```r
library(tidyverse)
library(car)
library(ggeffects)
```

```r
#importing datasets from tidytuesday

tuition_cost <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da

salary_potential <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste

diversity_school <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste
  select("name", "total_enrollment") %>%
  distinct()
```

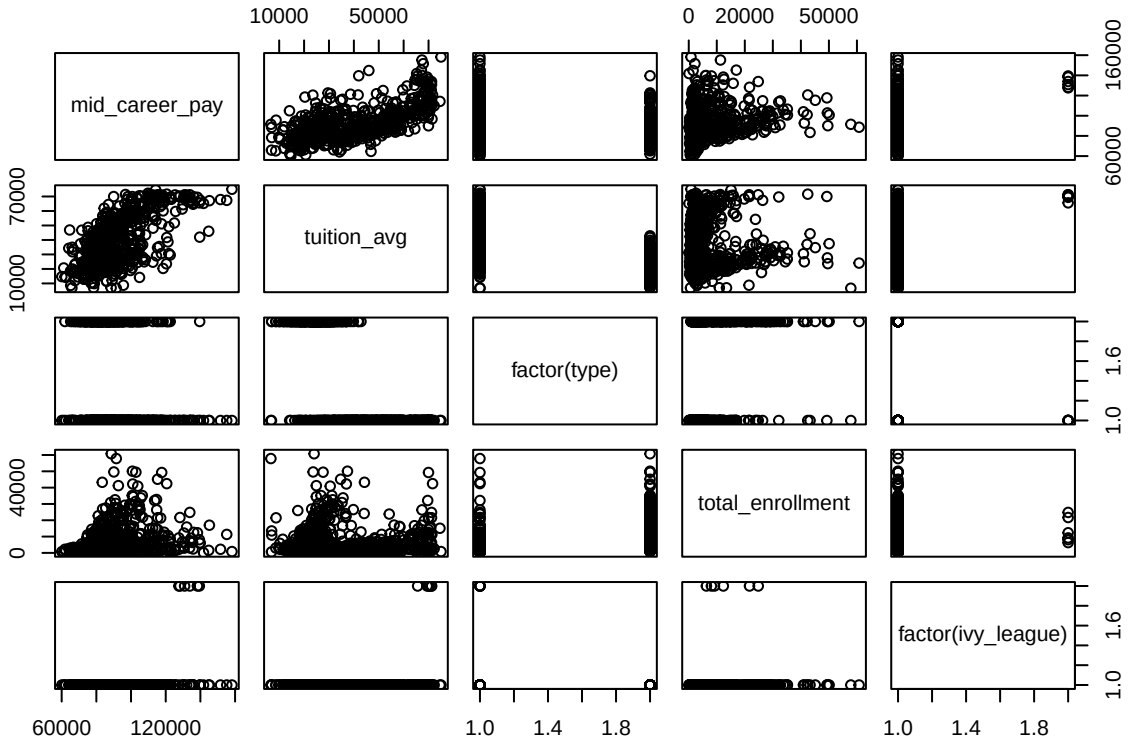
```r
#creating our desired dataset

full_data <- full_join(tuition_cost, salary_potential, by = "name") %>%
  full_join(diversity_school, by = "name") %>%
  #merging the three datasets by school name
  drop_na(name, type, out_of_state_total, mid_career_pay, total_enrollment) %>%
  #removing any schools that do not have available the data we want
  mutate("ivy_league" = ifelse(name %in% c("Brown University", "Cornell University", "Dartmouth College
  #creating an indicator variable to designate whether a school is an ivy league or not
  mutate("tuition_avg" = (in_state_total + out_of_state_total)/2)
#creating a tuition average variable by finding the mean instate and out of state tuitions
```

```r
#writing as a csv to share
write_csv(full_data, file = "college_scoreboard.csv")
```

First Model (MLR)

```r
#matrix plot to check for relationships
pairs(mid_career_pay ~ tuition_avg + factor(type) + total_enrollment + factor(ivy_league), data = full_
```




```
#first model with untransformed data
college1_mlr <- lm(mid_career_pay ~ tuition_avg + type + total_enrollment + ivy_league, data = full_data
summary(college1_mlr)
```

```
##
## Call:
## lm(formula = mid_career_pay ~ tuition_avg + type + total_enrollment +
##     ivy_league, data = full_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27240  -6753  -1996   4106  53358
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.309e+04  1.784e+03  29.761  < 2e-16 ***
## tuition_avg      8.236e-01  3.610e-02  22.817  < 2e-16 ***
## typePublic       1.097e+04  1.362e+03   8.052 4.03e-15 ***
## total_enrollment 2.843e-01  5.177e-02   5.491 5.78e-08 ***
## ivy_league1      1.871e+04  4.454e+03   4.200 3.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10650 on 635 degrees of freedom
## Multiple R-squared:  0.5553, Adjusted R-squared:  0.5525
## F-statistic: 198.2 on 4 and 635 DF,  p-value: < 2.2e-16
```

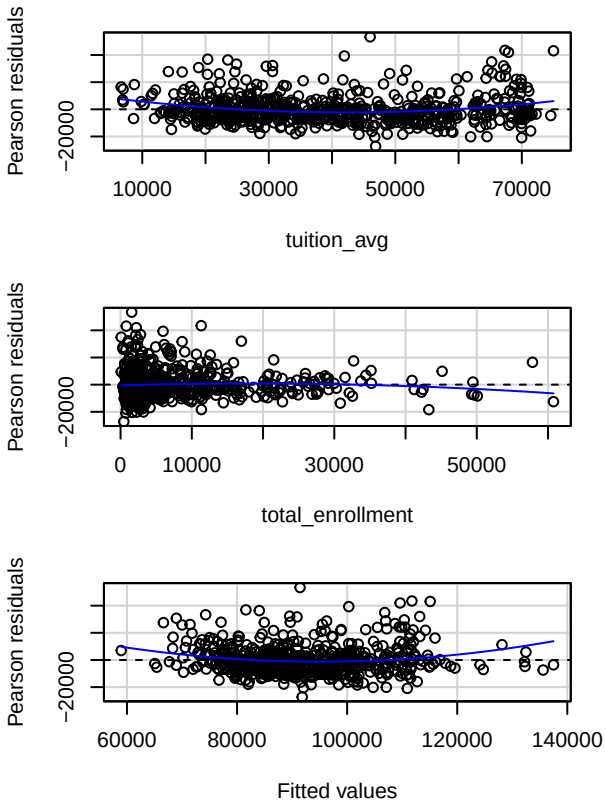
Multiple regression linear model:

$y_i = 53091.45 + 0.824(\text{tuition\_avg}) + 10965.41(\text{typePublic}) + 0.284(\text{total\_enrollment}) + 18706.94(\text{ivy\_league})$

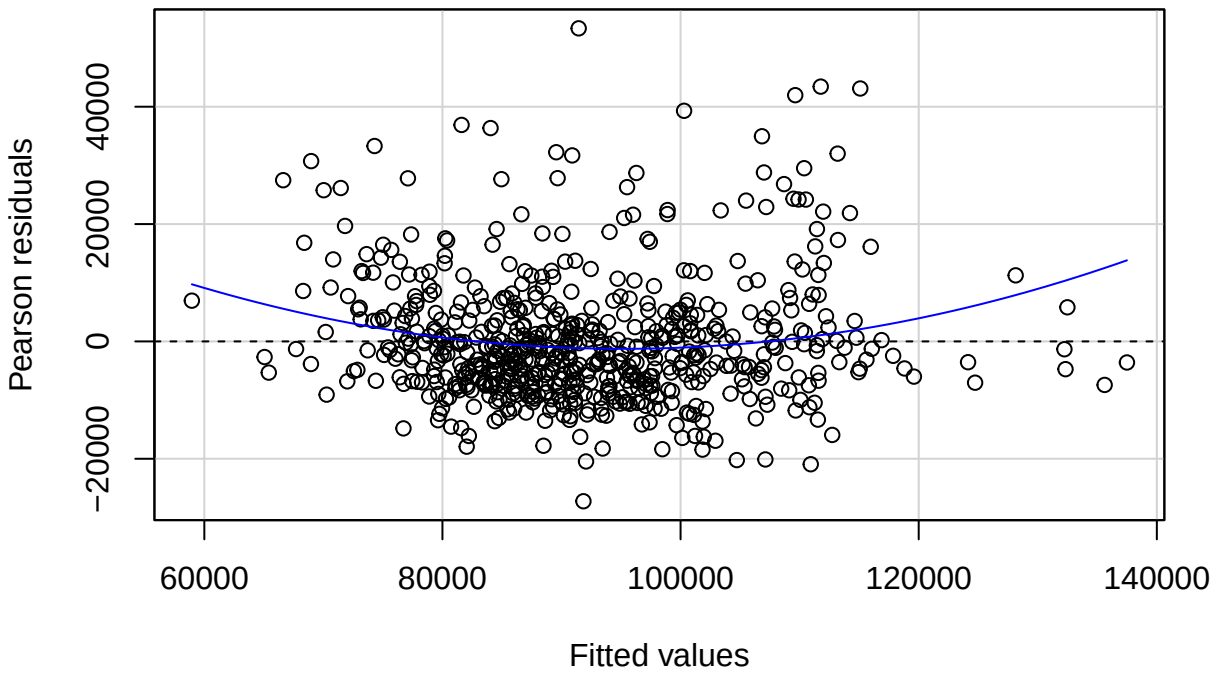
All p-values are extremely low. $R^2 = 0.555$

First Model (Residual Plots)

```
#checking assumptions with residual plots, clear right skew in total_enrollment but other plots look ok
residualPlots(college1_mlr, test = FALSE)
```




```
residualPlot(college1_mlr)
```



The residuals for tuition, type, ivy league, and the fitted model seem evenly distributed, but very far from 0. There is a lot of variation in this model. The enrollment residuals are skewed right.

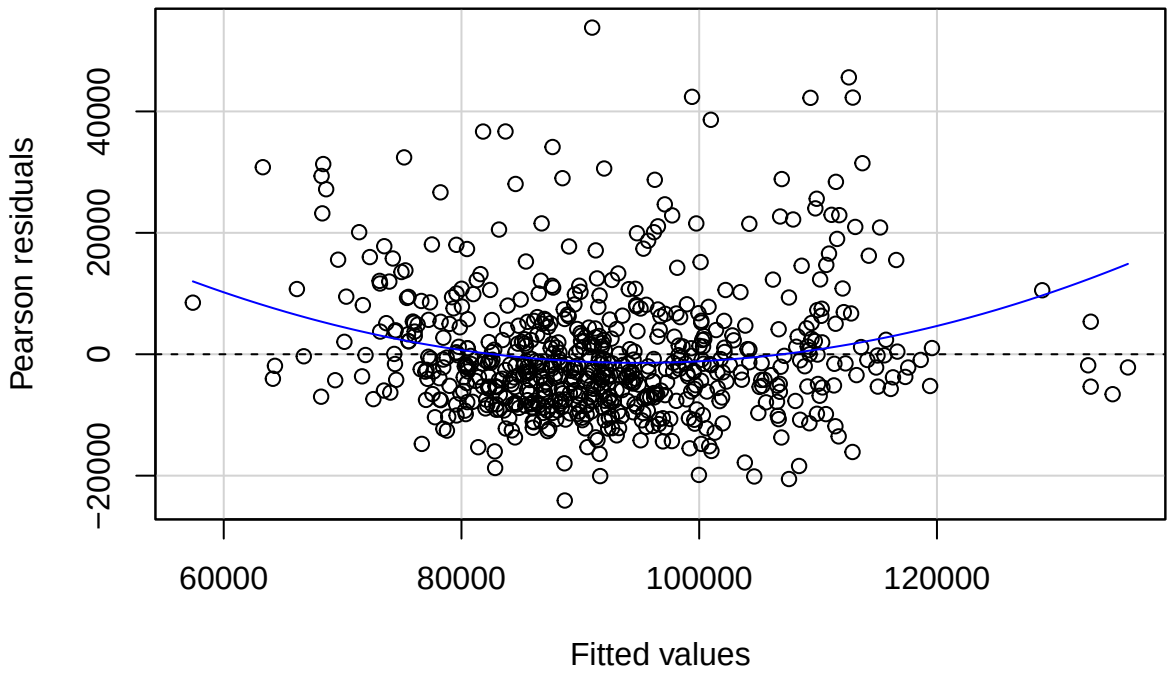
2nd Model (log(enrollment))

```
#fitting second model with natural transformation on total_enrollment
college2_mlr <- lm(mid_career_pay ~ tuition_avg + type + log(total_enrollment) + ivy_league, data = full
summary(college2_mlr)
```

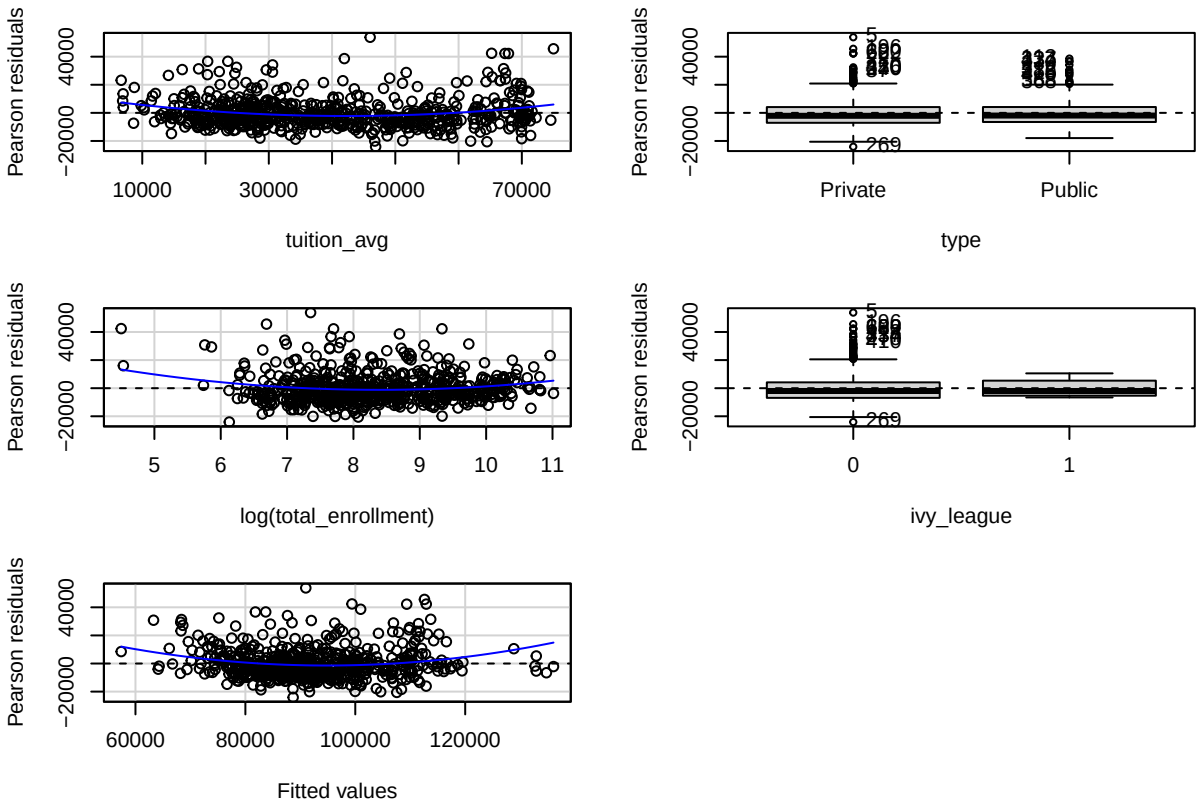
```
##
## Call:
## lm(formula = mid_career_pay ~ tuition_avg + type + log(total_enrollment) +
##     ivy_league, data = full_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24091  -6871  -1989   4156  53805
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.612e+04  3.550e+03  10.177  < 2e-16 ***
## tuition_avg             8.013e-01  3.764e-02  21.288  < 2e-16 ***
## typePublic              9.980e+03  1.484e+03   6.725 3.93e-11 ***
## log(total_enrollment) 2.446e+03  4.725e+02   5.177 3.03e-07 ***
## ivy_league1             1.814e+04  4.475e+03   4.053 5.69e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10680 on 635 degrees of freedom
## Multiple R-squared:  0.553,  Adjusted R-squared:  0.5502
## F-statistic: 196.4 on 4 and 635 DF,  p-value: < 2.2e-16
```

```
#checking assumptions with residual plots, no more skew but residuals are very large
residualPlot(college2_mlr)
```




```
residualPlots(college2_mlr)
```




```
##                         Test stat Pr(>|Test stat|)
## tuition_avg                5.2312        2.292e-07 ***
## type
## log(total_enrollment)      3.5327        0.0004412 ***
## ivy_league
## Tukey test                 6.7807        1.196e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$y_i = 36120 + 0.801(\text{tuition\_avg}) + 9980(\text{typePublic}) + 2446(\log(\text{total\_enrollment})) + 18140(\text{ivy\_league})$

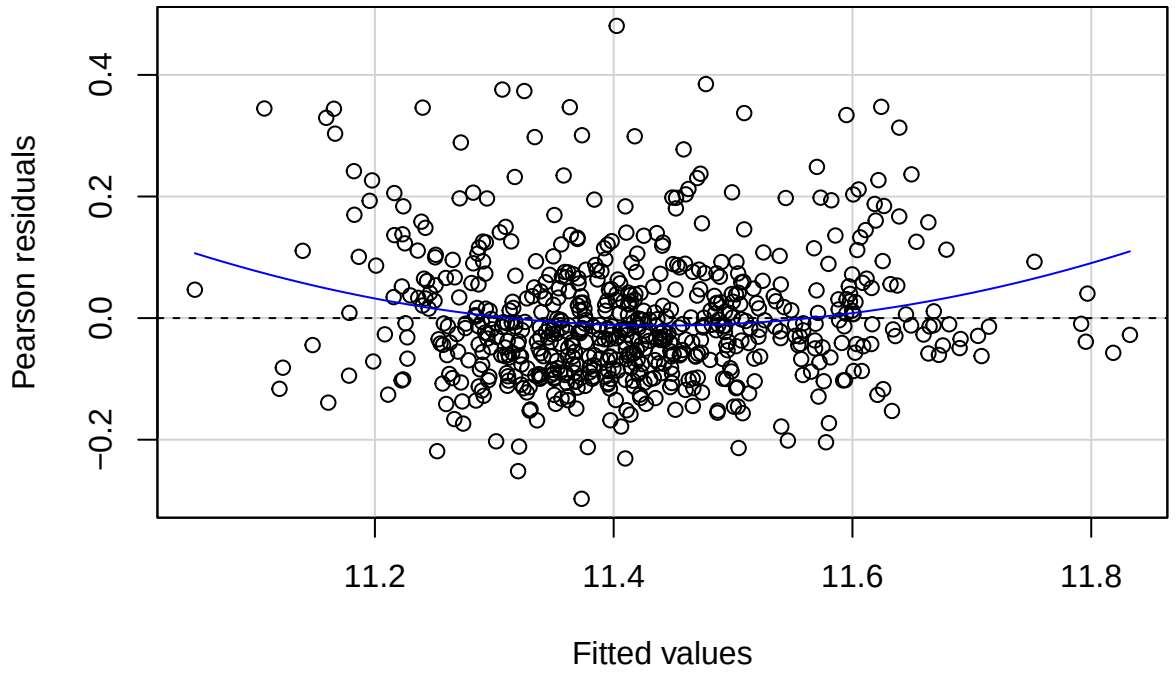
P-values still look good, and residual plots are not skewed. R^2 value is slightly lower at .553.

```
#fitting third model with natural log on response variable as well
college3_mlr <- lm(log(mid_career_pay) ~ tuition_avg + type + log(total_enrollment) + ivy_league, data =
summary(college3_mlr)
```

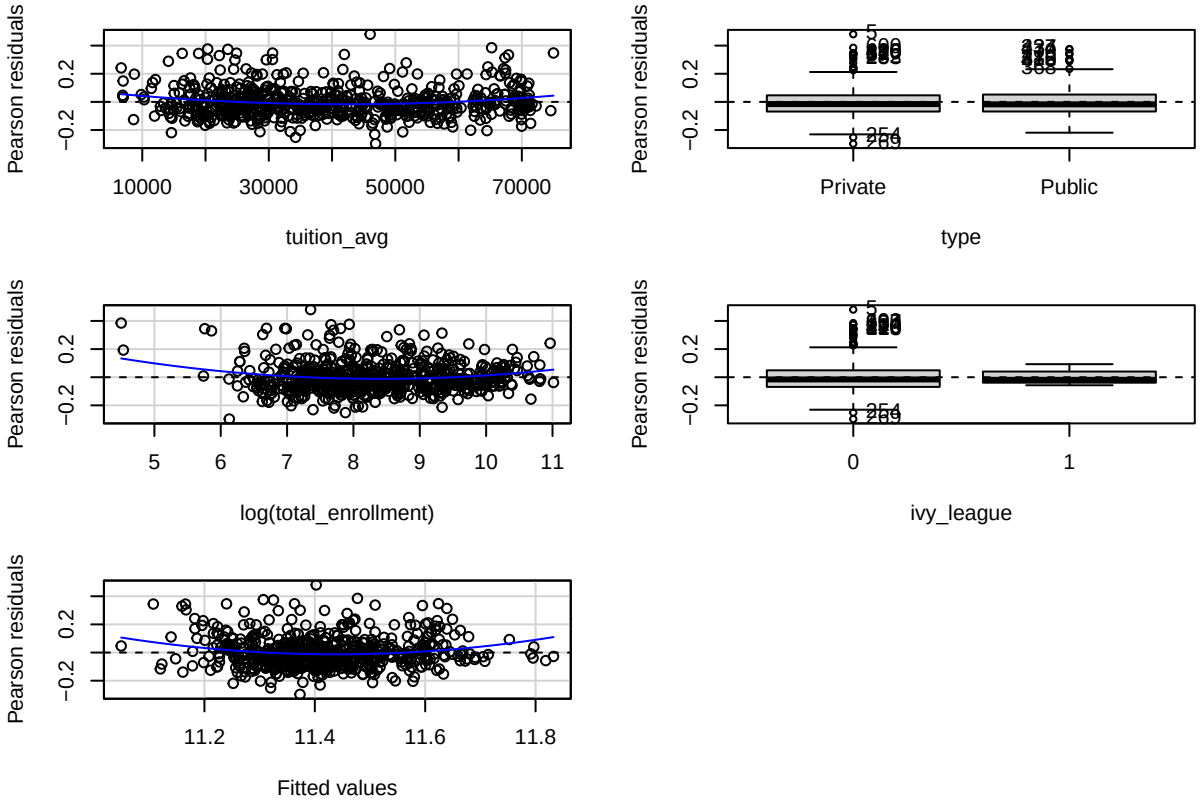
```
##
## Call:
## lm(formula = log(mid_career_pay) ~ tuition_avg + type + log(total_enrollment) +
##     ivy_league, data = full_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29719 -0.06828 -0.01604  0.04855  0.48058
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           1.080e+01  3.601e-02 299.940  < 2e-16 ***
## tuition_avg           8.332e-06  3.818e-07  21.822  < 2e-16 ***
## typePublic            1.009e-01  1.505e-02   6.699 4.62e-11 ***
## log(total_enrollment) 2.986e-02  4.793e-03   6.229 8.56e-10 ***
## ivy_league1           1.376e-01  4.540e-02   3.032  0.00253 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1083 on 635 degrees of freedom
## Multiple R-squared:  0.5672, Adjusted R-squared:  0.5645
## F-statistic: 208.1 on 4 and 635 DF,  p-value: < 2.2e-16
```

```
#checking assumptions with residual plots, everything looks fine
residualPlot(college3_mlr)
```



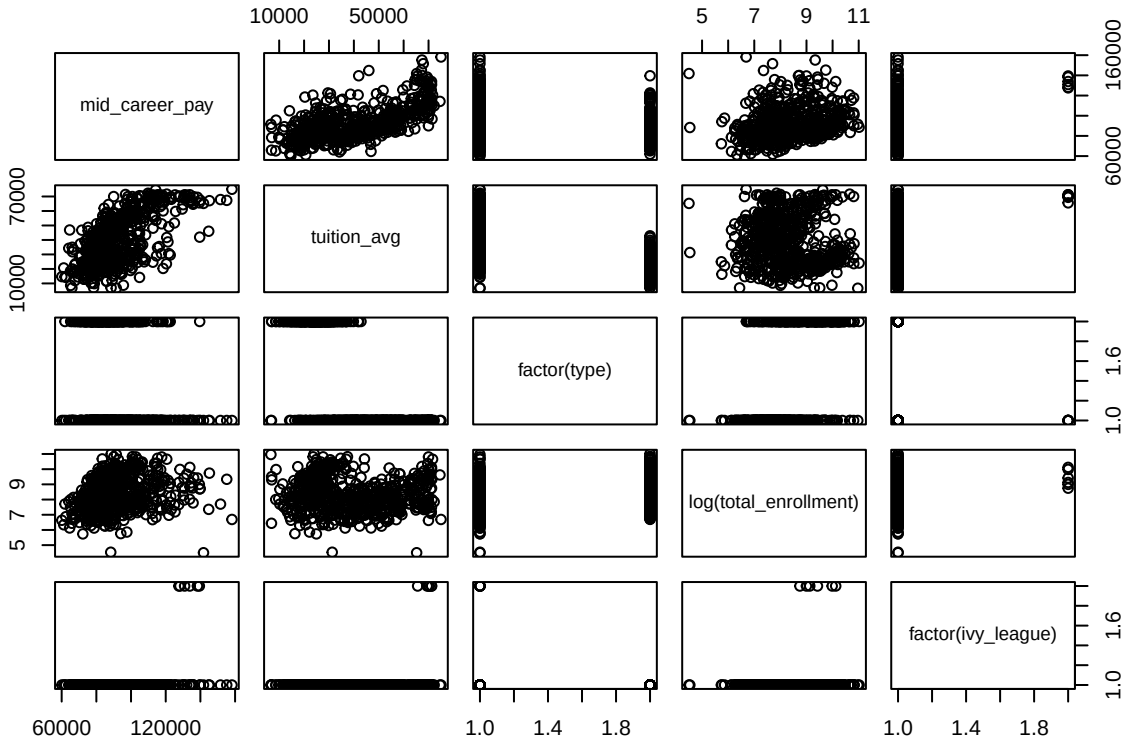

```
residualPlots(college3_mlr)
```




```
##                       Test stat Pr(>|Test stat|)
## tuition_avg              3.8223        0.0001452 ***
## type
## log(total_enrollment)    3.4996        0.0004987 ***
## ivy_league
## Tukey test               5.0098        5.449e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multicollinearity?

```
#using correlation matrices and VIF to check for multicolinearity, nothing concerning found
pairs(mid_career_pay ~ tuition_avg + factor(type) + log(total_enrollment) + factor(ivy_league), data =
```



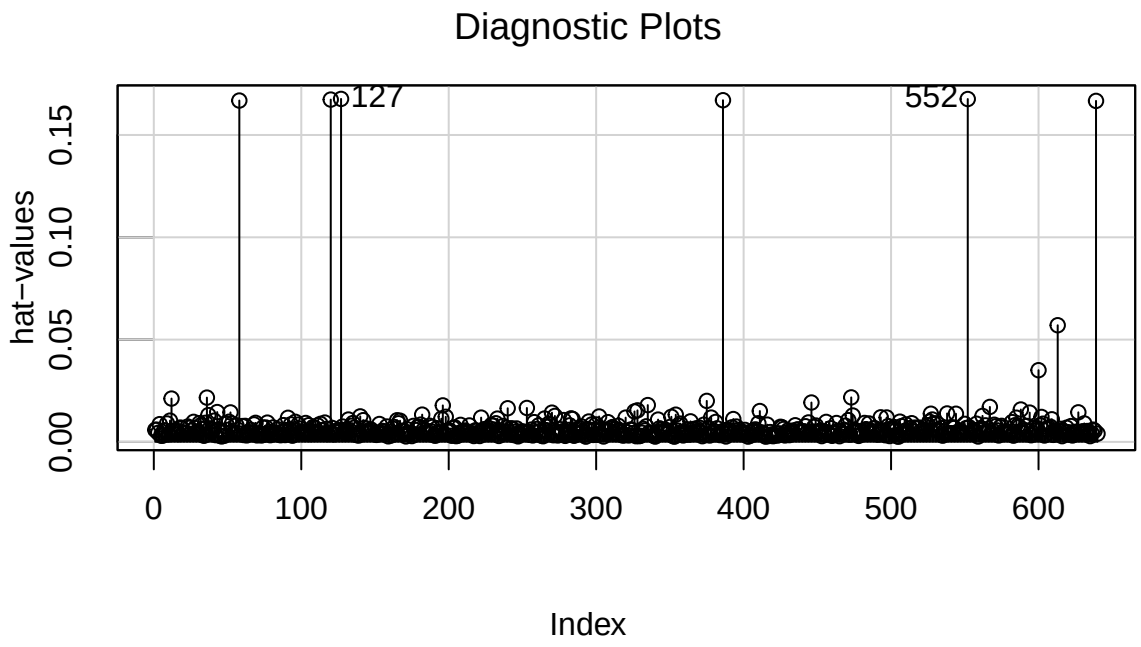

```
vif(college3_mlr)
```

```
##           tuition_avg                  type log(total_enrollment)
##              2.278492              2.908639              1.538768
##           ivy_league
##              1.044654
```

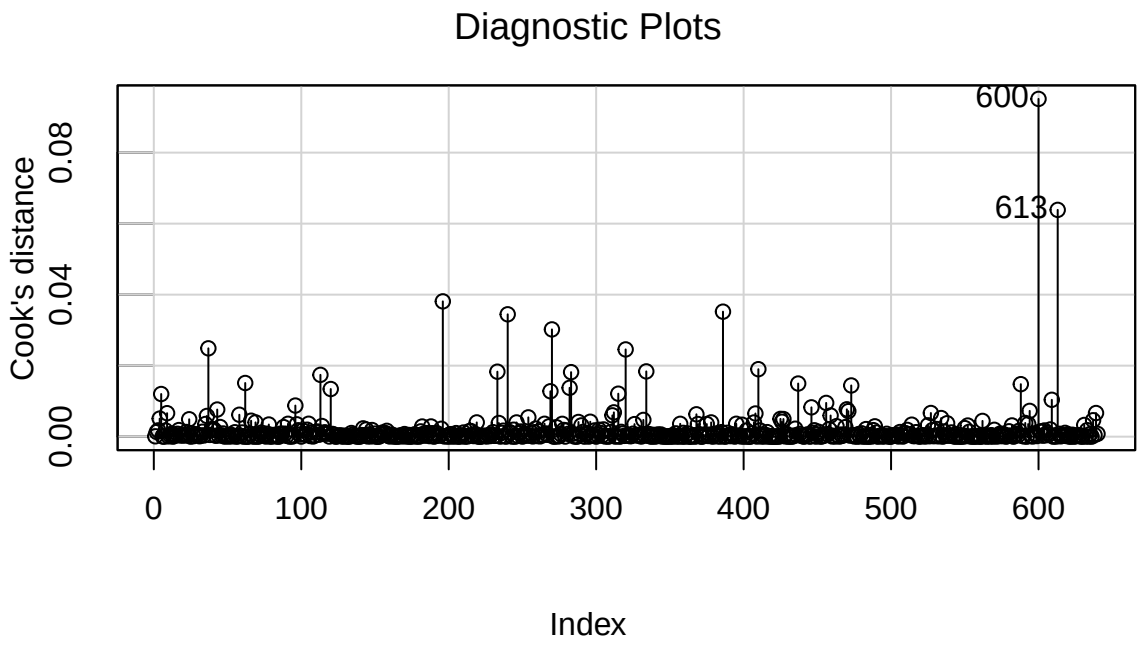
VIF: no values greater than 5! looks good

Outliers?

```
#using leverage and Cook's distance to look for influential points, nothing concerning found
infIndexPlot(college3_mlr, vars = "hat")
```




```
infIndexPlot(college3_mlr, vars = "Cook")
```



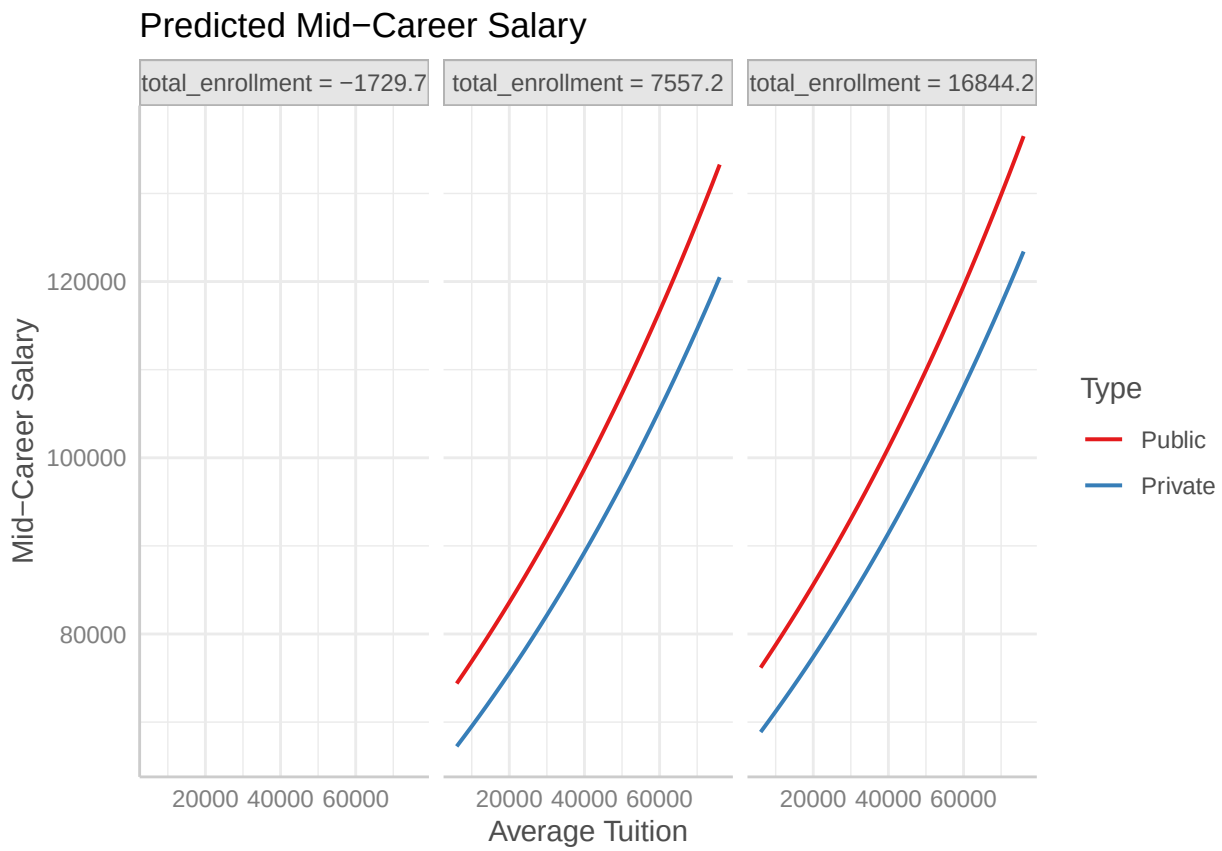

```
#graphing the model prediction with all predictors
model_pred1 <- ggpredict(college3_mlr, terms = ~tuition_avg + type + total_enrollment)
```

```
## Warning in log(total_enrollment): NaNs produced
```

```
## Model has log-transformed response. Back-transforming predictions to
##   original response scale. Standard errors are still on the log-scale.
```

```
plot(model_pred1, show_ci = FALSE) +
  labs(x = "Average Tuition",
       y = "Mid-Career Salary",
       title = "Predicted Mid-Career Salary",
       color = "Type")
```

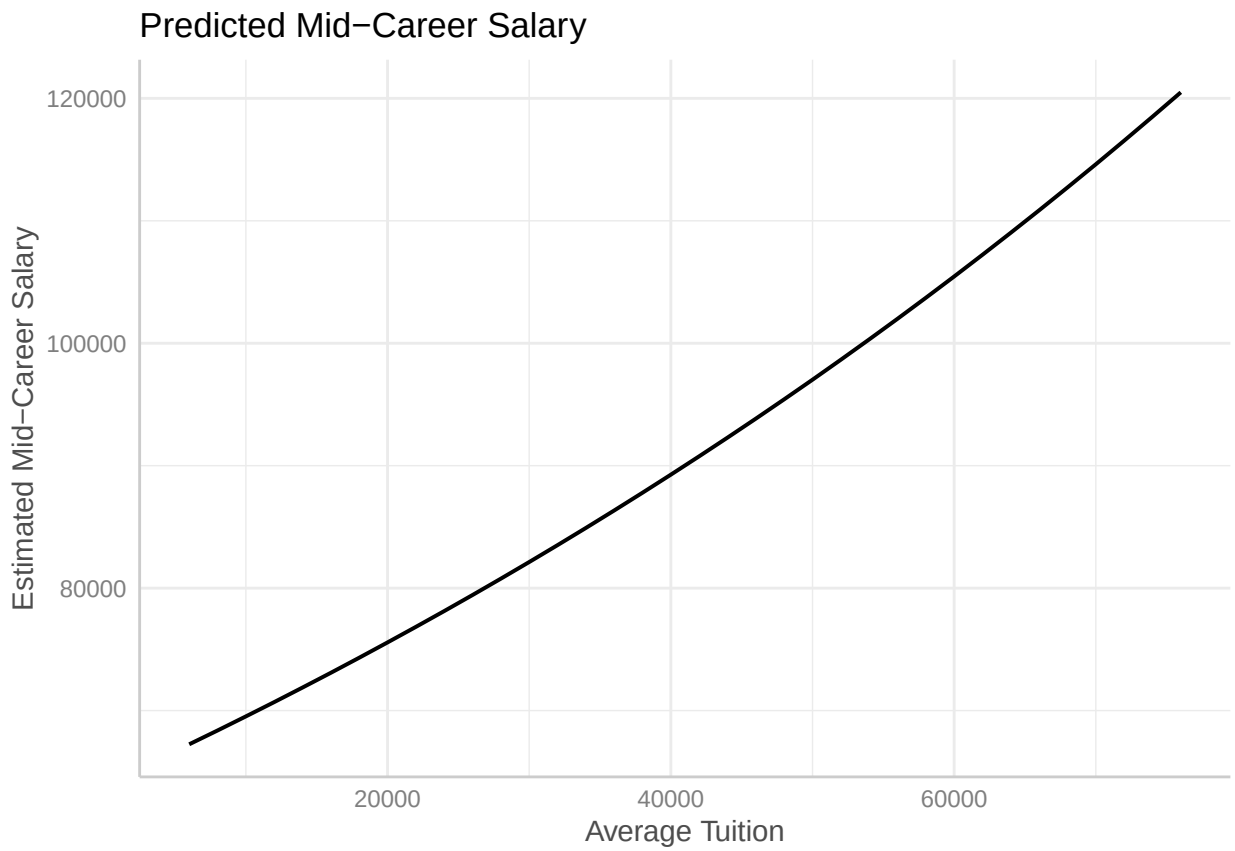
```
## Warning: Removed 72 rows containing missing values (`geom_line()`).
```




```
#graphing model prediction just tuition vs mid-career pay
model_pred2 <- ggpredict(college3_mlr, terms = ~tuition_avg)
```

```
## Model has log-transformed response. Back-transforming predictions to
##   original response scale. Standard errors are still on the log-scale.
```

```
plot(model_pred2, show_ci = FALSE) +
  labs(x = "Average Tuition",
       y = "Estimated Mid-Career Salary",
       title = "Predicted Mid-Career Salary")
```




```
#creating 95% confidence intervals for the coefficients
confint(college3_mlr, level = 0.95)
```

```
##                                 2.5 %       97.5 %
## (Intercept)           1.072878e+01 1.087019e+01
## tuition_avg           7.582081e-06 9.081608e-06
## typePublic            7.128940e-02 1.304119e-01
## log(total_enrollment) 2.044358e-02 3.926884e-02
## ivy_league1           4.847840e-02 2.267668e-01
```