Probabilistic Modeling and Statistical Computing Fall 2015

December 1, 2015

Constructing Tests

Recall maximum likelihood estimators: general method for constructing estimation formulae, often leads to identification of good statistics ("data reduction").

Is there a similar approach for constructing hypothesis tests?

Need to balance two objectives: significance level (should be kept small) and power (should be large).

Simple Hypotheses

Assume that both H_0 and H_a consist of single, well specified distributions.

- $H_0 : N(\mu_0, \sigma_0^2)$ vs. $H_a : N(\mu_1, \sigma_1^2)$
- Exponential distribution. $H_0 : \lambda = \lambda_0$ vs. $H_a : \lambda = \lambda_1$
- Multinomial distribution: $H_0 : (p_1, \ldots, p_n)$ vs. $H_a : (q_1, \ldots, q_n)$

Find a test of H_0 against H_a that has given significance level α and maximum power.
Find a test statistic and a critical value!

Likelihood Ratio

Assume that both H_0 and H_a consist of single, well specified distributions with pdf's or pmf's $f_0(x), f_a(x)$. The **likelihood function** for H_0 is

$$L_0(x_1, \ldots, x_n) = f_0(x_1) \times \cdots \times f_0(x_n)$$

and similarly for H_a.

Likelihood Ratio

For a sample $(x_1, \ldots, x_n)$, the likelihood ratio is

$$T = \frac{L_0(x_1, \ldots, x_n)}{L_a(x_1, \ldots, x_n)} = \frac{f_0(x_1) \times \cdots \times f_0(x_n)}{f_a(x_1) \times \cdots \times f_a(x_n)}$$

Likelihood Ratio

The likelihood ratio is

$$T = \frac{L_0(x_1, \ldots, x_n)}{L_a(x_1, \ldots, x_n)} = \frac{f_0(x_1) \times \cdots \times f_0(x_n)}{f_a(x_1) \times \cdots \times f_a(x_n)}$$

Interpretation: If T is small, the sample is more likely to come from the alternative distribution. If T is large, the sample is more likely to come from the null distribution.

Likelihood Ratio Test

Likelihood Ratio Test: Given a critical value C, reject H_0 if $T < C$. The significance level is $\mathcal{P}(T < C|H_0)$. The power is $\mathcal{P}(T < C|H_a)$.

Neyman - Pearson Lemma

Of all tests of H_0 versus H_a with given significance level α, the likelihood ratio test has the largest power (the lowest type II error probability).

This tells one how to find a test statistic. It does not tell us how to find the critical value C.

Example: Exponential Distribution

Consider data coming from an exponential distribution with rate $= \lambda$.

$H_0 : \lambda = \lambda_0$ versus $H_a : \lambda = \lambda_a > \lambda_0$

Given a sample $(x_1, \ldots, x_n)$.
Likelihood function for H_0:

$$L_0(x_1, \ldots, x_n) = \lambda_0^n e^{-\lambda_0 x_1} e^{-\lambda_0 x_2} \ldots e^{-\lambda_0 x_n} = \lambda_0^n e^{-\lambda_0 \sum_i x_i}$$

and similarly for H_a.

Example: Exponential Distribution

Consider data coming from an exponential distribution.
Likelihood ratio for this case:

$$T = \frac{L_0(x_1, \dots, x_n)}{L_a(x_1, \dots, x_n)} = \frac{\lambda_0^n e^{-\lambda_0 \sum_i x_i}}{\lambda_a^n e^{-\lambda_a \sum_i x_i}}$$

$$= \left(\frac{\lambda_0}{\lambda_a}\right)^n e^{(-\lambda_0 + \lambda_a) \sum_i x_i}$$

Reject H_0 if $T < C$, where C depends on α. This means reject H_0 if $\tilde{T} = \sum_i x_i < c_1$, since T depends only on $\tilde{T}$.

Where are we now? What is left?

The **likelihood ratio test** uses the test statistic $\tilde{T} = \sum_i x_i$ and rejects H_0 if $\tilde{T}$ is small, $\tilde{T} < c_1$.

Need to find a relation between critical value c_1 and desired significance level α.

To do this, need the distribution of $\tilde{T}$ if H_0 is true.

This can be done analytically or by a simulation.

Critical Region and Power

Compute the critical region of the most powerful test and its power as a function of n.

Fact: $\tilde{T} = \sum_i X_i$ has a Γ distribution, shape parameter $= n$, rate parameter λ.

c_1 = lower α quantile for a $\Gamma(n, \lambda_0)$ distribution.

```
c1 <- qgamma(alpha, shape = n, rate
= lambda0, lower.tail = T)

power <- cgamma(c1, shape = n, rate
= lambdaA, lower.tail = T)
```

Critical Region and Power

Compute the critical region of the most powerful test and its power as a function of n.

Fact: $\tilde{T} = \sum_i X_i$ has a Γ distribution, shape parameter $= n$, rate parameter λ.

c_1 = lower α quantile for a $\Gamma(n, \lambda_0)$ distribution.

```
c1 <- qgamma(alpha, shape = n, rate
= lambda0, lower.tail = T)
```

```
power <- cgamma(c1, shape = n, rate
= lambdaA, lower.tail = T)
```

Sample Variance

Recall the variance of X:

$$\begin{aligned} var(X) &= \mathcal{E}((X - \mathcal{E}(X))^2) \\ &= \mathcal{E}(X^2) - \mathcal{E}(X)^2 \end{aligned}$$

Unbiased plug-in version: Given a sample $x_1, \ldots, x_n$, define

Sample Variance

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Sample Covariance

Recall the covariance of X and Y:

$$\begin{aligned} cov(X, Y) &= \mathcal{E}((X - \mathcal{E}(X))(Y - \mathcal{E}(Y))) \\ &= \mathcal{E}(XY) - \mathcal{E}(X)\mathcal{E}(Y) \end{aligned}$$

Unbiased plug-in version: Given a sample of pairs $(x_1, y_1), \ldots, (x_n, y_n)$, define

Sample Covariance

$$cov_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Sample Correlation

Recall the correlation coefficient of X and Y:

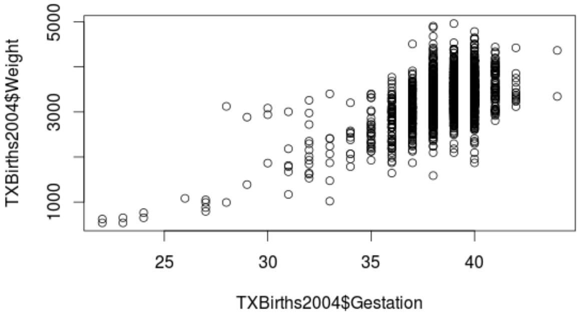
$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

Plug-in version: Given a sample of pairs $(x_1, y_1), \ldots, (x_n, y_n)$, define

Sample Correlation Coefficient

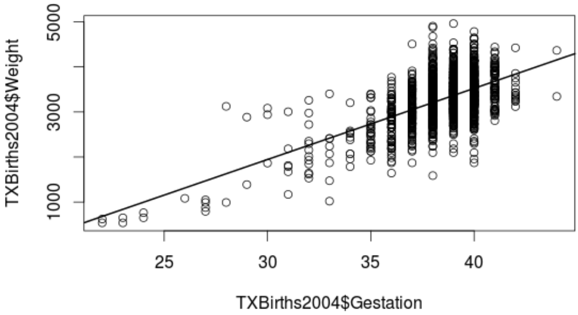
$$r = r_{xy} = \frac{cov_{x,y}}{s_x s_y}$$

Weight $\sim$ Gestation



Fitting a Line

Summarize this plot with a straight line:



Set-Up: Minimizing Residuals

Given n pairs $(x_1, y_1), \ldots, (x_n, y_n)$.
We want to find a straight line $y = \alpha + \beta x$ such that

$$y_i \approx \alpha + \beta x_i \quad (i = 1, 2, \ldots, n)$$

Residuals: $r_i = y_i - (\alpha + \beta x_i)$

Least squares: Minimize $F_2(\alpha, \beta) = \sum_i r_i^2$
Least absolute values:
Minimize $F_1(\alpha, \beta) = \sum_i |r_i|$
LASSO: Pick $\lambda > 0$.
Minimize $F(\alpha, \beta, \lambda) = \sum_i r_i^2 + \lambda(|\alpha| + |\beta|)$

Set-Up: Minimizing Residuals

Given n pairs $(x_1, y_1), \ldots, (x_n, y_n)$.
We want to find a straight line $y = \alpha + \beta x$ such that

$$y_i \approx \alpha + \beta x_i \quad (i = 1, 2, \ldots, n)$$

Residuals: $r_i = y_i - (\alpha + \beta x_i)$

Least squares: Minimize

$$F_2(\alpha, \beta) = \sum_i r_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$

Solution

- Unless all points are on a vertical line, there exists a unique solution.
- Formula for α, β: See *textbook*
- The optimal straight line satisfies $\bar{y} = \alpha + \beta\bar{x}$ and $\beta = r\frac{s_y}{s_x}$
- **R** implementation via `lm`, *linear model*

Some Notation

- The x_i come from an **explanatory variable**
- The y_i are values of the **response variable**
- Given α, β, the $\hat{y}_i = \alpha + \beta x_i$ are **predicted values** or **fits**
- The $r_i = y_i - \hat{y}_i$ are **residuals**
- *Explanatory variables may not be causes for responses*
- *Explanatory variables are not necessarily independent variables, response variables are not necessarily dependent variables.*

Regression toward the Mean

Recall

$$\beta = r\frac{s_y}{s_x}, \quad \bar{y} = \alpha + \beta\bar{x}, \quad y_i = \alpha + \beta x_i$$

Therefore:

$$\hat{y}_i - \bar{y} = \beta(x_i - \bar{x}) = r\frac{s_y}{s_x}(x_i - \bar{x})$$

$$\Longrightarrow \frac{\hat{y}_i - \bar{y}}{s_y} = r\frac{x_i - \bar{x}}{s_x}$$

So $\boxed{x_i - \bar{x} \approx s_x \implies \hat{y}_i - \bar{y} \approx r s_y}$*: "Regression toward the mean"*

Variation Explained

One can show

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

The LHS is s_y^2.
RHS $=$ variation of the residuals (unexplained by the regression) $+$ variation of the predictions.
One can show: $\frac{\sum_i (\hat{y}_i - \bar{y})^2}{n-1} = r^2 s_y^2$.

Therefore, $r^2 = var(\hat{y}_i)/var(y_i)$ $=$ "variation explained by the regression".

Examining Residuals

- Plot residuals against fitted values. *Look for curvature, outliers.*
- Histogram / QQ plot of residuals. *Look for bell-shape / skewedness / heavy tails*
- If available, time plot of residuals. *Trends due to changes in measurements?*

`plot(lm(...))` does all this and more.

Theoretical Assumptions

- Each y_i comes from a $N(\mu_i, \sigma^2)$ distribution
- $\mu_i = \alpha + \beta x_i$ and the x_i are known exactly.
- The σ^2 are all the same
- The y_i are independent

Statistical tasks:

- Estimate α, β, σ^2, CIs, hypothesis tests
- Estimate $\mu_i = \mathcal{E}(Y_i)$, CI
- Predict Y for a new x value

Basic Facts

- The MLEs $\hat{\alpha}, \hat{\beta}$ for α, β are exactly the least-squares estimates.
- Unbiased estimator: $\hat{\sigma^2} = \frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2$
- $\hat{\alpha}$, $\hat{\beta}$ have normal distributions.
- Can use t-tests for hypotheses about α and β. CIs are t-test based.
- Can make CIs for $\mathcal{E}(Y_i) = \mu_i$.
- *Prediction intervals for the Y_i are much wider.*

Multiple Linear Regression

Allow for more than one explanatory variable:

$$y_i \approx \alpha_+ \beta_i x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

or in matrix notation

$$y \approx X\beta$$

Example: Indicator variables Try to explain birth weight by including information about gender.
New variable: $g = 0/1$ for male/female babies.

Bootstrap - Basic Idea

Sample complete rows with replacement and build many regression models. Observe variability of the estimates.

Example: Make a bootstrap confidence interval for correlations between weight and gestation.

Example: Make a bootstrap confidence interval for the slope relating weight and gestation.