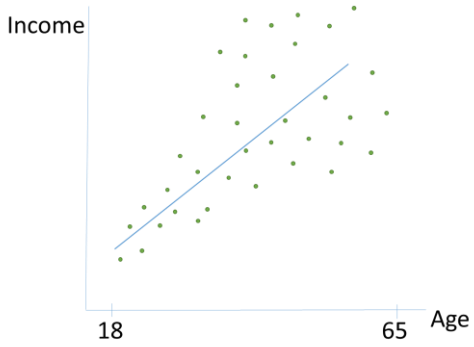
Prof. C. M. Dalton
ECN 209A
Spring 2015
Practice Problems (After HW1, HW2, before HW3)

CORRECTED VERSION

Question 1. Draw and describe a relationship with heteroskedastic errors. Support your claim with a brief explanation of the relationship between two variables and draw a representative scatterplot.

Income and Age

The relationship between age and income likely has heteroskedastic errors. At young ages, just after high school, the differences between income levels are small, but as people get older, the difference can increase dramatically. As a results, the variance of the error around the OLS line will get larger as Age increases.



Question 2

a) List and define all assumptions for multiple OLS regression.
These are all listed in section 6.5

b) Do heteroskedastic errors violate any of the above assumptions? Explain.
See p160. Not a problem for the β, but you will want to use heteroskedasticity-robust standard errors to make sure your standard errors are correct. (Stata: "regress indvar depvar, robust")

Question 3

a) Define adjusted R-squared. (formula)

P197

b) What does adjusted R-squared account for compared with standard R-squared? E.g. Why do we bother adjusting?

The fact that adding more regressors will automatically increase the amount of variation you explain. It discounts the R^2 to account for that.

Question 4 The cost of attending your college has once again gone up. Although you have been told that education is investment in human capital, which carries a return of roughly 10% a year, you (and your parents) are not pleased. One of the administrators at your university does not make the situation better by telling you that you pay more because the reputation of your institution is better than that of others. To investigate this hypothesis, you collect data randomly for 100 national universities and liberal arts colleges from the 2000-2001 *U.S. News and World Report* annual rankings. Next you perform the following regression

$$\underset{}{\textit{Cost}} = \underset{(580.2)}{7{,}311.17} + \underset{(524.4)}{3{,}985.20} \times \textit{Reputation} - \underset{(0.03)}{0.20} \times \textit{Size} + \underset{(3502.7)}{8{,}406.79} \times \textit{Dpriv} - \underset{(204.1)}{416.38} \times \textit{Dlibart} - \underset{(1200.0)}{2{,}376.51} \times \textit{Dreligion}$$

R^2=0.72, *SER* = 3,773.35 and coefficient standard errors are in ().

where *Cost* is Tuition, Fees, Room and Board in dollars,
Reputation is the index used in *U.S. News and World Report* (based on a survey of university presidents and chief academic officers), which ranges from 1 ("marginal") to 5 ("distinguished"),
Size is the number of undergraduate students, and
Dpriv, *Dlibart*, and *Dreligion* are binary variables indicating whether the institution is private, a liberal arts college, and has a religious affiliation.

a. Interpret the results. Do the coefficients have the expected sign?
HERE IS THE CHANGE
An increase in reputation by one category, increases the cost by roughly $3,985. (Significant at 1%, t=7.6)

The increasing the size of the college/university by 1000 students, the lowers the cost by $200. (Significant at 1%, t=6.67) OR An increase of 10,000 students results in a $2,000 lower cost. (Significant at 1%, t=6.67) (Whichever you think is a more interesting way to tell the story—increasing by 1 student changes the cost by 20 cents is not very compelling, and probably not the decision a policy maker is thinking about.)

Private schools charge roughly $8,406 more than public schools. (Significant at 5%, t=2.4)

A school with a religious affiliation is approximately $2,376 cheaper, presumably due to subsidies, (Significant at 5%, but not 1%, with t=1.98) and a liberal arts college also charges roughly $416 less. (Significant at 5% but not 1% with t=2.04)

There are no observations close to the origin, so there is no direct interpretation of the intercept. Other than perhaps the coefficient on liberal arts colleges, all coefficients have the expected sign.

b. What is the forecasted cost for a liberal arts college, which has no religious affiliation, a size of 1,500 students and a reputation level of 4.5? (All liberal arts colleges are private.)

Approximately $32,935.

c. To save money, you are willing to switch from a private university to a public university, which has a ranking of 0.5 less and 10,000 more students. What is the effect on your cost? Is it substantial (support this)?

Roughly $12,400. (New predicted cost is $20,536) Over four years of education, this implies nearly $50,000, it is a substantial amount of money for the average household.
CORRECTION: if all liberal arts colleges are private, then the public institute cannot be liberal arts. Thus, including also liberal arts dummy= 0, this value should be $20,951.)

d. Do you have a reason to suspect imperfect multicollinearity in the independent variables above? Why or why not? Describe what happens to an independent variable's estimated coefficient if imperfect multicollinearity is present.

We've already observed that all liberal arts colleges are private. Not all private colleges are liberal arts (ie MIT) but these variables are an example of imperfect multicollinearity. (This also might be true for religious affiliation.)

If multicollinearity is present, see bottom of p205. Here, it might be contributing to why the t-stats on private and liberal arts are smaller than others in the regression.

e. Eliminating the *Size* and *Dlibart* variables from your regression, the estimation regression becomes

Cost = 5,450.35 + 3,538.84 × *Reputation* + 10,935.70 × *Dpriv* – 2,783.31 × *Dreligion*;
R^2=0.68, *SER* = 3,792.68

Why do you think that the effect of attending a private institution has increased now?

Private institutions are smaller, on average, and some of these are liberal arts colleges. Both of these variables had negative coefficients.

This is due to omitted variable bias. The ρ_{xu} is negative, because 1. Correlation between private and size is negative (private turns to 1, size decreases) 2. Correlation between cost and size is negative (as size increases, cost decreases) The bias-inducing term is positive, which means the estimated coefficient on private now is being estimated "plus some stuff" The term is too high because of the omitted variable of size messing everything up!

f. Describe a variable that you could put into the regression (but won't!) that would be perfectly collinear with another variable(s) in the original regression. Also, why won't you add it?!

For example, adding a variable = 1 if a university/college is public. Perfectly multicollinear with private.
AHHH DUMMY VARIABLE TRAP!

These questions deal with Chpt 8, Nonlinear Regression

Question 5 Sports economics typically looks at winning percentages of sports teams as one of various outputs, and estimates production functions by analyzing the relationship between the winning percentage and inputs. In Major League Baseball (MLB), the determinants of winning are quality pitching and batting. All 30 MLB teams for the 1999 season. Pitching quality is approximated by "Team Earned Run Average" (ERA), and hitting quality by "On Base Plus Slugging Percentage" (OPS).

Summary of the Distribution of Winning Percentage, On Base Plus Slugging Percentage, and Team Earned Run Average for MLB in 1999

	Average	**Standard deviation**	**Percentile**						
			10%	**25%**	**40%**	**50% (median)**	**60%**	**75%**	**90%**
Team ERA	4.71	0.53	3.84	4.35	4.72	4.78	4.91	5.06	5.25
OPS	0.778	0.034	0.720	0.754	0.769	0.780	0.790	0.798	0.820
Winning Percentage	0.50	0.08	0.40	0.43	0.46	0.48	0.49	0.59	0.60

Your regression output is:

$$\widehat{Winpct} = -0.19 - 0.099 \times teamera + 1.490 \times ops$$
$$(0.08)\ (0.008) \qquad (0.126)$$

R^2=0.92, SER = 0.02.

(a) Interpret the regression. Are the results statistically significant and important?

Lowering the team ERA by one results in a winning percentage increase of roughly ten percent.
Increasing the OPS by 0.1 generates a higher winning percentage of approximately 15 percent.
The regression explains 92 percent of the variation in winning percentages.
Both slope coefficients are statistically significant, and given the small differences in winning percentage, they are also important.

(b) There are two leagues in MLB, the American League (AL) and the National League (NL). One major difference is that the pitcher in the AL does not have to bat. Instead there is a "designated hitter" in the hitting line-up. You are concerned that, as a result, there is a different effect of pitching and hitting in the AL from the NL. To test this hypothesis, you allow the AL regression to have a different intercept and different slopes from the NL regression. You therefore create a binary variable for the American League (*DAL*) and estimate the following specification:

$$\widehat{Winpct} = -0.29 + 0.10 \times DAL - 0.100 \times teamera + 0.008 \times (DAL \times teamera) + 1.622 * ops - 0.187 * (DAL \times ops)$$
$$(0.12)\ (0.24) \qquad (0.008) \qquad (0.018) \qquad (0.163) \qquad (0.160)$$

With R^2=0.92, SER = 0.02.

What is the regression for winning percentage in the AL and NL?
NL: $\widehat{Winpct} = -0.29 - 0.100 \times teamera + 1.622 \times ops$
AL : $\widehat{Winpct} = -0.19 - 0.092 \times teamera + 1.435 \times ops$.

Next, calculate the t-statistics and say something about the statistical significance of the AL variables. Since you have allowed all slopes and the intercept to vary between the two leagues, what would the results imply if all coefficients involving DAL were statistically significant?

The t-statistics for all variables involving DAL are, in order of appearance in the above regression, 0.42, 0.44, and –1.17. None of the coefficients is statistically significant individually. If these were statistically significant, then this would indicate that the coefficients vary between the two leagues. Hence these coefficients might suggest that the introduction of the designated hitter might not have changed the relationship.

Question 6 Suppose you have the following result when you run a scatterplot of your dependent variable with one of your independent variable:

a) Give two suggestions of how you might modify your OLS regression equation from the base of $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

1. *You could include a squared term of X_1 in addition to X_1 itself. This would allow the effects of X_1 to have a curved shape, instead of strictly linear.*
   $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X^2_{1i_{1i}} + \beta_3 X_{2i} + u_i$$
2. *You could use the log function to transform the X_1 variable.*
   $$Y_i = \beta_0 + \beta_1 \ln(X_{1i}) + \beta_2 X_{2i} + u_i$$
   Now, instead of a one unit change in X_1, which would mean that an increase by one unit has the same effect everywhere, we are accounting for a percentage change. A one unit change in the beginning has a different percentage change than one unit at the upper part of the distribution. Now the effect is allowed to change more like the picture.

b) For both of the suggestions above, explain how to interpret the effect on Y of a change in X_1.
1. *Polynomial- because the regression function is quadratic, this effect depends on the initial district income. Give an initial value of X_1 and make a one unit change. Then plug each of these into: $\beta_0 + \beta_1 X_1 + \beta_2 (X_1)^2$ and $\beta_0 + \beta_1 (X_1 + 1) + \beta_2 (X_1 + 1)^2$. Calculate the difference.*
2. *A log-linear model, where a 1% change in X_1 is associated with a change in Y of 0.01xβ_1*