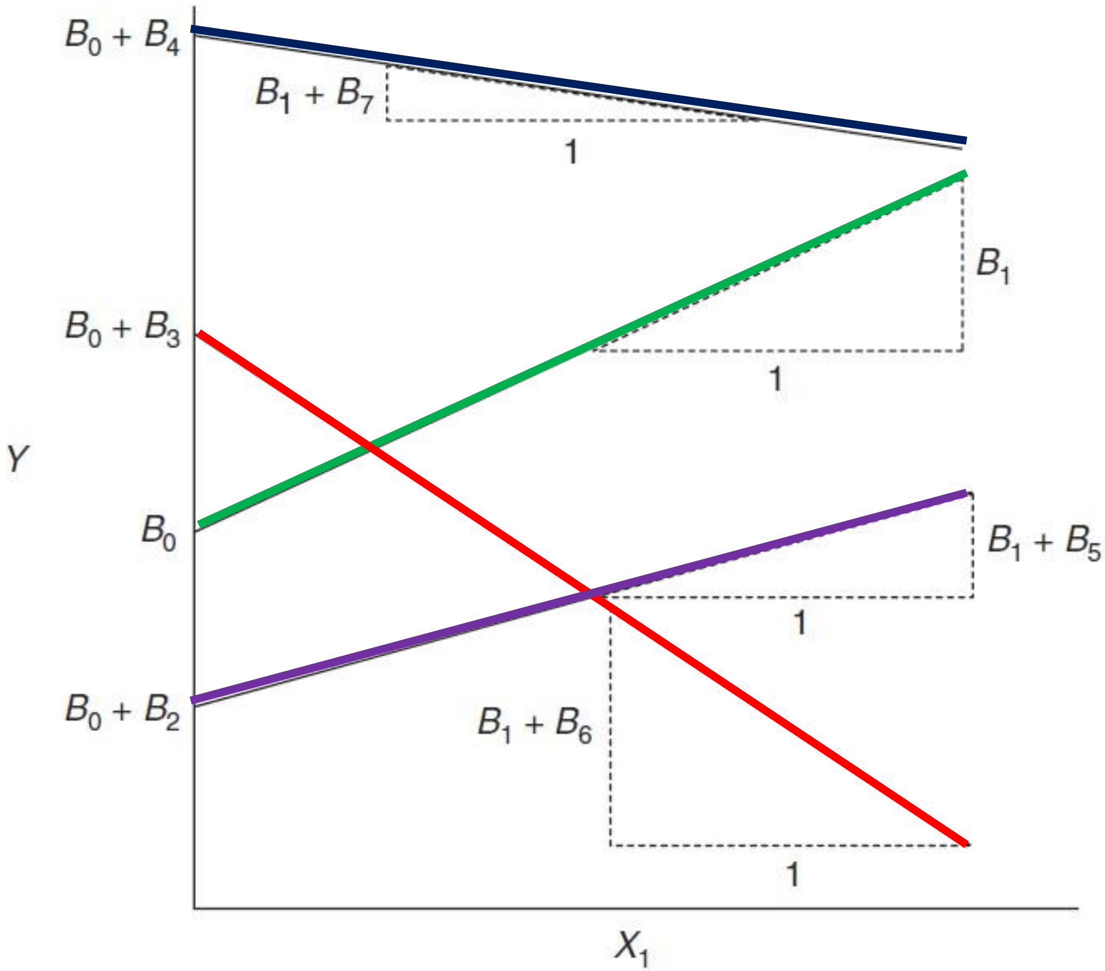
level 3: $X_3 = 1 \quad \hat{Y} = B_0 + B_1X_1 + B_3X_3 + B_6X_3X_1 = (B_0 + B_3) + (B_1 + B_6)X_1$

level 4: $X_4 = 1 \quad \hat{Y} = B_0 + B_1X_1 + B_4X_4 + B_7X_4X_1 = (B_0 + B_4) + (B_1 + B_7)X_1$

- Each level of the indicator variable now has an effect on both the **Y-intercept** and **slope** of the regression function with respect to the variable X1.

Manipulating Variables in Regression

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_2X_1 + B_6X_3X_1 + B_7X_4X_1$$



Manipulating Variables in Regression

- When interactions are between two or more continuous variables the regression equation becomes more complicated.

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_1X_2$$

- The regression function indicates that the **relationship between X_1 and Y** is dependent on the value of **X_2**, and conversely, that the **relation between X_2 and Y** is dependent on the value of **X_1**.

- There are **numerous cases** when the effect of one variable on Y depends on the value of one or more independent variables.

- Although they **may not explain** a great deal of the **variability** in the response, they may represent **important theoretical** aspects of the relation and should be **included** in the regression.

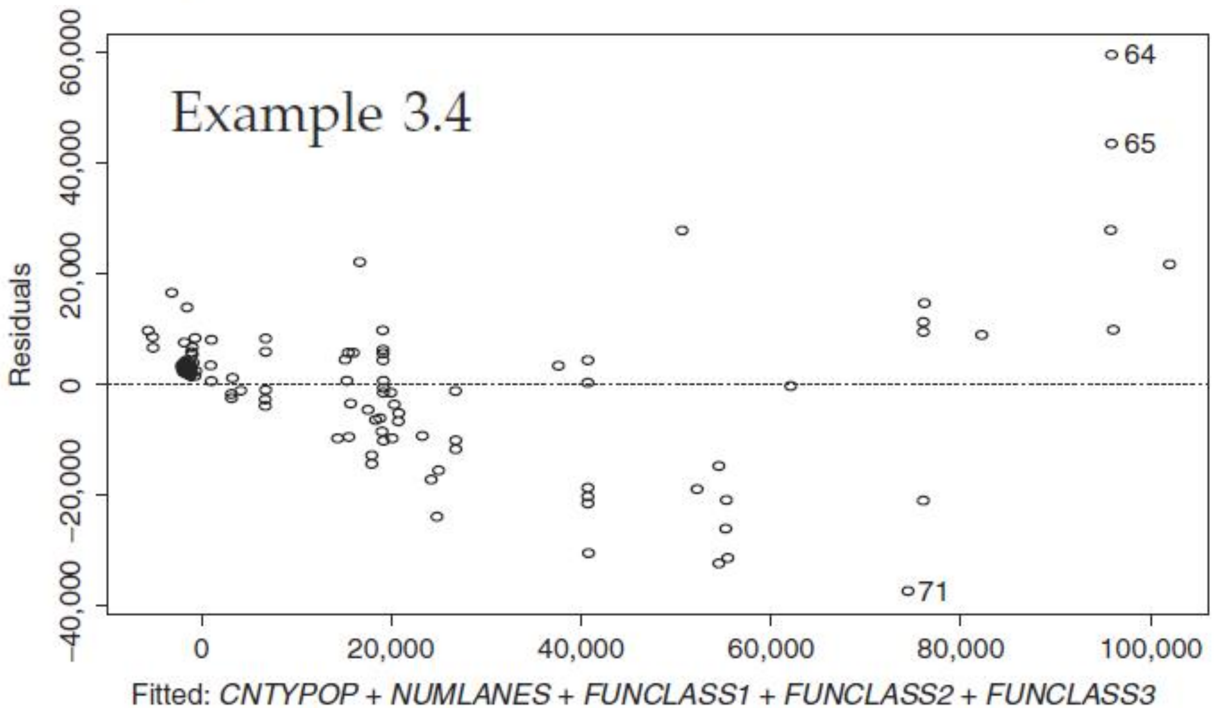
Checking Regression Assumptions

Summary of Ordinary Least Squares Linear Regression Model Assumptions

	Statistical Assumption	Mathematical Expression
1.	Functional form	$Y_i = \beta_0 + \beta_1 X_{1i} + e_i$
2.	Zero mean of disturbances	$E[\varepsilon_i] = 0$
3.	Homoscedasticity of disturbances	$VAR[\varepsilon_i] = \sigma^2$
4.	Nonautocorrelation of disturbances	$COV[\varepsilon_i, \varepsilon_j] = 0$ if $i \neq j$
5.	Uncorrelatedness of regressor and disturbances	$COV[X_i, \varepsilon_j] = 0$ for all i and j
6.	Normality of disturbances	$\varepsilon_i \approx N(0, \sigma^2)$
	Continuous Dependent Variable *Y*	

Checking Regression Assumptions

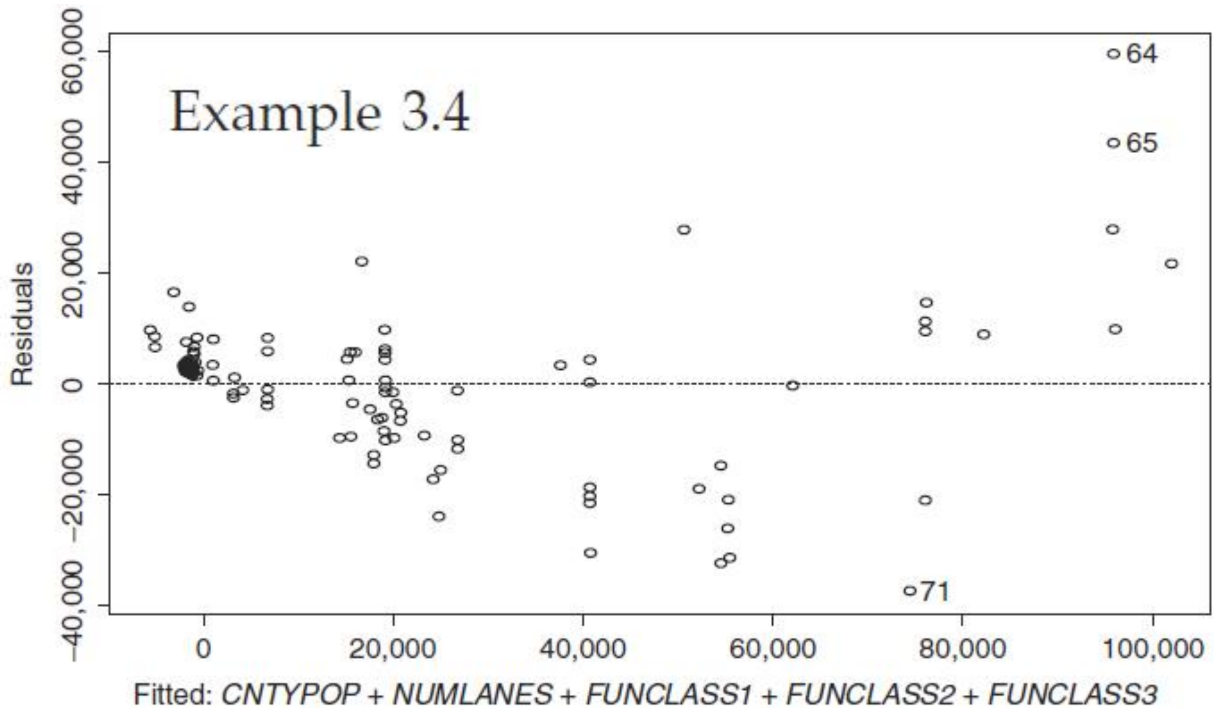
- **Linearity** is checked informally using several plots:
  - plots of **independent variables** on X-axis vs. **disturbances** on the Y-axis
  - plots of model **predicted values** on the X-axis vs. **disturbances** on the Y-axis.



NUMLANES

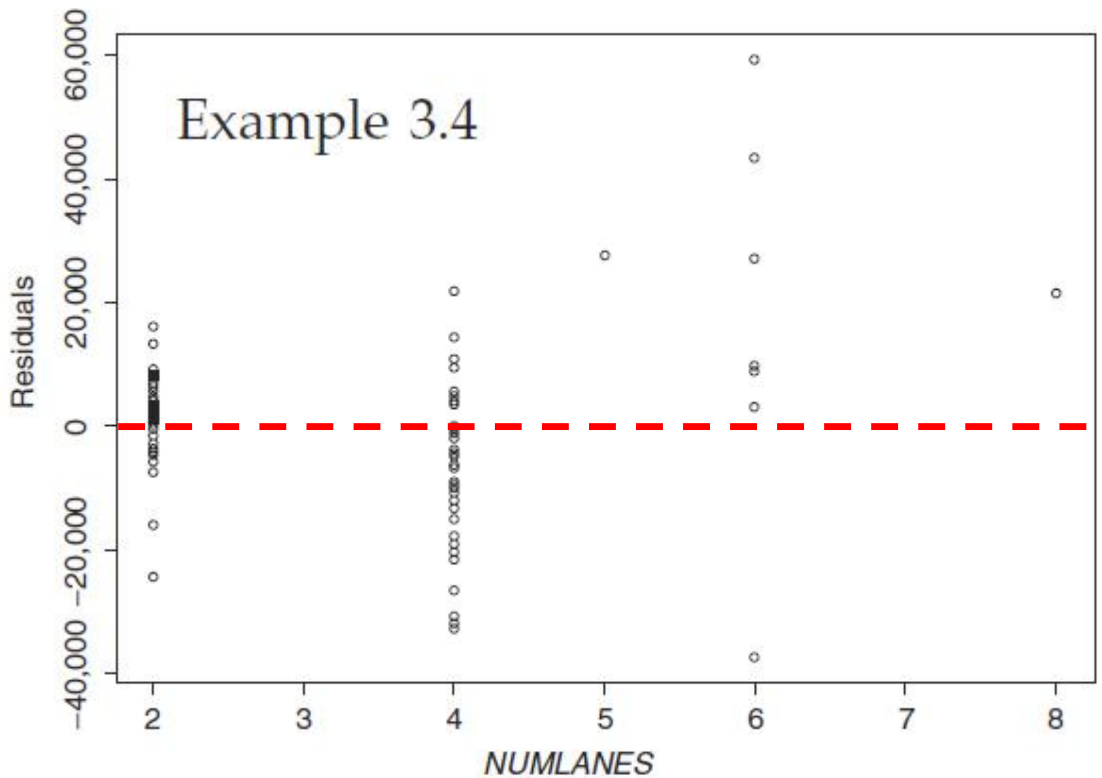
Checking Regression Assumptions

- The **U-shape** of disturbances is a clear indicator of one or more nonlinear effects in the regression.



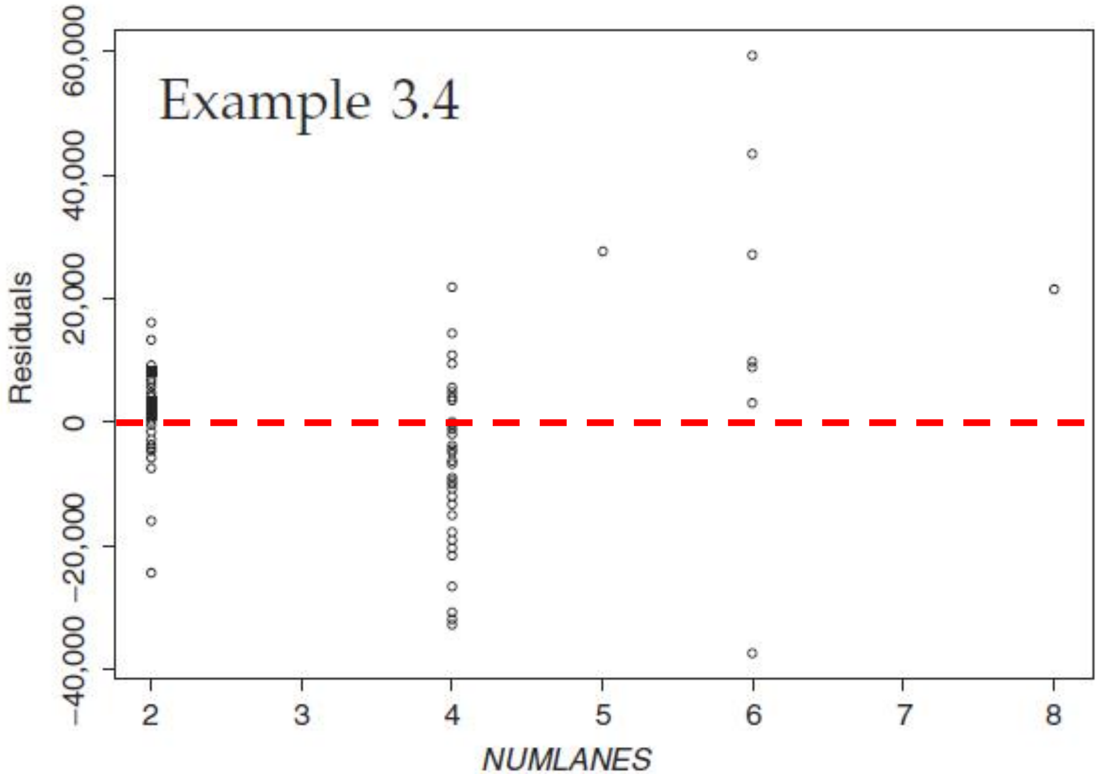
Checking Regression Assumptions

- To determine which independent variable(s) is contributing to the **nonlinearity**, plots of individual **independent variables** vs. the model **disturbances** are constructed and examined



Checking Regression Assumptions

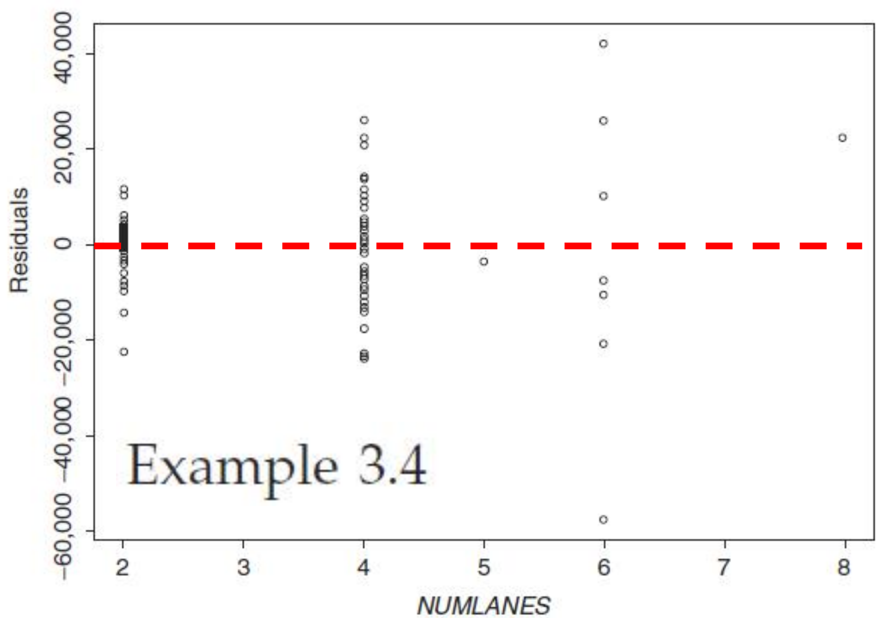
- The nonlinearity between NUMLANES and model disturbances suggests that treating the variable NUMLANES as a continuous variable may not be the best specification of this variable.



Checking Regression Assumptions

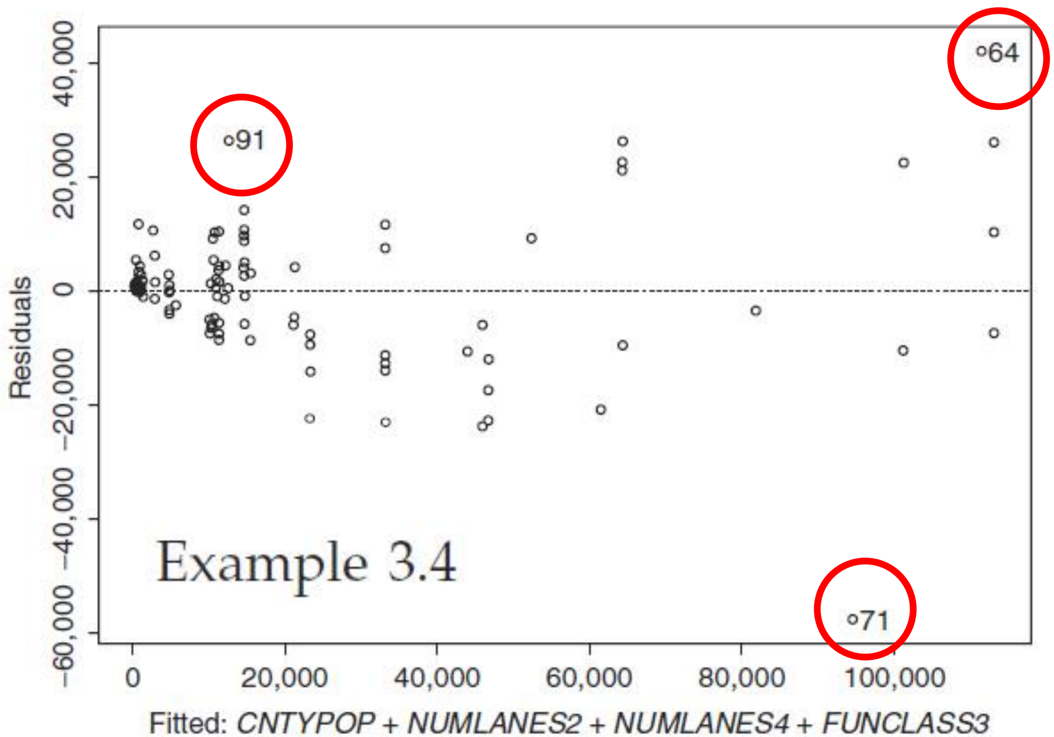
Least Squares Estimated Parameters (Example 3.2)

Parameter	Parameter Estimate	Standard Error of Estimate	*t*-value	*P*(> |*t*|)
Intercept	58698.910	5099.605	11.510	<0.0001
CNTYPOP	0.025	0.004	5.859	<0.0001
NUMLANES2	–58718.141	5134.858	–11.435	<0.0001
NUMLANES4	–48867.728	4685.006	–10.431	<0.0001
FUNCLASS3	31349.211	3689.281	8.497	<0.0001



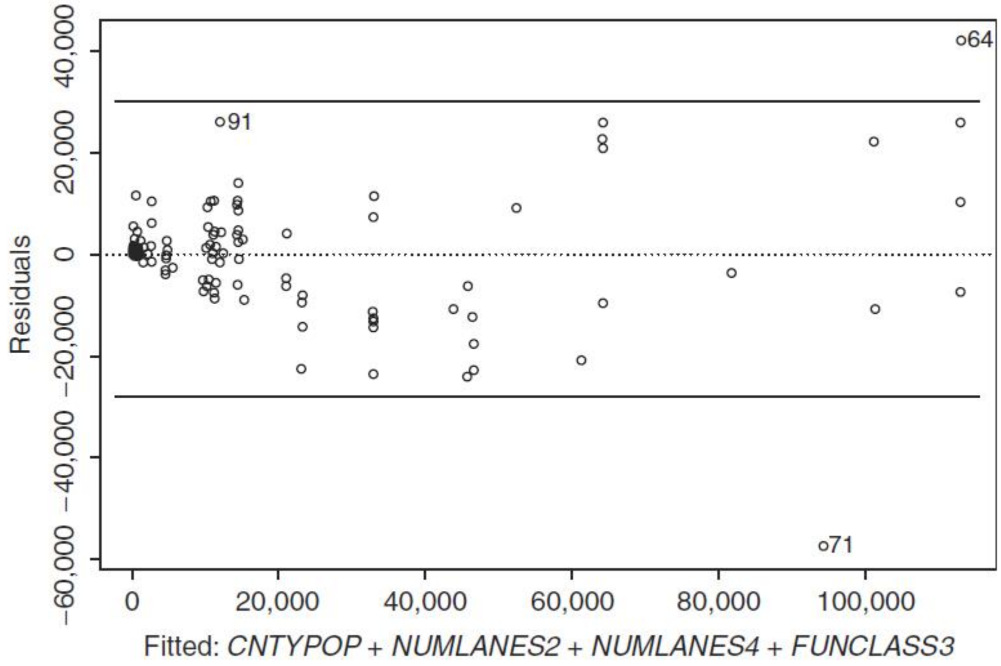
Checking Regression Assumptions

- Although there is still a slight U-shaped pattern to the disturbances, the effect is not as pronounced and both positive and negative disturbances can be found along the entire fitted regression line..



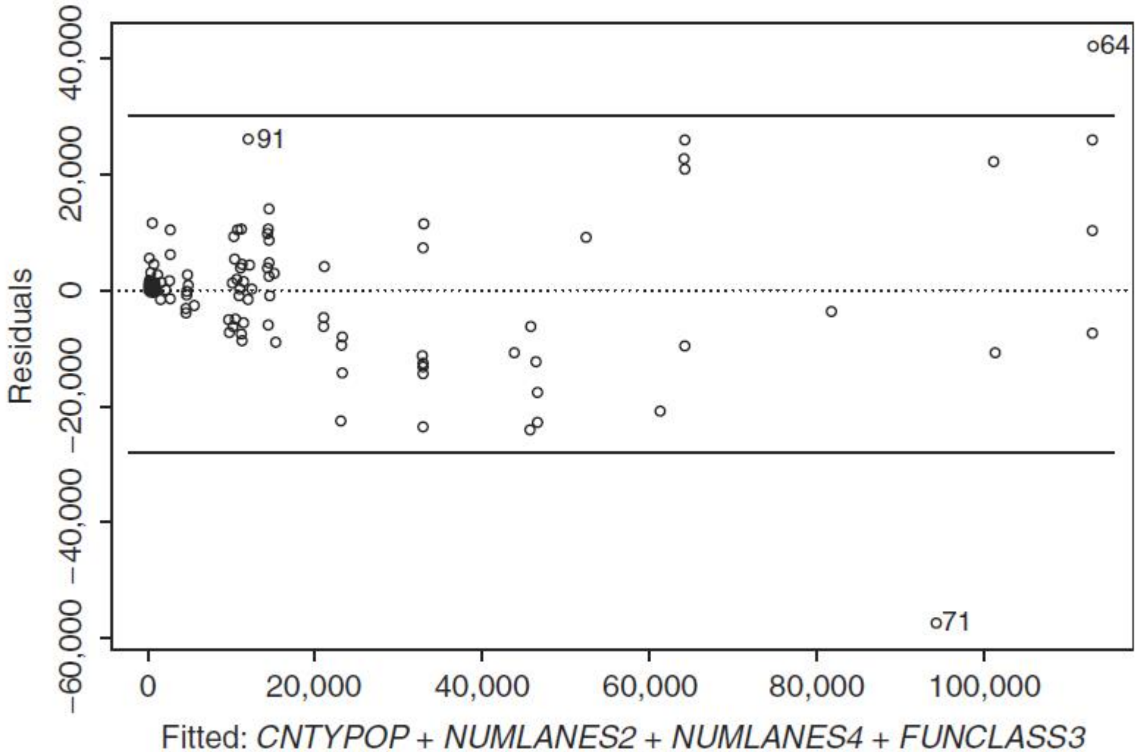
Checking Regression Assumptions

- When disturbances are not **homoscedastic**, they are said to be **heteroscedastic**.
  - A plot of model **fitted values** vs. **disturbances** is typically inspected first.
  - If heteroscedasticity is detected, then plots of the **disturbances** vs. **independent variables** should be conducted to identify the culprit.



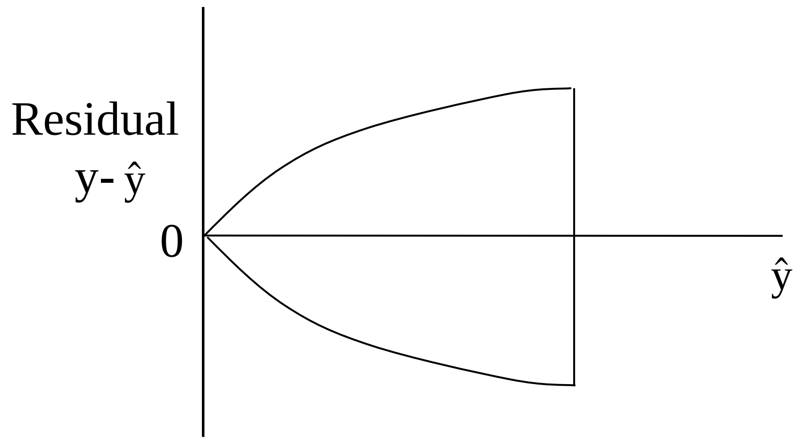
Checking Regression Assumptions

- The horizontal lines shown on the plot show a fairly **constant band** of equidistant disturbances around the regression function. In other words, the disturbances do not become **systematically larger or smaller** across fitted values of Y



Checking Regression Assumptions

- DETECTING UNEQUAL VARIANCES

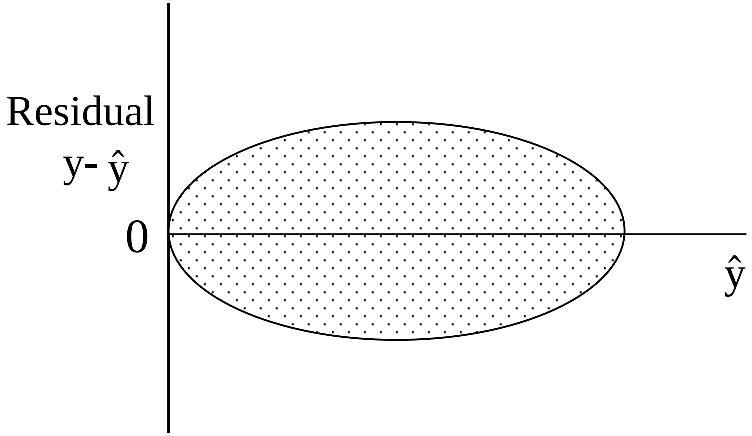


This pattern (actually, it is more rounded) may result if y is best represented as **a poisson** random variable. This happens if one is trying to **model count or arrival/departure data** (counts per unit area or time such as traffic counts). In this case, it may help to **fit √y** instead of y to the independent variables.

Another possible pattern is as follows:

Checking Regression Assumptions

DETECTING UNEQUAL VARIANCES

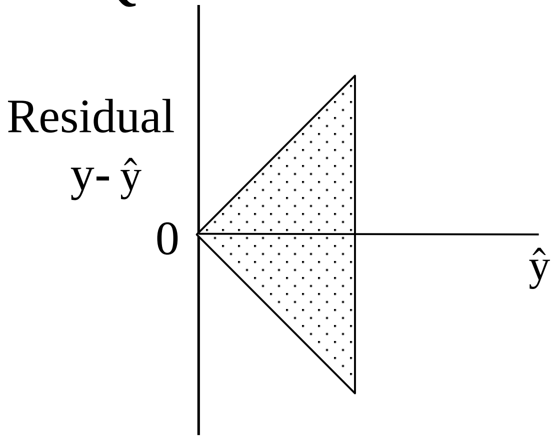


This pattern results if **y is a percentage or proportion,** $\hat{p} = \frac{y}{n}$. In this case, $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ is small when p is near 0 or 1 and reaches a maximum when p is equal to 0.5. Therefore, the plot will appear as shown above. In this case, it is useful to fit **$\sin^{-1}\sqrt{y}$ (instead of y),** where y is expressed in radians. This will offer a stabilizing influence on the residuals.

A third common situation is as follows:

Checking Regression Assumptions

- DETECTING UNEQUAL VARIANCES



This type of pattern tends to occur when the response variable y follows a multiplicative model (note that this is different from the first pattern in that the first pattern is supposed to be more rounded). Unlike the additive model discussed so far, in this model the dependent variable is written as the product of its mean and the random error component.

$y = [E(y)].\varepsilon$

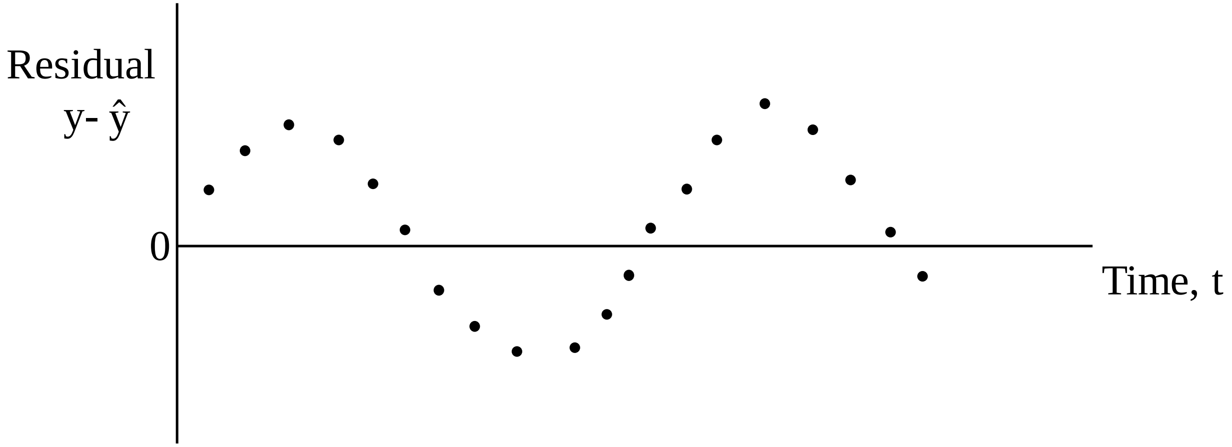
The variance of this response will grow proportionally to the square of the mean, i.e., $Var(y) = [E(y)]^2\sigma^2$, where σ^2 is the variance of ε. The appropriate transformation for this type of **data/model is log (y).** Why?

Checking Regression Assumptions

- Often, heteroscedasticity is easily detected. In many applications, disturbances that are an **increasing** function of fitted **values of Y** are often encountered.
- Remedial measures for dealing with heteroscedasticity include
  - **transformations** on the response variable, Y,
  - **weighted least squares** (WLS),
  - **ridge regression**
  - **generalized least squares**

Checking Regression Assumptions

- **Correlated disturbances** can result when observations are dependent across individuals, time, or space.
- Traffic volumes recorded every 15 min are typically correlated across observations.
- Correlation of disturbances across time is called serial correlation. The standard plot for detecting serial correlation is a plot of disturbances vs. time, or a plot of disturbances vs. ordered observations (over space).



Checking Regression Assumptions

- **Exogeneity assumption:** all independent variables in the regression are exogenous.
- By ignoring the "feed back effect" caused by endogeneity, statistical inferences can be strongly biased.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Y = current military expenditures, \$ (millions)

X_1 = number of past military conflicts

X_2 = gross domestic product, \$ (millions)

Checking Regression Assumptions

- An assumption imposed simply to allow regression parameter inferences to be drawn is that disturbances are approximately **normally** distributed
- Summary statistics of the disturbances, including minimum, first and third quartiles, median, and maximum values of the disturbances.
- **Histograms** of the disturbances. A histogram of the disturbances should reveal the familiar **bell-shape**d normal curve..
- Normal probability **quantile-quantile (Q-Q) plots** of the disturbances.
- Nonparametric methods such as the **chi square** goodness-of-fit (GOF) test or the **Kolmogorov–Smirnov GOF** test

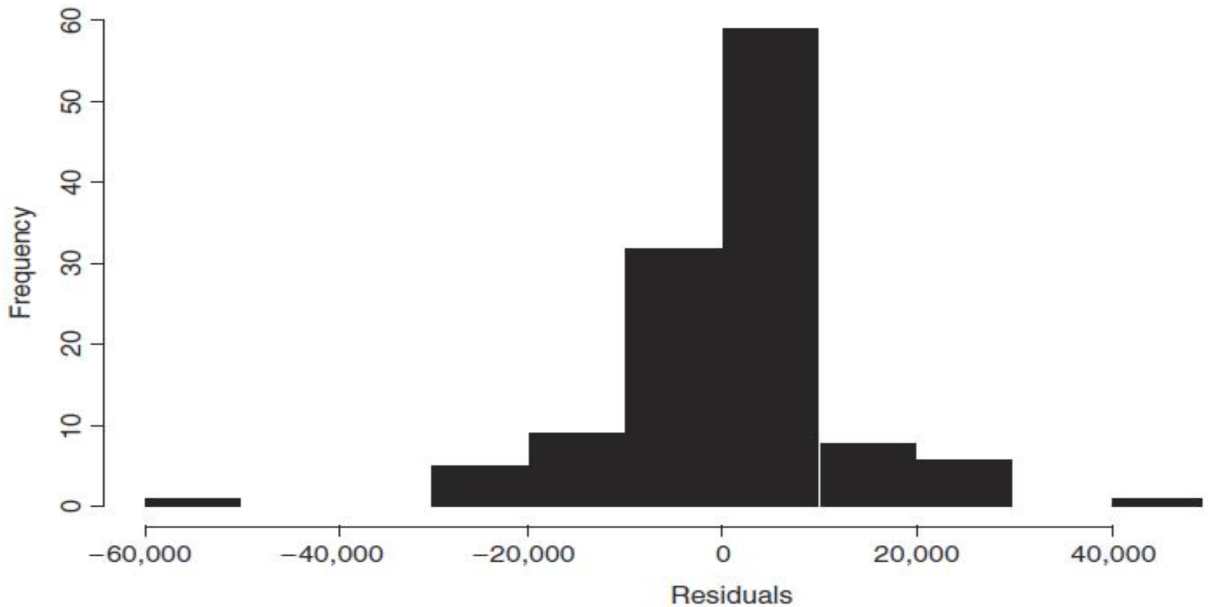
Checking Regression Assumptions

Consider the regression model in Example 3.4. A check of the homoscedasticity assumption produced no evidence to reject it. Because the data were collected at different points in space and not over time, serial correlation is not a concern. The normality assumption is now checked. Inspection of standard summary statistics provides an initial glimpse at the normality assumption.

Min	1st Quartile	Median	3rd Quartile	Max
–57878	-4658	671.5	3186	42204

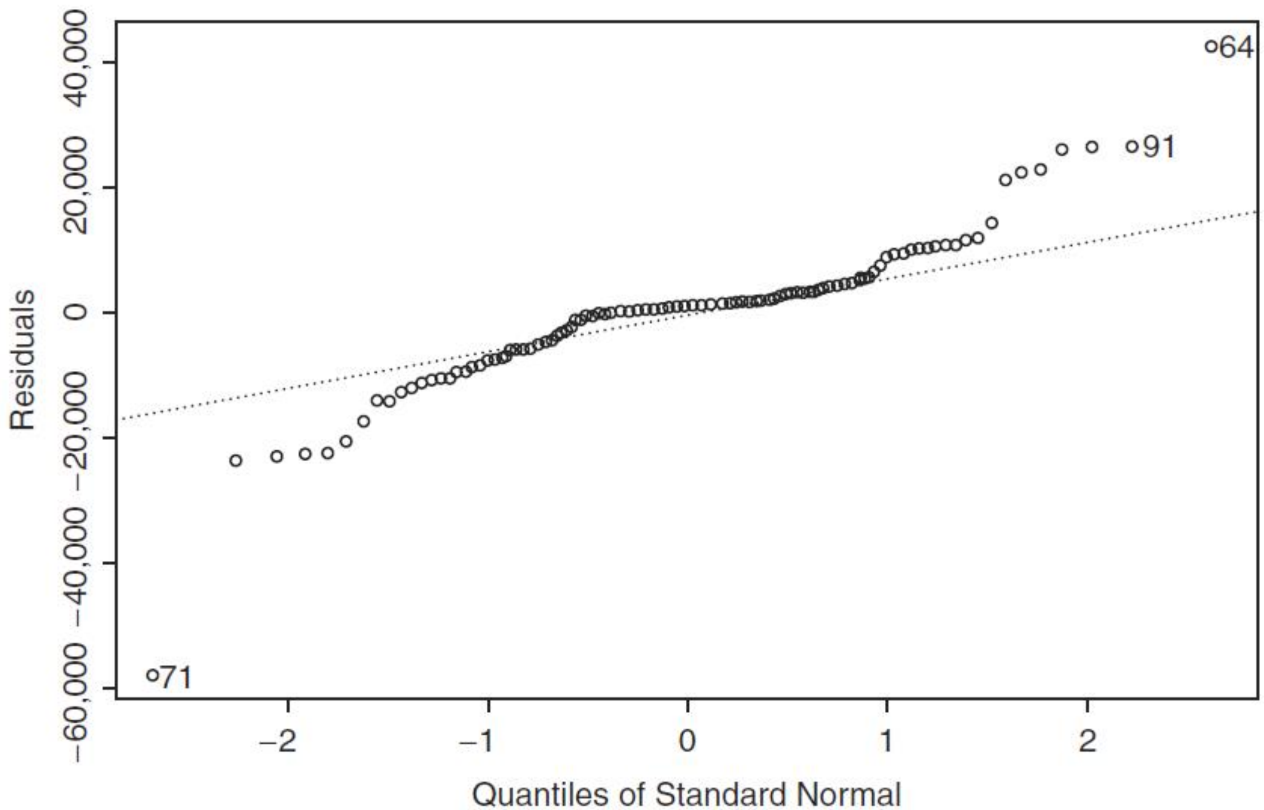
Checking Regression Assumptions

A histogram of the disturbances is shown in Figure 3.7. The Y-axis shows the number of observations, and the X-axis shows bins of disturbances. Disturbances both positive and negative appear to be outlying with respect to the bulk of the data. The peak of the distribution does not appear evenly distributed around zero, the expected result under approximate normality. What cannot be determined is whether the departures represent serious, extreme, and significant departures, or whether the data are consistent with the assumption of approximate normality. In aggregate the results are inconclusive.



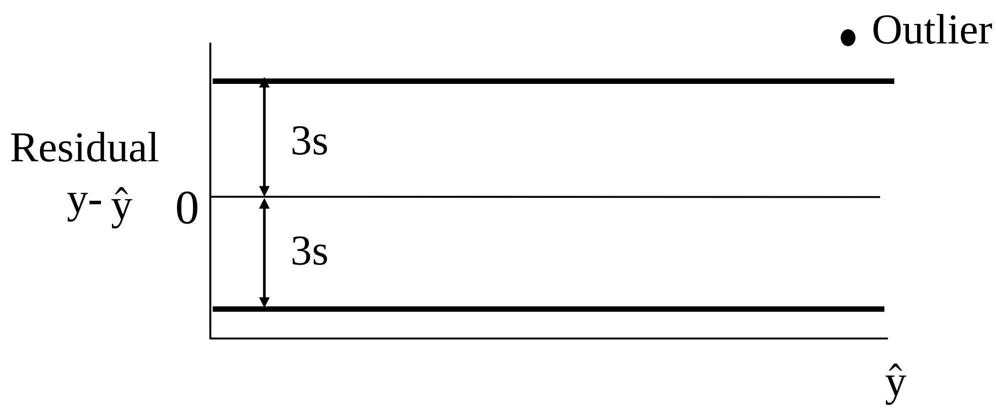
Checking Regression Assumptions

- Tails of the disturbances' distribution appear to depart from normality. Several observations, specifically observations 71, 91, and 64, appear to seriously depart from normality



Regression Outliers

- **Residual plots** can also be used to detect **outliers**, values of y that appear to be in total disagreement with the model
- As almost all values of y should lie within **3σ of e(y)** (for a normal distribution, **99.9%** of the observations lie within 3 standard deviations of the mean), one would expect the absolute value of residuals to be no greater than 3s
- A residual that is larger than 3s (in absolute value) is considered to be an outlier
- To detect outliers, construct a plot as shown below:

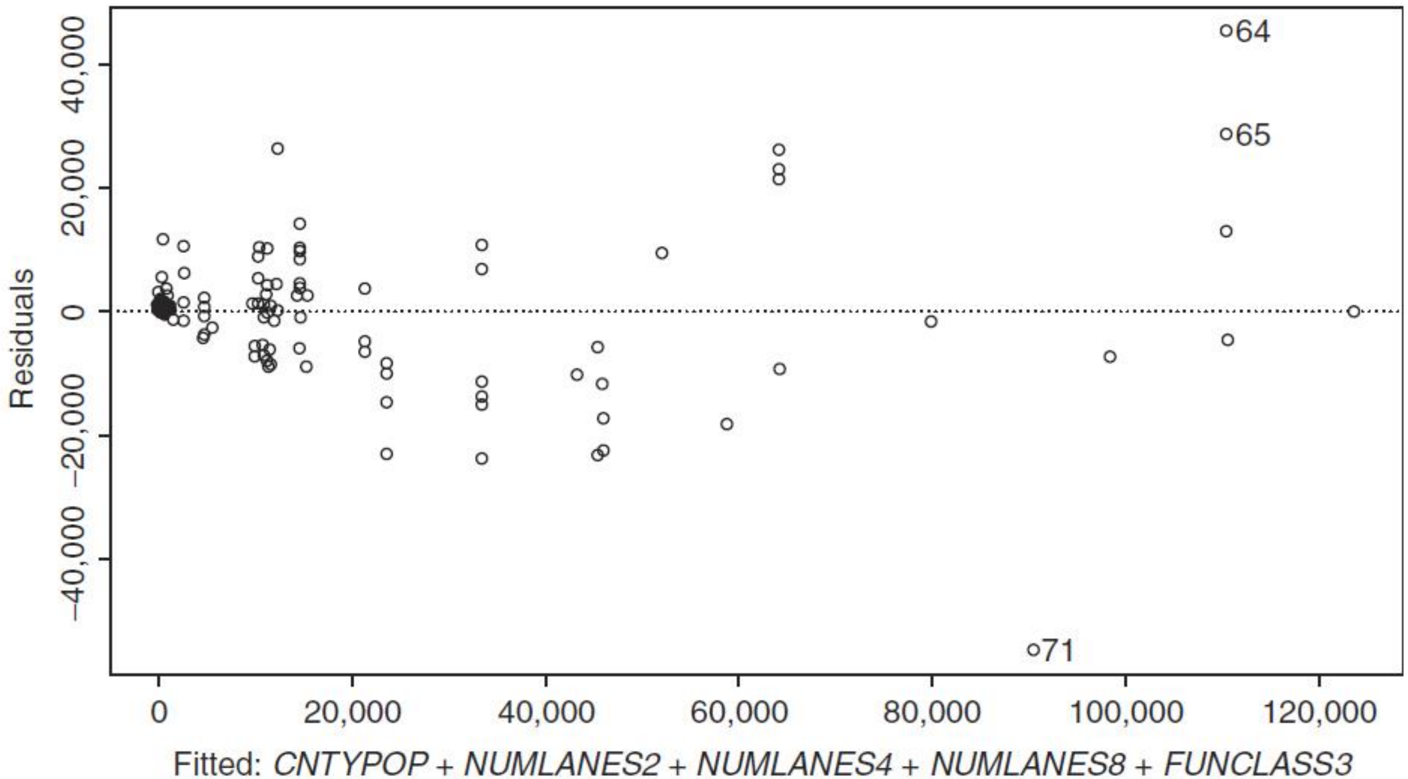


Regression Outliers

- There are many possible scenarios that could give rise to outliers
  - **Misspecification error**. A specified model may be inappropriate and fail to account for some important effects, particularly with respect to influential cases.
  - **Coding error.** An influential data point (or points) was recorded incorrectly during data collection.
  - **Data collection error.** Influential observations were the result of malfunctioning equipment, human error, or other errors that occurred during data collection.
  - **Calculation error.** Often there is a significant amount of data manipulation that occurs prior to analysis.

Regression Outliers

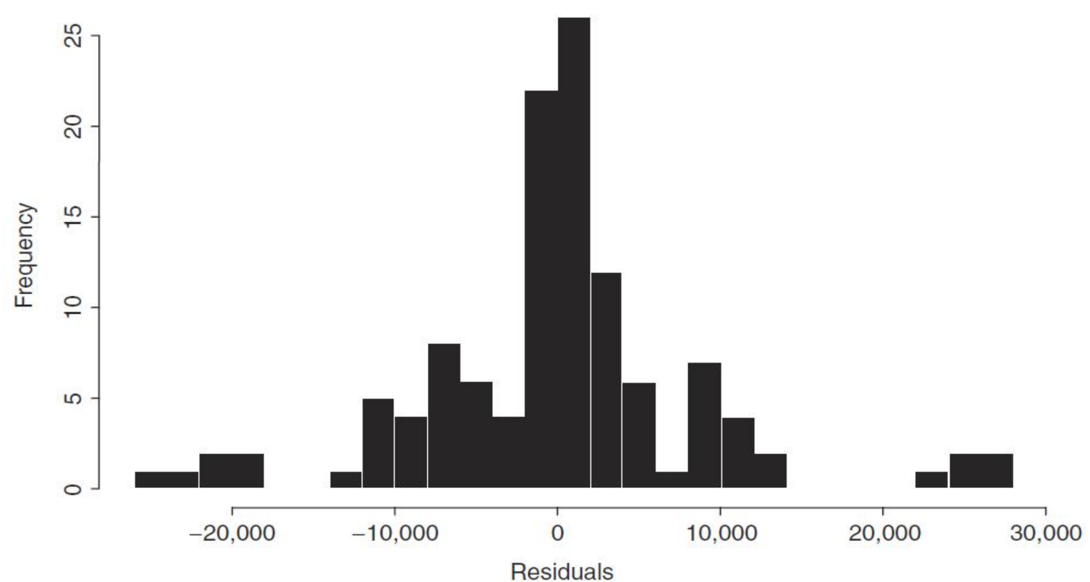
- Removing outliers without proper justification raises the possibility that data have been manipulated to support a particular hypothesis or model



Regression Outliers

Least Squares Estimated Parameters (Example 3.6)

Parameter	Parameter Estimate	Standard Error of Estimate	*t*-value	$P(>\|t\|)$
Intercept	59771.4183	4569.8595	13.0795	<0.0001
CNTYPOP	0.0213	0.0029	7.2198	<0.0001
NUMLANES2	–59274.8788	4569.1291	–12.9729	<0.0001
NUMLANES4	–48875.1655	4269.3498	–11.4482	<0.0001
NUMLANES8	22261.1985	10024.2329	2.2207	0.0284
FUNCLASS3	31841.7658	2997.3645	10.6233	<0.0001
R-squared		0.8947		
F-statistic	192.1 on 5 and 113 degrees of freedom, the *p*-value is <0.0001			



Model Goodness-of-Fit Measures

- GOF: **R-squared**, and **adjusted R-squared**.

- To develop the R-squared GOF statistic, some basic notions are required:

- The sum of square errors (disturbances) is given by:

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

- the regression sum of squares is given by

$$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$$

Model Goodness-of-Fit Measures

- The total sum of squares is given by

$$\mathrm{SST} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

- It also can be shown algebraically that **SST = SSR + SSE.**
- The coefficient of determination, R-squared, is defined as

$$R^2 = \frac{[\mathrm{SST} - \mathrm{SSE}]}{\mathrm{SST}} = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

- The coefficient of determination, R-squared, is defined as

$$R^2{}_{\text{adjusted}} = 1 - \frac{\frac{\mathrm{SSE}}{n-p}}{\frac{\mathrm{SST}}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\mathrm{SSE}}{\mathrm{SST}}$$

Model Goodness-of-Fit Measures

- The R^2 and R^2_{adjusted} measures provide only **relevant comparisons** with previous models that have been estimated on the phenomenon under investigation
- The absolute values of R^2 and R^2 adjusted measures **are not sufficient** measures to judge the quality of a model.
- Relatively large values of R^2 and R^2 adjusted can be caused by data artifacts.
- The R^2 value is 0.8947. The collection of independent variables accounts for about 89% of the uncertainty or variation in AADT.
- This is considered to be a good result because previous studies have achieved R-squared values of around 70%. It is only with respect to other models on the same phenomenon that R-squared comparisons are meaningful.

Multicollinearity in the Regression

- Estimated parameters vary widely from one sample to the next, perhaps resulting in **counterintuitive signs**.
- The **standard interpretation** of a regression parameter does **not apply**: one cannot simply adjust the value of an independent variable by one unit to assess its affect on the response, because the value of the correlated independent variable will change also.

Regression Model-Building Strategies

- (1) The simple regression analyses (one variable at a time) were conducted to identify the variables that were significantly. The **insignificant variables** were **discarded** in the following steps.
- (2) The **stepwise regression analysis** was then conducted to select variables from the set of significant variables. The Pearson correlation parameters between the selected variables were calculated. Some selected variables were collinear or nearly collinear with each other.
- (3) In such cases, the linear regression model was developed with each variable separately. The R^2 of each regression model was compared. The variable that produced the best R^2 was kept in the final model. The final selected variables were used to develop the linear regression models.