Fall, 2023 *Thursday, Oct. 12*

Stat 414 - Day 11
Maximum Likelihood Estimation

Last Time

- vcov gives us the variance-covariance matrix of the estimated coefficients
  - the square roots of the diagonal values $= SE(\hat{\beta}_j)$
  - the size of these SEs depends on σ, n, and $SD(X)$'s
  - if our estimate of σ is off, so will be these standard errors …
- If we don't have good candidates for weights (or not appropriate, including correlated errors), an alternative approach is stick with OLS but use HC "sandwich" standard errors
  - These use the squared residuals to tells us about the error variance structure

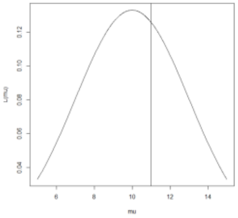
Example: Airfares from San Luis Obispo to a "random" sample of 12 major U.S. cities as found March 31, 2017 on Travelocity.com for travel on May 8-May 12, 2017 are found in airfare.txt.

(a) Fit the "null model" with no predictors

```
model0 = lm(price ~ 1, data = airfare)
summary(model0); sum(residuals(model0)^2); sigma(glsmodel)
```

(b) What is the estimated intercept? How is it found? How is σ estimated?

So far, we have estimated our parameters using "least squares estimation" and focused on minimizing the sum of squared errors as judged by the residuals. There are alternatives to least squares estimation. One such method is *maximum likelihood estimation*. The *likelihood* is the probability density function (pdf) of a random variable, but viewed as a function of the parameter(s). We want to choose parameter values that maximize the likelihood of observing the data we have. We are trying to match what we observe with what we expect to see.



First consider just one variable (e.g., airfare prices). Suppose our data follow a normal distribution $L(\mu, \sigma; y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}$.

Suppose we didn't know that μ = 10, but we observed x = 11. We could think about trying different values of μ (moving this curve left and right) and stopping when the peak of the curve is highest at our observed value. In this toy example, we would report $\hat{\mu}_{MLE}$ = 11.

With independent observations, the joint likelihood will be the product $L(\mu, \sigma; y) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{\left(\frac{-\Sigma(y_i-\mu)^2}{2\sigma^2}\right)}$. Our goal is to *maximize* this function, for which we would use differentiation, reporting the values of $\hat{\mu}$ and $\hat{\sigma}$. For most likelihoods, it is much simpler to work with the *log likelihood* and the MLEs estimates are the same.

(c) Write out the expression for the log likelihood. How would we find the values of μ and that maximize this function?

Fitting a linear model with maximum likelihood

```
#install.packages("nlme")
library(nlme)
model0ML = gls(price ~ 1, method = "ML", data = airfare )
sum(residuals(model0ML)^2);
sigma(model0ML)
```

(d) Is the equation the same? Is the residual standard error the same? How is it calculated now?

How do assess the fit of the model?

Rather than SSError, we will focus on the value of the “achieved” likelihood.

(e) Substitute our estimates back in to see how large we were able to make the likelihood function.

(f) Confirm this value in R with logLik(model0). What are the degrees of freedom?

Now consider the model that also includes the distances.

```
model1ML = gls(price ~ distance, method="ML", data = airfare)
summary(model1ML)
```

(g) Use the above output to find the value of the likelihood function and then confirm with R. Does this model do “better” than the previous model?

More measures of model fit (aka information criteria)

Want to maximize the log likelihood but can also penalize you for the number of parameters

- Want large $R^2_{\text{adjusted}} = R^2 - p/(n - p)(1 - R^2) = 1 - \text{SSE}/(n - p)/(\text{SST}/(n - 1))$

Here: p = number of parameters (intercept, slopes, σ)

- Want (2)(log)likelihood values to be large (or small *deviance* = -2 log-likelihood)
- Want small BIC = -2 x log-likelihood + p x ln(n)
- Want small AIC = -2 x log-likelihood + 2p

(h) Verify the AIC value in R using AIC(model0) and AIC(model1). Which model has the better value?

```
AIC(model0); AIC(model1)
```

(i) **Inference***: What was our previous method (in general) for comparing these two models/ deciding whether the distance variable should be included in the model?*

(j) Carry out the ‘likelihood ratio test.’

```
anova(model0ML, model1ML)
```

Definition*:* The **likelihood ratio test** compares nested models by using the statistic $-2(L_0 - L_1)$ which asymptotically follows a chi-square distribution with df = difference in number of parameters in the two models.

To find the p-value for the likelihood ratio test, adding distance to the model:

```r
1 - pchisq(2*(75.144 - 69.026), 1)
```