Lecture 10: F-Tests, R^2, and Other Distractions

36-401, Section B, Fall 2015

1 October 2015

Contents

1 The F Test

The F distribution with a, b degrees of freedom is *defined* to be the distribution of the ratio

$$\frac{\chi^2_a/a}{\chi^2_b/b}$$

when χ^2_a and χ^2_b are independent.

Since χ^2 distributions arise from sums of Gaussians, F-distributed random variables tend to arise when we are dealing with ratios of sums of Gaussians. The outstanding examples of this are ratios of variances.

1.1 F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$

Let's consider testing the null hypothesis $\beta_1 = 0$ against the alternative $\beta_1 \neq 0$, in the context of the Gaussian-noise simple linear regression model. That is, we won't question, in our mathematics, whether or not the assumptions of that model hold, we'll presume that they all do, and just ask how we can tell whether $\beta_1 = 0$.

We have said, *ad nauseam*, that under the unrestricted model,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

with

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

This is true no matter what β_1 is, so, in particular, it continues to hold when $\beta_1 = 0$ but we estimate the general model anyway.

The null model is that

$$Y = \beta_0 + \epsilon$$

with $\epsilon \sim N(0, \sigma^2)$, independent of X and independently across measurements. It's an exercise from 36-226 to show (really, remind!) yourself that, in the null model

$$\hat{\beta}_0 = \overline{y} \sim N(\beta_0, \sigma^2/n)$$

It is another exercise to show

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = s^2_Y$$

and

$$\frac{ns^2_Y}{\sigma^2} \sim \chi^2_{n-1}$$

However, s^2_Y is not independent of $\hat{\sigma}^2$. What *is* statistically independent of $\hat{\sigma}^2$ is the difference

$$s^2_Y - \hat{\sigma}^2$$

and

$$\frac{n(s^2_Y - \hat{\sigma}^2)}{\sigma^2} \sim \chi^2_1$$

I will not pretend to give a proper demonstration of this. Rather, to make it plausible, I'll note that $s^2_Y - \hat{\sigma}^2$ is the extra mean squared error which comes from estimating only one coefficient rather than two, that each coefficient kills one degree of freedom in the data, and the total squared error associated with one degree of freedom, over the entire data set, should be about $\sigma^2\chi^2_1$.

Taking the previous paragraph on trust, then, let's look at a ratio of variances:

$$\frac{s_Y^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = \frac{n(s_Y^2 - \hat{\sigma}^2)}{n\hat{\sigma}^2} \quad (1)$$

$$= \frac{\frac{n(s_Y^2 - \hat{\sigma}^2)}{\sigma^2}}{\frac{n\hat{\sigma}^2}{\sigma^2}} \quad (2)$$

$$= \frac{\chi^2_1}{\chi^2_{n-2}} \quad (3)$$

To get our F distribution, then, we need to use as our test statistic

$$\frac{s_Y^2 - \hat{\sigma}^2}{\hat{\sigma}^2}\frac{n-2}{1} = \left(\frac{s_Y^2}{\hat{\sigma}^2} - 1\right)(n-2)$$

which will have an $F_{1,n-2}$ distribution under the null hypothesis that $\beta_1 = 0$.

Note that the only random, data-dependent part of this is the ratio of $s_Y^2/\hat{\sigma}^2$. We reject the null $\beta_1 = 0$ when this is too large, compared to what's expected under the $F_{1,n-2}$ distribution. Again, this is the distribution of the test statistic *under the null* $\beta_1 = 0$. The variance ratio will tend to be larger under the alternative, with its expected size growing with $|\beta_1|$.

Running this F test in R The easiest way to run the F test for the regression slope on a linear model in R is to invoke the `anova` function, like so:

```
anova(lm(y~x))
```

This will give you an analysis-of-variance table for the model. The actual object the function returns is an `anova` object, which is a special type of data frame. The columns record, respectively, degrees of freedom, sums of squares, mean squares, the actual F statistic, and the p value of the F statistic. What we'll care about will be the first row of this table, which will give us the test information for the slope on X.

To illustrate more concretely, let's revisit our late friends in Chicago:

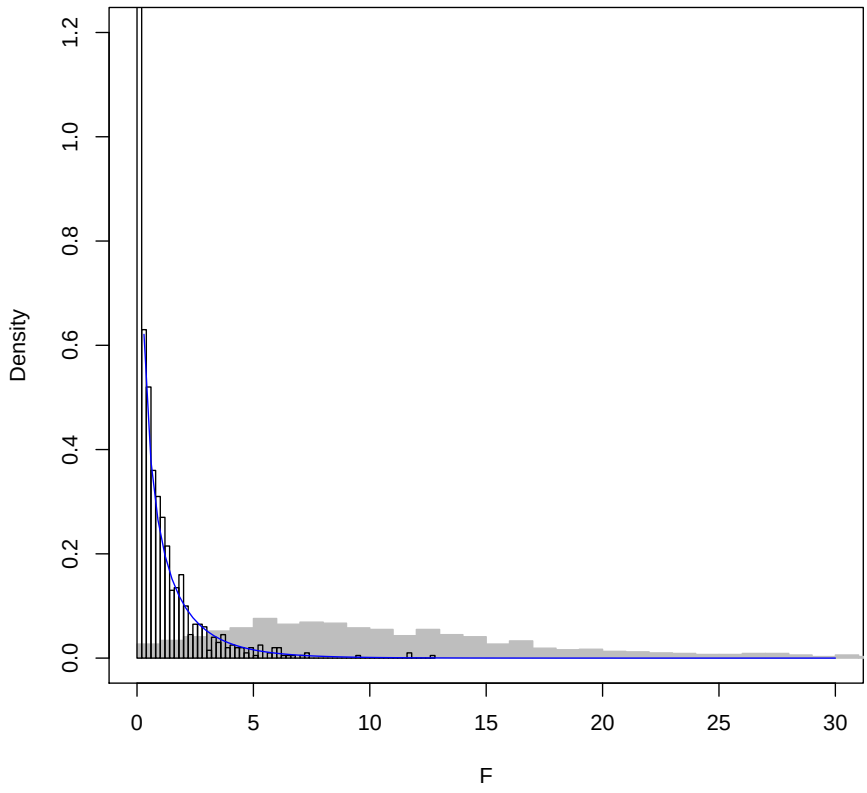
```
library(gamair); data(chicago)
death.temp.lm <- lm(death ~ tmpd, data=chicago)
anova(death.temp.lm)
```

```
## Analysis of Variance Table
##
## Response: death
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## tmpd          1  162473  162473  803.07 < 2.2e-16 ***
## Residuals 5112 1034236     202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As with `summary` on `lm`, the stars are usually a distraction; see Lecture 8 for how to turn them off.

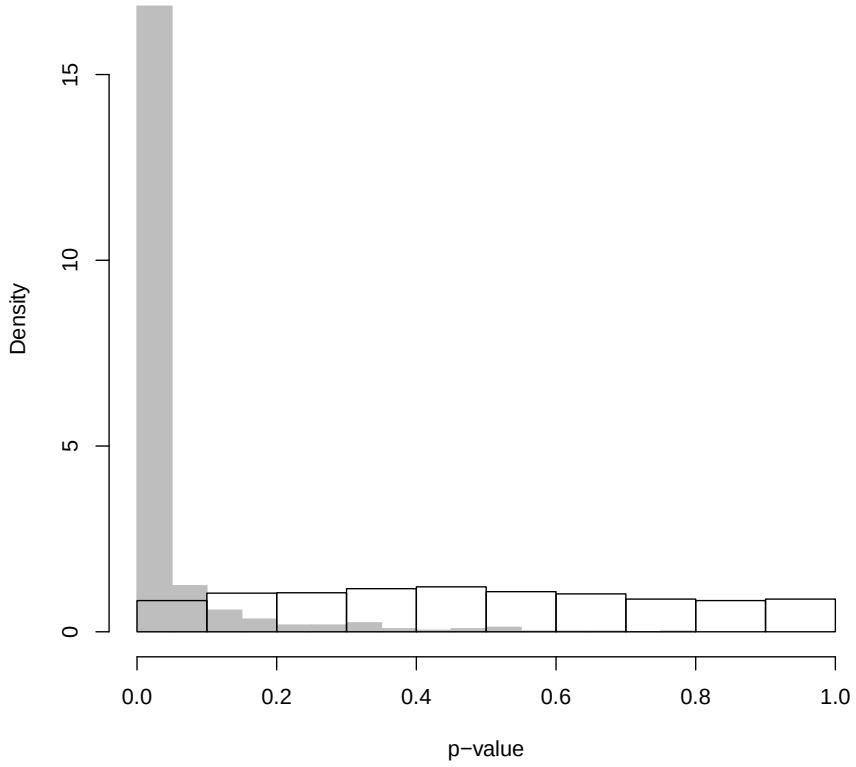
```
# Simulate a Gaussian-noise simple linear regression model
# Inputs: x sequence; intercept; slope; noise variance; switch for whether to
  # return the simulated values, or the ratio of s^2_Y/\hat{\sigma}^2
# Output: data frame or coefficient vector
sim.gnslrm <- function(x, intercept, slope, sigma.sq, var.ratio=TRUE) {
  n <- length(x)
  y <- intercept + slope*x + rnorm(n,mean=0,sd=sqrt(sigma.sq))
  if (var.ratio) {
    mdl <- lm(y~x)
    hat.sigma.sq <- mean(residuals(mdl)^2)
    # R uses the n-1 denominator in var(), but we need the MLE
    s.sq.y <- var(y)*(n-1)/n
    return(s.sq.y/hat.sigma.sq)
  } else {
    return(data.frame(x=x, y=y))
  }
}

# Parameters
beta.0 <- 5
beta.1 <- 0  # We are simulating under the null!
sigma.sq <- 0.1
x.seq <- seq(from=-5, to=5, length.out=42)
```

FIGURE 1: *Code setting up a simulation of a Gaussian-noise simple linear regression model, returning either the actual simulated data frame, or just the variance ratio* $s_Y^2/\hat{\sigma}^2$.




```
# Run a bunch of simulations under the null and get all the F statistics
  # Actual F statistic is in the 4th column of the output of anova()
f.stats <- replicate(1000, anova(lm(y~x, data= sim.gnslrm(x.seq, beta.0, beta.1,
                                                  sigma.sq, FALSE)))[1,4])
# Store histogram of the F statistics, but hold off on plotting it
null.hist <- hist(f.stats, breaks=50, plot=FALSE)
# Run a bunch of simulations under the alternative and get all the F statistics
alt.f <- replicate(1000, anova(lm(y~x, data=sim.gnslrm(x.seq, beta.0, -0.05,
                                              sigma.sq, FALSE)))[1,4])
# Store a histogram of this, but again hold off on plotting
alt.hist <- hist(alt.f, breaks=50, plot=FALSE)
# Create an empty plot
plot(0, xlim=c(0,30), ylim=c(0,1.2), xlab="F", ylab="Density", type="n")
# Add the histogram of F under the alternative, then under the null
plot(alt.hist, freq=FALSE, add=TRUE, col="grey", border="grey")
plot(null.hist, freq=FALSE, add=TRUE)
# Finally, the theoretical F distribution
curve(df(x,1,length(x.seq)-2), add=TRUE, col="blue")
```


FIGURE 2: *Comparing the actual distribution of F statistics when we simulate under the null model (black histogram) to the theoretical $F_{1,n-2}$ distribution (blue curve), and to the distribution under the alternative $\beta_1 = -0.05$.*




```
# Take the simulated F statistics and convert to p-values
p.vals <- pf(f.stats, 1, length(x.seq)-2, lower.tail=FALSE)
alt.p <- pf(alt.f, 1, length(x.seq)-2, lower.tail=FALSE)
hist(alt.p, col="grey", freq=FALSE, xlab="p-value", main="", border="grey",
     xlim=c(0,1))
plot(hist(p.vals, plot=FALSE), add=TRUE, freq=FALSE)
```

FIGURE 3: *Distribution of p-values from repeated simulations, under the null hypothesis (black) and the alternative (grey). Notice how the p-values under the null are uniformly distributed, while under the alternative they are bunched up towards small values at the left.*

Assumptions In deriving the F distribution, it is absolutely vital that all of the assumptions of the Gaussian-noise simple linear regression model hold: the true model must be linear, the noise around it must be Gaussian, the noise variance must be constant, the noise must be independent of X and independent across measurements. The *only* hypothesis being tested is whether, maintaining all these assumptions, we must reject the flat model $\beta_1 = 0$ in favor of a line at an angle. In particular, the test never doubts that the right model is a straight line.

The "general linear test" As a preview of coming attractions, we can look at what happens when we compare a linear, Gaussian-noise model with p parameters to a restricted Gaussian-noise linear model model with only q free parameters. Each model gives us an estimate of the noise variance, say $\hat{\sigma}^2_A$ and $\hat{\sigma}^2_0$ (respectively); these are just the mean squared residuals in each model. It will not surprise you to learn that, under the null that the smaller, restricted model is true

$$\frac{n(\hat{\sigma}^2_0 - \hat{\sigma}^2_A)}{\sigma^2} \sim \chi^2_{p-q}$$

while

$$\frac{n\hat{\sigma}^2_A}{\sigma^2} \sim \chi^2_{n-p}$$

The F statistic for testing the restriction of the full model to the sub-model is therefore

$$\frac{\hat{\sigma}^2_0 - \hat{\sigma}^2_A}{\hat{\sigma}^2_A} \frac{n-p}{p-q}$$

and it has an $F_{p-q,n-p}$ distribution.

ANOVA You will notice that I made no use of the ponderous machinery of analysis of variance which the textbook wheels out in connection with the F test. Despite (or because) of all of its size and complexity, this is really just a historical relic. In serious applied work from the modern (say, post-1985) era, I have never seen any study where filling out an ANOVA table for a regression, etc., was at all important.

There is more to be said for analysis of variance where the observations are divided into discrete, categorical groups, and one wants to know about differences between groups vs. variation within a group. In a few weeks, when we see how to handle categorical predictor variables, it will turn out that this useful form of ANOVA is actually a special case of linear regression.

1.2 The Likelihood Ratio Test

The F test is a special case of a much more general procedure, the **likelihood ratio test**, which works as follows. We start with a general model, where the parameter is a vector $\theta = (\theta_1, \theta_2, \ldots \theta_p)$. We contemplate a restriction, where

$\theta = (\theta_1, \theta_2, \ldots \theta_q, 0, \ldots 0)$, $q < p$. (See below on other possible restrictions.) The restricted sub-model is the null hypothesis, and the full model is the alternative.

Both the restricted model and the full model have maximum likelihood estimators; call these $\hat{\theta}$ and $\hat{\Theta}$, respectively. Let's write L for the log-likelihood function, so $L(\hat{\theta})$ is the maximized log-likelihood under the restricted null model, and $L(\hat{\Theta})$ is the maximized log-likelihood under the unrestricted, alternative, full model. Then

$$\Lambda \equiv L(\hat{\Theta}) - L(\hat{\theta})$$

is the log of the **likelihood ratio** between the models (because $\log a/b = \log a - \log b$). Λ is the test statistic in the likelihood ratio test[1].

Under some "regularity" conditions, which I'll sketch in a moment, there is a simple asymptotic distribution for Λ *under the null hypothesis*. As $n \to \infty$

$$2\Lambda \sim \chi^2_{p-q}$$

Let me first try to give a little intuition, then hand-wave at the math, and then work things through for test $\beta_1 = 0$ vs. $\beta_1 \neq 0$.

The null model is, as I said, a restriction of the alternative model. Any possible parameter value for the null model is also allowed for the alternative. This means the parameter space for the null, say ω, is a strict subset of that for the alternative, $\omega \subset \Omega$. The maximum of the likelihood over the larger space *must* be at least as high as the maximum over the smaller space:

$$L(\hat{\Theta}) = \max_{\theta \in \Omega} L(\theta) \geq \max_{\theta \in \omega} L(\theta) = L(\hat{\theta})$$

Thus, $\Lambda \geq 0$. What's more surprising is that its distribution doesn't change with n (asymptotically), and that depends on the difference in the number of free parameters. Because the MLE is consistent, under the null the estimates of $\theta_{q+1}, \theta_{q+2} \ldots \theta_p$ in $\hat{\Theta}$ all converge to zero, because those parameters *are* zero under the null. In fact, they get closer and closer to zero, but end up making larger and larger contributions to L, because L grows with n. The two effects cancel out, and each free parameter ends up contributing one χ^2_1 term.

Why χ^2? Well, for large n, $\hat{\theta}$ and $\hat{\Theta}$ both have Gaussian distributions around the true θ, and the contributions to the log-likelihood end up depending on the squares of parameter estimates. Since the square of a Gaussian is proportional to a χ^2, it's not surprising that we get something χ^2-ish, though it is nice how everything cancels out. I defer a fuller explanation to the option §5.

Sketch of the regularity conditions where the likelihood-ratio test has a χ^2 null First, the MLE must be consistent for both models, and must have a Gaussian distribution around the true parameter (for large n). Second, the restricted model has to "lie in the interior" of the unrestricted, alternative model, and not on the boundary. That is, it must make sense in the alternative model for all the zeroed-out parameters to be either positive or negative. (This

[1] Some people, being a bit pedantic, call it the log-likelihood-ratio test.

would be violated, for instance, if one of the parameters set to zero by the null were a variance.) And that's *mostly* it. Again, see §5 for more mathematical details.

Testing $\beta_1 = 0$ What's the log-likelihood at the MLE of the simple linear model? Dredging up the log-likelihood function from Lecture 6,

$$L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma^2}}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \quad (4)$$

But

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Substituting into Eq. 4,

$$L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \hat{\sigma}^2 - \frac{n\hat{\sigma}^2}{2\hat{\sigma^2}} = -\frac{n}{2}(1 + \log 2\pi) - \frac{n}{2}\log \hat{\sigma}^2$$

So we get a constant which doesn't depend on the parameters at all, and then something proportional to $\log \hat{\sigma}^2$.

The intercept-only model works similarly, only *its* estimate of the intercept is $\overline{y}$, and its noise variance, $\hat{\sigma}_0^2$, is just the sample variance of the y_i:

$$L(\overline{y}, 0, s_Y^2) = -\frac{n}{2}(1 + \log 2\pi) - \frac{n}{2}\log s_Y^2$$

Putting these together,

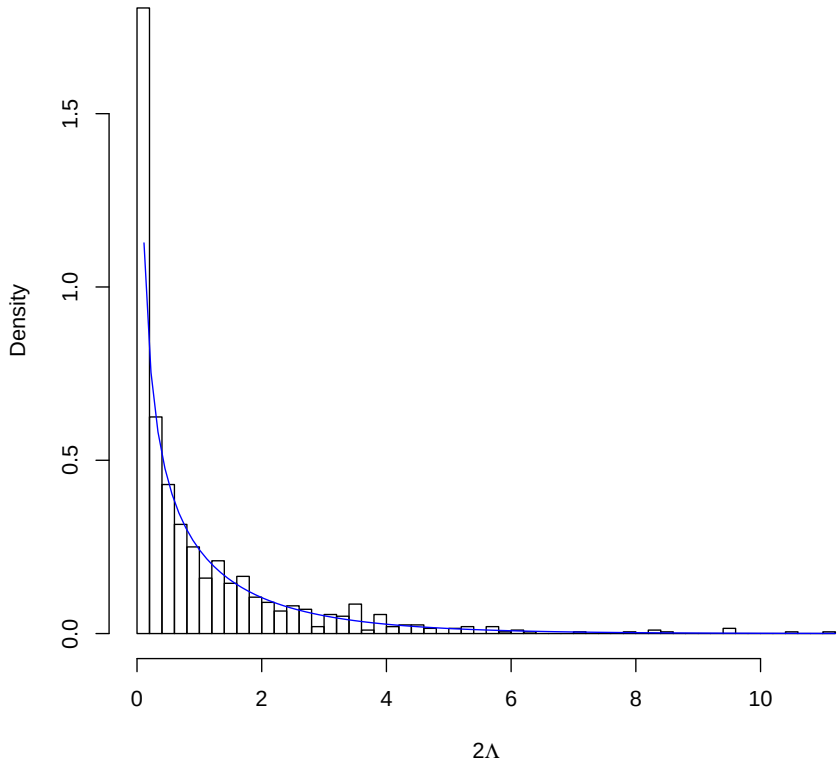
$$L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) - L(\overline{y}, 0, s_Y^2) = \frac{n}{2}\log \frac{s_Y^2}{\hat{\sigma}^2}$$

Thus, under the null hypothesis,

$$\frac{n}{2}\log \frac{s_Y^2}{\hat{\sigma}^2} \sim \chi_1^2$$

Figure 4 shows a simulation confirming this bit of theory.

Connection to F tests The ratio $s_Y^2/\hat{\sigma}^2$ is, of course, the F-statistic, up to constants not depending on the data. Since, for this problem, the likelihood ratio test and the F test use equivalent test statistics, if we fix the same size or level α for the two tests, they will have exactly the same power. In fact, even for more complicated linear models — the "general linear tests" of the textbook — the F test is always equivalent to a likelihood ratio test, at least when the presumptions of the former are met. The likelihood ratio test, however, applies to problems which do not involve Gaussian-noise linear models, while the F test is basically only good for them. If you can only remember *one* of the two tests, remember the likelihood ratio test.




```
# Simulate from the model 1000 times, capturing the variance ratios
var.ratios <- replicate(1000, sim.gnslrm(x.seq,beta.0,beta.1,sigma.sq))
# Convert variance ratios into log likelihood-ratios
LLRs <- (length(x.seq)/2)*log(var.ratios)
# Create a histogram of 2*LLR
hist(2*LLRs, breaks=50, freq=FALSE, xlab=expression(2*Lambda), main="")
# Add the theoretical chi^2_1 distribution
curve(dchisq(x,df=1), col="blue", add=TRUE)
```

FIGURE 4: *Comparison of log-likelihood ratios (black histogram) with theoretical χ^2_1 distribution (blue). Note we are simulating under the null hypothesis $\beta_1 = 0$. Can you add a histogram of the distribution under the alternative, and make histograms of p-values, as in Figures 2 and 3?*

Other constraints Setting $p - q$ parameters to zero is really a special case of imposing $p - q$ linearly independent constraints on the p parameters. For instance, requiring $\theta_2 = \theta_1$ while $\theta_3 = -2\theta_1$ is just as much a two-parameter restriction as fixing $\theta_2 = \theta_3 = 0$. This is because we could transform to a new set of parameters, say $\psi_1 = \theta_1$, $\psi_2 = \theta_2 - \theta_1$, $\psi_3 = \theta_3 + 2\theta_1$, where the restrictions *are* $\psi_2 = \psi_3 = 0$, and we can transform back to the θ parameters without loss of information. So the theory of the likelihood ratio test applies whenever we have linearly independent constraints.

More generally, that theory applies under the following (admittedly rather complicated) conditions:

- Under the null model, θ must obey equations $f_1(\theta) = 0$, $f_2(\theta) = 0$, ... $f_{p-q}(\theta) = 0$.
- Any θ which obeys those equations is in the null model.
- There is an *invertible* function g where, writing $\psi = g(\theta)$, in the null model, ψ always has $\psi_{q+1}, \ldots \psi_p = 0$, and under the alternative, ψ is unrestricted.

Basically, we need to be able to come up with a change of coordinates where the restrictions amount to fixing some coordinates to zero, but leaving the others alone.

2 What the F Test Really Tests

The textbook (§2.7–2.8) goes into great detail about an F test for whether the simple linear regression model "explains" (really, predicts) a "significant" amount of the variance in the response. What this really does is compare two versions of the simple linear regression model. The null hypothesis is that all of the assumptions of that model hold, *and* the slope, β_1, is exactly 0. (This is sometimes called the "intercept-only" model, for obvious reasons.) The alternative is that all of the simple linear regression assumptions hold[2], with $\beta_1 \neq 0$. The alternative, non-zero-slope model will always fit the data better than the null, intercept-only model (why?); the F test asks if the improvement in fit is larger than we'd expect under the null[3].

There are situations where it is useful to know about this precise quantity, and so run an F test on the regression. It is hardly ever, however, a good way to check whether the simple linear regression model is correctly specified, because neither retaining nor rejecting the null gives us information about what we really want to know.

[2]To get an exact F distribution for the test statistic, we also need the Gaussian-noise assumptions, but under the weaker assumptions of uncorrelated noise, we'll often approach an F distribution asymptotically.

[3]This is also what the likelihood ratio test of §1.2 is asking, just with a different notion of measuring fit to the data (likelihood vs. squared error). Everything I'm about to say about F tests applies, suitably modified, to likelihood ratio tests.

Suppose first that we retain the null hypothesis, i.e., we do not find any significant share of variance associated with the regression. This could be because (i) there is no such variance — the intercept-only model is right; (ii) there is some variance, but we were unlucky; (iii) the test doesn't have enough power to detect departures from the null. To expand on that last point, the power to detect a non-zero slope is going to increase with the sample size n, decrease with the noise level σ^2, and increase with the magnitude of the slope $|\beta_1|$. As $\sigma^2/n \to 0$, the test's power to detect any departures from the null $\to 1$. If we have a very powerful test, then we can reliably detect departures from the null. If we don't find them, then, we can be pretty sure they're not there. If we have a low-power test, not detecting departures from the null tells us little[4]. If σ^2 is too big or n is too small, our test is inevitably low-powered. Without knowing the power, retaining the null is ambiguous between "there's no signal here" and "we can't tell if there's a signal or not". It would be more useful to look at things like a confidence interval for the regression slope, or, if you must, for σ^2. Of course, there is also possibility (iv), that the real relationship is nonlinear, but the best linear approximation to it has slope (nearly) zero, in which case the F test will have no power to detect the nonlinearity.

Suppose instead that we reject the null, intercept-only hypothesis. *This does not mean that the simple linear model is right.* It means that the latter model predicts better than the intercept-only model — too much better to be due to chance. The simple linear regression model can be absolute garbage, with every single one of its assumptions flagrantly violated, and yet better than the model which makes all those assumptions *and* thinks the optimal slope is zero.

Figure 5 provides simulation code for a simple set up where the true regression function is nonlinear and the noise around it has non-constant variance. (Indeed, regression curve is non-monotonic and the noise is multiplicative, not additive.) Still, because a tilted straight line is a much better fit than a flat line, the F test delivers incredibly small p-values — the largest, when I simulate drawing 200 points from the model, is around 10^{-35}, which is about the probability of drawing any particular molecule from 3 billion liters of water. This is the math's way of looking at data like Figure 6 and saying "If you want to run a flat line through this, instead of one with a slope, you're crazy"[5]. This is, of course, true; it's just not an answer to "Is simple linear model right here?"

2.1 The Essential Thing to Remember

Neither the F test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$ nor the likelihood ratio test nor the Wald/t test of the same hypothesis tell us *anything* about the correctness of the simple linear regression model. All these tests *presume* the simple linear regression model with Gaussian noise is true, and check a special case (flat

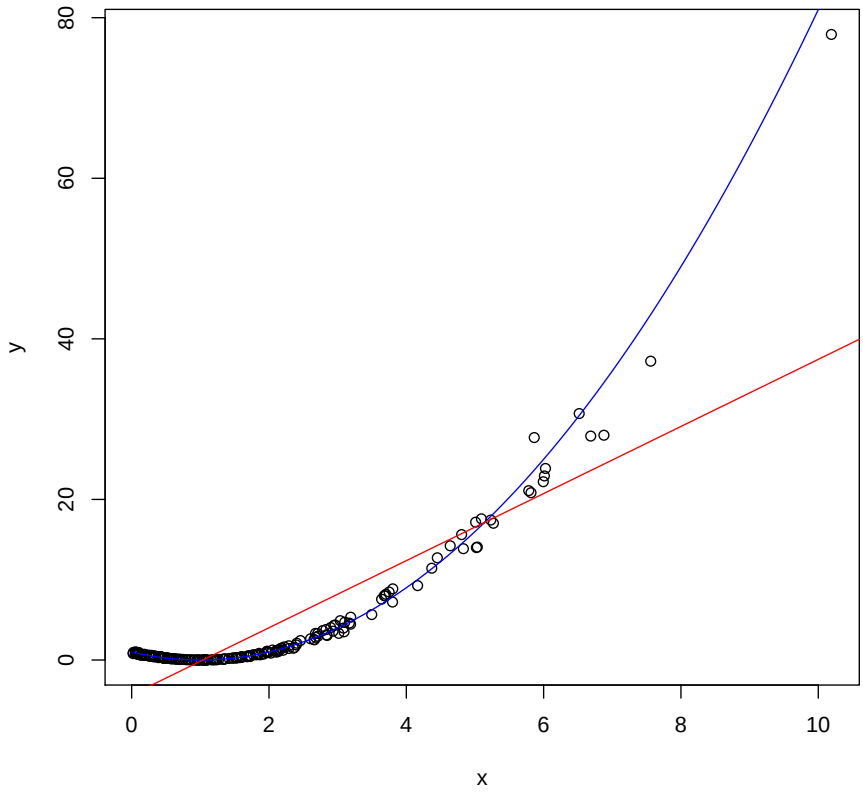
[4] Refer back to the discussion of hypothesis testing in Lecture 8.

[5] Similarly, when on p. 1.1 we're told the p-value is $\leq 2.2 \times 10^{-16}$, that doesn't mean that there's overwhelming evidence for the simple linear model, it again means that it'd be really stupid to prefer a flat line to a titled one.

```
# Simulate from a non-linear, non-constant-variance model
  # Curve: Y = (X-1)^2 * U
  # U ~ Unif(0.8, 1.2)
  # X ~ Exp(0.5)
# Inputs: number of data points; whether to return data frame or F test of
  # a simple linear model

sim.non.slr <- function(n, do.test=FALSE) {
    x <- rexp(n,rate=0.5)
    y <- (x-1)^2 * runif(n, min=0.8, max=1.2)
    if (! do.test) {
        return(data.frame(x=x,y=y))
    } else {
        # Fit a linear model, run F test, return p-value
        return(anova(lm(y~x))[["Pr(>F)"]][1])
    }
}
```

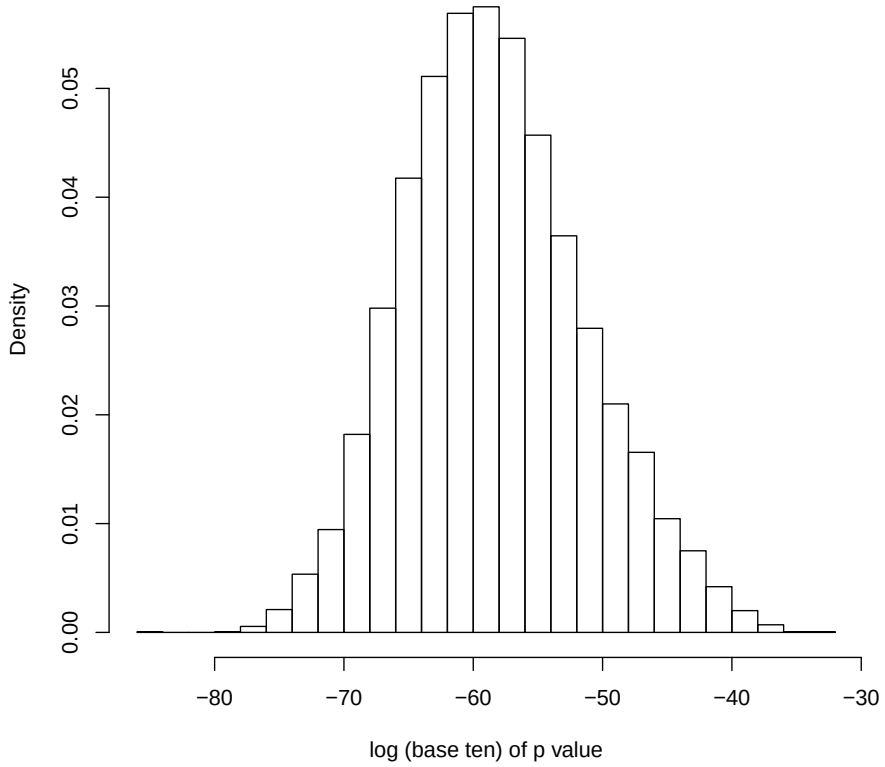
FIGURE 5: *Code to simulate a non-linear model with non-constant variance (in fact, multiplicative rather than additive noise).*

line) against the general one (titled line). They do not test linearity, constant variance, lack of correlation, or Gaussianity.




```
not.slr <- sim.non.slr(n=200)
plot(y~x, data=not.slr)
curve((x-1)^2, col="blue", add=TRUE)
abline(lm(y~x,data=not.slr), col="red")
```

FIGURE 6: *200 points drawn from the non-linear, heteroskedastic model defined in Figure 5 (black dots); the true regression curve (blue curve); the least-squares estimate of the simple linear regression (red line). Anyone who's read Lecture 6 and looks at this can realize the linear model is badly wrong here.*




```
f.tests <- replicate(1e4, sim.non.slr(n=200, do.test=TRUE))
hist(log10(f.tests),breaks=30,freq=FALSE, main="",
     xlab="log (base ten) of p value")
```

FIGURE 7: *Distribution of p values from the F test for the simple linear regression model when the data come from the non-linear, heteroskedastic model of Figure 5, with sample size of $n = 200$. The p-values are all so small that rather than plotting them, I plot their logs in base 10, so the distribution is centered around a p-value of 10^{-60}, and the* largest*, least-significant p-values produced in ten thousand simulations were around 10^{-35}.*

3 R^2

R^2 has several definitions which are equivalent when we estimate a linear model by least squares. The most basic one is the ratio of the sample variance of the fitted values to the sample variance of Y.

$$R^2 \equiv \frac{s^2_{\hat{m}}}{s^2_Y} \tag{5}$$

Alternatively, it's the ratio between the sample covariance of Y and the fitted values, to the sample variance of Y:

$$R^2 = \frac{c_{Y,\hat{m}}}{s^2_Y} \tag{6}$$

Let's show that these are equal. Clearly, it's enough to show that the sample variance of $\hat{m}$ equals its covariance with Y. The key observations are that (i) that each $y_i = \hat{m}(x_i) + e_i$, while (ii) the sample covariance between e_i and $\hat{m}(x_i)$ is exactly zero. Thus

$$c_{Y,\hat{m}} = c_{\hat{m}+e,\hat{m}} = s^2_{\hat{m}} + c_{e,\hat{m}} = s^2_{\hat{m}}$$

and we see that, *for linear models estimated by least squares*, Eqs. 5 and 6 do in fact always give the same result.

That said, what is $s^2_{\hat{m}}$? Since $\hat{m}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$,

$$s^2_{\hat{m}} = s^2_{\hat{\beta}_0 + \hat{\beta}_1 X} = s^2_{\hat{\beta}_1 X} = \hat{\beta}_1^2 s^2_X$$

We thus get a third expression for R^2:

$$R^2 = \hat{\beta}_1^2 \frac{s^2_X}{s^2_Y} \tag{7}$$

From this, we can derive yet a fourth expression:

$$R^2 = \left(\frac{c_{XY}}{s_X s_Y}\right)^2 \tag{8}$$

which we can recognize as the squared correlation coefficient between X and Y (hence the square in R^2). A noteworthy feature of this equation is that it shows we get exactly the same R^2 whether we regress Y on X, or regress X on Y.

A final expression for R^2 is

$$R^2 = \frac{s^2_Y - \hat{\sigma}^2}{s^2_Y} \tag{9}$$

Since $\hat{\sigma}^2$ is the sample variance of the residuals, and the residuals are uncorrelated (in sample) with $\hat{m}$, it's not hard to show that the numerator is equation to $s^2_{\hat{m}}$.

"Adjusted" R^2 As you remember, $\hat{\sigma}^2$ has a slight negative bias as an estimate of σ^2. One therefore sometimes sees an "adjusted" R^2, using $\frac{n}{n-2}\hat{\sigma}^2$ in place of $\hat{\sigma}^2$, that being an unbiased estimate of σ^2.

Limits for R^2 From Eq. 7, it is clear that R^2 will be 0 when $\hat{\beta}_1 = 0$. On the other hand, if all the residuals are zero, then $s^2_Y = \hat{\beta}_2^1 s^2_X$ and $R^2 = 1$. It is not too hard to show that R^2 can't possible be bigger than 1, so we have marked out the limits: a sample slope of 0 gives an R^2 of 0, the lowest possbile, and all the data points falling exactly on a straight line gives an R^2 of 1, the largest possible.

3.1 Theoretical R^2

Suppose we knew the true coefficients. What would R^2 be? Using Eq. 5, we'd see

$$
\begin{aligned}
R^2 &= \frac{\mathrm{Var}\left[m(X)\right]}{\mathrm{Var}\left[Y\right]} && (10)\\
&= \frac{\mathrm{Var}\left[\beta_0 + \beta_1 X\right]}{\mathrm{Var}\left[\beta_0 + \beta_1 X + \epsilon\right]} && (11)\\
&= \frac{\mathrm{Var}\left[\beta_1 X\right]}{\mathrm{Var}\left[\beta_1 X + \epsilon\right]} && (12)\\
&= \frac{\beta_1^2 \mathrm{Var}\left[X\right]}{\beta_1^2 \mathrm{Var}\left[X\right] + \sigma^2} && (13)
\end{aligned}
$$

Since all our parameter estimates are consistent, and this formula is continuous in all the parameters, the R^2 we get from our estimate will converge on this limit.

As you will recall from lecture 1, even if the linear model is totally wrong, our estimate of β_1 will converge on $\mathrm{Cov}\left[X, Y\right]/\mathrm{Var}\left[X\right]$. Thus, *whether or not the simple linear model applies*, the limiting theoretical R^2 is given by Eq. 13, provided we interpret β_1 appropriately.

3.2 Distraction or Nuisance?

Unfortunately, a lot of myths about R^2 have become endemic in the scientific community, and it is vital at this point to immunize you against them.

1. The most fundamental is that R^2 *does not measure goodness of fit.*
   (a) R^2 *can be arbitrarily low when the model is completely correct.* Look at Eq. 13. By making $\mathrm{Var}\left[X\right]$ small, or σ^2 large, we drive R^2 towards 0, even when every assumption of the simple linear regression model is correct in every particular.

Greetings, Redditors! May I trouble you to read to the end before commenting? (To Audiendi: I don't know who you are; I won't try to find out; you wouldn't be in trouble if I did.)

   (b) R^2 *can be arbitrarily close to 1 when the model is totally wrong.* For a demonstration, the R^2 of the linear model fitted to the simulation in §2 is 0.745. There is, indeed, no limit to how high R^2 can get when the true model is nonlinear. All that's needed is for the slope of the best linear approximation to be non-zero, and for $\mathrm{Var}[X]$ to get big.

2. R^2 is also pretty useless as a measure of predictability.

   (a) R^2 *says nothing about prediction error.* Go back to Eq. 13, the ideal case: even with σ^2 exactly the same, and no change in the coefficients, R^2 can be anywhere between 0 and 1 just by changing the range of X. Mean squared error is a *much* better measure of how good predictions are; better yet are estimates of out-of-sample error which we'll cover later in the course.

   (b) R^2 *says nothing about interval forecasts.* In particular, it gives us no idea how big prediction intervals, or confidence intervals for $m(x)$, might be.

3. R^2 cannot be compared across data sets: again, look at Eq. 13, and see that exactly the same model can have radically different R^2 values on different data.

4. R^2 cannot be compared between a model with untransformed Y and one with transformed Y, or between different transformations of Y. More exactly: it's a free country and no one will stop you from doing that, but it's meaningless; R^2 can easily go down when the model assumptions are better fulfilled, etc.

5. The one situation where R^2 can be compared is when different models are fit to the same data set with the same, untransformed response variable. Then increasing R^2 is the same as decreasing in-sample MSE (by Eq. 9). In that case, however, you might as well just compare the MSEs.

6. It is very common to say that R^2 is "the fraction of variance explained" by the regression. This goes along with calling R^2 "the coefficient of determination". These usages arise from Eq. 9, and have nothing to recommend them. Eq. 8 shows that if we regressed X on Y, we'd get exactly the same R^2. This in itself should be enough to show that a high R^2 says nothing about explaining one variable by another. It is also extremely easy to devise situations where R^2 is high even though neither one could possible explain the other[6]. Unless you want to re-define the verb "to explain" in

[6] Imagine, for example, regressing deaths in Chicago on the number of Chicagoans wearing t-shirts each day. For that matter, imagine regressing the number of Chicagoans wearing t-shirts on the number of deaths. For thousands of examples with even less to recommend them as explanations, see `http://www.tylervigen.com/spurious-correlations`.

> terms of R^2, there is no connection between it and anything which might be called a scientific explanation[7].

Using adjusted R^2 instead of R^2 does absolutely nothing to fix any of this.

At this point, you might be wondering just what R^2 is good for — what job it does that isn't better done by other tools. The only honest answer I can give you is that I have never found a situation where it helped at all. If I could design the regression curriculum from scratch, I would never mention it. Unfortunately, it lives on as a historical relic, so you need to know what it is, and what mis-understandings about it people suffer from.

4 The Correlation Coefficient

As you know, the correlation coefficient between X and Y is

$$\rho_{XY} = \frac{\mathrm{Cov}\left[X, Y\right]}{\sqrt{\mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]}}$$

which lies between -1 and 1. It takes its extreme values when Y is a linear function of X.

Recall, from lecture 1, that the slope of the ideal linear predictor β_1 is

$$\frac{\mathrm{Cov}\left[X, Y\right]}{\mathrm{Var}\left[X\right]}$$

so

$$\rho_{XY} = \beta_1 \sqrt{\frac{\mathrm{Var}\left[X\right]}{\mathrm{Var}\left[Y\right]}}$$

It's also straightforward to show (Exercise 1) that if we regress $Y/\sqrt{\mathrm{Var}\left[Y\right]}$, the *standardized* version of Y, on $X/\sqrt{\mathrm{Var}\left[X\right]}$, the standardized version of X, the regression coefficient we'd get would be ρ_{XY}.

In 1954, the great statistician John W. Tukey wrote (Tukey, 1954, p. 721)

> Does anyone know when the correlation coefficient is useful, as opposed to when it is used? If so, why not tell us?

Sixty years later, the world is still waiting for a good answer[8].

[7] Some people (e.g., Weisburd and Piquero 2008; Low-Décarie *et al.* 2014) have attempted to gather all the values of R^2 reported in scientific papers on, say, ecology or crime, to see if ecologists or criminologists have gotten better at explaining the phenomena they study. I hope it's clear why these exercises are pointless.

[8] To be scrupulously fair, Tukey did admit there was one clear case where correlation coefficients were useful; they are, as we have just seen, basically the slopes in simple linear regressions. But even so, as soon as we have multiple predictors (as we will in two weeks), regression will no longer match up with correlation. Also, *covariances* are useful, but why turn a covariance into a correlation?

5 More on the Likelihood Ratio Test

This section is optional, but *strongly* recommended.

We're assuming that the true parameter value, call it θ, lies in the restricted class of models ω. So there are q components to θ which matter, and the other $p-q$ are fixed by the constraints defining ω. To simplify the book-keeping, let's say those constraints are all that the extra parameters are zero, so $\theta = (\theta_1, \theta_2, \ldots \theta_q, 0, \ldots 0)$, with $p-q$ zeroes at the end.

The restricted MLE $\widehat{\theta}$ obeys the constraints, so

$$\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots \widehat{\theta}_q, 0, \ldots 0) \tag{14}$$

The unrestricted MLE does not have to obey the constraints, so it's

$$\widehat{\Theta} = (\widehat{\Theta}_1, \widehat{\Theta}_2, \ldots \widehat{\Theta}_q, \widehat{\Theta}_{q+1}, \ldots \widehat{\Theta}_p) \tag{15}$$

Because both MLEs are consistent, we know that $\widehat{\theta}_i \to \theta_i$, $\widehat{\Theta}_i \to \theta_i$ if $1 \leq i \leq q$, and that $\widehat{\Theta}_i \to 0$ if $q+1 \leq i \leq p$.

Very roughly speaking, it's the last extra terms which end up making $L(\widehat{\Theta})$ larger than $L(\widehat{\theta})$. Each of them tends towards a mean-zero Gaussian whose variance is $O(1/n)$, but their impact on the log-likelihood depends on the square of their sizes, and the square of a mean-zero Gaussian has a χ^2 distribution with one degree of freedom. A whole bunch of factors cancel out, leaving us with a sum of $p-q$ independent χ^2_1 variables, which has a χ^2_{p-q} distribution.

In slightly more detail, we know that $L(\widehat{\Theta}) \geq L(\widehat{\theta})$, because the former is a maximum over a larger space than the latter. Let's try to see how big the difference is by doing a Taylor expansion around $\widehat{\Theta}$, which we'll take out to second order.

$$\begin{aligned} L(\widehat{\theta}) &\approx L(\widehat{\Theta}) + \sum_{i=1}^{p} (\widehat{\Theta}_i - \widehat{\theta}_i) \left(\left. \frac{\partial L}{\partial \theta_i} \right|_{\widehat{\Theta}} \right) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (\widehat{\Theta}_i - \widehat{\theta}_i) \left(\left. \frac{\partial^2 L}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\Theta}} \right) (\widehat{\Theta}_j - \widehat{\theta}_j) \\ &= L(\widehat{\Theta}) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (\widehat{\Theta}_i - \widehat{\theta}_i) \left(\left. \frac{\partial^2 L}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\Theta}} \right) (\widehat{\Theta}_j - \widehat{\theta}_j) \end{aligned} \tag{16}$$

All the first-order terms go away, because $\widehat{\Theta}$ is a maximum of the likelihood, and so the first derivatves are all zero there. Now we're left with the second-order terms. Writing all the partials out repeatedly gets tiresome, so abbreviate $\partial^2 L / \partial \theta_i \partial \theta_j$ as $L_{,ij}$.

To simplify the book-keeping, suppose that the second-derivative matrix, or **Hessian**, is diagonal. (This should seem like a swindle, but we get the same conclusion without this supposition, only we need to use a lot more algebra — we diagonalize the Hessian by an orthogonal transformation.) That is, suppose

$L_{,ij} = 0$ unless $i = j$. Now we can write

$$L(\widehat{\Theta}) - L(\widehat{\theta}) \approx -\frac{1}{2}\sum_{i=1}^{p}(\widehat{\Theta}_i - \widehat{\theta}_i)^2 L_{,ii} \tag{17}$$

$$2\left[L(\widehat{\Theta}) - L(\widehat{\theta})\right] \approx -\sum_{i=1}^{q}(\widehat{\Theta}_i - \widehat{\theta}_i)^2 L_{,ii} - \sum_{i=q+1}^{p}(\widehat{\Theta}_i)^2 L_{,ii} \tag{18}$$

At this point, we need a fact about the asymptotic distribution of maximum likelihood estimates: they're generally Gaussian, centered around the true value, and with a shrinking variance that depends on the Hessian evaluated at the true parameter value; this is called the **Fisher information**, F or I. (Call it F.) If the Hessian is diagonal, then we can say that

$$\widehat{\Theta}_i \rightsquigarrow \mathcal{N}(\theta_i, -1/nF_{ii}) \tag{19}$$

$$\widehat{\theta}_i \rightsquigarrow \mathcal{N}(\theta_1, -1/nF_{ii}) \; 1 \leq i \leq q \tag{20}$$

$$\widehat{\theta}_i = 0 \; q+1 \leq i \leq p \tag{21}$$

Also, $(1/n)L_{,ii} \to -F_{ii}$.

Putting all this together, we see that each term in the second summation in Eq. 18 is (to abuse notation a little)

$$\frac{-1}{nF_{ii}}(\mathcal{N}(0,1))^2 L_{,ii} \to \chi^2_1 \tag{22}$$

so the whole second summation has a χ^2_{p-q} distribution. The first summation, meanwhile, goes to zero because $\widehat{\Theta}_i$ and $\widehat{\theta}_i$ are actually strongly correlated, so their difference is $O(1/n)$, and their difference squared is $O(1/n^2)$. Since $L_{,ii}$ is only $O(n)$, that summation drops out.

A somewhat less hand-wavy version of the argument uses the fact that the MLE is really a vector, with a multivariate normal distribution which depends on the inverse of the Fisher information matrix:

$$\widehat{\Theta} \rightsquigarrow \mathcal{N}(\theta, (1/n)F^{-1}) \tag{23}$$

Then, at the cost of more linear algebra, we don't have to assume that the Hessian is diagonal.

6 Concluding Comment

The tone I have taken when discussing F tests, R^2 and correlation has been dismissive. This is deliberate, because they are grossly abused and over-used in current practice, especially by non-statisticians, and I want you to be too proud (or too ashamed) to engage in those abuses. In a better world, we'd just skip over them, but you will have to deal with colleagues, and bosses, who learned

their statistics in the bad old days, and so have to understand what they're doing wrong. ("Science advances funeral by funeral".)

In all fairness, the people who *came up* with these tools were great scientists, struggling with very hard problems when nothing was clear; they were inventing all the tools and concepts we take for granted in a class like this. Anyone in this class, me included, would be doing very well to come up with *one* idea over the whole of our careers which is as good as R^2. But we best respect our ancestors, and the tradition they left us, when we improve that tradition where we can. Sometimes that means throwing out the broken bits.

7 Further Reading

Refer back to lecture 7 on diagnostics for ways of *actually* checking whether the relationship between Y and X is linear (along with the other assumptions of the model). We will come back to the topic of conducting formal tests of linearity, or other parametric regression specifications, in 402.

Refer back to lecture 8, on parametric inference, for advice on when it is actually interesting to test the hypothesis $\beta_1 = 0$.

Full mathematical treatments of likelihood ratio tests can be found in many textbooks, e.g., Schervish (1995) or Gouriéroux and Monfort (1989/1995, vol. II). The original proof that it has a χ^2_{p-q} asymptotic distribution was given by Wilks (1938). Vuong (1989) provides an interesting and valuable treatment of what happens to the likelihood ratio test when *neither* the null nor the alternative is strictly true, but we want to pick the one which is *closer* to the truth; that paper also develops the theory when the null is not a restriction of the alternative, but rather the two hypotheses come from distinct statistical models.

People have been warning about the fallacy of R^2 to measure goodness of fit for a long time (Anderson and Shanteau, 1977; Birnbaum, 1973), apparently without having much of an impact. (See Hamilton (1996) for a discussion of how academic communities can keep on teaching erroneous ideas long after they've been shown to be wrong, and some speculations about why this happens.)

That R^2 has got nothing to do with *explaining* anything has also been pointed out, time after time, for decades (Berk, 2004). A small demo of just how silly "variance explained" can get, using the Chicago data, can be found at `http://bactra.org/weblog/874.html`. Just what it *does* means to give a proper scientific explanation, and what role statistical models might play in doing so, is a topic full of debate, not to say confusion. Shmueli (2010) attempts to relate some of these debates to the practical conduct of statistical modeling. Personally, I have found Salmon (1984) very helpful in thinking about these issues.

Exercises

To think through or to practice on, not to hand in.

1. Define $\tilde{Y} = Y/\sqrt{\mathrm{Var}\left[Y\right]}$ and $\tilde{X} = X/\sqrt{\mathrm{Var}\left[X\right]}$. Show that the slope of the optimal linear predictor of $\tilde{Y}$ from $\tilde{X}$ is ρ_{XY}.

2. Work through the likelihood ratio test for testing regression through the origin ($\beta_0 = 0$) against the simple linear model ($\beta_0 \neq 0$); that is, write Λ in terms of the sample statistics and simplify as much as possible. Under the null hypothesis, 2Λ follows a χ^2 distribution with a certain number of degrees of freedom: how many?

References

Anderson, Norman H. and James Shanteau (1977). "Weak inference with linear models." *Psychological Bulletin*, **84**: 1155–1170. doi:10.1037/0033-2909.84.6.1155.

Berk, Richard A. (2004). *Regression Analysis: A Constructive Critique*. Thousand Oaks, California: Sage.

Birnbaum, Michael H. (1973). "The Devil Rides Again: Correlation as an Index of Fit." *Psychological Bulletin*, **79**: 239–242. doi:10.1037/h0033853.

Gouriéroux, Christian and Alain Monfort (1989/1995). *Statistics and Econometric Models*. Themes in Modern Econometrics. Cambridge, England: Cambridge University Press. Translated by Quang Vuong from *Statistique et modèles économétriques*, Paris: Économica.

Hamilton, Richard F. (1996). *The Social Misconstruction of Reality: Validity and Verification in the Scholarly Community*. New Haven: Yale University Press.

Low-Décarie, Etienne, Corey Chivers and Monica Granados (2014). "Rising complexity and falling explanatory power in ecology." *Frontiers in Ecology and the Environment*, **12**: 412–418. doi:10.1890/130230.

Salmon, Wesley C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Schervish, Mark J. (1995). *Theory of Statistics*. Berlin: Springer-Verlag.

Shmueli, Galit (2010). "To Explain or to Predict?" *Statistical Science*, **25**: 289–310. doi:10.1214/10-STS330.

Tukey, John W. (1954). "Unsolved Problems of Experimental Statistics." *Journal of the American Statistical Association*, **49**: 706–731. URL `http://www.jstor.org/pss/2281535`.

Vuong, Quang H. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica*, **57**: 307–333. URL http://www.jstor.org/pss/1912557.

Weisburd, David and Alex R. Piquero (2008). "How Well Do Criminologists Explain Crime? Statistical Modeling in Published Studies." *Crime and Justice*, **37**: 453–502. doi:10.1086/524284.

Wilks, S. S. (1938). "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *Annals of Mathematical Statistics*, **9**: 60–62. URL http://projecteuclid.org/euclid.aoms/1177732360. doi:10.1214/aoms/1177732360.