Homework 2 Solutions

1. Load data

```r
setwd("/Users/Daniel/Dropbox/Teaching/CourseR/data/")
d <- read.csv("CollegeScorecard.csv", na = c("NULL", "PrivacySuppressed"))
```

2. Process data

```r
d <- d[ ,c("INSTNM", "STABBR", "SAT_AVG_ALL", "GRAD_DEBT_MDN_SUPP",
           "md_earn_wne_p10")]
names(d) <- c("Institution", "State", "SAT", "Debt", "Earnings")

d$SAT_c <- d$SAT - mean(d$SAT, na.rm = TRUE)
mean(d$SAT_c, na.rm = TRUE)
```

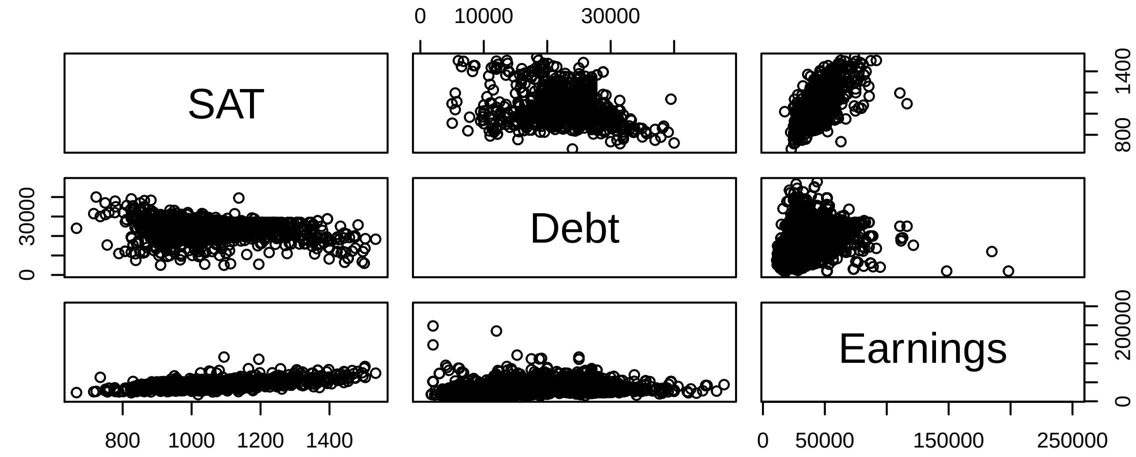
```
## [1] -9e-15
```

```r
d$Debt_c <- d$Debt - mean(d$Debt, na.rm = TRUE)
mean(d$Debt_c, na.rm = TRUE)
```

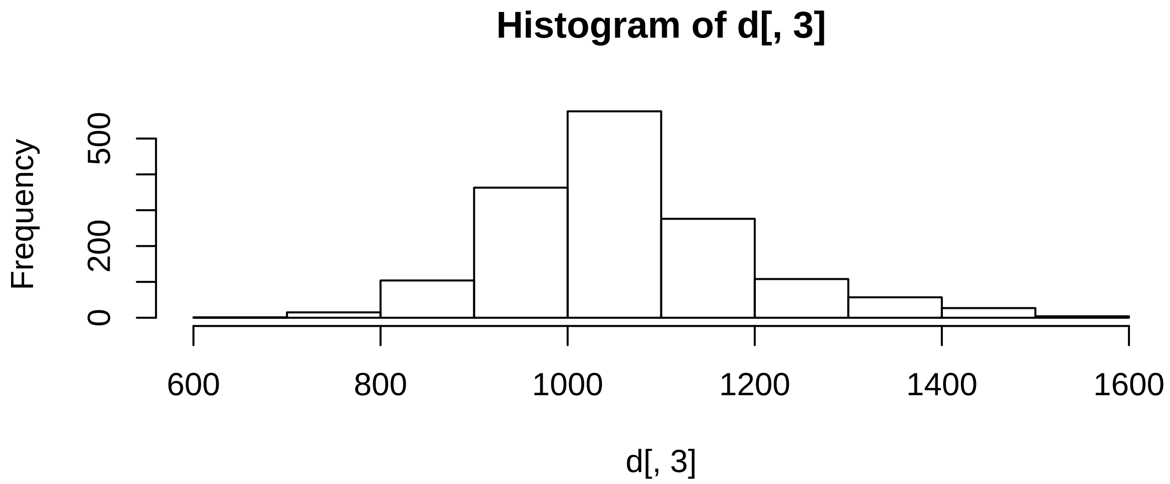
```
## [1] -1.7e-12
```

3. Create plots

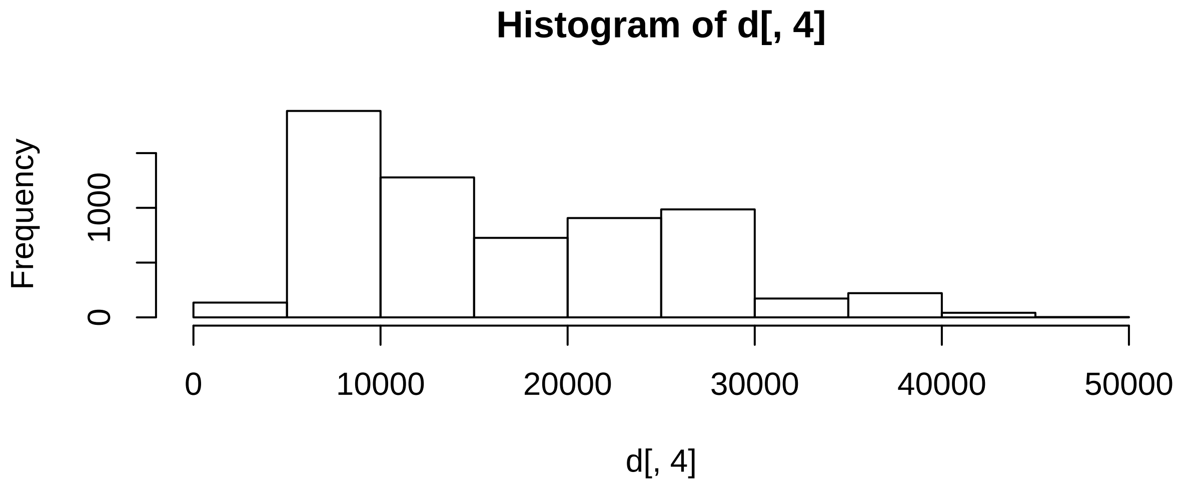
```r
pairs(d[ ,3:5])
```



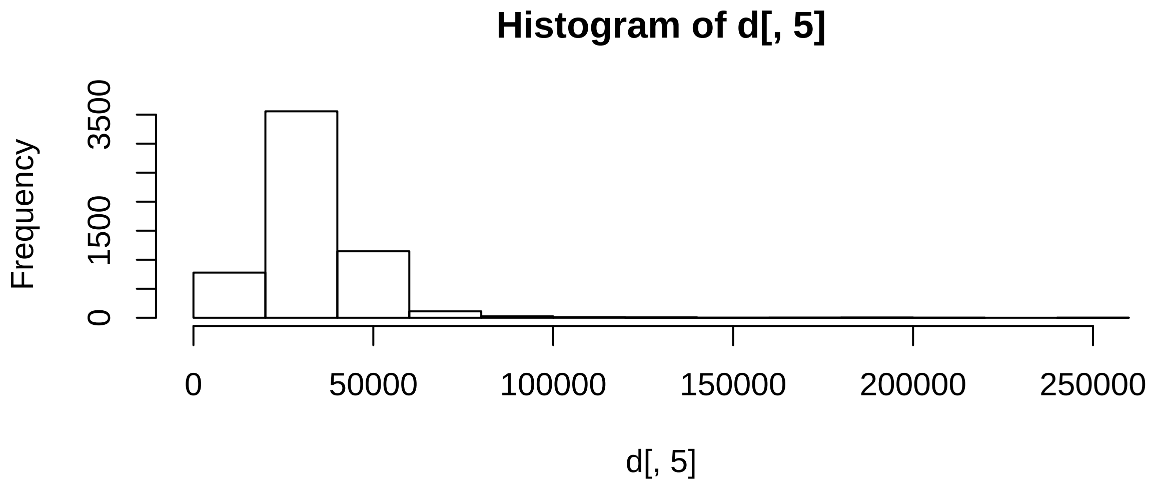
```r
hist(d[ ,3])
```



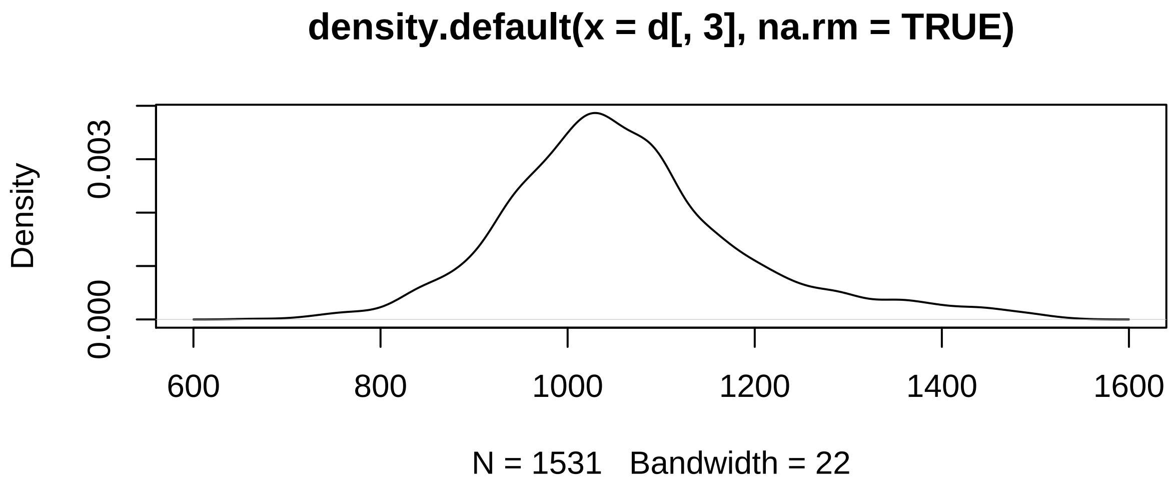
```
hist(d[ ,4])
```



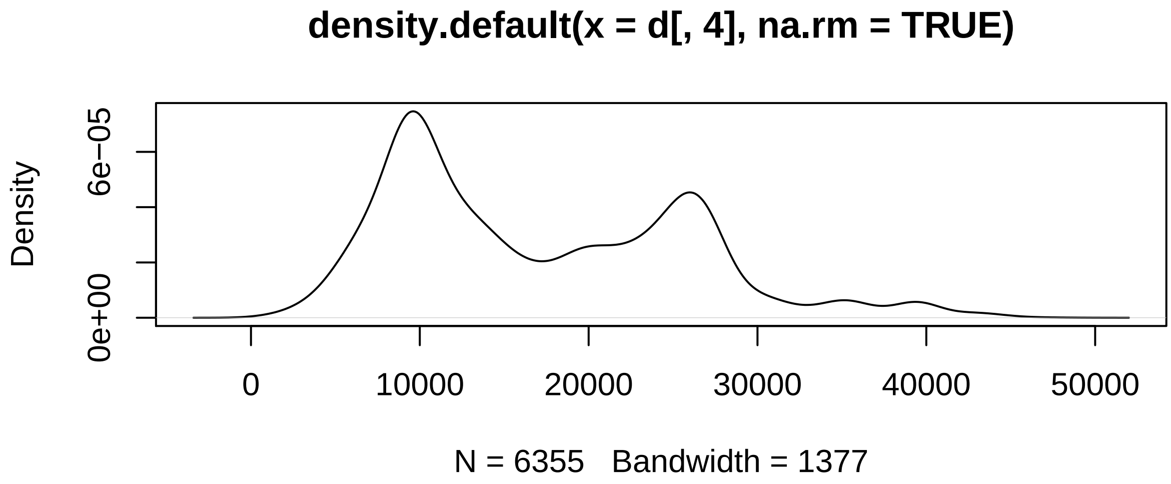
```
hist(d[ ,5])
```



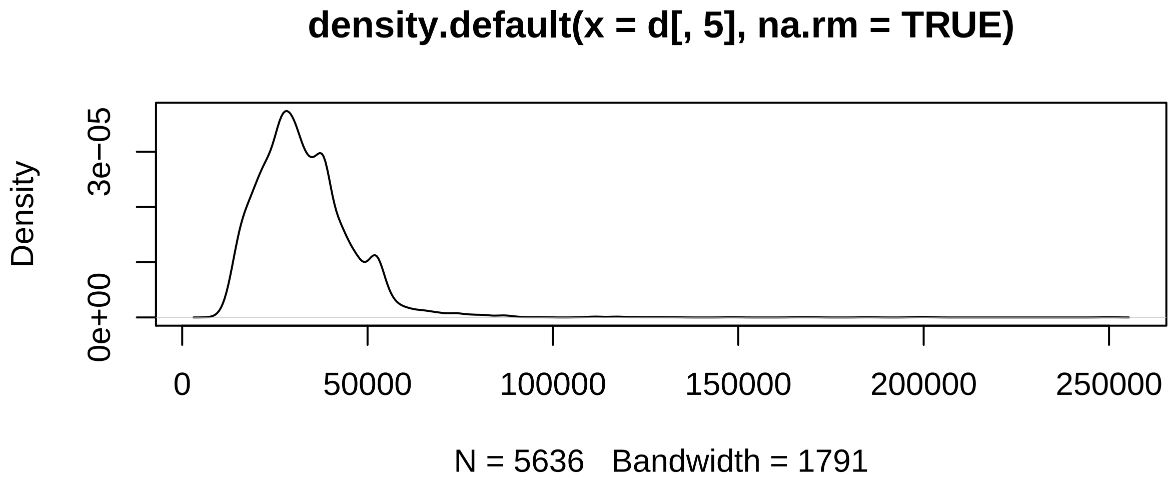
```
plot(density(d[ ,3], na.rm = TRUE))
```



```
plot(density(d[ ,4], na.rm = TRUE))
```



```
plot(density(d[ ,5], na.rm = TRUE))
```



4. Fit preliminary models

```
mA <- lm(Earnings ~ SAT, data = d)
summary(mA)
```

```
##
## Call:
## lm(formula = Earnings ~ SAT, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23195  -4932   -902   3554  71927
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9181.28    1667.70   -5.51  4.3e-08 ***
## SAT            49.04       1.56   31.42  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7790 on 1473 degrees of freedom
##   (6329 observations deleted due to missingness)
## Multiple R-squared:  0.401,  Adjusted R-squared:  0.401
## F-statistic:  987 on 1 and 1473 DF,  p-value: <2e-16
```

```
mB <- lm(Earnings ~ Debt, data = d)
summary(mB)
```

```
##
## Call:
## lm(formula = Earnings ~ Debt, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29936  -7248   -863   4922 175310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.16e+04   3.61e+02    59.9   <2e-16 ***
## Debt        6.83e-01   1.81e-02    37.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11200 on 5050 degrees of freedom
##   (2752 observations deleted due to missingness)
## Multiple R-squared:  0.22,   Adjusted R-squared:  0.22
## F-statistic: 1.42e+03 on 1 and 5050 DF,  p-value: <2e-16
```

```
mC <- lm(SAT ~ Debt, data = d)
summary(mC)
```

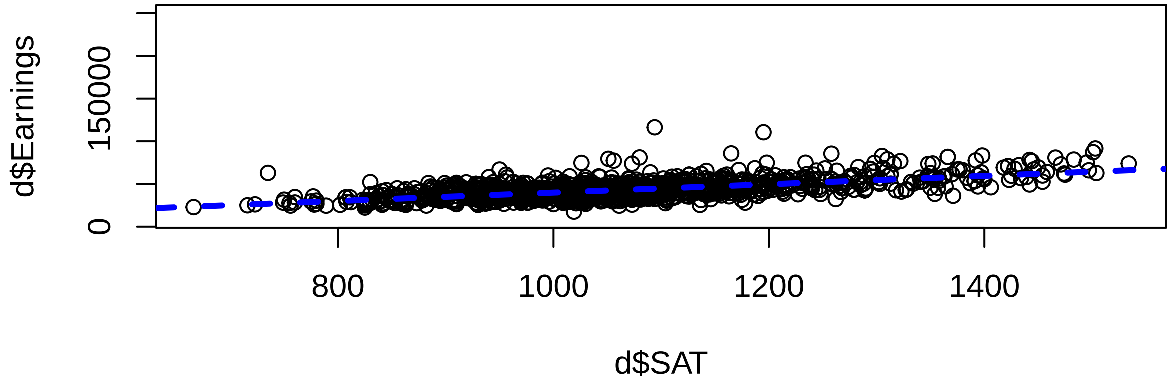
```
##
```

```
## Call:
## lm(formula = SAT ~ Debt, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -395.0  -85.8  -10.8   65.7  441.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.20e+03   1.70e+01   70.62  < 2e-16 ***
## Debt        -5.69e-03   7.06e-04   -8.06  1.6e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 127 on 1500 degrees of freedom
##   (6302 observations deleted due to missingness)
## Multiple R-squared:  0.0415, Adjusted R-squared:  0.0409
## F-statistic: 64.9 on 1 and 1500 DF,  p-value: 1.56e-15
```

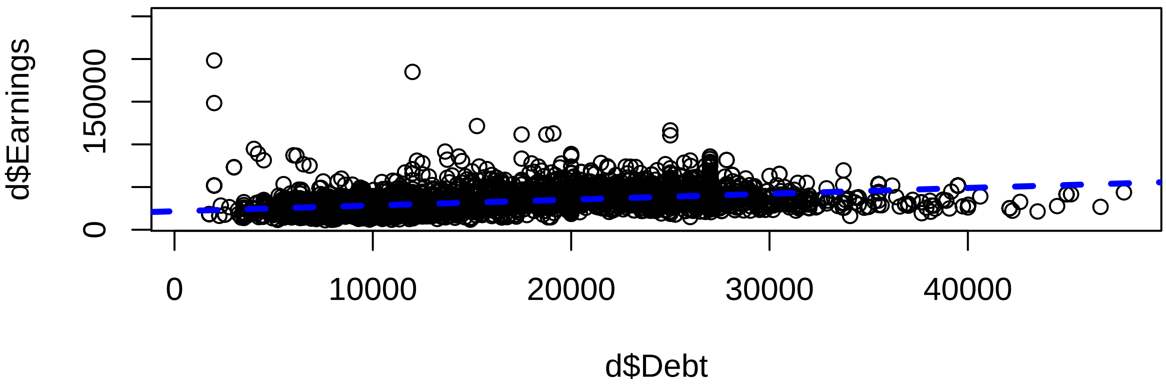
5. Plot regression models

The results of the coefficents are pretty straightforward to interpret. They represent the expected change in the outcome given a 1 unit increase in the predictor. Note that the coefficient for `Debt` is very small, because it represents the expected change in the outcome given a one dollar increase in debt. The intercepts range from difficult to interpret, to impossible. For example, the SAT scale does not go below 200, yet the intercept in Model A represents the expected earnings for schools with an average SAT score of 0. After centering the variables and refitting the models (code below), the intercepts represent the expected value of the outcome when the school has an average level of the predictor.

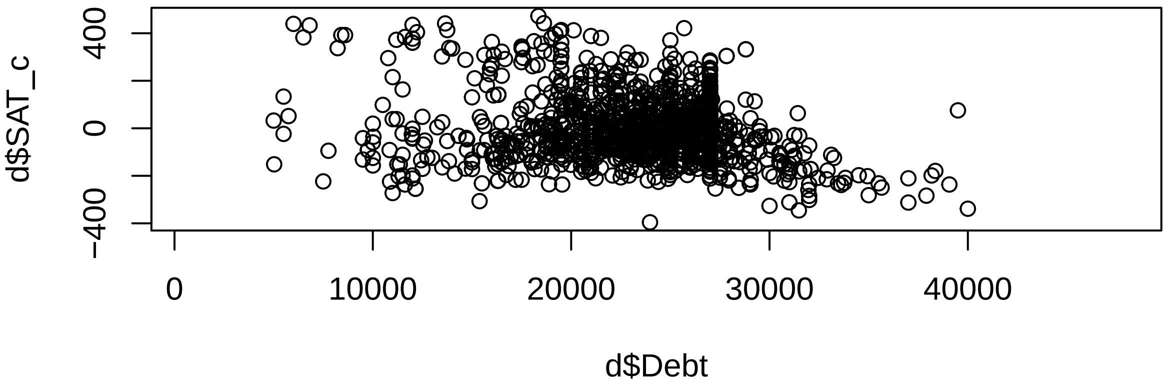
```r
plot(d$Earnings ~ d$SAT)
abline(coef(mA)[1], coef(mA)[2], col = "blue", lwd = 3, lty = 2)
```



```
plot(d$Earnings ~ d$Debt)
abline(coef(mB)[1], coef(mB)[2], col = "blue", lwd = 3, lty = 2)
```



```
plot(d$SAT_c ~ d$Debt)
abline(coef(mC)[1], coef(mC)[2], col = "blue", lwd = 3, lty = 2)
```



```
mA <- lm(Earnings ~ SAT_c, data = d)
summary(mA)
```

```
##
## Call:
## lm(formula = Earnings ~ SAT_c, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23195  -4932   -902   3554  71927
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42885.13     202.88   211.4   <2e-16 ***
## SAT_c          49.04       1.56    31.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7790 on 1473 degrees of freedom
##   (6329 observations deleted due to missingness)
## Multiple R-squared:  0.401,  Adjusted R-squared:  0.401
## F-statistic:  987 on 1 and 1473 DF,  p-value: <2e-16
```

```
mB <- lm(Earnings ~ Debt_c, data = d)
summary(mB)
```

```
##
## Call:
## lm(formula = Earnings ~ Debt_c, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29936  -7248   -863   4922 175310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.33e+04   1.58e+02   210.2   <2e-16 ***
## Debt_c      6.83e-01   1.81e-02    37.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11200 on 5050 degrees of freedom
##   (2752 observations deleted due to missingness)
## Multiple R-squared:  0.22,   Adjusted R-squared:  0.22
## F-statistic: 1.42e+03 on 1 and 5050 DF,  p-value: <2e-16
```

```
mC <- lm(SAT_c ~ Debt_c, data = d)
summary(mC)
```

```
##
## Call:
## lm(formula = SAT_c ~ Debt_c, data = d)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -395.0  -85.8  -10.8   65.7  441.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.771852   5.647319    6.87  9.7e-12 ***
## Debt_c      -0.005687   0.000706   -8.06  1.6e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 127 on 1500 degrees of freedom
##   (6302 observations deleted due to missingness)
## Multiple R-squared:  0.0415, Adjusted R-squared:  0.0409
## F-statistic: 64.9 on 1 and 1500 DF,  p-value: 1.56e-15
```

6. Fit the multiple regression model

```r
mr <- lm(Earnings ~ SAT_c + Debt_c, data = d)
summary(mr)
```

```
##
## Call:
## lm(formula = Earnings ~ SAT_c + Debt_c, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21404  -4796  -1051   3616  71753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.23e+04   3.66e+02  115.71   <2e-16 ***
## SAT_c       4.97e+01   1.61e+00   30.86   <2e-16 ***
## Debt_c      9.16e-02   4.61e-02    1.99    0.047 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7780 on 1452 degrees of freedom
##   (6349 observations deleted due to missingness)
## Multiple R-squared:  0.401,  Adjusted R-squared:  0.401
## F-statistic:  487 on 2 and 1452 DF,  p-value: <2e-16
```

Extra Credit. Predictor-residual plot

To compute the predictor residual plot, first predict SAT scores for each observation from their Debt.

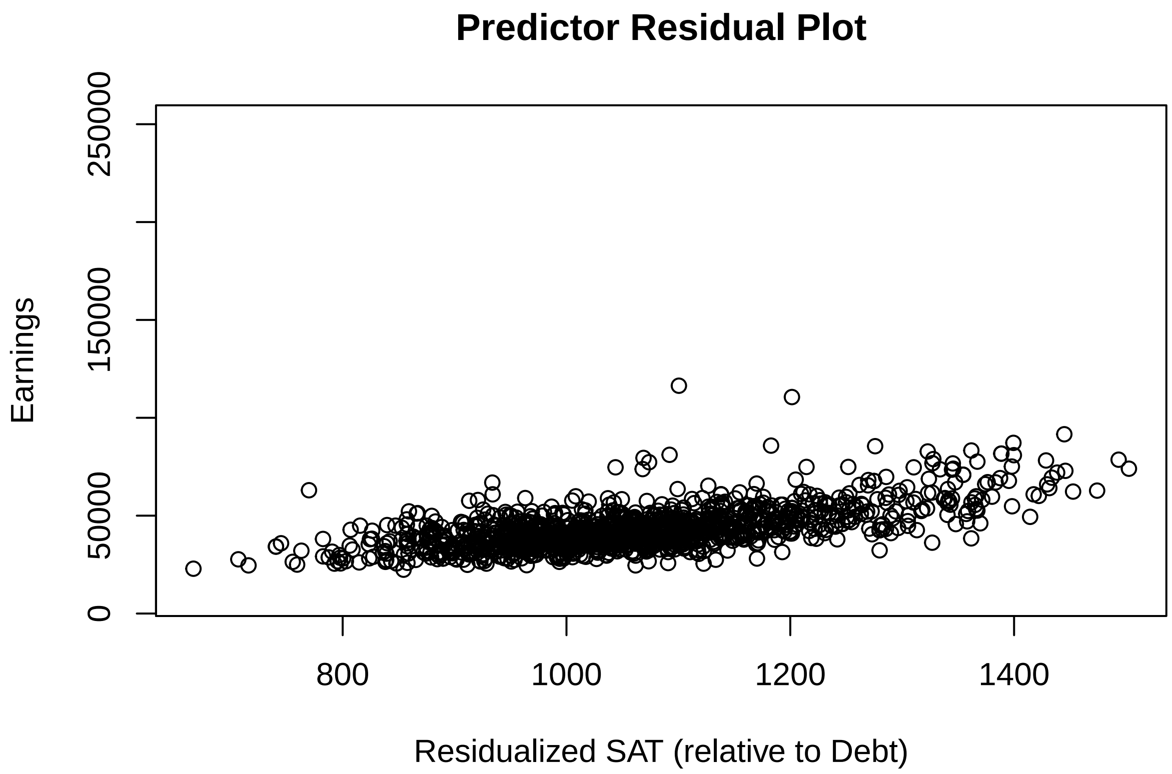
```r
predSAT <- coef(mC)[1] + coef(mC)[2]*d$Debt_c
```

Next compute the residual - i.e., the difference between the predicted and observed SAT score.

```
resSAT <- d$SAT - predSAT
```

Finally, plot the relation between the residualized SAT variable and Earnings. Note that I've added a few additional argument to the plot to put an overall title and label the x and y axes.

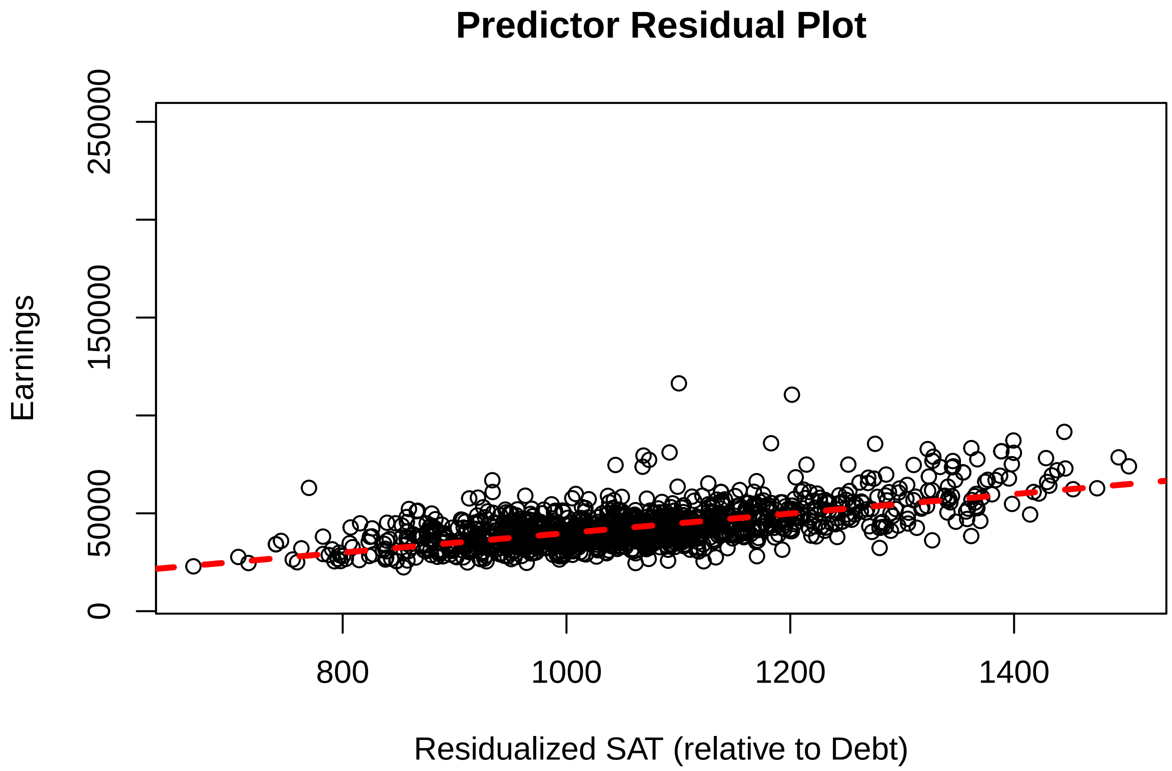
```
plot(resSAT, d$Earnings,
    main = "Predictor Residual Plot",
    xlab = "Residualized SAT (relative to Debt)",
    ylab = "Earnings")
```



The plot above is the relation between `Earnings` and `SAT_c` while controlling for (i.e., residualzing for) `Debt_c`. To actually plot the residualzed line, you'll need to fit one additional model.

```
resLine <- lm(d$Earnings ~ resSAT)
```

Then you can just use `abline()` like normal



```
abline(coef(resLine)[1], coef(resLine)[2], col = "red", lwd = 3, lty = 2)
```