

Homework #9 – Due Monday, Nov. 30
COR1-GB.1305 – Statistics and Data Analysis

The file `Magazine.CSV` contains data on advertising costs and characteristics of magazines. The response variable is `PageCost`, which represents the cost of a full-page color ad in the magazine. `Circ` is the circulation of the magazine (in thousands), `MedIncome` is the median income of the readers, and `%Male` is the percentage of the readers who are male. The square root of the circulation is given in `SqrtCirc`.

In this problem, we will fit a regression model to predict the mean page cost of a magazine with `Circ` of 10000, `MedIncome` of \$40,000, and `%Male` of 50. We will not necessarily use all three predictors, only the ones that are useful for predicting `PageCost`.

- (a) First, we will find an appropriate set of predictor variables.
 - (i) Run a multiple regression of `PageCost` on the original predictor variables (`Circ`, `MedIncome` and `%Male`). Before running it, click on Graphs, and check the box for Residuals plots: Four in one. Note that the residuals versus fit plot shows structure: a generally upward-sloping pattern, with three outliers at the right dragging things down. Identify the Magazines corresponding to the three outliers (all of which have a very large circulation).
 - (ii) To investigate further, generate a scatterplot of `PageCost` versus `Circ`. Note that the plot is “bunched up” at the left, and “stretched out” at the right, and also a bit curved. In what way do the points identified as outliers in (a) deviate from the pattern in the plot here?
 - (iii) To try to improve the linear relationship, let’s try working with the square root of Circulation (`SqrtCirc`) rather than the circulation itself. Plot `PageCost` versus `SqrtCirc`. Based on the plot, explain why it seems more appropriate to use `SqrtCirc` as an explanatory variable in a linear regression rather than `Circ`.
 - (iv) Now, run a multiple regression of `PageCost` on `SqrtCirc`, `MedIncome` and `%Male`. Plot the residuals versus fitted values. Does it look better than in (i)?
- (b) Next we will investigate the regression model of `PageCost` on `SqrtCirc`, `MedIncome` and `%Male`.
 - (i) Based on the p -value for the Analysis of Variance F test for this model, does the regression seem to be useful for predicting `PageCost`? Does this mean that all variables are useful?
 - (ii) Which coefficients in the regression are statistically significant?
 - (iii) Based on the p -values for the regression coefficients, which variables seem to be useless for predicting `PageCost`?
- (c) Now, we will try to simplify the model by deleting useless predictors. Re-run the regression for `PageCost`, this time with just the two significant explanatory variables you found in part (b).
 - (i) Did the R^2 go down by much compared to the regression in (b)? Is the F -statistic still significant? What does this suggest about the deleted predictor variable?
 - (ii) Are the coefficients of both variables statistically significant?

- (d) Finally, we will use the simplified multiple regression model to predict **PageCost**.
- (i) Get a 95% confidence interval for the mean page cost of a magazine with a **SqrtCirc** of 100, and a median income of \$40,000. To do this, after running the regression click on *Stat* \Rightarrow *Regression* \Rightarrow *Regression* \Rightarrow *Predict*. Then, enter 100 in the first line under **SqrtCirc** and enter 40000 in the first line under **MedIncome**.
 - (ii) Report the 95% prediction interval.
 - (iii) What is the difference between the prediction interval and the confidence interval?