CIS9660:
Data Mining for Business Analytics

2. Introduction to Linear Regression

Imp 2 4 5 6 7 8 9 14 15
16 17 18 19 20 21 22 23
24 25 27 28 29 30-38 41

(Gujarati, D.N., 2011. Econometrics by example. New York: Palgrave Macmillan. Chapter 1,4,5,6,7)

Linear Regression Model

- The LRM (Linear Regression Model) in its general form may be written as:

$$Y_i = B_1 + B_2 X_{2i} + B_3 X_{3i} + \ldots + B_k X_{ki} + u_i$$

  - This is known as the population or true model
  - The variable Y is known as the dependent variable
  - The X variables are known as the explanatory variables, predictors, covariates, or regressors
  - u is known as an error term, which is a catchall for all those variables that cannot be introduced in the model for a variety of reasons
  - The subscript i denotes the ith observation
  - B_1 is known as the intercept and B_2 to B_k are known as the coefficients Collectively, they are called regression coefficients or regression parameters
  - In regression analysis our primary objective is to explain the mean, or average, behavior of Y in relation to the regressors, that is, how mean Y responds to changes in the values of the X variables
  - How many regressors are included in the model depends on the nature of the problem and will vary from problem to problem

Estimation of the Linear Regression Model

Having obtained the data, the important question is: how do we estimate the LRM? Suppose we want to estimate a wage function of a group of workers. To explain the hourly wage rate (Y), we may have data on variables such as gender, ethnicity, union status, education, work experience, and many others, which are the X regressors. Further, suppose that we have a random sample of 1,000 workers. How then do we estimate the model?

A commonly used method to estimate the regression coefficients is the method of ordinary least squares (OLS). To explain this method, we rewrite the model as follows:

$$\begin{aligned} u_i &= Y_i - (B_1 + B_2 X_{2i} + B_3 X_{3i} + \ldots + B_k X_{ki}) \\ &= Y_i - \mathbf{BX} \end{aligned}$$

where the error term is the difference between the actual Y value and the Y value obtained from the regression model. One way to obtain estimates of the B coefficients would be to minimize the squared error term ui:

$$\sum u_i^2 = \sum (Y_i - B_1 - B_2 X_{2i} - B_3 X_{3i} - \ldots - B_k X_{ki})^2$$

We call $\sum u_i^2$ the error sum of squares (ESS).

Estimation of the Linear Regression Model

We will denote the estimated B coefficients with a lower-case b, and therefore the estimating regression can be written as:

$$Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + \ldots + b_k X_{ki} + e_i$$

which may be called the sample regression model, the counterpart of the population model given earlier.

- We call the b coefficients the estimators of the B coefficients
- We call e_i the residual, an estimator of the error term u_i
- The values of b coefficients will change from sample to sample
- The (population) regression coefficients or parameters, the B coefficients, are fixed numbers, although we do not know what they are. On the basis of the sample, we try to obtain the best guesses of them

R^2: A Measure of Goodness of Fit of the Estimated Regression

The coefficient of determination, denoted by R^2, is an overall measure of goodness of fit of the estimated regression line.

- That is, it gives the proportion or percentage of the total variation in the dependent variable Y (SST) that is explained by all the regressors
- R^2 = SSE/SST = 1 – SSR/SST
- We can think of each observation as being made up of an explained part, and an unexplained part, $y_i = \hat{y}_i + \hat{u}_i$ We then define the following :

  $\sum (y_i - \bar{y})^2$ is the total sum of squares (SST)

  $\sum (\hat{y}_i - \bar{y})^2$ is the explained sum of squares (SSE)

  $\sum \hat{u}_i^2$ is the residual sum of squares (SSR)

  Then $\text{SST} = \text{SSE} + \text{SSR}$

  - $\hat{y}_i$ is the predicted value of y_i calculated with the value of x_i and coefficients;
  - $\bar{y}$ is the mean of the observed value of all y;
  - y_i is the observed value of y_i.

- R^2 lies between 0 and 1. The closer it is to 1, the better is the fit, and the closer it is to 0, the worse is the fit.

R^2: A Measure of Goodness of Fit of the Estimated Regression

- One disadvantage of R^2 is that it is an increasing function of the number of regressors. That is, if you add a variable to model, the R^2 values increases. So sometimes researchers play the game of "maximizing" R^2, (include a large number of regressors to increase the value of R^2).
- To avoid this temptation, we use an R^2 measure that explicitly takes into account the number of regressors included in the model. Such an R^2 is called an adjusted R^2, denoted as $\overline{R}^2$(R-bar squared),
- The term "adjusted" means adjusted for the degrees of freedom, which depend on the number of regressors (k) in the model
- The adjusted R^2 is calculated as:

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{SST}$$

  - ✓ The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance
  - ✓ N is the number of observations in the sample

How to interpret the R^2 and the adjusted R^2 in Practice

- $\boldsymbol{R^2}$ is the fraction of the sample variance of Y_i explained by (or predicted by) X_i.
- An R^2 or an adjusted R^2 near 1 means that the regressors are jointly good at predicting the values of the dependent variable in the sample, and an R^2 or an adjusted R^2 near 0 means that they are not. This makes these statistics useful summaries of the predictive ability of the regression.
- However, it is easy to read more into them than they deserve.

How to interpret the R^2 and the adjusted R^2 in Practice

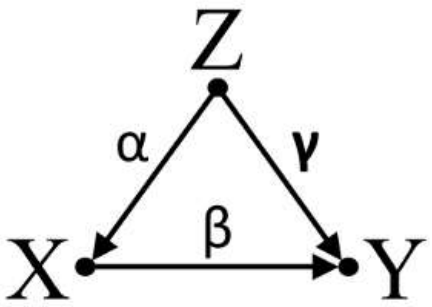
There are four potential pitfalls to guard against when using the R^2 or adjusted R^2 :

- An increase in the R^2 or adjusted R^2 does not necessarily mean that an added variable is statistically significant.
- A high R^2 or adjusted R^2 does not mean that the regressors are a true cause of the dependent variable. (Correlation versus Causality)
- A high R^2 or adjusted R^2 does not necessarily mean that you have the most appropriate set of regressors, nor does a low R^2 or adjusted R^2 necessarily mean that you have an inappropriate set of regressors.

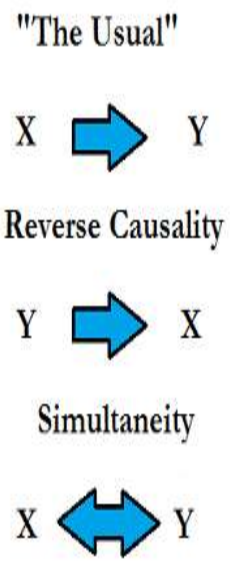
(James H. Stock and Mark W. Watson. *Introduction to Econometrics*. 2006. Print)

Correlation vs. Causality

- Correlation does not always imply causation because:
  - **Omitted variable bias:**
    - There is a variable that is not among the explanatory or response variables in a study, and yet may cause both the change in the depended and independent variables.



  - **Reverse causality**:
    - Reverse causality refers either to a direction of cause-and-effect contrary to a common presumption or to a two-way causal relationship in, as it were, a loop..



An Illustrative Example for Causality Analysis

The superintendent of an elementary school district must decide whether to hire additional teachers and she wants your advice. If she hires the teachers, she will reduce the number of students per teacher (the student-teacher ratio) by two. She faces a tradeoff. Parents want smaller classes so that their children can receive more individualized attention. But hiring more teachers means spending more money, which is not to the liking of those paying the bill! So she asks you: If she cuts class sizes, what will the effect be on student performance?

(James H. Stock and Mark W. Watson. *Introduction to Econometrics*. 2006. Print)

We sharpen the superintendent's question: If she reduces the average class size, what will the effect be on standardized test scores in her district?

$$\beta_{ClassSize} = \frac{\text{change in TestScore}}{\text{change in ClassSize}} = \frac{\Delta \text{TestScore}}{\Delta \text{ClassSize}},$$

The above equation is the definition of the slope of a straight line relating test scores and class size. This straight line can be written as:

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize,$$

- Where β_0 is the intercept of this straight line, and $\beta_{\text{ClassSize}}$ is the slope
- If you knew β_0 and $\beta_{\text{ClassSize}}$, you would be able to determine the change in test scores at a district associated with a change in class size
- What does it mean if we find $\beta_{\text{ClassSize}} = 0.6$? Does that make sense to you?

Suppose you have a sample of n districts:

- Y_i is the average test score in the i th district;
- X_i is the average class size in the i th district, and
- u_i denote the other factors influencing the test score in the i th district.
- Then the previous equation can be written more generally as:

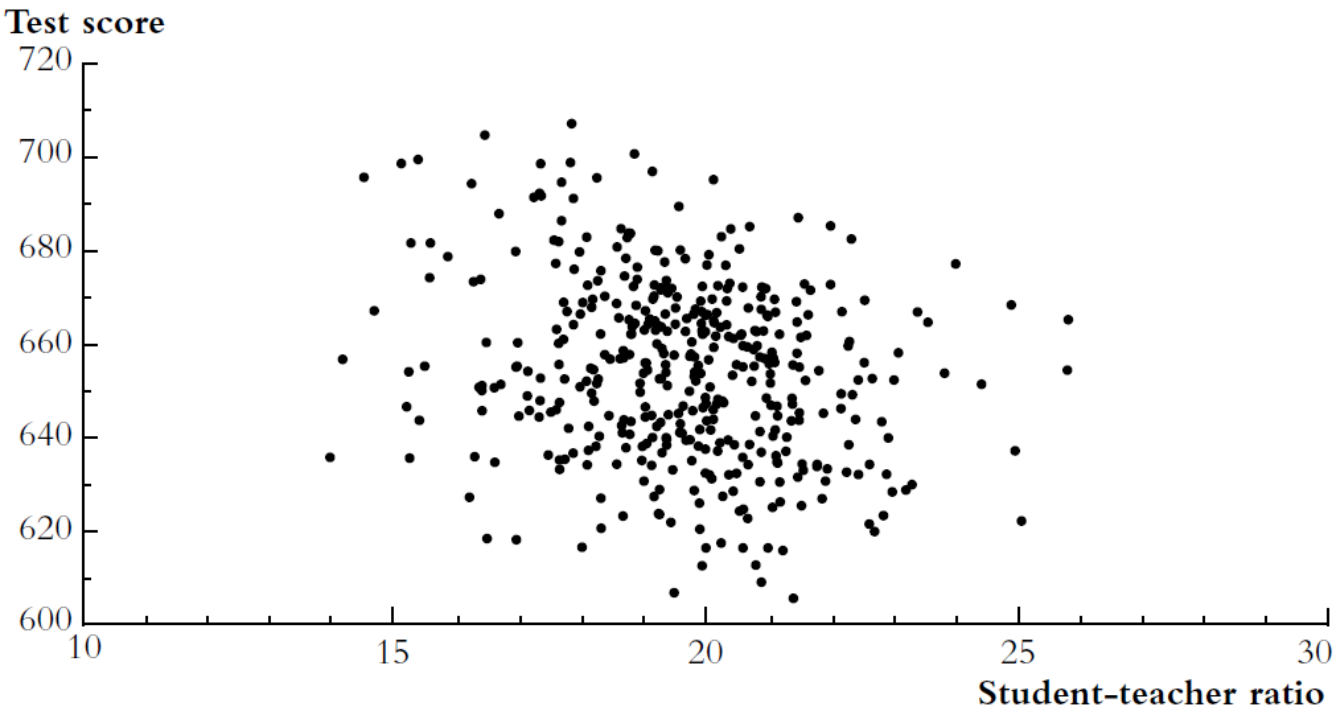
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- $\beta_0 + \beta_1 X_i$, is the population regression line or the population regression function. This is the relationship that holds between Y and X on average over the population.
- The intercept β_0 and the slope β_1 are the coefficients or parameters of the population regression line.
- The slope β_1 is the change in Y associated with a unit change in X.
- The intercept β_0 is the value of the population regression line when X = 0.
- The term u_i is the error term, which contains all the other factors besides X that determine the value of the dependent variable, Y, for a specific observation, X.

How to find the “best fit” line?

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)

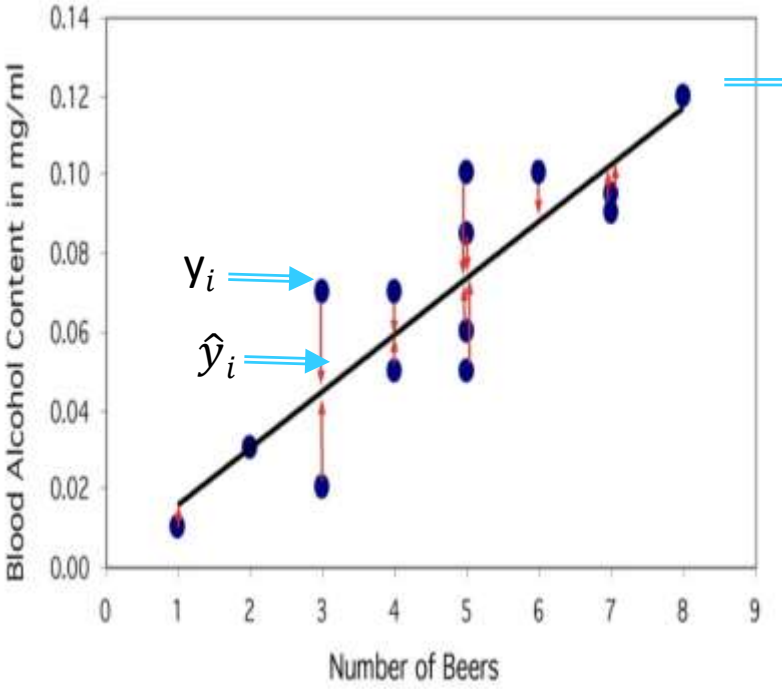
Data from 420 California school districts.



(James H. Stock and Mark W. Watson. *Introduction to Econometrics*. 2006. Print)

How to find the “best fit” line?

- Ordinary least squares (OLS)
- Intuitively, the OLS estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by **the sum of the squared mistakes made in predicting Y given X**.



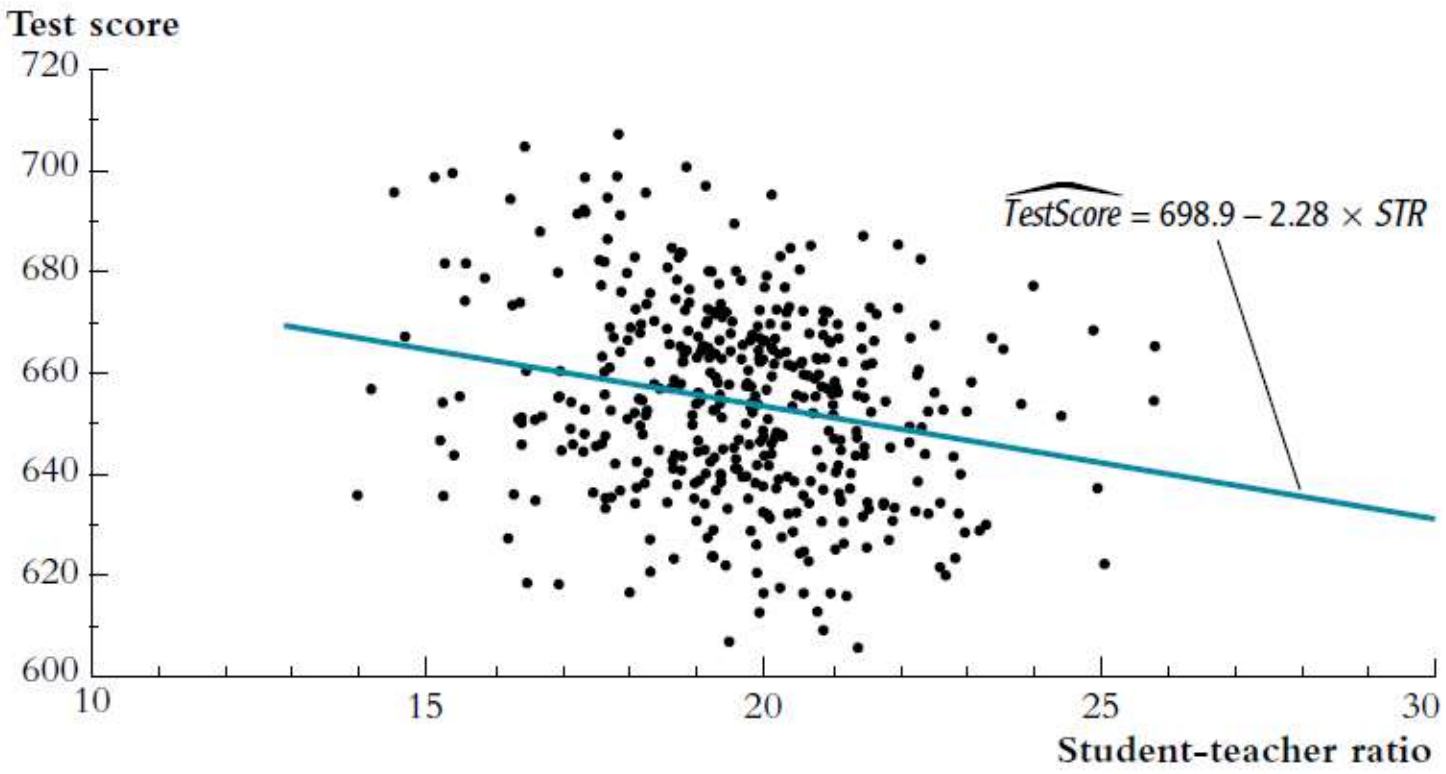
Minimize $\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

$$\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}.$$

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



- The slope of -2.28 means that an increase in the student-teacher ratio by one student per class is, on average, associated with a decline in districtwide test scores by 2.28 points on the test.

Regression with Multiple Regressors

When you propose model to the superintendent, she tells you that something is wrong with this formulation. She points out that class size is just one of many facets of elementary education, and that two districts with the same class sizes will have different test scores for many reasons, e.g., quality of their teachers, background of their students, how lucky the students were on test day, etc.

One approach would be to list the most important factors and to introduce them explicitly into the model.

- For now, we simply lump all these "other factors" together and write the relationship for a given district as:

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + \text{other factors.}$$

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic I: multicollinearity**
  - One of the assumptions of the classical linear regression model (CLRM) is that there is no perfect linear relationship among the regressors
  - Possible sources of **perfect collinearity**:
    - Having two or more perfectly correlated predictor variables (e.g. the dummy variable trap—including a dummy for both female and male in the model)
    - There is too little data available compared to the number of parameters to be estimated (e.g. fewer data points than regression coefficients)
  - Solutions to perfect multicollinearity
    - If you try to estimate one regression with perfect collinearity, the software will do one of two things: Either it will drop one of the occurrences or it will refuse to calculate the OLS estimates and give an error message.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic I: multicollinearity**
  - In practice, perfect linear relationship(s) among regressors is a rarity, but in many applications the regressors may be highly collinear. This case may be called imperfect collinearity
  - Perfect multicollinearity is a problem that often signals the presence of a logical error. In contrast, imperfect multicollinearity is not necessarily an error, but rather just a feature of OLS, your data, and the question you are trying to answer.
  - If the regressors are imperfectly multicollinear, then the coefficients on at least one individual regressor will be imprecisely estimated.
  - How to identify imperfect multicollinearity: Correlation Matrix
  - Question: If you try to estimate one regression with imperfect collinearity, do you have to drop one of the occurrences?

Critical Evaluation of the Classical Linear Regression Model

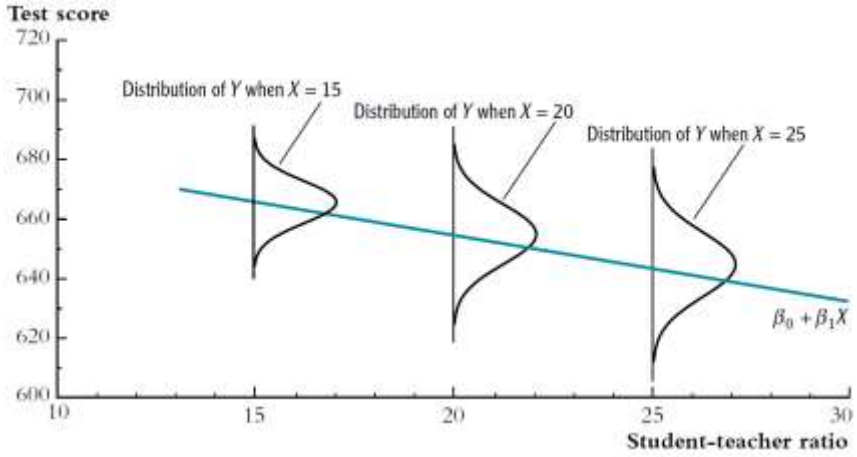
- **Regression diagnostic I: multicollinearity**
  - Consequences of imperfect collinearity:
    1) OLS estimators are still BLUE (Best Linear Unbiased Estimator), but they have large variances and covariances. As a result, the confidence intervals tend to be wider and we may not reject the "zero null hypothesis". (the t ratios of one or more coefficients tend to be statistically insignificant).
    2) Even though some regression coefficients are statistically insignificant, the R^2 value may be very high.
    3) The OLS estimators and their standard errors can be sensitive to small changes in the data.
    4) Adding a collinear variable to the chosen regression model can alter the coefficient values of the other variables in the model.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 2: heteroscedasticity** in the error term
  - Definition:
    - The classical linear regression model (CLRM) assumes that the error term u_i has homoscedasticity (equal variance) across observations. For instance, in studying consumption expenditure in relation to income, this assumption would imply that low-income and high-income households have the same disturbance variance even though their average level of consumption expenditure is different. If the assumption of homoscedasticity is not satisfied, there is the problem of heteroscedasticity

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 2: heteroscedasticity** in the error term
  - How to identify heteroscedasticity?
    - Regression with a single regressor: Plot the squared residuals against the regressor
    - Regression with multiple regressors: Plot the squared residuals against the fitted dependent variable
    - Breusch-Pagan test (a statistical test that is used to test for heteroskedasticity in a linear regression model): If the test has a p-value below an appropriate threshold (e.g. $p < 0.10$) then the null hypothesis of homoskedasticity is rejected and heteroskedasticity assumed



Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 2: heteroscedasticity** in the error term
  - Example:

    To help clarify with an example, we use the example of earnings of male versus female college graduates. Let $MALE_i$ be a binary variable that equals 1 for male college graduates and equals 0 for female graduates. The binary variable regression model relating someone's earnings to his or her gender is:

    $$Earnings_i = b_0 + b_1 MALE_i + u_i$$

    We can also write the Equation as two separate equations, one for men and one for women:

    $$Earnings_i = b_0 + u_i \text{ (women)}$$
    $$Earnings_i = b_0 + b_1 + u_i \text{ (men)}$$

    Thus, for women, u_i is the deviation of the i th woman's earnings from the population mean earnings for women (b_0), and for men, u_i is the deviation of the i th man's earnings from the population mean earnings for men ($b_0 + b_1$). In this example, the error term is homoskedastic if the variance of the population distribution of earnings is the same for men and women; if these variances differ, the error term is heteroskedastic.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 2: heteroscedasticity** in the error term
  - Consequences of heteroscedasticity:
    1) Heteroscedasticity does not alter the unbiasedness and consistency properties of OLS estimators. But OLS estimators are no longer of minimum variance or efficient. That is, they are not best linear unbiased estimators (BLUE); they are simply linear unbiased estimators (LUE).
    2) As a result, the t and F tests based under the standard assumptions of CLRM may not be reliable, resulting in incorrect conclusions regarding the statistical significance of the estimated regression coefficients.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 2: heteroscedasticity**
  - Which is more realistic, heteroskedasticity or homoskedasticity?
    - In practice, heteroskedasticity arises in many econometric applications. At a general level, economic theory rarely gives any reason to believe that the errors are homoskedastic. It therefore is prudent to assume that the errors might be heteroskedastic unless you have compelling reasons to believe otherwise
  - Solutions to heteroscedasticity
    - The simplest thing is always to use the heteroskedasticity-robust standard errors.
    - For historical reasons, many software programs use the homoskedasticity-only standard errors as their default setting, so it is up to the user to specify the option of heteroskedasticity-robust standard errors.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 3: autocorrelation**
  - Definition:
    - A common problem in regression analysis involving time series data (what is this?) is autocorrelation.
    - Recall that one of the assumptions of the classical linear regression model is that the error terms, u_t, are uncorrelated – that is the error term at time t is not correlated with the error term at time (t – 1) or any other error term in the past. If the error terms are correlated, we have autocorrelation problem.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 3: autocorrelation**
  - Consequences of autocorrelation:
    1) The OLS estimators are still unbiased and consistent. They are still normally distributed in large samples. But they are no longer efficient. That is, they are no longer BLUE (best linear unbiased estimator).
    2) If the autocorrelation is positive, standard errors tend to be smaller, and the results of the t or F tests will be inflated or biased in a positive manner. This inflation increases the Type I error rate (i.e., too often showing an effect when there actually is none).
    3) If the correlation between the errors is negative, the standard errors will be too large, causing the t or F test statistics to be smaller. This in turn, will increase Type II error rates (i.e., failing to show an effect when there actually is one).

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 3: autocorrelation**
  - How to identify autocorrelation?
    - Durbin–Watson test
      - A test statistic used to detect the presence of autocorrelation at lag 1 in the residuals (prediction errors) from a regression analysis (e.g., $u_t = \rho u_{t-1} + v_t$
      - Durbin–Watson d statistic is calculated as:

$$d = \frac{\sum_{t=2}^{t=n} (e_t - e_{t-1})^2}{\sum_{t=1}^{t=n} e_t^2}$$

      - How to understand Durbin–Watson d statistic:
        1) The d value lies between 0 and 4
        2) The closer it is to zero, the greater is the evidence of positive autocorrelation
        3) The closer it is to 4, the greater is the evidence of negative autocorrelation.
        4) If d is about 2, there is no evidence of positive or negative (first-) order autocorrelation
        5) Durbin–Watson test is not applicable, if the model contains lagged value(s) of the dependent variable (e.g., $\ln C_t = A_1 + A_2 \ln DPI_t + A_3 \ln W_t + A_4 R_t + A_5 \ln C_{t-1} + u_t$)

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 3: autocorrelation**
  - Solutions to autocorrelation
    - If the sample size is reasonably large, we can use the robust standard errors or HAC (heteroscedasticity and autocorrelation consistent) standard errors, which do not require any special knowledge of the nature of autocorrelation.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 4: model specification errors**
  - One of the assumptions of the classical linear regression model (CLRM) is that the model used in analysis is "correctly specified". By correct specification we mean one or more of the following:
    1) The model does not exclude any "core" variables.
    2) The model does not include superfluous variables.
    3) The functional form of the model is suitably chosen.
    4) There are no errors of measurement in the regressand and regressors.
    5) Outliers in the data, if any, are taken into account.
    6) The probability distribution of the error term is well specified.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 4: model specification errors**
  - Omission of "core" variables: Sometimes relevant variables are omitted because we do not have the data, or because we have not studied the underlying theory carefully, or because of carelessness. Whatever the reason, omission of important or "core" variables has the following consequences.
    1) If the omitted variables are correlated with the variables included in the model, the coefficients of the estimated model are biased.
    2) If the omitted variables are not correlated with the variables included in the model, the intercept of the estimated model is biased.
    3) The disturbance variance is incorrectly estimated.
    4) The variances of the estimated coefficients and the estimated standard errors are biased, leading to misleading conclusions about the statistical significance of the estimated parameters.
    5) Forecasts based on the incorrect model and the forecast confidence intervals based on it will be unreliable.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 4: model specification errors**
  - Omission of "core" variables: Example 1

DV: Test Scores	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	353.192***	38.542***	148.015***	271.628**	-39.322**
	(5.723)	(1.589)	(4.778)	(89.272)	(12.092)
Class Size	-4.376***	-4.549***		-2.642	-2.893***
	(0.205)	(0.028)		(1.905)	(0.256)
Quality of Instruction		64.014***			64.012***
		(0.281)			(0.275)
Socio-Economic Status			45.791***	18.274	17.448***
			(2.153)	(19.960)	(2.687)
R-squared	0.313	0.987	0.312	0.313	0.988
N	1000	1000	1000	1000	1000

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 4: model specification errors**
  - Inclusion of superfluous (irrelevant or unnecessary) variables: Sometimes researchers add variables in the hope that the R^2 value of their model will increase in the mistaken belief that the higher the R^2 the better the model. This is called overfitting a model. But if the variables are not economically meaningful and relevant, such a strategy is not recommended because of the following consequences.
    - The estimated coefficients of such a model are generally inefficient – that is, their variances will be larger than those of the true model.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 4: model specification errors**
  - Misspecification of the functional form of a regression model:

    - An example, linear or log-linear (Cobb–Douglas)?

    In an analysis to understand the determinants of wages, we have data on average wages, GDP, average hours of work, and capital expenditure for the 50 states in the USA and Washington, DC, for 1995. In labor economics, researchers often choose the log of wages as the regressand (instead of the actual value of wages). This is because the distribution of wages across the population tends to be skewed, with many workers at the low end of the distribution and a few at the high end of the distribution. On the other hand, the distribution of log of wages tends to be more symmetrical and it also has homoscedastic variance.

Critical Evaluation of the Classical Linear Regression Model

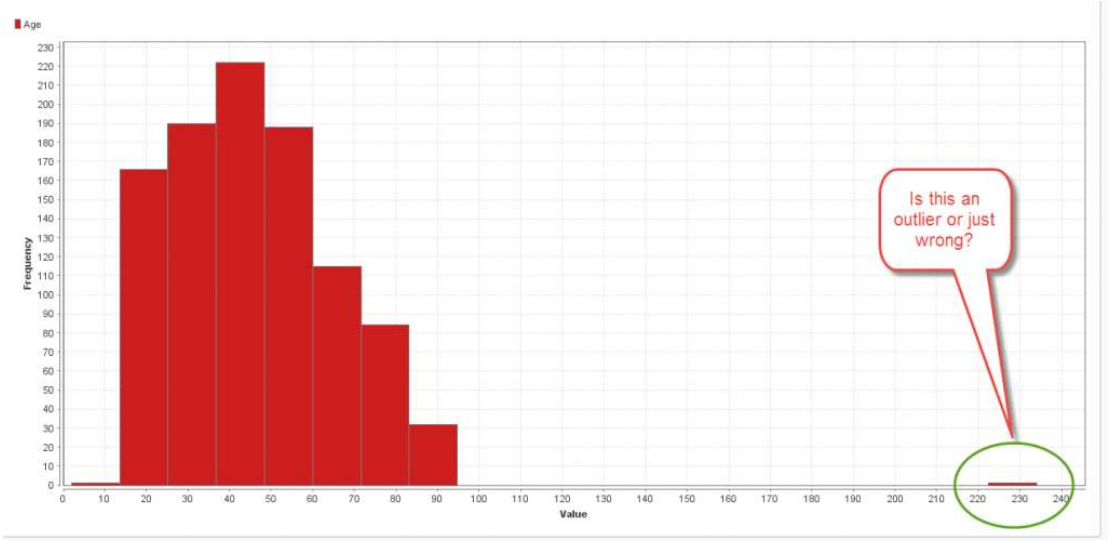
- **Regression diagnostic 4: model specification errors**
  - Errors of measurement: One of the assumptions of CLRM is that the values of the regressand as well as regressors are accurate. That is, they are not guess estimates, extrapolated, interpolated or rounded off in any systematic manner or recorded with errors.
    - If there are errors of measurement in the dependent variable, the estimated coefficients do not reflect the true relationship.
    - If there are errors of measurement in the independent variable, OLS estimators will be biased as well as inconsistent. Even such errors in a single regressor can lead to biased and inconsistent estimates of the coefficients of the other regressors in the model. And it is not easy to establish the size and direction of bias in the estimated coefficients.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 4: model specification errors**
  - Outliers, leverage and influence data: OLS gives equal weight to every observation in the sample. But this may create problems if we have observations that may not be "typical" of the rest of the sample. Such observations, or data points, are known as outliers, leverage or influence points.
    - Outliers: In the context of regression analysis, an outlier is an observation that is disproportionately distant from the bulk of the sample observations.
    - In this case such observation(s) can pull the regression line towards itself, which may distort the slope of the regression line.

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 4: model specification errors**
  - Outliers, leverage and influence data:
    - How to identify outliers?
      - The first step to detect outliers in R is to start with some descriptive statistics, and in particular with the minimum and maximum
      - Another basic way to detect outliers is to draw a histogram of the data
      - A simple method of detecting outliers is to plot the residuals or squared residuals from the estimated regression model.



How to deal with outliers?

1)If it is *an error*, remove it before estimating the parameters

2) If it is *a very important observation*, compare regression results with and without the outlier and use caution when making decisions

Critical Evaluation of the Classical Linear Regression Model

- **Regression diagnostic 4: model specification errors**
  - Probability distribution of the error term: The classical normal linear regression model assumes that the error term u_i in the regression model is normally distributed.
    - If the error term is not normally distributed, OLS estimators are still best linear unbiased estimators (BLUE). But we cannot reliably calculate confidence intervals for the model's forecasts using the t-distribution, especially for small sample sizes.
    - This assumption is thus critical if the sample size is relatively small.
    - How to test whether residuals are normally distributed?
      - One popularly used test is Shapiro-Wilk test normality test
    - How to deal with non-normally distributed residuals?
      - Transforming the dependent variable is one effective method
      - Enlarge the sample size

How to Select Variables?

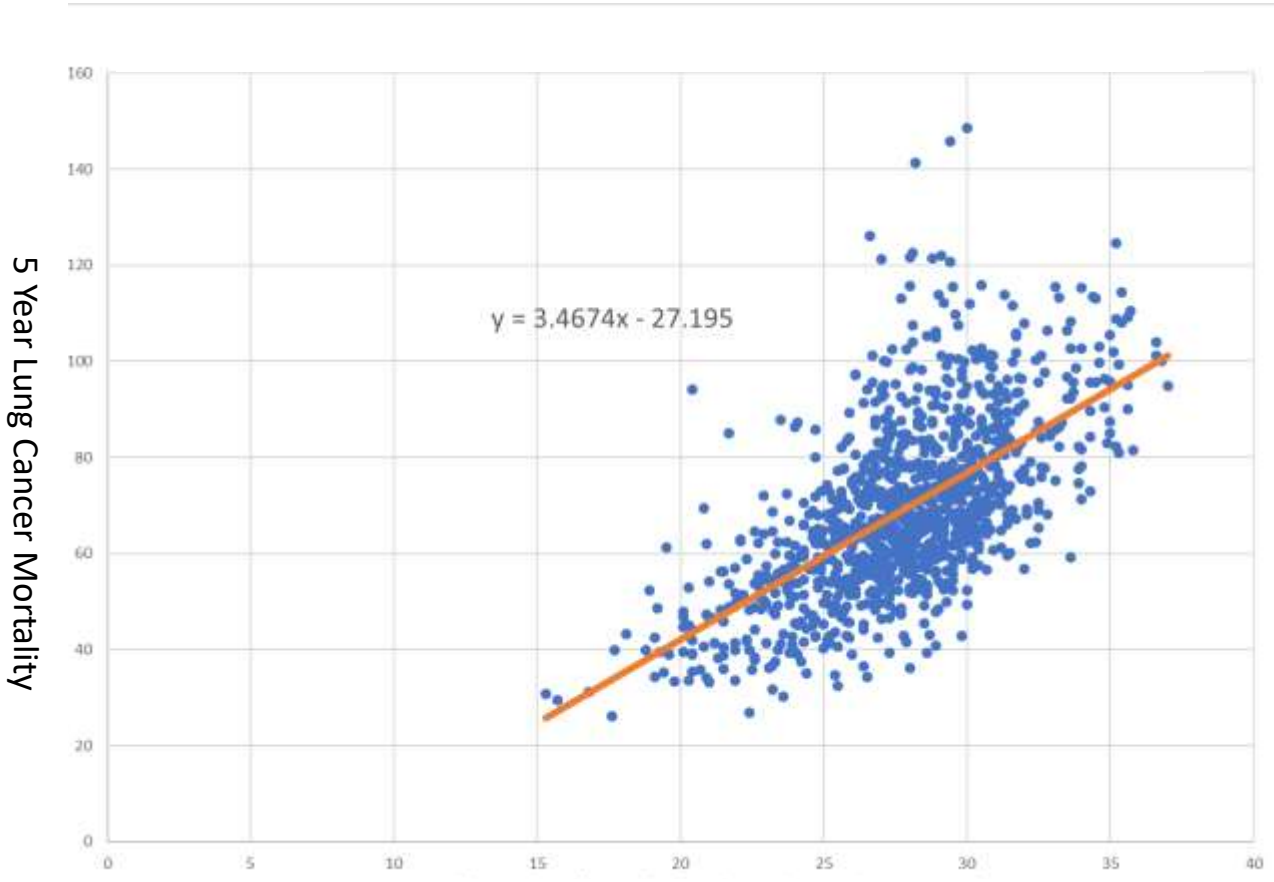
- Decisions about the regressors must weigh issues of omitted variable bias, data availability, data quality, and, most importantly, economic theory and the nature of the substantive questions being addressed.

- *A general approach to variable selection and model specification*:
  - Specify a "base" or "benchmark" model.
  - Specify a range of plausible alternative models, which include additional candidate variables.
  - Does a candidate variable change the coefficient of interest (β_1)?
  - Is a candidate variable statistically significant?
  - Use judgment, not a mechanical recipe…

Examples of Linear Regression

Does smoking cause lung cancer?

Examples of Linear Regression



Proportion Of Smokers And Nonsmokers Per County

Note: the coefficient 3.4674 is statistically significant at p=0.01

1. What is the dependent variable and what is the explanatory variable?
2. Why do we use a linear regression model?
3. How to interpret the result?
4. Based on the results, do you believe smoking causes lung cancer?
5. How to evaluate this linear regression?
6. How can you improve the model?

Why Regression Analysis?

- This example is about using data to measure causal effects
- Ideally, we would like an experiment
  - What would be an ideal experiment to estimate the effect of smoking on lung cancers?
- But almost always we only have observational (nonexperimental) data.
  - Proportion of smokers per county
  - Lung cancer mortality per county
- Most of the challenges arise from using observational to estimate causal effects
  - Confounding effects (omitted factors)
  - Simultaneous causality

Establishing causation

Lung cancer is clearly associated with smoking.

But it is hard to say whether there is a causal relationship between smoking and lung cancer based on the correlation.

What if a genetic mutation (lurking variable) caused
people to both get lung cancer and become addicted to smoking?

It took years of research and accumulated indirect evidence to reach the conclusion that smoking causes lung cancer.

Summary

- What is the basic form of linear regression?
- How to estimate the parameters of a linear regression?
- How to understand R^2 and adjusted R^2?
- Does correlation always imply causality? Why?
- How to evaluate a linear regression model?