Stat 414 — Day 3
Inference for Linear Models

Last Time

- Least squares estimation vs. Maximum likelihood estimation
- Measures of model fit should adjust for model complexity (number of parameters, e.g., R^2_{adj} and AIC (-2 log-likelihood + 2p) or BIC

Example 1: Pace of Life and Heart Disease

On a recent international trip, we noticed that we were walking down the street much faster than other people. This reminded me of a study by Levine (1990) on "pace of life" in different countries. One way he measured pace of life was "average walking speed of randomly chosen pedestrians." He then explored a possible association with incidence of heart disease (age-adjusted death rate due to ischemic heart disease). We will look at data for 36 U.S. cities, for walking speeds over a distance of 60 feet (measured during business hours on a clear summer day along a main downtown street, no units)

Figure 6. Regional differences in the pace of life are apparent in a map showing the 36 cities included in the American study. A faster pace corresponds to larger dots; the 18 fastest cities are represented by red dots, whereas the 18 slowest cities are represented by blue dots. Within each region measurements were made in nine cities—three large, three of medium size and three small.

Pace of Life

```
            City Heart Walk Talk Bank Watch    Region
1      Boston,MA    24   28   24   31    30 Northeast
2     Buffalo,NY    29   23   23   30    33 Northeast
```

```
3       NewYork,NY    31   24   18   29    32 Northeast
4 SaltLakeCity,UT    26   28   23   28    23      West
5     Columbus,OH    26   22   30   27    23   Midwest
6    Worcester,MA    20   25   24   26    27 Northeast
iscamsummary(PaceData$Heart)
missing       n     Min        Q1  Median       Q3      Max     Mean       SD    skew
      0      36      11        16      19       24       31   19.806    5.214   0.163
missing       n     Min        Q1  Median       Q3      Max     Mean       SD    skew
      0      36      12    18.750      22   23.250       30   21.417    4.285  -0.297
```

Applet Demo

Follow this link for the Two Quantitative Variables applet (with two pace of life variables)

(a) Does the basic regression model appear appropriate?

relationship looks linear, don't see skewness or strange observations, and don't see fanning or megaphone so equal variance ok

(b) Check the Show Regression Line box and write down the fitted model. Do you think the observed relationship is *statistically significant*? How are you deciding?

opinons were a bit mixed, do you think the observed relationship could have happened 'by chance alone'? (number of observations n = 36 cities)

(c) Write the null and alternative hypotheses that we are interested in testing, first in words, then in symbols

Null hypothesis: No relationship between walking speed and heart disease in the population; alternative hypothesis: Is a relationship between walking speed and heart disease in the population. Ho: $\beta_{walk} = 0$ Ha: $\beta_{walk} \neq 0$

Statistical significance

Step 1: Create a population to sample from.
- Scroll down and check the **Design Population** box.
- Use pull-down menu to select **Observed x**
- Press the **Create Population** button

The applet will create a population of 20,000 cites that matches our sample's characteristics, with normally distributed residuals with equal variances, but assumes a population slope of zero. (View the Population regression line.)
- Check the **Show Sampling Options** box.
- Specify **1000** samples
- Specify a sample size of **36**
- Use the pull-down menu to select the **slope** as the statistic
- Press the **Draw Samples** button

(d) Describe the shape, center, and SD of this null distribution of slopes of walking speed

approximately normal with mean about zero and SD of slopes from sample to sample about 0.20.

(e) Also describe what you learn about the behavior of the different simulated regression lines. How would you describe the graph to someone who can't see it?

bow tie shape, see-saw, cat whiskers

(f) Check the Show Original Regression Line box. What does this tell you about the statistical significance of the observed line?

Seems a little unlikely to get a regression line as steep as what we observed when the null hypothesis is true

(g) Calculate the *standardized statistic*

$$t = (observed\ slope - expected\ slope\ under\ null\ hypothesis)/SE(slope)$$

t = (.4231 - 0)/.20 = 2.12

- Use the Statistic pull-down menu to select t-statistic
- Check the box to **Overlay *t*-distribution**

(h) Does this distribution appear reasonably modelled by the t distribution?

yes, when we switched to the t-statistics for the simulated distribution, it looked a lot like a t-distribution which we overlaid

(i) Enter the value of the t-statistic that you found above and use the applet to count the number of samples with a t-statistic at least as extreme as the observed. How do the "simulation-based" and "theory-based" results compare?

Count how many samples have a t-statistic 2.11 or larger or -2.11 or smaller, we got a p-value of about 0.034. The simulation results and the theory-based (using the t-distribution) match because we created a world whee the regression model assumption assumptions are met.

- Return to the original sample (uncheck the **Design Population** box)
- Check the **Regression table** box

(j) What is reported for the slope, SE slope, t-statistic, and p-value. Any surprises?

SE of slope = 0.196, t = 2.16, p-value (two-sided) = 0.377. No suprises in this output. We conclude that because p-value is small (below .05), we will reject the null hypothesis -> we have convincing evidence that there is a genuine linear association (beta is not zero) in the population of all U.S. citiies.

(k) Check the box for 95% confidence interval for Slope **AND provide a detailed interpretation of this interval.**

(.025, .821) We are 95% confident that each one unit increase in walk speed is associated with an average increase of .025 to .821 in heart disease in the population. (Sorry about being sloppy with units on this one.)

(l) What is wrong with the following interpretation: "95% of the time the slope will be between 0.0254 and 0.8207"?

Is only one population slope, but lots of potential intervals from different samples. We can't say 'there is a 95% chance or 95% probability that the population slope is between .025 and .821' so we say 'confident' instead.

(m) What value for the degrees of freedom is reported with the confidence interval? Where does that number come from?

df = 34 = n - p = 36 -2 (intercept and slope)

- Check the **ANOVA table** box.

Definition

The *F*-test can be interpreted as a "drop in sums of squares" test.

$$F = MSModel/MSError = (SSE_{full} - SSE_{reduced})/(change\ in\ df)/MSE_{full}$$

When there is only one regression slope in the model, this is equivalent to the two-sided t-test for that coefficient, but generalizes very nicely to comparing any two models, including your model to the model with only the intercept ("overall *F*-test"). The models should be "nested" in the sense that you get from one model to the other by constraining regression coefficients (e.g., setting to zero) in the null hypothesis.

Fit a model with all four quantitative explanatory variables:

- Walk
- Bank: Average time a sample of bank clerks takes to make change for two $20 bills or to give $20 bills for change
- Talk: Recorded responses of postal clerks explaining the difference between regular, certified, and insured mail and divided the total number of syllables by the time of their response
- Watch: Proportion of men and women in downtown areas during business hours who were wearing a wrist watch

Carry out a *partial F-test* to compare the models. First get both ANOVA tables.

```
model1 <-lm(Heart ~ Walk, data = PaceData); anova(model1)
Analysis of Variance Table

Response: Heart
          Df Sum Sq Mean Sq F value Pr(>F)
Walk       1    115   115.0    4.68  0.038 *
Residuals 34    837    24.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2 <- lm(Heart ~ . - Region - City, data = PaceData); summary(model2)

Call:
lm(formula = Heart ~ . - Region - City, data = PaceData)

Residuals:
   Min     1Q Median     3Q    Max 
-8.066 -2.613  0.951  2.561  9.679 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   -5.955      7.309   -0.81    0.421  
Walk           0.394      0.192    2.05    0.048 *
Talk          -0.041      0.220   -0.19    0.853  
Bank           0.384      0.187    2.06    0.048 *
Watch          0.349      0.160    2.18    0.037 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.54 on 31 degrees of freedom
Multiple R-squared:  0.327, Adjusted R-squared:  0.24 
F-statistic: 3.77 on 4 and 31 DF,  p-value: 0.0131
anova(model2)
Analysis of Variance Table

Response: Heart
          Df Sum Sq Mean Sq F value Pr(>F)  
Walk       1    115   115.0    5.57  0.025 *
Talk       1      0     0.2    0.01  0.921  
Bank       1     98    97.6    4.72  0.037 *
Watch      1     98    98.5    4.77  0.037 *
Residuals 31    640    20.7                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#The period is convenient when you want to include all other variables, but here di
dn't want Region or "city" as variables in the model
```

(n) State the hypotheses you are testing, report the test statistic, degrees of freedom, p-value, and conclusion in context.

H_0: $\beta_{talk} = \beta_{bank} = \beta_{watch} = 0$ vs. Ha: at least one $\beta \neq 0$. The df numerator will be 3 (the number of parameters we are setitng to zero) and the df denomiator will be 31 (the degrees of freedom in the full model as we think that estimate or σ is more 'trustworthy'.

$$F = (836.60 - 640.31)/3/(20.655) \approx 3.17$$

with df = 3 and 31.

pf(3.17, 3, 31, lower.tail=FALSE) = 0.038

So we reject the null hypothesis and conclude that at least one of these variables is a useful addition to the model on top of the walk variable.

Example 2: Pace of Life cont - Maximum Likelihood

Let's redo the analysis using maximum likelihood estimation.

```
library(nlme)
model1ML <- gls(Heart ~ Walk, method = "ML", data = PaceData)
summary(model1ML)
Generalized least squares fit by maximum likelihood
  Model: Heart ~ Walk
  Data: PaceData
    AIC   BIC logLik
  221.4 226.2 -107.7

Coefficients:
             Value Std.Error t-value p-value
(Intercept) 10.745     4.271   2.516  0.0168
Walk         0.423     0.196   2.162  0.0377

 Correlation:
     (Intr)
Walk -0.981

Standardized residuals:
     Min       Q1      Med       Q3      Max
-1.87781 -0.68106 -0.09861  0.76904  2.14328

Residual standard error: 4.821
Degrees of freedom: 36 total; 34 residual
```

(a) What is "df model" here? What is the residual standard error? How does it compare to the least squares estimate"?

df for the model in this case is 3 (rather than 2): Intercept, Slope, and now $\sigma$$

Definition

Maximum Likelihood Estimation estimates σ and the β coefficients simultaneously. Then $\hat{\sigma}^2$ is found by $SSError/n$. This is a *biased* estimator of σ^2. *Restricted (or Residual) Maximum Likelihood (REML)* uses a different likelihood function that estimates the regression coefficients more "separately" from σ^2.

Fit the restricted maximum likelihood model

```
model1REML <- gls(Heart ~ Walk, method = "REML", data = PaceData)
summary(model1REML)
Generalized least squares fit by REML
  Model: Heart ~ Walk
  Data: PaceData
    AIC BIC logLik
```

```
  221.4 226 -107.7

Coefficients:
             Value Std.Error t-value p-value
(Intercept) 10.745     4.271   2.516  0.0168
Walk         0.423     0.196   2.162  0.0377

 Correlation:
     (Intr)
Walk -0.981

Standardized residuals:
     Min        Q1       Med        Q3       Max
-1.82490 -0.66187 -0.09584  0.74738  2.08289

Residual standard error: 4.96
Degrees of freedom: 36 total; 34 residual
```

(b) How has the residual standard error changed? Does it look familiar? How has the standard error of the slope coefficient changed?

REML estimates σ^2 by taking into account the number of parameters and so in fact ends up matching the OLS estimate of σ^2

(c) What are the (log) likelihood and AIC values for this model?

log like = -107.72, AIC = 221.44

Definition

A *Likelihood Ratio Test (LRT)* works just like a partial F-test. The difference in the log likelihood values between two models follows a *chi-square distribution* with df = difference in number of parameters between the full and reduced model.

$$2ln\left(L_{full}/L_{reduced}\right) = -2\ log\ likelihood\ reduced\ -\ (-2\ log\ likelihood\ full) = 2(L_1 - L_0)$$

Note: the expected value of a chi-square distribution is the degrees of freedom.

Intuitively, when the likelihood for the larger model is much greater than it is for the reduced model, we have evidence that the larger model is more closely aligned with the observed data, but we also need to take into account the additional parameters that had to be estimated in the larger model.

Often times, we won't care too much whether we are using ML or REML, but use the same method when comparing models. One rule is to use REML when you are more focused on the variances. Here, we aren't changing any assumptions about variances so we will use ML.

```
model2REML <- gls(Heart ~ Walk + Bank + Talk + Watch, method = "REML", data = PaceD
ata)
logLik(model2REML)
'log Lik.' -105.6 (df=6)
model2ML <- gls(Heart ~ Walk + Bank + Talk + Watch, method = "ML", data = PaceData)
logLik(model2ML)
```

```
'log Lik.' -102.9 (df=6)
logLik(model1ML)
'log Lik.' -107.7 (df=3)
```

(d) Calculate the test statistic. What are its degrees of freedom?

2*(107.71 - 102.89) = 9.84 compare to a chi-square distribution with df = 6 - 3 = 3

(e) Calculate the p-value from the chi-square distribution

```
teststat <- 9.54
pchisq(teststat, df= 3, lower.tail = FALSE)
[1] 0.02291
```

p-value from likelihood ratio test (LRT) is .0229 (partial f-test gave us a p-value of .037)

(f) How does the strength of evidence differ between OLS and ML estimation?

stronger with ML because smaller p-value

The parallel confidence intervals are often called “profile intervals” and they include the parameter values where the likelihood ratio test would fail to reject.

```
confint(model2ML)
                2.5 % 97.5 %
(Intercept) -20.27941 8.3704
Walk          0.01814 0.7700
Bank          0.01847 0.7502
Talk         -0.47129 0.3894
Watch         0.03581 0.6632
```

ok, here are the short cuts for both tests

```
anova(model1ML, model2ML)
         Model df   AIC   BIC logLik   Test L.Ratio p-value
model1ML     1  3 221.4 226.2 -107.7
model2ML     2  6 217.8 227.3 -102.9 1 vs 2   9.626   0.022
anova(model1, model2)
Analysis of Variance Table

Model 1: Heart ~ Walk
Model 2: Heart ~ (City + Walk + Talk + Bank + Watch + Region) - Region -
    City
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     34 837
2     31 640  3       196 3.17  0.038 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```