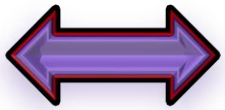
Simple Linear Regression
Part 4: Software & Diagnostic Plots

STAT 705: Regression and Analysis of Variance

Using Software

- Most of the calculations can be performed with SAS
- Other software can be used, but we will use SAS
- Software is a TOOL
  - It follows orders
  - It does not make any decisions
  - It cannot interpret results
- MAKE SURE your data has no mistakes
- MAKE SURE the computer is calculating what you intend

GARBAGE IN  **GARBAGE OUT**

A SAS Program

```
DATA example;
INPUT traffic lead;
DATALINES;
 8.1   227
 8.3   312
12.1   362
13.2   521
16.5   640
17.5   539
19.2   728
24.8   945
24.1   738
26.1   759
33.6  1263
22       .
;
SYMBOL1 V=DOT C=PURPLE;
PROC GPLOT DATA=example;
  PLOT lead*traffic;
  RUN;
PROC REG DATA=example;
 MODEL lead=traffic / P CLM CLI;
 RUN;
QUIT;
```

The DATA step defines the dataset name ('example') and the variable names ('traffic' and 'lead').

I use capital letters for SAS keywords, while lower case letters are names I chose for this example.

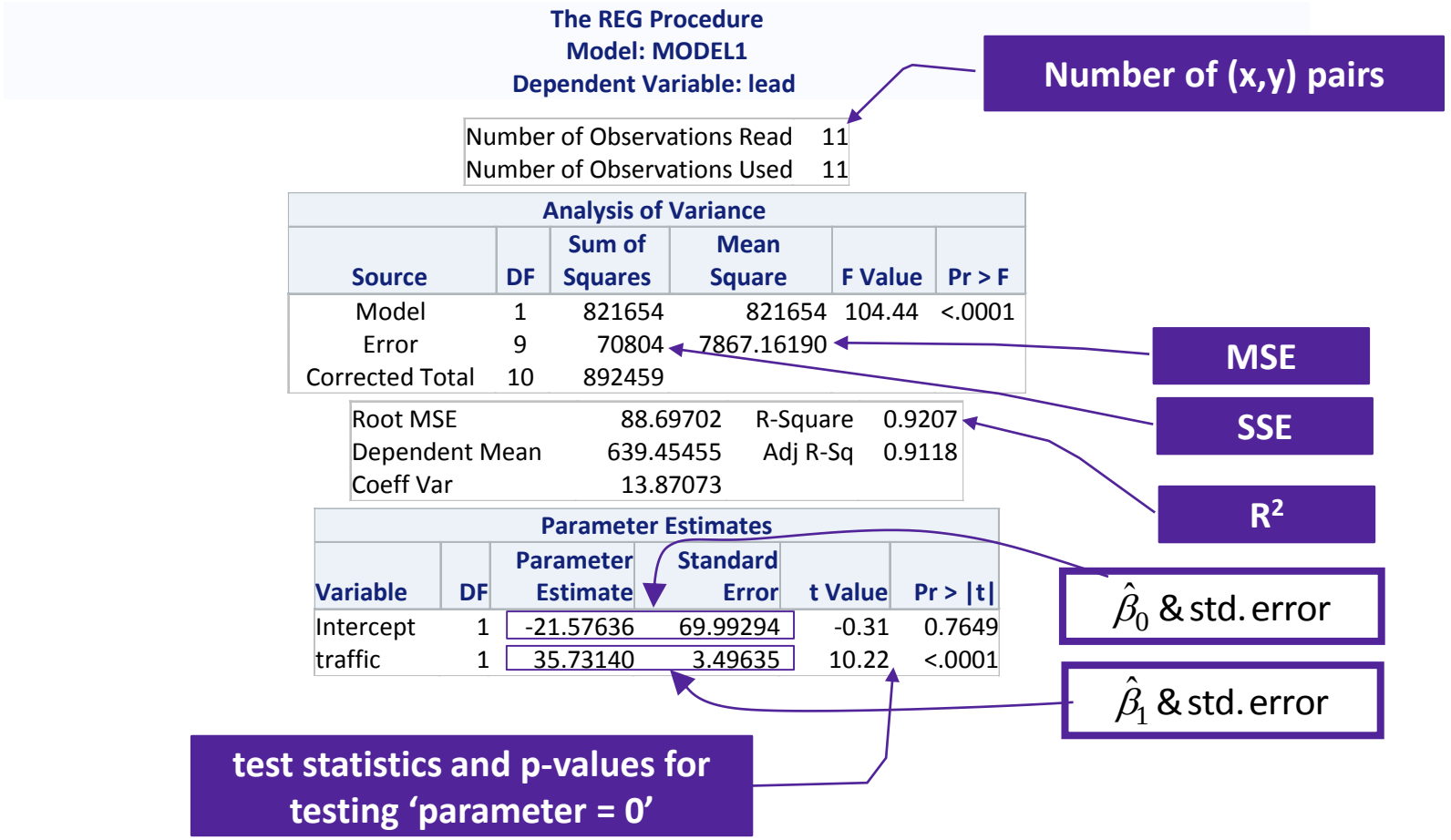
The last data line (22 and a period) are used to get a prediction interval and a confidence interval of the mean when X = 22.

Generate a scatterplot, using purple dots.
The format for the PLOT statement is Y*X.

PROC REG performs regression.
The format for the MODEL statement is Y=X.

Options for the model are after the slash.
P = predicted values
CLM = confidence limits for the mean
CLI = prediction limits for individual observations

SAS Output for PROC REG



The REG Procedure
Model: MODEL1
Dependent Variable: lead

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	821654	821654	104.44	<.0001
Error	9	70804	7867.16190		
Corrected Total	10	892459			

Root MSE	88.69702	R-Square	0.9207
Dependent Mean	639.45455	Adj R-Sq	0.9118
Coeff Var	13.87073		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.57636	69.99294	-0.31	0.7649
traffic	1	35.73140	3.49635	10.22	<.0001

Number of (x,y) pairs

MSE

SSE

R^2

$\hat{\beta}_0$ & std. error

$\hat{\beta}_1$ & std. error

test statistics and p-values for testing 'parameter = 0'

Output for Options CLM, CLI and P

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	227.0000	267.8480	45.1375	165.7399	369.9561	42.7144	492.9815	-40.8480
2	312.0000	274.9943	44.5761	174.1561	375.8324	50.4338	499.5547	37.0057
3	362.0000	410.7736	34.8699	331.8924	489.6548	195.1784	626.3688	-48.7736
4	521.0000	450.0781	32.5358	376.4769	523.6793	236.3582	663.7980	70.9219
5	640.0000	567.9917	27.6423	505.4606	630.5229	357.8270	778.1564	72.0083
6	539.0000	603.7231	26.9707	542.7111	664.7352	394.0054	813.4409	-64.7231
7	728.0000	664.4665	26.8549	603.7165	725.2166	454.8249	874.1082	63.5335
8	945.0000	864.5624	34.6466	786.1864	942.9383	649.1515	1080	80.4376
9	738.0000	839.5504	33.1445	764.5724	914.5284	625.3524	1054	-101.5504
10	759.0000	911.0132	37.6999	825.7302	996.2962	692.9943	1129	-152.0132
11	1263	1179	59.1819	1045	1313	937.7881	1420	84.0013
12	.	764.5144	29.4100	697.9845	831.0444	553.1255	975.9034	.

Sum of Residuals	0
Sum of Squared Residuals	70804
Predicted Residual SS (PRESS)	112161

Y values from data (X values are not printed)

option 'P'

option 'CLM'

option 'CLI'

Model Assumptions

- $\varepsilon_i \sim NIID\left(0, \sigma^2\right)$ encompasses three assumptions:
  1. Normal
  2. Independent
  3. Constant variance
- Should be checked before conducting any inference
- Can only be checked after fitting the model
- Methods
  - Interpret graphs (subjective)
  - Conduct formal hypothesis tests
    - Can be limited in scope, too sensitive

Checking Model Assumptions

Assumption: $\varepsilon_i \sim \text{NIID}(0, \sigma^2)$

1. Assess normality
   - Compare the residuals from the fitted model to the normal distribution
2. Assess independence
   - Consider the nature of the experiment. Non-independent data include time series data and some geographic data.
   - Objects that are "nearby" in either space or time may have values that are similar (so not independent)
3. Assess equality of variances
   - If this assumption is not satisfied, then inference that relies on the t-distribution may not be valid

more . . .

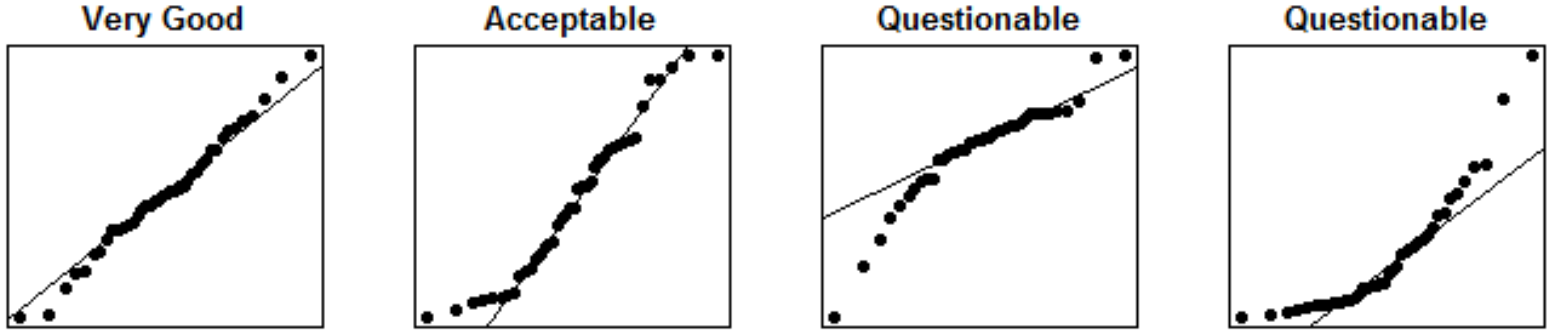
More on Model Assumptions

- Various diagnostic plots can be used to determine if these assumptions are grossly violated
- We can never be absolutely sure that these assumptions are valid for a particular population
- We look for evidence that one or more of the assumptions are grossly violated
- If the assumptions of the model are questionable, we may need to adjust the model or adjust the data (more on this later)

Assess Normality

- Normal probability plot … or … QQ plot
  - If points follow the line ⇨ residuals could be normal
  - Serious departures from the line ⇨ residuals probably not normal
- Formal hypothesis tests (e.g. Shapiro-Wilk), but these tend to be sensitive to the sample size

QQ plots



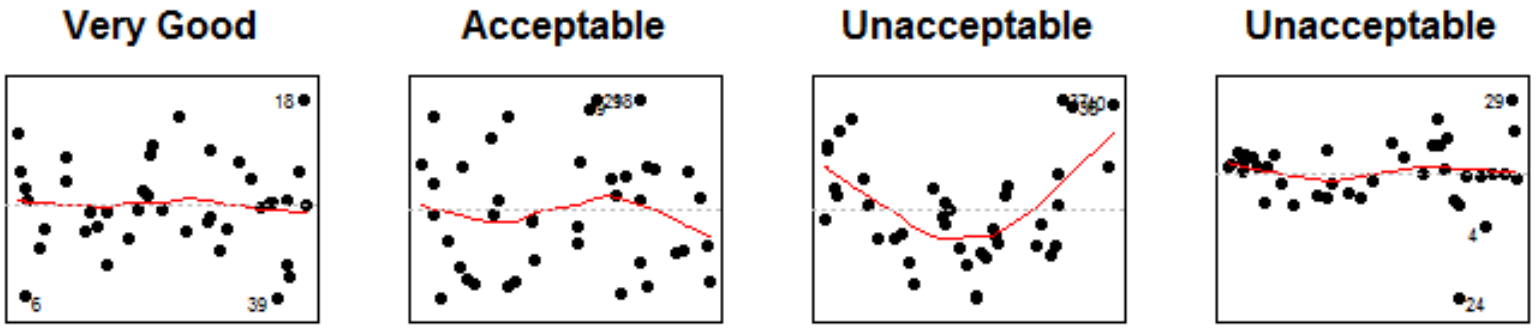
Theoretical (normal) quantiles on x-axis; Residual quantiles on y-axis

Assess Equality of Variances

- Homoscedastic ⇨ variances are equal
- Heteroscedastic ⇨ variances are not equal
- Assumptions on independence and variances can both be examined with a residual plot
  - On x-axis: observed Y, or fitted Y
  - On y-axis: residual, perhaps standardized
- There are formal hypothesis tests for equality of variances (e.g., Brown-Forsythe), but these require several observations for each value of X

Fitted vs. Residual Plots

- We want the points to show no obvious patterns
- Quadratic patterns (as in the 3rd graph) indicate the linear model is a poor fit to the data
- Wedge-shaped patterns (as in the 4th graph) indicate the variances are not all equal



Fitted values are on x-axis; Residuals are on y-axis.

Generating Diagnostic Plots in SAS

```
. . . DATA step goes here . . .

SYMBOL1 V=DOT C=PURPLE;
PROC REG DATA=example;
 MODEL lead=traffic / P CLM CLI;
 OUTPUT OUT=diagnostics
        RESIDUAL = resid
        PREDICTED = fitted;
 RUN;

PROC GPLOT DATA=diagnostics;
  PLOT resid*fitted;
  RUN;

PROC UNIVARIATE DATA=diagnostics NOPRINT;
 QQPLOT resid / NORMAL(MU=EST SIGMA=EST COLOR=BLACK);
   RUN;

QUIT;
```

In PROC REG, the OUTPUT statement creates a new SAS dataset. We have called the dataset 'diagnostics' and it contains the residuals and the predicted values from the fitted model. In the new dataset, these variables are called 'resid' and 'fitted', respectively.

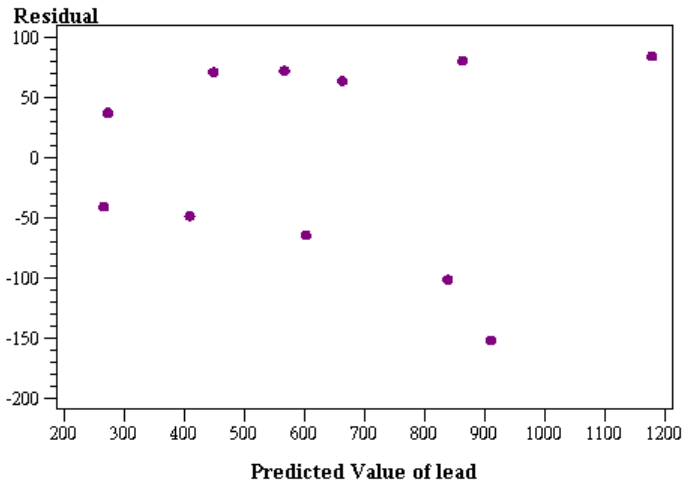
PROC GPLOT uses our newly-created dataset and generates the fitted vs. residual plot.

PROC UNIVARIATE generates a lot of printed output; the NOPRINT option suppresses that.

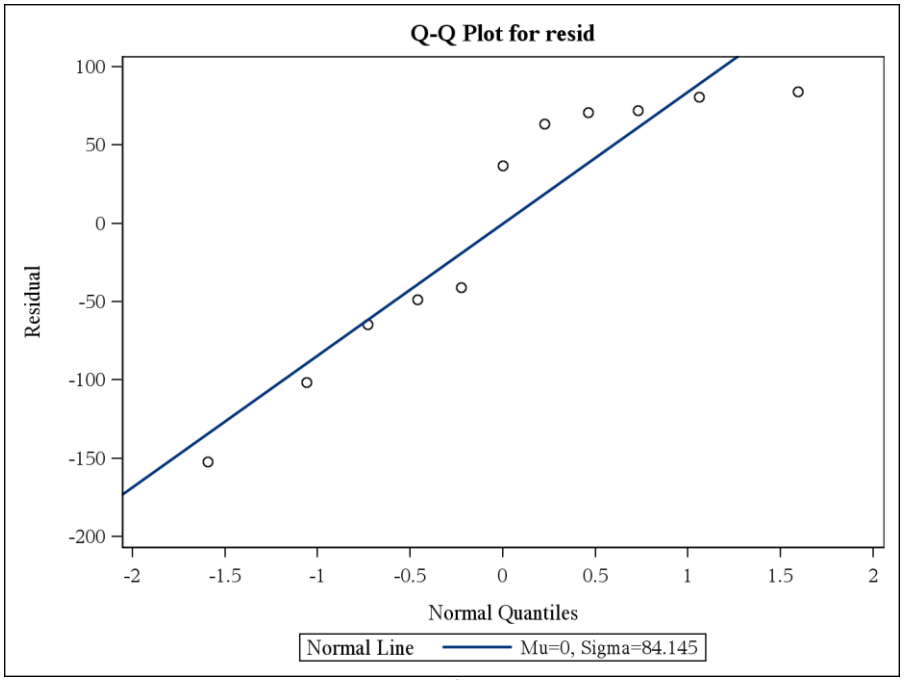
On the QQPLOT statement, the NORMAL option produces the line on the QQ plot.

Diagnostic Plots from SAS

Lead vs. Traffic Example



Suggests variances are not all the same. We will discuss possible corrective action in a later lesson.



Acceptable, considering the sample size is so small

Things You Should Know

- Access SAS and run the provided code
  - TrafficLead.sas
- Interpret normal probability plot
- Interpret residual plot