Chapter 9:
Inferences for Regression

Halima Bensmail

CS502

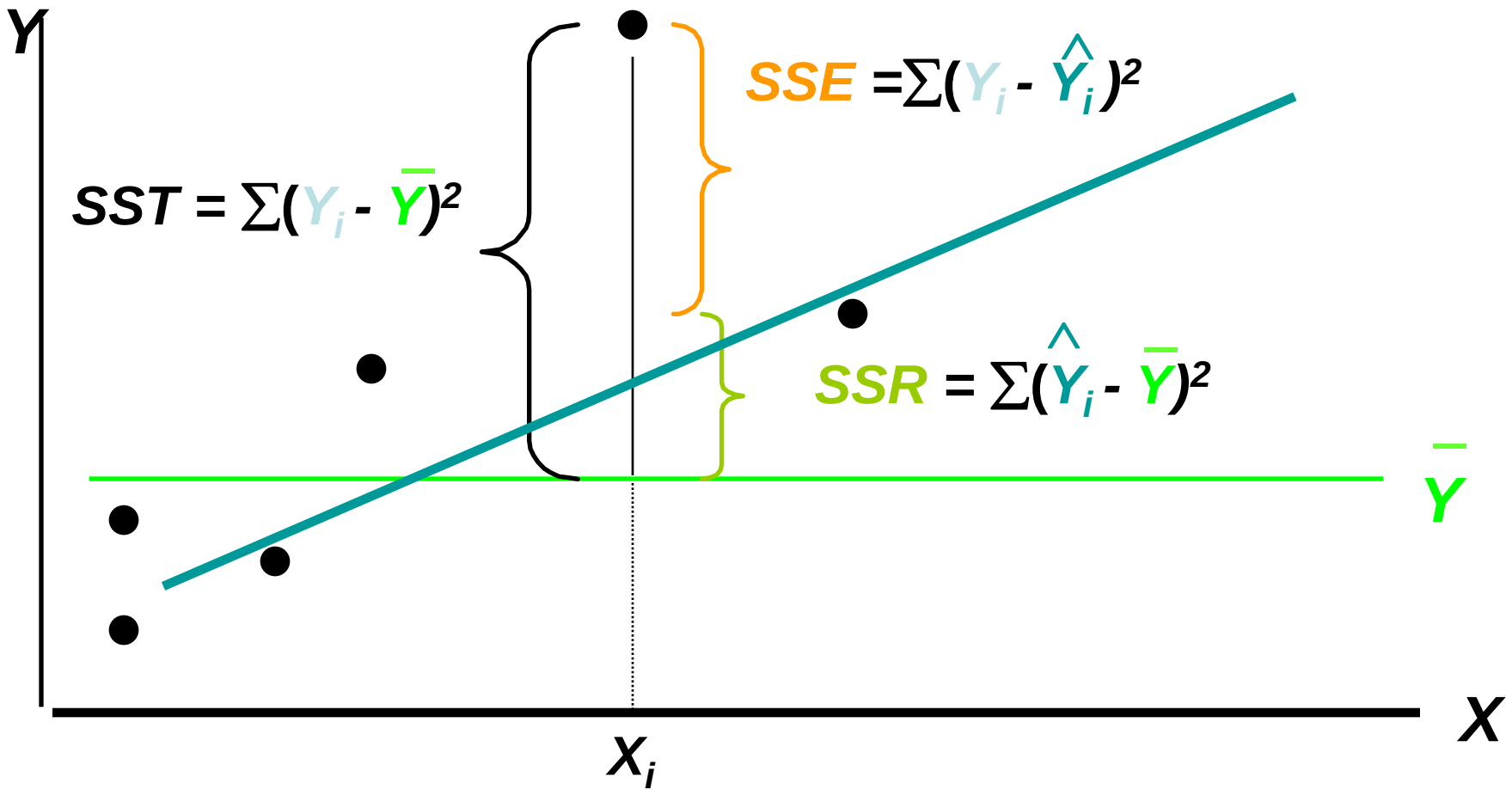
Monday 4-7pm

LAS Hall C

More on Linear regression

- *Variability*
- *Coefficient of determination*
- *Inferences on the slope*
- *Confidences interval*
- *Prediction*

Measures of Variation: The Sum of Squares



Measures of Variation: The Sum of Squares

- ***SST*** = ***SSR*** + ***SSE***

- **Total Sample Variability**
- = **Explained Variability** +
- **Unexplained Variability**

Measures of Variation: The Sum of Squares

- SST = Total Sum of Squares $\Sigma(Y_i - \bar{Y})^2$
  - Measures the variation of the Y_i values around their mean, $\bar{Y}$
- SSR = Regression Sum of Squares $\Sigma(\hat{Y}_i - \bar{Y})^2$
  - Explained variation attributable to the relationship between *X* and *Y*
- SSE = Error Sum of Squares $\Sigma(Y_i - \hat{Y}_i)^2$
  - Variation attributable to factors other than the relationship between *X* and *Y*

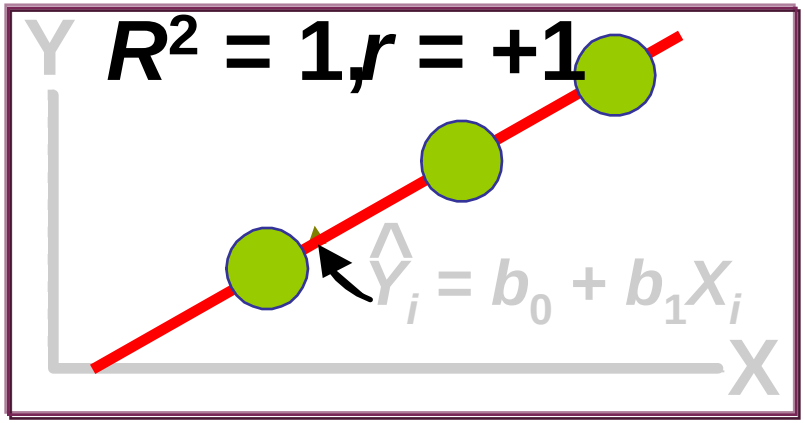
The Coefficient of Determination

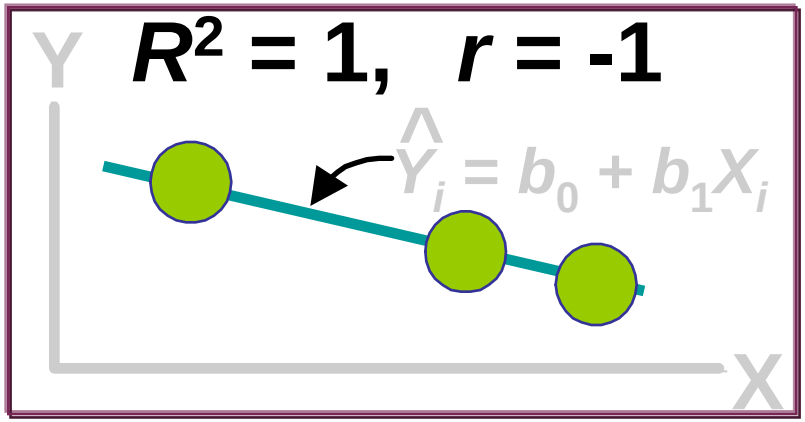
- $$R^2 = \frac{SSR}{SST} = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}}$$
- Measures the proportion of variation in Y that is explained by the independent variable X in the regression model

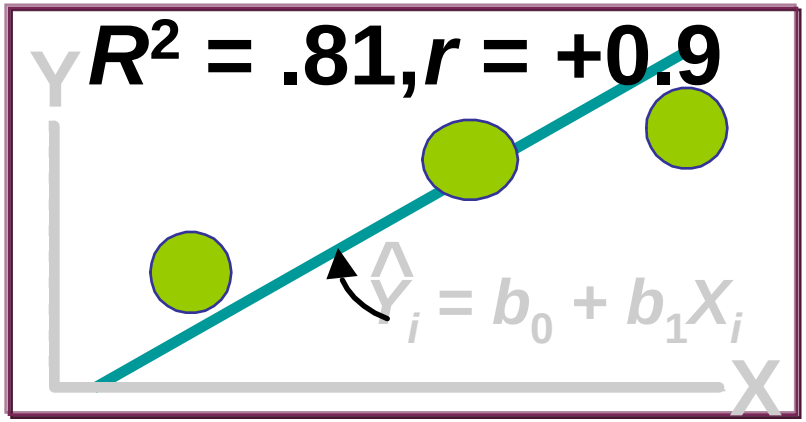
$0 < R^2 < 1$

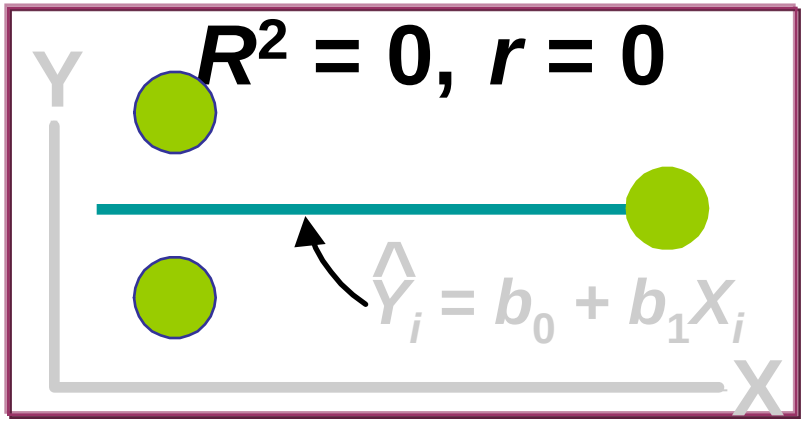
The closer to 1 the value of R^2 is, the better the regression line is

Coefficients of Determination (R^2) and Correlation (r)









Example: Produce Store

Store	Size(SF)	Annual Sales ($000)
1	1,726	3,681
2	1,542	3,395
3	2,816	6,653
4	5,555	9,543
5	1,292	3,318
6	2,208	5,563
7	1,313	3,760

Measures of Variation: Produce Store Example

Summary of Fit

RSquare	0.941981
RSquare Adj	0.930378
Root Mean Square Error	611.7515
Mean of Response	5130.429
Observations (or Sum Wgts)	7

R^2 = .94

94% of the variation in annual sales can be explained by the variability in the size of the store as measured by square footage

Root Mean Square Error of the Regression Estimate

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y - \hat{Y}_i\right)^2}{n-2}}$$

- ***MSE=SSE/(n-2)*** $= \Sigma(Y_i - \hat{Y}_i)^2 / (n-2)$
- The standard deviation of the variation of observations around the regression equation

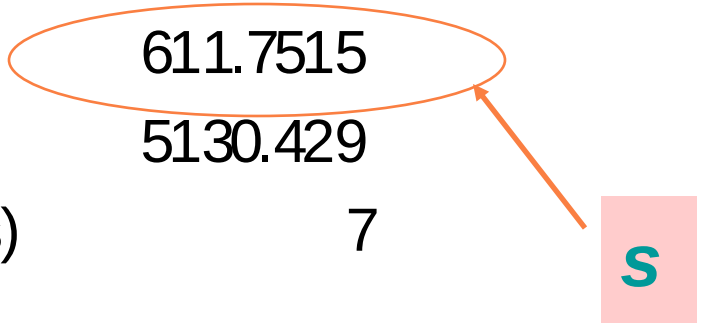
Store	Size(SF)	Annual Sales ($000)	Pred	(Y-Pred)	(Y-Pred)^2
1	1,726	3,681			
2	1,542	3,395			
3	2,816	6,653			
4	5,555	9,543			
5	1,292	3,318			
6	2,208	5,563			
7	1,313	3,760			

SSE?

Measures of Variation: Produce Store Example

Summary of Fit

RSquare	0.941981
RSquare Adj	0.930378
Root Mean Square Error	611.7515
Mean of Response	5130.429
Observations (or Sum Wgts)	7



Example: Produce Store

Store	Size	Annual Sales ($000)
1	1,726	3,681
2	1,542	3,395
3	2,816	6,653
4	5,555	9,543
5	1,292	3,318
6	2,208	5,563
7	1,313	3,760

Estimated Regression Equation:

$$\hat{Y} = 1636.415 + 1.487 X_i$$

The slope of this model is 1.487.

Does Square Footage Affect Annual Sales?

Which means:
Is Square Footage important to Predict Annual Sales?

Confidence Interval

- Range of values in which β_1 is thought to be
- Recall:the sampling distribution of $\bar{y}$ is $N(\mu, \frac{\sigma^2}{n})$

then the C.I. for $\mu = \bar{y} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

- C.I. for $\beta_1 = b_1 \pm t_{\frac{\alpha}{2},n-2} \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}} = b_1 \pm t_{\frac{\alpha}{2},n-2} \times SE_{b_1}$

STANDARD ERRORS FOR SLOPE AND INTERCEPT

The standard error of the slope b_1 of the least-squares regression line is

$$\mathrm{SE}_{b_1} = \frac{s}{\sqrt{\sum (x_i - \overline{x})^2}}$$

The standard error of the intercept b_0 is

$$\mathrm{SE}_{b_0} = s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}}$$

C.I. Example

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1636.4147	451.4953	3.62	0.0151
Square Feet	1.4866337	0.164999	9.01	0.0003

b_1 SE_{b1}

95% CI of β_1 is
$(1.48 - t_{(0.025,\ 5)} * 0.16\ ;\ 1.48 - t_{(0.025,\ 5)} * 0.16)$
$(1.48 - 2.571 * 0.16\ ;\ 1.48 - 2.571 * 0.16)$

Inferences about the Slope: Confidence Interval Example

$$b_1 \pm t_{\frac{\alpha}{2},n-2} SE_{b_1}$$

At 95% level of confidence the confidence interval for the slope is (1.062, 1.911). Does not include 0.

Conclusion: There is a significant linear dependency of annual sales on the size of the store.

Inference about the Slope: *t* Test

- *t* Test for a Population Slope
  - Is there a linear dependency of *Y* on *X* ?
- Null and Alternative Hypotheses
  - H_0: $\beta_1 = 0$ (No Linear Dependency)
  - H_1: $\beta_1 \neq 0$ (Linear Dependency)
- Test Statistics

$$t = \frac{b_1 - \beta_1}{SE_{b_1}} \text{ where } SE_{b_1} = \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}} \quad df = n - 2$$

t Test Example

Parameter Estimates

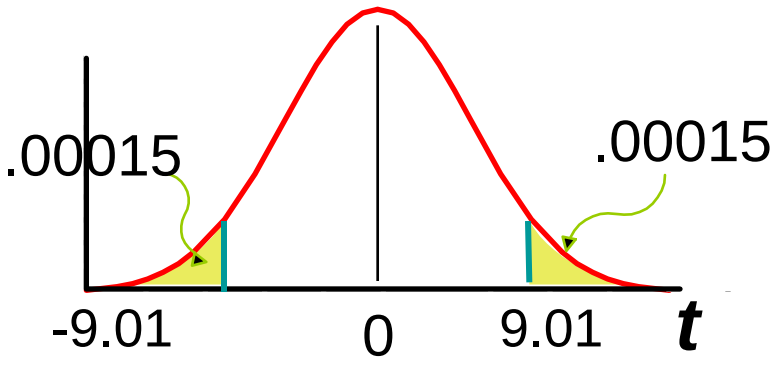
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1636.4147	451.4953	3.62	0.0151
Square Feet	1.4866337	0.164999	9.01	0.0003

b_1 SE_{b1} t P-value

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

$\alpha = .05$



t Test Example

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\alpha = .05$

- P-value=.0003
- Reject H_0
- There is evidence that square footage affects annual sales.

$H_0: \beta_1 \leq 0$

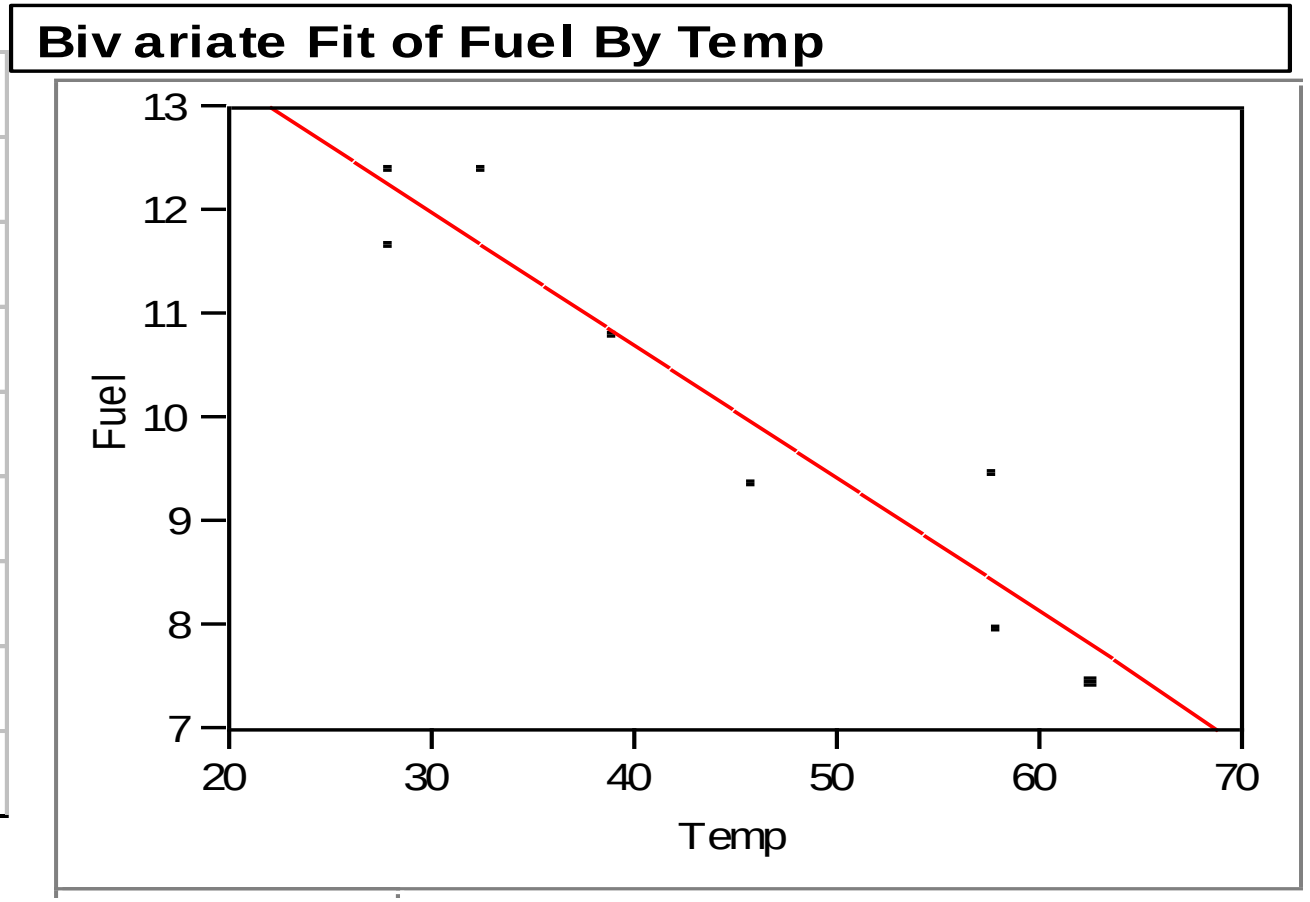
$H_1: \beta_1 > 0$

$\alpha = .05$

- P-value=.00015
- Reject H_0
- There is evidence that square footage affects annual sales.

Exercise: Fuel Consumption

y	x
12.4	28.0
11.7	28.0
12.4	32.5
10.8	39.0
9.4	45.9
9.5	57.8
8.0	58.1
7.5	62.5



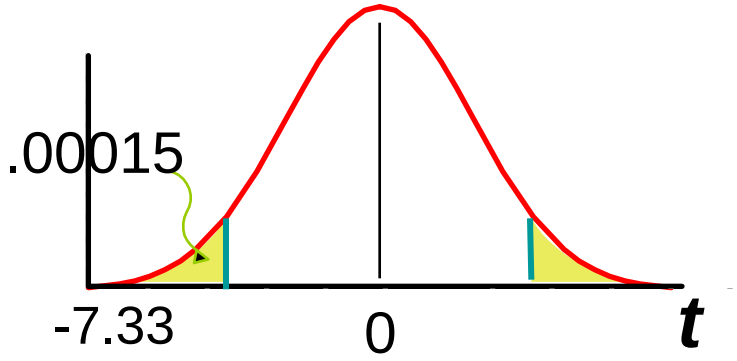
Exercise: Fuel Consumption

- Find 95% confidence interval for β_1
- Test the Hypothesis
  - H_0: $\beta_1=0$ vs. H_a: $\beta_1\neq0$
  - H_0: $\beta_1\geq0$ vs. H_a: $\beta_1<0$

Example: Fuel Consumption

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	15.837857	0.801773	19.75	<.0001	13.875989	17.799726
x	-0.127922	0.017457	-7.33	0.0003	-0.170638	-0.085205



Inferences about the Slope: Confidence Interval Example

$$b_1 \pm t_{\frac{\alpha}{2},n-2} SE_{b_1}$$

At 95% level of confidence the confidence interval for the slope is (-0.1706, -0.0852). Does not include 0.

Conclusion: There is a significant linear dependency of fuel consumption on the temperature.

Example: Fuel Consumption

- T-stat=(-0.12792) / 0.017457 = -7.327
- H_0: $\beta_1=0$ vs. H_a: $\beta_1\neq0$
  - P-value=0.0003

  Reject H_0:

  There is evidence that temperature affects fuel usage

- H_0: $\beta_1\geq0$ vs. H_a: $\beta_1<0$
  - P-value=0.00015
- Reject H_0
- There is evidence that temperature affects fuel usage

The ANOVA Table

	df	**SS**	**MS**	**F**	**Prob > F**
Regression	1	SSR	MSR =SSR/1	MSR/MSE	P-value of the F Test
Residuals	n-2	SSE	MSE =SSE/(n-2)		
Total	n-1	SST			

The ANOVA Table in JMP – Produce Store

Degrees of freedom

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	30380456	30380456	81.1791
Error	5	1871200	374239.92	Prob > F
C. Total	6	32251656		0.0003

Regression df

Error (residual) df

Total df

SSR

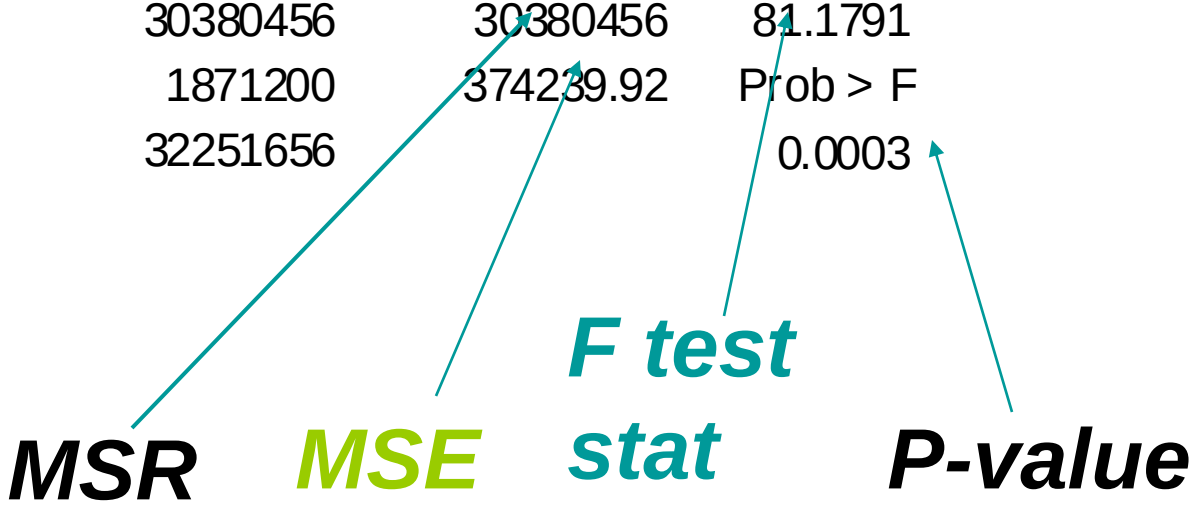
SSE

SST

The ANOVA Table in JMP – Produce Store

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	30380456	30380456	81.1791
Error	5	1871200	374239.92	Prob > F
C. Total	6	32251656		0.0003



Inferences about the Slope: *F* Test

- F Test for a Population Slope
  - Is there a linear dependency of *Y* on *X* ?
- Null and Alternative Hypotheses
  - H_0: $\beta_1 = 0$ (No Linear Dependency)
  - H_1: $\beta_1 \neq 0$ (Linear Dependency)
- Test Statistic

$$F = \frac{SSR/1}{SSE/(n-2)}$$

  - Numerator *d.f.=1*, denominator *d.f.=n-2*
  - F test needs (df1, df2, $\alpha = .05$)

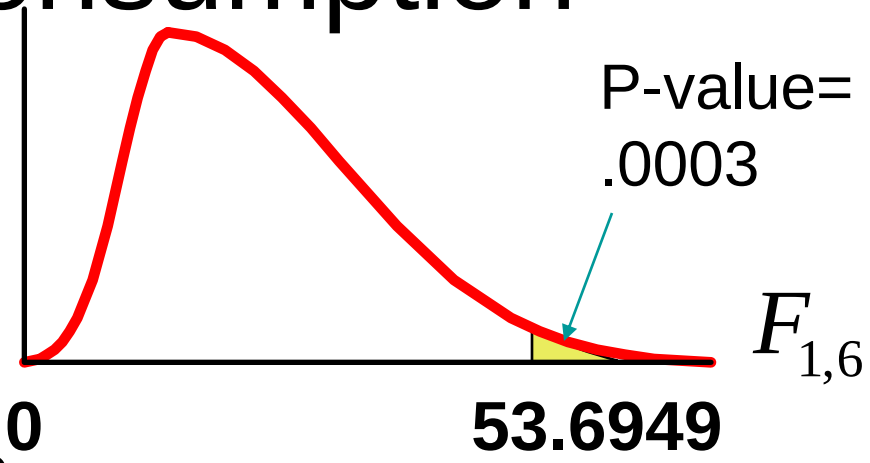
The ANOVA Table in JMP – Fuel Consumption

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	22.980816	22.9808	53.6949
Error	6	2.567934	0.4280	Prob > F
C. Total	7	25.548750		0.0003

The ANOVA Table in JMP – Fuel Consumption

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
$\alpha = .05$
numerator df = 1
Denominator df = 8-2=6



Decision: Reject H_0

Conclusion: There is evidence that temperature affects fuel consumption.

Relationship between a *t* Test and an *F* Test

- Null and Alternative Hypotheses
  - H_0: $\beta_1 = 0$ (No Linear Dependency)
  - H_1: $\beta_1 \neq 0$ (Linear Dependency)

$$\left(t_{n-2}\right)^2 = F_{1,n-2}$$

Estimation of Mean Values

- Using the Sales and Store size.
- If we want to consider all possible stores with 2000 square feet and we want to predict, on average, their annual sale.
- This is equivalent to calculate $\mu_{y|x=2000}$
- So if a relationship between X and Y exist, the best estimate of this point on the population regression line is given by
- $\hat{Y}_m = b_0 + b_1 X_m$

Prediction of Individual Values

- Now suppose that we are interested in a single store with 2000 square feet?

- So if a relationship between X and Y exist, the best prediction of the annual sale of this particular store is
- $Y_m=b_0+b_1 X_m$

CI and PI for Regression Response

CONFIDENCE AND PREDICTION INTERVALS FOR REGRESSION RESPONSE

A level C **confidence interval for the mean response** μ_y when x takes the value x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{\mu}}$$

Here $\mathrm{SE}_{\hat{\mu}}$ is the standard error for estimating a mean response.

A level C **prediction interval for a single observation** on y when x takes the value x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$$

The standard error $\mathrm{SE}_{\hat{y}}$ for estimating an individual response is larger than the standard error $\mathrm{SE}_{\hat{\mu}}$ for a mean response to the same x^*.

In both cases, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

STANDARD ERRORS FOR PREDICTION

The standard error for predicting the mean response when the explanatory variable x takes the value x^* is

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

The standard error for predicting an individual response when $x = x^*$ is

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

Estimation of Mean Values

Confidence Interval Estimate for $\mu_{Y|X=X_i}$:

The Mean of Y given a particular X_i

Size of interval vary according to distance away from mean, $\bar{X}$

Standard error of the estimate

$$\hat{Y}_i \pm t_{n-2} s \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

t value from table with df=n-2

Prediction of Individual Values

Prediction Interval for Individual Response Y_i at a Particular X_i

Addition of 1 increases width of interval from that for the mean of Y

$$\hat{Y}_i \pm t_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

Example: Produce Stores

Data for 7 Stores:

Store	Square Feet	Annual Sales ($000)
1	1,726	3,681
2	1,542	3,395
3	2,816	6,653
4	5,555	9,543
5	1,292	3,318
6	2,208	5,563
7	1,313	3,760

Consider a store with 2000 square feet.

Regression Equation Obtained:

$$\hat{Y} = 1636.415 + 1.487 X_i$$

Estimation of Mean Values: Example

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Find the 95% confidence interval for the average annual sales for stores of 2,000 square feet

Predicted Sales $\hat{Y} = 1636.415 + 1.487 X_i = 4610.45 (\$000)$

$$\hat{Y}_i \pm t_{n-2} s \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} = 4610.45 \pm 612.66$$

Prediction Interval for *Y* : Example

Prediction Interval for Individual $Y_{X=X_i}$

Find the 95% prediction interval for annual sales of one particular store of 2,000 square feet

Predicted Sales $\hat{Y} = 1636.415 + 1.487 X_i = 4610.45 (\$000)$

$$\hat{Y}_i \pm t_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} = 4610.45 \pm 1687.68$$

Interval Estimates for Different Values of *X*