Statistical Inference 1 - Homework 5

Adriana Souza, Lilly Raud, Kevin Hunt, Saad Usmani, Beau Britain, Mike McCormack

11/22/2017

Guidelines

- This is a group project. You may work in groups with up to 6 people. You only need to turn in one homework assignment for each group, but make sure that everyone's name is listed on the assignment.
- Also, even if you don't write the code for every part of the assignment, you should practice the skills in each section.
- This assignment focuses on simulating data that violates various assumptions of the linear regression model.
- Some sample code has been included, but you will need to try many different starting values for the parameters, sample size, and distributional assumptions. The included code is just to give you a starting point; it should not be considered sufficient to answer all the questions in each part. (And you are welcome to ignore the sample code and use your own)

Questions

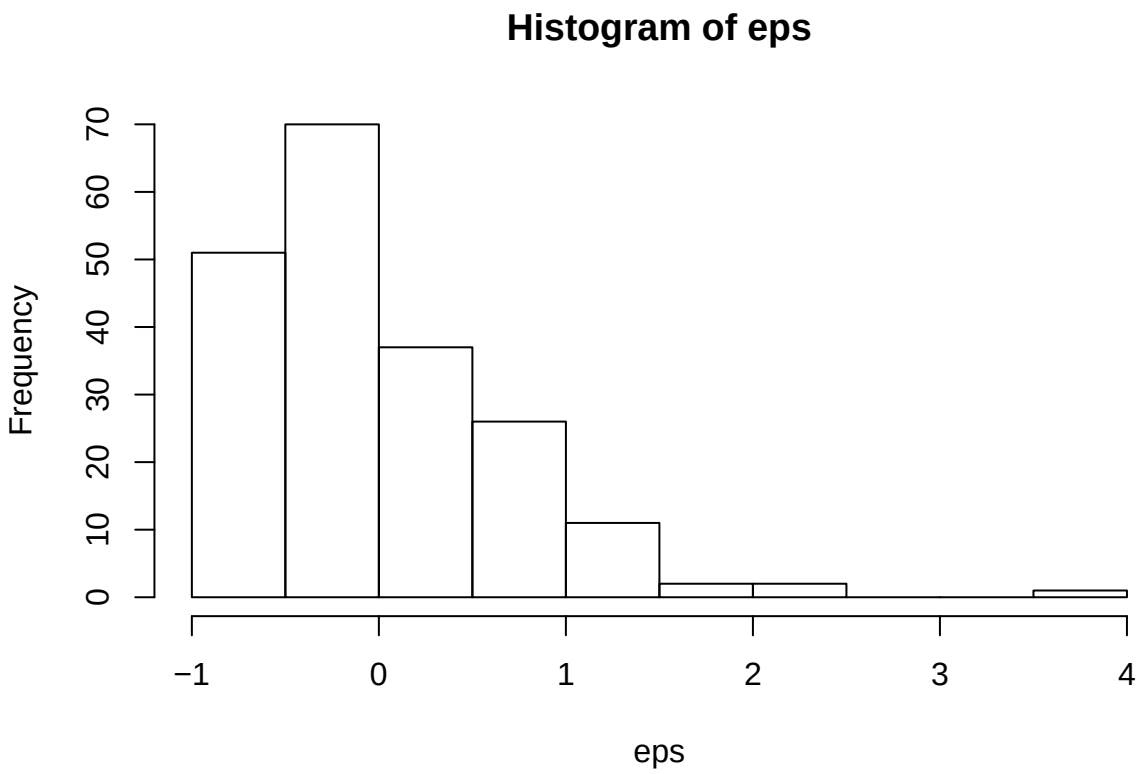
1. Nonlinear relationship
   a. Simulate data for a variety of different non-linear relationships (e.g. polynomial, exponential, sinusoidal).
   b. Try simulations with a small sample size (e.g. 20), a medium sample size (e.g. n= 100), and a large sample size (e.g. n = 5000).
   c. For each simulation,
      i. Predict y-hat at several different locations using a confidence interval.
      ii. Predict the beta coefficients for a linear model using a confidence interval.
      iii. Find the MSE (to estimate sigma^2)
      iv. Test to see whether the beta(s) are significant
   d. Which of the above tasks were affected by the nonlinear relationship?
   e. After you have experimented with the effects of different model structures, true parameter values, and sample sizes, let's repeat the simulation but test yourself to see whether you can detect non-linearity.
      i. Have R randomly choose whether to simulate data from a true linear model or a true nonlinear model.
      ii. Simulate data accordingly and display informal/ formal diagnostics as appropriate.
      iii. Based on the diagnostics, predict whether the problem areas you mentioned in part d will be affected or not. (Note: You are not predicting whether the assumptions are violated– just whether they are violated to such an extent that your ability to use the model is compromised)

Aside: Many of the issues you end up facing with a nonlinear relationship can also be seen if an important predictor is excluded from the model. If you have extra time, feel free to play with this issue as well (optional).

```
# Sample code to get started
n <- 20
b0 <- 10
b1 <- 2

x <- runif(n, 0, 10)
eps <- 3 * rnorm(n)
y <- b0 - b1 * (x - 5)^2 + eps
```

2. Non-normal errors
    a. Simulate errors from a variety of different non-normal distributions (e.g. gamma, poisson). Make sure to shift the errors over so that they are still centered at 0.
    b. Try simulations with a small sample size (e.g. 20), a medium sample size (e.g. n= 100), and a large sample size (e.g. n = 5000).
    c. For each simulation,
        i. Predict y-hat at several different locations using a confidence interval.
        ii. Predict the beta coefficients for a linear model using a confidence interval.
        iii. Find the MSE (to estimate sigma^2)
        iv. Test to see whether the beta(s) are significant (t-tests)
    d. Which of the above tasks were affected by the violation of assumptions?
    e. After you have experimented with the effects of different model structures, true parameter values, and sample sizes, let's repeat the simulation but test yourself to see whether you can detect non-normality.
        i. Have R randomly choose whether to simulate data with normal or nonnormal errors
        ii. Simulate data accordingly and display informal/ formal diagnostics as appropriate.
        iii. Based on the diagnostics, predict whether the problem areas you mentioned in part d will be affected or not. (Note: You are not predicting whether the assumptions are violated– just whether they are violated to such an extent that your ability to use the model is compromised)

```r
# Sample code to get started
n <- 200
b0 <- 10
b1 <- 2
eps_alpha <- 2
eps_beta <- 1/4


x <- runif(n, 0, 10)
eps <- (rgamma(n, eps_alpha, 1/eps_beta) - eps_alpha * eps_beta) * 2
y <- b0 - b1 * x + eps

hist(eps)
```




```
var(eps)
```

```
## [1] 0.4600411
```

```
mean(eps)
```

```
## [1] -0.003509722
```

3. Heterogeneous Variances
    a. Simulate errors from a variety of different relationships with X (e.g. eps = 2 * sqrt(x))
    b. Try simulations with a small sample size (e.g. 20), a medium sample size (e.g. n= 100), and a large sample size (e.g. n = 5000).
    c. For each simulation,
        i. Predict y-hat at several different locations using a confidence interval.
        ii. Predict the beta coefficients for a linear model using a confidence interval.
        iii. Find the MSE (to estimate sigma^2– does that even make sense here?)
        iv. Test to see whether the beta(s) are significant (t-tests)
    d. Which of the above tasks were affected by the violation of assumptions?
    e. After you have experimented with the effects of different model structures, true parameter values, and sample sizes, let's repeat the simulation but test yourself to see whether you can detect heteroskedacity.
        i. Have R randomly choose whether to simulate errors with constant or non-constant variance
        ii. Simulate data accordingly and display informal/ formal diagnostics as appropriate.
        iii. Based on the diagnostics, predict whether the problem areas you mentioned in part d will be affected or not. (Note: You are not predicting whether the assumptions are violated– just whether they are violated to such an extent that your ability to use the model is compromised)

```r
# Sample code to get started
n <- 200
b0 <- 10
b1 <- 10

x <- runif(n, 0, 10)
eps <- rnorm(n, sd = 0.5 * x^2)
y <- b0 + b1 * x + eps
```

4. Correlated Errors
    a. Simulate errors from a variety of different correlation structures.
    b. Try simulations with a small sample size (e.g. 20), a medium sample size (e.g. n= 100), and a large sample size (e.g. n = 5000).
    c. For each simulation,
        i. Predict y-hat at several different locations using a confidence interval.
        ii. Predict the beta coefficients for a linear model using a confidence interval.
        iii. Find the MSE (to estimate sigma^2)
        iv. Test to see whether the beta(s) are significant (t-tests)
    d. Which of the above tasks were affected by the violation of assumptions?
    e. After you have experimented with the effects of different model structures, true parameter values, and sample sizes, let's repeat the simulation but test yourself to see whether you can detect correlated errors.
        i. Have R randomly choose whether to simulate data with correlated or uncorrelated errors.
        ii. Simulate data accordingly and display informal/ formal diagnostics as appropriate.
        iii. Based on the diagnostics, predict whether the problem areas you mentioned in part d will be affected or not. (Note: You are not predicting whether the assumptions are violated– just whether they are violated to such an extent that your ability to use the model is compromised)

```r
# Sample code to get started
n <- 200
b0 <- 10
b1 <- 10
rho <- 0.9
sigma <- 2

x <- runif(n, 0, 10)

eps <- arima.sim(model = list(ar = rho), n = n)

y <- b0 + b1 * x + eps

# OR...

eps <- rep(0, n)
e.ind <- rnorm(n, mean = 0, sd = (sigma / sqrt(1-rho^2)))
eps[1] <- e.ind[1]
for (i in 2:n) {
  eps[i] <- rho * eps[i-1] + e.ind[i]
}

y <- b0 + b1 * x + eps
```

5. Multicollinearity
    a. Simulate predictors that are correlated with a variety of different correlation structures.

```
#bla
```

```
b.  Try simulations with a small sample size (e.g. 20), a medium sample size (e.g. n= 100), and a large
```

```
#bla
```

```
c.  For each simulation,
    i.  Predict y-hat at several different locations using a confidence interval.
    ii. Predict the beta coefficients for a linear model using a confidence interval.
    iii. Find the MSE (to estimate sigma^2)
    iv. Test to see whether the beta(s) are significant (t-tests)
d. Which of the above tasks were affected by the violation of assumptions?
e. After you have experimented with the effects of different model structures, true parameter values, ar
    i. Have R randomly choose whether to simulate data with correlated or uncorrelated predictor variabl
    ii. Simulate data accordingly and display informal/ formal diagnostics as appropriate.
    iii. Based on the diagnostics, predict whether the problem areas you mentioned in part d will be af
```

```
n <- 20
b0 <- 10
b1 <- 3
b2 <- 7

simga <- 2

x1 <- runif(n, 0, 10)
x2 <- x1 + rnorm(n)
cor(x1, x2)
```

```
## [1] 0.9222514
```

```
eps <- rnorm(n, sd = sigma)
y <- b0 + b1 * x1 + b2 * x2 + eps
```

6. Put it all together: Combine the code from the previous 5 parts. Have R randomly choose whether to generate data that violates one (or more) of the assumptions, or whether all the assumptions are valid. Show appropriate diagnostics and test yourself to see if you can predict whether there are problem areas or not. Repeat the simulation several times and record your accuracy at detecting the different problem areas.