MULTIPLE REGRESSION

Chapter 4

Topics

1. What is regression analysis?
2. An example (James' procedure).
   a. Logic and theory.
   b. Examining the data.
   c. Correlations.
   d. Finding the best model.
3. Evaluating the classical assumptions.

1. What is Regression Analysis?

- It is a statistical method for studying or evaluating the relationship between one or more independent variables and a single dependent variable.
  - We can use regression to predict values of the dependent variable.
  - We do not ”prove” causality, we simply test whether the relationships are significantly different from zero.
  - We generally assume the relationships to be linear.
  - Because we cannot explain everything with the independent variables, we include an error term.

Causality

'PLEASE PASS THE SALT';
Examining the Motivational Variables, Idiosyncratic Dynamics And Historic Precedents Associated With the Utterance

BYLINE: By Michael Pacanowsky

SECTION: Outlook; C1

LENGTH: 2561 words

STRONGLY ROOTED in the English speech community is the belief that the utterance, "Please pass the salt," is efficacious in causing salt to move from one end of a table to the source of the utterance. In his "Canterbury Tales," Chaucer notes:

Shee I askked

The salde to passe.

Ne surprissed was I

Tha shee didde (4, p. 318).

Similarly, Dickens writes:

"Old Heep did not become disgruntled at my obstinence. 'Please pass the salt, Davey,' he repeated coldly. I vacillated for a moment longer. Then I passed the salt, just as he knew I would (5, p. 278)."

The question of whether the movement of salt is causally dependent on the utterance of the phrase, "Please pass the salt," has occupied the attention of numerous philosophers (3, 9, 20). Empirical resolution of the validity of this belief, however, was not undertaken until the classic work of Hovland, Lumsdine and Sheffield (8) on the American soldier. Since then, numerous social scientists have explored the antecedent conditions that give rise to this apparent regularity. In this article, we will summarize those efforts that shed some light on the complex phenomenon known as salt passage.

Seminal study upon which all others rest

Army recruits, each sitting alone at one end of a table with salt at the other end, to repeat the utterance, "Please pass the salt," every five minutes for 12 hours. The average distance the salt traveled was .5 inch, which the experimenters explained was due to measurement error. The result of these two studies was, therefore, to demonstrate the importance of the presence of other people in the salt passage phenomenon.

Subsequent research

JANIS AND FESHBACH (10) found that other utterances were just as effective as "Please pass the salt"

No significant differences in the extent of compliance were found due to the utterances, "Please pass the salt," "Would you mind passing the salt?," "Could I have the salt down here, buddy?," and "Salt!" Janis and Feshbach noted that in every successful utterance, the word "salt" was found.

Moderating variables

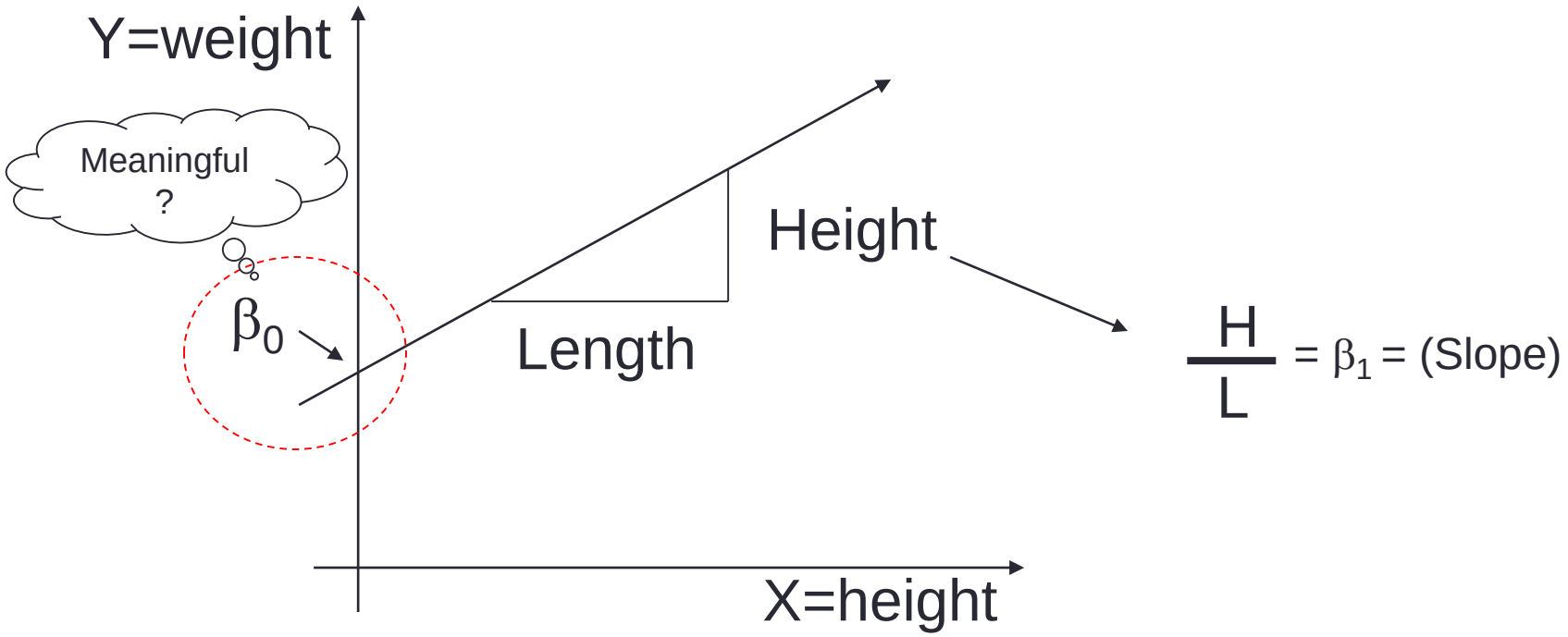
Festinger (6) tested the effects of substance uncertainty on salt passage. Subjects were placed at a table where salt was loosely piled on a napkin, while sugar was placed in a salt shaker. When a confederate said, "Please pass the salt," the overwhelming number of subjects passed the sugar. From this study, Festinger concluded that the salt shaker, not the salt itself, was the crucial factor in salt passage.

Future research

Third, future research needs to be concerned with the effects of situational variables on salt passage.Kelley's "presence of steak" variable and Milgram's "high threat" variable are suggestive. Effects of information-rich environments, overcrowding, presence of armed conflict, and so on would seem to mediate the salt passage phenomenon.

Simple Regression

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

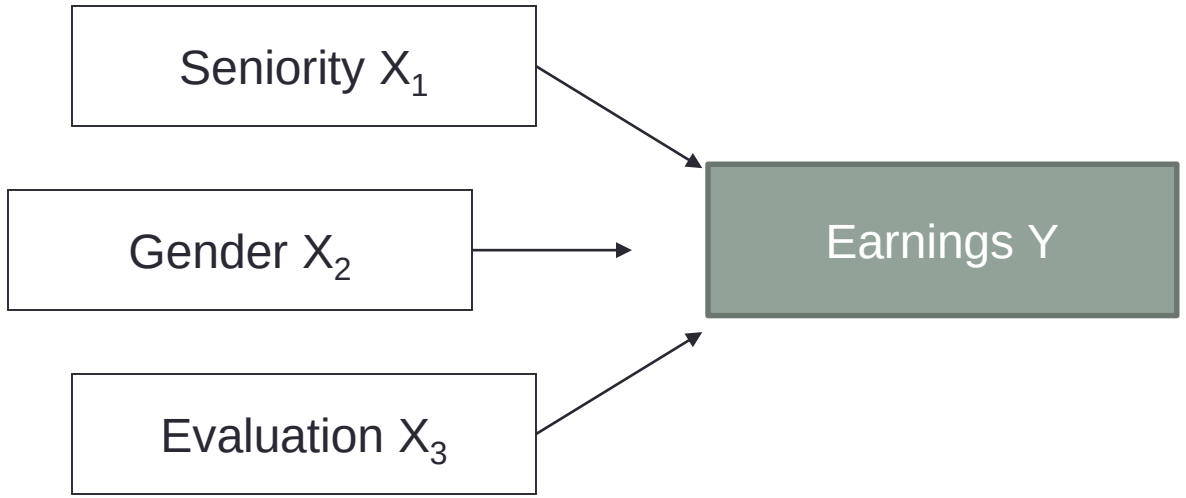


Dangers

- Too long model:
  - The model becomes less precise.

- Too short model:
  - The estimated parameters (betas) become biased.
  - The variance in the error term is not correctly estimated.
  - Hypothesis tests and confidence intervals can be misleading.

2a. From Theory …

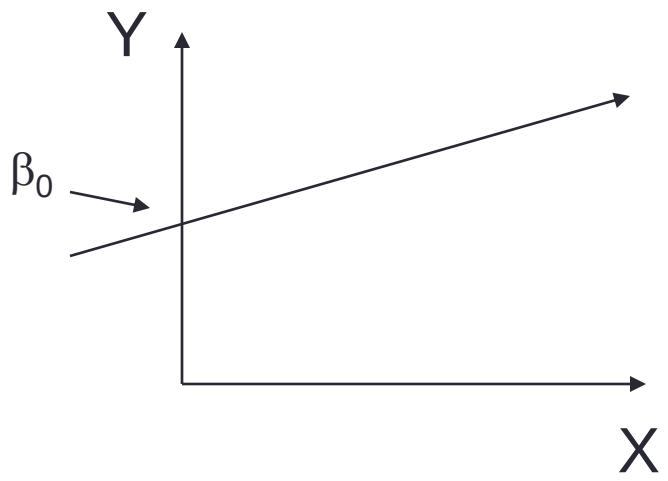


- The choice of independent variables should be theoretically based.
  - ✓ theory + logic.

… to Equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

- Where:
  - Y = daily earnings,
  - β_0 = the Y intercept,
  - β = the slope coefficients,
  - X = the variables,
  - ε = error



$$\text{Earnings} = \beta_0 + \beta_1 \text{Seniority} + \beta_2 \text{Gender} + \beta_3 \text{Evaluation} + \varepsilon$$

2b. Examining the Data

- Sample size.
- The normal distribution and normality.
- Relationships between variables.
- Outliers.
- Missing data.

Sample Size

- Remember that Power = f(effect size, alpha, sample size).
- Effect size = ?
- Alpha RoT = 0.05.
- Sample size = 31
- RoT
- <30 only simple regression.
- 15-20 observations per independent variable.
- With stepwise regression, minimum 50/X variable.

Ramifications of Small Sample

- n=31 in our study – so what?
- We are ‘overfitting’ the variate to the sample.

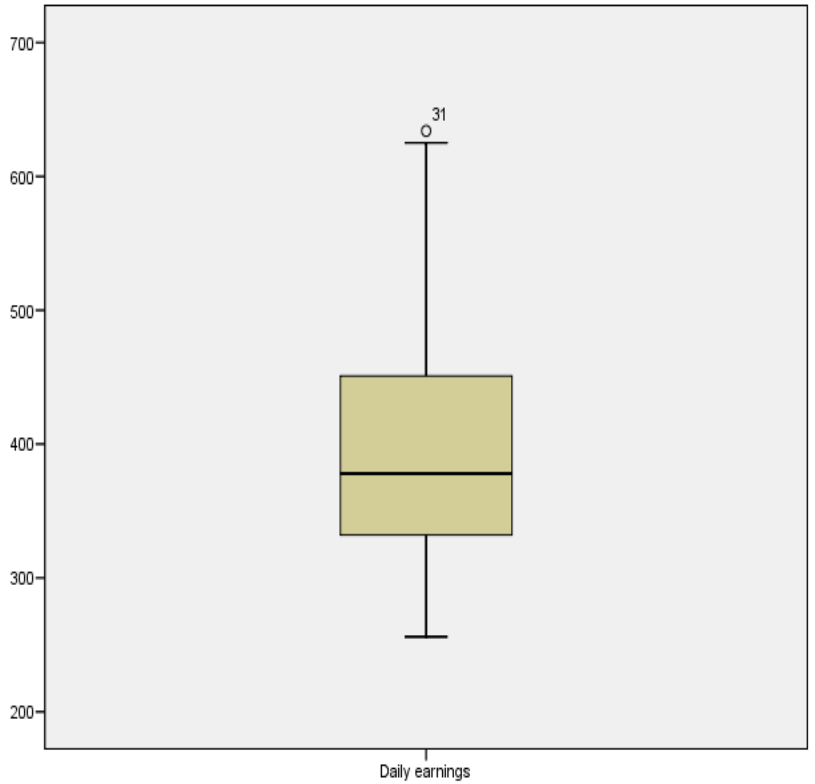
… Which means…

- We lose generalizability!

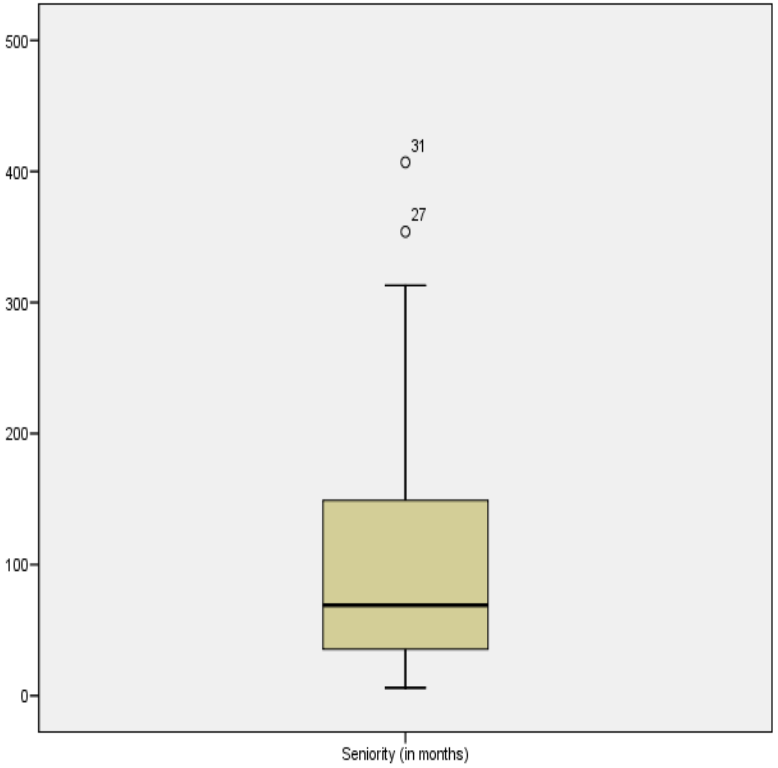
Descriptives

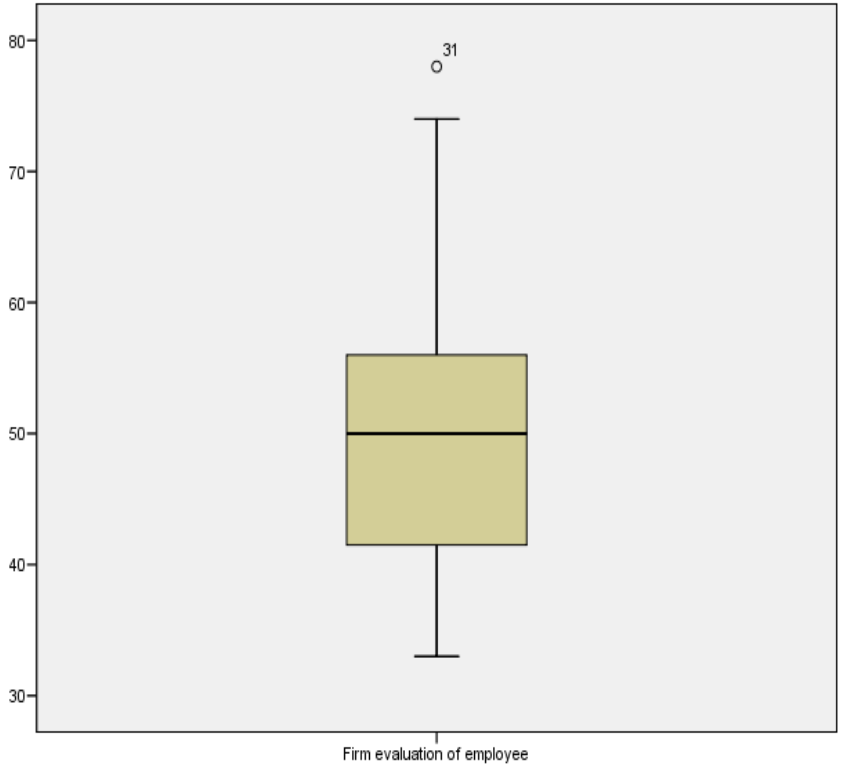
IMPORTANT
The variables do not need to be normally distributed. The residuals need to be normally distributed.

- Skewness and Kurtosis are OK, but tests of normality indicate not normal.
- IMPORTANT: Box plots indicate that 31 is a consistent outlier.



Descriptives continued

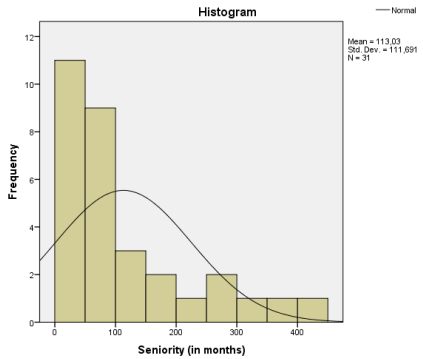


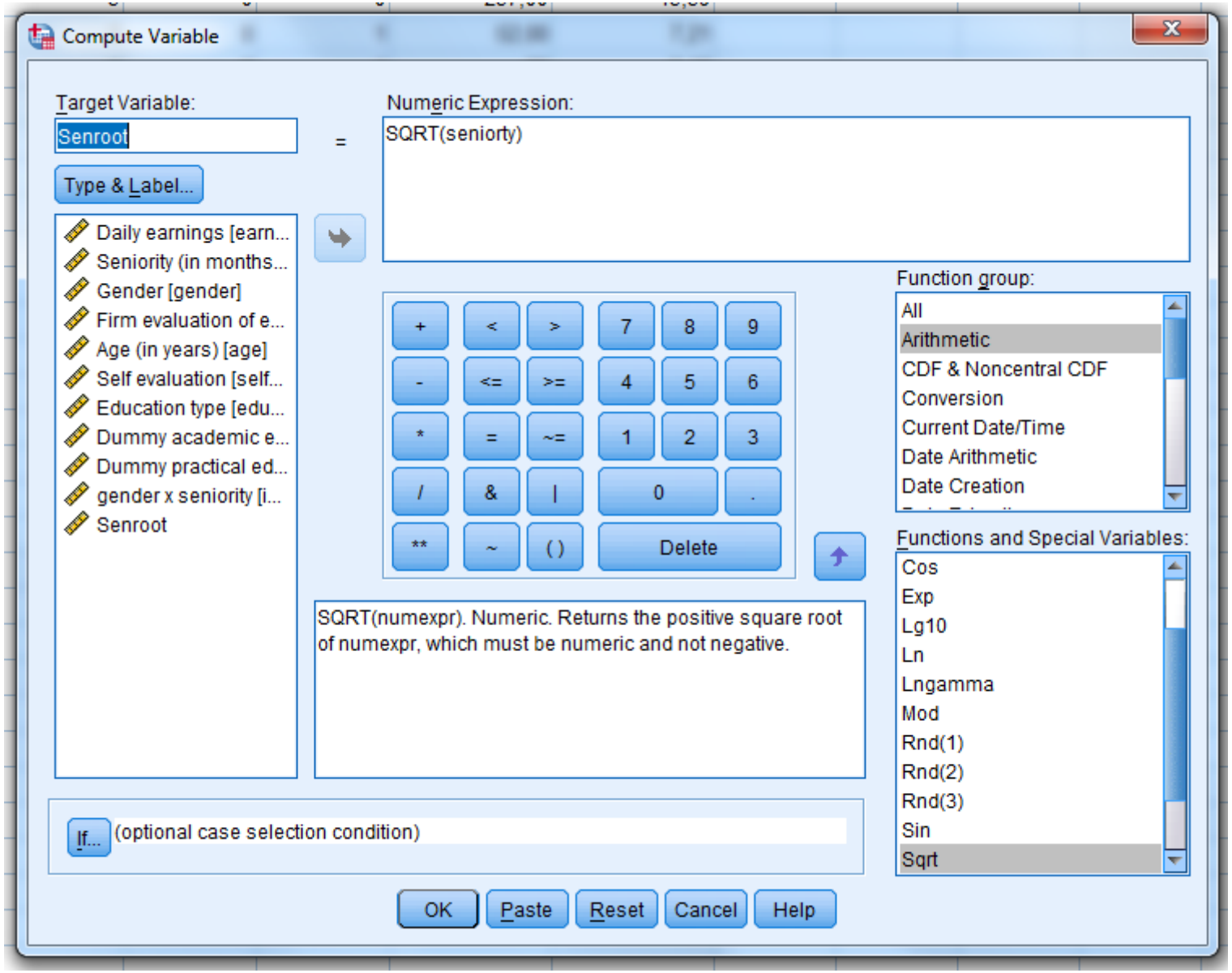


Transformation

- Flat Distribution: inverse (1/Y or 1/X).
- Negative Skew: square or cube.
- Positive Skew: log or square root.

- Positive skew means the data is clumped to the left of the histogram.



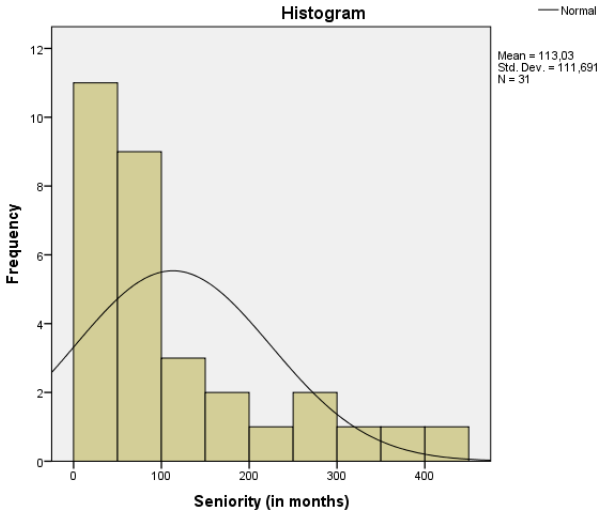

Compute Variable
Target Variable:
Senroot
Type & Label...
Daily earnings [earn...
Seniority (in months...
Gender [gender]
Firm evaluation of e...
Age (in years) [age]
Self evaluation [self...
Education type [edu...
Dummy academic e...
Dummy practical ed...
gender x seniority [i...
Senroot
=
Numeric Expression:
SQRT(seniorty)
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) Delete
SQRT(numexpr). Numeric. Returns the positive square root of numexpr, which must be numeric and not negative.
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Functions and Special Variables:
Cos
Exp
Lg10
Ln
Lngamma
Mod
Rnd(1)
Rnd(2)
Rnd(3)
Sin
Sqrt
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help

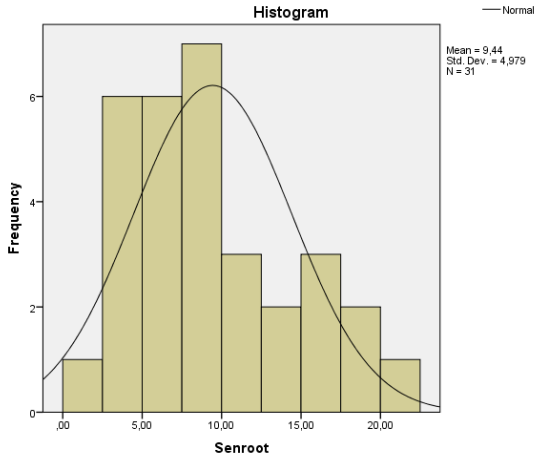
Transformation

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Seniority (in months)	,230	31	,000	,825	31	,000
Senroot	,145	31	,094	,933	31	,052

a. Lilliefors Significance Correction





Positive skew means bunched to the left

Transformed Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	70,344	24,746		2,843	,008
	Seniority (in months)	,288	,070	,311	4,089	,000
	Gender	9,769	8,355	,048	1,169	,253
	Firm evaluation of employee	5,858	,651	,698	9,000	,000

a. Dependent Variable: Daily earnings

Interpretation?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	32,638	20,826		1,567	,129
	Senroot	5,358	1,626	,258	3,296	,003
	Gender	6,247	8,791	,031	,711	,483
	Firm evaluation of employee	6,280	,671	,748	9,364	,000

a. Dependent Variable: Daily earnings

2c. Correlations

Low (ROT: significantly correlated with the dependent variable)

All coefficients are the correct sign

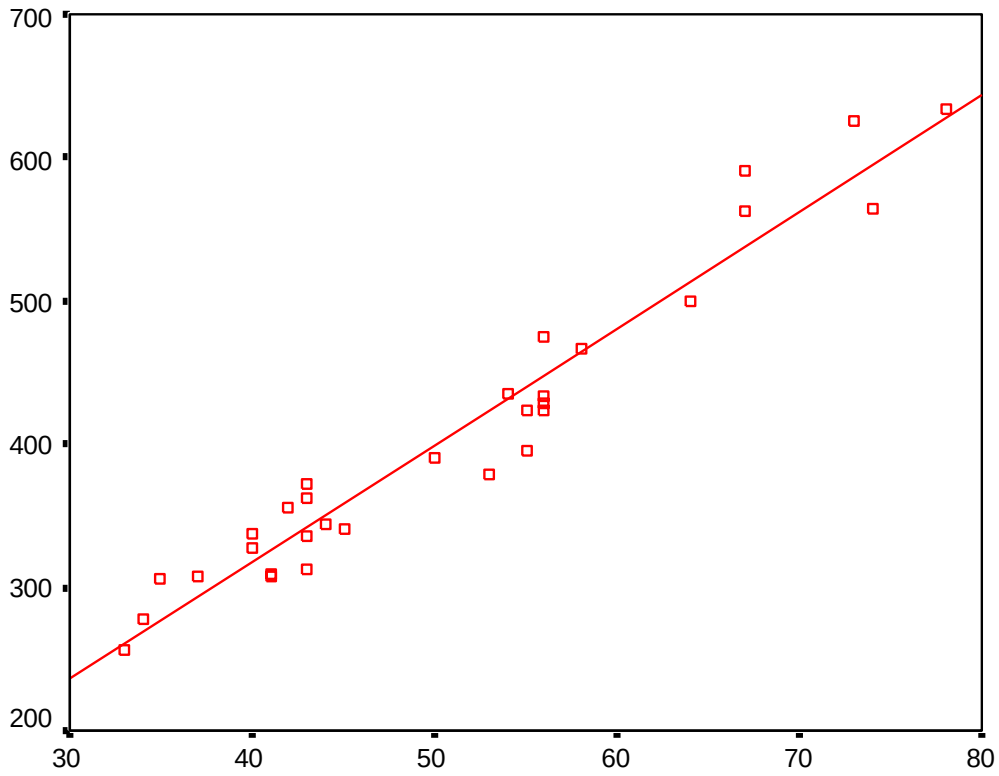
Correlations

		Daily earnings	Seniority (in months)	Gender	Firm evaluation of employee
Daily earnings	Pearson Correlation	1	.900**	.169	.972**
	Sig. (2-tailed)	.	.000	.363	.000
	N	31	31	31	31
Seniority (in months)	Pearson Correlation	.900**	1	-.071	.849**
	Sig. (2-tailed)	.000	.	.703	.000
	N	31	31	31	31
Gender	Pearson Correlation	.169	-.071	1	.206
	Sig. (2-tailed)	.363	.703	.	.267
	N	31	31	31	31
Firm evaluation of employee	Pearson Correlation	.972**	.849**	.206	1
	Sig. (2-tailed)	.000	.000	.267	.
	N	31	31	31	31

**. Correlation is significant at the 0.01 level (2-tailed).

Danger for multicollinearity (ROT>0.9), therefore, run VIF statistics

Scatterplot

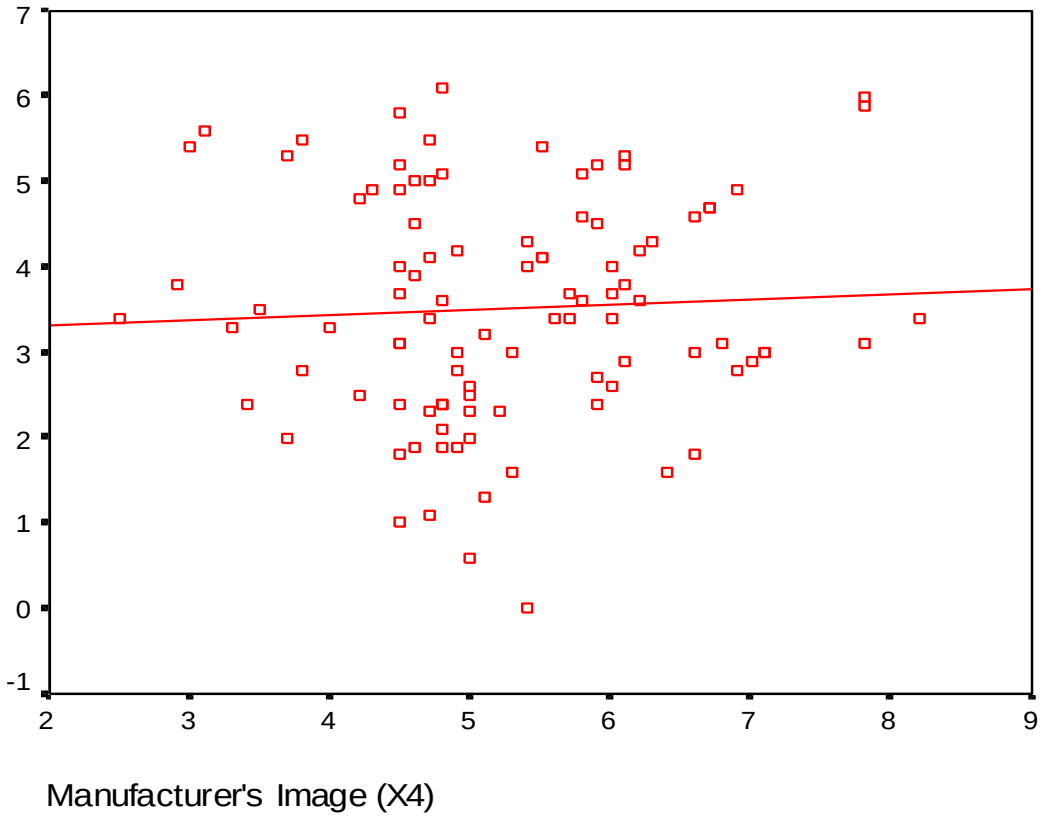


With a correlation coefficient of 0.972 is no surprise that the data are tightly grouped along the line.

Scatterplot – Low Correlation

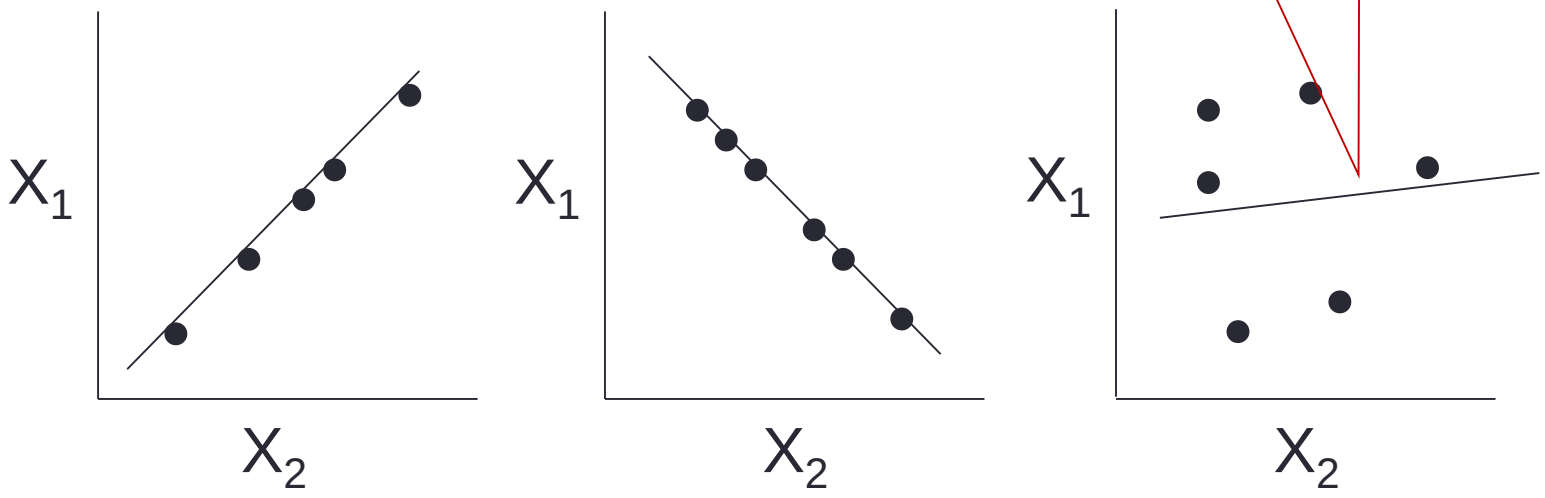
This is from a different model!

Correlation coefficient = .050



Slope

Beware the fit line. Just because it is there does not mean anything!

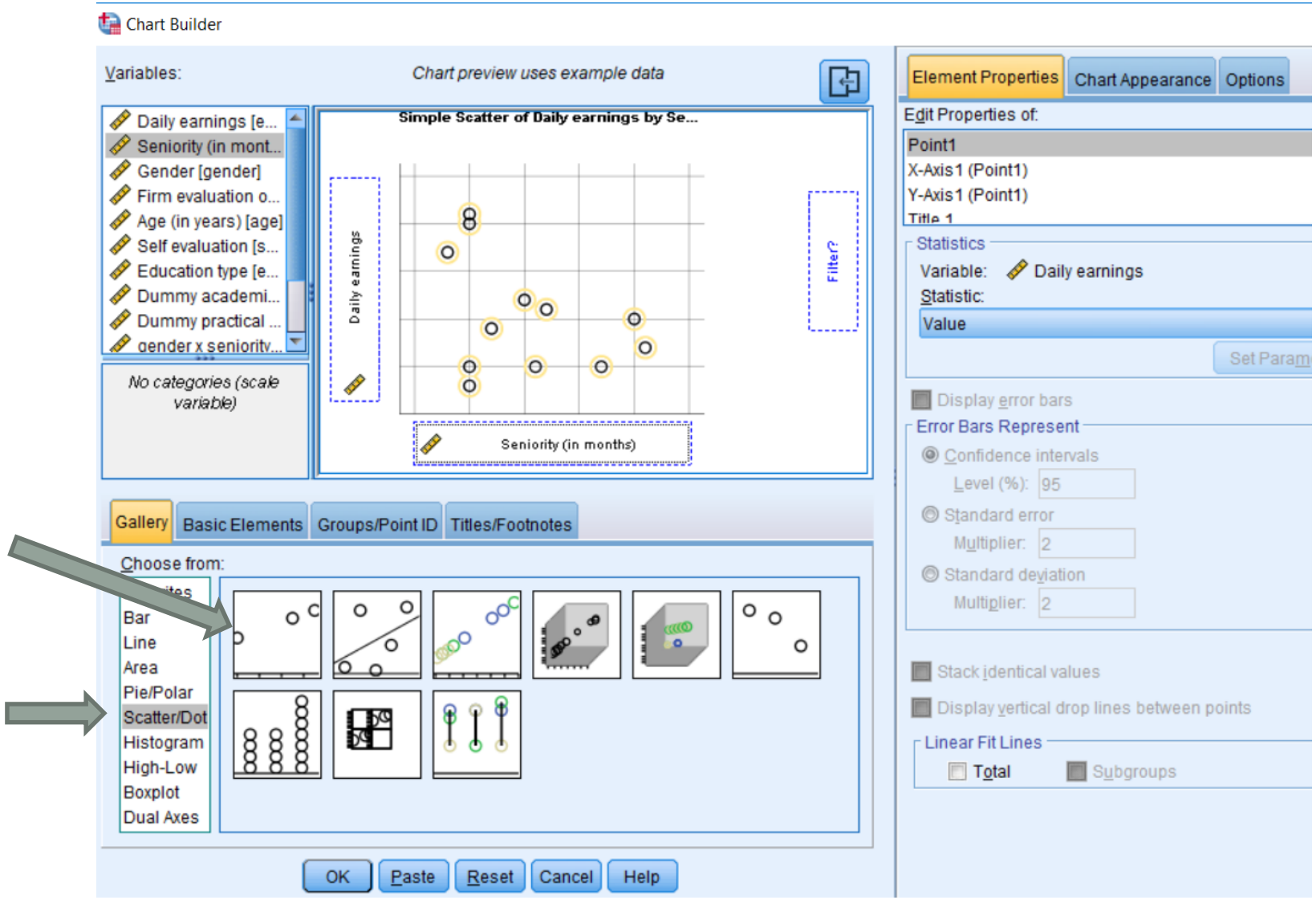


Perfect positive correlation +1

Perfect negative correlation -1

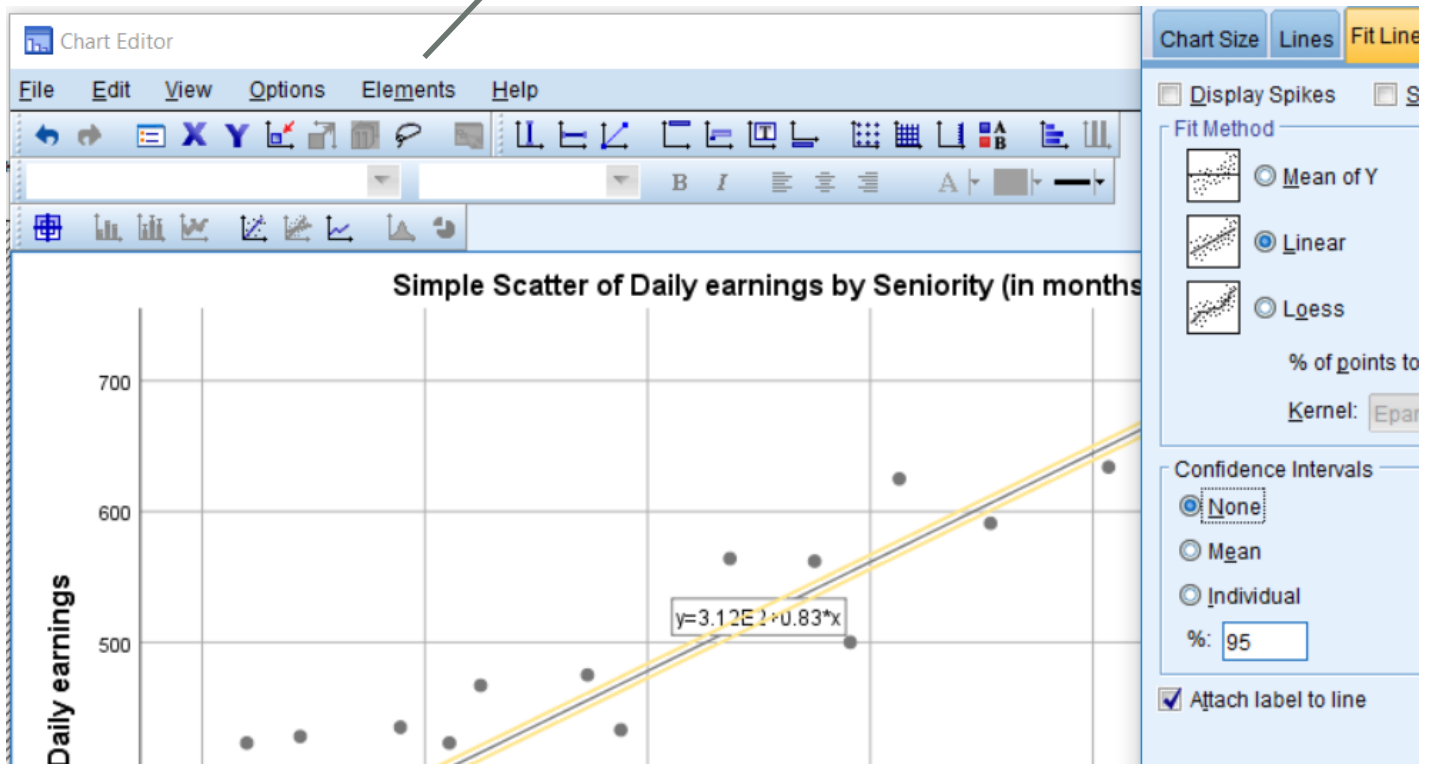
No correlation 0

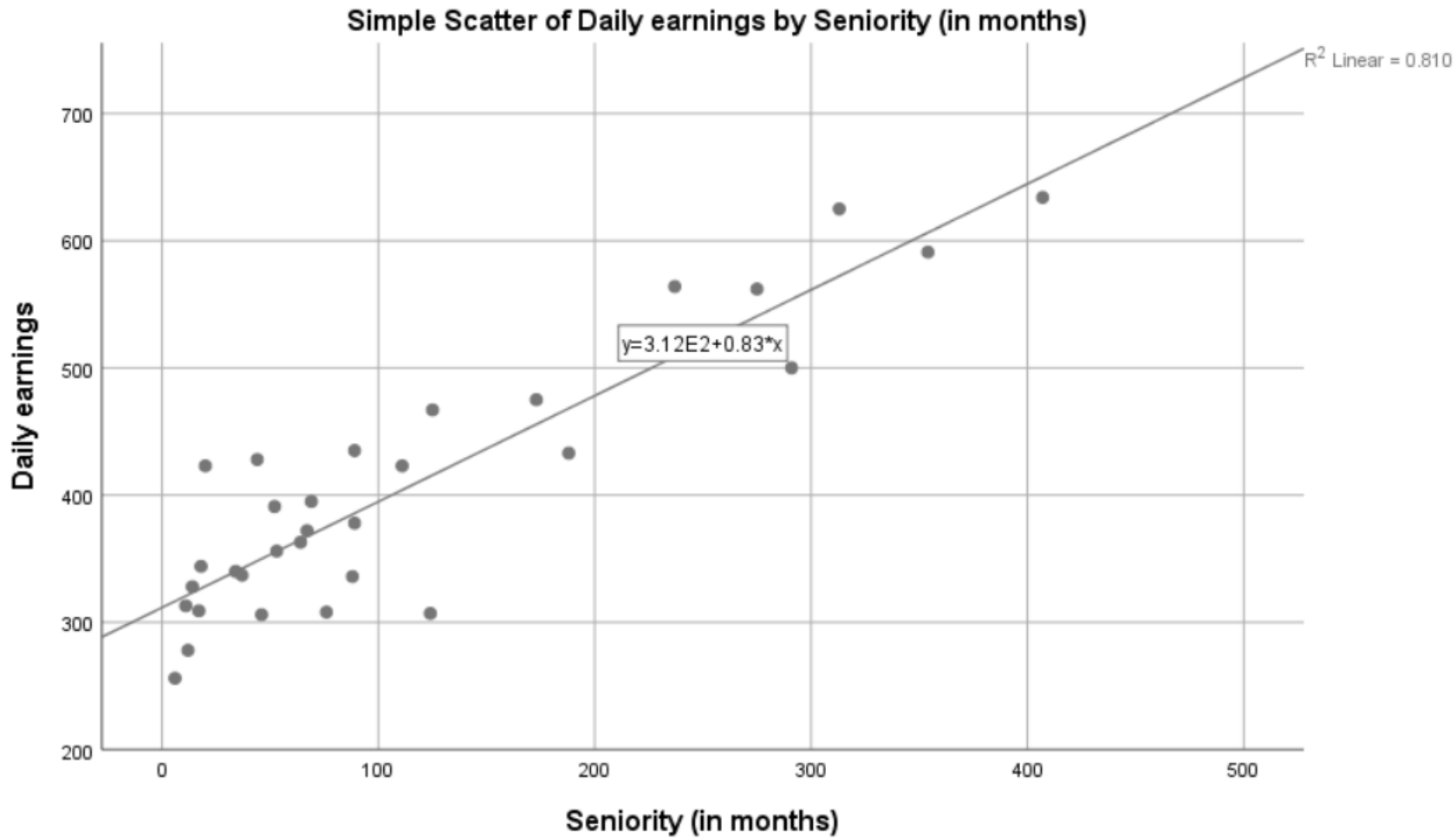
Graphs – Chart builder



Double-click on graph
Choose elements
Fit line at total

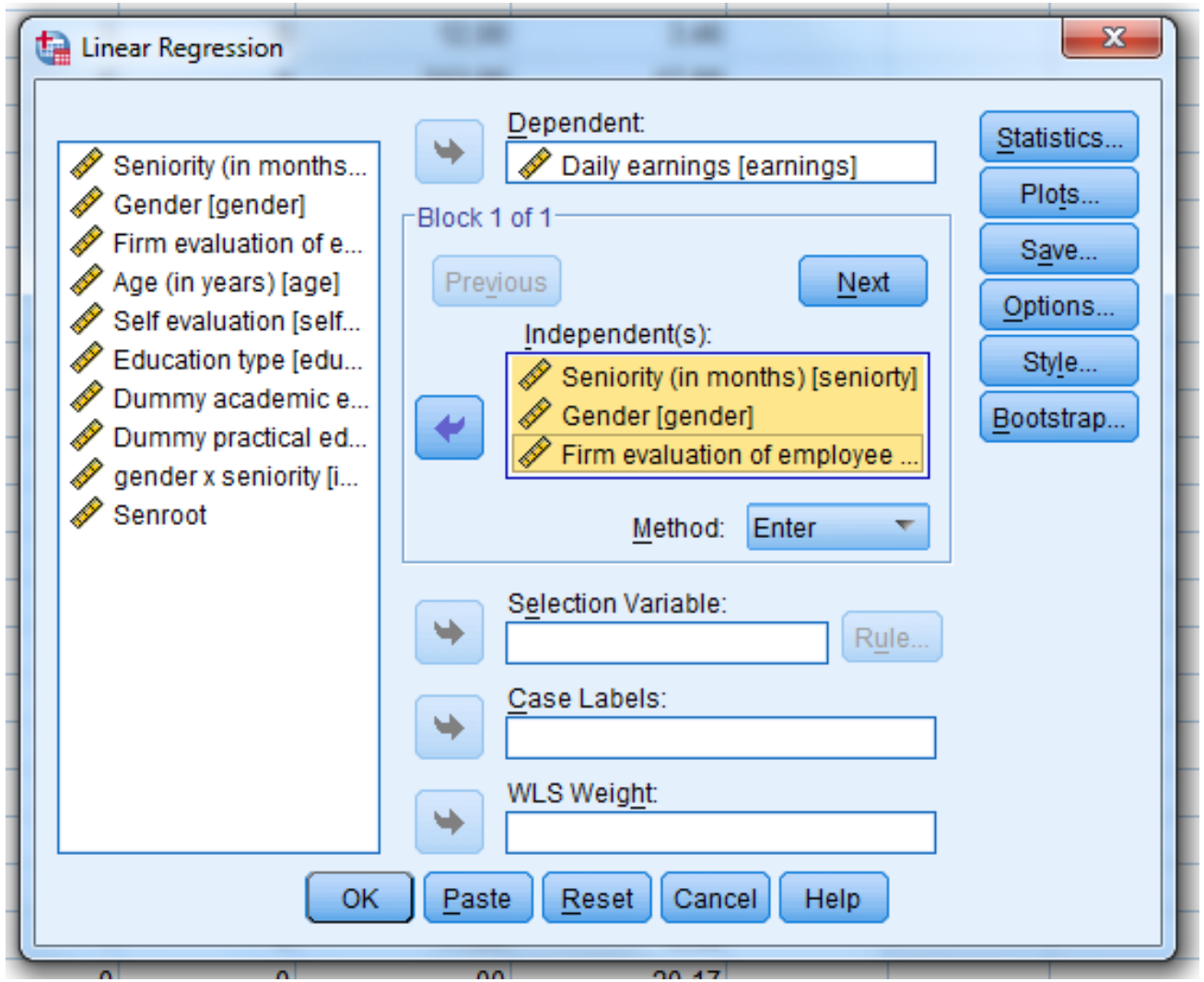
Choose fit line at total




Simple Scatter of Daily earnings by Seniority (in months)
R² Linear = 0.810
y=3.12E2+0.83*x
Daily earnings
Seniority (in months)
700
600
500
400
300
200
0
100
200
300
400
500


Simple Scatter of Daily earnings by Firm evaluation of employee
R² Linear = 0.944
Double-click to activate
y=-9.01+8.16*x
Daily earnings
700
600
500
400
300
200
30
40
50
60
70
80
Firm evaluation of employee


Simple Scatter of Daily earnings by Age (in years)
Double-click to activate
R^2 Linear = 0.327
y=1.93E2+5.22*x
Daily earnings
Age (in years)
200
300
400
500
600
20
30
40
50
60


Linear Regression
Seniority (in months...
Gender [gender]
Firm evaluation of e...
Age (in years) [age]
Self evaluation [self...
Education type [edu...
Dummy academic e...
Dummy practical ed...
gender x seniority [i...
Senroot
Dependent:
Daily earnings [earnings]
Block 1 of 1
Previous
Next
Independent(s):
Seniority (in months) [seniorty]
Gender [gender]
Firm evaluation of employee ...
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help

2d. Finding the Best Model

R^2 – The coefficient of determination (explained variance)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.983[a]	.966	.962	20.011

a. Predictors: (Constant), Firm evaluation of employee, Gender, Seniority (in months)

Refer to this when discussing individual equations.

Refer to this when comparing equations.

Evaluating the Entire Equation

Use this one when referring to the t-statistic cutoff values.

Small (<0.05) indicates significance.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	309532.5	3	103177.501	257.660	.000[a]
	Residual	10811.886	27	400.440		
	Total	320344.4	30			

a. Predictors: (Constant), Firm evaluation of employee, Gender, Seniority (in months)

b. Dependent Variable: Daily earnings

Use these two when referring to the F-statistic cutoff values.

The F-statistic indicates the significance of the entire equation (ignoring the parts).

F-Distribution $\alpha = 0.05$

Numerator ν_1

Denominator ν_2	1	2	3	4	. . .
.	.	.	.	.	. . .
.	.	.	.	.	. . .
.	.	.	.	.	. . .
25	4.24	3.39	2.99	2.76	. . .
26	4.23	3.37	2.98	2.74	. . .
27	4.21	3.35	**2.96**	2.73	. . .
28	4.20	3.34	2.95	2.71	. . .
29	4.18	3.33	2.93	2.70	. . .
.	.	.	.	.	. . .
.	.	.	.	.	. . .
.	.	.	.	.	. . .

Critical cutoff value.

Evaluating the Parameters (Xs)

Look for ”big” error terms.

Compare standardized effect sizes to assess relative importance.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	70.344	24.746		2.843	.008
	Seniority (in months)	.288	.070	.311	4.089	.000
	Gender	9.769	8.355	.048	**1.169**	.253
	Firm evaluation of employee	5.858	.651	.698	9.000	.000

a. Dependent Variable: Daily earnings

Look for significant t-values (ROT>2).

Small (<0.05) indicates significance.

Finding the Critical Cutoff Value

- We use n-k-1 to determine the critical value in the table, where
  - N is the number of observations (31),
  - K is the number of parameters (3).
- 31-3-1=27
- The critical value for a one-sided test at a 5% level of significance is 1.703.

t-Distribution

Level of Significance

One Sided	.	**5%**	**2.5%**	.	. . .
Two Sided	.	**10%**	**5%**	.	. . .
.	.	.	.	.	. . .
.	.	.	.	.	. . .
.	.	.	.	.	. . .
26	.	1.706	2.056	.	. . .
27	.	**1.703**	2.052	.	. . .
28	.	1.701	2.048	.	. . .
29	.	1.699	2.045	.	. . .
30	.	1.697	2.042	.	. . .
.	.	.	.	.	. . .
.	.	.	.	.	. . .
∞	.	1.645	1.960	.	. . .

Degrees of Freedom ν

Critical cutoff value.

Past 60 the critical cutoff is less than 2, thus the Rule of Thumb, t > 2

Rerunning the Model – no gender

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.982[a]	.965	.962	20.142

a. Predictors: (Constant), Firm evaluation of employee, Seniority (in months)

Very good.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	308985.1	2	154492.543	380.815	.000[a]
	Residual	11359.300	28	405.689		
	Total	320344.4	30			

a. Predictors: (Constant), Firm evaluation of employee, Seniority (in months)

b. Dependent Variable: Daily earnings

The Final Model

Equation coefficients.

Very good.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	60.265	23.347		2.581	.015		
	Seniority (in months)	.249	.062	.269	3.990	.000	.279	3.582
	Firm evaluation of employee	6.242	.565	.744	11.041	.000	.279	3.582

a. Dependent Variable: Daily earnings

The smaller effect.

The larger effect.

Less than 5 (10), so no apparent problem with multicollinearity (classic assumption).

Interpreting the Equation

Original equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Final equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

Which translates to:

$$\text{Earnings} = \beta_0 + \beta_1 \text{Seniority} + \beta_2 \text{Evaluation} + \varepsilon$$

Or:

$$\text{Earnings} = 60.265 + .249\text{Seniority} + 6.242\text{Evaluation} + \varepsilon$$

Interpretation:

Publishing pays!

3. The Classical Assumptions

- This is a fairly basic set of 7 assumptions required to hold for ”ordinary least squares” to be considered the best estimator of the regression model.
  - If the assumptions don’t hold, other estimation techniques may be better.
  - E.g. Weighted least squares, Two-stage least squares.

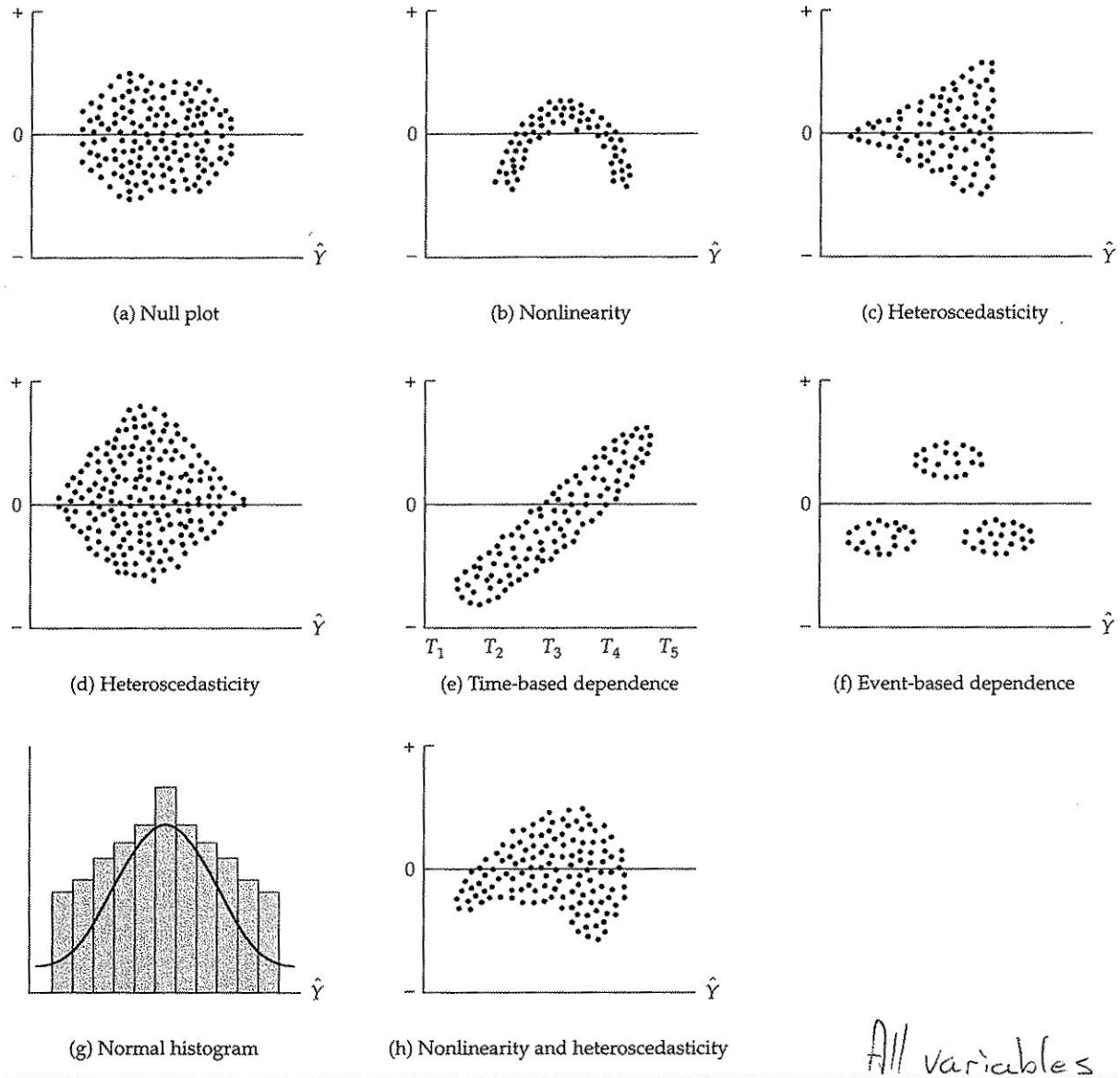
  - NOTE: We usually examine standardized residuals to allow for comparison across variables.
    - Studentized are a form of standardized.

The Assumptions

1. The regression model is linear in the coefficients.
   - Produce null plot and partial plots.
2. Constant variance of the error terms.
   - Plot studentized residuals against predicted values to look for heteroscedasticity.
3. Independence of the error terms.
   - Plot residuals against any possible sequencing variable (e.g. Time).
4. Normality of the error terms.
   - Test the residuals for a normal distribution (refer to examining data).

The Assumptions

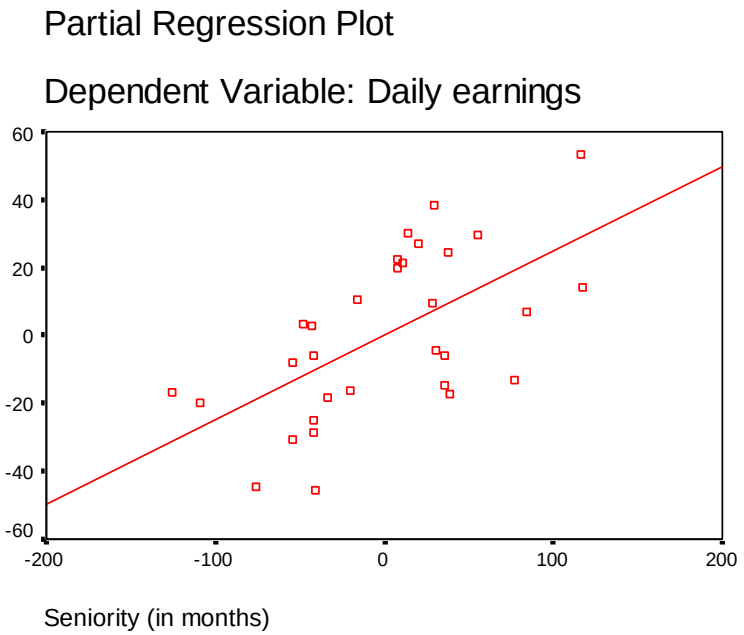
5. No explanatory variable is a perfect function of any other explanatory variable.
   - Test for multicollinearity.
6. The error term has a zero population mean.
   - This cannot be shown, however it is compensated for by the inclusion of the constant (β_0), so do not worry about it.
7. The explanatory variables are uncorrelated with the error term.
   - We assume this to be true unless we know otherwise.
   - An example of a violation of this is when introducing an interaction effect as the product of two variables. By definition we have violated this assumption and need to look for an alternative estimation procedure (e.g. 2SLS).

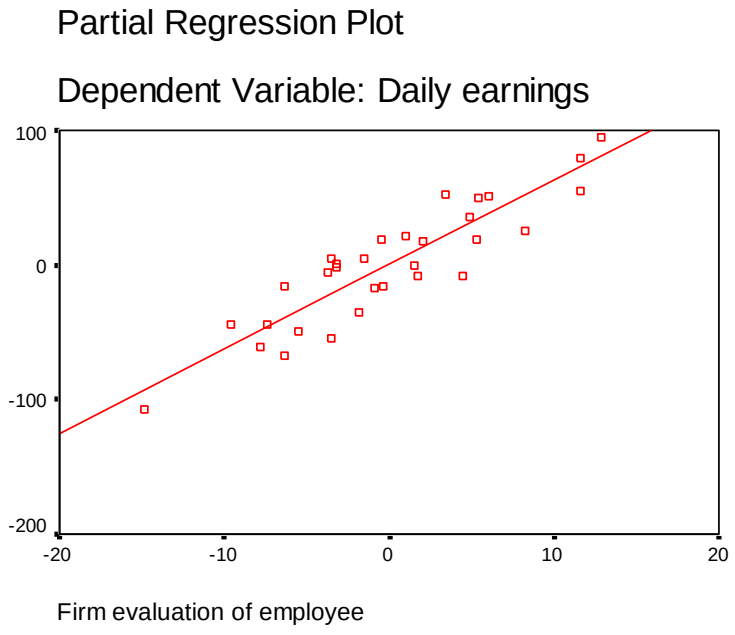


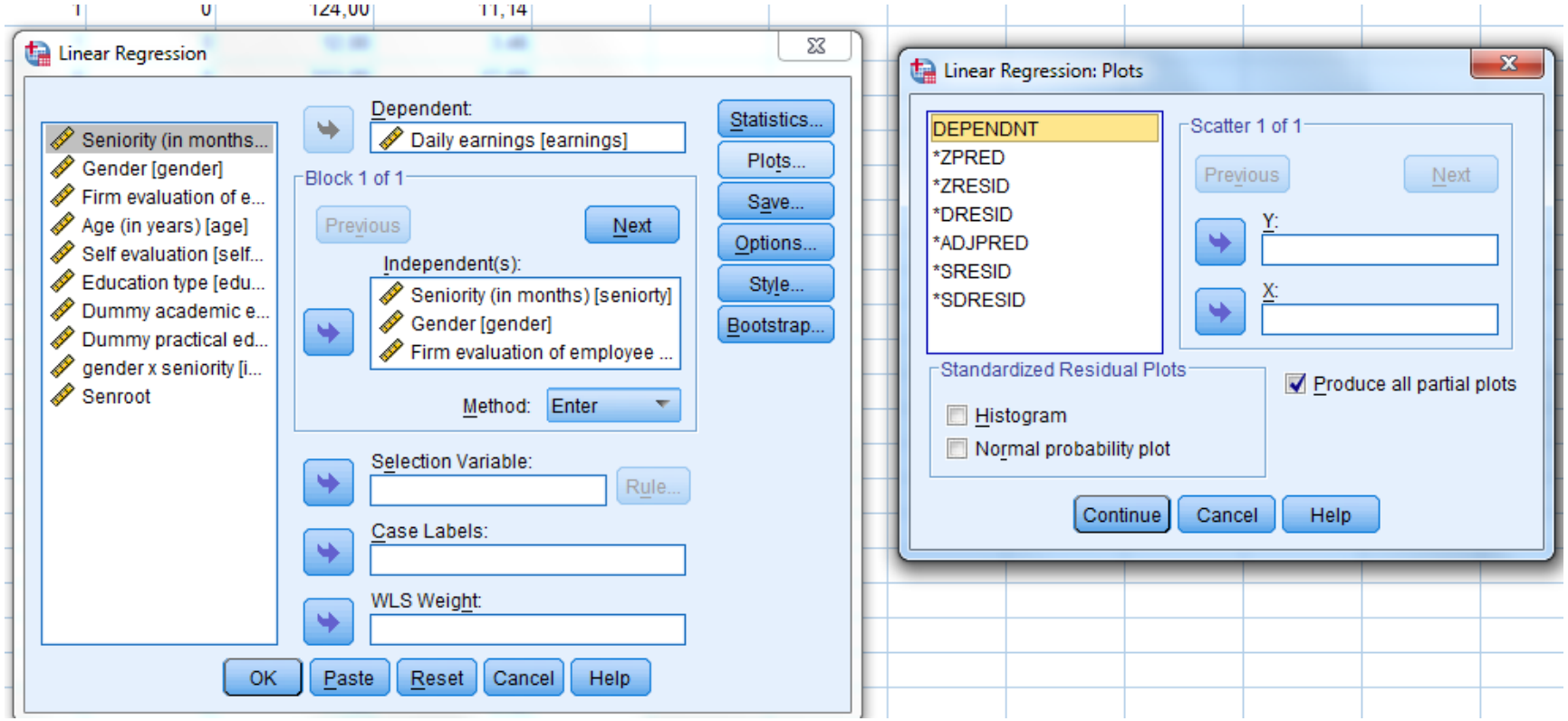
All variables

FIGURE 4-5 Graphical Analysis of Residuals

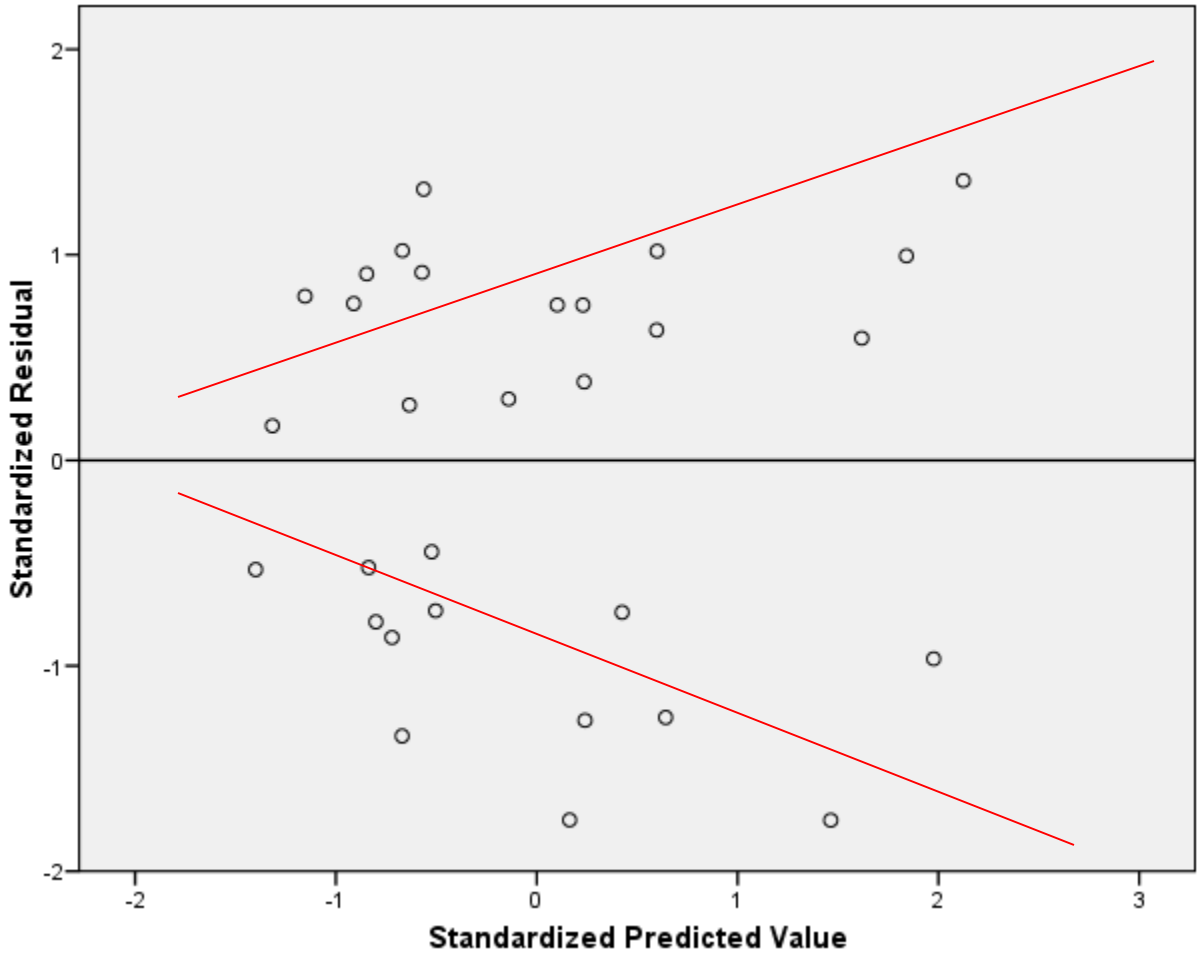
(1) Linearity – Partial Plots






Linear Regression
Seniority (in months...
Gender [gender]
Firm evaluation of e...
Age (in years) [age]
Self evaluation [self...
Education type [edu...
Dummy academic e...
Dummy practical ed...
gender x seniority [i...
Senroot
Dependent:
Daily earnings [earnings]
Block 1 of 1
Previous
Next
Independent(s):
Seniority (in months) [seniorty]
Gender [gender]
Firm evaluation of employee ...
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
X:
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
Continue
Cancel
Help

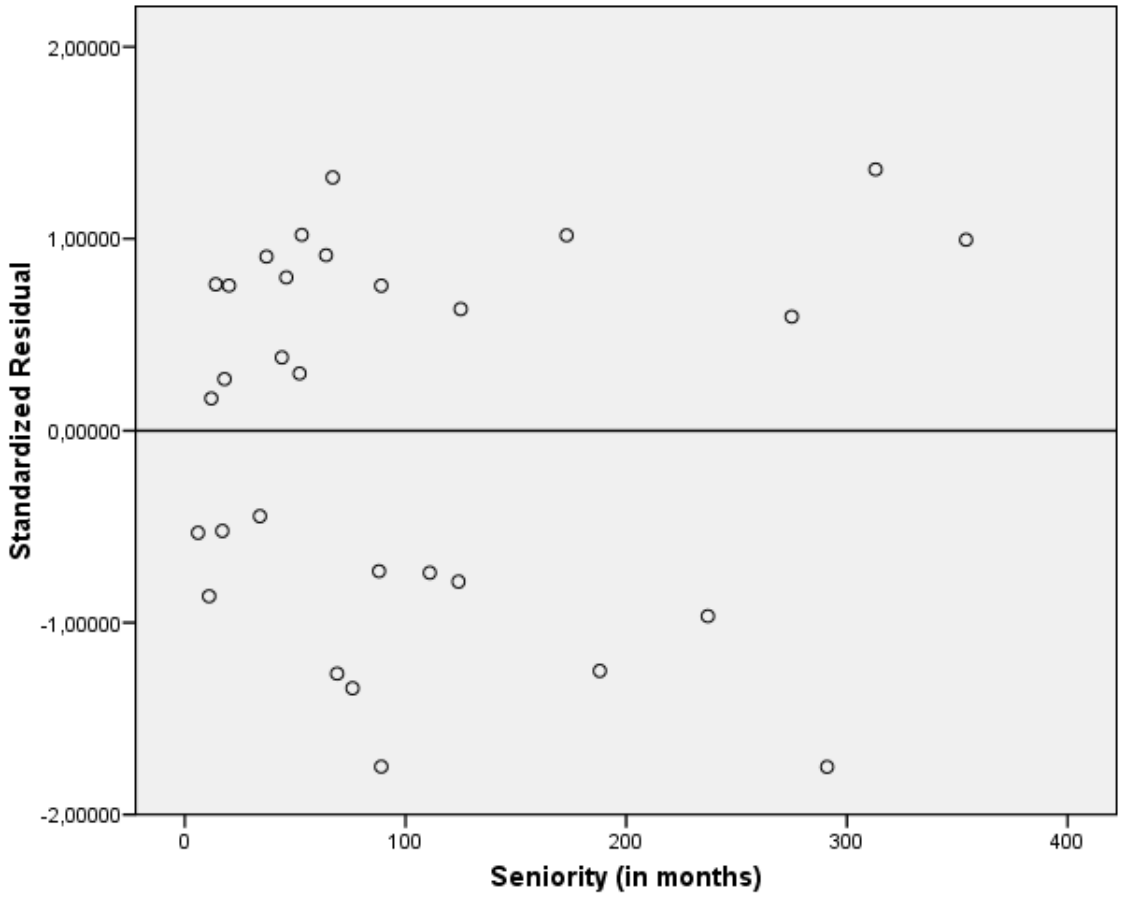
(2) Constant Variance of Error Terms



Hmm?

(3) Independence of the Error Terms

Plot standardized residuals against possible sequencing variables (e.g. time)

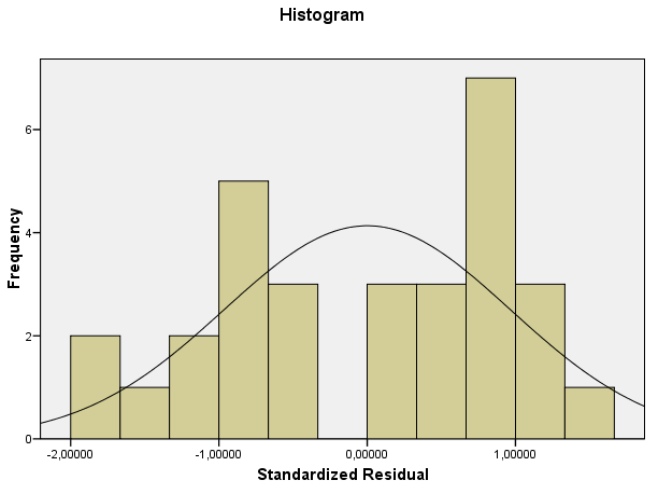


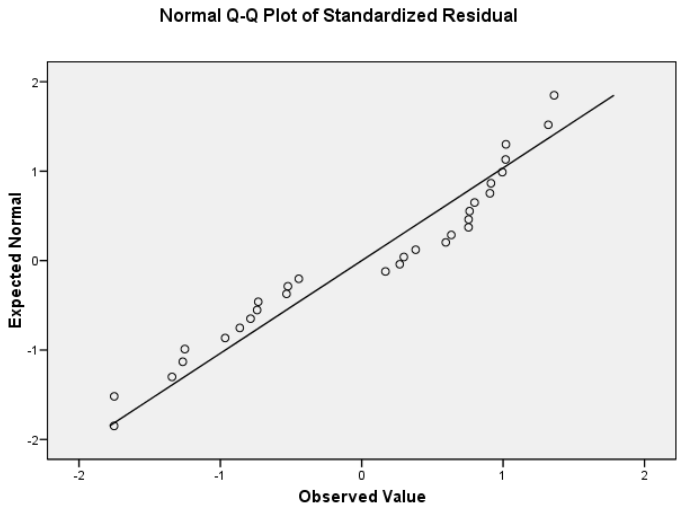
(4) Normality of Error Terms

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	,164	30	,038	,915	30	,020

a. Lilliefors Significance Correction





(5) Multicollinearity

- Multicollinearity is a gradual phenomenon so there are no fixed points.

- The correlation coefficient between X_1 and X_2 is 0.849.
  - Rule of thumb: Correlations over .9 indicate a risk of multicollinearity.

- The VIF (variance inflation factor) is 3.582 for each of the variables.
  - Rule of thumb: The VIF should be below 5 (10).

(6) Error has Zero Pop. Mean

- No test
  - Taken care of by the inclusion of the constant (β_0).

(7) X Var. Uncorrelated with Error

- Assumed to be true unless otherwise known.

- Example of violation:
- $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Product interaction effect

Advice: look for outliers in partial plots