NORTHERN ARIZONA UNIVERSITY
The W. A. Franke College of Business

What Do Regressions Estimate?

Working Paper Series—08-04 | February 2008

Pin Ng
Associate Professor of Economics

James Pinto
Professor of Economics

Both of

Northern Arizona University
The W. A. Franke College of Business
PO Box 15066
Flagstaff, AZ 86011-5066

What Do Regressions Estimate?

1. Introduction

The simple linear regression analysis can be found in a majority of introductory statistics textbooks on the market today. Almost all of them minimize the least squares criterion to obtain a *sample regression line*, which in turn acts as an estimate for some *unknown population regression line*. However, it is not always clear exactly what the population regression line measures in many of these books. Similarly in many academic research and practical applications, the least squares regression (sometimes called the L_2 regression) is often used to fit through some empirical data. In this case, very little or no clue can be found to determine what the population regression line that is being estimated is measuring.

Below are some interpretations that we have extracted from only a few of the introductory business statistics textbooks on the market today that attempt to offer interpretations for the least squares regression results but fall short in their attempts. Lind, Marchal and Wathen (2006, p.387) interpreted the intercept as "It is the estimated value of Y when $X = 0$" and, in the context of estimating the beta coefficient in finance, the slope coefficient as "when the S&P index increases by 1%, the stock price will increase by 1.5%**.**" Black (2004, p.485) stated: "One interpretation of the slope in this problem is that for every unit increase in x …, there is a $40.70 increase in the cost of the flight." In Albright, Winston and Zappe (2003, p.566), we found "The slope, 0.7623, indicates that the sales index tends to increase by about 0.76 for each 1-unit increase in the promotional expenses index". In Triola (2007, p.578), we saw: "The slope b_1 in the regression equation represents the marginal change in y that occurs when x changes by one unit." Doane and Seward (2008, p. 432) provided the following interpretation for the slope: "The slope ($b_1 = 54.039$) says that for each additional hour of flight, the Piper Cheyenne consumed about 54 pounds of fuel …"

What is wrong with the above interpretations? We need to first understand "What are we estimating when we fit a regression line through a scatter plot?" to answer this question. The answer to this question will dictate how one should interpret the results. What we are estimating in a regression depends on (1) what we want to estimate in the conditional relationship and (2) what the optimization criterion that we choose to use in computing an estimate for that aspect of the conditional relationship.

The next section illustrates the issues of the various alternative aspects of the conditional relationship one can estimate when performing a regression analysis and the corresponding available estimators. Section 3 formally sets up the framework for the correct interpretations of the various regression results while the last section conjectures about the possible reasons for the misuses and misinterpretations found in common textbooks and illustrates the necessary corrections needed to make the interpretations correct.

We hope that practitioners and educators will have a better understanding of what exactly we are estimating when performing a least squares regression or any regression for that matter, will be able to make more sensible interpretation of empirical results and teach students the correct way to interpret regression results after reading this paper.

2. What Are We Estimating in a Simple Regression?

In a simple regression setting, the population regression depicts some conditional relationship between the dependent and independent variables. For example, if one is interested in the average of the dependent variable *Y* for a given value of the independent variable *X*, the population regression represents the conditional *mean* function. The least squares (LS) regression computes the sample conditional *mean* of *Y* for a specific value of *X*, and it will be a natural estimate for the conditional mean function. On the other hand, if one is interested in estimating the conditional *median* relationship, the population regression of interest becomes the conditional *median* function. In this case, the least absolute deviation (LAD) criterion will result in the sample conditional median of *Y* for a fixed *X*, which will be the more natural candidate for the sample regression as an estimate for the conditional median function. Likewise, if estimating the *quantile* or *percentile* of the population conditional relationship is of interest, then the quantile regression (QR) introduced by Koenker and Bassett (1978), which minimizes an asymmetric risk function, should be used as the sample regression.

To illustrate all these various aspects of the population conditional relationship that one might be interested at, we plotted a simulated data set in Figure 1 with 200 pairs of $(Y_i, X_i)_{i=1}^n$ on a response variable *Y* and a covariate *X* from the model

$$Y_i = \beta_0 + \beta_1 X_i + \gamma(X_i)\varepsilon_i$$

where $\beta_0 = 1$, $\beta_1 = 1$ and $X_i \sim U(0,1)$ is randomly generated from a uniform distribution between 0 and 1. The $\varepsilon_i \sim N(0,0.1^2)$ is randomly generated from a normal distribution with a mean of 0 and a standard deviation of 0.1. Heteroscedasticity in the error term is modeled as linear in the independent variable as $\gamma(X_i) = 3X_i$. Since the error distribution is symmetric and centered at 0, its mean and median are both 0. Hence, the conditional mean function coincides with the conditional median function. Also superimposed in Figure 1 are the population conditional mean function (solid red line), the sample LS regression line (dash red line), the population conditional median function (solid black line), the sample LAD regression line (dash black line), the various population conditional quantile functions (solid gray lines) and the sample QR lines (dash gray lines). Figure 2 is an enlargement of the boxed region in Figure 1.

In this scenario, both the LS regression and LAD regression provide natural estimates for the conditional mean and median functions, respectively. However, if other conditional quantile functions are of interest to the researcher, neither the LS nor LAD regressions will be appropriate. QR regressions should be used instead.

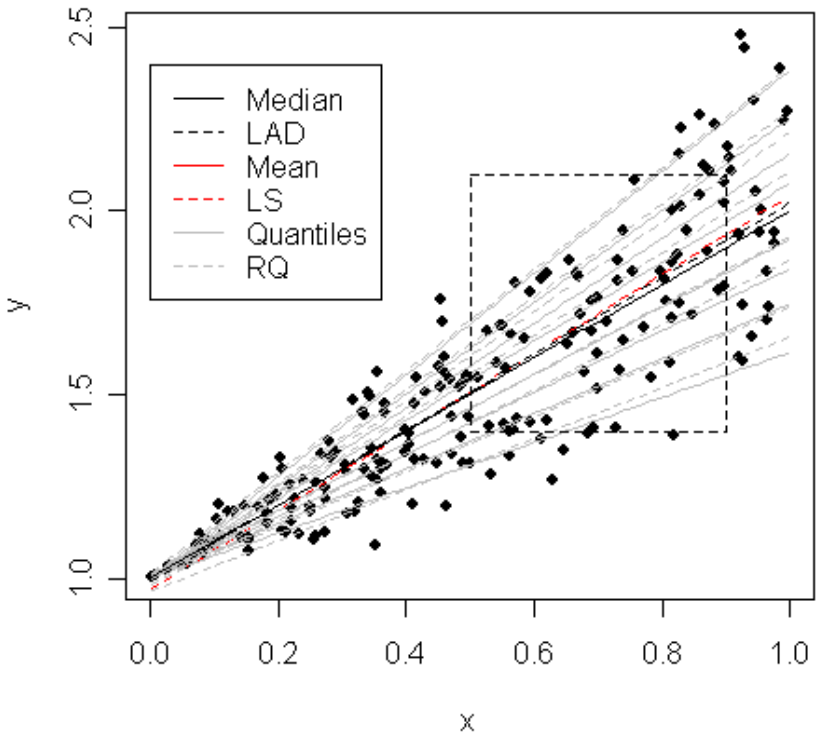

Figure 1. The data represent observations from a simple regression model with heteroscedastic symmetric errors. The black solid line represents the population conditional median function while the black dash line is the least absolute deviation regression, which is a natural estimate for the conditional median. The red solid line, which overlaps the black solid line, is the population conditional mean while the red dash line is the least squares regression, which estimates the conditional mean. The gray solid lines are the various population conditional quantile functions that are estimated by the quantile regressions represented by the gray dash lines.

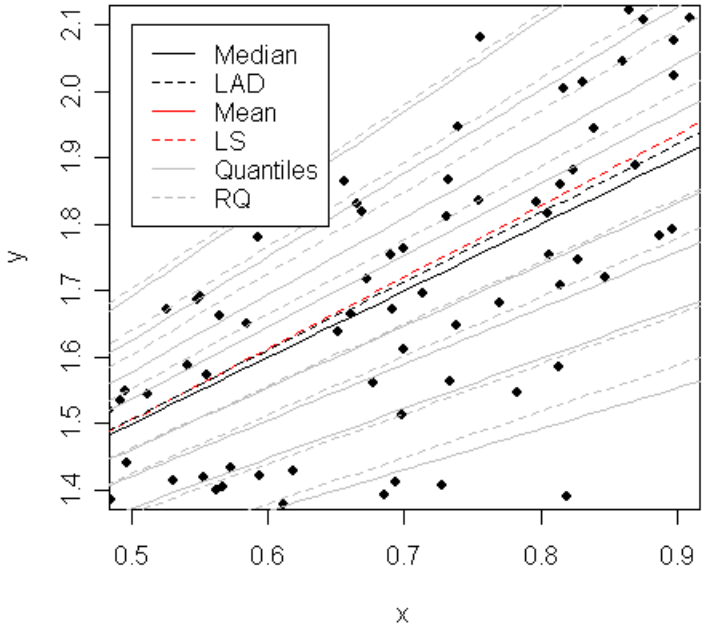

Figure 2. This is an enlargement of the boxed region in Figure 1.

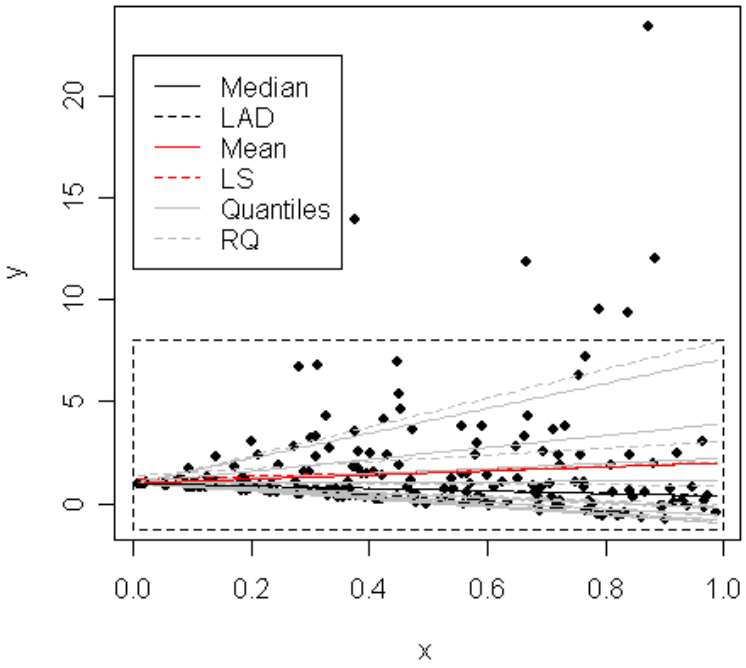

Figure 3. The data represent a simple regression model with heteroscedastic errors that are right-skewed.

The model presented in Figure 3 is similar to Figure 1 except that $\varepsilon_i \sim \chi^2(1)$ is randomly generated from a chi-square distribution with one degree-of-freedom and re-centered at its mean at 1 to illustrate the effect of a skewed error distribution on the conditional functions. Figure 4 is the magnification of the boxed region in Figure 3. As we can see from Figure 4, the conditional mean has a positive slope while the conditional median function has a negative slope. If the intention is to estimate the conditional median, the LAD regression is a natural candidate and provides a reasonably good estimate. The LS regression provides a good estimate for the conditional mean, but it is a terrible estimate for the conditional median. Again, if any of the other conditional quantile functions are of interest, one will have to use the QR regression lines.

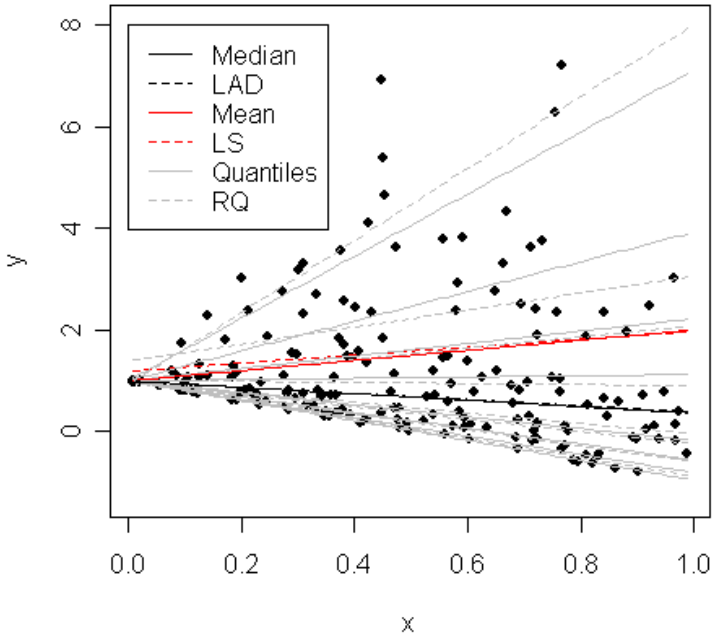

Figure 4. This is a magnification of the box-region in Figure 3.

We have seen from Figure 1 through Figure 4 that "How one should interpret the intercept and slope coefficients of the sample regression line is dependent upon what conditional relationship one is interested in estimating and what the criterion is being used in the regression optimization problem (the risk function in the context of decision theoretic paradigm)." To pinpoint exactly what has gone wrong in the interpretations in the various examples cited above, we formally set up the model in the next section.

3. The Correct Interpretations

In a univariate setting, for any real valued random variable, *Y*, with a finite second moment and characterized by its right-continuous distribution function, $F(y) = P(Y \le y)$, it is well know that the population mean, $\mu_Y = E(Y)$, minimizes the following risk function[1],

[1] See e.g. Lehmann (1983, p. 54).

$$\phi(a) = E(Y-a)^2 = \int (y-a)^2 f(y)dy = \int (y-a)^2 dF(y). \quad (1)$$

If we replace the distribution function above with the empirical distribution function, $F_n(y) = n^{-1}\sum_{i=1}^{n} I(Y_i \le y)$ with $I(\cdot)$ representing the indicator function, the empirical risk function becomes,

$$\phi_n(a) = \int (y-a)^2 dF_n(y) = n^{-1}\sum_{i=1}^{n}(Y_i - a)^2 . \quad (2)$$

This empirical risk function is minimized by the sample mean, $\bar{Y}$. We can see that the sample mean is a logical estimate for the population mean.

If the squared error loss function in (1) is replaced by the absolute error loss function, the population median, $\xi_{0.5} = F^{-1}(0.5) = \inf\{y : F(y) \ge 0.5\}$ minimizes the following new risk function,

$$\phi(a) = E|Y-a| = \int |y-a| f(y)dy = \int |y-a| dF(y). \quad (3)$$

Likewise, the sample median, $\hat{\xi}_{0.5}$ minimizes the following sample counterpart of (3),

$$\phi_n(a) = \int |y-a| dF_n(y) = n^{-1}\sum_{i=1}^{n}|Y_i - a| \quad (4)$$

and provides a logical estimate of the population median.

For an asymmetric error loss function, $\rho_\tau(u) = |u| + (2\tau - 1)u$ with $0 < \tau < 1$, the risk function,

$$\phi(a) = E\rho(Y-a) = 2(\tau-1)\int_{-\infty}^{a}(y-a)dF(y) + (2\tau)\int_{a}^{\infty}(y-a)dF(y) \quad (5)$$

is minimized by the τ-th population quantile of *Y*, $\xi_\tau = F^{-1}(\tau) = \inf\{y : F(y) \ge \tau\}$.[2] The τ-th sample quantile, which minimizes

$$\begin{aligned} \phi_n(a) &= \int \rho(y-a) dF_n(y) = n^{-1}\sum_{i=1}^{n}\rho(Y_i - a) \\ &= n^{-1}\sum_{Y_i - a > 0} 2\tau(Y_i - a) + \sum_{Y_i - a < 0} 2(\tau-1)(Y_i - a) \end{aligned} \quad (6)$$

becomes a natural estimate. Note that the loss function $\rho(u)$ assigns a weight of 2τ to the positive residuals and $2(\tau-1)$ to the negative residuals, and (6) becomes the absolute loss function in (4) when $\tau = 0.5$. The role of the loss function $\rho(u)$ can be easily visualized in Figure 5 for $\tau = 1/4$ so that

[2] Koenker (2005) provides a detailed derivation of the solution.

$\rho(u)$ assigns a weight three times as much for negative residuals as for positive residuals. In this case, $a = Y_{(2)}$ yields the smallest value for (6)-.

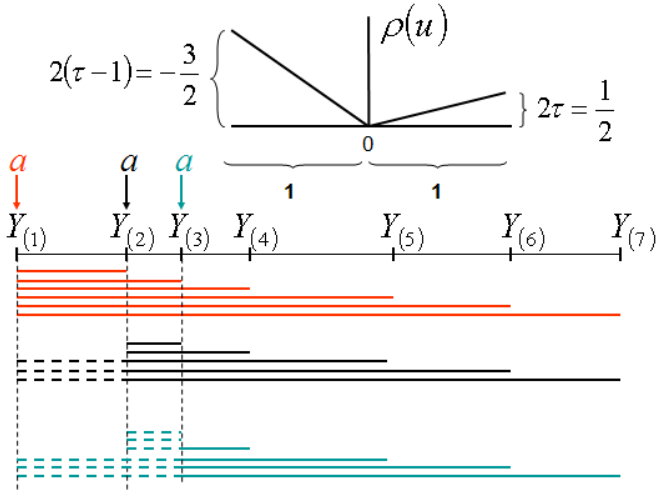

Figure 5. The risk function $\rho(u)$ for the 0.25th quantile regression assigns 3 times as much weight to negative residuals (represented by the dash line segments) as to positive residuals (represented by the solid line segments). At $a = Y_{(2)}$, the sum of the weighted residuals in black is the smallest compared to the sums in red and blue that correspond to $a = Y_{(1)}$ and $a = Y_{(3)}$, respectively.

Typically, sample median or quantiles are computed from ordering of the raw data. Figure 5 illustrates that sample quantiles can also be obtained from solving an optimization problem.

In a simple linear regression setup given *n* pairs of observations $(Y_i, X_i)_{i=1}^n$, the problem is usually posted as "finding the best estimates for the population parameters (β_0, β_1)" in the linear model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i . \qquad (7)$$

What is assumed, implicitly, is that there is an unobserved *true population regression line*, $\beta_0 + \beta_1 X_i$ with an intercept parameter β_0 and a slope parameter β_1, which describes the *true* linear relationship between the dependent variable *Y* and the explanatory variable *X*. The fact that the *n* pairs of data points (Y_i, X_i) do not fall exactly on the population regression line necessitates the introduction of the disturbance (or error) term, ε_i, into the model in (7). Usually the disturbance term is assumed to have a mean of zero and independent of the covariate *X*.

The most popular method to obtain the estimates for (β_0, β_1) is to obtain a pair of intercept and slope $(\hat{\beta}_0^{LS}, \hat{\beta}_1^{LS})$ which solves the problem

$$\min_{(\beta_0,\beta_1)\in R^2} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2 . \tag{8}$$

The least squares linear regression line, $\hat{Y}_i^{LS} = \hat{\beta}_0^{LS} + \hat{\beta}_1^{LS} X_i$ becomes an estimate of the population regression line, $\beta_0 + \beta_1 X_i$. The estimators $\left(\hat{\beta}_0^{LS}, \hat{\beta}_1^{LS}\right)$, are called the least squares estimators, because they are obtained by minimizing the least squares deviation from the fitted line to the data point in the response variable *Y* expressed in equation (7). Usually what is not explicitly addressed is the question "What does the least squares regression line, $\hat{Y}_i^{LS}$, really estimate?" The usual answer provided is "It estimates the population linear regression line." If one presses further and asks, "What exactly does this population linear regression line measure?" one will then discover that, by comparing equation (2) to (8), the population linear regression line that is being estimated by the least squares regression line is really the *conditional mean function*, $\mu_{Y|X}(X) = E(Y \mid X) = \beta_0 + \beta_1 X$, of the response *Y* given the covariate *X*. Therefore, the sample conditional mean function, $\hat{Y}_i^{LS}$, provides a natural estimate for the population conditional mean function.

If one replaces the squared error loss with the absolute error loss function and estimates (β_0, β_1) by the solutions to

$$\min_{(\beta_0,\beta_1)\in R^2} \sum_{i=1}^{n} \left|Y_i - \beta_0 - \beta_1 X\right| \tag{9}$$

one will obtain the least absolute deviation estimators $\left(\hat{\beta}_0^{LAD}, \hat{\beta}_1^{LAD}\right)$. If one carries out the questions and answers exercise above by comparing equation (4) to (9), one will realize that the LAD regression line, $\hat{Y}_i^{LAD} = \hat{\beta}_0^{LAD} + \hat{\beta}_1^{LAD} X_i$, provides an estimate for the so-called *true* population linear regression line, which is, in fact, the *conditional median* of the response *Y* given the covariate *X*. In fact, one can carry out this exercise one-step further by using the asymmetric loss function and the solutions to

$$\min_{(\beta_0,\beta_1)\in R^2} \sum_{i=1}^{n} \rho_\tau\left(Y_i - \beta_0 - \beta_1 X\right) =:$$

$$\min_{(\beta_0,\beta_1)\in R^2} \sum_{Y_i-\beta_0-\beta_1 X_i>0} 2\tau\left(Y_i - \beta_0 - \beta_1 X\right) + \sum_{Y_i-\beta_0-\beta_1 X_i<0} 2(\tau-1)\left(Y_i - \beta_0 - \beta_1 X\right)$$

as the estimators for (β_0, β_1). The estimators obtained $\left(\hat{\beta}_0^{\tau}, \hat{\beta}_1^{\tau}\right)$ are the τ-th regression quantiles, introduced in Koenker and Bassette (1978). What exactly does the regression line $\hat{\beta}_0^{\tau} + \hat{\beta}_1^{\tau} X_i$ estimate? Denoting the conditional distribution function of *Y* given *X* as $F_{Y|X}(y)$, the τ-th quantile (or 100τ-th percentile) regression line $\hat{Y}_i^{\tau} = \hat{\beta}_0^{\tau} + \hat{\beta}_1^{\tau} X_i$ provides an estimate for the τ-th *conditional quantile* (or 100τ-th *conditional percentile*) function, $F_{Y|X}^{-1}(\tau) = \beta_0^{\tau} + \beta_1^{\tau} X$ of *Y* given *X*, assuming that the τ-th conditional quantile is linear in *X*. That is, 100τ % of the data points will lie above the τ-th quantile

regression line $\hat{\beta}_0^{\tau} + \hat{\beta}_1^{\tau} X_i$ while $100(1-\tau)\%$ will fall below it. Note that the LAD estimators are just a special case of the τ-th regression quantiles when $\tau = 0.5$, so half of the data points fall above the LAD regression line while the other half fall below it.

It should now be obvious that the so-called true population linear regression line depends on which loss function is being used in the risk function to define the estimates of β_0 and β_1. When the squared loss function is used in the optimization problem, the resulting sample regression line provides an estimate of the conditional mean function. The sample regression line is an estimate of the conditional median function if the absolute loss function is used, and it becomes the quantile regression function and provides an estimate of the conditional quantile function if the asymmetric loss function $\rho(u)$ is used.

4. Misuses and Misinterpretations

4.1 Misuses of Estimators

In the previous sections, we have demonstrated that "The sample regression line estimate is really determined by the loss function that has been used to define the optimization problem." Given the same data set, the LS estimators can very well be estimating a very different *true population regression line* than the LAD estimators. If the disturbance term has an asymmetric distribution, and the intention is to estimate the conditional median, then LAD estimators should be used instead of the LS estimators. The LS estimators should be used if the intention is to estimate the conditional mean function if there are no outliers in the dependent variable. If there are outliers in the response variable *Y* and the intention is to estimate the measure of central tendency, the LAD estimators will be preferable to the LS estimators due to its robustness property.

Suppose we are given a data set in which the response variable is income and the covariate is years of education, and we are interested in the behavior of the upper quantile income, say the 95th percentile, given a specific years of education. We will naturally want to compute the 0.95-th regression quantiles $(\hat{\beta}_0^{0.95}, \hat{\beta}_1^{0.95})$ and use the .95-th quantile regression line $\hat{\beta}_0^{0.95} + \hat{\beta}_1^{0.95} X_i$ as the estimate for the 95th percentile income line instead of using the LS or LAD regression lines to estimate the *center* behavior. In all cases, the LS regression should not be used blindly as in a Povlovian fashion to estimate any linear relationship between the response and a covariate.

4.2 Misinterpretations

Since LS estimators are the most widely used estimators, we will focus on some common mistakes made in interpreting the estimating results obtained from the LS regression.

After a LS regression line is computed, a common incorrect interpretation that we have often encountered is "Given the value of the explanatory variable $X = X_i$, the estimated value of the dependent variable *Y* equals $\hat{Y}_i^{LS} = \hat{\beta}_0^{LS} + \hat{\beta}_1^{LS} X_i$." We have already shown in Section 3 that the LS regression line $\hat{Y}_i^{LS} = \hat{\beta}_0^{LS} + \hat{\beta}_1^{LS} X_i$ estimates the conditional mean of *Y* for a given value of $X = X_i$. Therefore, the correct interpretation should be "Given the value of explanatory variable $X = X_i$, the

estimated **average** value of the dependent variable *Y* equals $\hat{\beta}_0^{LS} + \hat{\beta}_1^{LS} X_i$." The response variable *Y* is a random variable, and we do not estimate a random variable. We estimate a parameter or, to be more exact, a conditional function in this case. It is, however, perfectly all right to interpret the fitted regression value as "The predicted value of the dependent variable *Y* equals $\hat{Y}_i^{LS} = \hat{\beta}_0^{LS} + \hat{\beta}_1^{LS} X_i$ given the value of explanatory variable $X = X_i$", because it makes sense to predict the future value of a random variable *Y*. Likewise, the LAD regression line, $\hat{Y}_i^{LAD} = \hat{\beta}_0^{LAD} + \hat{\beta}_1^{LAD} X_i$, provides an estimate of the conditional median for a specific value of the covariate, *X* while the QR regression line, $\hat{\beta}_0^{\tau} + \hat{\beta}_1^{\tau} X_i$, provides estimate for the τ -th conditional quantile.

Another common mistake that we have encountered occurs in the interpretation of the estimators $\left(\hat{\beta}_0^{LS}, \hat{\beta}_1^{LS}\right)$. Sometimes the claim is made that "$\hat{\beta}_0^{LS}$ estimates the value of the response variable *Y* when the covariate $X = 0$, and $\hat{\beta}_1^{LS}$ estimates the change in *Y* as a result of an incremental change in *X*." Again, given the fact that the LS regression line estimates the conditional mean of *Y* instead of the value of *Y*, the correct interpretation should be "$\hat{\beta}_0^{LS}$ estimates the **expected** value of the response variable *Y* when the covariate $X = 0$ (with the usual caveat that only if it makes sense to have *X* = 0 in the context of the problem) and $\hat{\beta}_1^{LS}$ provides an estimate for the change in the **expected** values of *Y* as a result of an incremental change in *X*." We postulate that a likely cause for such common misinterpretations is that when one takes the partial derivative of equation (7) with respect to *X*, the slope parameter is $\beta_1 = \frac{\partial Y_i}{\partial X_i}$. However, equation (7) depicts the relationship between the *n* observation pairs $\left(Y_i, X_i\right)$ not the functional relationship between the random variables *Y* and *X*. Therefore, one should take the partial derivative of the conditional mean function $\mu_{Y|X}\left(X\right) = E\left(Y \mid X\right) = \beta_0 + \beta_1 X$ with respect to *X* instead and obtain $\beta_1 = \frac{\partial E\left(Y \mid X\right)}{\partial X}$. The slope parameter should then be interpreted as the change in the expected value of *Y* because of an incremental change in *X*.

Similarly, the LAD estimators $\hat{\beta}_0^{LAD}$ should be interpreted as "The estimated median of the response variable *Y* when the covariate *X* is 0 (only if it makes sense for *X* = 0). In addition, $\hat{\beta}_1^{LAD}$ should be interpreted as "The estimated change in the median of the dependent variable as a result of an incremental change in the independent variable."

So exactly what are the mistakes in the examples cited in the Introduction? The **bolded** text inside the parentheses below represents the missing components in the interpretations that we fill in.

For Lind, Marchal and Wathen (2006, p.387), $\hat{\beta}_0^{LS}$ should be interpreted as "It is the estimated **(average)** value of *Y* when *X* = 0" and the slope as "when the S&P index increases by 1%, the stock price will increase by **(an estimated)** 1.5% **(on average).**" In Black (2004, p.485), "One interpretation of the slope in this problem is that for every unit increase in *x* …, there is a**(n estimated)** $40.70 increase in the

(average) cost of the flight." In Albright, Winston and Zappe (2003, p.566), "The slope, 0.7623, indicates that the sales index tends to increase by **(an estimated average of)** about 0.76 for each 1-unit increase in the promotional expenses index". Triola (2007, p.578) should interpret $\hat{\beta}_1^{LS}$ as "The slope b_1 in the regression equation represents the **(estimated)** marginal change in **(the average value of)** *y* that occurs when *x* changes by one unit." In Doane and Seward (2008, p. 432), "The slope ($b_1 = 54.039$) says that for each additional hour of flight, the Piper Cheyenne consumed a**(n estimated average of)** about 54 pounds of fuel …"

The examples in the prior paragraph represent only a portion of the wrong interpretations we have encountered. Of the 30 books randomly selected from the professional libraries of the authors, fifteen or 50% were found to contain incorrect interpretations. Such a high percentage is alarming, because the understanding of this topic by beginning business statistics students is, at best, incomplete.

5 Conclusion

We have demonstrated in the paper that the solutions to the intercept and slope coefficients in a simple regression model depend on the definition of the loss function used in the optimization problem. There is not a universal *population linear regression line*. If a squared error loss function is used, the population regression line being estimated turns out to be the conditional mean function while an absolute loss function yields the conditional median as the estimated population regression line. When an asymmetric loss function as that defined in Koenker and Bassett (1978) is used, the population regression line being estimated becomes the conditional quantile (percentile) function.

Also highlighted are a few mistakes commonly made in interpreting the least squares regression results, and we have discussed some potential misuses of the least square regressions. We hope that by pointing out these specific mistakes, practitioners and educators will have a better understanding of what exactly we are estimating when we perform a least square regression or any regression for that matter.

References

Albright, S. C., W. L. Winston and C. Zappe (2003), *Data Analysis & Decision Making with Microsoft Excel*, 2nd edition, Thomson Brooks/Cole.

Black, K. (2004), *Business Statistics For Contemporary Decision Making*, 4th edition, Wiley.

Doane, David P. and Lori E. Seward, *Essential Statistics in Business and Economics*, New York: McGraw-Hill Irwin.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press.

Lehmann, E. L. (1983), *Theory of Point Estimation*, John Wiley & Sons.

Lind, D. A., W. G. Marchal and S. A. Wathen (2006), *Basic Statistics for Business and Economics*, 5th edition, Boston: McGraw-Hill Irwin.

Triola, M. F. (2007), *Elementary Statistics Using Excel*, 3rd edition, Pearson/Addison Wesley.