Math 1710 Class 37

Chi Square, Regression Inference
Dr. Back

Nov. 23, 2009

Distribution

5- number summary: 33, 72, 82, 91, 100
mean 79.07 std. dev. 14.83

Distribution

5- number summary: 33, 72, 82, 91, 100
mean 79.07 std. dev. 14.83

Approximate Guide to How You Did:

Scores	**Freq.**	Letter
96-100	6	A+
90-95	19	A
88-89	6	A-
80-86	19	B+
71-79	17	B
63-69	10	B-
54-58	4	C
41-42	3	D
33-37	2	F

Distribution

5- number summary: 33, 72, 82, 91, 100
mean 79.07 std. dev. 14.83

Approximate Guide to How You Did:

Scores	**Freq.**	Letter
96-100	6	A+
90-95	19	A
88-89	6	A-
80-86	19	B+
71-79	17	B
63-69	10	B-
54-58	4	C
41-42	3	D
33-37	2	F

We may not use these letters.

Chi Square Distribution as an RV

Let $X_1, \ldots X_d$ be d independent standard normal random variables.

Chi Square Distribution as an RV

Let $X_1, \ldots X_d$ be d independent standard normal random variables.

The **Chi Square distribution with d degrees of freedom** is given by:

$$\chi^2 = X_1^2 + X_2^2 + \ldots X_d^2$$

Chi Square Distribution as an RV

Let $X_1, \ldots X_d$ be d independent standard normal random variables.

The **Chi Square distribution with d degrees of freedom** is given by:

$$\chi^2 = X_1^2 + X_2^2 + \ldots X_d^2$$

Chi Square Distribution Formula

$$f(x) = \frac{1}{\Gamma(\frac{d}{2})2^{\frac{d}{2}}} x^{\frac{d}{2}-1} e^{\frac{-x}{2}}$$

where d is the number of degrees of freedom.

Chi Square Distribution as an RV

Let $X_1, \ldots X_d$ be d independent standard normal random variables.
The **Chi Square distribution with d degrees of freedom** is given by:

$$\chi^2 = X_1^2 + X_2^2 + \ldots X_d^2$$

Chi Square Distribution Formula

$$f(x) = \frac{1}{\Gamma(\frac{d}{2})2^{\frac{d}{2}}} x^{\frac{d}{2}-1} e^{\frac{-x}{2}}$$

where d is the number of degrees of freedom.
The mean is d and the variance is $2d$.

Chi Square Hypotheses

Test of Independence: Two Categorical Variables.

- H_0: The two categorical variables are independent.
- H_a: There is an association between the variables.

Chi Square Hypotheses

Test of Homogeneity: Multiple samples/populations.
and Another Categorical Variable.

- H_0: The categorical variable has the same distribution within each population.
- H_a: The distributions differ among some of the populations.

(Which Population? could be viewed as a cat. var.)

Chi Square Hypotheses

Same Chi Square statistic,
Degrees of Freedom = (numrows-1)(numcols-1) for both indep. and homog.

Chi Square Hypotheses

Same Chi Square statistic,
Degrees of Freedom = (numrows-1)(numcols-1) for both indep. and homog.
The concept of “what is a population” does not have a clear answer, so we will not require you to distinguish between indep. and homog. on exams or homework.
(You need to know lots about the design to truly tell these apart.)

Chi Square Hypotheses

The concept of “what is a population” does not have a clear answer, so we will not require you to distinguish between indep. and homog. on exams or homework.
(You need to know lots about the design to truly tell these apart.)
In fact indep. and homog. have different exact mmathematical models, but both are approximated by the same χ^2.

Chi Square Hypotheses

Goodness of Fit: Frequencies of One Cat Var
AND a hypothesized distribution.

- H_0: The cat. var. follows the hypothesized distribution.
- H_a: The cat. var. doesn't.

Degrees of Freedom =number of cells -1.

Why $\frac{(Obs-Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment.
Each cell corresponds to a particular value of each of the two categorical variables.

Why $\frac{(Obs-Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment.

Notation:

n	The total number of observed subjects.
$\hat{p}$	The observed proportion in the cell of interest.
p	The true proportion in the cell of interest.

Why $\frac{(Obs - Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment. Then

Observed $n\hat{p}$.

Expected np.

$$\frac{(Obs - Exp)^2}{Exp} = \frac{(n\hat{p} - np)^2}{np} = \left(\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}\right)^2 q$$

Why $\frac{(Obs-Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment. Then

Observed $n\hat{p}$.

Expected np.

$$\frac{(Obs - Exp)^2}{Exp} = \frac{(n\hat{p} - np)^2}{np} = \left(\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}\right)^2 q$$

Thus except for the extra q (often near 1), we have the square of a Z-statistic!

Why $\frac{(Obs-Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment.
Then

Observed $n\hat{p}$.

Expected np.

$$\frac{(Obs-Exp)^2}{Exp} = \frac{(n\hat{p}-np)^2}{np} = \left(\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}\right)^2 q$$

Thus except for the extra q (often near 1), we have the square of a Z-statistic!
The degrees of freedom being less than the number of cells may be viewed as the result of careful analysis of all those extra q factors.

2 × 2 Tables

In this case let

p_1: true proportions of column 1 in the top cell of that column.

p_2: true proportions of column 2 in the top cell of that column.

2 × 2 Tables

In this case let

p_1: true proportions of column 1 in the top cell of that column.

p_2: true proportions of column 2 in the top cell of that column.

The null hypothesis in an independence test can also be expressed as

$$p_1 = p_2.$$

2×2 Tables

In this case let

p_1: true proportions of column 1 in the top cell of that column.

p_2: true proportions of column 2 in the top cell of that column.

The null hypothesis in an independence test can also be expressed as

$$p_1 = p_2.$$

The Chi Square statistic will be the square of the Z statistic for the 2-sided 2-sample proportion test with the same H_0.

Chi Square as Multiple Proportion

If X and Y are the categorical variables the contingency table describes, then H_0 in a test of independence can also be expressed as for each value j of Y, all the proportions

$$P(X = i|Y = j)$$

are equal, regardless of the value of i.

Degrees of Freedom

Why df=(numrows-1)(numcols-1) for test of independence? (a 3 by 4 example: df=2*3=6)

Degrees of Freedom

Why df=(numrows-1)(numcols-1) for test of independence? (a 3 by 4 example: df=2*3=6)

Imagine these marginals . . .

known	known	known		300
known	known	known		300
				300
100	200	300	400	1000

Degrees of Freedom

Idea: Consider all 3 by 4 contingency tables with the same marginal totals. These determine the expected values. How well the observed values match the expected tests the null hypothesis of independence.

Degrees of Freedom

The idea is that if we specify the observed values in the 6 known cells, then the rest of the cells are determined by the marginal totals. In other words df=6 corresponds to 6 cells can vary freely, the rest are determined.

Degrees of Freedom

Suppose we know the values in the 2×3 upper left:

40	80	100		300
30	110	150		300
				300
100	200	300	400	1000

Degrees of Freedom

Then the remaining values are determined by subtraction

40	80	100	80	300
30	110	150	10	300
30	10	50	310	400
100	200	300	400	1000

C

Regression Inference Questions

Basic Setup:

1) Data (x_i, y_i), $1 \leq i \leq n$ leads to line of regression

$$\hat{y} = b_0 + b_1 x$$

2) Assume an ideal line

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

3) Together with an error process ϵ_i folowing an $N(0, \sigma)$ law (independent for each i.)

4) So that individual observations come from

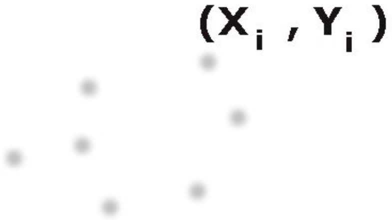
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Regression Inference Questions

Start with some data (x_i, y_i)

Regression Inference Questions

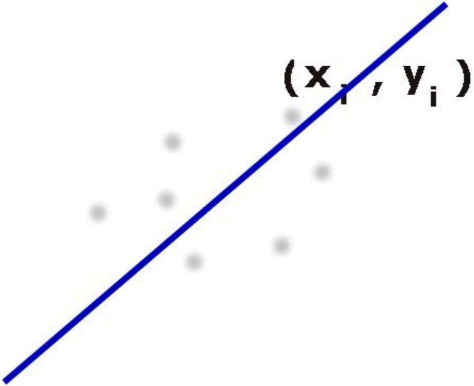
Start with some data (x_i, y_i)



Regression Inference Questions

These give the line of regression

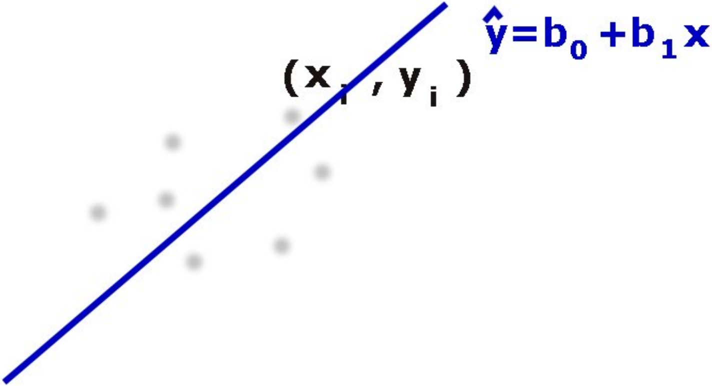
$$\hat{y} = b_0 + b_1 x_i$$



Regression Inference Questions

These give the line of regression

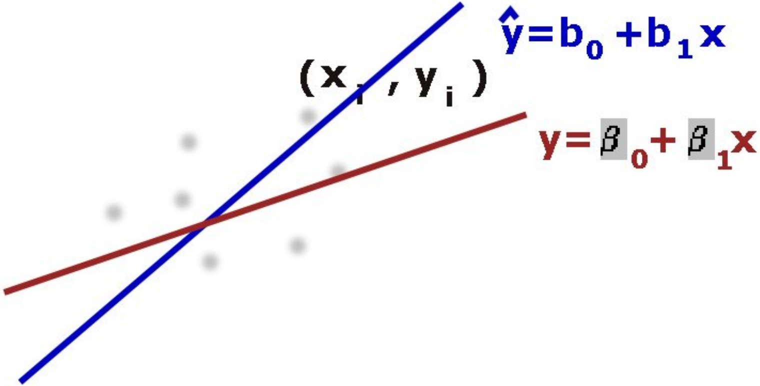
$$\hat{y} = b_0 + b_1 x_i$$



Regression Inference Questions

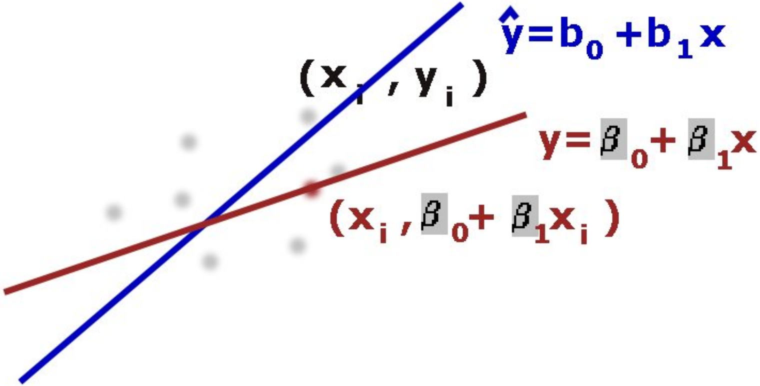
There is also an ideal line

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



Regression Inference Questions

We think of the data point (x_i, y_i) as arising by first plugging in x_i into the ideal line

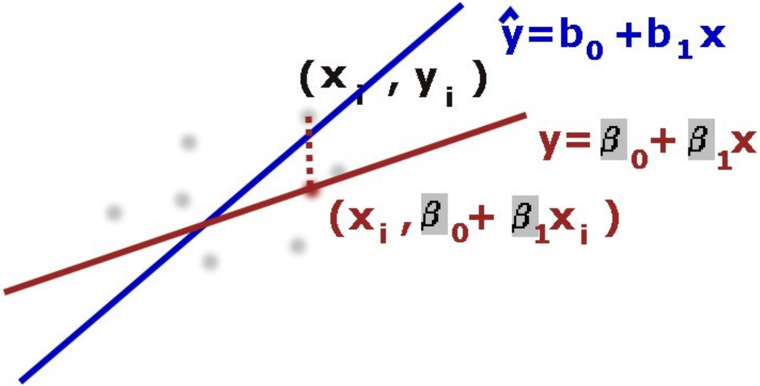


Regression Inference Questions

Then randomly choosing an offset

$$\epsilon_i \sim N(0, \sigma)$$

to see how far off the ideal line we go.

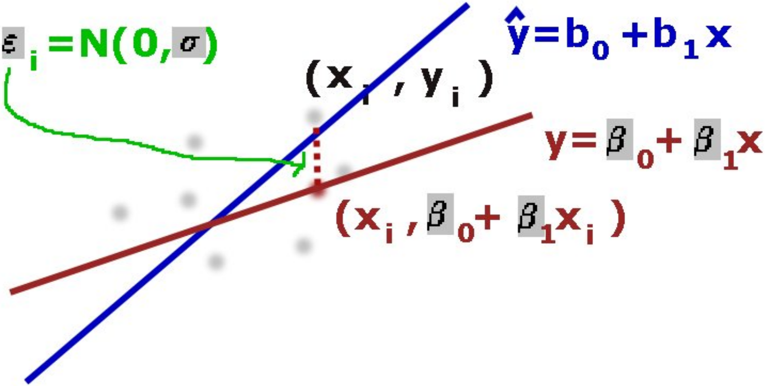


Regression Inference Questions

Then randomly choosing an offset

$$\epsilon_i \sim N(0, \sigma)$$

to see how far off the ideal line we go.



Regression Inference Questions

The errors must be independent and all have the same standard deviation σ in the ideal model.

Regression Inference Questions

Natural Questions in Regression:

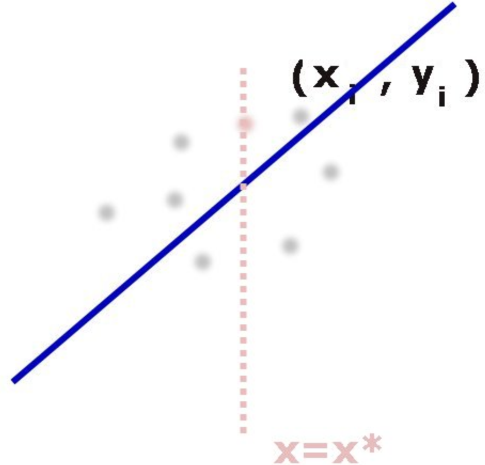
1) Estimate β_0 and β_1.
2) Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.
3) Estimate σ, the standard deviation of the error process.

For a given value x^* of x:

4) How accurately does the regression estimate $b_0 + b_1x^*$ approximate an actual y observation when $x = x^*$.
5) How accurately does the regression estimate $b_0 + b_1x^*$ approximate the average of a lot of y observations when $x = x^*$.

Regression Inference Questions

For $x = x^*$, we can try to predict one value of y leading to a prediction interval when $x = x^*$



Regression Inference Questions

Or we can try to predict the average of many values of y leading to a confidence interval for the mean response μ_y when $x = x^*$