Lab 14: 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)

```
## # A tibble: 6 x 6
##    Beds InPatientDays AllPatientDays PatientRevenue Rural     NurseSalaries
##   <dbl>         <dbl>          <dbl>          <dbl> <chr>             <dbl>
## 1   244           128            385          23521 Non-Rural          5230
## 2    59           155            203           9160 Rural              2459
## 3   120           281            392          21900 Non-Rural          6304
## 4   120           291            419          22354 Non-Rural          6590
## 5   120           238            363          17421 Non-Rural          5362
## 6    65           180            234          10531 Rural              3622
```

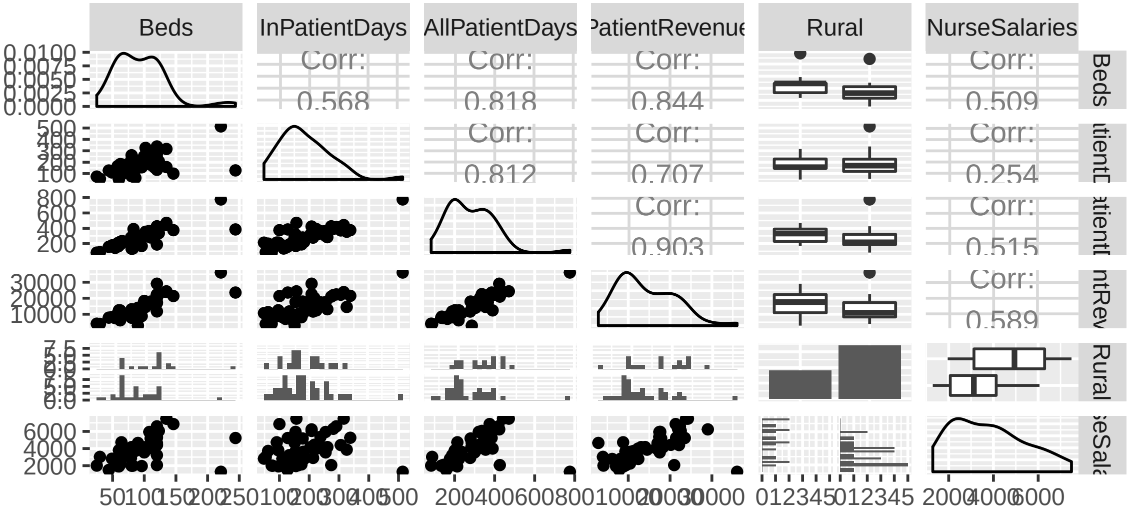
1. Make a pairs plot of the data.

```r
library(GGally)
```

```
##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

```r
ggpairs(nursing)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2. Based on your pairs plot, perform an initial check of the conditions of linearity, equal variance, and no outliers/high leverage observations. You don't need to do anything about these outlying/high leverage observations yet, we'll deal with them later.

Based on the plots in the bottom row, showing the response vs. each of the explanatory variables, the conditions of linearity and constant variance of the residuals look OK.

There are several outlying and high leverage observations. High leverage observations are those with explanatory variable values that stand out from the other points. For example, in the plot of InPatientDays vs. AllPatientDays, there is one observations that is far from the others; this is a high leverage observation since we are thinking about two explanatory variables (InPatientDays and AllPatientDays). Across all of the plots, it appears that there are about 2 high leverage observations.

Additionally, there is an outlying observation in the scatter plot of AllPatientDays vs. NurseSalaries, with a value of AllPatientDays around 400 and a small value of NurseSalaries around 2000. This is not a high leverage observation because there are many observations with AllPatientDays about 400, but it is an outlier because its nurse salary is very different from the nurse salaries of other observations with that value of AllPatientDays. Similarly, some outliers are visible in the plot of PatientRevenue vs. NurseSalaries.

3. Based on your pairs plot, make some statements about which of the explanatory variables have the strongest association with nursing salaries.

It appears that AllPatient and PatientRevenue have the strongest association with nursing salaries, based on the scatter plots.

4. Fit 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.

```r
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
## -4458.2  -715.8    53.4   882.4  2177.1
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2950.09214  613.51531   4.809 1.67e-05 ***
## Beds             -4.68139    8.72466  -0.537   0.5942
## InPatientDays    -5.53517    3.90686  -1.417   0.1633
## AllPatientDays    2.83597    4.54695   0.624   0.5359
## PatientRevenue    0.15015    0.06565   2.287   0.0268 *
## RuralRural     -934.57193  419.13494  -2.230   0.0307 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1271 on 46 degrees of freedom
## Multiple R-squared:  0.4712, Adjusted R-squared:  0.4137
## F-statistic: 8.199 on 5 and 46 DF,  p-value: 1.344e-05
```

5. Calculate the variance inflation factors (VIF) for the coefficient estimates in this model. Do these indicate potential issues with multicollinearity? What is the interpretation of the VIF for Beds? For AllPatientDays?

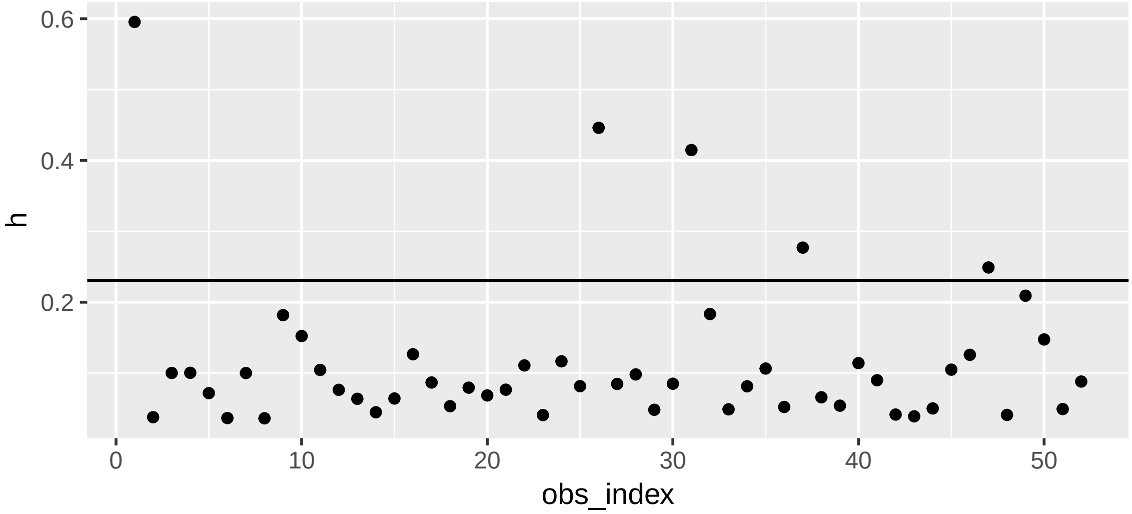
```r
vif(lm_fit)
```

```
##           Beds  InPatientDays AllPatientDays PatientRevenue          Rural
##       4.013106       3.652011       9.537951       6.621293       1.280639
```

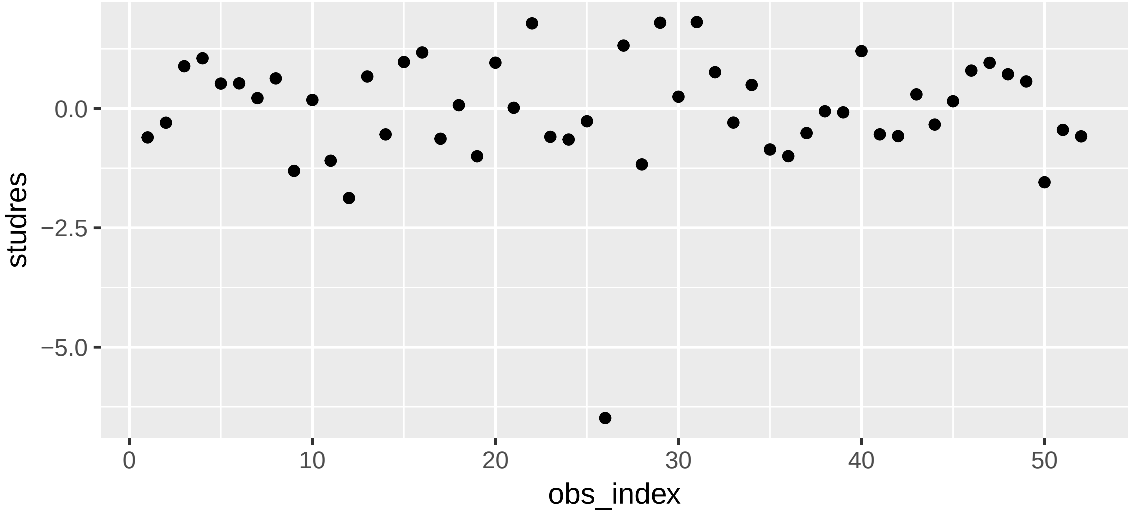
The variance inflation factors for AllPatientDays and PatientRevenue are both greater than 5, which is a typical threshold for indicating problems with multicollinearity. The VIF of 9.5 for AllPatientDays means that the variance of the coefficient estimate for AllPatientDays is about 9.5 times larger than it would be if AllPatientDays were not correlated with the other explanatory variables. In turn this means that our confidence interval for the coefficient of AllPatientDays is about 3 times wider than it would have been if AllPatientDays were not correlated with the other explanatory variables. Similarly, a confidence interval for the coefficient of Beds is about 2 times wider than it would have been if Beds were not correlated with the other explanatory variables.

6. Make plots showing the leverage, studentized residual, and Cook's distance for each observation. Do these diagnostics suggest that any observations are worth investigating further?

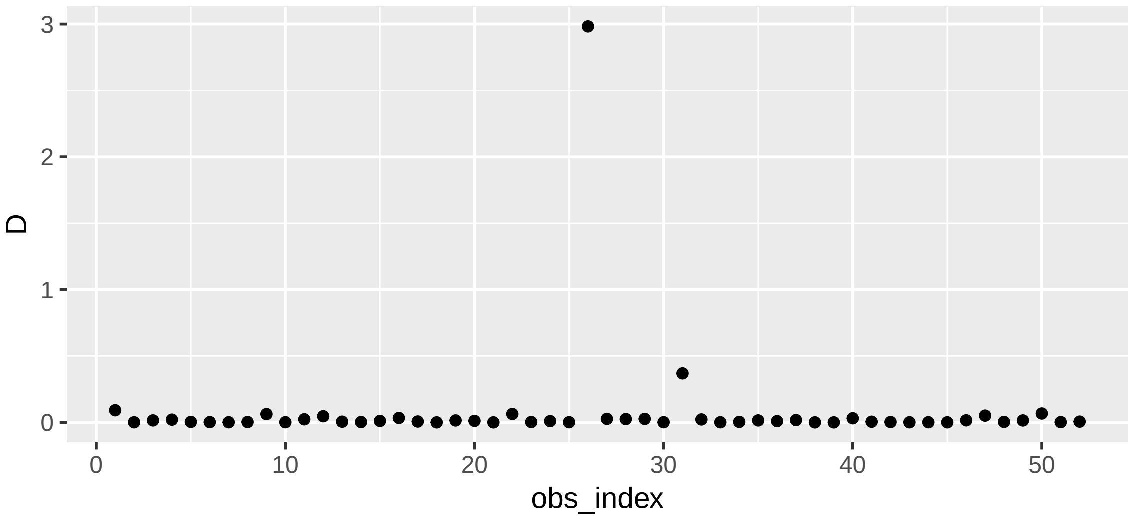
```r
# Code for Number 6
nursing <- nursing %>%
  mutate(
    obs_index = row_number(),
    h = hatvalues(lm_fit),
    studres = rstudent(lm_fit),
    D = cooks.distance(lm_fit)
  )

ggplot(data = nursing, mapping = aes(x = obs_index, y = h)) +
  geom_hline(yintercept = 2*6/ nrow(nursing))+
  geom_point()
```



```
ggplot(data = nursing, mapping = aes(x = obs_index, y = studres)) +
  geom_point()
```



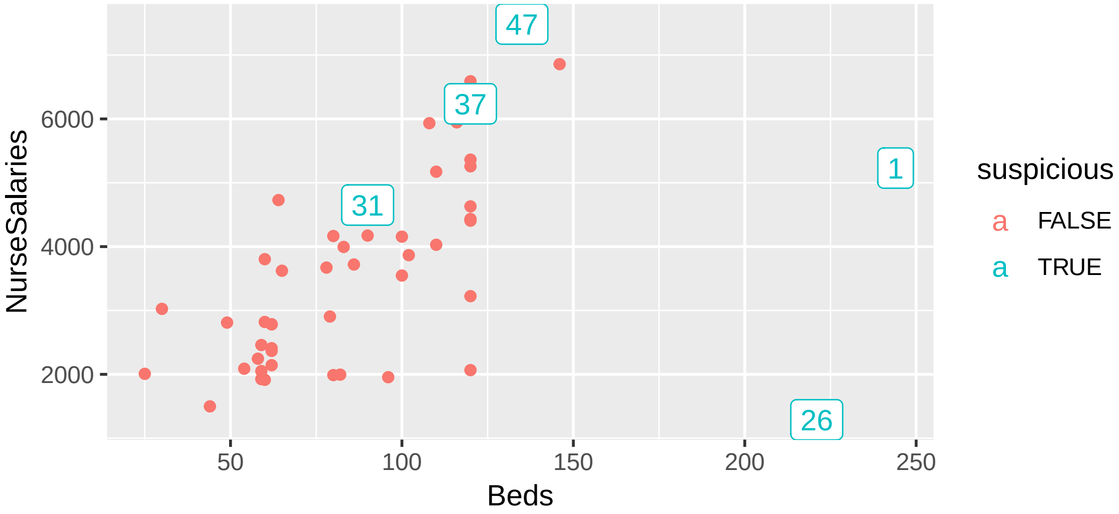
```
ggplot(data = nursing, mapping = aes(x = obs_index, y = D)) +
  geom_point()
```



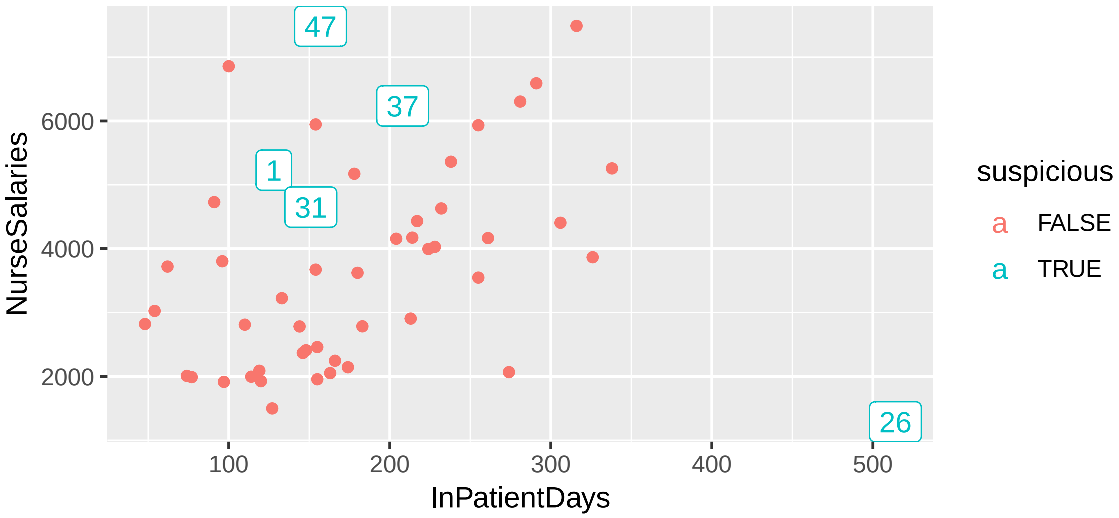
All three plots suggest that we should investigate observation number 26 in more depth. Additionally, the plot of leverages suggests that we might investigate observations 1, 31, 37, and 47.

7. Make scatter plots of each quantitative explanatory variable vs. the response, highlighting any observations you identified above.

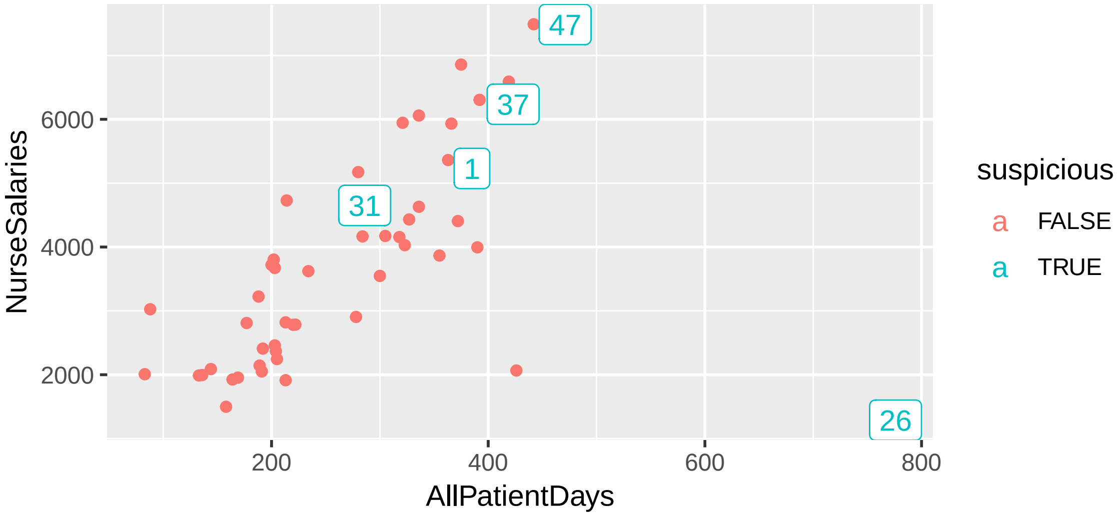
```
obs_to_investigate <- c(1, 26, 31, 37, 47)

nursing <- nursing %>%
  mutate(
    suspicious = row_number() %in% obs_to_investigate
  )

ggplot(data = nursing, mapping = aes(x = Beds, y = NurseSalaries, color = suspicious)) +
  geom_point() +
  geom_label(data = nursing %>% filter(suspicious), mapping = aes(label = obs_index))
```



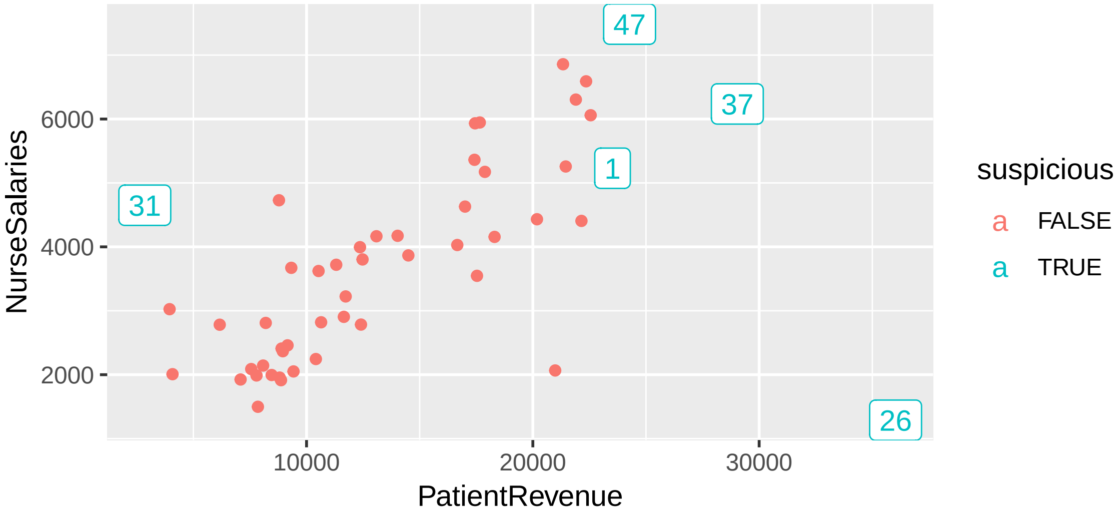
```
ggplot(data = nursing, mapping = aes(x = InPatientDays, y = NurseSalaries, color = suspicious)) +
  geom_point() +
  geom_label(data = nursing %>% filter(suspicious), mapping = aes(label = obs_index))
```



```r
ggplot(data = nursing, mapping = aes(x = AllPatientDays, y = NurseSalaries, color = suspicious)) +
  geom_point() +
  geom_label(data = nursing %>% filter(suspicious), mapping = aes(label = obs_index))
```



```r
ggplot(data = nursing, mapping = aes(x = PatientRevenue, y = NurseSalaries, color = suspicious)) +
  geom_point() +
  geom_label(data = nursing %>% filter(suspicious), mapping = aes(label = obs_index))
```



8. We want to conduct an analysis of the data both with and without the outliers. To start, let's leave the outliers in. Perform an all subsets regression to identify a set of models that have roughly equivalent performance using the full data set.

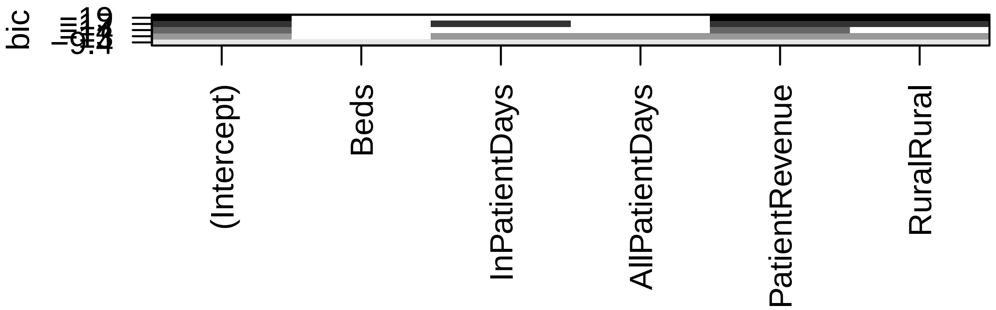
```
library(leaps)
candidate_models <- regsubsets(NurseSalaries ~ Beds + InPatientDays + AllPatientDays + PatientRevenue + Rural,
summary(candidate_models)
```

```
## 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] -14.290637 -18.817125 -16.751064 -13.053199  -9.426401
```

```
plot(candidate_models)
```



9. Fit any candidate models that you identified in part 7 and print out the model summaries. What do these models indicate about which variables are associated with the response?

```
fit1 <- lm(NurseSalaries ~ PatientRevenue + Rural, data = nursing)
summary(fit1)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ PatientRevenue + Rural, data = nursing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4729.8  -685.3    67.5   832.3  2125.1
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.859e+03  5.333e+02   5.361 2.22e-06 ***
## PatientRevenue  1.189e-01  2.633e-02   4.517 3.96e-05 ***
## RuralRural     -1.126e+03  3.822e+02  -2.946  0.00492 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1261 on 49 degrees of freedom
## Multiple R-squared:  0.4456, Adjusted R-squared:  0.4229
## F-statistic: 19.69 on 2 and 49 DF,  p-value: 5.298e-07
```

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

```
##
## Call:
## lm(formula = NurseSalaries ~ PatientRevenue + Rural + InPatientDays,
##     data = nursing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4305.8  -773.9   -39.2   920.5  2228.2
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2941.42946  532.76801   5.521 1.34e-06 ***
## PatientRevenue    0.15714    0.03882   4.048 0.000187 ***
## RuralRural     -962.85447  398.50432  -2.416 0.019538 *
## InPatientDays    -3.98110    2.99060  -1.331 0.189411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1251 on 48 degrees of freedom
## Multiple R-squared:  0.4653, Adjusted R-squared:  0.4319
## F-statistic: 13.92 on 3 and 48 DF,  p-value: 1.167e-06
```

Both models we identified as having low BIC provided very strong evidence of a positive association between patient revenue and mean nurse salaries in the population of nursing homes similar to those in this study, after accounting for the effects of the setting (rural or urban) and the number of in-patient days.

Additionally, we have strong evidence of an association between the setting (rural or urban) and nurse salaries after accounting for the effect of patient revenue; there is only moderately strong evidence of this association after also accounting for in patient days.

Neither of the models we considered provides evidence of an association between any of the other explanatory variables in the data set and nurse salaries in the population of nursing homes, after accounting for patient revenue and urban/rural setting.

10. Create a version of the data set that does not include any suspect observations. Go through the all subsets regression process again with your new data set to identify candidate models based on the filtered data set.

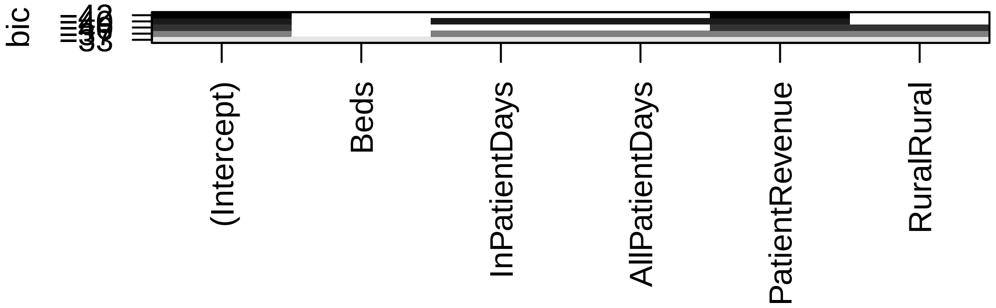
```
# We had previously indicated that observations 1, 26, 31, 37, and 47 were "suspicious"

nursing_no_suspicious <- nursing %>%
  filter(
    !suspicious
  )

candidate_models_no_suspicious <- regsubsets(NurseSalaries ~ Beds + InPatientDays + AllPatientDays + PatientRe
summary(candidate_models_no_suspicious)
```

```
## Subset selection object
## Call: regsubsets.formula(NurseSalaries ~ Beds + InPatientDays + AllPatientDays +
##     PatientRevenue + Rural, data = nursing_no_suspicious)
## 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_no_suspicious)$bic
```

```
## [1] -41.53909 -39.68480 -40.08103 -36.91592 -33.23388
```

```
plot(candidate_models_no_suspicious)
```



11. Fit any candidate models that you identified in part 9 and print out the model summaries. What do these models indicate about which variables are associated with the response?

```
fit3 <- lm(NurseSalaries ~ PatientRevenue, data = nursing_no_suspicious)
summary(fit3)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ PatientRevenue, data = nursing_no_suspicious)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3362.7  -541.1  -120.2   531.6  2048.4
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    703.21285  354.22988   1.985   0.0532 .
## PatientRevenue   0.22516    0.02467   9.126 8.41e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 931.4 on 45 degrees of freedom
## Multiple R-squared:  0.6492, Adjusted R-squared:  0.6414
## F-statistic: 83.29 on 1 and 45 DF,  p-value: 8.407e-12
```

```
fit4 <- lm(NurseSalaries ~ PatientRevenue + InPatientDays + AllPatientDays, data = nursing_no_suspicious)
summary(fit4)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ PatientRevenue + InPatientDays +
##     AllPatientDays, data = nursing_no_suspicious)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3622.7  -421.4    45.9   539.7  1565.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    562.74986  386.47606   1.456  0.15263
## PatientRevenue   0.16823    0.05828   2.886  0.00607 **
## InPatientDays   -6.89321    3.01490  -2.286  0.02722 *
## AllPatientDays   8.17855    4.05348   2.018  0.04989 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 891.6 on 43 degrees of freedom
## Multiple R-squared:  0.6929, Adjusted R-squared:  0.6714
## F-statistic: 32.33 on 3 and 43 DF,  p-value: 4.249e-11
```

12. Sum up your process, noting which of your findings if any were consistent across the different models you fit and which results were more dependent on the other explanatory variables included in the model or whether outlying/influential observations were included.

I probably just have just asked #12, not both 12 and 9. The answer below just my answer to number 9, modified to also include discussion of results from the models in number 11.

Based on the full data set, both models we identified as having low BIC provided very strong evidence of a positive association between patient revenue and mean nurse salaries in the population of nursing homes similar to those in this study, after accounting for the effects of the setting (rural or urban) and the number of in-patient days. This was also the case in the models we identified after removing 5 outlying or high leverage observations; here, one of the selected models included only patient revenue as an explanatory variable, and the other accounted for inpatient days and all patient days.

Based on the full data set, we have strong evidence of an association between the setting (rural or urban) and nurse salaries after accounting for the effect of patient revenue; there is only moderately strong evidence of this association after also accounting for in patient days. However, after removing 5 outlying or high leverage observations, the rural variable was not selected for inclusion in our models, suggesting that once these outliers were removed, there was not strong evidence of an association between an urban or rural setting and nurse salaries, after accounting for the other explanatory variables in the model.

One of the models we considered after removing the outlying and high leverage observations included InPatientDays and AllPatientDays as explanatory variables, and offered some evidence of an association between these variables and nurse salaries after accounting for patient revenue. However, BIC did not lead us to select these variables for any of our other models, indicating only week evidence of an association between these variables and the response overall, after accounting for the other explanatory variables.

None of the models we considered included beds as an explanatory variable, indicating that there is no evidence of an association between beds and nurse salaries after accounting for the other explanatory variables.

13. Using your version of the data set without the suspicious observations, fit a model that has only Beds as an explanatory variable, and print the model summary. Also fit a model that has includes Beds, PatientRevenue, InPatientDays, and AllPatientDays as explanatory variables, and print the model summary. What is the interpretation of the coefficient estimate labeled as Beds in each of these models?

```
fit_one <- lm(NurseSalaries ~ Beds, data = nursing_no_suspicious)
summary(fit_one)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ Beds, data = nursing_no_suspicious)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2958.30  -586.34    29.46   710.60  1901.48
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  316.919    463.244   0.684    0.497
## Beds          39.228      5.096   7.698 9.53e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1033 on 45 degrees of freedom
## Multiple R-squared:  0.5684, Adjusted R-squared:  0.5588
## F-statistic: 59.25 on 1 and 45 DF,  p-value: 9.527e-10
```

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

```
##
## Call:
## lm(formula = NurseSalaries ~ Beds + PatientRevenue + InPatientDays +
##     AllPatientDays, data = nursing_no_suspicious)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3600.9  -472.4   -88.1   504.5  1579.1
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     464.39565  446.37439   1.040   0.3041
## Beds              4.66055   10.28106   0.453   0.6527
## PatientRevenue    0.14710    0.07505   1.960   0.0567 .
## InPatientDays    -6.72578    3.06548  -2.194   0.0338 *
## AllPatientDays    7.97850    4.11518   1.939   0.0593 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 899.9 on 42 degrees of freedom
## Multiple R-squared:  0.6944, Adjusted R-squared:  0.6652
## F-statistic: 23.85 on 4 and 42 DF,  p-value: 2.41e-10
```

Interpretation of beds, model with only beds as an explanatory variable: We estimate that a 1 unit increase in beds is associated with an increase in nursing salary of about 39 units in the population of nursing homes similar to those in this study.

Interpretation of beds, model with 4 explanatory variables: After accounting for the effects of patient revenue, in patient days, and all patient days, we estimate that a 1 unit increase in beds is associated with an increase in nursing salary of about 4.7 units in the population of nursing homes similar to those in this study.

14. Based on each of your model fits in part 13, conduct a test of whether in the population, the coefficient of Beds is equal to 0. In each case, state your conclusion in context in terms of the strength of evidence against the null hypothesis.

Based on model with only beds as an explanatory variable:

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

The p-value for the test is 9.53e-10. The data offer extremely strong evidence against the null hypothesis of no association between the number of beds and nursing salaries.

Based on model with multiple explanatory variables:

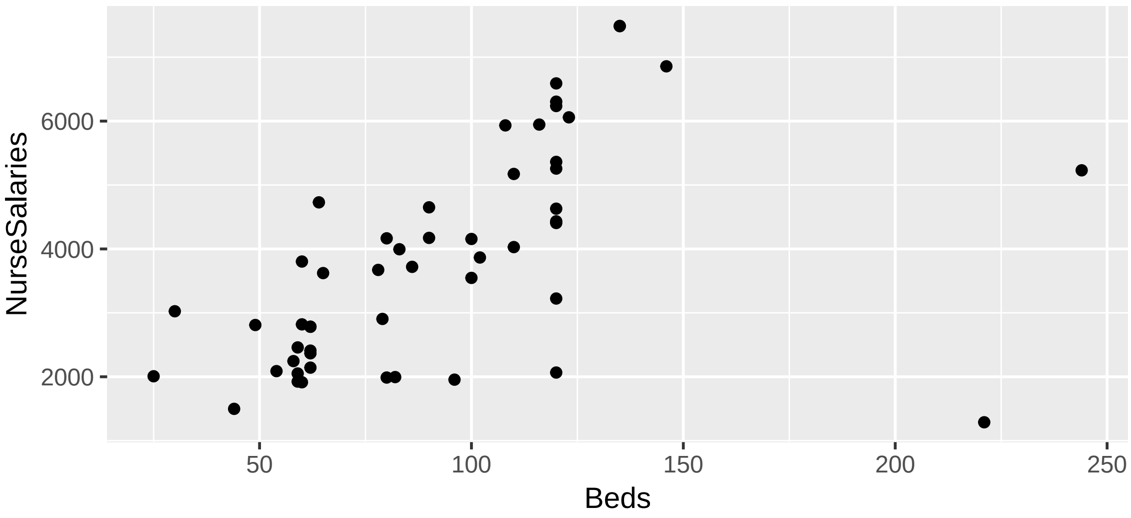
$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

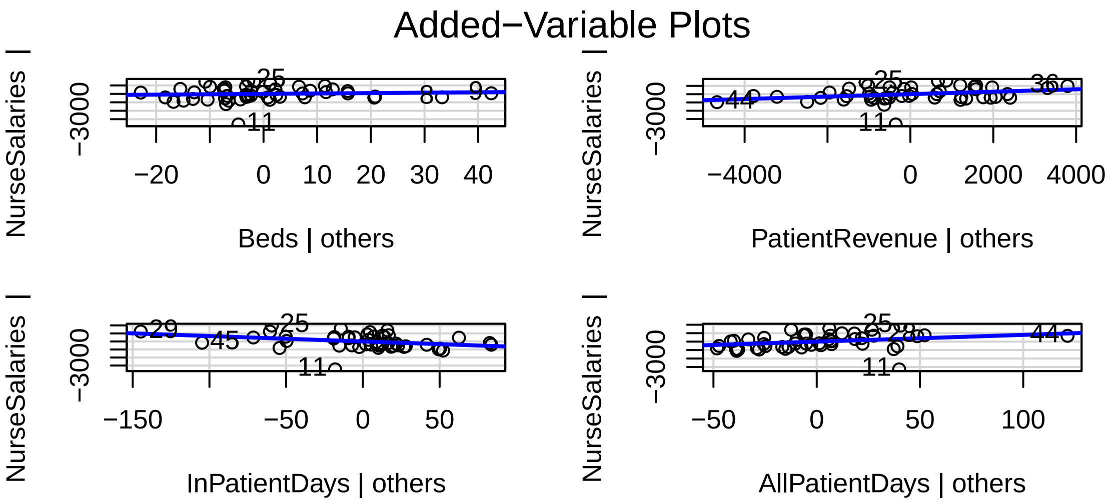
The p-value for the test is 0.65. The data offer no evidence against the null hypothesis of no association between the number of beds and nursing salaries after accounting for the effects of patient revenue, in-patient days, and all patient days.

15. Based on the version of the data set without suspicious observations, create a scatter plot of NurseSalaries vs. Beds. Also create added variables plots based on the model fit from part 13 with 4 explanatory variables. How do these plots relate to what we saw in parts 13 and 14?

```
ggplot(data = nursing, mapping = aes(x = Beds, y = NurseSalaries)) +
  geom_point()
```



```
avPlots(fit_several)
```



The scatter plot of beds vs. nurse salaries shows an association between these variables. This is the relationship estimated by the model that includes only beds as an explanatory variable.

The added variables plot for beds in the model with 4 explanatory variables shows only a weak relationship between beds and nurse salaries after the other explanatory variables have been accounted for. In this added variables plot, the horizontal axis is the residual from a regression where beds is the response variable and PatientRevenue, InPatientDays, and AllPatientDays are the explanatory variables. Essentially, this represents the new information contributed by beds after the other explanatory variables have been accounted for. The vertical axis in the added variables plot is the residual from a linear regression model where the response is nurse salaries and the explanatory variables are patient revenue, in-patient days, and all patient days. The residuals from this model represent what's "left over" in nursing salaries after accounting for patient revenue, in-patient days, and all patient days. The fact that there is not a strong relationship between these residuals in the added variables plot indicates that after accounting for patient revenue, inpatient days, and all patient days, there is not a strong relationship between beds and nurse salaries. This is reflected in the coefficient estimate and hypothesis test results discussed in parts 13 and 14.