NONLINEAR REGRESSION:

BUCKLEY'S USED CARS

Jack Buckley has owned a large used car lot outside Melbourne, Australia for over 30 years. As a business person he likes to keep track of how many cars a salesperson sells per week.

BUCKLEY'S USED CARS

Jack Buckley has owned a large used car lot outside Melbourne, Australia for over 30 years. As a business person he likes to keep track of how many cars a salesperson sells per week.

Jack would like to examine the relationship between how many total cars have been sold by each salesperson and how many weeks each salesperson has worked for Buckley's Used Cars.

BUCKLEY'S USED CARS

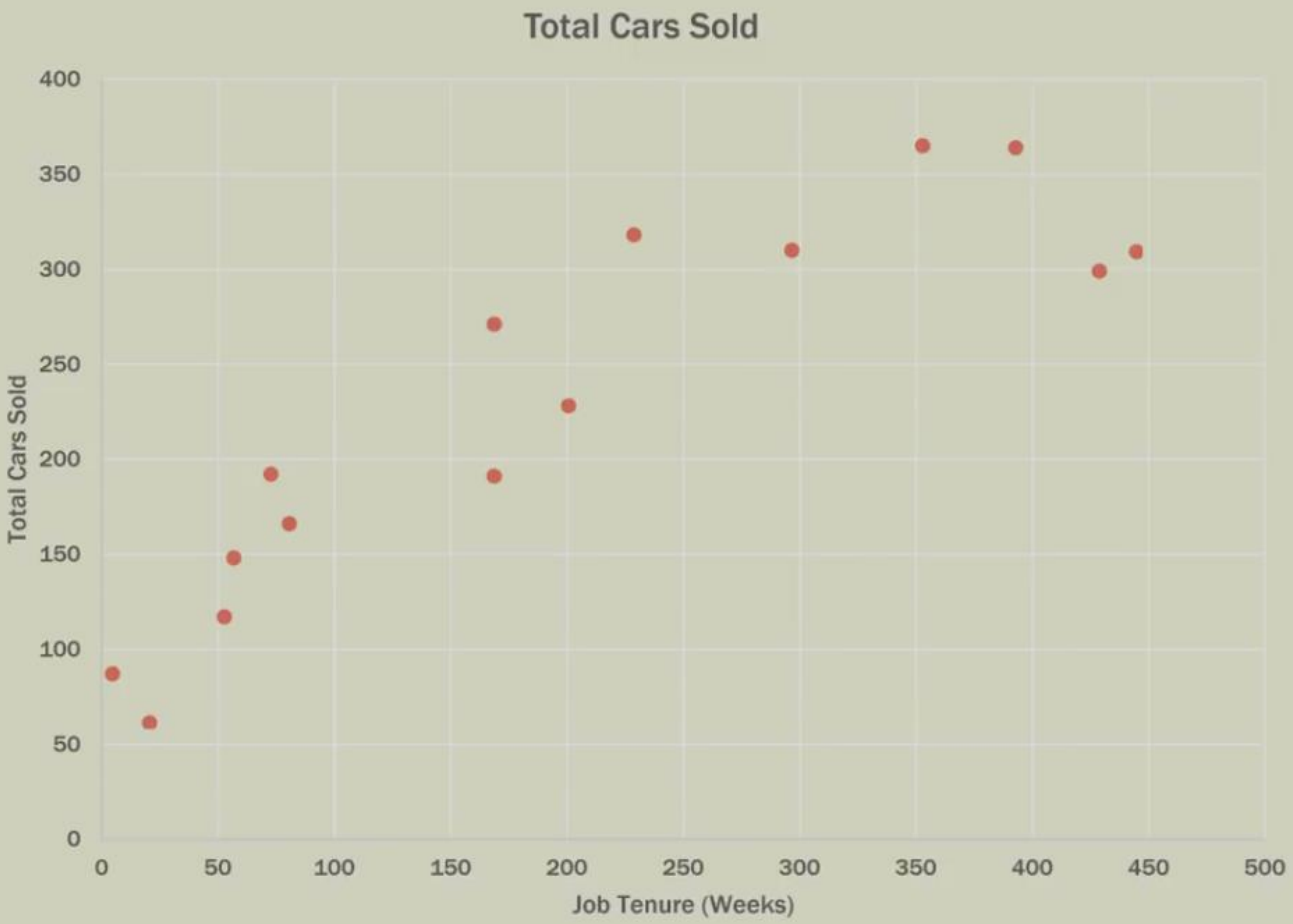
Jack Buckley has owned a large used car lot outside Melbourne, Australia for over 30 years. As a business person he likes to keep track of how many cars a salesperson sells per week.

Jack would like to examine the relationship between how many total cars have been sold by each salesperson and how many weeks each salesperson has worked for Buckley's Used Cars.

GOAL: Produce a model that MINIMIZES ERROR but will also be good for NEW DATA.

Job Tenure (Weeks)	Total Cars Sold
168	272
428	300
296	311
392	365
80	167
56	149
352	366
444	310
168	192
200	229
4	88
52	118
20	62
228	319
72	193

Job Tenure (Weeks)	Total Cars Sold
168	272
428	300
296	311
392	365
80	167
56	149
352	366
444	310
168	192
200	229
4	88
52	118
20	62
228	319
72	193




Total Cars Sold vs Job Tenure
y = 0.5823x + 114.5
R² = 0.8019
Total Cars Sold
Job Tenure (Weeks)
0
50
100
150
200
250
300
350
400
450
500

LINEAR REGRESSION OUTPUT

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.895504014
R Square	0.801927439
Adjusted R Square	0.786691089
Standard Error	45.94352485
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	1	111097.1028	111097.1028	52.63251344	6.41213E-06
Residual	13	27440.49718	2110.807475		
Total	14	138537.6			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	114.4963248	19.78813487	5.786109985	6.31907E-05	71.74665841	157.2459911
Job Tenure (Weeks)	0.582282138	0.080261341	7.254826906	6.41213E-06	0.408888052	0.755676224

LINEAR REGRESSION OUTPUT

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.895504014
R Square	0.801927439
Adjusted R Square	0.786691089
Standard Error	45.94352485
Observations	15

At first glance this looks like a pretty good model!

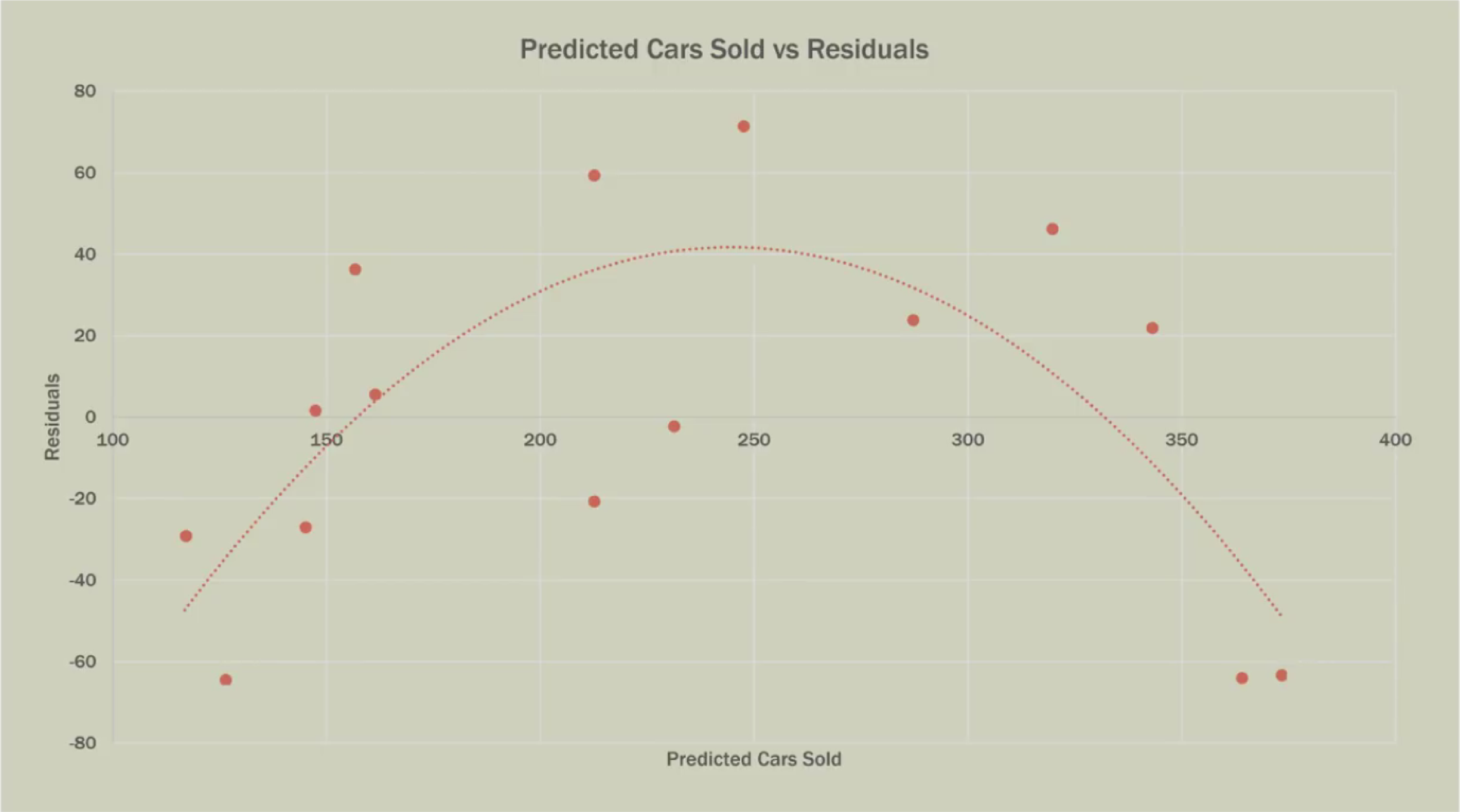
But is it the best model? Let's examine residuals.

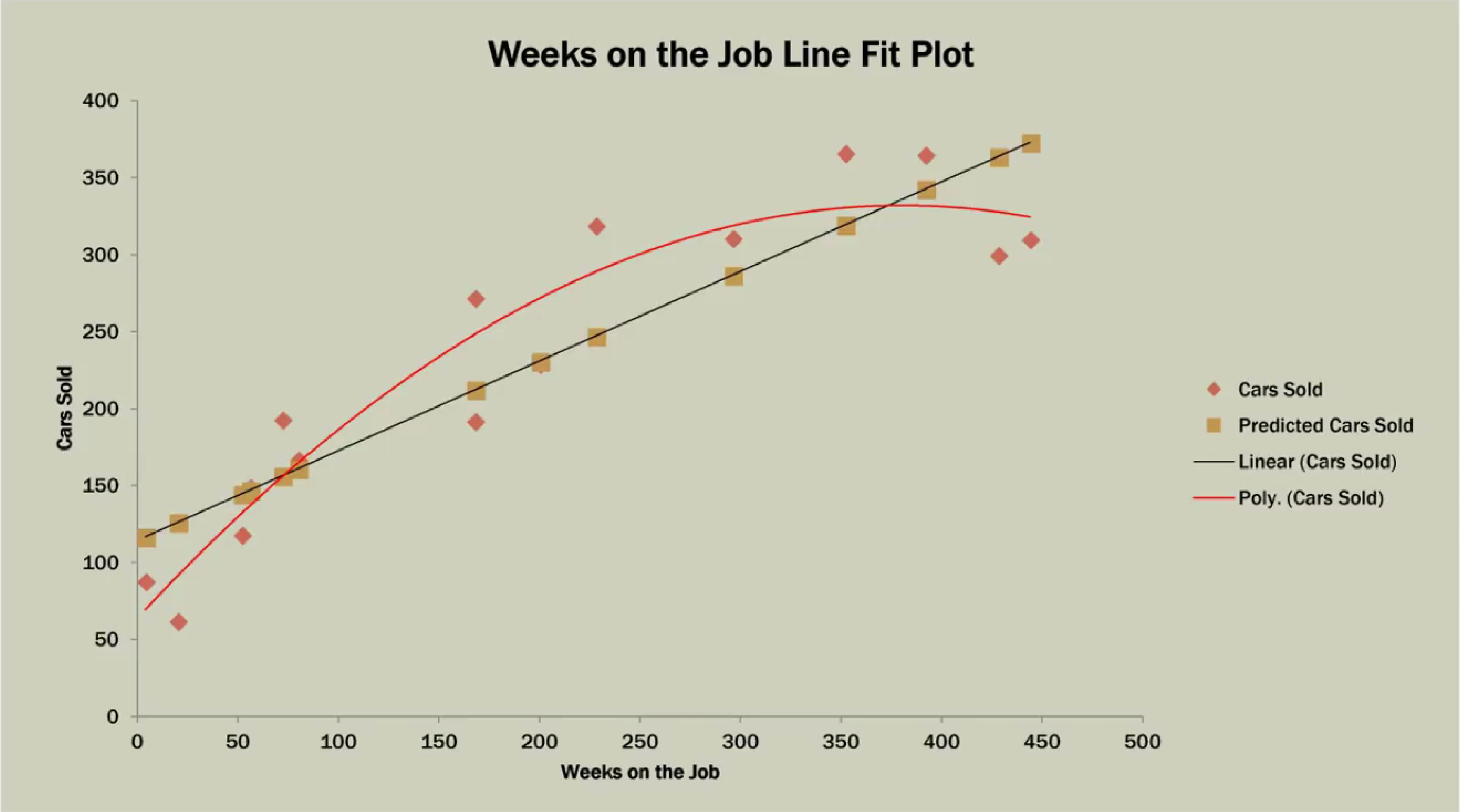
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	111097.1028	111097.1028	52.63251344	6.41213E-06
Residual	13	27440.49718	2110.807475		
Total	14	138537.6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	114.4963248	19.78813487	5.786109985	6.31907E-05	71.74665841	157.2459911
Job Tenure (Weeks)	0.582282138	0.080261341	7.254826906	6.41213E-06	0.408888052	0.755676224


Predicted Cars Sold vs Residuals
80
60
40
20
0
-20
-40
-60
-80
Residuals
100
150
200
250
300
350
400
Predicted Cars Sold


Predicted Cars Sold vs Residuals
Residuals
Predicted Cars Sold
80
60
40
20
0
-20
-40
-60
-80
100
150
200
250
300
350
400


Weeks on the Job Line Fit Plot
Cars Sold
400
350
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
450
500
Weeks on the Job
Cars Sold
Predicted Cars Sold
Linear (Cars Sold)
Poly. (Cars Sold)

THE NONLINEAR MODEL

- Based off the original scatterplot and even more so the residual plot, it appears that a linear model is **not the best choice to model the data**

THE NONLINEAR MODEL

- Based off the original scatterplot and even more so the residual plot, it appears that a linear model is **not the best choice to model the data**
- Too much reducible error!

THE NONLINEAR MODEL

- Based off the original scatterplot and even more so the residual plot, it appears that a linear model is **not the best choice to model the data**
- Too much reducible error!
- The residual plot shows a definite curvature

THE NONLINEAR MODEL

- Based off the original scatterplot and even more so the residual plot, it appears that a linear model is **not the best choice to model the data**
- Too much reducible error!
- The residual plot shows a definite curvature
- This indicates that the best model may be non-linear
- Polynomial regression adds extra independent variables that are powers of the original variable; x, x^2, x^3 etc.

QUADRATIC REGRESSION MODEL

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

- We now include in the model the square of the independent variable, x_1^2

QUADRATIC REGRESSION MODEL

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

- We now include in the model the square of the independent variable, x_1^2
- This will allow our model to capture the curvature in the scatterplot
- Quadratic models are pretty flexible, since higher order models such as x_1^3 can be inserted to the model

QUADRATIC REGRESSION MODEL

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

- We now include in the model the square of the independent variable, x_1^2
- This will allow our model to capture the curvature in the scatterplot
- Quadratic models are pretty flexible, since higher order models such as x_1^3 can be inserted to the model
- CAREFUL…adding terms can lead to overfitting

Weeks on the Job	Total Cars Sold
168	272
428	300
296	311
392	365
80	167
56	149
352	366
444	310
168	192
200	229
4	88
52	118
20	62
228	319
72	193

Weeks on the Job	Total Cars Sold
168	272
428	300
296	311
392	365
80	167
56	149
352	366
444	310
168	192
200	229
4	88
52	118
20	62
228	319
72	193

Weeks on the Job	Weeks^2	Total Cars Sold
168	28224	272
428	183184	300
296	87616	311
392	153664	365
80	6400	167
56	3136	149
352	123904	366
444	197136	310
168	28224	192
200	40000	229
4	16	88
52	2704	118
20	400	62
228	51984	319
72	5184	193

QUADRATIC MODEL OUTPUT

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95263863
R Square	0.90752036
Adjusted R Square	0.892107086
Standard Error	32.67505089
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	125725.6926	62862.8463	58.87914511	6.25571E-07
Residual	12	12811.90741	1067.658951		
Total	14	138537.6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	63.85096934	19.62804309	3.253048154	0.006917223	21.08513723	106.6168014	21.08513723	106.6168014
Weeks on the Job	1.409452543	0.23064056	6.111035036	5.24735E-05	0.906929932	1.911975155	0.906929932	1.911975155
Weeks^2	-0.001852148	0.000500369	-3.701561402	0.003027121	-0.002942359	-0.000761937	-0.002942359	-0.000761937

QUADRATIC MODEL OUTPUT

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95263863
R Square	0.90752036
Adjusted R Square	0.892107086
Standard Error	32.67505089
Observations	15

Much better fit than the linear model!

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	125725.6926	62862.8463	58.87914511	6.25571E-07
Residual	12	12811.90741	1067.658951		
Total	14	138537.6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	63.85096934	19.62804309	3.253048154	0.006917223	21.08513723	106.6168014	21.08513723	106.6168014
Weeks on the Job	1.409452543	0.23064056	6.111035036	5.24735E-05	0.906929932	1.911975155	0.906929932	1.911975155
Weeks^2	-0.001852148	0.000500369	-3.701561402	0.003027121	-0.002942359	-0.000761937	-0.002942359	-0.000761937

LINEAR VS QUADRATIC

Linear Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.895504014
R Square	0.801927439
Adjusted R Square	0.786691089
Standard Error	45.94352485
Observations	15

Quadratic Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95263863
R Square	0.90752036
Adjusted R Square	0.892107086
Standard Error	32.67505089
Observations	15

LINEAR VS QUADRATIC

Linear Model

SUMMARY OUTPUT

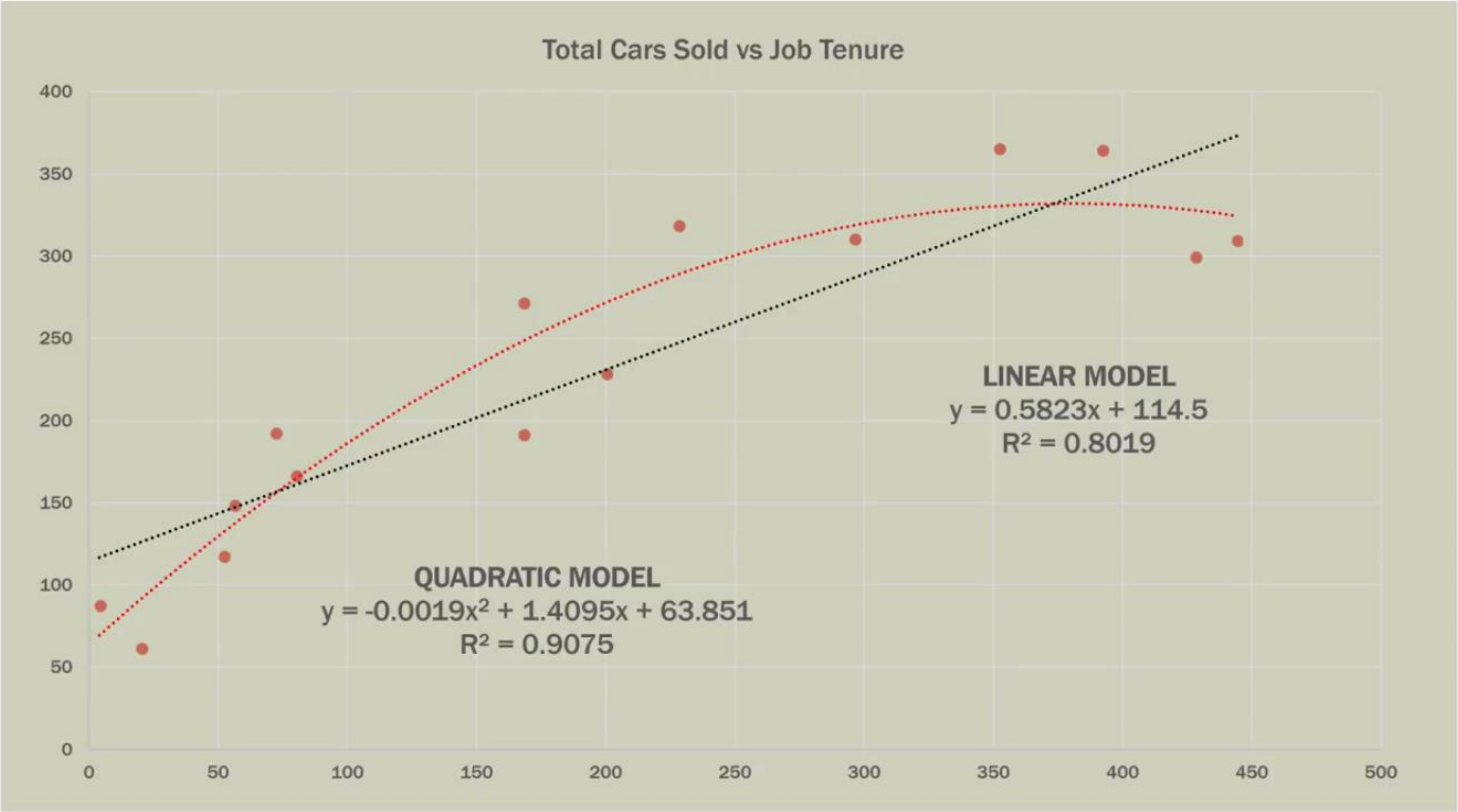
Regression Statistics	
Multiple R	0.895504014
R Square	0.801927439
Adjusted R Square	0.786691089
Standard Error	45.94352485
Observations	15

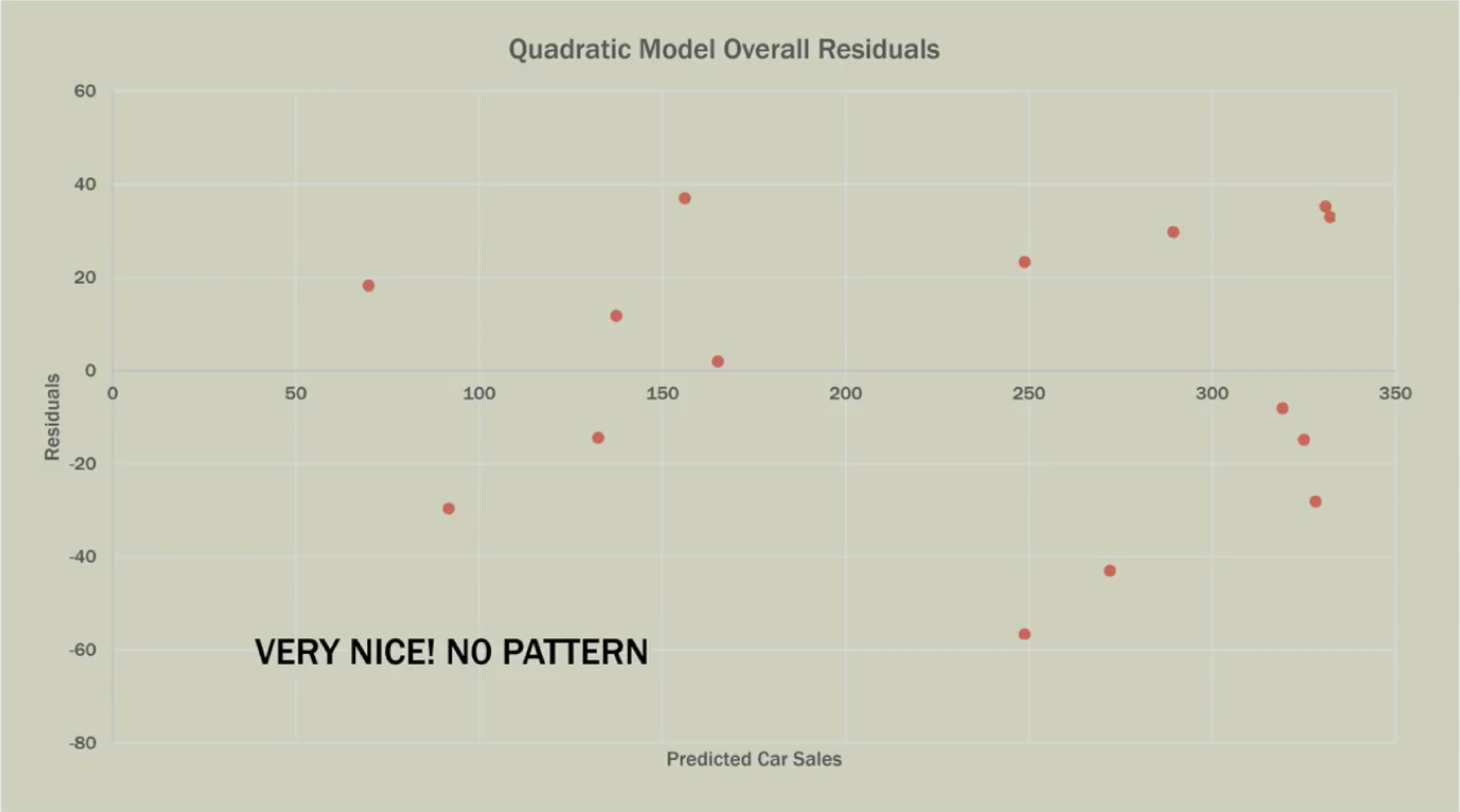
Quadratic Model

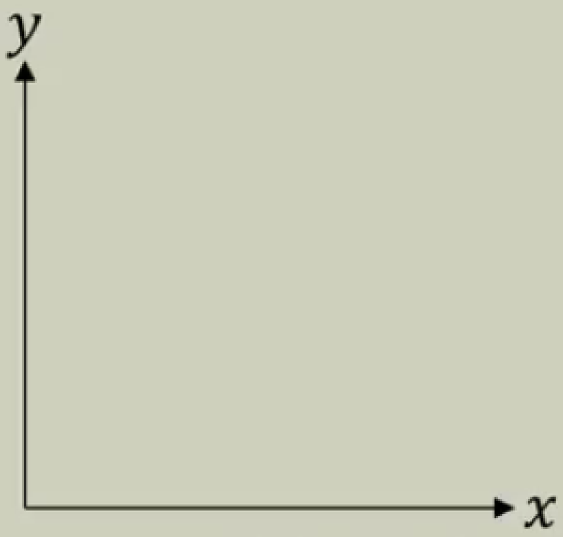
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95263863
R Square	0.90752036
Adjusted R Square	0.892107086
Standard Error	32.67505089
Observations	15

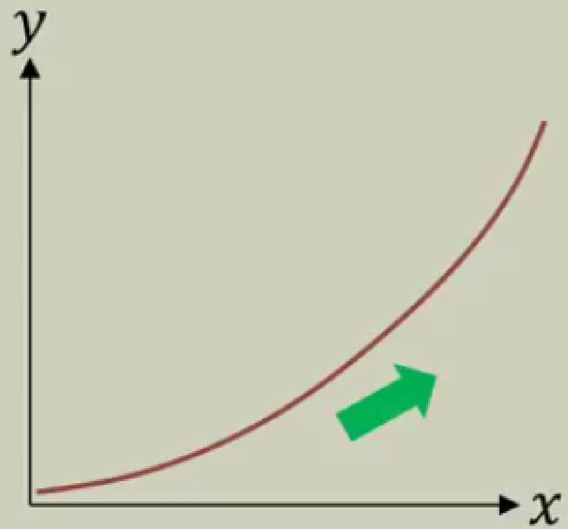
- **Higher R Square and Adjusted R Square which indicates more variance is explained in the quadratic model**
- **Lower standard error in the quadratic model indicates the observations fit tighter around the quadratic regression line**


Total Cars Sold vs Job Tenure
400
350
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
450
500
LINEAR MODEL
y = 0.5823x + 114.5
R² = 0.8019
QUADRATIC MODEL
y = -0.0019x² + 1.4095x + 63.851
R² = 0.9075

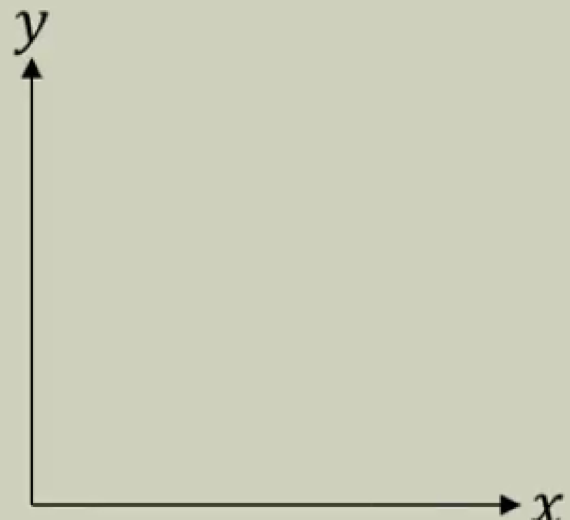

Quadratic Model Overall Residuals
60
40
20
0
-20
-40
-60
-80
0
50
100
150
200
250
300
350
Residuals
Predicted Car Sales
VERY NICE! NO PATTERN



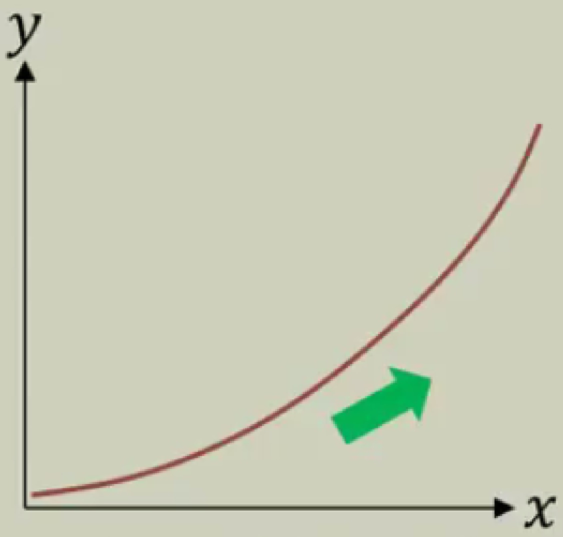
$b_1 > 0, b_2 > 0$



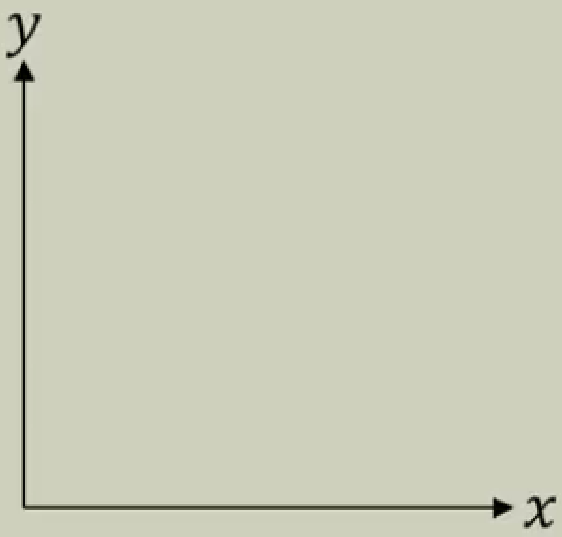
$b_1 > 0, b_2 > 0$



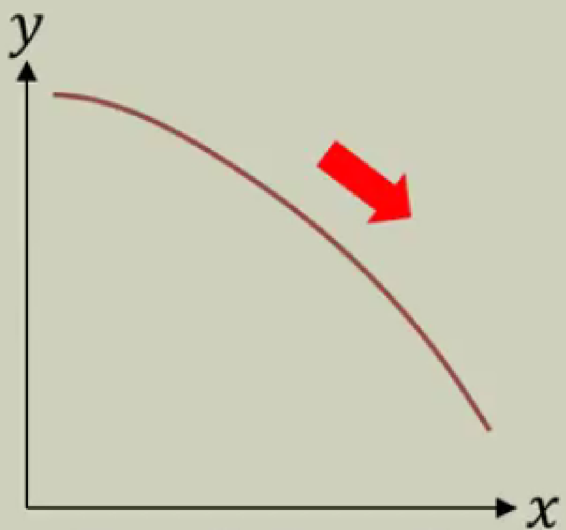
$b_1 < 0, b_2 < 0$



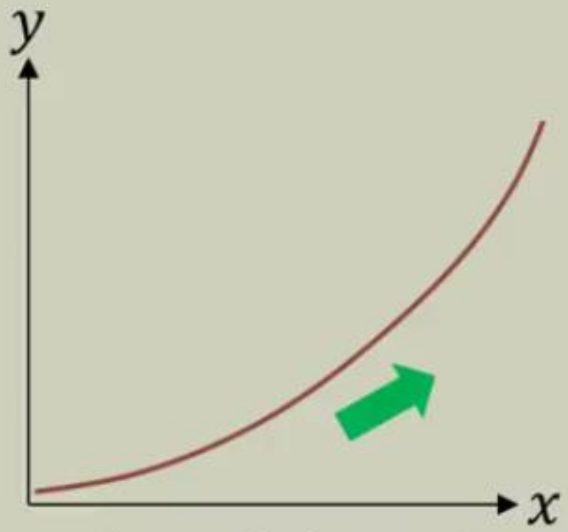
$b_1 > 0, b_2 > 0$



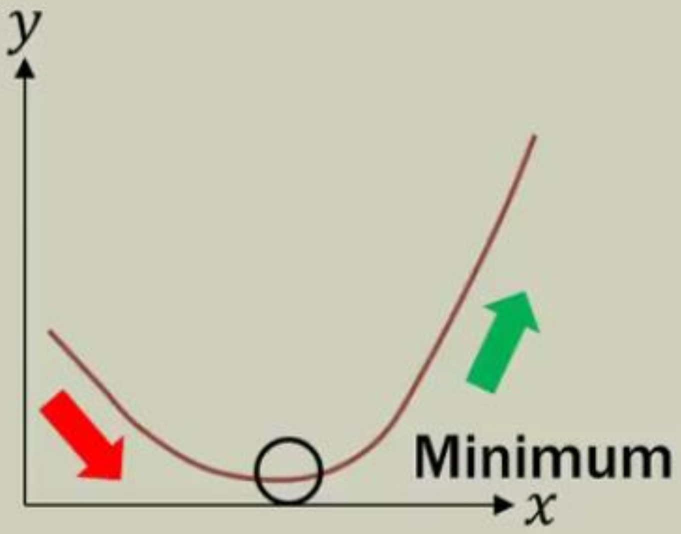
$b_1 < 0, b_2 > 0$



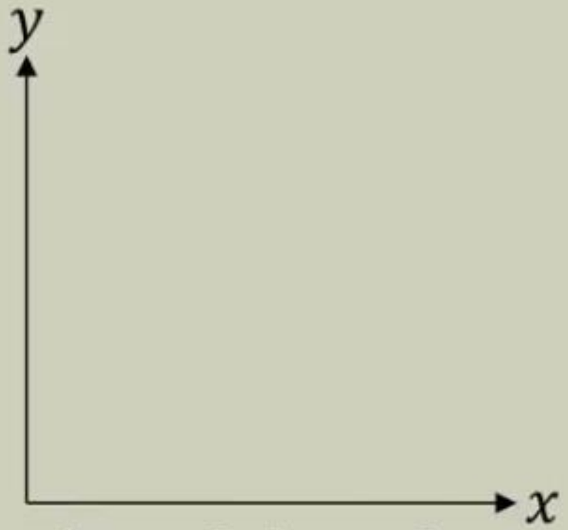
$b_1 < 0, b_2 < 0$



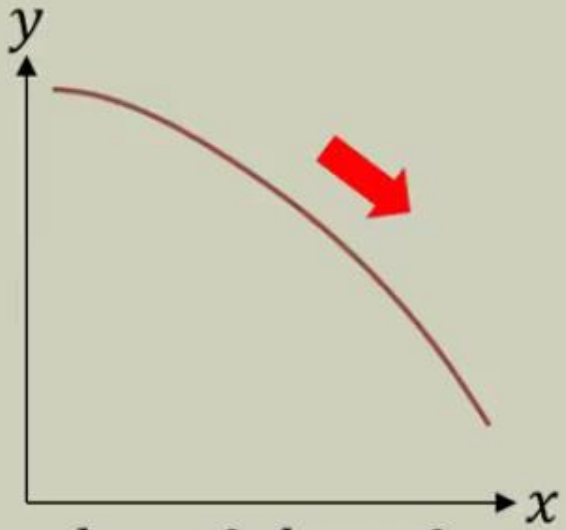
$b_1 > 0, b_2 > 0$



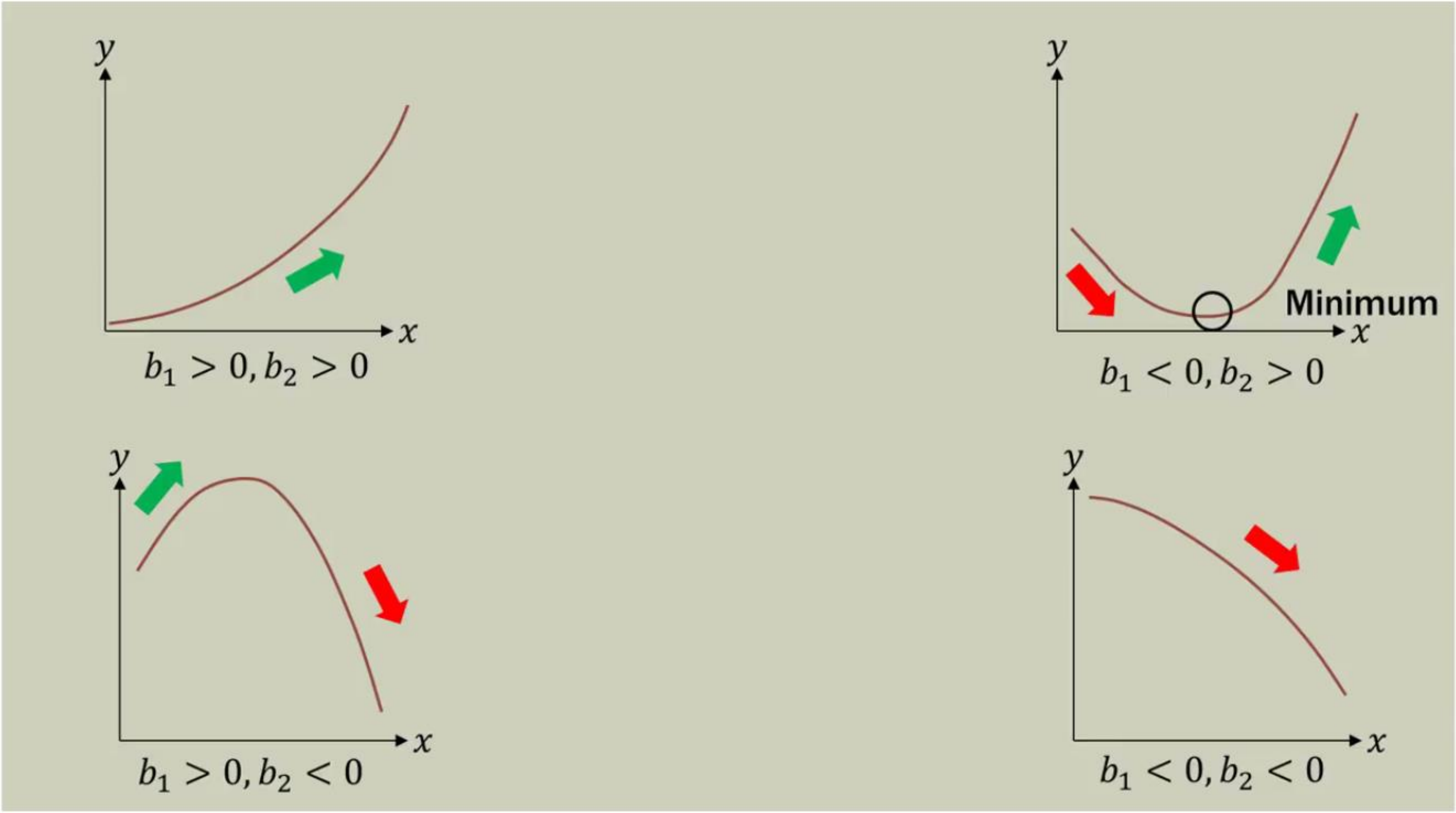
$b_1 < 0, b_2 > 0$



$b_1 > 0, b_2 < 0$



$b_1 < 0, b_2 < 0$


y
x
$b_1 > 0, b_2 > 0$
y
Minimum
x
$b_1 < 0, b_2 > 0$
y
x
$b_1 > 0, b_2 < 0$
y
x
$b_1 < 0, b_2 < 0$

FINAL NOTES

- In this example the non-linear quadratic model was a much better choice
- The residuals from the linear model indicated a curvilinear relationship between the original two variables
- Implementing a nonlinear quadratic model accomplished several things:
  - More explained variance
  - Tighter fit of the observations around the regression line
  - Reduced model error
  - A residual plot that no longer had a curvilinear shape
- This will naturally lead to better confidence and prediction intervals