χ^2- goodness of fit test

Some questions:

- Are the dice we roll in our experiments in class really fair ?
- Is getting Dengue(D) or severe form of Dengue (DSS) independent of `BICARB1` reading ?

Rephrase:

- How well the distribution of the data fit the model ?
- Does one variable affect the distribution of the other ?

χ^2- goodness of fit test

Specific Question:

- To understand how "close" are the observed values to those which would be expected under the fitted model ?

Towards Answer:

- In this case we seek to determine whether the distribution of results in a sample could plausibly have come from a distribution specified by a null hypothesis.
- The test statistic is calculated by comparing the observed count of data points within specified categories relative to the expected number of results in those categories (under Null).

χ^2- goodness of fit test

- Let T be a random variable with finite range $\{c_1, c_2, \ldots, c_k\}$ for which

$$P(T = c_j) = p_j > 0 \text{ for } 1 \leq j \leq k.$$

- Let $X_1, X_2, \ldots, X_n$ be the sample from the distribution T and let

$$Y_j = |\{j : X_j = c_j\}| \text{ for } 1 \leq j \leq k..$$

Y_j is the number of sample points whose outcome was c_j

- Then the statistic

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

χ^2- goodness of fit test

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- $\mathbf{X}^2$- has χ^2_{k-1} degrees of freedom, assymptotically as $n \to \infty$.
- Null Hypothesis: Distribution comes from Multinomial with parameters $p_1, p_2, \ldots, p_k$
- Alternate Hypothesis: Distribution comes from Multinomial with parameters with at least one parameter different from $p_1, p_2, \ldots, p_k$

χ^2- goodness of fit test

Example:

We divide the political parties in India into 3 large alliances: NDA, UPA, and Third-Front. In the previous election the support had been 38%, 32% and 30% support respectively. Super-Nation TV channel takes a sample of 100 people and finds that there are 35 for NDA, 40 for UPA and 25 for Third-Front. It concludes that the vote share has not changed. Is this hypothesis correct ?

χ^2- goodness of fit test

- Null Hypothesis: Vote Share is $(38, 32, 30)$
- Level of Significance: 0.05
- Data: Sample Vote share is $(35, 40, 25)$

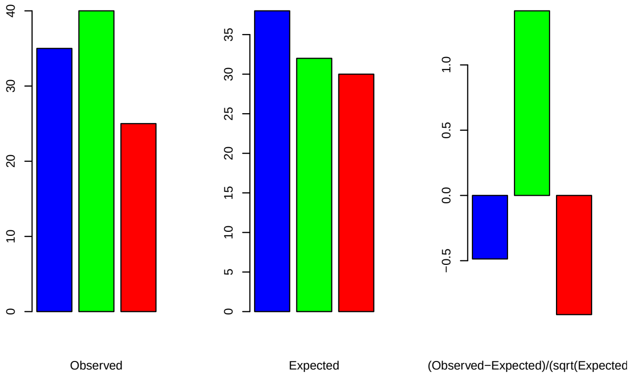
χ^2- goodness of fit test

Example Contd.:

```
> x = c(35,40,25)
> prob = c(38,32,30)
> prob = prob/sum(prob)
> n = sum(x)
> z = (x-n*prob)/((sqrt(n*prob)))
```

χ^2- goodness of fit test

Example Contd.:



χ^2- goodness of fit test

Example Contd.:

```
> Xsquared = sum(((x-n*prob)^2)/(n*prob))
> Xsquared
[1] 3.070175
> pchisq(Xsquared, df = 3 -1, lower.tail=FALSE)
[1] 0.2154368
```

Since *p*-value is not smaller than 0.05 we do not reject the null hypothesis.

χ^2- goodness of fit test

Example Contd.: We can use in built R function

```
> chisq.test(x,p=prob)

        Chi-squared test for given probabilities

data:  x
X-squared = 3.0702, df = 2, p-value = 0.2154
```

χ^2- goodness of fit test

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- Large values of $\mathbf{X}^2$ indicate that the observed counts don't match expected counts.
- Large values of $\mathbf{X}^2$ indicates evidence that Null is not correct.

χ^2- goodness of fit test

- Test Statistic:

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- Decide on level of significance: α
- Compute p-value:

$$\mathbb{P}(\chi^2_{k-1} \geq X^2)$$

- Reject Null Hypotheis:

if p-value is less than α

Contigency Tables

- Bivariate Data is often presented as a two-way table.

- For example in Dengue Data from Manipal Hospital

```
> y = read.table("dengueb.csv", header=TRUE)
```

```
> head(y)

  DIAGNO BICARB1
1    DSS    16.2
2    DSS    22.0
3    DSS    16.0
4    DSS    21.3
5    DSS    19.0
6    DSS    18.7
```

```
> tail(y)

   DIAGNO BICARB1
45      D    22.0
46      D    16.6
47      D    18.3
48      D    23.0
49      D    24.0
50      D    21.0
```

Contigency Tables

- Bivariate Data is often presented as a two-way table.
- For example in Dengue Data from Manipal Hospital

```
           Diagnosis
Cat.Marker  D DSS
         0  0   6
         1 17  15
         2  8   4
```

where we have grouped values of Marker to be $0, 1, 2$ depending on the values being less than or equal to 16, between 16 and 21, and greater than 21.

χ^2- test of independence

Specific question:

- Does one variable affect the distribution of the other ?

Notation:

- Let n_r be the number of rows in the table.
- Let n_c be the number of columns in the table.
- Let $n = n_r n_c$ be the total number of observations.

Model:

- Let $T \equiv (p_{ij})$ with $1 \leq i \leq n_r, 1 \leq j \leq n_c$ be a probability distribution on $\{(i,j) : 1 \leq i \leq n_r \text{ and } 1 \leq j \leq n_c\}$
- Let $p_i^R = \sum_{j=1}^{n_c} p_{ij}$ and $p_j^C = \sum_{i=1}^{n_r} p_{ij}$

χ^2- test of independence

- Null Hypothesis: Variables are independent i.e

$$p_{ij} = p_i^R p_j^C \text{ for all } 1 \leq i \leq n_r \text{ and } 1 \leq j \leq n_c$$

- Alternate Hypothesis: Variables are not independent

χ^2- test of independence

- Let y_{ij} record the frequency in the (i,j) cell.
- Let

$$\hat{p}_i^R = \frac{\sum_{j=1}^{n_c} y_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} y_{ij}} \text{ and } \hat{p}_j^C = \frac{\sum_{i=1}^{n_r} y_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} y_{ij}}$$

Let

$$\hat{p}_{ij} = \hat{p}_i^R \hat{p}_j^C$$

and

$$\mathbf{X}^2 := \sum_{i=1}^{n_r}\sum_{j=1}^{n_c} \frac{(y_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

χ^2- test of independence

- Test Statistic:

$$\mathbf{X}^2 := \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} \frac{(y_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

  is χ^2_q distributed assymptotically as $n \to \infty$ with $q = (n_r - 1)(n_c - 1)$ degrees of freedom.

- Decide on level of significance: α
- Compute p-value:

$$\mathbb{P}(\chi^2_q \geq X^2)$$

- Reject Null Hypotheis:

if p-value is less than α

χ^2- test of independence

For example in Dengue Data from Manipal Hospital:

```
> T = table(Cat.Marker, Diagnosis)
> T

          Diagnosis
Cat.Marker  D DSS
         0  0   6
         1 17  15
         2  8   4
```

Can we test if the Marker value is independent of the characterisation of Dengue as normal or severe ?

χ^2- test of independence

For example in Dengue Data from Manipal Hospital:

```
> chisq.test(T)

        Pearson's Chi-squared test

data:  T
X-squared = 7.4583, df = 2, p-value = 0.02401
```

Simple Linear Regression: Relationship in Bivariate Data

- Key: conditional mean of response variable given the predictor cariable is a linear function.
- Model: For data points (x_i, y_i) with $1 \leq i \leq n$,

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

  where ε_i assumed to be mean 0 and variance σ^2 Normal random variables.
- Observe only (x_i, y_i) for $1 \leq i \leq n$.

Simple Linear Regression: Relationship in Bivariate Data

- Find β_0, β_1 such that

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

  is minimized.

- Can be solved: Calculus and Linear Algebra

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \text{correlation}(x, y)\frac{S_x^2}{S_y^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Observations:

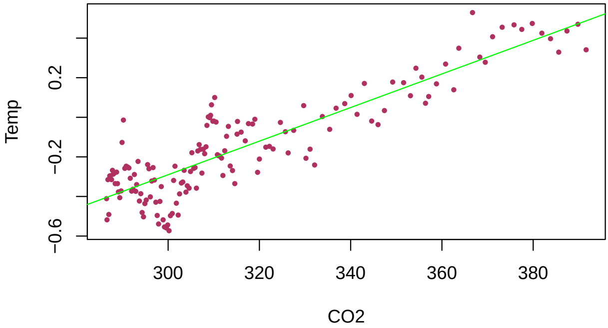
- Slope of line is function of Correlation in standarised scale.
- Line passes through $(\bar{x}, \bar{y})$
- Roles of y and x are not interchangeable.

Data Set: annualtemp.csv

```
> y = read.csv("annual_temp.csv", header=TRUE)

> head(y)

  Year   Temp   CO2   CH4   NO2 Irradiance Nino_SST Volcano
1 1861 -0.411 286.5 838.2 288.9   1361.097 26.74233 0.00281
2 1862 -0.518 286.6 839.6 288.9   1360.987 26.39426 0.00859
3 1863 -0.315 286.8 840.9 289.0   1360.837 26.16013 0.01318
4 1864 -0.491 287.0 842.3 289.1   1360.753 26.28774 0.00707
5 1865 -0.296 287.2 843.8 289.1   1360.691 26.32374 0.00302
6 1866 -0.295 287.4 845.5 289.2   1360.600 26.31218 0.00128

> tail(y)

    Year  Temp   CO2    CH4   NO2 Irradiance Nino_SST Volcano
146 2006 0.425 381.9 1784.5 320.0   1361.005 27.25267 0.00342
147 2007 0.397 383.8 1790.4 320.8   1360.939 26.66768 0.00454
148 2008 0.329 385.6 1797.8 321.7   1360.849 26.43034 0.00374
149 2009 0.436 387.4 1802.7 322.4   1360.822 27.50094 0.00402
150 2010 0.470 389.8 1807.7 323.2   1360.841 26.80601 0.00449
151 2011 0.341 391.6 1813.1 324.2   1361.083 26.39182 0.00370
```

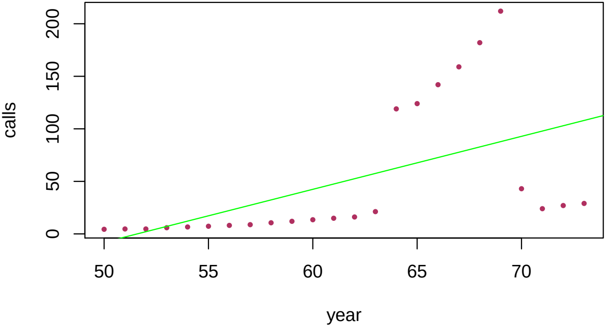
Simple Linear Regression: Effect of CO2 on Temperature

```
> plot(Temp ~ CO2,  data=y, pch=19,cex=.5, col="maroon")
> abline(lm(Temp ~ CO2, data=y), col="green")
```



Simple Linear Regression: Belgium Phone Calls

```
> require(MASS)
> plot(calls ~ year,  data=phones, pch=19,cex=.5, col="maroon")
> abline(lm(calls ~ year, data=phones), col="green")
```



Simple Linear Regression

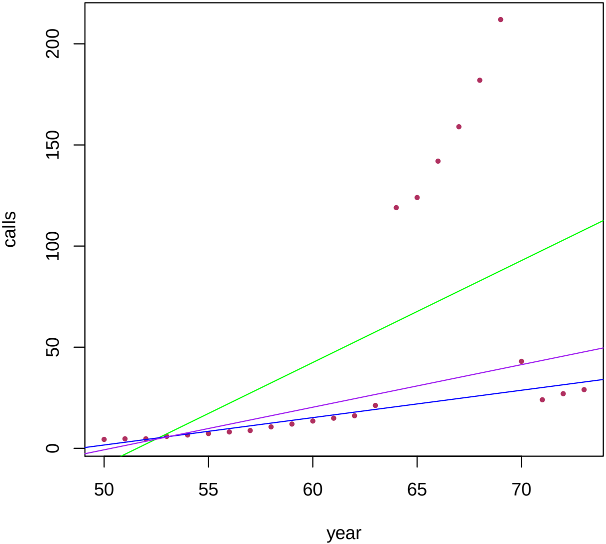
- Our own absolute deviation line:

```
> ABSMINLINE = function(x)
+ { with (phones, sum(abs(calls- x[1] -x[2]*year)))
+ }
> OPTIMAL = optim(c(0,0), fn = ABSMINLINE)
```

- Plotted `lm, rlm, absline`

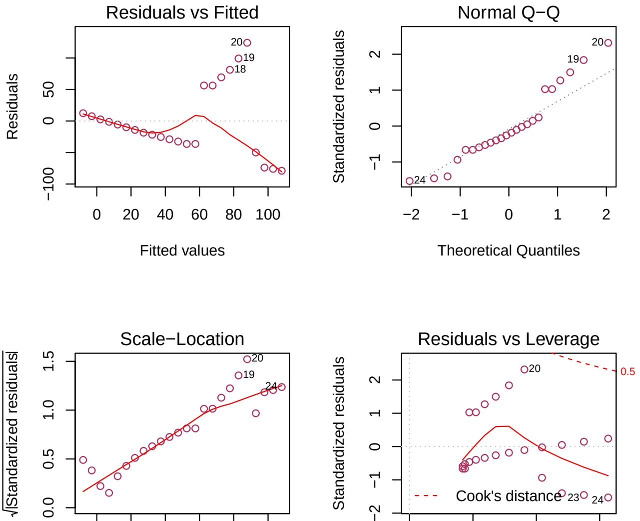
```
> abline(lm(calls ~ year, data=phones), col="green")
> abline(OPTIMAL$par, col="blue")
> abline(rlm(calls ~ year, data=phones), col="purple")
```

Simple Linear Regression



are there better fits ?: Shown you these four graphs.

```
> par(mfrow=c(2,2))
> plot(lm(calls ~ year,  data=phones),  col="maroon")
```



Simple Linear Regression

- Model:

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

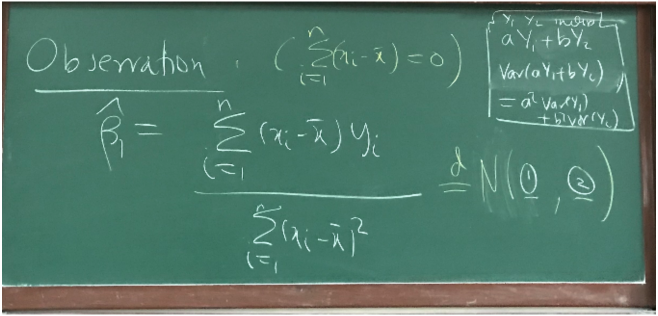
with ε_i being i.i.d Normal$(0, \sigma^2)$.

- Estimators:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \text{correlation}(x, y)\frac{S_x^2}{S_y^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Simple Linear Regression: September 17th



Simple Linear Regression: September 17th

- We had shown that

$$E[\hat{\beta}_1] = \beta_1$$

  and

- if

$$\text{RSS} \equiv \text{Residual Sum of Squares} := \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

  Then

$$E[RSS] = (n-2)\sigma^2$$

Simple Linear Regression: September 17th

- $$\hat{\beta}_1 \sim \text{Normal}(\beta_1, \frac{\sigma^2}{S_{xx}^2})$$
- $$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-2}$$
- RSS and $\hat{\beta}_1$ are independent and thus
$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{RSS}}{(n-2)S_{xx}^2}}} \sim t_{n-2}.$$

Simple Linear Regression: Testing

$$T := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{RSS}}{(n-2)S_{xx}^2}}} \sim t_{n-2}.$$

- Decide on level of significance: α

- Null Hypothesis: $\beta_1 = b$
- Alternate Hypothesis: $\beta_1 \neq b$

Simple Linear Regression: Testing

- Test Statistic:

$$T := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{RSS}}{(n-2)S_{xx}^2}}}$$

- Decide on level of significance: α

- Compute p-value:

$$\mathbb{P}(\mid t_{n-2} - b \mid \geq \mid T - b \mid)$$

- Reject Null Hypotheis:

if p-value is less than α

Simple Linear Regression: Confidence Interval

$$T := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{RSS}}{(n-2)S_{xx}^2}}} \sim t_{n-2}$$

The interval

$$\left(\hat{\beta}_1 - t_{n-2}(0.25)\sqrt{\frac{\text{RSS}}{(n-2)S_{xx}^2}}, \hat{\beta}_1 + t_{n-2}(0.25)\sqrt{\frac{\text{RSS}}{(n-2)S_{xx}^2}}\right)$$

–the 95% confidence interval for β_1 with

$$\mathbb{P}(t_{n-2} \geq t_{n-2}(0.25)) = 0.025.$$

Similar to Analysis of Variance One way

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Total Sum of squares = RSS + Regression Sum of squares

In Short:

$$\text{SS}_{\text{total}} = \text{RSS} + \text{SS}_{\text{Reg}}$$

Similar to Analysis of Variance One way

$$\text{SS}_{\text{total}} = \text{RSS} + \text{SS}_{\text{Reg}}$$

- As

$$\hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$$

  $\hat{\beta}_1 \sim 0$ is near zero then $\text{SS}_{\text{Reg}} \sim 0$ and
  $\hat{\beta}_1 \neq \sim 0$ is near zero then SS_{Reg} is large.

- Therefore SS_{Reg} can be used to test for $\hat{\beta}_1 = 0$.

Simple Linear Regression

$$\text{SS}_{\text{total}} = \text{RSS} + \text{SS}_{\text{Reg}}$$

- SS_{total} there are n sample points and one derived value $\bar{y}$ – $n-1$ degrees of freedom.
- RSS there are n sample points and two estimated values $\hat{\beta}_0, \hat{\beta}_1$ – $n-2$ degrees of freedom.
- SS_{Reg} has one degree of freedom.

Simple Linear Regression

- RSS and SS_{Reg} are independent.
- $$\frac{\text{SS}_{\text{Reg}}}{\sigma^2} \sim \chi^2_1 \qquad \text{and} \qquad \frac{(n-2)\text{RSS}}{\sigma^2} \sim \chi^2_{n-2}$$
- $$\frac{\text{SS}_{\text{Reg}}}{\frac{\text{RSS}}{n-2}} \sim F(1, n-2).$$

Simple Linear Regression: Testing

$$F := \frac{\mathsf{SS}_{\mathsf{Reg}}}{\frac{\mathsf{RSS}}{n-2}}$$

- Decide on level of significance: α

- Null Hypothesis: $\beta_1 = 0$
- Alternate Hypothesis: $\beta_1 \neq 0$

Simple Linear Regression: Testing

$$F := \frac{\mathrm{SS}_{\mathrm{Reg}}}{\frac{\mathrm{RSS}}{n-2}}$$

- Decide on level of significance: α
- Compute *p*-value:

$$\mathbb{P}(F(1, n-2) \geq F)$$

- Reject Null Hypotheis:

if *p*-value is less than α