USC
School of Business

Data Driven Decision Making: Simple Linear Regression Analysis

GSBA 545, Fall 2021

Professor Dawn Porter

Simple Linear Regression

- Simple Linear Regression (SLR) Model
- Regression Model Assumptions
  - Normality, Independence, Linearity, & Constant Variance
- Inference
  - F & T-testing
- Confidence & Prediction Intervals

USC School of Business

Simple Linear Regression Model

	Price	Square Feet
1	225,000	868
2	212,000	1021
3	210,000	1164
4	330,000	1598
5	165,000	888
6	300,000	1210
7	320,000	1295
8	210,000	1360
9	255,000	1440
10	229,000	1567
11	296,000	1767
12	450,000	1796
13	448,000	1940
14	285,000	1963
15	418,000	2022
16	319,000	2038
17	345,000	2690
18	272,000	2126
19	342,000	2163
20	455,000	2190
21	580,000	2320
22	496,000	2420
23	575,000	2452
24	625,000	2690
25	495,000	2930

65% of change in price is explained by square feet.

Summary of Fit

RSquare	0.6545
RSquare Adj	0.6507
Root Mean Square Error	91,047.3845
Mean of Response	390574.4681
Observations (or Sum Wgts)	94.0000

expected error

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1.4447e+12	1.445e+12	174.2785
Error	92	7.6265e+11	8.2896e+9	Prob > F
C. Total	93	2.2073e+12		<.0001*

true % explained is R-squared & not adj r-squared.

→ overall model.

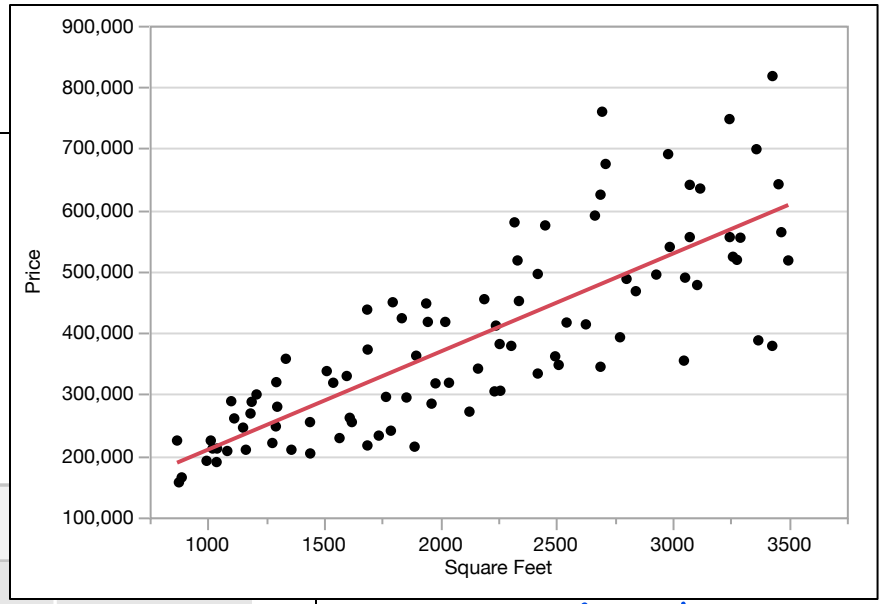
model quantitative VS quant data

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	50,598.6872	27,411.6632	1.85	0.0681
Square Feet	159.4259	12.0764	13.20	<.0001*

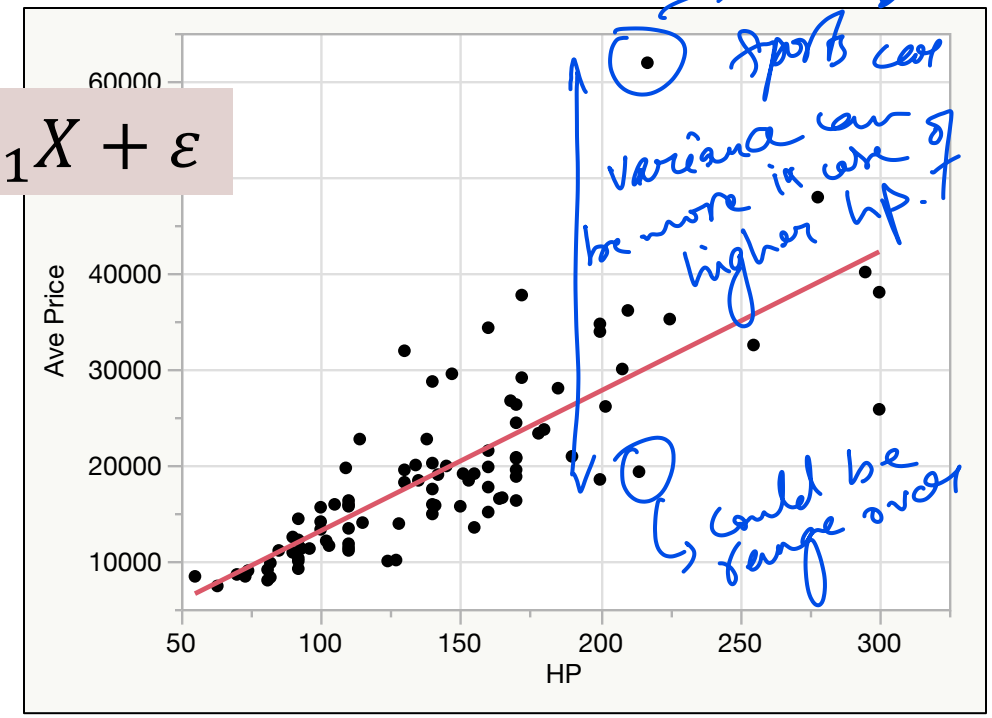
(159.43 ± 2σ) would be effect of 1 sq ft 98% of time

Sq. feet is a significant predictor of price



Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



Ave$	HP
15900	140
33900	200
29100	172
37700	172
30000	208
15700	110
20800	170
23700	180
26300	170
34700	200
40100	295
13400	110
11400	110
15100	160
15900	110
16300	170
16600	165

β_0 is the *Y*-intercept, or mean of *Y* when *X* is 0.*

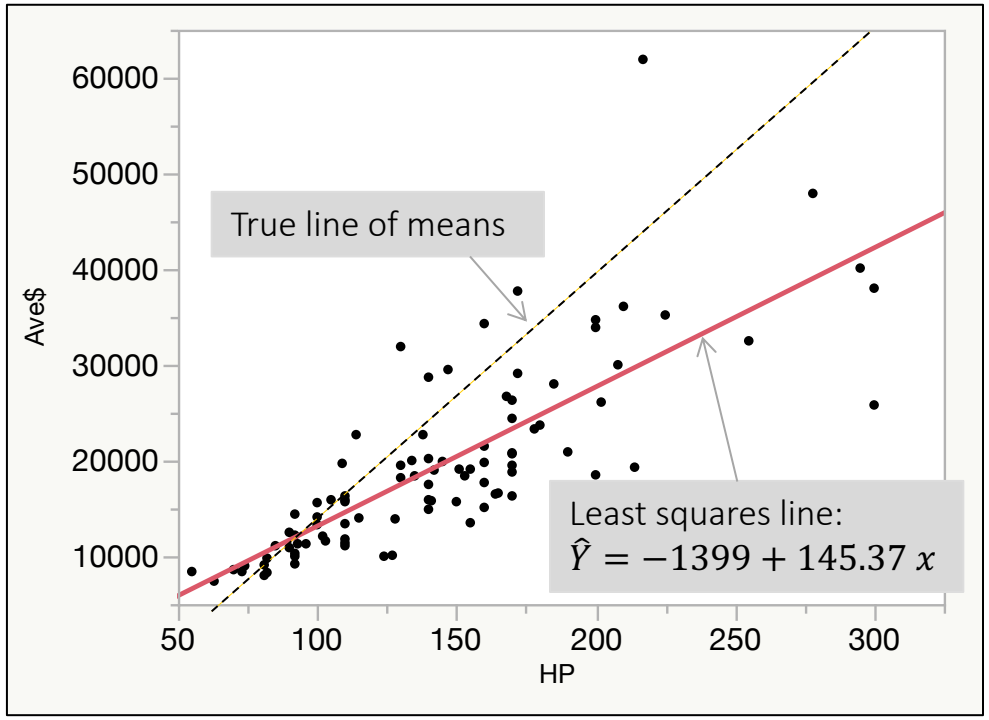
β_1 is the slope, change in the mean of *Y* per unit change in *X*.

ε is error term describing leftover effect on *Y*.

*Note: be careful: You need to have data where *X* is 0 for this to make sense.

Least Squares Estimation

Prediction (x = 210): $\hat{Y} = b_0 + b_1 x = -1399 + 145.37(210) = \$29{,}129$



Slope (b_1)

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$= r_{XY}\frac{s_Y}{s_X} = 145.37$$

Y-Intercept (b_0)

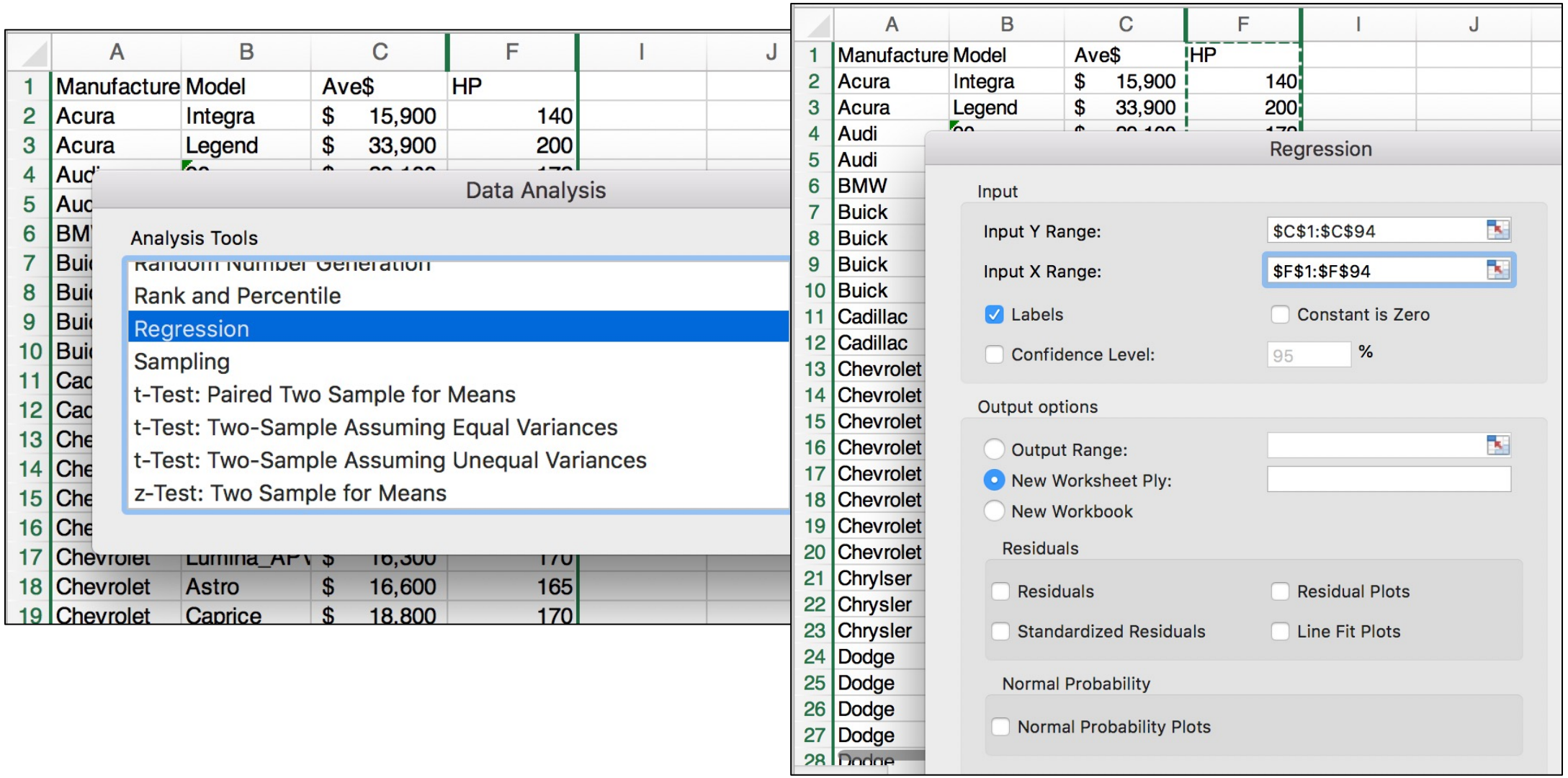
$$\bar{Y} = 19509.68, \bar{X} = 143.83$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$= 19509.68 - (145.37)(143.83)$$

$$= -1399$$

Ave$ and HP: Excel



	Manufacture	Model	Ave$	HP
2	Acura	Integra	$ 15,900	140
3	Acura	Legend	$ 33,900	200
18	Chevrolet	Astro	$ 16,600	165
19	Chevrolet	Caprice	$ 18,800	170

Least Squares Estimation: Excel

Regression Statistics	
Multiple R	0.788
R Square	0.621
Adj R Square	0.617
Std Error	5976.953
Observations	93

% of *Ave Price* explained using *HP*

How much we expect to be off, on average, when predicting *Ave Price*

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	5333140406	5333140406	149.287	6.837E-21
Residual	91	3250880884	35723966		
Total	92	8584021290			

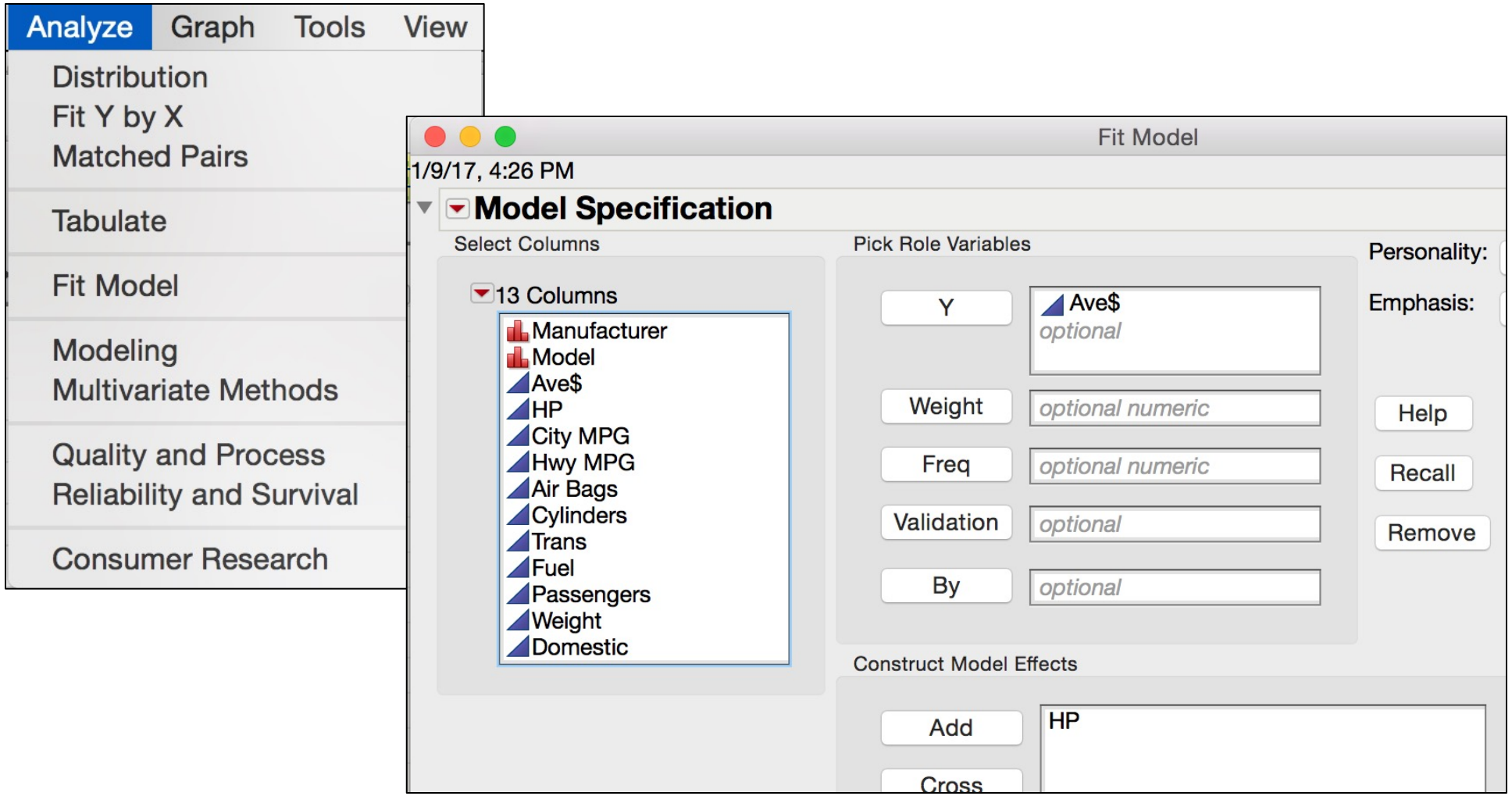
Coefficients for regression equation

Determines if entire model is significant or useful for predicting *Ave Price*

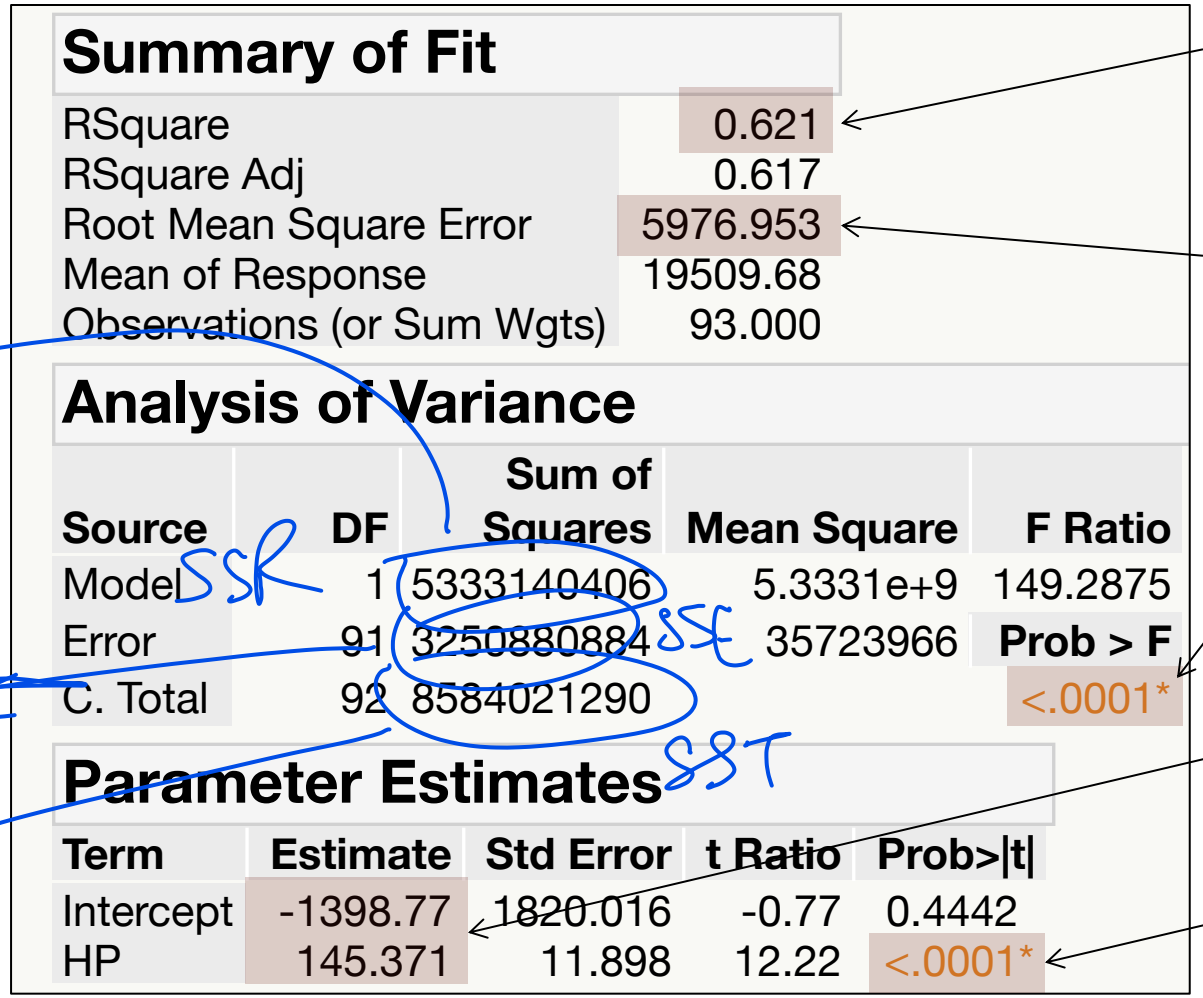
Determines if *HP* is significant or useful for predicting *Ave Price*

	Coefficients	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-1398.769	1820.016	-0.769	0.444	-5014.008	2216.470
HP	145.371	11.898	12.218	0.000	121.738	169.005

Ave$ and HP: JMP



Least Squares Estimation: JMP



Summary of Fit

RSquare	0.621
RSquare Adj	0.617
Root Mean Square Error	5976.953
Mean of Response	19509.68
Observations (or Sum Wgts)	93.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5333140406	5.3331e+9	149.2875
Error	91	3250880884	35723966	Prob > F
C. Total	92	8584021290		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1398.77	1820.016	-0.77	0.4442
HP	145.371	11.898	12.22	<.0001*

% of *Ave Price* explained using *HP*

How much we expect to be off, on average, when predicting *Ave Price*

Determines if entire model is significant or useful for predicting *Ave Price*

Coefficients for regression equation

Determines if *HP* is significant or useful for predicting *Ave Price*

Ave$ and HP: Python

```
# Import packages
import pandas as pd
import numpy as np
import statsmodels.formula.api as smf

# Import data
carPrice = pd.read_excel('CarPrice.xlsx', sheet_name='CarPrice')
print(f'Initial name of columns:\n {carPrice.columns}')

# Change column names
carPrice.columns = ['Manufacturer','Model','Price','City_MPG','Hwy_MPG','HP','Weight','Domestic']
print(f'\n\nDataset\n {carPrice.head()}')

result = smf.ols('Price ~ HP', data = carPrice).fit()
result.summary()
```

```
Initial name of columns:
 Index(['Manufacturer', 'Model', 'Ave$', 'City MPG', 'Hwy MPG', 'HP', 'Weight',
       'Domestic'],
      dtype='object')


Dataset
   Manufacturer    Model  Price  City_MPG  Hwy_MPG   HP  Weight  Domestic
0         Acura  Integra  15900        25       31  140    2705         0
1         Acura   Legend  33900        18       25  200    3560         0
2          Audi       90  29100        20       26  172    3375         0
3          Audi      100  37700        19       26  172    3405         0
4           BMW     535i  30000        22       30  208    3640         0
```

Ave$ and HP: (ANOVA) Python

Python doesn't automatically create a full ANOVA table or report the RMSE (Standard Error) value, so a little more code is necessary.

```
import statsmodels.api as sm
anova = sm.stats.anova_lm(result, typ=1)
print('Analysis of Variance')
anova.columns=['DF',' Sum_of_Squares','Mean_Square','F_Statistic','Prob>F']
anova
```

```
Analysis of Variance
```

	DF	Sum_of_Squares	Mean_Square	F_Statistic	Prob>F
HP	1.0	5.333140e+09	5.333140e+09	149.287468	6.837464e-21
Residual	91.0	3.250881e+09	3.572397e+07	NaN	NaN

```
# MSE
print(f'Mean Square Error: {result.mse_resid:.7}')

# RMSE
import math
print(f'Root Mean Square Error: {math.sqrt(result.mse_resid):.7}')
```

```
Mean Square Error: 3.572397e+07
Root Mean Square Error: 5976.953
```

Least Squares Estimation: Python

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.621
Model:	OLS	**Adj. R-squared:**	0.617
Method:	Least Squares	**F-statistic:**	149.3
Date:	Fri, 11 Sep 2020	**Prob (F-statistic):**	6.84e-21
Time:	17:27:45	**Log-Likelihood:**	-939.65
No. Observations:	93	**AIC:**	1883.
Df Residuals:	91	**BIC:**	1888.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1398.7691	1820.016	-0.769	0.444	-5014.008	2216.470
HP	145.3712	11.898	12.218	0.000	121.738	169.005

% of *Ave Price* explained using *HP*

Determines if entire model is significant or useful for predicting *Ave Price*

How much we expect to be off, on average, when predicting *Ave Price*

```
# MSE
print(f'Mean Square Error: {result.mse_resid:.7}')

# RMSE
import math
print(f'Root Mean Square Error: {math.sqrt(result.mse_resid):.7}')
```

```
Mean Square Error: 3.572397e+07
Root Mean Square Error: 5976.953
```

Coefficients for regression equation

Determines if *HP* is significant or useful for predicting *Ave Price*

Model Assumptions

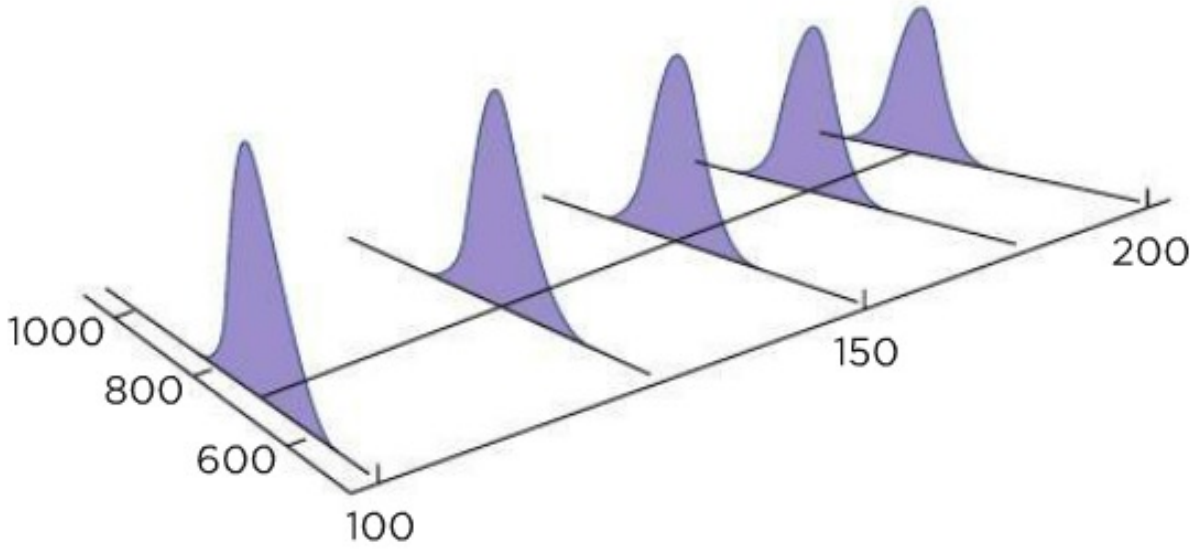
Assumptions about the model error terms

1. **Normality:** Error terms follow a normal distribution for all values of *x*.
2. **Independence:** Values of error terms are statistically independent of each other.
3. **Linearity:** Linear in parameters.
4. **Constant Variance:** (Homoscedasticity) Variance of error terms, σ^2, is the same for all values of *x*.

↳ like the case of cross-sectional data mentioned above, not best for time series data.

↳ Therefore, heteroskedasticity is a problem for regression model.

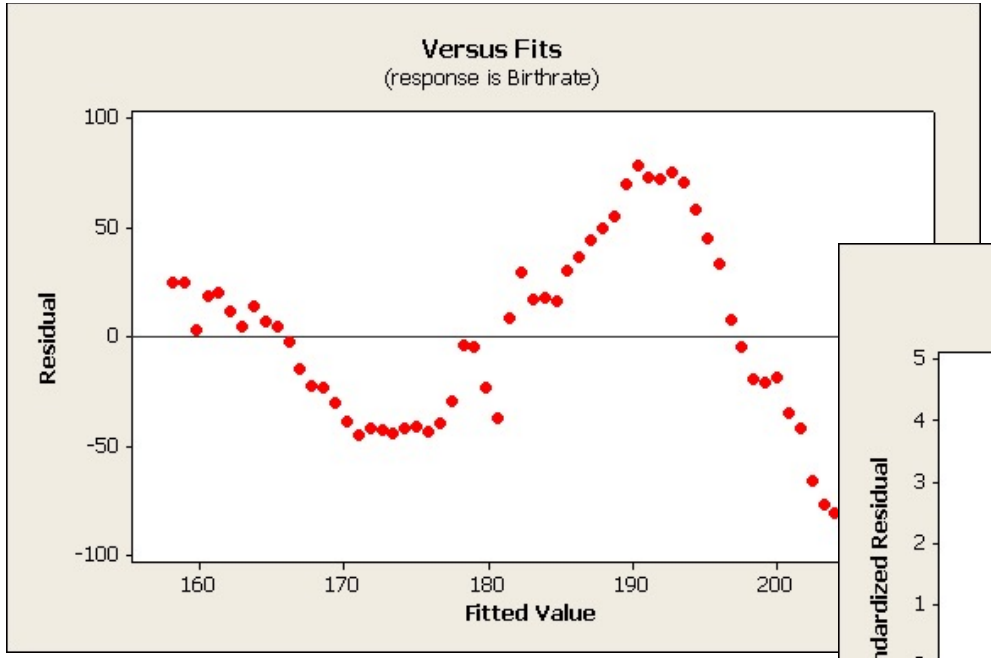
Normality

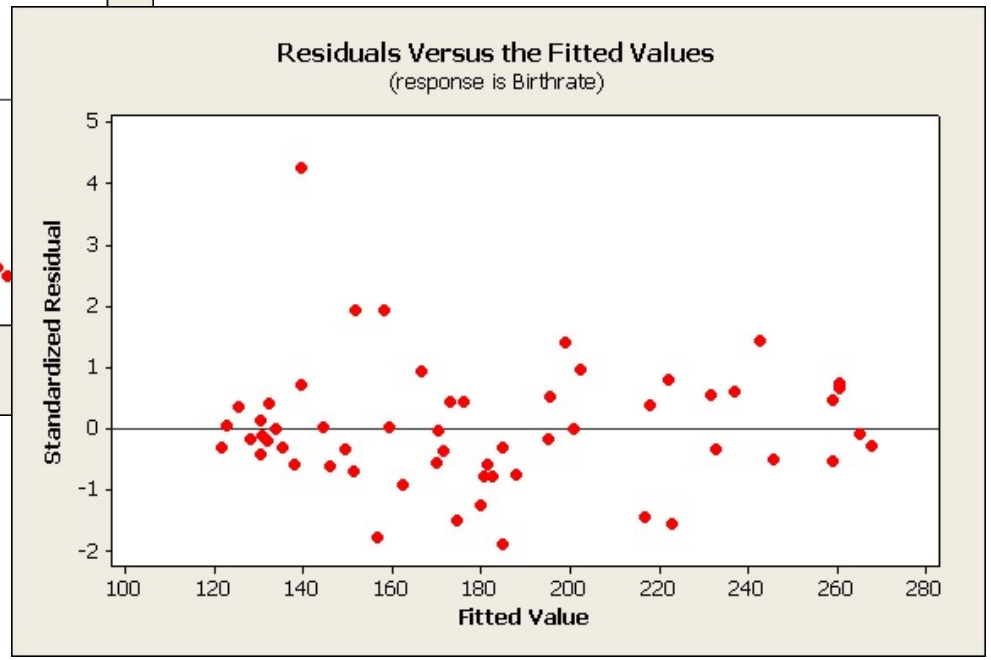


An approximate normal distribution of *Y* values is assumed at each level of *X*.
This allows us to create confidence and prediction intervals.

The errors should be closer to the line & ∴ follow normal dist^n.

Independence

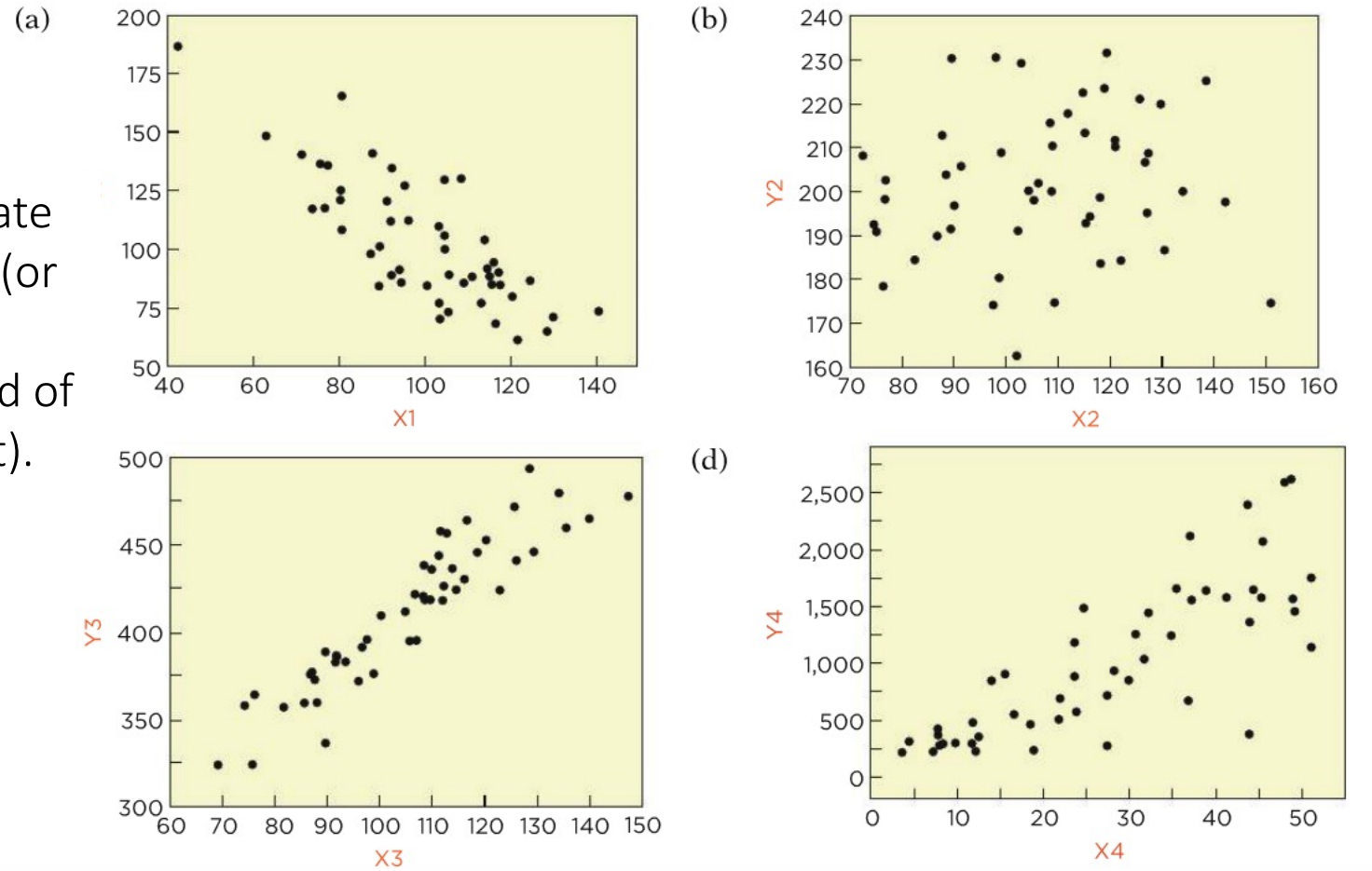




*This is mainly an issue with data organized as a *time series*.

Linearity

All of these plots indicate some level of linearity (or at least there doesn't appear to be some kind of curved pattern evident).



Random Error Variation

Variation of actual Y from predicted Y

- Measured by **standard error of estimate**
- Sample standard deviation of error terms
- Denoted by "*Standard Error*" in Excel and "Root Mean Square Error (**RMSE**)" in JMP and other programs.

Affects several factors

- Parameter significance
- Prediction accuracy

$$Std\ Error = RMSE = S_e = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-k-1}} = \sqrt{\frac{\sum e_i^2}{n-k-1}}$$

Measures of Variation

1. Total Sum of Squares (*SST*)

   Measures variation of observed Y_i around mean, $\bar{Y}$

2. Explained Variation (*SSR*): Sum of Squares explained by Model

   Variation due to relationship between *X* & *Y*

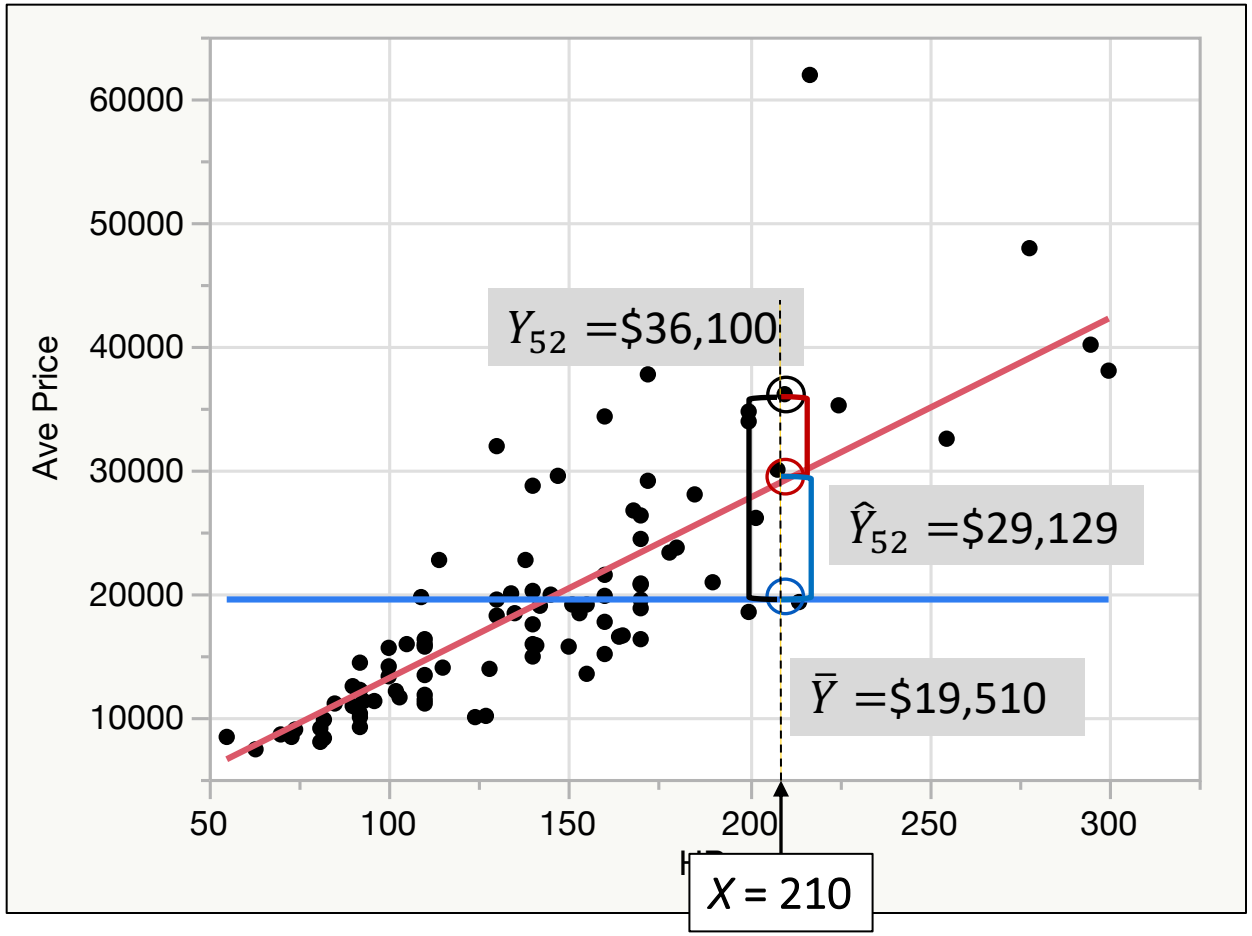
3. Unexplained Variation (*SSE*): Sum of Squares due to Error

   Variation due to other factors

$$SST = SSR + SSE$$

Measures of Variation

Prediction (x = 210): $\hat{Y} = b_0 + b_1 x = -1399 + 145.37(210) = \$29{,}129$



$Y_{52} - \hat{Y}_{52} = \$6971$

$e_{52} = \$6971$

$Y_{52} - \bar{Y} =$ $16,590

$\hat{Y}_{52} - \bar{Y} =$ $9619

So the model improved our prediction over using just the mean by $9619 for that car.

Correlation Coefficient

The **simple correlation coefficient** measures the strength of the linear relationship between *y* and *x* and is denoted by *r*.

$$r = \sqrt{R^2} = \sqrt{0.621} = 0.788$$

Summary of Fit

RSquare	0.621
RSquare Adj	0.617
Root Mean Square Error	5976.953
Mean of Response	19509.68
Observations (or Sum Wgts)	93.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5333140406	5.3331e+9	149.2875
Error	91	3250880884	35723966	Prob > F
C. Total	92	8584021290		<.0001

F-test for Overall Model

Testing H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$ at the α level of significance.

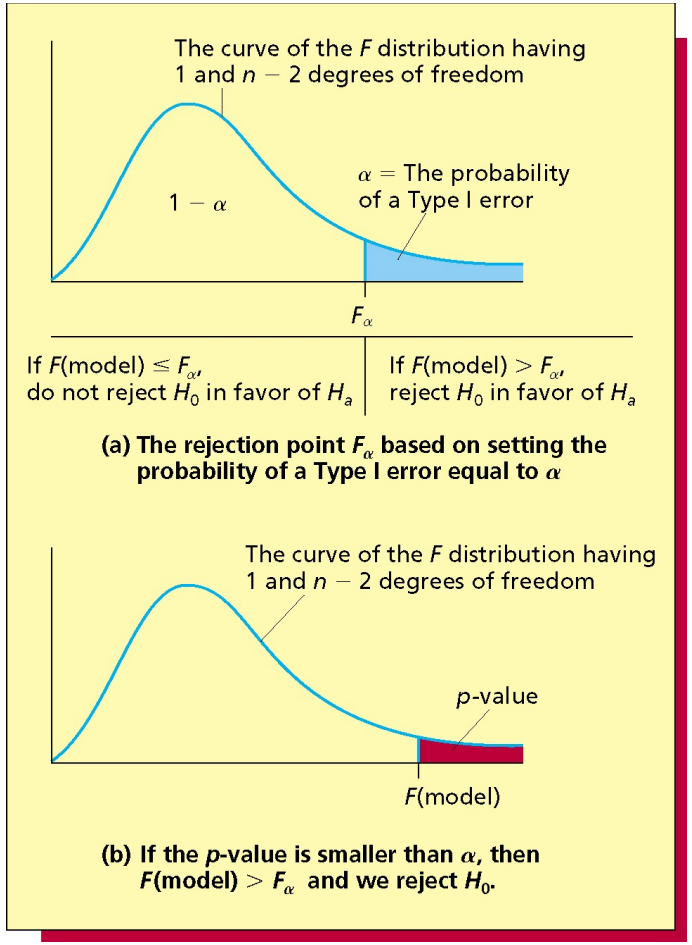
Test Statistic:

$$F = \frac{MSR}{MSE} = \frac{Explained\ variation/k}{Unexplained\ variation/(n-k-1)}$$

Reject H_0 if

$F(\text{model}) > F_\alpha$* or

p-value $< \alpha$



* F_α is based on df for both numerator and denominator.

F-test for SLR: Ave$ (Excel)

ANOVA					
	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	5333140406	5333140406	149.287	6.837E-21
Residual	91	3250880884	35723966		
Total	92	8584021290			

$$F = \frac{MSR}{MSE} = \frac{5{,}333{,}140{,}406}{35{,}723{,}966}$$

$$= 149.287 > 3.946 = F_{0.05,1,91}$$

p-value ≈ 0.000 < 0.05 = α

⇒ Reject, so *HP* is useful

F-test for SLR: Ave$ (JMP)

Summary of Fit

RSquare	0.621
RSquare Adj	0.617
Root Mean Square Error	5976.953
Mean of Response	19509.68
Observations (or Sum Wgts)	93.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5333140406	5.3331e+9	149.2875
Error	91	3250880884	35723966	Prob > F
C. Total	92	8584021290		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1398.77	1820.016	-0.77	0.4442
HP	145.371	11.898	12.22	<.0001

$$F = \frac{MSR}{MSE} = \frac{5.3331e+9}{35723966}$$

$$= 149.29 > 3.946 = F_{0.05,1,91}$$

p-value ≈ 0.000 < 0.05 = α

⇒ Reject, so *HP* is useful

F-test for SLR: Ave$ (Python)

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.621
Model:	OLS	**Adj. R-squared:**	0.617
Method:	Least Squares	**F-statistic:**	149.3
Date:	Fri, 11 Sep 2020	**Prob (F-statistic):**	6.84e-21

```
import statsmodels.api as sm
anova = sm.stats.anova_lm(result, typ=1)
print('Analysis of Variance')
anova.columns=['DF',' Sum_of_Squares','Mean_Square','F_Statistic','Prob>F']
anova
```

```
Analysis of Variance
```

	DF	Sum_of_Squares	Mean_Square	F_Statistic	Prob>F
HP	1.0	5.333140e+09	5.333140e+09	149.287468	6.837464e-21
Residual	91.0	3.250881e+09	3.572397e+07	NaN	NaN

$$F = \frac{MSR}{MSE} = \frac{5.3331e+09}{3.572397e+07}$$

$$= 149.29 > 3.946 = F_{0.05,1,91}$$

p-value ≈ 0.000 < 0.05 = α

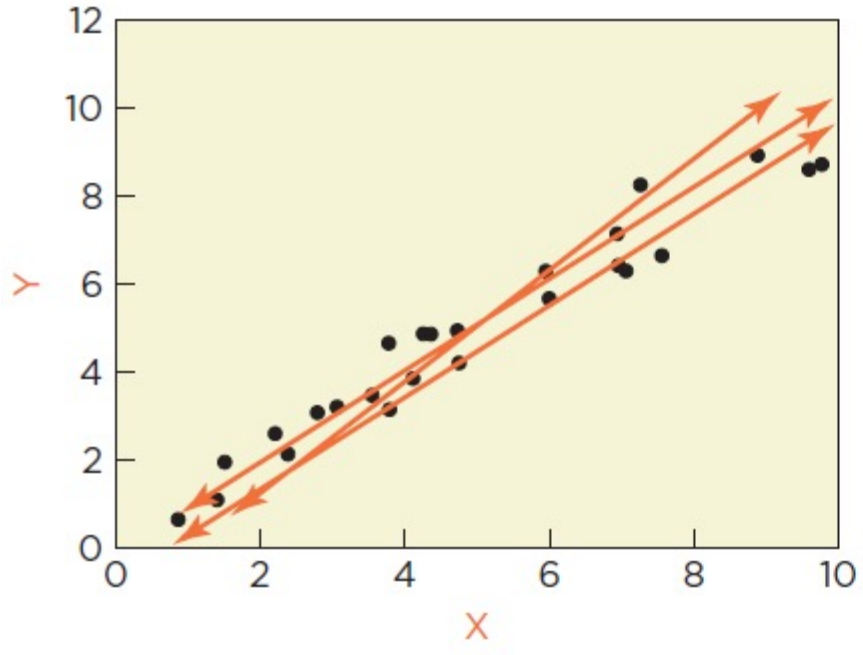
⇒ Reject, so *HP* is useful

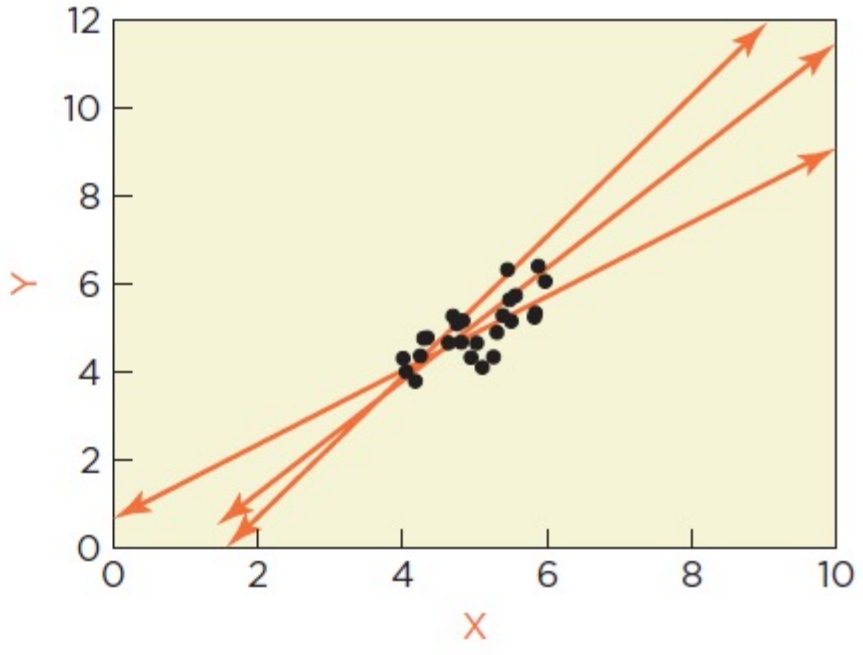
Standard Error for Slope: s_{b_1}

Describes the possible sample-to-sample variability of b_1.

- As *RMSE* increases, so does s_{b_1}
- As *n* increases, s_{b_1} decreases
- As s_x (std deviation of *X*) increases, s_{b_1} decreases

$$s_{b_1} = \frac{RMSE}{\sqrt{n-1}} \times \frac{1}{s_x}$$





Slope Significance Test

If regression assumptions hold, we can **reject** $H_0: \beta_1 = 0$ in favor of $H_a: \beta_1 \neq 0$ at the α level of significance if and only if the corresponding *p*-value is less than α (usually 0.05).

Test Statistic

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

95% Confidence Interval for β_1

$$\left[b_1 \pm t_{0.025,n-k-1} s_{b_1}\right]$$

* t_α, $t_{\alpha/2}$ and *p*-values based on $n - k - 1$ degrees of freedom, found as *DF Error* in JMP and *DF Residuals* in Python.

Slope Significance: Excel

Regression Statistics	
Multiple R	0.788
R Square	0.621
Adj R Square	0.617
Std Error	5976.953
Observations	93

$$t = \frac{b_1}{s_{b_1}} = \frac{145.371}{11.898} = 12.218 > 1.986 = t_{0.025,91}$$

p-value ≈ 0.000 < 0.05 = α

⇒ Reject, so *HP* is useful

ANOVA

	df	*SS*	*MS*	*F*	*Sig F*
Regression	1	5333140406	5333140406	149.287	6.837E-21
Residual	91	3250880884	35723966		
Total	92	8584021290			

	Coefficients	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-1398.769	1820.016	-0.769	0.444	-5014.008	2216.470
HP	145.371	11.898	12.218	0.000	121.738	169.005

Slope Significance: JMP

Summary of Fit

RSquare	0.621
RSquare Adj	0.617
Root Mean Square Error	5976.953
Mean of Response	19509.68
Observations (or Sum Wgts)	93.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5333140406	5.3331e+9	149.2875
Error	91	3250880884	35723966	Prob > F
C. Total	92	8584021290		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1398.77	1820.016	-0.77	0.4442
HP	145.371	11.898	12.22	<.0001

$$t = \frac{b_1}{s_{b_1}} = \frac{145.371}{11.898}$$

$$= 12.22 > 1.986 = t_{0.025,91}$$

p-value ≈ 0.000 < 0.05 = α

⇒ Reject, so *HP* is useful

Slope Significance: Python

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.621
Model:	OLS	**Adj. R-squared:**	0.617
Method:	Least Squares	**F-statistic:**	149.3
Date:	Fri, 11 Sep 2020	**Prob (F-statistic):**	6.84e-21
Time:	17:27:45	**Log-Likelihood:**	-939.65
No. Observations:	93	**AIC:**	1883.
Df Residuals:	91	**BIC:**	1888.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1398.7691	1820.016	-0.769	0.444	-5014.008	2216.470
HP	145.3712	11.898	12.218	0.000	121.738	169.005

$$t = \frac{b_1}{s_{b_1}} = \frac{145.371}{11.898}$$

$$= 12.22 > 1.986 = t_{0.025,91}$$

p-value ≈ 0.000 < 0.05 = α

⇒ Reject, so *HP* is useful

Prediction Intervals

Prediction ($X = x_0$)

$$\hat{Y} = b_0 + b_1 x_0$$

Ave$ (*HP* = 210): $\hat{Y} = b_0 + b_1 X = -1399 + 145.37(210) = \$29{,}129$

A **95% prediction interval** for an individual value of *Y* is

95% PI: $\left[\hat{Y} \pm t_{0.025, df\ Error}\, S_e^*\right]$

95% PI for **Ave$** when *HP* = 210:

$$[29129 \pm 1.986(5977)] = \boxed{[\$17{,}259, \$40{,}999]}$$

* In JMP, Python, and other programs, S_e is denoted *RMSE,* or *Root Mean Square Error.*

Confidence Intervals

Prediction ($X = x_0$)

$$\hat{Y} = b_0 + b_1 x_0$$

Ave$ (*HP* = 210): $\hat{Y} = b_0 + b_1 X = -1399 + 145.37(210) = \$29{,}129$

A **95% confidence interval** for the *mean* value of *Y* is

95% CI: $\left[\hat{Y} \pm t_{0.025, df\ Error} \frac{S_e^*}{\sqrt{n}}\right]$

95% CI for the ***mean*** **Ave$** when *HP* = 210:

$$\left[29129 \pm 1.986\left(\frac{5977}{\sqrt{93}}\right)\right] = \boxed{[\$27{,}898, \$30{,}360]}$$

* In JMP, Python, and other programs, S_e is denoted *RMSE,* or *Root Mean Square Error.*