Elements of statistics (MATH0487-1)

Prof. Dr. Dr. K. Van Steen

University of Liège, Belgium

December 10, 2012

Outline I

Outline II

Outline III

Outline IV

Outline V

- Homogeneity test

9. Introduction to Linear Regression
   - Correlation Analysis
   - Simple Linear Regression
   - Centering
   - Inference on Regression Coefficients
   - Checking Model Assumptions
   - Relation between Experience and Wage

10. Frequently Asked Questions

Testing Hypothesis with Table Data

Several Types of χ^2-tests:

- The following tests give rise to test statistics that follow a χ^2-distribution under their appropriate null (hypothesis)
  - Test of Goodness of fit
  - Test of independence
  - Test of homogeneity or (no) association

Recall: Properties of the χ^2-distribution

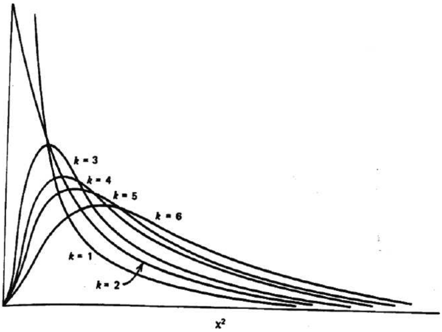
- Derived from the normal distribution

$$\begin{aligned} \chi_1^2 &= (\frac{y-\mu}{\sigma})^2 = Z^2 \\ \chi_k^2 &= Z_1^2 + Z_2^2 + \cdots + Z_k^2 \end{aligned}$$

  where $Z_1, \ldots, Z_k$ are all standard normal random variables

- k denotes the degrees of freedom
- A χ_k^2 random variable has
  - mean = k
  - variance = 2k
- Since a normal random variable can take on values in the interval $(-\infty, \infty)$, a chi-square random variable can take on values in the interval $(0, \infty)$

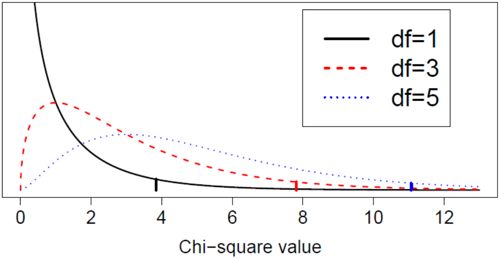
A family of χ^2-distributions



k = degrees of freedom

Critical values of χ^2

- We generally use only a one-sided test for the χ^2 distribution
- Area under the curve to the right of the cutoff for each curve is 0.05
- Increasing critical value with increasing number of degrees of freedom



The χ^2-table

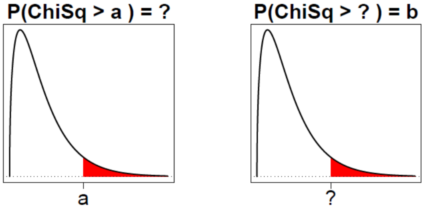
							u							
d	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995	.999
1	0.0^4393[b]	0.0^3157[c]	0.0^3982[d]	0.00393	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.0100	0.0201	0.0506	0.103	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.81
3	0.0717	0.115	0.216	0.352	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

[a] $\chi^2_{d,u}$ = uth percentile of a χ^2 distribution with d degrees of freedom.
[b] = 0.0000393
[c] = 0.000157
[d] = 0.000982

Source: Reproduced in part with permission of the Biometrika Trustees, from Table 3 of *Biometrika Tables for Statisticians*, Volume 2, edited by E. S. Pearson and H. O. Hartley, published for the Biometrika Trustees, Cambridge University Press, Cambridge, England, 1972.

For χ^2 random variables with degrees of freedom $=$ df we'll use

1. `pchisq(a, df, lower.tail=F)` to find $P(\chi_{df} \geq a) =?$



2. `qchisq(b, df, lower.tail=F)` to find $P(\chi_{df} \geq ?) = b$

The χ^2 goodness-of-fit test

Determine whether or not a sample of observed values of some random variable is compatible with the hypothesis that the sample was drawn from a population with a specified distributional form, i.e.

- Normal
- Binomial
- Poisson
- etc...

Here, the expected cell counts would be derived from the distributional assumption under the null hypothesis

The χ^2 goodness-of-fit test

$\chi^2 = \sum_{i=1}^{k}[\frac{(O_i - E_i)^2}{E_i}]$ where

- $O_i = i^{th}$ observed frequency
- $E_i = i^{th}$ expected frequency in the i^{th} cell of a table

- Degrees of freedom $= (\#$ categories $- 1)$

Note: This test is based on frequencies (cell counts) in a table, not proportions

Example: Handgun survey

- Survey 200 adults regarding handgun bill:
  - Statement: "I agree with a ban on handguns"
  - Four categories: Strongly agree, agree, disagree, strongly disagree

- Can one conclude that opinions are equally distributed over four responses?

Example: Handgun survey

Response (count)	1 Strongly agree	2 agree	3 disagree	4 Strongly disagree
Responding (O_i) Expected (E_i)	102 50	30 50	60 50	8 50

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right] \\
&= \frac{(102-50)^2}{50} + \frac{(30-50)^2}{50} + \frac{(60-50)^2}{50} + \frac{(8-50)^2}{50} \\
&= 99.36
\end{aligned}
$$

$\text{df} = 4 - 1 = 3$

Example: Handgun survey

- Critical value: $\chi^2_{4-1,0.05} = \chi^2_{3,0.05} = 7.81$
- Since $99.36 > 7.81$, we conclude that our observation was unlikely by chance alone ($p < 0.05$)
- Based on these data, opinions do not appear to be equally distributed among the four responses

The χ^2 test of independence

- Test the null hypothesis that two criteria of classification are independent
- $r \times c$ contingency table

		Criterion 1					
		1	2	3	$\cdots$	c	Total
	1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2c}	$n_{2.}$
Criterion 2	3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3c}	$n_{3.}$
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	n_{r3}	$\cdots$	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$\cdots$	$n_{.c}$	n

The χ^2 test of independence

- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} [\frac{(O_i - E_i)^2}{E_i}]$$

- Degrees of freedom $= (r-1)(c-1)$
  where r is the number of rows and c is number of columns
- Assume the marginal totals are fixed

The χ^2 test of no association (homogeneity)

- Test the null hypothesis that the samples are drawn from populations that are homogenous with respect to some factor
  - i.e. no association between group and factor
- Same test statistic as χ^2 test of independence
- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} [\frac{(O_i - E_i)^2}{E_i}]$$

- Degrees of freedom $= (r-1)(c-1)$
  where r is the number of rows and c is number of columns

Example: treatment response

		Response to Treatment		
	Treatment	Yes	No	Total
Observed Numbers	A	37	13	50
	B	17	53	70
	Total	54	66	120

- Test H_0 that there is no association between the treatment and response
- Calculate what numbers of "Yes" and "No" would be expected assuming the probability of "Yes" was the same in both treatment groups
- Condition on total the number of "Yes" and "No" responses

Example: treatment response

- Expected proportion with "Yes" response $= \frac{54}{120} = 0.45$
- Expected proportion with "No" response $= \frac{66}{120} = 0.55$

		Response to Treatment		
	Treatment	Yes	No	Total
Observed	A	37 (22.5)	13 (27.5)	50
(Expected)	B	17 (31.5)	53 (38.5)	70
	Total	54	66	120

- Get expected number of Yes responses on treatment A:

$$\frac{54}{120} \times 50 = 0.45 \times 50 = 22.5$$

- Using a similar approach you get the other expected numbers

Example: treatment response

$$\begin{aligned}
\text{Test statistic: } \chi^2 &= \sum_{i=1}^{k}[\frac{(O_i - E_i)^2}{E_i}] \\
&= \frac{(37-22.5)^2}{22.5} + \frac{(13-27.5)^2}{27.5} \\
&\quad + \frac{(17-31.5)^2}{31.5} + \frac{(53-38.5)^2}{38.5} \\
&= 29.1
\end{aligned}$$

- Degrees of freedom = (r-1)(c-1) = (2-1)(2-1) = 1
- Critical value for $\alpha = 0.001$ is 10.82 so we see $p < 0.001$
- Reject the null hypothesis, and conclude that the treatment groups are not homogenous (similar) with respect to response
- Response appears to be associated with treatment

Quanitfying Associations

Goal: Express the strength of the relationship between two variables

- Metric depends on the nature of the variables
- For now, we'll focus on continuous variables (e.g. height, weight)
- Important! **association does not imply causation**

To describe the relationship between two continuous variables, use:

- Correlation analysis
  - Measures *strength* and *direction* of the linear relationship between two variables
- Regression analysis
  - Concerns prediction or estimation of outcome variable, based on value of another variable (or variables)

Correlation Analysis

- Plot the data (or have a computer to do so)
- Visually inspect the relationship between two continous variables
- Is there a linear relationship (correlation)?
- Are there outliers?
- Are the distributions skewed?

Correlation Coefficients

- Measures the strength and direction of the **linear** relationship between two variables X and Y
- Population correlation coefficient:

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\cdot\mathrm{var}(Y)}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]\cdot \mathrm{E}[(Y-\mu_Y)^2]}}$$

- Sample correlation coefficient:
  (obtained by plugging in sample estimates)

$$r = \frac{\text{sample } \mathrm{cov}(X,Y)}{\sqrt{s_X^2\cdot s_Y^2}} = \frac{\sum_{i=1}^{n}\frac{(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}}{\sqrt{\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1}\cdot\sum_{i=1}^{n}\frac{(Y_i-\bar{Y})^2}{n-1}}}$$

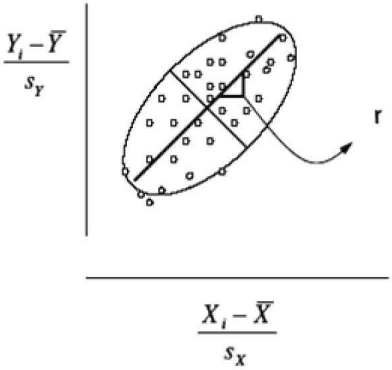
Correlation Coefficients

The correlation coefficient, ρ, takes values between -1 and $+1$

- -1: Perfect negative linear relationship
- 0: No linear relationship
- $+1$: Perfect positive relationship

Correlation Coefficients

- Plot standardized Y versus standardized X
- Observe an ellipse (elongated circle)
- Correlation is the slope of the major axis

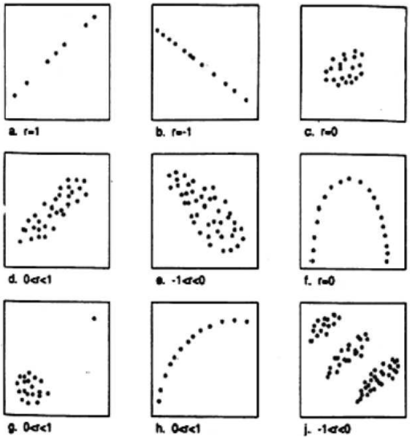


To Remember

- Other names for r
  - Pearson correlation coefficient
  - Product moment of correlation

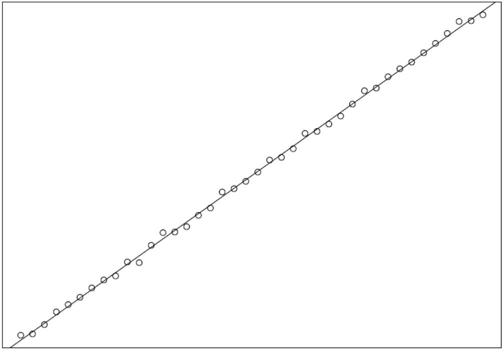
- Characteristics of r
  - Measures *linear* association
  - The value of r is independent of units used to measure the variables
  - The value of r is sensitive to outliers
  - r^2 tells us what proportion of variation in Y is explained by linear relationship with X

To Remember



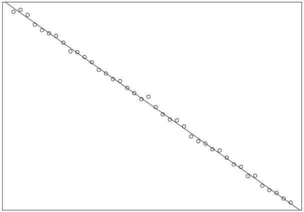
Examples

Perfect positive correlation, r $\approx$ 1

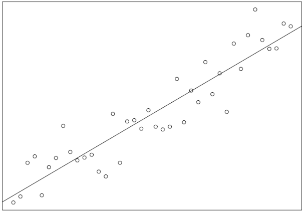


Examples

Perfect negative correlation, r ≈ -1

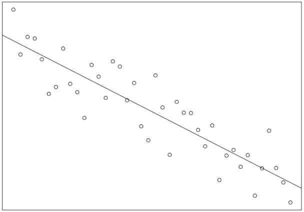


Imperfect positive correlation, 0< r <1

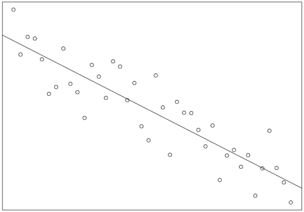


Examples

Imperfect negative correlation, -1<r <0

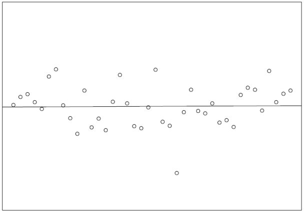


Imperfect negative correlation, -1<r <0

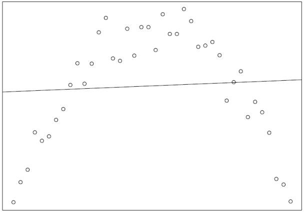


Examples

No relation, r $\approx$ 0



Some relation but little *linear* relationship, r $\approx$ 0



Association and Causality

- In general, association between two variables means there is some form of relationship between them
    - The relationship is not necessarily causal
    - Association does not imply causation, no matter how much we would like it to
- Example: Hot days, ice cream, drowning

Bradford Hill's Criteria for Causality

- Strength: magnitude of association
- Consistency of association: repeated observation of the association in different situations
- Specificity: uniqueness of the association
- Temporality: cause precedes effect
- Biologic gradient: dose-response relationship
- Biologic plausibility: known mechanisms
- Coherence: makes sense based on other known facts
- Experimental evidence: from designed (randomized) experiments
- Analogy: with other known associations

Simple Linear Regression (SLR)

Linear regression can be used to study a continuous outcome variable as a linear function of a predictor variable

Example: 60 cities in the US were evaluated for numerous characteristics, including:

Outcome variable (y) the % of the population with low income

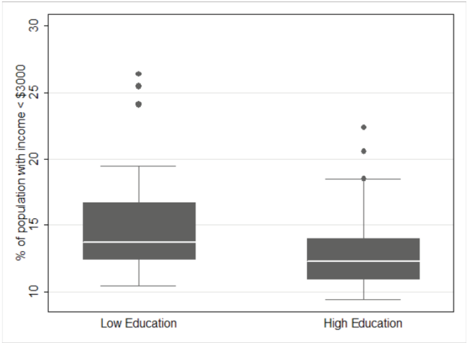
Predictor variable (x) median education level

Linear regression can help us to model the association between median education and % of the population with low income

Example

Boxplot of % low income by education level:

Education level is coded as a binary variable with values 'low' and 'high'



Simple Linear Regression: Regression Line

- Mean in low education group: 15.7%
- Mean in high education group: 13.2%

The two means could be compared by a t-test or ANOVA, but regression provides a unified equation:

$$\begin{aligned}\hat{y}_i &= \beta_0 + \beta_1 x_i \\ \hat{y}_i &= 15.7 - 2.5x_i\end{aligned}$$

where

- $x_i = 1$ for high education and 0 for low education (x is called a dummy variable or indicator variable that designates group)
- $\hat{y}_i$ is our estimate of the mean % low income for the given the value of education
- what about the β's?

Regression Analysis represented by Regression Line Equation

In simple linear regression, we use the equation for a line

$$y = mx + b$$

but we write it slightly differently:

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\begin{aligned} \beta_0 &= \text{y-intercept (value y when x=0)} \\ \beta_1 &= \text{slope of the line (rise/run)} \end{aligned}$$

Interpretation of Regression Model Components

$$\begin{aligned}\hat{y}_i &= \beta_0 + \beta_1 x_i \\ \hat{y}_i &= 15.7 - 2.5x_i\end{aligned}$$

- $x_i = 0$ (low education)

$$\begin{aligned}\hat{y}_i &= 15.7 - 2.5 \times 0 \\ &= 15.7 = \beta_0\end{aligned}$$

- $x_i = 1$ (high education)

$$\begin{aligned}\hat{y}_i &= 15.7 - 2.5 \times 1 \\ &= 13.2 = \beta_0 + \beta_1\end{aligned}$$

Interpretation of Regression Model Components

Intercept

- β_0 is the mean outcome for the **reference group**, or the group for which $x_i = 0$.
- Here, β_0 is the average percent of the population that is low income for cities with low education.

Slope

- β_1 is the **difference** in the mean outcome between the two groups (when $x_i = 1$ vs. when $x_i = 0$)
- Here, β_1 is **difference** in the average percent of the population that is low income for cities with high education compared to cities with low education.

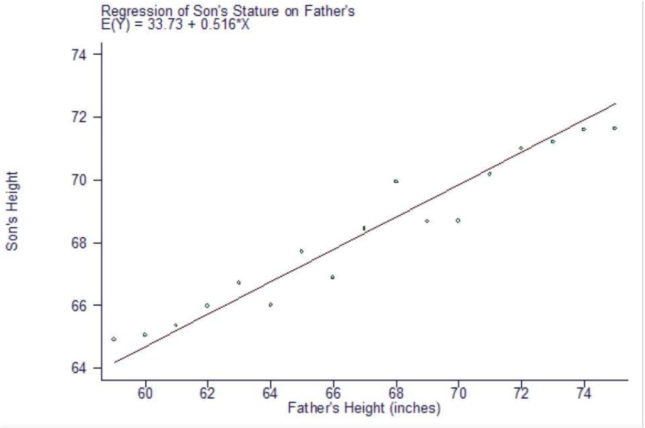
Why is Linear Regression so popular?

- Linear regression can refer to *simple* linear regression (one predictor) or *multiple* linear regression (more than 1 predictor)
- Linear regression naturally extends to quadratic, cubic ... regression to investigate curvilinear relationships
- Linear regression is flexible, since it can deal with
  - binary X
  - continuous X
  - categorical X
  - confounders
  - interactions (leading to k-order regression models)

Example: Galton's study on height

- 1000 records of heights of family groups
- Really tall fathers tend on average to have tall sons but not quite as tall as the really tall fathers
- Really short fathers tend on average to have short sons but not quite as short as the really short fathers
- There is a regression of a sons height toward the mean height for sons

Example: Galton's study on height



Regression Formalism I

- Probability model: Independent responses $y_1, y_2, \dots, y_n$ are sampled from
$$y_i \sim N(\mu_i, \sigma^2)$$

- Systematic model: $\mu_i = E(y_i|x_i) = \beta_0 + \beta_1 x_i$
where
$$\begin{aligned}\beta_0 &= \textit{intercept}\\ \beta_1 &= \textit{slope}\end{aligned}$$

Regression Formalism II

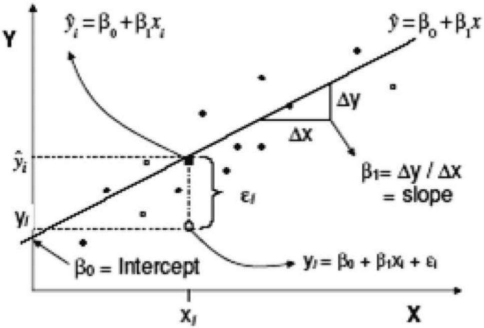
- Systematic: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Probability (random): $\epsilon_i \sim N(0, \sigma^2)$

- The response y_i is a linear function of x_i plus some random, normally distributed error, ϵ_i
- data $=$ signal $+$ noise

Regression Formalism: Model Assumptions

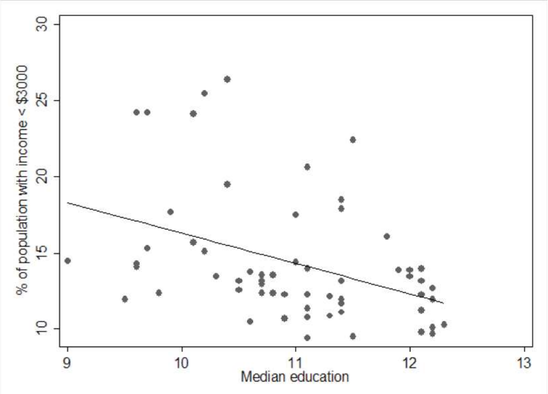
- The regression formalism naturally leads to four model assumptions:
  - The relationship is linear
  - The errors have the same variance
  - The errors are independent of each other
  - The errors are normally distributed
- When we *satisfy the assumptions*, it means that we have used all of the information available from the patterns in the data.
- When we *violate an assumption*, it usually means that there is a pattern to the data that we have not included in our model, and we could actually find a model that fits the data better.

Regression Formalism: Geometric Interpretation



Another example: City education versus income

When education is a continuous variable (not binary)



Another example: City education versus income

Using the continuous variable for median education in city i (x_i):

$$\begin{aligned} E(y_i|x_i) &= \beta_0 + \beta_1 x_i \\ E(y_i|x_i) &= 36.2 - 2.0x_i \end{aligned}$$

When $x_i = 0$

$$\begin{aligned} E(y_i|x_i) &= 36.2 - 2.0(0) \\ &= 36.2 = \beta_0 \end{aligned}$$

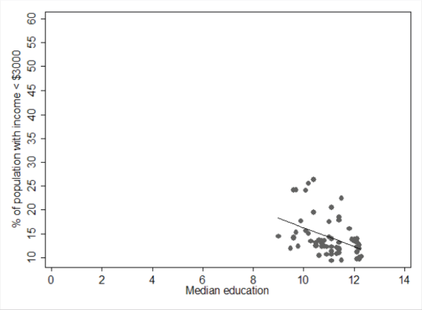
When $x_i = 1$

$$\begin{aligned} E(y_i|x_i) &= 36.2 - 2.0(1) \\ &= 34.2 = \beta_0 + \beta_1 \end{aligned}$$

When $x_i = 2$

$$\begin{aligned} E(y_i|x_i) &= 36.2 - 2.0(2) \\ &= 32.2 = \beta_0 + \beta_1 \times 2 \end{aligned}$$

Where is our Intercept?



The intercept isn't in the range of our observed data. This means:

- The intercept isn't very interpretable since the average of y when $x = 0$ was never observed
- Possible solution: we might want to *center* our x variable

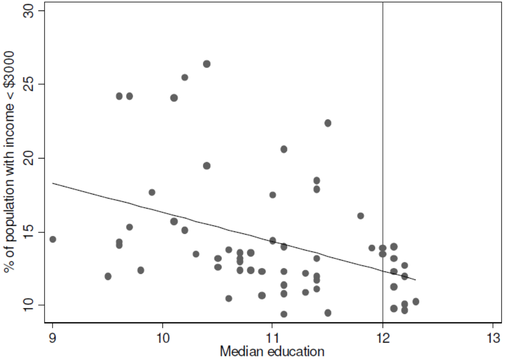
Need for Centering

As in the "City education versus income" example:

- β_0 makes no sense!
- We don't observe any cities with median education = 0
- We can change X to fix this problem by a process called **centering**
  1. Pick a value of X (c) within the range of the data
  2. For each observation, generate $X_{centered} = X_i - c$
  3. Redo the regression with $X_{centered}$

Centering

Use $c = 12$, as education level to center with:



Centering - New Regression Equation / Interpretation

$$\hat{Y}_i = \beta_0 + \beta_1 (X_{Centered\ i})$$

$$\hat{Y}_i = \beta_0 + \beta_1 (X_i - 12)$$

$$\hat{Y}_i = 12.2 - 2.0(X_i - 12)$$

- β_1 has not changed
- β_0 now corresponds to average of y when $X_{\text{centered } i}=0$ or, equivalently, $X_i=12$ **(not $X_i=0$)**
- Note: with $X_i=0$, we have

$$\hat{Y}_i = 12.2 - 2.0(0 - 12)$$
$$= 12.2 + 24 = 36.2$$

Using Sample Data to Estimate the Truth

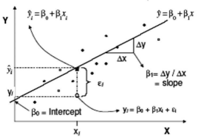
- So far, we have presented our **fitted** regression line as

$$\hat{y}_i = \beta_0 + \beta_1 x_i,$$

without having said anything about how to obtain the "best" such regression line.

Using Sample Data to Estimate the Truth

- Note: Sometimes linear regression is referred to as "least squares regression". This has to do with the fact that a criterium of "minimizing squared deviations from the mean" is often used to estimate the parameters of the regression model. However, other estimation methods exist (beyond the scope of this course).



- Hence, since we actually used a *sample* to *estimate* the *population regression line*, a more accurate notation would have been

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Drawing Conclusions about Population Associations

- β_0: changes depending on centering of X, which doesn't affect association of interest

- Real concern: is X associated with Y?
- Assess by **testing** β_1:
  Does $\beta_1 = 0$ in the population from which this sample was drawn?
  - Hypothesis testing
  - Confidence interval

Hypothesis Testing in Regression

Formulation of null hypothesis, alternative hypothesis, derivation of test statistic:

- H_0: $\beta_1 = 0$
- H_0: $\beta_1 \neq 0$
- Test statistic:
$$t_{obs} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$
- df = n-k-1
  - n = number of observations
  - k = number of predictors (X's)

Would you have expected this statistic to follow a t-distribution?

Would you have expected this statistic to follow a t-distribution?

Summary table:Confidence intervals for difference in means

Population Distribution	Sample Size	Population Variances	95% Confidence Interval
Normal	Any	known	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
	Any	unknown, $\sigma_1^2 = \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm t_{0.025,n_1+n_2-2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$
	Any	unknown, $\sigma_1^2 \neq \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm t_{0.025,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Not Normal/ Unknown	Large	known	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
	Large	unknown, $\sigma_1^2 = \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$
	Large	unknown, $\sigma_1^2 \neq \sigma_2^2$	$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
	Small	Any	Non-parametric methods

Example: Education

- H_0: $\beta_1 = 0$
- Test statistic: $$t_{obs} = \frac{-2.0 - 0}{0.59} = -3.36$$

- df = n-k-1 = 60-1-1 = 58
  - n = number of observations = 60
  - k = number of predictors (X's) = 1

- Calculate our p-value

```
2*pt(-3.36, df=58)
[1] 0.001383108
```

- p-value=0.001

Example: Education

- If there were no association between median education and percentage of disadvantaged citizens in the population, there would be about a 1% chance of observing data as or more extreme than ours.

- The null probability is very small, so:
  - reject the null hypothesis
  - conclude that median education level and percentage of disadvantaged citizens are associated in the population

The use of Confidence Intervals

It becomes easy once you have a pivotal quantity identified:

We calculate the CI using the usual formula:

$$\hat{\beta}_1 \pm t_{CR} \mathrm{SE}\left(\hat{\beta}_1\right)$$

df of t_{CR} = n-k-1

For the education example, the 95% CI for β_1 is:

$$-2.0 \pm 2.021 \times 0.59$$

$$\Rightarrow (-3.2, -0.8)$$

The use of Confidence Intervals

- We are '95% confident' that the true population **decrease** in percentage of low income citizens per additional year of median education is between **3.2 and 0.8**

- Since this interval does not contain 0, we believe percentage of low income citizens and median education are associated among cities in the United States

Statistical Modeling

General Approach:

General approach for most statistical modeling:

- Define the population of interest
- State the scientific questions & underlying theories
- Describe and explore the observed data
- Define the model
  - Probability part (models the randomness / noise)
  - Systematic part (models the expectation / signal)

General Approach (continued):

- Estimate the parameters in the model
  - Fit the model to the observed data
- Make inferences about covariates
- Check the validity of the model
  - Verify the model assumptions
- Re-define, re-fit, and re-check the model if necessary
- Interpret the results of the analysis in terms of the scientific questions of interest

Model Validity Checks

- Check that the assumptions of the model hold
- Plots
- Residual Checking
- Global Model Checks (adjusted R^2, AIC, BIC)

Model Validity Checks

What do we have to check?

Model Systematic:

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

Probability:

$$\varepsilon_i \sim N(0,\sigma^2)$$

Assumptions

- L Linear relationship
- I Independent observations
- N Normally distributed around line
- E Equal variance across X's

Model Validity Checks

How do we have to check?

- Simply plotting the data can be one of the most powerful model checking techniques
- From a simple plot of Y on X that includes the fitted regression line, we can check:
  - linearity, normality, equal (constant) variance, outliers, etc.

Residual Plots

- Y-axis
  - residuals
  - standardized residuals
    - standardized residuals are Z values, so extreme observations are obvious
- X-axis
  - continuous X
  - fitted values
    - fitted values are a linear combination of X's

LINE: Linear relationship

- Is the model correct?
  - Is this the right line?
  - Are there **outliers** for which the model may be wrong?
- Assess with graphs
  - 1 continuous X:
    - graph Y vs. X with line
    - residual plot
  - 2+ continuous X's:
    - adjusted variable plot

LINE: Linear relationship

- In applied statistics, a **partial regression plot** attempts to show the effect of adding an additional variable to the model (given that one or more independent variables are already in the model).
- Partial regression plots are also referred to as added variable plots or adjusted variable plots.
- Partial regression plots are formed by:
  1. Computing the residuals of regressing the response variable against the independent variables but omitting Xi
  2. Computing the residuals from regressing Xi against the remaining independent variables
  3. Plotting the residuals from 1. against the residuals from 2.

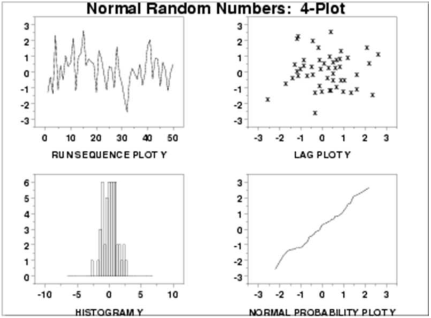
LINE: Independent observations

- The relevant question here is: Are all the subjects surveyed independent of one another?
- In order to answer this question, one needs information about how the data were collected . . .

Can the "independence assumption" be assessed graphically?

LINE: Independent observations

Can the "independence assumption" be assessed graphically?



If the lag plot (Y_i versus Y_{i-1}) is without structure, then the randomness assumption holds . . .

LINE

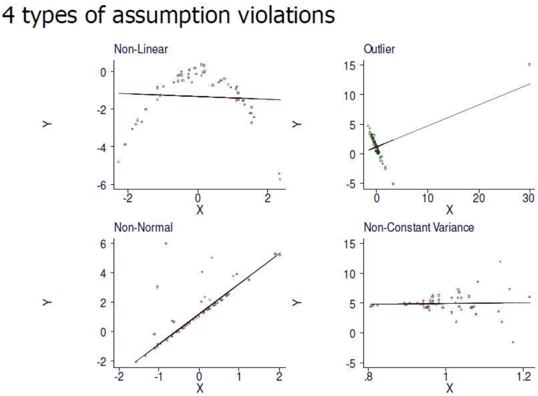
- At ***every value of X***, the observed points should follow a roughly normal distribution centered at the fitted value of Y.
- Assess with residual plots

LINE

- At ***every value of X***, the observed points should follow a roughly normal distribution with the same variance across all X's

- Assess with residual plots

Plotting y versus x



Standardized residuals:

- The residuals, ε_i , are the differences between the observed values, y_i, and their fitted values: $$\varepsilon_i = y_i - \hat{y}_i$$
- Since our model states: $\varepsilon_i \sim N(0,\sigma^2)$
- We know that the *standardized* residuals,

$$\frac{\varepsilon_i - 0}{\sqrt{\hat{\sigma}^2}} \quad \text{where} \quad \hat{\sigma}^2 = MSE$$

should follow a standard normal distribution

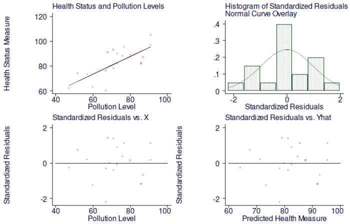
Graphical Model Validity Checks: Residual Analysis

If the model fits the data well, we expect:

- A histogram of the standardized residuals should look normal.
  - Check for asymmetry and outliers.
- A plot of the residuals vs. X should look like a random scatter (no systematic relationship)
- A plot of the residuals vs. $\hat{y}_i$ (the fitted values) should also look like a random scatter.

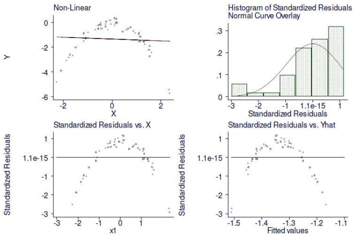
Example 1: Residual Analysis on Health Status

Example: Relationship between health status and pollution in 20 geographic areas



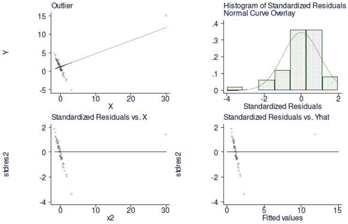
- Regression scatterplot looks good
- Standard Residuals appear fairly normally distributed
- Standard Residuals vs X appear randomly scattered (i.e. no apparent patterns & no extreme outliers)
- Standard Residuals vs predicted values appear randomly scattered (i.e. no apparent patterns & no extreme outliers)

Example 2: Non-linearity



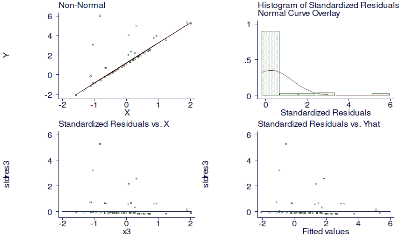
- Regression scatterplot shows non-linear relationship
- Standard Residuals don't look normally distributed
- Standard Residuals vs X shows non-linear relationship
- Standard Residuals vs predicted values shows non-linear relationship

Example 3: Outliers



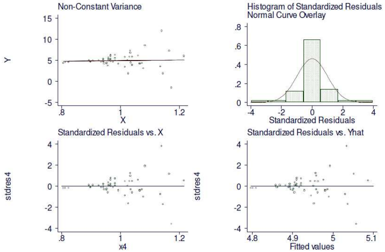
- Regression scatterplot shows outlier
- Standard Residuals look normal but 'large' residual present
- Standard Residuals vs X shows a pattern & the outlier
- Standard Residuals vs Y shows a pattern & the outlier

Example 4: Non-normality



- Regression scatterplot shows non-even spread
- Standard Residuals don't look normally distributed
- Standard Residuals vs X shows non-even spread
- Standard Residuals vs predicted values shows non-even spread

Example 5: Heteroscedasticity



- Regression scatterplot shows increasing variability
- Standard Residuals do look normally distributed
- Standard Residuals vs X shows increasing variability
- Standard Residuals vs predicted values shows increasing variability

Minimal Practice: Fit a regression line and . . .

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Experience}_i$$

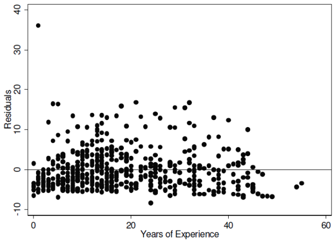
$$\Rightarrow \quad \hat{Y}_i = 8.38 + 0.04\text{Experience}_i$$

Check model assumptions

For instance, plot residuals versus x:

- Used to assess remaining relationships within data
  - assumption of "linearity"

- Line has been "flattened"
- Residuals (or error terms) are centered at 0: horizontal line shows "flattened" regression line



Check model assumptions

Compute standardized residuals:

- With the actual residuals, it's hard to tell which points are extreme
- Standardized residuals are

$$\text{Standardized Residual}_i = \frac{\text{residual}_i}{SD_{residuals}} = Z_i$$

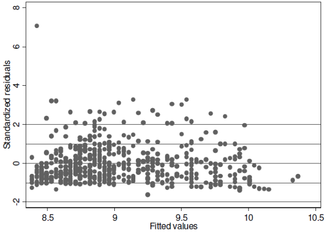
- |sres|>2 about 5%
- |sres|>3 about 1%

Note that for a normal distribution, About 68.27% of the values lie within 1 standard deviation of the mean. Similarly, about 95.45% of the values lie within 2 standard deviations of the mean. Nearly all (99.73%) of the values lie within 3 standard deviations of the mean

Check model assumptions

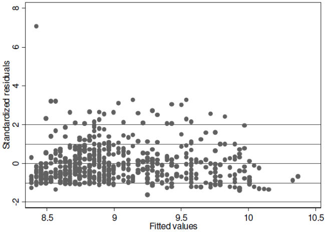
Plot standardized residuals versus x:

- Plotting against X is fine when there's only one continuous X in model
- When multiple continuous X's are in model
  - plotting residuals against fitted values is like plotting against all the X's at once
  - if problems are seen, one can plot residuals against each X to see which causes problem



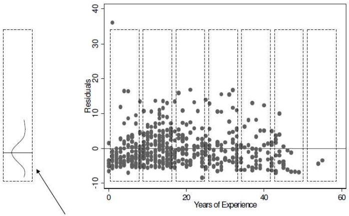
Model fit: Residual pattern for people with very little experience?

Caution: Outliers Check



- Outliers far from the pattern of the rest of the X's *may* affect the line
- the regression line always goes through (mean $\overline{X}$, mean $\overline{Y}$)
- an outlier near the mean $\overline{X}$ will not influence the line very much
- an outlier far from the mean $\overline{X}$ can draw the line towards itself

Caution: Normality Check

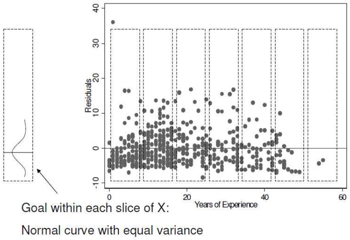


Goal within each slice of X:

Normal curve centered at 0

- Parameters estimates are still correct, but CI's are misleading
- Including additional predictors sometimes solves this problem
- Another solution is to transform Y
  - ln(Y) or sqrt(Y) draws in data skewed to *high* values
  - 1/Y or 1/sqrt(Y) draws in data skewed to *low* values
  - use transformed Y instead of original Y
  - interpret parameters according to transformed Y!

Caution: Homescedasticity Check



- Again, parameter estimates are valid, but CI's are misleading
- Adding additional parameters may solve the problem

Recapitulation

Questions?

Acknowledgements

Most slides on formal statistical inference, chi-square testing and regression are based on an excellent course series of Sandy Eckel, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA.