Chapter 8

Linear Regression

8.1 Motivation

Linear regression is probably the most widely used, and useful, statistical technique for solving environmental problems. Linear regression models are extremely powerful, and have the power to empirically tease out very complicated relationships between variables. Generally speaking, the technique is useful, among other applications, in helping explain observations of a dependent variable, usually denoted y, with observed values of one or more independent variables, usually denoted x_1, x_2, ... A key feature of all regression models is the error term, which is included to capture sources of error that are not captured by other variables. Linear regression models have been heavily studied, and are very well-understood. They are only appropriate under certain assumptions, and they are often misused, even in published journal articles. These notes are intended to provide you with a broad overview of linear regression, but are not intended to exhaust all details.

The basic principle of linear regression can be illustrated with a very simple example. [Example 3.1 in Manly]. Chlorophyll-a (denoted C) is a widely used indicator of lake water quality. High concentrations of Chlorophyll-a are associated with eutrophication, which is affected by the level of nitrogen in the water. We are specifically interested in the effect an increase in nitrogen would have on the Chlorophyll-a in a lake. Using Manly's data (in the course web - called Chlorophyll.xls), the plot of Chlorophyll-a vs. Nitrogen, with a fitted linear regression line, is given in figure 8.1. But something is wrong

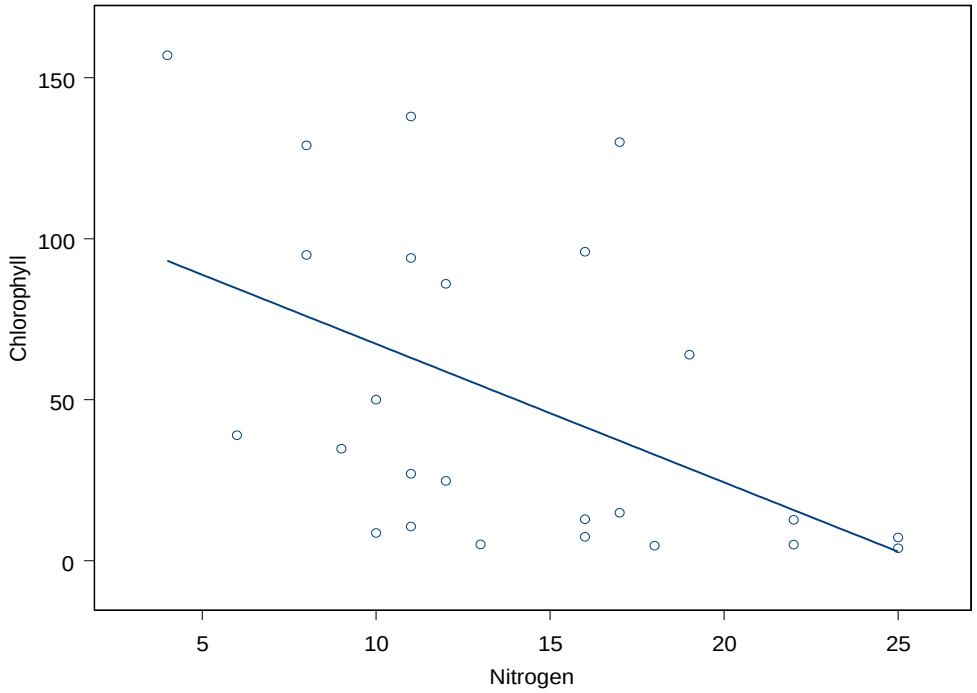


Figure 8.1: Chlorophyll-A vs. nitrogen concentration, with a fitted regression line.

here, the figure suggests that more nitrogen leads to lower Chlorophyll-a, which runs counter to our intuition. Perhaps we are omitting an important variable that might help explain the observed level of Chlorophyll-a in a lake.

In fact, high phosphorus (P) *and* high nitrogen (N) levels are associated with high Chlorophyll-a. Therefore, both variables must be included in the regression, even if we are only interested in the effect of N on Chlorophyll-a. Manly uses the following linear model to represent the relationship between C, P, and N:

$$C_i = \beta_0 + \beta_1 P_i + \beta_2 N_i + \epsilon_i \tag{8.1}$$

where C is chlorophyll-a, N is nitrogen, P is phosphorous, β_0, β_1, and β_2 are parameters that are unknown (to be estimated from the data), and ϵ is an error term. The error term will pick up an variation in the data that is unexplained by P and N. Before we estimate the parameters of this equation, let's think a little about it; here are some observations and questions:

- Suppose $P = 0$ and $N = 0$, what would we expect the level of C to be?
- What should be the sign of β_1 and β_2?
- Is our linear specification appropriate?
- What justification do we have for an additive error term?

Given that we believe the linear model we wrote above, how can we estimate the unknown parameters, β_0, β_1, and β_2? That's the purpose of this lecture. We want to choose the β's to create the "best fit" of this model to our data. The estimated parameters that accomplish this are: $\hat{\beta}_0 = -9.386$, $\hat{\beta}_1 = 0.333$, and $\hat{\beta}_2 = 1.200$. The $\hat{}$ indicates that the thing is an estimated parameter. We haven't said anything about how these numbers were chosen; that is the topic of the next set of lecture notes.

Suppose someone is thinking of developing a golf course near a lake, which is expected to increase the concentration of nitrogen in the lake by a small margin. By how much can we expect Chlorophyll-a to increase per unit increase in nitrogen? The answer, is simply our estimate of β_2, which is 1.2; a one unit increase in nitrogen leads to about a 1.2 unit increase in Chlorophyll-a[1].

[1]Note that in general it is poor practice to simply think of the possible variables that might explain something and include them into a linear model. But for the purposes of explaining the technique, we'll let Manly's example slide.

8.2 Example & Question

U.S. Gasoline Market: It seems plausible that the gas price has some effect on gas consumption. From a policy perspective, the price elasticity of demand could be extremely important, when, say, the government is contemplating increasing the gasoline tax. Our question is: how responsive is gas consumption to gas price? We might also be interested in how income or other characteristics affect gas consumption.

8.3 Evidence or Data

The first panel in figure 8.2 shows the per capita gas consumption in the U.S. through time. But this doesn't give us any information about how responsive gas consumption is to gas price. Your first instinct might be the following: Since we wish to know how gas price affects gas consumption, let's just use our data (from the Economic Report of the President, and posted in the course shared file called gasmarket.xls) and plot gas price vs. gas consumption [2]. The problem with this approach is that we have excluded many other variables that are likely to affect the consumption of gasoline. In fact, we are likely to get very misleading results using this approach (see figure that shows a positive relationship between gas price and gas consumption - oops!).

What other variables are likely to affect gas consumption (besides price)? Income, price of new cars, and price of used cars are probably appropriate. Figure 8.3 gives plots of all four independent variables (price, income, price of new cars, price of used cars) through time. The remainder of these notes is devoted to discussing, in a general way, how these types of data can be combined to answer questions such as the one posed above.

Note that the data presented here differ in units from the data described in class. The qualitative patterns are the same, but the estimated coefficients differ.

[2]The astute student will note that we have conformed to the way demand curves are typically drawn. This orientation of the axes is misleading, since gas price actually affects gas consumption, not the other way around.

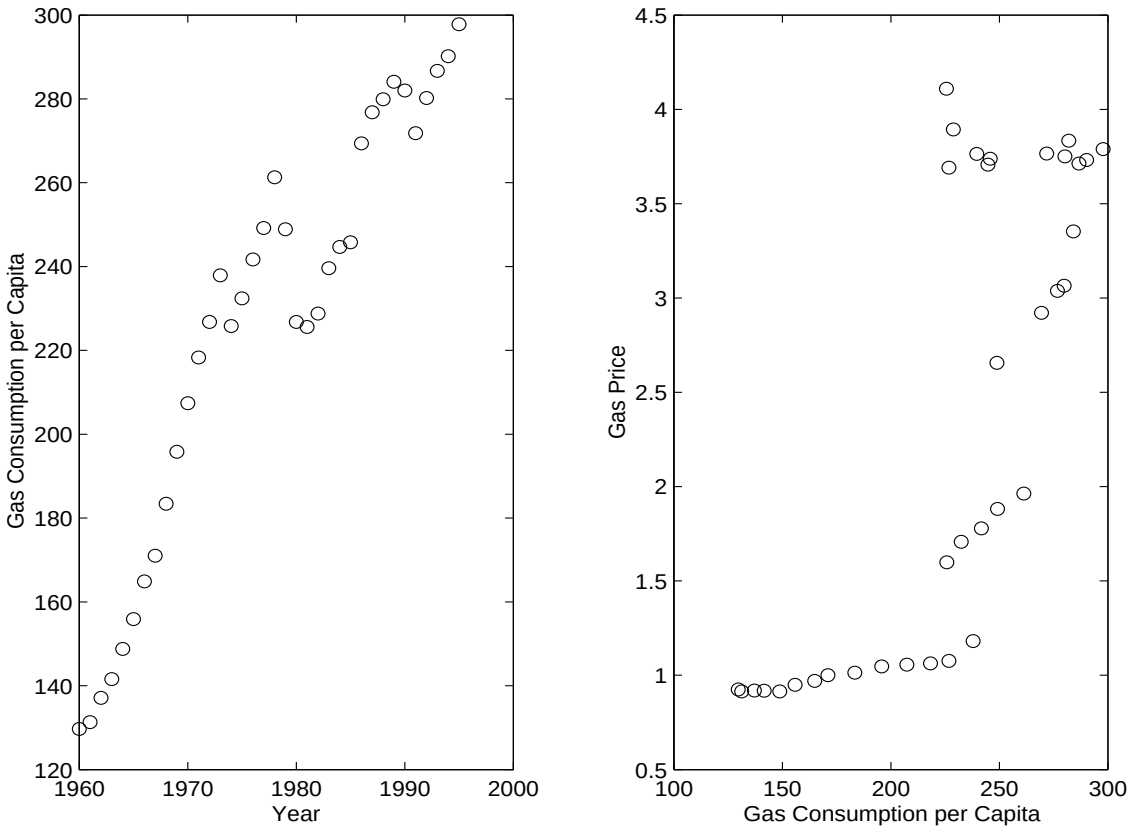


Figure 8.2: Gas consumption through time (left) and as related to gas price (right).

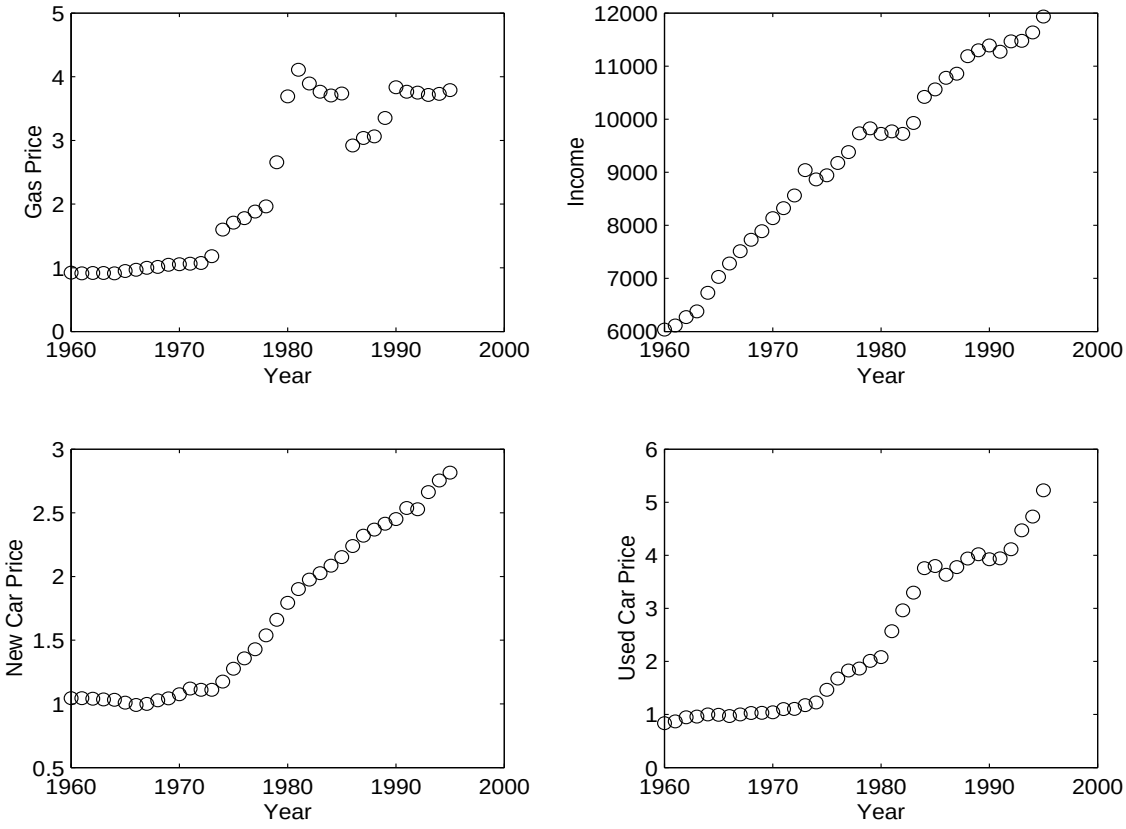


Figure 8.3: Gas price, income, price of new cars, and price of old cars throuh time.

8.4 Technique

In this section, we discuss the Ordinary Least Squares (OLS) estimator within the context of the Classical Linear Regression Model. But first, let's define what we mean by an "estimator":

Definition 8.4.1
An estimator is a rule or strategy for using data to estimate an unknown parameter, and is defined before the data are drawn[3].

It should be clear that some estimators are better than others. We won't go into a great deal of detail here about different estimators, but you should know what an estimator is, and that it is just one way to estimate an unknown parameter; other ways exist.

Whenever we want to use data to parameterize a model (such as the Chlorophyll-a model or the U.S. gasoline market model), we must choose an estimator with which to choose the "best" parameters for that model. It turns out that one such estimator, the OLS estimator, is widely (though not universally) applicable. Most of the discussion in this section is devoted to a discussion of the assumptions under which the OLS estimator is valid.

8.4.1 What is "Linear" Regression?

What do we mean when we say "linear regression". Regression refers to the fact that although observed data are variable, they tend to "regress" towards their mean. Linear refers to the type of equation we use in our models. To use the OLS estimator, the model must be linear in parameters. The standard regression equation, $Y = \alpha + \beta X + \epsilon$ is linear in the parameters α and β. The regression equation $Y = \gamma \frac{X^2}{log(X)} + \theta Z^3 + \epsilon$ is also linear in parameters (γ and θ), and could be estimated using the OLS estimator. What we mean by "could be estimated by OLS" is that we can use the technique of ordinary least squares regression to find the "best" parameters to achieve a certain criterion.

On the other hand, suppose we are modeling fish stock dynamics, and we wish to use the "Ricker Model". The Ricker Model is given below:

$$R_{t+1} = S_t e^{\phi(1-S_t)} \epsilon_t \tag{8.2}$$

[3]Adapted from Greene. Econometric Analysis

where R is the number of fish "recruits", S is the number of spawners and ϕ is a parameter to be estimated. This model is NOT linear in the parameter, ϕ, so it cannot be estimated using the OLS estimator. However, suppose we take the log of both sides:

$$log(R_{t+1}) = log(S_t) + \phi(1 - S_t) + log(\epsilon) \tag{8.3}$$

Now we have a model that is linear in parameters, so we can estimate the model using OLS.

8.4.2 Assumptions for CLRM

In order to use the OLS estimator, the following five basic assumptions must hold:

1. The dependent variable (usually denoted Y) can be expressed as a function of a specific set of independent variables, where the function is linear in unknown coefficients or parameters, and an additive error (or disturbance) term. The coefficients are assumed to be constants but are unknown. Violations of this assumption are called "specification errors", some of which are listed below:

    - Wrong set of regressors - omitting relevant independent variables or including variables that do not belong.
    - Nonlinearity - when the relationship is not linear in parameters
    - Changing parameters - when the parameters do not remain constant during the period in which the data were collected.

2. The expected value of the disturbance term is zero; i.e. the mean of the distribution from which the disturbance term is drawn is zero. A violation of this assumption introduces bias in the intercept of the regression equation.

3. The disturbance terms (there is one for every row of data) all have the same variance and are not correlated with one another. This assumption is often violated with one of the following problems:

    - Heterskedasticity - when the disturbances do not all have the same variance (often the case in cross-sectional data); constant variance is called "homoskedasticity",

- Autocorrelated Errors - when the disturbances are correlated with one another (often the case in time-series data)

4. For standard statistical inference (the next lecture) we usually assume $e_i \sim N(0, \sigma^2)$.

5. It is possible to repeat the sample with the same independent variables. Some common violations are:

   - Errors in variables - when there is measurement error in the independent variables.
   - Autoregression - when a lagged value of the dependent variable is an independent variable
   - Simultaneous equations - when several dependent variables are determined jointly by several relationships.

6. The number of observations is greater than the number of independent variables, and that there are no exact linear relationships between the independent variables.

8.4.3 Properties of Estimators

There are certain properties that we want our estimators to embody. The two main properties are unbiasedness and efficient. The bias of an estimate is the distance that estimate is from the true value of the thing it is estimating. For example, suppose you are considering purchasing an instrument that measures the salt concentration in water samples. You can think of the instrument as an "estimator" of the salt concentration. Machines typically introduce some small measurement error, and you want to know how accurate is the brand of machine you are considering purchasing. To test the machine's accuracy, you conduct the following experiment:

1. Take in a sample with a known salt concentration of, say, 50 ppm,

2. Use the machine to measure the salt concentration,

3. Repeat the measurement, say 10000 times,

4. Plot a histogram of the measurements.

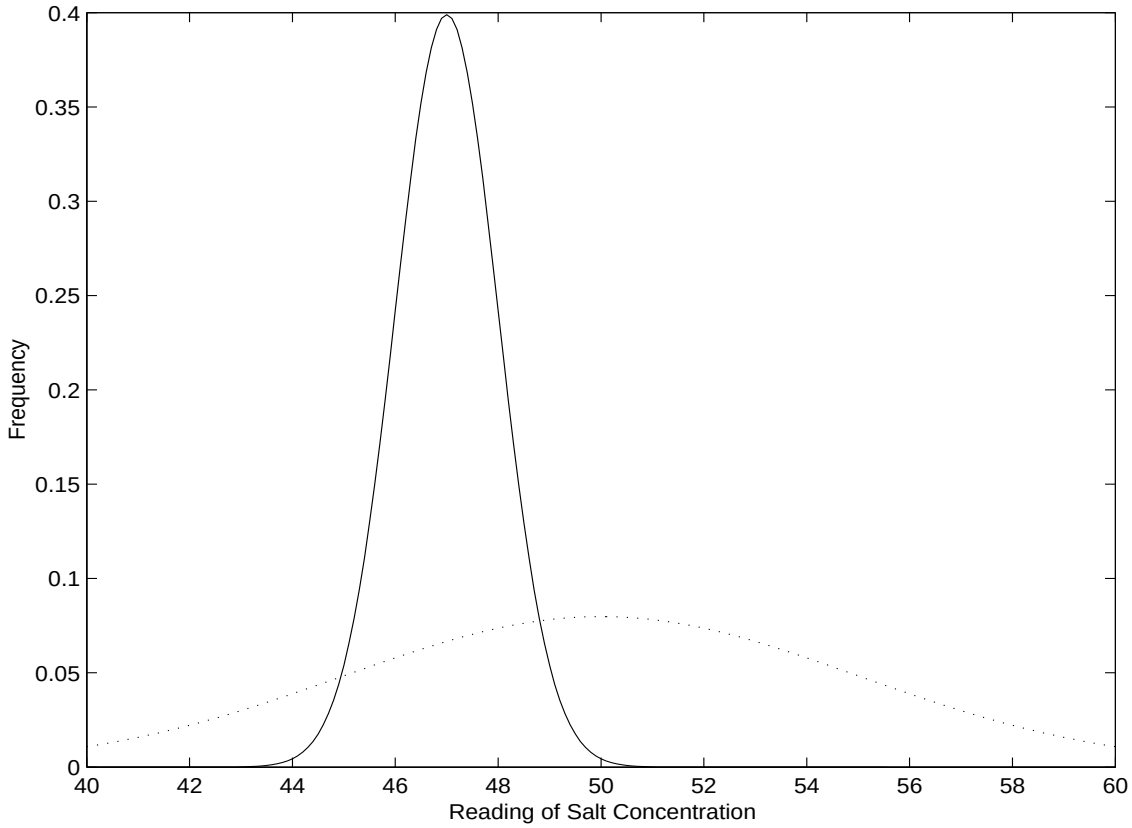


Figure 8.4: A biased estimator of salt concentration (solid line) and an unbiased but less efficient estimator (dotted line).

Figure 8.4 shows the histograms for (1) a biased machine (because the expected value of the reading is not equal to the known true value), and (2) an unbiased machine. The second desirable property, "efficiency", means that an estimator should have the smallest variance. In the example above, the first machine (the biased one) has a smaller variance, and in that respect is a better estimator than the first. The example illustrates that often we will have to make a tradeoff between biasedness and efficiency. Under the assumptions outlined above, the OLS estimator is the "best" unbiased estimator, which means that of all the possible unbiased estimators, the OLS estimator has the minimum variance.

8.4.4 Correlation vs. Causation

Explanatory variables may "explain" the dependent variable well, but this does not necessarily imply a causal link. This highlights the problem with "data mining", where many many relationships are tested to see which one has the best "fit". A high fit does not necessarily mean that the essential structure of the relationship between two or more variables has been accurately modeled. This makes it extremely difficult to use the model for any sort of predictions.

The most appropriate way to select a model is to first use basic principles or theory to construct a structural relationship between the variables of interest. For example, just because chicken production and global CO_2 measurements are nearly perfectly correlated over time, doesn't mean that if chicken production increases in the future, so will global CO_2.

True causality is extremely difficult to tease out statistically. One method is called "Granger Causality", which essentially asks whether the explanatory variable happens prior to the dependent variable; if so, a causal link is more defensible. We won't go into a great deal of detail on this subject, but it is important to remember that *correlation does not imply causation.*

8.4.5 Plotting the Residuals

Plotting the residuals is a cheap and effective way to assess model performance and specification. Remember, under our assumptions, the expected value of any one error is zero, and the variance or "spread" of the residuals should remain constant for all residuals. Put simply, the residuals should not exhibit any "pattern". If they do, something is probably wrong, and a great deal of effort should be taken to try to uncover the source of this pattern.

Several plots provided by S-Plus are helpful in assessing the fit and assumptions of a model. Some particularly useful plots are as follows (with examples from the Chlorophyll-a example above):

1. Residuals vs. Fit (figure 8.5). This plot is used to assess whether there is unexplained structure in the model (e.g. whether you have misspecified the model or something was measured with error). Under the assumptions of the CLRM, this should appear as random noise. Sadly, it fails this test, since it appears as though small fitted values also tend to have negative residuals.

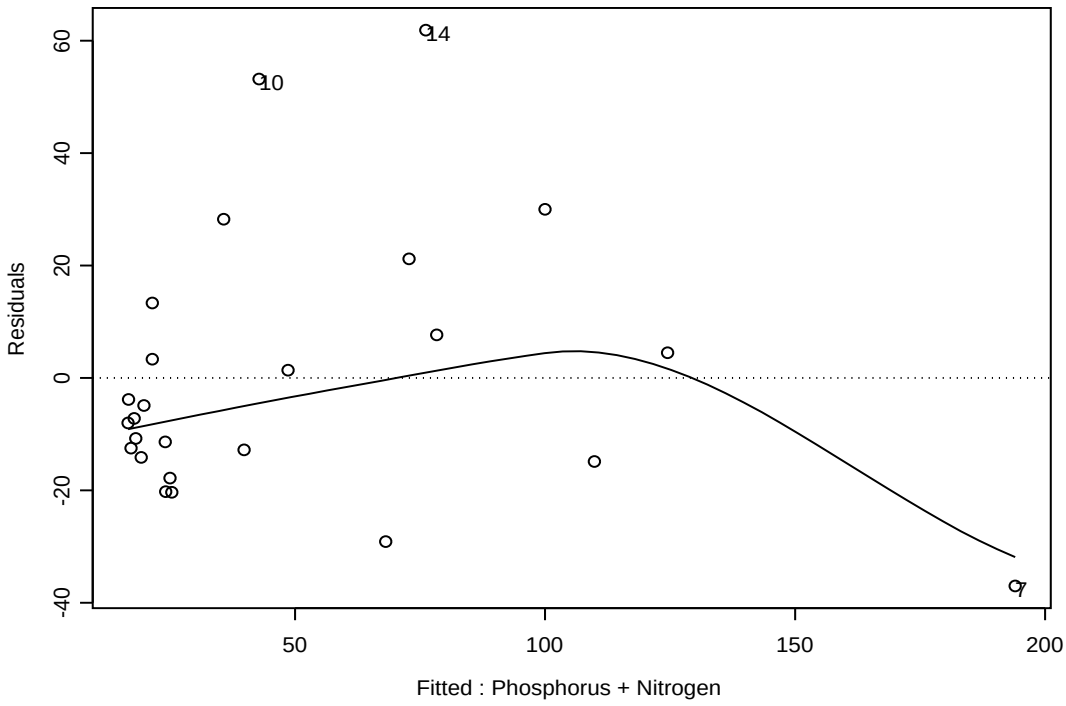


Figure 8.5: Residuals of the chlorophyll-a regression plotted against the predicted values.

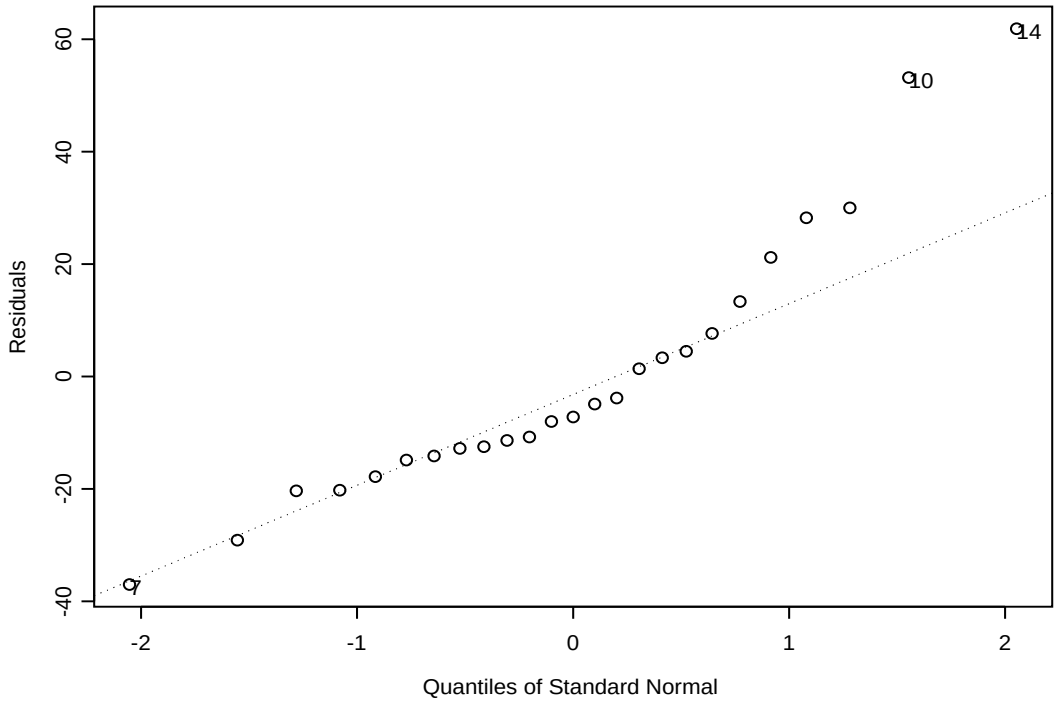


Figure 8.6: Quantile plot of residuals from the chlorophyll-a regression. If the residuals were normally distributed then the points would fall along the dotted line.

2. Normal Quantile Plot of Residuals (figure 8.6). Recall our assumption that the errors of a linear regression model are normally distributed. As introduced earlier, the residuals will fall on the diagonal line if they are normally distributed. The normal quantile plot for the chlorophyll example is in figure 8.6, which suggests that the residuals closely follow the normal distribution, except for large values, which are more likely for our residuals than they are for the normal distribution. See figure 8.7 for the histogram of our residuals. Many other plots are available,

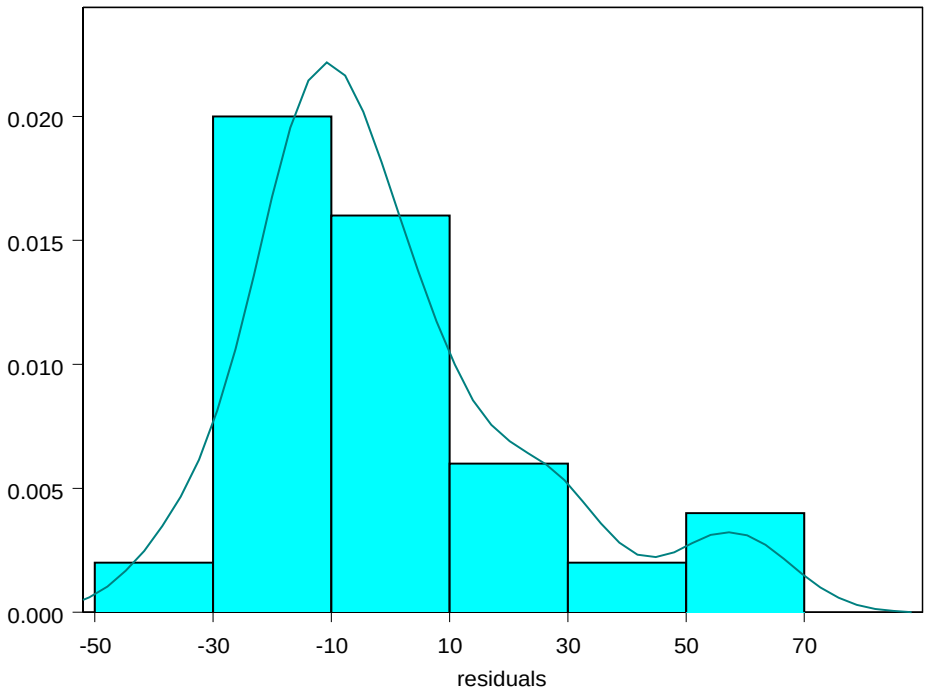


Figure 8.7: Histogram and density plot of the residuals from the chlorophyll-a regression.

and the S-Plus online help is extremely useful in this regard[4].

8.4.6 Consequences of Violating Assumptions

Violating any of the assumptions of the CLRM causes the estimates, or our confidence in the estimates, to be incorrect. The following is a brief attempt to summarize the consequences of common violations, and what can be done about it.

Problem	How Detect?	Consequence	Possible Corrections
Autocorrelation	Durbin-Watson	Unbiased; wrong inf.	GLS
Heteroskedasticity	Plot resids[5]	Unbiased; wrong inf.	GLS
Contemporaneous Corr.[6]	Plot; Hausman Test	Biased	I.V.
Multicolinearity	Correlation table	Usually OK	omit?
Omitted Variables	Theory[7]	Biased	Add variable(s)
Wrong Regressors	Theory	Unbiased; extra noise	Omit Variables
True nonlinear	Plot residuals	Wrong inf.	Non-linear model

8.4.7 A Practical Guide to Model Specification

1. Start with theory. What variables do you need? How should they enter? What is the source of error, and how does it enter? The alternative is "data mining", which refers to a situation when you allow the data to drive the functional form of the model. Data mining is appropriate in some situations, but your first line of inquiry should be into the theoretical underpinnings of the process you are trying to model. Absolutely do not fit a model to data and then do statistical inference on the same data - this is "double dipping". You may be able to split your data sample, fit the model to part and do statistical inference on the other part.

[4]See Help-Online Manuals-Guide to Statistics vol I. Then look up linear regression.

[5]Formal tests include: Goldfeld-Quandt, Breusch-Pagan, and White.

[6]When e_n is correlated with X_n. Measurement error gives rise to contemporaneous correlation. A form of measurement error exists when $E(e_i) \neq 0$, in which case the intercept term will be biased.

[7]Plotting residuals and getting to know your data may give rise to understanding of what omitted variables cause outlying $\hat{e}_i$'s.

2. Check the assumptions for the CLRM to make sure the OLS estimator is an appropriate estimator for the problem you have specified.
3. Collect and plot your data. Look for outliers, inconsistencies, etc. Get to know your data.
4. Estimate the model, and run F tests (description coming in a later lecture) to test restrictions of the model. At this point, you may want to try a Box-Cox transform (of your variables- thus keeping the model linear in parameters) to ensure the errors are normally distributed.
5. Check the R^2 statistic and the "adjusted" R^2 statistic to get an idea of how much variation your model explains.
6. Plot residuals. If there appears to be a pattern (anything other than a random scatter), you have a misspecification problem (that is, one of the assumptions of the CLRM does not hold).
7. Seek alternative explanations. What else may cause this pattern in the data?

8.5 Application of Technique

So far, we have not discussed how the estimator works, and that remains for the next set of notes. Let's decide which regressors (independent variables) we will use in our regression - we already know, by assumption (1) we must include all relevant regressors (in fact, the plot of gas consumption vs. price provides a perfect example of this - the regression line would slope up, suggesting that people purchased more gas as the price increased). From demand theory in economics, the primary constituents of demand are price, income, and prices of substitutes and compelements. To proceed, let's try to explain gas consumption per capita (G) with (1) gas price (Pg), (2) income (Y), (3) new car prices (Pnc), and (4) used car prices (Puc) in the regression. Now we need to decide on a functional form for our model. The simplest functional form is linear in the regressors:

$$G = \beta_0 + \beta_1 Pg + \beta_2 Y + \beta_3 Pnc + \beta_4 Puc + \epsilon \tag{8.4}$$

Another common specification in economics is the log-log specification:

$$log(G) = \beta_0 + \beta_1 \log(Pg) + \beta_2 \log(Y) + \beta_3 \log(Pnc) + \beta_4 \log(Puc) + \epsilon \tag{8.5}$$

One reason the log-log specification is commonly used is that the parameter estimates can be interpreted as elasticities (the estimate of β_1, called $\hat{\beta}_1$, is intepreted as the % change in gas consumption with a 1% change in gas price).

Now we'll estimate both equations above, and provide the parameter estimates and some summary statistics below:

Model	β_0	β_1	β_2	β_3	β_4	R^2	p-val: F
Linear	-0.09 (.08)	-0.04 (.002)	0.0002 (.000)	-0.10 (.11)	-0.04 (.08)	.97	.000
Log-Log	-12.34 (.000)	-0.06 (.08)	1.37 (.000)	-0.13 (.33)	-0.12 (.16)	.96	.000

Now, suppose a parameter estimate turns out to be zero. That suggests that changes in the variable to which the parameter is attached do not have any effect on the dependent variable. For example, suppose our best estimate of β_1 in equation 8.4 was 0 (instead of -0.04). That would suggest that, in fact, price has no effect on gas consumption. Since there is still variability in the data, the value in parenthesis, called the "p-value" for a particular parameter, gives the probability that the parameter is as high (or low) as the parameter estimate simply by chance. Don't worry too much about this now, but just remember that a very low p-value (say ≤ 0.05) means that we have a great deal of confidence that the coefficient is truly different from zero. A high p-value suggests that the true value of the parameter may be zero (and therefore, it's associated variable may have no effect).

The second summary statistic is the R^2 value which gives a measure of the goodness of fit of our model. R^2 values of, say, $\geq .7$ are very high (it ranges from 0 to 1), and usually mean the model fits very well. We'll return to this later.

Finally, the p-value of the F-statistic is analogous to the p-value for each parameter estimate, except that it refers to the model as a whole. Low p-values ($p < .05$ or so) for the F-statistic mean that the parameter estimates, taken as a whole, do provide some explanatory power for the dependent variable.

We'll conclude by answering our gas market question: What is the effect of a price change on the quantity of gasoline purchased. Holding all other things constant (income and car prices), we can say (using model 8.4 above) that a one unit increase in price results in about a .04 unit decrease in the quantity purchased (since the units of gas per pop are thousands of gallons, a \$0.10 increase in the gas price leads to approximately a 4 gallon decrease

in gas consumption per capita. Higher income has a positive effect on gas purchases, as expected, and the prices of both new and used cars has a negative effect on gas prices.

8.5.1 Confidence intervals for the predicted marginal effect

On the basis of a multiple linear regression analysis, we concluded that a \$0.10 increase in the gas price corresponds to about a 4 gallon decrease in per capita gas consumption. But how precise is this number? On thing we may be particularly interested in is whether we are even certain that the true value is indeed positive at all (for reasons that only show up in the error term, some people might actually drive more after the gas price increase). The question of precision of our estimate is best answered by calculating a confidence interval for the true value of this response. We'll compute a 90% confidence interval. Recall our linear model specification (where the dependent variable is G, gas consumption per capita):

$$G = \beta_0 + \beta_1 Pg + \beta_2 Y + \beta_3 Pnc + \beta_4 Puc + \epsilon \tag{8.6}$$

where, as reported in a previous set of notes, the estimate of β_1 is -.04237 with a standard error of .00984.

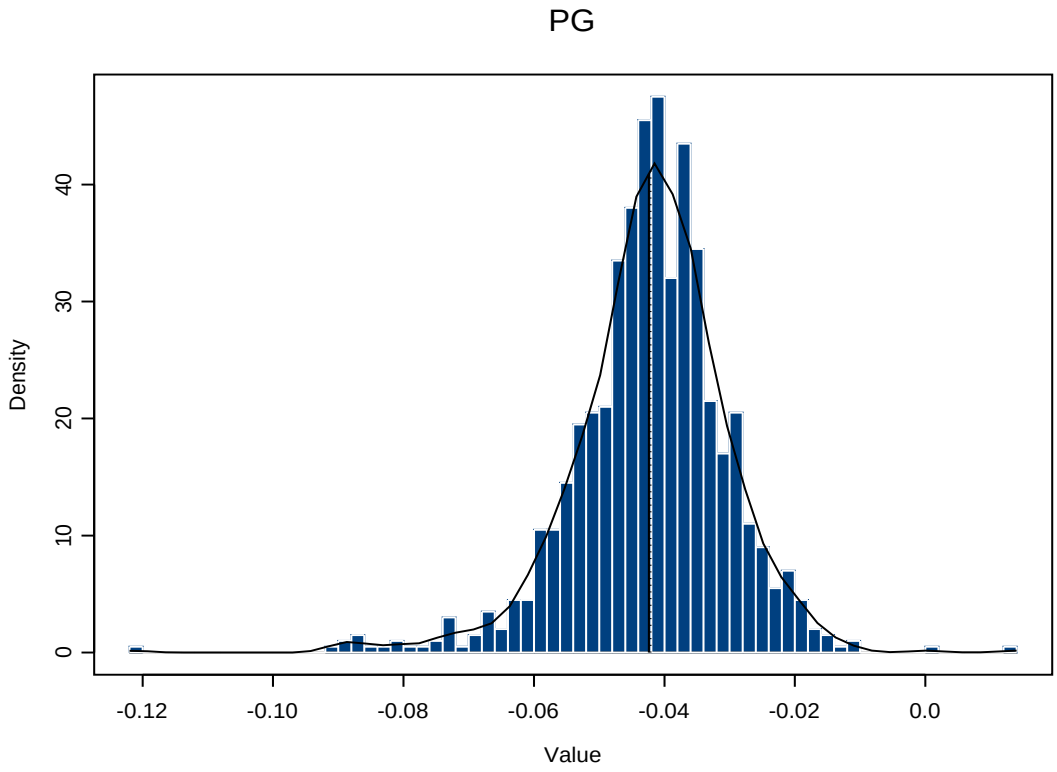
If we assume that the errors from our regression are normally distributed, then our confidence interval for any given parameter (β_1 in this case), is based on the t distribution. The critical t statistic for a 90% confidence interval with 32 degrees of freedom (number of data points (37) minus number of parameters(5)) is 1.7. The associated confidence limits are:

$$L_l = -.04237 - (1.7)(.00984) = -.0591 \tag{8.7}$$

$$L_u = -.04237 + (1.7)(.00984) = -.0256 \tag{8.8}$$

In other words, we expect the true response to be contained in the interval [-.0591, -.0256] about 90% of the time. This suggests that there really is an important negative response of consumption to gas price increases; a 10 cent increase in gas price will correspond to something like a 3 to 6 gallon decrease in per capita gas consumption.

Bootstrapping the confidence intervals in S-Plus (using the script "coef(lm(G.POP $\sim$ PG+Y+PNC+PUC))" in the Expression box), gives empirical percentile confidence interval for β_1 of [-.063, -0.026]. The distribution of β_1 from 1000 bootstrapped samples is shown in figure 8.5.1.


PG
Density
0
10
20
30
40
-0.12
-0.10
-0.08
-0.06
-0.04
-0.02
0.0
Value

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0898	0.0508	-1.7687	0.0868
GasPrice	-0.0424	0.0098	-4.3058	0.0002
Income	0.0002	0.0000	23.4189	0.0000
New.Car.Price	-0.1014	0.0617	-1.6429	0.1105
Used.Car.Price	-0.0432	0.0241	-1.7913	0.0830

Table 8.1: Splus output for the regression of per-capita gas consumption on gas price, income, and new and used car prices.

8.5.2 Significance of regression terms

Recall that the model for gas consumption was

$$G = \beta_0 + \beta_1 P + \beta_2 I + \beta_3 N + \beta_4 U, \tag{8.9}$$

where G is the per-capita gas consumption, P is the price of gas, I is the average income, N is the average price of a new car, and U is the price of a used car. Running the model in Splus returns the output in table 8.1. The column labelled `Pr(>|t|)` gives the P–value for the null hypothesis that the coefficient equals zero. Thus we might conclude that there is very strong evidence that gas price and income affect gas consumption, and rather weak evidence that car prices have an effect.

In a report you want to include much of this information in tabular format. Note that you only need two of SE, t, and P; the usual convention is SE and P. You also want to give the variables natural names, instead of whatever you had in the computer dataset, and you want to make sure that you explain the statistical test used to generate the P-values. Thus you might present something like table 8.2.

	Estimate	SE	P
Intercept	-0.090	0.051	0.09
Gas Price	-0.042	0.0098	0.0002
Income	0.0002	0.0000	¡0.0001
New Car Price	-0.10	0.062	0.1
Used Car Price	-0.043	0.024	0.08

Table 8.2: Estimates and standard errors of the coefficients in the gas consumption regression (eq. 8.9). The P–values are for a two–sided t–test against the hypothesis that the coefficient is zero ($df = 31$).