ECON 4101 Econometrics
Midterm Exam

Pranav Singh

February 20, 2017

Problem 1

```r
# Given information:
x.bar <- 500.7297
x.sd <- 49.24917
y.bar <- 216.7784
y.sd <- 30.27246
xy.cor <- 0.6809816
n.x <- 37
n.y <- 37
```

```r
# (a) Fit simple linear regression model: y = b0 + b1*x
b1 <- xy.cor * y.sd^2/x.sd^2  # b1 = Cov(x,y) / Var(x)
b0 <- y.bar - b1 * x.bar
cat("Estimated intercept coefficient:", b0)
```

```
## Estimated intercept coefficient: 87.94267
```

```r
cat("Estimated slope coefficient:", b1)
```

```
## Estimated slope coefficient: 0.257296
```

The Coefficient of Determination (R^2) indicates the proportion of the variation of the dependent variable that can be explained by the variation of the independent variables. In linear least squares regression with an intercept term, R^2 equals the square of the Pearson correlation coefficient between the observed responses y_i and the predeicted values $\widehat{f}(x_i)$. For univariate linear least squares regression, such as the case in this problem, this is equivalent to the squared Pearson correlation coefficient between y and x. Thus, the proportion of the variation of y that can be explained by the variation of x is equal to $COR^2(x, y) = 0.4637359$

Problem 2

In regression analysis, the Global F-Test tests the following hypotheses for the model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_n = 0$$
$$H_1 : \exists i \geq 1 \text{ for which } \beta_i \neq 0$$

Its test statistic is given by $F = \frac{MSR}{MSE} = \frac{R^2/k}{(1-R^2)/(n-(k+1))}$, where k is the number of predictor variables in the linear regression model. The critical value for the test at a given significance level α is $F_{\alpha,\mathrm{df}_r,\mathrm{df}_e}$, where $\mathrm{df}_r = k$ and $\mathrm{df}_e = n - (k+1)$ are the degrees of freedom of the regression and the errors, respectively. The critical value can be found using numerical software or a table of the CDF of the F-distribution. We reject H_0 at the α significance level when the obtained test statistic $F \geq F_{\alpha,\mathrm{df}_r,\mathrm{df}_e}$. The Global F-Test is considered a joint test in multiple regression analysis because it tests whether all the slope coefficients are equal to 0. That is, it's concerned with the joint probability distribution of all the slope parameters. In contrast, the individual t-tests are only concerned with a single coefficient each. For multiple regression analysis, as the number of predictor variables in the model increases, the probability of at least one of the individual t-tests yielding a false positive increases as well. The Global F-test helps combat this issue via its joint nature. Furthermore, two correlated predictors may appear statistically insignificant themselves via the t-test but be jointly significant via the F-test, the reason being that the multicollinearity can lead to the correlated variables inflating the standard errors of each other's estimated slope coefficients.

Prroblem 3

Gauss-Markov Assumptions

1. Linearity
   - The dependent variable is assumed to be a linear function of the independent variables
   - i.e. $y = \beta_0 + \sum \beta_i f_i(x_i) + \epsilon$ is allowed, but, for example, $y = \beta_0\beta_1 + \beta_1^2 x_1$ is not allowed.
2. Strict exogeneity
   - i.e. Zero Conditional Mean of Error Term: $E(\epsilon_i | x_1, \ldots, x_n) = 0$
3. Random sample (x_i, y_i)
   - The data is randomly sampled from some populations.
4. Full rank
   - The sample predictors data matrix $\mathbf{X}$ must be non-singular, i.e. it must have full rank.
   - Otherwise $\mathbf{X}$ is not invertible and the ordinary least quares (OLS) estimator cannot be computed.
   - **Perfect collinearity** between any of the regressors violates this assumption. i.e. when the set of predictors is not linearly independent.
5. Spherical Assumptions
   a) Constant Variance: $\text{VAR}(\epsilon|\mathbf{X}) = \sigma^2\mathbf{I}$, with $sigma^2 > 0$
   b) No Serial Correlation (aka "Auto-Correlation"): $\text{COV}(\epsilon_i, \epsilon_j) = 0$

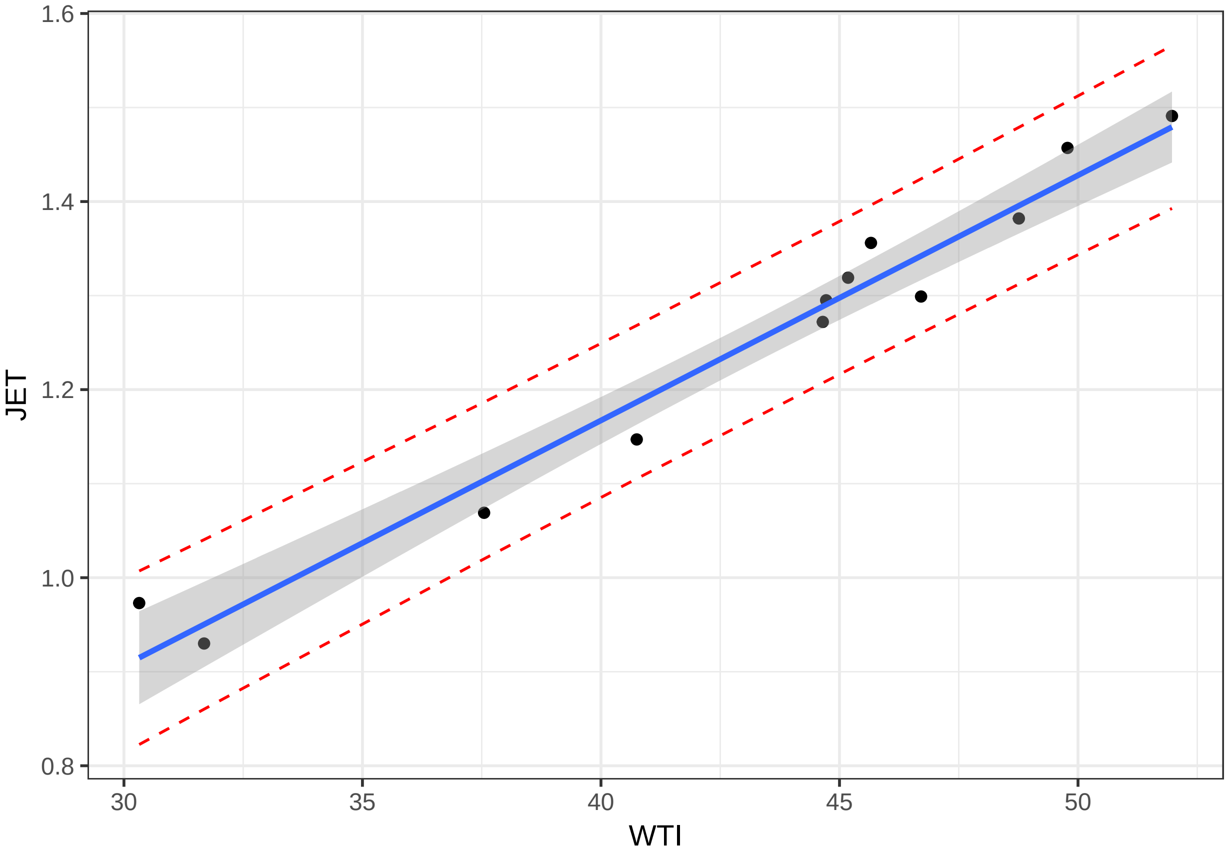
When the Gauss-Markov Assumptions are satisfied, the Gauss-Markov theorem states that the best linear unbiased estimator (BLUE) is given by the ordinary least squares (OLS) estimator.

Problem 4

```
text <- "MONTH WTI JET
JAN 31.68 0.930
FEB 30.32 0.973
MAR 37.55 1.069
APR 40.75 1.147
MAY 46.71 1.299
JUN 48.76 1.382
JUL 44.65 1.272
AUG 44.72 1.295
SEP 45.18 1.319
OCT 49.78 1.457
NOV 45.66 1.356
DEC 51.97 1.491"
con <- textConnection(text)
petroleum <- read.csv(con, sep = " ")
summary(petroleum)
```

```
##      MONTH        WTI              JET
##  APR    :1   Min.   :30.32   Min.   :0.930
##  AUG    :1   1st Qu.:39.95   1st Qu.:1.127
##  DEC     :1   Median :44.95   Median :1.297
##  FEB     :1   Mean   :43.14   Mean   :1.249
##  JAN     :1   3rd Qu.:47.22   3rd Qu.:1.363
##  JUL     :1   Max.   :51.97   Max.   :1.491
##  (Other):6
```

```
lm.fit <- lm(JET ~ WTI, petroleum)
preds <- predict(lm.fit, newdata = NULL, interval = 'prediction')
df <- cbind(petroleum, preds)
ggplot(df, aes(x = WTI, y = JET)) +
  geom_point() +
  geom_smooth(method = 'lm', se = T, level = 0.95) +
  geom_line(aes(y = lwr), color = 'red', linetype = 'dashed') +
  geom_line(aes(y = upr), color = 'red', linetype = 'dashed')
```



In the graph above, the blue shaded region depicts the confidence interval, while the red dashed lines depict the prediction interval for our fitted linear regression model. The 95% confidence intervals for the parameters β_0 and β_1 are, in order:

```
confint(lm.fit, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -0.02545157 0.27433334
## WTI          0.02263411 0.02950391
```

```
summary(lm.fit)
```

```
##
## Call:
```

```
## lm(formula = JET ~ WTI, data = petroleum)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.043124 -0.023813 -0.004406  0.021282  0.058147
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 0.124441   0.067272    1.85      0.0941 .
## WTI         0.026069   0.001542   16.91 0.000000011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03496 on 10 degrees of freedom
## Multiple R-squared:  0.9662, Adjusted R-squared:  0.9628
## F-statistic:   286 on 1 and 10 DF,  p-value: 0.00000001099
```

The tiny p-value of the F-statistic for the fitted model suggests we have ample evidence to reject the null hypothesis of no relationship between WTI and JET at the 0.1% significance level. In other words, we are highly confident that there is indeed a relationship between the prices of crude oil (WTI) and jet fuel (JET).

We will estimate the elasticity of jet fuel prices using our linear functional form:

```
elasticity <- lm.fit$coefficients[2] * mean(petroleum$WTI)/mean(petroleum$JET)
elasticity
```

```
##       WTI
## 0.9003809
```

We interpret the above result as indicating that a 1% increase in the price of crude oil is associated with a 0.9003809% increase in the price of jet fuel.

Problem 5

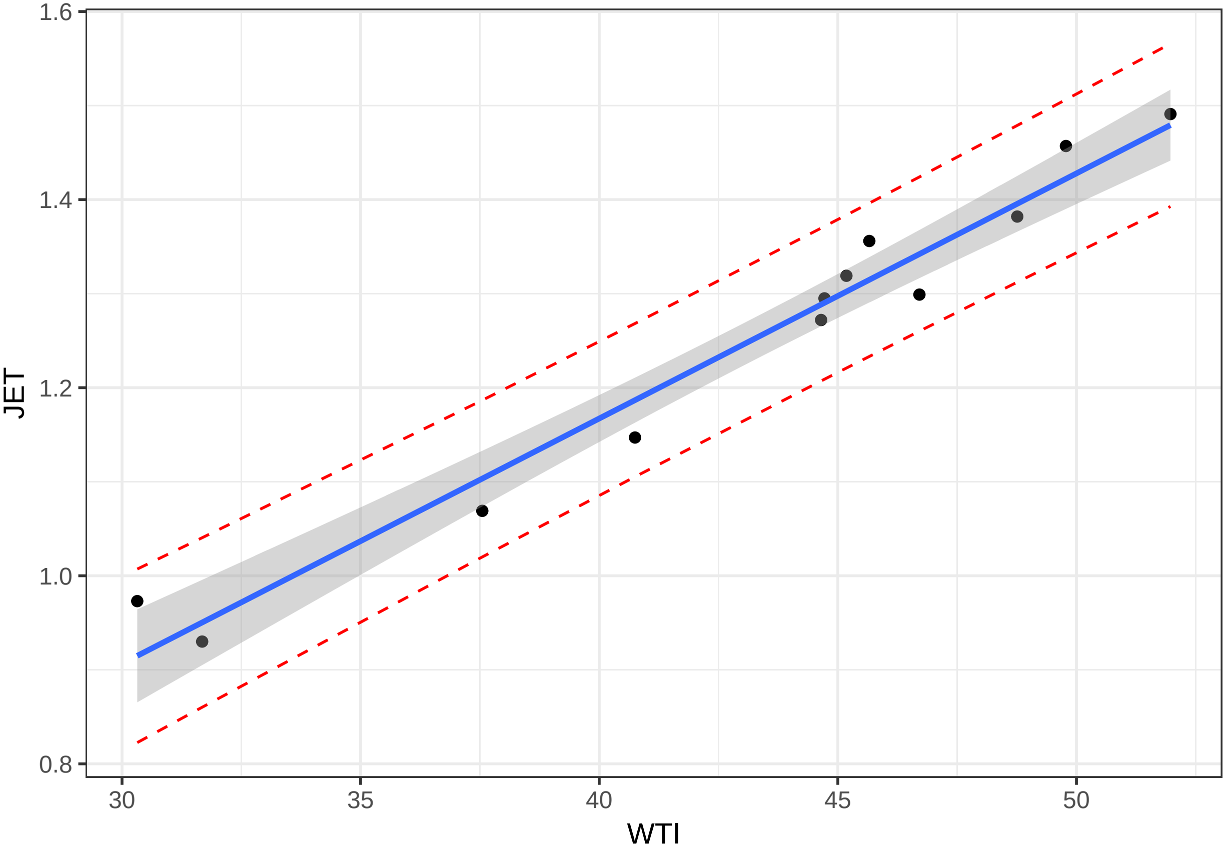
In a February 2015 article in Seeking Alpha, Adrian Wong reported that price correlations between crude oil and various petroleum products, “are close to 100%”. Jet kerosene is one such petroleum product that is distilled from crude oil. So naturally, the question arose of understanding the price correlation between crude oil and jet kerosene. Since crude oil is a preqrequisite in the supply chain of jet kerosene production, we hypothesize that an increase in the price of crude oil is associated with an increase in price of jet kerosene.

We were given data for the monthly average prices of crude oil and jet kerosene for the months in the year 2016. We choose the simple linear regression model with P_C = the average daily price of one barrel of crude oil as the predictor of the response P_J = the average daily price of one gallon of jet fuel. The linear functional form of this model can be written as $P_J = \beta_0 + \beta_1 P_C + \epsilon$, where we assume the error term to have expectation zero (i.e. $E(\epsilon) = 0$). Our first hypothesis can then be rephrased as $\beta_1 > 0$. We further hypothesize that the intercept term $\beta_0 = 0$.

The crude oil (WTI) and jet kerosene (JET) prices in the data yielded the following five-number summary:

```
##       WTI             JET
##  Min.   :30.32   Min.   :0.930
##  1st Qu.:39.95   1st Qu.:1.127
##  Median :44.95   Median :1.297
##  Mean   :43.14   Mean   :1.249
##  3rd Qu.:47.22   3rd Qu.:1.363
##  Max.   :51.97   Max.   :1.491
```

Using the statistical programming language `R`, the model stated above was fit to the sample data. The following graph shows the data samples, the line fit by our model; the shaded region depicts the confidence interval, while the red dashed lines depict the prediction interval for our fitted linear regression model.



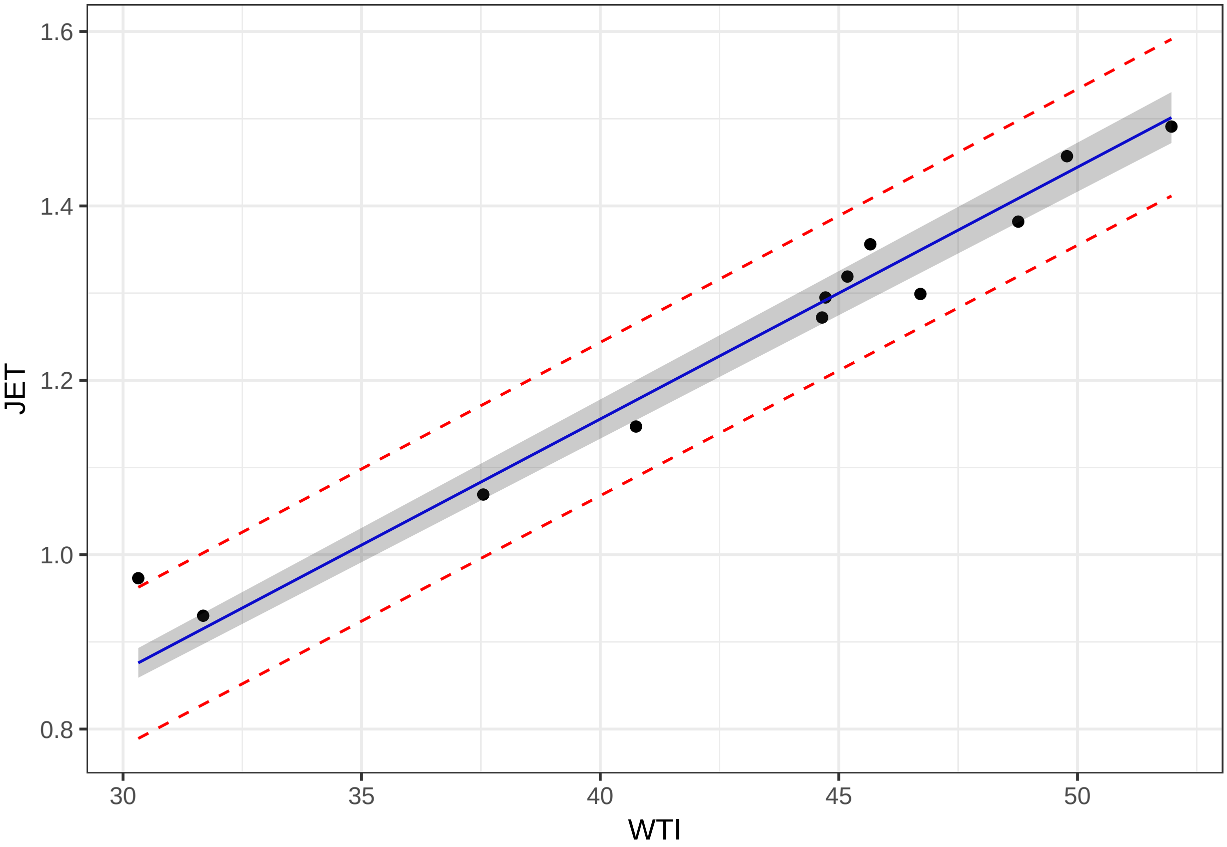
The 95% confidence intervals for the parameter estimates β_0 and β_1 were found to be, in order:

```
##                   2.5 %     97.5 %
## (Intercept) -0.02545157 0.27433334
## WTI          0.02263411 0.02950391
```

```
## 
## Call:
## lm(formula = JET ~ WTI, data = petroleum)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.043124 -0.023813 -0.004406  0.021282  0.058147 
## 
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)    
## (Intercept) 0.124441   0.067272    1.85      0.0941 .  
## WTI         0.026069   0.001542   16.91 0.000000011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.03496 on 10 degrees of freedom
## Multiple R-squared:  0.9662, Adjusted R-squared:  0.9628 
## F-statistic:   286 on 1 and 10 DF,  p-value: 0.00000001099
```

The tiny p-value of the F-statistic for the fitted model (shown in the above model summary) indicates we are highly confident that there is indeed a relationship between the prices of crude oil (WTI) and jet fuel (JET). Moreover, the p-value of the t-test statistic for the slope intercept term confirms our hypothesis that the direction of this relationship is positive, at least at the 0.001% significance level (i.e. without a reasonable doubt). In contrast, our estimate of the intercept term was statistically insignificant at even the 5% significance level. This implies that we cannot reject the hypothesis that $\beta_0 = 0$. That is, the data doesn't significantly contradict (but nor does it directly support) our hypothesis that the true value of β_0 is in fact 0. Estimating the elasticity from the linear functional form of the model, our analysis suggests that a 1% increase in the price of crude oil is associated with a 0.9003809% increase in the price of jet fuel.

Since the intercept term was deemed statistically insignificant, and since our initial assumption was that its true value is 0, it is worth fitting a simple linear model without an intercept term. Doing so yielded the following graph and model summary:



```
##
## Call:
## lm(formula = JET ~ 0 + WTI, data = petroleum)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05038 -0.02005 -0.00361  0.01585  0.09710
##
## Coefficients:
##      Estimate Std. Error t value            Pr(>|t|)
## WTI 0.0288884  0.0002555   113.1 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.03862 on 11 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.9991
## F-statistic: 1.279e+04 on 1 and 11 DF,  p-value: < 0.00000000000000022
```

Using the coefficient of determination R^2 as our measure of goodness-of-fit, we find that the model without an intercept term better fits the data than the one with an intercept term. Moreover, the fact that $R^2 \approx 1$ for the model without an intercept term lends credence to Adrian Wong's reporting that the price correlations between crude oil and varios petroleum products "are close to 100%".

The economic implication of our analysis is that changes in price of crude oil are a strong harbinger of downstream changes in price of jet kerosense. Moreover, the changes in price are in the same direction for both goods. Further research should consider possible confounding biases in our model. That is, is the apparent causal relationship between crude oil and jet kerosene prices due to the fact that they are both directly caused by some other variable? Also, more granular pricing data would help in assessing to what extent is there a lag from a change in crude oil price to a change in jet kerosene price.

Problem 6

I'd like to reproduce the following paper:

Kilian, Lutz, and Daniel P. Murphy. "The Role Of Inventories And Speculative Trading In The Global Market For Crude Oil." *Journal of Applied Econometrics* 29.3 (2013): 454-78. Web.