Week 9: *Uncertainty*

EMSE 6035: Marketing Analytics for Design Decisions

John Paul Helveston

October 23, 2024

Pilot Analysis Report

Due 11/03 (that's 10 days from now)

Week 9: *Uncertainty*

1. Computing uncertainty

2. Reshaping data

BREAK

3. Cleaning pilot data

4. Estimating pilot data models

Week 9: *Uncertainty*

1. Computing uncertainty

2. Reshaping data

BREAK

3. Cleaning pilot data

4. Estimating pilot data models

Maximum likelihood estimation

$$\tilde{u}_j = \boldsymbol{\beta}'\mathbf{x}_j + \tilde{\varepsilon}_j$$

$$= \boxed{\beta_1} x_{j1} + \boxed{\beta_2} x_{j2} + \ldots + \tilde{\varepsilon}_j$$

Weights that denote the *relative* value of attributes $x_{j1}, x_{j2}, \ldots$

Estimate β_1 , β_2 , … , by minimizing the negative log-likelihood function:

$$\text{minimize} - \ln(\mathcal{L}) = -\sum_{j=1}^{J} y_j \ln\left[P_j(\boldsymbol{\beta}|\mathbf{x})\right]$$

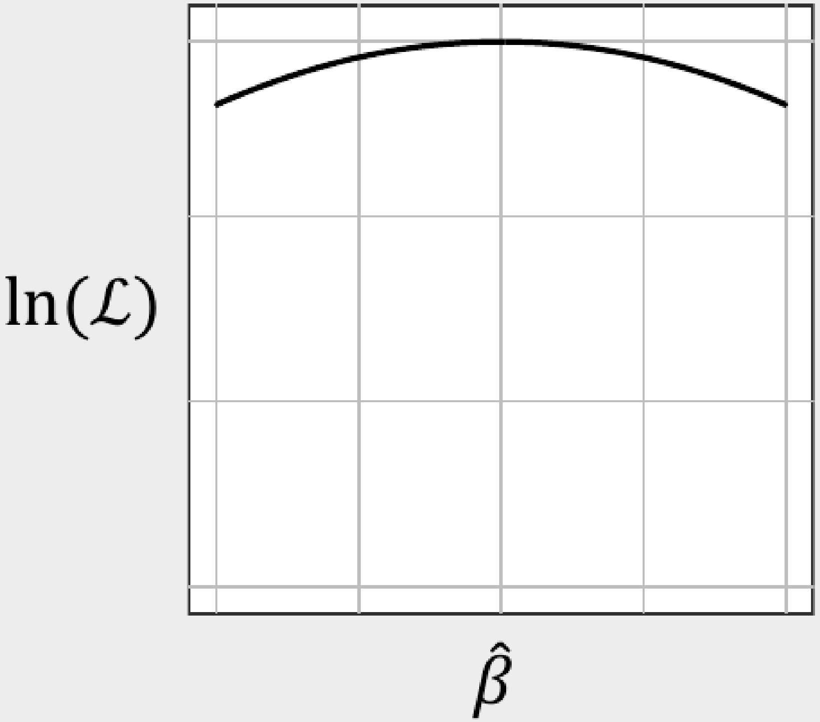
with respect to $\boldsymbol{\beta}$

$y_j = 1$ if alternative j was chosen

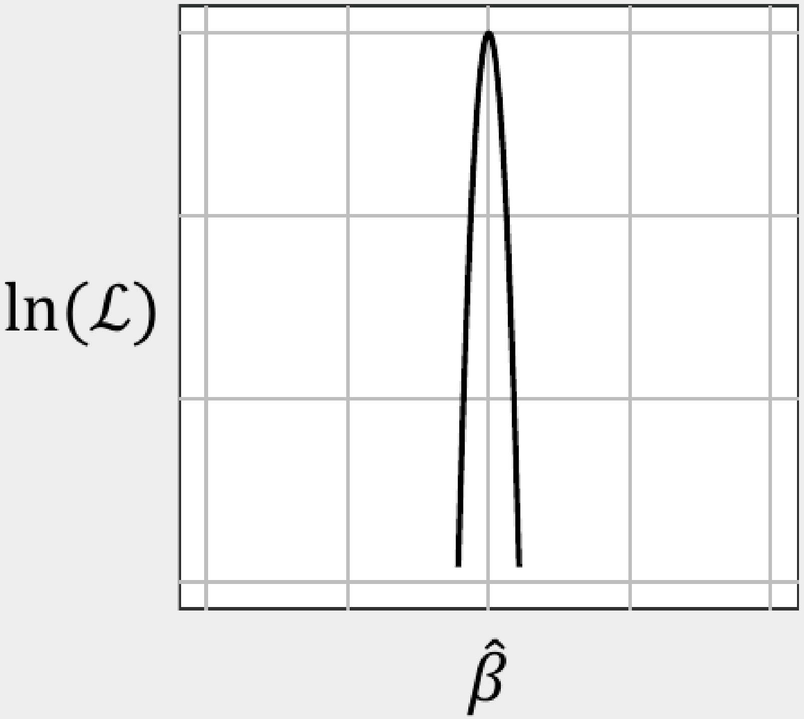
$y_j = 0$ if alternative j was not chosen

The certainty of $\widehat{\boldsymbol{\beta}}$ is inversely related to the curvature of the log-likelihood function

Greater variance in $\ln(\mathcal{L})$,
Less certainty in $\hat{\beta}$



Less variance in $\ln(\mathcal{L})$,
Greater certainty in $\hat{\beta}$



The *curvature* of the log-likelihood function is related to the hessian

$$\Sigma_{\boldsymbol{\beta}} = -\Big[\overbrace{\nabla^2_{\boldsymbol{\beta}} \ln(\mathcal{L})}^{\text{Hessian}}\Big]^{-1}$$

Covariance of $\hat{\boldsymbol{\beta}}$

The *curvature* of the log-likelihood function is related to the hessian

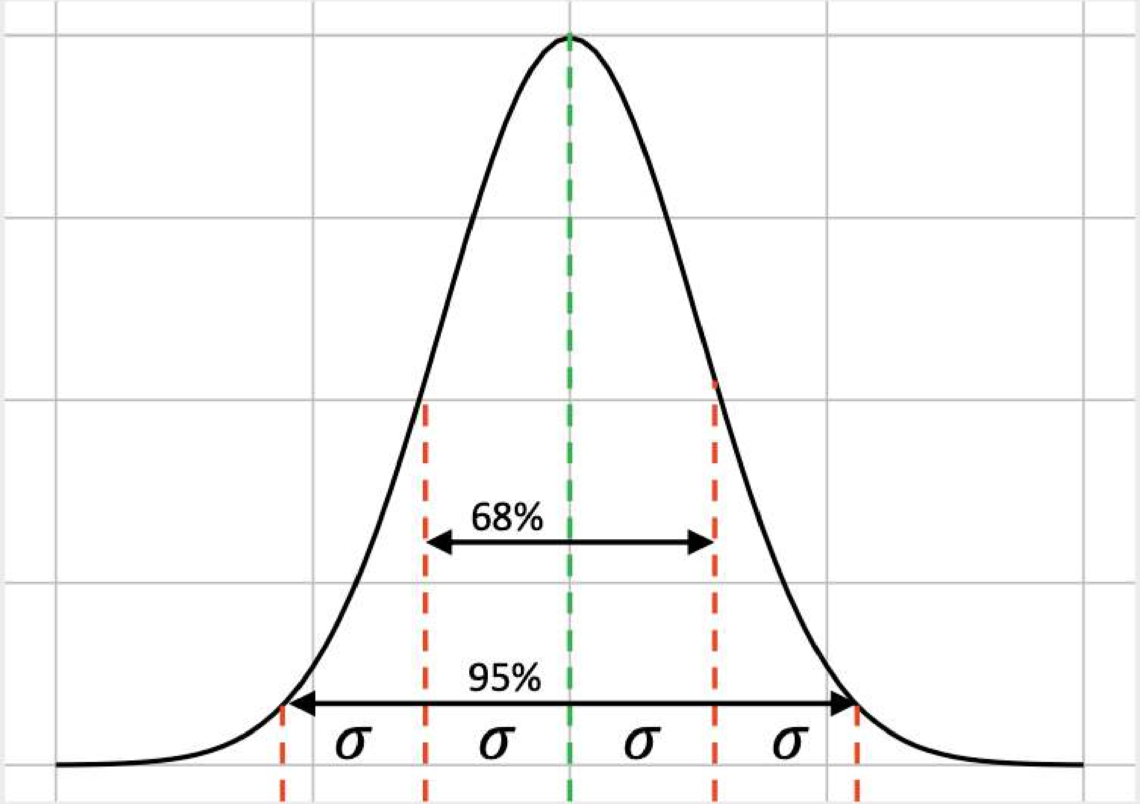
Hessian

$$\sum_{\boldsymbol{\beta}} = -\left[\nabla^2_{\boldsymbol{\beta}} \ln(\mathcal{L})\right]^{-1} = \begin{bmatrix} \sigma^2_{11} & \cdots & \sigma^2_{m1} \\ \vdots & \ddots & \vdots \\ \sigma^2_{1n} & \cdots & \sigma^2_{mn} \end{bmatrix}$$

Covariance of $\widehat{\boldsymbol{\beta}}$

Usually report parameter uncertainty ("standard errors") with σ values

Est.	Std. Err.
$\hat{\beta}_1$	σ_1
$\hat{\beta}_2$	σ_2
$\vdots$	$\vdots$
$\hat{\beta}_m$	σ_m



$\hat{\beta}$

A 95% confidence interval is approximately $[\hat{\beta} - 2\sigma, \hat{\beta} + 2\sigma]$

Practice Question 1

Suppose we estimate a model and get the following results:

$$\hat{\beta} = \begin{bmatrix} -0.4 \\ 0.5 \end{bmatrix}$$

$$\nabla^2_\beta \ln(\mathcal{L}) = \begin{bmatrix} -6000 & 60 \\ 60 & -700 \end{bmatrix}$$

a) Use the hessian to compute the standard errors for $\hat{\beta}$

b) Use the standard errors to compute a 95% confidence interval around $\hat{\beta}$

Simulating uncertainty

We can use the coefficients and hessian from a model to obtain draws that reflect parameter uncertainty

```
beta <- c(-0.7, 0.1, -4.0)

hessian <- matrix(c(
    -6000,   50,   60,
       50, -700,   50,
       60,   50, -300),
    ncol = 3, byrow = TRUE)
```

```
covariance <- -1*solve(hessian)
draws <- MASS::mvrnorm(10^5, beta, covarianc

head(draws)
```

```
#>            [,1]       [,2]      [,3]
#> [1,] -0.6906754 0.18627139 -3.971442
#> [2,] -0.6908079 0.22441160 -3.951020
#> [3,] -0.6957499 0.07853724 -3.980281
#> [4,] -0.7008129 0.06657027 -4.074518
#> [5,] -0.7034466 0.06253057 -4.014958
#> [6,] -0.6909423 0.08707439 -3.975708
```

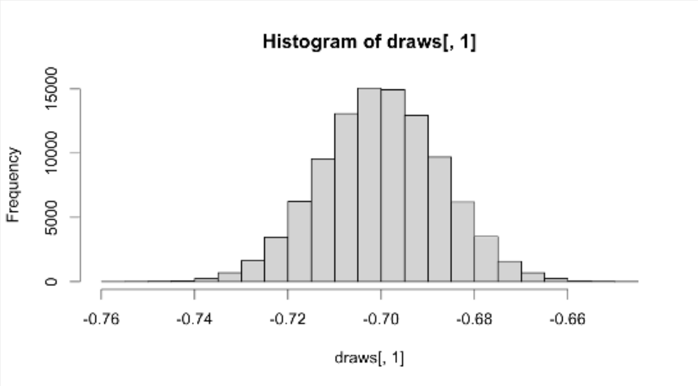
Simulating uncertainty

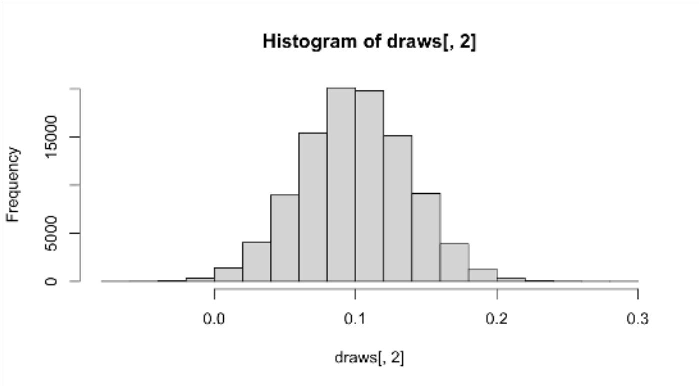
We can use the coefficients and hessian from a model to obtain draws that reflect parameter uncertainty

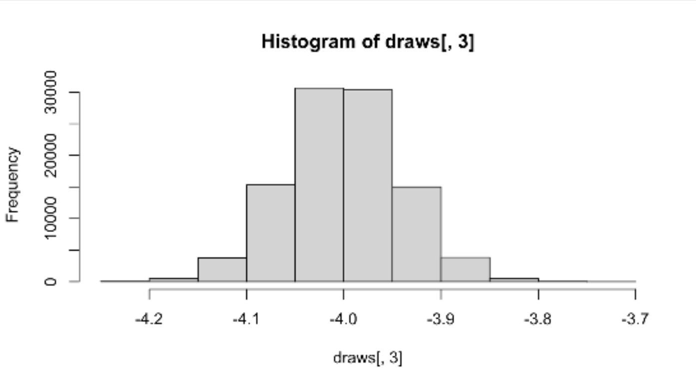
```r
hist(draws[, 1])
```

```r
hist(draws[, 2])
```

```r
hist(draws[, 3])
```







Practice Question 2

Suppose we estimate the following utility model describing preferences for cars:

$$u_j = \alpha p_j + \beta_1 x_j^{mpg} + \beta_2 x_j^{elec} + \varepsilon_j$$

a) Generate 10,000 draws of the model coefficients using the estimated coefficients and hessian. Use the `mvrnorm()` function from the `MASS` library.

b) Use the draws to compute the mean and 95% confidence intervals of each parameter estimate.

The estimated model produces the following results:

Parameter	Coefficient
α	-0.7
β_1	0.1
β_2	-0.4

Hessian:

$$\begin{bmatrix} -6000 & 50 & 60 \\ 50 & -700 & 50 \\ 60 & 50 & -300 \end{bmatrix}$$

Download the logitr-cars repo from GitHub

Computing and visualizing uncertainty

1. Open `logitr-cars`
2. Open `code/5.1-uncertainty.R`

Week 9: *Uncertainty*

1. Computing uncertainty

2. Reshaping data

BREAK

3. Cleaning pilot data

4. Estimating pilot data models

Federal R&D Spending by Department

```
#> # A tibble: 6 × 15
#>    year   DHS   DOC   DOD   DOE   DOT   EPA   HHS Interior  NASA   NIH   NSF Other  USD
#>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl
#> 1  1976     0   819 35696 10882  1142   968  9226     1152 12513  8025  2372  1191  183
#> 2  1977     0   837 37967 13741  1095   966  9507     1082 12553  8214  2395  1280  179
#> 3  1978     0   871 37022 15663  1156  1175 10533     1125 12516  8802  2446  1237  196
#> 4  1979     0   952 37174 15612  1004  1102 10127     1176 13079  9243  2404  2321  205
#> 5  1980     0   945 37005 15226  1048   903 10045     1082 13837  9093  2407  2468  188
#> 6  1981     0   829 41737 14798   978   901  9644      990 13276  8580  2300  1925  196
```

Federal R&D Spending by Department

"Wide" format

```
#> # A tibble: 6 × 15
#>    year   DHS   DOC   DOD   DOE   DOT   EPA   HHS Inte
#>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <
#> 1  1976     0   819 35696 10882  1142   968  9226
#> 2  1977     0   837 37967 13741  1095   966  9507
#> 3  1978     0   871 37022 15663  1156  1175 10533
#> 4  1979     0   952 37174 15612  1004  1102 10127
#> 5  1980     0   945 37005 15226  1048   903 10045
#> 6  1981     0   829 41737 14798   978   901  9644
```

"Long" format

```
#> # A tibble: 6 × 3
#>   department  year rd_budget_mil
#>   <chr>      <dbl>         <dbl>
#> 1 DOD         1976         35696
#> 2 NASA        1976         12513
#> 3 DOE         1976         10882
#> 4 HHS         1976          9226
#> 5 NIH         1976          8025
#> 6 NSF         1976          2372
```

Federal R&D Spending by Department

"Wide" format

```
#> # A tibble: 6 × 15
#>     year   DHS   DOC   DOD   DOE   DOT   EPA   HHS Inte
#>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <
#> 1  1976     0   819 35696 10882  1142   968  9226
#> 2  1977     0   837 37967 13741  1095   966  9507
#> 3  1978     0   871 37022 15663  1156  1175 10533
#> 4  1979     0   952 37174 15612  1004  1102 10127
#> 5  1980     0   945 37005 15226  1048   903 10045
#> 6  1981     0   829 41737 14798   978   901  9644
```

```
#> [1] 42 15
```

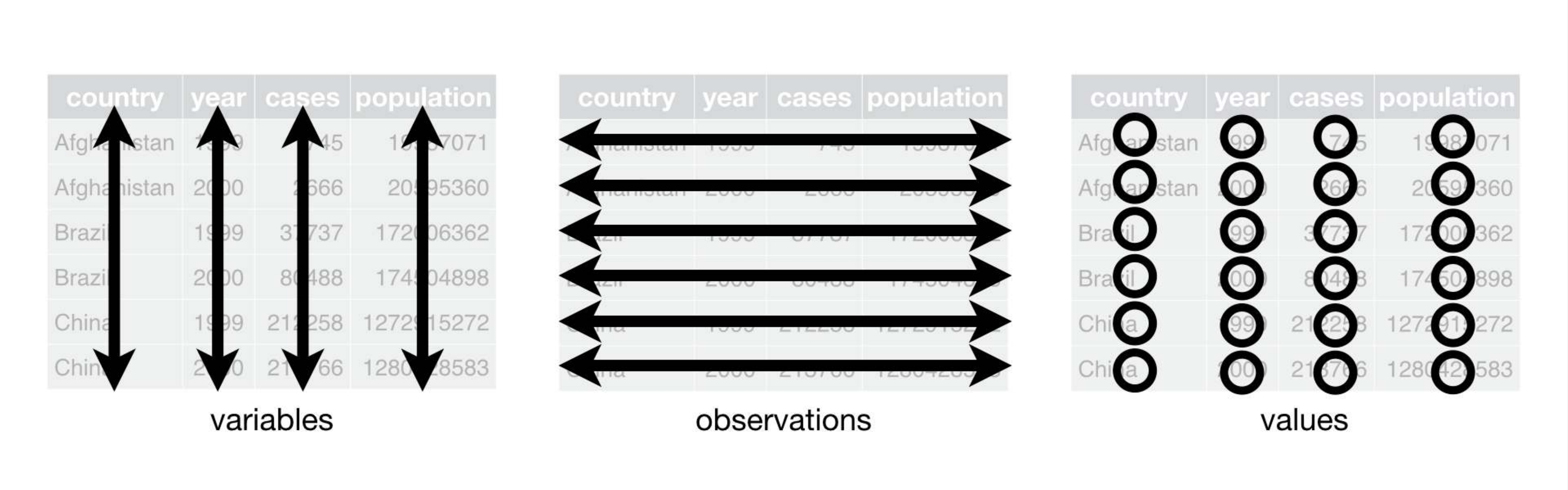
"Long" format

```
#> # A tibble: 6 × 3
#>   department  year rd_budget_mil
#>   <chr>      <dbl>         <dbl>
#> 1 DOD         1976         35696
#> 2 NASA        1976         12513
#> 3 DOE         1976         10882
#> 4 HHS         1976          9226
#> 5 NIH         1976          8025
#> 6 NSF         1976          2372
```

```
#> [1] 588   3
```

Tidy data = "Long" format

- Each **variable** has its own **column**
- Each **observation** has its own **row**



Tidy data

- Each **variable** has its own **column**
- Each **observation** has its own **row**

```
#> # A tibble: 6 × 3
#>   department  year rd_budget_mil
#>   <chr>      <dbl>         <dbl>
#> 1 DOD         1976         35696
#> 2 NASA        1976         12513
#> 3 DOE         1976         10882
#> 4 HHS         1976          9226
#> 5 NIH         1976          8025
#> 6 NSF         1976          2372
```

"Long" format

```
#> # A tibble: 6 × 3
#>   department  year rd_budget_mil
#>   <chr>      <dbl>         <dbl>
#> 1 DOD         1976         35696
#> 2 NASA        1976         12513
#> 3 DOE         1976         10882
#> 4 HHS         1976          9226
#> 5 NIH         1976          8025
#> 6 NSF         1976          2372
```

"Wide" format

```
#> # A tibble: 6 × 15
#>    year   DHS   DOC   DOD   DOE   DOT   EPA   HHS Inte
#>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <
#> 1  1976     0   819 35696 10882  1142   968  9226
#> 2  1977     0   837 37967 13741  1095   966  9507
#> 3  1978     0   871 37022 15663  1156  1175 10533
#> 4  1979     0   952 37174 15612  1004  1102 10127
#> 5  1980     0   945 37005 15226  1048   903 10045
#> 6  1981     0   829 41737 14798   978   901  9644
```

Do the names describe the values?

Yes: "Long" format

```
#> # A tibble: 6 × 3
#>   department  year rd_budget_mil
#>   <chr>      <dbl>         <dbl>
#> 1 DOD         1976         35696
#> 2 NASA        1976         12513
#> 3 DOE         1976         10882
#> 4 HHS         1976          9226
#> 5 NIH         1976          8025
#> 6 NSF         1976          2372
```

No: "Wide" format

```
#> # A tibble: 6 × 8
#>    year   DHS   DOC   DOD   DOE   DOT   EPA   HHS
#>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1  1976     0   819 35696 10882  1142   968  9226
#> 2  1977     0   837 37967 13741  1095   966  9507
#> 3  1978     0   871 37022 15663  1156  1175 10533
#> 4  1979     0   952 37174 15612  1004  1102 10127
#> 5  1980     0   945 37005 15226  1048   903 10045
#> 6  1981     0   829 41737 14798   978   901  9644
```

Quick practice 1: "long" or "wide" format?

Description: Tuberculosis cases in various countries

```
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

Quick practice 2: "long" or "wide" format?

Description: Word counts in LOTR trilogy

```
#> # A tibble: 9 × 4
#>   Film                       Race   Female  Male
#>   <chr>                      <chr>   <dbl> <dbl>
#> 1 The Fellowship Of The Ring Elf      1229   971
#> 2 The Fellowship Of The Ring Hobbit     14  3644
#> 3 The Fellowship Of The Ring Man         0  1995
#> 4 The Return Of The King     Elf       183   510
#> 5 The Return Of The King     Hobbit      2  2673
#> 6 The Return Of The King     Man       268  2459
#> 7 The Two Towers             Elf       331   513
#> 8 The Two Towers             Hobbit      0  2463
#> 9 The Two Towers             Man       401  3589
```

Quick practice 3: "long" or "wide" format?

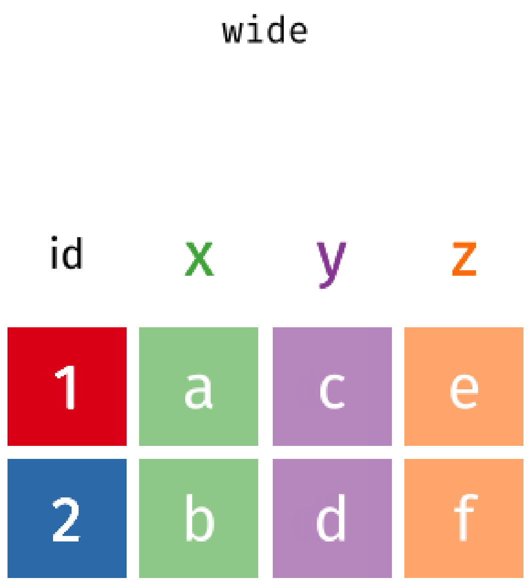
Description: Word counts in LOTR trilogy

```
#> # A tibble: 15 × 4
#>    Film                       Race   Gender Word_Count
#>    <chr>                      <chr>  <chr>       <dbl>
#>  1 The Fellowship Of The Ring Elf    Female       1229
#>  2 The Fellowship Of The Ring Elf    Male          971
#>  3 The Fellowship Of The Ring Hobbit Female         14
#>  4 The Fellowship Of The Ring Hobbit Male         3644
#>  5 The Fellowship Of The Ring Man    Female          0
#>  6 The Fellowship Of The Ring Man    Male         1995
#>  7 The Return Of The King     Elf    Female        183
#>  8 The Return Of The King     Elf    Male          510
#>  9 The Return Of The King     Hobbit Female          2
#> 10 The Return Of The King     Hobbit Male         2673
#> 11 The Return Of The King     Man    Female        268
#> 12 The Return Of The King     Man    Male         2459
#> 13 The Two Towers             Elf    Female        331
#> 14 The Two Towers             Elf    Male          513
#> 15 The Two Towers             Hobbit Female          0
```

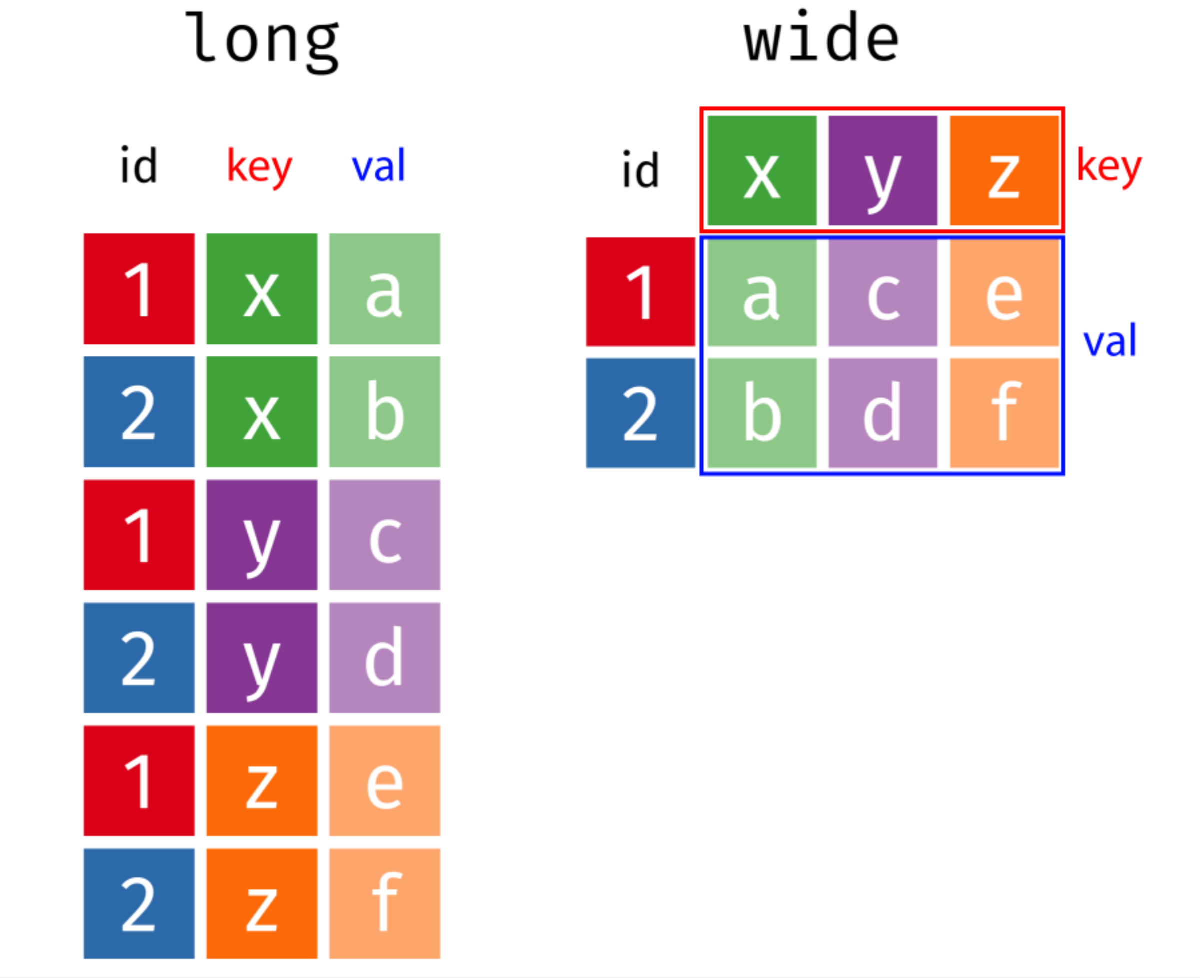
Reshaping data with `pivot_longer()` and `pivot_wider()`

Reshaping data

```
pivot_longer()
pivot_wider()
```



From "long" to "wide" with `pivot_wider()`



From "long" to "wide" with `pivot_wider()`

```
head(fed_spend_long)
```

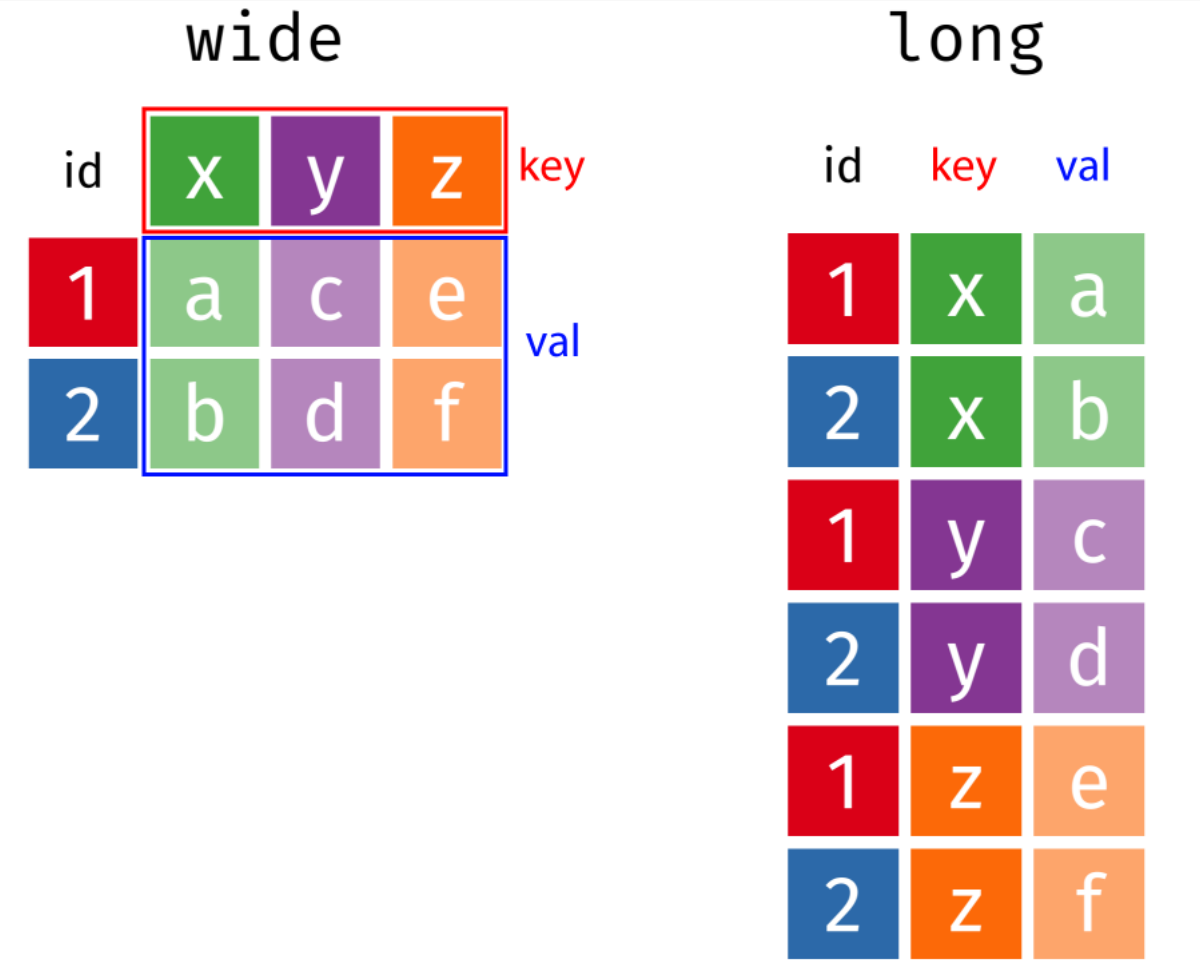
```
#> # A tibble: 6 × 3
#>   department  year rd_budget_mil
#>   <chr>      <dbl>         <dbl>
#> 1 DOD         1976         35696
#> 2 NASA        1976         12513
#> 3 DOE         1976         10882
#> 4 HHS         1976          9226
#> 5 NIH         1976          8025
#> 6 NSF         1976          2372
```

```
fed_spend_wide <- fed_spend_long %>%
    pivot_wider(
        names_from = department,
        values_from = rd_budget_mil)

head(fed_spend_wide)
```

```
#> # A tibble: 6 × 15
#>    year   DOD  NASA   DOE   HHS   NIH   NSF
#>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <
#> 1  1976 35696 12513 10882  9226  8025  2372
#> 2  1977 37967 12553 13741  9507  8214  2395
#> 3  1978 37022 12516 15663 10533  8802  2446
#> 4  1979 37174 13079 15612 10127  9243  2404
#> 5  1980 37005 13837 15226 10045  9093  2407
#> 6  1981 41737 13276 14798  9644  8580  2300
```

From "wide" to "long" with `pivot_longer()`



From "wide" to "long" with `pivot_longer()`

```
head(fed_spend_wide)
```

```
#> # A tibble: 6 × 15
#>    year   DOD  NASA   DOE   HHS    NI
#>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl
#> 1  1976 35696 12513 10882  9226   802
#> 2  1977 37967 12553 13741  9507   821
#> 3  1978 37022 12516 15663 10533   880
#> 4  1979 37174 13079 15612 10127   924
#> 5  1980 37005 13837 15226 10045   909
#> 6  1981 41737 13276 14798  9644   858
```

```
fed_spend_long <- fed_spend_wide %>%
    pivot_longer(
        names_to = "department",
        values_to = "rd_budget_mil",
        cols = DOD:Other)

head(fed_spend_long)
```

```
#> # A tibble: 6 × 3
#>    year department rd_budget_mil
#>   <dbl> <chr>              <dbl>
#> 1  1976 DOD                35696
#> 2  1976 NASA               12513
#> 3  1976 DOE                10882
#> 4  1976 HHS                 9226
#> 5  1976 NIH                 8025
#> 6  1976 NSF                 2372
```

Can also set `cols` by selecting which columns *not* to use

```r
names(fed_spend_wide)
```

```
#>  [1] "year"      "DOD"       "NASA"
```

```r
fed_spend_long <- fed_spend_wide %>%
    pivot_longer(
        names_to = "department",
        values_to = "rd_budget_mil",
        cols = -year)

head(fed_spend_long)
```

```
#> # A tibble: 6 × 3
#>    year department rd_budget_mil
#>   <dbl> <chr>              <dbl>
#> 1  1976 DOD                35696
#> 2  1976 NASA               12513
#> 3  1976 DOE                10882
#> 4  1976 HHS                 9226
#> 5  1976 NIH                 8025
#> 6  1976 NSF                 2372
```

Your turn: Reshaping Data

15:00

Open the `practice.Rmd` file.

Run the code chunk to read in the following two data files:

- `pv_cell_production.xlsx`: Data on solar photovoltaic cell production by country
- `milk_production.csv`: Data on milk production by state

Now modify the format of each:

- If the data are in "wide" format, convert it to "long" with `pivot_longer()`
- If the data are in "long" format, convert it to "wide" with `pivot_wider()`

Break

05:00

Week 9: *Uncertainty*

1. Computing uncertainty

2. Reshaping data

BREAK

3. Cleaning pilot data

4. Estimating pilot data models

Download the demo-choice-based-conjoint repo

Cleaning surveydown survey data

1. Open `survey.Rproj`
2. Open `code/data_cleaning.R`

Your Turn

20:00

As a team, pick up where you left off last week and create a `choiceData` data frame in a "long" format

Week 9: *Uncertainty*

1. Computing uncertainty

2. Reshaping data

BREAK

3. Cleaning pilot data

4. Estimating pilot data models

Estimating pilot data models

1. Open `survey.Rproj`
2. Open `code/modeling.R`

Your Turn

As a team:

1. Use your `choiceData` data frame to estimate preliminary choice models.
2. Interpret your model coefficients **with uncertainty**.