HW1 - Josh Janda

Josh Janda

06 February, 2020

Problem 1

The data set prostate from the faraway library, is from a study on 97 men with who were due to receive a radical prostatectomy. Make a numerical and graphical summary of the data.

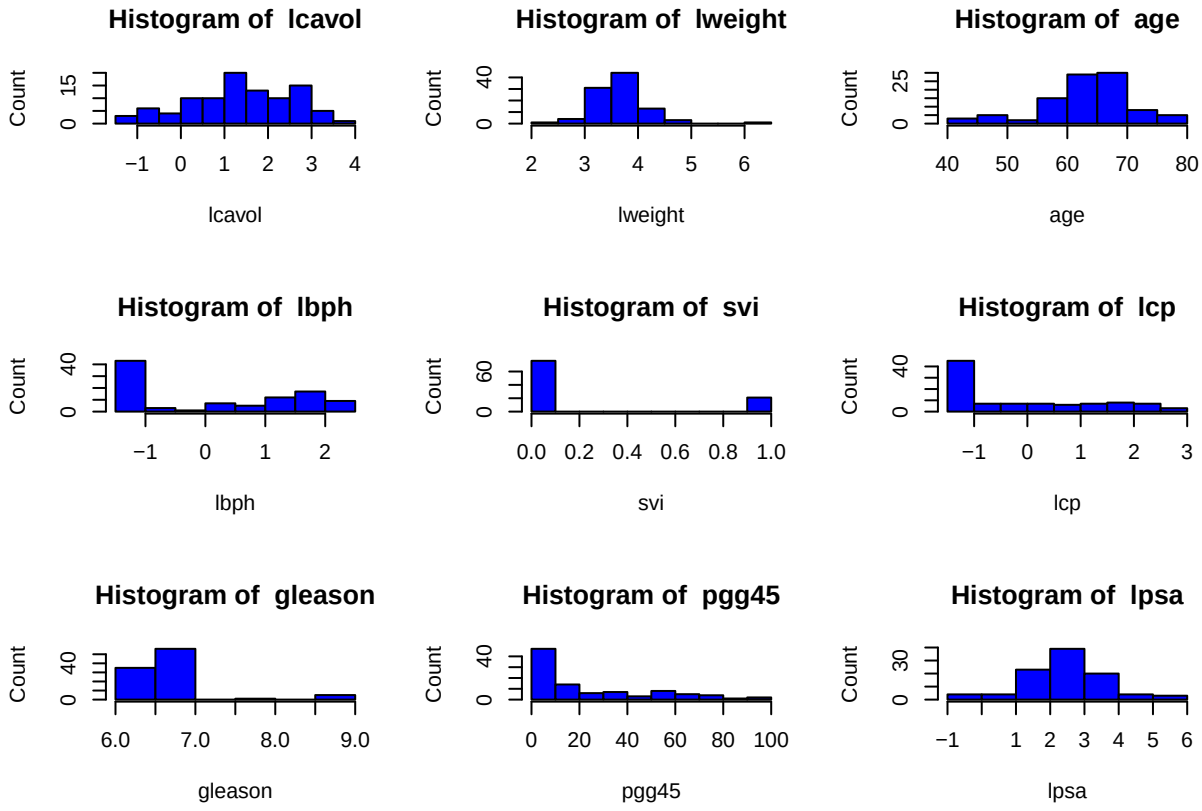
```
#import libraries
library(faraway)
library(tidyverse)
library(GGally)
```

```
prostate_data = prostate

#numerical summary
summary(prostate_data)
```

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
##  Min.   :-0.4308  
##  1st Qu.: 1.7317  
##  Median : 2.5915  
##  Mean   : 2.4784  
##  3rd Qu.: 3.0564  
##  Max.   : 5.5829
```

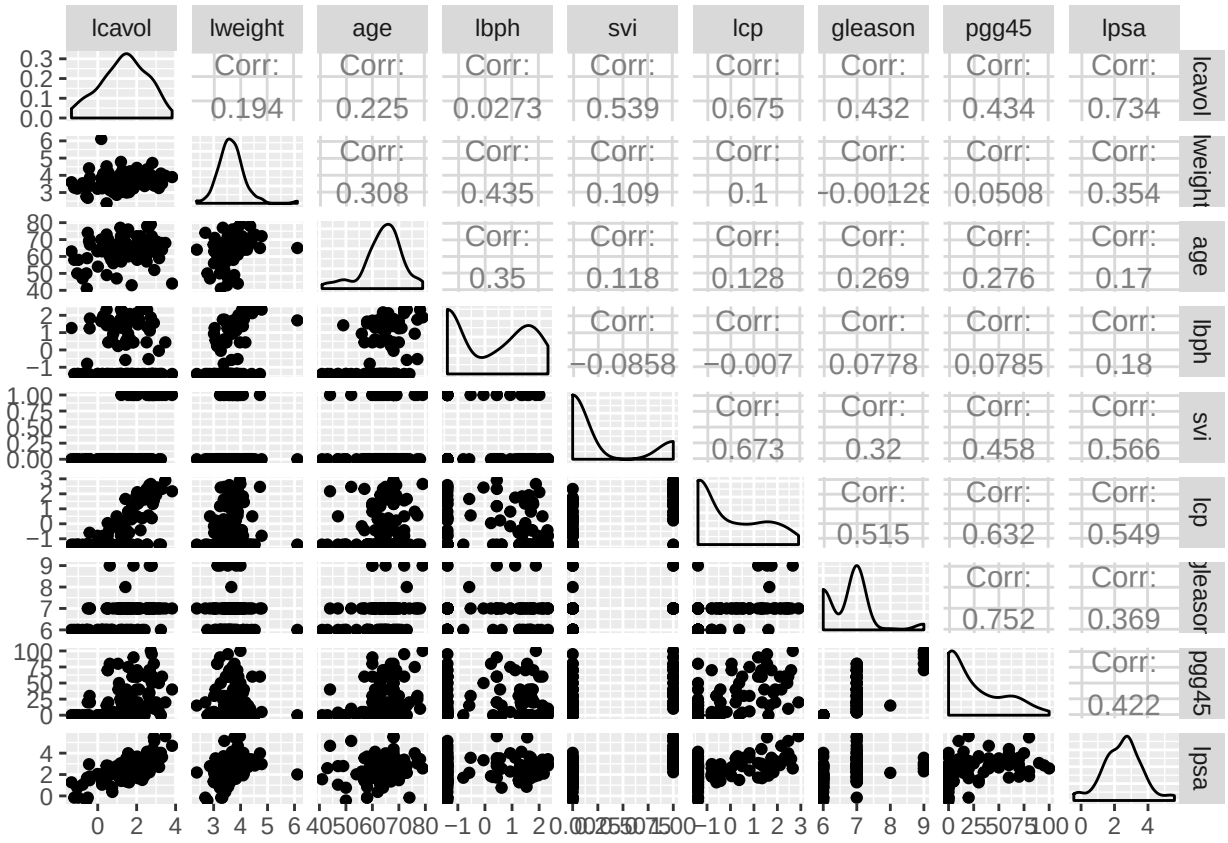
The data above provides the summary statistics of the `prostate` dataset provided by faraway. We are able to see the mean and the 0th, 25th, 50th, 75th, and 100th percentile for each column. With the use of summary statistics, we are able to get a better understanding of the data we are working with and any potential outliers. For the log variables, we are able to see that minimum values are negative which means their true minimum values are less than one. Looking at age, we can see the age range for this dataset was between 41 and 79 years old. This makes sense as males become more susceptible to prostate cancer at around age 40. One potential outlier can be seen in the `pgg45` variable, which has a min of 0 and a max of 100. The mean and median for this variable are 24.38 and 15, respectively. This is indicating that the 100 maximum value is an outlier in the data.

```r
par(mfrow = c(3, 3))

for (i in 1:ncol(prostate_data)){

  data = prostate_data[, i]
  column_name = colnames(prostate_data)[i]
  hist(data, col = "blue",
       xlab = paste(column_name), ylab = "Count",
       main = paste("Histogram of ", column_name))

}
```



The plot grid above displays the histogram of each column. Column names starting with `l` are seen to have a more normally shaped distribution, which is due to the values of the columns being logarithms of the true values. For the other variables, they do not seem to take on any unique distribution. They, however, are mostly skewed to the right due to most of the density being at the minimum point of the data for that variable. Using histograms, we are also able to see outliers in the data by looking at skewedness.

```r
ggpairs(prostate_data, progress = FALSE)
```



Looking at the plot matrix above, we are actually able to see scatter plots, density plots, and correlation between each of the variable interactions. The variables with a linear relationship can be seen to have a higher correlation coefficient. An example of variables with a strong linear relationship are `pgg45` and `gleason`, with a correlation of ~.752. An example of variables with almost no linear relationship are `gleason` and `lweight`, with a correlation coefficient of ~-.0013. Obviously, with the scatter matrix, we are able to gain a great deal of information regarding linear relationships between each feature in the dataset. We are also able to achieve the same style of plot created above, through the use of density plots on the diagonal.

Problem 2

a. Show that for the SLR model, the coefficient of determination R^2 is equal to the square of the correlation coefficient r_{xy}^2

This can be shown through demonstration of modeling in R. I will create a simple regression of `lcavol` on `age`. Referencing the previous plot above, the scatter matrix, the correlation between these variables is .225. This should be the same value obtained for R^2 on the regression model.

```
prob_2_model = lm(lcavol ~ age, data = prostate_data)
corr_lcavol_age = cor(prostate_data$lcavol, prostate_data$age)
r2_lcavol_age = 1 - sum(prob_2_model$residuals^2) / sum((prostate_data$lcavol
                                                         - mean(prostate_data$lcavol))^2)
```

The R^2 value obtained by this model is 0.0506249.

The calculated R^2 model using the formula $R^2 = 1 - \frac{RSS}{TSS}$ is 0.0506249. This is the same output as above.

The correlation between `lcavol` and `age` is 0.2249999. Since this number is equal to r_{xy}, if you take the square of it you will get the same value as the R^2 of this model since it is a simple linear regression.

Doing so, we get a value of r_{xy}^2 being $.225^2 \approx 0.0506249 =$ `r2_lcavol_age`.

This proves that for a simple linear regression with an intercept term we get the same value of R^2 and r_{xy}^2

Problem 3

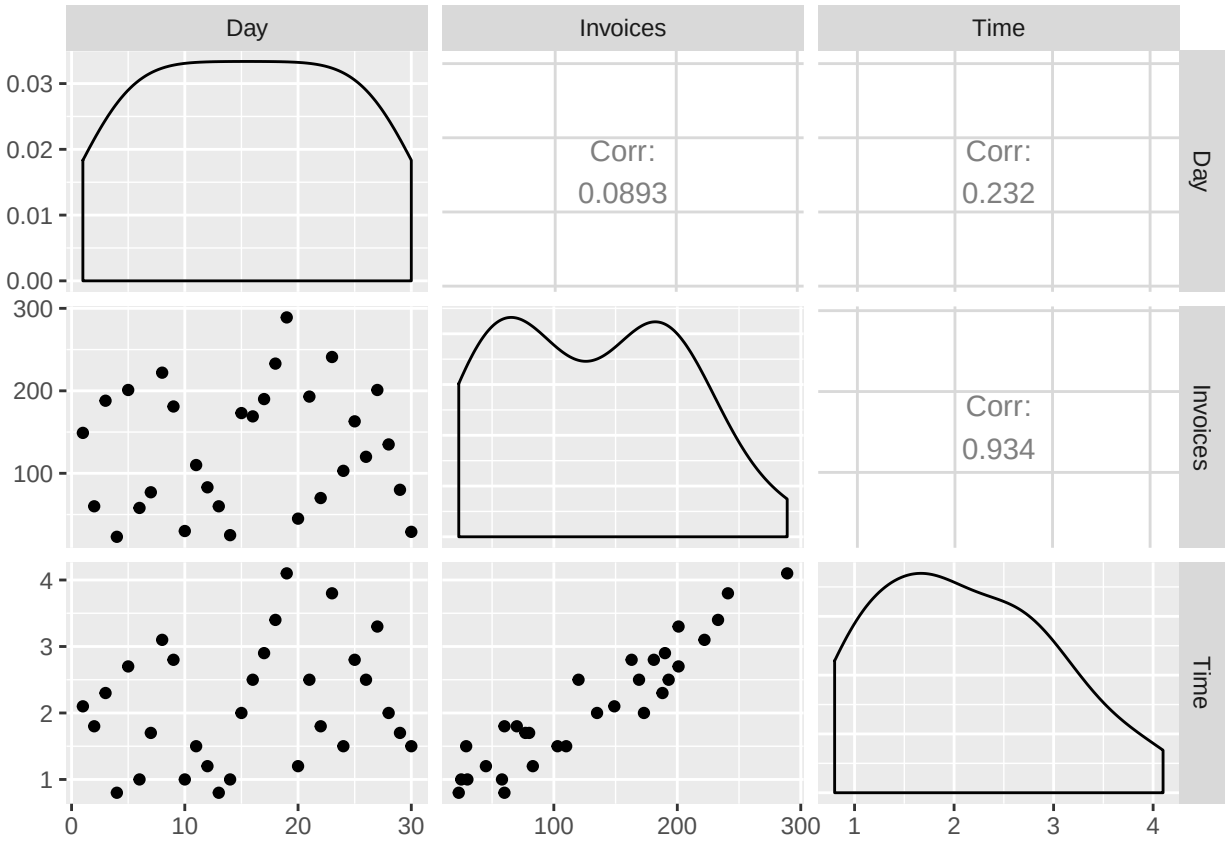
The manager of the purchasing department of a large company would like to develop a regression model to predict the average amount of time it takes to process a given number of invoices. Over a 30-day period, data are collected on the number of invoices processed and the total time taken (in hours). The data are available in the file invoices.txt. The following model was fit to the data: $Y = \beta_0 + \beta_1 x + e$, where Y is the processing time and x is the number of invoices.

a. Plot the data and comment on the results

```
invoices = read_tsv("invoices.txt", col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   Day = col_double(),
##   Invoices = col_double(),
##   Time = col_double()
## )
```

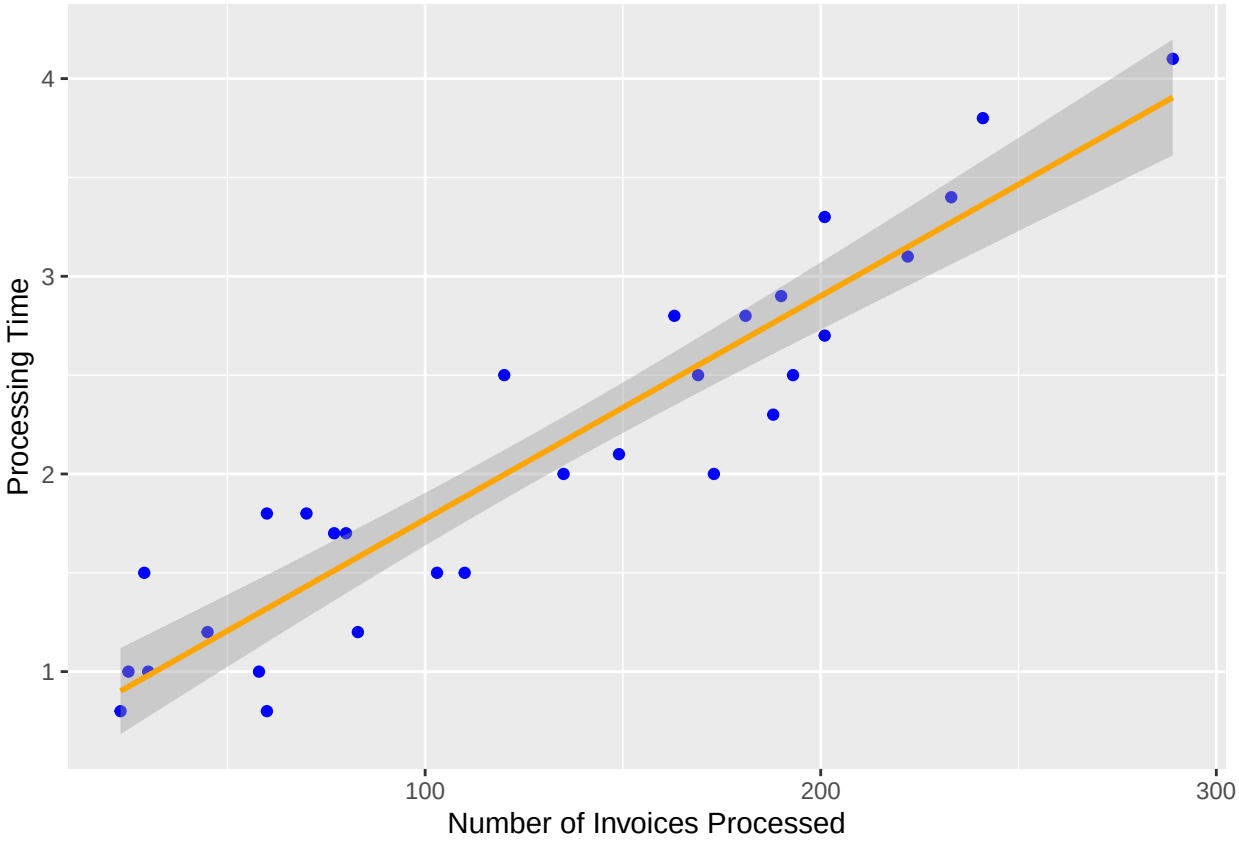
```
ggpairs(invoices, progress = FALSE)
```



Looking at the scatter matrix above, I can see the scatter plots between each variable, the density plots of each variable, and lastly the correlation between each variable. The variables with the strongest linear relationship are `time` and `invoices` with a correlation of *.934*. The density plots show us similar information to a histogram, although it is assuming the data is continuous over the interval. For the `time` variable, the data is right skewed. This means that the time it takes for invoices to be processed is more frequently shorter than longer. For the number of invoices, the data is also right skewed with a slight dip in frequency in the middle.

Overall, with the scatter matrix I am able to see the two variables with the strongest linear relationship as well as understand the distribution of each variable.

```
ggplot(data = invoices, aes(x = Invoices, y = Time)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", formula = y ~ x, color = "orange") +
  labs(x = "Number of Invoices Processed", y = "Processing Time")
```



Looking at the scatter plot above, I focus on the relationship between `Invoices` and `Time`. I have included a regression line which is fit to the following formula: $Time = \beta_0 + \beta_1 Invoices + e$. Overall, this plot demonstrates the strong linear relationship between these two variables by the fitting regression line going through the data points without any obvious outliers (in fact, most of the data stays within the 95% confidence interval provided).

b. Find a 95% confidence interval for the start-up time, β_0

```
prob_3_model = lm(Time ~ Invoices, data = invoices)
b0_95_ci = confint(prob_3_model, "(Intercept)", level = .95)
```

The estimated value for β_0 ($\hat{\beta}_0$) is 0.6417099

Using the formula to find a 95% confidence interval for β_0 ($\hat{\beta}_0 \pm t_{.025,28} * SE_{\hat{\beta}_0}$) we get the following values:

- 95% Lower bound for $\hat{\beta}_0$: 0.3912496
- 95% Upper bound for $\hat{\beta}_0$: 0.8921701

c. Suppose that a best practice benchmark for the average processing time for an additional invoice is 0.01 hours (pr. 0.6 minutes). Test the null hypothesis $H_0 : \beta_1 = 0.01$ against a two-sided alternative. Interpret your result.

```r
beta_1_hat = coef(prob_3_model)[2]
se_beta_1_hat = summary(prob_3_model)$coefficients[2, 2]
```

First, let's setup the null and alternative hypothesis.

$$H_0 : \beta_1 = .01 \;\; H_1 : \beta_1 \neq .01$$

The obtained value for $\hat{\beta}_1$ is 0.0112916. For testing this claim, I will be using values of $\alpha = .01, .05, .10$.

To find the t-score for testing our claim, we use the following formula:

$$t = \frac{\hat{\beta}_1 - .01}{SE_{\hat{\beta}_1}}$$

Using that formula, we can directly solve for t.

$$t = \frac{0.0112916 - .01}{8.1840203 \times 10^{-4}}$$

```r
beta_1_t_score = (beta_1_hat - .01 ) / se_beta_1_hat
prob_t = pt(beta_1_t_score, df = 28, lower.tail = FALSE)
```

Once t is solved for, we obtain a value of 1.5782513.

Now, we can solve for the probability of this t score using the t distribution CDF. Once the probability is solved for, we must multiply by two since it is a two sided test.

$$P(t \geq 1.5782513) = 2 * 0.0628701 = 0.1257402$$

So, we have a p-value of 0.1257402. I can reject the null hypothesis that $\beta_1 = .01$ at $\alpha = .01,\ .05,\ and\ .10$.

d. Find a point estimate and a 95% prediction interval for the time taken to process 130 invoices.

```r
pt_est_130_invoices = predict(prob_3_model,
                              newdata = data.frame(Invoices = 130),
                              interval = "none")

invoices_130_95_pred_interval = predict(prob_3_model,
                                        newdata = data.frame(Invoices = 130),
                                        interval = "prediction",
                                        level = .95)
```

The formula used to find a point estimate of the time taken to process 130 invoices is: $\hat{Time} = \hat{\beta}_0 + \hat{\beta}_1 * 130$

The point estimate for the time taken to process 130 invoices is 2.1096236.

The formula used to find a prediction interval is similar to a confidence interval. However, it provides a wider prediction due to the interval being confident on a single observation, rather than a mean observation. (95% on average it will take x time to do 130 invoices (confidence) vs 95% it will take x time to do these 130 invoices (prediction))

The formula for the prediction interval is:

$$\hat{Time} \pm t_{.025,28} * \hat{\sigma} * \sqrt{1 + \frac{1}{30} + \frac{130 - \bar{x}}{S_{xx}}}$$

Where $\hat{\sigma} = \sqrt{\frac{RSS}{28}}$ and S_{xx} is the standard deviation of x (invoices).

With that said, here is the 95% prediction interval for processing time of 130 invoices.

The 95% lower bound for this interval is: 1.4229473

The 95% upper bound for this interval is: 2.7962999

Problem 5

Fit the following model to the data: $PriceChange = \beta_0 + \beta_1 LoanPaymentsOverdue + e$.

a. Calculate the R2 and adjusted R2 for the SLR model. Provide an interpretation of both

quantities.

```
indicators = read_csv("indicators.csv")
```

```
## Parsed with column specification:
## cols(
##   MetroArea = col_character(),
##   PriceChange = col_double(),
##   LoanPaymentsOverdue = col_double()
## )
```

```
indicators_model = lm(PriceChange ~ LoanPaymentsOverdue, data = indicators)

indicators_r2 = summary(indicators_model)$r.sq
indicators_adj_r2 = summary(indicators_model)$adj.r.sq
```

The R^2 for this model is 0.2791527. This means that ~27.9% of the variance in the *PriceChange* variable is explained by *LoanPaymentsOverdue*.

The $Adj.R^2$ for this model is 0.2340997. Note that it is lower than the value obtained for R^2. While similar to R^2, $Adj.R^2$ adjusts for the number of variables being fit in the model. If you include more features in the model, $Adj.R^2$ will put a penalty on the model for that. If the added variable adds greater performance to the model, $Adj.R^2$ will increase. If the added variable does not improve the model, $Adj.R^2$ will decrease. In this case, this value of $Adj.R^2$ shows that this one feature improves our model greater than chance (~23.4% greater, since without any features we should have an $Adj.R^2$ of zero).

b. Find a 95% confidence interval for the slope of the regression model, β_1. On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.

```
indicators_slope_95_ci = confint(indicators_model, "LoanPaymentsOverdue", level = .95)
```

The 95% lower bound for the slope of this model is -4.1634543

The 95% upper bound for the slope of this model is -0.3335853

With these values both being negative, as well as being 95% confident the true value of β_1 is contained in this interval, this confidence interval shows evidence of a significant negative linear association at a significance level of $\alpha = .05$

c. Use the fitted regression model to estimate E(Y|X = 4). Find a 95% confidence interval for E(Y|X = 4). Is 0% a feasible value for E(Y|X = 4)? Give a reason to support your answer.

```
expected_y_given_x = predict(indicators_model,
                             newdata = data.frame(LoanPaymentsOverdue = 4),
                             interval = "confidence", level = .95)
```

Using the regression model to estimate E(Y|X = 4) is the same as computing the point estimate when X is equal to 4. The confidence interval, as mentioned in the previous problem, is saying that we are 95% confidenct that when X is 4 we will on average get that associated value of Y. In a prediction interval, we are focused on that specific observation and it's associated estimate and prediction interval.

For this question, E(Y|X = 4) = -4.4795854

The 95% lower bound for E(Y|X = 4) is -6.6488492

The 95% upper bound for E(Y|X = 4) is -2.3103215

With that estimated mean value and associated 95% lower and upper bounds, 0% is not a feasible value for E(Y|X = 4) because it is not contained within the 95% confidence interval. Since 0 is not in the interval, we can reject the hypothesis that E(Y|X = 4) = 0 at $\alpha = .05$.