Worksheet: Multiple Regression and Variable Selection

Part 1: What Not To Do. Party in power and economic performance.

Go to https://projects.fivethirtyeight.com/p-hacking/

(a) Suppose you are a Democratic data analyst with an agenda: You want to show that the economy performs better when Democrats are in power.

- Choose "Democrats" for the political party. The horizontal axis of the plot now measures the amount of power held by Democrats, and the vertical axis the performance of the economy. Your goal is to find statistically significant evidence of an association between these variables (p-value as small as you can make it), with a positive slope
- By changing the settings for which politicians are included, how economic performance is measured, and the options for weighting politicians by how powerful they are and whether or not recessions are excluded, manipulate the variables used until you have found statistically significant evidence of a positive association between these variables.

You win! Case proved, write it up and get published.

(b) Suppose you are a Republican data analyst with an agenda: You want to show that the economy performs better when Democrats are in power.

- Choose "Republicans" for the political party. The horizontal axis of the plot now measures the amount of power held by Republicans, and the vertical axis the performance of the economy. Your goal is to find statistically significant evidence of an association between these variables (p-value as small as you can make it), with a positive slope
- By changing the settings for which politicians are included, how economic performance is measured, and the options for weighting politicians by how powerful they are and whether or not recessions are excluded, manipulate the variables used until you have found statistically significant evidence of a positive association between these variables.

You win! Case proved, write it up and get published.

What's the point?

- You can find "statistically significant" evidence of anything if that is your goal and you are flexible enough in your data analysis. That doesn't mean your conclusions are correct.
- Formally, a p-value only measures the strength of evidence against the null hypothesis of the test *if the analysis was pre-specified* before looking at the data. If the test or the model you fit was dependent on the data in any way, the p-value is unreliable as an indicator of strength of evidence.
- Our goal is not to find statistically significant results. Our goal is to present an honest discussion of what the data can and cannot tell us about the world, complete with limitations of our analysis. A result is only convincing if it shows up in a variety of reasonable analyses of the data.
- We *must* present results from all reasonable models for the data based on a variety of reasonable decisions about what variables are included in the model and how those variables are defined.
- Any time someone has a really complicated data set and they present only a few findings from a single model, you should be very suspicious.

Part 2: What To Do. Nursing Salaries.

We have data about 52 licensed nursing home facilities in New Mexico, collected by the Department of Health and Social Services of the State of New Mexico. Let's use these data to estimate the relationship between the salaries of nurses at a given facility (`NurseSalaries`, our response variable) and a variety of other characteristics of the facility. The variables in the data set are:

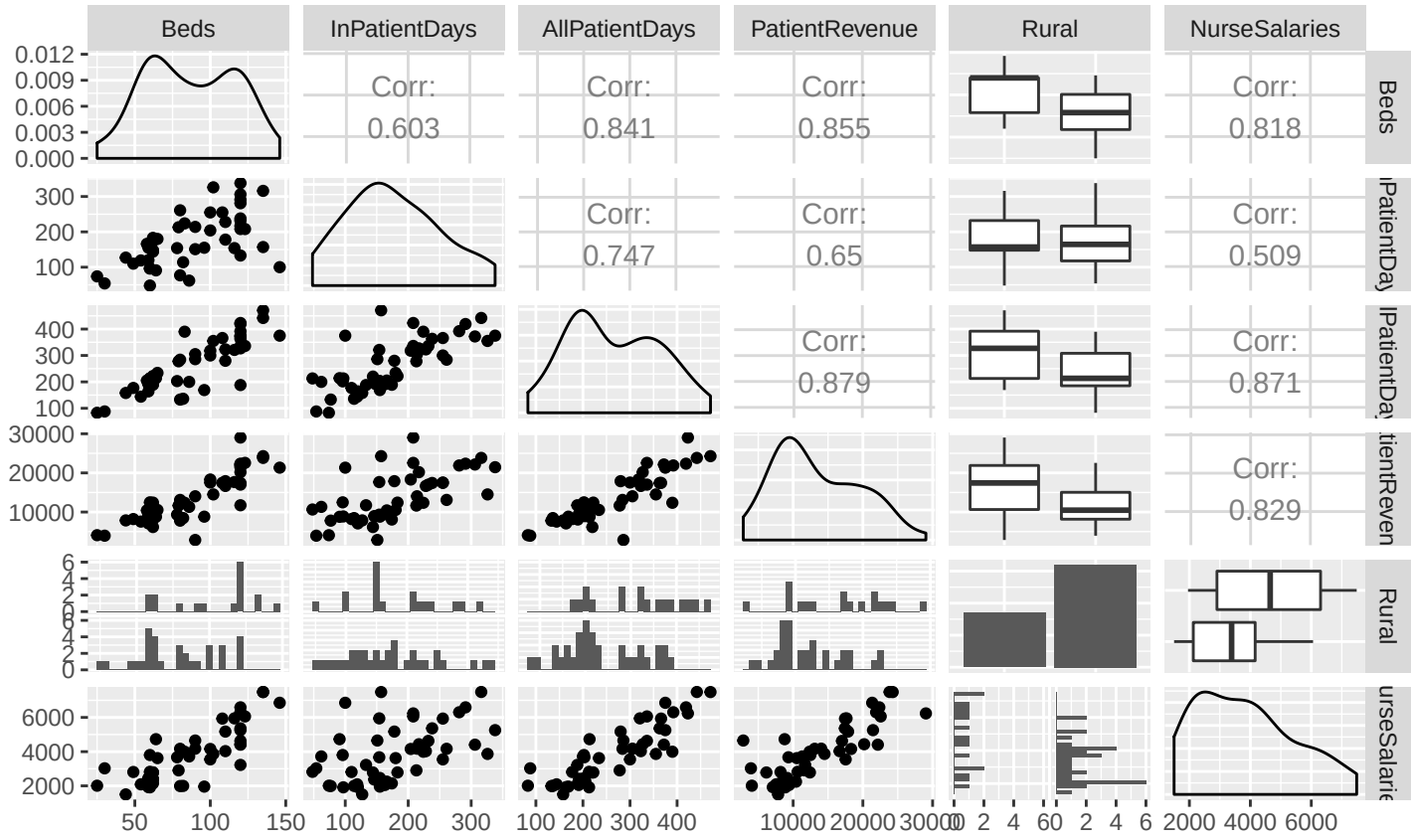
- `Beds`: Number of beds in the nursing home
- `InPatientDays`: Annual medical in-patient days (in hundreds)
- `AllPatientDays`: Annual total patient days (in hundreds)
- `PatientRevenue`: Annual patinet care revenue (in hundreds of dollars)
- `Rural`: Either "Rural" or "Non-Rural"
- `NurseSalaries`: Annual nursing salaries (in hundreds of dollars)

I have removed three outlying/high leverage observations. In order to focus on other aspects of the analysis, for today we will ignore these data points (ordinarily, we should check and see whether our conclusions depend on whether those observations are included).

```
## # A tibble: 6 x 6
##    Beds InPatientDays AllPatientDays PatientRevenue Rural     NurseSalaries
##   <dbl>         <dbl>          <dbl>          <dbl> <chr>             <dbl>
## 1    59           155            203           9160 Rural              2459
## 2   120           281            392          21900 Non-Rural          6304
## 3   120           291            419          22354 Non-Rural          6590
## 4   120           238            363          17421 Non-Rural          5362
## 5    65           180            234          10531 Rural              3622
## 6   120           306            372          22147 Rural              4406
```

Here is a pairs plot of the data.

```
library(GGally)
ggpairs(nursing)
```



1. Based on the pairs plot, perform an initial check of the conditions of linearity, equal variance, and no outliers/high leverage observations. Also check and see whether there are any indications of potential problems with multicollinearity.

2. Based on the pairs plot, which of the explanatory variables appear to have the strongest association with nursing salaries?

Here is a model that has NurseSalaries as the response, all other variables in the data set as explanatory variables, and does not include any interaction terms. Also shown are the variance inflation factors (VIF) for the coefficient estimates in this model.

```
lm_fit <- lm(NurseSalaries ~ Beds + InPatientDays + AllPatientDays + PatientRevenue + Rural, data = nursing)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ Beds + InPatientDays + AllPatientDays +
##     PatientRevenue + Rural, data = nursing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1312.0  -480.3  -192.4   675.0  1698.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     51.95128  455.81747   0.114  0.90979
## Beds            10.61995    7.16125   1.483  0.14537
## InPatientDays   -6.77619    2.23580  -3.031  0.00412 **
## AllPatientDays  13.21671    2.83591   4.660 3.05e-05 ***
## PatientRevenue   0.04464    0.03907   1.142  0.25958
## RuralRural     -17.02622  252.20773  -0.068  0.94649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 713.4 on 43 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8104
## F-statistic: 42.04 on 5 and 43 DF,  p-value: 1.744e-15
```

```
vif(lm_fit)
```

```
##           Beds  InPatientDays AllPatientDays PatientRevenue          Rural
##       4.380440       2.586838       7.196287       5.526630       1.387749
```

3. Do the variance inflation factors indicate potential issues with multicollinearity? What does the VIF for Beds mean for the size of a confidence interval for β_1 in the model?

4. Below are results from an all subsets regression. Based on these results, which models have roughly equivalent performance?

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.6.3
```

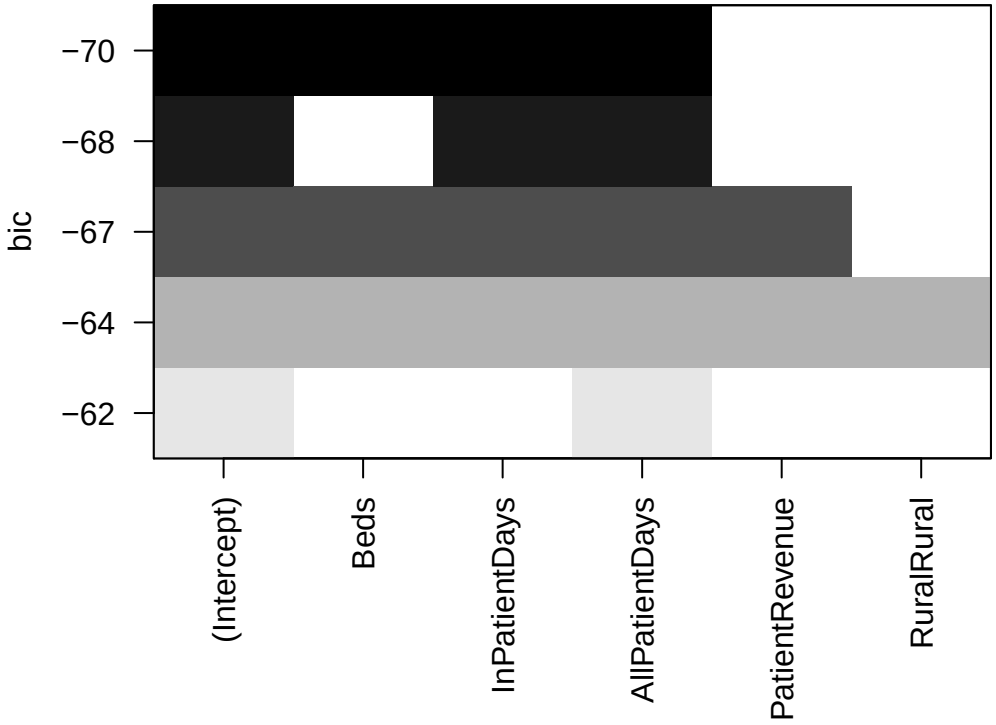
```
candidate_models <- regsubsets(NurseSalaries ~ Beds + InPatientDays + AllPatientDays + PatientRevenue + Rural,
summary(candidate_models)
```

```
## Subset selection object
## Call: regsubsets.formula(NurseSalaries ~ Beds + InPatientDays + AllPatientDays +
##     PatientRevenue + Rural, data = nursing)
## 5 Variables  (and intercept)
##                Forced in Forced out
## Beds               FALSE      FALSE
## InPatientDays      FALSE      FALSE
## AllPatientDays     FALSE      FALSE
## PatientRevenue     FALSE      FALSE
## RuralRural         FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Beds InPatientDays AllPatientDays PatientRevenue RuralRural
## 1  ( 1 ) " "  " "           "*"             " "             " "
## 2  ( 1 ) " "  "*"           "*"             " "             " "
## 3  ( 1 ) "*"  "*"           "*"             " "             " "
## 4  ( 1 ) "*"  "*"           "*"             "*"             " "
## 5  ( 1 ) "*"  "*"           "*"             "*"             "*"
```

```
summary(candidate_models)$bic
```

```
## [1] -61.93374 -68.40657 -69.84360 -67.41191 -63.52529
```

```
plot(candidate_models)
```


bic
−70
−68
−67
−64
−62
(Intercept)
Beds
InPatientDays
AllPatientDays
PatientRevenue
RuralRural

5. Here are summaries of the model fits for the three best models in part 4. Summarize what these models have to say about the associations between the explanatory and response variables in the data set.

```
fit1 <- lm(NurseSalaries ~ InPatientDays + AllPatientDays, data = nursing)
summary(fit1)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ InPatientDays + AllPatientDays,
##     data = nursing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1876.7  -479.6  -174.1   535.6  1590.9
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     172.858    313.889   0.551  0.58451
## InPatientDays    -7.137      2.168  -3.292  0.00192 **
## AllPatientDays   18.712      1.649  11.348 6.28e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 739.2 on 46 degrees of freedom
## Multiple R-squared:  0.8049, Adjusted R-squared:  0.7964
## F-statistic:  94.9 on 2 and 46 DF,  p-value: < 2.2e-16
```

```
fit2 <- lm(NurseSalaries ~ Beds + InPatientDays + AllPatientDays, data = nursing)
summary(fit2)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ Beds + InPatientDays + AllPatientDays,
##     data = nursing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1345.6  -493.9  -231.6   678.0  1756.4
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -95.948    322.976  -0.297  0.76778
## Beds             14.315      6.296   2.274  0.02780 *
## InPatientDays    -6.804      2.081  -3.269  0.00207 **
## AllPatientDays   14.801      2.335   6.340 9.75e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 707.8 on 45 degrees of freedom
## Multiple R-squared:  0.825,  Adjusted R-squared:  0.8134
## F-statistic: 70.72 on 3 and 45 DF,  p-value: < 2.2e-16
```

```
fit3 <- lm(NurseSalaries ~ Beds + InPatientDays + AllPatientDays + PatientRevenue, data = nursing)
summary(fit3)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ Beds + InPatientDays + AllPatientDays +
##     PatientRevenue, data = nursing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1318.2  -484.4  -198.6   669.8  1697.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     31.78007  340.30998   0.093  0.92602
## Beds            10.68348    7.01838   1.522  0.13511
## InPatientDays   -6.82842    2.07384  -3.293  0.00196 **
## AllPatientDays  13.27368    2.67666   4.959  1.1e-05 ***
## PatientRevenue   0.04446    0.03854   1.154  0.25488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 705.2 on 44 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8147
## F-statistic: 53.77 on 4 and 44 DF,  p-value: < 2.2e-16
```