

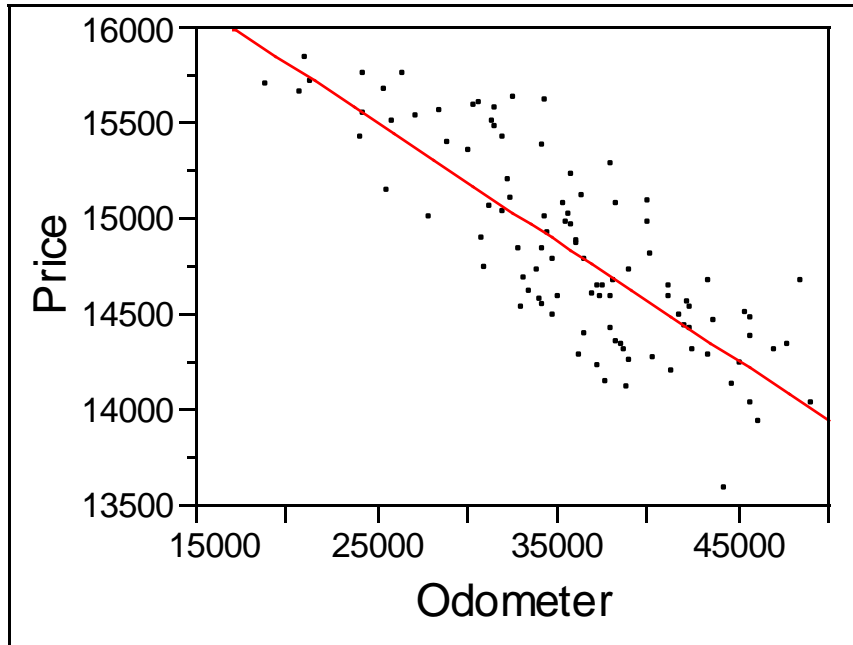
Lecture 6

Stat102, Summer 2008

- Estimating the conditional mean of y given x (Section 3.5.1), and CI^s
- Predicting an individual value of y given x (Section 3.5.2), and CI^s
- Some cautions in interpreting regression results (Section 3.7)

Example; Car Prices

- A used-car dealer wants to understand how odometer reading affects the selling price of used cars.
- The dealer randomly selects 100 three-year old Ford Tauruses that were sold at auction during the past month. Each car was in top condition and equipped with automatic transmission, AM/FM cassette tape player and air conditioning.
- carprices.JMP contains the price and number of miles on the odometer of each car.



Car Prices

Linear Fit

$$\text{Price} = 17067 - 0.0623 \text{ Odometer}$$

Summary of Fit

RSquare	0.650
Root Mean Square Error	$S_e = 303.1$
Mean of Response	$\bar{y} = 14823$
Observations	100

Moments

Mean	$\bar{x} = 36009$
Std Dev	$S_x = 6597$
Std Err Mean	659
N	100

Note: $S_x^2 = 6597^2 = 4352409$

Two prediction problems

- a) The used-car dealer has an opportunity to bid on a lot of cars offered by a rental company. The rental company has very many Ford Tauruses, all equipped with automatic transmission, air conditioning and AM/FM cassette tape players. **All of these cars** have about **40,000 miles** on the odometer. The dealer would like an estimate of the **average** selling price of **all of these cars**.
- b) The used-car dealer is about to bid on a 3-year old Ford Taurus equipped with automatic transmission, air conditioner and AM/FM cassette tape player and with 40,000 miles on the odometer. The dealer would like to predict the selling price of an **individual** car.

Prediction Problem (a): Prediction of the Mean for a Given x

- Goal is to estimate the conditional **mean** of selling price given odometer reading, $x_m=40,000$. *ie*, estimate $\mu_{Y|x_m}$ at $x_m=40,000$.
- The estimate itself is just $\hat{\mu}_{Y|x_m} = \hat{y} = b_0 + b_1 x_m$.
- For our example

$$\begin{aligned}\hat{\mu}_{Y|x_m} &= \hat{y} = 17067 - 0.0623x_m \\ &= 17067 - 0.0623 \times 40,000 \\ &= 14,575\end{aligned}$$

- Assuming the ideal simple linear regression model holds, and **if** σ_e^2 were known the variance of this estimate would be

$$\sigma_m^2 = \sigma_e^2 \left(\frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1)s_x^2} \right).$$

This follows from a theoretical calculation like those in the notes for Lectures 3 & 5.

- Hence the **standard error** is

$$s_m = s_e \sqrt{\frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1)s_x^2}} = \dots = 35.47.$$

- The sampling distribution of $\hat{\mu}_{Y|x_m}$ is normal
- The relevant t-statistic has $n - 2$ DF, since the estimate s_e used in the formula for s_m involves estimating the **two** coefficients b_0 & b_1 .
- The 95% CI for $\mu_{Y|x_m}$ at $x_m=40,000$ is

$$\hat{\mu}_{Y|x_m} \pm t_{\alpha/2;n-2} s_m = 14,575 \pm 1.98 \times 35.47 = 14575 \pm 70.$$

Prediction Problem (b):

Prediction of the value of **one new observation** at a given x

- Goal is to estimate the selling price of another car with a given odometer reading, $x_p=40,000$.
- The estimate itself is just $\hat{\mu}_{Y|x_p} = \hat{y} = b_0 + b_1x_p$.
- This is the **same estimate as** that for the conditional mean at x_p
- For our example, as before,

$$\begin{aligned}\hat{\mu}_{Y|x_p} &= \hat{y} = 17067 - 0.0623x_p \\ &= 17067 - 0.0623 \times 40,000 \quad . \\ &= 14,575\end{aligned}$$

- Y is normally distributed about its mean.

- Assuming the ideal simple linear regression model holds, the prediction (*individual*) standard error is

$$s_p = s_e \sqrt{\mathbf{1} + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{(n-1)s_x^2}} = \dots = 305.2.$$

This follows from theoretical calculations like those for s_m .

- Note that s_p is always bigger than s_m , and usually is MUCH BIGGER (due to the extra **1** in the formula); in fact

$$s_p^2 = s_e^2 + s_m^2$$

- The explanation is that this SE must allow for the variability of the individual car about $\mu_{Y|x_p}$, as well as the variability of $\hat{\mu}_{Y|x_p}$ about the true $\mu_{Y|x_p}$.
- The 95% Pred. interval at $x_p=40,000$ is

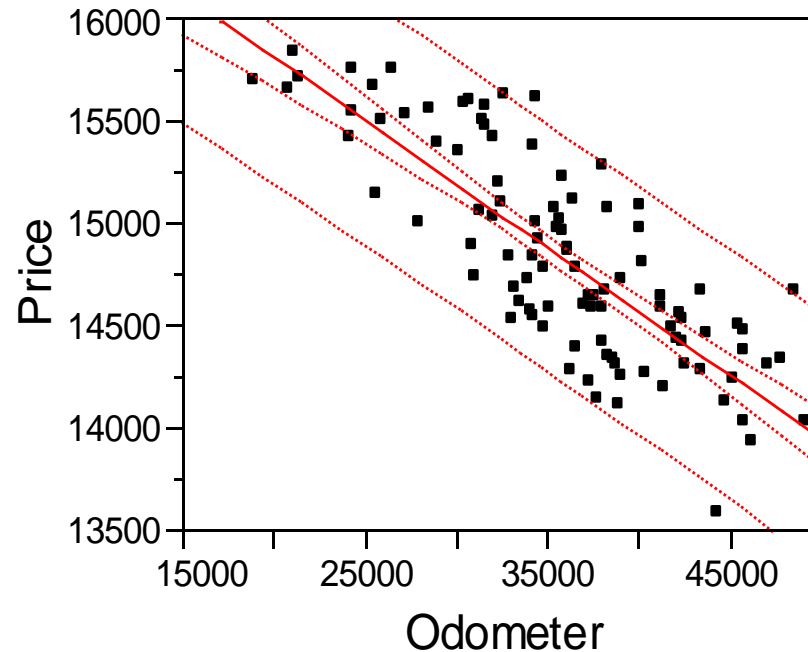
$$\hat{\mu}_{Y|x_p} \pm t_{\alpha/2; n-2} s_p = 14575 \pm 1.98 \times 305.2 = 14575 \pm 604.$$

Note: There are **$n-2$** DF. Also note again that this is wider than the CI for the mean.

Prediction Intervals and CIs for mean response in JMP

- After Fit Y by X, click red triangle next to Linear Fit.
- Clicking Confid Curves Fit displays the confidence intervals for the mean response.
- Clicking Confid Curves Indiv displays the prediction intervals.
- You can use the crosshair tool (under Tools) to *approximate* the confidence or prediction interval for a particular x value.
- OR calculate from the formulas
- Or go to Analyze>Fit Model (we'll describe this later)

CI's for the Mean and for Individual Predictions, in JMP



- The narrower bands correspond to the 95% CI^s for the mean.
- The wider bands are for the 95% Individual Prediction CI^s.
- Note that most of the points lie inside these bands. *WHY?*

You can see how the mean CI^s get wider; the individual CI^s also do, but imperceptibly.

Association vs. Causality

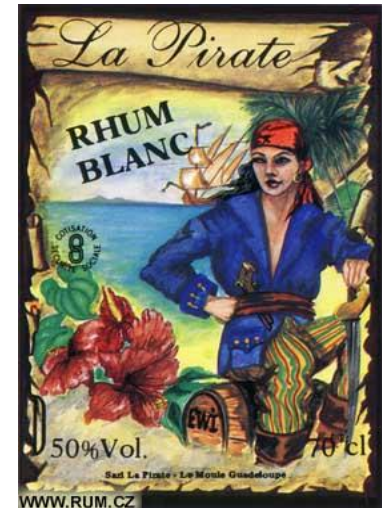
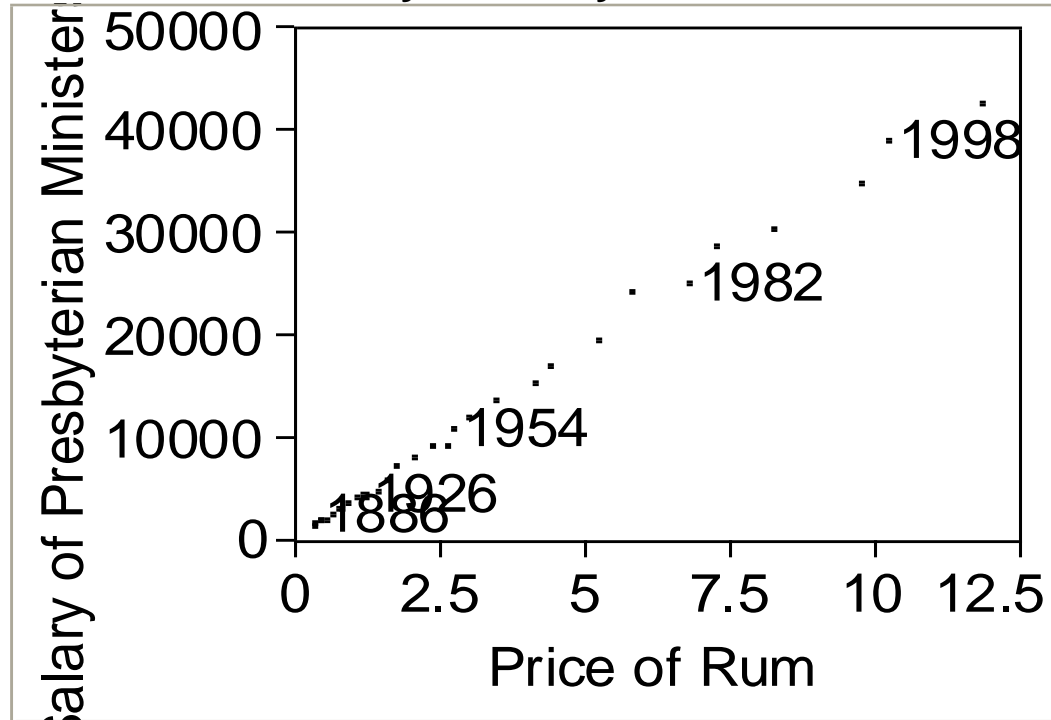
See Section 3.7.1

- A high R^2 means that x has a strong linear relationship with y – there is a strong association between x and y .
- Possible explanations for high R^2 :
 - X causes Y
 - Reverse is true.
 - Neither of the above, BUT due to lurking (confounding) variables related to both x and y which are the common cause of x and y

Example

- Over the last 100 years, the price of rum in Havana (x) has provided excellent predictions of the salary of Presbyterian ministers in Massachusetts (y).
- WHY?

Bivariate Fit of Salary of Presbyterian Ministers in MA By Price of Rum



Are the Presybterian ministers benefiting from the rum trade, or are they supporting it?

Rum and Presbyterian Ministers (cont)

- Over the last 100 years, the price of rum in Havana (x) has provided excellent predictions of the salary of Presbyterian ministers in Massachusetts (y).
- **WHY?**
 - Are the ministers benefiting from the rum trade? (change in x “causes” change in y)
 - Are they supporting it? (y “causes” x)
 - Are there confounding variable(s)?
 - z changes; and changing z causes change in both x and y

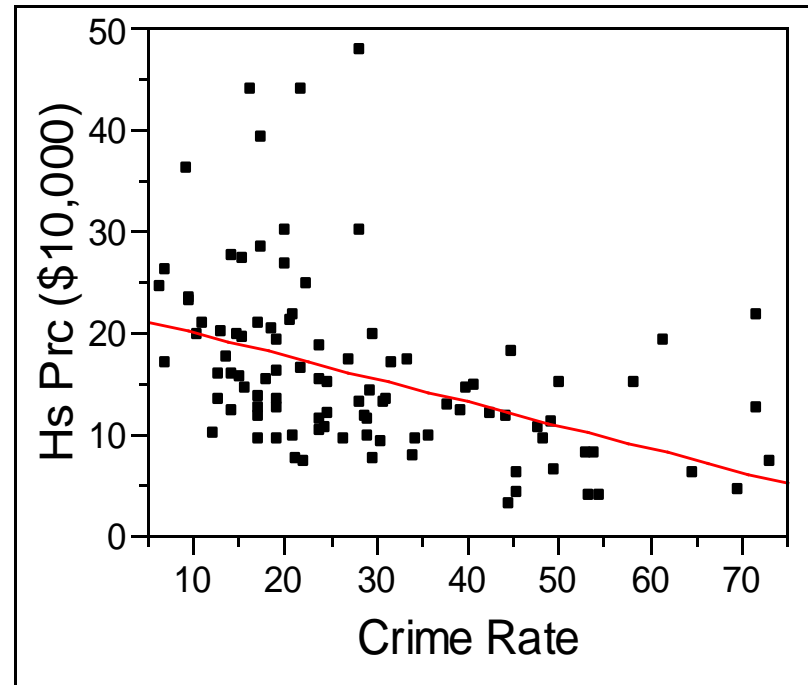
Example

To Illustrate Section 3.7.1 & 3.7.2

- A community in the Philadelphia area is interested in how crime rates affect property values.
- If low crime rates increase property values, the community may be able to cover the costs of increased police protection by gains in tax revenues from higher property values.
- Data on the average housing price (\$10000) and crime rate (per 1000 population) for communities in Pennsylvania near Philadelphia for 1996 are shown in Philadelphia_Crime_Rate.JMP.

Do Property Values Go Down as Crime Goes Up?

Plot Shows Avg House Price (\$10000) vs Crime rate (crimes/1000 pop)
in PA communities in the Philadelphia region



Note: One observation has been excluded from this plot **and** from the following analyses. It is for “Center City Philadelphia”.

Do Property Values Go Down as Crime Goes Up? (cont)

One data point excluded (for Center City Philadelphia)

Summary of Fit

RSquare	0.184
Root Mean Square Error	7.89
Mean of Response	15.85
Observations	98

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1348	1348	21.68
Error	96	5970	62.2	Prob > F
C. Total	97	7318		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	22.52	1.64	13.73	<.0001
Crime Rate	-0.229	0.0492	-4.66	<.0001

Cause and Effect

(Section 3.7.1)

- Can you deduce a cause-and-effect relationship from these data?
- What are other possible explanations for the association between housing prices and crime rate, other than that high crime rates cause low housing prices?

Extrapolation

See Section 3.7.2

- When constructing estimates of $\mu_{y|x_{new}}$ or predicting individual values of a dependent value based on x_{new} , caution must be used if x_{new} is outside the range of the observed x 's. The data does not provide information about whether the simple linear regression model continues to hold outside of the range of the observed x 's.
- **Example:** Does the simple linear regression model fit in this data provide a reasonably accurate prediction of the average house price in Center City?
- **Hint:** The CrimeRate in Center City is 366. [*Why is it so high?*] So the L-S prediction of House Price (\$10,000) is

$$22.52 - .229 \times 366 = -61.3$$

Typical Questions (Based on Philadelphia_crime_rate.jmp)

1. What proportion of the variability in a community's average House Price is explained by the linear regression on crime rate? (Give a numerical measure.)
2. What is the 90% CI for the slope? Does this “convincingly demonstrate” that Crime Rate is negatively related to House Price?
3. Based on this regression, what is the 95% CI for the mean of Y among communities having crime rate 30?
4. Based on this regression, what is a 95% interval for the average house price (y) in a single, given community having crime rate 30?

Harder Questions (requiring additional thought)

5. Consider the average community with crime rate 40 and the average community with crime rate 50. Find a 95% CI for the difference in their mean values of y . (!)

6. Based on this data can you find a 95% interval for the price of a given house in a community having crime rate 30? If so, what is it? If not, why not – and what additional data would enable you to find such an interval? (!)

7. For the data excluding Center City Philadelphia, does the ideal simple linear regression population model appear to hold? Discuss. And, if you think it may not hold, would you suggest any alternate ways of analyzing or interpreting the data for which the model would seem more appropriate? (!)