Lecture 9: Linear Regression

- Model
- Inferences on regression coefficients
- $R^2$
- Residual plots
- Handling categorical variables
- Adjusted $R^2$
- Model selection
- Forward/Backward/Stepwise

November 4, 2010

Example: Lung Function in CF pts

- A study on lung function in patients with cystic fibrosis
- PEmax (maximal static expiratory pressure, cm H$_2$O) is the response variable
- A potential list of explanatory variables relate to body size or lung function: age, sex, height, weight, BMP (body mass as a percentage of the age-specific median), FEV1 (forced expiratory volume in 1 second), RV (residual volume), FRC (functional residual capacity), TLC (total lung capacity)
- For now, let’s consider age alone
- Quantify this relationship by postulating a model of the form

$$y = a + \beta x + e, \ e \sim N(0, \sigma^2)$$
Linear Regression Model

\[ y = \alpha + \beta x + e , \quad e \sim N(0, \sigma^2) \]

- \( y \): dependent/response/outcome variable
- \( x \): independent/explanatory/predictor variable
- \( e \): error term
- \( \alpha, \beta \): coefficients/regression coefficients/model parameters
  - \( \alpha \): intercept
  - \( \beta \): slope, describes the magnitude of association between \( X \) and \( Y \)
- For any given \( x \), \( y = \text{constant} + \text{normal random variable} \)
- The values \( x \) are considered to be measured without error

Assumptions

- For a specified value of \( x \), the distribution of the \( y \) values is normal with mean \( y = \alpha + \beta x \) and standard deviation \( \sigma \)
- For any specified value of \( x \), \( \sigma \) is constant
- This assumption of constant variability across all values of \( x \) is known as homoscedasticity

Residuals

- Use the data from the sample to estimate \( \alpha \) and \( \beta \), the coefficients of the regression line

\[ y = \alpha + \beta x + e , \quad e \sim N(0, \sigma^2) \]

- Call the estimators \( a \) and \( b \)
  \[ \hat{y} = a + bx \]

- The discrepancies between the observed and fitted values are called residuals
  \[ d = y - \hat{y} = y - a - bx \]

Fitting the Model

- One mathematical technique for fitting a straight line to a set of points is known as the method of least squares
- To apply this method, note that each data point \((x_i, y_i)\) lies some vertical distance \(d_i\) from an arbitrary line \((d_i\) is measured parallel to the vertical axis)
- Ideally, all residuals would be equal to 0
- Since this is impossible, we choose another criterion: we minimize the sum of squared

\[ S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \]
Fitting the Model

- The resulting line is the least squares line
- Using calculus, it can be shown that
  \[ b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]
  \[ a = \bar{y} - b\bar{x} \]

- Once \( a \) and \( b \) are known, we can substitute various values of \( x \) into the regression and compute \( y \).

Goodness of Fit

- After estimating the model parameters, we need to evaluate how well the model fits the data
- Three criteria:
  - Inference about beta
  - \( R^2 \)
  - Residual plots

- These concepts will hold for more complex cases, such as multiple regression, logistic regression, and Cox regression

Inference about \( \beta \)

- Because the parameter \( \beta \) describes the relationship between \( X \) and \( Y \), inference about \( \beta \) tells us about the strength of the linear relationship.
- After estimating the model parameters, we can do hypothesis testing and build confidence intervals for \( \beta \).
- The standard error of \( b \) in a simple linear regression is estimated as
  \[ \hat{s.e.}(b) = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \]

- To test the hypotheses \( H_0: \beta = 0 \), we calculate the test statistic
  \[ t = \frac{b}{\hat{s.e.}(b)} \]

- Under \( H_0 \), this has a t distribution with \( n-2 \) df
- If the true population slope is equal to 0, there is no linear relationship between \( x \) and \( y \); \( x \) is of no value in predicting \( y \)
- 100(1-\( \alpha \)) CI for \( \beta \):
  \[ b \pm t_{n-2, 1-\frac{\alpha}{2}} \hat{s.e.}(b) \]

- We can also carry out a similar procedure for \( \alpha \)
Example: Cystic Fibrosis Patients

```R
> install.packages("ISwR")
> library(ISwR)
> data(cystfibr)
> attach(cystfibr)
> my.model = lm(pemax~age)
> summary(my.model)

Call:
  lm(formula = pemax ~ age)

Residuals:
    Min     1Q Median     3Q    Max

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.408     16.657   3.026  0.00601 **
age            4.055      1.088   3.726  0.00111 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.97 on 23 degrees of freedom
Multiple R-squared: 0.3764,  Adjusted R-squared: 0.3492
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109
```

**Example: CF**

- We reject $H_0$ and conclude that the population slope is not equal to 0. $P$emax increases as age increases.
- Check:
  
  $50.408/16.657=3.026$
  
  $(1-\text{pt}(3.026,23))*2 = 0.00601$

- A 95% confidence interval for beta is
  
  $50.408 +/- 2.069*(16.657)$
  
  $(15.9, 84.9)$

Plotting the Regression Line

```R
plot(age,pemax,cex=2,pch=20)
names(my.model)
abline(my.model$coeff[1],my.model$coeff[2],lw=3)
```

R\(^2\)

- Another measure is $R^2$, sometimes called the coefficient of determination:

  $R^2 = \frac{\text{Reg SS}}{\text{Total SS}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$

- This is the proportion of variation explained by the model
- It is also the square of Pearson’s correlation coefficient

```R
> cor(pemax,age)^2
[1] 0.3763505
```
Residual Plots

• We’ve been assuming that the association between $X$ and $Y$ in the population is truly linear.
• Even if the association is nonlinear, these methods may still fit a line without detecting a problem. In this case, inferences from the model will not be correct.
• Previously we defined a point’s residual:
  \[ d_i = y_i - \hat{y}_i = y_i - a - bx_i \]
  
  • Because of the assumptions of linear regression, we expect all the residuals to be normally distributed with the same mean (0) and the same variance.
  • Violations of the linear regression assumptions can often be detected on a residual plot.

Residual Plots

• Plot the predicted $y$-values on the $x$-axis and the residuals on the $y$-axis.
• Are the residuals normally distributed with constant variance?

Residual Plots

• Another example:

Example: Cystic Fibrosis Patients

Does this model violate the assumption for constant variance?
Linear Regression

- Which models are ‘linear’?
  - \( y = a + bx \)
  - \( y = bx \)
  - \( y = a + b_1x_1 + b_2x_2 \)
  - \( y = a + b \log(x) \)
  - \( y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k \)
  - \( \log(y) = a + bx \)

- In fact, linear regression is not so restrictive

Summary: Simple Linear Regression

- Linear model
  \[
  y = \alpha + \beta x + e, \quad e \sim N(0, \sigma^2)
  \]
- Method of Least Squares
  \[
  S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - b x_i)^2
  \]
- Testing for significance of coefficients
  \[
  t = \frac{b}{s.e.(b)}
  \]

Multiple Linear Regression

- If knowing the value of a single explanatory variable improves our ability to predict a continuous response, we might suspect that information about additional variables could also be used to our advantage
- To investigate the more complicated relationship among a number of different variables, we use multiple linear regression analysis
  \[
  y = a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e
  \]
  \[
  e \sim N(0, \sigma^2)
  \]

- The intercept \( \alpha \) is the mean value of the response when all \( k \) explanatory variables are equal to 0
- The slope \( \beta_j \) is the change in \( y \) that corresponds to a one-unit increase in \( x_j \), given that all other explanatory variables remain constant
- The model is no longer a simple but something multidimensional
Least Squares

- Again, we define the “best” line by minimization of the sum of squared residuals

\[ S = \sum_{i=1}^{n} \alpha_i^2 = \sum_{i=1}^{n} (y_i - [a + b_1x_{i1} + \cdots + b_kx_{ik}])^2 \]

- Unfortunately, there is no simple formulas for the coefficients
- There is an elegant solution but this requires more mathematical notations
- Hypothesis testing for the coefficients is done the same way

A Single Predictor Model

```r
> my.model = lm(pemax ~ age)
> summary(my.model)

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.408     16.657   3.026  0.00601 **
age            4.055      1.088   3.726  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.97 on 23 degrees of freedom
Multiple R-squared: 0.3764, Adjusted R-squared: 0.3492
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109
```

- Age is a significant predictor of PEmax
- PEmax = 50.4 + 4.06 * age

Visualizing Data

- Before performing any analysis, it is good to view the data

```r
> plot(cystfibr)
pairs(cystfibr,gap=0)
```

- You can see the close relationship between age and height and weight

A Two-Predictor Model

```r
> my.model = lm(pemax ~ age + height)
> summary(my.model)

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.8600    68.2493   0.262    0.796
age           2.7178     2.9325   0.927    0.364
height        0.3397     0.6900   0.492    0.627
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.43 on 22 degrees of freedom
Multiple R-squared: 0.3831, Adjusted R-squared: 0.3271
F-statistic: 6.832 on 2 and 22 DF,  p-value: 0.00492
```

- PEmax = 17.9 + 2.72 * age + 0.40 * height
- How to interpret the coefficients?
- Which terms are significant here?
Inference for Coefficients

• We test the following hypothesis:
  $H_0: \beta_j = 0$ and all other $\beta$’s $\neq 0$
  $H_1: \beta_j \neq 0$ and all other $\beta$’s $\neq 0$

• The test statistic

\[ t = \frac{b_j}{\text{s.e.}(b_j)} \]

follows a $t$-distribution with $(n-k-1)$ df under the null

• $k$ is the number of explanatory variables
• $n$ is the number of data points

Adjusted $R^2$

- Age explained 37.6% of the variability in PEmax
- Age and height explained 38.3% of the variability in PEmax.
- The inclusion of an additional variable in a regression model can never cause $R^2$ to decrease
- To get around this problem, we use the **adjusted $R^2$** to penalize for the added complexity of the model
- Here, adjusted $R^2$ decreased. We conclude that this model is not an improvement over the age-only model
- Adjusted $R^2$ no longer has the same interpretation as $R^2$

F-test

- We perform inference about them together to determine whether the model demonstrates a statistically significant relationship between any predictor variable and the outcome variable.

\[
y = \alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k + e
\]

\[ e \sim N(0, \sigma^2) \]

- $H_0: \beta_1=\beta_2=\cdots = \beta_k = 0$ vs $H_1: \text{at least one } \beta_i \neq 0$
- We use the F-test to test this hypothesis

F-test

- Total sum of squares can be decomposed into **Regression** sum of squares (part explained by the model) and **Residual** sum of squares (remaining part)

\[
\text{Total SS} = \text{Reg SS} + \text{Res SS}
\]

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

- We normalize by the degrees of freedom to get regression and residual mean sum of squares. The ratio of these two values follows an $F$-distribution with $(k, n-k-1)$ df.

\[ F = \frac{\text{Reg MS}}{\text{Res MS}} \]
Categorical Variables

- Since “sex” is a categorical variable with two categories, we can represent a patient’s category by creating a variable that takes values “0” and “1” for men and women, respectively.
- This is not a continuous variable, but that is not a problem.

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e \]

- A binary variable created to represent a categorical variable with two categories is called a **dummy variable**, since its value (1 vs. 0) is arbitrarily chosen as numeric quantity.

```
my.model = lm(pemax ~ age + sex)
summary(my.model)
```

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 59.027     | 18.146  | 3.253    | 0.00365 ** |
| age        | 3.843      | 1.096   | 3.507    | 0.00199 ** |
| sex        | -12.632    | 10.944  | -1.154   | 0.26081    |

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.78 on 22 degrees of freedom
Multiple R-squared: 0.412, Adjusted R-squared: 0.3585

- PEmax = 59 + 3.84 * age – 12.6 * sex
- For men, PEmax = 59 + 3.84 * age
- For women, PEmax = 46.4 + 3.84 * age
- So \( b_2 \) is the difference between the predicted values of a man and a woman of the same age.

Categorical Variables

- What if there are multiple categories?
- Examples:
  - Geographic location: "Northeast", "South", "Midwest", etc.
  - Race: "White", "Black", "Asian", "American Indian".
- How about \( X = 0 \) “White”, \( X = 1 \) “Black”, \( X = 2 \) “Asian”, \( X = 3 \) “American Indian”?
- We need to create multiple dummy variables
  \( X_1 = \) “Black”, \( X_2 = \) “Asian”, \( X_3 = \) “American Indian”
- One category, by default, is the reference category
- So, for a multiple-category variable, the number of dummy variables needed is one fewer than the number of categories.

Interaction Terms

- In the regression models above, each explanatory variable is related to the outcome independently of all others.
- Sometimes the effect of an explanatory variable depends on the level of another explanatory variable.
- If this occurs, it is called an **interaction effect**
- If we believe interaction effects are present, we include them in the model.

\[ y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e \]
Interaction Terms

- For instance, we can see whether the effect of age is different for men and women
  
  \[ \hat{y} = a + b_1x_1 + b_2x_2 \]

  - men: \( \hat{y} = a + b_1x_1 \)
  - women: \( \hat{y} = (a + b_2) + b_1x_1 \)

A Full Model

- We can do inference about all the model parameters together (F-test), or we can do inference about them separately (t-tests).

  \[
  \text{my.model} = \text{lm(pemax ~ age+sex+height+weight+bmp+fev1+rv+frc+tlc)}
  \]

  \[
  \text{summary(my.model)}
  \]

  Coefficients:
  Estimate  Std. Error t value  Pr(>|t|)
  (Intercept)  176.0582   225.8912   0.779    0.448
  age          -2.5420     4.8017  -0.529    0.604
  sex          -3.7368    15.4598  -0.242    0.812
  height       -0.4463     0.9034  -0.494    0.628
  weight       -2.9928    2.0080   1.490    0.157
  bmp          -1.7449     1.1552  -1.510    0.152
  fev1          1.0807     1.0809   1.000    0.333
  rv            0.1970     0.1962   1.004    0.331
  frc           0.1970     0.4997   0.377    0.711
  tlc           0.1970     0.4997   0.377    0.711

  Residual standard error: 25.47 on 15 degrees of freedom
  Multiple R-squared: 0.6373, Adjusted R-squared: 0.4197
  F-statistic: 2.929 on 9 and 15 DF, p-value: 0.03195

CF Example

- Age was a statistically significant before
- Note that none of the variables are significant now!
- But the joint F-test is still significant; there must be some effect
- From the p-values in the full model, you cannot tell whether a variable would be significant in a reduced model
- A predictor may become non-significant when there is a highly correlated predictor

Model Selection

- How do we select the ‘best’ model?
- As a general rule, we prefer to include only those explanatory variables that help us to predict the response \( y \), the coefficients of which can be accurately estimated
- To study the full effect of each explanatory variable on the response, it would be necessary to perform a separate regression analysis for each possible combination of the variables
- While thorough, the all possible models approach is usually extremely time-consuming
- More frequently, we use a stepwise approach to choose a “best-fitting” model
Forward Selection

- Two commonly used procedures are **forward selection** and **backward elimination**
- Forward selection begins with no variables in the model and introduces variables one at a time
- The model is evaluated at each step
- For example, we might begin by including the single variable that yields the largest $R^2$
- We next add the variable that increases $R^2$ the most (the increase must be statistically significant)
- We continue this procedure until none of the remaining variables explains a significant amount of the additional variability in $y$

Backward Selection

- Backward elimination begins by including all explanatory variables in the model
- We drop the variable that contributes the least to the overall $R^2$
- The process continues until each remaining variable explains a significant portion of the variability in $y$
- Your software probably uses a more sophisticated penalty
- Example: AIC (Alkaike information criterion): $2k + n \times \log (RSS/n)$
- This find the model that best explains the data with a minimum of free parameters

CF Example

- Model from the original paper

```
> summary(my.model)

Coefficients:  Estimate  Std. Error t value Pr(>|t|)
(Intercept) 126.3336     34.7199   3.639 0.001536 **
weight       1.5365      0.3644   4.216 0.000387 ***
fev1         1.1086      0.5144   2.155 0.042893 *
bmp         -1.4654      0.5793  -2.530 0.019486 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.44 on 21 degrees of freedom
Multiple R-squared: 0.57,    Adjusted R-squared: 0.5086
F-statistic: 9.279 on 3 and 21 DF,  p-value: 0.0004180
```

- It is somewhat arbitrary that weight ended up in the model
- You may not get the same result, depending on your selection criteria
- $PE_{max}$ is probably connected to the patient's physical size (age, height, or weight).

Stepwise Model

- A stepwise procedure allows variables that have been dropped from the model to re-enter at a later time
- Usually the p-value criteria are relaxed in these procedures
- In a forward stepwise method, one may have
  - $p=0.1$ to include or remove a variable
  - $p=0.1$ to include and $p=0.2$ to remove a variable
- It is possible to end up with different final models, depending on which strategy is used
- The decision is usually made based on a combination of statistical and nonstatistical considerations
- Key aspects are simplicity of the model and whether the model can be easily interpreted