**Midterm 2**

"I have not used any resources from outside the class or discussed the exam with anyone." - Martin Zanaj

**Data Hprice**

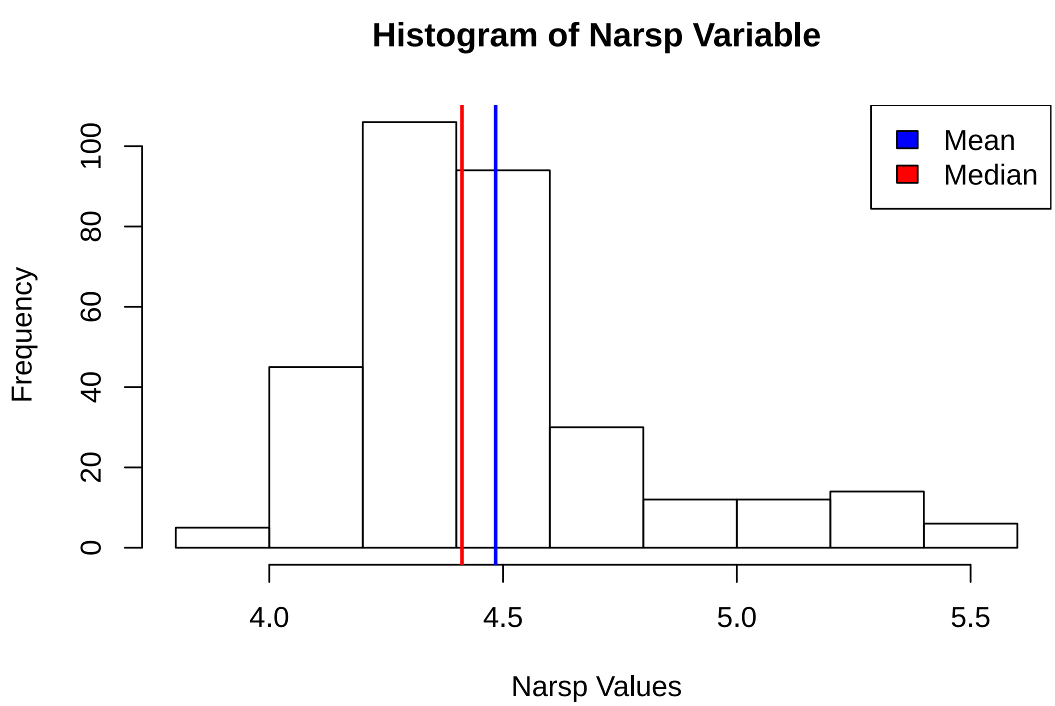
Hprice is a dataset on housing prices in 36 US metropolitan statistical areas (MSAs) over 9 years from 1986 to 1994. The data has 324 observations and 8 variables.

| Variable | Type | Description |
|---|---|---|
| narsp | Quantitative | natural log average sale price in thousands of dollars |
| ypc | Quantitative | average per capita income |
| perypc | Quantitative | percentage growth in per capita income |
| regtest | Quantitative | Regulatory environment index |
| rcdum | Categorical | Rent control (0=no, 1=yes) |
| ajwtr | Categorical | Adjacent to a coastline (0=no, 1=yes) |
| msa | Categorical | indicator for the MSA (1-36) |
| time | Categorical? | Year 1=1986 to 9= 1994 |

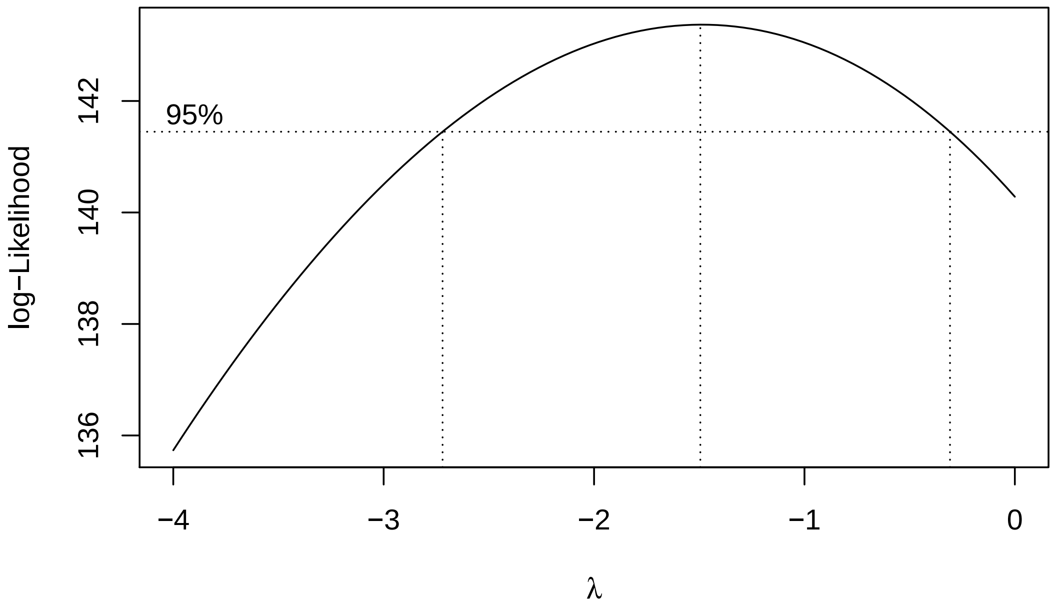
1. The summary function allows to see descriptive statistics for all variables. In the case of *narsp*, the mean is larger than the median. This is common of a **right skewed** scenario.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.920   4.264   4.412   4.484   4.575   5.563
```

2. A plot that allows one to visualize the distribution of a variable is the histogram. The histogram of the variable *narsp* does indeed confirm the previous finding- right skewed with a mean greater than the mean. Hence, the results are **consistent**.



3. We want to pick the lambda tha maximizes the log-likelihood function. From the plot, the 95% confidence interval for $\lambda$ (-2.8, -0.3) does not include 1, so a transformation is appropriate. The estimated value for the optimal $\lambda$ = **-1.47**. In this case, we can use the same value for the $\lambda$ for the transformation, or we could use an approximation to make interpretation easier. The rounded value of -1 is within the confidence interval. We can transform the data using $\lambda$ **=-1**, which corresponds to the **inverse transformation** (transformed vlaue= 1/original value).




4. In both models, all predictors are significant including the intercetp. In terms of R^2 the first model (original) seems to have a higher R^2 coefficient. The transformed model has smaller standard errors for its coefficients that then original model. The sign of all coefficient (except intercept) has changed in the transformed model. The reciprocal reverses order among values of the same sign: largest becomes smallest. Now, the resulting formula becomes y= 1/(b0+b1X1+b2X2+…+bnXn). The slope b1, b2,.., bn represent the expected change in average y (natural log average sale price in thousands of dollars) associated with a 1-unit increase in X.

```
##
## Call:
## lm(formula = narsp ~ ypc + perypc + regtest + rcdum + time, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31051 -0.11653 -0.01862  0.07919  0.57618
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.661e+00  8.458e-02  31.460  < 2e-16 ***
## ypc          7.180e-05  4.291e-06  16.735  < 2e-16 ***
## perypc      -1.387e-02  5.091e-03  -2.725 0.006794 **
```

```
## regtest      2.973e-02  3.112e-03   9.555  < 2e-16 ***
## rcdum1       1.587e-01  3.199e-02   4.960 1.15e-06 ***
## time        -1.886e-02  5.103e-03  -3.695 0.000258 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1656 on 318 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7508
## F-statistic: 195.7 on 5 and 318 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = (1/narsp) ~ ypc + perypc + regtest + rcdum + time,
##     data = hprice)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0257995 -0.0042241  0.0003853  0.0054470  0.0184695
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.072e-01  4.117e-03  74.602  < 2e-16 ***
## ypc         -3.293e-06  2.089e-07 -15.767  < 2e-16 ***
## perypc       6.561e-04  2.478e-04   2.647  0.00852 **
## regtest     -1.303e-03  1.515e-04  -8.602 3.65e-16 ***
## rcdum1      -6.309e-03  1.557e-03  -4.051 6.42e-05 ***
## time         6.913e-04  2.484e-04   2.783  0.00571 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008063 on 318 degrees of freedom
## Multiple R-squared:  0.727,  Adjusted R-squared:  0.7227
## F-statistic: 169.3 on 5 and 318 DF,  p-value: < 2.2e-16
```

**Data Divusa**, represents the divorce rates in the USA from 1920-1996. The data has 77 observations and 7 variables.

| Variable | Type | Description |
|---|---|---|
| year | Quantitative | the year from 1920-1996 |
| divorce | Quantitative | divorce per 1000 women aged 15 or more |
| unemployed | Quantitative | unemployment rate |
| femlab | Quantitative | percent female participation in labor force aged 16+ |
| marriage | Quantitative | marriages per 1000 unmarried women aged 16+ |
| birth | Quantitative | births per 1000 women aged 15-44 |
| military | Quantitative | military personnel per 1000 population |

5. Backward elimination strategy, using alpha=.05, yields the model with predictors: **year**, **femlab**, **marriage**, **birth**, **military**.

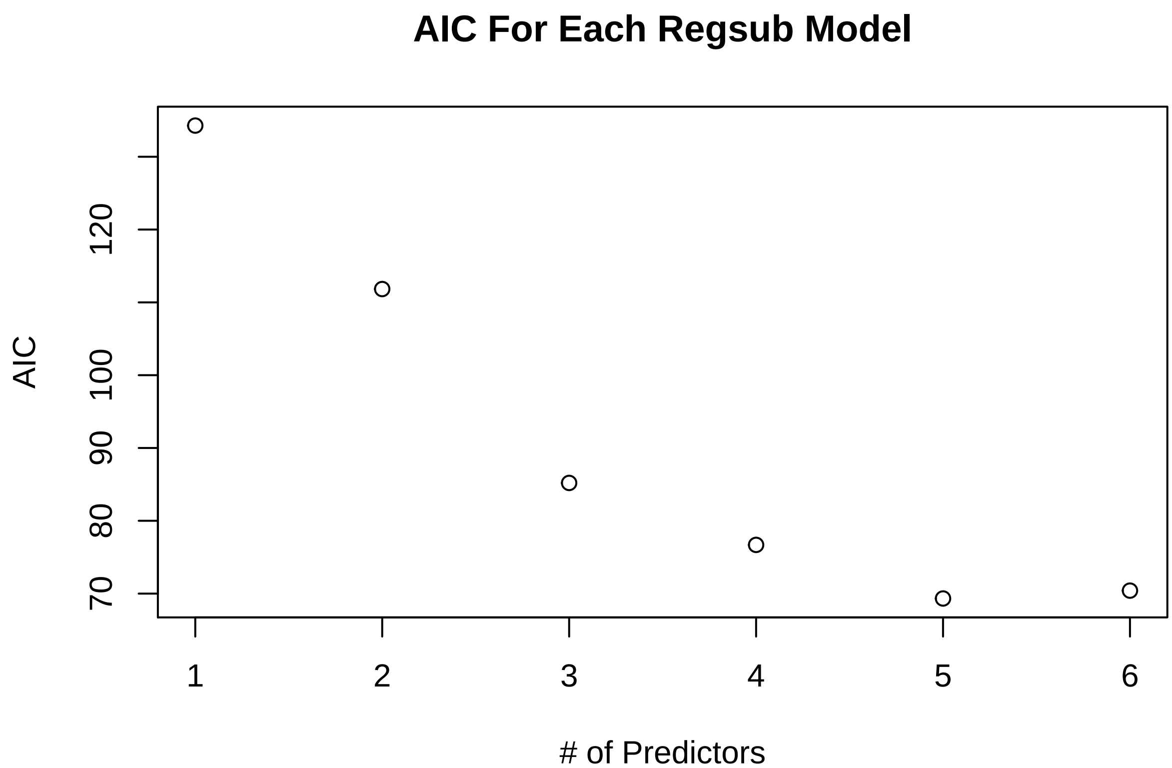
```
##
## Call:
## lm(formula = divorce ~ year + unemployed + femlab + marriage +
```

```
##     birth + military, data = divusa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9087 -0.9212 -0.0935  0.7447  3.4689
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 380.14761   99.20371   3.832 0.000274 ***
## year         -0.20312    0.05333  -3.809 0.000297 ***
## unemployed   -0.04933    0.05378  -0.917 0.362171
## femlab        0.80793    0.11487   7.033 1.09e-09 ***
## marriage      0.14977    0.02382   6.287 2.42e-08 ***
## birth        -0.11695    0.01470  -7.957 2.19e-11 ***
## military     -0.04276    0.01372  -3.117 0.002652 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.513 on 70 degrees of freedom
## Multiple R-squared:  0.9344, Adjusted R-squared:  0.9288
## F-statistic: 166.2 on 6 and 70 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = divorce ~ year + femlab + marriage + birth + military,
##     data = divusa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7586 -1.0494 -0.0424  0.7201  3.3075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 405.61670   95.13189   4.264 6.09e-05 ***
## year         -0.21790    0.05078  -4.291 5.52e-05 ***
## femlab        0.85480    0.10276   8.318 4.29e-12 ***
## marriage      0.15934    0.02140   7.447 1.76e-10 ***
## birth        -0.11012    0.01266  -8.700 8.43e-13 ***
## military     -0.04120    0.01360  -3.030  0.00341 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.511 on 71 degrees of freedom
## Multiple R-squared:  0.9336, Adjusted R-squared:  0.929
## F-statistic: 199.7 on 5 and 71 DF,  p-value: < 2.2e-16
```

6. The table shows the predictors in order of importance (best fit) according to the regsubsets function and their respective AIC score. The plot is able to give a graphical representation of the table. The optimal model is the one with the smallest AIC (69.33) and predictors: **femlab**, **birth**,**marriage**,**year**, & **military**.

| Model | Predictor | AIC |
|---|---|---|
| 1 | femlab | 134.28 |
| 2 | femlab+birth | 111.83 |
| 3 | femlab+birth+marriage | 85.2 |
| 4 | femlab+birth+marriage+year | 76.69 |
| 5 | femlab+birth+marriage+year+military | 69.33 |
| 6 | femlab+birth+marriage+year+military+unemployed | 70.41 |




```
## Subset selection object
## Call: regsubsets.formula(divorce ~ year + unemployed + femlab + marriage +
##     birth + military, data = divusa)
## 6 Variables  (and intercept)
##            Forced in Forced out
## year           FALSE      FALSE
## unemployed     FALSE      FALSE
## femlab         FALSE      FALSE
## marriage       FALSE      FALSE
## birth          FALSE      FALSE
## military       FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          year unemployed femlab marriage birth military
## 1  ( 1 ) " "  " "        "*"    " "      " "   " "
## 2  ( 1 ) " "  " "        "*"    " "      "*"   " "
```

```
## 3  ( 1 ) " "  " "        "*"    "*"      "*"   " "
## 4  ( 1 ) "*"  " "        "*"    "*"      "*"   " "
## 5  ( 1 ) "*"  " "        "*"    "*"      "*"   "*"
## 6  ( 1 ) "*"  "*"        "*"    "*"      "*"   "*"
```

```
## Start:  AIC=134.28
## divorce ~ femlab
##
##          Df Sum of Sq    RSS    AIC
## <none>                 418.1 134.28
## - femlab  1    2024.4 2442.5 268.19
```

```
##
## Call:
## lm(formula = divorce ~ femlab, data = divusa)
##
## Coefficients:
## (Intercept)       femlab
##     -3.6553       0.4387
```

```
## Start:  AIC=111.83
## divorce ~ femlab + birth
##
##          Df Sum of Sq     RSS    AIC
## <none>                 304.38 111.83
## - birth   1    113.73  418.10 134.28
## - femlab  1    865.16 1169.54 213.48
```

```
##
## Call:
## lm(formula = divorce ~ femlab + birth, data = divusa)
##
## Coefficients:
## (Intercept)       femlab        birth
##     6.37560      0.35985     -0.07864
```

```
## Start:  AIC=85.2
## divorce ~ femlab + birth + marriage
##
##            Df Sum of Sq     RSS     AIC
## <none>                   209.84  85.196
## - marriage  1     94.54  304.38 111.834
## - birth     1    194.92  404.76 133.781
## - femlab    1    949.45 1159.29 214.805
```

```
##
## Call:
## lm(formula = divorce ~ femlab + birth + marriage, data = divusa)
##
## Coefficients:
## (Intercept)       femlab        birth     marriage
##     -1.5455       0.4134      -0.1163       0.1261
```

```
## Start:  AIC=76.69
## divorce ~ femlab + birth + marriage + year
##
##            Df Sum of Sq    RSS     AIC
## <none>                   183.08  76.691
## - year      1    26.761 209.84  85.196
## - marriage  1   105.757 288.84 109.798
## - femlab    1   137.509 320.59 117.829
## - birth     1   183.446 366.53 128.140
```

```
##
## Call:
## lm(formula = divorce ~ femlab + birth + marriage + year, data = divusa)
##
## Coefficients:
## (Intercept)       femlab        birth     marriage         year
##    302.4928       0.7261      -0.1131       0.1344      -0.1619
```

```
## Start:  AIC=69.33
## divorce ~ femlab + birth + marriage + year + military
##
##            Df Sum of Sq    RSS     AIC
## <none>                   162.12  69.330
## - military  1    20.957 183.08  76.691
## - year      1    42.054 204.18  85.089
## - marriage  1   126.643 288.77 111.779
## - femlab    1   158.003 320.13 119.718
## - birth     1   172.826 334.95 123.203
```

```
##
## Call:
## lm(formula = divorce ~ femlab + birth + marriage + year + military,
##     data = divusa)
##
## Coefficients:
## (Intercept)       femlab        birth     marriage         year     military
##    405.6167       0.8548      -0.1101       0.1593      -0.2179      -0.0412
```

```
## Start:  AIC=70.41
## divorce ~ femlab + birth + marriage + year + military + unemployed
##
##              Df Sum of Sq    RSS     AIC
## - unemployed  1     1.925 162.12  69.330
## <none>                     160.20  70.410
## - military    1    22.231 182.43  78.417
## - year        1    33.199 193.40  82.912
## - marriage    1    90.468 250.66 102.884
## - femlab      1   113.214 273.41 109.572
## - birth       1   144.897 305.10 118.015
##
## Step:  AIC=69.33
## divorce ~ femlab + birth + marriage + year + military
##
```

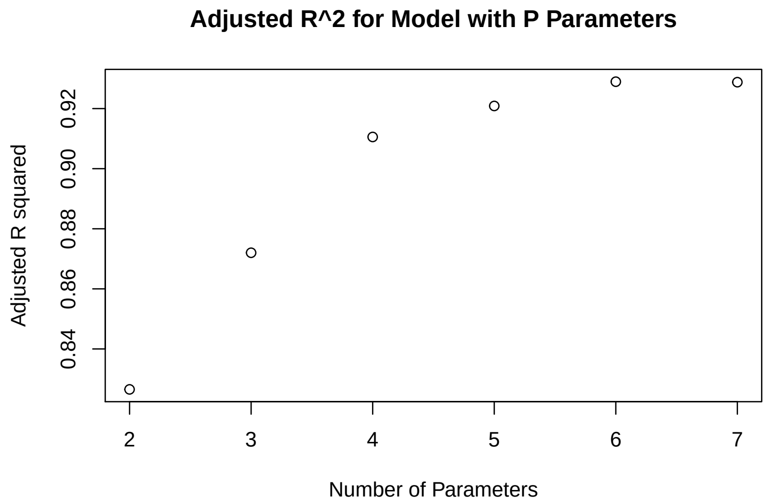
```
##            Df Sum of Sq    RSS     AIC
## <none>                  162.12  69.330
## - military  1    20.957 183.08  76.691
## - year      1    42.054 204.18  85.089
## - marriage  1   126.643 288.77 111.779
## - femlab    1   158.003 320.13 119.718
## - birth     1   172.826 334.95 123.203
```

```
##
## Call:
## lm(formula = divorce ~ femlab + birth + marriage + year + military,
##     data = divusa)
##
## Coefficients:
## (Intercept)       femlab        birth     marriage         year     military
##    405.6167       0.8548      -0.1101       0.1593      -0.2179      -0.0412
```

```
## Start:  AIC=70.41
## divorce ~ year + unemployed + femlab + marriage + birth + military
##
##              Df Sum of Sq    RSS     AIC
## - unemployed  1     1.925 162.12  69.330
## <none>                    160.20  70.410
## - military    1    22.231 182.43  78.417
## - year        1    33.199 193.40  82.912
## - marriage    1    90.468 250.66 102.884
## - femlab      1   113.214 273.41 109.572
## - birth       1   144.897 305.10 118.015
##
## Step:  AIC=69.33
## divorce ~ year + femlab + marriage + birth + military
##
##            Df Sum of Sq    RSS     AIC
## <none>                  162.12  69.330
## - military  1    20.957 183.08  76.691
## - year      1    42.054 204.18  85.089
## - marriage  1   126.643 288.77 111.779
## - femlab    1   158.003 320.13 119.718
## - birth     1   172.826 334.95 123.203
```

```
##
## Call:
## lm(formula = divorce ~ year + femlab + marriage + birth + military,
##     data = divusa)
##
## Coefficients:
## (Intercept)         year       femlab     marriage        birth     military
##    405.6167      -0.2179       0.8548       0.1593      -0.1101      -0.0412
```

7. Adjusted R^2- The model with the highest Adjusted R^2 is the model with a total of 6 parameters. This means that the final model will have 5 predictors, and such ones being (in order of importance)- **femlab**,**birth**,**marriage**,**year**, and **military**. All predictors are significant.The R^2 is exceptionally high 93%, which signifies a good fit. The standard errors are all small (within .10). Overall, this model is a good fit.




```
##
## Call:
## lm(formula = divorce ~ femlab + birth + marriage + year + military,
##     data = divusa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7586 -1.0494 -0.0424  0.7201  3.3075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 405.61670   95.13189   4.264 6.09e-05 ***
## femlab        0.85480    0.10276   8.318 4.29e-12 ***
## birth        -0.11012    0.01266  -8.700 8.43e-13 ***
## marriage      0.15934    0.02140   7.447 1.76e-10 ***
## year         -0.21790    0.05078  -4.291 5.52e-05 ***
## military     -0.04120    0.01360  -3.030  0.00341 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.511 on 71 degrees of freedom
## Multiple R-squared:  0.9336, Adjusted R-squared:  0.929
## F-statistic: 199.7 on 5 and 71 DF,  p-value: < 2.2e-16
```

8. Least Absolute Deviations & OLS are similar in terms of intercetp, coeffiecients sign/values, and standard errors. The majority of the predicotrs are similar in both models. The major difference pertains to the predictors **military** & **unemployed**. In the OLS only the last one is insignificant (alpha=5%); whereas in the LAD model *unemployed* is significant & *military* is insignificant.

```
## Warning in summary.rq(lad_fit, se = "nid"): 3 non-positive fis
```

```
##
## Call: rq(formula = divorce ~ ., data = divusa)
##
```

```
## tau: [1] 0.5
##
## Coefficients:
##             Value     Std. Error t value   Pr(>|t|)
## (Intercept) 349.26601 105.32648    3.31603   0.00145
## year         -0.18522   0.05594   -3.31130   0.00147
## unemployed   -0.08944   0.03735   -2.39430   0.01933
## femlab        0.74767   0.11091    6.74110   0.00000
## marriage      0.09962   0.03265    3.05094   0.00322
## birth        -0.09707   0.01460   -6.64849   0.00000
## military     -0.03910   0.02547   -1.53518   0.12925
```

9. **Yes**, LASSO can be used for variable selection. Thorugh LASSO one can perform regularization and feature selection. It penalizes the coefficients of the regression variables shrinking some of them to zero. After the shrinking procees, he variables that still have a non-zero coefficient are selected to be part of the model. The extent to which the regularization is taken depends on lambda parameter. In short,LASSO helps to increase the model interpretability by eliminating irrelevant variables that are not associated with the response variable. Hence, LASSO indirectly performs variable selection for us.

   **No**, Ridge regression does perform variable slection. Rather, it "shrinks" all predictor coefficient estimates toward zero. These estimates might get very close to zero, but they might never become equal to zero.

10. In a scenarion where the errors are correlated, but all other assumptions are met one can use **Generalized Least Squares**. One example, where one might have correlated data is temporal data. If the errors are correlated, one can calculated this by computing the correlation between succesive pairs of residuals. Some models that can aid with this are a simple **AR(1)**, or a more sophisticated **ARMA** model.

11. KNN is a non-parametric machine learning algorithim-does not make any assumption about the distribution of the data. It calculates the distance between a specific number of observations (k); it finds the k closest neighbors, and finally it assigns an observation to the category that make up the majority of the neighbors. Now, **bias** can be tought of as the systematic error in a determinate estimatation. Ideally, one wants low bias- the less the bias (error) the closer to the truth. **Variance**, can be thought of as the concept of estimating the function f(.) from different datasets (of the same population). Ideally, we do not want the f(.) estimated function to differ too much from one another- we do not want the f() to be too specific to the dataset at hand. Instead, we want the estimated f(.) to do well with other data as well. Hence, it is desirable to have low variance as well. Now, ideally we want a model with low bias and low variance, but unfortunately these are two competing resources, and as a result we need to compromise between the two. In the case of KNN, the choice of K will determine the degree of this compromise. As we choose smaller and smaller K's, our bias will eventually become non-existent, but on the other hand, our variance will sky-rocket high. This model will be no good. If we pick a K that is too big, we will end up getting a samller variance, but our bias will increase given we will introduce error into our estimate. Hence, the best strategy is to pick a K such that bias & variance are not too high/nor too low, but optimal. The optimal scenario depends on the data and on the application. Usually such K is choosen through cross-validation & tends to be between 5-30 (generally, not always). Finally, this is called a trade-off because you are never picking the most flexible model, nor the least flexible model for they will favor only one side of the formula. Instead, you attempt to favor an optimal balance where there is neither overfitting nor underfitting.