STAT302 Methods of Data Analysis 1
Module 6: Decomposition of Variance

Austin Brown

University of Toronto, Department of Statistical Sciences

October 17, 2024

Review/Misc

Course review

Let's review what we have covered so far.

- Why don't we just use least squares on a data set and report the estimates? We don't need any regression model here.
- Every data set is different and there are specific answers depending on your data set but not universal answers to questions to every project.
- Weak relationship does not equal no relationship!

Notation

You will see different notation throughout the exam and literature such as a regression model defined by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

with $\epsilon \sim N(0, \sigma^2)$. The notation may be slightly different but the concept is exactly the same. We used extra notation to learn to be careful not to mix up what is the random and what are fixed numbers computed from the data.

- Sheather's textbook uses just se to denote the estimated standard error.
- Rencher's textbook uses s^2 to denote $\hat{\sigma}^2$.
- etc.

Course review

So far we have learned how to build a single regression model and perform statistical inference on the population with this model with points estimates, confidence intervals, and prediction intervals.

The next big topic is to learn the basics of how to determine if we should remove predictors or select/reduce a regression model. The topic is often called **model selection**.

Review: statistical hypothesis testing

- Null hypothesis H_0 and alternative hypothesis H_A
- The level of the test α (error tolerance)
- Test statistic, its sampling distribution, and then its computed value from the data
- Computed p-value
- Conclusion

Lecture 1: Hypothesis tests for invididual coefficients

Learning objectives

- Learn how to perform a hypothesis test for an individual coefficient for a predictor
- Implement this in R

Model selection / inference

Consider the population

$$Y = -1 + .5x_1 + 0x_2 + e$$

$e \sim N(0, .5)$. Then really the population follows

$$Y = -1 + .5x_1 + e.$$

and **there is no linear relationship (on average) between the average response and the predictor** x_2. At the same time, the least squares fit on a data set may be result in nonzero estimated coefficients.

Question: Any ideas on how to draw conclusions about this from a data set?

Confidence intervals for invdividual coefficients

We learned $(1-\alpha)$ % confidence intervals for coefficients of predictors. If the confidence interval contains 0, then this provides us statistical evidence for a 0 coefficient. However, if $\alpha = .05$ or $\alpha = .1$, the level of statistical evidence changes. We also need to be careful how we interpret a confidence interval.

Test for individual coefficients: hypothesis

The **null hypothesis** $H_0 : \beta_i = \beta_i^0$ versus the **alternative hypothesis** $H_A : \beta_i \neq \beta_i^0$

- Usually test $\beta_i^0 = 0$.
- Can also perform one-sided alternatives like $\beta_i > 0$, etc.

Test for individual coefficients: test statistic

Under the regression model assumptions and the null hypothesis H_0, the random test statistic has sampling distribution

$$T = \frac{\hat{\beta}_i - \beta_i^0}{\hat{se}(\hat{\beta}_i)} \sim t_{n-(p+1)}.$$

Under H_0, we assume to know β_i^0 and so we do not have an unknown value here. We can compute an observed test statistic from the data.

(Updated) Test for individual coefficients: p-value

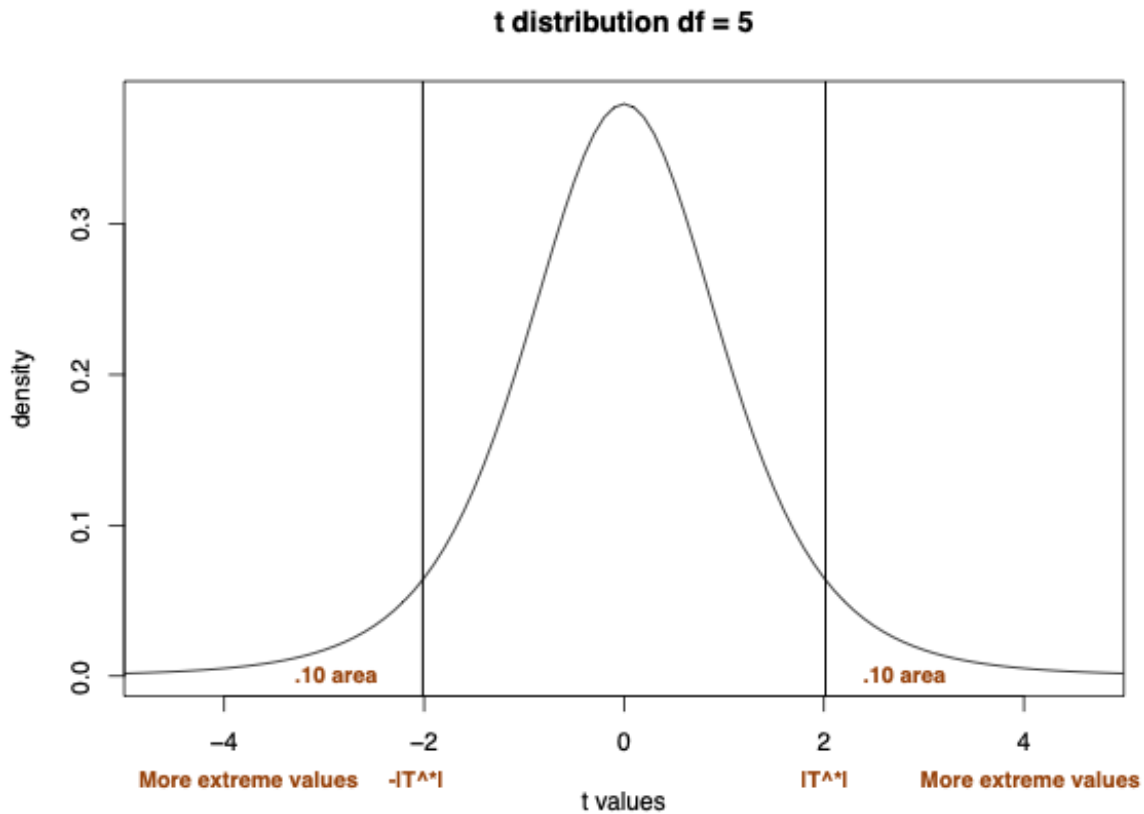
The p-value is the probability that T is as at least as extreme or more extreme than the observed test statistic:

$$\begin{aligned}\text{p-value} &= P(|T| \geq |T^*|) \\ &= P(T \geq |T^*|) + P(T \leq -|T^*|) \\ &= 2P(T \geq |T^*|).\end{aligned}$$

If the p-value is smaller than α, then this is evidence against the null hypothesis H_0.

(Updated) p-value example

Here the p-value for the two-sided hypothesis test is $.20$.



Test for individual coefficients: conclusions

- If the p-value is smaller than α:
  Performing the hypothesis test $H_0 : \beta_i = \beta_i^0$ versus the alternative $H_A : \beta_i \neq \beta_i^0$ at significance level α, we reject the null hypothesis and conclude there is statistical evidence that $\beta_i \neq \beta_i^0$.
- If the p-value is larger than α:
  Performing the hypothesis test $H_0 : \beta_i = \beta_i^0$ versus $H_A : \beta_i \neq \beta_i^0$ at significance level α, we fail to reject the null hypothesis.

Hypothesis test example

Consider $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, $e \sim N(0, \sigma^2)$. We wish to test for a linear relationship between the response and x_1 with a level (allowed error) of $.05$. From the data, a regression is fit and the diagnostics checked. Then a p-value is computed $.002$.

Performing the hypothesis test $H_0 : \beta_1 = 0$ versus the alternative $H_A : \beta_1 \neq 0$ at significance level $.05$, the p-value is $.002$ and we reject the null hypothesis. We conclude there is statistical evidence of a linear relationship between the average response and the predictor x_1.

Simple linear regression: worked example

```
load("module6.Rdata")
x = data1$x
y = data1$y

# Compute least squares estimates
s_xx = sum( (x - mean(x))^2 )
x_bar = mean(x)
y_bar = mean(y)
beta_1_hat = sum((y - y_bar) * (x - x_bar)) / s_xx
beta_0_hat = y_bar - beta_1_hat * x_bar

# Compute estimate to sigma2
resid = y - beta_0_hat - beta_1_hat * x
RSS = sum( resid^2 )
sigma2_hat = RSS / 98

# Compute estimated standard errors
se_beta_1 = sqrt( sigma2_hat/s_xx )
se_beta_0 = sqrt( sigma2_hat/100 + sigma2_hat * x_bar^2 /s_xx )

# Compute test statistics and p-values
test_stat1 = beta_1_hat / se_beta_1
2 * pt(abs(test_stat1), df = 98, lower.tail=FALSE)

test_stat0 = beta_0_hat / se_beta_0
2 * pt(abs(test_stat0), df = 98, lower.tail=FALSE)
```

Simple linear regression: worked example with R

Compare the previous code output to:

```
fit = lm(y ~ x, data = data1)
summary(fit)
```

Multiple linear regression: worked example

```
load("module6.Rdata")
fit = lm(y ~ group + x2, data = data2)
summary(fit)
```

- How would we interpret dropping an indicator variable?

Lecture 2: F tests

Learning objectives

- Learn how to perform a hypothesis test comparing regression models using F tests.
- Understand the partial F test and overall F test
- Implement these in R

F-tests and t-tests

You may test if each coefficient is 0 with an error level α using a t-test and these errors will add up each test. The F-test is capable of testing if multiple variables are 0 at the same time (the hypothesis tests and conclusions are different however).

F distribution visualization

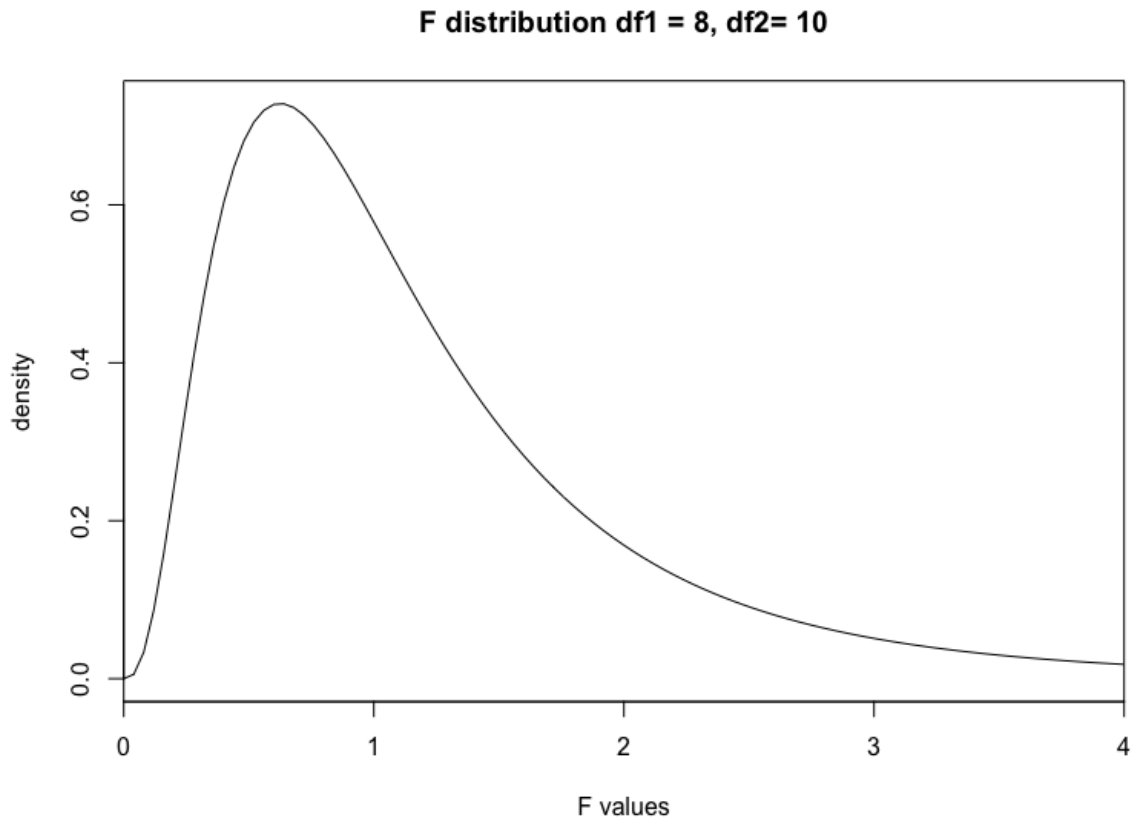
$F \sim F(df1, df2)$ is a random variable with $F(df1, df2)$ distribution with parameters $df1, df2 > 0$.

- F is non-negative
- The distributon can take many different shapes.
- It is a ratio of independent Chi-square distributions divided by their degree of freedom parameter

$$F = \frac{\chi^2_{df1}/df1}{\chi^2_{df2}/df2} \sim F(df1, df2).$$

F distribution example/visualization

Plot of $F \sim F(8, 10)$:



Partial F test

Consider we have fit a regression model and its diagnostics are acceptable. Compare two regression models

- **Full model:** Includes all p predictors.
- **Reduced model:** Includes only a subset of predictors of the full model's predictors. The **null model** is when the reduced model is only the intercept.

We can always reorder the predictors so that the reduced model uses the first r predictors: $\beta_1, \ldots, \beta_r$. This will simplify notation.

Partial F test

The reduced model here:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e, e \sim N(0, \sigma^2).$$

- **Null hypothesis:** $H_0 : \beta_{r+1} = \cdots = \beta_p = 0$
- **Alternative hypothesis:** H_A : At least one (possibly more) of $\beta_{r+1}, \ldots, \beta_p$ is nonzero.

The partial F hypothesis test in words:

- **Null hypothesis:** the reduced model is sufficient
- **Alternative hypothesis:** the reduced model is not sufficient

Difference between F tests

- **Partial F test:** the reduced model can be any subset of predictors from the full model.
- **Overall F test:** the reduced model is the null model which contains only the intercept.

The overall F test gives evidence that there is at least one nonzero coefficient in our model. We will see more motivation for this when introduce multicolinearity.

Partial F test: simple example

Consider a **full model**:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

and a possible **reduced model**

$$Y = \beta_0 + \beta_1 x_1 + e$$

so $\beta_2 = \beta_3 = 0$.

We wish to test the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ versus the alternative H_A : At least one of β_2, β_3 is nonzero.

Partial F test: full example

Consider a **full model**:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$$

and a possible **reduced model**

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e$$

so $\beta_{r+1} = \cdots = \beta_p = 0$.

Null hypothesis: $H_0 : \beta_{r+1} = \cdots = \beta_p = 0$
Alternative hypothesis: H_A : At least one of $\beta_{r+1}, \ldots, \beta_p$ is nonzero.

Partial F test: test statistic

- RSS_{full}: the residual sum of squares using the full model fit:

$$RSS_{full} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  where the fit uses the full model.
- $\text{df}_{full} = n - (p + 1)$
- $RSS_{reduced}$: the residual sum of squares using the reduced model fit
- SST: the sum of squares total is RSS_{null}

$$SST = RSS_{null} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

- $\text{df}_{reduced} = n - (r + 1)$
- MSR: The mean squares residual is $MSR = \text{RSS}_{full}/\text{df}_{full}$ is just $\hat{\sigma}^2$ for the full model

Partial F test: test statistic

- SS_{reg}: Sum of squares regression is

$$SS_{reg} = RSS_{reduced} - RSS_{full}$$

- df_{reg}: # of predictors not in the reduced model. It is the difference $\text{df}_{reduced} - \text{df}_{full} = (n-r-1) - [n-p-1] = p-r$.
- MS_{reg}: the mean squares regression is SS_{reg}/df_{reg}

Partial F test: test statistic

If the reduced model has a much larger RSS than the full model, then the full model "fits the data better". This gives us evidence against the reduced model. In other words, if

$$SS_{reg} = RSS_{reduced} - RSS_{full}$$

is large.

Question: How large is too large?

Partial F test: test statistic sampling distribution

Under the regression model, the random test statistic is the **ratio of mean sums of squares** with sampling distribution given by

$$
\begin{aligned}
F &= \frac{(\mathsf{RSS}_{reduced} - \mathsf{RSS}_{full})/(\mathsf{df}_{reduced} - \mathsf{df}_{full})}{\mathsf{RSS}_{full}/\mathsf{df}_{full}} \\
&= \frac{\mathsf{SS}_{reg}}{\hat{\sigma}^2(p-r)} \\
&= \frac{MS_{reg}}{MSR} \\
&\sim F(p-r, n-p-1).
\end{aligned}
$$

(Updated) Partial F test: p-value

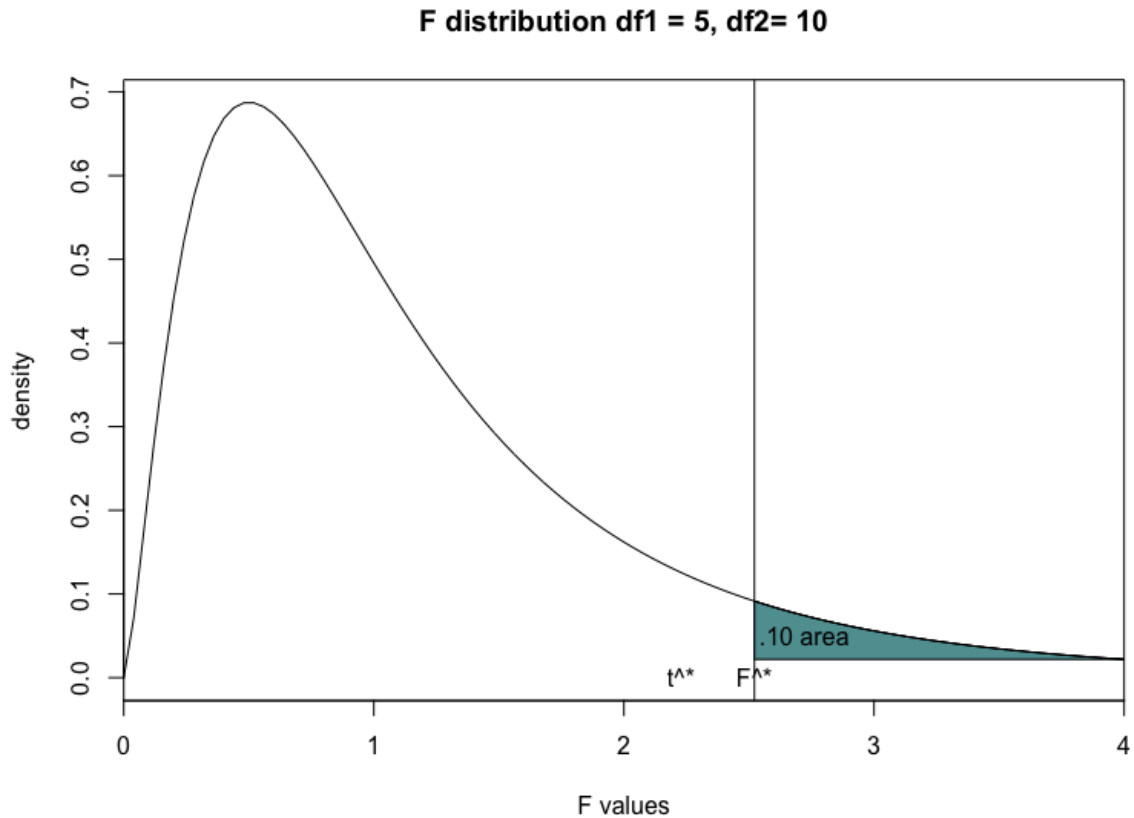
The p-value is the probability that F is as at least as large or larger (extreme or more extreme) than observed test statistic F^* from the data:

$$\text{p-value} = P(F \geq F^*)$$

If the p-value less than the level of the hypothesis test α, then we have evidence against the null hypothesis and so against the reduced model.

p-value visualization

Here the p-value is .10.



Partial F test: conclusions

- If the p-value is smaller than the level:
  At significance level α, we reject the null hypothesis and conclude there is statistical evidence that at least one of $\beta_{r+1}, \ldots, \beta_p$ are nonzero. There is statistical evidence of a significant linear relationship between the average response and at least one of the predictors $x_{r+1}, \ldots, x_p$.
- If the p-value is larger than the level: we fail to reject the null hypothesis.

Partial F test: procedure

1. Fit a full regression model
2. Check assumptions via diagnostics
3. Determine a reduced regression model to test
4. Specify the null hypothesis and alternative hypothesis and level
5. Compute the test statistic value and compute the p-value
6. Draw conclusions

Lecture 3: Implementing F tests

Learning objectives

- Implement partial and overall F tests using R

Lecture 3: Considerations

Considerations

General (rough) guideline of tests in an analysis:

- F test / t-test / confidence interval -> chosen model -> report confidence interval for the coefficients for the chosen model

Always use diagnostics to verify modeling assumptions before using a hypothesis test.

- The results rely on the regression model assumptions and if these fail then the result is unreliable.
- We generally check the full model for the assumptions and then perform tests on submodels.

Caution when using partial F test:

- Cannot compare models from different datasets
- Cannot compare models that are not subsets of one another

Lecture 3: F test for simple linear regression

Simple linear regression: worked example

$H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ at level $\alpha = .05$

Question: Which F test is this?

Simple linear regression: worked example

```
x = data1$x
y = data1$y

# Compute least squares estimates
s_xx = sum( (x - mean(x))^2 )
x_bar = mean(x)
y_bar = mean(y)
beta_1_hat = sum((y - y_bar) * (x - x_bar)) / s_xx
beta_0_hat = y_bar - beta_1_hat * x_bar

resid_full = y - beta_0_hat - beta_1_hat * x
RSS_full = sum( resid_full^2 )

resid_null = y - y_bar
RSS_null = sum( resid_null^2 )

SS_reg = RSS_null - RSS_full

MS_reg = SS_reg/1
MSR = RSS_full/98

test_stat = MS_reg/MSR
pf(test_stat, df1 = 1, df2 = 98, lower.tail=FALSE)
```

Question: What is the conclusion?

Simple linear regression: worked example

Compare this to using summary output. Also, compare to the t-test.

```
fit = lm(y ~ x, data = data1)
summary(fit)
```

Simple linear regression: worked example

```
fit_full = lm(y ~ x)
fit_null = lm(y ~ 1)

anova(fit_null, fit_full)
```

Question: What does each part of this output mean?

Simple linear regression: worked example

The t test and F test are equivalent when you drop 1 predictor.

```
summary(fit_full)
```

Lecture 3: F test for multiple linear regression

Multiple linear regression: worked example

```
fit_full = lm(y ~ x1 + group + x2 + x1:x2, data = data2)
fit_reduced = lm(y ~ x1 + group + x2, data = data2)
anova(fit_reduced, fit_full)
```

Multiple linear regression: worked example

Compare this to the summary output:

```
summary(fit_full)
```

The t test and F test are equivalent when you drop 1 predictor.

Lecture 3: activity

Multiple linear regression: worked example

Activity: use the partial F test to statistical evidence for the true population relationship.

```
load("module6.Rdata")
head(data2)
```

Module takeaways

1. What do the components of the sum of squares decomposition measure?
2. How are these components measured/computed?
3. What is the difference between the hypothesis for the ANOVA test and the Partial F test?
4. How is the decomposition used differently for the ANOVA test and the Partial F test?
5. How do we conduct each decomposition-based test?
6. What are the conclusions of each decomposition-based test?

References I