Question 1.

Linear regression would be appropriate for predicting my monthly electric bill. Predictors that have the potential to be explanatory would be daily temperatures, hours at home per day, school days in the month, workdays in the month, and work from home days in the month.

Question 2.

Overall, my thought process was to begin by creating a linear regression model using all factors so that I could determine the significance levels of each factor. Having removed the non-significant factors from the initial model, I would build multiple additional models that used varying factor combinations. These additional models would be cross-validated and compared to each other using Adjusted R^2 to determine the "best" model. The "best" model would then be used to predict crime for the data point provided. I outlined this multi-step process below:

Step 1 - Load Dataset

With limited observations, I created one dataset and didn't split it into training/validation datasets (this will be addressed with cross-validation in a later step).

```
# Load crime data into a data frame
data_df <- read.table("uscrime.txt", header=TRUE)
```

Step 2 - Determining Significant Factors

To understand which factors have significance, I created a model that used all available factors.

```
# Model using all available factors
model_all <- lm(Crime ~., data_df, model = TRUE)
summary(model_all)
```

Model_all produced the following summary:

```
Call:
lm(formula = Crime ~ ., data = data_df)

Residuals:
   Min     1Q Median     3Q    Max 
-395.7  -98.1   -6.7  113.0  512.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.98e+03   1.63e+03   -3.68  0.00089 ***
M            8.78e+01   4.17e+01    2.11  0.04344 *  
So          -3.80e+00   1.49e+02   -0.03  0.97977    
Ed           1.88e+02   6.21e+01    3.03  0.00486 ** 
Po1          1.93e+02   1.06e+02    1.82  0.07889 .  
Po2         -1.09e+02   1.17e+02   -0.93  0.35883    
```

```
LF          -6.64e+02   1.47e+03   -0.45  0.65465   
M.F          1.74e+01   2.04e+01    0.86  0.39900   
Pop         -7.33e-01   1.29e+00   -0.57  0.57385   
NW           4.20e+00   6.48e+00    0.65  0.52128   
U1          -5.83e+03   4.21e+03   -1.38  0.17624   
U2           1.68e+02   8.23e+01    2.04  0.05016 . 
Wealth       9.62e-02   1.04e-01    0.93  0.36075   
Ineq         7.07e+01   2.27e+01    3.11  0.00398 **
Prob        -4.86e+03   2.27e+03   -2.14  0.04063 * 
Time        -3.48e+00   7.17e+00   -0.49  0.63071   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 209 on 31 degrees of freedom
Multiple R-squared:  0.803,   Adjusted R-squared:  0.708 
F-statistic: 8.43 on 15 and 31 DF,  p-value: 3.54e-07
```

From this summary we can determine that only M, Ed, Po1, U2, Ineq, and Prob were shown to have significance estimating Crime when all factors were included.

Step 3 - Multiple Model Creation

Given the factors that were determined to be significant, I created five models using various combinations of these significant factors:

```
# Create models using variety of factor configurations - only significant
factors

# all significant factors
model_1 <- lm(Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data_df)

# all but Po1
model_2 <- lm(Crime ~ M + Ed + U2 + Ineq + Prob, data_df)

# all but U2
model_3 <- lm(Crime ~ M + Ed + Po1 + Ineq + Prob, data_df)

# * significance and above
model_4 <- lm(Crime ~ M + Ed + Ineq + Prob, data_df)

# ** significance
model_5 <- lm(Crime ~ Ed + Ineq, data_df)
```

Step 4 - Model Comparison

To compare the models built in Step 3, I created a function that runs cv.lm against each model and calculates adjusted R^2. The adjusted R^2 will be used to choose the “best” model.

```
# Assess each model using cross-validation and return Adj. R^2
cv_model_assessment <- function(model) {
  # Use cv.lm to perform CV on model
  cvmodel <- cv.lm(data_df, model, m = 5, printit = FALSE)

  # Sum of the Squared Errors
  ssres <- attr(cvmodel,"ms")*nrow(data_df)

  # Total Sum of Squared differences between data and its mean
  sstot <-  sum((data_df$Crime - mean(data_df$Crime))^2)

  # Calculate R^2
  r_sq <- 1 - (ssres/sstot)

  # Calculate Adj. R^2
  adj_r_sq_num <- (1-r_sq)*(nrow(data_df)) # numerator (1-R^2)*(n-1)
  adj_r_sq_den <- nrow(data_df)-(ncol(model$model)-1)-1 # denominator
(n-k-1)
  adj_r_sq <- 1 - (adj_r_sq_num/adj_r_sq_den)

  return(adj_r_sq)
}

# Find Adj. R^2 based on cv.lm for each model
ma_adj_r_sq <- cv_model_assessment(model_all)
m1_adj_r_sq <- cv_model_assessment(model_1)
m2_adj_r_sq <- cv_model_assessment(model_2)
m3_adj_r_sq <- cv_model_assessment(model_3)
m4_adj_r_sq <- cv_model_assessment(model_4)
m5_adj_r_sq <- cv_model_assessment(model_5)

# Display the Adj. R^2 values
ma_adj_r_sq # 0.111
m1_adj_r_sq # 0.575
m2_adj_r_sq # 0.053
m3_adj_r_sq # 0.553
m4_adj_r_sq # -0.11
m5_adj_r_sq # -0.0895
```

Based on the adjusted R^2 for each model, I've chosen model_1 - the model that incorporated all significant factors - as the "best" model:

```
# Chose "best" model - model_1
# model_1 uses all significant factors from model_all
summary(model_1)

Call:
lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = data_df)

Residuals:
   Min     1Q Median     3Q    Max 
-470.7  -78.4  -19.7  133.1  556.2 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -5040.5      899.8   -5.60  1.7e-06 ***
M              105.0       33.3    3.15   0.0031 ** 
Ed             196.5       44.8    4.39  8.1e-05 ***
Po1            115.0       13.8    8.36  2.6e-10 ***
U2              89.4       40.9    2.18   0.0348 *  
Ineq            67.7       13.9    4.85  1.9e-05 ***
Prob         -3801.8     1528.1   -2.49   0.0171 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 201 on 40 degrees of freedom
Multiple R-squared:  0.766,	Adjusted R-squared:  0.731 
F-statistic: 21.8 on 6 and 40 DF,  p-value: 3.42e-11
```

Step 5 - Predict Crime for Data Point using Model

Using model_1, I predicted the crime value for the data point provided:

```
# Data point to estimate Crime for - note I'm only bringing in the
variables model_1 requires
new_dp <- data.frame(M = 14.0, U2 = 3.6, Ineq = 20.1, Prob = 0.04, Po1 =
12.0, Ed = 10.0)
predict.lm(model_1, new_dp)
```

Model_1 produces a prediction for Crime of 1,304.

APPENDIX --- Full R Script

```r
# Clear the environment
rm(list = ls())

# Comment in set.seed(33) to repeat results
set.seed(33)

# Load DAAG library for access to cv.lm
require(DAAG)

# Load crime data into a data frame
data_df <- read.table("uscrime.txt", header=TRUE)

# Model using all available factors
model_all <- lm(Crime ~., data_df)
summary(model_all)

# Create models using variety of factor configurations - only significant
factors
model_1 <- lm(Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data_df) # all
significant factors
model_2 <- lm(Crime ~ M + Ed + U2 + Ineq + Prob, data_df)       # all but
Po1
model_3 <- lm(Crime ~ M + Ed + Po1 + Ineq + Prob, data_df)      # all but
U2
model_4 <- lm(Crime ~ M + Ed + Ineq + Prob, data_df)            # *
significance and above
model_5 <- lm(Crime ~ Ed + Ineq, data_df)                       # **
significance

# Assess each model using cross-validation and return Adj. R^2
cv_model_assessment <- function(model) {
  # Use cv.lm to perform CV on model
  cvmodel <- cv.lm(data_df, model, m = 5, printit = FALSE)

  # Sum of the Squared Errors
  ssres <- attr(cvmodel,"ms")*nrow(data_df)

  # Total Sum of Squared differences between data and its mean
  sstot <-  sum((data_df$Crime - mean(data_df$Crime))^2)

  # Calculate R^2
  r_sq <- 1 - (ssres/sstot)

  # Calculate Adj. R^2
  adj_r_sq_num <- (1-r_sq)*(nrow(data_df)) # numerator (1-R^2)*(n-1)
  adj_r_sq_den <- nrow(data_df)-(ncol(model$model)-1)-1 # denominator
(n-k-1)
  adj_r_sq <- 1 - (adj_r_sq_num/adj_r_sq_den)

  return(adj_r_sq)
}

# Find Adj. R^2 based on cv.lm for each model
ma_adj_r_sq <- cv_model_assessment(model_all)
m1_adj_r_sq <- cv_model_assessment(model_1)
m2_adj_r_sq <- cv_model_assessment(model_2)
m3_adj_r_sq <- cv_model_assessment(model_3)
m4_adj_r_sq <- cv_model_assessment(model_4)
m5_adj_r_sq <- cv_model_assessment(model_5)

# Display the Adj. R^2 values
ma_adj_r_sq # 0.111
m1_adj_r_sq # 0.575
m2_adj_r_sq # 0.053
m3_adj_r_sq # 0.553
m4_adj_r_sq # -0.11
m5_adj_r_sq # -0.0895

# Chose "best" model - model_1
# model_1 uses all significant factors from model_all
summary(model_1)

# Data point to estimate for - note I'm only bringing in the variables
model_1 requires
new_dp <- data.frame(M = 14.0, U2 = 3.6, Ineq = 20.1, Prob = 0.04, Po1 =
12.0, Ed = 10.0)
predict.lm(model_1, new_dp)
```