Statistics 262: Intermediate Biostatistics

Regression & Survival Analysis

Jonathan Taylor & Kristin Cobb
This course is an applied course, and we will look at the following types of data (to be explained, of course!)

- Linear Regression Models
- ANOVA
- Mixed Effects
- Generalized Linear Models
- Survival Models: (Non-, Semi-)Parametric
Logistics – People

- Instructor: Jonathan Taylor
  (jonathan.taylor@stanford.edu)

- Instructor: Kristin Cobb
  (kcobb@stanford.edu)

- TA: Pei Wang (wp57@stanford.edu)

- TA: Eric Bair (ebair@stat.stanford.edu)
Logistics: Computer labs

The following times are reserved for us (all the way over at the Fleischmann computer lab...)

- April 20th, 11:30am-1:00pm
- April 27th, Noon-1:00pm
- May 11th, 10:30am-12:30pm
- May 18th, 10:30am-12:30pm
Logistics: Evaluation

This is our proposed marking scheme:

- 6 assignments: 60%.
- take home final (40 %) OR project (40 %).
Logistics: Computing environment

- Although I prefer \texttt{R}, I will use \texttt{SAS} for examples in class.
- Students can submit assignments in either \texttt{R} or \texttt{SAS}.
What is a linear regression model?

Dumb example: Suppose you wanted to test the following hypothesis

\[ H: \text{Tall people tend to marry tall people.} \]

How could we do this?
Collect height information from married couples in the form of $(M_i, F_i), 1 \leq i \leq n$ for some number $n$ samples.

Plot pairs $(M_i, F_i)$.

If there is a “relationship” between $M$’s and $F$’s, you might conclude that $H$ is true.

What is relationship? A linear regression model says that

$$F_i = \beta_0 + \beta_1 M_i + \varepsilon_i$$

where $\varepsilon_i$ is “random error.”
Fitting a model

- Data seem to indicate linear relationship is good.
- How good is the fit?
- Can we “quantify” the quality of fit as in two-sample $t$-test?
- This is the basis of simple linear regression.
Usually, we care about more than one “predictor” at a time – *multiple* linear regression.

Example: predicting birth weight of babies based given mother’s habits during pregnancy: i.e. smoking / alcohol / exercise / diet.

We can estimate “effects” for many variables: are some of them less important than others?

With many variables there are many possible models: which is the “best”? 
Suppose you work for a pharmaceutical company developing a drug to lower cholesterol and you want to test the hypothesis:

Does the drug affect cholesterol, i.e. does it decrease cholesterol on average?

To test its effect, you might consider a case / control experiment with a placebo on a genetically pure strain of mice.

Observe \((D_i, P_i), 1 \leq i \leq n\) (could have different numbers of cases / placebos).
ANOVA: more general

- Question is $\mu_D$ (average cholesterol in cases) different than $\mu_P$ (average cholesterol in placebos)?

- This example is just a two sample $t$-test, but what if you had different dose levels / more than one drug / more than one strain?

- Other issues: in genetically pure mice, “average” effect is well-defined. What about in human studies?

- How do we generalize the results of the study to the entire population? (*random, mixed* vs. *fixed* effects models).
In HRP/STATS 261 you have seen some examples of a generalized linear model:

- logistic regression: trying to predict probabilities based on covariates;
- log-linear models: modelling count data / contingency tables;
- multiple linear regression models are also “generalized linear models”;
- these models can be put into a general class: GLMs.
Poisson regression

In a given HIV+ population, we might be obtain a viral genotype of patients at two different time points.

As HIV mutates in response to drug treatment, the virus undergoes mutations in between the two visits.

How quickly do mutations develop? Are there some covariates that influence this rate of mutation?

How do we model this?
Raw Data

The image contains a scatter plot with the x-axis labeled as 'T' and the y-axis labeled as 'N'. The plot displays various data points, each represented by a circle. The x-axis ranges from 5 to 40, and the y-axis ranges from 0 to 8.
Poisson regression model

- Basic Model (no other covariates):

  \[ N_i \sim \text{Poisson}(e^{\beta_0 + \beta_1 T_i}), \quad 1 \leq i \leq n \]

  where \( N_i \) is the number of mutations for subject \( i \), and \( T_i \) is the time between visits.

- This implies that

  \[ \log(E(N_i)) = \beta_0 + \beta_1 T_i, \quad \text{Var}(N_i) = e^{\beta_0 + \beta_1 T_i}. \]

- This is a “log” link, with “variance function”

  \[ V(x) = x. \]
Poisson regression fit
The “main” topic of this course is survival data. Survival data refers to a fairly broad class of data: basically, the object of study are “survival times” of individuals and the observations have three components:

- An observed time, \( T_i, 1 \leq i \leq n \).
- Observed covariates \( X_{ij}, 1 \leq i \leq n, 1 \leq j \leq p \).
- A “censoring” variable \( \delta_i, 1 \leq i \leq n \).
What is “δ”? In a given study, not everyone “fails”, eventually funding dries up and data has to be published!

Alternatively, people leave the area, or are otherwise “lost to follow-up.” However, knowing that they left the study “alive” has some information in it.

With much at stake, we do not want to throw away this data.

The difficulty comes in incorporating δ into the “likelihood.”
Kidney infection data

Given two different techniques of catheter placement in patients (surgically, percutaneously), physicians want to decide if one technique is better at holding off infections.

- Observations: time on study, presence of infection, catheter placement.
- Censoring can bias usual estimates of CDF: Kaplan-Meier estimate gives an unbiased estimate.
- Can we determine whether “survival” experience are different in the two groups? *(log-rank test)*
Kidney data

\[ \begin{align*}
\text{P}(T > t) & \quad \text{for} \quad t = 0, 5, 10, 15, 20, 25 \\
0 & \quad 0.0, 0.2, 0.4, 0.6, 0.8, 1.0
\end{align*} \]
Other survival models

- We could model time as coming from a parametric family of distributions: Gumbel, Weibull, etc.

- Alternatively, we can fit a semi-parametric model like Cox’s proportional hazards model (which is actually not appropriate in this case....)
Moving on

- End of introduction.
- After break: regression!
Simple linear regression

- Specifying the model.
- Fitting the model: least squares.
- Inference about parameters.
- Diagnostics.
Returning to husband and wife data:

\[ F_i = \beta_0 + \beta_1 M_i + \varepsilon_i \]

- Assumption: \( E(\varepsilon_i | M_i) = 0 \) (often: \( \varepsilon_i | M_i \sim N(0, \sigma^2) \))
- Regression equation: \( E(F_i | M_i) = \beta_0 + \beta_1 M_i \)
- Errors are independent across pairs of couples.
- This fully specifies joint distribution of \( (F_1, \ldots, F_n) \) given \( (M_1, \ldots, M_n) \).
Least squares regression chooses the line that minimizes

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (F_i - \beta_0 - \beta_1 M_i)^2.$$ 

(Conditional) mean can be estimated for any given height $M$ as

$$\hat{F}(M) = \hat{\beta}_0 + \hat{\beta}_1 \cdot M.$$ 

where $(\hat{\beta}_0, \hat{\beta}_1)$ are the minimizers of SSE.
Estimating $\sigma^2$

- Strength of association will depend on variability in data, as in two sample $t$-test.
- Natural estimate of $\sigma^2$
  \[
  \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left( F_i - \hat{F}(M_i) \right)^2 .
  \]
- Under normality assumption
  \[
  \frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi_{n-2}^2}{n-2} .
  \]
LIBNAME DATADIR 'C:sas';

PROC IMPORT OUT=DATADIR.height
    DATAFILE = 'C:heights.csv' DBMS=CSV REPLACE;
RUN;

PROC REG DATA=DATADIR.height;
    MODEL WIFE=HUSBAND;
RUN;
Inference: CIs

- Under the normality assumption, any linear combination of \((\hat{\beta}_0, \hat{\beta}_1)\) are also normally distributed.

- Can be used to construct confidence interval for predicted mean:

\[
\hat{F}(M) \pm t_{n-2,1-\alpha/2} \cdot \hat{SD}(\hat{F}(M))
\]

- As well as a new observation

\[
\hat{F}(M) \pm t_{n-2,1-\alpha/2} \cdot \sqrt{\hat{SD}(\hat{F}(M))^2 + \hat{\sigma}^2}
\]
SAS: predicted mean, CIs

PROC REG DATA=DATADIR.height;
   MODEL WIFE=HUSBAND / P CLM CLI;
RUN;

- \texttt{P} refers to “predicted” mean.
- \texttt{CLM} refers to “mean.”
- \texttt{CLI} refers to new “individual.”
Inference: Hypothesis tests

- We can test our “hypothesis” formally in terms of the coefficient $\beta_1$.
- If the regression function is truly linear, then $\beta_1 = 0$ means height of husband has no effect on height of wife.
- Under $H_0 : \beta_1 = 0$

$$\frac{\hat{\beta}_1}{SD(\hat{\beta}_1)} \sim t_{n-2}.$$
PROC REG DATA=DATADIR.height;
    MODEL WIFE=HUSBAND;
    MYTEST: TEST WIFE=0;
    OTHERTEST: TEST WIFE=1;
RUN;

Output has a page for MYTEST and OTHERTEST.

If there is more than one covariate, more than one variable can be specified per TEST statement.
PROC REG DATA=DATADIR.height;
   MODEL WIFE=HUSBAND;
   PLOT WIFE*HUSBAND;
RUN;
Least Squares Fit
Diagnostics

- How do we tell if our assumptions are justified?
- Can we check if we have left a higher order term out?
- Is the variance constant across couples?
- Are the residuals close to being normally distributed?
SAS: residual plot

```
PROC REG DATA=DATADIR.height;
  MODEL WIFE=HUSBAND;
  PLOT R.*P. / VREF=0;
RUN;
```

- \( R_i \) refers to “residual” \( R_i = F_i - \hat{F}(M_i) \).
- \( P \) refers to “predicted” value \( \hat{F}(M_i) \).

Can detect missing higher order terms, nonconstant variance.
Residual Plot: Count data
Residual Plot
PROC REG DATA=DATADIR.height;
   MODEL WIFE=HUSBAND;
   PLOT R.*NQQ.;
RUN;

NQQ refers to “quantile-quantile”.
Can detect departures from normality.