Linear Regression

Linear regression is another method used to assess association between two (or more) variables.

Linear regression attempts to describe the nature of the association by constructing a “best-fitting” mathematical model.

When using linear regression we assume that the variables are associated in a linear fashion, and we attempt to find that line that explains the association.
Linear Regression Model:

For possibly associated variables $X$ and $Y$ we assume the model

$$Y = \alpha + \beta \cdot X + \varepsilon.$$  

$\alpha$ and $\beta$ are unknown “coefficient” values.

$X$ and $Y$ indicate the random variables we observe.

$\varepsilon$ is assumed to be a mean-zero Normal random variable.

The equation $Y = \alpha + \beta \cdot X$ defines the average association between $X$ and $Y$. Sometimes write $E(Y|X = x) = \alpha + \beta \cdot x$.

The $\varepsilon$ is the random error thrown in on top that defines the “real-life” values. It is the part of $Y$ not explained by $X$. 
Linear Regression vs. Correlation

- Correlation is a measure of the **strength of the association**. $\rho^2$ can be loosely described as the percent of variation in the Y’s that is explained by the X’s.

- Linear regression attempts to describe the **form of the association**.

- The slope parameter $\beta$ is related to $\rho$:

$$\beta = \frac{\sigma_Y}{\sigma_X} \rho$$
Find the line of best fit

For every possible line, we assess how closely it fits the observed data by calculating the vertical (Y) distances to all the points, $e_i$ called the “residuals”.

The line of best fit is defined as the line that minimizes the sum of the squared residuals.

More formally: the best fitting line $y = a + bx$, is defined by the $a$ and $b$ that minimize

$$
\sum_i e_i^2 = \sum_i \left( y_i - (a + bx_i) \right)^2.
$$
Finding line of best fit (continued)

The coefficients of the best fitting line, $a$ and $b$, estimates of $\alpha$ and $\beta$, respectively, can be calculated via the formulae

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}.$$ 

Note: $b$ is related to $r$: 

$$b = \frac{s_y}{s_x} r$$
Example: attachment level and cigarettes/day

smoking amount and attachment level (28 smokers)

SPSS output

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>2.319</td>
<td>.635</td>
<td>3.653</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>cigarettes smoked/day</td>
<td>.067</td>
<td>.032</td>
<td>.380</td>
<td>.046</td>
</tr>
</tbody>
</table>

This says $a = 2.319$, $b = .067$, and the best fitting line is

$$Y = 2.319 + 0.067\times X,$$

where $Y = \text{mean attachment level}$ and $X = \text{cigs smoked/day}$. 
**Interpretation of regression coefficients**

The best fitting line is

\[ Y = 2.319 + 0.067X, \]

where \( Y \) = mean attachment level and \( X \) = cigs smoked/\text{day}.  

This can be interpreted as saying that “each extra cigarette smoked per day is associated with an additional 0.067 mm of attachment loss.”

Also “each pack smoked per day (20 cigs) is associated with an additional 0.067\times20 = 1.34 \text{ mm loss of attachment.”}"

**Use estimated model for estimation or prediction**

Best estimate of mean attachment level for people who smoke 30 cigarettes/day is

\[ 2.319 + 0.067\times30 = 4.329 \text{ mm.} \]
Goodness of fit

A key measure of the strength of the association is the “mean squared error”, MSE, which is basically the average of the squared residuals.

$$MSE = \frac{1}{n-2} \sum_i e_i^2 = \frac{1}{n-2} \sum_i (y_i - (a + bx_i))^2$$

If this value is small with respect to the sample variance of the y’s, then we consider our regression model to be a worthwhile explanation of the association. See notes pp 146-8, or Rosner §11.8

The MSE is also used to estimate the standard error of $b$.

$$SE(b) = \frac{\sqrt{MSE}}{\sqrt{(n-1)s_x^2}}$$

Note:

1. $SE(b)$ ↓ as $MSE$ ↓

2. $SE(b)$ ↓ as $s_x$ ↑

That is, we get better estimates of $\beta$ when the line is a good fit, and when the X’s are more spread out.
Interpretation of Correlation Coefficient with respect to Regression

- Let $\hat{y}_i = a + b \cdot x_i$ be the predicted values of $y$ from the regression equation.

- Note that $y_i = \hat{y}_i + e_i$

- Then the square of the Pearson correlation is equal to the proportion of the total variation of the $y_i$ that is explained by the prediction line:

$$r^2 = \frac{\text{Var}(\hat{y}_i)}{\text{Var}(y_i)}$$

- If the correlation is perfect ($r = 1$) that says that all the variation in the $y_i$ is explained by the prediction line.
Inference for regression coefficients

If the $\epsilon_i$ are Normally distributed then $b/SE(b) \sim t_{n-2}$.

- A $(1-\alpha)100\%$ confidence interval for $\beta$ is
  
  $b \pm t_{n-2,1-\alpha/2}SE(b)$.

- Can test $H_0: \beta = 0$ vs $H_1: \beta \neq 0$ using the statistic
  
  $t = \frac{b}{SE(b)}$.

**Example: Cigarettes and attachment level**

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1</td>
<td>2.319</td>
<td>.635</td>
</tr>
<tr>
<td>(constant)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cigarettes smoked/day</td>
<td>.067</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Dependent Variable: mean attachment level*

We can test whether there is evidence that self-reported cigarettes smoked/day is related to attachment; that is, test $H_0: \beta = 0$ vs $H_1: \beta \neq 0$.

$$t = \frac{0.067}{0.032} = 2.098.$$  

Compare to $t_{26}$ distribution. Since $t$ is greater than $t_{26,.975} = 2.06$, reject at 0.05 level (p-value $= P(|t_{26}| > 2.098) = 0.046$).

A 95% confidence interval for $\beta$ is

$$0.067 \pm 2.06 \cdot 0.032 = (0.001, 0.133).$$
Confidence intervals for regression-based predictions

- The best guess (point estimate) for $Y$ given $X = x$ is
  \[ y = a + bx. \]

- For estimation of the average $Y$ given $X = x$:
  \[
  SE(Y_{\text{ave}} \mid X = x) = \sqrt{MSE \cdot \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_x^2} \right)},
  \]
  and the $(1 - \alpha)100\%$ confidence interval is
  \[
  a + bx \pm t_{n-2,1-\alpha/2} SE(Y_{\text{ave}} \mid X = x).
  \]

- For the prediction of a single $Y$ given $X = x$:
  \[
  SE(Y \mid X = x) = \sqrt{MSE \cdot \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) \cdot s_x^2} \right)},
  \]
  and the $(1 - \alpha)100\%$ confidence interval is
  \[
  a + bx \pm t_{n-2,1-\alpha/2} SE(Y \mid X = x).
  \]
To find the SE for our best guess of the average attachment level of people who smoke 30 cigarettes/day we need the MSE. Can get it from this SPSS “Model Fit” output:

\[
\text{ANOVA}
\]

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>7.896</td>
<td>1</td>
<td>7.896</td>
<td>4.401</td>
<td>.046(^{a})</td>
</tr>
<tr>
<td>Residual</td>
<td>46.645</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>54.541</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}\) Predictors: (Constant), self report cigs./day

\(^{b}\) Dependent Variable: mean attachment level (mm)

\[
SE(Y_{\text{ave}} \mid X = 30) = \sqrt{1.794 \cdot \left( \frac{1}{28} + \frac{(30 - 18.179)^2}{(27) \cdot 8.051^2} \right)} = 0.455
\]

So 95% confidence interval for average MAL of people who smoke 30 cigarettes/day is

\[4.329 \pm 2.06 \cdot 0.455 = (3.392 \ 5.266)\]
Pointwise 95% confidence bands

- Created by computing the confidence intervals for regression predictions at each X value in the data.

- Inner longer dashed lines are confidence bands for average MAL

- Outer small dash lines are confidence bands for individual MAL

- The precision of the prediction is better near the center of the data.
Multiple Linear Regression

The linear regression model can be extended to include multiple independent variables.

\[ Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_k \cdot X_k + \varepsilon. \]

- Coefficients are still estimated by minimizing squared residuals.

- Residuals are now the difference between the observed Y value and a predicted value that is computed from a more complicated prediction formula which includes multiple variables.

- In this more complicated situation there are no simple formulae for the estimates. A computer is usually used to calculate the coefficient estimates.
Example: Attachment as a function of Cigarette use and Age

With multiple regression we can also put age along with cigarette use into the regression model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>95% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>1</td>
<td>-1.065</td>
<td>1.377</td>
<td>.446</td>
</tr>
<tr>
<td></td>
<td>.071</td>
<td>.029</td>
<td>.402</td>
</tr>
<tr>
<td></td>
<td>.074</td>
<td>.028</td>
<td>.440</td>
</tr>
</tbody>
</table>

- The coefficients can be considered to be “adjusted” for each other.

- The estimate of the effect of cigarettes is now more precise (smaller std. error) as we have removed some of the residual variation by including age in the model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>36.116</td>
<td>25</td>
<td>1.445</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>54.541</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Overall goodness of fit is better (smaller MSE) with the addition of a good explanatory variable to the model.

Notes:
- a. Dependent Variable: mean attachment level (mm)
- b. Predictors: (Constant), age (yrs), self report cigs./day
- b. Dependent Variable: mean attachment level (mm)
Using categorical variables as predictor variables

Categorical variables can be used as predictors in the regression context using “dummy” variables.

For modelling the effect of a variable $X$ with $k$ categories, we create a set of variables $X_i$

- Each $X_i$ corresponds to one of the $k$ categories
  - $X_i = 1$ if the observation is in category $i$
  - $X_i = 0$ if the observation is not in category $i$
- We put all the $X_i$’s in the regression model except one.
- The left-out category/$X_i$ is the “comparison” category.

In the resulting model
\[
Y = \alpha + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \ldots + \beta_k \cdot X_k + \varepsilon,
\]

The comparison category is category 1.

- The mean of the comparison category is $Y = \alpha$
- The mean of group 2 is $Y = \alpha + \beta_2$
- The parameter $\beta_i$ is interpreted as the difference between group $i$ and the comparison category
Example: Gum data

Put the “group” variable into a regression model with group as the independent variable and the outcome is change in DMFS.

There are three categories of group (gum A, gum B, gum C) so create categorical variables:

\[ X_1 = \begin{cases} 1 & \text{for children who got gum A} \\ 0 & \text{for children who did not get gum A} \end{cases} \]

\[ X_2 = \begin{cases} 1 & \text{for children who got gum B} \\ 0 & \text{for children who did not get gum B} \end{cases} \]

We have not created a variable indicating membership in group C, as group C will be our comparison category.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>B</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.63</td>
<td>0.66</td>
<td>3.98</td>
<td>0.000</td>
<td>1.32</td>
<td>3.93</td>
<td></td>
</tr>
<tr>
<td>X_1</td>
<td>-3.35</td>
<td>1.06</td>
<td>-3.15</td>
<td>0.002</td>
<td>-5.46</td>
<td>-1.24</td>
<td></td>
</tr>
<tr>
<td>X_2</td>
<td>-3.44</td>
<td>0.97</td>
<td>-3.58</td>
<td>0.001</td>
<td>-5.37</td>
<td>-1.54</td>
<td></td>
</tr>
</tbody>
</table>

The coefficient for \( X_1 \), -3.345, is interpreted as the mean difference between gum A and gum C.
Example: adjusted t-test

We can use a regression model to compute an independent-samples t-test with equal variances by using measure of interest as the dependent variable, and an indicator variable for group membership as a single independent variable.

Here we are using NHANES3 data to compare mean attachment levels between former smokers and current smokers.

### Independent Samples Test

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>df</th>
<th>Sig.</th>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean attachment loss</td>
<td>-.18</td>
<td>98</td>
<td>.86</td>
<td>-0.06</td>
<td>0.35</td>
<td>-0.76</td>
</tr>
</tbody>
</table>

### Linear Regression Model

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>2.010</td>
<td>0.25</td>
<td>8.0</td>
<td>.000</td>
</tr>
<tr>
<td>current smoker</td>
<td>-0.06</td>
<td>0.35</td>
<td>-0.18</td>
<td>0.86</td>
</tr>
</tbody>
</table>

dependent variable: mean attachment loss
Example: adjusted t-test (continued)

Note that Age explains a lot of the variation in the mean attachment loss variable.

If we adjust for the effect of age using the linear regression model, then we remove the variability associated with age and get a more precise test.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>B</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>-0.74</td>
<td>0.41</td>
<td>-1.81</td>
<td>0.074</td>
</tr>
<tr>
<td>current smoker</td>
<td>0.65</td>
<td>0.29</td>
<td>2.22</td>
<td>0.029</td>
</tr>
<tr>
<td>decade</td>
<td>0.50</td>
<td>0.07</td>
<td>7.69</td>
<td>0.000</td>
</tr>
</tbody>
</table>

We now see a significant association between current/former smoking and attachment loss.
Correlated Data

• Most standard statistical inference methods rely on the assumption that the observations are independent.

• With many common types of data this assumption may be violated
  
  o Observations on multiple teeth within the same mouth (e.g. 4 third molars for each patient)
  
  o Cluster randomized clinical trial, where instead of randomly assigning each participant, the study randomized groups of participants (e.g. dental offices, classrooms)
  
  o Cluster sampling, randomly sample dental offices and then sample observations/participants within each sampled dental office

• In most cases involving correlated data the standard estimators will be unbiased.

• The usual standard error estimates can be incorrect, which can lead to incorrect p-values and confidence intervals.
50,000,000 ELVIS FANS CAN’T BE WRONG

ELVIS' GOLD RECORDS - Volume 2

A FOOL SUCH AS I
I NEED YOUR LOVE TONIGHT
WEAR MY RING AROUND YOUR NECK
DOING' THINK IT'S TIME
I BEG OF YOU
A BIG HUNK O' LOVE
DON'T
DRY WISH CAME TRUE
ONE NIGHT
I COT STRING

LPM-2075
Strategies for analyzing Correlated Data

• Reduce data to summaries over independent units.
  
  o For third molars, the independent unit would be patient. So instead of reporting whether each *tooth* was removed, can use data that indicates whether each *patient* had at least one third molar removed.

  o For attachment loss instead of looking at individual periodontal sites can summarize over patient (mean attachment level)

  o For cluster randomized trials could summarize to the cluster level; for example, average number of new caries for all children in classroom

• Note this strategy reduces the sample size (n) to the number of independent units rather than number of observations

• With this strategy one cannot compare characteristics that vary within the independent units.

  o Could not compare whether impacted versus non-impacted third molars were removed more if some patients had some of each.
Generalized Estimating Equations

• Generalized Estimating Equations (GEE) is a regression-based approach
• Works with “clustered” data
• The clusters are groups of observations that may be correlated with other observations within the same cluster, but are assumed to be independent from observations in other clusters.
• A common GEE procedure estimates the parameters as before; that is, the same estimates one would get assuming all observations are independent
• A Robust Variance Estimator is used to estimate correct standard errors that account for the correlation between individuals within clusters.