Inferential and Predictive Statistics for Business

Module 4

ILLINOIS

Gies College of Business

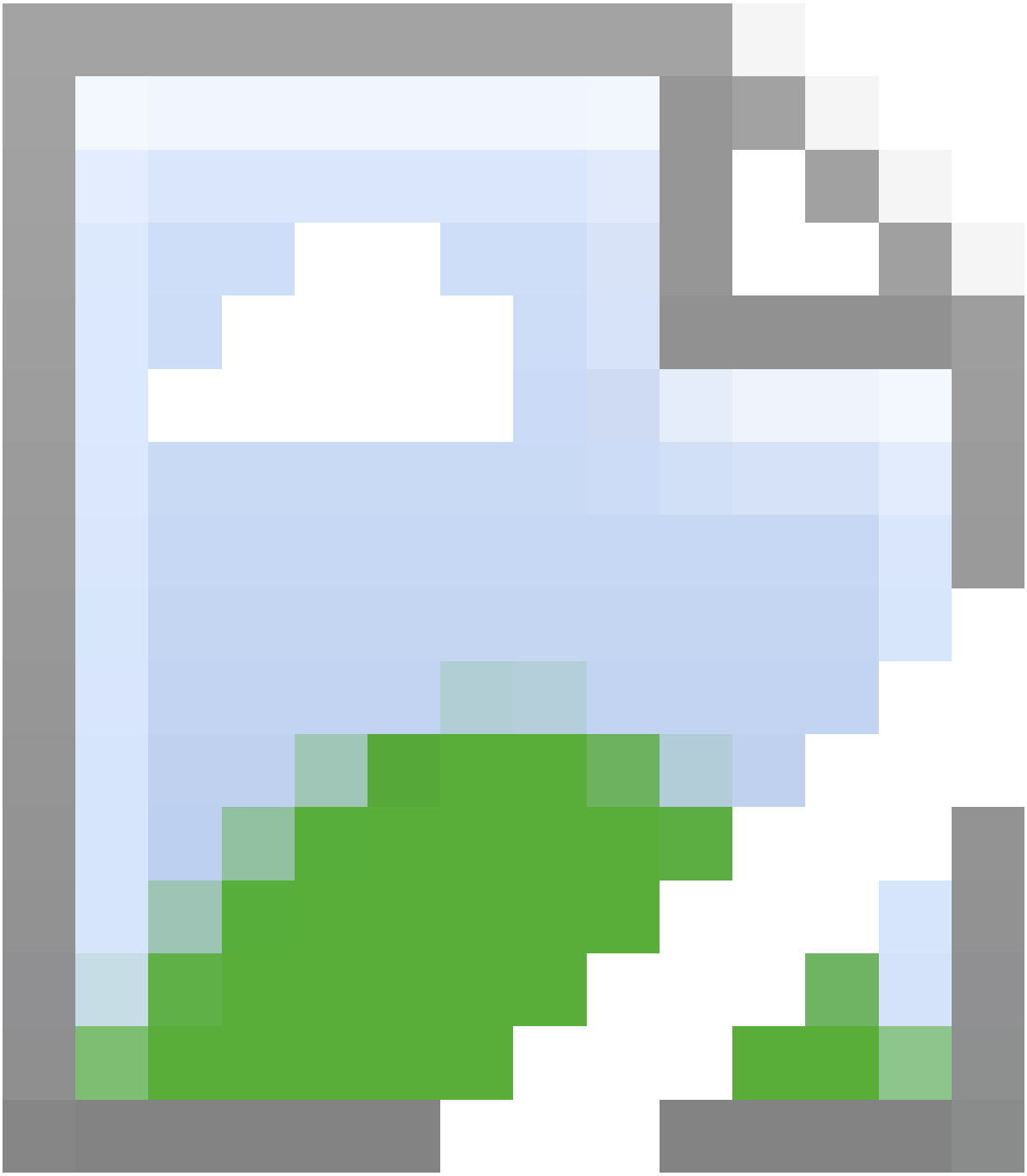


Table of Contents

1. Preface
2. Module 4 Inferential and Predictive Statistics for Business
    1. Lesson 4-1 Model Introduction
        1. Lesson 4-1.1 Model Introduction
        2. Lesson 4-1.2 Developing Multiple Regression Model in Excel
        3. Lesson 4-1.3 Multiple Regression: Excel Output Analysis
    2. Lesson 4-2 Qualitative Data – Two Categories
        1. Lesson 4-2.1 Qualitative Data – Two Categories
        2. Lesson 4-2.2 Dummy Variable in Excel – Two Categories
    3. Lesson 4-3 Qualitative Data – Many Categories
        1. Lesson 4-3.1 Qualitative Data – Many Categories
        2. Lesson 4-3.2 Dummy Variable in Excel – Many Categories

Preface

Thank you for choosing a Gies eBook.

This Gies eBook is based on an extended video lecture transcript made from Module 4 of Professor Fataneh Taghaboni-Dutta's *Inferential and Predictive Statistics for Business* on Coursera. The Gies eBook provides a reading experience that covers all of the information in the MOOC videos in a fully accessible format. The Gies eBook can be used with any standards-based e-reading software supporting the ePUB 3.0 format.

Each Gies eBook is broken down by lessons that are navigable using our e-reader's table of contents feature. Within each lesson the following sequence of content will always occur:

- Lesson title
- A link to the web-based videos for each lesson (You must be online to view.)

Within the lesson, every time there is a slide change or a switch to the next informative video scene, you will be presented with:

- Thumbnail image of the current slide or video scene
- Any text present on the slide in the video is recreated below the thumbnail in a searchable, screen reader-ready format.
- Extended text description of the important visuals such as graphs and charts presented in the slides.
- Any tabular data from the video is recreated and properly labeled for screen reader navigation and reading.
- All math equations are presented in MathML that provides both content and presentation if on screen.
- Transcript that captures all of the original speech in the video labeled by the person speaking.

All Gies eBooks are designed with accessibility and usability as a priority. This design is intended to serve all readers in a flexible manner regardless of their choice of digital reading tools.

If you have any questions or suggestions for improvement for this Gies eBook, please contact Giesbooks@illinois.edu

Copyright © 2019 by Fataneh Taghaboni-Dutta

All rights reserved.

Published by the Gies College of Business at the University of Illinois at Urbana-Champaign, and the Board of Trustees of the University of Illinois

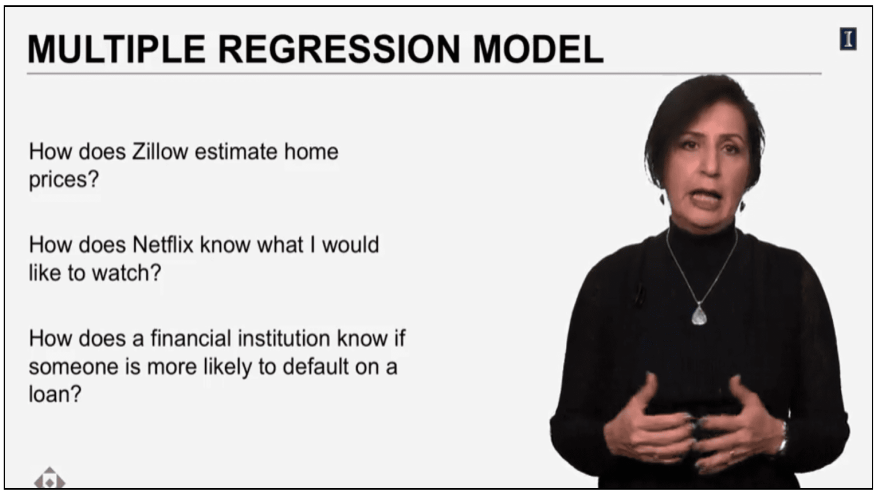
Module 4 Inferential and Predictive Statistics for Business

Lesson 4-1 Model Introduction

Lesson 4-1.1 Model Introduction

Media Player for Video

Multiple Regression Model (1 of 2) - Slide 1



How does Zillow estimate home prices?

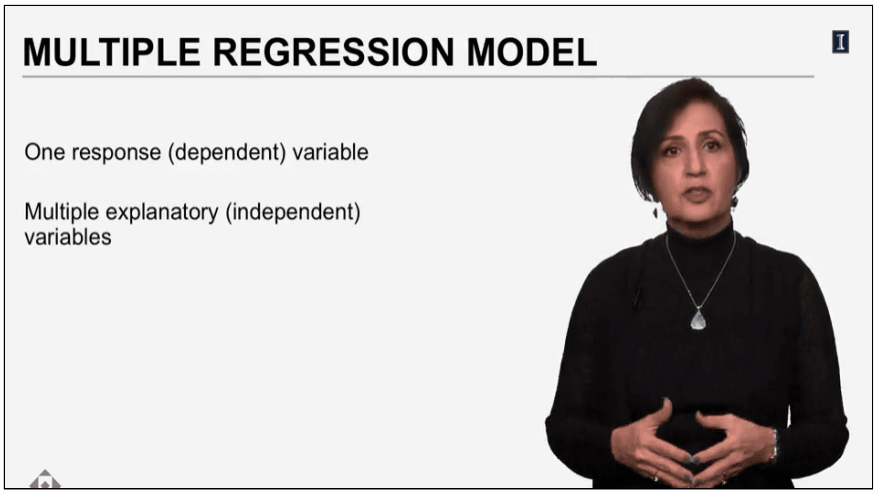
How does Netflix know what I would like to watch?

How does a financial institution know if someone is more likely to default on a loan?

Transcript

How does Zillow estimate home prices? How does Netflix know what I would like to watch? How does a financial institution know if someone is more likely to default on a loan? These are similar questions to this in the last module when you learned about simple linear regression. In this module, we will learn about multiple regression, which will take us one step further in our analysis.

Multiple Regression Model (2 of 2) - Slide 2



One response (dependent) variable

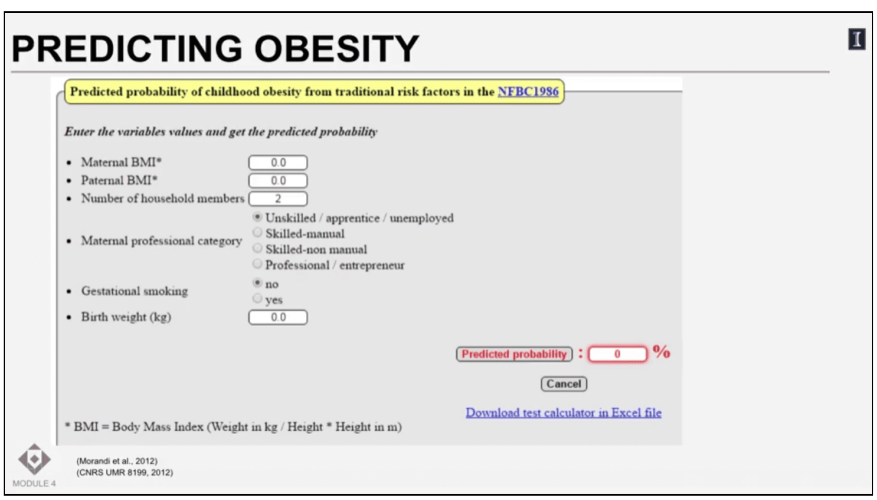
Multiple explanatory (independent) variables

Transcript

The main difference in multiple regression as compared to simple linear regression is the ability to have more than one explanatory variable in our analysis. We still have one response variable, y, but now, we can have many explanatory variables, which are our x's.

There is no limit to the number of independent variables a model can use, which means multiple regression models can handle more complex situations than the simple linear regression models.

Predicting Obesity - Slide 3



This slide contains a fillable form that calculates predicted probability of childhood obesity.

(Morandi et al., 2012)

(CNRS UMR 8199, 2012)

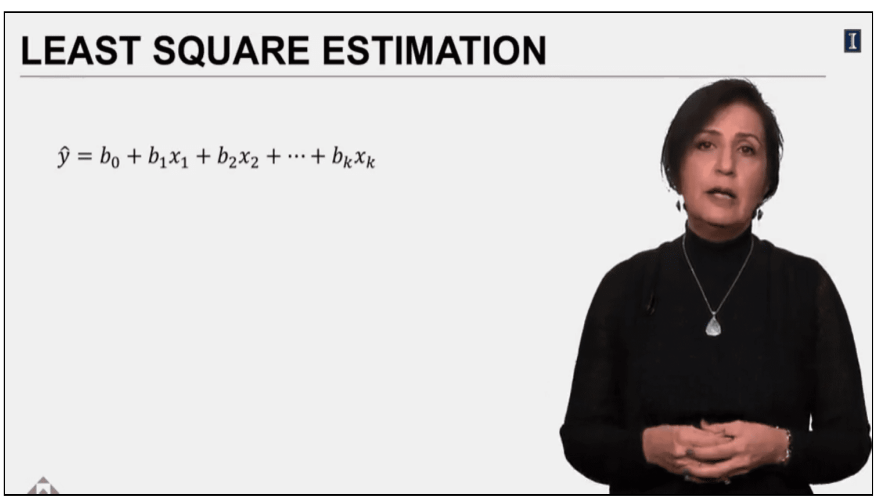
Transcript

A study done on newborns established a predictive equation for childhood and adolescent obesity. The goal of the study was to know which baby was the greatest risk so preventive measures can change this outcome for the newborn. The study is looking at the probability of a baby becoming obese based on several factors: mother's BMI, father's BMI, number of people living in a household, mother's professional category, if the mother smoked during pregnancy, and finally, the baby's birth weight.

So, in total, there are seven independent variables that are being used to predict the response variable. Some of these variables are numerical values, the number of people in a household for example or birth weight. And some are categorical like whether the mother smoked or not. In this module, we will learn how to find which variables are good to have in the prediction model as well as know how to develop the formula which will would allow us to do the prediction.

By the way, this is a really fun calculator to use, the link is provided in the reference section, go ahead and play with this, you may be surprised by the results, at least for me some were counterintuitive, that's what made it fun.

Least Square Estimation - Slide 4

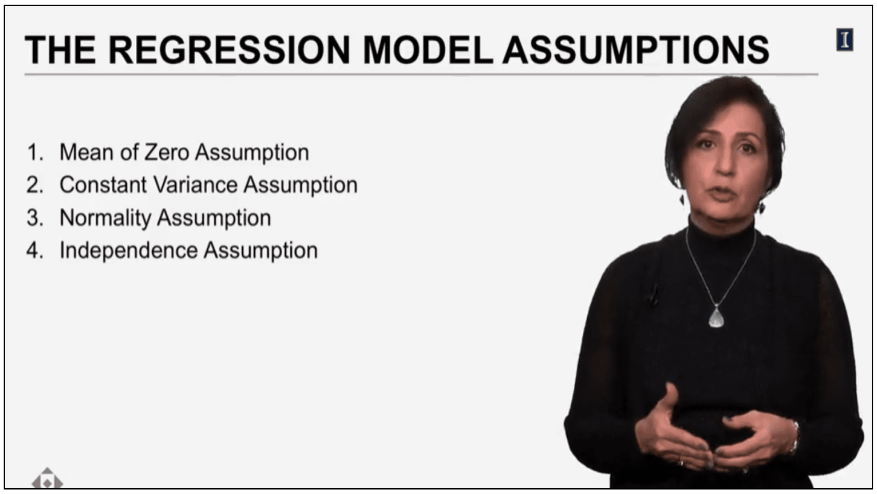


$\hat{y} = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k$

Transcript

Multiple regression would also use the least square method to find the regression equation. But now, it has a slope for each x-variable. ŷ is the point estimate of the mean value of the dependent variable, when the values of the independent variables are x1, x2, x3 and so on. Remember, ŷ is an estimate so you'll have some errors, just as we discussed for simple linear regression.

The Regression Model Assumptions - Slide 5

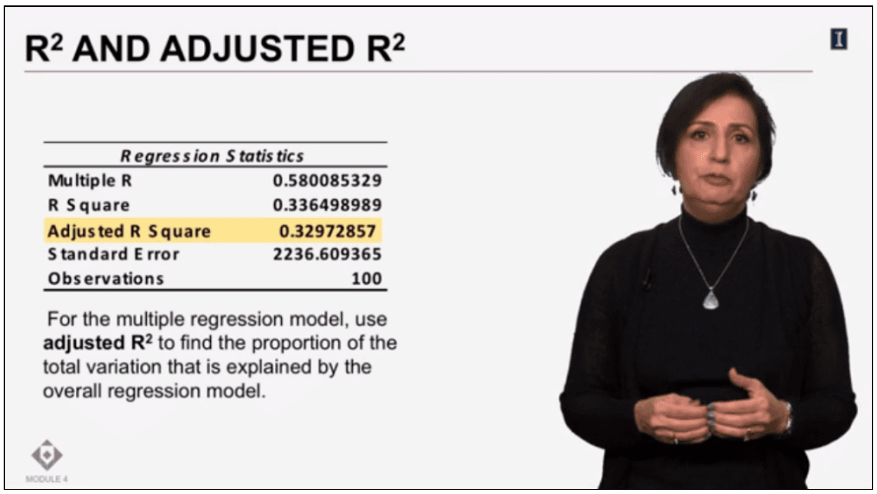


1. Mean of Zero Assumption
2. Constant Variance Assumption
3. Normality Assumption
4. Independence Assumption

Transcript

The assumptions for single linear regression, which we learned in the last module, must hold for multiple regression as well. The validity of the assumptions is checked by analyzing the residuals, those are the errors in the prediction, just as we did for simple linear regression.

r^2 and Adjusted r^2 - Slide 6



Regression statistics

Multiple R	0.580085329
R Square	0.336498989
Adjusted R Square	0.32972857
Standard Error	2236.609365
Observations	100

Adjusted R Square is highlighted in the table.

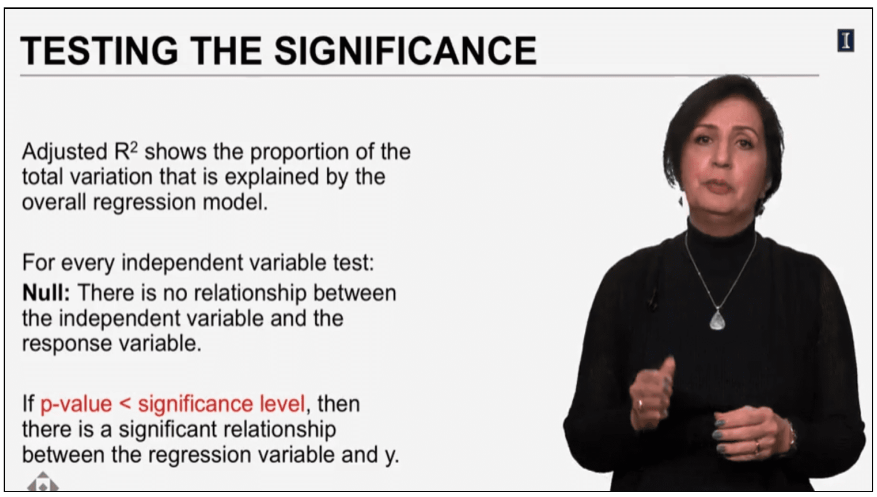
For the multiple regression model, use **adjusted** r^2 to find the proportion of the total variation that is explained by the overall regression model.

Transcript

There was one element of Excel output in the top table that I never reviewed when we were learning about simple linear regression, and that is adjusted R-squared. We have learned that R-squared is the proportion of the total variations that is explained by the overall regression model. When we have multiple regression, the valuable R-squared will increase even if the new variables being added to the model has no relationship to y. Adjusted R-squared corrects this tendency in R-squared.

As you can see here, adjusted square is smaller than R-squared and this will always be the case. While we used R-squared in simple linear regression to know what percent of variations were explained by the regression model, in multiple regression, it will be use the adjusted R-squared to find the proportion.

Testing the Significance - Slide 7



$\text{Adjusted } R^2$ shows the proportion of the total variation that is explained by the overall regression model.

For every independent variable test: **Null**: There is no relationship between the independent variable and the response variable.

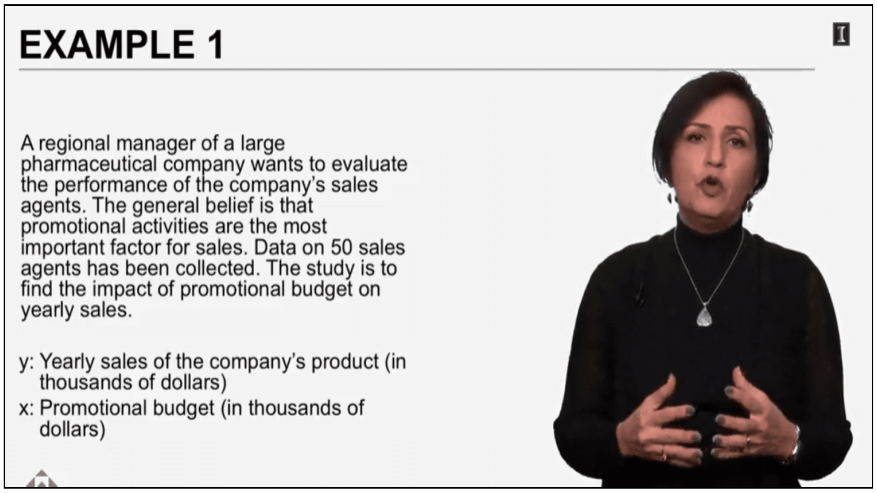
If **p-value < significance level**, then there is a significant relationship between the regression variable and y.

Transcript

While adjusted R-squared value will tell us how good the regression model is in explaining the variations, the significance of each independent variable is assessed separately. In multiple regression, it is customary to test the significance of every independent variable used in the model.

Again, the null hypothesis for each independent variable state that no relationship exists between the independent variable and the response variable. Therefore, if p-value is less than significance level, then we would be rejected the null hypothesis, meaning that there is a significant relationship between the regression variable and the response variable.

Example 1 - Slide 8



A regional manager of a large pharmaceutical company wants to evaluate the performance of the company's sales agents. The general belief is that promotional activities are the most important factor for sales. Data on 50 sales agents has been collected. The study is to find the impact of promotional budget on yearly sales.

y: Yearly sales of the company's product (in thousands of dollars)

x: Promotional budget (in thousands of dollars)

Transcript

Let's begin with an example. A regional manager of a large pharmaceutical company wants to evaluate the performance of the company's sales force. The general belief is that marketing and advertising is the most important factor for sales. Data on 50 sales agents has been collected. The study objective is to find the impact of promotional budget on yearly sales. So, we have one explanatory variable here and we will run a simple linear regression.

Excel Output: Simple Linear Regression - Slide 9

EXCEL OUTPUT – SIMPLE LINEAR REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.63587376
R Square	0.40433544
Adjusted R Square	0.39192577
Standard Error	1080.06248
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	38008353.3	38008353.3	32.5823	6.97799E-07
Residual	48	55993678.5	1166534.97		
Total	49	94002031.8			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	8043.79161	387.114414	20.7788481	1.2E-25	7265.445917	8822.137311
Promotional Budget	0.5793468	0.10149578	5.70808782	7E-07	0.375275869	0.78341774

Summary Output

Regression Statistics - Slide 9

Multiple R	0.63587376
R Square	**0.40433544**
Adjusted R Square	0.39192577
Standard error	1080.06248
Observations	50

ANOVA - Slide 9

	df	SS	MS	F	Significance F
Regression	1	38008353.3	38008353.3	32.5823	6.97799E-07
Residual	48	55993678.5	1166534.97		
Total	49	94002031.8			

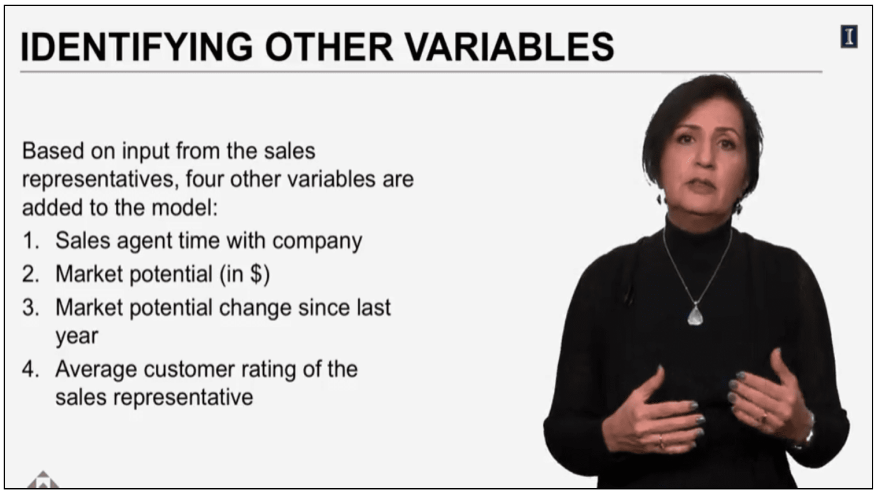
Other Statistics - Slide 9

	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	8043.79161	387.114414	20.7788481	1.2E-25	7265.445917	8822.137311
Promotional budget	0.5793468	0.10149578	5.70808782	**7E-07**	0.375275869	0.78341774

Transcript

P-value for promotional budget is very small, thus there is a significant relationship between promotional budget and sales; however, only 40% of the variation is explained by this variable. This is not satisfactory; the general manager has asked his analysts to develop a better model.

Identifying Other Variables - Slide 10



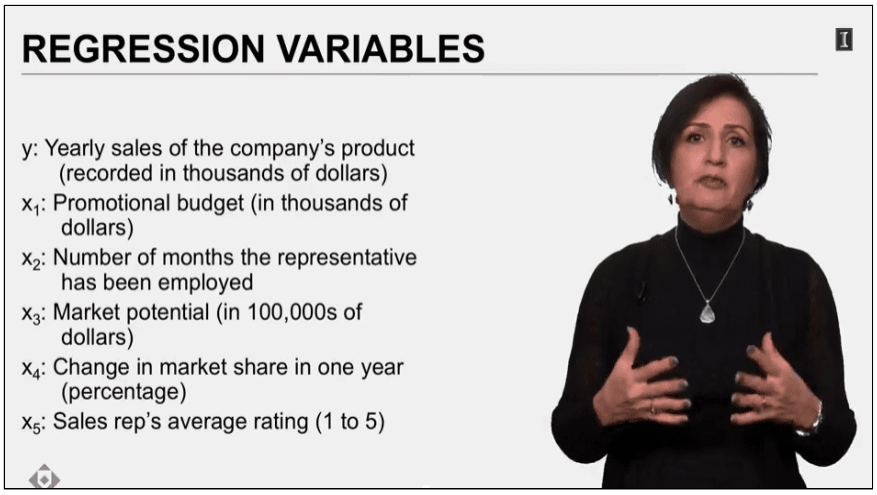
Based on input from the sales representatives, four other variables are added to the model:

1. Sales agent time with company
2. Market potential (in $)
3. Market potential change since last year
4. Average customer rating of the sales representative

Transcript

Based on input from the sales representative, four other variables are added to the model: sales agent time with the company, market potential in dollars, market potential change since last year in dollars and average customer rating of the sales representative.

Regression Variables - Slide 11



y : Yearly sales of the company's product (recorded in thousands of dollars)

x_1 : Promotional budget (in thousands of dollars)

x_2 : Number of months the representative has been employed

x_3 : Market potential (in 100,000s of dollars)

x_4 : Change in market share in one year (percentage)

x_5 : Sales rep's average raiting (1 to 5)

Transcript

So, now y is assumed to be influence by five different independent variables, we now have to collect data for all these variables and then we can run the multiple regression model.

Excel Output: Multiple Regression - Slide 12

EXCEL OUTPUT – MULTIPLE REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8562096
R Square	0.7330949
Adjusted R Square	0.7027648
Standard Error	75.512825
Observations	50

This model is better than the simple linear regression with R = 0.40

ANOVA

	df	SS	MS	F	Significance F
Regression	5	689124.0983	137824.8	24.171	1.29955E-11
Residual	44	250896.2198	5702.187		
Total	49	940020.318			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	520.08453	73.8516445	7.042288	1E-08	371.2463239	668.92274
Promotional Budget	3.0146845	0.840772413	3.585613	0.0008	1.32021902	4.7091499
Time	5.2706174	1.132557877	4.653729	3E-05	2.988097025	7.5531379
Market Potential	0.0191646	0.006074822	3.154757	0.0029	0.006921585	0.0314076
Market Share Change	-74.322586	98.78631602	-0.752357	0.4558	-273.4133239	124.76815
Customer Rating	51.121931	17.83372083	2.866588	0.0063	15.18042783	87.063433

Summary Output

Regression Statistics - Slide12

Multiple R	0.8562096
R Square	0.7330949
Adjusted R Square	**0.7027648**
Standard error	75.512825
Observations	50

ANOVA - Slide 12

	df	SS	MS	F	Significance F
Regression	5	689124.0983	137824.8	24.171	1.29955E-11
Residual	44	250896.2198	5702.187		
Total	49	940020.318			

Other Statistics - Slide 12

	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	520.08453	73.8516445	7.042288	1E-08	371.2463239	668.92274
Promotional budget	3.0146845	0.840772413	3.585613	0.0008	1.32021902	4.7091499
Time	5.2706174	1.132557877	4.653729	3E-05	2.988097025	7.5531379
Market potential	0.0191646	0.006074822	3.154757	0.0029	0.006921585	0.0314076
Market Share change	−74.322586	98.78631602	−0.752357	0.4558	−273.4133239	124.76815
Customer rating	51.121931	17.83372083	2.866588	0.0063	15.18042783	87.063433

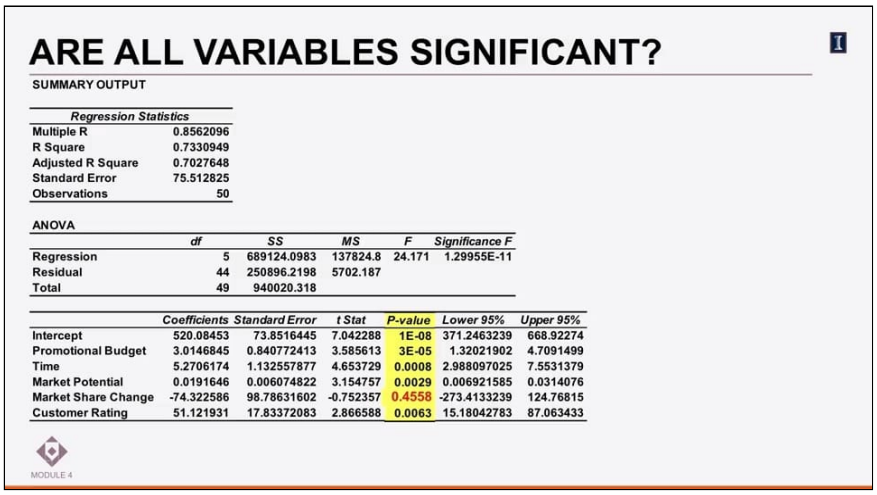
This model is better than the simple linear regressions with R = 0.40

Transcript

Here's the output for the multiple regression – here, the adjusted R-squared is 0.70, which means the model now can explain 70% of the variations seen in sales, as compared to the simple linear regression where we could only explain 40% of the variations. So, the new model is definitely better than the simple linear regression model.

Note that I'm using the adjusted R-squared, as I explained earlier, R-squared will increase even if the new variables being added to the model has no relationship to y. Adjusted R-squared corrects this tendency in R-squared.

Are All Variables Significant? - Slide 13



ARE ALL VARIABLES SIGNIFICANT?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8562096
R Square	0.7330949
Adjusted R Square	0.7027648
Standard Error	75.512825
Observations	50

ANOVA

	df	SS	MS	F	Significance F
Regression	5	689124.0983	137824.8	24.171	1.29955E-11
Residual	44	250896.2198	5702.187		
Total	49	940020.318			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	520.08453	73.8516445	7.042288	1E-08	371.2463239	668.92274
Promotional Budget	3.0146845	0.840772413	3.585613	3E-05	1.32021902	4.7091499
Time	5.2706174	1.132557877	4.653729	0.0008	2.988097025	7.5531379
Market Potential	0.0191646	0.006074822	3.154757	0.0029	0.006921585	0.0314076
Market Share Change	-74.322586	98.78631602	-0.752357	0.4558	-273.4133239	124.76815
Customer Rating	51.121931	17.83372083	2.866588	0.0063	15.18042783	87.063433

MODULE 4

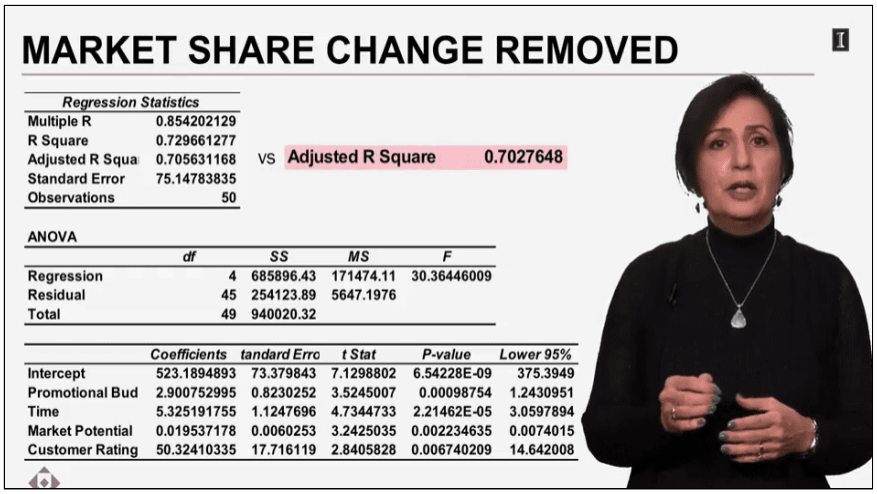
This slide contains the same tables shown on Slide 12 - Excel Output: Multiple Regression with the P-value column of the third table: Other Statistics - Slide 12 highlighted.

Transcript

While the model is better, we still need to make sure that we only have variables in our model that have an effect on the response variable. This means that we need to evaluate each independent variables' p-value to establish its significance. To do this, we will focus on the third table, and on the p-values of each variable.

Any variable with a p-value less than 0.05 would be considered significant. We will use this as the level of significance here; you could use 0.01 or anything else. If the p-value is greater than 0.05, then the variable doesn't have a significant relationship with the response variable sales. In this case, the variable market share change has a p-value that is greater than the value of 0.05, thus it's not significant. When you find a variable that it's not significant, we will remove it from the model and run the analysis again.

Market Share Change Removed - Slide 14



Summary Output

Regression Statistics - Slide 14

Multiple R	0.854202129
R Square	0.729661277
Adjusted R Square	0.705631168
Standard error	75.14783835
Observations	50

Next to the Adjusted R square is a note: **vs. Adjusted R square 0.7027648**.

ANOVA - Slide 14

	df	SS	MS	F
Regression	4	685896.43	171474.11	30.36446009
Residual	45	254123.89	5647.1976	
Total	49	940020.32		

Other Statistics - Slide 14

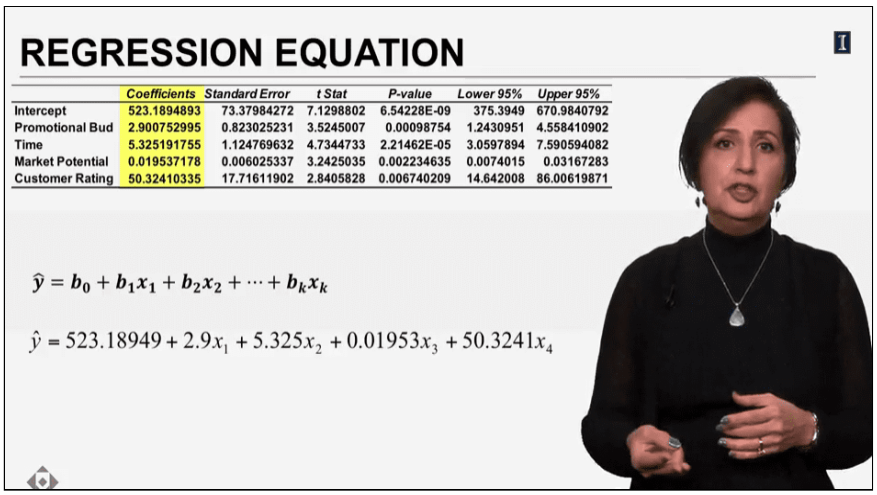
	Coefficients	Standard error	t-stat	p-value	Lower 95%
Intercept	523.1894893	73.379843	7.1298802	6.54228E-09	375.3949
Promotional budget	2.900752995	0.8230252	3.5245007	0.00098754	1.2430951
Time	5.325191755	1.1247696	4.7344733	2.21462E-05	3.0597894
Market potential	0.019537178	0.0060253	3.2425035	0.002234635	0.0074015
Customer rating	50.32410335	17.716119	2.8405828	0.006740209	14.642008

Transcript

This is the result of the analysis without the market share change, the adjusted R-squared has not changed much, actually has improved a bit, compared to before it's 0.702. So, the regression model has not suffered from removing this variable because this variable was not useful in explaining the variations in sales.

Now, look at the p-values for all the remaining variables, everything now has a significant relationship with the response variable of sales.

Regression Equation - Slide 15



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	523.1894893	73.37984272	7.1298802	6.54228E-09	375.3949	670.9840792
Promotional Bud	2.900752995	0.823025231	3.5245007	0.00098754	1.2430951	4.558410902
Time	5.325191755	1.124769632	4.7344733	2.21462E-05	3.0597894	7.590594082
Market Potential	0.019537178	0.006025337	3.2425035	0.002234635	0.0074015	0.03167283
Customer Rating	50.32410335	17.71611902	2.8405828	0.006740209	14.642008	86.00619871

Other Statistics - Slide 15

	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	523.1894893	73.379843	7.1298802	6.54228E-09	375.3949	670.9840792
Promotional budget	2.900752995	0.8230252	3.5245007	0.00098754	1.2430951	4.558410902
Time	5.325191755	1.1247696	4.7344733	2.21462E-05	3.0597894	7.590594082
Market potential	0.019537178	0.0060253	3.2425035	0.002234635	0.0074015	0.03167283
Customer rating	50.32410335	17.716119	2.8405828	0.006740209	14.642008	86.00619871

The Coefficients column is highlighted

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k$$

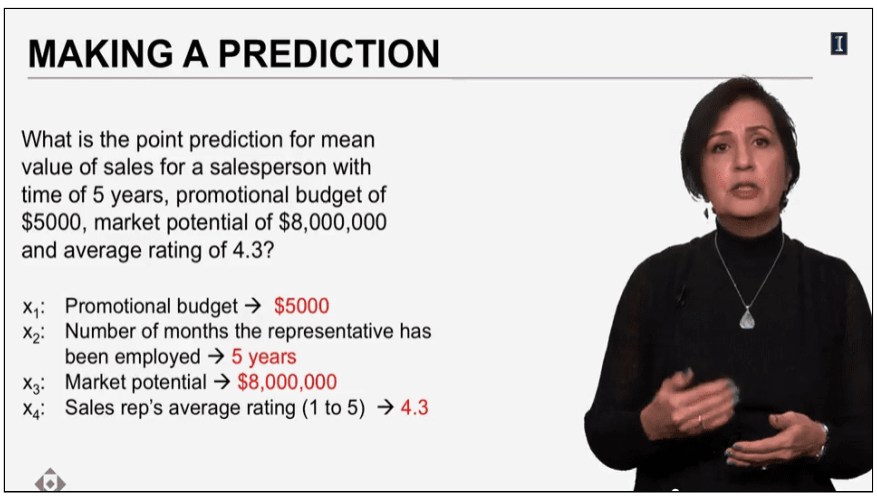
$$\hat{y} = 523.18949 + 2.9x_1 + 5.325x_2 + 0.01953x_3 + 50.3241x_4$$

Transcript

Now that we're satisfied with this model we can write the regression equation by looking at numbers to place in b0 and all the coefficients of the independent variables, which we will find in the third table of Excel output.

Thus, this is our regression equation for sales and its relationship with length of sales agent's tenure, the promotional budget, market potential and average customer ratings of sales agents.

Making a Prediction (1 of 2) - Slide 16



What is the point prediction for mean value of sales for a salesperson with time of 5 years, promotional budget of $5000, market potential of $8,000,000 and average rating of 4.3?

$X_1:$ Promotional budget $\rightarrow$**$5000**

$X_2:$ Number of months the representative has been employed $\rightarrow$**5 years**

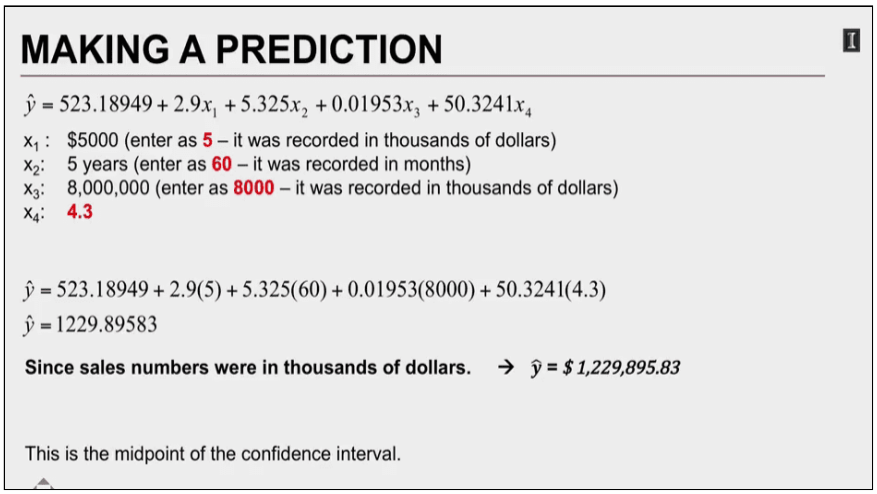
$X_3:$ Market potential $\rightarrow$**$8,000,000**

$X_4:$ Sales rep's average rating (1 to 5) $\rightarrow$**4.3**

Transcript

Now you can use this equation and make prediction. What is the point prediction value of sales for a salesperson with five years of experience, promotional budget of $5,000 and market potential of 8 million and average rating of 4.3? X, which is sales reps promotional budget here is 5,000, that's x1; x2 is the number of months the representative has been employed and it's in 60 months because we had it recorded months; market potential for sales agent is 8 million and sales reps has a rating of 4.3. Now, place these values in the regression equation.

Making a Prediction (2 of 2) - Slide 17



$\hat{y} = 523.18949 + 2.9x_1 + 5.325x_2 + 0.01953x_3 + 50.3241x_4$

x_1 : $\$5000$ (enter as **5** - it was recorded in thousands of dollars)

x_2 : 5 years (enter as **60** - it was recorded in months)

x_3 : $8{,}000{,}000$ (enter as **8000** - it was recorded in thousands of dollars)

x_4 : **4.3**

$\hat{y} = 523.18949 + 2.9(5) + 5.325(60) + 0.01953(8000) + 50.3241(4.3)$

$\hat{y} = 1229.89583$

Since sales number were in thousands of dollars. $\rightarrow$ $\hat{y} =$ ***$1229.89583***

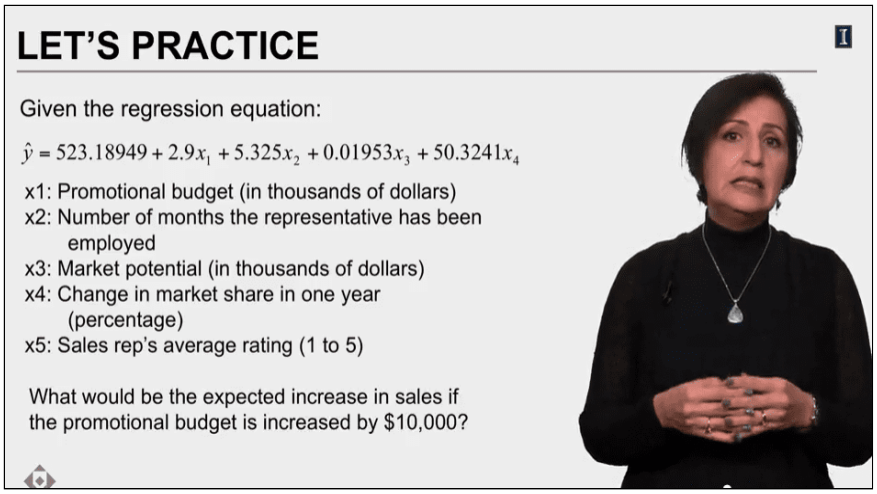
This is the midpoint of the confidence interval.

Transcript

Given the regression equation and the values of the independent variables, the estimated value of y sales is 1,229.89583, just remember to express it in a correct way.

We have recorded and analyzed the data by expressing the sales in thousands of dollars, so we need to multiply the prediction value by 1,000, which gives us $1,229,895.83. And this is the midpoint of the confidence interval.

Let's Practice - Slide 18



Given the regression equation: $\hat{y} = 523.18949 + 2.9x_1 + 5.325x_2 + 0.01953x_3 + 50.3241x_4$

x_1 : Promotional budget (in thousands of dollars)

x_2 : Number of months the representative has been employed

x_3 : Market potential (in thousands of dollars)

x_4 : Change in market share in one year (percentage)

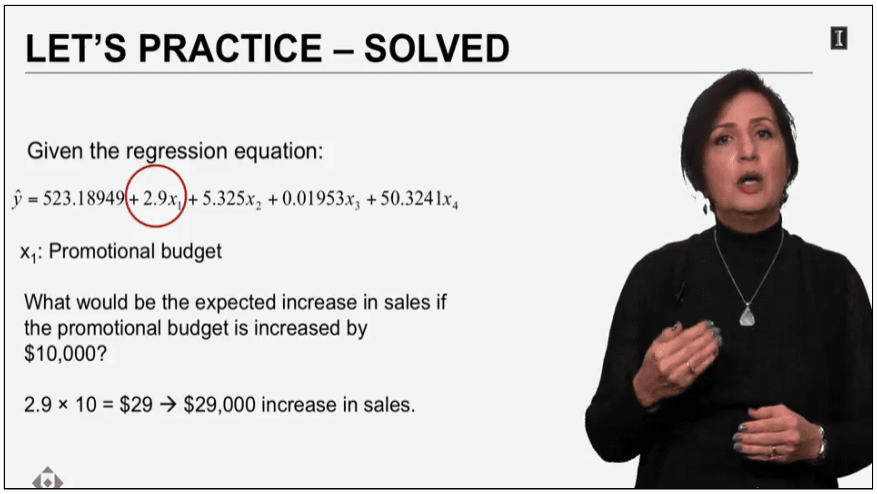
x_5 : Sales rep's average rating (1 to 5)

What would be the expected increase in sales if the promotional budget is increased by $10,000?

Transcript

Now let's practice, given the regression equation we just developed, what would be the expected increase in sales if the promotional budget is increased by $10,000?

Let's Practice - Solved - Slide 19



Given the regression equation: $\hat{y} = 523.18949 + 2.9x_1 + 5.325x_2 + 0.01953x_3 + 50.3241x_4$

$+2.9x_1$ is circled.

x_1 : Promotional budget

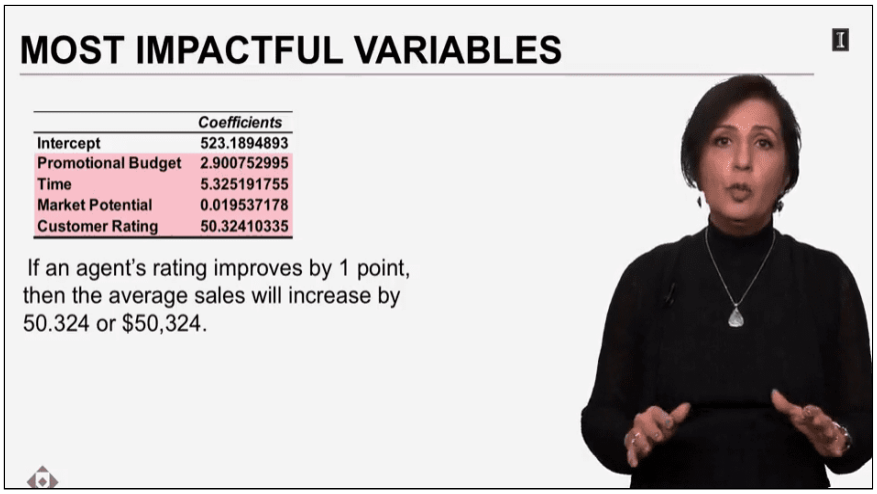
What would be the expected increase in sales if the promotional budget is increased by $10,000?

2.9 × 10 = $29 → $29,000 increase in sales.

Transcript

x1 is the promotional budget and has a positive coefficient of 2.9. To answer the question asked, $10,000 increase would be entered as ten, and that would result in an additional value to the sales by $29,000. Now you may ask yourself, what factors have the greatest impact on the sales? We can find this out by looking at the coefficients.

Most Impactful Variables - Slide 20



Other Statistics - Slide 20

	Coefficients
Intercept	523.1894893
Promotional budget	2.900752995
Time	5.325191755
Market potential	0.019537178
Customer rating	50.32410335

Promotional Budget, Time, Market Potential, and Customer Rating are highlighted.

If an agent's rating improves by 1 point, then the average sales will increase by 50.324 or $50,324.

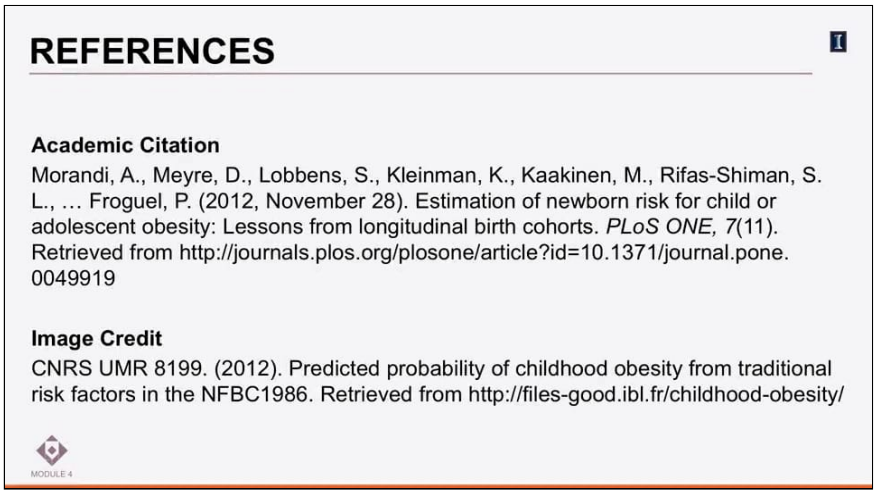
Transcript

Focusing on the coefficients on independent variables, we see that the customer rating has the largest coefficient followed by the time the agent has stayed with the company. If agents' rating improved by one point, then the average sales would be increase by 50.324 or $50,324.

The analysis has made us aware that the better customer rating has the most positive impact on the performance of the agent, more than the promotional budget which originally was thought to be the most important factor. Furthermore, we see as the agent's length of staying increases, so does their ability to sell more, looking at this data, it may be a good idea to provide the work environment where sales' agents are happy employees who stay on and their professional satisfaction results in superior customers service. All of which would be very good for the company's bottom line.

These are the type of revealing insights we get from multiple regression, ability to identify what variables influence the response variable and to what degree. As always, please watch the Excel demo videos for more examples.

References - Slide 21



Academic Citation:

Morandi, A., Meyre, D., Lobbens, S., Kleinman, K., Kaakinen, M., Rifas-Shiman, S. L., ... Froguel, P. (2012, November 28). *Estimation of newborn risk for child or adolescent obesity: Lessons from longitudinal birth cohorts.* PLoS ONE, 7(11). Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049919.

Image Credit

CNRS UMR 8199. (2012) Predicted probability of childhood obesity from traditional risk factors in the NFBC1986. Retrieved from http://files-good.ibl.fr/childhood-obesity/.

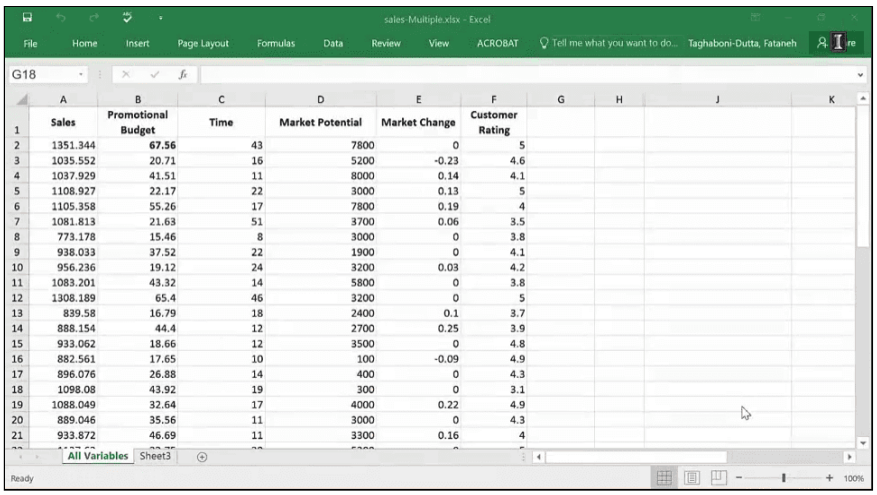
Transcript

No instruction provided on this slide

Lesson 4-1.2 Developing Multiple Regression Model in Excel

Media Player for Video

Sales Multiple Excel Worksheet (1 of 7) - Slide 22



Sales	Promotional Budget	Time	Market Potential	Market Change	Customer Rating
1351.344	67.56	43	7800	0	5
1035.552	20.71	16	5200	-0.23	4.6
1037.929	41.51	11	8000	0.14	4.1
1108.927	22.17	22	3000	0.13	5
1105.358	55.26	17	7800	0.19	4
1081.813	21.63	51	3700	0.06	3.5
773.178	15.46	8	3000	0	3.8
938.033	37.52	22	1900	0	4.1
956.236	19.12	24	3200	0.03	4.2
1083.201	43.32	14	5800	0	3.8
1308.189	65.4	46	3200	0	5
839.58	16.79	18	2400	0.1	3.7
888.154	44.4	12	2700	0.25	3.9
933.062	18.66	12	3500	0	4.8
882.561	17.65	10	100	-0.09	4.9
896.076	26.88	14	400	0	4.3
1098.08	43.92	19	300	0	3.1
1088.049	32.64	17	4000	0.22	4.9
889.046	35.56	11	3000	0	4.3
933.872	46.69	11	3300	0.16	4

This slide shows an Excel spreadsheet with the Mulitple Sales data set. There are six columns: Sales, Promotional Budget, Time, Market Potential, Market Change, and Customer Rating.

Download the Multiple Sales Excel file (Refer to Data - First worksheet)

Transcript

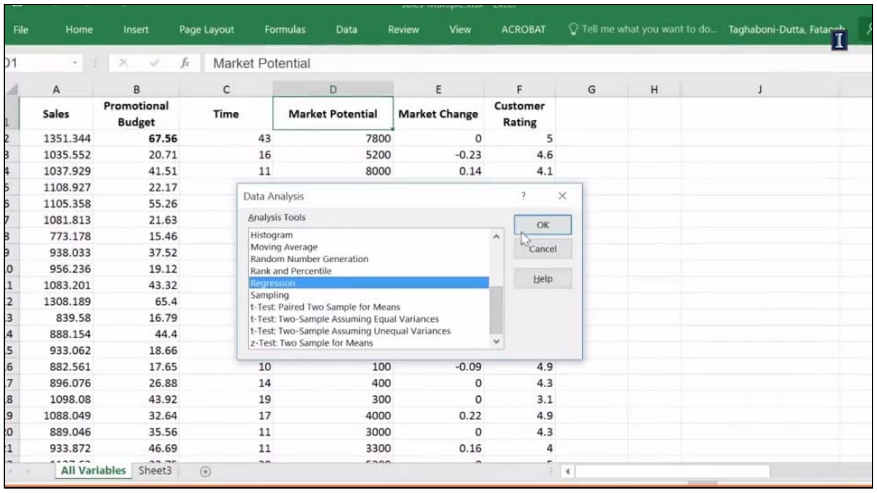
In this video, I'm going to illustrate how to run linear regression when we have multiple x-values. So, this is an example where we think that the sales representative can be predicted by how much promotional budget they have, how long they have been with the company in terms of number of months, the market potential where they are working in, the market change since the last period and how they are rated by their customers, their average rating by their customers.

So, these values are in thousands of dollars, so the sales are in thousands of dollars, promotional budget is in thousands of dollars and market potential is in thousands of dollars. And time is in number of months. And customer rating can be any value between zero and five.

So, to do this, first, we want to go to data analysis. But before I do the data analysis, I want to tell you something. In Excel, it's required that you would have all your x's in a continuous manner, so you cannot have, for example, promotional budget in your first column sales, which is your y-value in the second column and then start having more x's. It doesn't allow you, you have to be able to pick all the

x's in a continuous manner, so make sure before you run the analysis, you rearrange your worksheets such that the x's are all in consecutive columns.

Sales Multiple Excel Worksheet (2 of 7) - Slide 23



This slide shows the Data Analysis pop-up window in Excel with the Regression option highlighted.

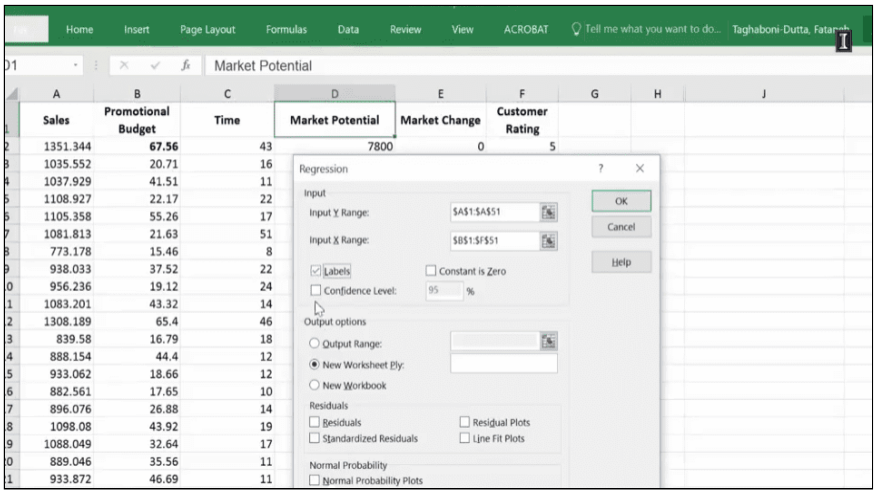
Transcript

So, once you have that, you go to data analysis. It really is a very simple difference between the two that we have done, between simple linear regression and what we will do in multiple regression. We select regression, and once again, it will ask you what is the y-variable.

So, you're going to click on the first cell, pick the entire column and then when it comes to x-variables, you have to simply select all of them so here is columns b through f and then you do the same thing, control shift down and select the rest of it. So, it basically says that it has selected B1 to F51, so this includes everything that we have identified as an independent variable, the promotional budget time, market potential, market change, and customer rating.

Again, I recommend that you would have headings for these because otherwise they will just show up as variable one, variable two, variable three, and that doesn't really make much of a sense.

Sales Multiple Excel Worksheet (3 of 7) - Slide 24



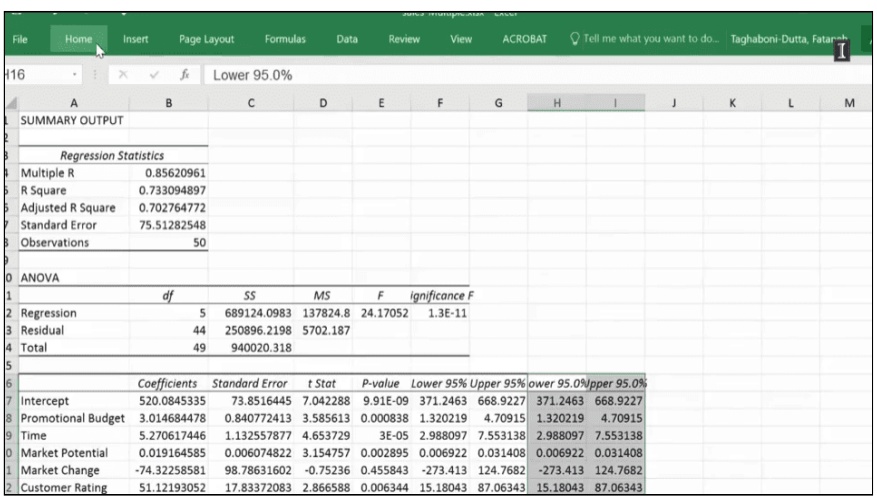
Sales	Promotional Budget	Time	Market Potential	Market Change	Customer Rating
1351.344	67.56	43	7800	0	5
1035.552	20.71	16			
1037.929	41.51	11			
1108.927	22.17	22			
1105.358	55.26	17			
1081.813	21.63	51			
773.178	15.46	8			
938.033	37.52	22			
956.236	19.12	24			
1083.201	43.32	14			
1308.189	65.4	46			
839.58	16.79	18			
888.154	44.4	12			
933.062	18.66	12			
882.561	17.65	10			
896.076	26.88	14			
1098.08	43.92	19			
1088.049	32.64	17			
889.046	35.56	11			
933.872	46.69	11			

This slide shows the Regression pop-up window in Excel. The Input y range is $A$1:$A$51, the Input x range is $B$1:$F$51. The Labels checkbox is checked, the Confidence level is set to 95%, and New Worksheet is selected under Output.

Transcript

So, I'm going to click on labels and I'm going to put in new worksheet, and here's our output.

Sales Multiple Excel Worksheet (4 of 7) - Slide 25



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.85620961
R Square	0.733094897
Adjusted R Square	0.702764772
Standard Error	75.51282548
Observations	50

ANOVA

	df	SS	MS	F	ignificance F
Regression	5	689124.0983	137824.8	24.17052	1.3E-11
Residual	44	250896.2198	5702.187		
Total	49	940020.318			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Upper 95.0%
Intercept	520.0845335	73.8516445	7.042288	9.91E-09	371.2463	668.9227	371.2463	668.9227
Promotional Budget	3.014684478	0.840772413	3.585613	0.000838	1.320219	4.70915	1.320219	4.70915
Time	5.270617446	1.132557877	4.653729	3E-05	2.988097	7.553138	2.988097	7.553138
Market Potential	0.019164585	0.006074822	3.154757	0.002895	0.006922	0.031408	0.006922	0.031408
Market Change	-74.32258581	98.78631602	-0.75236	0.455843	-273.413	124.7682	-273.413	124.7682
Customer Rating	51.12193052	17.83372083	2.866588	0.006344	15.18043	87.06343	15.18043	87.06343

This slide shows the Summary Output generated by Excel.

Regression Statistics - Slide 25

Multiple R	0.85620961
R Square	0.733094897
Adjusted R Square	**0.702764772**
Standard error	75.51282548
Observations	50

ANOVA - Slide 25

	df	SS	MS	F	Significance F
Regression	5	689124.0983	137824.8	24.17052	1.3E-11
Residual	44	250896.2198	5702.187		
Total	49	940020.318			

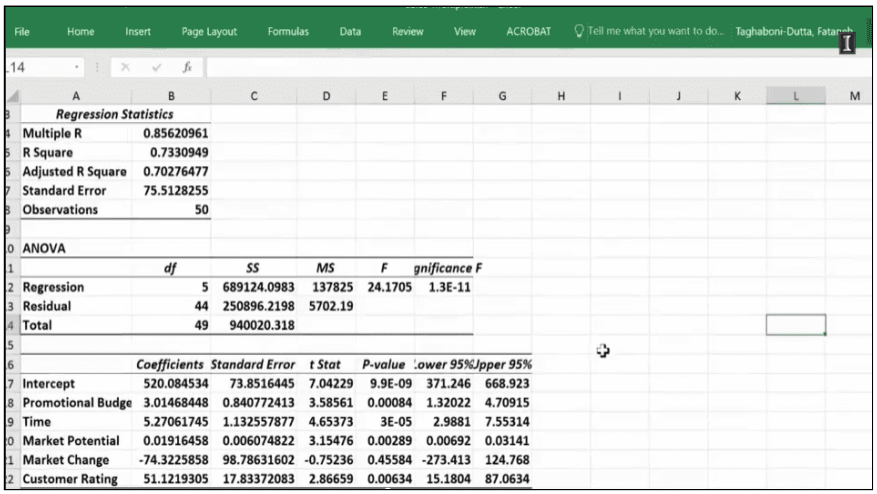
Other Statistics - Slide 25

	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	520.0845335	73.8516445	7.042288	9.91E-09	371.2463	668.9227
Promotional Budget	3.014684478	0.840772413	3.585613	0.000838	1.320219	4.70915
Time	5.270617446	1.132557877	4.653729	3E-05	2.988097	7.553138
Market potential	0.019164585	0.006074822	3.154757	0.002895	0.006922	0.031408
Market Share change	−74.32258581	98.78631602	−0.75236	0.455843	−273.413	124.76815
Customer rating	51.12193052	17.83372083	2.866588	0.006344	15.18043	87.06343

Transcript

And again, as you can see, the numbers are kind of squished in so I'm going to double click on every column to expand it, so I just put my cursor here, double click, double click and I think now everything is visible. And as I have mentioned before, upper and lower repeats itself twice so you can go ahead and ignore it or you can be like me and always remove it. Ok, so now what do we see?

Sales Multiple Excel Worksheet (5 of 7) - Slide 26



Regression Statistics	
Multiple R	0.85620961
R Square	0.7330949
Adjusted R Square	0.70276477
Standard Error	75.5128255
Observations	50

ANOVA

	df	SS	MS	F	gnificance F
Regression	5	689124.0983	137825	24.1705	1.3E-11
Residual	44	250896.2198	5702.19		
Total	49	940020.318			

	Coefficients	Standard Error	t Stat	P-value	ower 95%	Upper 95%
Intercept	520.084534	73.8516445	7.04229	9.9E-09	371.246	668.923
Promotional Budge	3.01468448	0.840772413	3.58561	0.00084	1.32022	4.70915
Time	5.27061745	1.132557877	4.65373	3E-05	2.9881	7.55314
Market Potential	0.01916458	0.006074822	3.15476	0.00289	0.00692	0.03141
Market Change	-74.3225858	98.78631602	-0.75236	0.45584	-273.413	124.768
Customer Rating	51.1219305	17.83372083	2.86659	0.00634	15.1804	87.0634

This slide contains the same Summary Output shown on [Slide 25 - Sales Multiple Excel Worksheet (4 of 7)](). The Adjusted R Square in the first table is highlighted. The P-value column in the third table is also highlighted.

Transcript

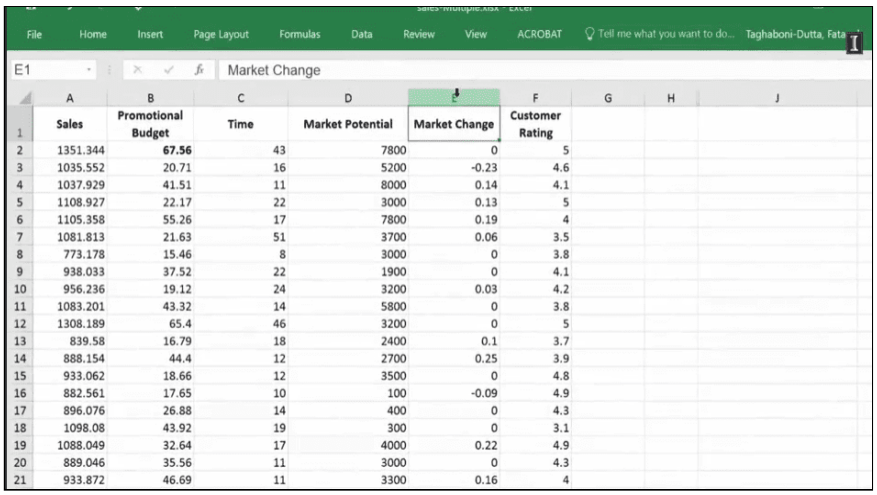
First of all, how is this as a model. As a model, because it's a multiple regression, and in multiple regression, naturally the R-squared will go up as you increase the number of independent variables. So, in multiple regression, we will focus on adjusted R-squared. So, our adjusted R-squared says that about 70% of variations that you see are explained by variables that you have identified. One of the things that you would see is that we no longer can just talk about correlation here because it's now multiple of things that are being included, so correlation is a bad regression model itself, rather than the correlation of each individual x-variable.

So, the first thing that we want to do is that we want to look at the p-values of every coefficient that we have here. So, focusing on the p-values, any p-value that shows a value greater than 0.05 because that's our level of significance, means that it does not have any strong relationship with the y-value that we are trying to predict. And here is the market change. So, we do not reject the hypothesis that these two things, market change and sales, do not have a relationship, we do not reject it. So, what do you have to do first before you would use the model is to remove such variables. These variables are basically noise, they're not adding anything, and it could be because market potential itself is capturing what we are looking for, a market change is a redundant variable, in which case you would say that there is a collinearity between them, which means market change

somewhat can be explained by the market potential itself, so it's a redundant variable or it's a variable that's neither useful in predicting our sales.

So, the first thing you need to do is clean up our model, and we're going to remove it. Now, if I had more than one, I would start removing the one with the highest value, then move onto the next one and move onto the next one. You will remove these one by one until you have only variables that have p values that are significant, and in this case, that means p-values are less than 0.05.

Sales Multiple Excel Worksheet (6 of 7) - Slide 27



Sales	Promotional Budget	Time	Market Potential	Market Change	Customer Rating
1351.344	67.56	43	7800	0	5
1035.552	20.71	16	5200	-0.23	4.6
1037.929	41.51	11	8000	0.14	4.1
1108.927	22.17	22	3000	0.13	5
1105.358	55.26	17	7800	0.19	4
1081.813	21.63	51	3700	0.06	3.5
773.178	15.46	8	3000	0	3.8
938.033	37.52	22	1900	0	4.1
956.236	19.12	24	3200	0.03	4.2
1083.201	43.32	14	5800	0	3.8
1308.189	65.4	46	3200	0	5
839.58	16.79	18	2400	0.1	3.7
888.154	44.4	12	2700	0.25	3.9
933.062	18.66	12	3500	0	4.8
882.561	17.65	10	100	-0.09	4.9
896.076	26.88	14	400	0	4.3
1098.08	43.92	19	300	0	3.1
1088.049	32.64	17	4000	0.22	4.9
889.046	35.56	11	3000	0	4.3
933.872	46.69	11	3300	0.16	4

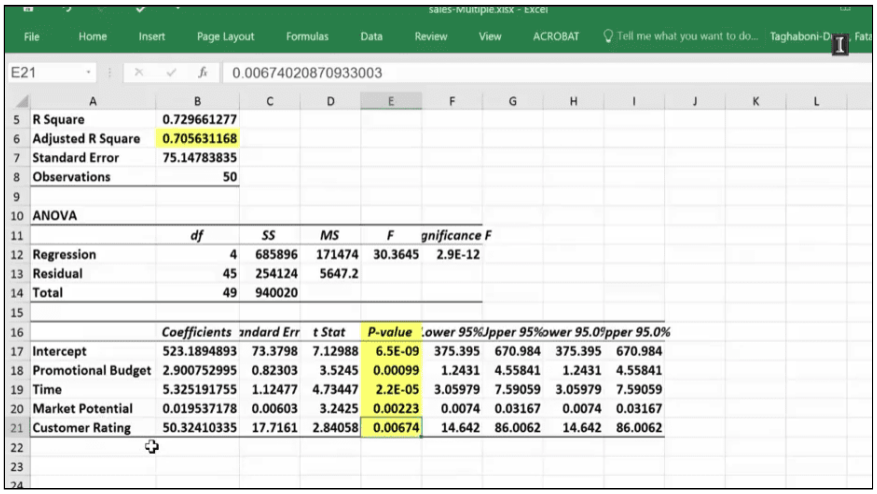
This slide shows the professor removing the Market Change column. Customer Rating is now located in Column E.

Transcript

So, what I'm going to do is that I'm going to go back to my variables and this time, I'm going to get rid of this column altogether. So, you can either just remove it or create another sheet that doesn't include it. So, in this case I'm just going to remove it and it's gone.

So, now I'm going to run my model with just these four variables. So, I have reduced my number of independent variables from five to four. Go back to data, data analysis, click regression and you will see because I'm in the same sheet, it has the same values, but remember now I have to change this. I no longer have columns b through f, but I have column e, so I'm just going to change it here and not go through the process over again, and it's going to put it in a new worksheet again.

Sales Multiple Excel Worksheet (7 of 7) - Slide 28



E21 0.00674020870933003

5	R Square	0.729661277							
6	Adjusted R Square	0.705631168							
7	Standard Error	75.14783835							
8	Observations	50							
9									
10	ANOVA								
11		df	SS	MS	F	gnificance F			
12	Regression	4	685896	171474	30.3645	2.9E-12			
13	Residual	45	254124	5647.2					
14	Total	49	940020						
15									
16		Coefficients	ndard Err	t Stat	P-value	ower 95%	Upper 95%	ower 95.0%	pper 95.0%
17	Intercept	523.1894893	73.3798	7.12988	6.5E-09	375.395	670.984	375.395	670.984
18	Promotional Budget	2.900752995	0.82303	3.5245	0.00099	1.2431	4.55841	1.2431	4.55841
19	Time	5.325191755	1.12477	4.73447	2.2E-05	3.05979	7.59059	3.05979	7.59059
20	Market Potential	0.019537178	0.00603	3.2425	0.00223	0.0074	0.03167	0.0074	0.03167
21	Customer Rating	50.32410335	17.7161	2.84058	0.00674	14.642	86.0062	14.642	86.0062

This slide contains a new Summary Output with the Adjusted R Square cell and the P-value column highlighted.

The new Adjusted R Square is 0.705631168.

The new P-values are:

- 6.5E-09
- 0.00099
- 2.2E-05
- 0.00223
- 0.00674

Download the Multiple Sales Excel file (Refer to Stat Output - Second worksheet)

Transcript

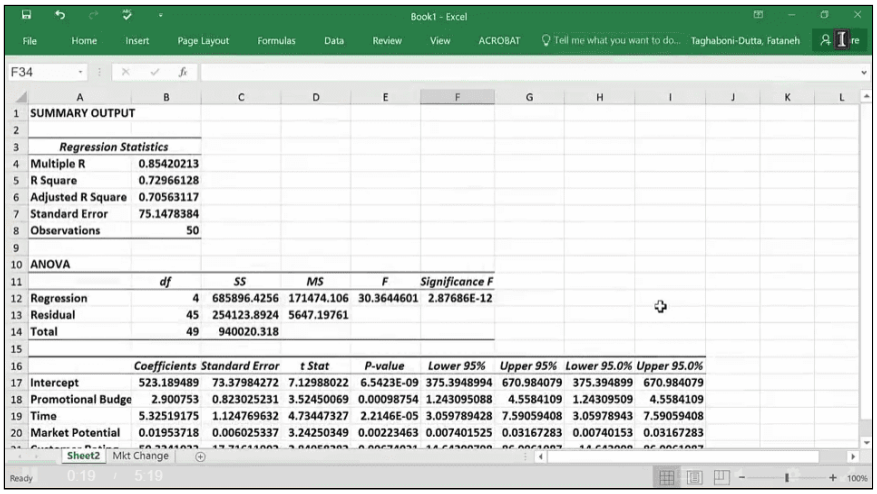
So, this is our new output. So, what do we see? So, adjusted R-squared is 7.05, the last one was 7.02, so actually, our model improved just a little bit. Sometimes you will see that it may not improve at all, but it will stay the same. It is still better. It is better to get noise out of your model. Now, let's look at our p-values.

The p-values are all significant, they are all less than 0.05. So, now we only have things in our model that is less than 0.05, the level of significance, therefore, we have a good model and this model now can use this for prediction.

Lesson 4-1.3 Multiple Regression: Excel Output Analysis

Media Player for Video

Excel Worksheet (1 of 3) - Slide 29



1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.85420213							
5	R Square	0.72966128							
6	Adjusted R Square	0.70563117							
7	Standard Error	75.1478384							
8	Observations	50							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	685896.4256	171474.106	30.3644601	2.87686E-12			
13	Residual	45	254123.8924	5647.19761					
14	Total	49	940020.318						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	523.189489	73.37984272	7.12988022	6.5423E-09	375.3948994	670.984079	375.394899	670.984079
18	Promotional Budge	2.900753	0.823025231	3.52450069	0.00098754	1.243095088	4.5584109	1.24309509	4.5584109
19	Time	5.32519175	1.124769632	4.73447327	2.2146E-05	3.059789428	7.59059408	3.05978943	7.59059408
20	Market Potential	0.01953718	0.006025337	3.24250349	0.00223463	0.007401525	0.03167283	0.00740153	0.03167283

This slide contains the Summary Output generated by Excel.

Regression Statistics - Slide 29

Multiple R	0.85420213
R Square	0.72966128
Adjusted R Square	0.70563117
Standard error	75.1478384
Observations	50

ANOVA - Slide 29

	df	SS	MS	F	Significance F
Regression	4	685896.4256	171474.106	30.3644601	2.87686E-12
Residual	45	254123.8924	5647.19761		
Total	49	940020.318			

Other Statistics - Slide 29

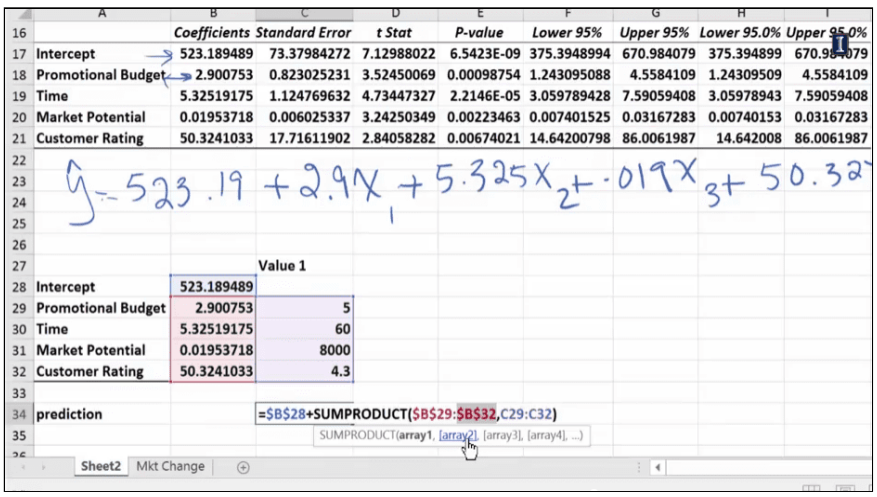
	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	523.189489	73.37984272	7.12988022	6.5423E-09	375.3948994	670.984079
Promotional Budget	2.900753	0.823025231	3.52450069	0.00098754	1.243095088	4.5584109
Time	5.32519175	1.124769632	4.73447327	2.2146E-05	3.0589789428	7.59059408
Market potential	0.01953718	0.006025337	3.24250349	0.00223463	0.007401525	0.03167283

[Download the Multiple Sales Excel file (Refer to Stat Output - Second worksheet)](#)

Transcript

So, now that we have a regression model that we feel like it's a good model for prediction, I'm going to show you how we're going to use the output for the purpose of prediction. And for that matter, I'm going to focus on the last table of this output.

Excel Worksheet (2 of 3) - Slide 30



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	523.189489	73.37984272	7.12988022	6.5423E-09	375.3948994	670.984079	375.394899	670.984079
Promotional Budget	2.900753	0.823025231	3.52450069	0.00098754	1.243095088	4.5584109	1.24309509	4.5584109
Time	5.32519175	1.124769632	4.73447327	2.2146E-05	3.059789428	7.59059408	3.05978943	7.59059408
Market Potential	0.01953718	0.006025337	3.24250349	0.00223463	0.007401525	0.03167283	0.00740153	0.03167283
Customer Rating	50.3241033	17.71611902	2.84058282	0.00674021	14.64200798	86.0061987	14.642008	86.0061987

$\hat{y} = 523.19 + 2.9x_1 + 5.325x_2 + .019x_3 + 50.32$

		Value 1
Intercept	523.189489	
Promotional Budget	2.900753	5
Time	5.32519175	60
Market Potential	0.01953718	8000
Customer Rating	50.3241033	4.3
prediction		=$B$28+SUMPRODUCT($B$29:$B$32,C29:C32)

SUMPRODUCT(array1, [array2], [array3], [array4], ...)

Sheet2 | Mkt Change

$$\hat{y} = 523.19 + 2.9x_1 + 5.325x_2 + 0.019x_3 + 50.32x_4.$$

Transcript

So, focusing on the last part of this output, these are the parts that are important to my regression equation. So, if I were to write this regression equation, it would look like this. ŷ is the intercept and the intercept comes from right here and I'm going to just round it a little, 1.9 plus 2.9 times x1, promotional budget is x1, and it has a coefficient of 2.9 and I'm going to go down to the time, because the coefficient of 5.325 times x2 plus market potential has 0.019 and its x3 and customer rating is 50.32 x4. The coefficient will tell me which one has the highest impact if I could change it.

So, for example, if I change x1 by one value, the total sales would go up by 2.9, and of course, sales are in thousands, so 2.9 would represent 2,900. So, if I look at all the coefficients, I would see that this has the highest potential for change [Underlines the customer rating coefficient 50.32], which means that if I could get my employees to have a better customer service, if they go from a three-point rating to a four-point rating, I'm going to see a difference of $50,000 almost. The next thing is the number of years that they have been [Underlines the Time coefficient 5.325]. So, these are the coefficients telling me where do I get my most bang for my buck.

So, if you are going to manipulate one of these variables, you should look for variables that have the highest impact. So, now let me show you how you would actually do the prediction. For that, I like to

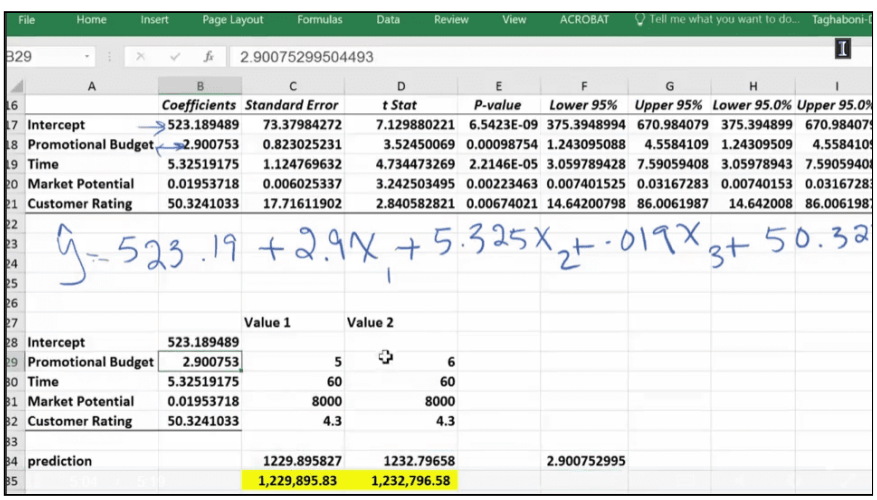
copy my values here down, so I just copy them and then I will paste them right here. So, these are the coefficients and I can enter now different values.

So, value one, for example, is going to be, so, there is nothing that I would enter for intercept, but promotional budget, let's say I'm going to consider someone who has a $5,000 promotional budget and because my values were in thousands, I'm going to enter this as five, this person has been with the company for five years, so that would be 60 months because time with the company has been recorded in number of months. The market potential that they are in has an $8 million market potential and again it is done in thousands, so I'm going to enter this as 8,000, so 8,000 is going to represent 8 million, and this person has a customer rating of 4.3.

So, if I want to know now what is the prediction, I'm going to record it right here. The prediction is going to be intercept this value plus this value multiplied by this value and then this with this and this with this. So, the best thing I can do is, first of all, make a sum product. Sum product is a function in Excel which will do exactly what we want to do. It will take array 1 and one by one it will multiple it by its array 2 and sum them up as it goes along. So, basically, it will take 2.9 multiplied by 5 plus 5.32 times 60 and so on.

Now, if I want to repeat this formula over and over again as I change values, one thing we want to do is that we want to lock the location of our coefficient. So, I'm going to put my cursor back on B28 and then click F4 and it will lock it, and same thing with B29, I'm going to click on B29 and then say F4 and B32 F4, and in a minute, you will see now that I have locked it, I can just enter new values and I will get new sum products. So, this will tell me that the prediction is 1.229 because this was in thousands, so maybe I want to just take this back and multiple it by a thousand, even show it in dollars so it will be easier to understand.

Excel Worksheet (3 of 3) - Slide 31



	A	B	C	D	E	F	G	H	I
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	523.189489	73.37984272	7.129880221	6.5423E-09	375.3948994	670.984079	375.394899	670.98407
18	Promotional Budget	2.900753	0.823025231	3.52450069	0.00098754	1.243095088	4.5584109	1.24309509	4.558410
19	Time	5.32519175	1.124769632	4.734473269	2.2146E-05	3.059789428	7.59059408	3.05978943	7.5905940
20	Market Potential	0.01953718	0.006025337	3.242503495	0.00223463	0.007401525	0.03167283	0.00740153	0.0316728
21	Customer Rating	50.3241033	17.71611902	2.840582821	0.00674021	14.64200798	86.0061987	14.642008	86.006198
22									
23	$y = 523.19 + 2.9X_1 + 5.325X_2 + .019X_3 + 50.32$								
24									
25									
26									
27			Value 1	Value 2					
28	Intercept	523.189489							
29	Promotional Budget	2.900753	5	6					
30	Time	5.32519175	60	60					
31	Market Potential	0.01953718	8000	8000					
32	Customer Rating	50.3241033	4.3	4.3					
33									
34	prediction		1229.895827	1232.79658		2.900752995			
35			1,229,895.83	1,232,796.58					

This slide contains the same data as [Slide 30 - Excel Worksheet (2 of 3)](#) with the predictions highlighted.

[Download the Multiple Sales Excel file (Refer to Prediction - Third worksheet)](#)

Transcript

I'm going to go to home and, so this would be my prediction. And this is, again, a point estimate of the prediction interval. I'm not giving you the interval, although there is a 95% confidence interval that we have for each one of them, so we can create the prediction interval, but as I said, Excel doesn't automate that and it's a little cumbersome so I'm just going to ignore it for now.

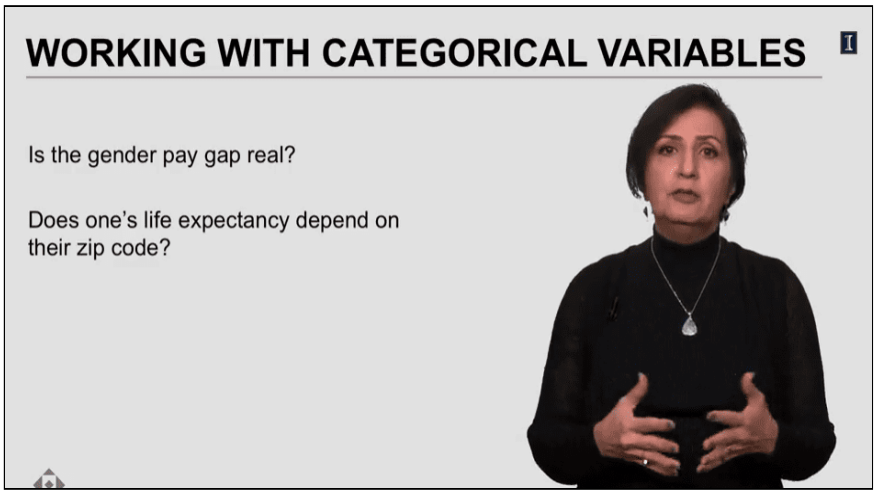
So, now what happens if I said, okay, I'm going to give this same person a little bit more promotional budget. So, I'm going to increase their promotional budget from 5,000 to 6,000. So, what will happen? Well, you should be able to tell that my sales should go up by the coefficient. The rest of the things are the same. So, it's the same person with the same number of months, so I'm going to just copy these, and now because I have locked my locations of coefficients, it can just take this column and multiply it by values that I see in this column, so, I'm just going to drag this out and you would see that it has gone up. And how much has gone up, it should be by that coefficient. So, I'm going to just show it to you that that is the case. If I take the difference between this increase of 1 with this one, you would get 2.9, exactly like you see here, as you would expect it. So, if I enter 60, it will go up by 10 times that. So, basically, now we can take this model and make point estimates of our prediction for values that we give.

Lesson 4-2 Qualitative Data – Two Categories

Lesson 4-2.1 Qualitative Data – Two Categories

Media Player for Video

Working with Categorical Variables - Slide 32



Is the gender pay gap real?

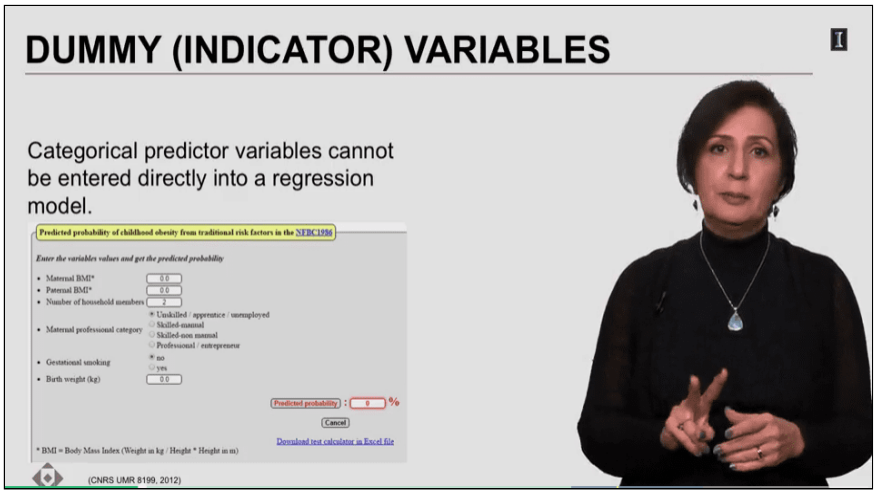
Does one's life expectancy depend on their zip code?

Transcript

So far, we have only looked at including quantitative data in our regression models. However, we may wish to include descriptive or qualitative data as well. For example, we may want to include the gender of the respondent in our data if we think that gender of the subject has an effect on the person's pay or does one's life expectancy depend on their zip code? Both of these are examples of studies that have been published, both gender and zip code are examples of categorical data.

Recall that even though zip codes is a number, it is not a numerical value, it doesn't represent a quantity, it represents a category. In this lesson, we will learn how incorporate categorical variables in regression analysis.

Dummy (Indicator) Variables (1 of 2) - Slide 33



Categorical predictor variables cannot be entered directly into a regression model.

This slide contains an image of the Obesity Calculator form introduced in Slide 3 - Predicting Obesity.

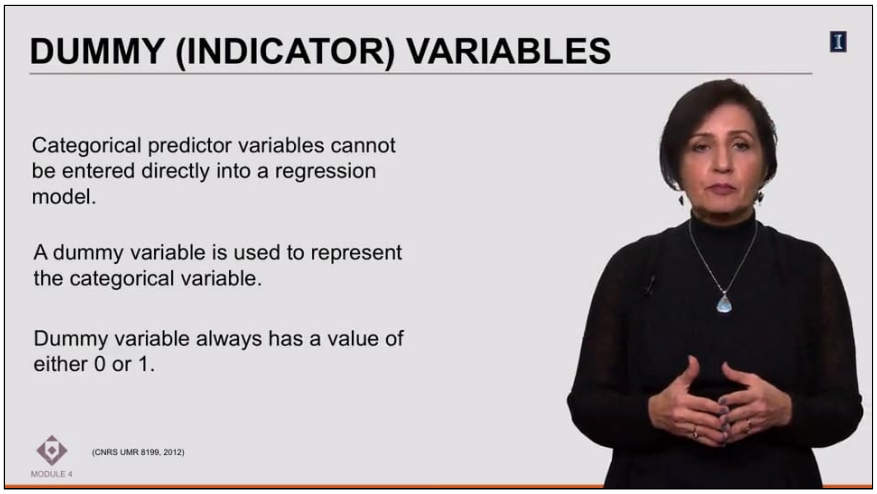
(CNRS UMR 8199, 2012)

Transcript

Categorical predictor variables cannot be entered directly into regression model. We have to take an additional step before being able to run the analysis. Let's take a look at the obesity calculator again. I shared this with you in the last lesson.

In this study, they had several categorical variables such as gestational smoking which could take on a value of a yes or no; or the maternal professional categories which could be one of the four given choices. These variables first need to be coded such as their meaning is represented by a number.

Dummy (Indicator) Variables (2 of 2) - Slide 34



Categorical predictor variables cannot be entered directly into a regression model.

A dummy variable is used to represent the categorical variable.

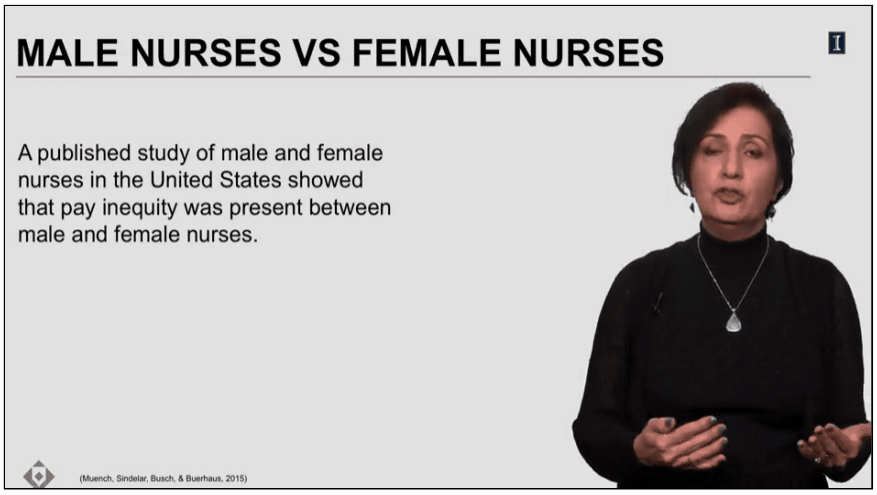
Dummy variable always has a value of either 0 or 1

Transcript

To do this, we create a dummy variable. A dummy variable always has a value of either zero or one. For example, if the mother smoked during the pregnancy then the dummy variable will have a value of one and zero. Otherwise, to represent yes or no.

And of course, when a variable has more than two possibilities as it is in the case of maternal professional category, the use of dummy variables will require a bit more work. For now, let's start with an example where there are two possibilities for the indicator or dummy variable. And later, we will learn how to use dummy variables when there are more than two possibilities.

Male Nurses vs. Female Nurses - Slide 35



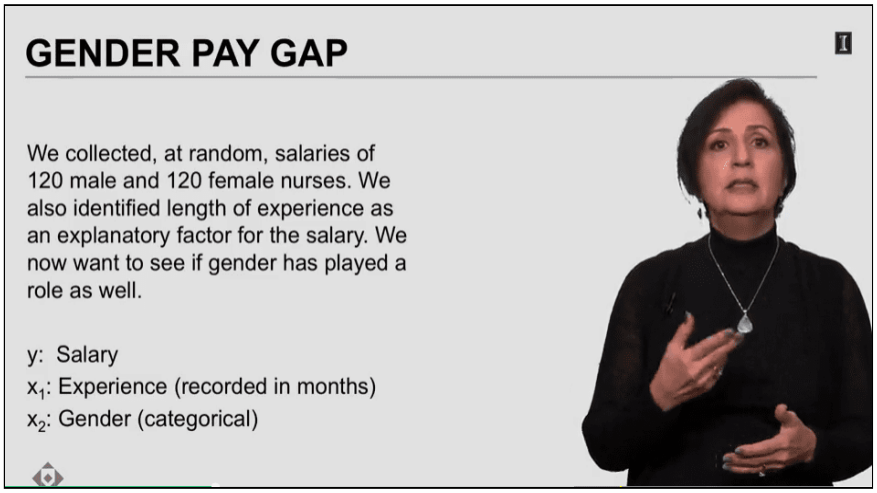
A published study of male and female nurses in the United States showed that pay inequity was present between male and female nurses.

(Muench, Sindelar, Busch, & Buerhaus, 2015)

Transcript

A published study of male and female nurses in the United States showed that pay inequity was present in male and female nurses. What this study found was that the gender was an explanatory variable for predicting a nurse's salary. Let me use similar data that became used to show how you could check on the correlation between gender and pay.

Gender Pay Gap - Slide 36



We collected, at random, salaries of 120 male and 120 female nurses. We also identified length of experience as an explanatory factor for the salary. We now want to see if gender has played a role as well.

y: Salary

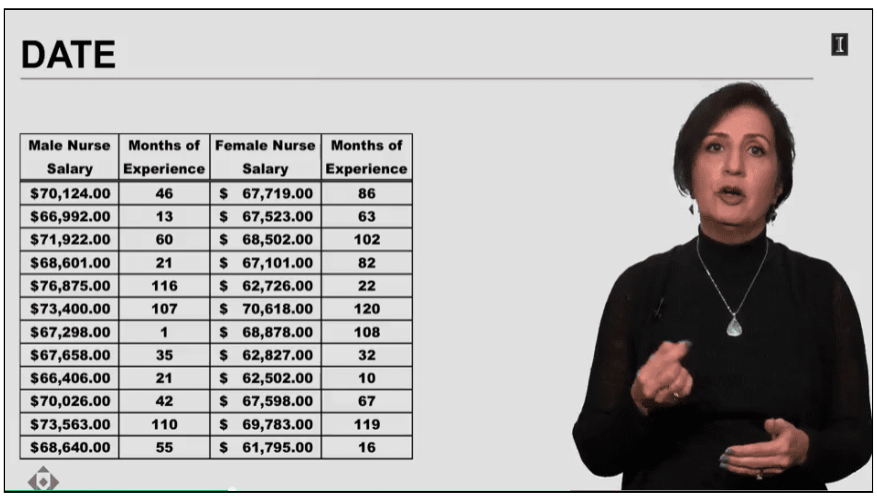
x_1 : Experience (recorded in months)

x_2 : Gender (categorical)

Transcript

We collected, at random, salaries of 120 male and 120 female nurses. We also have identified length of experience as an explanatory factor for the salary. We now want to see if gender has played a role as well. So, here variables are: for y, it's salary that's being predicted, x1 is experience (recorded in months), x2 represents gender which is categorical and has two possibilities, male or female.

Date - Slide 37



Male Nurse Salary	Months of Experience	Female Nurse Salary	Months of Experience
$70,124.00	46	$ 67,719.00	86
$66,992.00	13	$ 67,523.00	63
$71,922.00	60	$ 68,502.00	102
$68,601.00	21	$ 67,101.00	82
$76,875.00	116	$ 62,726.00	22
$73,400.00	107	$ 70,618.00	120
$67,298.00	1	$ 68,878.00	108
$67,658.00	35	$ 62,827.00	32
$66,406.00	21	$ 62,502.00	10
$70,026.00	42	$ 67,598.00	67
$73,563.00	110	$ 69,783.00	119
$68,640.00	55	$ 61,795.00	16

This slide contains a table showing a data set containing the annual salary and months of experience for a set of nurses, sorted by gender

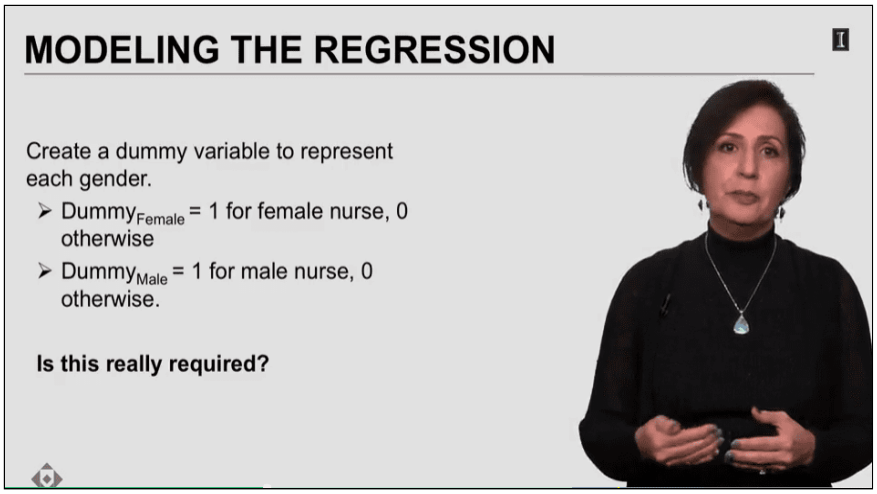
Nurse salary and months of experience

Male Nurse Salary	Months of Experience (Male)	Female Nurse Salary	Months of Experience (Female)
$70,124.00	46	$67,719.00	86
$66,992.00	13	$67,523.00	63
$71,922.00	60	$68,502.00	102
$68,601.00	21	$67,101.00	82
$76,875.00	116	$62,726.00	22
$73,400.00	107	$70,618.00	120
$67,298.00	1	$68,878.00	108
$67,658.00	35	$62,827.00	32
$66,406.00	21	$62,502.00	10
$70,026.00	42	$67,598.00	67
$73,563.00	110	$69,783.00	119
$68,640.00	55	$61,795.00	16

Transcript

This is just a small snapshot of the data we have. The data here shows that the months of the experience and salaries of male and the female nurses. Before we can run regression on this, we have to code each male or female nurse with a numerical value of 0 or 1.

Modeling the Regression - Slide 38



Create a dummy variable to represent each gender.

$$\text{Dummy}_{\text{Female}} = 1 \text{ for female nurse, } 0 \text{ otherwise}$$

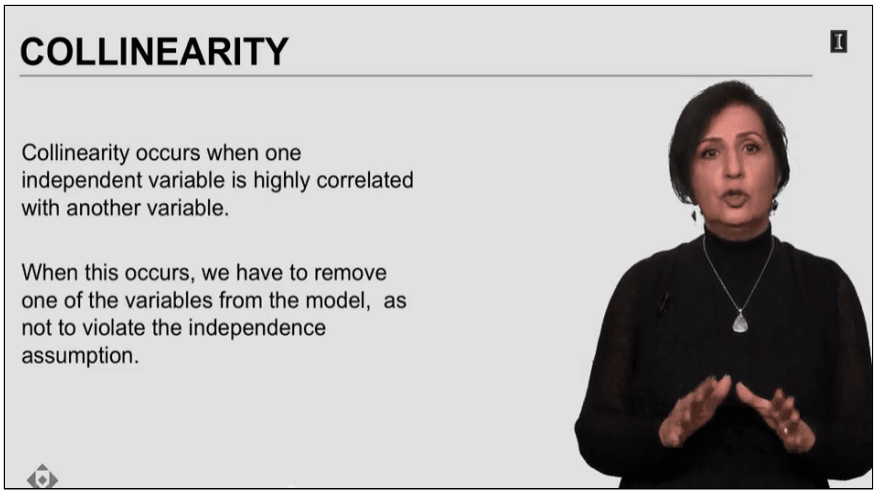
$$\text{Dummy}_{\text{Male}} = 1 \text{ for male nurse, } 0 \text{ otherwise}$$

Is this really required?

Transcript

We can create the dummy variable to represent each gender. For example, I can have a dummy female which I can set to 1, for a female nurse 0 otherwise and a dummy male which is equal 1 for male nurses and 0 otherwise. But is this really how we should do this? The answer is no. And let me explain why by explaining what is known as a collinearity.

Collinearity - Slide 39



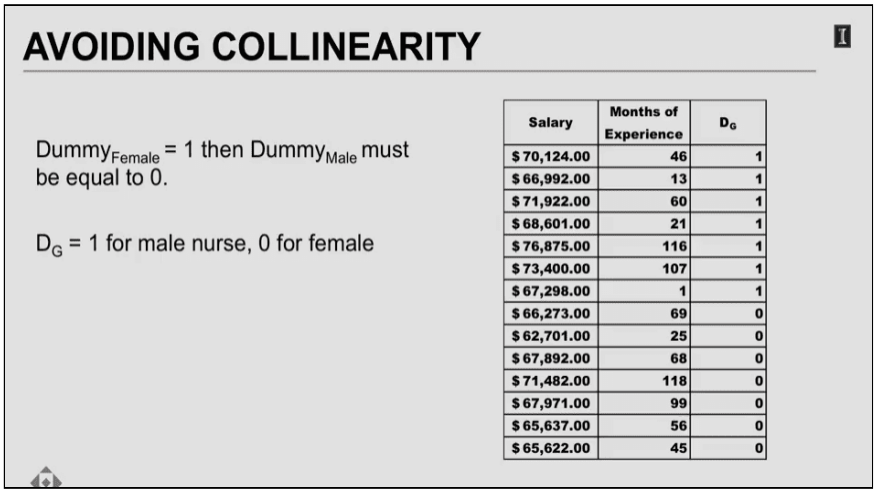
Collinearity occurs when one independent variable is highly correlated with another variable.

When this occurs, we have to remove one of the variables from the model, as not to violate the independence assumption.

Transcript

In multiple regressions, collinearity occurs when one independent variable is highly correlated with another variable. When this occurs, we have to remove one of the variables from the model, as not to violate the independence assumption. The way we created a dummy variable in the last slide for a nurse's gender will cause collinearity and this is why.

Avoiding Collinearity - Slide 40



Salary	Months of Experience	D_G
$70,124.00	46	1
$66,992.00	13	1
$71,922.00	60	1
$68,601.00	21	1
$76,875.00	116	1
$73,400.00	107	1
$67,298.00	1	1
$66,273.00	69	0
$62,701.00	25	0
$67,892.00	68	0
$71,482.00	118	0
$67,971.00	99	0
$65,637.00	56	0
$65,622.00	45	0

This slide contains a table showing annual salary, months of expereince, and the dummy value for each nurse

$\text{Dummy}_{\text{Female}} = 1$ then $\text{Dummy}_{\text{Male}}$ must be equal to 0

$D_G = 1$ for male nurse, 0 for female

Nurse salary and month of experience

Salary	**Months of Experience**	D_G
$70,124.00	46	1
$66,992.00	13	1
$71,922.00	60	1
$68,601.00	21	1
$76,875.00	116	1
$73,400.00	107	1
$67,298.00	1	1
$66,273.00	69	0
$62,701.00	25	0
$67,892.00	68	0
$71,482.00	118	0
$67,971.00	99	0
$65,637.00	56	0
$65,622.00	45	0

Transcript

If dummy for female is 1, then it means that you are looking at the salary that belongs to a female nurse, then for this subject, the value dummy male must be 0. So, if we code the nurse's gender this way, the two dummy variables would be completely correlated and violates the independent assumption. To do this correctly we would create just one dummy variable to represent the nurse's gender.

I'm going to call this D sub g: D for dummy and g for gender. I am now going to set a value to 1 if the nurse is a male nurse and 0 if the nurse is a female.

The table you see now is a snapshot of what the coded data would look like. Now that all the values in the data set are numeric, we can run the regression in analysis.

Gender Gap: Excel Output - Slide 41

GENDER GAP: EXCEL OUTPUT

Regression Statistics	
Multiple R	0.969796537
R Square	0.940505323
Adjusted R Square	0.940003258
Standard Error	850.7891425
Observations	240

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	61149.40381	138.0981426	442.796715	0	60877.34716	61421.46
Months of Experier	82.30084538	1.668613537	49.3228921	1.2737E-126	79.01363665	85.588054
Gender	4667.701219	110.7148371	42.159672	4.041E-112	4449.590331	4885.8121

Regression Statistics - Slide 41

Multiple R	0.969796537
R Square	0.940505323
Adjusted R Square	**0.940003258**
Standard error	850.7891425
Observations	240

Other Statistics - Slide 41

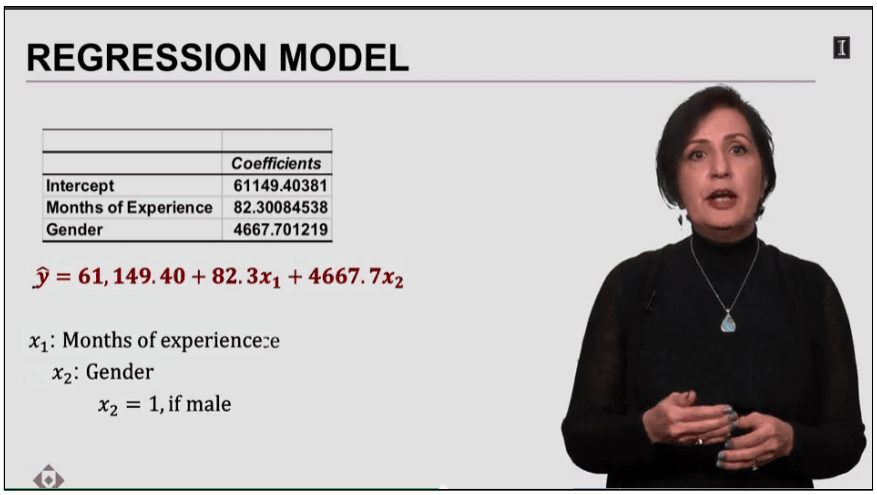
	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	61149.40381	138.0981426	443.796715	0	60877.34716	61421.46
Months of experience	83.30084538	1.668613537	49.3228921	1.2737E-126	79.01363665	85.588054
Gender	4667.701219	110.7148371	42.159672	4.041E-112	4449.590331	4885.8121

The P-value column in the "Other Statistics - Slide 41" table is highlighted.

Transcript

First, we get a very good model. The explanatory variables included in the model explain 94% of variation seen in the salary data, but are both gender and month of experience significant? For that, we would look at the bottom table of the output and focus on the p-values of these variables. Both variables have very small piece which implies that they both have a significant relationship with salary. So, now let's see what all of these means.

Regression Model - Slide 42



This slide contains the Coefficients column of the Other Statistics table presented on [Slide 41 - Gender Gap: Excel Output](#).

$$\hat{y} = 61{,}149.40 + 82.3x_1 + 4667.7x_2$$

$$x_1 : \text{Months of experience}$$

$$x_2 : \text{Gender}$$

$$x_2 = 1, \text{ if male}$$

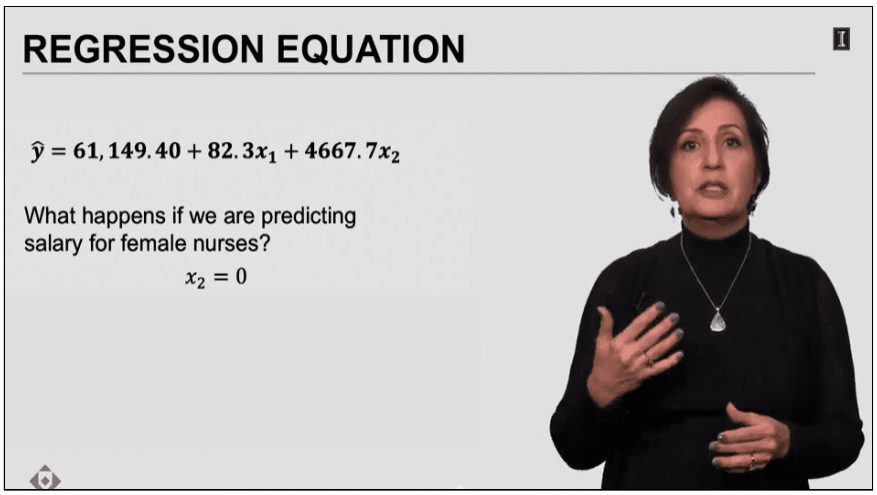
Transcript

In a perfect world, gender should have no effect on the salary, but here we saw that it had, such is life. Based on this analysis what can be said about the pay gap then?

The coefficient of independent variables along with the intercept make up the regression equation expressed as you see here. Recall that x1 represents months of experience, this can be any number depending on how long the individuals have been a nurse.

The other variable in the equation is x2 gender and this is an interesting one. This is the one that has only two possibilities, this variable is equal to 1 if the nurse is a male and 0 if the nurse is a female.

Regression Equation (1 of 2) - Slide 43



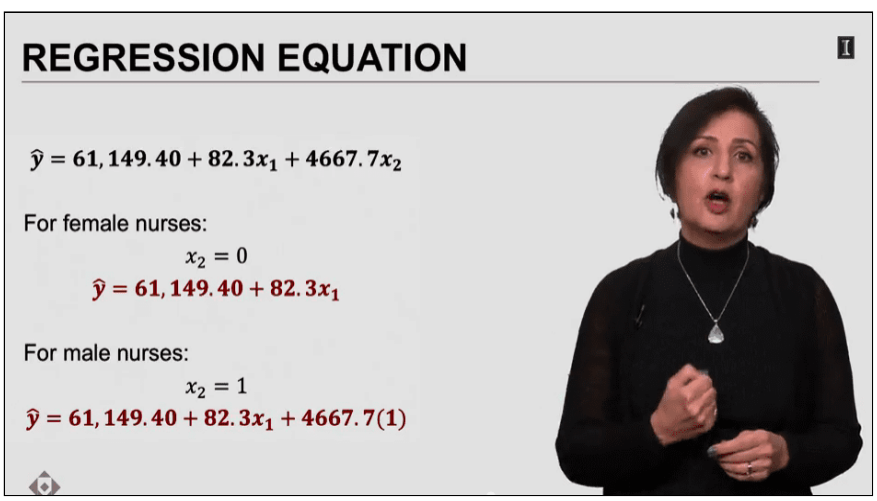
$\hat{y} = 61{,}149.40 + 82.3x_1 + 4667.7x_2$

What happens if we are predicting salary for female nurses? $x_2 = 0$

Transcript

In order to understand the coefficient for gender variable ask yourself what would happen if we wanted to predict salary for a female nurse. Since the gender variable is 0 for female nurses then that part of the equation gets eliminated and this will be the equation for female nurses.

Regression Equation (2 of 2) - Slide 44



$\hat{y} = 61{,}149.40 + 82.3x_1 + 4667.7x_2$

For female nurses: $x_2 = 0$

$\hat{y} = 61{,}149.40 + 82.3x_1$

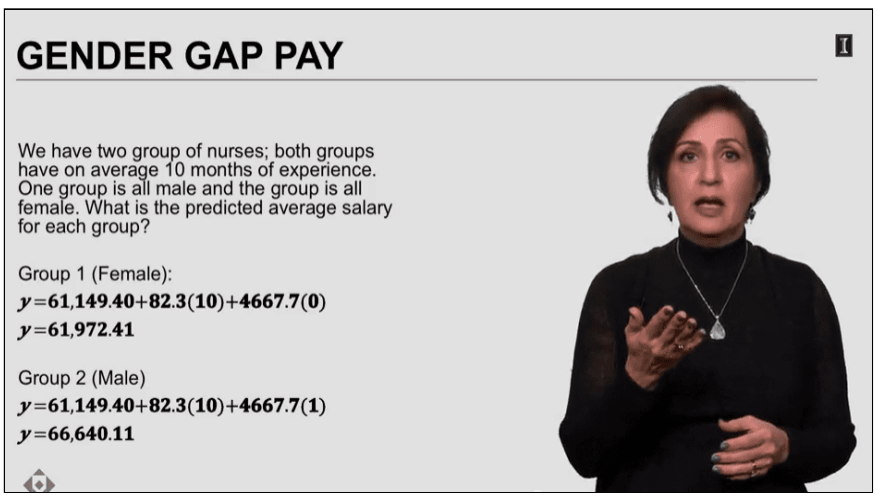
For male nurses: $x_2 = 1$

$\hat{y} = 61{,}149.40 + 82.3x_1 + 4667.7(1)$

Transcript

And for male nurses, x2 is 1, and the last coefficient is multiplied by 1. You see what just happened? By multiplying the last term by 1 we just gave all the male nurses a constant salary bonus of $4,667.70.

Gender Gap Pay (1 of 2) - Slide 45



We have two group of nurses; both groups have on average 10 months of experience. One group is all male and the group is all female. What is the predicted average salary for each group?

Group 1 (Female): **y = 61,149.40 + 82.3(10) + 4667.7(0) = 61,972.41**

Group 2 (Male): **y = 61,149.40 + 82.3(10) + 4667.7(1) = 66,640.11**

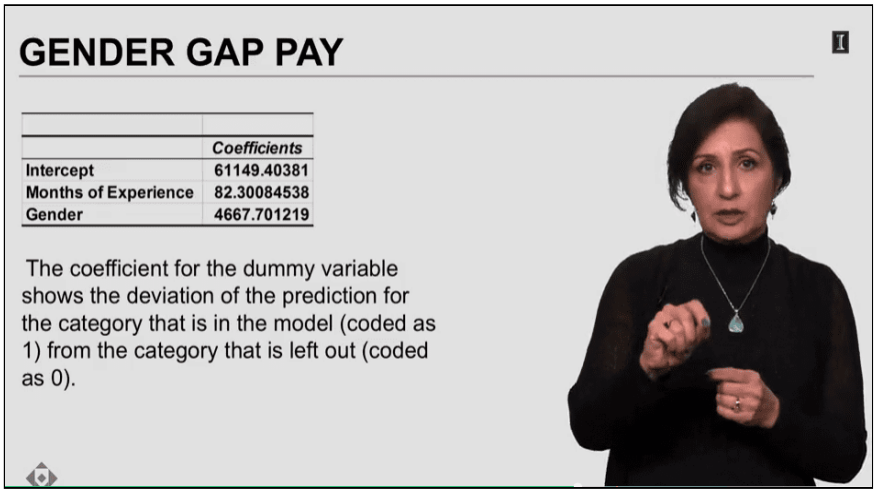
Transcript

Now, let's use the equation. We have two groups of nurses; both groups have on average ten months of experience. One group is all male and one group is all female, what is estimate average salary for each group?

Placing ten in x1 and zero in x2. Predicted salary for female nurses will be about $61,972 and placing 10 in x1 and 1 in x2 for male nurses, the expected salary for male nurses is about $66,640 which,

again, is the $4,667.70 more than the female nurses.

Gender Gap Pay (2 of 2) - Slide 46



This slide contains the Coefficients column of the Other Statistics table presented on [Slide 41 - Gender Gap: Excel Output](#).

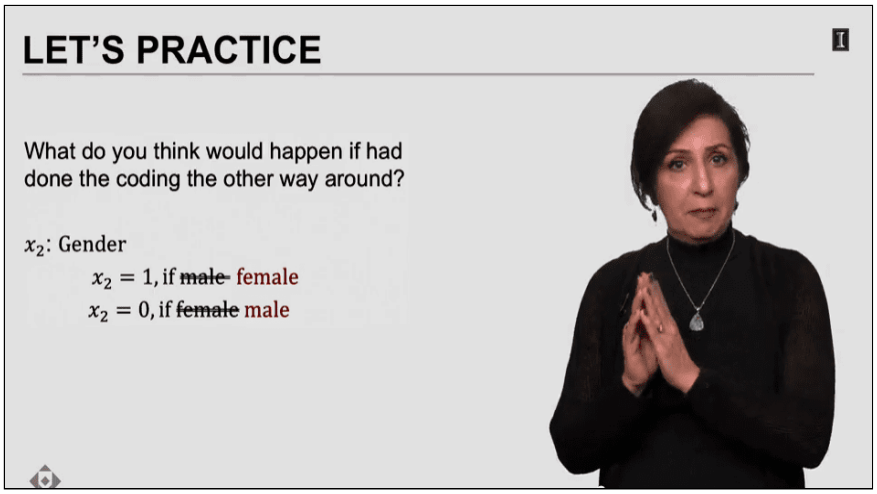
The coefficient for the dummy variable shows the deviation of the prediction for the category that is in the model (coded as 1) from the category that is left out (coded as 0).

Transcript

So, the coefficient for gender variable shows the difference in the salary for male nurses from the female nurses. In this case that it's added gender bonus of $4,667 for males.

The takeaway here is for all models that have categorical data of the two possibilities, the coefficient of dummy variable shows the deviation of the prediction for the category that is in the model that is coded as 1 from the category that is left out that it's coded as 0.

Let's Practice - Slide 47



What do you think would happen if had done the coding the other way around?

x_2 : Gender

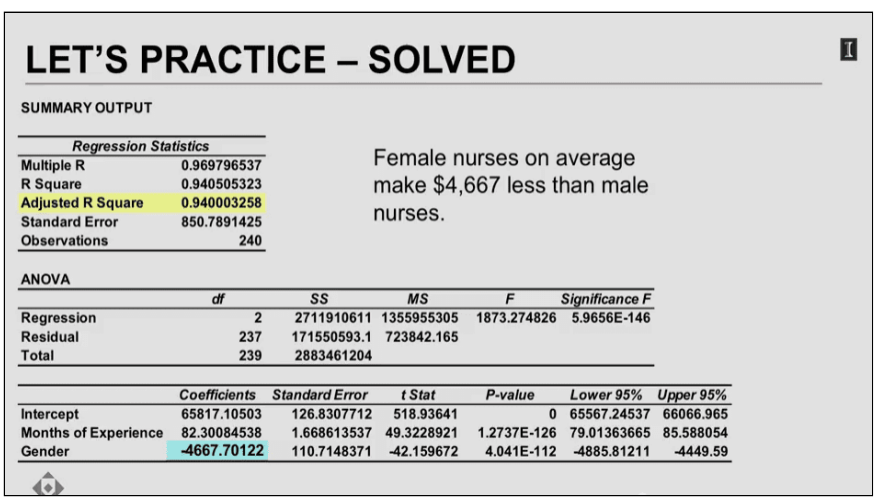
$x_2 = 1$: if ~~male~~ **female**

$x_2 = 0$: if ~~female~~ **male**

Transcript

So, let's practice. What do you think would happen if I had done the coding the other way around? For x2 that was for gender, I would have said x2 is equal to 1 if it's a female nurse and x2 is equals to 0 if it's a male nurse.

Let's Practice - Solved - Slide 48



Summary Output

Regression Statistics - Slide 48

Multiple R	0.969796537
R Square	0.940505323
Adjusted R Square	**0.940003258**
Standard error	850.7891425
Observations	240

ANOVA - Slide 48

	df	SS	MS	F	Significance F
Regression	2	2711910611	1355955305	1873.274826	5.9656E-146
Residual	237	171550593.1	723842.165		
Total	239	2883461204			

Other Statistics - Slide 48

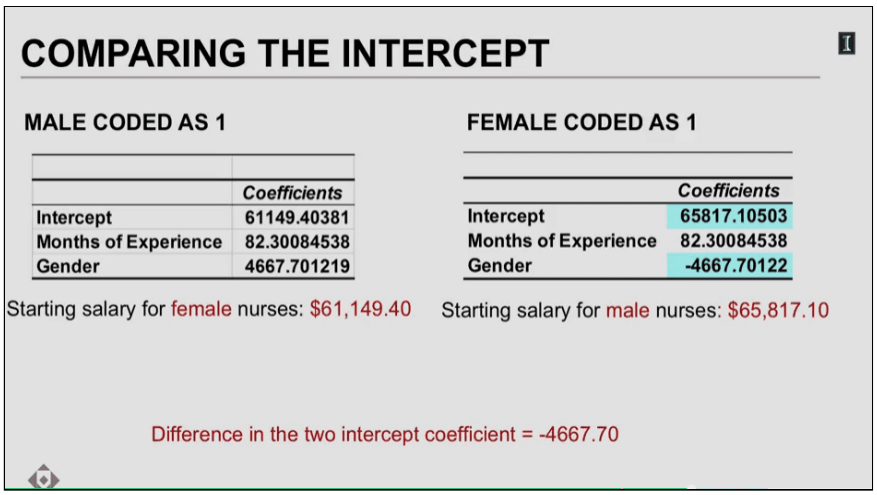
	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	65817.10503	126.8307712	518.93641	0	65567.24537	66066.965
Months of experience	82.30084538	1.668613537	49.3228921	1.2737E-126	79.01363665	85.588054
Gender	**−4667.70122**	110.7148371	−42.159672	4.041E-112	−4885.81211	−4449.59

Female nurses on average make $4,667 less than male nurses.

Transcript

Adjusted R-squared is the same as before, what changed are the two coefficients. First, it's the coefficient of gender, valued wise it is the same as before, but now it's negative. So, a female nurse gets a deduction of $4,667.60. But there is another coefficient that has changed as well and that is the coefficient of the intercept. I'm guessing you didn't see this coming. Let me explain what just happened.

Comparing the Intercept - Slide 49



Male coded as 1

	Coefficients
Intercept	61149.40381
Months of experience	82.30084538
Gender	4667.701219

Starting salary for **female** nurses: **$61,149.40**

Female coded as 1

	Coefficients
Intercept	**65817.10503**
Months of experience	82.30084538
Gender	**-4667.70122**

Starting salary for **male** nurses: **$65,817.10**

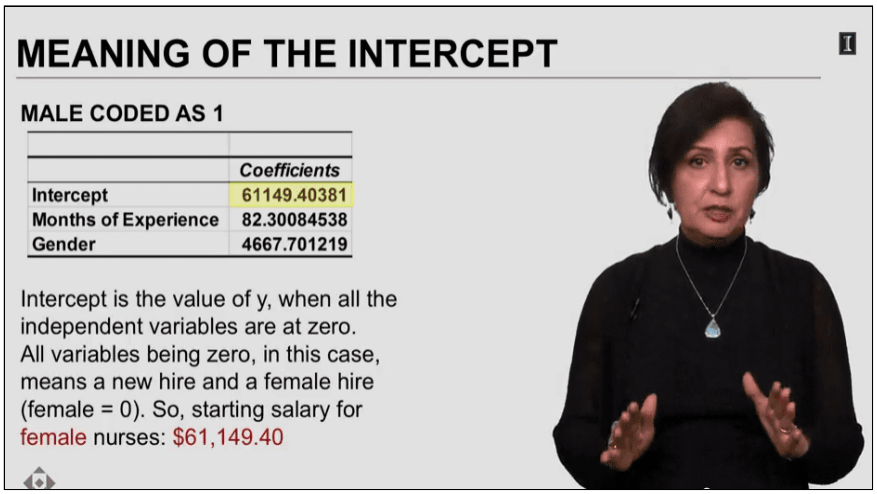
Difference in the two-intercept coefficient = −4667.70

Transcript

The intercepts are different when we look at the two different ways of coding the gender. The best way for me to show you what just happened is by just subtracting the two and you see that the magic number of 4,667.70 gender effect reappears.

This is the estimated starting salary of the female nurse, and this is the estimated salary of male nurses which is $4,667 more than the female nurses. Now, that you see the numbers, let me explain the principles of it.

Meaning of the Intercept - Slide 50



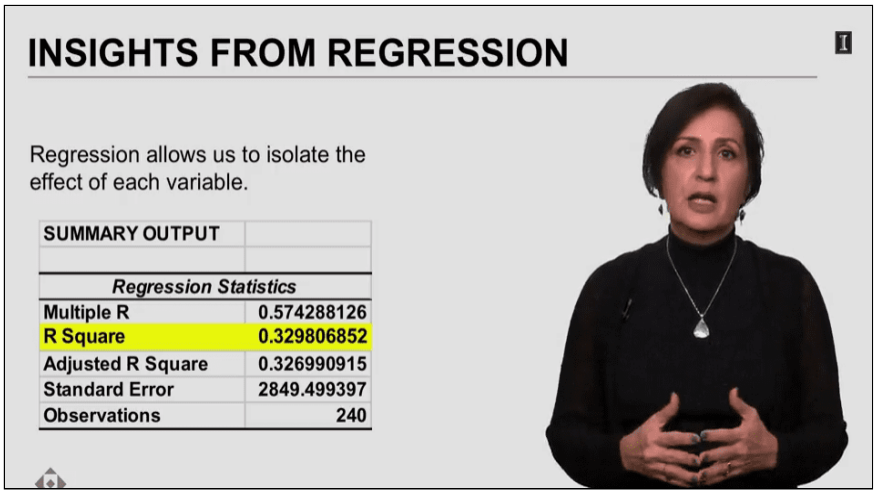
This slide contains the Male Coded as 1 table presented on Slide 49 - Comparing the Intercept with the Intercept value highlighted.

Intercept is the value of y, when all the independent variables are at zero. All variables being zero, in this case, means a new hire and a female hire (female = 0). So, starting salary for **female** nurses: **$61,149.40**

Transcript

The coefficient value of intercept is the value of y when all independent variables are set at 0. All variables being 0, in this case, would mean this nurse is a new hire and it's a female nurse, female is equal to female 0. Does the intercept in the model when we code the dummy variable as 1 for male nurses shows the starting salary for female? And that would be $61,149.40, and it would be the other way around when we code female nurses as 1.

Insights from Regression - Slide 51



Regression allows us to isolate the effect of each variable.

Summary Output

Regression Statistics - Slide 51

Multiple R	0.574288126
R Square	**0.329806852**
Adjusted R square	0.326990915
Standard error	2849.499397
Observations	240

Transcript

Could we have done the study by hypothesis testing within the two groups? The answer is yes, but regression gives us more information than just the ability to reject the null hypothesis that male and female nurses have equal pay. Regression allows us to isolate the effect of each variable.

So, in this example, we were able to, in one shot, isolate the two sources of variation for differences in pay, experience and gender. And we can say that collectively they explain 94% of the variations. If we want to know how much of the difference is explained by just gender then we can run a simple linear regression with just the gender dummy variable.

And I did just that, I ran the analysis with only gender. So, I get this summary output. This was ran as a simple lineal regression with just one independent variable. So, looking at R-squared I see that about 33% of the salary variations are due to gender. Regression not only is a great predictive model, but it's also a great a way of isolating the variables that influence all variable interest, in this case, salary. By knowing this, we can then set out the change that independent variables so that our outcomes are better for us. These are why the exercise, eat nuts, get rest, etcetera, etcetera because someone has done the studies that show exactly what and how these variables affect our health. It is no different in business and we can hardness this power for this very powerful statistical tool to improve business outcomes.

References - Slide 52

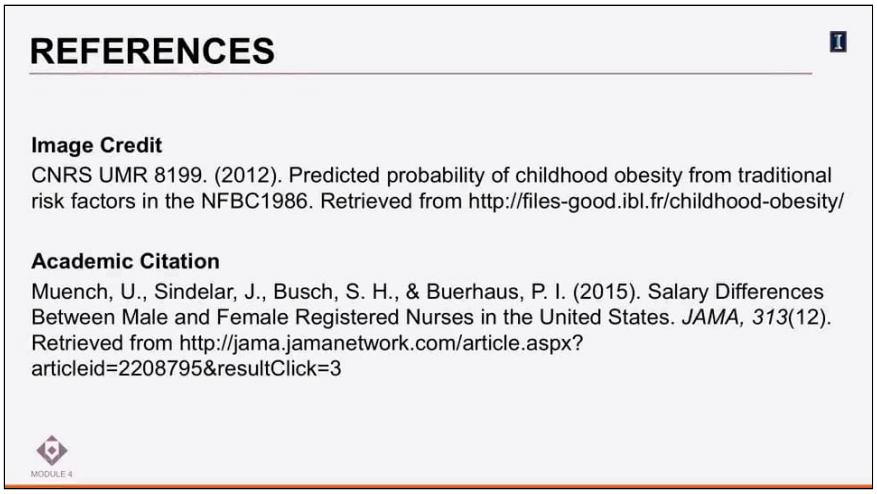


Image Credit

CNRS UMR 8199. (2012). Predicted probability of childhood obesity from traditional risk factors in the NFBC1986. Retrieved from http://files-good.ibl.fr/childhood-obesity/.

Academic Citation:

Muench, U., Sindelar, J., Busch,S.H., & Buerhaus, P.I. (2015). *Salary Differences Between Male and Female Registered Nurses in the United States.* JAMA, 313(12). Retrieved from http://jama.jamanetwork.com/article.aspx?articleid=2208795&resultClick=3.

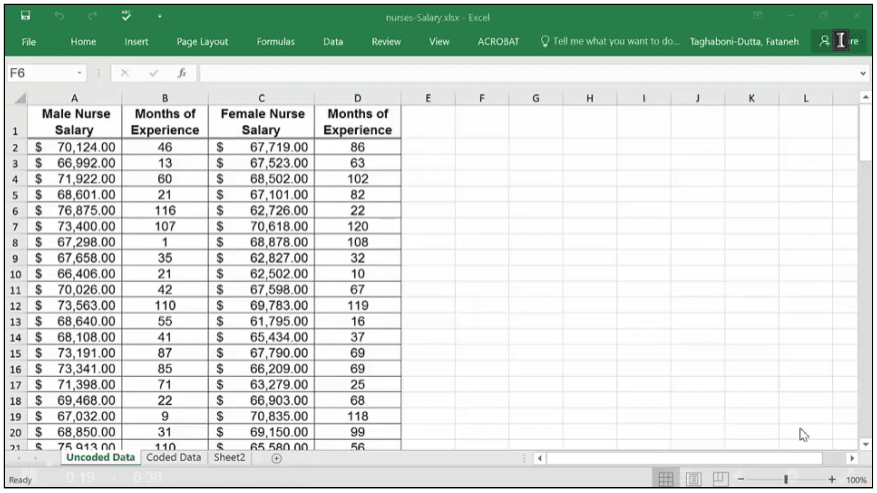
Transcript

No instruction provided on this slide

Lesson 4-2.2 Dummy Variable in Excel – Two Categories

Media Player for Video

Nurses Salary Excel Worksheet (1 of 6) - Slide 53



Male Nurse Salary	Months of Experience	Female Nurse Salary	Months of Experience
$ 70,124.00	46	$ 67,719.00	86
$ 66,992.00	13	$ 67,523.00	63
$ 71,922.00	60	$ 68,502.00	102
$ 68,601.00	21	$ 67,101.00	82
$ 76,875.00	116	$ 62,726.00	22
$ 73,400.00	107	$ 70,618.00	120
$ 67,298.00	1	$ 68,878.00	108
$ 67,658.00	35	$ 62,827.00	32
$ 66,406.00	21	$ 62,502.00	10
$ 70,026.00	42	$ 67,598.00	67
$ 73,563.00	110	$ 69,783.00	119
$ 68,640.00	55	$ 61,795.00	16
$ 68,108.00	41	$ 65,434.00	37
$ 73,191.00	87	$ 67,790.00	69
$ 73,341.00	85	$ 66,209.00	69
$ 71,398.00	71	$ 63,279.00	25
$ 69,468.00	22	$ 66,903.00	68
$ 67,032.00	9	$ 70,835.00	118
$ 68,850.00	31	$ 69,150.00	99
$ 75,913.00	110	$ 65,590.00	56

Excel spreadsheet containing the Nurses Salary data set with the following data in columns A through D: Male Nurse Salary, Months of Experience (male), Female Nurse Salary, Months of Experience (female.)

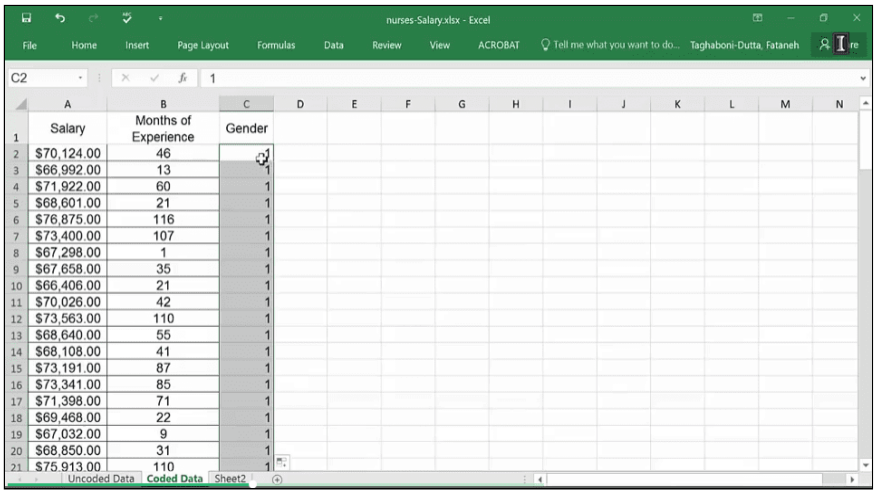
Download the Nurses Salary Excel file (Refer to Uncoded data - First worksheet)

Transcript

In this video, I'm showing you how to use dummy variables, to model independent variables that are categorical in nature. In this example, we are comparing nurses' salary for male nurses and female nurses. We are wondering if their gender has anything to do with their salaries. Now, gender is a categorical variable, which means it can either be female or male, it has two possibilities, but I can't enter that into regression analysis. Regression analysis requires numerical values, which means we have to code the gender of the nurse as a 0—1 to make it either a male nurse or a female nurse. You got to do exactly that.

So, this is my uncoded data. It shows the male nurses' salaries in this column, their months of experience, and then for females and then for their months of experience [Columns A, B, C, and D]. So, what I'm going to do is that I'm going to create another tab called coded data and here, I'm just going to say "salary", "months of experience", and their "gender" [Columns A, B, and C].

Nurses Salary Excel Worksheet (2 of 6) - Slide 54



Salary	Months of Experience	Gender
$70,124.00	46	1
$66,992.00	13	1
$71,922.00	60	1
$68,601.00	21	1
$76,875.00	116	1
$73,400.00	107	1
$67,298.00	1	1
$67,658.00	35	1
$66,406.00	21	1
$70,026.00	42	1
$73,563.00	110	1
$68,640.00	55	1
$68,108.00	41	1
$73,191.00	87	1
$73,341.00	85	1
$71,398.00	71	1
$69,468.00	22	1
$67,032.00	9	1
$68,850.00	31	1
$75,913.00	110	1

The slide shows another worksheet with the coded data of the Nurses Salary. The coded data includes three columns: Salary, Months of experience, and Gender. Here, the Gender column values are selected.

Download the Nurses Salary Excel file (Refer to Coded data - Second worksheet)

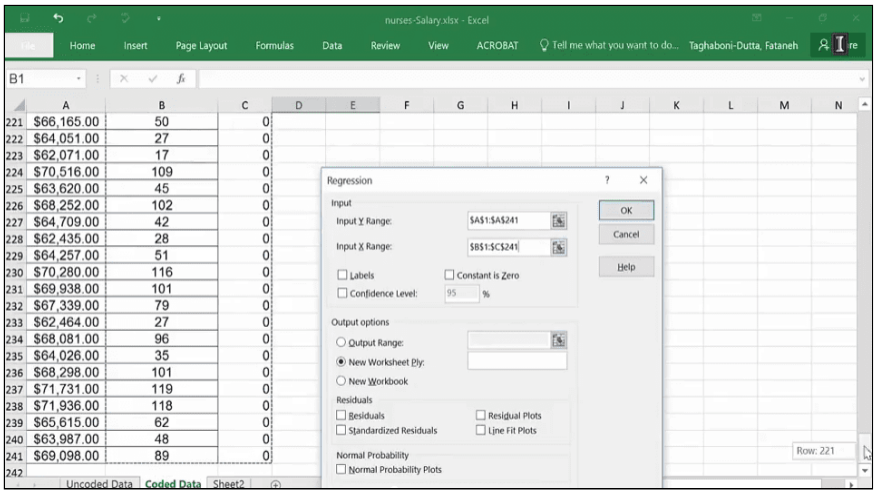
Transcript

And then, go back to my uncoded data, pick the male salaries along with their months of experience. Control-shift-down, pick all of it, copy this, so, control-C, go back to my coded data, place my cursor in the first sell, control-V, expand the column to include everything. So, then these are the salaries of our male nurses. So, if I just – control-shift-down – you will see that it goes down to 121, that represents the 120 male nurses that are in our data set. So, now what I want to do, is that I want to get the female nurses and bring them here. But before I do that, to make my life easier, I'm going to mention here that these are the male nurses, and I'm going to code them as 1, and, I'm just going to put my cursor back on here, get the cross here and double-click and all 120 will be now coded as one, which means they're male.

And I will now grab the female numbers and bring them and just simply paste them here. So, I'm going to go back to my uncoded data, scroll back up, pick the salaries and the months of experience, control-shift-down, I pick all of them, control-C, coded data, place my cursor here and just say control-V [Puts cursor in cell A 122]. So, the female nurses also have their values transferred here. Now, I'm going to put the value 0 here, because these are female nurses, and as soon as I enter the very first female as a 0, I can put my cross here and double click, and now everything is coded.

So, if you want to double check, you can just go up here, control-shift-down, and you will see that everything has a coding.

Nurses Salary Excel Worksheet (3 of 6) - Slide 55



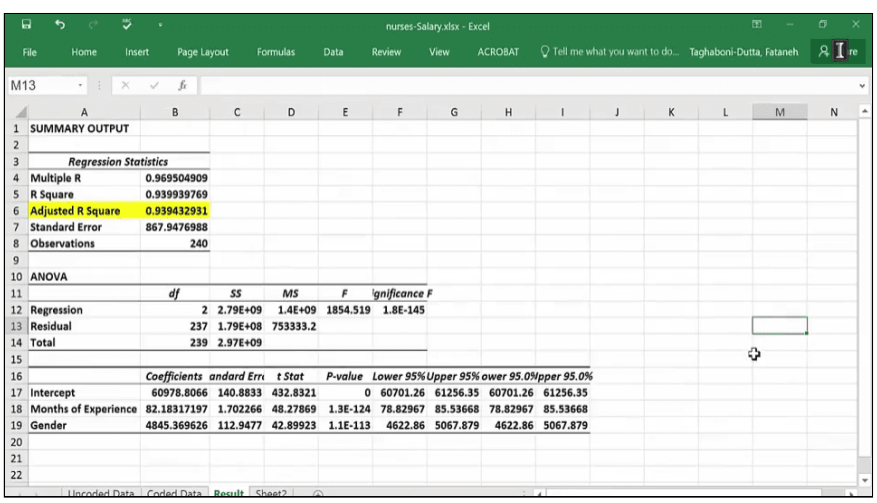
This slide shows the Regression pop-up window; the Input y range is $A$1:$A$241, the Input x range is $B$1:$C$241, and the Confidence level is set to 95%. In the Output section, the New worksheet ply option is selected.

Transcript

So, now you are ready to do your data analysis, because everything is numerical, whether they're categorical or not. So, go to data, data analysis, pick regression.

At this point, you're not doing anything different than any other time, we will put the y-values here, so, these are salaries, it picks all of it, then go up, pick both, months of experience and gender, I'm going to make sure that it pick my labels, pick all of it and say labels, and I'm going to put it in a new sheet.

Nurses Salary Excel Worksheet (4 of 6) - Slide 56



This slide contains the Summary Output generated by Excel.

Regression Statistics - Slide 56

Multiple R	0.969504909
R Square	0.939939769
Adjusted R Square	**0.939432931**
Standard error	867.9476988
Observations	240

ANOVA - Slide 56

	df	SS	MS	F	Significance F
Regression	2	2.79E+09	1.4E+09	1854.519	1.8E-145
Residual	237	1.79E+08	753333.2		
Total	239	2.97E+09			

Other Statistics - Slide 56

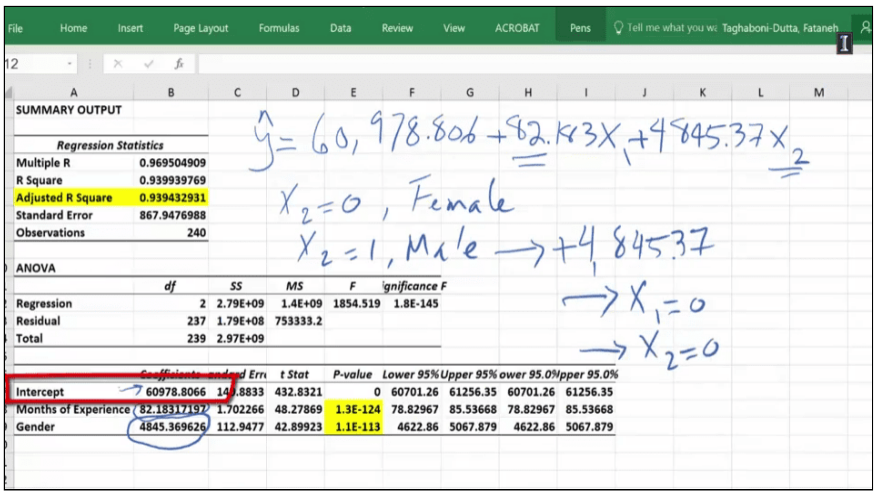
	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	60978.8066	140.8833	432.8321	0	60701.26	61256.35
Months of Experience	82.18317197	1.702266	48.27869	1.3E-124	78.82967	85.53668
Gender	4845.369626	112.9477	42.89923	1.1E-113	4622.86	5067.879

Download the Nurses Salary Excel file (Refer to Regression Output - Third worksheet)

Transcript

Okay, so first of all, how is our regression model? Our regression model is very good, almost 94% of the variations we see in the salary is due to the variables that we have identified. Are they both very good? Well, let's see if they're both significant by looking at their p-value, and indeed, they are extremely small, so, both gender and experience have turned out to be significant relationship with the salaries.

Nurses Salary Excel Worksheet (5 of 6) - Slide 57



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.969504909
R Square	0.939939769
Adjusted R Square	0.939432931
Standard Error	867.9476988
Observations	240

ANOVA

	df	SS	MS	F	ignificance F
Regression	2	2.79E+09	1.4E+09	1854.519	1.8E-145
Residual	237	1.79E+08	753333.2		
Total	239	2.97E+09			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	60978.8066	14 .8833	432.8321	0	60701.26	61256.35	60701.26	61256.35
Months of Experience	82.18317197	1.702266	48.27869	1.3E-124	78.82967	85.53668	78.82967	85.53668
Gender	4845.369626	112.9477	42.89923	1.1E-113	4622.86	5067.879	4622.86	5067.879

This slide contains the same table presented on Slide 56 - Nurses Salary Excel Worksheet (4 of 6) along with the following equations:

$$\hat{y} = 60{,}978.806 + 82.183x_1 + 4845.37x_2.$$

$$x_2 = 0, \text{Female}$$

$$x_2 = 1, \text{Male}$$

Intercept Coefficient 60978.8066 in the Other Statistics - Slide 56 table is circled; the Adjusted R Square of the Regression Statistics - Slide 56 table and the P-value column of the Other Statistics - Slide 56 table are highlighted.

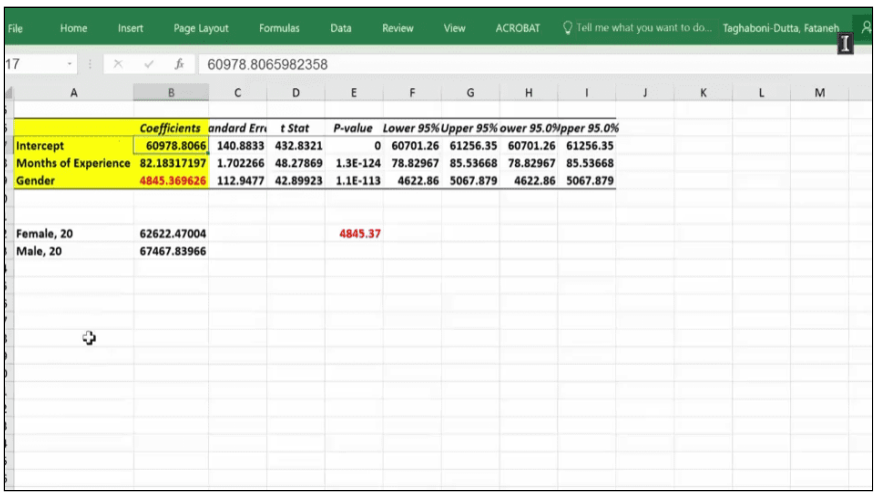
Transcript

So, what do we have here as a regression equation? Our regression equation is your salary is a coefficient of intercept is 60,978.806, so that's coming from here, then months of experience right here is 82.183 times x1 plus the last one, which is gender and it's right here, is 4,845.37 almost, x2.

x1 represents months of experience, so, for every month of experience the salary goes up by $82. This is what it's showing here. What is x2? x2 is gender, and for female x2 is 0, so, if you are looking at the female nurse, the last term is going to be eliminated from our equation altogether, but if x2 equal to 1, and it's a male, then the last term gets multiply by 1. Essentially what you're saying is that whatever the female nurses' salary is, males will get an additional $4,845 extra due to their gender, and, this is what you will get.

So, what does the intercept means? So, remember, intercept is when x1 is equal to 0 and x2 is equal to 0, so x1 is 0 when somebody is a brand-new nurse with no experience, and x2 is 0 when they are female nurse. So, basically, the value that you see here is the starting salary of the female nurse, so, this is the point of intercept. Intercept is being all the variables are set to equal to 0, and when the variable of months of experience is 0 is a fresh brand-new nurse and is a female nurse. So, now you can start predicting with this model because is a good enough model that you can do prediction.

Nurses Salary Excel Worksheet (6 of 6) - Slide 58



	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	60978.8066	140.8833	432.8321	0	60701.26	61256.35	60701.26	61256.35
Months of Experience	82.18317197	1.702266	48.27869	1.3E-124	78.82967	85.53668	78.82967	85.53668
Gender	4845.369626	112.9477	42.89923	1.1E-113	4622.86	5067.879	4622.86	5067.879
Female, 20	62622.47004			4845.37				
Male, 20	67467.83966							

This slide contains the third Other Statistics table presented on Slide 56 - Nurses Salary Excel Worksheet (4 of 6) with the Coefficients column highlighted.

Female with 20 months experience salary is 62622.47004. Male with 20 months experience salary is 67467.83966. Difference in salaries is 4845.37. Gender Coefficient is 4845.369626.

Download the Nurses Salary Excel file (Refer to Regression Output - Third worksheet)

Transcript

To use this for prediction, you need to focus on the last table and the coefficient, so, let's scroll down and pay attention to our coefficients.

So, now let's use this model to predict the salary for a female that has 20 months of experience and a male nurse that has 20 months of experience. For a female nurse, the number is going to be the coefficient plus the months of experience coefficient multiplied by 20 and I didn't do the last variable because it would be multiplied by 0, by the way.

For the male, I'm going to take the coefficient add to it, the coefficient for months of experience, multiplied by 20 plus the last coefficient and it's redundant, but I'm going to write multiplied by 1 and these are the differences between the two. And of course, if I took the differences of these two, I would get exactly what I was supposed to get, the difference gender makes in the salary, these two values are exactly the same, that's the gender difference.

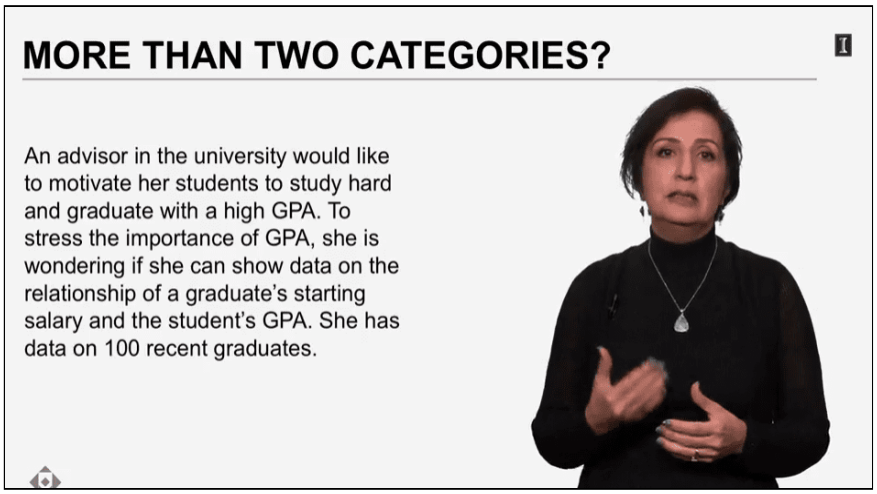
What would have happened if I had codded it the other way? And I have shown you that in my PowerPoint, the only thing that would change is that intercept will change, intercept will then represent the starting salaries for a male nurse, and then the gender coefficient will become a negative value. Try that on your own and see how it will change.

Lesson 4-3 Qualitative Data – Many Categories

Lesson 4-3.1 Qualitative Data – Many Categories

Media Player for Video

More than Two Categories? - Slide 59

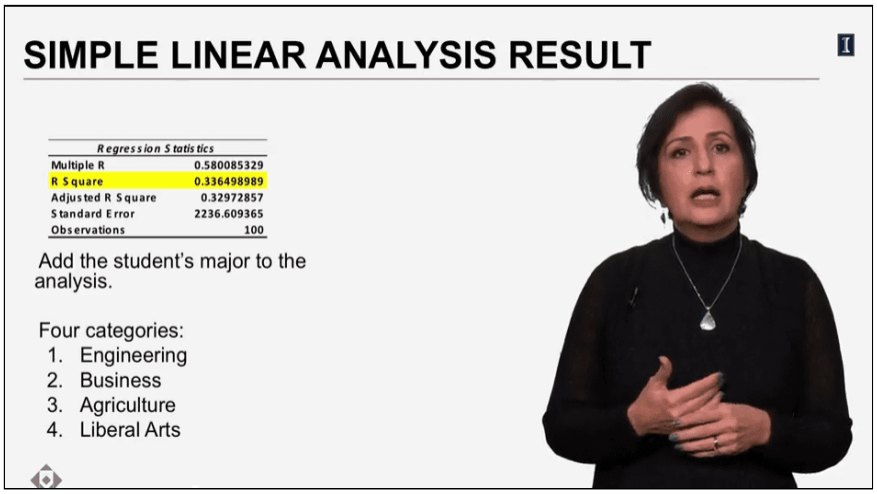


An advisor in the university would like to motivate her students to study hard and graduate with a high GPA. To stress the importance of GPA, she is wondering if she can show data on the relationship of a graduate's starting salary and the student's GPA. She has data on 100 recent graduates.

Transcript

In the last lesson, we learned how to code for categorical variables that had only two possibilities, like gender. What do we do when there are more than two choices. In this lesson, I will take you through what is needed to run the regression model when we have categorical variables with more than two possibilities.

Simple Linear Analysis Result - Slide 60



Regression Statistics - Slide 60

Multiple R	0.580085329
R Square	**0.336498989**
Adjusted R Square	0.32972857
Standard error	2236.609365
Observations	100

Add the student's major to the analysis.

Four categories:

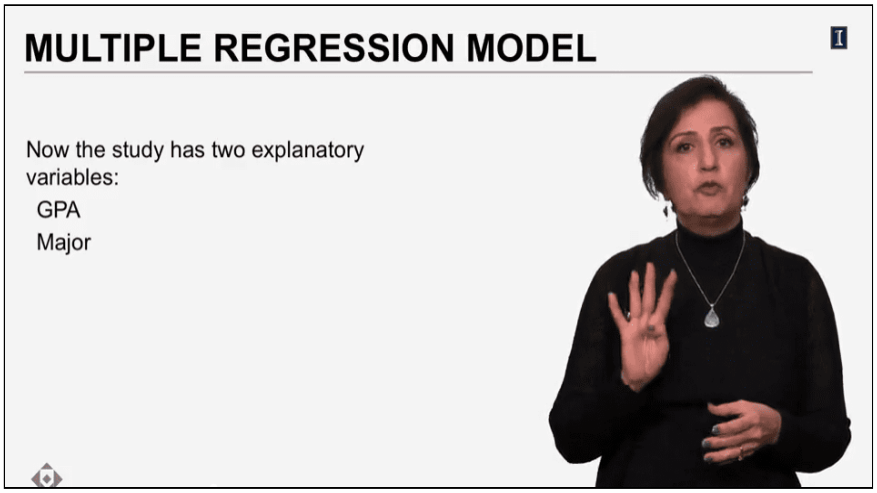
1. Engineering
2. Business
3. Agriculture
4. Liberal Arts

Transcript

Let's go back to an example we looked at in the last module when we learned about simple linear regression, which was an advisor in the university would like to motivate her students to study hard and graduate with high GPA. To stress the importance of GPA, she's wondering if she can show data on relationship of graduate's starting salary and the student's GPA. She has data on 100 recent graduates. While GPA had a really small p-value and, thus, highly significant, it only explains about 33.6 of the variation.

The advisor was not happy to say that unknown factors was controlling almost 67% of the variation, so she decided to add another variable to her analysis, which she believe also contributed to a graduate's salary, and that was the graduate's major. There are lots of majors in the university, so she decided, for now, to simplify her analysis by categorizing the student in four possibilities; engineering, business, agriculture, and liberal arts.

Multiple Regression Model - Slide 61



Now the study has two explanatory variables:

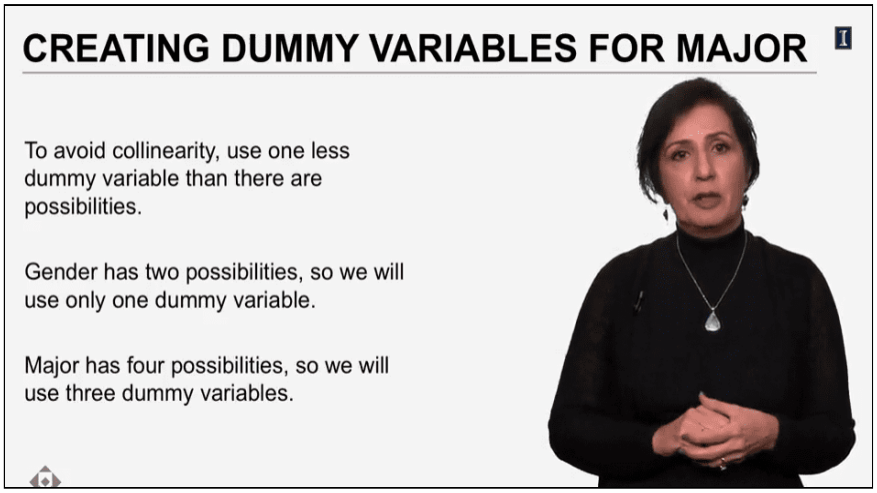
GPA

Major

Transcript

Now, this study has two explanatory variables; GPA and major. GPA is a numerical value and major is a categorical variable with four possibilities. So, we first have to construct dummy variables for major, and only then we can run the analysis.

Creating Dummy Variables for Major (1 of 2) - Slide 62



To avoid collinearity, use one less dummy variable than there are possibilities.

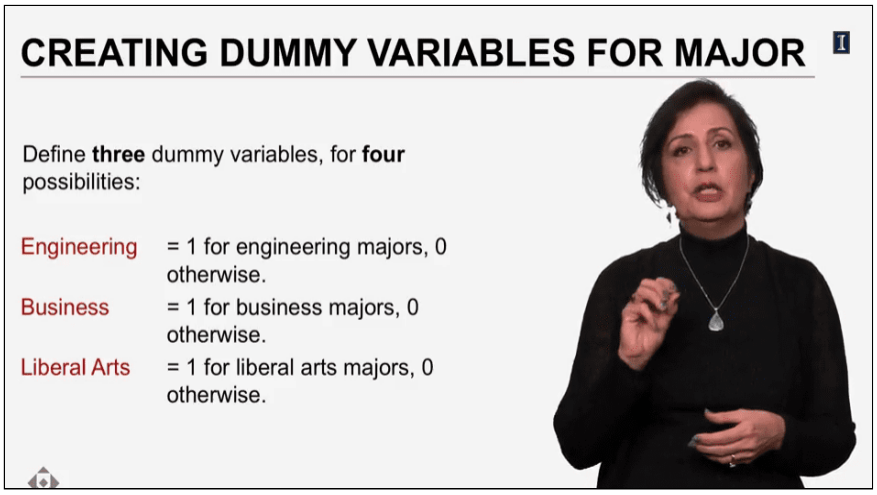
Gender has two possibilities, so we will use only one dummy variable.

Major has four possibilities, so we will use three dummy variables.

Transcript

Like before, we should be mindful of collinearity and to avoid collinearity use one less dummy variable than there are possibilities. In case of gender, there are only two possibilities and we use only one dummy variable.

Creating Dummy Variables for Major (2 of 2) - Slide 63



Define **three** dummy variables, for **four** possibilities:

Engineering = 1 for engineering majors, 0 otherwise.

Business = 1 for business majors, 0 otherwise.

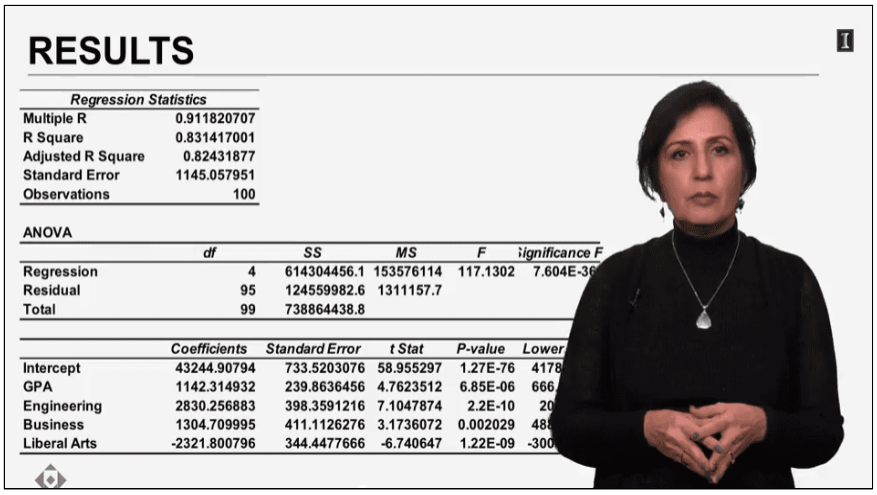
Liberal Arts = 1 for liberal arts majors, 0 otherwise.

Transcript

In this case, there are four possibilities so we will use three dummy variables. Define three dummy variables for four possibilities. It would be engineering is equal to 1 for engineering majors and 0 otherwise. Business will be 1 for business majors and 0 otherwise, and liberal arts will be 1 for liberal arts majors and 0 otherwise. So, if a student has 0 for all these values, then the student has to be in the college of agriculture, this is a major left out of the model.

Again, my numbers will look different based on which one is left out, results and the analysis will be the same. So, don't worry about which ones to code as one and keep in the model.

Results - Slide 64



Regression Statistics - Slide 64

Multiple R	0.911820707
R Square	0.831417001
Adjusted R Square	0.82431877
Standard error	1145.057951
Observations	100

ANOVA - Slide 64

	df	SS	MS	F	Significance F
Regression	4	614304456.1	153576114	117.1302	7.604E-36
Residual	95	124559982.6	1311157.7		
Total	99	738864438.8			

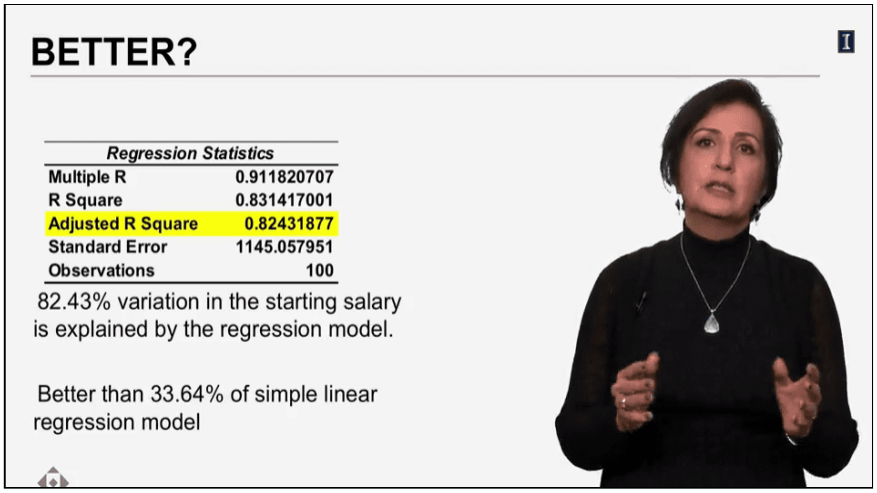
Other Statistics - Slide 64

	Coefficients	Standard error	t-stat	p-value
Intercept	43244.90794	733.5203076	58.955297	1.27E-76
GPA	1142.314932	239.8636456	4.7623512	6.85E-06
Engineering	2830.256883	398.3591216	7.1047874	2.2E-10
Business	1304.709995	411.1126276	3.1736072	0.002029
Liberal Arts	−2321.800796	344.4477666	−6.740647	1.22E-09

Transcript

Here's the complete analysis of our data. Too much to look at, at once, so let's review this piece-by-piece.

Better? - Slide 65



This slide contains the Regression Statistics table presented on Slide 64 - Results with the Adjusted R Square highlighted.

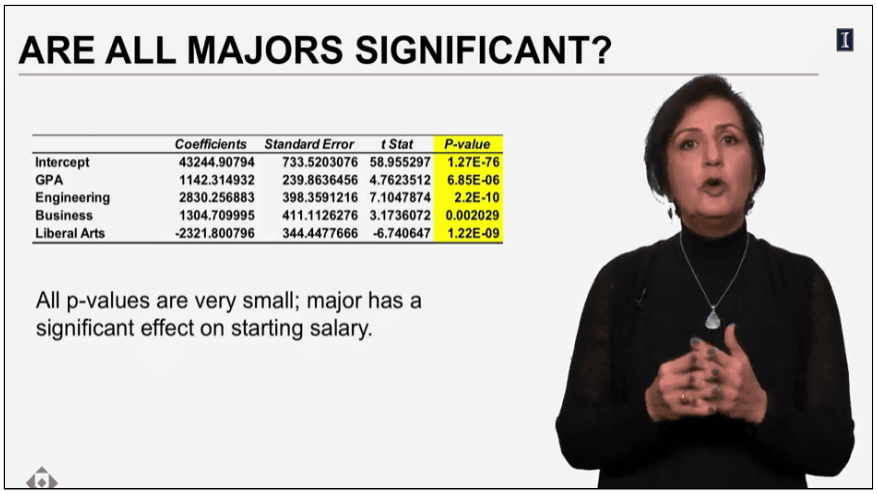
82.43% variation in the starting salary is explained by the regression model.

Better than 33.64% of simple linear regression model

Transcript

So, did our prediction model improve as compared to the simple linear regression? Let's look at the adjusted R-squared. As you can see, 82.43% of the variations in the starting salary is explained by the regression model. This is definitely better than the 33.64% for the simple linear regression model, which only considered GPA.

Are All Majors Significant? - Slide 66



This slide contains the Other Statistics table presented on [Slide 64 - Results]() with the P-value column highlighted.

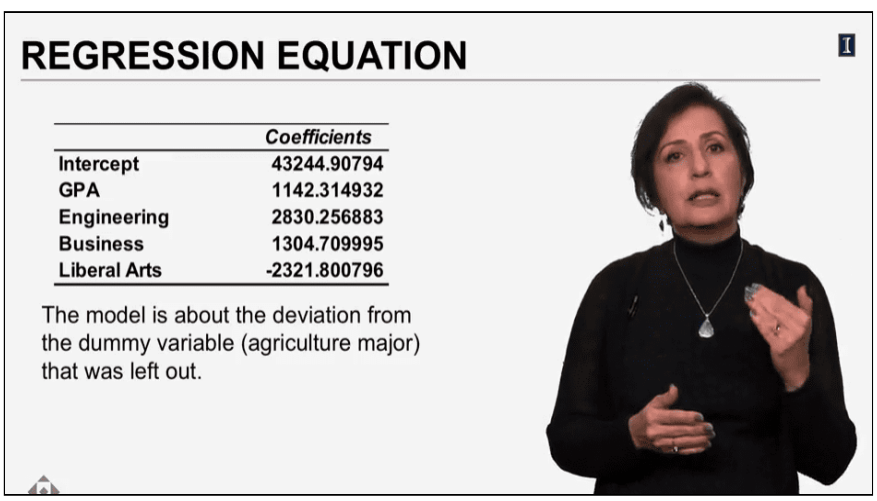
All p-values are very small; major has a significant effect on starting salary.

Transcript

Now we can turn our attention to the individual variables that are in the model. For that, we will look at the p-values of each. All p-values are very small so we conclude that all majors have a significant effect on starting salary.

Now let's move on to using the results for prediction.

Regression Equation - Slide 67



This slide contains a table showing the Coefficients vaues.

Other Statistics - Slide 67

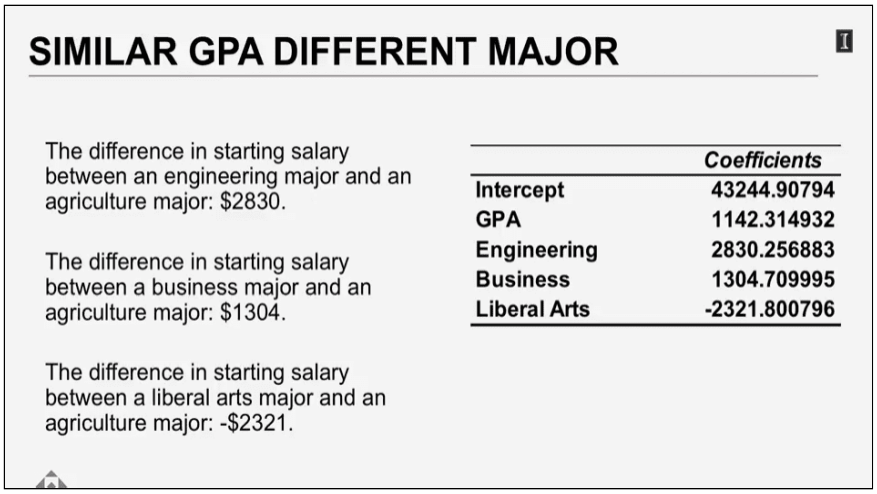
	Coefficients
Intercept	43244.90794
GPA	1142.314932
Engineering	2830.256883
Business	1304.709995
Liberal Arts	−2321.800796

The model is about the deviation from the dummy variable (agriculture major) that was left out.

Transcript

Just as before, the model is about the deviation from the dummy variable, agriculture major students, that were left out.

Similar GPA Different Major (1 of 2) - Slide 68



The difference in starting salary between an engineering major and an agriculture major: $2830.

The difference in starting salary between a business major and an agriculture major: $1304.

The difference in starting salary between a liberal arts major and an agriculture major: -$2321.

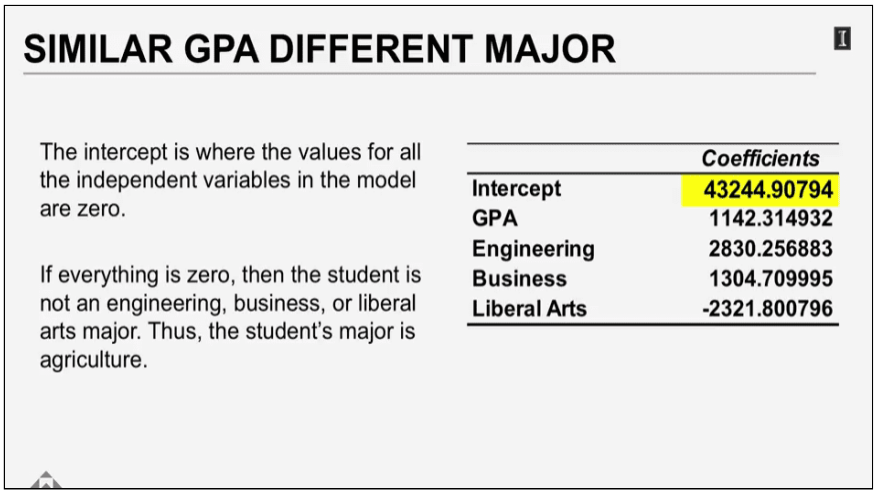
This slide contains the Other Statistics table presented on Slide 67 - Regression Equation.

Transcript

Now we will explore what exactly this means. For students who have the same GPA the coefficient for engineering major is 2,830, and this is being a positive number value, means that majoring in engineering as compared to agriculture adds about $2,830 to the person's salary.

Same goes for business, compared to ag majors, business majors get a positive bump of about $1,304. The difference in salary between liberal arts majors and an agriculture major is a negative value of 2,321, which means they earn about this much less than the students who major in agriculture, holding the GPAs the same, of course.

Similar GPA Different Major (2 of 2) - Slide 69



The intercept is where the values for all the independent variables in the model are zero.

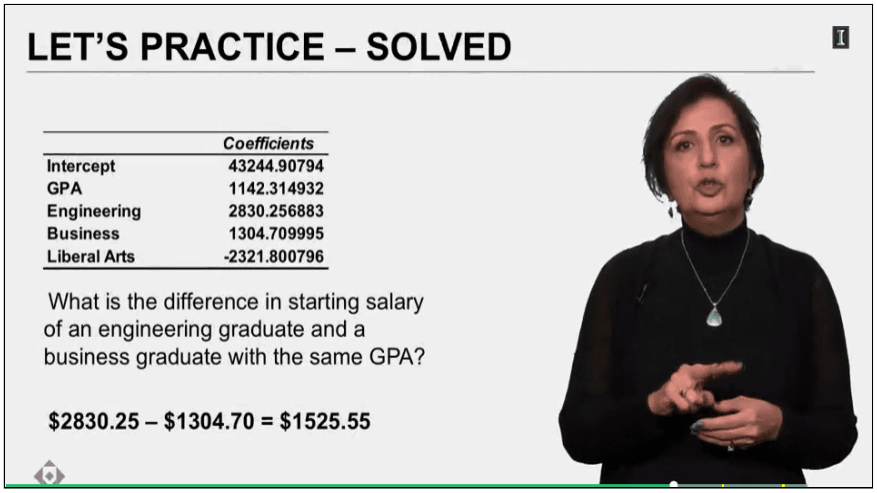
If everything is zero, then the student is not an engineering, business, or liberal arts major. Thus, the student's major is agriculture.

This slide contains the Other Statistics table presented on [Slide 67 - Regression Equation] with the Intercept highlighted.

Transcript

The intercept is where the values for all independent variables in the model is 0. Everything is 0, then the student is not in engineering, is not in business and is not a liberal arts major, thus, the student's major is agriculture.

Let's Practice - Solved (1 of 3) - Slide 70



This slide contains the Other Statistics table presented on Slide 67 - Regression Equation.

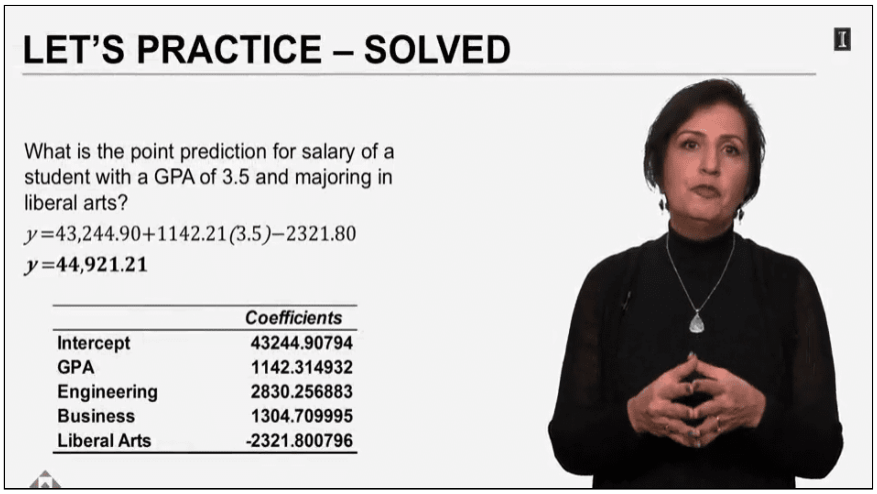
What is the difference in starting salary of an engineering graduate and a business graduate with the same GPA?

$2830.25 − 1304.70 = $1525.55

Transcript

What is the difference in starting salary of an engineering graduate and business graduate with the same GPA?

Let's Practice - Solved (2 of 3) - Slide 71



What is the point prediction for salary of a student with a GPA of 3.5 and majoring in liberal arts?

y = 43,244.90 + 1142.21(3.5) − 2321.80

y = 44,921.21

This slide contains the Other Statistics table presented on Slide 67 - Regression Equation.

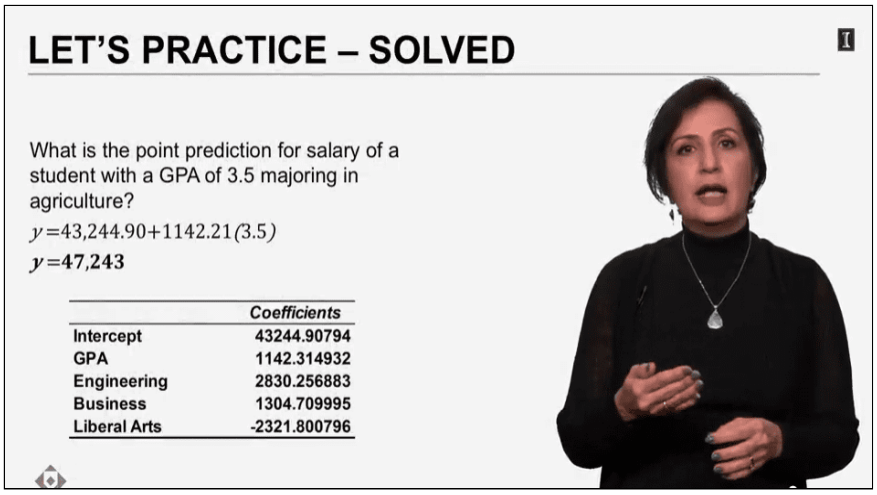
Transcript

It is the difference between the two coefficients, which means the engineering on average will make $1,525.55 more than a business graduate. Now let's look at this one. What is the point prediction for salary of a student with a GPA of 3.5 majoring in liberal arts?

Take the intercept and add to that the coefficient of the GPA times 3.5. Then add liberal arts coefficient to get 44,921.21.

What is the point prediction for a salary of a student with a GPA of 3.5 majoring in agriculture?

Let's Practice - Solved (3 of 3) - Slide 72



What is the point prediction for salary of a student with a GPA of 3.5 majoring in agriculture?

y = 43,244.90 + 1142.21(3.5)

y = 47,243

This slide contains the Other Statistics table presented on [Slide 67 - Regression Equation](#).

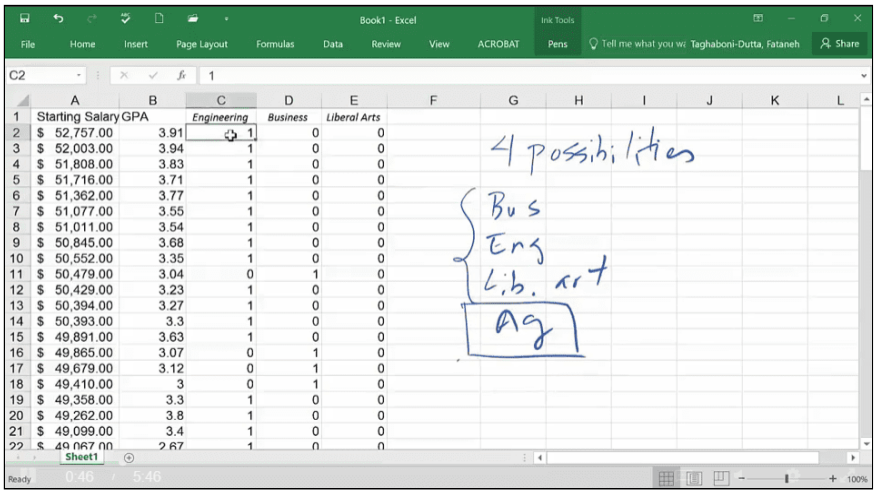
Transcript

Take the intercept, add to that the coefficient of GPA times 3.5. All other coefficients will be multiplied by 0, so the salary is 47,243. Once again, you see that you're comparing effect of many different variables at the same time. Regression allows us to know the collective impact, as well as the individual impact that these variables have on the response variable. One of the challenges in regression is to select the right independent variables which doesn't violate the model's assumption.

Lesson 4-3.2 Dummy Variable in Excel – Many Categories

Media Player for Video

Excel Worksheet (1 of 8) - Slide 73



	Starting Salary	GPA	Engineering	Business	Liberal Arts
2	$ 52,757.00	3.91	1	0	0
3	$ 52,003.00	3.94	1	0	0
4	$ 51,808.00	3.83	1	0	0
5	$ 51,716.00	3.71	1	0	0
6	$ 51,362.00	3.77	1	0	0
7	$ 51,077.00	3.55	1	0	0
8	$ 51,011.00	3.54	1	0	0
9	$ 50,845.00	3.68	1	0	0
10	$ 50,552.00	3.35	1	0	0
11	$ 50,479.00	3.04	0	1	0
12	$ 50,429.00	3.23	1	0	0
13	$ 50,394.00	3.27	1	0	0
14	$ 50,393.00	3.3	1	0	0
15	$ 49,891.00	3.63	1	0	0
16	$ 49,865.00	3.07	0	1	0
17	$ 49,679.00	3.12	0	1	0
18	$ 49,410.00	3	0	1	0
19	$ 49,358.00	3.3	1	0	0
20	$ 49,262.00	3.8	1	0	0
21	$ 49,099.00	3.4	1	0	0
22	$ 49,067.00	2.67	1	0	0

This slide contains an Excel sheet with the GPA Major Salary data set. There are five columns: Starting Salary, GPA, Engineering, Business, Liberal Arts.

To the right of these columns is writen:

4 possibilities

- Bus
- Eng
- Lib. art
- Ag

[Download the GPA Majors Salary Excel file (Refer to Data - First worksheet)]

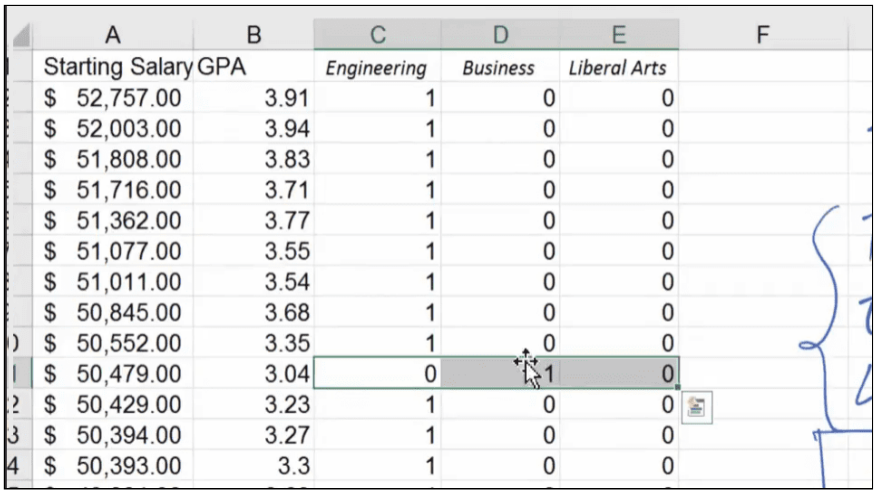
Transcript

So, here's the example on the categorical data, like major, has more than two possibilities. Here, the adviser had broken the majors to four possibilities: business, engineering, liberal arts and ag.

So, if you can code a student as either business, engineer or liberal arts and ag, all you need to do is to just pick three. And again, it doesn't matter which one you pick, and in this case, I picked business, engineer and liberal arts to code, and I left ag out; which means all of our data will be compared to students who are left out and those will be the students who have majored in agriculture.

So, if the student has a value for 1 in engineering, it means that they are an engineering student, if they have a value for business it means that they are a business student. And, as you can see, the rest of it would be 0 here, we have assumed that they are not doing double majors in business and engineering and so on.

Excel Worksheet (2 of 8) - Slide 74



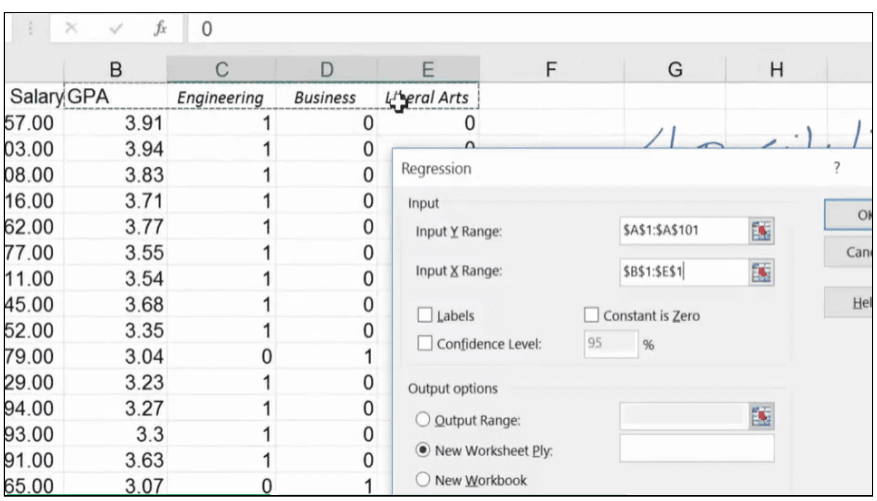
Starting Salary	GPA	Engineering	Business	Liberal Arts
$ 52,757.00	3.91	1	0	0
$ 52,003.00	3.94	1	0	0
$ 51,808.00	3.83	1	0	0
$ 51,716.00	3.71	1	0	0
$ 51,362.00	3.77	1	0	0
$ 51,077.00	3.55	1	0	0
$ 51,011.00	3.54	1	0	0
$ 50,845.00	3.68	1	0	0
$ 50,552.00	3.35	1	0	0
$ 50,479.00	3.04	0	1	0
$ 50,429.00	3.23	1	0	0
$ 50,394.00	3.27	1	0	0
$ 50,393.00	3.3	1	0	0

This slide shows a randomly selected row of data where the Starting Salary is $50,479, the GPA is 3.04, Engineering is 0, Business is 1, and Liberal Arts is 0.

Transcript

And we will have students, like this one, who has 0 for all the variables, so this person must be an ag major. So, if they are not an engineer, a business or liberal arts, the only thing remaining is that they are ag. So, whenever you have a categorical variable you need to take the number of possibilities, minus 1, and include that many dummy variables in your model. So, here we have four majors, meaning 4 minus 1, or three dummy variables, and then we code it as 0's and 1's, and then you are ready to run the problem.

Excel Worksheet (3 of 8) - Slide 75



This slide contains the Regression pop-up window; the Input y range is $A$1:$A$101, the Input x range is $B$1:$E$1, and the Confidence level is set to 95%. In the Output section, the New worksheet ply option is selected.

Transcript

So, we would go to "Data", we would go to "Data analysis", pick "Regression", input your y-value. Prediction is for the starting salaries, so the y-value is right here. I'll pick all of the data that I have and the input is the GPA as well as the person's major, and I pick all of that. I make sure that it knows that I have labels included and then I say, "Okay".

Excel Worksheet (4 of 8) - Slide 76

Adjusted R Square

	A	B	C	D	E	F	G	H
1	SUMMARY OUTPUT							
2								
3	*Regression Statistics*							
4	Multiple R	0.911821						
5	R Square	0.831417						
6	Adjusted R Square	0.824319						
7	Standard Error	1145.058						
8	Observations	100						
9								
10	ANOVA							
11		*df*	*SS*	*MS*	*F*	*Significance F*		
12	Regression	4	6.14E+08	1.54E+08	117.1301611	7.60373E-36		
13	Residual	95	1.25E+08	1311158				
14	Total	99	7.39E+08					
15								
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	
17	Intercept	43244.91	733.5203	58.9553	1.2694E-76	41788.68602	44701.12987	
18	GPA	1142.315	239.8636	4.762351	6.85279E-06	666.1253891	1618.504475	
19	Engineering	2830.257	398.3591	7.104787	2.20151E-10	2039.414037	3621.099729	
20	Business	1304.71	411.1126	3.173607	0.002029098	488.5482389	2120.871752	
21	Liberal Arts	-2321.8	344.4478	-6.74065	1.21523E-09	-3005.61607	-1637.985522	

This slide contains the Summary Output generated by Excel.

Regression Statistics - Slide 76

Multiple R	0.911821
R Square	0.831417
Adjusted R Square	0.824319
Standard error	1145.058
Observations	100

ANOVA - Slide 76

	df	SS	MS	F	Significance F
Regression	4	6.14E+08	1.54E+08	117.1301611	7.60373E-36
Residual	95	1.25E+08	1311158		
Total	99	7.39E+08			

Other Statistics - Slide 76

	Coefficients	Standard error	t-stat	p-value	Lower 95%	Upper 95%
Intercept	43244.91	733.5203	58.9553	1.2694E-76	41788.68602	44701.12987
GPA	1142.315	239.8636	4.762351	6.85279E-06	666.1253891	1618.504475
Engineering	2830.257	398.3591	7.104787	2.2015E-10	2039.414037	3621.099729
Business	1304.71	411.1126	3.173607	0.002029098	488.5482389	2120871752
Liberal Arts	−2321.8	344.4478	−6.74065	1.21523E-09	-3005.61607	-1637.985522

Download the GPA Majors Salary Excel file (Refer to Regression Output - Second worksheet)

Transcript

And here's our analysis. First of all, Adjusted R-Squared is significantly higher than what it used to be, so 82.4% of variations in the starting salary of these graduates can be explained by their GPA and their major. So, now let's look at the variables that we have and make sure that every single variable is significant in our model, and for that, I will go to the last table.

Excel Worksheet (5 of 8) - Slide 77

	A	B	C	D	E	F	G
6		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
7	Intercept	43244.91	733.5203	58.9553	1.2694E-76	41788.68602	44701.12987
8	GPA	1142.315	239.8636	4.762351	6.85279E-06	666.1253891	1618.504475
9	Engineering	2830.257	398.3591	7.104787	2.20151E-10	2039.414037	3621.099729
0	Business	1304.71	411.1126	3.173607	0.002029098	488.5482389	2120.871752
1	Liberal Arts	-2321.8	344.4478	-6.74065	1.21523E-09	-3005.61607	-1637.985522
2							
3			Student 1	Student 2	Student 3	Student 4	
4	Intercept	43244.91					
5	GPA	1142.315					
6	Engineering	2830.257					
7	Business	1304.71					
8	Liberal Arts	-2321.8					
9							
0	Point Estimate of Prediction						
1							
2							
3							
4							

This slide shows the Other Statistics table presented on Slide 76 - Excel Worksheet (4 of 8).

The following table is added:

Student Data - Slide 77

		Student 1	Student 2	Student 3	Student 4
Intercept	43244.91				
GPA	1142.315				
Engineering	2830.257				
Business	1304.71				
Liberal Arts	−2321.8				

Below the student data table, a new cell is added and labeled "Point Estimate of Prediction"

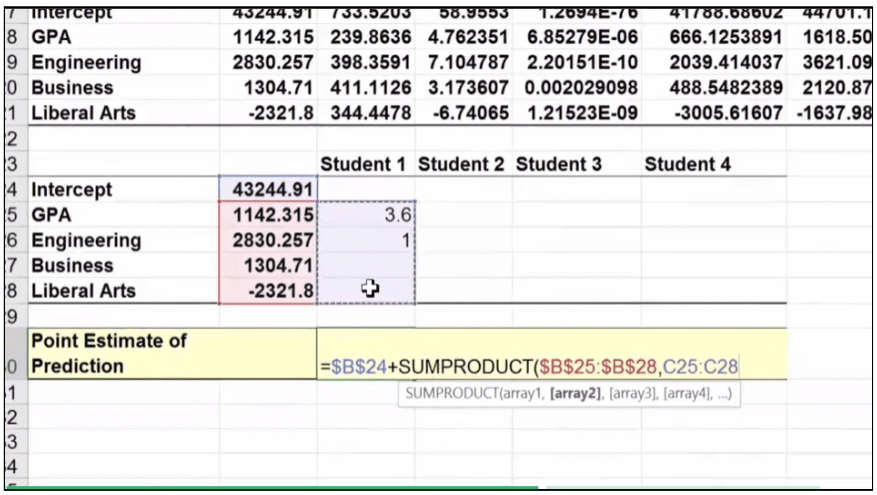
Download the GPA Majors Salary Excel file (Refer to Regression Output - Second worksheet)

Transcript

So, looking at the p-value, I see that all the values are less than 0.05 so, every variable that we have identified, GPA and major, has a significant relationship with a starting salary, what you are trying to predict. And now we are ready to make the predictions.

So, let's take these values like I like to do, copy them down here, and now I'm going to say: student number 1, student number 2, number 3 and student number 4. I am going to predict point estimate of prediction. So, I'm going to make the prediction here for each type of student [Selects B24].

Excel Worksheet (6 of 8) - Slide 78



7	Intercept	43244.91	733.5203	58.9553	1.2694E-76	41788.68602	44701.1
8	GPA	1142.315	239.8636	4.762351	6.85279E-06	666.1253891	1618.50
9	Engineering	2830.257	398.3591	7.104787	2.20151E-10	2039.414037	3621.09
0	Business	1304.71	411.1126	3.173607	0.002029098	488.5482389	2120.87
1	Liberal Arts	-2321.8	344.4478	-6.74065	1.21523E-09	-3005.61607	-1637.98
2							
3			Student 1	Student 2	Student 3	Student 4	
4	Intercept	43244.91					
5	GPA	1142.315	3.6				
6	Engineering	2830.257	1				
7	Business	1304.71					
8	Liberal Arts	-2321.8					
9							
0	Point Estimate of Prediction		=$B$24+SUMPRODUCT($B$25:$B$28,C25:C28				
1							
2							
3							
4							

SUMPRODUCT(array1, **[array2]**, [array3], [array4], ...)

In the Student 1 column, 3.6 has been entered in GPA row and a 1 has been entered in the Engineering row.

The following formula has been entered into the Point Estimate of Prediction value cell: =$B$24 + SUMPRODUCT($B$25:$B$28, C25:C28)

Transcript

So, let's say this student has a GPA of 3.6 and is an engineering student, so that means that they are not business or liberal arts, so this is all I need to provide. So, I'm going to do the prediction right here, which is my intercept and I'm going to copy my formula. So, I'm going to lock the cells for the coefficients for my model. So, I'm going to press F4 to lock it, plus, and then I'm going to use the function I told you about, Sum product. So, I'm going to pick Sum product, and Sum product is going to take these variables, remember, from GPA onwards, and I'm going to say F4 again to lock them [Selects B25 through B28].

Second array is basically what's here and, if the cell is empty, is just going to replace it with 0, so you're not going to get any problems there. And this is the value we get.

Excel Worksheet (7 of 8) - Slide 79

	A	B	C	D	E	F	G
6		*Coefficients*	*standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
7	Intercept	43244.91	733.520308	58.95529748	1.2694E-76	41788.68602	44701.1298
8	GPA	1142.315	239.863646	4.76235125	6.85279E-06	666.1253891	1618.50447
9	Engineering	2830.257	398.359122	7.104787439	2.20151E-10	2039.414037	3621.09972
0	Business	1304.71	411.112628	3.173607201	0.002029098	488.5482389	2120.87175
1	Liberal Arts	-2321.8	344.447767	-6.740646976	1.21523E-09	-3005.61607	-1637.98552
2							
3			Student 1	Student 2	Student 3	Student 4	
4	Intercept	43244.91					
5	GPA	1142.315	3.6	2.6			
6	Engineering	2830.257	1	1			
7	Business	1304.71					
8	Liberal Arts	-2321.8					
9							
0	Point Estimate of Prediction		$50,187.50	$ 49,045.18			
1							
2			=C30-D30				
3							
4							
5							

This slide shows the Other Statistics table presented on [Slide 76 - Excel Worksheet (4 of 8)](#).

In the Student 2 column, 2.6 has been entered in GPA row and a 1 has been entered in the Engineering row.

Student Data - Slide 79

		Student 1	**Student 2**	**Student 3**	**Student 4**
Intercept	43244.91				
GPA	1142.315	**3.6**	2.6		
Engineering	2830.257	**1**	1		
Business	1304.71				
Liberal Arts	−2321.8				

Point Estimate of Prediction for Student 1 is **$50,187.50**.

Point Estimate of Prediction for Student 2 is **$49,045.18**.

Transcript

Okay, let's now focus on this column, and then I will change the values, so you can see how these coefficients will matter. So, let's assume the next student has a GPA of 2.6 and still an engineering student. What would the expected salary be?

So, if I just dragged this out, you would see that the salary difference is what the GPA coefficient is. So, for every complete point that it drops the student's salary goes down by $1,142. So, that's why you see the difference between these two be. So, if I just quickly do that you would see that, that's the difference you will see, exactly what you would see as the coefficient.

Excel Worksheet (8 of 8) - Slide 80

	A	B	C	D	E	F	G
6		*Coefficient*	*standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
7	Intercept	43244.91	733.520308	58.95529748	1.2694E-76	41788.68602	44701.1298
8	GPA	1142.315	239.863646	4.76235125	6.85279E-06	666.1253891	1618.50447
9	Engineering	2830.257	398.359122	7.104787439	2.20151E-10	2039.414037	3621.09972
0	Business	1304.71	411.112628	3.173607201	0.002029098	488.5482389	2120.87175
1	Liberal Arts	-2321.8	344.447767	-6.740646976	1.21523E-09	-3005.61607	-1637.98552
2							
3			Student 1	Student 2	Student 3	Student 4	
4	Intercept	43244.91					
5	GPA	1142.315	3.6	2.6	2.6	4	
6	Engineering	2830.257	1	1			
7	Business	1304.71					
8	Liberal Arts	-2321.8				1	
9							
0	Point Estimate of Prediction		$50,187.50	$ 49,045.18	$ 46,214.93	$ 45,492.37	
1							
2			$ 1,142.31	$ 2,830.26		$ 722.56	
3							
4							

This slide shows the Other Statistics table presented on Slide 76 - Excel Worksheet (4 of 8).

In the Student 3 column, 2.6 has been entered in GPA row

In the Student 4 column, 4 has been entered in GPA row and a 1 has been entered in the Liberal Arts row.

Student Data - Slide 80

		Student 1	**Student 2**	**Student 3**	**Student 4**
Intercept	43244.91				
GPA	1142.315	**3.6**	2.6	2.6	4
Engineering	2830.257	**1**	1		
Business	1304.71				
Liberal Arts	−2321.8				1

Point Estimate of Prediction for Student 1 is **$50,187.50**.

Point Estimate of Prediction for Student 2 is **$49,045.18**.

Point Estimate of Prediction for Student 3 is **$46,214.93**.

Point Estimate of Prediction for Student 4 is **$45,492.37**.

Download the GPA Majors Salary Excel file (Refer to Regression Output - Second worksheet)

Transcript

So, what if this student is a 2.6, but is an ag major? So, if it's an ag major then it doesn't get any coefficient get multiplied by 1. So, it's just this value. So, I can just drag this out and you would see the difference between an Ag major and a GPA of the same engineering major, that is a difference that you would see as your engineering coefficient being removed. So, again, if I look at the difference between an engineering student with a 2.6 and a student in ag, you would see that you would get the value. So, the engineering student had this additional value added to that number, and what the ag major is getting is that intercept plus the bonus they get for their GPA.

And, if I take this value and now I say, it's a 2.6, and the student is a liberal arts major; what you would see is, again, the coefficient of liberal arts is going to reduce this person's salary by $2,000.32. So, again, if I look at the difference here, is the difference between this value, and this value. Remember, everybody is being compared to an ag major here. So, this is what you would get. So, this is how we would use it. Now, of course, if I have someone who has a little higher GPA, so let's say this person is now a 3.6, the value would go up. And, if it's a 4.0 the value would go up also. This is how we can use regression model for prediction.