STA302/STA1001, Week 3

Mark Ebden, 21–26 September 2017

With grateful acknowledgment to Alison Gibbs and Becky Lin

Today's class

- The Confidence Interval in Linear Regression
- Hypothesis testing on β_0 and β_1
- Regression Analysis of Variance
- Reference: Simon Sheather §§2.2, 2.3, 2.5

Computing Labs with R installed

Robarts has a Computer Lab open whenever the library itself is open:

- `https://mdl.library.utoronto.ca/technology/computer-lab`
- Monday to Friday 8:30 am to 11 pm
- Saturday 9 am - 10 pm
- Sunday 10 am - 10 pm

There are also four IIT (Information & Instructional Technology) labs:

- In Sidney Smith Hall, Carr Hall, and in Ramsay Wright
- Need Help with an IIT lab? Phone: 416-946-HELP (4357)
- Email: `iit@artsci.utoronto.ca`
- Walk-in: Come to Sidney Smith Room 572 (IIT Office), Monday to Friday, 8:45 am - 5:00 pm

More about the IIT Computer Labs

The four are:

- Sidney Smith Hall room 561 (lower level) (49 seats) - 100 St. George Street: 8:45 am to 7 pm
- Carr Hall room 325 (3rd floor) (30 seats) - 100 St. Joseph Street: 8:45 am to 9 pm
- Ramsay Wright room 107 (20 seats) - 25 Harbord Street: 8:45 am to 9 pm
- Ramsay Wright room 109 (24 seats) - 25 Harbord Street: 8:45 am to 9 pm

Before dropping in, click the links at left here to ensure the room hasn't been booked: `http://lab.chass.utoronto.ca/schedules.php`

More about the IIT Computer Labs

Logging in:

- You must use a valid UTORid and password to log in to lab computers
- If you have trouble logging in, please verify your UTORid credentials at `https://www.utorid.utoronto.ca` (click on the "verify" link under the yellow "Problems with your UTORid?" heading). If your UTORid username and password do not work, reset your password on this page.
- For more help, contact the IIT labs, or reach the Information Commons helpdesk at 416-978-HELP (4357) or `help.desk@utoronto.ca`

More about the IIT Computer Labs

Printing:

- Printing is available in the Sidney Smith and Ramsay Wright labs, but not Carr Hall
- You must have a TCard with sufficient value stored on it. A card reader attached to the print release station will debit the print job cost from your TCard at the time of printing

Saving Data:

- Data is not saved on the lab computers
- Back-up your data frequently, and ensure you have an appropriate storage and/or back-up method for your files (e.g. use a USB key or email materials to yourself)

A note about correlation

In Week 2, we introduced the assumption that the e_i's are uncorrelated. This means that:

$$\rho_{ij} = \frac{\text{cov}(e_i, e_j)}{\sigma_i \sigma_j} = 0 \quad \forall\, i \neq j$$

where ρ_{ij} indicates the linear correlation between any two of the e's

Lack of correlation is a gentler assumption than independence:

- Two independent random variables will have correlation 0, but not necessarily vice versa
- Consider for example $X \sim \text{Unif}(-1, 1)$ and $Y = X^2$, which are dependent but $\text{cov}(X, Y) = \mathbb{E}(X^3) = 0$

Towards a Confidence Interval

For a chosen value of x^*,

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Because $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates,

$$\mathbb{E}(\hat{y}^*) = \beta_0 + \beta_1 x^*$$

And, using our equations from Week 2,

$$\begin{aligned}
\text{var}(\hat{y}^*) &= \text{var}(\hat{\beta}_0) + \ (x^*)^2 \text{var}(\hat{\beta}_1 x^*) \ + \ 2x^* \text{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\
&= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] + \ \frac{(x^*)^2 \sigma^2}{S_{xx}} \ - \ \frac{2x^* \sigma^2 \bar{x}}{S_{xx}} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]
\end{aligned}$$

Towards a Confidence Interval

Now bringing in our assumption from Tuesday that the errors are normally distributed:

$$\hat{y}^* \sim \mathcal{N}\left(\beta_0 + \beta_1 x^*,\ \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]\right)$$

Equivalently we can write this as

$$Z = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{\sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim \mathcal{N}(0, 1)$$

Towards a Confidence Interval

We don't generally know σ^2, but can estimate using the mean square error, S^2, as in question 3 from last week. This changes our Z score into a T score:

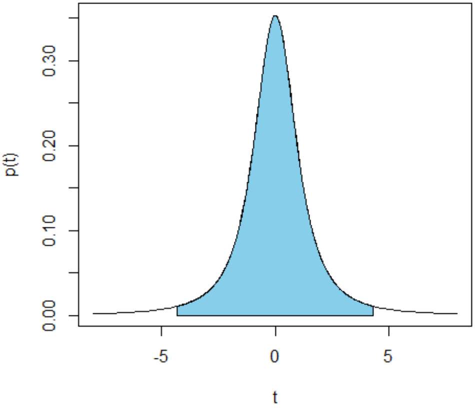
$$T = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \ \sim \ t_{n-2}$$

This distribution tells us that for a given value of x^*:

- The difference between $\hat{y}^*$ and the population regression line's ordinate, $\mathbb{E}(Y|X = x^*) = \beta_0 + \beta_1 x^*$, follows a (scaled) t_{n-2} distribution

A Confidence Interval

What upper- and lower bounds on $\hat{y}^*$ can be expected to encompass the population regression line, i.e. encompass the true $\mathbb{E}(Y^*)$, 95% of the time?



The answer is called a 95% confidence interval.

R code to shade a graph

```
c1 = qt(0.025,2) # Left bound of shaded region
c2 = qt(0.975,2)
x0 = 8 # Highest t-score to plot
myseq = seq(c1, c2, 0.01)
cx <- c(c1,myseq,c2) # vector of x-points to outline shaded region
cy <- c(0,dt(myseq,2),0)
curve(dt(x,2),xlim=c(-x0,x0),xlab='t',ylab='p(t)')
polygon(cx,cy,col='skyblue') # connect the dots
```

You don't need to know the `curve` and `polygon` commands

Quantiles of t_{n-2}

We'll represent the quantile function, $F^{-1}(p)$, of the t distribution by $t(1-p, \nu)$, where p is the cumulative probability and ν is the number of degrees of freedom.

For our 95% confidence interval:

- In the lower bound we'll set $p = \alpha/2 = 0.05/2$
- In the upper bound we'll set $p = 1 - \alpha/2 = 0.975$

Thus we're interested in two cases: $t(\alpha/2, n-2)$ and $t(1-\alpha/2, n-2)$.

Equivalently, because the t distribution is symmetric, and because $\alpha = 0.05$, we're interested in $\pm\, t(0.025, n-2)$.

Specifying the Confidence Interval

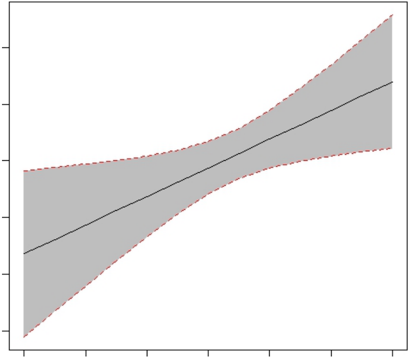
From our expression for T (slide 10), we see that the two limits of the confidence interval are given by:

$$\hat{y}^* \pm t(0.025, n-2)\, S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

or equivalently:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) \pm t(0.025, n-2)\, S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Plot of Pointwise Confidence intervals



Exercise: Produce this kind of plot for a small data set:

$$\{(2,1),(4,3),(6,4)\}$$

Don't worry about shading, but you should know how to plot the three lines: upper, mean, lower.

What about Confidence Intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$?

Developing on question #3

Our estimator of σ^2 in question #3 from last week, S^2, is the Mean Square Error (MSE).

Our means and variances are expressed in terms of σ, which is unknown, hence the importance of question #3.

For example, the variance of $\hat{\beta}_1$ was found to be

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

However, we use S in place of σ to get:

$$\widehat{\text{var}\left(\hat{\beta}_1\right)} = \frac{S^2}{S_{xx}}$$

Standard error

The square root of this is known as the *standard error* (the estimate of the standard deviation of a parameter) in regression. So,

$$\mathrm{se}\left(\hat{\beta}_1\right) = \sqrt{\frac{S^2}{S_{xx}}}$$

and of course

$$\mathrm{se}\left(\hat{\beta}_0\right) = \sqrt{S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

You're already used to more simply referring to standard error as the standard deviation of a sampling distribution.

Recap of our guesses about β_1

We've shown how to estimate the mean and variance of $\hat{\beta}_1$.

Then, following the same kind of logic we used in the confidence intervals for $\hat{y}^*$, we can show that:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\text{se}\left(\hat{\beta}_1\right)} \sim t_{n-2}$$

And thus the bounds of the confidence interval are:

$$\hat{\beta}_1 \pm t(0.025, n-2)\,\text{se}(\hat{\beta}_1)$$

Similarly, for $\hat{\beta}_0$:

$$\hat{\beta}_0 \pm t(0.025, n-2)\,\text{se}(\hat{\beta}_0)$$

More than one conception of standard error

1. A familiar way to find standard error:

- Collect n observations of some phenomenon
- Measure the sample variance, s^2
- $\text{se} = \sigma/\sqrt{n}$ and $\widehat{\text{se}} = s/\sqrt{n}$
- Some authors (but not Rice for example) say directly: $\text{se} = s/\sqrt{n}$

2. In regression analysis:

- Estimate the variance of the ith predictor estimate, i.e. $\widehat{\text{var}\left(\hat{\beta}_i\right)}$
- $\text{se} = \sqrt{\widehat{\text{var}\left(\hat{\beta}_i\right)}}$
- i.e. we're concerned with the s.d. of a parameter that stemmed from linear regression, not from a sampling distribution
- If you don't like conflating two terms, you may refer to one as the "s.e. of the regression"

Today's class

- The Confidence Interval in Linear Regression
- **Hypothesis testing on β_0 and β_1**
- Regression Analysis of Variance
- Reference: Simon Sheather §§2.2, 2.3, 2.5

Suppose we want to test whether β_1 is likely to be a particular value, β_1^0. For example, perhaps $\beta_1^0 = 0$.

This is an example of the kind of problem on which we can apply a *hypothesis test*

Hypothesis testing

We establish a pair of hypotheses:

- H_0 (null hypothesis): $\beta_1 = \beta_1^0$
- H_1 or H_a (alternative hypothesis): $\beta_1 \neq \beta_1^0$

A statistical hypothesis evaluates the compatibility of H0 with the data. We can evaluate H_0 by answering:

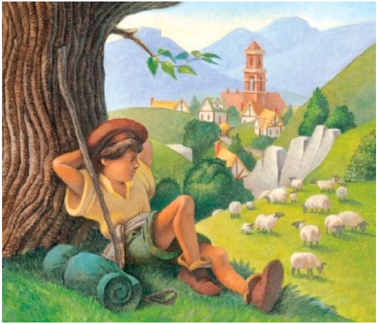
- Is our estimated $\hat{\beta}_1$ plausible/probable if H_0 is true?
- Is the difference between β_1^0 and our estimated $\hat{\beta}_1$ *large* compared to experimental noise?

The outcome here is binary:

- Reject H_0 (accept H_1), or don't reject H_0 (some authors would say "accept H_0")
- Therefore, whenever we run a hypothesis test, we run the risk of drawing one of two kinds of false conclusion (next slide)

What can go wrong with statistical hypothesis testing?

Decision	H_0 True	H_0 False
Do not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct



Error rates

The type I error rate is defined as:

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ is true})$$

The type II error rate is defined as:

$$\beta = P(\text{Don't reject } H_0 | H_1 \text{ is true})$$

It's perhaps unfortunate for us that this represents another β, by coincidence. Not to be confused with our familiar β_0 or β_1 in STA302.

Statistical hypotheses and power

Power (a.k.a. sensitivity) is defined as:

$$\begin{aligned}
\text{power} &= 1 - \beta \\
&= 1 - P(\text{Don't reject } H_0 | H_1 \text{ is true}) \\
&= P(\text{Reject } H_0 | H_1 \text{ is true})
\end{aligned}$$

The probability that a fixed-level α test will reject H_0 when a particular alternative value of the parameter is true is called the *power* of the test to detect that alternative.

How to decide which hypothesis is more likely

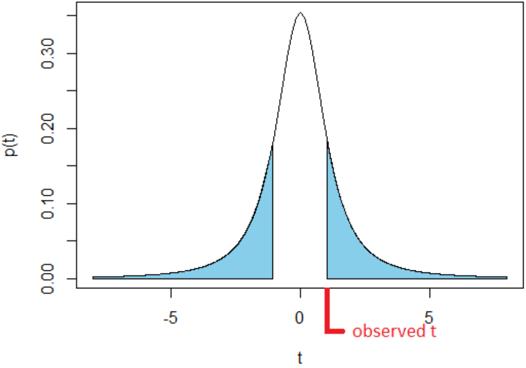
- You've encountered several statistics which measure central tendency, variability, etc, in an effort to describe/summarize some data
- When a statistic is used in hypothesis testing, it's known as the *test statistic*
- And when this statistic follows a t-distribution under the null hypothesis, our hypothesis test is an example of a t-**test**, a.k.a. Student's t-test
- These should usually be two-sided (we prepare for the test statistic's being abnormally high or low) but you do see one-sided tests as well (when the analyst says they have good reason to only check for one or the other of the high/low cases)

Key point: Temporarily assume H_0 is true. Then t_{observed} would be an observation from a t_{n-2} distribution. Is the t_{observed} you saw actually a reasonable-looking sample from that distribution?

The Student's t-test

This is one kind of testing that reports a "p-value". Based on the density function $p(t)$, and the observed statistic t_{observed}:

$$
\begin{aligned}
p\text{-value} &= P\left(t \text{ is as extreme or more extreme than } t_{\text{observed}} \mid H_0 \text{ true}\right) \\
&= P\left(|t| \geq |t_{\text{observed}}| \mid H_0 \text{ true}\right) \quad \leftarrow \text{for a two-sided } t\text{-test}
\end{aligned}
$$



From the p-value to the results of a hypothesis test

We ask whether there is any contradiction between H_0 and the observed data

- The p-value is the probability under the null hypothesis of obtaining a result as extreme or more extreme than the observed result
- A small p-value implies evidence against the null hypothesis
- A large p-value implies no evidence against the null hypothesis

If the p-value is large does this imply that the null hypothesis is true?

What does the p-value say about the probability that the null hypothesis is true? Try using Bayes' rule to figure this out.

How small is small?

One approach:

- Set a significance level, α, before conducting the test
- A popular choice is $\alpha = 0.05$
- If the *p*-value is below α, you reject the null hypothesis (and accept H_1)
- An advantage of this approach is that it gets you to think about the problem and the data carefully before data are collected. What α would you really like?

However:

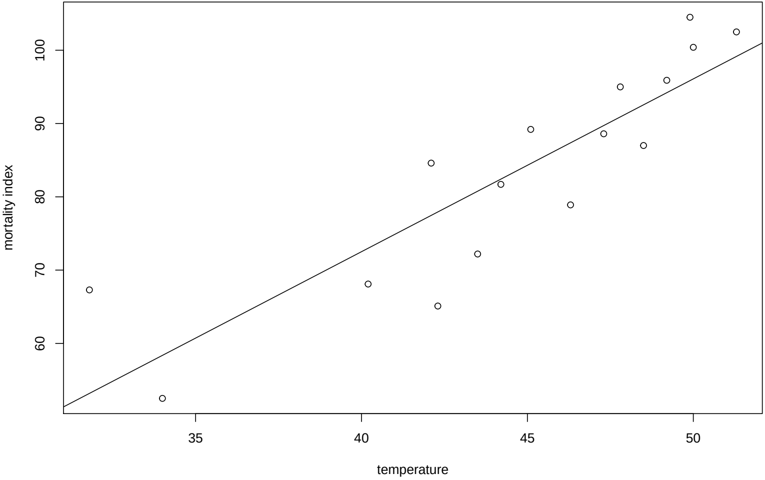
- This approach can be considered wasteful, since *p*-values of 0.04 and 10^{-4} yield the same result
- Ronald Fisher tended to report the *p*-value and let it speak for itself

R combines the best of both worlds, as we'll see

Procedure for a t test

1. Assume the null hypothesis, H_0
2. Calculate your T statistic given H_0
3. Was your observed result plausible? Yes/no: accept H_0/H_1

Returning to the temperature/mortality dataset



R has already calculated our p-value

```r
summary(myFit)
```

```
##
## Call:
## lm(formula = M ~ T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8358  -5.6319   0.4904   4.3981  14.1200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.7947    15.6719  -1.391    0.186
## T             2.3577     0.3489   6.758  9.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our p-value affects our interpretation

Interpreting b_0 or b_1 when their p-value is low:

- What does the slope mean? For each unit increase in X, Y can be expected to increase by b_1X
- What does the intercept mean? The b_0 has meaning when you are studying very small values of X. It tells you what Y might be when X is around 0

Interpreting b_0 or b_1 when their p-value is high:

- We can say very little in such cases

Extra information: the two-sample t-test

Suppose that there is a clinical trial, in which subjects are randomized to treatments A or B with equal probability. Let μ_A be the mean response in the group receiving drug A and μ_B be the mean response in the group receiving drug B. The null hypothesis is that there is no difference between A and B; the alternative claims there is a clinically meaningful difference between them.

$$H_0 : \mu_A = \mu_B \text{ versus } H_1 : \mu_A \neq \mu_B$$

We want to know if the standard treatment is better than the experimental treatment, or vice versa

The two-sample t-test

Let's assume the patient data are independent random samples from a normal distribution with means μ_A and μ_B but the same variance.

Let's use $\bar{y}_A - \bar{y}_B$ as our test statistic. The distribution is

$$\bar{y}_A - \bar{y}_B \sim \mathcal{N}\left(\mu_A - \mu_B, \sigma^2(1/n_A + 1/n_B)\right).$$

So,

$$\frac{(\bar{y}_A - \bar{y}_b) - \delta_\mu}{\sigma\sqrt{1/n_A + 1/n_B}} \sim \mathcal{N}(0, 1)$$

and we can set δ_μ to zero and continue as per slides 28–30.

Today's class

- The Confidence Interval in Linear Regression
- Hypothesis testing on β_0 and β_1
- **Regression Analysis of Variance**
- Reference: Simon Sheather §§2.2, 2.3, 2.5

Regression Analysis of Variance

How well does the regression line summarize the data?

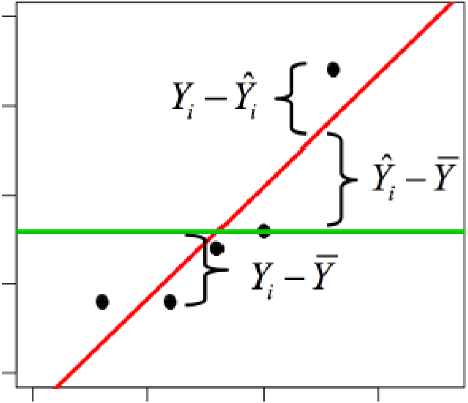
Decomposition of sums of squares:

$$
\begin{aligned}
y_i &= \hat{y}_i + \hat{e}_i \\
&= b_0 + b_1 x_i + \hat{e}_i \\
&= \bar{y} - b_1\bar{x} + b_1 x_i + \hat{e}_i \\
y_i - \bar{y} &= b_1(x_i - \bar{x}) + \hat{e}_i
\end{aligned}
$$

Squaring both sides, and summing, leads to:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} b_1^2(x_i - \bar{x})^2 + \sum_{i=1}^{n} \hat{e}_i^2$$

The building blocks of ANOVA



Analysis of variance

a.k.a. ANOVA or "Decomposition of SS", where SS = sum of squares

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} b_1^2(x_i - \bar{x})^2}_{\text{SSReg}} + \underbrace{\sum_{i=1}^{n} \hat{e}_i^2}_{\text{RSS}}$$

SST ("Total SS"):

- Also known as *Corrected SS*
- This is by comparison with the "uncorrected SS", which is just $\sum_{i=1}^{n} y_i^2$

SSReg ("Model SS" or Regression SS):

- It is the amount of variation in y's explained by the regression line

RSS ("Residual sum of squares", or Error sum of squares):

- The method of least squares minimized this

Exercise

Show that

$$b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum (\hat{y}_i - \bar{y})^2$$

The ANOVA Table

We usually summarize these quantities as:

Source	SS	d.f.	MS = SS/df
Regression line	$b_1^2 S_{xx} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	$b_1^2 S_{xx}$
Error	$\sum_{i=1}^{n} \hat{e}_i^2$	$n - 2$	S^2
Total	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$	–

Coefficient of Determination

$$R^2 = \frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}, \quad 0 \leq R^2 \leq 1$$

R^2 gives the percent of variation in y's that is explained by the regression line

In the Montreal Protocol dataset, we have $R^2 \approx \frac{203119}{203993} \approx 99.6\%$

R^2 is useful, but:

- No absolute rules about how big it should be
- Not resistant to outliers (we'll see this next week)
- Not meaningful for models with no intercept
- We can get a very high R^2 by overfitting (complicated model, may fit well for data you have but won't work well on other data)

Means

Mean square of regression $=$ MSReg $=$ SSReg $/\ 1 = b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$

Think of MSReg as an estimator, $\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$

$$\mathbb{E}(\text{MSReg}) = \sigma^2 + \beta_1^2 S_{xx}$$

MSE "Mean Square Error" $=$ RSS$/n - 2 = \sum_{i=1}^n \hat{e}_i^2/(n-2)$

$$\mathbb{E}(\text{MSE}) = \sigma^2$$

Reminder of distribution theory

If $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$, and U and V are independent, then

$$\frac{U/\nu_1}{V/\nu_2} \sim \ ?$$

ANOVA - F statistic

- This idea, due to Ronald Fisher, is about comparing variations
- Fisher introduced the method in his 1925 book "Statistical Methods for Research Workers"
- This statistical procedure enables us to answer several questions at once
- Before, the prevailing method was to test one thing at a time
- In the 1925 book, he included one F table for various numerator and denominator degrees of freedom
  - The table gave the critical values for only the 5% points
  - As use of the method spread, so did the use of the 5% level (Stephen Stigler, *Fisher and the 5% level*, 2008)

A new hypothesis test

If $\beta_1 = 0$, $\mathbb{E}(\text{MSReg}) = \mathbb{E}(\text{MSE})$.

Moreover, if $\beta_1 = 0$, then $\frac{\text{MSReg}}{\sigma^2} \sim \chi^2(1)$ and $\frac{\text{MSE}(n-2)}{\sigma^2} \sim \chi^2(n-2)$

Therefore, if $\beta_1 = 0$,

$$\frac{\frac{\text{MSReg}}{\sigma^2}/1}{\frac{\text{MSE}(n-2)}{\sigma^2}/(n-2)} \sim F_{1,n-2}$$

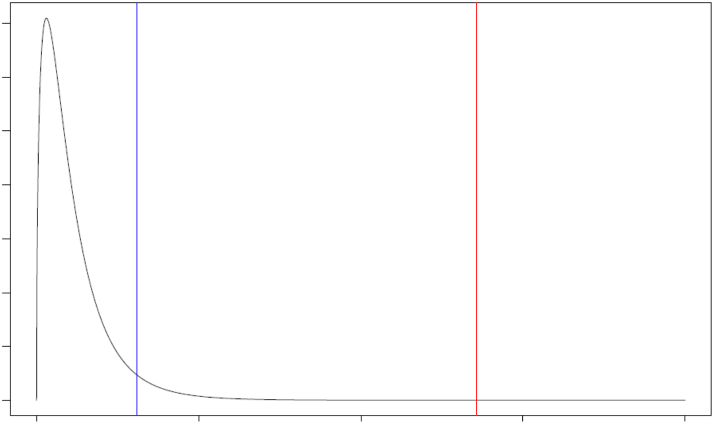
This opens up another test of $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

What is the test statistic?

We can use as our test statistic $F_{\text{obs}} = \frac{\text{MSReg}}{MSE}$:

- Under H_0, this is an observation from an F distribution with 1 and $n-2$ degrees of freedom
- $\beta_1 \neq 0$ gives larger values of F_{obs}, so deviations from $\beta_1 = 0$ are in the right tail of the F distribution
- On the Montreal Protocol data, we get a high F_{obs}, leading to again get $p < 0.001$. This is strong evidence that β_1 isn't 0.

Example



F versus t

In general, the square of a r.v. with a t_m distribution results in a r.v. with an $F_{1,m}$ distribution.

This approach is more useful in multiple linear regression (more than one predictor), which we'll do after the midterm.

For now, an exercise for you: Show, in general, that $t^2_{\text{obs}} = F_{\text{obs}}$

Next steps

- Solutions to HW #1 to be posted very soon – last chance to try them without peaking!
- Next TA office hours: tomorrow morning

Exercises:

- Try today's plotting exercise, and the proofs
- Try the seven questions at the back of Chapter 2 in Simon Sheather's textbook
- Use R where it would make things easier