**Final Project-Regression Xue Yu (2377173)**

**Q1**

We want to assess the influence of the anthropometric and fitness variables on a reaction time measurement taken on the applicants. The data that we have are listed below.

Define:

: Reaction time to a Visual Stimulus (in sec.);

: Height to Applicant (cm.);

: Weight to Applicant (kg.);

: Shoulder Width (cm.);

: Pelvic Width (cm.);

: Minimum Chest Circumference (cm.);

: Thigh Skinfold Thickness (mm.);

: Pulse Rate (count per min.);

: Diastolic Blood Pressure;

: Number of Chinups Applicant Can Complete;

: Maximum Breathing Capacity (liters);

: Pulse Rate after 5 Mins. Recovery from Treadmill Running;

: Maximum Treadmill Speed (Individually Set for Each Applicant);

: Treadmill Endurance Time: Maximum Time to Exhaust While Running Treadmill (min.);

: Total Body Fat Measurement.

***Establish Models***

* **Complete Straight Line Model including all variables**



1. Fit the model



Since there are no repeated data, we do overall test, assuming that

Vs. 

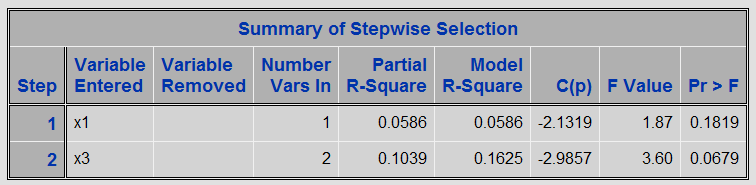
1. We use Stepwise Selection Method to do some tests and get results.

First, look at for different situations for different SLENTRY and SLSTAY.

|  |  |  |
| --- | --- | --- |
| Variables Enter and Stay | SLENTRY(SLE) | SLSTAY(SLS) |
| None | 0.05 | 0.05 |
| None | 0.1 | 0.05 |
| None | 0.1 | 0.1 |
| None | 0.15 | 0.15 |
| , | 0.2 | 0.2 |

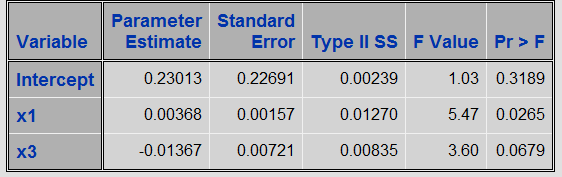
From table above, we find that setting SLE and SLS too strictly will lead to no variables enter and stay in model. This fact tells us that these variables may have little linear relationships with response factor, or we can say that these data we are given may have little meanings to predict reaction time through linear model. However, in theory, we need them, so for tests below, the predetermined SLE is 0.2, so is SLS.

Next, based on predetermined SLE=0.2 and SLS=0.2, we have



From stepwise procedure result,  and come into this model and stay. However, Model R-Square are not very large, which means at most about 16% of the variability in reaction time can be explained through this model. C (p) is fine in this model.

Then we get parameter estimators which are listed below.



Regression Function is

.

1. Conclusion

For all 14 variables, onlyand enter and stay in the model, so we eliminate other variables and just keep and to establish a new model called refined straight line model and focus on these two variables.

* **Refined Straight Line Model**

* + - **Model 1-and **



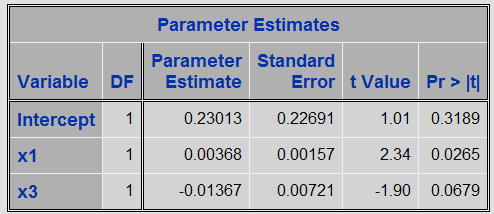
1. Fit the model



Since there are no repeated data, we do overall test, assuming that

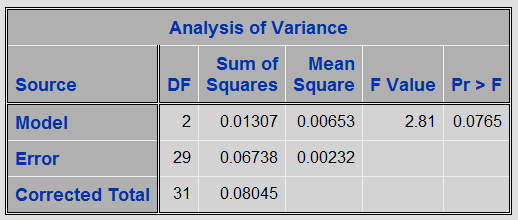
Vs. 

1. We do some tests and get results.

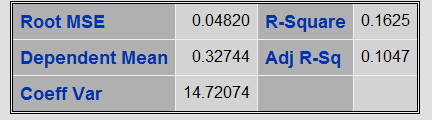


Regression function is





Since Pr>F is 0.0765, which is <0.1, we don’t rejectand think that the overall regression is significant, for both and, value is <0.1, and are significant or we can say that height and shoulder width are important to predict reaction time. Then check Adj R-Square.



Since Adj R-Square is 0.1047, which is a very small number, then we do some changes to original model in order to improve model. Adding interaction term may help.

* + - **Model 2-,and**

****

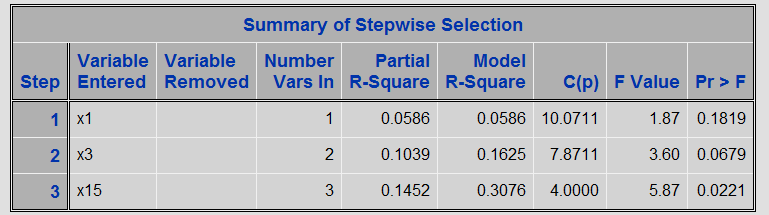
1. Fit the model

****

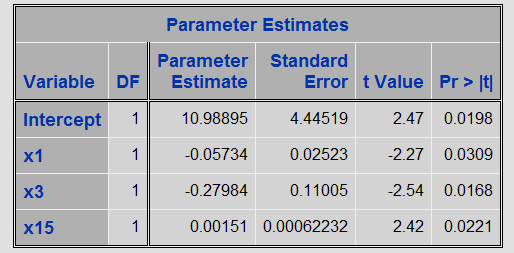
Since there are still no repeated data, we do overall test, assuming that

Vs. 

1. Based predetermined SLE=0.2, SLS=0.2, we get these 3 variable all enter and stay in the model.

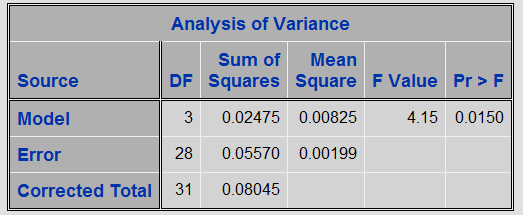


Then we get parameter estimators.



Regression function is





Since Pr>F is 0.0150, which is <0.05, we reject.

From the result above, we get Adj R-Square is 0.3076, which is greater than Adj R-Square in model 1. We can’t say that this model is better than model 1, we can just conclude that more variation in can be expressed through this model than that one. Interaction of height and shoulder width is useful to predict reaction time to a visual stimulus. However, model assumptions need to be checked.

We keep trying to improve model by adding second-order polynomial terms into model, so model 3 is established.

* + - **Model 3-,,,and**

****

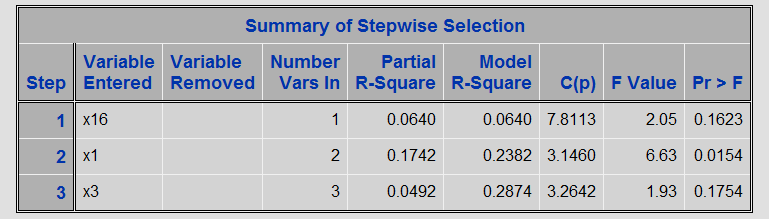
1. Fit the model

****

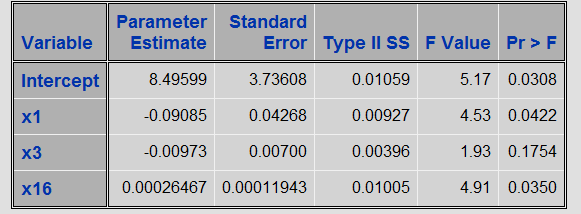
Since there are still no repeated data, we do overall test, assuming that

Vs. 

1. We still use stepwise selection method to tests and results are listed below.



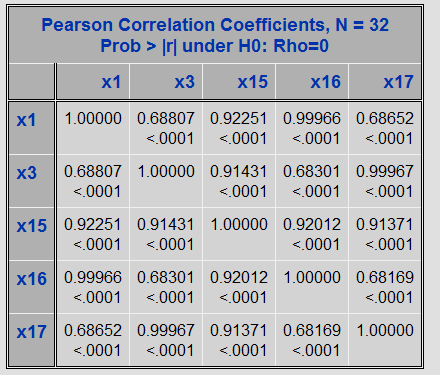
From the stepwise selection method, we get that, and enter and stay the model.



Regression function is

****

Interaction term is eliminated in this model, so we consider correlation matrix.



From correlation matrix, we find that interaction term is highly correlated with, so we just need to keep one term from these two in the model. We check model assumption before comparing model 2 with model 3.

***Check Models***

* + - **Model 2-,and**

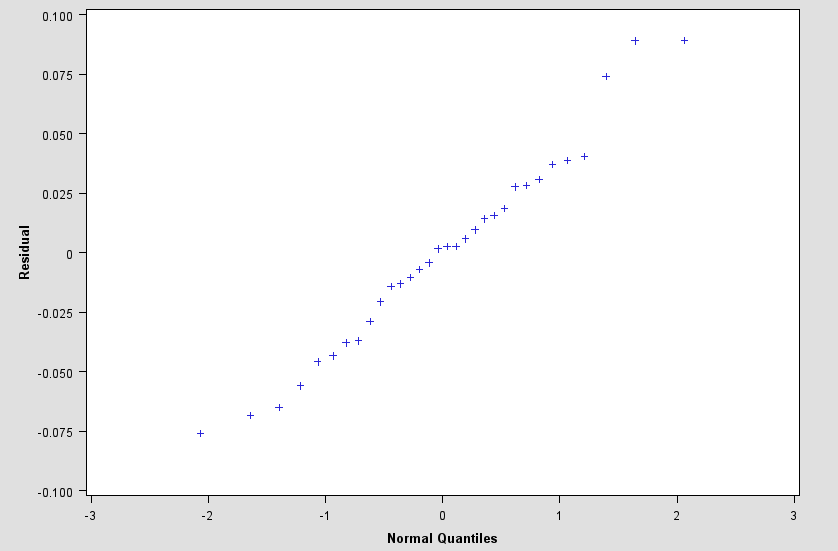
****

The Model Assumptions are

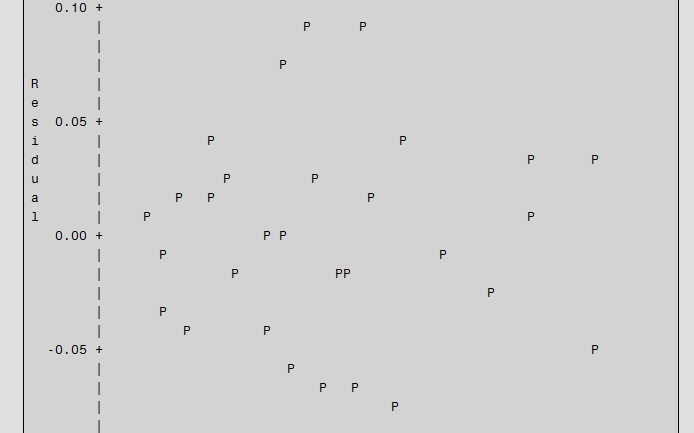
1. Linear model is sufficient;
2.  and all share the same , ;
3.  …are independent.

Firstly, we check the assumption (2); we divide that into two parts.

In order to check whetherfollow normal distribution, we use QQ-plot.

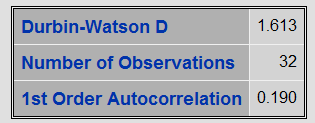


From the QQ-plot, we find that it’s approximately a straight line, which means that normality is satisfied.



From residual plot, all points all fall into interval, which means no obvious model defects. All share the same. All in all, assumption (2) is satisfied.

Then we check Assumption (3) .We use Durbin-Watson tests since no “time” factor is present.



We get =1.613, =32, =3.

Assuming that

: Vs. 

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 1.24 | 1.65 |
|  | 1.15 | 1.55 |
|  | 1.04 | 1.43 |

Since =1.613>1.43 and =2.387>1.43, we should not reject that: at significance level and thinkare not auto correlated. Because we have checked that normality and known that all errors follow normal distribution, we can conclude that  are independent at level .In other words, assumption (3) is satisfied.

The data collected are not repeated, we cannot use Lack of fit to check assumption (1).However, since we assume assumption (2) and (3) are satisfied, it’s common that assumption (1) is satisfied.

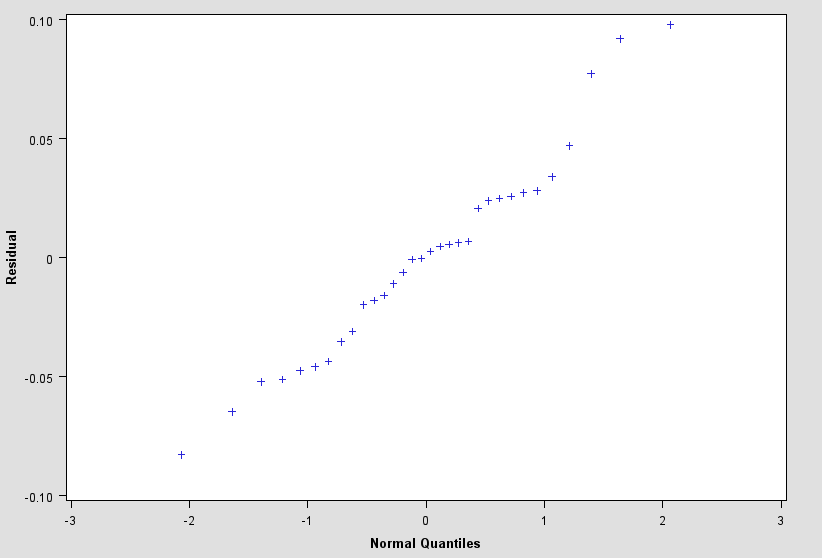
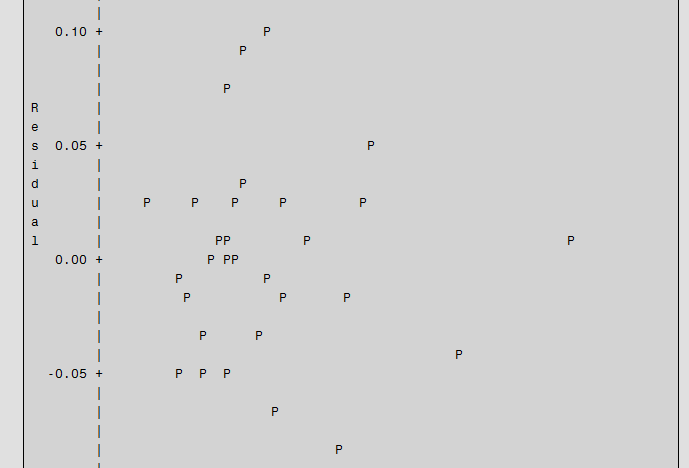
* + - **Model 3-,,,and**

****

The Model Assumptions are

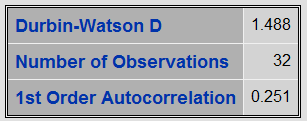
1. Linear model is sufficient;
2. and all share the same , ;
3. …are independent.

Firstly, we check the assumption (2). QQ-plot and residual plot are listed below.

From the QQ-plot and residual plot, according to what we use to analyze assumption conditions in model 2, we conclude that assumption (2) is satisfied.

Then we check Assumption (3) .We use Durbin-Watson tests since no “time” factor is present.



We get =1.488, =32, =5.

Assuming that

: Vs. 

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 1.11 | 1.82 |
|  | 1.02 | 1.71 |
|  | 0.92 | 1.60 |

Since ifor, reject at level; if and, do not reject  at level. However, =1.488 and =2.512, test is to be inconclusive.

We just can use simplified test to decide. Since =1.488<=1.82, we reject at level. We conclude that are auto correlated. In other words, for this model, assumption (3) is not satisfied.

In summary, model assumptions are not satisfied for model 3.

***Transform Models***

Then we do more transformations to response and variables and give a brief view.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Variables enter and stay  at significant level 0.2 | Adj R-Square | Normality Satisfied? |
|  | None |  |  |
|  | None |  |  |
|  |  | 0.0942 | Yes |
|  |  | 0.1096 | Yes |
|  |  | 0.0944 | Yes |
|  |  | 0.1072 | Yes |
| … |  |  |  |

Without considering other conditions, Adjusted R-Square gives us a criterion to give up these models above.

***Compare Models***

Now, we have 3 models, we compare them below:

|  |  |  |
| --- | --- | --- |
| Model | Adjusted R-Square | Model Assumptions |
| Model 1 | 0.1047 | Un-checked |
| Model 2 | 0.3076 | Yes |
| Model 3 | 0.2874 | No |

For 3 models, model 2 is relatively “better” than model 1 and model 3. Since it gets the largest Adjusted R-Square value and all the model assumptions are satisfied, model 3 contains more terms, but it doesn’t pass assumption check.

***Conclsion***

We get data about anthropometric and physical fitness measurements on police deparement applicants. We try to assess the influence of the anthropometric and fitness variables on a reaction time measurement taken on the applicants.

First, we include all 14 variables into model, after using stepwise selection method, we find that only height and shoulder width have obvious effects on reaction time comparing with other factors.

Then we focus on these two factors and do serious tests.

On one hand , we consider interaction of height and shoulder width, by adding this term into model, we improve model which just includes height and shoulder width.

We get model

.

However, only 30.76% variation in reaction time can be explained by this model , 70% remains unexplained.

What’s more ,this model just can helps us to know that there is some relationhship between a person’s height , shoulder width and his or her reaction time, other factors listed in this survey may have no effects. But, what’s the exect relationship? We don’t know, more research work need to be done to get a deeper sight.

On the other hand, in order to keep improving model, we consider adding second-order polynomial terms, but we get a similar adjusted R-Square value and model assumpations are not satisfied. This fact tells us that this new model is not adequate.

We also establish transform models of taking log of reaction time, reciprocal of reaction time,square root of hight or shoulder width and so on. Although these models are more or less express relationship between these 3 factors, a comparatively simper model just with interaction is preferred.

If we can get repeated data, lack of fit tests can be done to check adequacy of this model. More data collected will help us to find more appreciated relationship between reaction time and anthropometric and fitness factors.

Finally, one issue here we should pay attention to that is we set predeterminded significant level as 0.2, however, which may lead to that when we use stepwise selection method, it’s “too easy” for a variable to enter and stay, problem exists and we should find some methods to fix it.

**Q2**

We get data that was developed from a salary survey of computer professionals in a large corporation. We try to measure the effects of experience, education and management responsibility on salary using regression analysis.

Define:

: Salary;

: Experience, measured in years;

: Education, coded as 1 for completion of high school, 2 for B.S., and 3 for an advanced degree;

: MGT, coded as 1 for a person with management responsibility and 0 otherwise.

***Establish Models***

* **Straight line model without interaction terms**



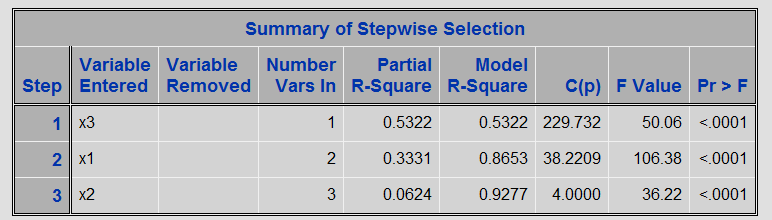
1. Fit the model



Since there are no repeated data, we do overall test, assuming that

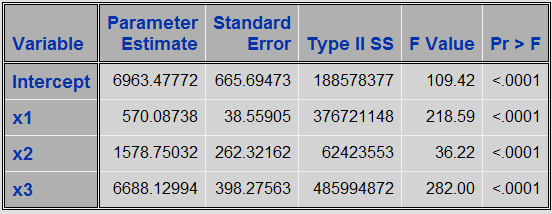
Vs. 

1. We use Stepwise Selection Method to do some tests and get results.



From table above, we find that all variables enter and stay in the model. All variables are significant. And C (p) is 4, which is not too big and closed to 3-the number of variables in the model.

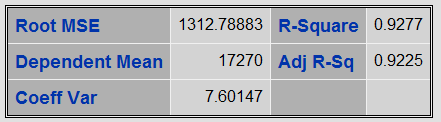
Then we get parameter estimators.



Regression function is

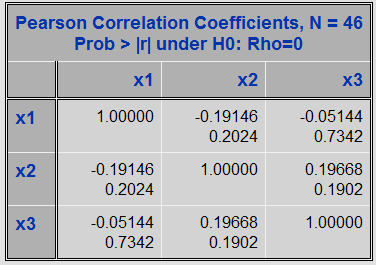


Check Adj-R-Square which is 0.9225.



This table tells us about 92.25% variation in salary can be expressed through experience, education and a person’s management responsibility.

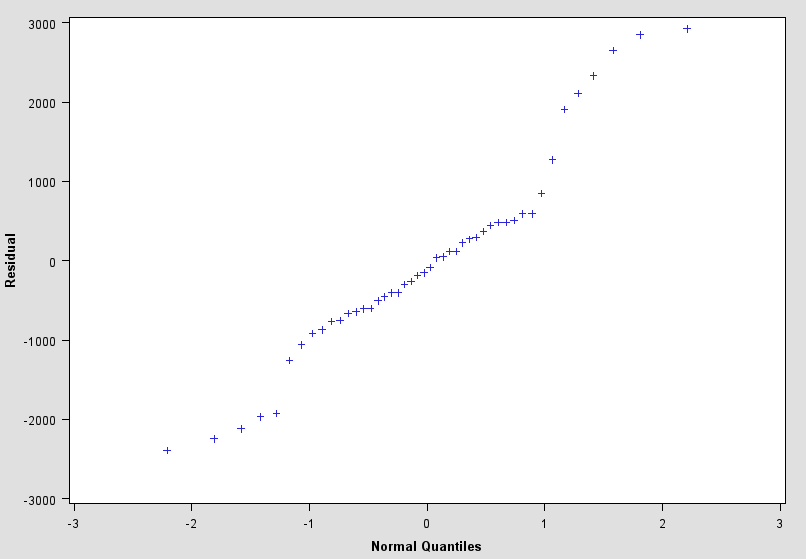
Now, we have 3 variables in the model, in order to check whether all these 3 are necessary, we obtain correlation matrix.



From table above, we can see that, andare not tightly correlated, so we will keep all variables in the model.

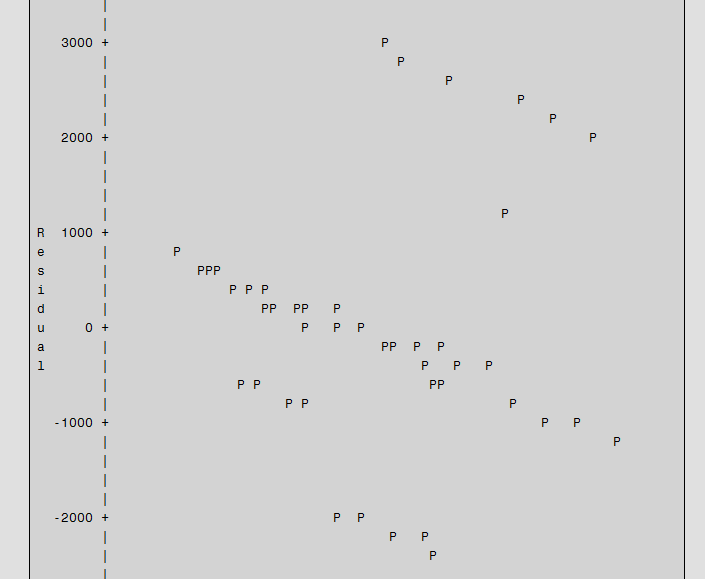
1. Simply check model assumptions.

QQ-plot is listed below.



Check Normality of model by using QQ-plot, since it’s closed to a straight line, we should think that normality is satisfied.

However, residual plot shows some pattern. Downward trend is guessed and we need statistical analysis of them.



1. Conclusion

In this model, we find that experience, education and management responsibility all have effects on salary and they are not tightly corrected with each other. All these 3 variables are necessary and should be included in the model.

From the results listed, we get Adj R-Square is 0.9225, which means the fitted model is statistically significant. However, about 8% of the variation remains unexplained; more work can be done on this problem. What’s more, residual plot shows some pattern which means problems about model assumptions may exist. We will add interaction terms in order to improve original model.

* **Straight line model with interaction terms (1)**



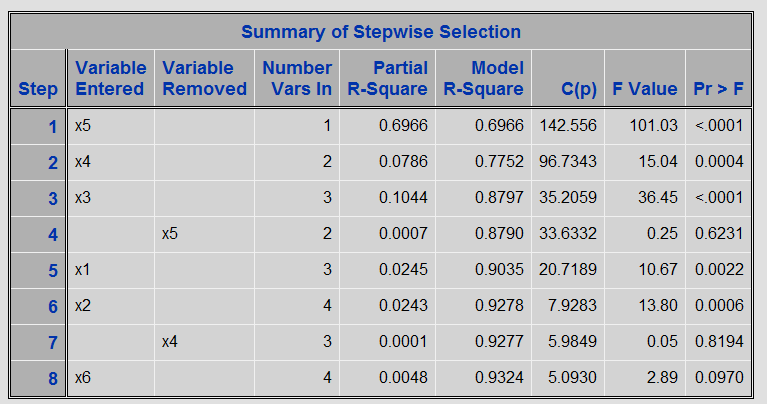
1. Fit the model



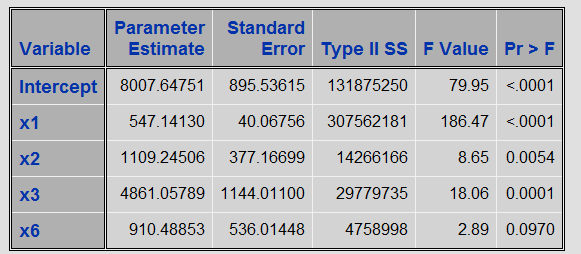
Assuming that

Vs. 

1. We use Stepwise Selection Method to do some tests and get results.



From table above, we find that, and terms finally stay in the model. Final C (p) is 5.0930, which is fine since it’s not too big and closed to 4.



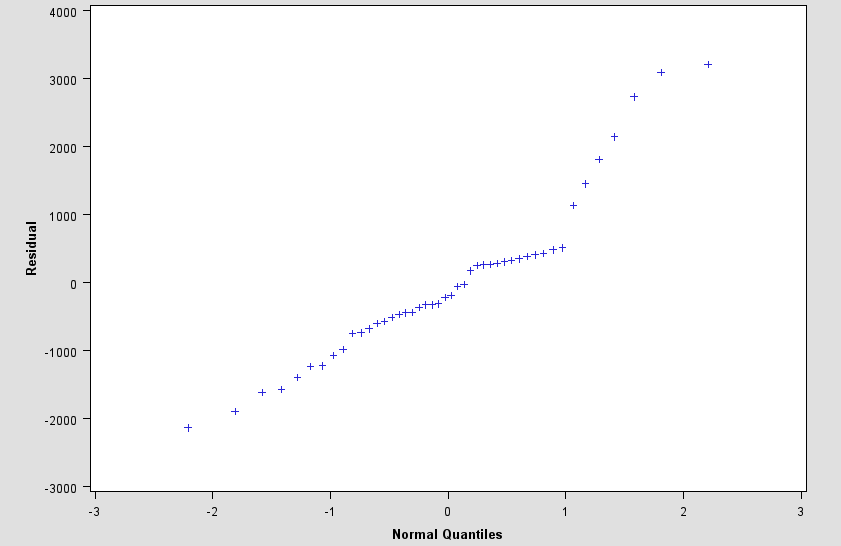
Regression function is



Get Adj-R-Square which is 0.9324 from summary of stepwise selection table.

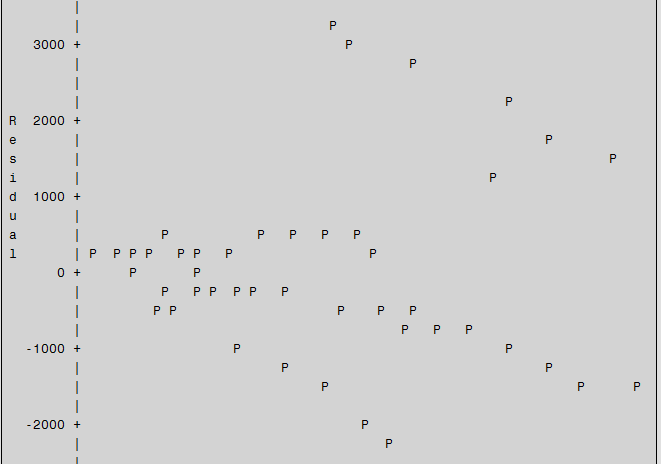
We see that although we add interaction terms in the model, Adj R-square is not improved a lot.

1. Simply check model assumptions.



Check Normality of model by using QQ-plot, since it’s closed to a straight line, we should think normality is satisfied.

However, residual plot still shows some pattern.



1. Conclusion

In this model, we add interaction terms into model, we find that interaction of education and MGT plays a more important role than other 2 interaction terms. However, Adj R-Square is not improved a lot.

* **Straight line model with interaction terms (2)**



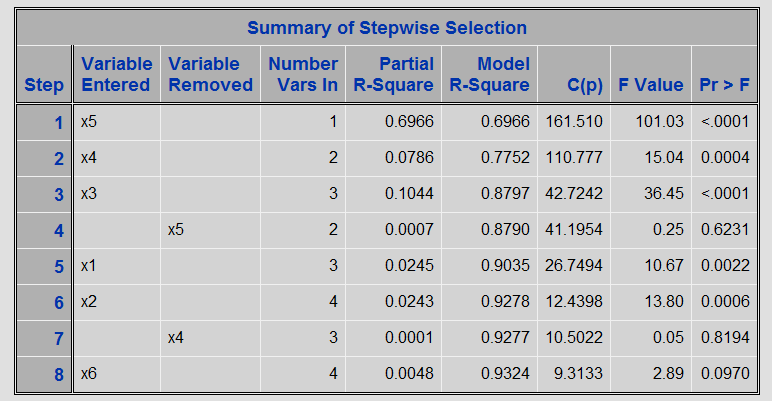
1. Fit the model



Assuming that

Vs. 

1. We use Stepwise Selection method to do some tests and get results.



From table above, we find that term does not enter this model at predetermined level, and final C (p) is 9.3133, which is too big.

1. Conclusion

In this model, we add term into model, we find that this term is not necessarily used to predicate salary. So we give up this model.

***Compare Models***

Now, we have decided 2 models, we compare them below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Variables finally stay  in the model | Regression Function | Adjusted  R-Square |
| Without interaction terms |  |  | 0.9225 |
| With interaction terms |  |  | 0.9324 |

Based on what we have done, straight line without interaction terms is chosen according to reasons list below.

1. Model without interaction terms in a simple model.
2. It’s easy to collect data and has practical meaning since it’s “cheaper” technically.
3. Two models have similar Adjusted R-Square .Without considering other factors, adding interaction terms does not improve model a lot.

Then we focus on checking model assumptions.

***Check Model***

* **Straight line model without interaction terms**



Since we have obtained regression function which is

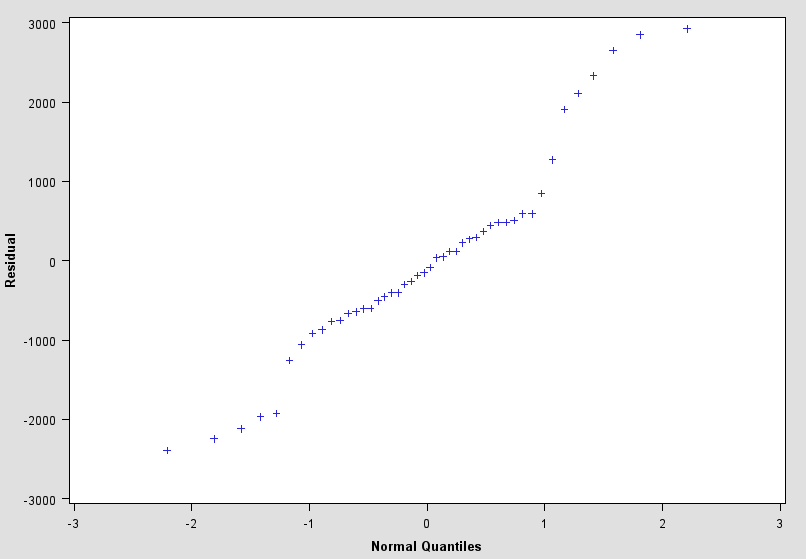


The Model Assumptions are

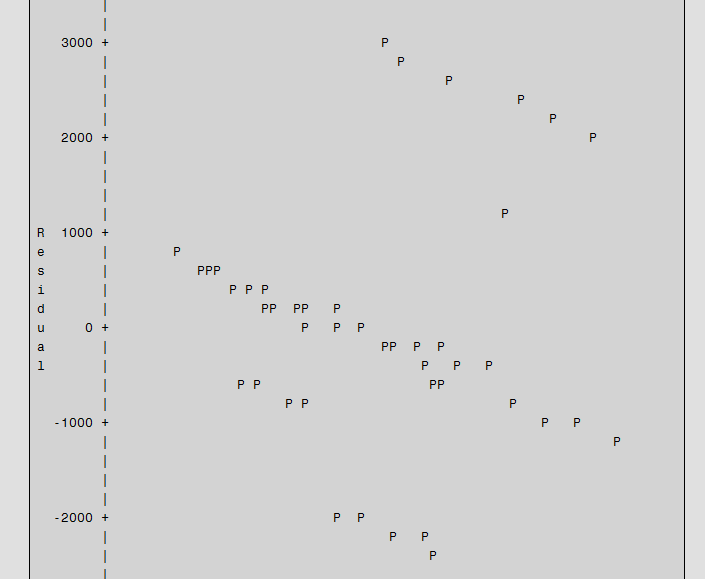
1. Linear model is sufficient;
2.  and all share the same , ;
3.  …are independent.

Firstly, we check the assumption (2); we divide that into two parts.

In order to check whetherfollow normal distribution, we use QQ-plot.



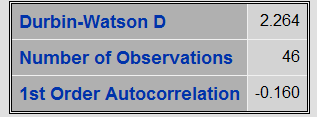
From the QQ-plot, we find that it’s approximately a straight line, which means that normality is satisfied.



From residual plot, it shows 3 descending bands. We may guess some pattern may exist. However, based what we have learned, we don’t know the exact problem; we may try some transformations to fix it.

All in all, assumption (2) is considered satisfied.

We use Durbin-Watson tests since no “time” factor is present.



We get =2.264, =46, =3.

: Vs. 

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 1.38 | 1.67 |
|  | 1.3 | 1.58 |
|  | 1.20 | 1.48 |

Since =2.264>1.48 and =1.736>1.48, we should not reject that: at significance level and thinkare not auto correlated. Because we have checked that normality and known that all errors follow normal distribution, we can conclude that  are independent at level .In other words, assumption (3) is satisfied.

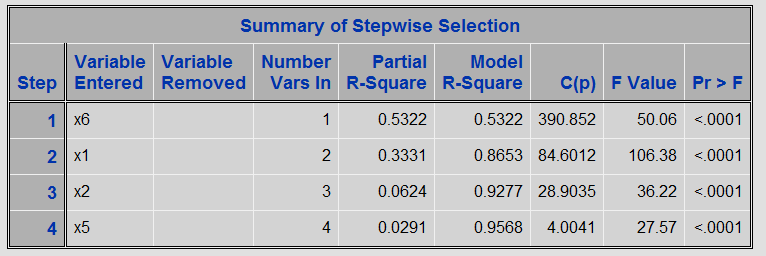
The data collected are not repeated, we cannot use Lack of fit to check assumption (1).However, since we assume assumption (2) and (3) are satisfied, it’s common that assumption (1) is satisfied.

***Transform Model***

Then we do transformation by adding, and then give a brief view.

Model

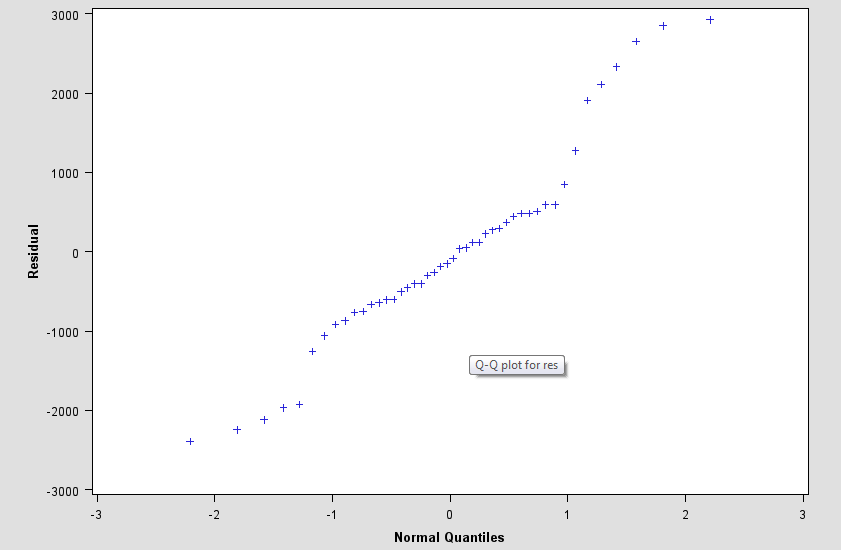
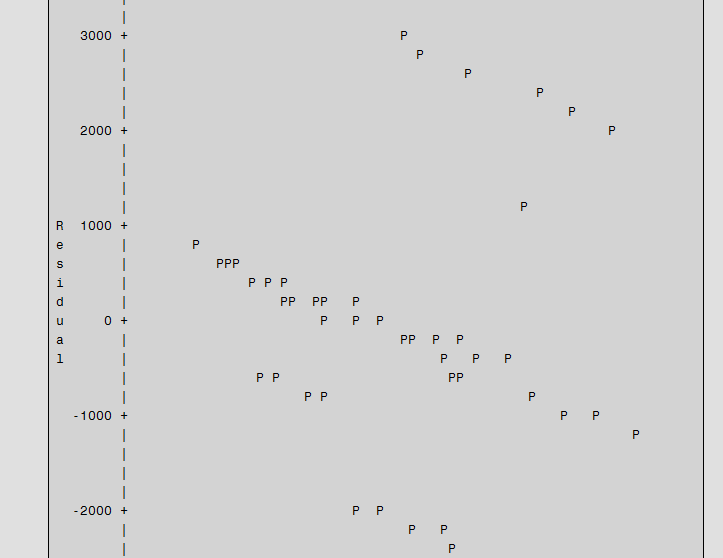




Termsenter and stay in the model, since we keep, we keep, and Model R-square is 0.9568.

Although R-Square of this model is greater than our original model, it does not have remarkably change. What’s more, in practically, it’s a complicate model since it contains too many terms.

Model assumptions need to be checked. We just quickly see QQ-plot and residual plot.

***Conclusion***

We get data from a salary of computer professionals in a large corporation; and we are interested in determining the effect of experience, education and a person’s management responsibility on salary.

We start from linear regression model without any interaction terms; results show that this linear relationship exists.



If we know a person’s working years, education level and with or without management responsibility these conditions, we may calculate his or her salary. However, what we get is not the exact relationship, it’s just estimation.

Model assumption has been checked.

Then we consider co-effects of experience, education and management responsibility, we establish a new model by adding interaction terms, but model is improved a little.

We also focus on second-order polynomial model and do some tests. Results show that more variation in salary can be explained in the model. However, this model contains too many terms, which means it will become very “expensive” to run experiments practically.

If based on what we have now, I suggest straight line model since it’s a simple model and we can easily collect data. And it has already explained about 93% variation in salary, to estimate salary with experience, education and management responsibility, that’s enough. In practically, it’s a “cheapest” model comparatively.

If we can get repeated data, lack of fit test can be done to check adequacy of linear model. More data collected will help us to find more appreciated relationship between salary and a person’s working experience, education level, and management responsibility.