STAT302 Methods of Data Analysis 1
Module 7: Decomposition of Variance 2

Austin Brown

University of Toronto, Department of Statistical Sciences

October 24, 2024

Announcements/Review

- Reading week
- Project Part 2 and NQA extension deadline.
- Review last lecture

Model selection with F-tests outline

1. Full model
2. Check assumptions on the full model
3. F-tests/t-tests on reduced models until there is no evidence to drop more predictors
4. Conclude with a final model
5. Check assumptions on the final model
6. Finally, use the final model: confidence intervals, interpretation using the estimated coefficients, and prediction.

Overview

Why use the F-test and not just t-test? This lecture covers some more motivation for the F-test over the t-test.

Lecture 1: overall F test and the coefficient of determination

Learning objectives

- Decomposition of variance
- Overall F test
- Coefficient of determination R^2 and adjusted version

Lecture 1: overall F test

Motivating example

Should we drop x_1 and x_2 with the t-tests?

```
> load("module7.Rdata")

> fit = lm(y ~ x1 + x2, data = data)
> summary(fit)

Call:
lm(formula = y ~ x1 + x2, data = data)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3521     0.5667  -2.386    0.019 *
x1            0.8446     0.5654   1.494    0.139
x2            1.2097     1.1337   1.067    0.289
---

Residual standard error: 0.1056 on 97 degrees of freedom
Multiple R-squared:  0.6065,  Adjusted R-squared:  0.5984
F-statistic: 74.76 on 2 and 97 DF,  p-value: < 2.2e-16
```

Lecture 1: ANOVA tables for overall F-tests

Multiple linear regression: overall F test

- **Null hypothesis:** $H_0 : \beta_1 = \cdots = \beta_p = 0$
- **Alternative hypothesis:** H_A : At least one of $\beta_1, \ldots, \beta_p$ is nonzero.

Rejecting the null hypothesis provides evidence for the current model.

ANOVA tables

There are 3 types of ANOVA tables for multiple linear regression and we will not discuss them further: ANOVA type 1/2/3. This is a topic for later courses in design of experiments. We will do things the long way instead and will cover the ANOVA table for the overall F-test.

Simple linear regression: Overall F test ANOVA table

ANOVA table for the overall F-test.

Source	**Df**	**Sum squares**	**Mean squares**	**F value**
Regression	1	SS_{reg}	$MS_{reg} = SS_{reg}/1$	MS_{reg}/MSR
Residual	$n-2$	RSS	$MSR = RSS/(n-2)$	
Total	n - 1	SST		

What are these values here:

$$RSS =$$

$$SS_{reg} =$$

$$SST =$$

Multiple linear regression: ANOVA table

ANOVA table for the overall F-test.

Source	**Df**	**Sum squares**	**Mean squares**	**F value**
Regression	p	SS_{reg}	$MS_{reg} = SS_{reg}/p$	MS_{reg}/MSR
Residual	$n-p-1$	RSS	$MSR = RSS/(n-p-1)$	
Total	n - 1	SST		

Simple least squares: decomposition of sums of squares

$$SST = \sum_i (y_i - \overline{y})^2$$

$$=$$

$$=$$

$$= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \overline{y})^2$$

$$= RSS + SS_{reg}.$$

Simple regression: decomposition of sums of squares

Recall from the least squares solutions that $\sum_i (y_i - \hat{y}_i)(x_i - \bar{x}) = 0$. So

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \hat{\beta}_1 \sum_i (y_i - \hat{y}_i)(x_i - \overline{x}) = 0.$$

So then we showed that

$$SS_{reg} = SST - RSS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2.$$

Multiple regression: decomposition of sums of squares

We can also show that in MLR,

$$SST = RSS + SS_{reg}$$

and

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2.$$

In terms of degrees of freedom, we have $n - 1 = n - p - 1 + p$.

Multiple regression: decomposition of sums of squares

Lecture 1: coefficient of determination

New concept: model scoring

A new concept this lecture is to give a regression model a "score". Motivation can be to consider two people analyze the same data set and arrive at two different fits.

- Which model is better?
- Can be used to select a model
- RSS can be seen as a score, R^2 and R^2_{adj} are also ways to score a model

Question: What is wrong with using the RSS?

Coefficient of determination

From the SST decomposition

$$SST = SS_{reg} + RSS$$
$$1 = \frac{SS_{reg}}{SST} + \frac{RSS}{SST}$$
$$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{RSS}{SST}$$

The coefficient of determination R^2 is a measure of goodness of the model.

- Measures the strength of the relationship between the response and the predictors.
- $0 \leq R^2 \leq 1$
- It is scale-free (independent of units) due to dividing by SST.
- Complex models can have a large R^2 even if the predictors are not statistically significant.

Coefficient of determination

If the regression fits the data well, the RSS is small. Then also

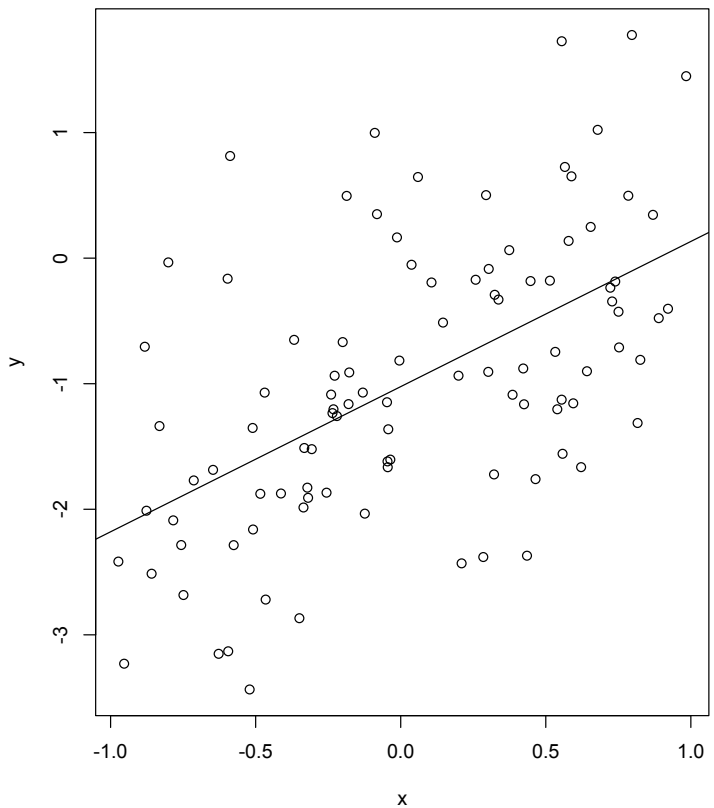
$$0 \leq 1 - \frac{RSS}{SST} \leq 1$$

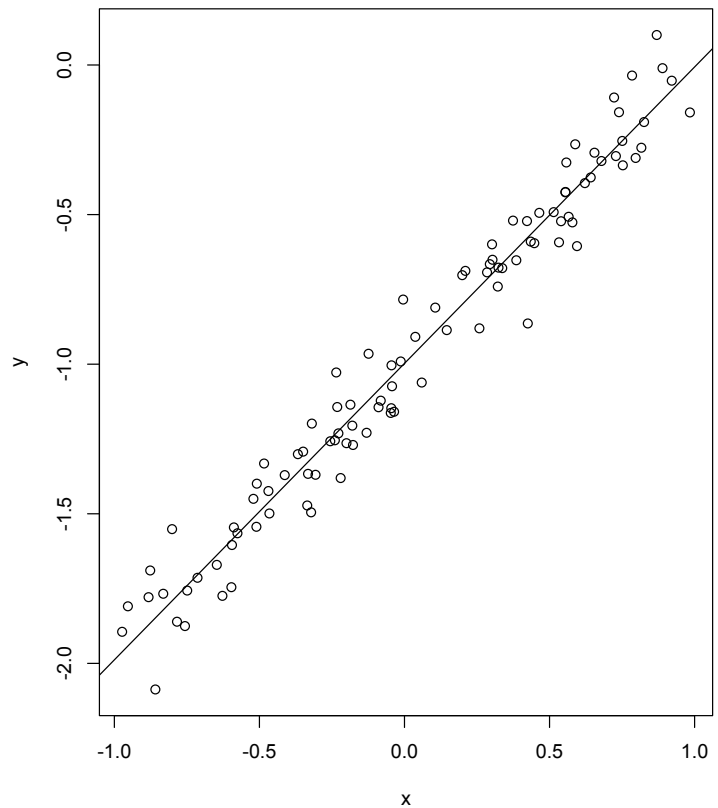
is large.

For noisy data, this may not be close to 1.

Example: coefficient of determination

Which one has higher R^2?





New concept: penalize model complexity

While complex models may have smaller RSS, they can also be more difficult to interpret. We can penalize complex models more and create a new score.

Adjusted coefficient of determination

The coefficient of determination R^2_{adj} is a measure of goodness of the model.

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

- Can be used to compare between two regression models. Generally better methods are available.
- Lacks a simple interpretation like R^2
- Penalizes complexity of the model

R example

```
load("module7.Rdata")

x1 = data$x1
x2 = data$x2
y = data$y

fit = lm(y ~ x1 + x2)
summary(fit)

# Compute R^2
RSS = sum( resid(fit)^2 )
SST = sum( (y - mean(y))^2 )
R2 = 1 - RSS/SST

# Compute overall F statistic
n = length(y)
sigma2_hat = RSS / (n - 3)
SS_reg = SST - RSS
F_stat = SS_reg / (2 * sigma2_hat)
```

R example

- Is the R^2 good here?
- Why is that this F test is rejecting but the t tests are not?

Lecture 2: relationships between predictors

Learning objectives

- Colinearity of predictors
- Multicolinearity of predictors
- Drawbacks for F tests for model selection
- Exploratory data analysis: scatterplot matrix, sample correlation

Recall

What can go wrong with our interpretation of the "rate of change" interpretation of the fitted coefficients to our predictors in polynomial regression and with the existence of interaction terms?

Lecture 2: collinearity and multicollinearity

Collinearity and multicollinearity of predictors

Consider we have some predictors in our dataset $X_1 = x_1, X_1 = x_2, X_1 = x_3$. We can think of the predictors as random say X_1, X_2, X_3 and $X_1 = x_1$ means we observe the random predictor at value x_1. Ideally the predictors are uncorrelated. However, we can have these properties:

collinearity exists when two predictors say X_1 and X_2 are highly correlated (highly linearly related).

multicollinearity is when more than two predictors say X_1, X_2, and X_3 are highly correlated (highly linearly related).

Multicollinearity is often used/said to mean 2 or more for simplicity.

Loss of interpretation

We lose our nice interpretation with highly correlated predictors.
We cannot really fix all other predictors and only increase 1.

Prediction and inference paradigms

If our only goal is prediction, we might not really care if the model is approximating the population model or if we can interpret it. We might just want to predict as best as we can. Multicollinearity is as much of a concern here.

For inference and interpretation of the coefficients of the predictors, this is a different story. Highly correlated predictors is something that can effect both of these.

Regression on predictors

We can think of the predictors as random X_1, X_2, X_3. The response

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

with $\epsilon \sim N(0, \sigma^2)$ is random given **fixed** predictor values $X_1 = x_1, X_2 = x_2, X_3 = x_3$.

We can also think of X_3 being a random response given other fixed predictor values and regress in this way:

$$X_3 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \epsilon'$$

with normal error ϵ'.

Regression on predictors

Consider

$$X_3 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \epsilon'$$

with normal error ϵ'.

- What if we perform a regression analysis here and conclude there is evidence $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$.
- What if the R^2 is large?

Lecture 2: assessing correlation among predictors

Multicollinearity: assessment

Roughly, we are going to identify multicollinearity by doing regression analysis on a predictor against other predictors.

Assessing highly correlated predictors: Exploratory data analysis

- Compute sample correlation.
- Look at the Scatterplot matrix
  - can only assess collinearity
- Use lm to fit of one predictor against another.
  - can assess multicollinearity with R^2.

Regression analysis

Fit a regression of predictor j against all other predictors:

$$X_j = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + e$$

where $e \sim N(0, \nu^2)$ and x_j is not included as a predictor.

- Define and compute R_j^2 is the R^2 here.
- Continue with a standard regression analysis (Model diagnostics, etc.).

Example

```
pairs(data)

fit2 = lm(x2 ~ x1)

par(mfrow = c(2, 2))
plot(fit2)

summary(fit2)
```

Lecture 2: types of relationships between predictors

Types of relationships between predictors

- **Structural:** polynomial regression, interaction terms
- **Data-based:** observational data, poorly designed experiment, etc.

Structural relationships

Often it is unwanted correlation between predictors. Other times it is implemented purposefully such as interaction terms and polynomial terms.

- Polynomial terms:

$$\hat{y} = -1 + .5x - 1.5x^2$$

- Interaction terms:

$$\hat{y} = -1 - .5x_1 + .5x_2 - .5x_1x_2$$

Structural relationships

We lose the nice interpretation. Can we fix x_2 and x_1x_2 and only increase x_1?

$$\hat{y} = -1 - .5x_1 + .5x_2 - .5x_1x_2$$

We can no longer fix all other variables due to the high correlation and so we do not have the same interpretation.

Collinearity: exact linear releationship

Exact linear relationship

$$X_2 = \alpha_0 + \alpha_1 X_1$$

- Can cause $X^T X$ to not be invertible and then there are not unique minimizers to to the residual sum of squares.
- Can cause instability in $(X^T X)^{-1}$ and so the variance of $\hat{\beta}_i$ can become very large.

Collinearity: approximate linear relationships

Not precisely linear but say a simple linear regression model between them:

$$E(X_2|X_1 = x) = \alpha_0 + \alpha_1 x$$

and

$$\text{Var}(X_2|X_1 = x) = \nu^2$$

Multicollinearity: approximate linear relationships

Not precisely linear but say a simple multiple regression model between them:

$$E(X_3|X_1 = x_1, \cdots, X_p = x_p) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

and

$$\mathsf{Var}(X_3|X_1 = x_1, X_2 = x_2) = \nu^2$$

Multicollinearity: how to interpret?

You will need to discuss clearly and be careful of your interpretations in regression analyis with the presence of multicollinearity amongst the predictors. This is the case when the correlation is very high. We have seen this already in our interpretations with interaction terms.

Lecture 2: activity

Back to menu pricing for a new restaurant in NYC

From [Sheather, 2009]:
Imagine that you have been asked to join the team supporting a new chef who plans to create a new Italian restaurant in Manhattan. The aims of the restaurant are to provide the highest quality Italian food utilizing state-of-the-art décor while setting a new standard for high-quality service in Manhattan. You have been told that the restaurant is going to be located no further south than the Flatiron District and it will be either east or west of Fifth Avenue.

You have been asked to determine the pricing of the restaurants dinner menu such that it is competitively positioned with other high-end Italian restaurants in the target area.

Activity

- Identify the restaurants in the data set which, given the customer ratings, are (i) unusually highly priced; and (ii) unusually lowly priced.
- Explore relationships between the predictors (collinearity and multicollinearity possibilities).
- Which predictor variables should we be careful with our interpretations?

Lecture 3: assessing and addressing correlation in the predictors

Learning objectives

- Addressing collinearity and multicollinearity
- What to do
- Variance inflation factors
- Real-data example

Overview

We learned basic model selection with hypothesis testing. Now we will learn some flaws in this approach when the predictors are highly correlated and what to do. Later we will learn some more modern approaches to model selection.

What to do with the presence of multicollinearity

What to do if multicollinearity is identifited.

- **Identify and discuss.** Always discuss the limitations of your regression analysis if evidence of multicollinearity is identified.
- **Gather more data.** With more data, the estimation is better. However, this is almost always impossible to do.
- **Do nothing.** The existence of multi/collinearity does not necessarily mean the variances of $\hat{\beta}_i$ are inflated. $\hat{\beta}$ is still unbiased, but need to be very careful when removing predictors.
- **Drop a predictor.** Drop one of the multi/collinear variables. This can cause $\hat{\beta}$ to be biased if the predictor is incorrectly dropped.
- **Change form/combine predictors.** Can center/combine predictors like $\boldsymbol{x} - \overline{\boldsymbol{x}}$ but interpretation can become tricky.
- **Other techniques.** Beyond the scope of the course.

Back to our motivating example

Let us figure out what the problem is with this below example.

```
> load("module7.Rdata")

> fit = lm(y ~ x1 + x2, data = data)
> summary(fit)

Call:
lm(formula = y ~ x1 + x2, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3521     0.5667  -2.386    0.019 *
x1            0.8446     0.5654   1.494    0.139
x2            1.2097     1.1337   1.067    0.289
---

Residual standard error: 0.1056 on 97 degrees of freedom
Multiple R-squared:  0.6065,  Adjusted R-squared:  0.5984
F-statistic: 74.76 on 2 and 97 DF,  p-value: < 2.2e-16
```

Lecture 3: variance inflation factors

Inflation of the variance

Since

$$\text{Var}(\hat{\beta}_j) = \sigma^2 (X^T X)^{-1}_{j+1,j+1}$$

then multicollinearity can cause instability in $(X^T X)^{-1}$ and due to this, the variance can become large.

VIF

Say we have X_1, X_2, X_3 continuous predictors. Set up a regression model

$$E(X_3|\{\text{other predictors}\}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$$

Compute the R_3^2 for this regression model.

If R_3^2 is large then there is evidence of multicollinarity.

VIF

It can be shown that

$$\text{Var}(\hat{\beta}_j) \propto \frac{\sigma^2}{1 - R_j^2}$$

So if R_j^2 is large, then the variance of individual $\hat{\beta}_j$ can inflate.

- A large variance pulls the test statistic for the t-tests towards 0 and can cause the t-tests to fail to reject.
- Causes the confidence intervals to be large and the marginal t-tests to fail.

VIF

Define the **variance inflation factor**

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R_j^2}$$

The folklore is that is to use a cutoff of something like $VIF > 5$.

Implementations in R

The first way is to simply use the lm and compute the VIF. There is a function in the package "car". This also uses a "generalized VIF" that can handle categorical variables.

```
install.packages("car") # Installs the package

library(car) # load the packge

vif(fit) # Compute the VIF/GVIF for all predictors at the same time
```

Lecture 3: challenges in hypothesis testing

Issues with multicollinearity

Suppose the population follows

$$Y = .5 + 1x_1 - 2x_2 + \epsilon$$

and so $\beta_1 > 0$ and at first you might think that there is a positive linera relationship between the (average) response and x_1. But say $X_1 = x_1$ and $X_2 = x_2 = x_1 + \epsilon'$, then

$$\begin{aligned} Y &= .5 + 1x - 2(1x + \epsilon') + \epsilon \\ &= (.5 - 2\epsilon') - 1x_1 + \epsilon \end{aligned}$$

- Y and x_1 can be strongly negatively associated even when $\beta_1 > 0$ due to strong multicollinearity.

Back to our motivating example

This is precisely what is happening in our example! The F-test picks up that at least one of the coefficients is non-zero but it does not know which one. This is because it tests both coefficients at the same time. However, the t-test does not pick this up as it only tests a single coefficient at a time.

Challenges with collinearity and multicollinearity

- The variance of the estimated coefficients can inflate or become large. So the estimated coefficients are less reliable.
- Hypothesis tests can fail. Predictors may fail the marginal/individual hypothesis tests. It may lead to incorrectly dropping a predictor from the model.
- Estimated coefficients can have unintuitive/incorrect signs (we saw this with interaction terms).
- We lose the nice interpretation of the estimated coefficients because we are unable to fix all other predictors while only increasing one predictor.

Challenges in hypothesis testing

- Dropping variables can be unreliable when there is collinearity or multicollinearity of predictors
- Just because you fail to reject the hypothesis test does not necessarily mean you will remove the predictor.
- If the hypothesis test is unreliable and you remove a predictor that should actually be included, then the least squares estimators can become biased.

Lecture 3: observational studies

Issues with multicollinearity

Examples:

- The variable was not directly measured or collected in an observational study
- The experiment was not designed/controlled properly
- A variable is dropped even though we should have included it

Issues with a observational studies

Read Sheather's examples

Lecture 3: activity

Activity: menu pricing in NYC

- Choose a model with hypothesis testing
- Determine which of the predictor variables Food, Décor and Service has the largest estimated effect on Price? Is this effect also the most statistically significant?
- If the aim is to choose the location of the restaurant so that the price achieved for dinner is maximized, should the new restaurant be on the east or west of Fifth Avenue?
- Does it seem possible to achieve a price premium for "setting a new standard for high-quality service in Manhattan" for Italian restaurants?
- The new chef is considering going with a cheaper restaurant design interior but high quality service. Use your model to predict provide a menu price for food, decor, and service rating of 25, 25, 25 and 25, 15, 25. Provide your recommendation.

Module takeaways

1. How are both coefficients of determination computed?
2. What is the difference between the two coefficients and why do we need two?
3. How are the coefficients of determination interpreted?
4. What does the VIF measure and why is that useful for detecting multicollinearity?
5. How do we identify multicollinearity and why is it important to identify?

References I

Simon Sheather. *A modern approach to regression with R*. Springer Science & Business Media, 2009.