Problem Set 03
Linear Models – Fall 2015
Due date: 2015-10-07

Problems sets are due at the start of class on the due date. Please hand write or type up and print the solutions; we will not accept e-mail solution sets except in exceptional circumstances. You may discuss problem sets with others, but must write up your own solutions. This means that you should have no need to look at other's final written solutions. Many of these problems come from a variety of textbooks, which are referenced in the problems. These are for citation purposes and not because you will need to consult the text itself (though you may feel free to do so).

I. General linear model

Consider the case where the spherical errors assumption is violated, and the variance of ϵ is instead given by:

$$\mathbb{E}(\epsilon\epsilon^t|X) = \sigma^2 V(X) \tag{1}$$

For some matrix $V(X)$. We decompose the variance into σ^2 and $V(x)$ because in what follows we assume that $V(x)$ is known beforehand but σ^2 is estimated from the data. Using the Cholesky decomposition of V, where $V = C^tC$, consider the transformed variants of X, ϵ and y:

$$\tilde{y} = Cy \tag{2}$$

$$\tilde{X} = Cx \tag{3}$$

$$\tilde{\epsilon} = C\epsilon \tag{4}$$

It is known (see Fumio Hayashi, section 1.6) that the transformed model $\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}$ follows all of the standard linear model assumptions if the original follows all but the spherical errors assumption. Transforming the model back into the original space yields the generalized least squares estimator:

$$\widehat{\beta}_{GLS} = (X^tV^{-1}X)^{-1}X^tV^{-1}y \tag{5}$$

1. Using the transformed model, derive formulae in terms of X, y, V, and s^2 for the confidence and prediction intervals in the general linear model.

2. Download the cleaned 2007 ASA flight data (`euler.stat.yale.edu/~tba3/stat612/psets/pset03/data/airline2007_pset03.Rds`). I have selected flights arriving at the top 25 airports. The dataset is also divided into three random, evenly sized groups: I, II, and III. Assume that $V(X)$ is a diagonal matrix that depends only on the arrival airport.

Estimate the form of $V(X)$ using the data from group I and fit the following linear model to the data in group II using both ordinary least squares and the generalized model suggested by your estimate of $V(X)$:

$$\text{arrDelay}_i = \text{dest}_i + \text{depDelay}_i + \epsilon_i \tag{6}$$

Use constrast sums for the arrival airport (i.e., no intercept) term. How different are the estimators $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_{GLS}$?

Now, on group III construct two-sided, 95%-prediction intervals for the arrival delays using both estimates of β. Calculate the error rate (proportion of observations outside of the prediction inveral) by airport. How much do these differ between the two estimators?

3. Take all of the flights that arrived at ORD (Chicago O'hare) airport on 2007-01-11, and order the data from the earliest arrival to the latest arrival. Fit the following linear model on the data and plot a time series of the residuals.

$$\text{arrDelay}_i = \beta_0 + \beta_1 \cdot \text{depDelay}_i + \epsilon_i \tag{7}$$

What assumptions of the classical regression does your plot suggest are violated in this model?

A common description of this type of model is that the covariance of the errors follows the following:

$$\mathbb{E}(\epsilon_i \epsilon_j) = \sigma^2 \cdot e^{-\phi|i-j|} \tag{8}$$

Using the estimate of σ^2 from the original regression model, derive a reasonable estimate of ϕ from the model output (Hint: look at the case where $|i-j|$ is equal to 1 and find the ϕ that yields the sample covariance).

Now, using the predicted ϕ, fit the ordinary least squares and generalized least squares model from arrivals to ORD (Chicago O'hare) airport on 2007-02-13. How much do the two estimated $\widehat{\beta}$ differ from one another? Now find the predicted values $\widehat{y}$ for the two estimates of β. How much do these differ? Which is a better fit of the data?

II. Breaking classical linear model assumptions

In these problems, we ask you to run simulations in R to estimate the effect of violating various assumptions in the classical multivariate linear model presented in class. *For these problems, please attatch your R code as well as the written solutions.*

1. Normality Construct a linear model with $p = 5$, $n = 1000$, X generated by independent random normals, and $\beta = (1, 1, 1, 0, 0)$. For the following error distributions, generate an independent noise vector ϵ and set $y = X\beta + \epsilon$ (with no intercept):

1. normal distribution with $\sigma = 1$, `rnorm`
2. continuous uniform distribution from -2 to 2, `runif`
3. t-distribution with 2 degrees of freedom, `rt`
4. discrete uniform over the set $\{-1, 1\}$, `sample`
5. standard Cauchy distribution, `rcauchy`

For each, calculate the probability that the standard, two-sided, 95%-confidence interval (which assumes normality) for β_1 covers the true β_1 by iterating the simulation (about 20000 times each should be enough). Use a fixed X matrix over the iterations. How do you feel about the importance of the normality assumption given this simulation?

2. Spherical errors Consider the following model for the error terms (setting $\epsilon_0 = 0$ for the base case) for some $\phi > 0$:

$$\delta_t \sim_{i.i.d.} \mathcal{N}(0, 1), \quad t = 1, \ldots, n \tag{9}$$

$$\epsilon_t = \phi \cdot \epsilon_{t-1} + \delta_t \tag{10}$$

Write out the variance-covariance matrix of ϵ for $n = 3$.

Now, again consider a linear model with $p = 5$, $n = 1000$, X generated by independent random normals, and $\beta = (1, 1, 1, 0, 0)$ using the constructed ϵ. For ϕ equal to 0.5, 0.9, and 1 run a simulation to determine the probability that the confidence interval for β_1 falls within the standard, two-sided, 95%-confidence interval (which assumes normality) for β_1. (You may need a double loop for this, so we only need 2500 iterations here). How do you interpret these results?

3. Endogeneity Finally, consider the autoregression model (we intialize $Y_0 = 0$):

$$Y_t = \beta Y_{t-1} + \epsilon_t \tag{11}$$

Where ϵ is a random variable with a standard multivariate normal distribution. Why does this model violate strict endogeneity?

Run a simulation with $n = 1000$ and β equal to 0.5, 0.9, and 1. Determine the expected bias of the ordinary least squares estimator (with no intercept) for β in each of these cases. How do you interpret these results?