ICD EPSY 8251 Workshop

Christopher David Desjardins

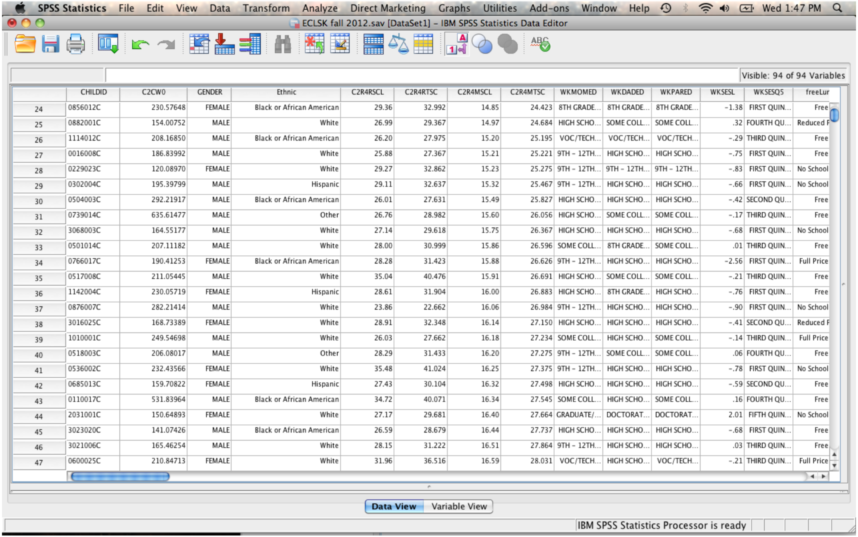
31 August 2012

Overview

Introduction to SPSS

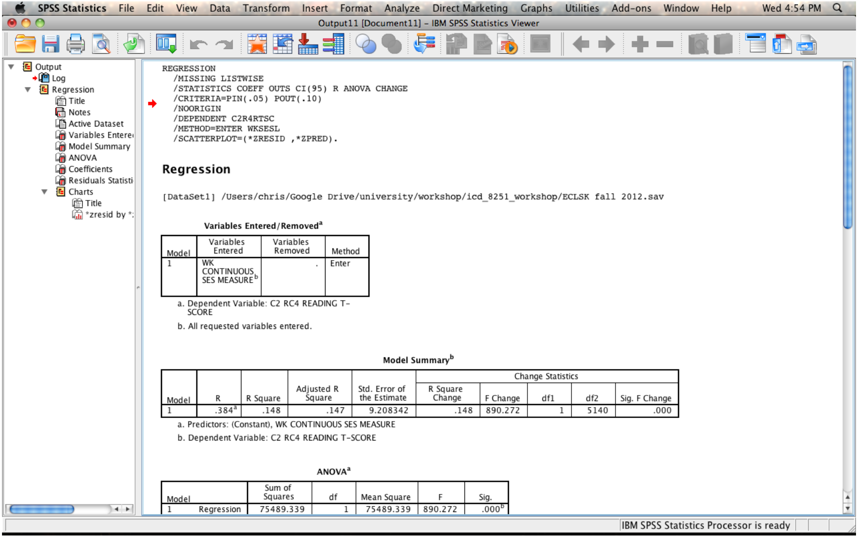
- Goals:
  - Understand the different elements of SPSS
    - Data Editor, Output Viewer, and Syntax Editor
    - Different menus
  - Be able to load data and create graphs

Data Editor

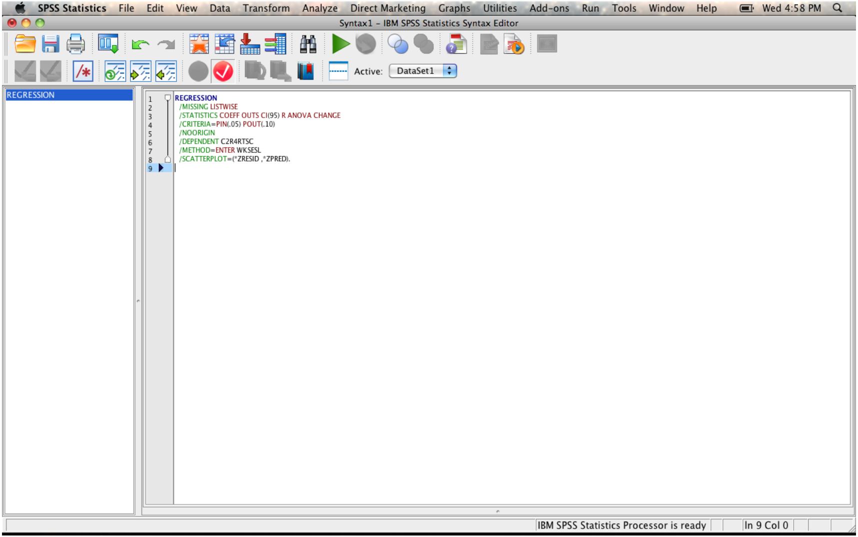


	CHILDID	C2CW0	GENDER	Ethnic	C2R4RSCL	C2R4RTSC	C2R4MSCL	C2R4MTSC	WKMOMED	WKDADED	WKPARED	WKSESL	WKSESQ5	freeLur
24	0856012C	230.57648	FEMALE	Black or African American	29.36	32.992	14.85	24.423	8TH GRADE...	8TH GRADE...	8TH GRADE...	-1.38	FIRST QUIN...	Free
25	0882001C	154.00752	MALE	White	26.99	29.367	14.97	24.684	HIGH SCHO...	SOME COLL...	SOME COLL...	.32	FOURTH QU...	Reduced F
26	1114012C	208.16850	MALE	Black or African American	26.20	27.975	15.20	25.195	VOC/TECH...	VOC/TECH...	VOC/TECH...	-.29	THIRD QUIN...	Free
27	0016008C	186.83992	MALE	White	25.88	27.367	15.21	25.221	9TH - 12TH...	HIGH SCHO...	HIGH SCHO...	-.75	FIRST QUIN...	Free
28	0229023C	120.08970	FEMALE	White	29.27	32.862	15.23	25.275	9TH - 12TH...	9TH - 12TH...	9TH - 12TH...	-.83	FIRST QUIN...	No School
29	0302004C	195.39799	MALE	Hispanic	29.11	32.637	15.32	25.467	9TH - 12TH...	HIGH SCHO...	HIGH SCHO...	-.66	FIRST QUIN...	No School
30	0504003C	292.21917	MALE	Black or African American	26.01	27.631	15.49	25.827	HIGH SCHO...	HIGH SCHO...	HIGH SCHO...	-.42	SECOND QU...	Free
31	0739014C	635.61477	MALE	Other	26.76	28.982	15.60	26.056	HIGH SCHO...	SOME COLL...	SOME COLL...	-.17	THIRD QUIN...	Free
32	3068003C	164.55177	MALE	White	27.14	29.618	15.75	26.367	HIGH SCHO...	HIGH SCHO...	HIGH SCHO...	-.68	FIRST QUIN...	No School
33	0501014C	207.11182	MALE	White	28.00	30.999	15.86	26.596	SOME COLL...	8TH GRADE...	SOME COLL...	.01	THIRD QUIN...	Free
34	0766017C	190.41253	FEMALE	Black or African American	28.28	31.423	15.88	26.626	9TH - 12TH...	HIGH SCHO...	HIGH SCHO...	-2.56	FIRST QUIN...	Full Price
35	0517008C	211.05445	MALE	White	35.04	40.476	15.91	26.691	HIGH SCHO...	SOME COLL...	SOME COLL...	-.21	THIRD QUIN...	Free
36	1142004C	230.05719	FEMALE	Hispanic	28.61	31.904	16.00	26.883	HIGH SCHO...	8TH GRADE...	HIGH SCHO...	-.76	FIRST QUIN...	Free
37	0876007C	282.21414	MALE	White	23.86	22.662	16.06	26.984	9TH - 12TH...	HIGH SCHO...	HIGH SCHO...	-.90	FIRST QUIN...	No School
38	3016025C	168.73389	FEMALE	White	28.91	32.348	16.14	27.150	HIGH SCHO...	HIGH SCHO...	HIGH SCHO...	-.41	SECOND QU...	Reduced F
39	1010001C	249.54698	MALE	White	26.03	27.662	16.18	27.234	SOME COLL...	HIGH SCHO...	SOME COLL...	-.14	THIRD QUIN...	Full Price
40	0518003C	206.08017	MALE	Other	28.29	31.433	16.20	27.275	9TH - 12TH...	SOME COLL...	SOME COLL...	.06	FOURTH QU...	Free
41	0536002C	232.43566	FEMALE	White	35.48	41.024	16.25	27.375	9TH - 12TH...	HIGH SCHO...	HIGH SCHO...	-.78	FIRST QUIN...	No School
42	0685013C	159.70822	FEMALE	Hispanic	27.43	30.104	16.32	27.498	HIGH SCHO...	HIGH SCHO...	HIGH SCHO...	-.59	SECOND QU...	Free
43	0110017C	531.83964	MALE	Black or African American	34.72	40.071	16.34	27.545	SOME COLL...	HIGH SCHO...	SOME COLL...	.16	FOURTH QU...	Free
44	2031001C	150.64893	FEMALE	White	27.17	29.681	16.40	27.664	GRADUATE/...	DOCTORAT...	DOCTORAT...	2.01	FIFTH QUIN...	No School
45	3023020C	141.07426	MALE	Black or African American	26.59	28.679	16.44	27.737	HIGH SCHO...	HIGH SCHO...	HIGH SCHO...	-.68	FIRST QUIN...	Free
46	3021006C	165.46254	MALE	White	28.15	31.222	16.51	27.864	9TH - 12TH...	HIGH SCHO...	HIGH SCHO...	.03	THIRD QUIN...	Free
47	0600025C	210.84713	FEMALE	White	31.96	36.516	16.59	28.031	VOC/TECH...	HIGH SCHO...	VOC/TECH...	-.21	THIRD QUIN...	Full Price

Output Viewer



Syntax Editor



File formats

- Data are saved in *.sav format.
- Output are saved in *.spv format.
- Syntax is saved in *.sps format
- You can directly save as Excel format or CSV
- For **R**, you can use the `foreign` library and the `read.spss()` function or save as a CSV and use the `read.csv()` command.

Variable View Fields

- `Name` and `Label`, descriptors of the variable.
- `Values`, value assigned to a level of the variable.
- `Type`, usually `Numeric` or `String`.
- `Width` and `Decimals`, affects the attributes of the cells in data view
- `Missing`, what is missing data coded as?
- `Columns` and `Align`, affects the attributes of the variable
- `Measure`
  - `Scale` - Interval or ratio scale
  - `Nominal` - Factor with no ordered levels
  - `Ordinal` - Factor with ordered levels

Menus in Data View

- For descriptives, regression, ANOVA, t-test, look in the `Analyze` menu.
- For graphing, look in the `Graphs` and `Analyze` menu.
- For data manipulation, look in the `Data` menu.

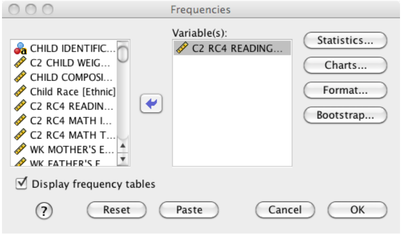
Data

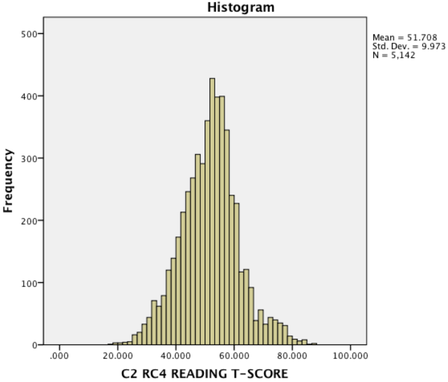
- The data used in this workshop (and throughout 8251) is from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99.
- The data focuses on children's early school experiences beginning with kindergarten and following children through middle school.
- The ECLS-K data provide descriptive information on children's status at entry to school, their transition into school, and their progression through 8th grade.
- The longitudinal nature of the ECLS-K data enables researchers to study how a wide range of family, school, community, and individual factors are associated with school performance.
- `http://nces.ed.gov/ecls/kindergarten.asp`

Graphs

- Histograms
- Stem-and-Leaf plot
- Boxplot
- Scatterplot

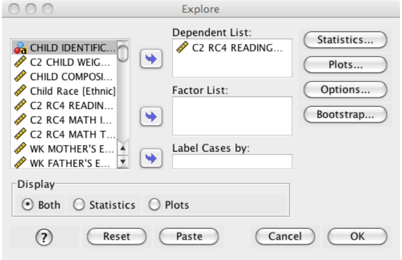
Analyze → Descriptive Statistics → Frequencies ...





Stem-and-Leaf plot

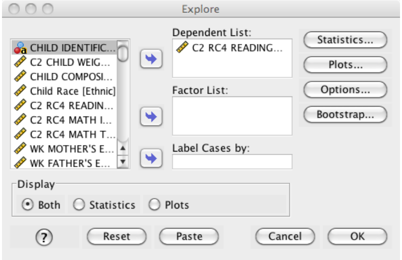
Analyze → Descriptive Statistics → Explore ...

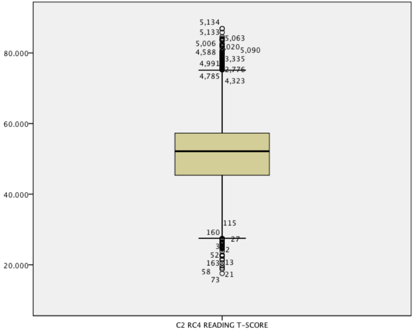



```
C2 RC4 READING T-SCORE Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    38.00 Extremes    (=<27)
    10.00        2 .  77
    37.00        2 .  888999
    59.00        3 .  0000111111
    82.00        3 .  2222222333333
    81.00        3 .  4444445555555
   126.00        3 .  666666666777777777777
   166.00        3 .  8888888888888999999999999999
   212.00        4 .  00000000000000000011111111111111111
   265.00        4 .  22222222222222222222333333333333333333333333
   311.00        4 .  444444444444444444444444444445555555555555555555555
   355.00        4 .  66666666666666666666666666666677777777777777777777777777777
   355.00        4 .  88888888888888888888888888888999999999999999999999999999999
   436.00        5 .  000000000000000000000000000000000011111111111111111111111111111111111111
   513.00        5 .  222222222222222222222222222222222222222222223333333333333333333333333333333333333333
3
   462.00        5 .  44444444444444444444444444444444444444444445555555555555555555555555555555555
   468.00        5 .  666666666666666666666666666666666666666666667777777777777777777777777777777777
   290.00        5 .  888888888888888888888888888889999999999999999999
   261.00        6 .  0000000000000000000000000111111111111111111
   126.00        6 .  222222222233333333333
   149.00        6 .  444444444444455555555555
    49.00        6 .  66666777
    68.00        6 .  88888999999
    41.00        7 .  0000111
    49.00        7 .  22222333
    28.00        7 .  4444&
   105.00 Extremes    (>=75)

 Stem width:    10.000
 Each leaf:        6 case(s)


 & denotes fractional leaves.
```

Boxplot

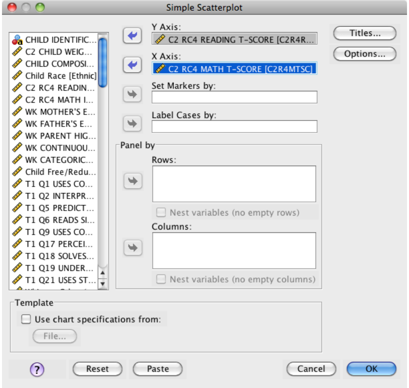
Analyze → Descriptive Statistics → Explore ...

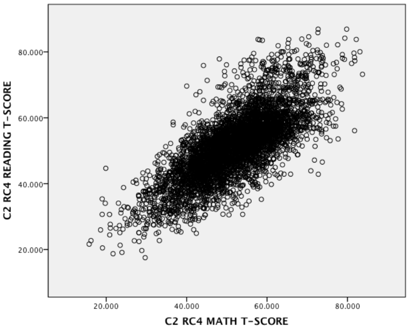




Scatterplot

Graphs → Legacy Dialogs → Scatter/Dot ...





Graphing in SPSS

- Spend the next 10 minutes exploring the ECLSK data with graphs.
- See if you can recreate the graphs just presented.

EDA

- Goals:
  - To understand, to be able to calculate by hand, and to be able to have SPSS provide:
    - Measures of central tendency and variability
    - Understand how to calculate the expectation and variance of a RV
    - Covariance
    - Correlation
  - To be able to connect the concepts of covariance and correlation

Measures of Central Tendency

- Mean, $\bar{x} = \sum_i^n \frac{y_i}{n}$
- Median, $\tilde{x} = P(X \leq m) \geq \frac{1}{2}$ and $P(X \geq m) \geq \frac{1}{2}$
- Mode, most frequent value in a data set
- What is the mean, median, and mode?

y	10	1	3	1	6

Measures of variability

- Sample variance (unbiased), $s^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$
- Standard deviation, $s = \sqrt{s^2}$
- What is s^2 and s?

y	7	7	9	2	1

Expectation of a random variable

- Let X be a random variable (rv)
- The expectation of a X is defined as:
  - Discrete case, $EX = \sum_x xf(x)$
  - Continuous case, $EX = \int_{-\infty}^{\infty} xf(x)dx$
- The EX is also known as the mean of X or the mean of the distribution of X
- For a normal distribution, $EX = \mu$ and for a binomial distribution, $EX = np$.

Variance of a rv

- The variance of a rv, $Var(X) = E[(X - EX)^2] = E[X^2] - (EX)^2$
- The variance of the normal distribution, $Var(X) = \sigma^2$ and the variance of the binomial distribution is $Var(X) = np(1 - p)$.
- We can then describe the normal distribution as, $X \sim N(\mu, \sigma^2)$, where μ refers to the mean and σ^2 refers to the variance..

Properties of Expectations

- **Property 1**. If $X = c$ and c is a constant rv, then $E(c) = c$ and $Var(c) = 0$.
- **Property 2**. Let a, b be constants and $Y = aX + b$, then $E(Y) = aE(X) + b$ and $Var(Y) = a^2 Var(X)$.
- **Property 3**. If $X_1, X_2, \cdots, X_n$ are independent rv, then $E[X_1, X_2, \cdots, X_n] = EX_1 * EX_2 \cdots EX_n$
- If the expectation of an estimator is equal to the population parameter of interest, then it is termed **unbiased**.
- Is the sample mean unbiased? note: $E(X_i) = \mu$ and X_i are independent and identically distributed.
  $E(\frac{\sum_i^n X_i}{n}) = \frac{1}{n}E(\sum_i^n X_i) = \frac{1}{n}\sum_i^n E(X_i) = \frac{1}{n}n\mu = \mu$
- This is why we divide by $n-1$ to calculate obtain an unbiased estimate of variance.

Covariance and correlation

- Covariance, $Cov(X, Y) = E[(X - EX)]E[(Y - EY)]$
  - In 8251, you'll mostly use, $Cov(X, Y) = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$
- Correlation, $\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$
- What is $Cov(X, Y)$ and ρ?

x	7	18	2	12	3
y	12	14	5	19	1

EDA in SPSS

- Spend the next 10 minutes, exploring the ECLSK data
- Make sure you know how to find descriptive statistics, covariances, correlations, and can create simple plots!
    - Analyze → Descriptive Statistics and Analyze → Correlate.
    - Graphs → Chart Builder and Graphs → Legacy Dialogs.

SLR

- Goals:
  - To be able to relate correlation to SLR
  - To understand the assumptions of SLR
  - To be able to calculate the intercept and slope
  - To be able to calculate the coefficient of determination
  - To be able to run a SLR in SPSS and interpret the output

Simple Linear Regression Model

SLR model

$$y|x = \beta_0 + \beta_1 x + e|x$$

- $E[y|x] = \beta_0 + \beta_1 x$, because $E[e|x] = 0$.
- The residuals, $e|x$, are distributed $N(0, \sigma^2)$.
- $\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\beta_1 = \frac{SXY}{SXX} = \rho \frac{sd(y)}{sd(x)}$, where $SXY = \sum_i^n (x_i - \bar{x})(y_i - \bar{y})$ and $SXX = \sum_i^n (x_i - \bar{x})^2$

x	7	18	2	12	3
y	12	14	5	19	1

What is β_0? What is β_1?

Assumptions of SLR

- Validity. The data you are analyzing should map to the research question you are trying to answer.
- The mean is a linear function of the predictor(s).
- Independence of residuals. That is that the residuals are uncorrelated with one another and are independent.
- Homogeneity of variance. Constant variance.
- Conditional distribution of $y|x$ is normal. Alternatively, the residuals are normally distributed. This is the weakest assumption.

Coefficient of determination

R^2 - Coefficient of determination

$1 - \frac{RSS}{SST}$, where $SST = \sum_i^n (y_i - \bar{y})^2$ and $RSS = \sum_i^n (y_i - \hat{y}_i)^2$

- Interpretation - The proportion of variability in Y that is explained by conditioning on X (i.e. the model).
- This value ranges from 0 to 1
- In SLR, R^2 is equal to the square of the correlation between Y and X.

R^2 - Coefficient of determination

$1 - \frac{RSS}{SST}$, where $SST = \sum_i^n (y_i - \bar{y})^2$ and $RSS = \sum_i^n (y_i - \hat{y}_i)^2$

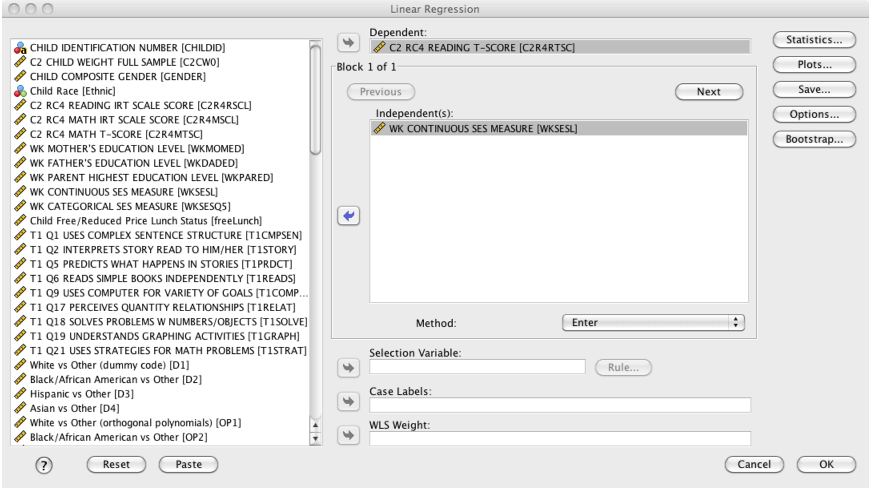
- Interpretation - The proportion of variability in Y that is explained by X (i.e. the model).
- This value ranges from 0 to 1
- In SLR, R^2 is equal to the square of the correlation between Y and X.
- What is SST, RSS, and R^2? What is $\sum_i^n y_i - \bar{y}$?

y_i	22	23	26	35	31
$\hat{y}_i$	28.5	26.67	26.67	28.5	26.67
$\bar{y}$	27.4				

Example SLR from ECLSK

- Response (i.e. dependent or outcome variable) is the Reading T-score.
- Predictor (i.e. independent variable or covariate) is SES as a continuous measure.
- $E[Reading|SES] = \beta_0 + \beta_1[SES]$
- `Analyze` $\rightarrow$ `Regression` $\rightarrow$ `Linear...`

Setting up the regression - SPSS



Output - SPSS

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.384[a]	.148	.147	9.208342

a. Predictors: (Constant), WK CONTINUOUS SES MEASURE

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	75489.339	1	75489.339	890.272	.000[b]
	Residual	435838.915	5140	84.794		
	Total	511328.253	5141			

a. Dependent Variable: C2 RC4 READING T-SCORE

b. Predictors: (Constant), WK CONTINUOUS SES MEASURE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	51.143	.130		394.002	.000
	WK CONTINUOUS SES MEASURE	5.173	.173	.384	29.837	.000

a. Dependent Variable: C2 RC4 READING T-SCORE

Regression and output - R

```
> m0=lm(C2R4RTSC~WKSESL,data=eclsk)
> summary(m0)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.1431     0.1298  394.00   <2e-16 ***
WKSESL        5.1729     0.1734   29.84   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.208 on 5140 degrees of freedom
Multiple R-squared: 0.1476, Adjusted R-squared: 0.1475
F-statistic: 890.3 on 1 and 5140 DF,  p-value: < 2.2e-16
```

SLR in SPSS

- Spend the next 10 minutes, exploring the ECLSK data
- Run a SLR with a continuous outcome and one predictor.
    - Interpret the output

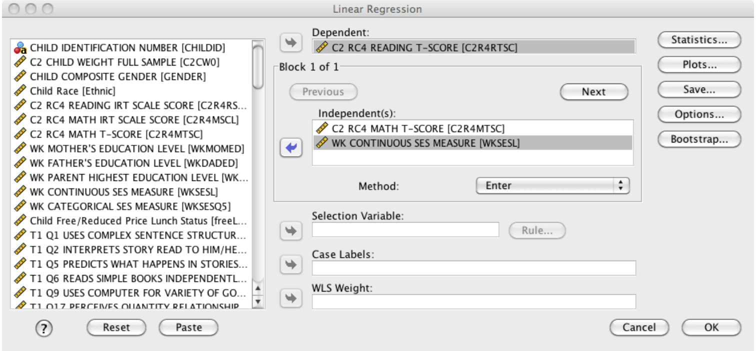
Regression and ANOVA

- Be able to run a multiple regression in SPSS and interpret the output
  - Using quantitative predictors and interactions.
  - Be able to check assumptions: Q-Q plot and residual plot
- Understand the relationship between ANOVA and MR
- Understand how to calculate a one-way ANOVA by hand
- Be able to run a one-way ANOVA in SPSS and interpret the output

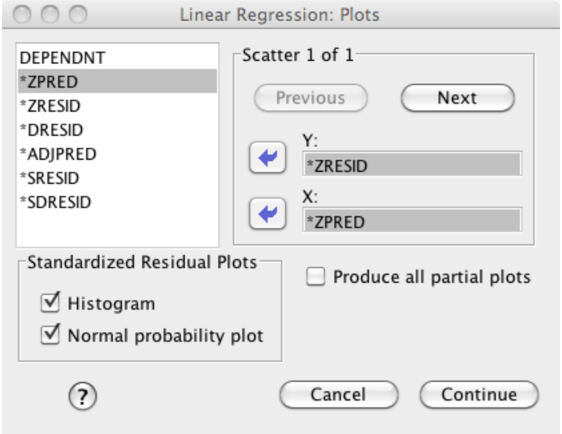
Multiple Regression in SPSS

- Running multiple regression in SPSS is done the same way as SLR.
- Solving for the βs is a little more difficult and most easily tackled with matrix algebra.
  - $Y = X\beta + e$
  - $\beta = (X^T X)^{-1} X^T Y$
  - You will not need to do this in 8251!
- Response (i.e. dependent or outcome variable) is the Reading T-score.
- Predictor (i.e. independent variable or covariate) is SES and Math T-score.
- $E[Reading|SES, MATH] = \beta_0 + \beta_1[SES] + \beta_2[MATH]$

Setting up the regression



Testing assumptions



Results

- $[\hat{Reading}|SES, MATH] = 13.864 + 1.410[SES] + .725[MATH]$
- What does 13.864 correspond to?
- How do we interpret 1.410? 0.725?
- Do we meet assumptions?

Interaction model

- Create an interaction term
  - `Transfrom` $\rightarrow$ `Compute Variable...`
- Refit model
- Is the interaction necessary?

Relationship between regression and ANOVA

- ANOVA (Analysis of Variance) examines whether groups differ in some outcome measure.
- The independent variable in an ANOVA is qualitative (categorical)
- A regression with a qualitative predictor with 2 levels will yield identical results to an ANOVA.
- Compare `Geneal Linear Model` → `Univariate`... to `Regression` → `Linear`... when using Reading T-Score as the DV and Gender as IV.
- So... ANOVA is a special case of regression.
- Using a predictor with more than 2 requires additional dummy variables.
- Also notice the relationship between T with 1 *df* and F.

Derivation of One-Way ANOVA

- See handout by Dr. Bob delMas

MR and ANOVA in SPSS

- Spend the next 10 minutes, exploring the ECLSK data
- Run a MR with two continuous outcomes and an interaction.
- Run a one-way ANOVA with Ethnic as the categorical predictor
- How many dummy variables would it take to make the regression output identical to the ANOVA output?

t-test and CIs

- Understand the relationship between a t-test to regression and ANOVA
- Be able to calculate the following t-test by hand and in SPSS
  - One-sample t-test
  - Two-sample t-test
    - Equal n and variance
    - Unequal n and equal variance
  - Be able to calculate by hand and in SPSS confidence intervals

Relationship between t-test and regression

- A one-sample t-test is equivalent to testing whether the intercept in a regression is equal to some value μ (often 0).
- A two-sample t-test is equivalent to testing whether the slope term in a regression for a qualitative variable is equal to some value μ (often 0).
- t-tests are used in regression to test whether parameters are significantly different than some value.
  - In SPSS, t is, by default, calculated using a μ of 0.
  - The distribution of a parameter estimate in a regression model under the null hypothesis has a t-distribution and this is why we calculate a t-statistic.

One-sample t-test

- $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$, where s is the sample standard deviation, $\bar{X}$ is the sample mean.
- $H_O = \mu$, the population mean is equal to μ
- $H_A \neq \mu$, the population mean is not equal to μ
- $df = n - 1$
- $T \sim T(df)$

Running a one-sample t-test

Suppose we have $X = \{2, 1, 2, 0, 2\}$ and we want to test whether μ is 0. What is H_O? What is T? What is df?

Is there evidence against the null hypothesis if the critical value for significance is 2.776?

Two-sample t-test - equal n and variance

- $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{2}(s^2_{X_1} + s^2_{X_2})} * \sqrt{\frac{2}{n}}}$
- $\sqrt{\frac{1}{2}(s^2_{X_1} + s^2_{X_2})}$ is the pooled standard deviation
- $df = n_1 + n_2 - 2$
- If $X_1 = \{4, 2, 3, 5, 3\}, X_2 = \{2, 1, 2, 0, 2\} \ldots$

  Test $H_O : \mu_1 = \mu_2$ against $H_A : \mu_1 \neq \mu_2$.

  Is there evidence against the null hypothesis if the critical value for significance is 2.306?

Two-sample t-test - unequal n

- $t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, where $S_{X_1X_2} = \sqrt{\frac{(n_1-1)s^2_{X_1}+(n_2-1)s^2_{X_2}}{n_1+n_2-2}}$
- $df = n_1 + n_2 - 2$
- If $X_1 = \{5, 4, 4, 5\}$ and $X_2 = \{3, 2, 2, 3, 3\}\ldots$

  Test $H_O : \mu_1 = \mu_2$ against $H_A : \mu_1 \neq \mu_2$.

  Is there evidence against the null hypothesis if the critical value for significance is 2.364?

Confidence Intervals

- CIs give identical information to hypothesis tests.
- If CIs don't overlap with our test value, we know it is significant at α
- Typically construct 95% CI
- For a Z-statistic
  - $z \pm 1.96 * SE$
  - Critical value, 1.96, corresponds to the 97.5^{th} quantile of a standard normal distribution
  - We use this when σ^2 is known. VERY UNCOMMON.
- For a t-statistic with a (1 - α)% CI
  - $\bar{X} \pm t_{\alpha/2} * SE$
  - Critical value, $t_{\alpha/2}$, corresponds to the $\alpha/2$ quantile of a t-distribution with df
  - Critical value found in books or R or let SPSS calculate it!

t-test and CI in SPSS

- Spend the next 10 minutes, exploring the ECLSK data
- Run a one-sample t-test and a two-sample t-test
- Make sure you request CIs.
- How are these calculated?

Wrap Up

- Today you learned . . .
  - SPSS
    - EDA
    - Simple Regression
    - Multiple Regression & ANOVA
    - T-tests and Confidence Intervals
    - Graphing & data management
  - Formulas
    - Measures of central tendency
    - Measures of variability & association
    - How to calculate a slope & intercept in SLR
    - How to perform a t-test and calculate confidence intervals

Now what?

- If this was a review, then relax and have a nice weekend
- If this was mostly a review, relax and have a nice weekend
- If this was all completely new, relax and have a nice weekend. But . . .
    - The more basic stuff (t-tests, confidence intervals, measures of variability/association/central tendency, hypothesis testing) you should spend some time to learn.
    - Don't feel overwhelmed
    - Dr. Harwell will review these things again. Remember this is just a primer!
    - Spend some time focusing on stuff that was especially difficult

How to succeed in stat courses

- Work in groups and help one another out
- Don't wait too long, seek out help from Dr. Harwell, the TA, or other students
- You will learn `R` in 8252. Take some time over the semester, if you have time, to familiarize yourself with it.
- Stop by 164 ICD if you have questions.