**Tor Thogersen**

**DA 460 – Fall 2017**

**Lab 8 - Handout 8 R and Handout 8 SAS**

**Part 8 – R Handout - Multiple linear regression**

1. **Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.**

Observational Study due to data is being collected based on student evaluations (opinion or observations) on how the teacher performed, thus we are not able to provided true causation between how the teacher looks and evaluations. We would need to rephrase the questions to find out if there is a correlation between how the teacher looks and +/- evaluations.

1. **Describe the distribution of score. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?**

Yes, with a -.6971 skewness level the histogram is left skewed, most of the reviews are positive between 3.5 to 5.0, this histogram shows students tend to have more positive evaluations then negative evaluations, this is what I expected that most students would rate teachers as good on average and choose higher or lower rankings only if warranted.

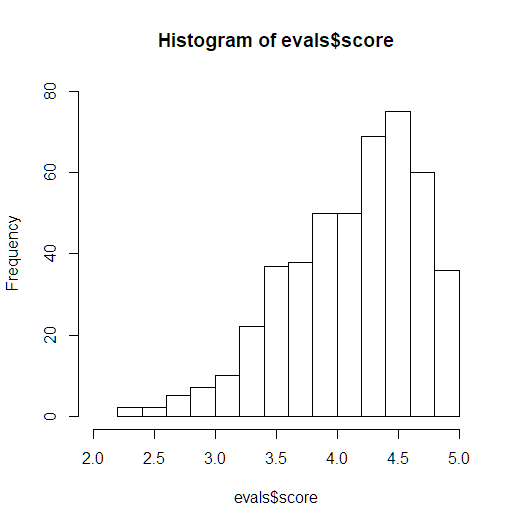
> hist(evals$score, probability = FALSE, xlim = c(2.0,5.0), ylim = c(0.0,80.0))

> skewness(evals$score)

[1] -0.697102

> mean(evals$score)

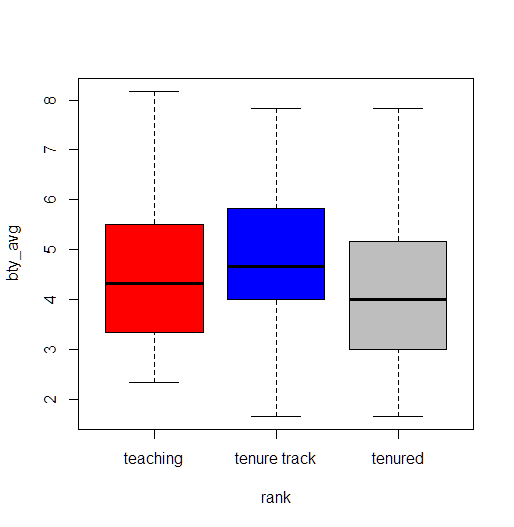
[1] 4.17473



1. **Excluding score, select two other variables and describe their relationship using an appropriate visualization (scatterplot, side-by-side boxplots, or mosaic plot).**

I choose the variables rank and avg beauty score using a side-by-side box plot, based on the data it shows that teacher on Tenure Track have the highest average bty\_avg and tenured professors have the lowest bty\_avg score.

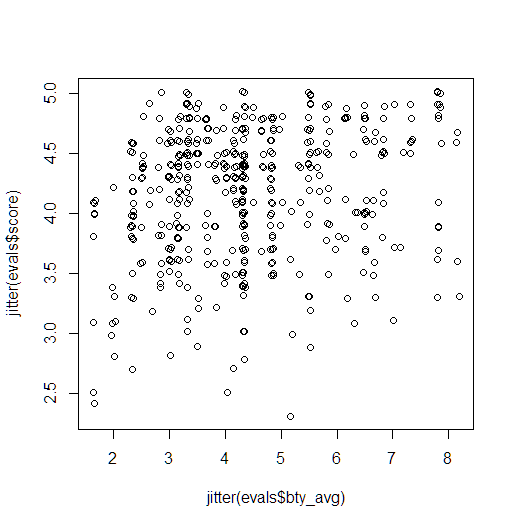
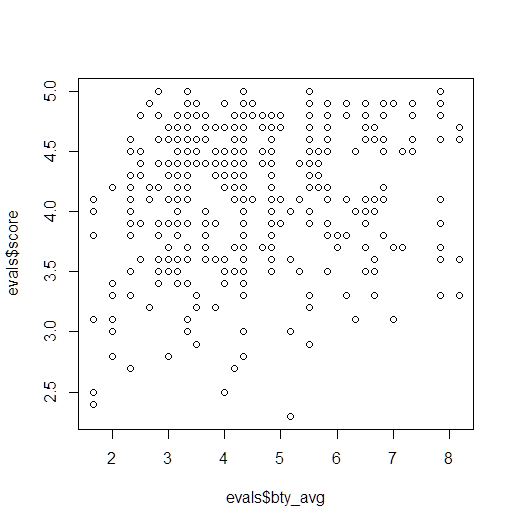
> boxplot(evals$bty\_avg~evals$rank, col = c("red", "blue", "grey"), ylab = "bty\_avg", xlab = "rank")



1. **Replot the scatterplot, but this time use the function jitter() on the y- or the x-coordinate. (Use ?jitter to learn more.) What was misleading about the initial scatterplot?**

Jitter function enables you to visualize the scatter plots better and see the intersect points on both y & x axis better, first scatter plot lines everything up on linear and vertical axis making you believe each intersect lines up perfectly.

> plot(evals$score ~ evals$bty\_avg)



> plot(jitter(evals$score)~jitter(evals$bty\_avg))

1. **Let’s see if the apparent trend in the plot is something more than natural variation. Fit a linear model called m\_bty to predict average professor score by average beauty rating and add the line to your plot using abline(m\_bty). Write out the equation for the linear model and interpret the slope. Is average beauty score a statistically significant predictor? Does it appear to be a practically significant predictor**

y-hat = 3.88034 + 0.06664 \* bty\_avg

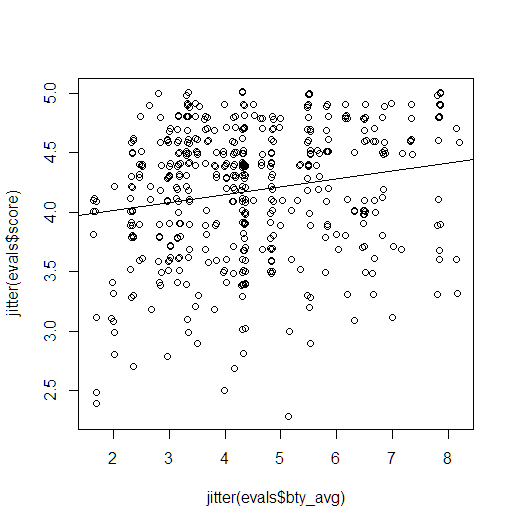
based on the results we have multiple factors to look at; first I checked correlation and show data is less than .8 so we are good, after performing the summary on lm:

yes, because p-value is less than alpha we have statistically significant predictor but because the incremental change is .06 the score will not change dramatically so it may not be a practical significant predictor.

> m\_bty <- lm(evals$score ~ evals$bty\_avg)

> plot(jitter(evals$score) ~ jitter(evals$bty\_avg))

> abline(m\_bty)



> cor(evals$score, evals$bty\_avg)

[1] 0.1871424

> summary(m\_bty)

Call:

lm(formula = evals$score ~ evals$bty\_avg)

Residuals:

Min 1Q Median 3Q Max

-1.9246 -0.3690 0.1420 0.3977 0.9309

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.88034 0.07614 50.96 < 2e-16 \*\*\*

evals$bty\_avg 0.06664 0.01629 4.09 5.08e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5348 on 461 degrees of freedom

Multiple R-squared: 0.03502, Adjusted R-squared: 0.03293

F-statistic: 16.73 on 1 and 461 DF, p-value: 5.083e-05

1. **Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these)**

First: plot does not have even disbursement at 0.0, the spread from 0 to -2.0 is greater than spread 0.0 to 1.0

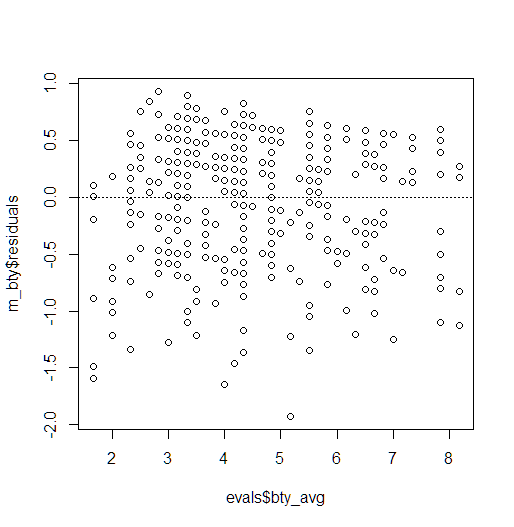
second: skew of -.706 which shows a left skewed chart that does have slight bell curve.

Third: the plot has slight curve to data but shows an upward trend

Overall the data fails to meet the normal residuals conditions, I believe they due to outliers that need to be researched and data needs to be cleaned up.

> plot(m\_bty$residuals ~ evals$bty\_avg)

> abline(h = 0, lty = 3)



> summary(m\_bty$residuals)

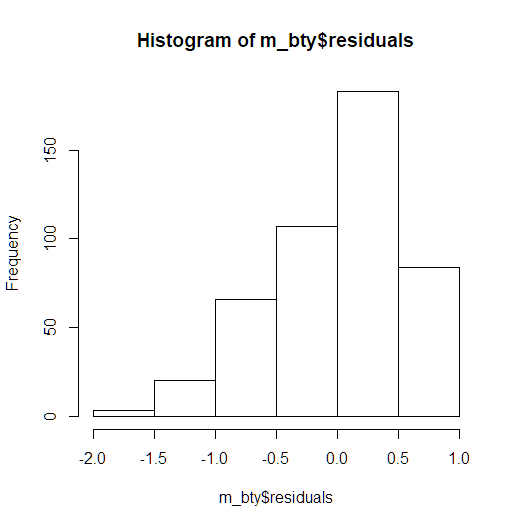
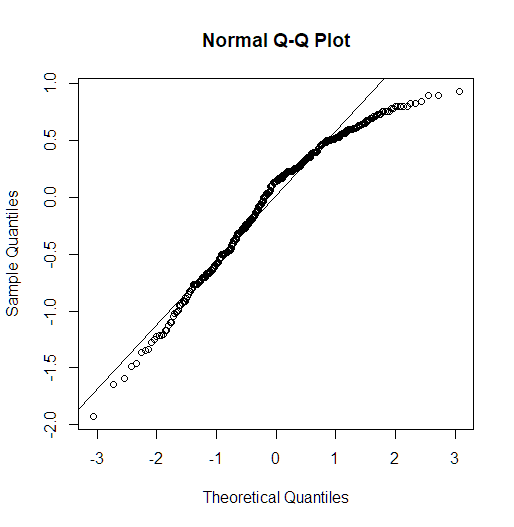
Min. 1st Qu. Median Mean 3rd Qu. Max.

-1.9247 -0.3690 0.1420 0.0000 0.3977 0.9309

> hist(m\_bty$residuals)

> skewness(m\_bty$residuals)

[1] -0.7067592

> qqnorm(m\_bty$residuals)

> qqline(m\_bty$residuals)

1. **P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.**

* Scatter plot show no grouping or patters and is randomaly spread out, we do see an uneven balance at 0.0 for the disbursement with may cause some concerns.
* Histogram has a left skew but is still showing a bell curve and normal distribution
* Normal probability plot follows the linear line with slight curve at the to end showing some outliers

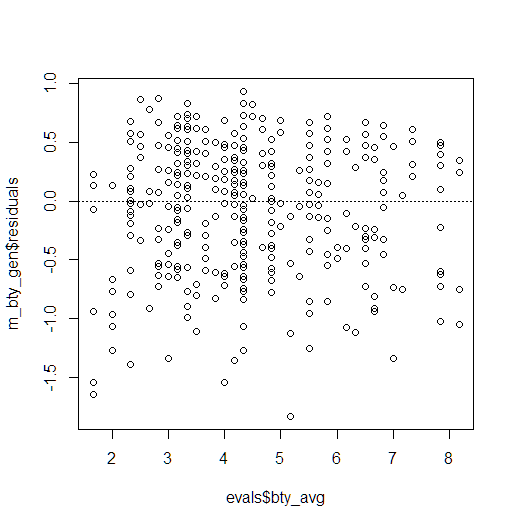
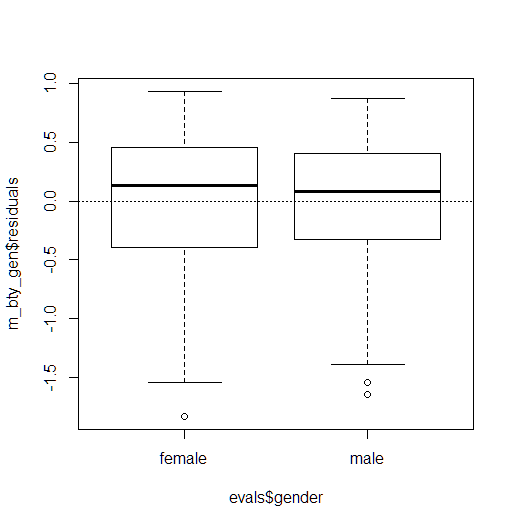
Overall, I would I believe the conditions have been met.

> plot(m\_bty\_gen$residuals ~ evals$bty\_avg)

> abline(h = 0, lty=3)

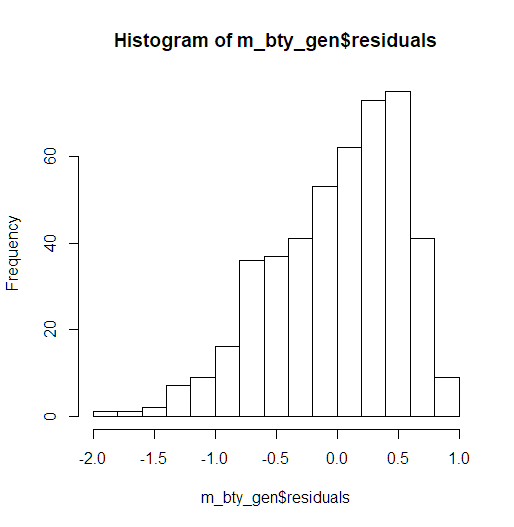
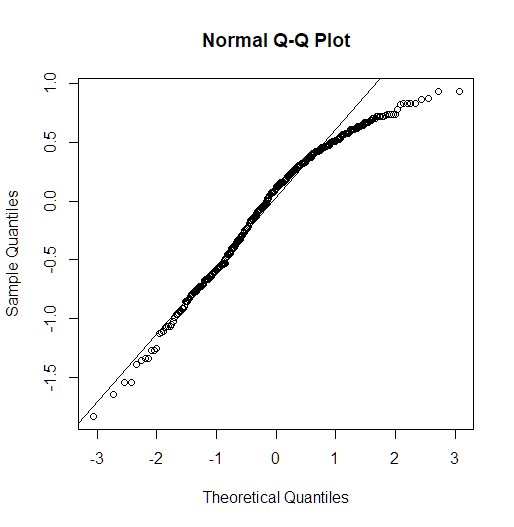
> plot(m\_bty\_gen$residuals ~ evals$gender)

> abline(h = 0, lty=3)

> qqnorm(m\_bty\_gen$residuals)

> qqline(m\_bty\_gen$residuals)



> hist(m\_bty\_gen$residuals)

1. **Is bty\_avg still a significant predictor of score? Has the addition of gender to the model changed the parameter estimate for bty\_avg?**

Yes, bty\_avg is increased in signifance due to lower p-value 6.48e-06

> summary(m\_bty\_gen)

Call:

lm(formula = score ~ bty\_avg + gender, data = evals)

Residuals:

Min 1Q Median 3Q Max

-1.8305 -0.3625 0.1055 0.4213 0.9314

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.74734 0.08466 44.266 < 2e-16 \*\*\*

bty\_avg 0.07416 0.01625 4.563 6.48e-06 \*\*\*

gendermale 0.17239 0.05022 3.433 0.000652 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5287 on 460 degrees of freedom

Multiple R-squared: 0.05912, Adjusted R-squared: 0.05503

F-statistic: 14.45 on 2 and 460 DF, p-value: 8.177e-07

1. **What is the equation of the line corresponding to males? (Hint: For males, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score?**

scoreˆ=β^0+β^1×bty\_avg+β^2×(1)

=β^0+β^1×bty\_avg

1. **Create a new model called m\_bty\_rank with gender removed and rank added in. How does R appear to handle categorical variables that have more than two levels? Note that the rank variable has three levels: teaching, tenure track, tenured.**

> m\_bty\_rank <- lm(score ~ bty\_avg + rank, data = evals)

> summary(m\_bty\_rank)

Call:

lm(formula = score ~ bty\_avg + rank, data = evals)

Residuals:

Min 1Q Median 3Q Max

-1.8713 -0.3642 0.1489 0.4103 0.9525

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.98155 0.09078 43.860 < 2e-16 \*\*\*

bty\_avg 0.06783 0.01655 4.098 4.92e-05 \*\*\*

ranktenure track -0.16070 0.07395 -2.173 0.0303 \*

ranktenured -0.12623 0.06266 -2.014 0.0445 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5328 on 459 degrees of freedom

Multiple R-squared: 0.04652, Adjusted R-squared: 0.04029

F-statistic: 7.465 on 3 and 459 DF, p-value: 6.88e-05

> names(m\_bty\_rank)

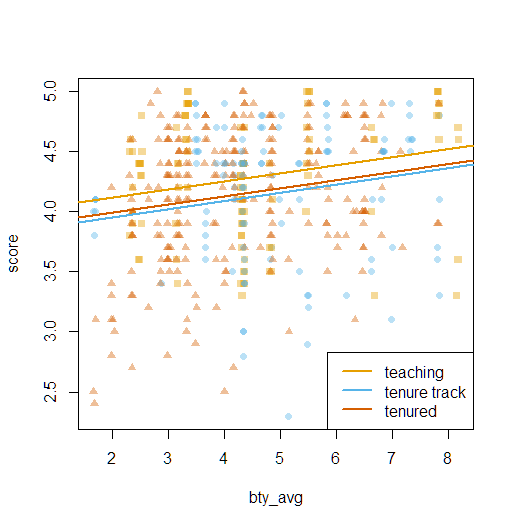
[1] "coefficients" "residuals" "effects" "rank"

[5] "fitted.values" "assign" "qr" "df.residual"

[9] "contrasts" "xlevels" "call" "terms"

[13] "model"

> multiLines(m\_bty\_rank)



1. **Which variable would you expect to have the highest p-value in this model? Why? Hint: Think about which variable would you expect to not have any association with the professor score.**

cls\_profssingle has the hightest P-value: 0.77806

> m\_full <- lm(score ~ rank + ethnicity + gender + language + age + cls\_perc\_eval + cls\_students + cls\_level + cls\_profs + cls\_credits + bty\_avg + pic\_outfit + pic\_color, data = evals)

> summary(m\_full)

Call:

lm(formula = score ~ rank + ethnicity + gender + language + age +

cls\_perc\_eval + cls\_students + cls\_level + cls\_profs + cls\_credits +

bty\_avg + pic\_outfit + pic\_color, data = evals)

Residuals:

Min 1Q Median 3Q Max

-1.77397 -0.32432 0.09067 0.35183 0.95036

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.0952141 0.2905277 14.096 < 2e-16 \*\*\*

ranktenure track -0.1475932 0.0820671 -1.798 0.07278 .

ranktenured -0.0973378 0.0663296 -1.467 0.14295

ethnicitynot minority 0.1234929 0.0786273 1.571 0.11698

gendermale 0.2109481 0.0518230 4.071 5.54e-05 \*\*\*

languagenon-english -0.2298112 0.1113754 -2.063 0.03965 \*

age -0.0090072 0.0031359 -2.872 0.00427 \*\*

cls\_perc\_eval 0.0053272 0.0015393 3.461 0.00059 \*\*\*

cls\_students 0.0004546 0.0003774 1.205 0.22896

cls\_levelupper 0.0605140 0.0575617 1.051 0.29369

cls\_profssingle -0.0146619 0.0519885 -0.282 0.77806

cls\_creditsone credit 0.5020432 0.1159388 4.330 1.84e-05 \*\*\*

bty\_avg 0.0400333 0.0175064 2.287 0.02267 \*

pic\_outfitnot formal -0.1126817 0.0738800 -1.525 0.12792

pic\_colorcolor -0.2172630 0.0715021 -3.039 0.00252 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.498 on 448 degrees of freedom

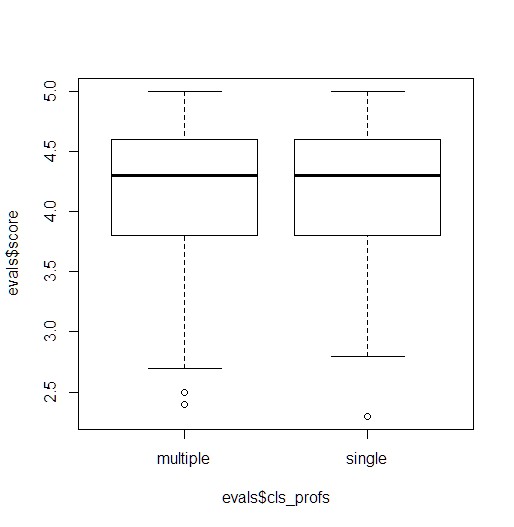
Multiple R-squared: 0.1871, Adjusted R-squared: 0.1617

F-statistic: 7.366 on 14 and 448 DF, p-value: 6.552e-14

1. **Check your suspicions from the previous exercise. Include the model output in your response**

cls\_profs has the highest pvalue and least association to scores.

> plot(evals$score ~ evals$cls\_profs)



1. **Interpret the coefficient associated with the ethnicity variable.**

Ethnicity pvalue = 0.11698 has the high pvalue and can be removed from the data model to increase the overall adjusted RSquare value.

1. **Drop the variable with the highest p-value and re-fit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficient estimates depend on the other variables that are included in the model.) If not, what does this say about whether or not the dropped variable was collinear with the other explanatory variables**

Adjusted R-squared: 0.1634 improved meaning the data model is more significant to the data.

> m\_full\_mod <- lm(score ~ rank + ethnicity + gender + language + age + cls\_perc\_eval +

+ cls\_students + cls\_level + cls\_credits + bty\_avg + pic\_outfit + pic\_color, data = evals)

> summary(m\_full\_mod)

Call:

lm(formula = score ~ rank + ethnicity + gender + language + age +

cls\_perc\_eval + cls\_students + cls\_level + cls\_credits +

bty\_avg + pic\_outfit + pic\_color, data = evals)

Residuals:

Min 1Q Median 3Q Max

-1.7836 -0.3257 0.0859 0.3513 0.9551

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.0872523 0.2888562 14.150 < 2e-16 \*\*\*

ranktenure track -0.1476746 0.0819824 -1.801 0.072327 .

ranktenured -0.0973829 0.0662614 -1.470 0.142349

ethnicitynot minority 0.1274458 0.0772887 1.649 0.099856 .

gendermale 0.2101231 0.0516873 4.065 5.66e-05 \*\*\*

languagenon-english -0.2282894 0.1111305 -2.054 0.040530 \*

age -0.0089992 0.0031326 -2.873 0.004262 \*\*

cls\_perc\_eval 0.0052888 0.0015317 3.453 0.000607 \*\*\*

cls\_students 0.0004687 0.0003737 1.254 0.210384

cls\_levelupper 0.0606374 0.0575010 1.055 0.292200

cls\_creditsone credit 0.5061196 0.1149163 4.404 1.33e-05 \*\*\*

bty\_avg 0.0398629 0.0174780 2.281 0.023032 \*

pic\_outfitnot formal -0.1083227 0.0721711 -1.501 0.134080

pic\_colorcolor -0.2190527 0.0711469 -3.079 0.002205 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4974 on 449 degrees of freedom

Multiple R-squared: 0.187, Adjusted R-squared: 0.1634

F-statistic: 7.943 on 13 and 449 DF, p-value: 2.336e-14

1. **Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write out the linear model for predicting score based on the final model you settle on.**

scoreˆ=β^0+β^1×ethnicity\_not\_minority+β^2×gender\_male+β^3×language\_non−englist+β^4×age+β^5+×class\_perceval+β^6×class\_credits\_one+β^7×bty\_avg+β^8×picture\_color\_colored

> m\_full\_final <- lm(score ~ gender+ ethnicity + language + age + cls\_perc\_eval + cls\_credits + bty\_avg + pic\_color, data = evals)

> summary(m\_full\_final)

Call:

lm(formula = score ~ gender + ethnicity + language + age + cls\_perc\_eval +

cls\_credits + bty\_avg + pic\_color, data = evals)

Residuals:

Min 1Q Median 3Q Max

-1.85320 -0.32394 0.09984 0.37930 0.93610

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.771922 0.232053 16.255 < 2e-16 \*\*\*

gendermale 0.207112 0.050135 4.131 4.30e-05 \*\*\*

ethnicitynot minority 0.167872 0.075275 2.230 0.02623 \*

languagenon-english -0.206178 0.103639 -1.989 0.04726 \*

age -0.006046 0.002612 -2.315 0.02108 \*

cls\_perc\_eval 0.004656 0.001435 3.244 0.00127 \*\*

cls\_creditsone credit 0.505306 0.104119 4.853 1.67e-06 \*\*\*

bty\_avg 0.051069 0.016934 3.016 0.00271 \*\*

pic\_colorcolor -0.190579 0.067351 -2.830 0.00487 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4992 on 454 degrees of freedom

Multiple R-squared: 0.1722, Adjusted R-squared: 0.1576

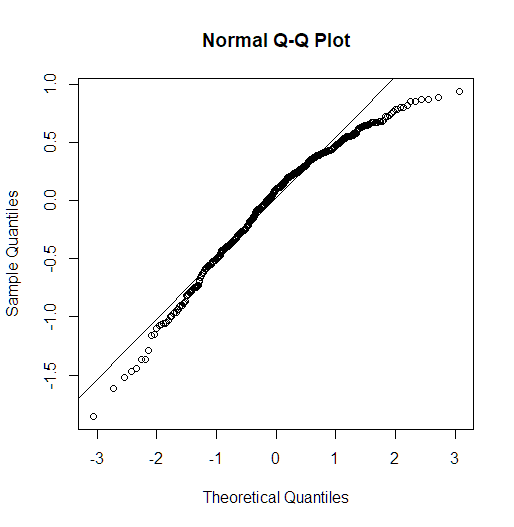
F-statistic: 11.8 on 8 and 454 DF, p-value: 2.58e-15

1. **Verify that the conditions for this model are reasonable using diagnostic plots.**

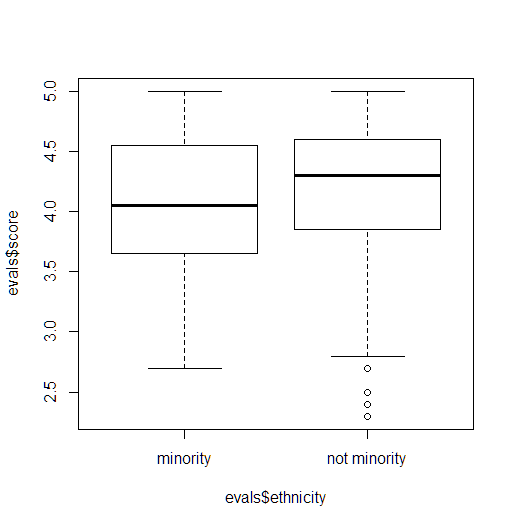
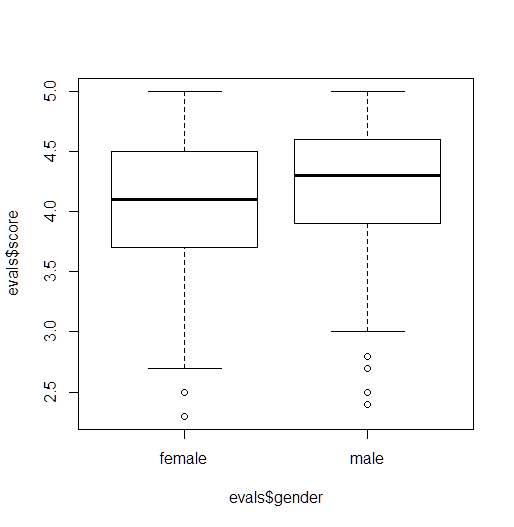
Overall, all residual values passed or met the normality / linear test

> qqnorm(m\_full\_final$residuals)

> qqline(m\_full\_final$residuals)

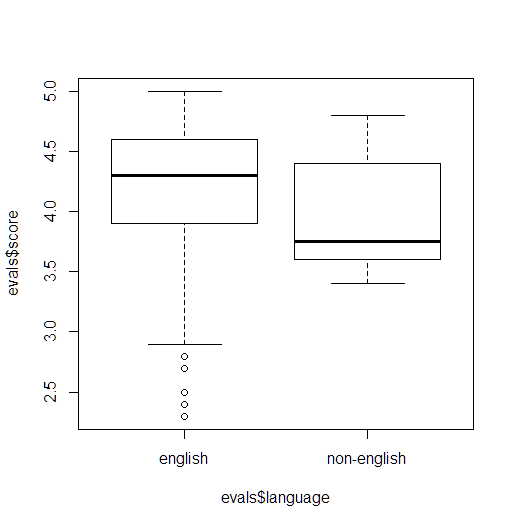
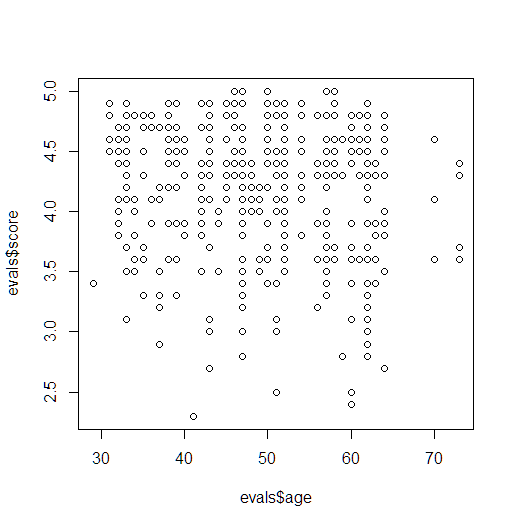


> plot(evals$score ~ evals$ethnicity)

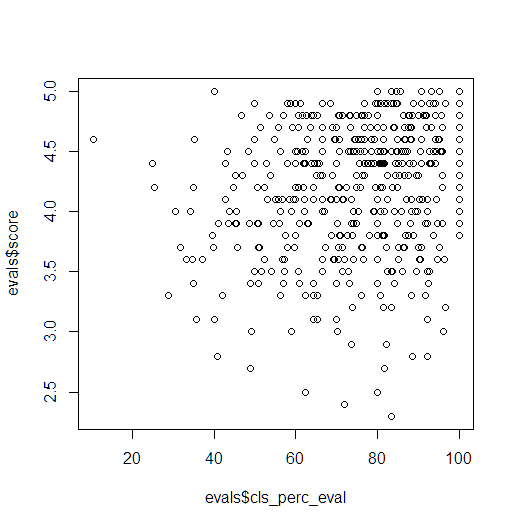
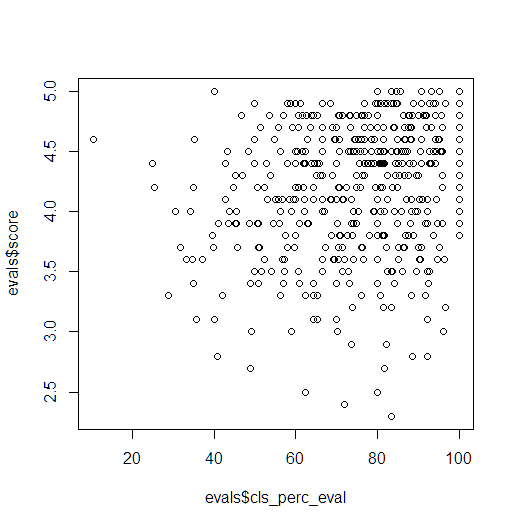
> plot(evals$score ~ evals$gender)

> plot(evals$score ~ evals$language)

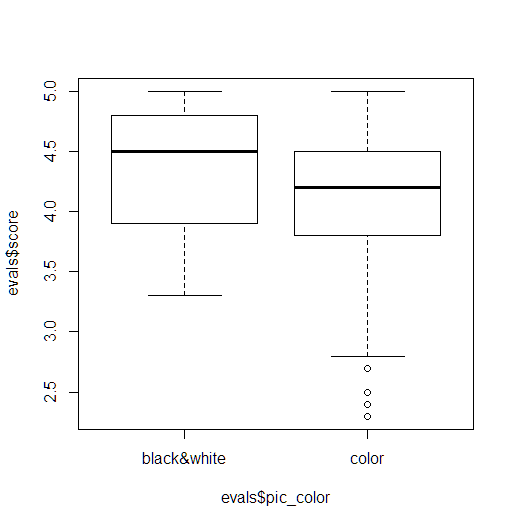
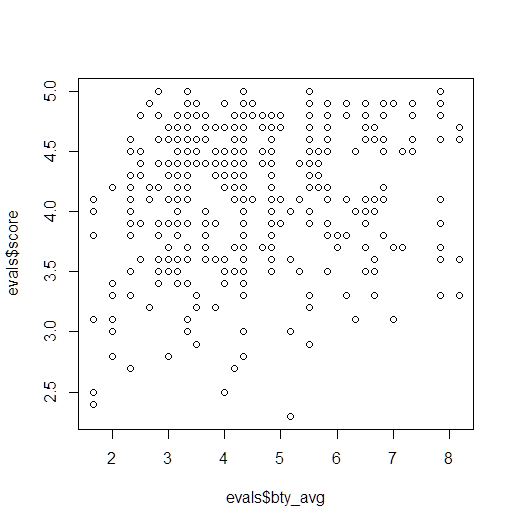
> plot(evals$score ~ evals$age)

> plot(evals$score ~ evals$cls\_perc\_eval)

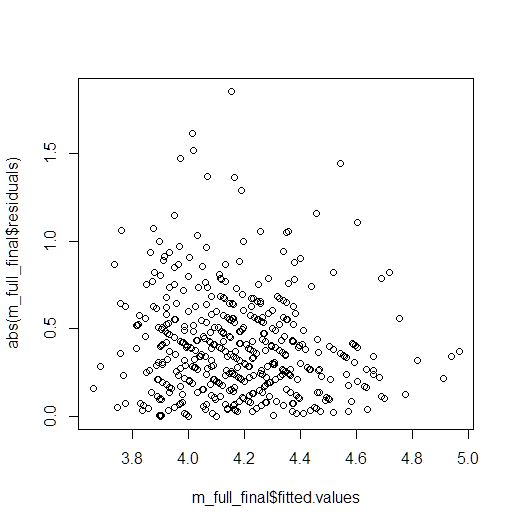
> plot(evals$score ~ evals$cls\_credits)

> plot(evals$score ~ evals$bty\_avg)



> plot(evals$score ~ evals$pic\_color)

> plot(abs(m\_full\_final$residuals) ~ m\_full\_final$fitted.values)



1. **The original paper describes how these data were gathered by taking a sample of professors from the University of Texas at Austin and including all courses that they have taught. Considering that each row represents a course, could this new information have an impact on any of the conditions of linear regression?**

No, because classes and are independent of each other so evaluations scores would not be impacted. More data may improve the study.

1. **Based on your final model, describe the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score.**

Teacher should have the following characteristics: Male, English as primary language, teach a one credit course, class photo (B&W), age (33 – 60), his class should have high participation in teacher evaluation.

1. **Would you be comfortable generalizing your conclusions to apply to professors generally (at any university)? Why or why not?**

no, I would like to perform similar studies at multiple university with different demographics to have a solid population sample.

**Part 8 – SAS Handout - Multiple linear regression**

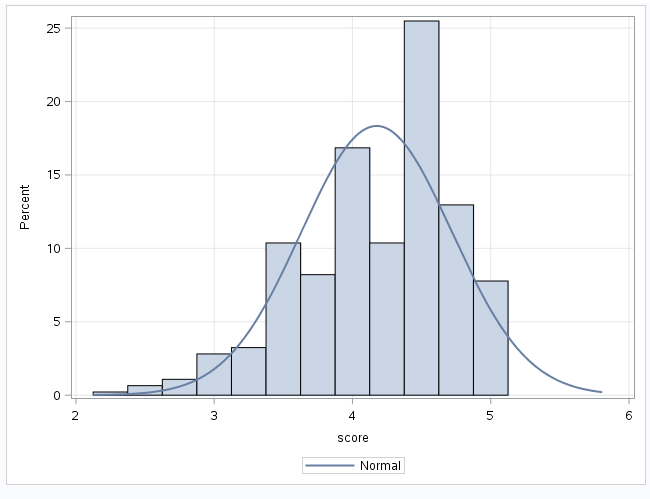
1. **Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.**

Observational Study due to data is being collected based on student evaluations (opinion or observations) on how the teacher performed, thus we are not able to provided true causation between how the teacher looks and evaluations. We would need to rephrase the questions to find out if there is a correlation between how the teacher looks and +/- evaluations.

1. **Describe the distribution of score. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?**

proc sgplot data=WORK.EVALS;  
 histogram score /;  
 density score;  
 yaxis grid;

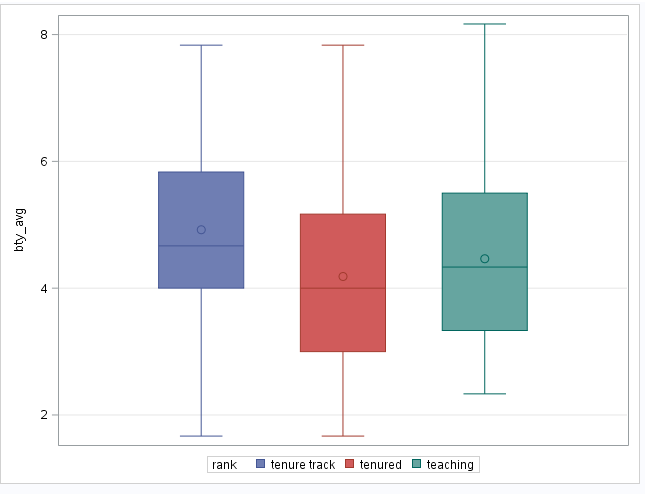
xaxix grid;  
run;



1. **Excluding score, select two other variables and describe their relationship using an appropriate visualization (scatterplot, side-by-side boxplots, or mosaic plot)**

I choose the variables rank and avg beauty score using a side-by-side box plot, based on the data it shows that teacher on Tenure Track have the highest average bty\_avg and tenured professors have the lowest bty\_avg score.

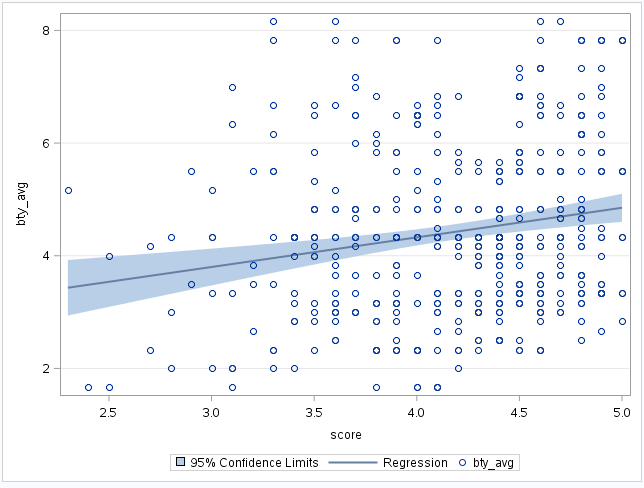
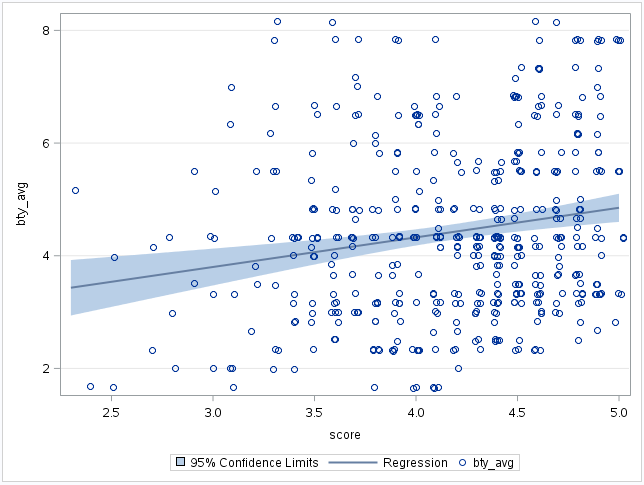
proc sgplot data=WORK.EVALS;  
 vbox bty\_avg / group=rank grouporder=Data name='Box';  
 yaxis grid;  
run;



1. **Replot the scatterplot, but this time use the JITTER option in the SCATTER statement. (Use ?jitter to learn more.) What was misleading about the initial scatterplot?**

Jitter function enables you to visualize the scatter plots better and see the intersect points on both y & x axis better, first scatter plot lines everything up on linear and vertical axis making you believe each intersect lines up perfectly.

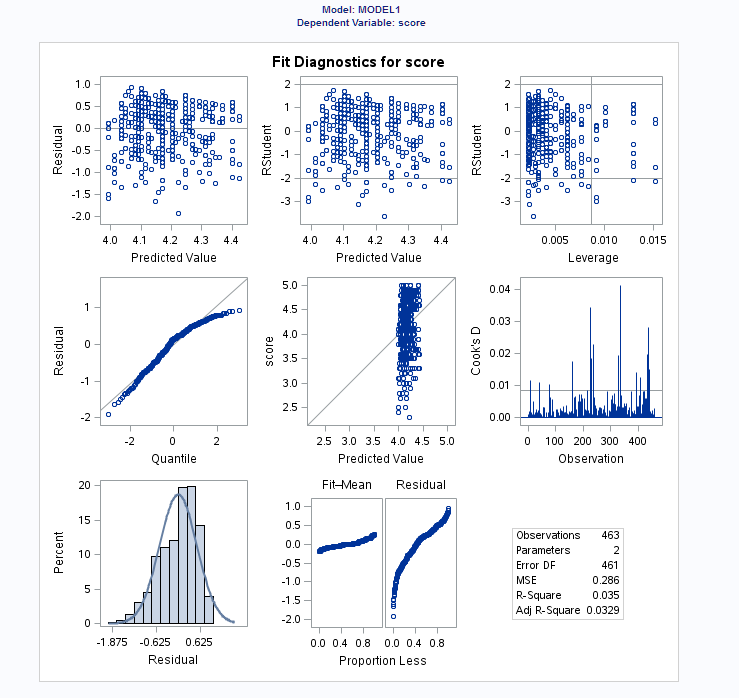
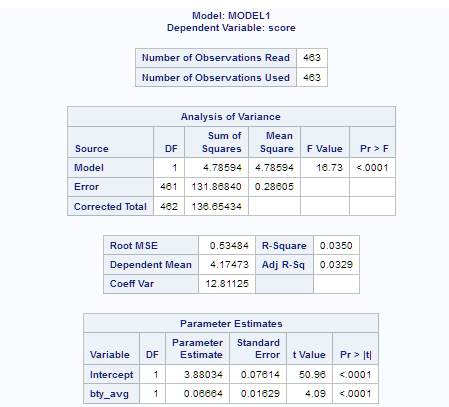
proc sgplot data=WORK.EVALS;  
 reg x=score y=bty\_avg / nomarkers CLM name='Regression';  
 scatter x=score y=bty\_avg /;  
 yaxis grid;  
run;

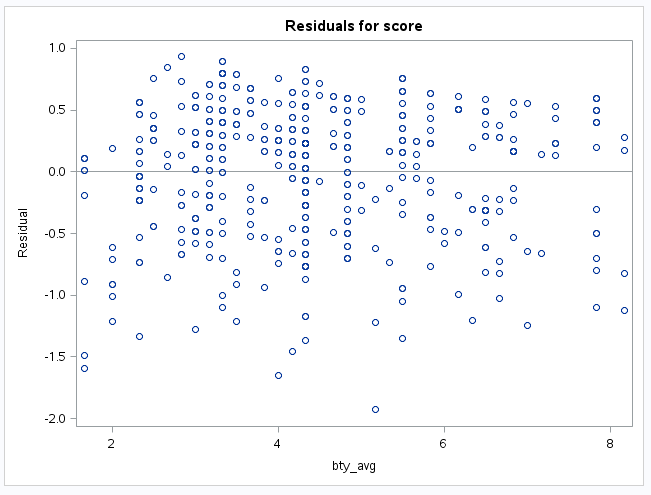
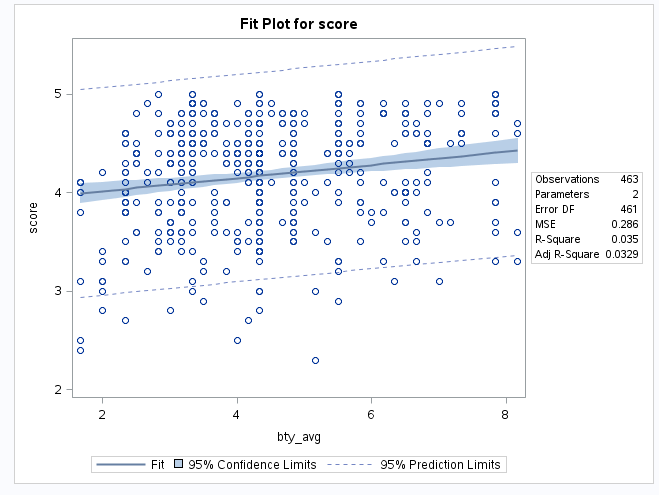
   
proc sgplot data=WORK.EVALS;  
 reg x=score y=bty\_avg / nomarkers CLM name='Regression';  
 scatter x=score y=bty\_avg / JITTER;  
 yaxis grid;  
run;

1. **Let’s see if the apparent trend in the plot is something more than natural variation. Fit a linear model called m\_bty to predict average professor score by average beauty rating and add the line to your plot using abline(m\_bty). Write out the equation for the linear model and interpret the slope. Is average beauty score a statistically significant predictor? Does it appear to be a practically significant predictor**

Yes, based on the fit diagnostic score the I do not see a discerning pattern, spread is good with lots of variability, distribution is left skewed with normal distribution, only issue would be the linear regression line has slight curve at the top due to outliers. I would say due to the limited increase of .06 for slope the practical significant is limited.

proc reg data=work.evals;  
 model score=bty\_avg;  
run;

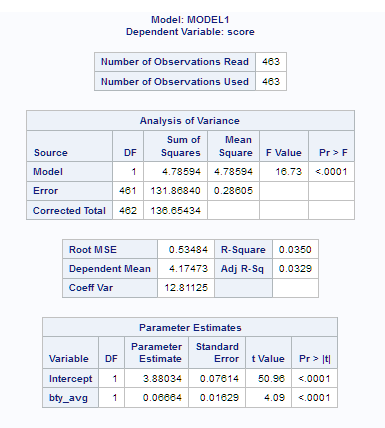
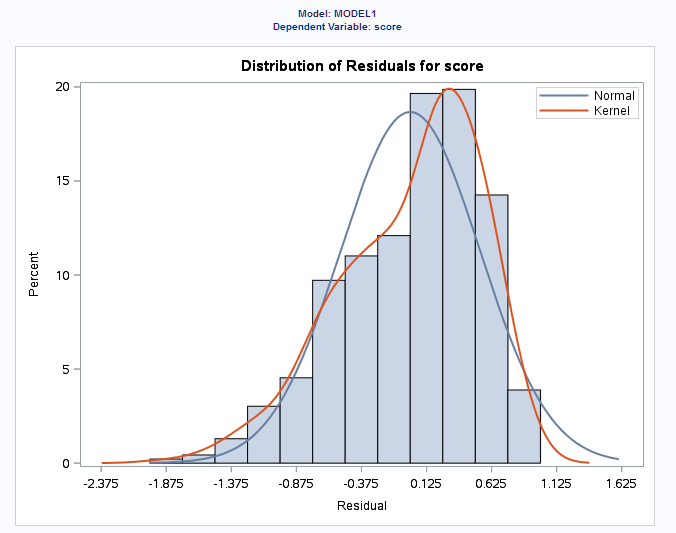


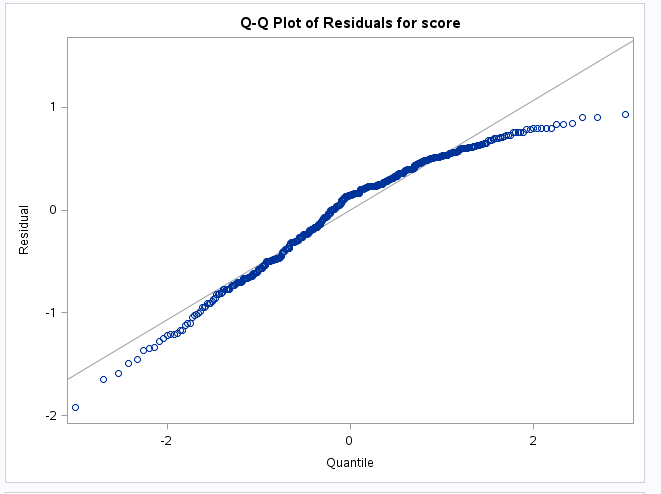
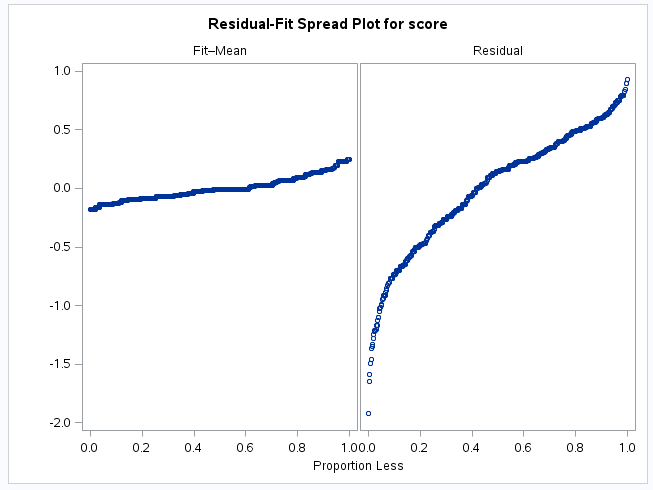


1. **Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these)**

Yes, based on the charts and plots, I see a good spread with variability, almost linear LSR line and normal distribution with left skewness.

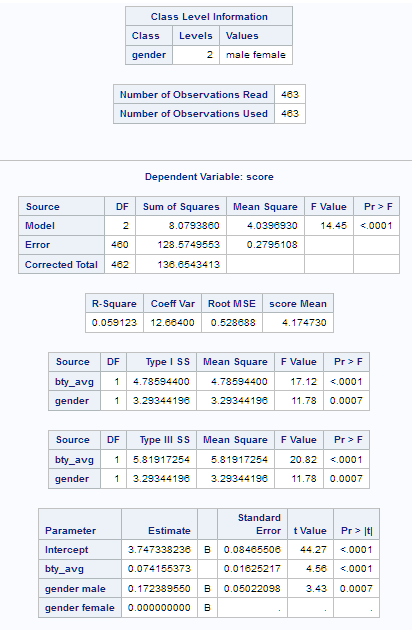
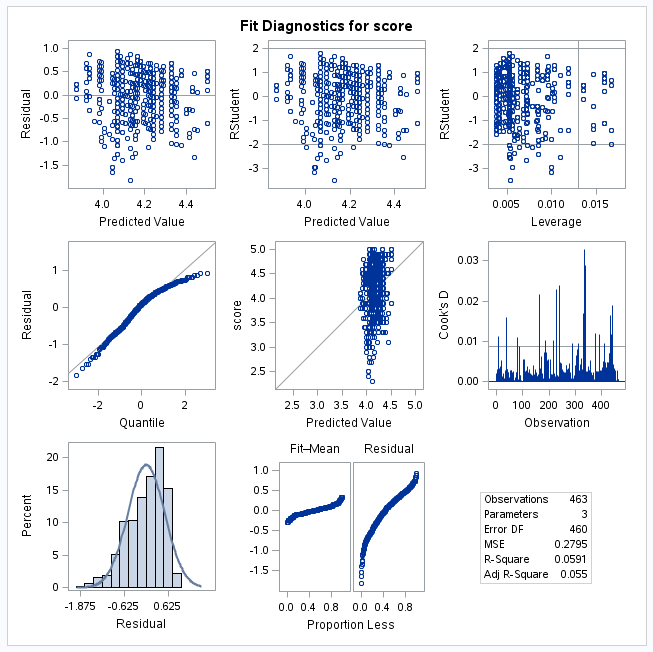
proc reg data=work.evals plots=diagnostics (unpack);  
 model score=bty\_avg;  
run;



1. **P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.**

Yes, based on the fit diagnostics for data model we have normality, linearity and variability.

1. **Is bty\_avg still a significant predictor of score? Has the addition of gender to the model changed the parameter estimate for bty\_avg?**

Bty\_avg is still the better parameter in the data model due to having smaller p-value and being even more significant now that we added the additional variable.

1. **What is the equation of the line corresponding to males? (Hint: For males, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score?**

scoreˆ=β^0+β^1×bty\_avg+β^2×(1)

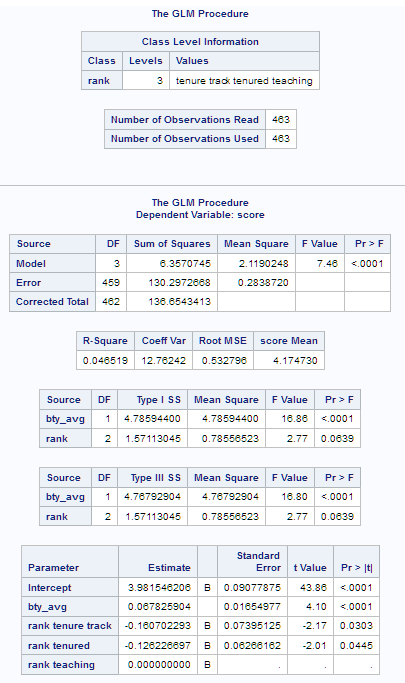
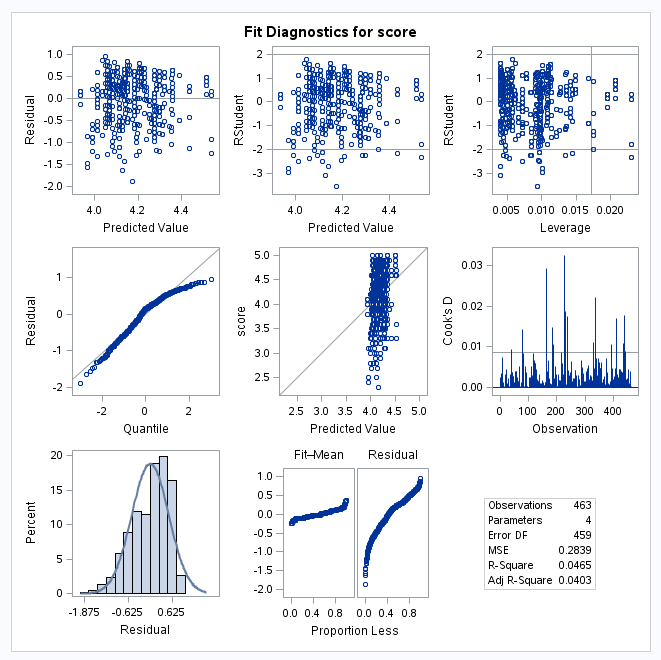
=β^0+β^1×bty\_avg

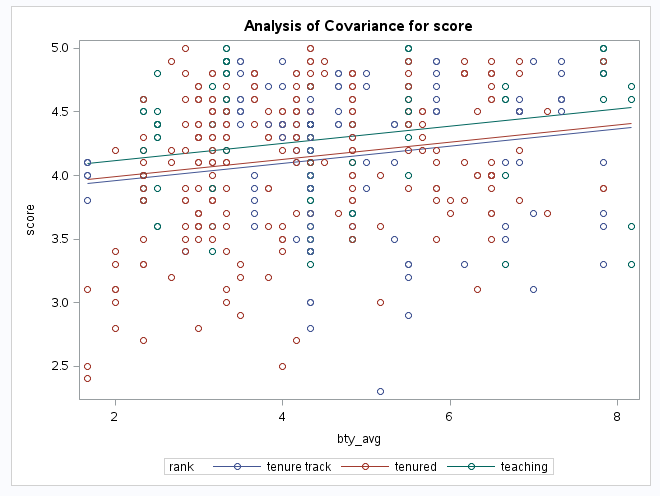
1. **Estimate a new model with gender removed and rank added in. How does SAS Appear to handle categorical variables that have more than two levels? Note that the rank variable has three levels: teaching, tenure track, tenured**

SAS simply codes the category that comes last alphabetically as the reference category.

proc glm data=evals plots=diagnostics;

class rank / ref = first;  
 model score=bty\_avg rank / solution;  
run;



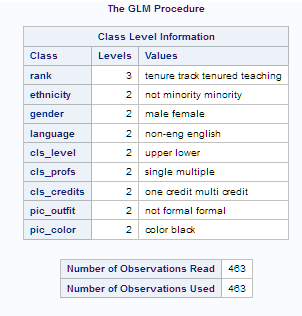
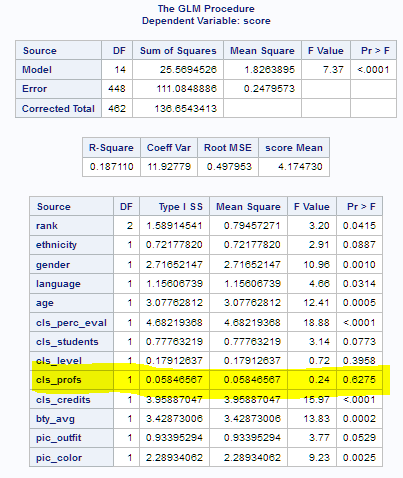
1. **Which variable would you expect to have the highest p-value in this model? Why? Hint: Think about which variable would you expect to not have any association with the professor score.**

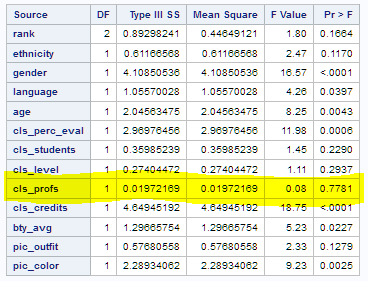
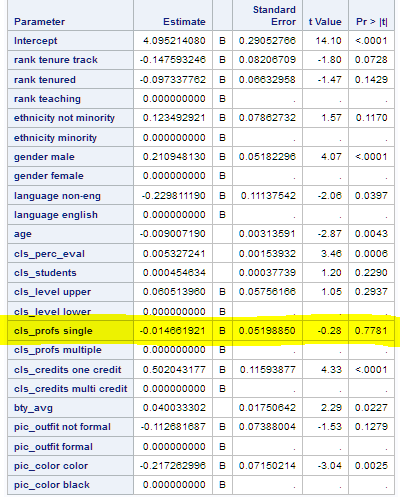
I believe “cls\_profs” variable will have the hightest p-value due to having least association with professor evalutation score.

1. **Check your suspicions from the previous exercise. Include the model output in your response**

I was right the “cls\_prof” variable had the highest p-value

proc glm data=evals;  
 class rank ethnicity gender language cls\_level cls\_profs  
 cls\_credits pic\_outfit pic\_color / ref=first;  
 model score=rank ethnicity gender language age cls\_perc\_eval   
 cls\_students cls\_level cls\_profs cls\_credits bty\_avg   
 pic\_outfit pic\_color / solution;  
run;

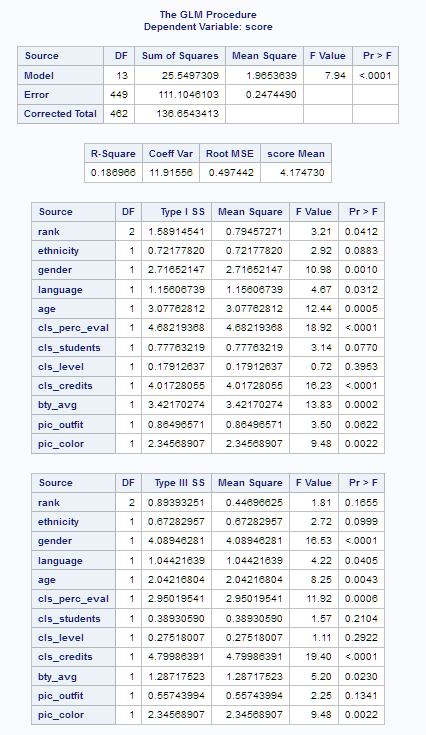
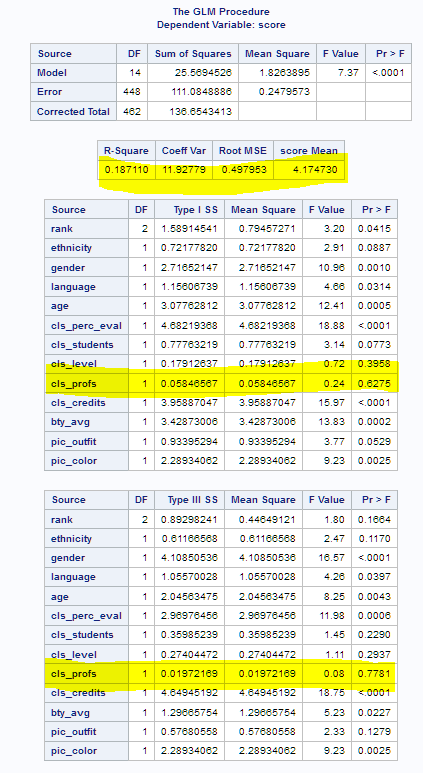
1. **Interpret the coefficient associated with the ethnicity variable.**

If your looking at pvalue < alpha as being significate and pvalues > alpha as not being significate, you would remove ethnicity variable because it’s p -value(s) are greater than alpha thus reducing the overall significane for this data model.

1. **Drop the variable with the highest p-value and re-fit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficient estimates depend on the other variables that are included in the model.) If not, what does this say about whether the dropped variable was collinear with the other explanatory variables**

After removing the highest p-value variable the adjusted RSquare value actually went down, it shows that it was not collinear value and cold be used within the data model or it just means we need to remove additional variables to create a better model.

proc glm data=evals;  
 class rank ethnicity gender language cls\_level cls\_credits pic\_outfit pic\_color / ref=first;  
 model score=rank ethnicity gender language age cls\_perc\_eval   
 cls\_students cls\_level cls\_credits bty\_avg   
 pic\_outfit pic\_color / solution;  
run;  
proc glm data=evals;  
 class rank ethnicity gender language cls\_level cls\_profs  
 cls\_credits pic\_outfit pic\_color / ref=first;  
 model score=rank ethnicity gender language age cls\_perc\_eval   
 cls\_students cls\_level cls\_profs cls\_credits bty\_avg   
 pic\_outfit pic\_color / solution;  
run;

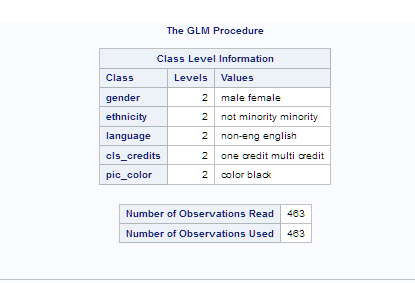
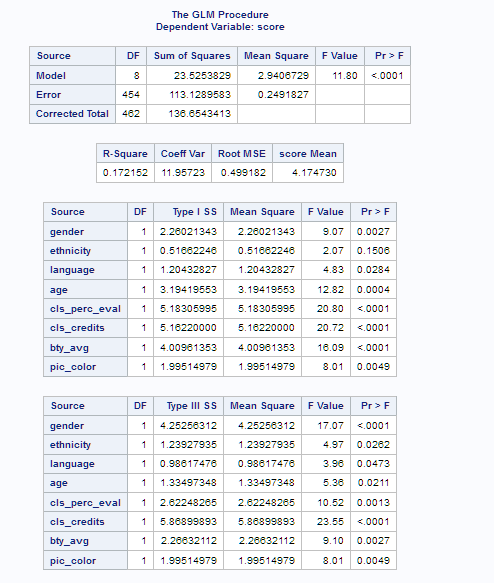
 

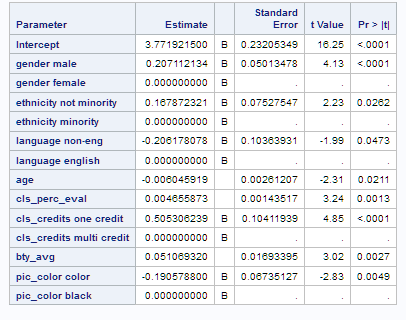
1. **Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write out the linear model for predicting score based on the final model you settle on.**

scoreˆ=β^0+β^1×ethnicity\_not\_minority+β^2×gender\_male+β^3×language\_non−englist+β^4×age+β^5+×class\_perceval+β^6×class\_credits\_one+β^7×bty\_avg+β^8×picture\_color\_colored

proc glmselect data=evals;  
 class rank ethnicity language cls\_level cls\_profs  
 cls\_credits pic\_outfit pic\_color / ref=first;  
 model score=rank ethnicity language age cls\_perc\_eval   
 cls\_students cls\_level cls\_profs cls\_credits bty\_avg   
 pic\_outfit pic\_color / selection=backward;  
run;

proc glm data=evals;  
 class gender ethnicity language cls\_credits   
 pic\_color / ref=first;  
 model score= gender ethnicity language age cls\_perc\_eval   
 cls\_credits bty\_avg pic\_color / solution;  
 run;

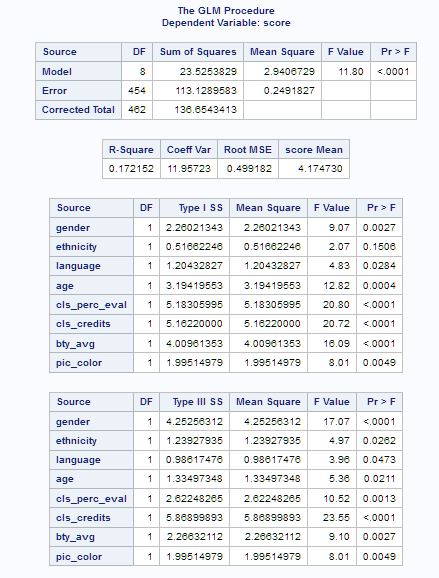
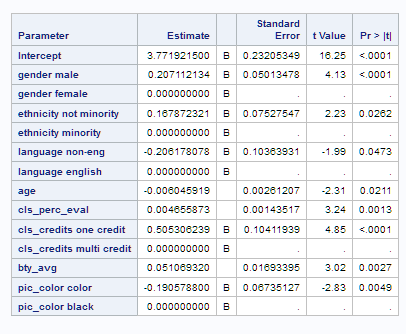


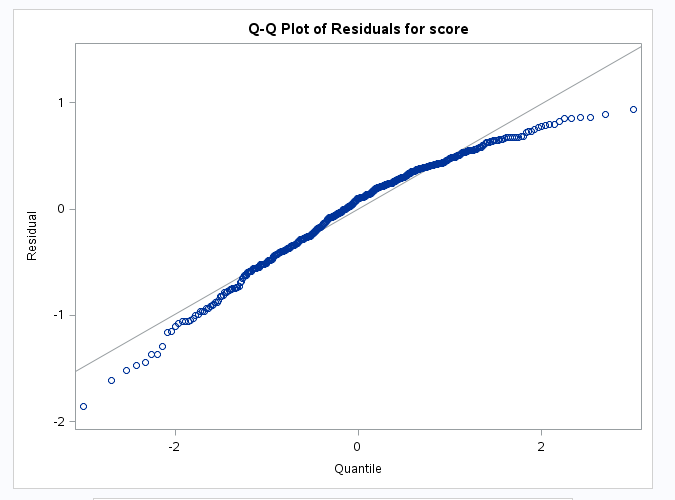
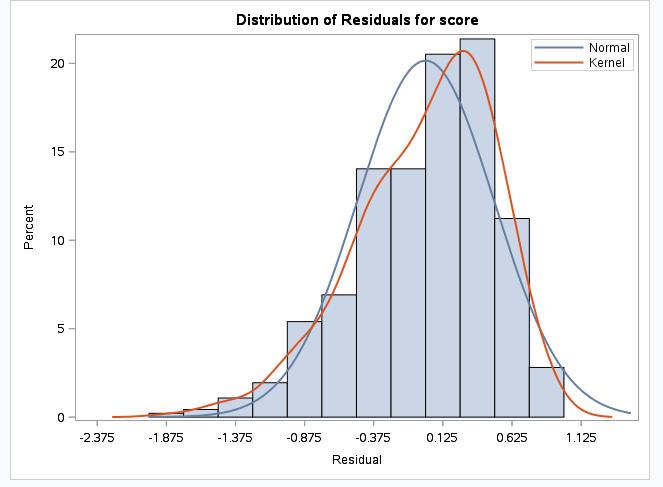
1. **Verify that the conditions for this model are reasonable using diagnostic plots.**

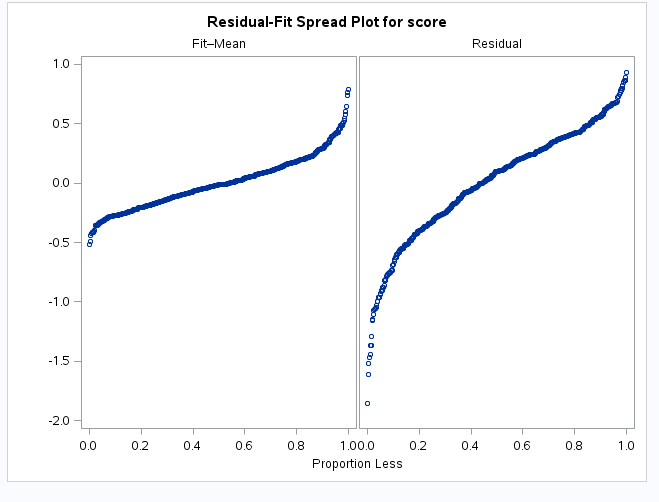
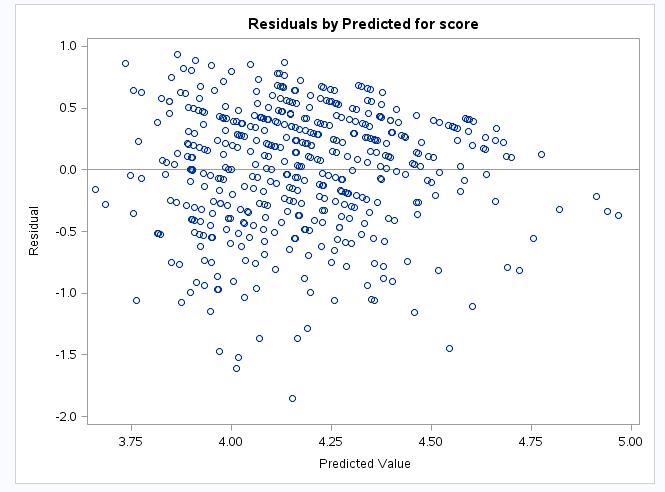
Normality – data has a bell shaped curve with left skewness but data is normally distributed

Linearity – residual plots follow the linear model with upward trend and is slightly curved do to outliers.

constant variability has passed with well formed spread of data with no formalized shape

1. **The original paper describes how these data were gathered by taking a sample of professors from the University of Texas at Austin and including all courses that they have taught. Considering that each row represents a course, could this new information have an impact on any of the conditions of linear regression?**

No, because classes and are independent of each other so evaluations scores would not be impacted. More data may improve the study.

1. **Based on your final model, describe the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score.**

Teacher should have the following characteristics: Male, English as primary language, teach a one credit course, class photo (B&W), age (33 – 60), his class should have high participation in teacher evaluation.

1. **Would you be comfortable generalizing your conclusions to apply to professors generally (at any university)? Why or why not?**

no, I would like to perform similar studies at multiple university with different demographics to have a solid population sample