Modeling: Variable Selection

Request: "Estimate the annual maintenance costs attributable to annual mileage on a car. Dollars per thousand miles driven will suffice."

This sounds like a regression problem! Let's sample some cars, and look at their costs and mileage over the past year.

The Results

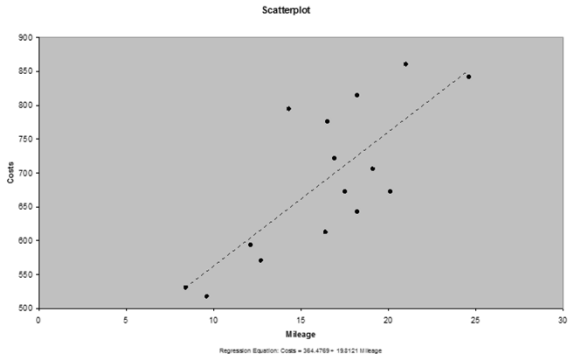
Regression: Costs

	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%

This all looks fine. And it's wrong!

Here's What the Computer Sees:



What it *doesn't* see is the age bias in the data: The cars to the left are mostly older cars, and the cars to the right are mostly newer.

An un(age)biased chart would have some lower points on the left, and some higher points on the right … and the regression line would be steeper.

Specification Bias

- … arises when you leave out of your model a potential explanatory variable that (1) has its own effect on the dependent variable, and (2) covaries systematically with an included explanatory variable.
- The included variable plays a double role, and its coefficient is a biased estimate of its pure effect.
- That's why, when we seek to estimate the pure effect of one explanatory variable on the dependent variable, we should use the most-complete model possible.

Seeing the Man Who isn't There

Yesterday, upon the stair,
I met a man who wasn't there
He wasn't there again today
I wish, I wish he'd go away...

Antigonish (1899), Hughes Mearns

When doing a regression study in order to estimate the pure effect of some variable on the dependent variable, the first challenge in the real (non-classroom) world is to decide for what variables to collect data. The "man who isn't there" can do you harm.

Let's return to the motorpool example, with Mileage as the only explanatory variable, and look at the residuals, i.e., the errors our current model makes in predicting for individuals in the sample.

Learning from our Mistakes

Take the "residuals" output

Costs	predicted	residual	Mileage
643	725.06	-82.06	18.2
613	689.39	-76.39	16.4
673	762.70	-89.70	20.1
531	530.90	0.10	8.4
518	554.67	-36.67	9.6
594	604.20	-10.20	12.1
722	699.30	22.70	16.9
861	780.53	80.47	21.0
842	851.85	-9.85	24.6
706	742.89	-36.89	19.1
795	647.79	147.21	14.3
776	691.38	84.62	16.5
815	725.06	89.94	18.2
571	616.09	-45.09	12.7
673	711.19	-38.19	17.5

Sort the observations from largest to smallest residual.

Costs	predicted	residual	Mileage
795	647.79	147.21	14.3
815	725.06	89.94	18.2
776	691.38	84.62	16.5
861	780.53	80.47	21.0
722	699.30	22.70	16.9
531	530.90	0.10	8.4
842	851.85	-9.85	24.6
594	604.20	-10.20	12.1
518	554.67	-36.67	9.6
706	742.89	-36.89	19.1
673	711.19	-38.19	17.5
571	616.09	-45.09	12.7
613	689.39	-76.39	16.4
643	725.06	-82.06	18.2
673	762.70	-89.70	20.1

Age
2
2
2
1
1
1
0
1
2
1
0
2
0
0
0

And see if something differentiates the observations near the top of the list from those near the bottom.

If so, consider adding that differentiating variable to your model!

We Can Do This Repeatedly

Our new model:

Regression: Costs			
	constant	Mileage	Age
coefficient	180.9150	26.6788	71.1309
std error of coef	73.2707	3.7041	19.0376
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		52.2696	
coefficient of determination		81.22%	
adjusted coef of determination		78.09%	

After sorting on the new residuals, 3 of the top 4 and 5 of the top 7 cars (those with the greatest positive residuals) are Hondas. 3 of the bottom 4 and 5 of the bottom 7 cars (those with the greatest negative residuals) are Fords.

This might suggest adding "make" as another new variable.

Make
1
1
0
1
0
1
1
0
0
1
0
0
0
1
0

Why Not Just Include the Kitchen Sink?

- Spurious correlation
  - The Dow, and women's skirts
- Collinearity
  - For example, age and odometer miles: Likely highly correlated
  - Computer can't decide what to attribute to each
    - Large standard errors of coefficients leads to large significance levels = no evidence either belongs.
    - But if either is included alone, strong evidence it belongs