Multiple Regression

Rob McCulloch

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへで

- 1. Multiple Regression Model
- 2. Estimates and Plug-in Prediction
- 3. Confidence Intervals and Hypothesis Tests

- 4. Fits, Resids, and R-squared
- 5. Categorical x and Dummy Variables
- 6. The SLR Slope and MLR sigmahat

1. Multiple Regression Model

The plug-in predictive interval for the price of a house given its size is quite large ($\pm 2 * 22.5 = \pm 45$).

How can we improve this?

If we know more about a house, we should have a better idea of its price !!

Our data has more variables than just size and price:

The first 7 rows are:

(price and size /1000)

Home	Nbhd	Offers	SqFt	Brick	Bedrooms	Bathrooms	Price	pricethou	sizethou
1	2	2	1790	No	2	2	114300	114.3	1.79
2	2	3	2030	No	4	2	114200	114.2	2.03
3	2	1	1740	No	3	2	114800	114.8	1.74
4	2	3	1980	No	3	2	94700	94.7	1.98
5	2	3	2130	No	3	3	119800	119.8	2.13
6	1	2	1780	No	3	2	114600	114.6	1.78
7	3	3	1830	Yes	3	3	151600	151.6	1.83

Suppose we know the number of bedrooms and bathrooms a house has as well as its size, then what would our prediction for price be ? The Multiple Regression Model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$ $\epsilon_i \sim N(0, \sigma^2)$

Y is a linear combination of the x's + error.

The ϵ is the same as in simple linear regression, ϵ is the part of Y you can't predict from x !!!!

Example:

In our housing example, we might want to relate price to size, nbed, and nbath.

y=price: thousands of dollars x1=nbed: number of bedrooms x2=nbath: number of bathrooms x2=size: thousands of square feet

 $price_i = \beta_0 + \beta_1 \operatorname{nbed}_i + \beta_2 \operatorname{nbath}_i + \beta_3 \operatorname{size} + \epsilon_i$

With 2 x variables our model is y = (a plane) + error.



5

Plotting in 3-D is interesting but tricky. Cars data: y=price, x1=mileage, x2=year.



Training data. 3D plot of (mileage, year) (scaled) vs price.

Later I'll show you a plot for the housing data with 4 variables that works!!!

2. Estimates and Plug-in Prediction

Multiple Regression Model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$$

The parameters of the model are $(\beta_0, \beta_1, \ldots, \beta_k, \sigma)$.

Given data, we will have estimates:

 $\hat{\beta}_j$: estimate of β_j .

 $\hat{\sigma}$: estimate of σ .

House Price Example

$$\mathsf{price}_i = eta_0 + eta_1 \,\mathsf{nbed}_i + eta_2 \,\mathsf{nbath}_i + eta_3 \,\mathsf{size} + \epsilon_i, \ \ \epsilon_i \sim N(0, \sigma^2)$$

Call: lm(formula = price ~ nbed + nbath + size, data = ddf) Residuals: Min 10 Median Max 30 -53.71 -15.63 -0.24 13.85 49.36 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -5.641 17,200 -0.328 0.743504 nbed 10,460 2.912 3.592 0.000472 *** nbath 13.546 4 219 3 211 0 001685 ** size 35.643 10.667 3.341 0.001102 ** Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.36 on 124 degrees of freedom Multiple R-squared: 0.4396,Adjusted R-squared: 0.426 F-statistic: 32.42 on 3 and 124 DF, p-value: 1.5356-15

$$\hat{eta}_0=-5.641$$
, $\hat{eta}_1=10.46$, $\hat{eta}_2=13.546$, $\hat{eta}_3=35.643$, $\hat{\sigma}=20.36$.

~

Prediction:

The plug-in estimate of the model for the conditional distribution of Y given the x's is

price = -5.64 + 10.46 nbed + 13.546 nbath + 35.643 size + ϵ $\epsilon \sim N(0, 20.36^2)$

Given nbed=3, nbath=2, and size = 2.2. price = $-5.64 + 10.46 * 3 + 13.546 * 2 + 35.643 * 2.2 + \epsilon$ = $131.2 + \epsilon$ $\sim N(131.2, 20.36^2)$ $\approx 131.2 \pm 2 * 20.36 = 131.2 \pm 40.72$

Note:

With just size, our plug-in predictive +/- was

 $2^{2}2.467 = 44.952$

With nbath and nbed added to the model the +/- is

 $2*\ 20.36 = 40.72$

The additional information makes our prediction more precise.

But not a whole lot, we still need better x's !!

Interpret:

 $\mathsf{price} = -5.64 + 10.46\,\mathsf{nbed} + 13.546\,\mathsf{nbath} + 35.643\,\mathsf{size} + \epsilon$

With size and nbath **held fixed**, if you add one bedroom the average price goes up by 10.46 (thousand dollars).

Note that it would not makes sense to add several bedrooms and imagine holding size and nbath fixed.

Note:

When we regressed price on size the coefficient was about 70.

Now the coefficient for size is about 36.

Without nbath and nbed in the regression, an increase in size can by associated with an increase in nbath and nbed in the background.

If all I know is that one house is a lot bigger than another I might expect the bigger house to have more beds and baths!

With nbath and nbed held fixed, the effect of size is smaller.

3. Confidence Intervals and Hypothesis Tests

Let $se(\hat{\beta}_j)$ denote the standard error of $\hat{\beta}_j$.

95% confidence interval for β_j : $\hat{\beta}_j \pm 2 \, se(\hat{\beta}_j)$

For small n - k - 1 (< 20 or 30) use: qt(.975,n-k-1) in R or =tinv(.05,n-k-1) in excel instead of 2.

House Price Example

$$price_i = \beta_0 + \beta_1 nbed_i + \beta_2 nbath_i + \beta_3 size + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2)$$

Call:

lm(formula = price ~ nbed + nbath + size, data = ddf)

Residuals:

Min 1Q Median 3Q Max -53.71 -15.63 -0.24 13.85 49.36

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)				
(Intercept)	-5.641	17.200	-0.328	0.743504				
nbed	10.460	2.912	3.592	0.000472	***			
nbath	13.546	4.219	3.211	0.001685	**			
size	35.643	10.667	3.341	0.001102	**			
Signif. code	es: 0 '**	**' 0.001 ' [,]	**' 0.01	'*' 0.05	'.' 0	.1'	,	1

Residual standard error: 20.36 on 124 degrees of freedom Multiple R-squared: 0.4396,Adjusted R-squared: 0.426 F-statistic: 32.42 on 3 and 124 DF, p-value: 1.535e-15

 $\hat{\beta}_3 = 35.74$, $se(\hat{\beta}_3) = 10.7$.

Confidence Interval: $35.64 \pm 2 * 10.7 = (14.24, 57.04)$.

Hypothesis Tests:



For small n, t thing.

IF the null is true, the t should look like a z !!!

House Price Example

price_i =
$$\beta_0 + \beta_1 \operatorname{nbed}_i + \beta_2 \operatorname{nbath}_i + \beta_3 \operatorname{size} + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2)$$

1

Call:

lm(formula = price ~ nbed + nbath + size, data = ddf)

Residuals:

Min 1Q Median 3Q Max -53.71 -15.63 -0.24 13.85 49.36

Coefficients:

	Estimate S	td. Error t	value	Pr(> t)		
(Intercept)	-5.641	17.200	-0.328	0.743504		
nbed	10.460	2.912	3.592	0.000472	***	
nbath	13.546	4.219	3.211	0.001685	**	
size	35.643	10.667	3.341	0.001102	**	
Signif. code	es: 0 '***	'0.001'**	, 0.01	'*' 0.05	'.' 0.1	۰,

Residual standard error: 20.36 on 124 degrees of freedom Multiple R-squared: 0.4396,Adjusted R-squared: 0.426 F-statistic: 32.42 on 3 and 124 DF, p-value: 1.535e-15

$$\hat{\beta}_3 = 35.74, \ se(\hat{\beta}_3) = 10.7.$$

Test $\beta_3 = 0: \ t = \frac{35.74 - 0}{10.7} = 3.34. \Rightarrow reject.$

p-values:

To test

$$H_0:\beta_3=0$$

we can just look at the p-value in the output.

The p-value is .0011, so we reject.

4. Fits, Resids, and R-squared

In multiple regression the fit is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$$

"estimate of the part of y related to the x's".

The residual,

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

"estimate of the part of y not related to the x's"

Recall that with $1 \times we$ can see



$$y_i = \hat{y}_i + e_i$$

With $2 \times$ variables we have a 3-D picture:



In multiple regression, the resids have sample mean 0 and are uncorrelated with each of the x's and the fitted values:

Table of corr	elations						
		SqFt	Bedrooms	Bathrooms	Price	Fitted Values	Residuals
SqFt		1.000					
Bedroo	oms	0.484	1.000				
Bathro	oms	0.523	0.415	1.000			
Price		0.553	0.526	0.523	1.000		
Fitted	Values	0.834	0.793	0.789	0.663	1.000	
Residu	lals	0.000	0.000	0.000	0.749	0.000	1.000

$$\mathbf{y}_{i} = \mathbf{\hat{y}}_{i} + \mathbf{e}_{i}$$

estimated x part of y

estimated part of y that has nothing to do with x's

This is the plot of the residuals from the multiple regression of price on size, nbath, nbed vs the fitted values. We see the 0 correlation.



The correlation is also 0, for each of the x's.

Key Intuition:

$$y_i = \hat{y}_i + e_i$$

Multiple regression pulls out of y everthing that looks linearly related to *each* x.

The part that is left over (the residuals), is linearly unrelated to each of the x's.

Since the fitted values are just a combination of the x's, they are uncorrelated with the residuals.

$$y_i = \hat{y}_i + e_i$$

 $cor(\hat{y}, e) = 0, \quad ar{e} = 0$
 $ar{\hat{y}} = ar{y}.$

Because the fits and the resids have 0 sample correlation, the variance of the sum is the sum of the variances:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2$$

Total variation =

variation explained by x's + unexplained variation.

R-squared:

$$\mathcal{R}^2 = rac{explained}{total} = rac{\sum (\hat{y_i} - ar{y})^2}{\sum (y_i - ar{y})^2}$$

 $0 \le R^2 \le 1$, the closer R^2 is to 1, the better the fit.

In our housing example:

Results of multiple regression for pricethou

Summary measures		
Multiple R	0.6630	
R-Square	0.4396	
Adj R-Square	0.4260	
StErr of Est	20.3565	

ANOVA Table

Source	df	SS	MS	F	p-value
Explained	3	40300.9877	13433.6626	32.4180	0.0000
Unexplained	124	51384.2266	414.3889		

Regression coefficients

-		Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
	Constant	-5.6408	17.2004	-0.3279	0.7435	-39.6852	28.4035
	Bedrooms	10.4599	2.9123	3.5916	0.0005	4.6956	16.2242
	Bathrooms	13.5461	4.2187	3.2110	0.0017	5.1962	21.8961
	sizethou	35.6427	10.6673	3.3413	0.0011	14.5292	56.7561

$$\mathsf{R}^2 = \frac{40301}{40301 + 51384} = .439$$

R² is also the square of the correlation between the fitted values and y:



Regression finds the linear combination of the x's which is most correlated with y.

27

Results of multiple r	egression fo	r pricethou	-0	or(ŷ v	() = 66	53
Summary measures		_		Oi (y, y) = .00))
Multiple R	0.6630	4				
R-Square	0.4396					
Adj R-Square	0.4260					
StErr of Est	20.3565					
ANOVA Table						
Source	df	SS	MS	F	p-value	
Explained	3	40300.9877	13433.6626	32.4180	0.0000	
Unexplained	124	51384.2266	414.3889			
Regression coefficie	nts					
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-5.6408	17.2004	-0.3279	0.7435	-39.6852	28.4035
Bedrooms	10.4599	2.9123	3.5916	0.0005	4.6956	16.2242
Bathrooms	13.5461	4.2187	3.2110	0.0017	5.1962	21.8961
sizethou	35.6427	10.6673	3.3413	0.0011	14.5292	56.7561

The overall F-test

The p-value beside "F" if testing the null hypothesis:

 $H_0: \beta_1 = \beta_2 = \cdots = \beta_k$ (all the slopes are 0) Results of multiple regression for pricethou Summary measures Multiple R 0.6630 R-Square 0.4396 Adj R-Square 0.4260 StErr of Est 20.3565 ANOVA Table SS Source df MS p-value Explained 3 40300 9877 13433 6626 32 4180 0 0000 Unexplained 124 51384.2266 414.3889 Regression coefficients Coefficient Std Err t-value p-value Lower limit Upper limit Constant -5.640817.2004 -0.32790.7435 -39.6852 28.4035 Bedrooms 2.9123 3.5916 0.0005 4.6956 16.2242 10.4599 Bathrooms 13.5461 4.2187 3.2110 0.0017 5.1962 21.8961 sizethou 10.6673 3.3413 0.0011 14.5292 56.7561 35.6427

We reject the null, at least some of the slopes are not 0.

5. Categorical x and Dummy Variables

Here, again, is the first 7 rows of our housing data:

Home	Nbhd	Offers	SqFt	Brick	Bedrooms	Bathrooms	Price	pricethou	sizethou
1	2	2	1790	No	2	2	114300	114.3	1.79
2	2	3	2030	No	4	2	114200	114.2	2.03
3	2	1	1740	No	3	2	114800	114.8	1.74
4	2	3	1980	No	3	2	94700	94.7	1.98
5	2	3	2130	No	3	3	119800	119.8	2.13
6	1	2	1780	No	3	2	114600	114.6	1.78
7	3	3	1830	Yes	3	3	151600	151.6	1.83

Does whether a house is brick or not affect the price of the house?

This is a categorical variable.

How can we use multiple regression with categorical x's ??!!

What about the neighborhood? (location, location, location !!)

Let's consider relating price to size *and* Brick.

Can we plot price vs. size and Brick?

In general, plotting with several numeric variables is hard. But when variables are categorical we can use some simple plots where we use the color and/or shape of the plot symbol to represent a variable.

Here we have to numeric variables: price, size, and one categorical variable: Brick.

Price vs size and Brick with Brick indicated by color..



Price vs size and Brick with Brick indicated by shape and color.



Price vs size and Brick with Brick indicated by separate plots.



Well, it really looks like Brick=Yes houses sell for more!!

How can we put Brick into a regression model !!??

Adding a Binary Categorical x

.

To add "brick" as an explanatory variable in our regression we create the dummy variable which is 1 if the house is brick and 0 otherwise:

										
Home	Nbhd	Offers	SqFt	Brick	Bedrooms	Bathrooms	Price	sizethou	pricethou	brickdum
1	2	2	1790	No	2	2	114300	1.79	114.3	0
2	2	3	2030	No	4	2	114200	2.03	114.2	0
3	2	1	1740	No	3	2	114800	1.74	114.8	0
4	2	3	1980	No	3	2	94700	1.98	94.7	0
5	2	3	2130	No	3	3	119800	2.13	119.8	0
6	1	2	1780	No	3	2	114600	1.78	114.6	0
7	3	3	1830	Yes	3	3	151600	1.83	151.6	1
8	3	2	2160	No	4	2	150700	2.16	150.7	0
9	2	3	2110	No	4	2	119200	2.11	119.2	0
10	2	3	1730	No	3	3	104000	1.73	104	0
11	2	3	2030	Yes	3	2	132500	2.03	132.5	1
12	2	2	1870	Yes	2	2	123000	1.87	123	1
13	1	4	1910	No	3	2	102600	1.91	102.6	0
14	1	5	2150	Yes	3	3	126300	2.15	126.3	1

"brick dumm

Note:

I created the dummy by using the excel formula:

```
=IF(Brick="Yes",1,0)
```

In R:

```
mc = read.table("midcity.txt",header=T)
bdum=as.numeric(mc$Brick)-1
```

> table(mc\$Brick)

No Yes 86 42 > table(bdum) bdum 0 1 86 42 As a simple first example, let's regress price on size and brick.

Here is our model

$$Price_i = \beta_0 + \beta_1 \operatorname{size}_i + \beta_2 \operatorname{brickdum}_i + \epsilon_i$$

How do you interpet β_2 ?

$$Price_i = \beta_0 + \beta_1 \operatorname{size}_i + \beta_2 \operatorname{brickdum}_i + \epsilon_i$$

What is the expected value of a brick house given the size=s:

Expected price: $\beta_0 + \beta_1 s + \beta_2$

What is the expected value of a non brick house given the size=s:

Expected price: $\beta_0 + \beta_1 s$

 β_2 is the expected difference in price between a brick and non-brick house.

Note:

You could also create a dummy which is 1 if a house is non-brick and 0 if it is brick.

You would get the fit, but the meaning of β_2 would change.

You can't put both dummies in though because given one, the information in the other is redundant.

Let's try it !!

Results of multiple regression for pricethou

Summary measures

Multiple R	0.6884
R-Square	0.4739
Adj R-Square	0.4655
StErr of Est	19.6441

ANOVA Table

Source	Source df		MS	F	p-value					
Explained	2	43448.6791	21724.3396	56.2964	0.0000					
Unexplained	125	48236.5352	385.8923							
Regression coefficients										
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit				
Constant	-9.4443	16.5771	-0.5697	0.5699	-42.2525	23.3639				
sizethou	66.0584	8.2653	7.9922	0.0000	49.7003	82.4165				
brickdum	23.4451	3.7098	6.3198	0.0000	16.1029	30.7873				

 $\pm 2 \hat{\sigma} = 39.3$, this is best we have done !!

What is the brick effect: $23.4 \pm 2(3.7) = 23.4 \pm 7.4$.

We can see the effect of the dummy by plotting the fitted values vs size.

The upper line is for the brick houses and the lower line is for the non-brick houses.



We can interpret β_2 as a shift in the intercept.

Notice that our model assumes that the price difference between a brick and non-brick house does not depend on the size!

The two variables do not "interact".

Sometimes we expect variables to interact.

Now let's add brick to the regression of price on size, nbath, and nbed.

Results of multiple regression for pricethou

Summary measures

Multiple R	0.7634
R-Square	0.5828
Adj R-Square	0.5692
StErr of Est	17.6345

ANOVA Table

Source	df	SS	MS	F	p-value	
Explained	4	53435.3823	13358.8456	42.9580	0.0000	
Unexplained	123	38249.8320	310.9742			
Regression coeffic	ients					
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-5.2794	14.9004	-0.3543	0.7237	-34.7739	24.2151
Bedrooms	10.8731	2.5237	4.3084	0.0000	5.8776	15.8686
Bathrooms	9.8184	3.6993	2.6541	0.0090	2.4959	17.1409
sizethou	35.8006	9.2409	3.8742	0.0002	17.5088	54.0923
brickdum	21.9091	3.3712	6.4989	0.0000	15.2361	28.5821

 $\pm 2\,\hat{\sigma} = 35.2$

Adding brick seems to be a good idea !!

Using a Categorical Variable with more than two categories (levels):

How do we use the variables "neighborhood" in our multiple regression model?

It is a categorical variable with more than two levels.

Recall that we are *very excited* about doing this!!!

Plot price vs size, Brick, and Neighborhood.



Let's just start by building a model with size and Neighborhood.

We start by creating a separate dummy variable for each of the three neighborhoods.

Create a dummy for each neighborhood:

•

Home	Nbhd	Offers	SqFt	Brick	Bedrooms	Bathrooms	Price	Nbhd_1	Nbhd_2	Nbhd_3
1	2	2	1790	No	2	2	114300	_0	_1	_0
2	2	3	2030	No	4	2	114200	0	1	0
3	2	1	1740	No	3	2	114800	0	1	0
4	2	3	1980	No	3	2	94700	0	1	0
5	2	3	2130	No	3	3	119800	0	1	0
6	1	2	1780	No	3	2	114600	1	0	0
7	3	3	1830	Yes	3	3	151600	0	0	1
8	3	2	2160	No	4	2	150700	0	0	1

eg. Nbhd_1 indicates if the house is in neighborhood 1 or not

Now we add any two of the three dummies. Given any two, the information in the third is redundant.

Let's first do price on size and neighborhood:

 $Price_i = \beta_0 + \beta_1 \operatorname{Nbhd}_1 + \beta_2 \operatorname{Nbhd}_2 + \beta_3 \operatorname{size}_i + \epsilon_i$

What is the expected price of a house with size=s in neighborhood 1?

 $\mathsf{Price}_i = \beta_0 + \beta_1 \,\mathsf{Nbhd}_1 + \beta_2 \,\mathsf{Nbhd}_2 + \beta_3 \,\mathit{size}_i + \epsilon_i$ Suppose size = s.

Expected price in neighborhood 1: $\beta_0 + \beta_1 + \beta_3 s$. Expected price in neighborhood 2: $\beta_0 + \beta_2 + \beta_3 s$. Expected price in neighborhood 3: $\beta_0 + \beta_3 s$.

 β_1 : how different (on average) 1 is than 3. β_2 : how different (on average) 2 is than 3.

The neighborhood corresponding to the dummy we leave out becomes the "base case" we compare to.

Let's try price on size and neighborhood !!

Results of multiple regression for pricethou

Summary measures

Multiple R	0.8277
R-Square	0.6851
Adj R-Square	0.6774
StErr of Est	15.2601

ANOVA Table

Source	df	SS	MS	F	p-value	
Explained	3	62809.1498	20936.3833	89.9053	0.0000	
Unexplained	124	28876.0645	232.8715			
Regression coefficie	nts					
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	62.7765	14.2477	4.4061	0.0000	34.5763	90.9766
Nbhd_1	-41.5353	3.5337	-11.7542	0.0000	-48.5294	-34.5412
Nbhd_2	-30.9666	3.3688	-9.1922	0.0000	-37.6344	-24.2988
sizethou	46.3859	6.7459	6.8762	0.0000	33.0340	59.7379

 $\pm 2\,\hat{\sigma} = 30.52$

neighborhood works great !!

Here is fits vs size.

Which line corresponds to which neighborhood ?

Where do you want to live ?

Again we are assuming that size and neighborhood do not interact.



Same plot done in ggpplot and including the data.



ok, let's try price on size, nbed, nbath, brick, and neighborhood.

Results of multiple regression for pricethou

Summary measures

Multiple R	0.8972
R-Square	0.8050
Adj R-Square	0.7954
StErr of Est	12.1547

ANOVA Table

Source	df	SS	MS	F	p-value
Explained	6	73809.1440	12301.5240	83.2669	0.0000
Unexplained	121	17876.0703	147.7361		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	52.0032	11.5181	4.5149	0.0000	29.2000	74.8063
Bedrooms	1.9022	1.9023	0.9999	0.3193	-1.8639	5.6682
Bathrooms	6.8269	2.5628	2.6638	0.0088	1.7532	11.9007
Nbhd_1	-34.0837	3.1690	-10.7554	0.0000	-40.3576	-27.8099
Nbhd_2	-29.2180	2.8637	-10.2030	0.0000	-34.8874	-23.5486
sizethou	35.9304	6.4044	5.6102	0.0000	23.2511	48.6097
Brick_Yes	18.5078	2.3963	7.7235	0.0000	13.7637	23.2519

 $\pm 2 \hat{\sigma} = 24.$ with just size it was +/- 45 !!!

Let's try dropping bedrooms.

Results of multiple regression for pricethou

Summary measures

Multiple R	0.8963
R-Square	0.8034
Adj R-Square	0.7954
StErr of Est	12.1547

ANOVA Table

Source	df	SS	MS	F	p-value
Explained	5	73661.4233	14732.2847	99.7203	0.0000
Unexplained	122	18023.7910	147.7360		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	53.6295	11.4027	4.7032	0.0000	31.0567	76.2023
Bathrooms	7.2304	2.5308	2.8569	0.0050	2.2204	12.2405
Nbhd_1	-35.3137	2.9205	-12.0916	0.0000	-41.0952	-29.5322
Nbhd_2	-30.1452	2.7094	-11.1262	0.0000	-35.5087	-24.7817
sizethou	37.9050	6.0924	6.2217	0.0000	25.8445	49.9656
Brick_Yes	18.3121	2.3883	7.6674	0.0000	13.5843	23.0400

 $\hat{\sigma}$ (and R-squared) is about the same without bedrooms.

Given our data,

If you were trying to predict the price of a house, would you pay money to know the number of bedrooms ??

If you were trying to predict the price of a house, and you already know the number of bathrooms, neighborhood, size, and whether it is brick,

would you pay money to know the number of bedrooms ??

Here is the regression for price on bedrooms. Call: lm(formula = price ~ nbed, data = ddf) Residuals: Min 10 Median 30 Max -48.671 -14.496 0.462 13.178 61.763 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 71.575 8.718 8.210 2.24e-13 *** nbed 19.465 2.804 6.941 1.83e-10 *** ___ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 22.94 on 126 degrees of freedom Multiple R-squared: 0.2766, Adjusted R-squared: 0.2709 F-statistic: 48.18 on 1 and 126 DF, p-value: 1.83e-10

Bedrooms has information about price, *but*, if you already have the information in the other variables, it may not have much *additional* information given what we can learn from n = 128 houses.

Regression finds a linear combination of the variables that is like y.



price vs size:

price vs combination of size, nbath, brick, nbhd



The residuals are the part of y not related to the x's.



In general to add a categorical x, you can create dummies, one for each possible category (or level as we sometimes call it).

Use all but one of the dummies.

It does not matter which one you drop for the fit, but the interpretation of the coefficients will depend on which one you choose to drop.

6. The SLR Slope and MLR sigmahat

Using the fact that the resids are uncorrelated with x we can derive the SLR slope estimate forumula.

$$\hat{y}_{2} = \hat{\beta}_{0} + \hat{\beta}_{1} x_{2} = [-\frac{1}{3} - \hat{\beta}_{1} = -\frac{1}{3} + \hat{\beta}_{1} x_{2};$$

 $e_{1} - e_{2} = e_{1} = y_{2} - [-\frac{1}{3} - \hat{\beta}_{1} = +\hat{\beta}_{1} x_{2}]$
 $= [-\frac{1}{3} - -\frac{1}{3}] - \hat{\beta}_{1} (x_{1} - x_{2})$

$$O = \sum (x_{1} - \overline{x})(y_{1} - \overline{y}) - \hat{\beta}_{1} \sum (x_{1} - \overline{x})^{2}$$

$$= \sum (x_{1} - \overline{x})(y_{1} - \overline{y}) - \hat{\beta}_{1} \sum (x_{1} - \overline{x})^{2}$$

$$\Rightarrow \vec{\beta} = \frac{z(x_1 - \bar{z})(y_1 - \bar{y})}{\overline{z(x_1 - \bar{z})^2}}$$

60

Let's also note the estimate of σ^2 in MLR.

61