Topic 22 - Linear Regression and Correlation

STAT 511

Professor Bruce Craig

Background Reading

Devore : Section 12.1 - 12.5

Overview

- Consider **one** population but **two** variables
- For each sampling unit observe X and Y
- Assume <u>linear</u> relationship between variables
- Regression/correlation assess association
- Relationship may be non-linear but linear in a particular region
- Can often transform Y and/or X to create linear association
  - Dose-response curve: Response with log(dose)
  - Lineweaver-Burk: 1/velocity with 1/concen

Overview

- A scatter plot allows visual assessment of relationship
- One variable (X) plotted on x-axis, other variable (Y) plotted on y-axis
- Linear regression determines "best" line through (X, Y) pairs

$$Y = b_0 + b_1 X$$

- Correlation describes the "tightness" of the linear fit

$$-1 \leq r \leq 1$$

Scatter Plot Example

The VO_2 max readings for 8 healthy adults following exercise are recorded. Does it appear that VO_2 max decreases with an increase in activity? Create a scatterplot to investigate the relationship.

Subject	VO_2 Max	Duration of Exercise
1	82	9.5
2	74	9.9
3	63	10.2
4	65	10.0
5	58	10.7
6	44	11.0
7	55	10.8
8	48	11.0

Basic Computations

- Have n pairs of (x, y) values
- Univariate summary statistics needed for analysis

Statistic	X	Y
Mean	$\overline{x}$	$\overline{y}$
Sum of Squares	$\mathsf{S}_{xx} = \sum (x - \overline{x})^2$	$\mathsf{S}_{yy} = \sum (y - \overline{y})^2$
Std Deviation	$s_x = \sqrt{\frac{\mathsf{S}_{xx}}{n-1}}$	$s_y = \sqrt{\frac{\mathsf{S}_{yy}}{n-1}}$

- Also need joint summary statistic

$$\mathsf{S}_{xy} = \sum (x - \overline{x})(y - \overline{y})$$

Basic Computations

- Sign of S_{xy} indicates direction of trend
- $(x - \overline{x})(y - \overline{y})$ positive
  - $x > \overline{x}$ and $y > \overline{y}$
  - $x < \overline{x}$ and $y < \overline{y}$
- $(x - \overline{x})(y - \overline{y})$ negative
  - $x > \overline{x}$ and $y < \overline{y}$
  - $x < \overline{x}$ and $y > \overline{y}$

Least Squares Estimation

- Many different ways to fit line
- Need criterion to assess "best" fit
- The least squares criterion estimates are

$$b_1 = \frac{\mathsf{S}_{xy}}{\mathsf{S}_{xx}} \qquad b_0 = \overline{y} - b_1 \overline{x}$$

- Estimates also labeled $\widehat{\beta}_0$ and $\widehat{\beta}_1$

Least Squares Estimation

- Given b_0 & b_1, the predicted value for $x^\star$

$$\widehat{y} = b_0 + b_1 x^\star$$

- Residual is $y - \widehat{y}$ and represents the vertical distance of y from fitted line
- Least squares minimizes the sum of these squared residuals

$$\mathsf{SSE} = \sum (y - \widehat{y})^2$$

- Can show estimates result in

$$\mathsf{SSE} = \mathsf{S}_{yy} - \frac{\mathsf{S}_{xy}^2}{\mathsf{S}_{xx}}$$

Residual Standard Deviation

- Describes the "closeness" of the data to fitted line
- How far above/below line y's tend to be
- Std deviation based on squared residuals

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

- Similar to s_y estimate except for $\hat{y}_i$ and $n-2$
- Use $n-2$ because variation about line (i.e., we're using $\hat{y}_i$ not $\overline{y}$ as the predicted value)
- Normal $\rightarrow$ approx 95% of obs within $\pm 2s$

Linear Model of X and Y

- Want to generalize sample to population
- Assume for value of X, can observe diff values of Y
  - If X indicates trt, similar to ANOVA
  - Dist of Y's assumed Normal with unknown mean
- Given X, have conditional dist of Y
- Model mean of $Y|X = x$ as linear function

$$\text{E}(Y|X = x) = \beta_0 + \beta_1 x$$
$$\downarrow$$
$$Y|X = x \sim \text{N}(\beta_0 + \beta_1 x, \sigma^2)$$

Assumptions

- The conditional distribution of Y is

  Normally Distributed

  $\mu_{Y|X=x} = \beta_0 + \beta_1 x$

  $\sigma_{Y|X=x}$ constant (can drop $X = x$)

- Consider ANOVA problem
  - Assume mean depends on treatment group
  - Assume constant variance
- ANOVA just special case of regression

Prediction

- Recall predicted value for $X = x_o$ is

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

- Must use caution in interpretation of $\hat{y}$
- If x_o within range of x's $\rightarrow$ interpolation
- If x_o outside range of x's $\rightarrow$ extrapolation
- Extrapolation should be avoided
  - No assurances still linear outside range

Prediction

- Standard error of $\widehat{y}$ depends on whether estimating

  a Conditional mean $\rightarrow$ point on regression line

  b Future observation $\rightarrow$ point may vary from line

$$\overset{(a)}{\sqrt{s^2\left(\frac{1}{n}+\frac{(x_o-\overline{x})^2}{S_{xx}}\right)}} \qquad \overset{(b)}{\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_o-\overline{x})^2}{S_{xx}}\right)}}$$

- Similar argument to confidence interval vs prediction interval

Example

Recall the VO_2 Max example. Construct a 95% CI for the mean VO_2 level when a healthy adult exercises for 10.5 minutes.

First, need to estimate the regression line. The summary statistics necessary for this calculation are shown below.

$$\sum y = 489 \qquad \sum x = 83.1$$

$$\sum y^2 = 31023 \qquad \sum xy = 5030.8 \qquad \sum x^2 = 865.43$$

From these,

$S_{yy} = 31023 - 489^2/8 = 1132.88$

$S_{xx} = 865.43 - 83.1^2/8 = 2.23$

$S_{xy} = 5030.8 - 489(83.1)/8 = -48.69$

so

$b_1 = -48.69/2.23 = -21.84$

and

$b_0 = (489/8) + 21.84(83.1/8) = 288.0$

The table below computes the pred values and residuals

x	y	Predicted	Residual	$(y-\widehat{y})^2$
9.5	82	80.520	1.480	2.190
9.9	74	71.784	2.216	4.917
10.2	63	65.232	-2.232	4.982
10.0	65	69.600	-4.600	21.160
10.7	58	54.312	3.688	13.601
11.0	44	47.760	-3.760	14.138
10.8	55	52.128	2.872	8.248
11.0	48	47.760	0.240	0.058
83.1	489		0.000	69.288

Since SSE = 69.288, $s = \sqrt{69.288/6} = 3.40$.

The pred value for $x = 10.5$ is $288.0 - 21.84(10.5) = 58.68$.

Interested in the average VO_2 level, so

$$\text{SE}(\widehat{y}) = 3.40\sqrt{\tfrac{1}{8} + \tfrac{(10.5-10.3875)^2}{2.23}} = 1.23.$$

Because we use s, the df is 6 so a 95% CI is

$$58.68 \pm 2.447(1.23) = (55.67, 61.69).$$

An interval which 95% of the time would contain the observed y for a person exercising $x = 10.5$ minutes is

$$58.68 \pm 2.447(3.615) = (49.83, 67.53).$$

Standard Error of b_1

- As with all estimates, b_1 subject to sampling error

- Standard error of b_1

$$s_{b_1} = \sqrt{\frac{s^2}{S_{xx}}}$$

- In situations where X's are under experimental control

  If S_{xx} made large $\rightarrow$ small SE

  Increase S_{xx} by increasing dispersion of x (spread out)

  If increase $n \rightarrow S_{xx}$ increases

Inference of β_1

- Need sampling distribution to construct CI or perform hypothesis test
- Given normality assumption, sampling distribution of b_1 is also normal

  CI: $b_1 \pm t_\alpha \mathsf{s}_{b_1}$ (df $= n-2$)

  Hypothesis test

  $$H_0 : \beta_1 = \beta_1^\star$$

  $$t_s = \frac{b_1 - \beta_1^\star}{\mathsf{s}_{b_1}}$$

- Can look at $\beta_1 = 0$ to see if there is a linear association

Standard Error of b_0

- Sometimes interested in intercept β_0
- Standard error of b_0

$$\mathsf{s}_{b_0} = \sqrt{s^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\mathsf{S}_{xx}} \right)}$$

- In situation where X is under experimental control

  If S_{xx} made large $\rightarrow$ small SE

  Increase S_{xx} by increasing dispersion

  If $\overline{x}$ close to zero $\rightarrow$ small SE

  If increase $n \rightarrow$ increase SS_X

Inference of β_0

- Need sampling distribution to construct CI or perform hypothesis test
- Given normality assumption, sampling distribution of b_0 is also normal

  CI: $b_0 \pm t_\alpha \mathsf{s}_{b_0}$ (df$= n-2$)

  Hypothesis tests

  $$H_0 : \beta_0 = \beta_0^\star$$

  $$t_s = \frac{b_0 - \beta_0^\star}{\mathsf{s}_{b_0}}$$

- Can compute joint confidence regions using F dist

Example

Recall the VO_2 Max problem. The standard errors for both b_0 and b_1 are

$$\mathsf{s}_{b_1} = \sqrt{\frac{(69.288/6)}{2.23}} = 2.27$$
$$\mathsf{s}_{b_0} = \sqrt{(69.288/6)(\frac{1}{8} + \frac{(83.1/8)^2}{2.23})} = 23.67$$

The 95% CI are

$$-21.84 \pm 2.447(2.27) = (-27.39, -16.29)$$
$$288.0 \pm 2.447(23.67) = (230.08, 345.92)$$

Since 0 is not in the CI for β_1, we can say that there is a linear association and it is a negative association. As the amount of exercise increases, there is a decrease in the VO_2 max.

We must be careful interpreting anything outside of the x range. We should not feel comfortable saying that the VO_2 max at rest is somewhere between 230.08 and 345.92 with 95% confidence.

The Correlation Coefficient

- Describes how close the data cluster about the line
- Describe direction and "tightness"
- Correlation coefficient is a dimensionless statistic

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

- Symmetry - can interchange X and Y with altering the value

The Correlation Coefficient

- Properties
  - r has same sign as b_1
  - r^2 known as coefficient of determination. % of total variation in Y explained by regression

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS}_Y}$$

  - If straight line fit $\rightarrow$ SSE = 0

    $r = \pm 1$ and $r^2 = 100\%$

  - If no linear association $\rightarrow$ SSE = SS_Y

    $r = 0$ and $r^2 = 0\%$

Coefficient of Determination

Determines % of variability in Y explained by linear relationship with X

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST})}$$

- Can use r^2 to approximate reduction in std dev

$$\frac{s}{s_Y} \approx \sqrt{1 - r^2}$$

- Prior to regression, the std dev of Y is s_Y
- After regressing X on Y the std dev is s
- $\sqrt{1 - r^2}$ approx reduction in std dev (i.e., closeness)

Hypothesis Test for ρ

- r is sample correlation coefficient
- Use ρ to denote the pop correlation coefficient
- Under linear model with normal errors

$$\rho = \beta_1 \frac{\sigma_X}{\sigma_Y} \rightarrow b_1 \sqrt{\frac{\text{SS}_X}{\text{SS}_Y}} = r$$

- Can do t-test to see if population linear association (same as $H_0 : \beta_1 = 0$)

$$t_s = r\sqrt{\frac{n-2}{1-r^2}}$$

Confidence Interval for ρ

- If sample size large, can construct CI for ρ
- Based on Fisher transformation of r

$$V = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \approx \mathrm{N}\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$

- Construct CI for V, then convert back to CI for ρ

$$\begin{aligned} c_1 &= v - z_{\alpha/2}\frac{1}{\sqrt{n-3}} \\ c_2 &= v + z_{\alpha/2}\frac{1}{\sqrt{n-3}} \\ &\downarrow \end{aligned}$$

$$\left(\frac{\exp(2c_1)-1}{\exp(2c_1)+1}, \frac{\exp(2c_2)-1}{\exp(2c_2)+1}\right)$$

Example - Transformation

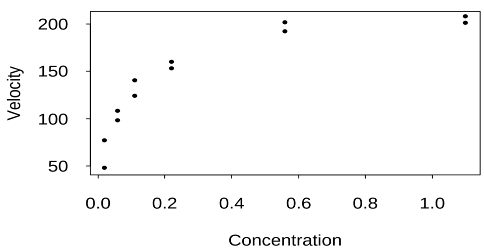
(Bates and Watts: Nonlinear Regression) Consider the data set used to describe the relationship between "velocity" of an enzymatic reaction (V) and the substrate concentration (C). Consider only the experiment where the enzyme is treated with Puromycin.

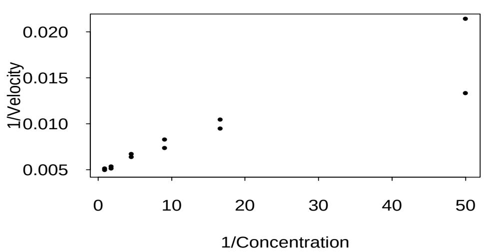
The common model used to describe the relationship between "velocity" and concentration is the Michaelis-Menten model

$$V = \frac{\theta_1 C}{\theta_2 + C}$$

where θ_1 is the maximum velocity of the reaction and θ_2 describes how quickly (in terms of increasing concentration) the reaction will reach maximum velocity.

Scatterplot





Since this is a non-linear model, one approach is to transform the variables so that a linear relationship exists. While this usually works quite well, one must be aware that the transformation changes the distribution of the data.

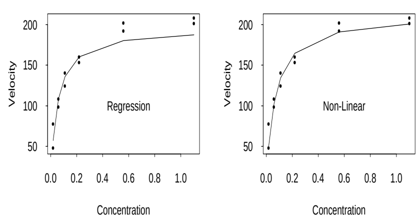
For example, with this model, we can rewrite it as a linear model if we look at the inverse concentration and inverse velocity.

$$\frac{1}{V} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\left(\frac{1}{C}\right)$$

One must be careful with this transformation for two reasons. First, since we are looking at inverse concentrations, very low concentrations will be highly influential in the regression analysis. Second, the equal variance assumption is often violated. The variance violation is more easily picked up when there are replicates. Also notice that the two observations at a very low concentration are now much further separated from the others making these observations more influential.

Since the variance appears constant in the untransformed plot, the better way to estimate the parameters is to use non-linear estimation methods. These procedures are beyond the scope of the class but one could view it as an iterative least squares approach where some initial estimates of the parameters are given, the predicted values are calculated using the untransformed model, and new parameter estimates are proposed until the residual sum of squares decreases to a minimum. The two plots below show the fitted line in reference to the original data using both the linear and non-linear approaches.

Method	$\widehat{\theta}_1$	$\widehat{\theta}_2$
Regression	195.80	0.0484
Non-Linear	212.70	0.0641



SAS Proc Reg

```
option nocenter ls=75;
goptions color=('none');

PROC IMPORT OUT=EXAMPLE
            DATAFILE= "U:\.www\datasets511\exp12-01.xls"
            DBMS=EXCEL2000 REPLACE;
     GETNAMES=YES;
RUN;

proc gplot data=example;
   plot y_*x_;
run;

proc reg data=example;
   model y_=x_  / clb clm cli p r;
   output out=a2  p=pred r=resid;

proc gplot data=a2;
   plot resid*x_/ vref=0;
run;
```


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	39.68595	39.68595	418.32	<.0001
Error	28	2.65634	0.09487		
Corrected Total	29	42.34230			

Root MSE	0.30801	R-Square	0.9373
Dependent Mean	2.84033	Adj R-Sq	0.9350
Coeff Var	10.84411		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.39774	0.16801	-2.37	0.0251
x_	1	3.07997	0.15059	20.45	<.0001

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	-0.74189	-0.05359
x_	1	2.77150	3.38843

Output Statistics

Obs	Dep Var y_	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	1.0200	0.8342	0.1131	0.6027	1.0658
2	1.2100	0.8958	0.1105	0.6696	1.1221
3	0.8800	1.0806	0.1028	0.8701	1.2912
4	0.9800	1.1730	0.0990	0.9702	1.3759
5	1.5200	1.3578	0.0917	1.1699	1.5458
6	1.8300	1.4502	0.0882	1.2695	1.6309
7	1.5000	1.7582	0.0772	1.6001	1.9164

Obs	95% CL Predict		Residual	Std Error Residual	Student Residual	-2-1 0 1 2
1	0.1622	1.5063	0.1858	0.287	0.648	| |* |
2	0.2256	1.5661	0.3142	0.288	1.093	| |** |
3	0.4155	1.7458	-0.2006	0.290	-0.691	| *| |
4	0.5103	1.8358	-0.1930	0.292	-0.662	| *| |
5	0.6995	2.0162	0.1622	0.294	0.552	| |* |
6	0.7939	2.1065	0.3798	0.295	1.287	| |** |
7	1.1078	2.4087	-0.2582	0.298	-0.866	| *| |