Math 1710 Class 23

Regression, Experimentation
Dr. Back

Oct. 21, 2009

Major Changes in Your Next Homework Assignment

Please check the website.

Highlight of Airfare and Distance

- The slope b_1 of $\widehat{Fare} = b_1 Distance + b_0$ is the *average* increase in fare per extra mile.
- $\widehat{\text{Fare}} = 177 + .079 \cdot \text{Distance}$ and $\widehat{\text{Distance}} = -644 + 6.13 \cdot \text{Fare}$ are **different** lines!
- (Note $\frac{1}{.079} \neq 6.13$.)
- If you want to compute r on a TI-83/84, the place to look is stat $\rightarrow$ calc $\rightarrow$ linreg. And ONCE, you need to set DiagnosticsOn in the Catalog.

$Var(d_i) = (1 - r^2)Var(y_i)$

Sample Variance: $\mathrm{Var}(x_i) = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$

$Var(d_i) = (1 - r^2) Var(y_i)$

Sample Variance: $\mathrm{Var}(x_i) = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2$

For paired data (x_i, y_i),

Sample Covariance: $\mathrm{Cov}(x_i, y_i) = \frac{1}{n-1} \Sigma (x_i - \bar{x})(y_i - \bar{y})$

$Var(d_i) = (1 - r^2) Var(y_i)$

Sample Variance: $\mathrm{Var}(x_i) = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$

For paired data (x_i, y_i),

Sample Covariance: $\mathrm{Cov}(x_i, y_i) = \frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})$

$$\text{So } r = \frac{Cov(x_i, y_i)}{s_x s_y}$$

$Var(d_i) = (1 - r^2)Var(y_i)$

Sample Variance: $\mathrm{Var}(x_i) = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$

For paired data (x_i, y_i),

Sample Covariance: $\mathrm{Cov}(x_i, y_i) = \frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})$

$$\text{So } r = \frac{Cov(x_i, y_i)}{s_x s_y}$$

Because $(a+b)^2 = a^2 + 2ab + b^2$

$$\mathrm{Var}(\mathrm{a_i} + \mathrm{b_i}) = \mathrm{Var}(a_i) + \mathrm{Var}(b_i) + 2\mathrm{Cov}(a_i, b_i).$$

$Var(d_i) = (1 - r^2) Var(y_i)$

Because $(a+b)^2 = a^2 + 2ab + b^2$

$$\mathrm{Var}(\mathrm{a_i} + \mathrm{b_i}) = \mathrm{Var}(a_i) + \mathrm{Var}(b_i) + 2\mathrm{Cov}(a_i, b_i).$$

Apply this to:

$$\mathrm{Var}(d_i) = \mathrm{Var}(y_i - \hat{y}_i)$$

$Var(d_i) = (1 - r^2) Var(y_i)$

Apply this to:

$$\mathrm{Var}(d_i) = \mathrm{Var}(y_i - \hat{y}_i)$$

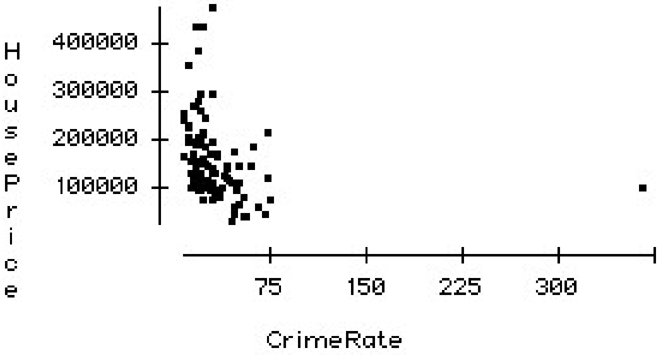
$$\begin{aligned}
\mathrm{Var}(y_i - (\bar{y} + b_1(x_i - \bar{x}))) =& \mathrm{Var}((y_i - \bar{y}) - b_1(x_i - \bar{x})) \\
=& s_y^2 + b_1^2 s_x^2 - 2b_1 \mathrm{Cov}(y_i - \bar{y}, x_i - \bar{x}) \\
=& s_y^2 + r^2 \frac{s_y^2}{s_x^2} s_x^2 - 2r \frac{s_y}{s_x}(r s_x s_y) \\
=& s_y^2 - r^2 s_y^2 \\
=& (1 - r^2) s_y^2.
\end{aligned}$$

Crime Rates by Locality

Crime Rates vs Housing Prices in Philadelphia 1996
Crime Rate is Crimes Per 1000
Housing Prices in Dollars

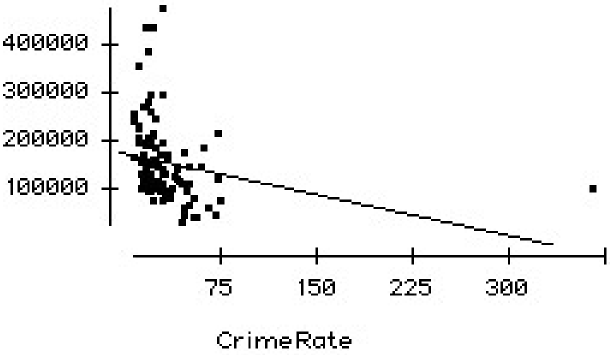
Crime Rates by Locality

scatterplot



Crime Rates by Locality

with regression line



$$\widehat{HP} = -577 \cdot CR + 177K \quad r^2 = .06$$

Crime Rates by Locality

regression display

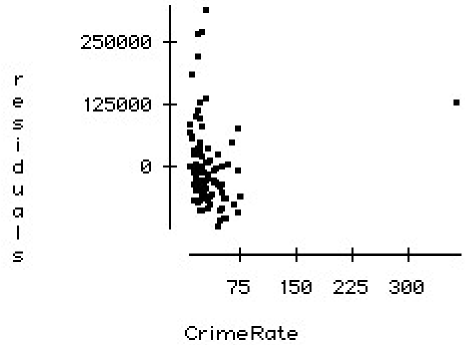
Dependent variable is: **HousePrice**
No Selector
111 total cases of which 12 are missing
R squared = 6.2%   R squared (adjusted) = 5.3%
s = 84.33e3 with 99 - 2 = 97 degrees of freedom

Source	**Sum of Squares**	**df**	**Mean Square**	**F-ratio**
Regression	45.9673e9	1	45.9673e9	6.46
Residual	689.739e9	97	7.11071e9	

Variable	**Coefficient**	**s.e. of Coeff**	**t-ratio**	**prob**
Constant	176629	11.25e3	15.7	≤ 0.0001
CrimeRate	-576.908	226.9	-2.54	0.0126

Crime Rates by Locality

Residuals



Crime Rates by Locality

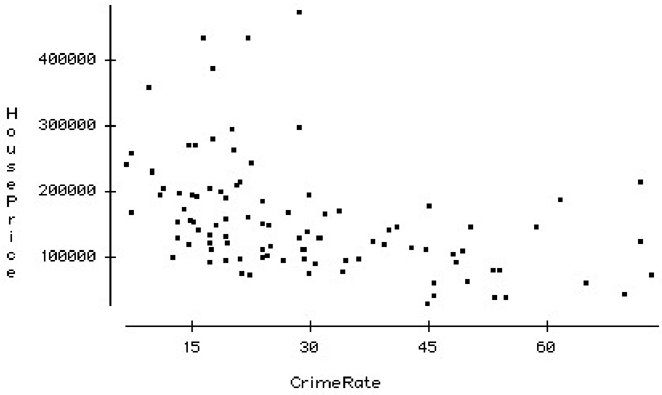
Now analyze without the Center City Outlier.

The point is we now get a very different regression line with a much higher r^2.

Outliers matter!

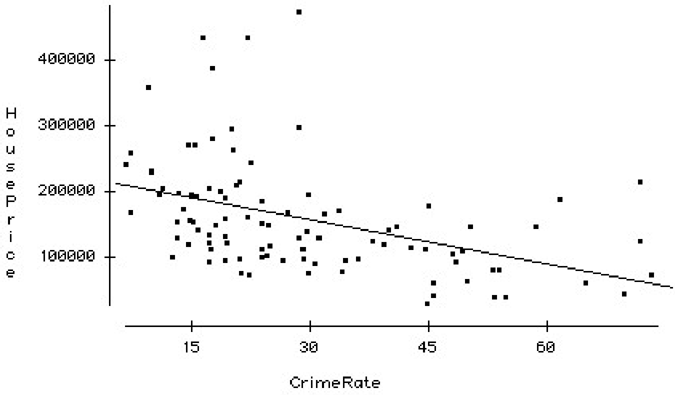
Crime Rates by Locality

scatterplot



Crime Rates by Locality

with regression line



$$\widehat{HP} = -2290 \cdot CR + 225K \quad r^2 = .18$$

Crime Rates by Locality

regression display

Dependent variable is:	**HousePrice**
cases selected according to	**Not Center City**

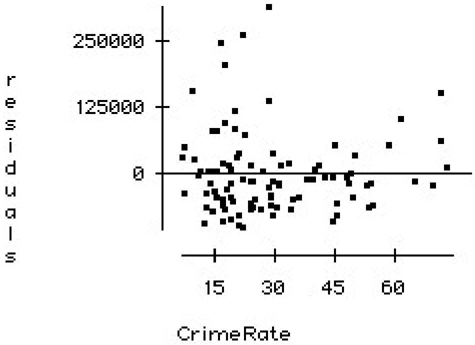
111 total cases of which 13 are missing
R squared = 18.4%   R squared (adjusted) = 17.6%
s = 7.886e4 with 98 - 2 = 96 degrees of freedom

Source	**Sum of Squares**	**df**	**Mean Square**	**F-ratio**
Regression	1.34831e11	1	1.34831e11	21.7
Residual	5.97038e11	96	6.21914e9	

Variable	**Coefficient**	**s.e. of Coeff**	**t-ratio**	**prob**
Constant	225234	1.64e4	13.7	≤ 0.0001
CrimeRate	-2288.69	491.5	-4.66	≤ 0.0001

Crime Rates by Locality

Residuals



Crime Rates by Locality

Now transform from CR to $\frac{1}{CR}$. The point here is that transforming your data to get a form more appropriate to the assumptions of simple linear regression is an important activity. We don't cover it, but chapter 10 of BVD is all about this. This is just 1 example to raise the possibility in your mind.

Crime Rates by Locality

Math 1710
Class 23

V1u

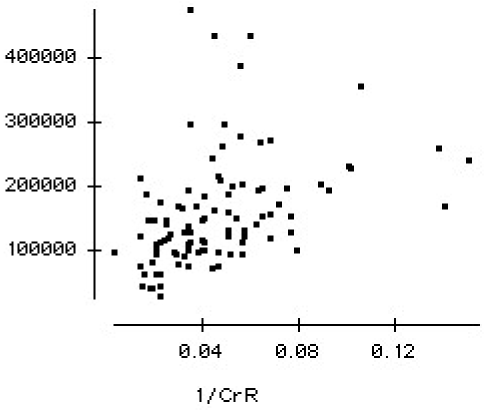
Admin

Airfare

Proof of $Var(d_i)$

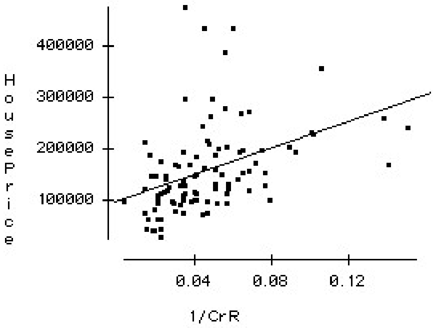
Philadelphia Crime Rates

scatterplot



Crime Rates by Locality

with regression line



$$\widehat{HP} = 1.3M \cdot \frac{1}{CR} + 97.9K \quad r^2 = .17$$

But Center City included.

Crime Rates by Locality

regression display

Dependent variable is: **HousePrice**
No Selector
111 total cases of which 12 are missing
R squared = 17.4% R squared (adjusted) = 16.5%
s = 79.16e3 with 99 - 2 = 97 degrees of freedom

Source	**Sum of Squares**	**df**	**Mean Square**	**F-ratio**
Regression	127.948e9	1	127.948e9	20.4
Residual	607.758e9	97	6.26555e9	

Variable	**Coefficient**	**s.e. of Coeff**	**t-ratio**	**prob**
Constant	97920.6	15.46e3	6.33	≤ 0.0001
1/CrR	1.30138e6	288e3	4.52	≤ 0.0001

Crime Rates by Locality

Residuals