ANOVA –
Categorical Data as Predictors

One-Way ANOVA

- ANOVA is used to test for differences among several means without increasing the Type I error rate, only doing one test

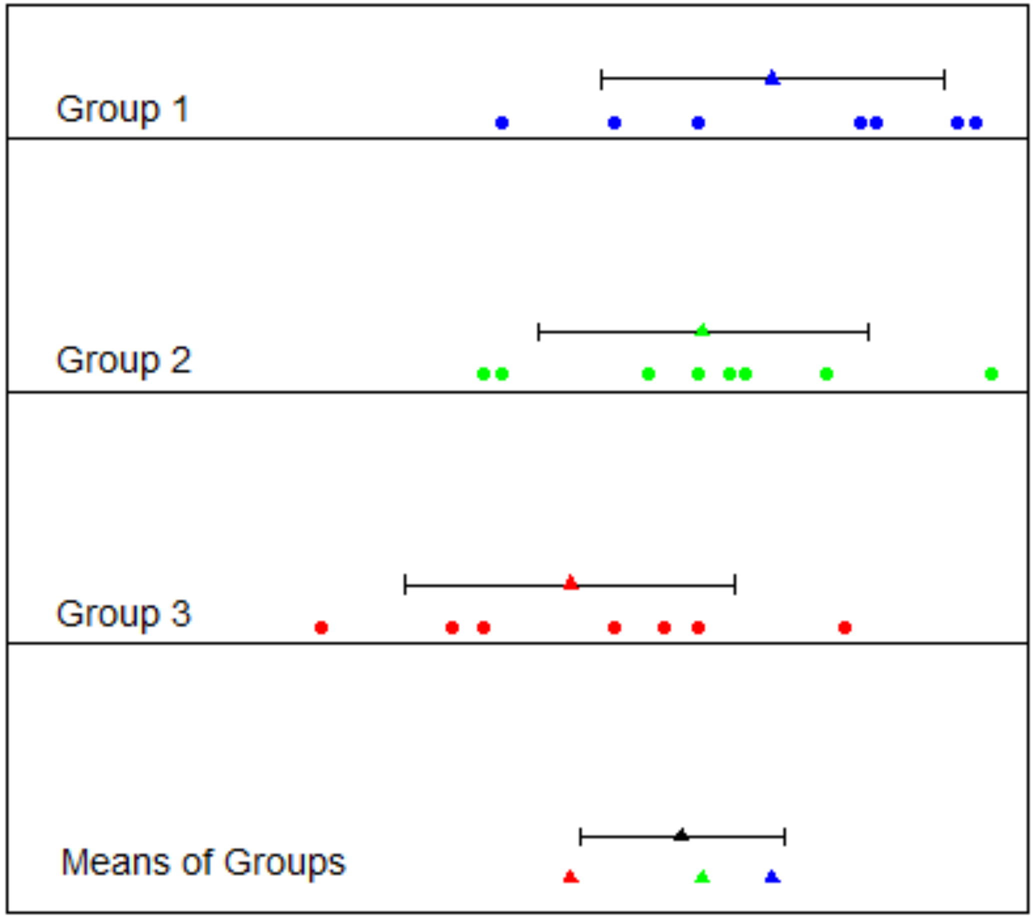
- The ANOVA uses data from all groups to estimate standard errors, which can increase the power of the analysis

- Basic Idea
  - Calculate the mean of the observations within each group
  - Compare the variance of these means to the average variance within each group
  - As the means become more different, the variance among the means increases

Why Look at Variance When Interested in Means?

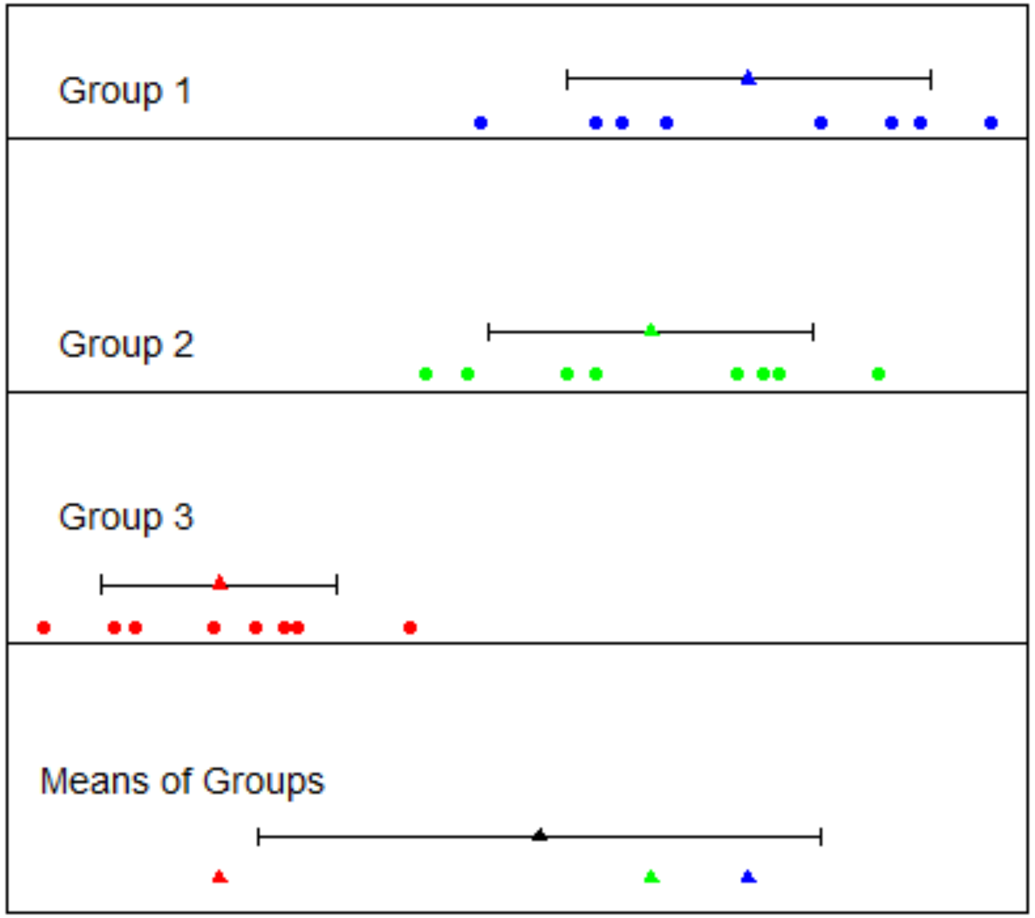
- The variability within each sample is approximately the same

- The variability in the mean values of the samples is consistent with the variability within the individual samples



Why Look at Variance When Interested in Means?

- The variability in the sample means is much larger than would be expected given the variability within each of the samples



One-Way ANOVA

- When there is only one categorical variable which denotes the groups and only one measurement variable (quantitative), a one-way ANOVA is carried out

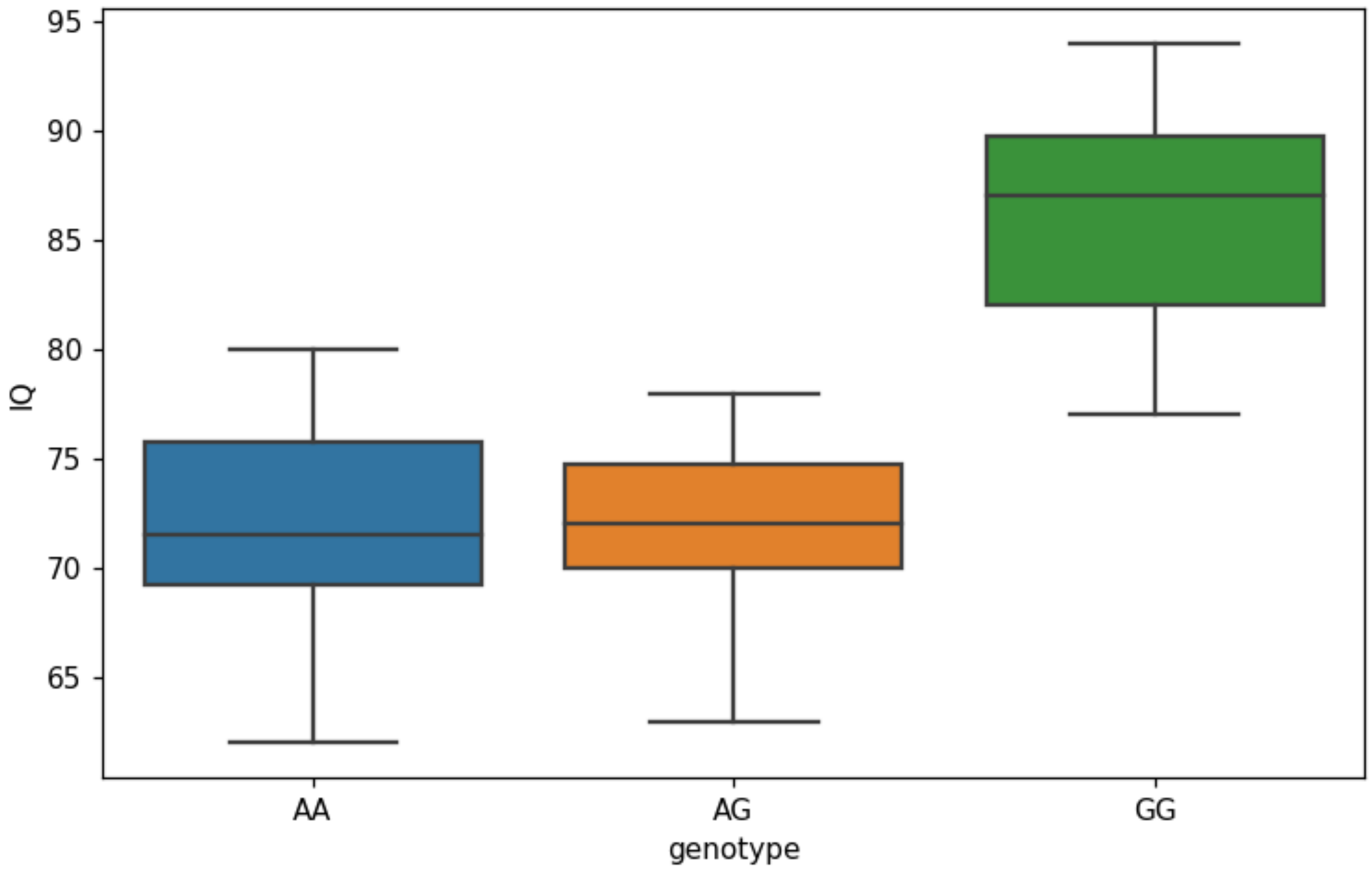
- For a one-way ANOVA the observations are divided into I mutually exclusive categories, giving the one-way classification

One-Way ANOVA Assumptions

- Assumptions
  - Each of the populations is Normally distributed with the same variance (homogeneity of variance)
  - The observations are sampled independently, the groups under consideration are independent

- ANOVA is robust to moderate violations of its assumptions, meaning that the probability values (P-values) computed in an ANOVA are sufficiently accurate even if the assumptions are moderately violated

One-Way ANOVA Example



- 54 observations
- 18 AA observations

mean IQ for AA = 71.9

- 18 AG observations

mean IQ for AG = 72.2

- 18 GG observations

mean IQ for GG = 86.1

Introduction of Notation

- Consider I groups, whose means we want to compare

- Let n_i, $i = 1, 2, \ldots, I$ be the sample size of group i

- For the simulated verbal IQ and genotype data, ($I = 3$), representing the three possible genotypes at the particular location in a gene of interest. Each person in this data set, as well as having a genotype, also has a verbal IQ score

One-Way ANOVA Example

- Want to examine if the mean verbal IQ score is the same across the 3 genotype groups

Null hypothesis is that the mean verbal IQ is the same in the three genotype groups:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

One-Way ANOVA Example

Within-Groups Variance

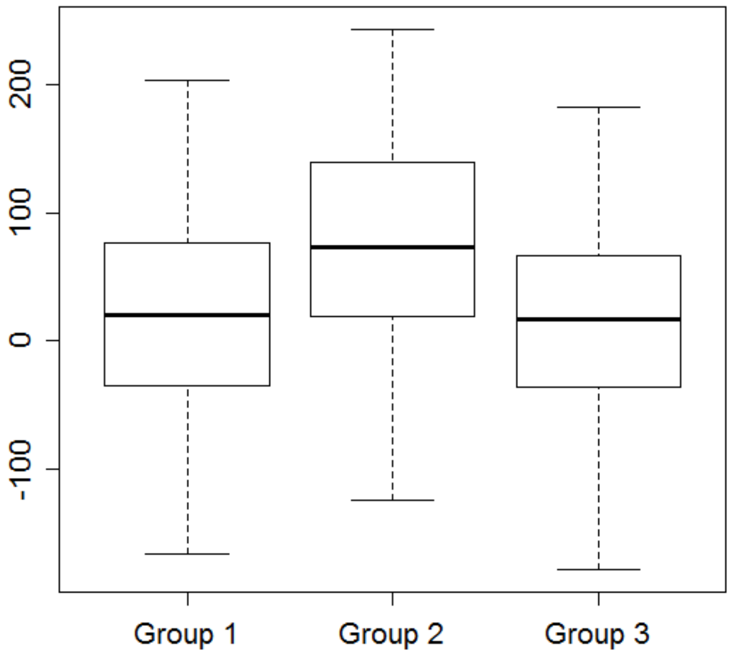
- Remember assumption that the population variances of the three groups is the same

- Under this assumption, the three variances of the three groups all estimate this common value, σ^2

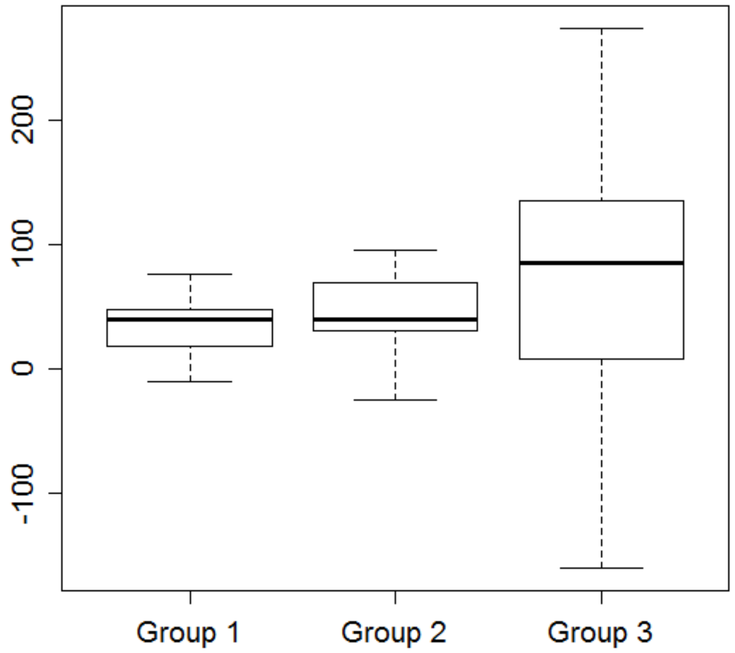
One-Way ANOVA Example

Within-Groups Variance

- Since the population variances are assumed to be equal the estimate of the population variance, derived from the separate within-group estimates, is valid whether or not the null hypothesis is true



EQUAL



NOT EQUAL

Within-Groups Variance

- For groups with equal sample size this is defined as the average of the variances of the groups

$$s_w^2 = \frac{1}{I}\sum_{i=1}^{I} s_i^2 \ = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\frac{(x_{ij} - \bar{x}_i)^2}{n_i - 1}\right)$$

x_{ij} = observation *j* in group *i*

$\bar{x}_1, \bar{x}_2, \bar{x}_3$ = sample means of the genotype groups AA, AG, GG

One-Way ANOVA Example

Between-Groups Variance

If the null hypothesis is true (which we assume):

- the groups can be considered as random samples from the same population

- assumed equal variances, and because the null hypothesis is true, then the population means are equal

- The means are observations from the same sampling distribution of the mean

Between-Groups Variance

- The sampling distribution of the mean has variance

$$\sigma^2/n$$

- This gives a second method of obtaining an estimate of the population variance (n = number of observations in each group)

- The observed variance of the treatment means is an estimate of σ^2/n and is given by:

$$\frac{s^2}{n} = \sum_{i=1}^{I} \frac{(\bar{x}_i - \bar{x})^2}{I-1}$$

Unequal Sample Sizes

- There are adjustments to these formulae when the sample sizes are not all equal in the groups:
  - Within Groups variance:

$$s_w^2 = \sum_{i=1}^{I} \frac{(n_i - 1)s_i^2}{N - I}$$

  - Between Groups Variance:

$$s_b^2 = \sum_{i=1}^{I} n_i \frac{(\bar{x}_i - \bar{x})^2}{I - 1}$$

One-Way ANOVA Example

- If the null hypothesis IS TRUE then
  - the between-groups variance s_b^2
  - and the within-groups variance s_w^2
  - are both estimates of the population variance σ^2
- If the null hypothesis is NOT TRUE then
  - the population means are not all equal
  - then s_b^2 will be greater then the population variance, σ^2
  - it will be increased by the *treatment* (genotype) differences

One-Way ANOVA Example

- To test the null hypothesis we compare the ratio of s_b^2 and s_w^2 using an F-test
- F statistic is given by: $$F = \frac{s_b^2}{s_w^2}$$

with I -1 and I(n-1) = N - I degrees of freedom

- Can also think of the F statistic as:

$$F = \frac{\text{variability due to treatment effect and variability due to chance}}{\text{variability due to chance}}$$

F Distribution and F-test

- The F distribution is the continuous distribution of the ratio of two estimates of variance

- The F distribution has two parameters: degrees of freedom numerator (top) and degrees of freedom denominator (bottom)

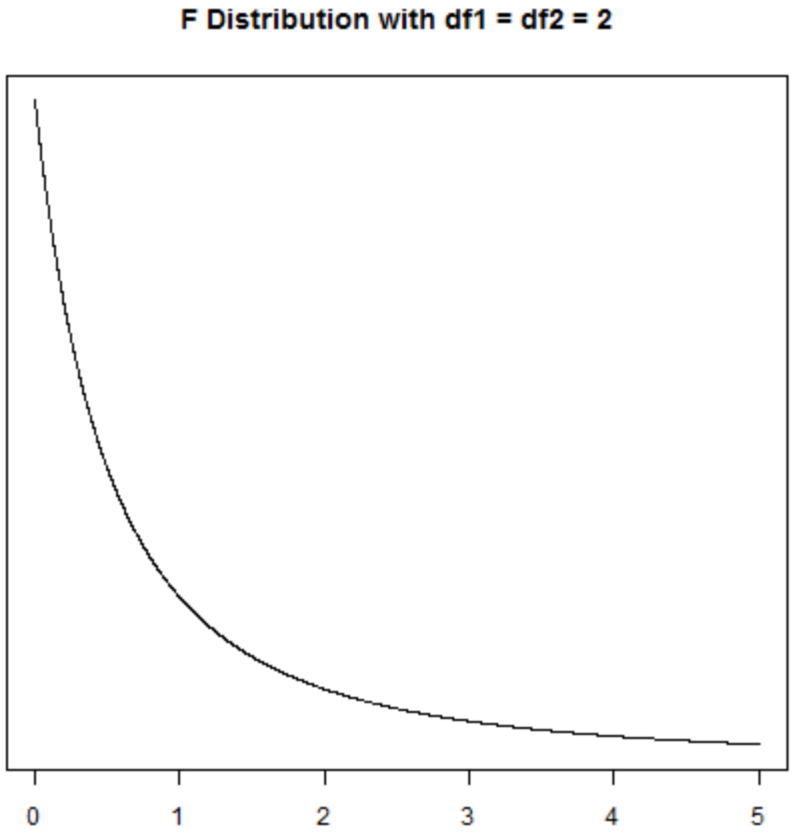
- The F-test is used to test the hypothesis that two variances are equal

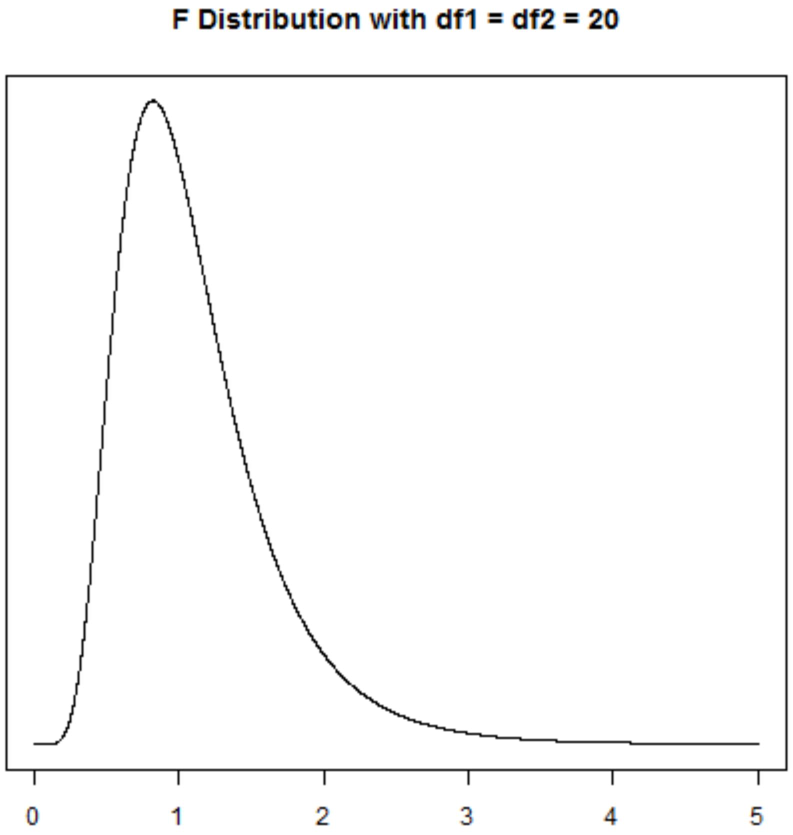
F Distribution and F-test

- The validity of the F-test is based on the requirement that the populations from which the variances were taken are Normal

- In the ANOVA, a one-sided F-test is used

F Distribution





One-Way ANOVA Example

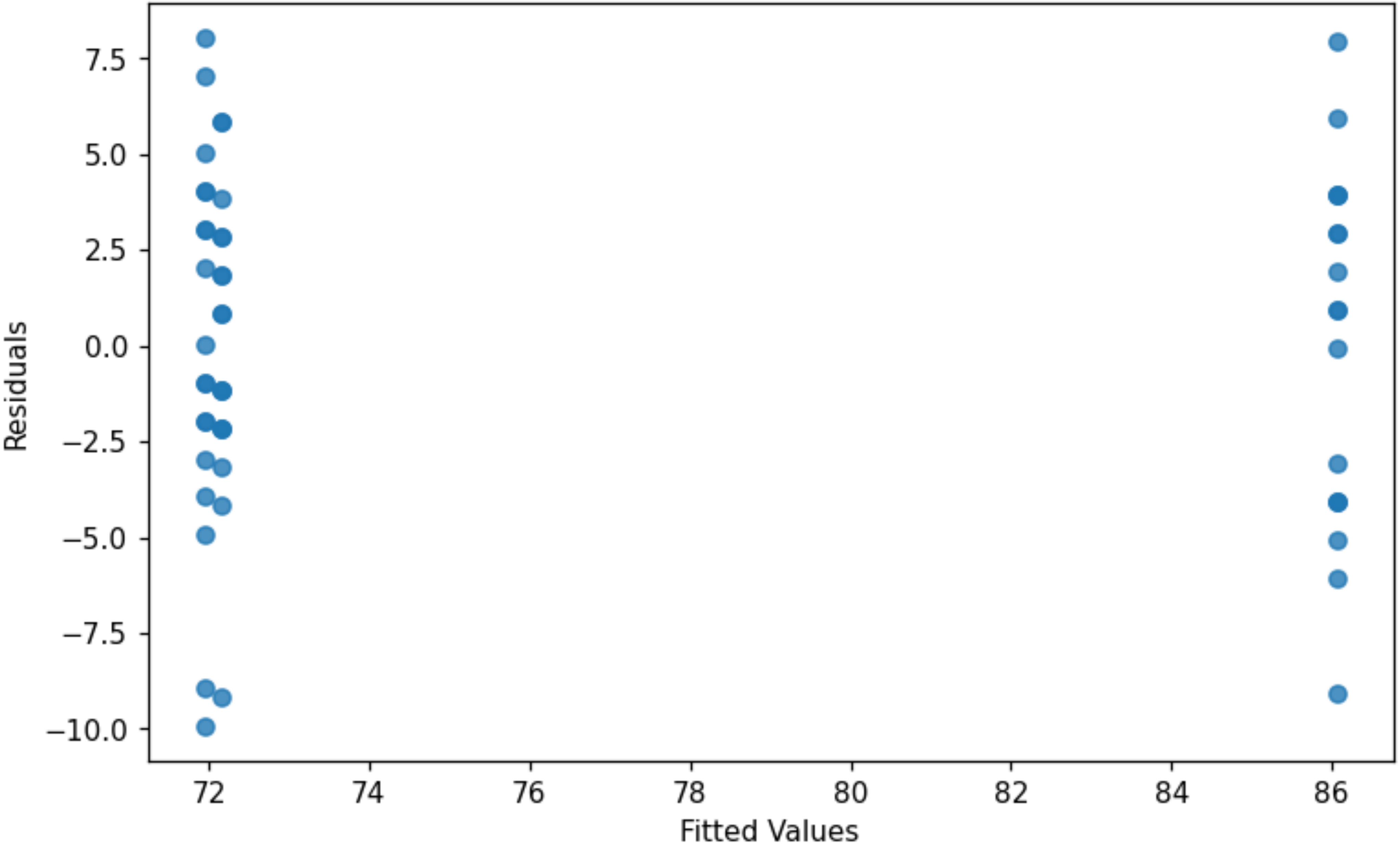
```
df.head()
Out[5]:
  genotype    IQ
0       AA  63.0
1       AA  67.0
2       AA  75.0
3       AA  76.0
4       AA  70.0

#create lm model output and then use anova on it, create using ols formula
from statsmodels.formula.api import ols
model= ols('IQ ~ C(genotype)', data=df).fit()
#or get dummy variables to get same results
df[["geno_AG","geno_GG"]]=pd.get_dummies(df["genotype"])[["AG","GG"]]
X = df[["geno_AG","geno_GG"]]
X = sm.add_constant(X)
y = df['IQ']
model1 = sm.OLS(y, X).fit()
```

One-Way ANOVA Assumption Checking

- Homogeneity of variance = homoscedasticity
  - The dependent variable (quantitative measurement) should have the same variance in each category of the independent variable (qualitative variable)
- ANOVA is robust for small to moderate departures from homogeneity of variance, especially with equal sample sizes for the groups
- Rule of thumb: the ratio of the largest to the smallest group variance should be 3:1 or less, but be careful, the more unequal the sample sizes the smaller the differences in variances which are acceptable
- Testing for homogeneity of variance
  - Examine boxplots of the data by group, will highlight visually if there is a large difference in variability between the groups
  - Plot residuals versus fitted values and examine scatter around zero,

residuals = observations – group mean

group mean = fitted value

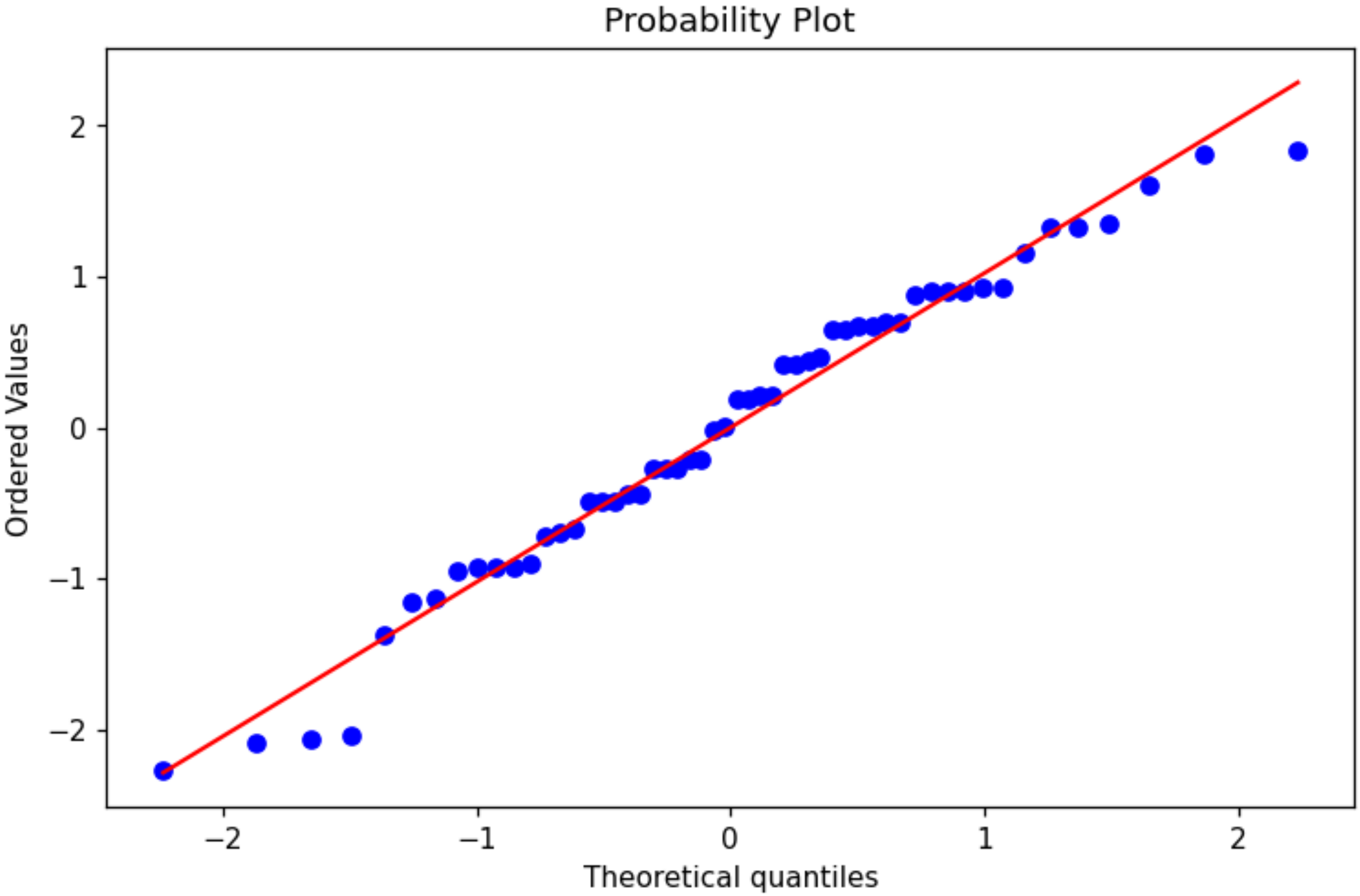

Residuals
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
72
74
76
78
80
82
84
86
Fitted Values

One-Way ANOVA Assumption Checking

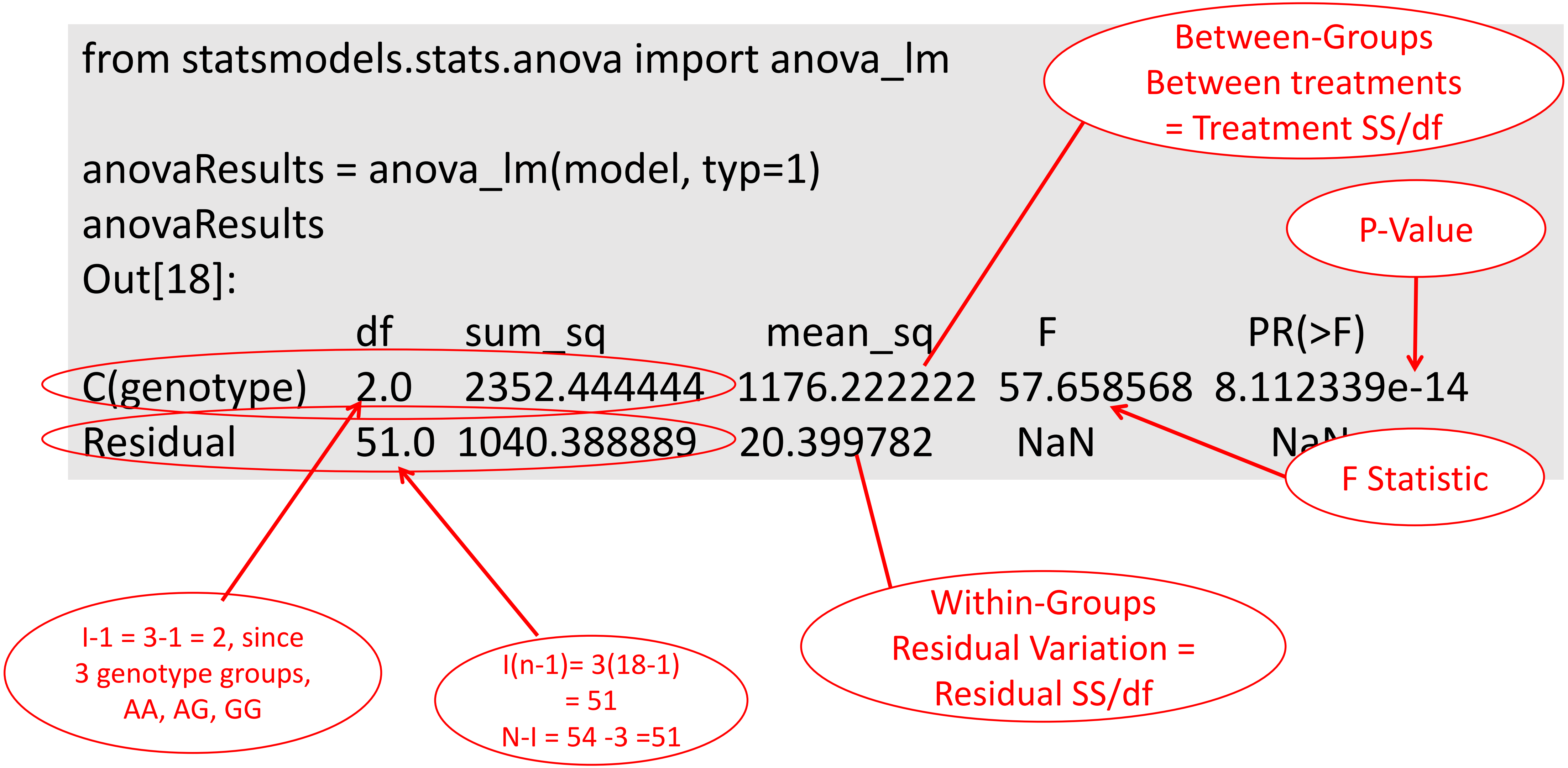
- Normality Assumption
  - The dependent variable (measurement, quantitative variable) should be Normally distributed in each category of the independent variable (qualitative variable)

- Again ANOVA is robust to moderate departures from Normality

- Checking the Normality assumption
  - Boxplots of the data
  - Quantile-Quantile plots (QQ plots) of the residuals, which should give a 45-degree line on a plot of observed versus expected values,


Probability Plot
Ordered Values
Theoretical quantiles
2
1
0
−1
−2
−2
−1
0
1
2

One-Way ANOVA Example



```
from statsmodels.stats.anova import anova_lm

anovaResults = anova_lm(model, typ=1)
anovaResults
Out[18]:
               df      sum_sq        mean_sq          F          PR(>F)
C(genotype)    2.0     2352.444444  1176.222222  57.658568  8.112339e-14
Residual       51.0  1040.388889   20.399782      NaN         NaN
```

Between-Groups
Between treatments
= Treatment SS/df

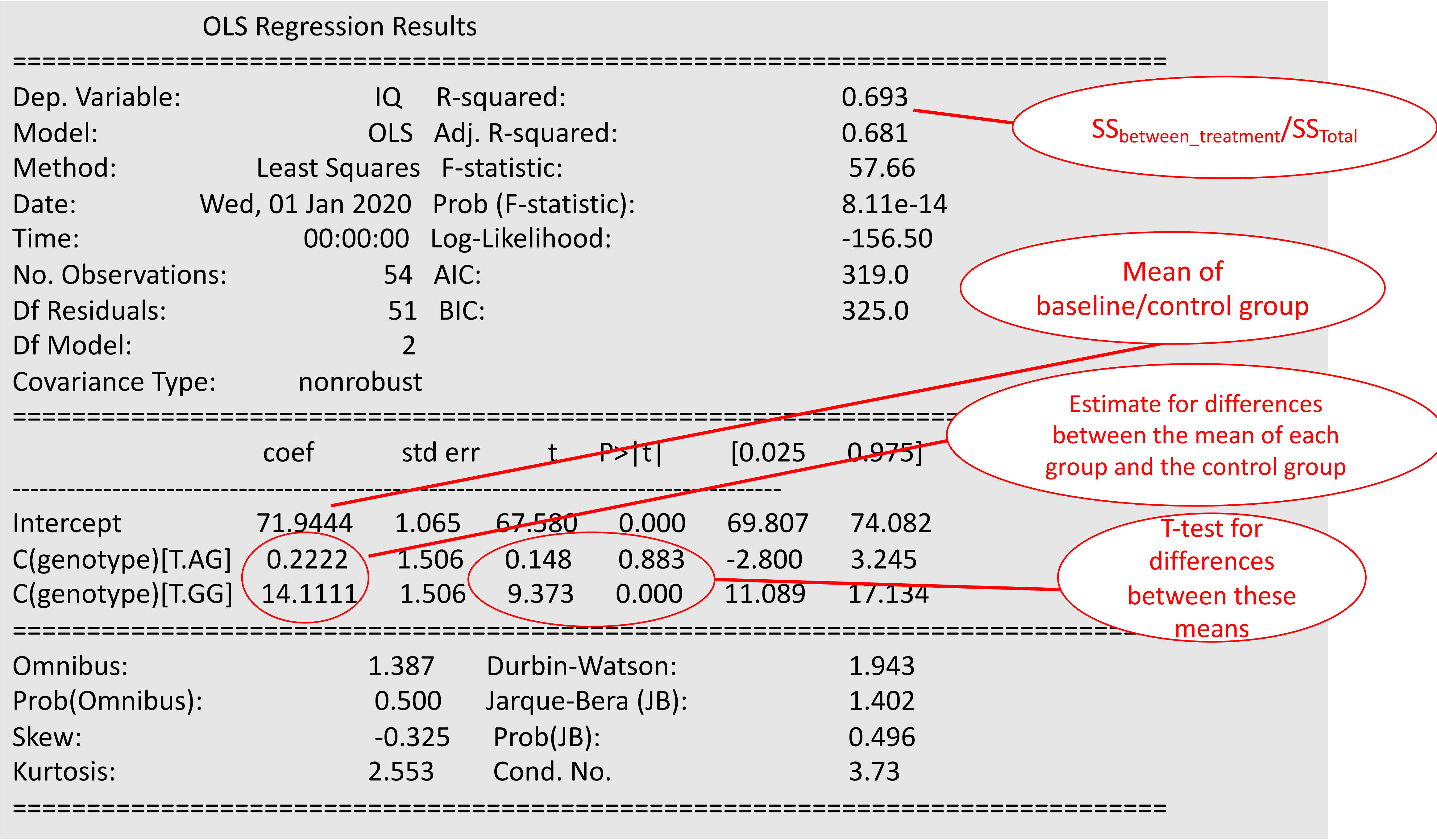
P-Value

F Statistic

Within-Groups
Residual Variation =
Residual SS/df

I-1 = 3-1 = 2, since
3 genotype groups,
AA, AG, GG

I(n-1)= 3(18-1)
= 51
N-I = 54 -3 =51



```
                    OLS Regression Results
==============================================================================
Dep. Variable:                     IQ   R-squared:                       0.693
Model:                            OLS   Adj. R-squared:                  0.681
Method:                 Least Squares   F-statistic:                     57.66
Date:                Wed, 01 Jan 2020   Prob (F-statistic):           8.11e-14
Time:                        00:00:00   Log-Likelihood:                -156.50
No. Observations:                  54   AIC:                             319.0
Df Residuals:                      51   BIC:                             325.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept         71.9444      1.065     67.580      0.000      69.807      74.082
C(genotype)[T.AG]  0.2222      1.506      0.148      0.883      -2.800       3.245
C(genotype)[T.GG] 14.1111      1.506      9.373      0.000      11.089      17.134
==============================================================================
Omnibus:                        1.387   Durbin-Watson:                   1.943
Prob(Omnibus):                  0.500   Jarque-Bera (JB):                1.402
Skew:                          -0.325   Prob(JB):                        0.496
Kurtosis:                       2.553   Cond. No.                         3.73
==============================================================================
```

$SS_{between\_treatment}/SS_{Total}$

Mean of baseline/control group

Estimate for differences between the mean of each group and the control group

T-test for differences between these means

What to do with a Significant ANOVA Result (F-test)

- If the ANOVA is significant and the null hypothesis is rejected, the only valid inference that can be made is that at least one population mean is different from at least one other population mean

- The ANOVA does not reveal which population means differ from which others

What to do with a Significant ANOVA Result (F-test)

- Only think about investigating differences between individual groups when the overall comparison of groups (ANOVA) is significant, or that you had intended particular comparisons at the outset

- Need to consider whether the groups are ordered or not

Two Way ANOVA

- Two way analysis of variance (ANOVA) without interactions is the same a regression with two categorical explanatory variables.
- Two way analysis of variance (ANOVA) with interactions is the same a regression with two categorical explanatory variables plus a third categorical explanatory variable for the interaction.
- When the explanatory variable is categorical, conceptually it is recoded using dummy or indicator variables.

Two Way ANOVA

- We can test more hypothesis in a two-way ANOVA:
- There is no difference in the means of factor A
- There is no difference in means of factor B
- There is no interaction between factors A and B
- The alternative hypothesis for the first two is: the means are not equal.
- The alternative hypothesis for the last one is: there is an interaction between A and B.

Assumptions of a Two-Way ANOVA

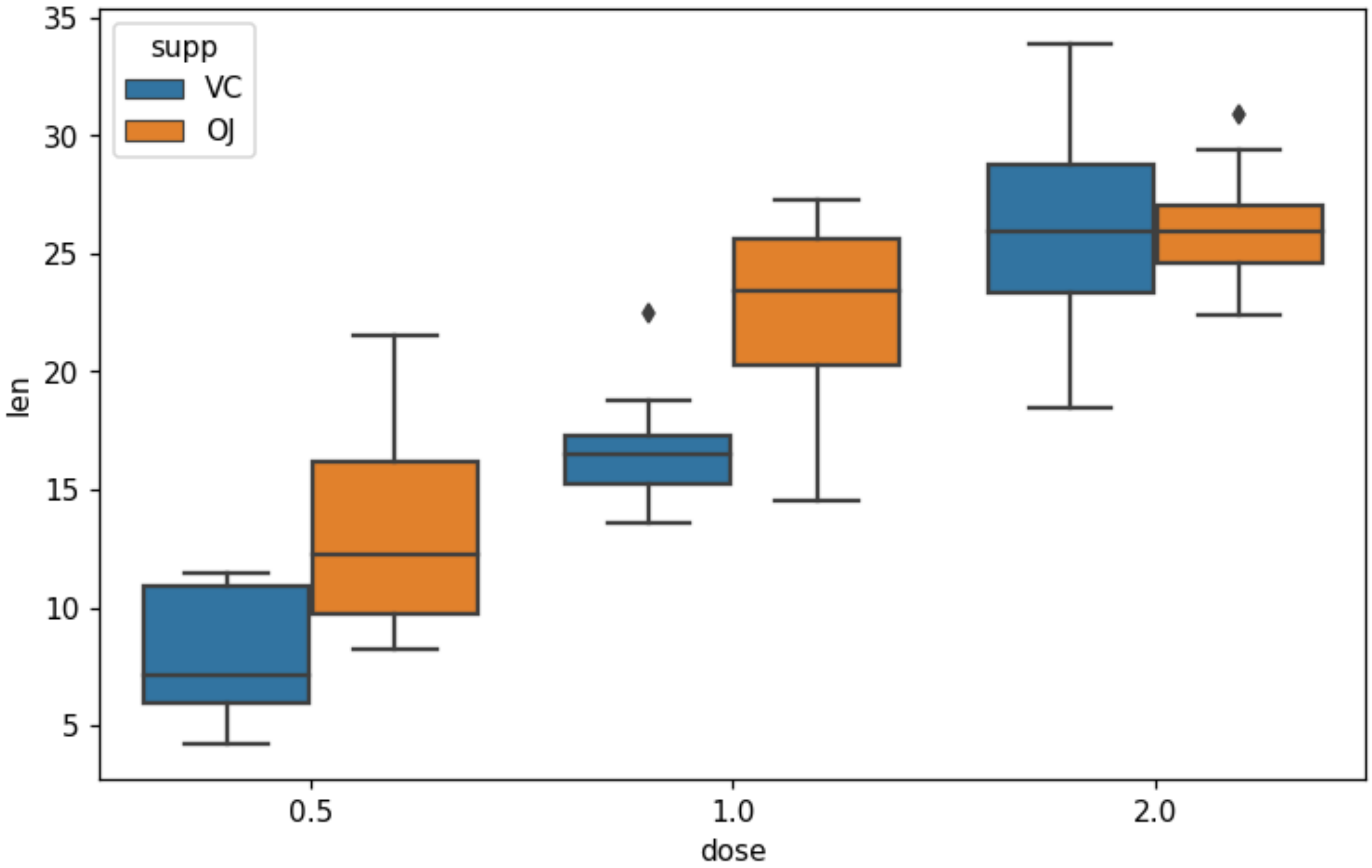
- The dependent variable should be continuous and the two independent variables should be in categorical, independent groups.
- Observations are sampled independently – that each sample has been drawn independently of the other samples
- Equality of Variance (homogeneity of variance) – That the variance of data in the different groups should be the same
- Normality – That each sample is taken from a normally distributed population

Example: ToothGrowth

- Tooth Growth dataset in R contains data from a study evaluating the effect of vitamin C on tooth growth in Guinea pigs.
- The experiment has been performed on 60 pigs, where each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods - orange juice or ascorbic acid (a form of vitamin C and coded as VC).

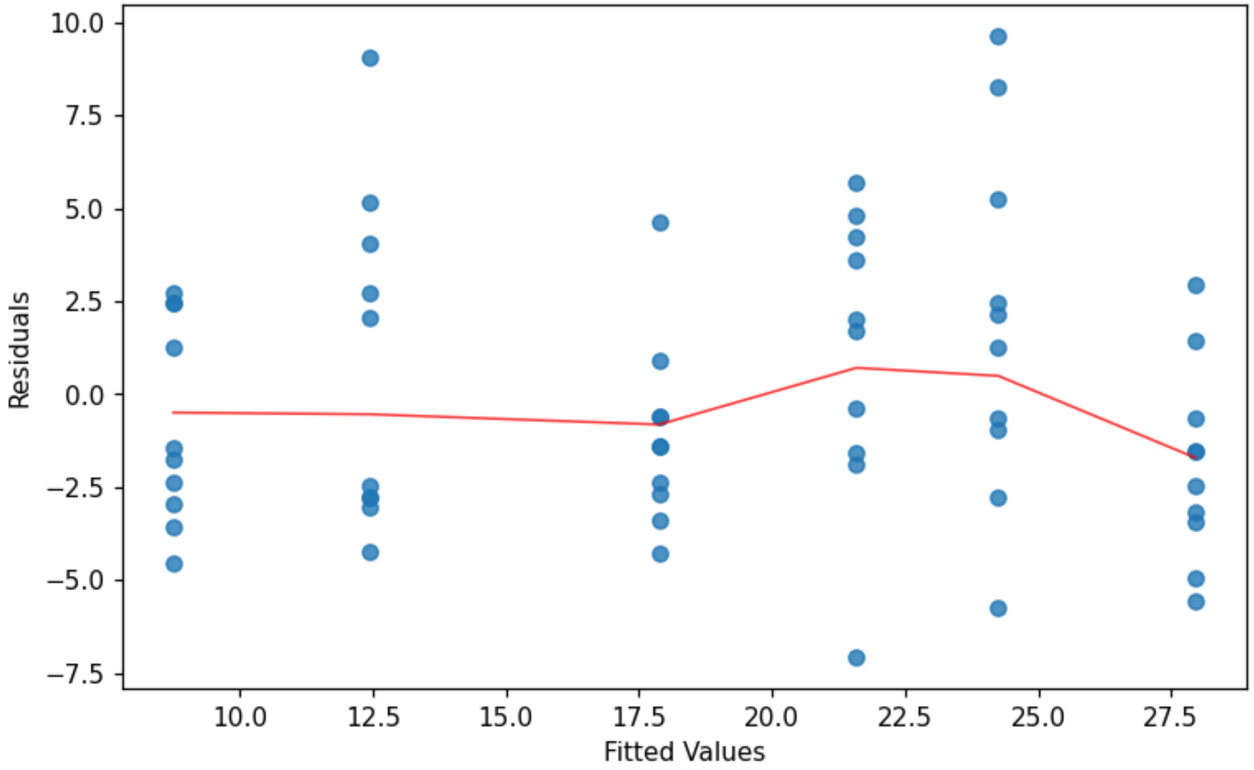
```
ToothGrowth.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 60 entries, 0 to 59
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   len     60 non-null     float64
 1   supp    60 non-null     object
 2   dose    60 non-null     float64
dtypes: float64(2), object(1)
memory usage: 1.5+ KB

TooothGrowth.head()
```

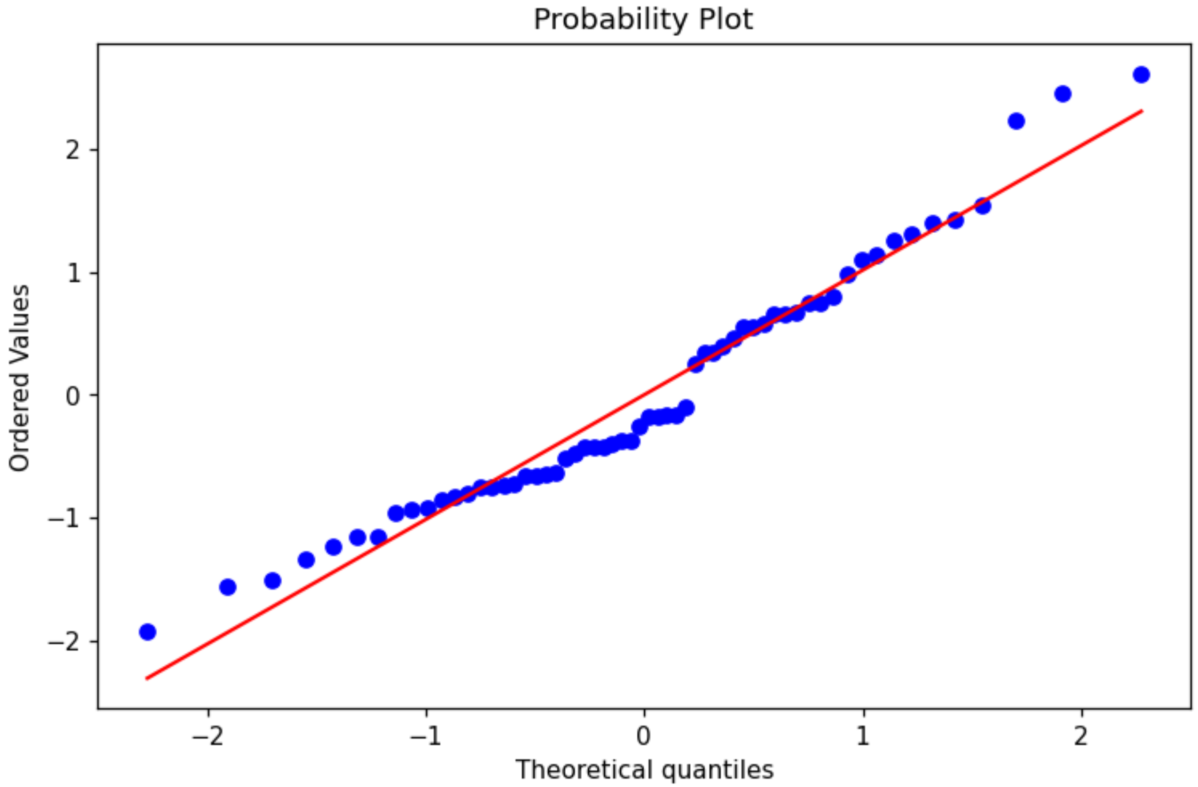
```
sns.boxplot(x="dose", y="len",hue='supp', data=ToothGrowth)
```



Example

```
model2= ols('len ~ C(supp)+C(dose)', data=ToothGrowth).fit()

#fitted values
model_fitted_vals = model2.fittedvalues
#model residuals
model_residuals = model2.resid
#standardised residuals
model_norm_residuals = model2.get_influence().resid_studentized_internal

sns.regplot(x=model_fitted_vals,y=model_residuals,
            ci=False,lowess=True,
            line_kws={'color': 'red', 'lw': 1, 'alpha': 0.8})
plt.xlabel("Fitted Values")
plt.ylabel("Residuals")

stats.probplot(model_norm_residuals, plot=sns.mpl.pyplot)
plt.show()
```


Residuals
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
Fitted Values


Probability Plot
Ordered Values
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical quantiles

Example

```
supp_dose= pd.crosstab(index=ToothGrowth['supp'], columns=ToothGrowth["dose"], margins=True)

supp_dose
Out[60]:
dose  0.5  1.0  2.0  All
supp
OJ     10   10   10   30
VC     10   10   10   30
All    20   20   20   60
#balanced design
anova2way = anova_lm(model2, typ=1)
anova2way
Out[61]:
             df      sum_sq         mean_sq          F            PR(>F)
C(supp)     1.0     205.350000     205.350000      14.016638   4.292793e-04
C(dose)     2.0     2426.434333    1213.217167     82.810935   1.871163e-17
Residual   56.0     820.425000     14.650446       NaN         NaN
```

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                    len   R-squared:                       0.762
Model:                            OLS   Adj. R-squared:                  0.750
Method:                 Least Squares   F-statistic:                     59.88
Date:                Wed, 01 Jan 2020   Prob (F-statistic):           1.78e-17
Time:                        00:00:00   Log-Likelihood:                -163.60
No. Observations:                  60   AIC:                             335.2
Df Residuals:                      56   BIC:                             343.6
Df Model:                           3
Covariance Type:            nonrobust
==================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          12.4550      0.988     12.603      0.000      10.475      14.435
C(supp)[T.VC]      -3.7000      0.988     -3.744      0.000      -5.680      -1.720
C(dose)[T.1.0]      9.1300      1.210      7.543      0.000       6.705      11.555
C(dose)[T.2.0]     15.4950      1.210     12.802      0.000      13.070      17.920
==============================================================================
Omnibus:                        3.615   Durbin-Watson:                   1.814
Prob(Omnibus):                  0.164   Jarque-Bera (JB):                3.366
Skew:                           0.575   Prob(JB):                        0.186
Kurtosis:                       2.853   Cond. No.                         4.22
==============================================================================
```

Interpretation

- From the ANOVA table (or coefficient table), we can conclude that both *supp* and *dose* are statistically significant.
- Therefore, this would imply that changing delivery methods (supp) or the dose of vitamin C, will impact significantly the mean tooth length.
- Not the above model is called **additive model**. It makes an assumption that the two factor variables are independent.
- If you think that these two variables have interaction effect then replace the plus symbol (+) by an asterisk (*)

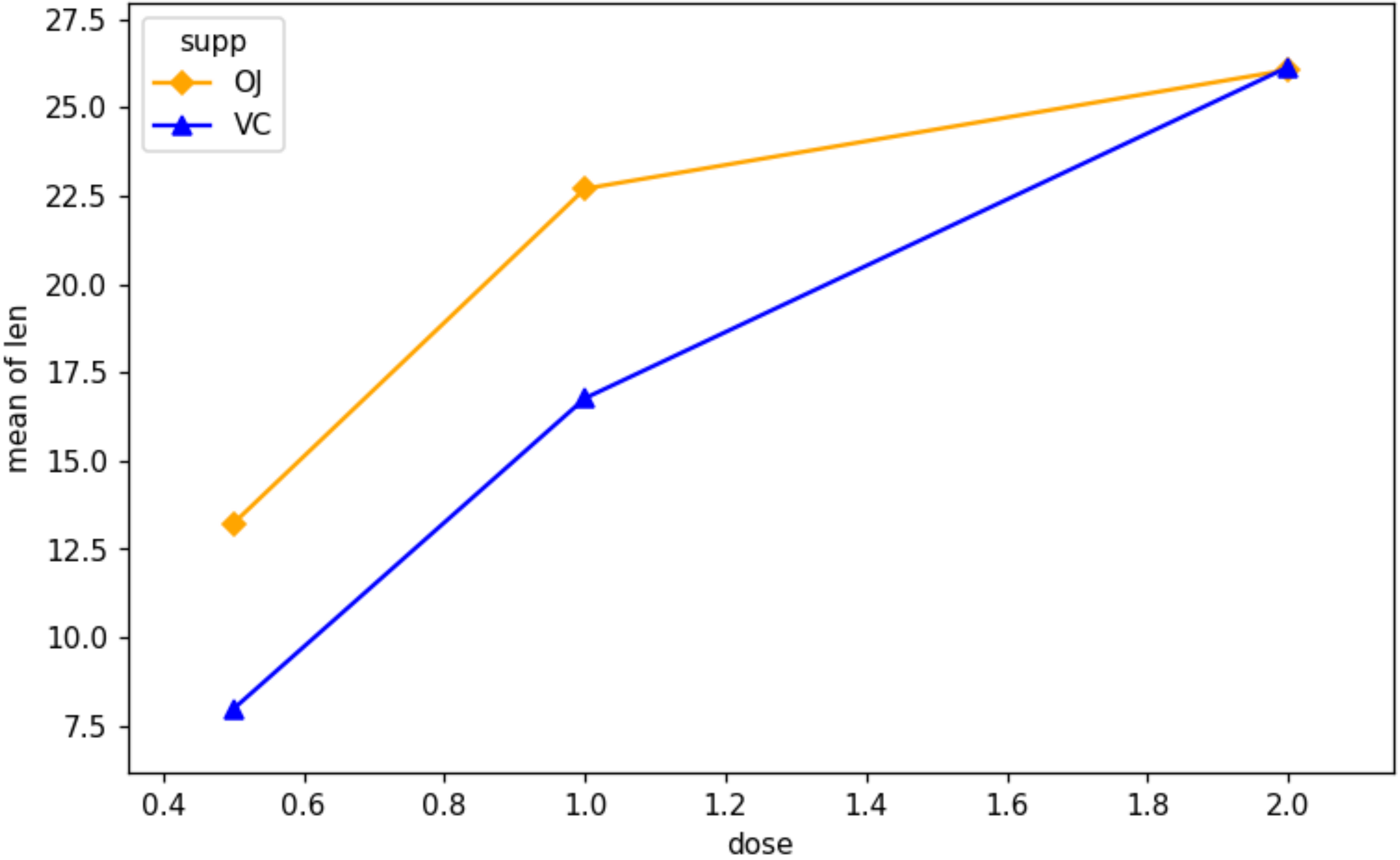
Example with interaction term

```
model2int= ols('len ~ C(supp)*C(dose)', data=ToothGrowth).fit()
model2int= ols('len ~ C(supp)+C(dose)+C(supp):C(dose)', data=ToothGrowth).fit()
# These two calls are equivalent
```

Interaction Plot in python

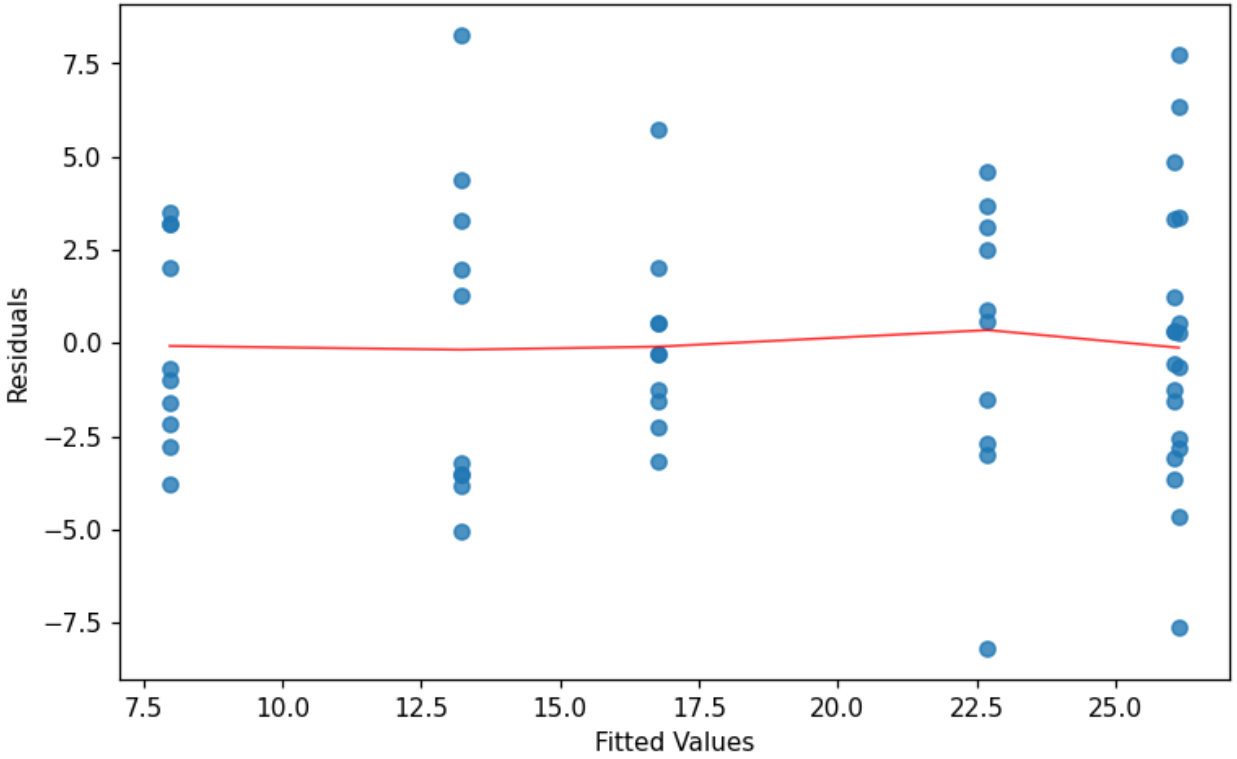
- Interaction_plot command in statsmodels package
- Parameters inside that package:
  - **x** : the factor to be plotted on x axis.
  - **trace**: the factor to be plotted as lines
  - **response**: a numeric variable giving the response

```
from statsmodels.graphics.factorplots import interaction_plot

interaction_plot(ToothGrowth['dose'], ToothGrowth['supp'], ToothGrowth['len'],
                colors=['orange','blue'], markers=['D','^'])
plt.show()
```

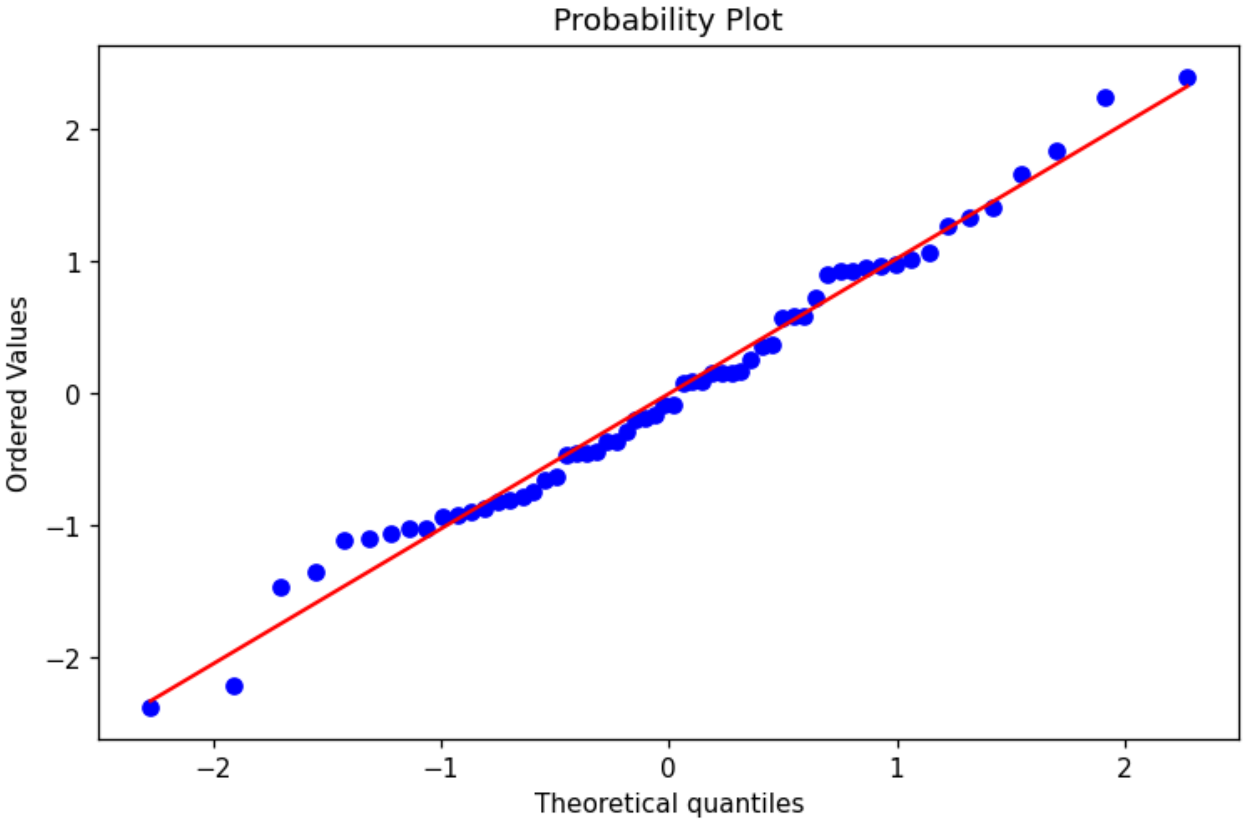

supp
OJ
VC
mean of len
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
dose

Interpreting Interaction Plots

- Parallel lines - No interaction occurs.
- Nonparallel lines - An interaction occurs. The more nonparallel the lines are, the greater the strength of the interaction.
- Can see that the lines are parallel between D0.5 and D1 for two different supplement types with OJ leading to longer teeth but then there is an interaction when the dose is increased to 2 and both supplements perform the same.



Residuals
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
Fitted Values



Probability Plot
Ordered Values
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical quantiles

Example with interaction term

```
model2int= ols('len ~ C(supp)*C(dose)', data=ToothGrowth).fit()
model2int= ols('len ~ C(supp)+C(dose)+C(supp):C(dose)', data=ToothGrowth).fit()
# These two calls are equivalent
```

```
anova2wayint = anova_lm(model2int, typ=1)
anova2wayint
                 df      sum_sq        mean_sq          F             PR(>F)
C(supp)          1.0    205.350000    205.350000    15.571979   2.311828e-04
C(dose)          2.0    2426.434333   1213.217167   91.999965   4.046291e-18
C(supp):C(dose)  2.0    108.319000    54.159500     4.106991    2.186027e-02
Residual         54.0   712.106000    13.187148     NaN         NaN
```

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                     len   R-squared:                       0.794
Model:                             OLS   Adj. R-squared:                  0.775
Method:                  Least Squares   F-statistic:                     41.56
Date:                 Wed, 01 Jan 2020   Prob (F-statistic):           2.50e-17
Time:                         00:00:00   Log-Likelihood:                -159.35
No. Observations:                   60   AIC:                             330.7
Df Residuals:                       54   BIC:                             343.3
Df Model:                            5
Covariance Type:             nonrobust
==================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                       13.2300      1.148     11.521      0.000      10.928      15.532
C(supp)[T.VC]                   -5.2500      1.624     -3.233      0.002      -8.506      -1.994
C(dose)[T.1.0]                   9.4700      1.624      5.831      0.000       6.214      12.726
C(dose)[T.2.0]                  12.8300      1.624      7.900      0.000       9.574      16.086
C(supp)[T.VC]:C(dose)[T.1.0]    -0.6800      2.297     -0.296      0.768      -5.285       3.925
C(supp)[T.VC]:C(dose)[T.2.0]     5.3300      2.297      2.321      0.024       0.725       9.935
==============================================================================
```

- From the ANOVA results and based on a significance level of 0.05, you can conclude that:
  - the p-value of supp is 2.311828e-04, which indicates that the levels of supp are associated with significant different tooth length.
  - the p-value of dose is 4.046291e-18, which indicates that the levels of dose are associated with significant different tooth length.
  - the p-value for the interaction between supp*dose is 0.02, which indicates that the relationships between dose and tooth length depends on the supp method.

```
TootḥGrowth.groupby('dose').mean()
Out[99]:
        len
dose
0.5   10.605
1.0   19.735
2.0   26.100

ToothGrowth.groupby('supp').mean()
Out[100]:
               len               dose
supp
OJ             20.663333  1.166667
VC             16.963333  1.166667

ToothGrowth.groupby(['dose', 'supp']).mean()
Out[101]:
                              len
dose         supp
0.5          OJ               13.23
             VC                7.98
1.0          OJ               22.70
             VC               16.77
2.0          OJ               26.06
    V        C                26.14
```

Unbalanced two-way ANOVA

- Balanced designs correspond to the situation where we have equal sample sizes within levels of our independent grouping levels.
- In experimental data, the experimenter will often set up the data with equal number of observations per cell.
- An **unbalanced design** has unequal numbers of samples in each group.
- typ ="2" can be used to compute two-way ANOVA test for unbalanced designs.

Unbalance ANOVA

```
anova2wayint = anova_lm(model2int, typ=2)
anova2wayint
Out[102]:
                        sum_sq     df          F        PR(>F)
C(supp)             205.350000    1.0  15.571979  2.311828e-04
C(dose)            2426.434333    2.0  91.999965  4.046291e-18
C(supp):C(dose)     108.319000    2.0   4.106991  2.186027e-02
Residual            712.106000   54.0        NaN           NaN
```

Unbalanced vs Balanced

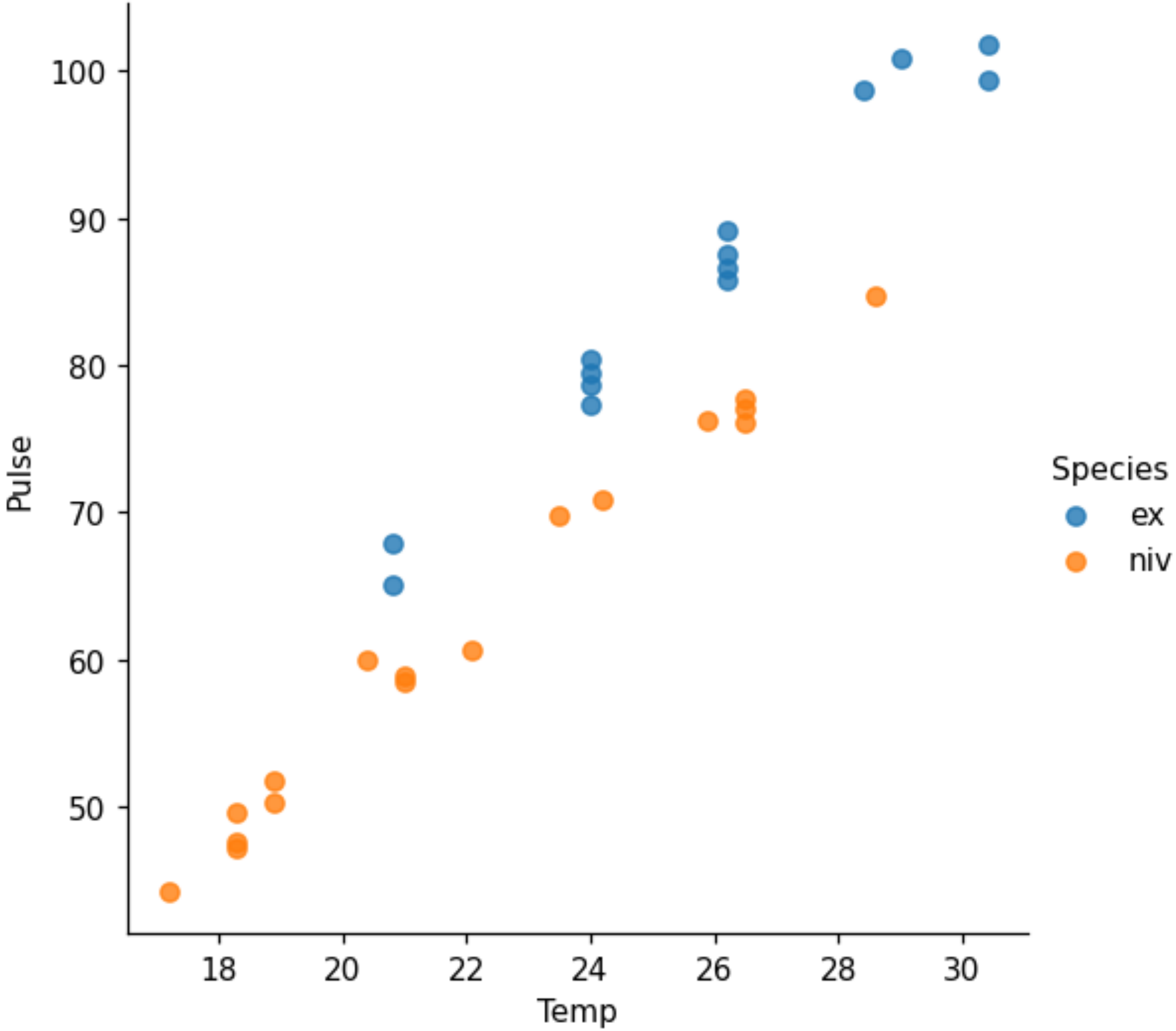
- Unbalanced design will always impact the power - the ability to detect significant differences. The power is limited by the size of the smallest cell.
- Unbalanced design usually impacts the ability to cleanly divide up the sums of squares and may end up with unexplained variance that is due to an effect but unable to say which effect. This is different than unexplained (residual) variance.
- Also it may mask an important relationship in the data.

ANCOVA

- Analysis of covariance (ANCOVA) is the same as a regression with one categorical and one continuous explanatory variables.
- ANCOVA evaluates whether the means of a dependent variable are equal across levels of a categorical independent variable, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates.
- The categorical variable divides the regressions into two or more sets.

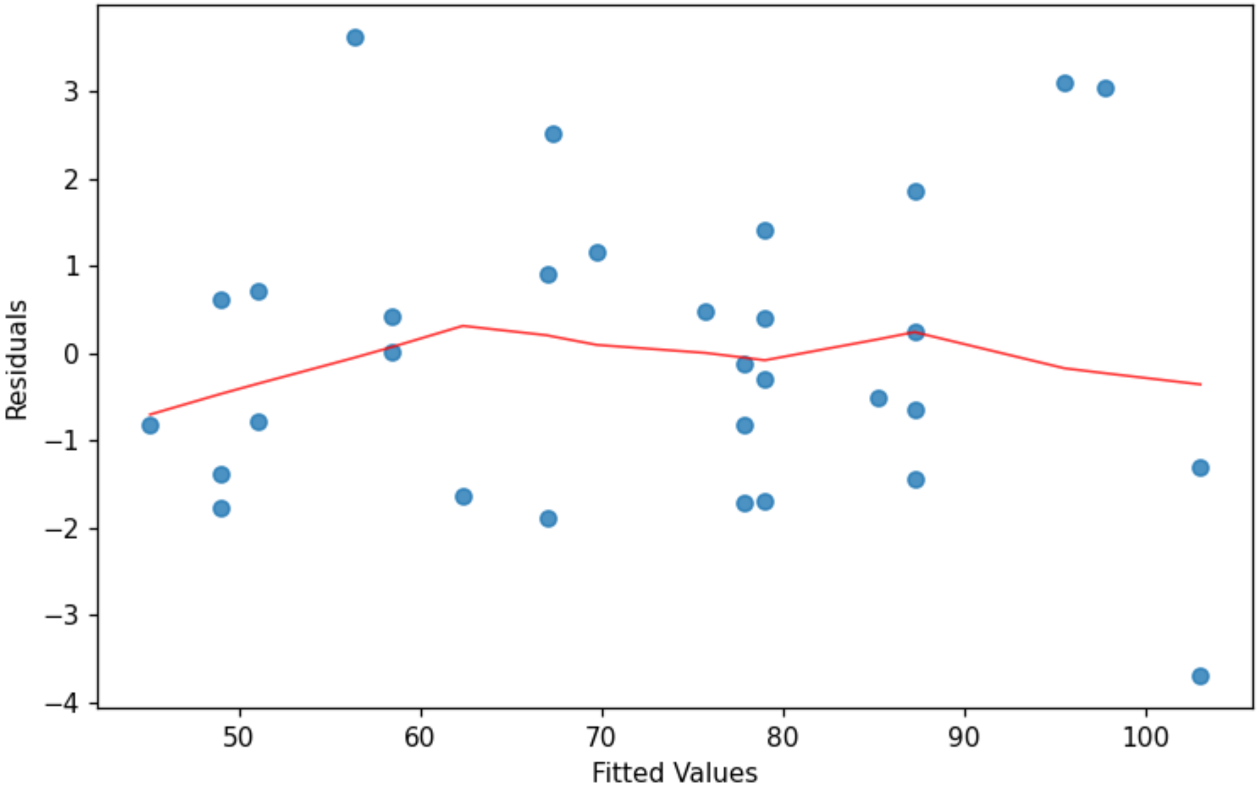
Example: crickets

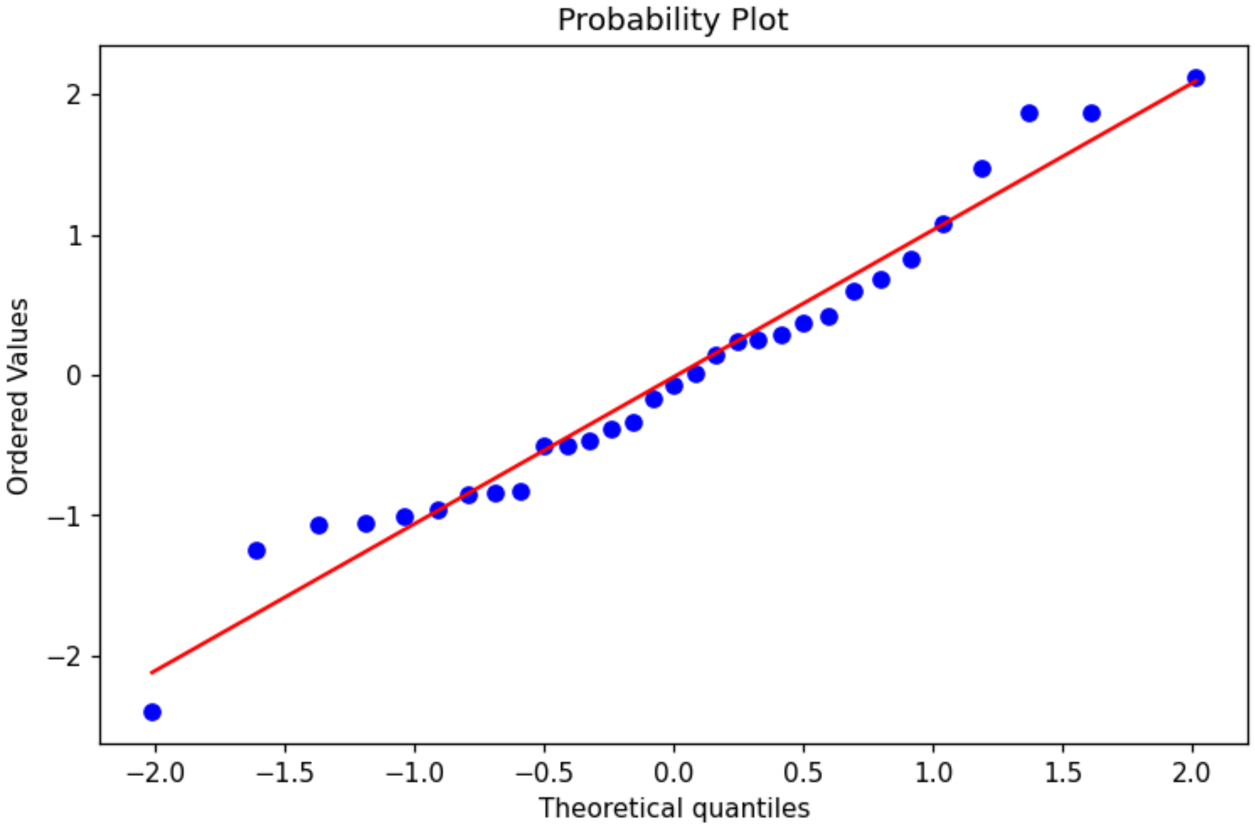
- Walker (1962) studied the mating songs of male tree crickets. Each wingstroke by a cricket produces a pulse of song, and females may use the number of pulses per second to identify males of the correct species.
- Walker wanted to know whether the chirps of the crickets *Oecanthus exclamationis* and *Oecanthus niveus* had different pulse rates.
- Measure the pulse rate of the crickets at a variety of temperatures from both species


100
90
80
70
60
50
Pulse
18
20
22
24
26
28
30
Temp
Species
ex
niv

Example: crickets - unbalanced

```
data['Species'].value_counts()
Out[115]:
niv    17
ex     14
Name: Species, dtype: int64



model= ols('Pulse ~ Temp * C(Species)', data=data).fit()
```


Residuals
3
2
1
0
−1
−2
−3
−4
50
60
70
80
90
100
Fitted Values


Probability Plot
Ordered Values
2
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Theoretical quantiles

```
                  OLS Regression Results
==============================================================================
Dep. Variable:                  Pulse   R-squared:                       0.990
Model:                            OLS   Adj. R-squared:                  0.989
Method:                 Least Squares   F-statistic:                     898.9
Date:                Wed, 01 Jan 2020   Prob (F-statistic):           3.77e-27
Time:                        00:00:00   Log-Likelihood:                -59.635
No. Observations:                  31   AIC:                             127.3
Df Residuals:                      27   BIC:                             133.0
Df Model:                           3
Covariance Type:            nonrobust
===================================================================================================
                               coef     std err         t       P>|t|     [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept                  -11.0408       4.151    -2.659      0.013    -19.559      -2.523
C(Species)[T.niv]           -4.3484       4.962    -0.876      0.389    -14.529       5.832
Temp                         3.7514       0.160    23.429      0.000      3.423       4.080
Temp:C(Species)[T.niv]      -0.2340       0.201    -1.165      0.254     -0.646       0.178
==============================================================================
Omnibus:                        0.829   Durbin-Watson:                   1.623
Prob(Omnibus):                  0.661   Jarque-Bera (JB):                0.615
Skew:                           0.334   Prob(JB):                        0.735
Kurtosis:                       2.828   Cond. No.                         531.
==============================================================================
```

Example: crickets - unbalanced

```
anova_model= anova_lm(model, typ=2)
anova_model
Out[120]:
                      sum_sq       df        F              PR(>F)
C(Species)            598.003953   1.0      189.788769     9.906686e-14
Temp                  4376.082568  1.0      1388.839184    9.350847e-25
Temp:C(Species)       4.275779     1.0      1.357006       2.542464e-01
Residual              85.074090    27.0        NaN         NaN
```

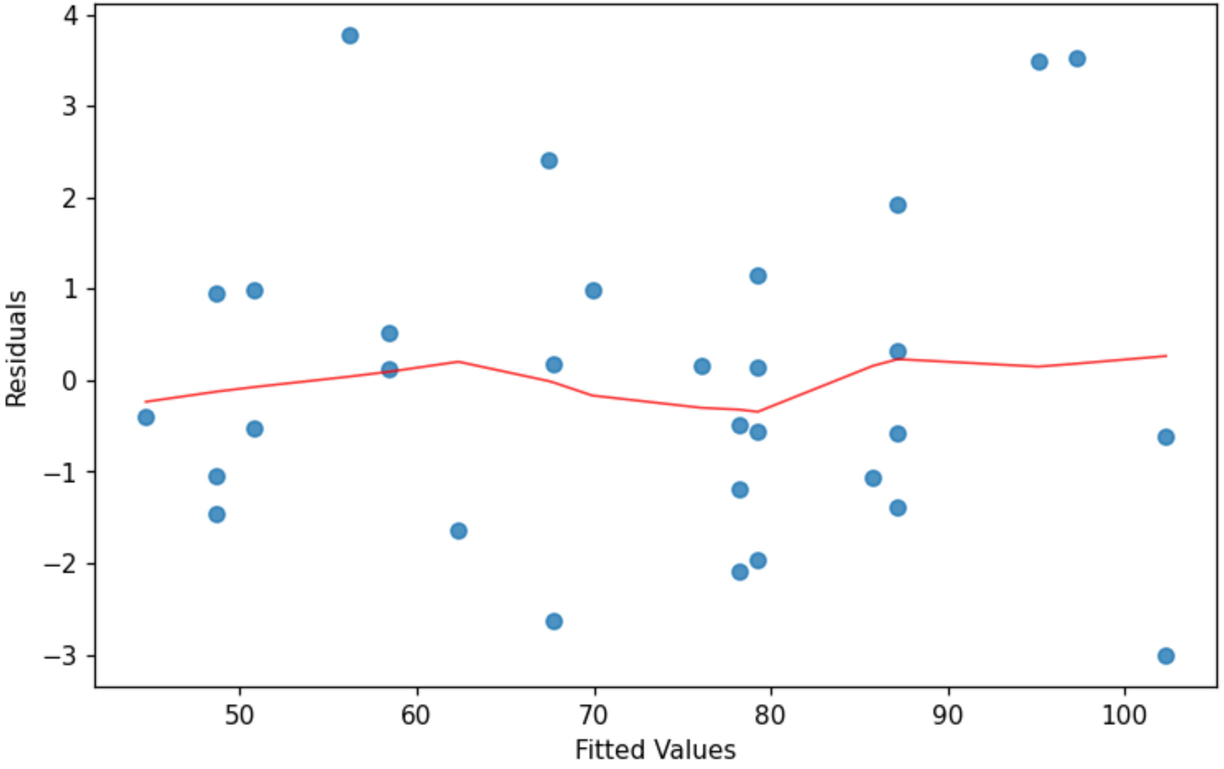
Example: crickets

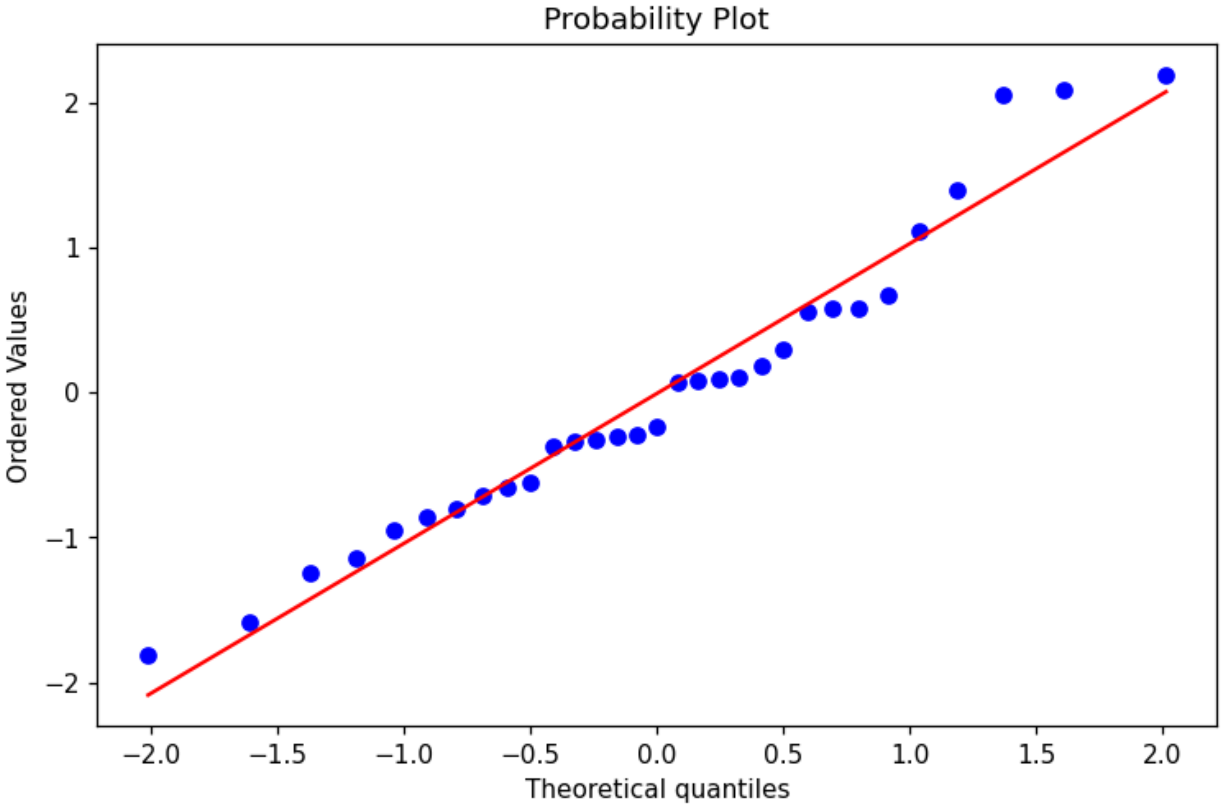
- From the ANOVA results and based on a significance level of 0.05, you can conclude that:
  - the p-value of temp is $9.351 x 10^{-25}$, which indicates that temperature and pulse have a significant relationship.
  - the p-value of Species is $9.907 x 10^{-14}$, which indicates that the levels of species are associated with significant different pulse rates.
  - the p-value for the interaction between temp*Species is 0.2542, which indicates that the interaction term is not significant, so the slope across species levels is not different.

Example: crickets

- Rerun without interaction term:

```
model2= ols('Pulse ~ Temp + C(Species)', data=data).fit()
```


Residuals
4
3
2
1
0
−1
−2
−3
50
60
70
80
90
100
Fitted Values


Probability Plot
Ordered Values
2
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Theoretical quantiles

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                  Pulse   R-squared:                       0.990
Model:                            OLS   Adj. R-squared:                  0.989
Method:                 Least Squares   F-statistic:                      1331
Date:                Wed, 01 Jan 2020   Prob (F-statistic):           1.76e-28
Time:                        00:00:00   Log-Likelihood:                -60.395
No. Observations:                  31   AIC:                             126.8
Df Residuals:                      27   BIC:                             131.1
Df Model:                           3
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept              -7.2109      2.551     -2.827      0.009     -12.436      -1.986
C(Species)[T.niv]     -10.0653      0.735    -13.689      0.000     -11.571      -8.559
Temp                    3.6028      0.097     37.032      0.000       3.403       3.802
==============================================================================
Omnibus:                        2.343   Durbin-Watson:                   1.509
Prob(Omnibus):                  0.310   Jarque-Bera (JB):                1.789
Skew:                           0.586   Prob(JB):                        0.409
Kurtosis:                       2.892   Cond. No.                         195.
==============================================================================
```

Example: crickets - unbalanced

```
anova_model2= anova_lm(model2, typ=2)
anova_model2
Out[124]:
                  sum_sq        df        F             PR(>F)
C(Species)    598.003953    1.0     187.399388   6.271533e-14
Temp          4376.082568   1.0     1371.354138  2.487732e-25
Residual      89.349869     28.0        NaN           NaN
```

Example: crickets

- From the ANOVA results and based on a significance level of 0.05, you can conclude that:
  - the p-value of temp is $2.488\text{x}10^{-25}$, which indicates that temperature and pulse have a significant relationship.
  - the p-value of Species is $6.272\text{x}10^{-14}$, which indicates that the levels of species are associated with significantly different pulse rates.
  - This leads to different intercepts among the groups

ANCOVA – three different models

1. Common slope and intercept

   $y = \beta_0 + \beta_1 x + \varepsilon$

   Here x is Temp and y is Pulse rate

2. Common slope – different intercepts

   $y = \beta_0 + \beta_1 x + \beta_2 z_2 + \varepsilon$

   Here $z_2$is an indicator variable for niv species

3. Separate lines – different intercepts and different slopes

   $y = \beta_0 + \beta_1 x + \beta_2 z_2 + \beta_3 x \times z_2 + \varepsilon$

Example: crickets

2. Common slope – different intercepts

$$y = - 7.2109 + 3.6028\, x - 10.0653 z_2 + \varepsilon$$

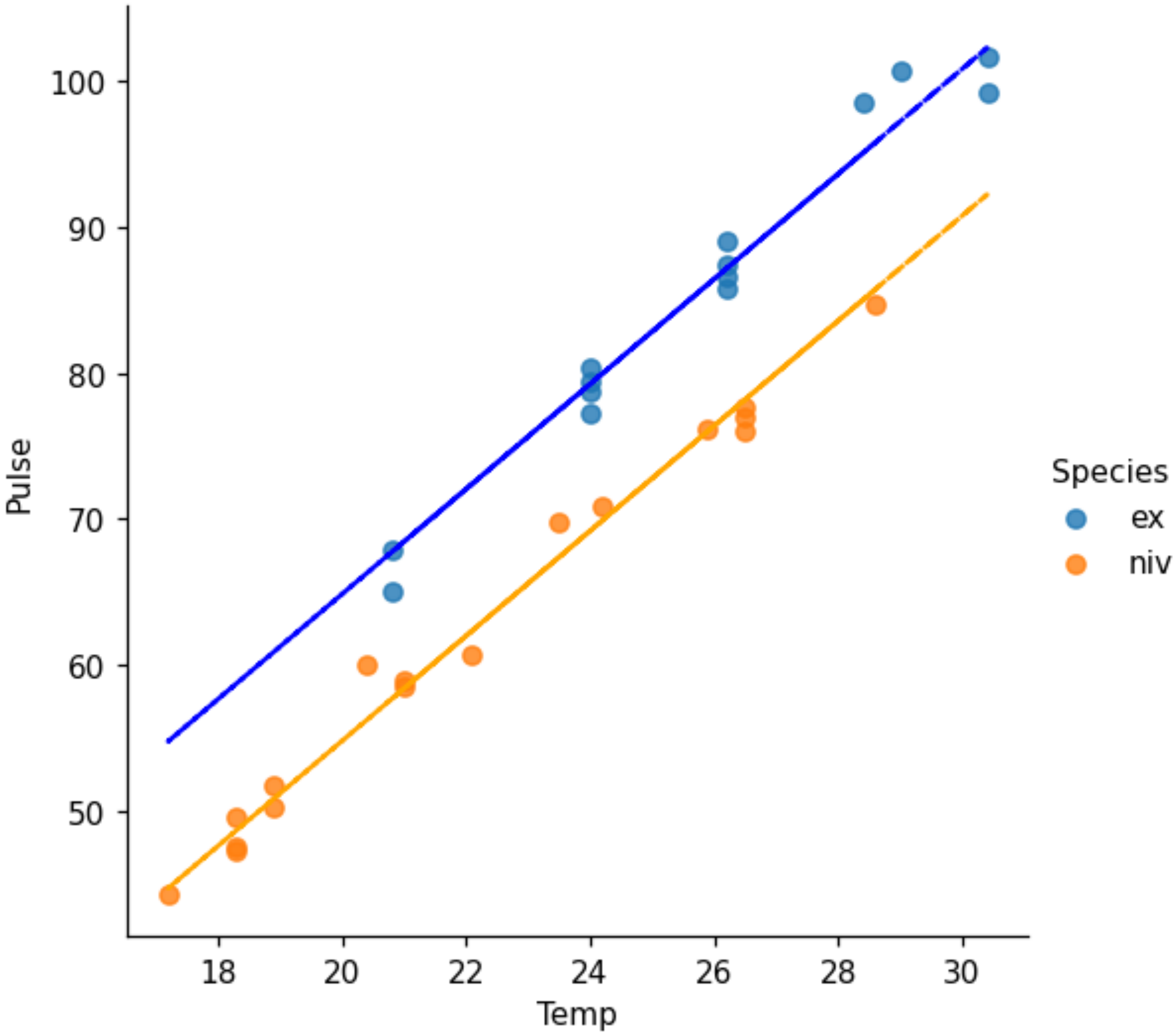
- When $z_2 = 0$:

$$y = - 7.2109 + 3.6028\, x + \varepsilon$$

- When $z_2 = 1$:

$$y = -17.2762 + 3.6028\, x + \varepsilon$$

```
sns.lmplot(x='Temp', y='Pulse', hue='Species',data=data,fit_reg=False)
x=data['Temp']
b, m =-7.2109, 3.6028,
plt.plot(x, b+ m*x, color='blue', linestyle='dashed')
b, m =-17.2762 , 3.6028
plt.plot(x, b+ m*x, color='orange', linestyle='dashed')
```


Pulse
100
90
80
70
60
50
18
20
22
24
26
28
30
Temp
Species
ex
niv