Linear Regression Analysis for Survey Data

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

Goals for this Lecture

- Linear regression
  - How to think about it for Lickert scale dependent variables
  - Coding nominal independent variables
- Linear regression for complex surveys
- Weighting
- Regression in JMP

Regression in Surveys

- Useful for modeling responses to survey questions as function of (external) sample data and/or other survey data
  - Sometimes easier/more efficient then high-dimensional multi-way tables
  - Useful for summarizing how changes in the Xs affect Y

(Simple) Linear Model

- General expression for a linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  - β_0 and β_1 are model parameters
  - ε is the error or noise term
- Error terms often assumed independent observations from a $N(0,\sigma^2)$ distribution
  - Thus $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
  - And $E\left(Y_i\right) = \beta_0 + \beta_1 x_i$

Linear Model

- Can think of it as modeling the expected value of y,

$$E(y \mid x) = \beta_0 + \beta_1 x$$

where on a 5-point Lickert scale, the ys are only measured very coarsely

- Given some data, we will estimate the parameters with coefficients

$$E(\hat{y} \mid x) \equiv \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ is the predicted value of y

Estimating the Parameters

- Parameters are fit to minimize the sums of squared errors:

$$SSE = \sum_{i=1}^{n}\left(y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_i\right]\right)^2$$

- Resulting OLS estimators:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Using Likert Scale Survey Data as Dependent Variable in Regression

- Likert scale data is categorical (ordinal)
- If use as dependent variable in regression, make the assumption that "distance" between categories is equal

	<u>**Coding**</u>
☐ Strongly agree	1
☐ Agree	2
☐ Neutral	3
☐ Disagree	4
☐ Strongly disagree	5

2-1=1

3-2=1

4-3=1

5-4=1

Coding imposes this –
Is it reasonable?

My Take

- Generally, I’m okay with assumption for 5-point Likert scale
  - Boils down to assuming “Agree” is halfway between “Neutral” and “Strongly agree”
- Not so much for Likert scales without neutral midpoint or more than 5 points
- If plan to analyze with regression, perhaps better to use numerically labeled scale with more points:

Strongly agree				Neither agree nor disagree				Strongly disagree
1	2	3	4	5	6	7	8	9

From Simple to Multiple Regression

- Simple linear regression: One Y variable and one X variable ($y_i = \beta_0 + \beta_1 x_i + \varepsilon$)
- Multiple regression: One Y variable and *multiple* X variables
  - Like simple regression, we're trying to model how Y depends on X
  - Only now we are building models where Y may depend on many Xs

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon$$

Using Multiple Regression to "Control for" Other Factors

- Often interested in the effect of one particular x on y
  - Effect of deployment on retention?
- However, other xs also affect y
  - Retention varies by gender, family status, etc.
- Multiple regression useful for isolating effect of deployment after accounting for other xs
  - "Controlling for the effects of gender and family status on retention, we find that deployment affects retention…"

Correlation Matrices
Useful Place to Start

- JMP: Analyze > Multivariate Methods > Multivariate

Correlations

	2a	2b	2c	2d	2e	2f	2g	2h	2i	
2a	1.0000	0.6615	0.3363	0.1057	0.4057	0.1659	0.2781	0.4134	0.3564	0.7
2b	0.6615	1.0000	0.2870	0.1004	0.3305	0.1343	0.1437	0.3590	0.3183	0.5
2c	0.3363	0.2870	1.0000	0.0616	0.2272	0.1290	0.0666	0.1259	0.0227	0.2
2d	0.1057	0.1004	0.0616	1.0000	0.1324	0.1391	0.0563	0.2080	0.1913	0.0
2e	0.4057	0.3305	0.2272	0.1324	1.0000	0.2922	0.4095	0.3287	0.3206	0.3
2f	0.1659	0.1343	0.1290	0.1391	0.2922	1.0000	0.3440	0.4836	0.2848	0.1
2g	0.2781	0.1437	0.0666	0.0563	0.4095	0.3440	1.0000	0.3569	0.2344	0.3
2h	0.4134	0.3590	0.1259	0.2080	0.3287	0.4836	0.3569	1.0000	0.2966	0.3
2i	0.3564	0.3183	0.0227	0.1913	0.3206	0.2848	0.2344	0.2966	1.0000	0.2
3a	0.7266	0.5709	0.2437	0.0752	0.3551	0.1924	0.3158	0.3982	0.2516	1.0
3b	0.4304	0.7040	0.2665	0.0397	0.2753	0.1519	0.1921	0.3379	0.1811	0.5
3c	0.3849	0.4556	0.3499	0.0980	0.3949	0.1501	0.3789	0.2866	0.2028	0.5
3d	0.2548	0.3015	0.2734	0.6815	0.2549	0.1824	0.1267	0.2743	0.2422	0.3
3e	0.3511	0.2999	0.2939	0.0185	0.7195	0.3453	0.3030	0.2517	0.2712	0.4
3f	0.1012	0.1529	0.0786	0.0140	0.2060	0.6816	0.3564	0.2690	0.2195	0.2
3g	0.3301	0.1056	0.0415	0.0276	0.3449	0.2466	0.6719	0.2807	0.1838	0.3
3h	0.3903	0.2945	0.1406	0.1517	0.2757	0.3317	0.2428	0.7096	0.2118	0.3
3i	0.2665	0.2702	-0.0302	0.1997	0.2367	0.2258	0.1847	0.2947	0.8628	0.2

Regression with Categorical Independent Variables

- How to put “male” and “female” categories in a regression equation?
  – Code them as indicator (dummy) variables
- Two ways of making dummy variables:
  – Male = 1, female = 0
    - Default in many programs
  – Male = 1, female = -1
    - *Default in JMP for nominal variables*

Coding Examples

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	80.405405	1.891707	42.50	<.0001*
Male_Ind	-0.475173	2.580267	-0.18	0.8544

0/1 coding

Compares calc_grade to a baseline group

Regression equation:

females: calc_grade=80.41 - 0.48 $\times$ 0

males: calc_grade=80.41 – 0.48 $\times$ 1

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	80.167819	1.290133	62.14	<.0001*
Gender[F]	0.2375864	1.290133	0.18	0.8544

-1/1 coding

Compares each group to overall average

Regression equation:

females: calc_grade=80.18 + 0.24 $\times$1

males: calc_grade=80.18 + 0.24 $\times$ (-1)

How to Code *k* Levels

- Two coding schemes: 0/1 and 1/0/-1
  - Use k-1 indicator variables
- E.g., three level variable: “a,” “b,”, & “c”
- 0/1: use one of the levels as a baseline
  - ➢ Var_a = 1 if level=a, 0 otherwise
  - ➢ Var_b = 1 if level=b, 0 otherwise
  - ➢ Var_c – *exclude as redundant (baseline)*
    - Example:

Variable	Var_a	Var_b
a	1	0
b	0	1
c	0	0
a	1	0
c	0	0
b	0	1
b	0	1

How to Code *k* Levels (cont'd)

- 1/0/-1: use the mean as a baseline
  - Variable[a] = 1 if variable=a, 0 if variable=b, -1 if variable=c
  - Variable[b] = 1 if variable=b, 0 if variable=a, -1 if variable=c
  - Variable[c] – *exclude as redundant*
    - Example

Variable	Variable[a]	Variable[b]
a	1	0
b	0	1
c	-1	-1
a	1	0
c	-1	-1
b	0	1
b	0	1

If Assumptions Met…

- ...can use regression to do the usual inference
  - Hypothesis tests on the slope and intercept
  - R-squared (fraction in the variation of y explained by x)
  - Confidence and prediction intervals, etc.
- However, one (usually unstated) assumption is data comes from a SRS…

Regression in Complex Surveys

- Problem:
  - Sample designs with unequal probability of section will likely result in incorrectly estimated slope(s)
  - If design involves clustering, standard errors will likely be wrong (too small)
- We won’t go into analytical details here
  - See Lohr chapter 11 if interested
- Solution: Use software (not JMP) that appropriately accounts for sample design
  - More at the end of the next lecture

A Note on Weights and Weighted Least Squares

- “Weighted least squares” often discussed in statistics textbooks as a remedy for unequal variances
  - Weights used are <u>not</u> the same as sampling weights previously discussed
- Some software packages also allow use of “weights” when fitting regression
  - Generally, these are “frequency weights” – again not the same as survey sampling weights
- Again, for complex designs, use software designed for complex survey analysis

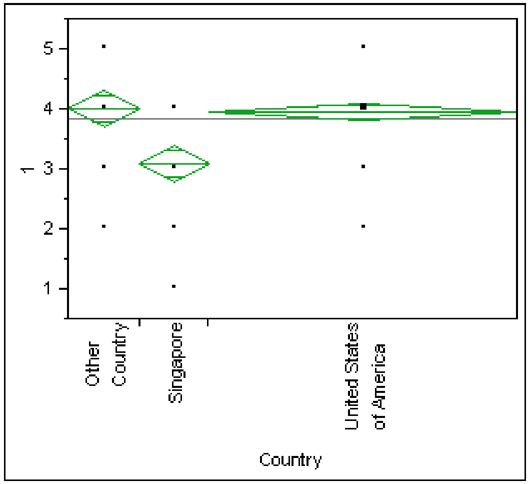
Population vs. Sample

- Sometimes have a census of data: can regression still be used?
  - Yes, as a way to <u>summarize</u> data
- I.e., statistical inference from sample to population no longer relevant
- But regression can be a parsimonious way to summarize relationships in data
  - Must still meet linearity assumption

Regression in JMP

- In JMP, use Analyze > Fit Model to do multiple regression
  - Fill in Y with (continuous) dependent variable
  - Put Xs in model by highlighting and then clicking “Add”
    - Use “Remove” to take out Xs
  - Click “Run Model” when done
- Takes care of missing values and non-numeric data automatically

From NPS New Student Survey:
Q1 by Country – ANOVA vs. Regression



Analysis of Variance

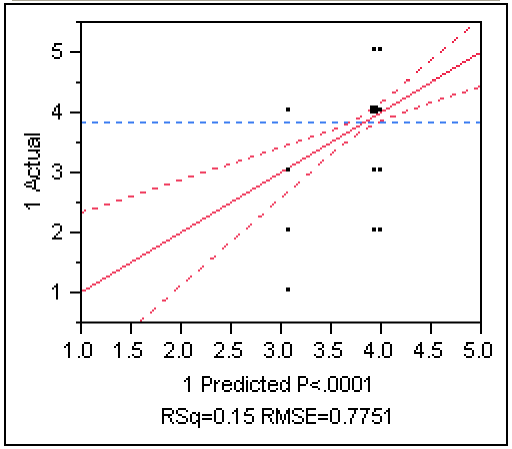
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Country	2	16.29297	8.14648	13.5610	<.0001*
Error	159	95.51568	0.60073		
C. Total	161	111.80864			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Other Country	26	4.00000	0.15200	3.6998	4.3002
Singapore	25	3.08000	0.15501	2.7738	3.3862
United States of America	111	3.94595	0.07357	3.8007	4.0912

Std Error uses a pooled estimate of error variance

Actual by Predicted Plot



Summary of Fit

RSquare	0.145722
RSquare Adj	0.134976
Root Mean Square Error	0.775066
Mean of Response	3.820988
Observations (or Sum Wgts)	162

Parameter Estimates

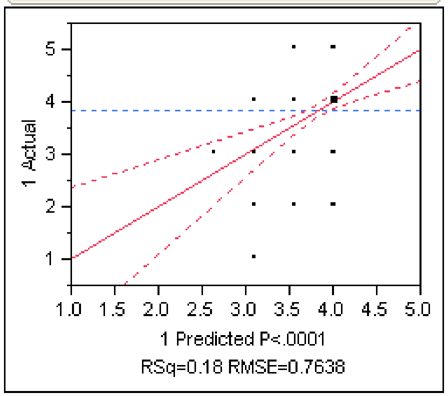
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.6753153	0.07641	48.10	<.0001*
Country[Other Country]	0.3246847	0.116362	2.79	0.0059*
Country[Singapore]	-0.595315	0.117678	-5.06	<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Country	2	2	16.292966	13.5610	<.0001*

From NPS New Student Survey: Q1 by Country and Gender

Actual by Predicted Plot



Summary of Fit

RSquare	0.175589
RSquare Adj	0.159936
Root Mean Square Error	0.763802
Mean of Response	3.820988
Observations (or Sum Wgts)	162

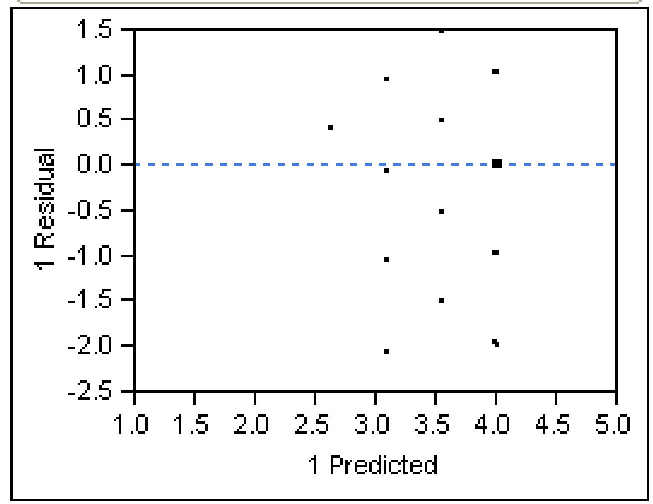
Parameter Estimates

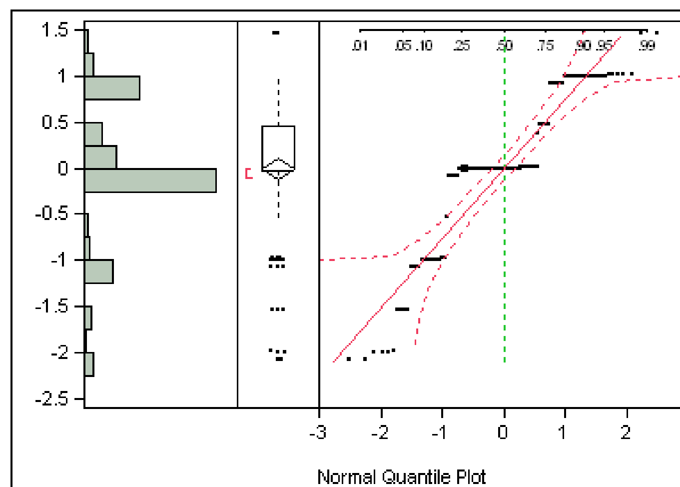
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.4721766	0.113486	30.60	<.0001*
Country[Other Country]	0.2946606	0.115355	2.55	0.0116*
Country[Singapore]	-0.606686	0.116065	-5.23	<.0001*
Sex[F]	-0.233163	0.097456	-2.39	0.0179*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Country	2	2	17.545942	15.0378	<.0001*
Sex	1	1	3.339396	5.7241	0.0179*

Residual by Predicted Plot





Regress Q1 on Country, Sex, Race, Branch, Rank, and CurricNumber

Summary of Fit

RSquare	0.11618
RSquare Adj	-0.01271
Root Mean Square Error	0.753535
Mean of Response	3.945946
Observations (or Sum Wgts)	111

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	3.5760051	0.190575	18.76	<.0001*
Sex[F]		-0.253621	0.106384	-2.38	0.0191*
Race[Asian American/Pacific Islander]		-0.172698	0.289673	-0.60	0.5525
Race[Black/African American]		0.0312715	0.239935	0.13	0.8966
Race[Hispanic/Latinos]		0.2333162	0.241025	0.97	0.3355
Race[Unknown]		-0.157008	0.272681	-0.58	0.5661
Military Branch[CIV]	Biased	-0.186943	0.596779	-0.31	0.7548
Military Branch[USA]	Biased	0.273696	0.278187	0.98	0.3277
Military Branch[USAF]	Biased	-0.471716	0.385382	-1.22	0.2239
Military Branch[USMC]	Biased	0.192187	0.315844	0.61	0.5443
Mil Rank[CIV]	Biased	0.0049813	0.339662	0.01	0.9883
Mil Rank[Junior]	Biased	0.1060277	0.232463	0.46	0.6493
Mil Rank[Mid]	Zeroed	0	0	.	.
CurricNumber[GSBPP]		0.0317088	0.149738	0.21	0.8327
CurricNumber[GSEAS]		-0.013023	0.132597	-0.10	0.9220
CurricNumber[GSOIS]		-0.144275	0.145385	-0.99	0.3235

Analysis of Variance

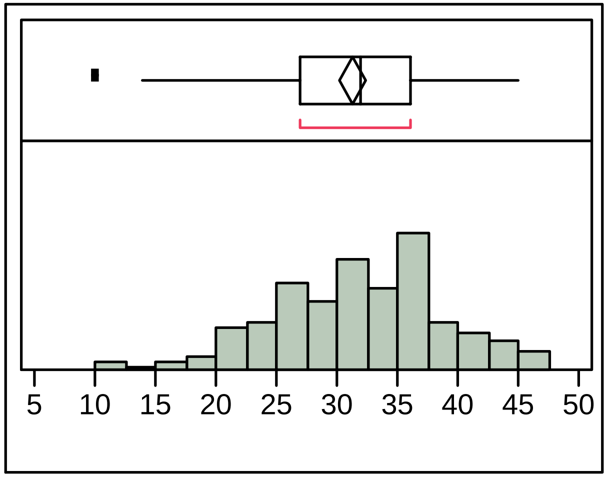
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	7.165506	0.511822	0.9014
Error	96	54.510169	0.567814	Prob > F
C. Total	110	61.675676		0.5598

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
Country	0	0	0.0000000	.	.	
Sex	1	1	3.2271699	5.6835	0.0191*	
Race	4	4	0.8145052	0.3586	0.8375	
Military Branch	4	3	1.2543237	0.7363	0.5329	LostDFs
Mil Rank	3	2	0.4174921	0.3676	0.6933	LostDFs
CurricNumber	3	3	0.7473494	0.4387	0.7258	

Make and Analyze a New Variable

- “In-processing Total” = sum(Q2a-Q2i)



Moments

Mean	31.290323
Std Dev	7.1523021
Std Err Mean	0.5744867
upper 95% Mean	32.425214
lower 95% Mean	30.155431
N	155

Quantiles

100.0%	maximum	45.000
99.5%		45.000
97.5%		45.000
90.0%		40.000
75.0%	quartile	36.000
50.0%	median	32.000
25.0%	quartile	27.000
10.0%		21.600
2.5%		14.900
0.5%		10.000
0.0%	minimum	10.000

Satisfaction with In-processing (1)

GSEAS worst at in-processing?

Summary of Fit

RSquare	0.053467
RSquare Adj	0.034662
Root Mean Square Error	7.027252
Mean of Response	31.29032
Observations (or Sum Wgts)	155

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	31.53424	0.594178	53.07	<.0001*
CurricNumber[GSBPP]	2.0309773	1.194393	1.70	0.0911
CurricNumber[GSEAS]	-2.555979	0.943298	-2.71	0.0075*
CurricNumber[GSOIS]	0.7865147	0.904941	0.87	0.3862

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
CurricNumber	3	3	421.21242	2.8432	0.0398*

Or are CIVs and USAF least happy?

Summary of Fit

RSquare	0.150805
RSquare Adj	0.090148
Root Mean Square Error	7.249163
Mean of Response	31.50943
Observations (or Sum Wgts)	106

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	27.285265	1.516813	17.99	<.0001*
CurricNumber[GSBPP]	2.353132	1.443114	1.63	0.1062
CurricNumber[GSEAS]	-0.953094	1.240644	-0.77	0.4442
CurricNumber[GSOIS]	-0.342795	1.310007	-0.26	0.7941
Military Branch[CIV]	-11.48528	4.356541	-2.64	0.0097*
Military Branch[USA]	6.4123699	2.348524	2.73	0.0075*
Military Branch[USAF]	-1.929405	3.62882	-0.53	0.5961
Military Branch[USMC]	2.0576925	2.846399	0.72	0.4715

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
CurricNumber	3	3	144.87868	0.9190	0.4347
Military Branch	4	4	693.78518	3.3006	0.0139*

Satisfaction with In-processing (2)

Or are Singaporians unhappy?

Summary of Fit

RSquare	0.032938
RSquare Adj	0.020214
Root Mean Square Error	7.079645
Mean of Response	31.29032
Observations (or Sum Wgts)	155

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	31.017034	0.71228	43.55	<.0001*
Country[Other Country]	1.9829665	1.084264	1.83	0.0694
Country[Singapore]	-2.475367	1.097029	-2.26	0.0255*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Country	2	2	259.48658	2.5886	0.0784

Making a new variable…

Summary of Fit

RSquare	0.157605
RSquare Adj	0.105319
Root Mean Square Error	6.76519
Mean of Response	31.29032
Observations (or Sum Wgts)	155

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	28.24856	1.045547	27.02	<.0001*
Type Student[Other FORNAT]	4.3767518	1.576432	2.78	0.0062*
Type Student[Singapore]	1.5252172	1.686527	0.90	0.3673
Type Student[US Air Force]	-3.354183	3.509584	-0.96	0.3408
Type Student[US Army]	5.0781409	1.931596	2.63	0.0095*
Type Student[US Marine Corps]	-0.142772	2.614642	-0.05	0.9565
Type Student[US Navy]	4.1157705	1.23088	3.34	0.0011*
CurricNumber[GSBPP]	1.7437146	1.188619	1.47	0.1445
CurricNumber[GSEAS]	-2.042984	1.016537	-2.01	0.0463*
CurricNumber[GSOIS]	0.1193437	0.948678	0.13	0.9001

Effect Tests

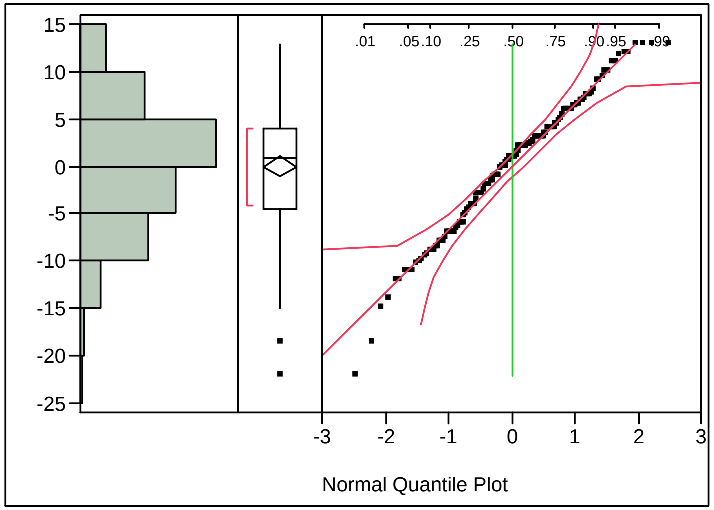
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Type Student	6	6	820.39209	2.9875	0.0088*
CurricNumber	3	3	212.44201	1.5472	0.2049

Satisfaction with In-processing (3)

- Final model?

Summary of Fit

RSquare	0.130639
RSquare Adj	0.095394
Root Mean Square Error	6.802609
Mean of Response	31.29032
Observations (or Sum Wgts)	155



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	28.052194	1.036317	27.07	<.0001*
Type Student[Other FORNAT]	4.9478063	1.547937	3.20	0.0017*
Type Student[Singapore]	0.489473	1.565631	0.31	0.7550
Type Student[US Air Force]	-3.71886	3.477345	-1.07	0.2866
Type Student[US Army]	5.614473	1.8104	3.10	0.0023*
Type Student[US Marine Corps]	0.2335206	2.407476	0.10	0.9229
Type Student[US Navy]	3.985781	1.22162	3.26	0.0014*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Type Student	6	6	1029.1625	3.7067	0.0018*

What We Have Just Learned

- Linear regression
  - How to think about it for Lickert scale dependent variables
  - Coding nominal independent variables
- Linear regression for complex surveys
- Weighting
- Regression in JMP