Regression Analysis

The Gotham City Motor Pool

Gotham City maintains a fleet of automobiles in a special motor pool. These cars are used by the various city agencies when special needs arise for temporary use by personnel who do not have their own assigned automobiles.

The manager of the pool is trying to determine what elements contribute to the maintenance and repair costs of the cars under his care. He surmises that components such as the mileage driven during a year, the age of the car, and possibly even the make of the car might help explain why annual maintenance and repair costs vary. Accordingly, he collects data on these variables for fifteen cars selected at random from the cars in the pool.

"Costs" refers to total maintenance and repair expenses for a car over the past 12 months.

"Mileage" is thousands (000s) of miles driven over the past 12 months.

"Age" is the car's age (in years) at the beginning of the past 12 months.

Conveniently, all of the cars in the fleet are either Fords or Hondas; this is encoded as "Make" = 0 for Fords, and "Make" = 1 for Hondas.

Motor Pool Data

Costs	Mileage	Age	Make (Ford=0, Honda=1)
$643	18.2	0	0
$613	16.4	0	1
$673	20.1	0	0
$531	8.4	1	1
$518	9.6	2	1
$594	12.1	1	0
$722	16.9	1	1
$861	21.0	1	0
$842	24.6	0	0
$706	19.1	1	0
$795	14.3	2	1
$776	16.5	2	1
$815	18.2	2	0
$571	12.7	2	0
$673	17.5	0	1

Looking at One Variable at a Time …

	A	B	C	D	E	F	G
1		Univariate statistics					
2			Costs	Mileage	Age	Make	
3		mean	688.866667	16.3733333	1	0.46666667	
4		standard deviation	111.678663	4.34370919	0.84515425	0.51639778	
5		standard error of the mean	28.8353068	1.12154089	0.21821789	0.13333333	
6							
7		minimum	518	8.4	0	0	
8		median	673	16.9	1	0	
9		maximum	861	24.6	2	1	
10		range	343	16.2	2	1	
11							
12		skewness	0.038	-0.214	0.000	0.149	
13		kurtosis	-1.189	-0.068	-1.615	-2.308	
14							
15		number of observations	15				
16							
17		t-statistic for computing					
18		95%-confidence intervals	2.1448				

	estimate ± margin of error
95%-confidence interval for mean annual cost per car (across the fleet)	$688.87 ± 2.1448·$28.84
95%-confidence interval for mean annual miles driven per car	16,373 ± 2.1448·1,121
95%-confidence interval for mean age of cars in the fleet	1.000 ± 2.1448·0.218
95%-confidence interval for fraction of fleet for which Make = Honda	46.67% ± 2.1448·13.33%

(Two-Dimensional) Tables

As a first step in his analysis of these data, the manager calculates the average maintenance and repair costs for new, one-year-old, and two-year-old cars. The results are:

- Although he is somewhat surprised by the results, the manager concludes that the age of the car does not significantly influence the repair and maintenance costs.

Age	Number of Cars	Average Maintenance and Repair Costs
0	5	$688.80
1	5	$682.80
2	5	$695.00

As a next step, the manager calculates the average costs for each make of car. The results are:

- He concludes that he should, in the future, give preference to purchasing Hondas since he would save $52 each per year in maintenance and repairs.

Make	Number of Cars	Average Maintenance and Repair Costs
Ford	8	$713.13
Honda	7	$661.14

Do you agree with the manager? How would you suggest that he analyze the data? What are your conclusions?

(Two-Dimensional) Tables Show Only Shadows

Not surprisingly, the new cars are more popular with city employees, and get taken out more often (and consequently are driven further) than the older cars during the year.

Age	Number of Cars	Average Maintenance and Repair Costs	Average Mileage
0	5	$688.80	19.200
1	5	$682.80	16.780
2	5	$695.00	13.140

And, for whatever reason, the Fords are more popular, and end up being driven further than the Hondas during the year.

Make	Number of Cars	Average Maintenance and Repair Costs	Average Mileage
Ford	8	$713.13	18.250
Honda	7	$661.14	14.220

We need a way to look at all the dimensions of a relationship at the same time. This is what regression analysis can do!

For example, we'll see (via regression) that Fords *of the same age, and driven the same distance,* as Hondas have maintenance and repair costs, on average, $47 ***less*** per year.

The Regression "Machine"

The regression model
(terminology, structure,
and assumptions)

relevant sample data
(no missing data)

regression results
(today's focus)

The Regression Model

$$\text{Costs} = \alpha + \beta_1 \cdot \text{Mileage} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Make} + \varepsilon$$

the
dependent
variable

explanatory variables
or
independent variables

residual term

coefficients

linear
mathematical structure

What we're talking about is sometimes explicitly called "linear regression analysis," since it assumes that the underlying relationship is linear (i.e., a straight line in two dimensions, a plane in three, and so on)!

Why Spend All This Time on Such a Limited Tool?

- Some interesting relationships *are* linear.
- All relationship are locally linear!
- Several of the most commonly encountered nonlinear relationships in management can be translated into linear relationships, studied using regression analysis, and the results then untranslated back to the original problem!

A Few Final Assumptions Concerning ε

- The validity of regression analysis depends on several assumptions concerning the residual term.
  - $E[\varepsilon] = 0$. This is purely a cosmetic assumption. The estimate of α will include any on-average residual effects which are different from zero.
  - ε varies normally across the population. While a substantive assumption, this is typically true, due to the Central Limit Theorem, since the residual term is the total of a myriad of other, unidentified explanatory variables. If this assumption is not correct, all statements regarding confidence intervals for individual predictions might be invalid.
- The following additional assumptions will be discussed later in the course.
  - StdDev[ε] does not vary with the values of the explanatory variables. (This is called the *homoskedasticity* assumption.) Again, if this assumption is not correct, all statements regarding confidence intervals for individual predictions might be invalid.
  - ε is uncorrelated with the explanatory variables of the model. The regression analysis will "attribute" as much of the variation in the dependent variable as it can to the explanatory variables. If some unidentified factor covaries with one of the explanatory variables, the estimate of that explanatory variable's coefficient (i.e., the estimate of its effect in the relationship) will suffer from "specification bias," since the explanatory variable will have both its own effect, and some of the effect of the unidentified variable, attributed to it. This is why, when doing a regression for the purpose of estimating the effect of some explanatory variable on the dependent variable, we try to work with the most "complete" model possible.

What Can Regression Analysis Do for You?

- Make predictions (based on available information)
- Estimate group means (for similar individuals)
- Measure effects (while controlling for other influences)

- Help evaluate/improve a model (of a relationship)

1. Make a Prediction

For an individual, predict the value of the dependent variable, given the values of some of the explanatory variables.

- Process: "Regress" the dependent variable onto the given explanatory variables. Then "Predict." Fill in the values of the explanatory variables. Hit the "Predict" button.
- Answer: (prediction) ± (~2)·(standard error of prediction)

- Example: Predict the maintenance cost of a one-year-old car (in the fleet) which will be driven 18,500 miles over the next year.

Make a Prediction

- Predict the annual maintenance cost of a one-year-old car (in the fleet) which will be driven 18,500 miles over the next year.
- Process: "Regress" Cost onto Mileage and Age. Then "Predict", filling in 18.5 for Mileage and 1 for Age. Hit the "Predict" button.
  - (prediction) ± (~2)·(standard error of prediction)
  - $745.60 ± 2.1788·$54.56 (we're 95%-confident that the prediction is within ±$118.88 of actual Costs, i.e., actual Costs end up somewhere between $626.74 and $864.47)

Regression: Costs			
	constant	Mileage	Age
coefficient	180.914993	26.678792	71.1309195
std error of coef	73.2707488	3.7041353	19.0375739
t-ratio	2.4691	7.2024	3.7363
significance	2.9541%	0.0011%	0.2841%
beta-weight		1.0377	0.5383
standard error of regression		52.2695799	
coefficient of determination		81.22%	
adjusted coef of determination		78.09%	
number of observations		15	
residual degrees of freedom		12	
t-statistic for computing			
95%-confidence intervals		2.1788	

Prediction, using most-recent regression			Make multiple predictions
	constant	Mileage	Age
coefficients	180.915	26.67879	71.13092
values for prediction		18.5	1
predicted value of Costs		745.6036	Predict
standard error of prediction		54.55551	
standard error of regression		52.26958	
standard error of estimated mean		15.62674	
confidence level	95.00%		
t-statistic	2.1788		
residual degr. freedom	12		
confidence limits	lower	626.7373	
for prediction	upper	864.4698	
confidence limits	lower	711.5558	
for estimated mean	upper	779.6513	

2. Estimate a Group Mean

For a group of similar individuals (i.e., individuals with the same values for several explanatory variables), estimate the mean value of the dependent variable.

- Process: "Regress" the dependent variable onto the given explanatory variables. Then "Predict." Fill in the values of the explanatory variables. Hit the "Predict" button.
- Answer: (prediction) ± (~2)·(standard error of estimated mean)

- Example: Estimate the mean annual maintenance cost of two-year-old Fords (*note the plural!*) in the fleet.

Estimate a Group Mean

- Estimate the mean annual maintenance cost of two-year-old Fords (*note the plural!*) in the fleet.
- Process: "Regress" Cost onto Age and Make. Then "Predict", filling in 2 for Age and 0 for Make. Hit the "Predict" button.
  - (prediction) ± (~2)·(standard error of estimated mean)
  - $722.72 ± 2.1788·$59.01

Regression: Costs			
	constant	Age	Make
coefficient	705.66422	8.5266055	-54.266055
std error of coef	52.7288569	37.4525058	61.2960511
t-ratio	13.3829	0.2277	-0.8853
significance	0.0000%	82.3740%	39.3384%
beta-weight		0.0645	-0.2509
standard error of regression		116.83827	
coefficient of determination		6.18%	
adjusted coef of determination		-9.45%	
number of observations		15	
residual degrees of freedom		12	
t-statistic for computing			
95%-confidence intervals		2.1788	

Prediction, using most-recent regression			Make multiple predictions
	constant	Age	Make
coefficients	705.6642	8.526606	-54.2661
values for prediction		2	0
predicted value of Costs		722.7174	Predict
standard error of prediction		130.8925	
standard error of regression		116.8383	
standard error of estimated mean		59.00574	
confidence level	95.00%		
t-statistic	2.1788		
residual degr. freedom	12		
confidence limits	lower	437.5271	
for prediction	upper	1007.908	
confidence limits	lower	594.155	
for estimated mean	upper	851.2799	

Sources of Prediction-Related Error (a two-slide technical digression)

There are two ways in which our prediction (Y_{pred}) for an individual might differ from reality (Y):

reality $\quad Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$

the prediction equation $\quad Y_{pred} = a + b_1 X_1 + \ldots + b_k X_k + 0$

standard error of the prediction		standard error of the estimated mean		standard error of the regression, StdDev(ε)
this	= $\sqrt{\quad}$	this^2	+	this^2

The Standard Error of the Regression

- Using the prediction equation, we predict for each sample observation.
- The difference between the prediction and the actual value of the dependent variable (i.e., the error) is an estimate of that individual's residual.
- StdDev(ε) is estimated from these.

Predicted values and residuals

Costs	predicted	residual	Mileage	Age	Make
643	646.9291	-3.9291	18.2	0	0
613	640.9970	-27.9970	16.4	0	1
673	703.2598	-30.2598	20.1	0	0
531	477.7736	53.2264	8.4	1	1
518	587.3092	-69.3092	9.6	2	1
594	540.0364	53.9636	12.1	1	0
722	729.7791	-7.7791	16.9	1	1
861	803.9010	57.0990	21.0	1	0
842	836.6744	5.3256	24.6	0	0
706	747.5703	-41.5703	19.1	1	0
795	726.6534	68.3466	14.3	2	1
776	791.8783	-15.8783	16.5	2	1
815	794.8457	20.1543	18.2	2	0
571	631.7833	-60.7833	12.7	2	0
673	673.6095	-0.6095	17.5	0	1

Indeed, the regression "process" simply finds the coefficient estimates which minimize the standard error of the regression (or equivalently, which minimize the sum of the squared residuals)!

3. Measure a Pure Effect

A one-unit difference in an explanatory variable, when everything else of relevance remains the same, is typically associated with how large a difference in the dependent variable?

- Process: "Regress" the dependent variable onto *all* of the relevant explanatory variables (i.e., use the "most complete" model available).
- Answer: (coefficient of explanatory variable)

  ± (~2)·(standard error of coefficient)

- Example: Estimate the "pure" impact of 1,000 miles of driving during the year on annual maintenance costs.

Measure a Pure Effect

- Estimate the "pure" impact of 1,000 miles of driving during the year on annual maintenance costs.
- Process: "Regress" Costs onto Mileage, Age, and Make (i.e., use the "most complete" model available, so Age and Make can be held constant).
  - (coefficient of explanatory variable) ± (~2)·(standard error of coefficient)
  - $29.65 ± 2.2010·$3.92

Regression: Costs				
	constant	Mileage	Age	Make
coefficient	107.340945	29.6477024	73.9582688	47.4337242
std error of coef	82.0422871	3.91510733	17.9148891	28.9836595
t-ratio	1.3084	7.5726	4.1283	1.6366
significance	21.7429%	0.0011%	0.1677%	12.9983%
beta-weight		1.1531	0.5597	0.2193
standard error of regression		48.9578919		
coefficient of determination		84.90%		
adjusted coef of determination		80.78%		
number of observations		15		
residual degrees of freedom		11		
t-statistic for computing				
95%-confidence intervals		2.2010		

Predictions, using most-recent regression							
		Make single prediction					
Predict							
	coefficients	values for prediction					
constant	107.340945						
Mileage	29.6477024	13	14	16	17	10	11
Age	73.9582688	0	0	1	1	2	2
Make	47.4337242	0	0	0	0	1	1
predicted value of Costs		492.7611	522.4088	655.6625	685.3102	599.1682	628.8159
standard error of prediction		60.62579	58.9124	52.53636	52.10449	54.80818	54.29203
standard error of regression		48.95789	48.95789	48.95789	48.95789	48.95789	48.95789
standard error of estimated mean		35.75768	32.76881	19.05765	17.83263	24.63861	23.46805
		new Ford		1-year-old Ford		2-year-old Honda	
		increase	29.6477	increase	29.6477	increase	29.6477

4. Modelling: Why Does The Dependent Variable Vary?

What fraction of the variance in the dependent variable (from one individual to another) can potentially be explained by the fact that the explanatory variables in our model vary (from one individual to another)?

- Process: "Regress" the dependent variable onto the explanatory variables currently under consideration. Then look at the "adjusted coefficient of determination" (synonymously, "the adjusted r-squared").

- Example: Why does annual maintenance expense vary across the cars in the fleet?

The Adjusted Coefficient of Determination

- Why does annual maintenance expense vary across the cars in the fleet?
- Partial Answer: Because Mileage, Age, and Make vary across the fleet.
- How much of an answer is that? Regress Costs onto those three variables.

Regression: Costs				
	constant	Mileage	Age	Make
coefficient	107.340945	29.6477024	73.9582688	47.4337242
std error of coef	82.0422871	3.91510733	17.9148891	28.9836595
t-ratio	1.3084	7.5726	4.1283	1.6366
significance	21.7429%	0.0011%	0.1677%	12.9983%
beta-weight		1.1531	0.5597	0.2193
standard error of regression		48.9578919		
coefficient of determination		84.90%		
adjusted coef of determination		80.78%		
number of observations		15		
residual degrees of freedom		11		
t-statistic for computing				
95%-confidence intervals		2.2010		

Variation in Mileage, Age, and Make can potentially explain 80.78% of the variation in Costs. The other 19.22% of the variation *must* be explained by other unidentified variables still lumped together in the residual term.

Why Does the Dependent Variable Vary?

- Why does annual maintenance expense vary across the cars in the fleet?
  - Variation in Mileage alone can explain 56.26% of the variation in Costs.
  - Variation in Age alone can explain nothing (read "0%" for negative values*).
  - Variation in the two together can explain 78.09% of the variation in Costs.

Regression: Costs		
	constant	Mileage
coefficient	364.476942	19.812076
std error of coef	76.8173302	4.54471998
t-ratio	4.7447	4.3594
significance	0.0383%	0.0774%
beta-weight		0.7706
standard error of regression		73.8638412
coefficient of determination		59.38%
adjusted coef of determination		56.26%
number of observations		15
residual degrees of freedom		13
t-statistic for computing		
95%-confidence intervals		2.1604

Regression: Costs		
	constant	Age
coefficient	685.766667	3.1
std error of coef	47.3006794	36.6389487
t-ratio	14.4980	0.0846
significance	0.0000%	93.3861%
beta-weight		0.0235
standard error of regression		115.862529
coefficient of determination		0.06%
adjusted coef of determination		-7.63%
number of observations		15
residual degrees of freedom		13
t-statistic for computing		
95%-confidence intervals		2.1604

Regression: Costs			
	constant	Mileage	Age
coefficient	180.914993	26.678792	71.1309195
std error of coef	73.2707488	3.7041353	19.0375739
t-ratio	2.4691	7.2024	3.7363
significance	2.9541%	0.0011%	0.2841%
beta-weight		1.0377	0.5383
standard error of regression		52.2695799	
coefficient of determination		81.22%	
adjusted coef of determination		78.09%	
number of observations		15	
residual degrees of freedom		12	
t-statistic for computing			
95%-confidence intervals		2.1788	

The Explanatory Power of the Model

- Names can vary: The {adjusted, corrected, unbiased} {coefficient of determination, r-squared} all refer to the same thing.
  - Without an adjective, the {coefficient of determination, r-squared} refers to a number slightly larger than the "correct" number, and is a throwback to pre-computer days.
- When a new variable is added to a model, which actually contributes nothing to that model (i.e., its true coefficient is 0), the adjusted coefficient of determination will, on average, remain unchanged.
  - Depending on chance, it might go up or down a bit.
  - *If negative, interpret it as 0%.
  - The thing without the adjective will always go up. That's obviously not quite "right."

5. Modelling: Relative Explanatory Importance

Considering a set of explanatory variables, rank them in order of importance in helping to explain why the dependent variable varies.

- Process: "Regress" the dependent variable onto the target set of explanatory variables. Rank the variables in order of the absolute values of their beta-weights, from largest (most important in explaining variation in the dependent variable) to smallest (least important in explaining variation).

- Example: Rank Mileage, Age, and Make in order of relative importance in helping to explain why Costs vary across the fleet.

Regression Analysis

Variation in What Most Helps to Explain Why the Dependent Variable Varies?

- Rank Mileage, Age, and Make in order of relative importance in helping to explain why Costs vary across the fleet.
  - One standard-deviation's-worth of variation in Mileage is associated with 1.1531 standard-deviation's-worth of variation in Costs.
  - One standard-deviation's-worth of variation in Age is associated with 0.5567 standard-deviation's-worth of variation in Costs.
  - One standard-deviation's-worth of variation in Make is associated with 0.2193 standard-deviation's-worth of variation in Costs.

Regression: Costs				
	constant	Mileage	Age	Make
coefficient	107.340945	29.6477024	73.9582688	47.4337242
std error of coef	82.0422871	3.91510733	17.9148891	28.9836595
t-ratio	1.3084	7.5726	4.1283	1.6366
significance	21.7429%	0.0011%	0.1677%	12.9983%
beta-weight		1.1531	0.5597	0.2193
standard error of regression		48.9578919		
coefficient of determination		84.90%		
adjusted coef of determination		80.78%		
number of observations		15		
residual degrees of freedom		11		
t-statistic for computing				
95%-confidence intervals		2.2010		

Therefore, variation in Mileage is more than twice as important as is variation in Age (and more than five times as important as is the fact that some cars in the fleet are Fords, and others are Hondas) in helping to explain why Costs vary (when all three of those variables are considered together).

The Beta-Weights

- You can't compare regression coefficients directly, since they may carry different dimensions.
- The beta-weights are dimensionless, and combine how much each explanatory variable varies, with how much that variability leads to variation in the dependent variable.
  - Specifically, they are the product of each explanatory variable's standard deviation (how much it varies) and its coefficient (how much its variation affects the dependent variable), divided by the standard deviation of the dependent variable (just to remove all dimensionality).

6. Modelling: Which Variables "Belong" in the (Current) Model?

How strong is the evidence that each explanatory variable has a non-zero coefficient (i.e., plays a predictive role) in the current model?

- Process: "Regress" the dependent variable onto the (current) set of explanatory variables. For each explanatory variable, examine the "significance level" (synonymously, the "p-value") of the sample data with respect to the null hypothesis that the true coefficient of that variable is zero.
  - The closer the significance level is to 0%, the stronger is the evidence against that null hypothesis (i.e., the stronger is the evidence that this variable does indeed belong in the current model).
- Example: How strong is the evidence that Mileage, Age, and Make each "belong" in the model which predicts Costs from all three?

Which Variables "Belong" in the (Current) Model?

- How strong is the evidence that Mileage, Age, and Make each "belong" in the model which predicts Costs from all three?
  - For Mileage and Age, "overwhelmingly strong".
  - For Make, a "little bit of supporting evidence", but not even "moderately strong".

Regression: Costs				
	constant	Mileage	Age	Make
coefficient	107.340945	29.6477024	73.9582688	47.4337242
std error of coef	82.0422871	3.91510733	17.9148891	28.9836595
t-ratio	1.3084	7.5726	4.1283	1.6366
significance	21.7429%	0.0011%	0.1677%	12.9983%
beta-weight		1.1531	0.5597	0.2193
standard error of regression		48.9578919		
coefficient of determination		84.90%		
adjusted coef of determination		80.78%		
number of observations		15		
residual degrees of freedom		11		
t-statistic for computing				
95%-confidence intervals		2.2010		

- With more data, if the true coefficient of Make is non-zero, the significance level will move towards 0%, and the evidence for inclusion will be stronger.
- With more data, if the true coefficient of Make is really zero, the significance level will stay well above 0%, and the estimate of the coefficient will move towards 0 (the "truth").

Regression Analysis: Summary

1. **Make a Prediction**
   For an individual, predict the value of the dependent variable, given the values of some of the explanatory variables.
   - Process: "Regress" the dependent variable onto the given explanatory variables. Then "Predict." Fill in the values of the explanatory variables. Hit the "Predict" button.
   - Answer: (prediction) ± (~2)·(standard error of prediction)

2. **Estimate a Group Mean**
   For a group of similar individuals (i.e., individuals with the same values for several explanatory variables), estimate the mean value of the dependent variable.
   - Process: "Regress" the dependent variable onto the given explanatory variables. Then "Predict." Fill in the values of the explanatory variables. Hit the "Predict" button.
   - Answer: (prediction) ± (~2)·(standard error of estimated mean)

3. **Measure a Pure Effect**
   A one-unit difference in an explanatory variable, when everything else of relevance remains the same, is typically associated with how large a difference in the dependent variable?
   - Process: "Regress" the dependent variable onto *all* of the relevant explanatory variables (i.e., use the "most complete" model available).
   - Answer: (coefficient of explanatory variable) ± (~2)·(standard error of coefficient)

4. **Modelling: Why Does the Dependent Variable Vary?**
   What fraction of the variance in the dependent variable (from one individual to another) can potentially be explained by the fact that the explanatory variables in our model vary (from one individual to another)?
   - Process: "Regress" the dependent variable onto the explanatory variables currently under consideration. Then look at the "adjusted coefficient of determination" (synonymously, "the adjusted r-squared").

5. **Modelling: Relative Explanatory Importance**
   Considering a set of explanatory variables, rank them in order of importance in helping to explain why the dependent variable varies.
   - Process: "Regress" the dependent variable onto the target set of explanatory variables. Rank the variables in order of the absolute values of their beta-weights, from largest (most important in explaining variation in the dependent variable) to smallest (least important in explaining variation).

6. **Modelling: Which Variables "Belong" in the (Current) Model?**
   How strong is the evidence that each explanatory variable has a non-zero coefficient (i.e., plays a predictive role) in the current model?
   - Process: "Regress" the dependent variable onto the (current) set of explanatory variables. For each explanatory variable, examine the "significance level" (synonymously, the "p-value") of the sample data with respect to the null hypothesis that the true coefficient of that variable is zero.
   - The closer the significance level is to 0%, the stronger is the evidence against that null hypothesis (i.e., the stronger is the evidence that this variable ***does*** indeed belong in the current model).