Data Analysis

Multiple regression

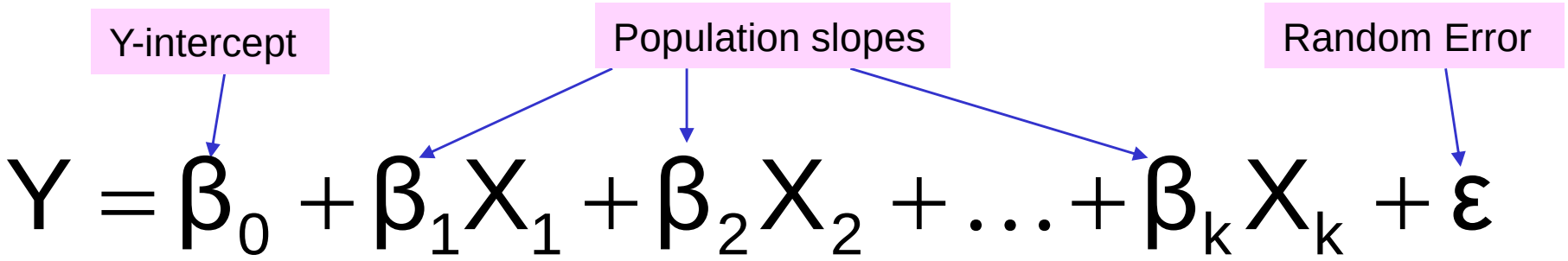
National Research University Higher School of Economics
Master's Program "Big Data Systems"

Fall 2018

Multiple linear regression

Idea: Examine the linear relationship between
1 dependent (Y) & 2 or more independent variables (X_i)

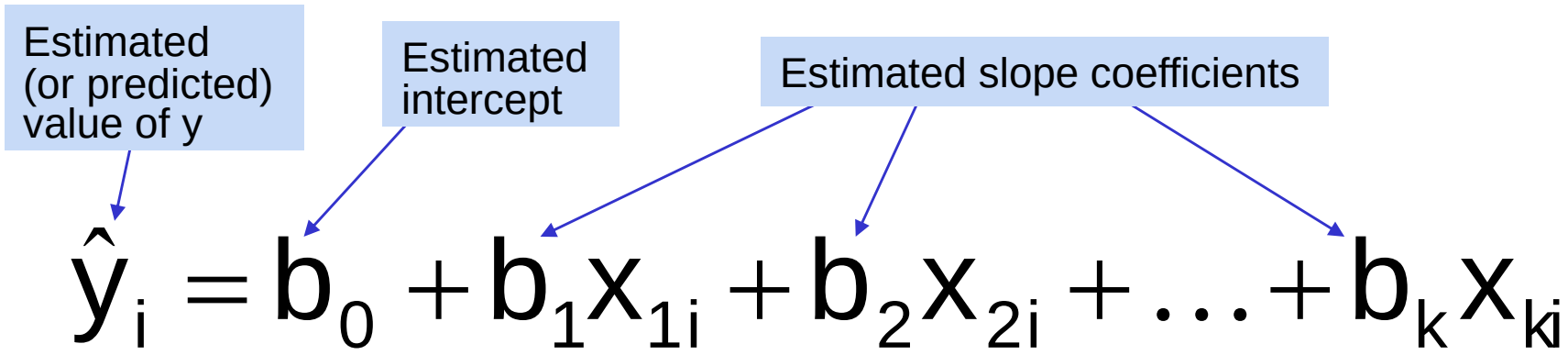
Multiple Regression Model with k Independent Variables:



Fitted model

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:



In this chapter we will always use a computer to obtain the regression slope coefficients and other regression summary measures.

Example

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1 (\text{Price}) + b_2 (\text{Advertising})$$

Multiple regression: things to remember

- Get the summary statistics of the multiple regression model
- Interpret coefficient of determination R^2 and the adjusted R^2
- Regression slopes, β_i:
  - Interpret the t-test for β_i. Confirm that analysis by examining confidence interval for β_i
  - Interpret overall F-test.
- Try to improve your model by removing the insignificant variables (in terms of t-test). To select what to remove, use:
  - Partial F-test
  - AIC criterion
- Play with the non-linear models using the above criteria.

The foregoing ideas are briefly illustrated in the remainder of the presentation.

Summary statistics

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Coefficient of determination

- Reports the proportion of total variation in y explained by all x variables taken together

$$R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

- This is the ratio of the explained variability to total sample variability

Coefficient of determination (2)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$R^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Variance of the model

- Consider the population regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

- The unbiased estimate of the variance of the errors is

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-K-1} = \frac{SSE}{n-K-1}$$

where $e_i = y_i - \hat{y}_i$

- The square root of the variance, s_e , is called the standard error of the estimate

Variance of the model (2)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$s_e = 47.463$

The magnitude of this value can be compared to the average y value

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Adjusted R^2

- R^2 never decreases when a new X variable is added to the model, even if the new variable is not an important predictor variable
  - This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
  - We lose a degree of freedom when a new X variable is added
  - Did the new X variable add enough explanatory power to offset the loss of one degree of freedom?

Adjusted R^2 (2)

- Used to correct for the fact that adding non-relevant independent variables will still reduce the error sum of squares

$$\overline{R}^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)}$$

(where n = sample size, K = number of independent variables)

  - Adjusted R^2 provides a better comparison between multiple regression models with different numbers of independent variables
  - Penalize excessive use of unimportant independent variables
  - Smaller than R^2

Adjusted R^2 (3)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$\overline{R}^2 = .44172$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Statistical inference for slopes

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

t-value for Price is t = -2.306, with p-value .0398

t-value for Advertising is t = 2.855, with p-value .0145

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Statistical inference for slopes (2)

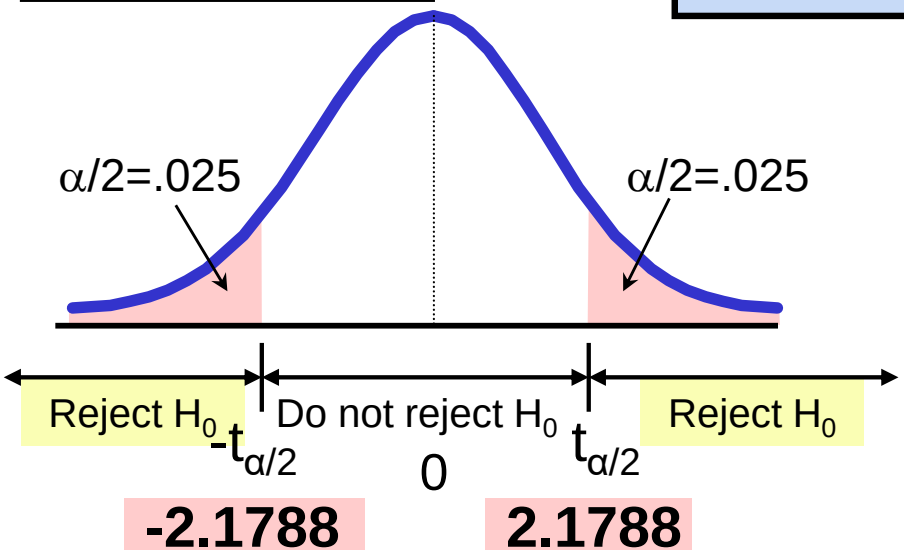
$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

d.f. = 15-2-1 = 12

α = .05

$t_{12,\ .025}$ = 2.1788



From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Price	**-24.97509**	**10.83213**	**-2.30565**	**0.03979**
Advertising	**74.13096**	**25.96732**	**2.85478**	**0.01449**

The test statistic for each variable falls in the rejection region (p-values < .05)

Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect pie sales at α = .05

Confidence intervals for slopes

- Confidence interval limits for the population slope β_j

$$b_j \pm t_{n-K-1,\alpha/2} S_{b_j}$$

where t has $(n - K - 1)$ d.f.

	Coefficients	*Standard Error*
Intercept	**306.52619**	**114.25389**
Price	**-24.97509**	**10.83213**
Advertising	**74.13096**	**25.96732**

Here, t has $(15 - 2 - 1) = 12$ d.f.

Example: Form a 95% confidence interval for the effect of changes in price (x_1) on pie sales:

$$-24.975 \pm (2.1788)(10.832)$$

So the interval is $-48.576 < \beta_1 < -1.374$

Confidence intervals for slopes (2)

Confidence interval for the population slope β_i

	Coefficients	*Standard Error*	…	*Lower 95%*	*Upper 95%*
Intercept	**306.52619**	**114.25389**	…	**57.58835**	**555.46404**
Price	**-24.97509**	**10.83213**	…	**-48.57626**	**-1.37392**
Advertising	**74.13096**	**25.96732**	…	**17.55303**	**130.70888**

Example: Excel output also reports these interval endpoints:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price

Total F test

- F-Test for Overall Significance of the Model
- Shows if there is a linear relationship between all of the X variables considered together and Y
- Use F test statistic
- Hypotheses:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)

H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

Total F test (2)

- Test statistic:

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

where F has k (numerator) and

$(n-K-1)$ (denominator)

degrees of freedom

- The decision rule is

$$\text{Reject} H_0 \text{ if } F > F_{k,n-K-1,\alpha}$$

Total F test (3)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$F = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F-Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

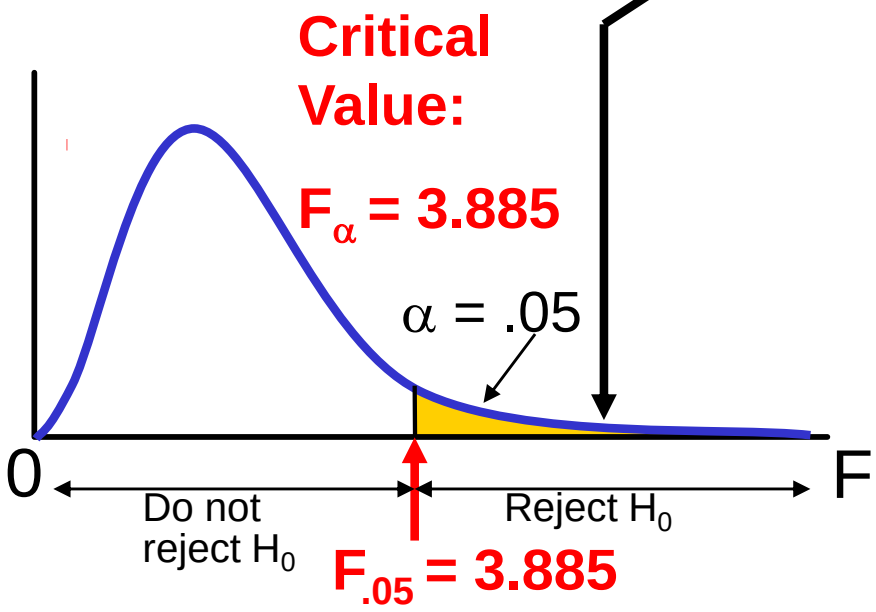
Total F test (4)

H_0: $\beta_1 = \beta_2 = 0$

H_1: β_1 and β_2 not both zero

$\alpha = .05$

$df_1 = 2$ $df_2 = 12$



Test Statistic:

$$F = \frac{MSR}{MSE} = 6.5386$$

Decision:

Since F test statistic is in the rejection region (p-value < .05), reject H_0

Conclusion:

There is evidence that at least one independent variable affects Y

Partial F test

- Consider a multiple regression model involving variables x_j and z_j, and the null hypothesis that the z variable coefficients are all zero:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \alpha_1 z_{1i} + \cdots \alpha_r z_{ri} + \varepsilon_i$$

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

$$H_1 : \text{ at least one of } \alpha_j \neq 0 \quad (j = 1, \ldots, r)$$

Partial F test (2)

- Goal: compare the error sum of squares for the complete model with the error sum of squares for the restricted model
  - First run a regression for the complete model and obtain SSE
  - Next run a restricted regression that excludes the z variables (the number of variables excluded is r) and obtain the restricted error sum of squares SSE(r)
  - Compute the F statistic and apply the decision rule for a significance level α

$$\text{Reject } H_0 \text{ if } \quad F = \frac{(SSE(r) - SSE)/r}{s_e^2} > F_{r,n-K-r-1,\alpha}$$

Akaike Information Criterion

- Based on maximal likelihood concept
- Formula

$$\text{AIC} = -\frac{2}{N}\ln(\text{likelihood}) + \frac{2}{N} \times \text{number of parameters}$$

- Decision rule: choose the model with minimal AIC.

Forecasting

- Given a population regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

- then given a new observation of a data point

$$(x_{1,n+1}, x_{2,n+1}, \ldots, x_{K,n+1})$$

the best linear unbiased forecast of $\hat{y}_{n+1}$ is

$$\hat{y}_{n+1} = b_0 + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_K x_{K,n+1}$$

- It is risky to forecast for new X values outside the range of the data used to estimate the model coefficients, because we do not have data to support that the linear model extends beyond the observed range.

Example

Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

$$= 306.526 - 24.975\,(5.50) + 74.131\,(3.5)$$

$$= 428.62$$

Predicted sales is 428.62 pies

Note that Advertising is in \$100's, so \$350 means that $X_2 = 3.5$

Multiple linear regression in R

- Regression statistics:

```
> da1=lm(Price ~ Mileage+Weight+HP,data=cars2)
> summary(da1)
```

- Plot of residuals:

```
> plot(fitted(da1), resid(da1))
```

- Partial F-test:

```
> da2=lm(Price~Weight+HP,data=cars2)
> anova(da2,da1)
```

- Selection by the AIC criterion:

```
> library(MASS)
> stepAIC(da1)
```

Non-linear regression

- The relationship between the dependent variable and an independent variable may not be linear
- Can review the scatter diagram to check for non-linear relationships
- Example: Quadratic model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

  - The second independent variable is the square of the first variable

Quadratic regression

- Testing the Quadratic Effect
  - Compare the linear regression estimate

$$\hat{y} = b_0 + b_1 x_1$$

  - with quadratic regression estimate

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

  - Hypotheses
    - H_0: $\beta_2 = 0$ (The quadratic term does not improve the model)
    - H_1: $\beta_2 \neq 0$ (The quadratic term improves the model)

Quadratic regression (2)

- Testing the Quadratic Effect

  Hypotheses

  - H_0: $\beta_2 = 0$ (The quadratic term does not improve the model)
  - H_1: $\beta_2 \neq 0$ (The quadratic term improves the model)

- The test statistic is

$$t = \frac{b_2 - \beta_2}{s_{b_2}}$$

$$d.f. = n - 3$$

where:

b_2 = squared term slope coefficient

β_2 = hypothesized slope (zero)

S_{b_2} = standard error of the slope

Testing for quadratic effect

- Testing the Quadratic Effect

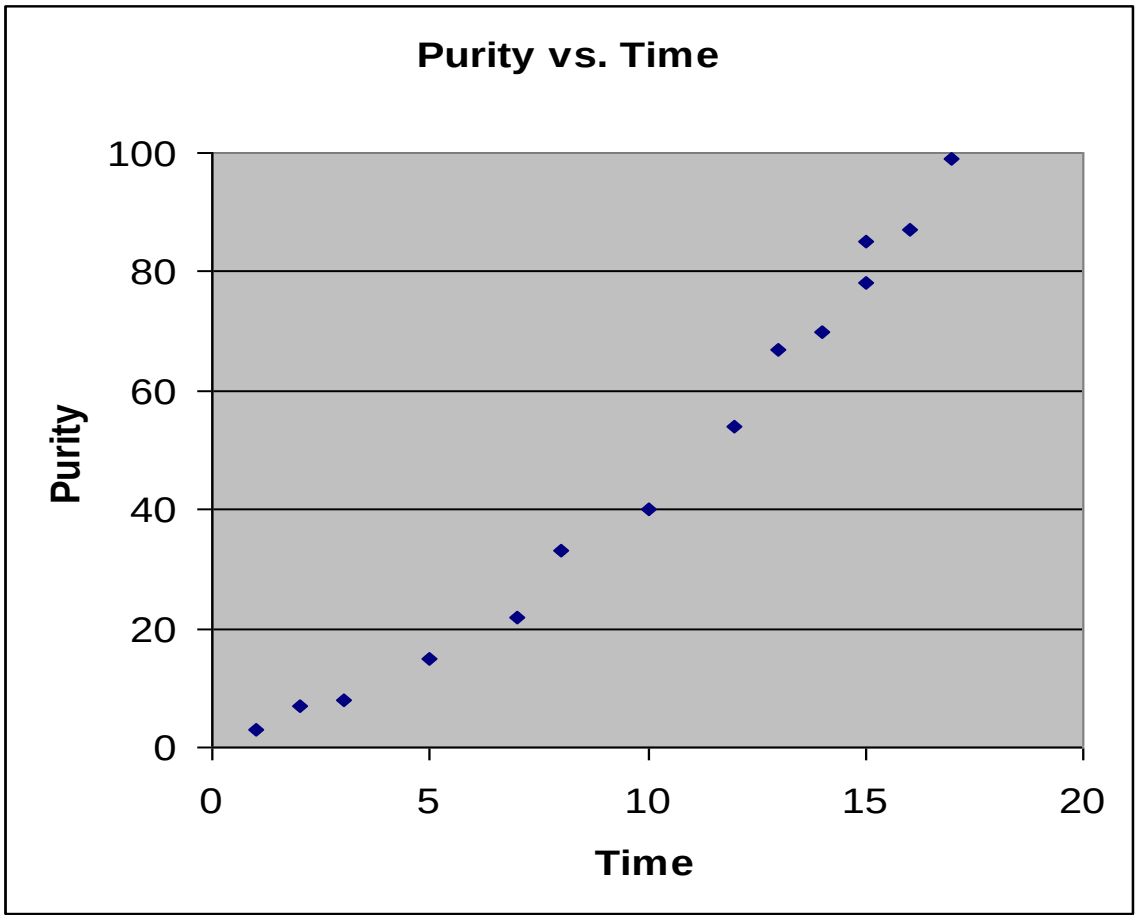
Compare R^2 from simple regression to $\overline{R}^2$ from the quadratic model

- If $\overline{R}^2$ from the quadratic model is larger than R^2 from the simple model, then the quadratic model is a better model

Quadratic regression: example

Purity	Filter Time
3	1
7	2
8	3
15	5
22	7
33	8
40	10
54	12
67	13
70	14
78	15
85	15
87	16
99	17

- Purity increases as filter time increases:



Quadratic regression: example (2)

- Simple regression results:

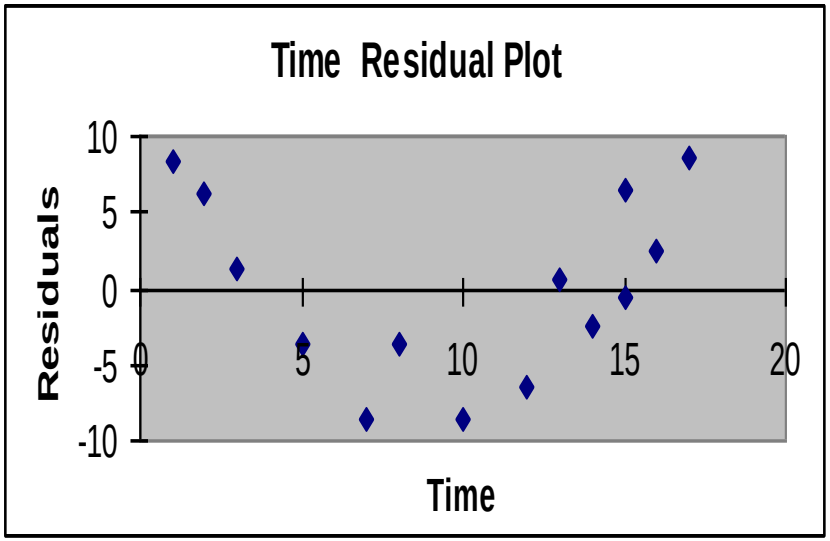
$\hat{y}$ = -11.283 + 5.985 Time

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-11.28267	3.46805	-3.25332	0.00691
Time	5.98520	0.30966	**19.32819**	2.078E-10

Regression Statistics	
R Square	**0.96888**
Adjusted R Square	0.96628
Standard Error	**6.15997**

F	*Significance F*
373.57904	2.0778E-10

t statistic, F statistic, and R^2 are all high, but the residuals are not random:



Quadratic regression: example (3)

- Quadratic regression results:

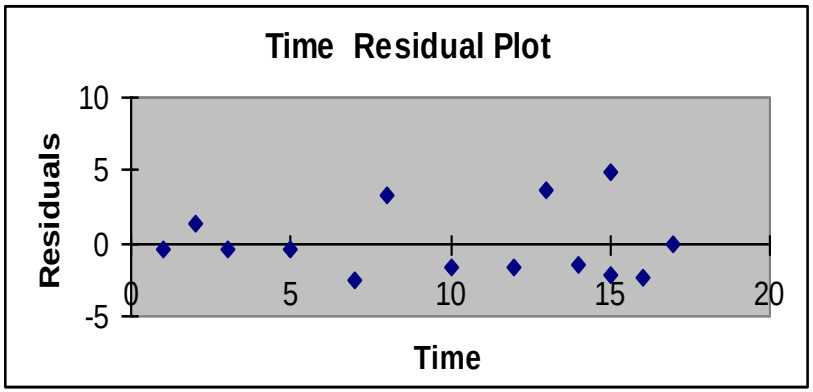
$$\hat{y} = 1.539 + 1.565\ \text{Time} + 0.245\ (\text{Time})^2$$

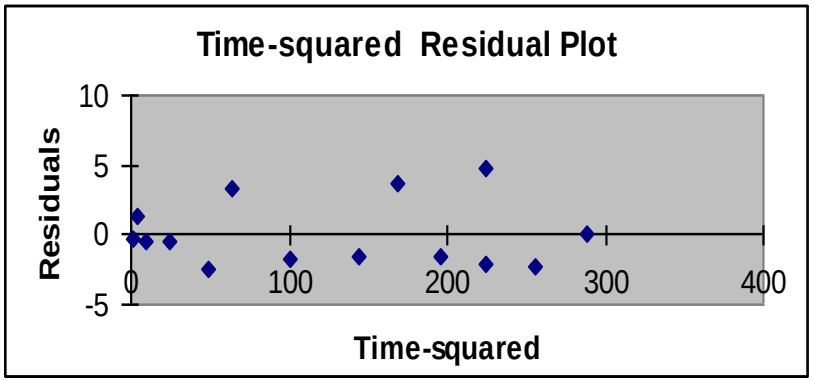
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.53870	2.24465	0.68550	0.50722
Time	1.56496	0.60179	2.60052	0.02467
Time-squared	0.24516	0.03258	**7.52406**	1.165E-05

Regression Statistics	
R Square	0.99494
Adjusted R Square	**0.99402**
Standard Error	**2.59513**

F	*Significance F*
1080.7330	2.368E-13

The quadratic term is significant and improves the model: $\overline{R}^2$ is higher and s_e is lower, residuals are now random





Dummy variables

- A dummy variable is a categorical independent variable with two levels:
  - yes or no, on or off, male or female
  - recorded as 0 or 1
- Regression intercepts are different if the variable is significant
- Assumes equal slopes for other variables
- If more than two levels, the number of dummy variables needed is (number of levels - 1)

Dummy variables: example

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Let:

y = Pie Sales

x_1 = Price

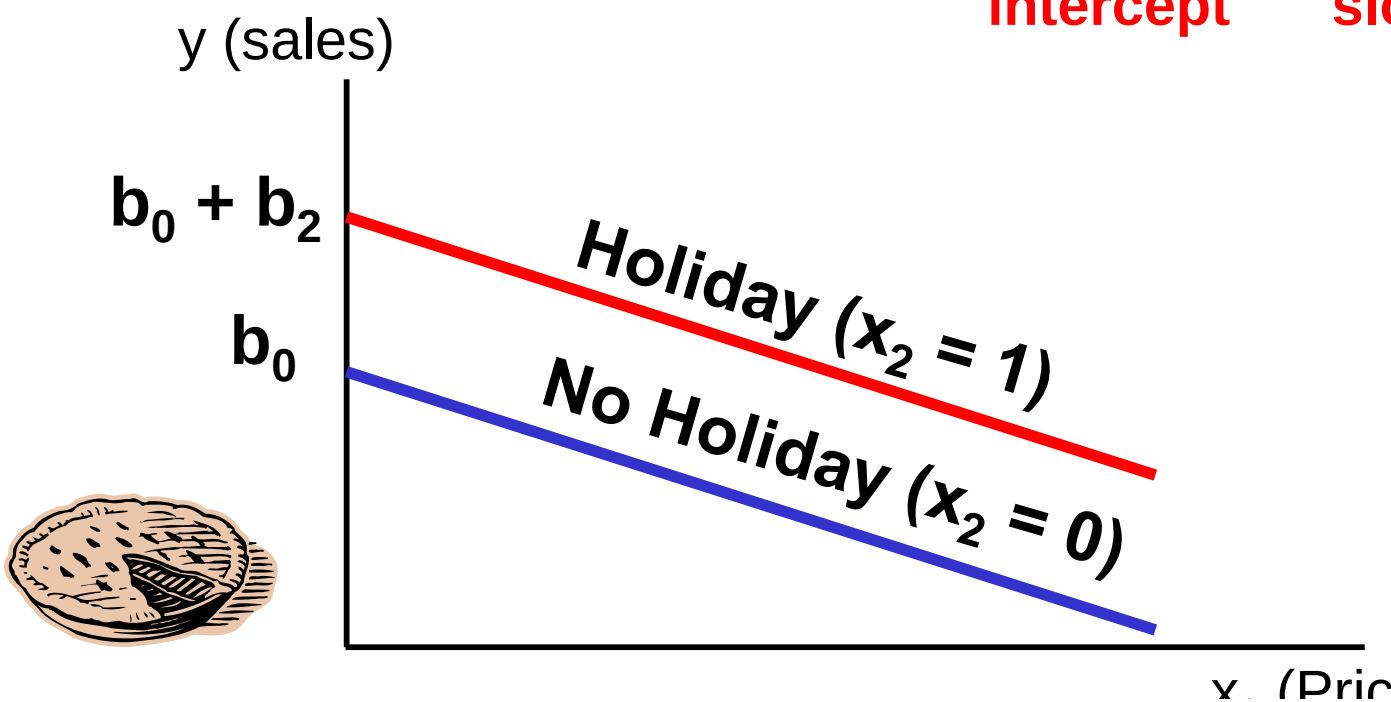
x_2 = Holiday (X_2 = 1 if a holiday occurred during the week)
(X_2 = 0 if there was no holiday that week)

Dummy variables: example (2)

$$\hat{y} = b_0 + b_1 x_1 + b_2(1) = (b_0 + b_2) + b_1 x_1 \quad \textbf{Holiday}$$

$$\hat{y} = b_0 + b_1 x_1 + b_2(0) = b_0 + b_1 x_1 \quad \textbf{No Holiday}$$

Different intercept **Same slope**



If $H_0: \beta_2 = 0$ is rejected, then "Holiday" has a significant effect on pie sales

Dummy variables: example (3)

Example: $Sales = 300 - 30(Price) + 15(Holiday)$

Sales: number of pies sold per week

Price: pie price in $

$$Holiday: \begin{cases} 1 & \text{If a holiday occurred during the week} \\ 0 & \text{If no holiday occurred} \end{cases}$$

$b_2 = 15$: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price

Residual analysis

- These residual plots are used in multiple regression:
  - Residuals vs. $\hat{y}_i$
  - Residuals vs. x_{1i}
  - Residuals vs. x_{2i}
  - Residuals vs. time (if time series data)

Use the residual plots to check for violations of regression assumptions

Acknowledgments

Slides are adapted from those accompanying Newbold's textbook

References

See Chapters 1-8 of [1] for more.

[1] J. Verzani.
Using R for Introductory Statistics, Second Edition.
Chapman & Hall/CRC The R Series. Taylor & Francis, 2014.