Examining the drivers of catwalk models' salary

Group 13: Joby George, Charles Klein,

David Smith, David Zhang

Professor V

STAT 3200, Section 2

Due: October 19, 2017

Pledge:

Introduction:[1] **Predicting salaries of catwalk models**

A fashion student aims to learn more about the factors that drive catwalk model salaries and acquired data from 231 models to complete an analysis. To assist the student to meet this goal, multiple linear and nonlinear regression can be used to determine the relationships between salary and each collected variable. The end goal of this process is to learn more about the relationship between catwalk models' salaries and our variables of interest.

Data Summary: Understanding the variables and their relationships with salary

Data was collected on three explanatory variables that the student believed to affect salary: Age, Years, and Beauty.

Variable Name	**Variable Description**
Age	Age (in years) of the model
Years	Years worked as a model
Beauty	A percentage (0-100) received from a panel of experts rating the physical attractiveness of the model

To better understand the relationship between these variables and model salary, three scatter plots were generated, each with an explanatory variable on the x-axis and salary on the y-axis. Additionally, quadratic and interaction variables (Age^2, $Years^2$, $Beauty^2$, Age*Years, Age*Beauty, Years*Beauty) were created to test whether interaction variables better modeled salary, or if the variables had quadratic relationships with salary. Scatter plots for all variables vs salary are available in **Appendix 1**. None of these plots showed a clear relationship between the explanatory variable and salary, linear or otherwise, as many observations are grouped together in the lower salary values. Next, the three original explanatory variables were checked for

multicollinearity. This was done by creating a matrix of scatter plots of each explanatory variable and determining which had high correlation coefficients (**Appendix 1**), as well as calculating each variable's Variance Inflation Factor (VIF). Age and years, intuitively, demonstrated high multicollinearity, with the correlation coefficient between the two equaling 0.9547. The VIF of years was found to be 12.141, and the VIF of age was 12.637. Age was removed from the variables to analyze, as it had a higher VIF, and experience modelling intuitively seems to be the stronger driver. After removing the variable age, the VIF of years and beauty was recalculated, and both were insignificant, about 1.031 each. The remaining variables to analyze salary were: years, beauty, years^2, beauty^2, and the interaction variable, years*beauty. The hypothesized model based off these drivers is salary = β_0 + β_1*years + β_2*beauty + β_3*years^2 + β_4*beauty^2 + β_5*years*beauty + ε. Checking for outliers with this model, 25 observations were outliers in the x direction (**see Appendix 2)**. These points had leverage values, a statistic that measures how far an x value is from the other x values in the dataset, greater than the cutoff point, 0.052. Additionally, 12 observations were outliers in the y direction (**see Appendix 3)**, because their studentized residuals are greater than 2 in absolute value. This alarmingly high number of outliers suggests caution in data analysis and interpretation.

Analysis: Predicting Model Salary

The preliminary model is as such: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1)^2 + \beta_4 (x_2)^2 + \beta_5 x_1 x_2 + \varepsilon$. In this equation, y represents salary, x_1 represents years and x_2 represents beauty. The analysis provided by SAS show the adjusted R^2 to be 0.132, the standard error to be 149.31 and the overall model F-test's p-value to be less than 0.0001. This indicates that the model, while statistically significant, has a low predictive power, for only 15% of the variation in salary can be explained

by the variation in our explanatory variables. In addition, the model has standard error of 149.31**,** which indicates high variation in the model's predictions. To improve upon the model iterative-based and criteria-based selections were used, and models were compared. Based on the forward regression analysis, the best model was salary = $\beta_{0\ +}\ \beta_{1*}$year + β_{2*} $(year)^2 + \varepsilon$. Based on the stepwise regression, the best model was salary = $\beta_0 + \beta_1(year)^2 + \varepsilon$. However, due to the restriction that higher order variables may not be used without their corresponding lower terms (in this case, the quadratic term without its linear term), we reject the stepwise regression's. Using criteria-based selection, the model with the highest adjusted r-squared was salary = $\beta_{0\ +}$ β_{1*}year + β_{2*} $(year)^2 + \varepsilon$, while the model with the lowest Mallow's C(p) was salary = $\beta_1(year)^2 + \varepsilon$ (rejected for the same reason as the stepwise selection), with salary = $\beta_0 + \beta_{1*}(year)^2 + \varepsilon$ being the second-best. Salary =$\beta_{0\ +}\ \beta_1$ year + $(year)^2 + \varepsilon$ appeared to have a higher standard error than other models that ranked well, but it had shorter prediction intervals, due to the lesser number of predictors, which is desirable. Thus, the chosen "best" linear model was salary = $\beta_0 + \beta_1$ year + $\beta_2(year)^2 + \varepsilon$, which had an adjusted R^2 of 0.135 and standard error of 149.1. Although is not statistically significant, due to the restriction, it must be present.

After examining the regression assumptions, it was evident that the model violated the error normality and constant variance assumptions, as can be seen in **Appendix 4**. The Q-Q plot is very curved, violating normality, and the residual plots show clear fanning out, showing a violation of the constant variance assumption. This indicates that the relationship between y and x_1 is not modeled well with linear regression.

In an attempt to fix the issues with the assumptions, the y variable was transformed by applying the square root and natural log functions. As can be seen in **Appendix 4**, the square root

transformation produced better results than the natural log transformation. The residuals became more randomly scattered, with a more "horizontal band" appearance, suggesting a more constant variance, and the QQ plot is straighter, indicating a greater normality. Thus, the better transformed model was $y^* = \beta_0 + \beta_1 x_1 + \beta_2(x_1)^2 + \varepsilon$, where $y^* = y^{0.5}$. However, this non-linear model produced an adjusted R^2 of 0.0990, ~0.04 less than that of the linear model, and a standard error of 6.1773, much less than that of the linear model. This nonlinear model also, while producing a more randomly scattered residual plot (**Appendix 4)**, did not produce an objectively "randomly scattered" residual plot; it still violates the constant variance assumption. Furthermore, while it has a straighter QQ plot, it did not produce an objectively "straight' QQ plot; it still violates the normality assumption. Therefore, the transformation is unnecessary and insignificant.

Finally, influential points were examined. There were no influential points according to Cook's distance (**Appendix 4**), since all of the values were less than 0.79, the value of the F-statistic. If a model was to be made using these predictors, it would be $y^* = 6.40192 - 0.48295x_1 + 0.18084(x_1)^2 + \varepsilon$, and its prediction equation would be $y_{hat} = 6.40192 - 0.48295x_1 + 0.18084(x_1)^2$

Conclusion: Age, years, and beauty are poor predictors of model salary

In conclusion, the analysis showed that when trying to predict model salary based on beauty, age, and years, using a multiple regression analysis, the models were not valid. Despite using both linear and nonlinear regression to predict salary, the regression assumptions of constant variance and normality were not met **(see Appendix 2 and 3**). These models are not adequate for predicting model salary.

Appendices:

Appendix 1: Exploratory data analysis, determining the relationship between independent and dependent variables.

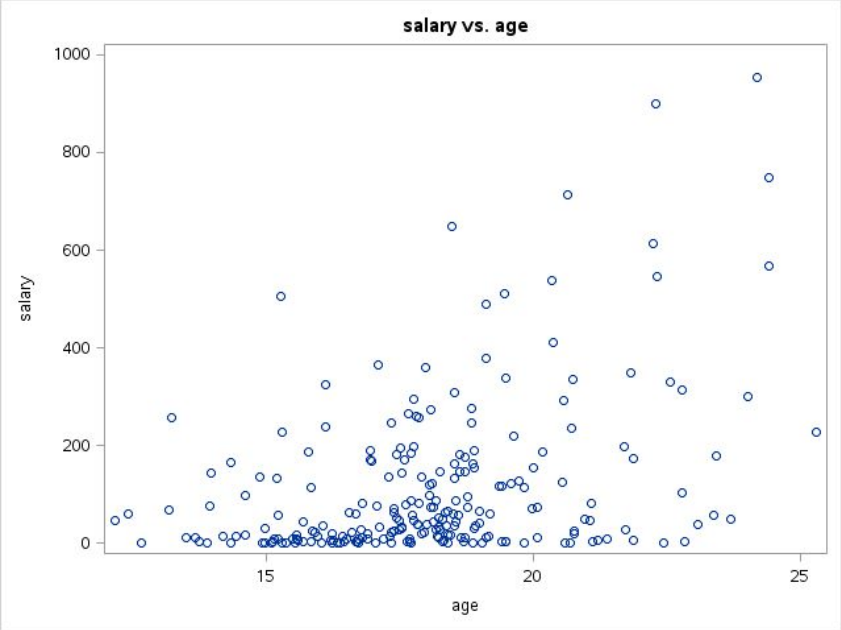

Figure 1: Scatterplot of salary vs age

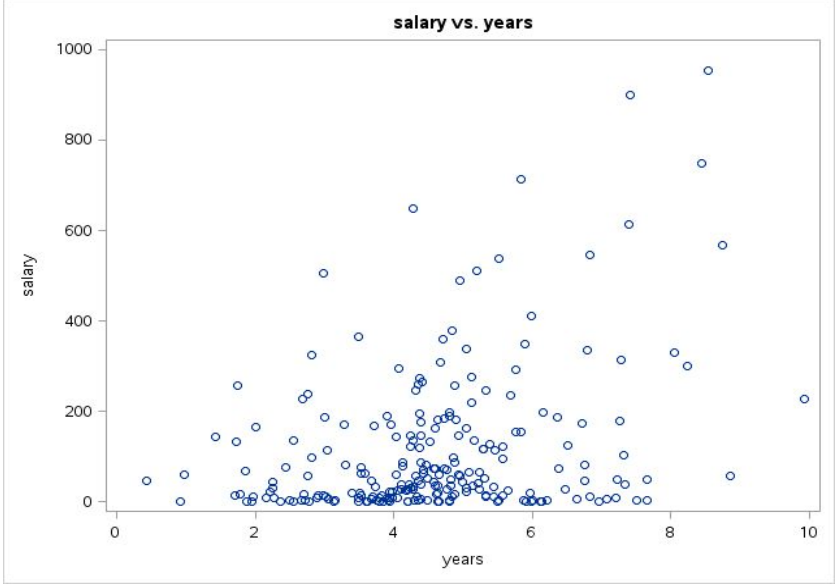

Figure 2: Scatterplot of salary vs years

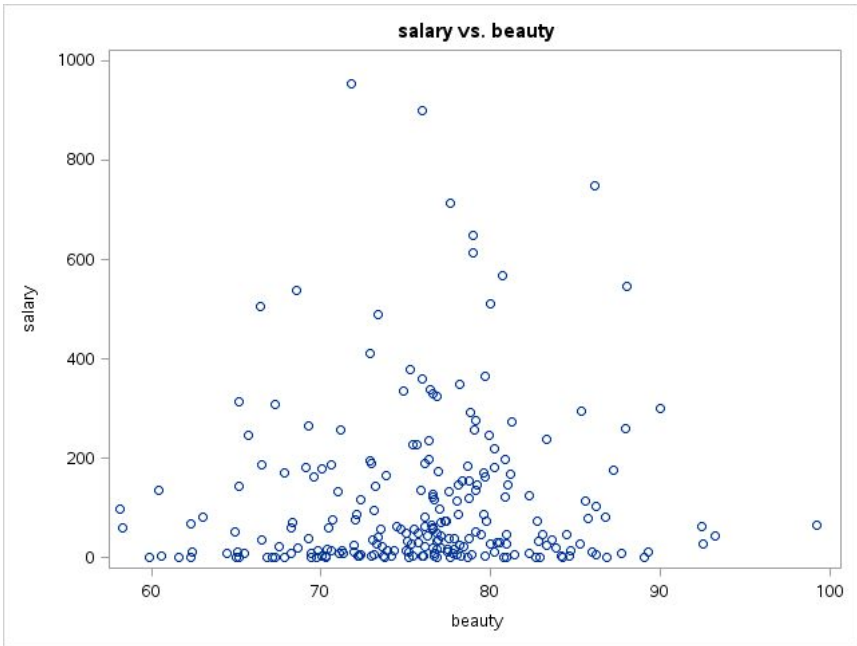

Figure 3: Scatterplot of salary vs beauty

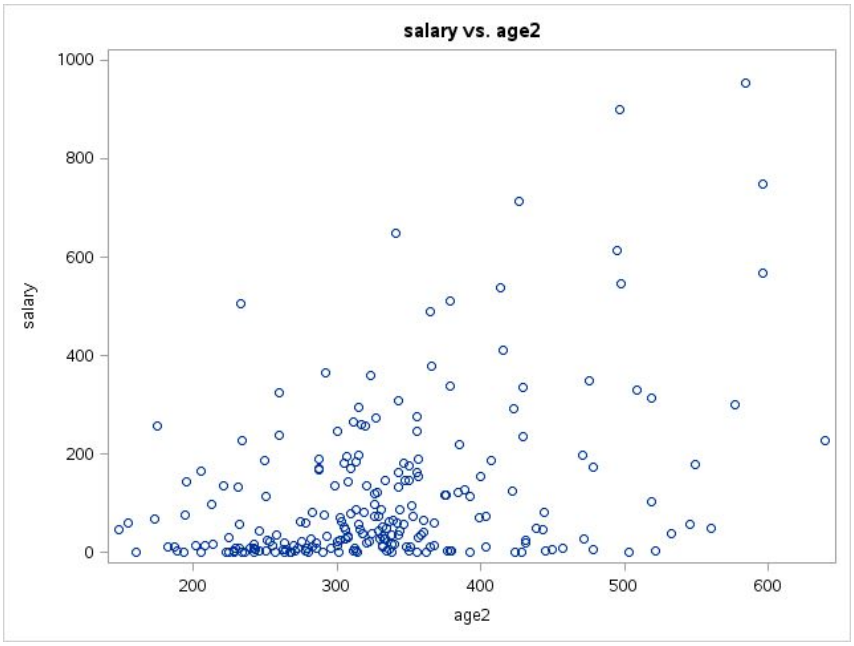

Figure 4: Scatterplot of salary vs age^2

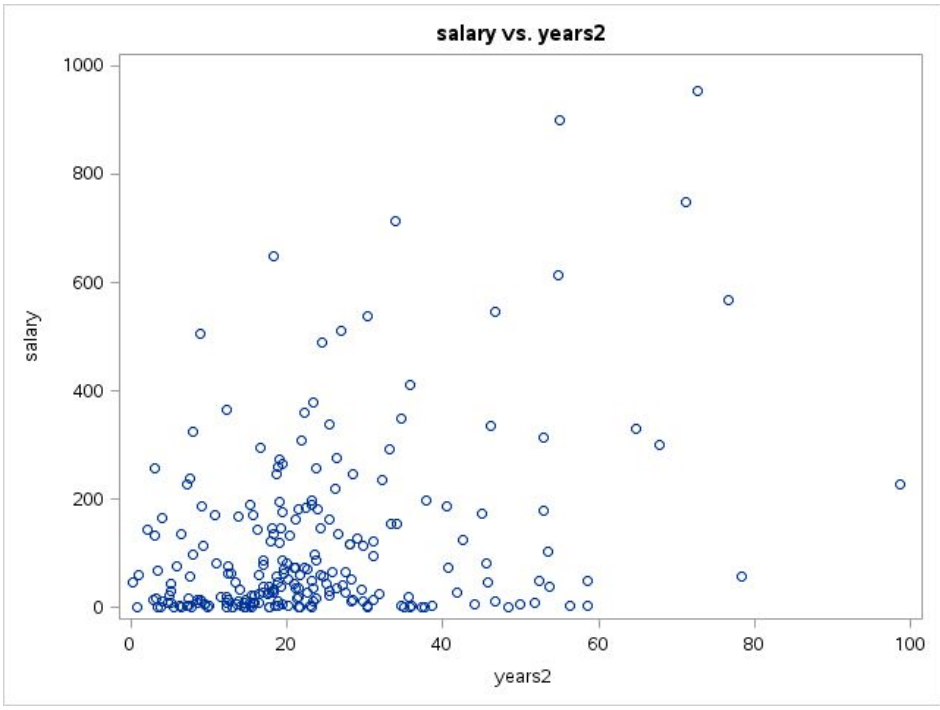

Figure 5: Scatterplot of salary vs years^2

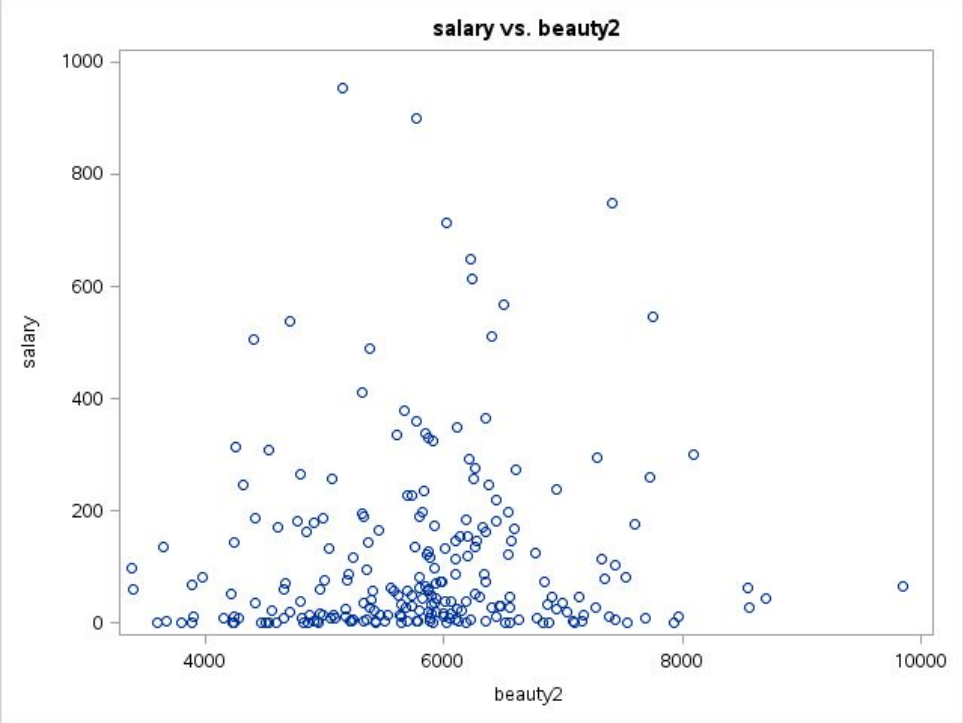

Figure 6: Scatterplot of salary vs beauty^2

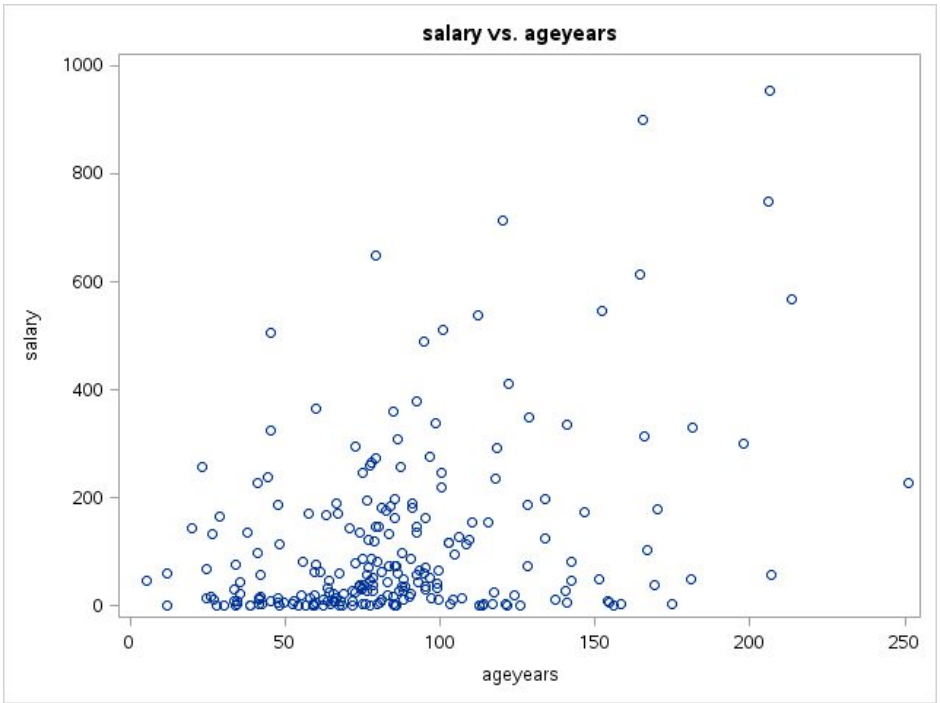

Figure 7: Scatterplot of salary vs age*years

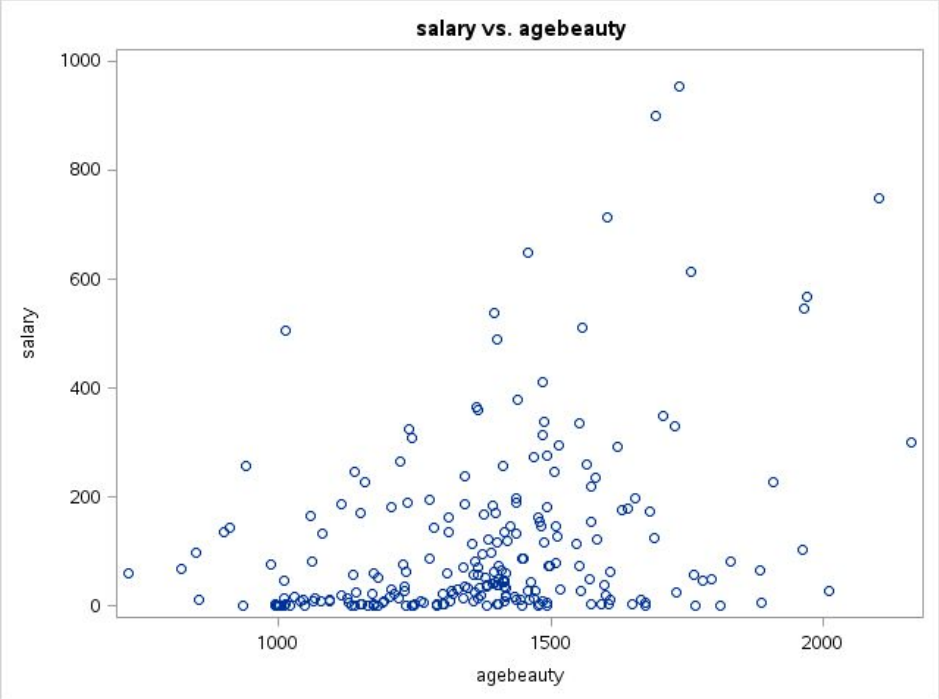

Figure 8: Scatterplot of salary vs age*beauty

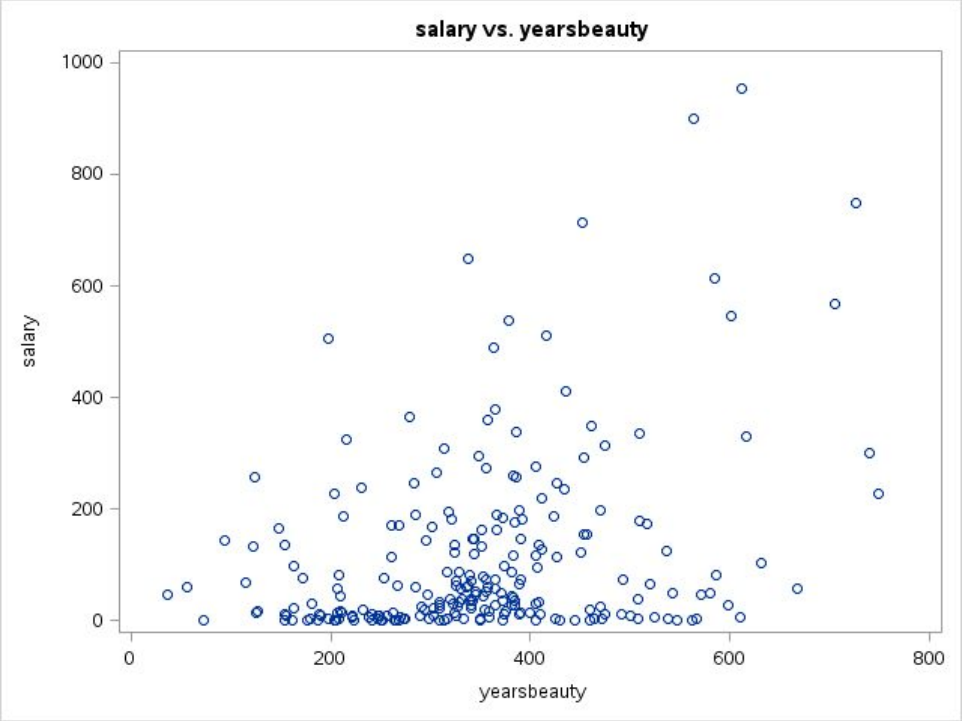

Figure 9: Scatterplot of salary vs years*beauty

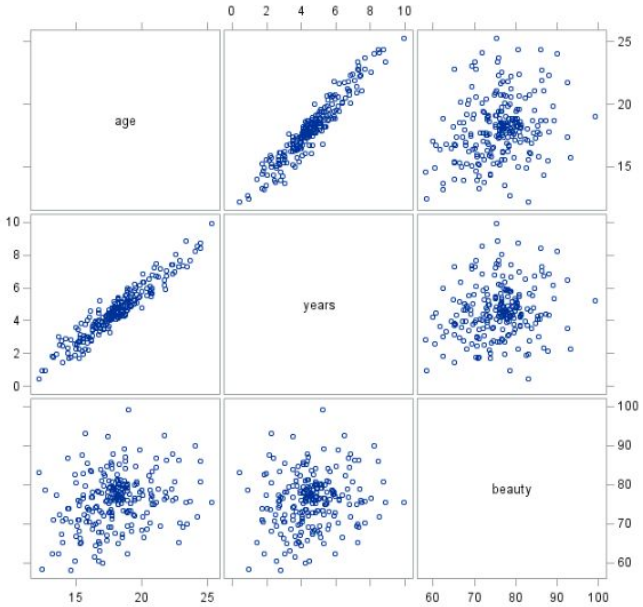

Figure 10: Testing Multicollinearity

Appendix 2: X Outlier analysis

Observation #
6
12
19
24
34
52
58
61
68
74
85
92
102
110
132
133
154
156
158
167

181
191
195
205
223

Appendix 3: Y Outlier Analysis

Y Outlier Analysis	
Observation	**Studentized Residual**
3	2.78
6	4.54
25	2.59
42	2.68
117	3.89
128	2.58
136	4.69
156	3.04
158	-2.10
171	2.36
192	3.03
199	3.95

Appendix 4: Residual and Outlier Analysis

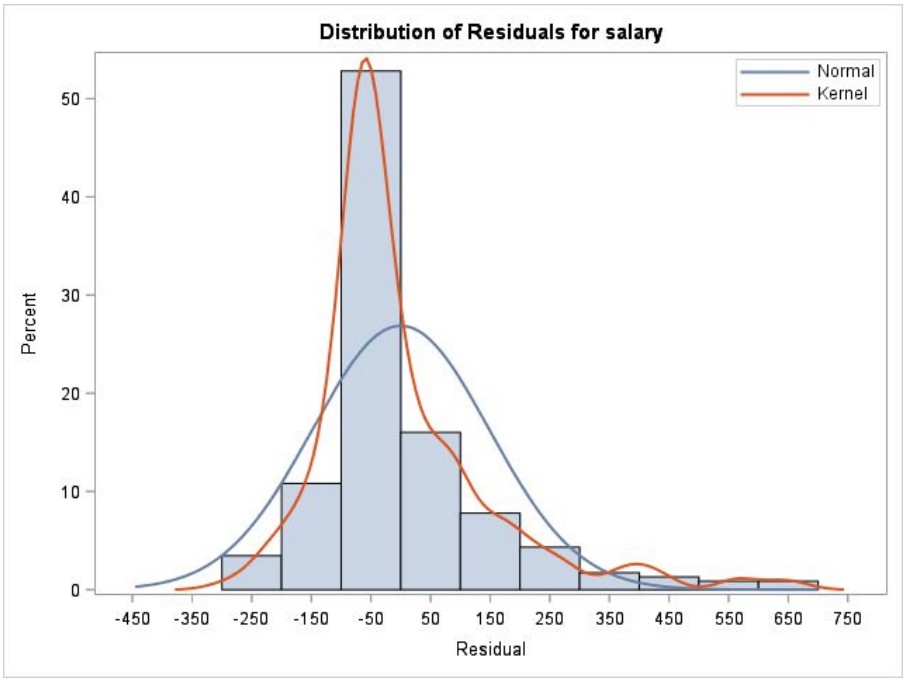

Figure 1: Distribution of residuals for best linear model, showing a skew right within the residuals

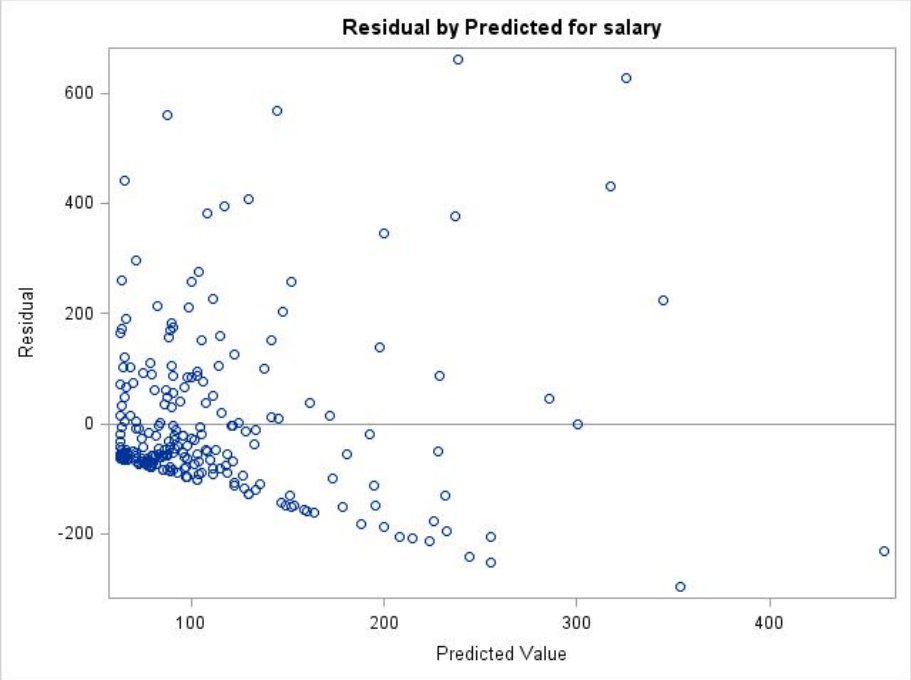

Figure 2: Residual versus predicted y (for best linear model) value showing violation of constant variance assumption

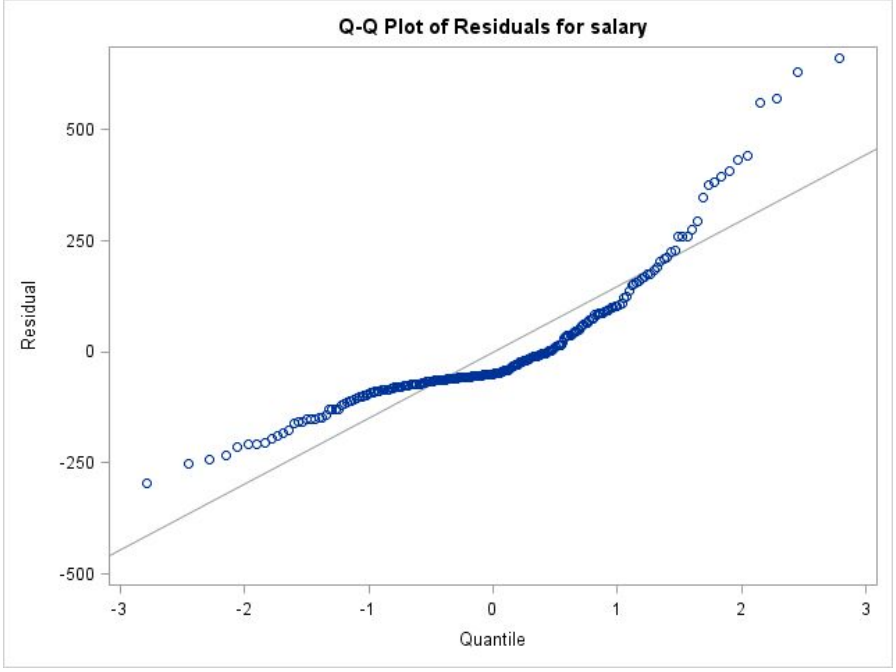

Figure 3: Q-Q plot of residuals (for best linear model) showing violation of normality assumption

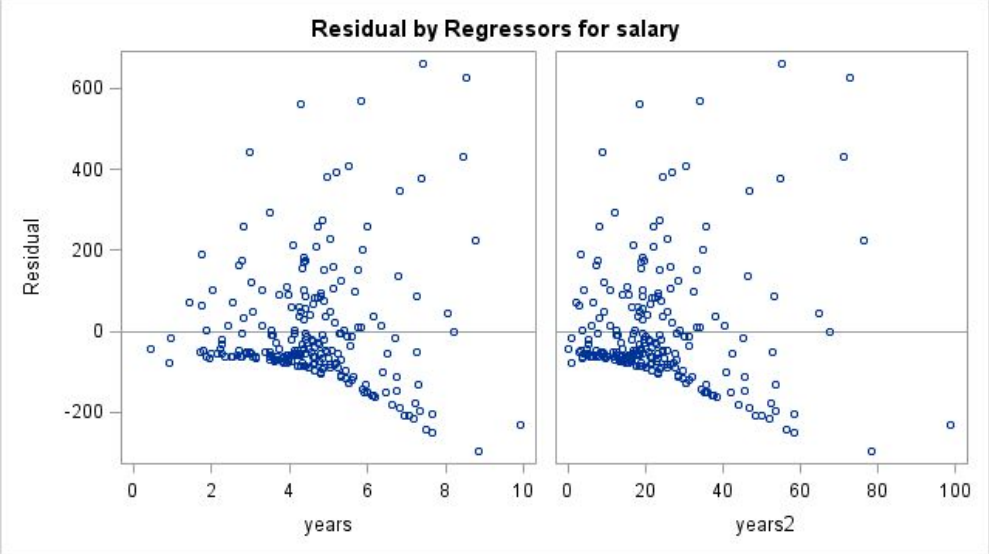

Figure 4: Scatterplot of residuals vs dependent variables, again showing violation of constant variance

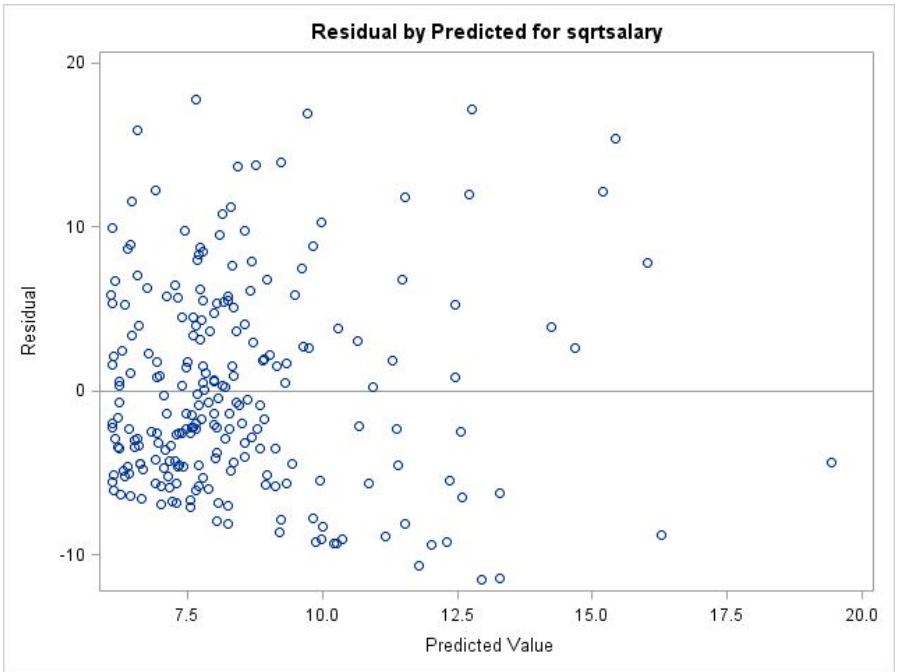

Figure 5: Residuals of sqrt(salary) versus predicted salary, showing violation of constant variance assumption

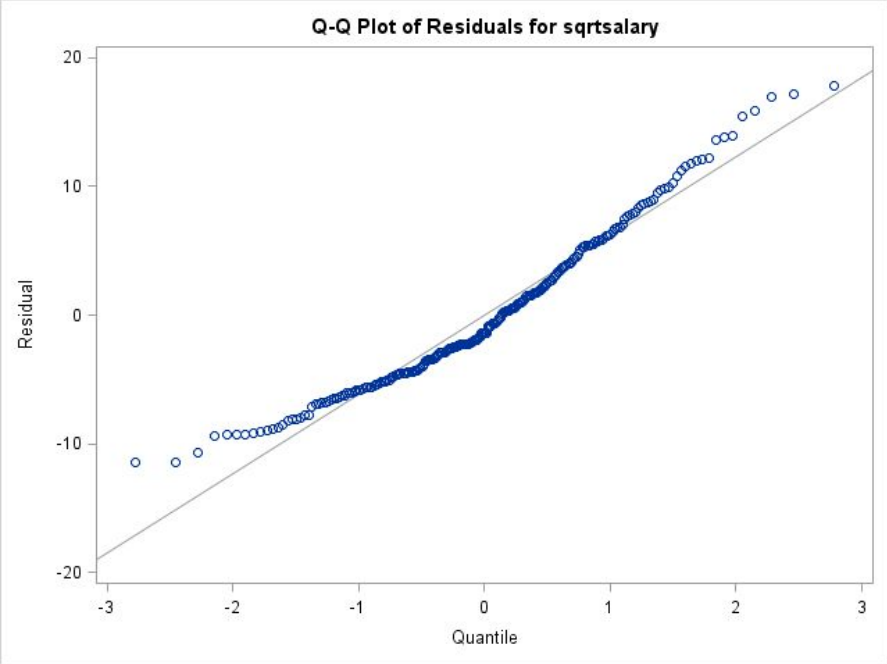

Figure 6: Q-Q plot of sqrt(salary) showing a violation of the normality assumption

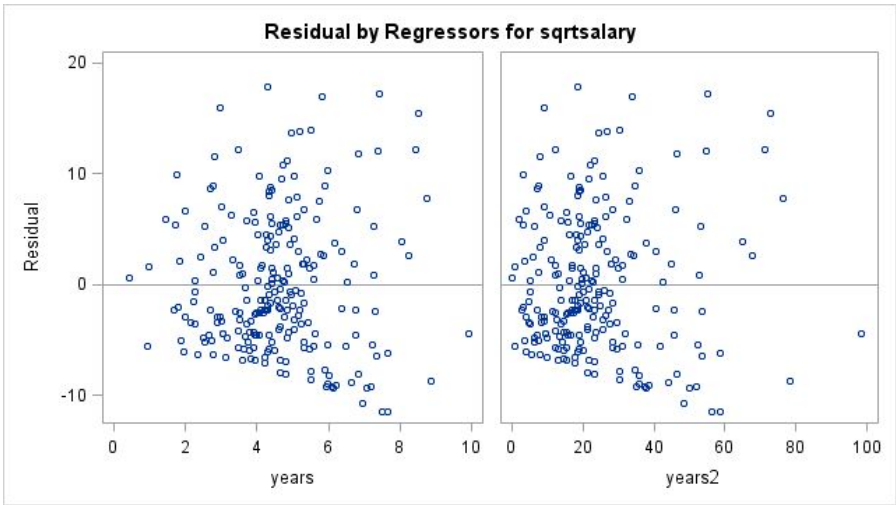

Figure 7: Residual plot of sqrt(salary) versus the x variables, showing the violation of constant variance assumption

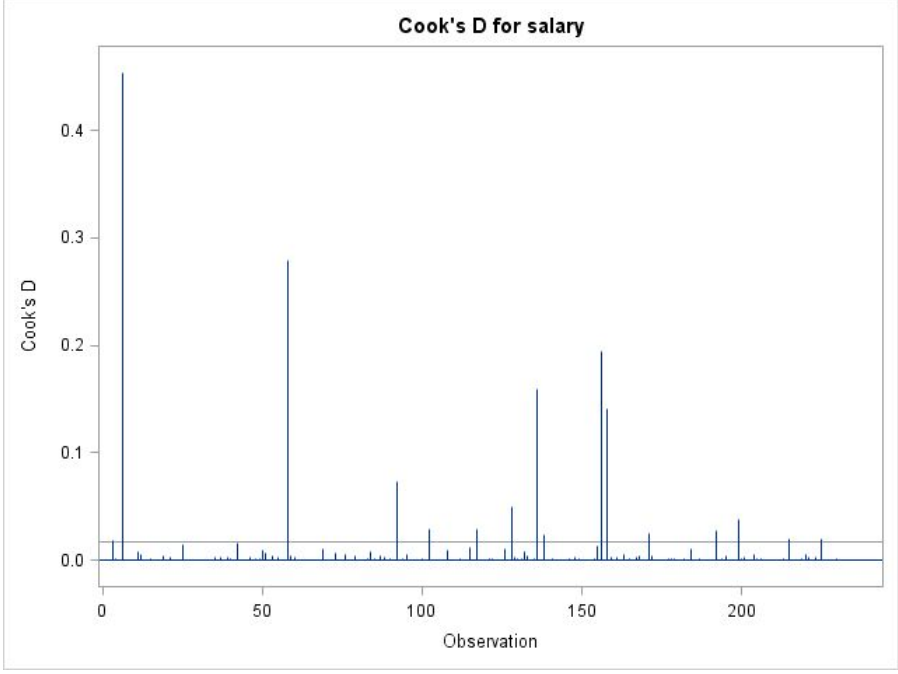

Figure 8: X Observations plotted against their Cook's D value, showing the presence of several x outliers

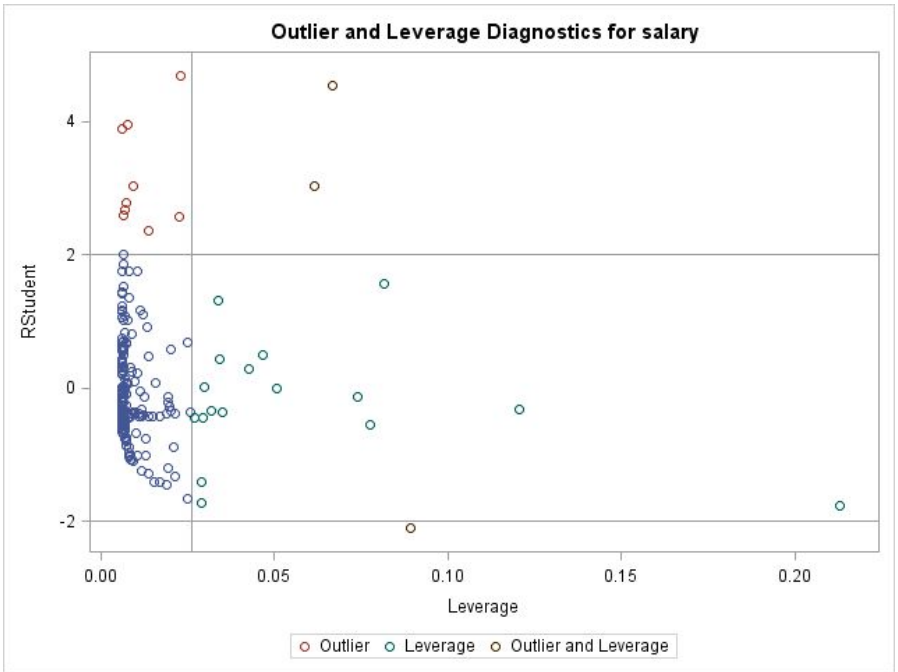

Figure 9: Outlier and Leverage diagnostics, showing influential points and outliers

Appendix 5: SAS Code used for the analysis

```
data fashion; /*creating a data set named “fashion” */
input salary age years beauty; /*inputting variables */
cards;
SALARY        AGE    YEARS          BEAUTY
3.70     16.67   3.15     78.25
...; /*data removed for printing simplification*/

proc sgplot data=fashion; /*creating scatter plots of salary vs. each ind. Variable */
scatter y=salary x=age;
title "salary vs. age";
run;

proc sgplot data=fashion; /*scatter plot*/
scatter y=salary x=years;
title "salary vs. years";
run;

proc sgplot data=fashion;
scatter y=salary x=beauty;
title "salary vs. beauty";
run;
```

```
data fashion2; /* new data set named “fashion2” for dummy variables*/
set fashion; /*using the same data as fashion*/
ageyears = age*years; /*interaction terms*/
agebeauty = age*beauty;
yearsbeauty = years*beauty;
age2 = age**2; /*quadradic terms*/
years2 = years**2;
beauty2 = beauty**2;
run;

proc sgplot data=fashion2; /*more scatterplots w/ quadratic and interaction terms */
scatter y=salary x=age2;
title "salary vs. age2";
Run;
proc sgplot data=fashion2;
scatter y=salary x=years2;
title "salary vs. years2";
Run;
proc sgplot data=fashion2;
scatter y=salary x=beauty2;
title "salary vs. beauty2";
Run;
proc sgplot data=fashion2;
scatter y=salary x=ageyears;
title "salary vs. ageyears";
Run;
proc sgplot data=fashion2;
scatter y=salary x=agebeauty;
title "salary vs. agebeauty";
Run;
proc sgplot data=fashion2;
scatter y=salary x=yearsbeauty;
title "salary vs. yearsbeauty";
run;
proc sgscatter data=fashion;
matrix age years beauty;
run;

proc corr data=fashion nosimple; /*multicollinearity plot*/
run;

proc reg data=fashion plots=none; /* obtains VIF of each variable*/
```

```
model salary = age years beauty / vif;
run;

proc reg data=fashion plots=none;
model salary = years beauty / vif;
run;

proc reg data=fashion2 plots(only)=(cooksd rstudentbypredicted /*testing for outliers/influential
points*/
rstudentbyleverage);
model salary = years beauty years2 beauty2 yearsbeauty/ r influence;
run;

data teststat; /*getting the t and f values to compare outliers with*/
y=(2*(5+1))/231;
t=quantile('T',0.975,231-5-2);
f=quantile('F',0.5,2+1,231-5-1);
run;

proc print data=teststat;
run;

proc reg data=fashion2 plots=none; /*selection by criteria*/
model salary = years beauty yearsbeauty years2 beauty2/ selection=adjrsq;
title "best equation";
run;

proc reg data=fashion2 plots=none;
model salary = years beauty yearsbeauty years2 beauty2/ selection=cp;
title "best equation";
run;

proc reg data=fashion2 plots=none; /*selection by iteration*/
model salary = years beauty yearsbeauty years2 beauty2/ selection=forward;
title "best equation";
run;

proc reg data=fashion2 plots=none;
model salary = years beauty yearsbeauty years2 beauty2/ selection=stepwise;
title "best equation";
run;

proc reg data=fashion2 plots=none; /*getting prediction intervals*/
```

```
model salary = years yearsbeauty years2 / cli clm;
title "salary vs. years ageyears years2";
run;

proc reg data=fashion2 plots=none;
model salary = years years2 / cli clm;
title "salary vs years years2";
run;

proc reg data=fashion2 plots(only)=(residualbypredicted /* checking for violations of
assumptions*/
residualplot qqplot residualhistogram);
model salary = years years2;
run;

proc reg data=fashion2 plots(only)=(cooksd rstudentbypredicted /*influential points*/
rstudentbyleverage);
model salary = years years2/ r influence;
Run;

data fashion3; /*new set for the transformation*/
 set fashion2;
 sqrtsalary=salary**0.5;
 lnsalary=log(salary);
run;

proc reg data=fashion3 plots(only)=(residualbypredicted residualplot qqplot /*checking violations
of assumptions for transformed models*/
  residualhistogram);
 model sqrtsalary = years years2;
 run;

proc reg data=fashion3 plots(only)=(residualbypredicted residualplot qqplot
  residualhistogram);
 model lnsalary = years years2;
 run;

proc reg data=fashion3 plots(only)=(cooksd rstudentbypredicted /*outlier/influential points */
rstudentbyleverage);
model salary = years years2/ r influence;
run;

data teststat2; /*getting t and f values to compare*/
```

```
y=(2*(2+1))/231;
t=quantile('T',0.975,231-2-2);
f=quantile('F',0.5,2+1,231-2-1);
run;
proc print data=teststat2;
run;

proc arima data=gold6;
identify var=logprice crosscor=sasdate noprint;
estimate input=sasdate p=(1);
outlier alpha=0.05;
run;
```