ECON 4101 Econometrics
Final Exam

Pranav Singh

May 1, 2017

Problem 1

1.1

Based off the fitted model, vacation miles traveled was positively correlated with income and age and negatively correlated wtih number of kids. Specifically, an increase of one unit in the annual household income was associated with an increase of 17.427 miles driven per year, an increase of one unit in the average age of adult members in the household was associated with an increae of 16.365 miles driven per year, and each additional kid in the household was associated with a decrease of 84.956 miles driven per year.

1.2

The residuals seem to be homeskedastic relative to the age variable but heteroskedastic relative to the income variable. The next step should be to formally test for heteroskedasticity via the Goldfeld-Quandt Test. For this test, the null hypothesis assumes homoskedastic errors and the alternative hypothesis assumes heteroskedastic errors. The test statistic corresponds to an F-test of equality of the variances of the two partitions modeled separately. The degrees of freedom for the numerator and denominator are one less than the number of samples in each partition, respectively. If the test statistic is less than the critical F-value with alpha=0.05 and the aforementioned degrees of freedom, then we reject the null hypothesis. Otherwise, we don't have sufficient evidence to warrant the claim of heteroskedastic errors.

Problem 2

2.1

Autocorrelation, or serial correlation, refers to the correlation between the errors in different time periods in a time series or panel data model. Stock price modeling is an example of an econometric model where autocorrelation is likely to exist.

2.2

An AR(1) model refers to an autoregressive process of order one. It is a time series model whose current value depends linearly on its most recent value plus an unpredictable disturbance (i.e. an error term representing the cumulative effect of a collection of uncorrelated random variables that has mean zero and constant variaance). That is, $x_t = \rho x_{t-1} + e_t$.

2.3

Two main consequences of autocorrelation are:

1. The least squares estimator is still linear unbiased, but no longer BLUE.
2. Standard errors are incorrect, therefore confidence intervals, and inference may be misleading.

Problem 3

The Durbin-Watson Test for Autocorrelation of an autoregressive process of order one takes as its null hypothesis the claim that the errors are uncorrelated and its alternative hypothesis the claim that the errors are autocorrelated. The test statistic for this test is:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

The decision rules are as follows:

$$\begin{aligned} d == 2 &\ : \text{No autocorrelation.} \\ d < 2 &\ : \text{Positive autocorrelation} \\ d > 2 &\ : \text{Negative autocorrelation} \end{aligned}$$

```
df <- fread('../../Data/part3.csv')

#Extract residuals and predicted values
e <- df$Residuals
n <- length(e)

k = 1 #number of lags

#Durbin-Watson statistic
dw1 <- sum(diff(e,k)^2)/sum(e^2) ; dw1
```

```
## [1] 1.089621
```

The resulting test statistic value of 1.0896205 is less than 2, so we interpret this as the model having positive autocorrelation.

Problem 4

The notation ARIMA$(p, d, q)(P, D, Q)[m]$ refers to a Seasonally-Adjusted Autoregressive Integrated Moving Average (ARIMA) model where the parameters (p, d, q) refer to the non-seasonal part of the ARIMA model and the parameters (P, D, Q) refer to the seasonal part of the ARIMA model. More specifically, for the non-seasonal part of the ARIMA model, p is the order (number of lags) of the autoregressive model, d is the degree of differencing (number of times the data have had past values subtracted), and q is the order of the moving average model. The uppercase versions of those parameters are similarly defined for the seasonal portion of the ARIMA model. Lastly, the parameter m refers to the number of periods in each season. Thus, ARIMA$(1, 0, 2)(1, 0, 0)[12]$ refers to an ARIMA model for which:

1. The non-seasonal portion has autoregressive order of 1, no differencing terms, and a moving average of order 2.
2. The seasonal portion has autoregressive order 1, no differencing terms, no moving average terms, and is formed using 12 seasonal periods.

Problem 5

```r
df <- fread('../../Data/part5.csv')
```

5.1

```r
summary(lm(auto ~ dtime, df))
```

```
##
## Call:
## lm(formula = auto ~ dtime, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65635 -0.14379  0.02776  0.15292  0.82457
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 0.484795   0.071449   6.785 0.00000176 ***
## dtime       0.007031   0.001286   5.467 0.00002834 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3273 on 19 degrees of freedom
## Multiple R-squared:  0.6113, Adjusted R-squared:  0.5909
## F-statistic: 29.88 on 1 and 19 DF,  p-value: 0.00002834
```

We see from the above fitted linear probability model that an increase of one minute in `dtime` is associated with an approximate 0.7031% increase in probability of a person choosing an automobile. Thus, the probability of a person choosing is an automobile given dtime=2 is 1.4062% greater than the case where dtime=0.

5.2

```r
mod2 <- glm(auto ~ dtime, df, family = binomial(link=logit))
summary(mod2)
```

```
##
## Call:
## glm(formula = auto ~ dtime, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6469  -0.3414  -0.1128   0.3967   2.3012
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.23758    0.75048  -0.317   0.7516
## dtime        0.05311    0.02064   2.573   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 29.065  on 20  degrees of freedom
## Residual deviance: 12.332  on 19  degrees of freedom
## AIC: 16.332
##
## Number of Fisher Scoring iterations: 6
```

```
odds.ratios <- exp(coef(mod2)); odds.ratios
```

```
## (Intercept)       dtime
##   0.7885374   1.0545455
```

We see from the above fitted logit model that an increase of one minute in `dtime` is associated with an approximate 5.45% increase in probability of a person choosing an automobile. Thus, the probability of a person choosing is an automobile given dtime=2 is 10.9% greater than the case where dtime=0.

5.3

```
df$prediction <- ifelse(predict(mod2, newdata=df, type='response') > 0.5, 1, 0)
accuracy <- mean(df$auto == df$prediction); accuracy
```

```
## [1] 0.9047619
```

Our logit model yielded 90.48% accuracy in predicting whether a person chooses an automobile given the difference in commute times between bus and auto.