Linear Regression
(model with categorical variable
and
model with transformation)

Regression with categorical variables

- Coding of categorical variables
  - 0/1 (this is **reference cell coding**): **default**
  - -1/1 (**deviations from means coding**)

- ANOVA is a special case of linear regression model with only categorical variables
- Interpretation of coefficient should be different
  - One unit increase in x … does not make sense for categorical variable
- Estimated coefficient represents difference between **Reference (a group coded as 0)** vs. **Comparison group (the other group coded as 1)**

Example with categorical variable: professor.csv

- **SALARY: continuous response**
- **Gender: categorical predictor**
- **TIME/ CITS/ PUBS: continuous predictors**

(1) SALARY ~ Gender + PUBS **(w/out interaction)**

(2) SALARY ~ Gender + PUBS + Gender*PUBS **(w/ interaction)**

- R automatically handles factor variable (with reference cell coding) and we can check which group is set as a reference group (coded as 0) from the output
- If interaction term is significant, it implies: **the effect of PUBS differs depending on the level of Gender**

Example with categorical variable: w/out interaction

```
summary(lm(SALARY~Gender+PUBS, data=professor)) ##

##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41911.98    2353.34  17.810  < 2e-16 ***
## GenderMale   7143.82    2533.25   2.820  0.00653 **
## PUBS          470.28      90.42   5.201 2.61e-06 ***
## ---
```

Average difference between Male and Female. If it is positive, it means that compared to females, males earn $7,143.82 more.

- **Gender=Female (reference group - coded as 0)**

$$\hat{Y} = 41911.9 + 470.28 * \text{PUBS}$$

- **Gender=Male (comparison group - coded as 1)**

$$\hat{Y} = (41911.9+7143) + 470.28 * \text{PUBS}$$

- Interpretation: On average, males earn 7143 more than females do when the number of publications is same

Example with categorical variable: w/ interaction

```
summary(lm(SALARY~Gender*PUBS, data=professor)) ##

## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      47680.3     2969.8  16.055   <2e-16 ***
## GenderMale       -1998.6     3937.9  -0.508    0.614
## PUBS               102.1      152.2   0.671    0.505
## GenderMale:PUBS    535.9      183.6   2.918    0.005 **
```

- Decide to include interaction -> include main effects
- **Gender=Female (reference group - coded as 0)**

$$\hat{Y} = 47680.3 + 102.1 * \text{PUBS}$$

- **Gender=Male (comparison group – coded as 1)**

$$\hat{Y} = (47680.3 - 1998.6) + (102.1 + 535.9) * \text{PUBS}$$

Regression with categorical variables: More than 2 levels

- What if a categorical variable has more than two levels in it? E.g., A1, A2, A3
- Same manner – one reference group (e.g., A1) and two estimates are **expected differences between (A2 vs. A1) and (A3 vs. A1)**
- Example with ToothGrowth data (used in ANOVA)

Regression with categorical variables: ToothGrowth.csv

```
## 'data.frame':    60 obs. of  3 variables:
##  $ Toothlength: num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ Supplement : Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2
##  $ Dose       : Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1
```

```
summary(lm(Toothlength ~ Dose, data=tooth))

##
## Call:
## lm(formula = Toothlength ~ Dose, data = tooth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6000 -3.2350 -0.6025  3.3250 10.8950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6050     0.9486  11.180 5.39e-16 ***
## Dose1         9.1300     1.3415   6.806 6.70e-09 ***
## Dose2        15.4950     1.3415  11.551  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The expected (average) difference between Dose2 and Dose 0.5 is 15.4950

On average, the difference between the average Toothlength for the group in Dose1 is 15.4950 larger than the expected Toothlength for group Dose 0.5

The expected (average) difference between Dose1 and Dose 0.5 is 9.1300.

On average, the difference between the average Toothlength for the group in Dose1 is 9.1300 larger than the expected Toothlength for group Dose 0.5

With log transformation

NOT THIS FOR LOG

- Additive change
  - ✓Same amount of change in Y regardless of x values
- Multiplicative change
  - ✓Change of Y depends on x values

- We expect to see β increase in log(Y) with one unit increase in X (-> hard to get what it means)
- Important to interpret with **original scale of Y**
  - Meaning of positive or negative β
  - Multiplicative change instead of additive change

Data: athletes.txt

- Data on 102 male and 100 female athletes collected at the Australian Institute of Sport, courtesy of Richard Telford and Ross Cunningham.

- Sport: Sport
- Sex : male or female
- Ht : Height in cm
- Wt : Weight in kg
- LBM : Lean body mass
- RCC : Red cell count
- WCC : White cell count
- Hc : Hematocrit
- Hg : Hemoglobin
- Ferr : Plasma ferritin concentration
- BMI : Body mass index = weight/height^2
- SSF : Sum of skin folds
- %Bfat : % body fat

Data: athletes.txt

- Response: Ferr (Plasma ferritin concentration)
- Predictors: all except Sport

❖ **Use original scale of Ferr**

- Stepwise selection through “proc reg” with .05 criteria for entering and removal of variables
- Check diagnostics plot
- Interpretation of estimated coefficients

❖ **Use log transformed Ferr**

- Try log transformation on Ferr
- Repeat the same thing with log(Ferr)
- Check diagnostics plot
- Interpretation of estimated coefficients

With log transformation

- When x is continuous
  - Y is expected to increase/ decrease with multiplicative factor being $e^{\hat{\beta}}$, with one unit increase in x

If $B^{1\wedge}$ is greater than 0, then $e^{B1\wedge} > 1$ **As X increases Y increases**

If $B^{1\wedge}$ is les than 0, then $e^{B1\wedge} < 1$ **As X increases Y decreases**

- When x is categorical
  - Expected Y of male (comparison group) is $e^{\hat{\beta}}$ times for the expected Y of female (reference group)