# Multifactor Studies w/o Replication

VARIANCE ESTIMATION

WHY WOULD DESIGNS HAVE NO REPLICATION

WHAT HYPOTHESES CAN STILL BE TESTED

# Background

# Replication is Key to Statistics: Sharing Strength

A key theme in statistics is to "share strength" across observations for the purpose of estimation/prediction

Instead of formally defining "sharing strength", here are some examples:

- We have observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. All of the $X_i = X$ (that is, all observations have the same "X" value). Then we can estimate $\mu\{Y|X\}$ via $\bar{Y}$
- We have observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. In MLR it is common that no two observations have the same "X" value. We can still estimate $\mu\{Y|X_i\}$ (that is, the mean of Y at $X_i$), but via a more complicated sample mean (e.g. Chapter 7.3.2)
- For one-way ANOVA, we estimate the overall variance by taking a weighted average of each of the variances of (e.g. Chapter 5.2.2)

# Replication is Key to Statistics: A Basic Example

A general observation about estimation: Suppose we have data from a distribution with:

unknown population mean $\mu$ and unknown population variance $\sigma^2$

Consider two scenarios

A. We have a single observation Y. We can estimate $\mu$ with $\hat{\mu} = Y$, $\sigma^2$ cannot be estimated

B. We have two observations $Y_1, Y_2$. If we additionally assume that these $Y_i$ have the same $\mu$ and $\sigma^2$, we can estimate $\mu$ with $\hat{\mu} = \frac{(Y_1+Y_2)}{2}$ and $\widehat{\sigma^2} = \sum_{i=1}^{2}(Y_i - \hat{\mu})^2$

→ The assumptions of equal mean and variance allows us to ”share strength”

We need at least two observations ”sharing strength” to estimate a variance

# Why Do We Need to Estimate the Variance?

Anytime we are trying to quantify uncertainty, we need an estimate of the variance

(well, technically, we just need an estimate of the variability)

**Examples:**

- Confidence Intervals: estimate $\pm\, \hat{\sigma} * (\text{quantile})$
- Testing hypotheses: $\text{Var}(\hat{\beta}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$, so we need an estimate of $\sigma^2$ in order to do inference in MLR

# Estimating the Variance in MLR w/o Replication

In MLR, it doesn't make sense to assume that each $Y_i$ has the same $\mu$

However, we can still "share strength" by assuming a model that encodes a principled way of combining observations even though they have different means

We can use all of the observations to estimate the *shared* model parameters:

- $\beta_0, \beta_1, \ldots, \beta_p$
- $\sigma^2$

even though perhaps no two observations have the same mean

Let's look at this in more detail..

# Estimating the Variance in MLR w/o Replication

As mentioned, there is often no replication with MLR

To estimate the variance, we instead do the following:

1. Take the $n$ observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ and estimate the model

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_p + \cdots + \beta_p x_p + \varepsilon$, where var(ε) = $\sigma^2$

which produces $\hat{\mu}\{Y|X\}$ via least squares

2. Estimate the variance via $\frac{1}{(n-p)} \sum_{i=1}^{n} (Y_i - \hat{\mu}\{Y_i|X_i\})^2$

Note that this requires that the variance estimate depends on the form for the model. Hence, it could be a poor estimate if…

- … the model doesn't capture the mean very well
- … the variances aren't constant as a function of $X$

# Estimating the Variance in MLR w/ Replication

Suppose instead that the $X_1$, …, $X_n$ are such that there is replication

(a previous example was the meat processing on slide 9 of 8b_linearRegressionDarren)

In this case, we can directly estimate the variance in a model-free way

The MLR model again:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_p + \cdots + \beta_p x_p + \varepsilon = \mu\{Y|X\} + \varepsilon$$

Instead of estimating $\mu\{Y|X\}$ via least squares, we can form an estimate via a restricted sample average (this is like basic example B.):

$$\check{\mu}\{Y|X\} = \frac{1}{n_X}\sum_{i=1}^{n} Y_i \, \mathbf{1}(X_i = X)$$

Where $n_X$ is the number of $X_i$ that equal $X$

# Estimating the Variance in MLR w/ Replication

Given the estimate of $\mu\{Y|X\}$ given by:

$$\check{\mu}\{Y|X\} = \frac{1}{n_X}\sum_{i=1}^{n} Y_i \, \mathbf{1}(X_i = X)$$

And the general model for the response:

$$Y = \mu\{Y|X\} + \varepsilon$$

We can estimate the variance without the linearity assumption nor the equal variances assumption:

$$\widehat{\sigma^2}(X) = \frac{1}{n_X - 1}\sum_{i=1}^{n}(Y_i - \check{\mu}\{Y|X_i\})^2 \, \mathbf{1}(X_i = X)$$

(of course, this is only defined at the original data $X_i$)

Alternatively, we can get an estimate of the variance without the linearity assumption but with the equal variances assumption by taking a weighted average of the $\widehat{\sigma^2}(X_i)$ with weights $n_{X_i}$

# Multifactor Studies w/o Replication

# Estimating the Variance in ANOVA w/o Replication

Due to the fact that ANOVA deals only with categorical explanatory variables (i.e. factors), without replication we cannot estimate $\sigma^2$

If there is no replication, there is a mean term in the model for every observation.

Hence, we are in a more complicated version of basic example A.:

A. We have a single observation $Y$. We can estimate $\mu$ with $\hat{\mu} = Y$, $\sigma^2$ cannot be estimated

In this case, it is impossible to estimate the variance unless some simplifying assumptions are made

# Why Would a Design Lack Replication?

Suppose we want to develop a new rocket engine for space craft

We want an engine that, for a fixed weight of fuel or metals, can fly the highest

We have narrowed down the options to

- Using one of two types of rocket fuels (F1,F2)
- Using one of two types of metal alloys for the shielding (A1, A2)

Our engineers cannot agree which of these factor levels will be best nor even if there might be an *interaction* between these factors

The only way to decide will be to make production-scale tests

However, production-scale tests for rockets are incredibly expensive

# Why Would a Design Lack Replication?

Using notation, we are looking to estimate

$$Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk},$$

Where we need to include the interaction term $\mu_{jk}$ due to the possibility that the metal alloy might react with the fuel type

Our company has just completed a round of investor funding and only has the money/time for 4 “test runs”. How can we allocate them in this experiment?

# Why Would a Design Lack Replication?

The four test runs could be allocated as:

1. Measure A1,F1 four times.

This would put us in the “one-sample” case from last semester. We could get a variance estimate, but we have no information about:

- A2
- nor F2
- nor the interaction between alloy and fuel

2. We could allocate 2 runs to A1,F1, and 2 runs to A2,F2

This would put us in the “two-sample” case from last semester.

We can test for differences between these two configurations, but have no information about A1,F2 nor A2,F1

# Why Would a Design Lack Replication?

The four test runs could be allocated as:

3. We could allocate 2 runs to A1,F1, 1 run to A2,F2, and 1 run to F1, A2

This does give us some information about 3 of the level combinations.

Under the additive model specification:

$$Y_{ijk} = \mu + \mu_j + \mu_k + \varepsilon_{ijk},$$

We can estimate each mean term with 1 left over degree of freedom each

Let's look at an example…

# Example

```
DATA example3;
    INPUT fuel $ alloy $ height;
    DATALINES;
    F1 A1 131
    F1 A1 122
    F2 A2 112
    F1 A2 117
    ;
```

No information about F2 A1 ($\mu_{21}$)

$$Y_{ijk} = \mu + \mu_j + \mu_k + \varepsilon_{ijk},$$

```
PROC GLM DATA=example3 PLOTS=ALL;
    CLASS fuel alloy;
    MODEL height = fuel alloy;
RUN;
```

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| fuel | 1 | 12.50000000 | 12.50000000 | 0.31 | 0.6772 |
| alloy | 1 | 60.16666667 | 60.16666667 | 1.49 | 0.4374 |

$$Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk},$$

```
PROC GLM DATA=example3 PLOTS=ALL;
    CLASS fuel alloy;
    MODEL height = fuel alloy fuel*alloy;
RUN;
```

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| fuel | 1 | 12.50000000 | 12.50000000 | 0.31 | 0.6772 |
| alloy | 1 | 60.16666667 | 60.16666667 | 1.49 | 0.4374 |
| fuel*alloy | 0 | 0.00000000 | . | . | . |

# Why Would a Design Lack Replication?

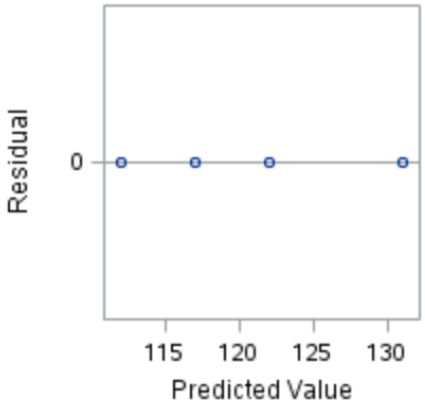
The four test runs could be allocated as:

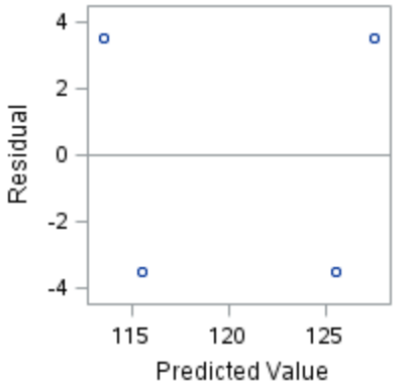
4. We could allocate 1 run to each combination

This does give us some information about all 4 of the level combinations.

Let’s look at an example…

# Example




```
DATA example4;
    INPUT fuel $ alloy $ height;
    DATALINES;
    F1 A1 131
    F1 A2 122
    F2 A1 112
    F2 A2 117
    ;
```

Information about all model parameters, but no variance estimate

$$Y_{ijk} = \mu + \mu_j + \mu_k + \varepsilon_{ijk},$$

```
PROC GLM DATA=example4 PLOTS=ALL;
    CLASS fuel alloy;
    MODEL height = fuel alloy;
RUN;
```

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| fuel | 1 | 144.0000000 | 144.0000000 | 2.94 | 0.3362 |
| alloy | 1 | 4.0000000 | 4.0000000 | 0.08 | 0.8228 |

$$Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk},$$

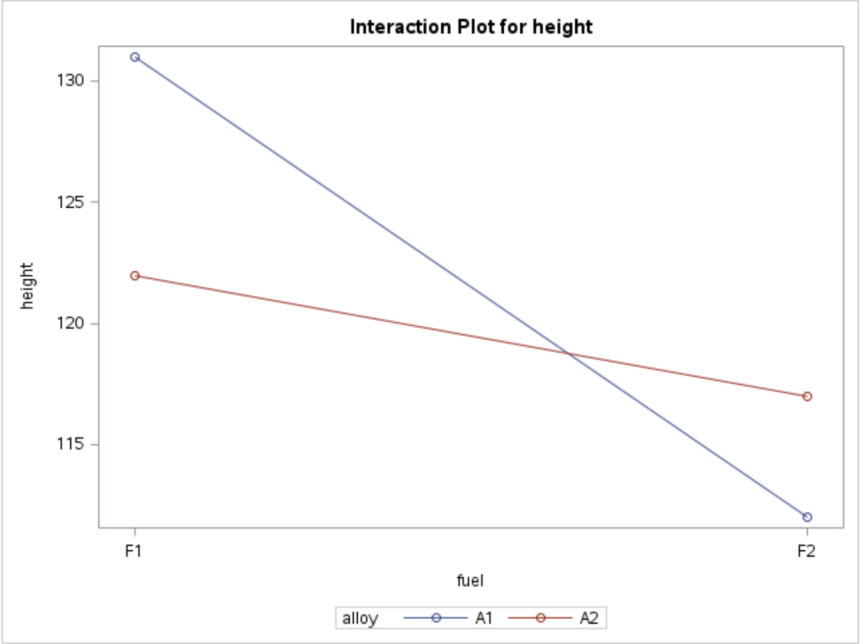
```
PROC GLM DATA=example4 PLOTS=ALL;
    CLASS fuel alloy;
    MODEL height = fuel alloy fuel*alloy;
RUN;
```

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| fuel | 1 | 144.0000000 | 144.0000000 | . | . |
| alloy | 1 | 4.0000000 | 4.0000000 | . | . |
| fuel*alloy | 1 | 49.0000000 | 49.0000000 | . | . |

# Example (Continued)

```
DATA example4;
    INPUT fuel $ alloy $ height;
    DATALINES;
    F1 A1 131
    F1 A2 122
    F2 A1 112
    F2 A2 117
    ;


PROC GLM DATA=example4 PLOTS=ALL;
    CLASS fuel alloy;
    MODEL height = fuel alloy fuel*alloy;
RUN;
```




We can still estimate the means, it is just with the observations themselves
Testing is impossible w/o variance estimate (as in basic example A.)

# Summary

In this example, cost/time considerations demand that we can only have 4 test runs

The most informative configuration would be 1 test run at each factor level combination

If we cannot rule out the interaction before running the analysis, then the best we can do is estimate each mean

If we have reason to believe that the additive model is sufficient, then the analysis can proceed as an additive/crossed analysis as before

(Note that we can actually still do testing in the interaction model if we have an auxiliary variance estimate)