Homework 2

Advanced Methods for Data Analysis (36-402)

Due Friday February 1, 2019 *at 3:00 pm* according to Canvas

See the syllabus for general instructions regarding homework; note that you should **always show all your work** and submit a writeup with R code.

1 A Refresher in Linear Regression

The data for this problem are in two files named `housetrain.csv` (containing training data) and `housetest.csv` (containing test data.) They are both comma-separated files with headers. You will need to read each of them into an *R* `data.frame`. The `read.csv` command will do this, if you get the syntax correct.

The data are from a census survey from several years ago. Each record (line in the file) corresponds to a small area called a *census tract*. The variables that appear in the data file have the following names:

- `Population`: The population of the census tract.
- `Latitude`: The number of degrees north of the equator where the census tract is located. South is negative, so latitude is between -90 and 90.
- `Longitude`: The number of degrees east of Greenwich where the census tract is located. West is negative, so longitude is between -180 and 180.
- `Median_house_value`: The median assessed value of houses (in thousands of dollars) in the census tract.
- `Median_household_income`: The median household income in the census tract.
- `Mean_household_income`: The average household income in the census tract.

The data are from census tracts in California and Pennsylvania.

The main goal of this problem is to model the relationship, if any, between `Median_house_value` (the response) and the other variables (potential predictors.) For all plots, use the option `,pch="."` because the default circles will overlap too much with such large data sets.

(a) Compute the correlation matrix between all of the variables in the training data set. Which potential predictors are most highly correlated (positively or negatively) with the response?

(b) Use the training data to fit the following models:

Model 0: The *null model* which says that conditional on all of the potential predictors, the values of Y_j are independent and identically distributed with some mean μ.

Model 1: A simple linear regression of the response on `Median_household_income`.

Model 2: A simple linear regression of the response on `Mean_household_income`.

Model 3: A multiple regression of the response on *both* `Median_household_income` and `Mean_household_income`.

Model 4: A regression of the response on `Median_household_income`, `Mean_household_income` and 5 simulated covariates of your choice that are *independent* of the response. Briefly mention how you simulated these extra covariates.

For each model, include a summary and do a residual analysis. Based on the analysis, do you think the underlying assumptions are reasonable? Give a brief justification.

(c) Explain why the coefficients of `Median_household_income` and `Mean_household_income` in Model 3 are both different from the coefficients of the same predictors in Models 1 and 2.

(d) How does the *training error* for Model 4 compare to the *training error* in Model 3. What do you expect that would happen to the training error if you include more and more covariates that are independent of the response?

(e) Compute the "test error" for each model (average squared error from predicting the test responses using the test predictors and the model fit with training data.) Which of the models are best? Discuss based on the test error and the analysis from Question (b).

(f) (*Extra Credit Question*) Linear regression is about projecting the response vector Y onto the space spanned by linear combinations of the $\mathbf{X}$ variables. So the residual vector ϵ is orthogonal to the fitted values $\widehat{Y} = HY$, where $H = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ (it is called the "hat matrix" or the "projection matrix"). Try the following:

1. Fit a regression of `Median_house_value`(Y) on `Population`(X1) and `Median_household_income`(X2). Call the residuals $\widehat{\epsilon}_1$.
2. Fit a regression of `Mean_household_income`(X3) on `Population`(X1) and `Median_household_income`(X2). Call the residuals $\widehat{\epsilon}_2$.
3. Fit a regression of $\widehat{\epsilon}_1$ on $\widehat{\epsilon}_2$.
4. Fit a regression of `Median_house_value`(Y) on `Population`(X1), `Median_household_income`(X2) and `Mean_household_income`(X3).

What do you notice? Can you intuitively explain why this is the case?

2 Relaxing Our Regression Assumptions

Consider arbitrary random variables $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$ with *absolutely no assumptions relating the two*, and consider linearly regressing Y on X (in the population), with regression coefficients defined by

$$\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X,Y), \quad \beta_0 = \mathbb{E}(Y) - \beta^T\mathbb{E}(X).$$

Conditional on X, our prediction for Y is hence $\beta_0 + \beta^T X$.

(a) Define the residual error $\epsilon = Y - \beta_0 - \beta^T X$. Prove that ϵ has mean zero, $\mathbb{E}(\epsilon) = 0$. Again, you are only allowed to use the definitions of β and β_0 above and properties of expectations.

(b) Prove that ϵ is uncorrelated with the predictor variables, $\mathrm{Cov}(\epsilon, X) = 0$.
Hint: Make sure you remember how to manipulate matrices and their transposes, and review Appendix F in Shalizi. You can use that $\mathrm{Cov}(Y,X)^T = \mathrm{Cov}(X,Y)$ *without a proof.*

(c) By construction, we have the relationship $Y = \beta_0 + \beta^T X + \epsilon$, i.e., we've written Y as a linear function of X plus an error term ϵ. This error term has mean zero by part (a). Does part (b) imply that the error term is independent of X? What in particular does this mean about the conditional variance $\mathrm{Var}(\epsilon|X)$? Need this be constant with X?

(d) Consider i.i.d. data (X_i, Y_i), $i = 1, \ldots n$, each with the same distribution as (X, Y). For simplicity you may assume from now on that $\mathbb{E}(X) = \mathbb{E}(Y) = 0$ (though this is not really necessary). Use the same sample notation as in Lecture, i.e., $\vec{Y} = (Y_1, \ldots Y_n)^T \in \mathbb{R}^n$ for the vector of outcomes, and

$$\mathbb{X} = \begin{bmatrix} X_1^T \\ X_2^T \\ \ldots \\ X_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$$

for the design matrix for a sample of size n.

Consider the least squares estimator

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\vec{Y}.$$

Compute $\mathbb{E}(\hat{\beta}|X^n)$, where $X^n = \{X_1, ...X_n\}$ is the set of inputs. Is $\mathbb{E}(\hat{\beta}|X^n)$ necessarily equal to $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X,Y)$? If not, under what assumptions will it be?
Hint: Conditional on the inputs X^n*, you can treat the design matrix* $\mathbb{X}$ *as a constant matrix; this is the*

"fixed design" setting.

(e) Compute $\text{Var}(\hat{\beta}|X^n)$. In your formula, you can denote the variance of $\vec{\epsilon} = \vec{Y} - \mathbb{X}\beta$ conditional on $X_1, \ldots, X_n$ by $\text{Var}(\vec{\epsilon}|X^n) = \Sigma$. What does your formula reduce to in the case that $\Sigma = \sigma^2 I$, where I is the $n \times n$ identity matrix (i.e., the case where the ϵ_i's have the same variance)?

(f) In this part, we investigate what happens when we do not include a variable in the model, and this variable is related to both Y and other covariates. In particular, suppose that $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, but we only observe (Y, X_1); so we run a linear regression of Y on X_1 and compute the ordinary least squares estimator $\widehat{\beta}_1$. Prove that $\widehat{\beta}_1$ is generally not a consistent estimator of β_1. Under what conditions would $\widehat{\beta}_1$ be consistent?

Recall that $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ *means that* $\forall \epsilon > 0$, $\mathbb{P}\left(|\widehat{\beta}_1 - \beta| > \epsilon\right) \to 0$ *as* $n \to \infty$.
Hints: You can proceed as follows:

1. *Identify what the estimator* $\widehat{\beta}_1$ *is in terms of* X_1, X_2 *and* ϵ.

2. *Can you think of a* ***very famous*** *law that would be useful to show convergence in probability? You may assume that if* $X_n \xrightarrow{p} X$, *then* $g(X_n) \to g(X)$ *for any continuous function (i.e. ratio). This is the so called continuous mapping theorem:* https://en.wikipedia.org/wiki/Continuous_mapping_theorem.