SIMPLE LINEAR REGRESSION: INFERENTIAL

The model $Y = aX + b$ now comes with assumptions on X, Y.

Use this to answer:

- How does X affect on Y?
- Predict average response $E[Y|X]$.

Model assumptions: ("LINER")

1. **Linearity:** There exist a, b such that for each x value, $E[Y] = ax + b$.
2. **Independence:** In sampling, the residuals are independent values from a $N(0, \sigma^2)$ distribution.
3. **Normality:** For each x value, Y is normally distributed.
4. **Equal variance (Homoscedasticity):** For each x value, Y has a common variance of σ^2.
5. **Random:** The data come from a well-designed random sample or randomized experiment.

Model assumptions: (more precise)

1. **Linearity:** There exist a, b such that $E[Y|X] = aX + b$.
2. **Independence:** In sampling, the residuals are independent values from a $N(0, \sigma^2)$ distribution.
3. **Normality:** $Y|(X = x)$ is normal for each x.
4. **Equal variance (Homoscedasticity):** $Var[Y|X = x] = \sigma^2$ for all x.
5. **Random:** The data come from a well-designed random sample or randomized experiment.

Model assumptions: (more precise)

$$Y|X \sim N(aX + b, \sigma^2)$$

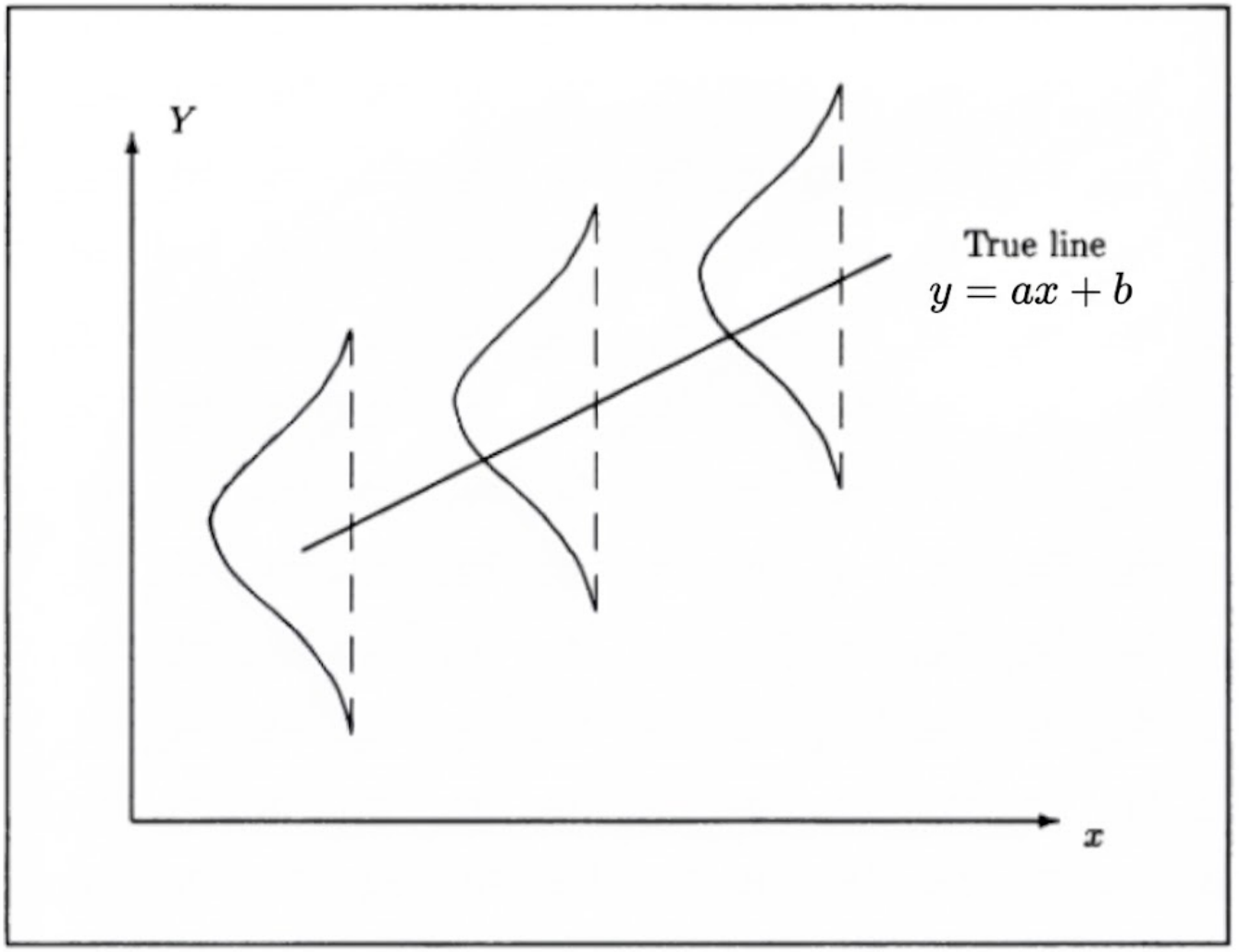


Figure 15.3: The assumptions of the linear regression model.

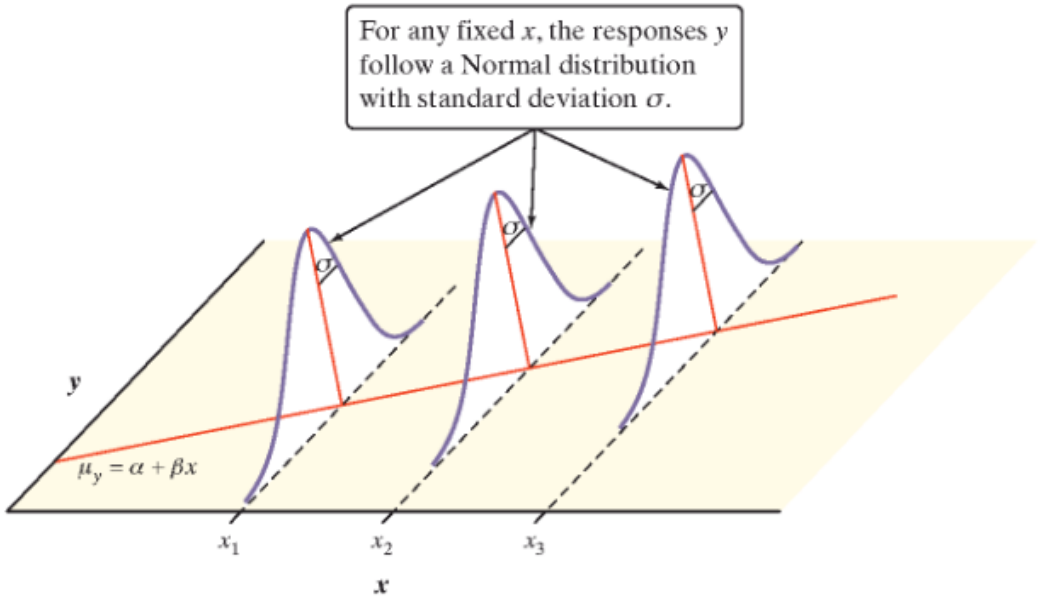


FIGURE 12.6 The regression model when the conditions for inference are met. The line is the population (true) regression line, which shows how the mean response μ_y changes as the explanatory variable x changes. For any fixed value of x, the observed response y varies according to a Normal distribution having mean μ_y and standard deviation σ.

Assumption 1: Linear. Relationship between variables X and Y is linear: $E[Y|X] = ax + b$

Check by:

- Scatterplot shows a clear linear form.

Assumption 2: Independence. Individual observations are independent of each other. When sampling without replacement, check the 10% condition.

Check by:

- If we think of x as time, then the next residual should not depend on the previous residual
- If we think of x as time, then there should not be seasonal/cyclical/wavy/sinusoidal pattern in residuals.

Assumption 3: Normal: For any fixed value of x, the response y varies according to a Normal distribution.

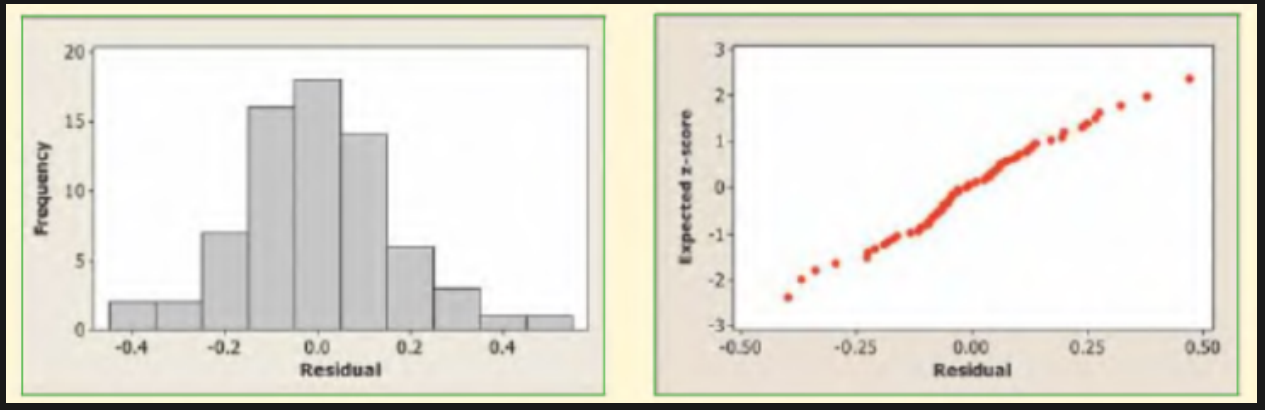
Check by:

- Histogram of residuals is single-peaked and somewhat bell-shaped.
- Normal Q-Q plot is very close to linear.

Assumption 3: Normal: For any fixed value of x, the response y varies according to a Normal distribution.

- Histogram of residuals is single-peaked and somewhat bell-shaped.
- Normal Q-Q plot is very close to linear.

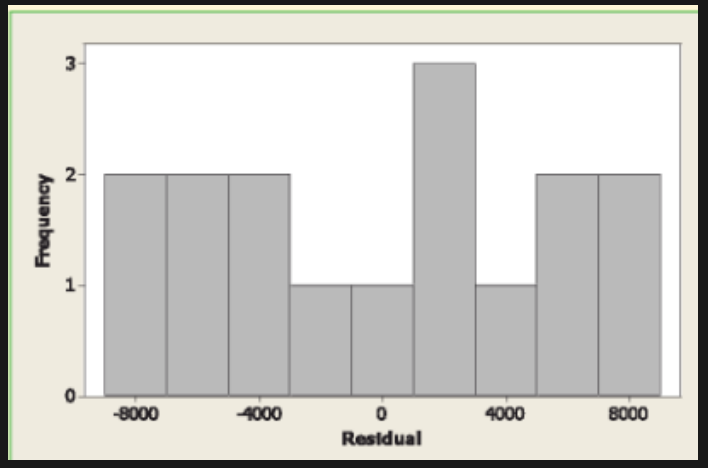
Okay:



Assumption 3: Normal: For any fixed value of x, the response y varies according to a Normal distribution.

- Histogram of residuals is single-peaked, somewhat bell-shaped.
- Normal Q-Q plot is very close to linear.

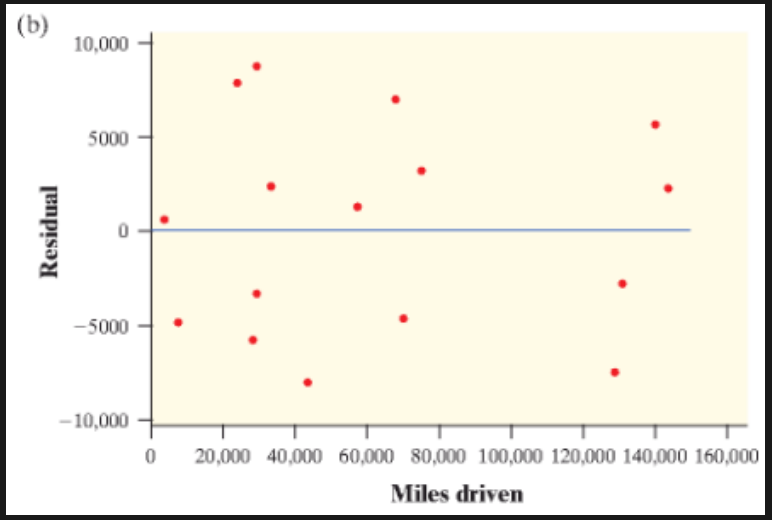
Okay:



Assumption 4: Equal SD: (Homoscedasticity): The standard deviation of y is the same for all values of x.

- Residual plot does not fan out more as you go left or right.

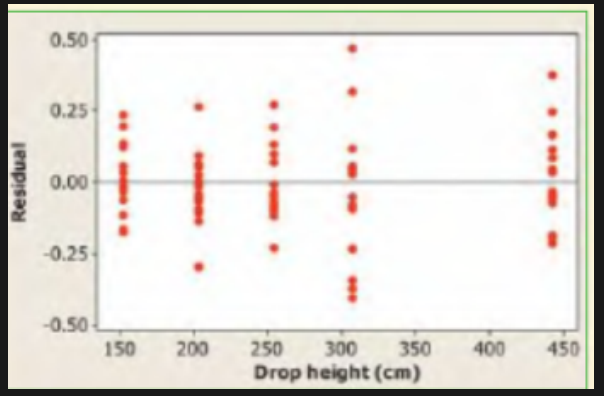
Good:



Assumption 4: Equal SD: (Homoscedasticity): The standard deviation of y is the same for all values of x.

- Residual plot does not fan out more as you go left or right.

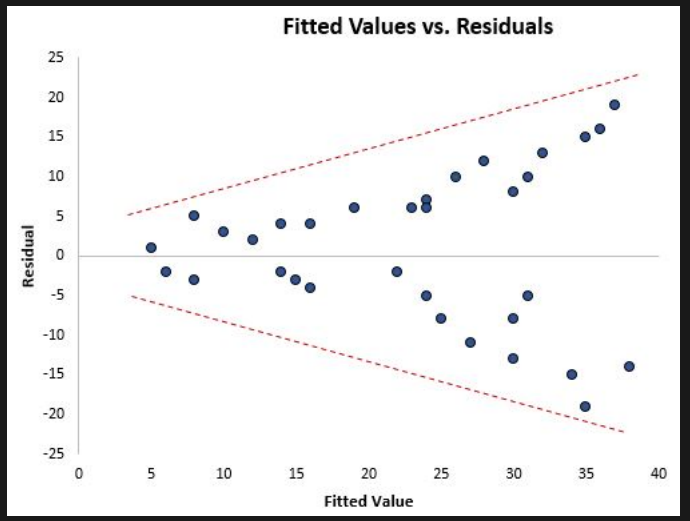
Okay:



Assumption 4: Equal SD: (Homoscedasticity): The standard deviation of y is the same for all values of x.

- Residual plot does not fan out more as you go left or right.

Assumption violated:



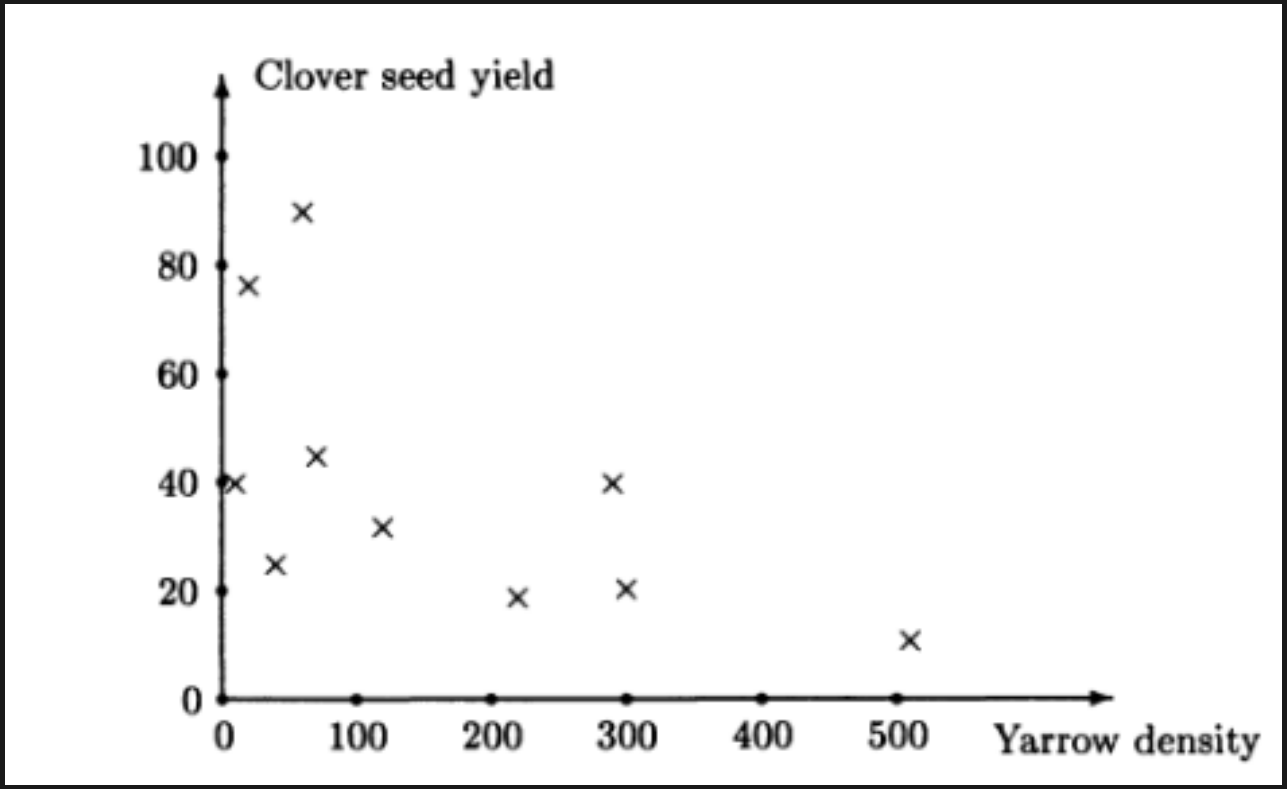
TEST HYPOTHESES ON SLOPE

- x has no effect on y
  - $\iff H_0 : \text{slope} = 0$
- "For every 1% increase in the market's return, the fund's expected return increases by exactly 1%"
  - $\iff H_0 : \text{slope} = 1$
- "The rate of improvement in test scores per additional hour of study time is identical for both tutored and untutored students." (Two sample)
  - $\iff H_0 : \text{slope}_{\text{tutored}} = \text{slope}_{\text{not tutored}}$

The purpose of the study was to estimate the effect of the agricultural weed yarrow on the yield of a white clover seed crop.

Number of yarrow flower stems per m^2	Clover seed yield in g/m^2
10	38.8
20	76.7
40	25.1
60	89.9
70	45.3
120	32.2
220	19.0
290	39.7
300	19.5
510	11.4

Scatterplot:



Enter data into SAS:

```
data yarrow_clover;
    input yarrow clover;
    datalines;
10 38.8
20 76.7
40 25.1
60 89.9
70 45.3
120 32.2
220 19.0
290 39.7
300 19.5
510 11.4
;
run;
```

Run regression analysis:

```
proc reg data=yarrow_clover;
    model log_clover = yarrow  / P;
    plot log_clover * yarrow / NOSTAT;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: clover

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2120.34039	2120.34039	4.58	0.0649
Error	8	3707.66361	463.45795		
Corrected Total	9	5828.00400			

Root MSE	21.52807	R-Square	0.3638
Dependent Mean	39.76000	Adj R-Sq	0.2843
Coeff Var	54.14505		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	55.21877	9.92874	5.56	0.0005
yarrow	1	-0.09426	0.04407	-2.14	0.0649

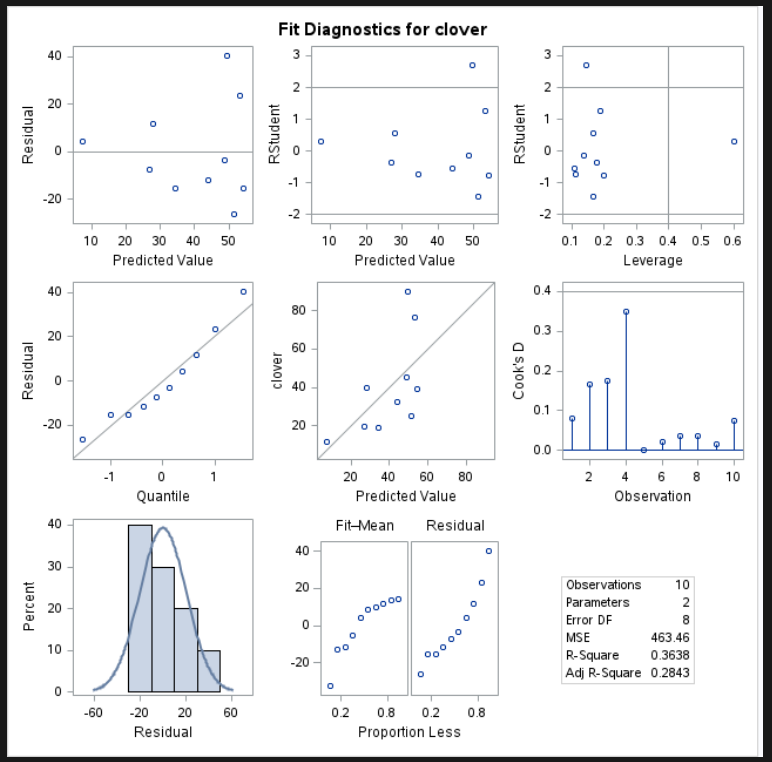
The model is $y = -3.86x + 317.46$.

The F-statistic is $F = 4.58$.

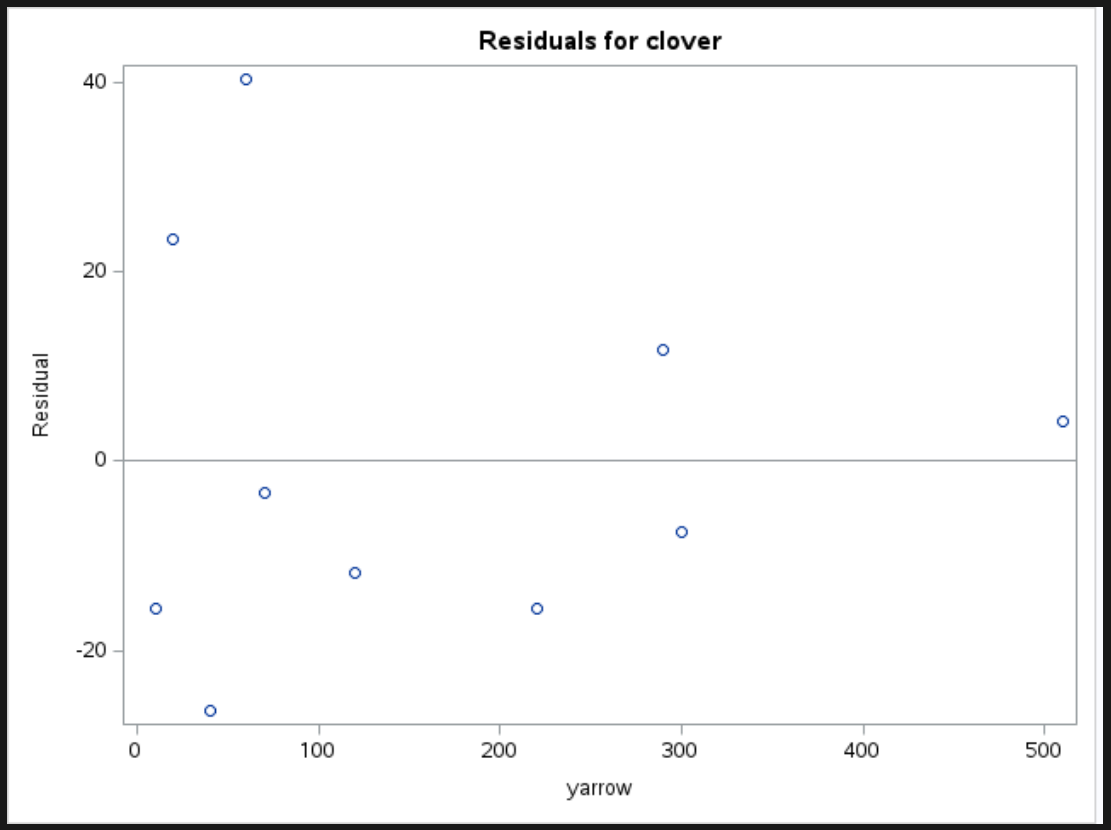
The corresponding P-value is $P = 0.0649$.

We are unable to reject the hypothesis $H_0 : b = 0$ that yarrow density has no effect on clover seed yield at the 5% significance level. (Assuming assumptions are checked).

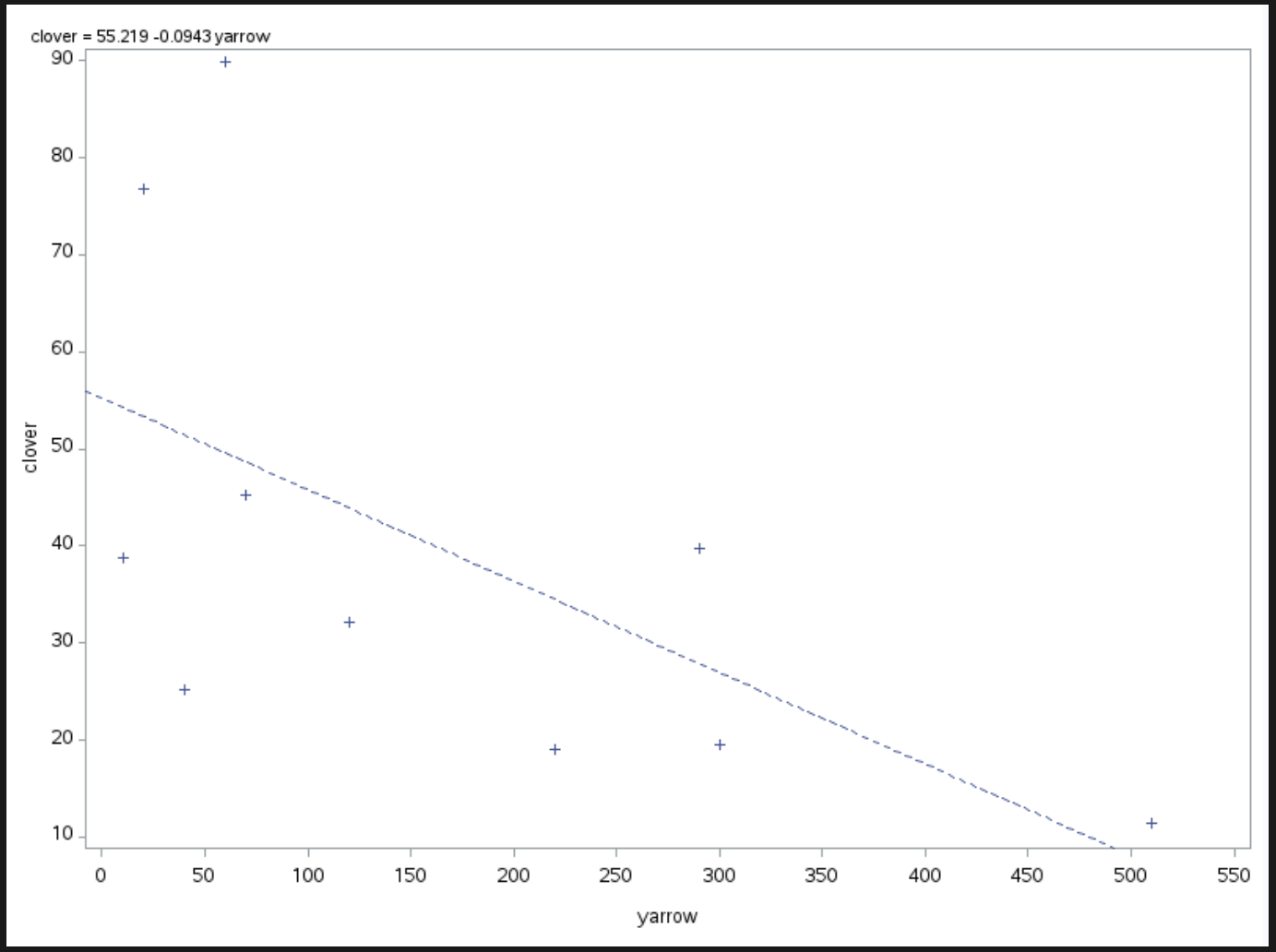
Upper left residual plot $\hat{y}$ vs $y - \hat{y}$: constant variance violated.



Residual plot x vs $y - \hat{y}$: Variance not constant, so assumption fails.



Lines fits worse on left side:



The fix: Replace y by $\log y$ to stabilize the variance.

Refitting the model:

$$\log y = ax + b$$

```
data yarrow_clover;
    input yarrow clover;
    log_clover = Log(clover);
    datalines;
10 38.8
20 76.7
40 25.1
60 89.9
70 45.3
120 32.2
220 19.0
290 39.7
300 19.5
510 11.4
;
```

```
proc reg data=yarrow_clover;
    model log_clover = yarrow  / P;
    plot log_clover * yarrow / NOSTAT;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: log_clover

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.00846	2.00846	9.71	0.0143
Error	8	1.65427	0.20678		
Corrected Total	9	3.66273			

Root MSE	0.45473	R-Square	0.5484
Dependent Mean	3.50350	Adj R-Sq	0.4919
Coeff Var	12.97945		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.97928	0.20972	18.97	<.0001
yarrow	1	-0.00290	0.00093086	-3.12	0.0143

The model is $\log y = 0.00290x + 3.979$

The F-statistic is $F = 9.71$.

The corresponding P-value is $P = 0.0143$.

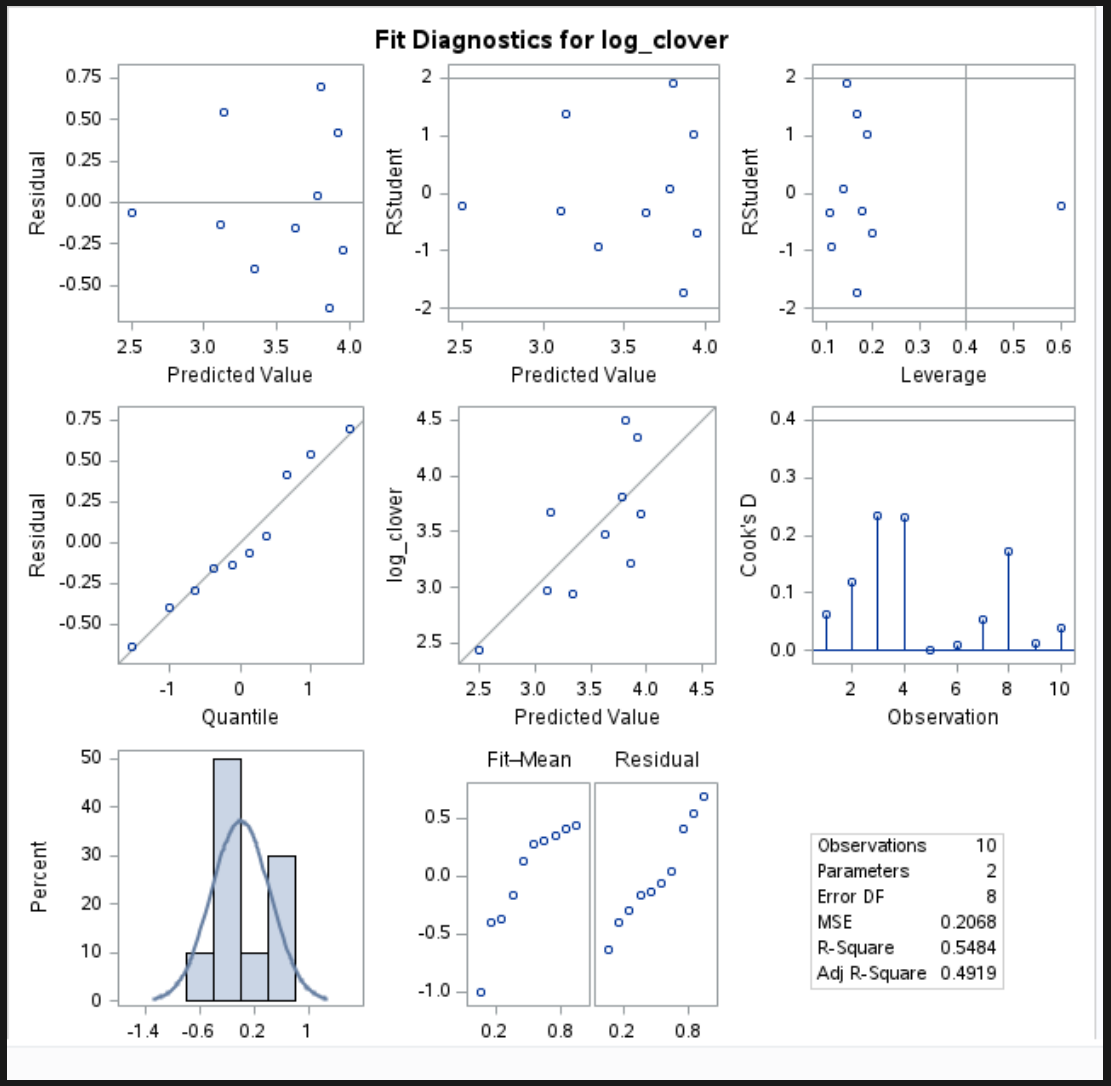
We **reject** the hypothesis $H_0 : b = 0$ that yarrow density has no effect on clover seed yield at the 5% significance level. (Assuming assumptions are checked).

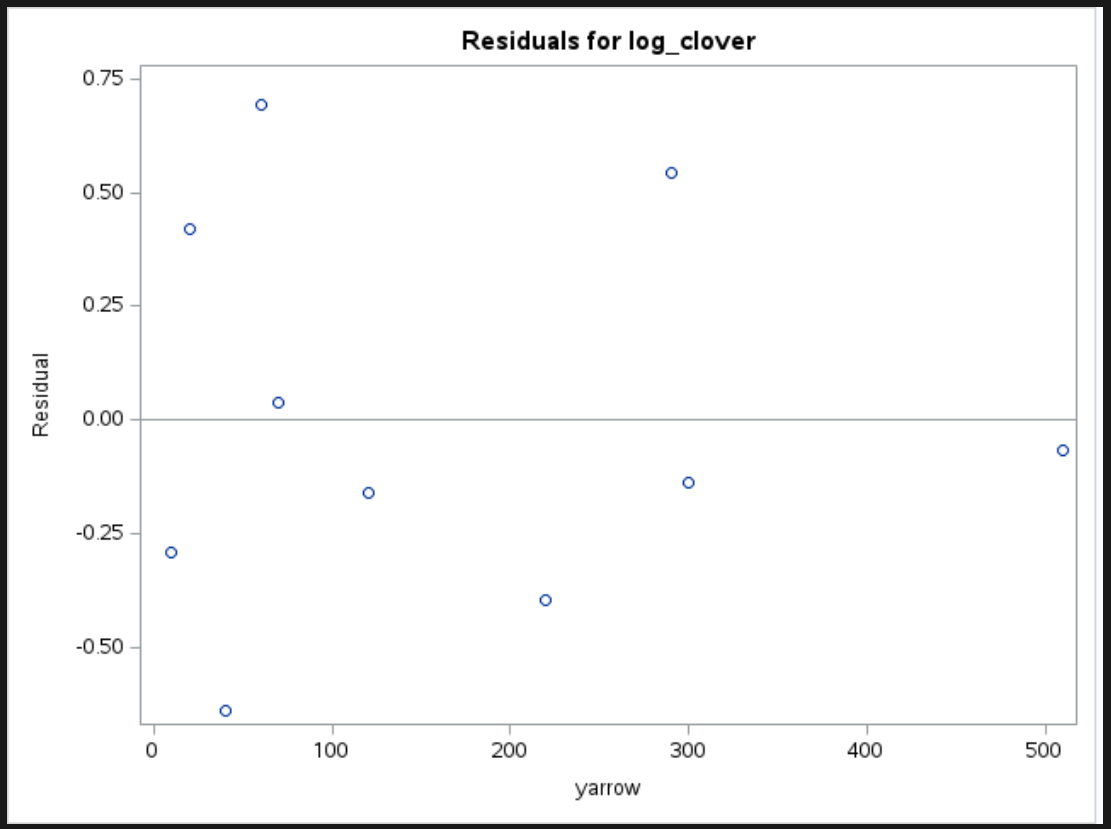
The REG Procedure
Model: MODEL1
Dependent Variable: log_clover

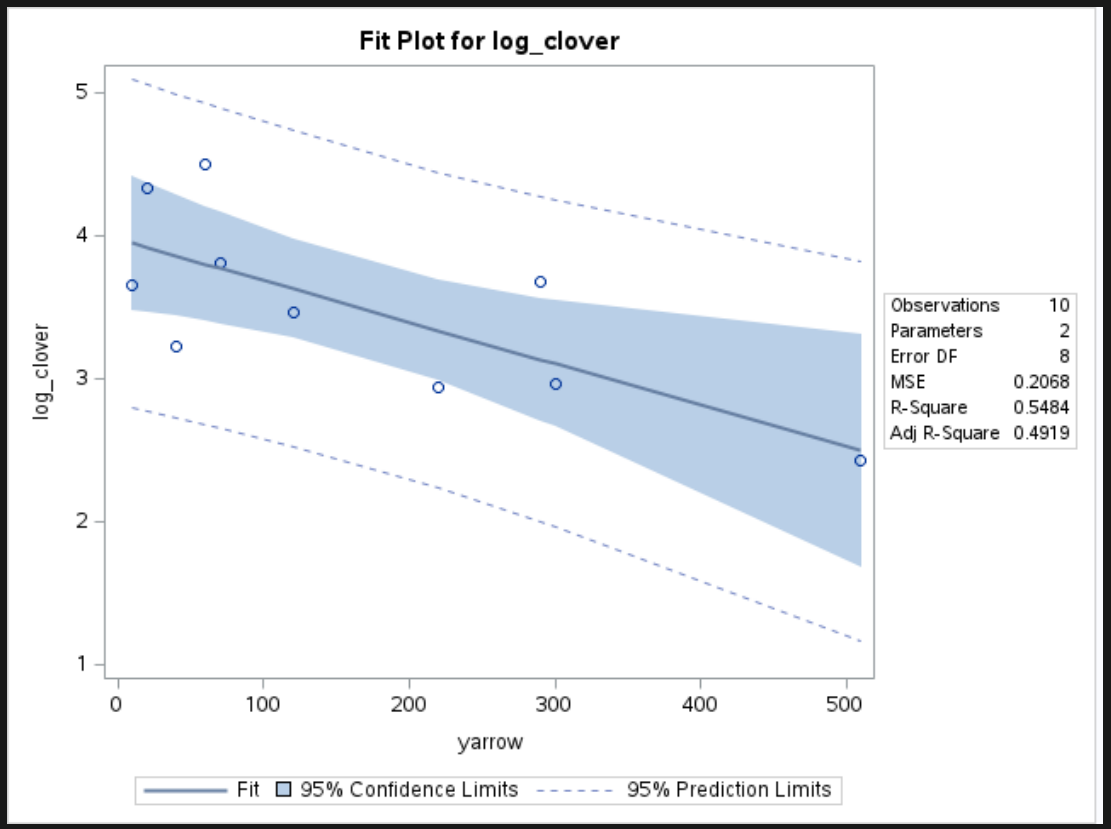
Output Statistics			
Obs	Dependent Variable	Predicted Value	Residual
1	3.66	3.9503	-0.2918
2	4.34	3.9213	0.4186
3	3.22	3.8632	-0.6404
4	4.50	3.8052	0.6935
5	3.81	3.7762	0.0371
6	3.47	3.6311	-0.1592
7	2.94	3.3410	-0.3966
8	3.68	3.1380	0.5434
9	2.97	3.1090	-0.1385
10	2.43	2.4997	-0.0661

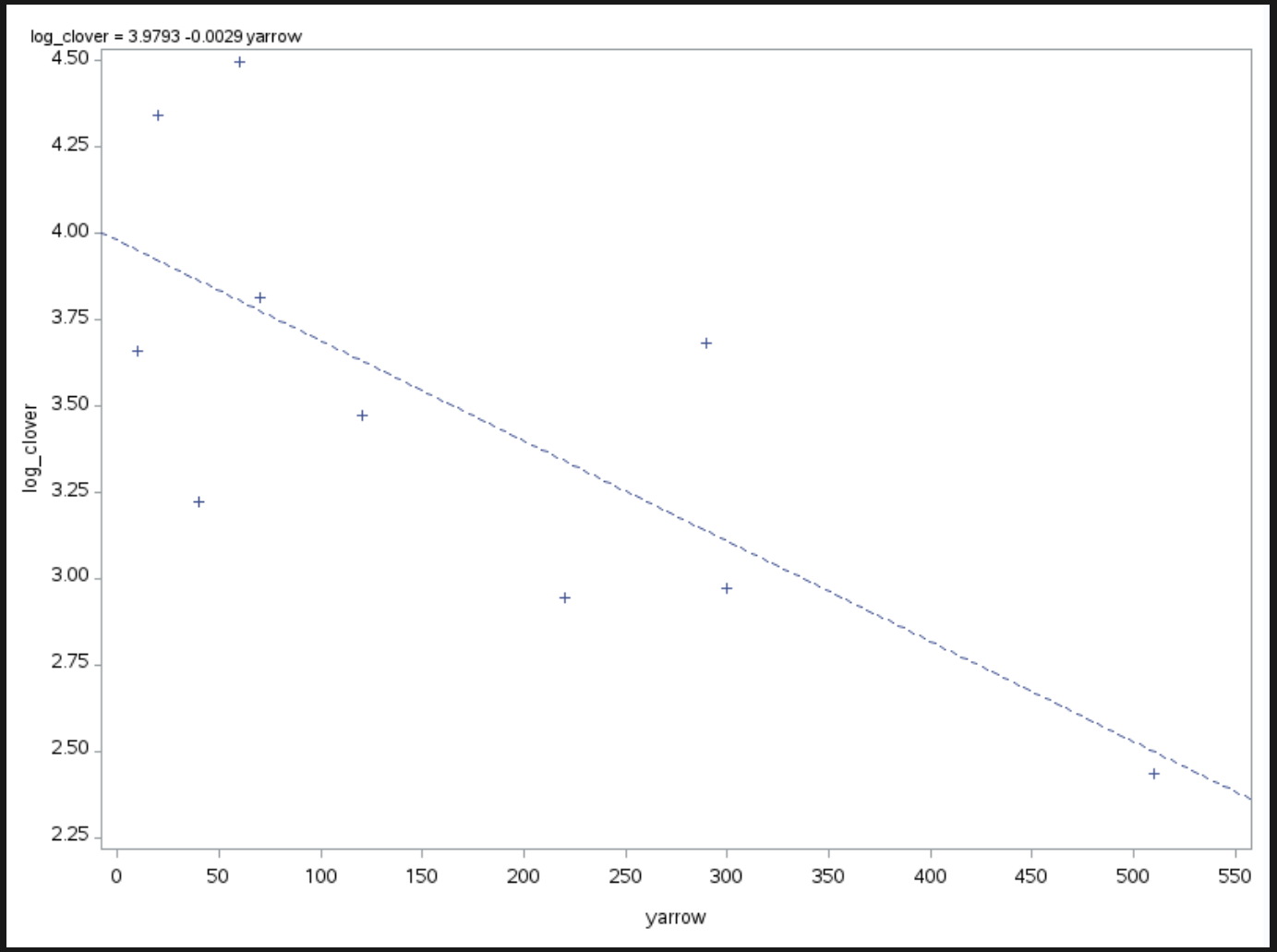
Sum of Residuals	0
Sum of Squared Residuals	1.65427
Predicted Residual SS (PRESS)	2.35834

Residuals and histogram look better now.




Residuals for log_clover
Residual
0.75
0.50
0.25
0.00
-0.25
-0.50
0
100
200
300
400
500
yarrow


Fit Plot for log_clover
log_clover
yarrow
5
4
3
2
1
0
100
200
300
400
500
Observations 10
Parameters 2
Error DF 8
MSE 0.2068
R-Square 0.5484
Adj R-Square 0.4919
Fit
95% Confidence Limits
95% Prediction Limits


log_clover = 3.9793 -0.0029 yarrow
log_clover
4.50
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
0
50
100
150
200
250
300
350
400
450
500
550
yarrow

Theory: how to get F-statistics?

$$F_{\mathrm{df}_M,\mathrm{df}_E} = \frac{\mathrm{SSM}/\mathrm{df}_M}{\mathrm{SSE}/\mathrm{df}_E} = \frac{\mathrm{MSM}}{\mathrm{MSE}} = \frac{2.00846}{0.20678} = 9.71$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.00846	2.00846	9.71	0.0143
Error	8	1.65427	0.20678		
Corrected Total	9	3.66273			

Theory: how to get F-statistics?

$$F_{\mathrm{df}_M, \mathrm{df}_E} = \frac{\mathrm{SSM}/\mathrm{df}_M}{\mathrm{SSE}/\mathrm{df}_E} = \frac{\|\hat{y} - \bar{y}\|^2/\mathrm{df}_M}{\|y - \hat{y}\|^2/\mathrm{df}_E}$$

$\|\hat{y} - \bar{y}\|^2$ is distributed as a sum of squares of df_M-many iid $N(0, \sigma^2)$'s

$\|y - \hat{y}\|^2$ is distributed as a sum of squares of df_E-many iid $N(0, \sigma^2)$'s

Notice the variances give us various statistics.

This is why it's called "analysis of variance" or ANOVA.

$$F_{\mathrm{df}_M,\mathrm{df}_E} = \frac{\mathrm{SSM}/\mathrm{df}_M}{\mathrm{SSE}/\mathrm{df}_E}$$

$$R^2 = \frac{\mathrm{SSM}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}/\mathrm{df}_E}{\mathrm{SST}/\mathrm{df}_T}$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.00846	2.00846	9.71	0.0143
Error	8	1.65427	0.20678		
Corrected Total	9	3.66273			

Problem. Manually compute F-statistics, R^2, adjusted R^2 from the ANOVA table.

Answer.

$$F_{\mathrm{df}_M, \mathrm{df}_E} = \frac{\mathrm{SSM}/\mathrm{df}_M}{\mathrm{SSE}/\mathrm{df}_E}$$

$$R^2 = \frac{\mathrm{SSM}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}/\mathrm{df}_E}{\mathrm{SST}/\mathrm{df}_T}$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.00846	2.00846	9.71	0.0143
Error	8	1.65427	0.20678		
Corrected Total	9	3.66273			

$$F_{\mathrm{df}_M,\mathrm{df}_E} = \frac{\mathrm{SSM}/\mathrm{df}_M}{\mathrm{SSE}/\mathrm{df}_E}$$

$$R^2 = \frac{\mathrm{SSM}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}/\mathrm{df}_E}{\mathrm{SST}/\mathrm{df}_T}$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.00846	2.00846	9.71	0.0143
Error	8	1.65427	0.20678		
Corrected Total	9	3.66273			

$$F_{1,8} = \frac{\mathrm{SSM}/1}{\mathrm{SSE}/8}$$

$$R^2 = \frac{\mathrm{SSM}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}/8}{\mathrm{SST}/9}$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.00846	2.00846	9.71	0.0143
Error	8	1.65427	0.20678		
Corrected Total	9	3.66273			

$$F_{1,8} = \frac{2.00846/1}{1.65427/8}$$

$$R^2 = \frac{2.00846}{3.66273} = 1 - \frac{1.65427}{3.66273}$$

$$R^2_{\text{adj}} = 1 - \frac{1.65427/8}{3.66273/9}$$

Adjusted R^2 is useful for comparing linear models with different number of parameters (next time).