Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2003

Module 3

Inference about the SRM

Mini-Review: Inference for a Mean

An ideal setup for inference about a mean assumes normality,

$$y_1,\ldots,y_n \text{ iid} \sim N(\mu_y,\sigma_y^2)$$

where $\bar{y}$ is used to estimate $\mu_{y.}$

What is meant by the sampling distribution of $\bar{y}$?

"Astonishing Fact #1" The sampling distribution of $\bar{y}$ is

$$\bar{y} \sim N(\mu_y,\sigma_y^2/n)$$

The standard error of $\bar{y}$ is $SE(\bar{y}) = s_y/\sqrt{n}$

$\bar{y} \pm 2\,SE(\bar{y})$ are approx 95% CI limits for μ_y

For testing H$_0$: $\mu_y = c$ vs H$_1$: $\mu_y \neq c$, t ratio = $(\bar{y} - c)/SE(\bar{y})$

if $|t \text{ ratio}| > 2$ *or* p-value < .05 *or* 95% CI does not contain c, reject H$_0$ at the .05 level of significance.

Sampling Distributions in Regression

For data $(x_1, y_1),\ldots, (x_n, y_n)$ generated with the SRM,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,\ldots, n$$
$$\varepsilon_1,\ldots,\varepsilon_n \text{ iid} \sim N(0,\sigma_\varepsilon^2)$$

What is meant by the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$?

How could you use the simulation in utopia.jmp to generate these sampling distributions?

"Astonishing Fact #2" The sampling distribution[1] of $\hat{\beta}_1$

1) has mean $E(\hat{\beta}_1) = \beta_1$

2) has standard deviation $SD(\hat{\beta}_1) = \dfrac{\sigma_\varepsilon}{\sqrt{n}} \times \dfrac{1}{SD(x)}$

3a) is exactly normal

3b) is approximately normal even if the errors $\varepsilon_1,\ldots,\varepsilon_n$ are not normally distributed

Note: The sampling distribution of $\hat{\beta}_0$ has the same properties but with a slightly different formula for $SD(\hat{\beta}_0)$

[1] More precisely, this result refers to the sampling distribution when $y_1,\ldots,y_n$ vary, but the values of $x_1,\ldots,x_n$ are treated as fixed constants that are the same for every sample. Things work out similarly even if the x_i are random.

Inference about β_0 and β_1

Typically the intercept β_0 and slope β_1 are estimated by $\hat{\beta}_0$ and $\hat{\beta}_1$, and the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are given[2] by JMP.

We'll denote them by $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$.

Confidence Intervals

Approximate 95% CI's for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm 2SE(\hat{\beta}_0) \quad \text{and} \quad \hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

Hypothesis Tests

For H_0: $\beta_0 = c$ vs H_1: $\beta_0 \neq c$,, t ratio $= \dfrac{\hat{\beta}_0 - c}{SE(\hat{\beta}_0)}$

For H_0: $\beta_1 = c$ vs H_1: $\beta_1 \neq c$,, t ratio $= \dfrac{\hat{\beta}_1 - c}{SE(\hat{\beta}_1)}$

If $|t \text{ ratio}| > 2$ *or* p-value < .05 *or* 95% CI does not contain c, reject H_0 at the .05 level of significance.

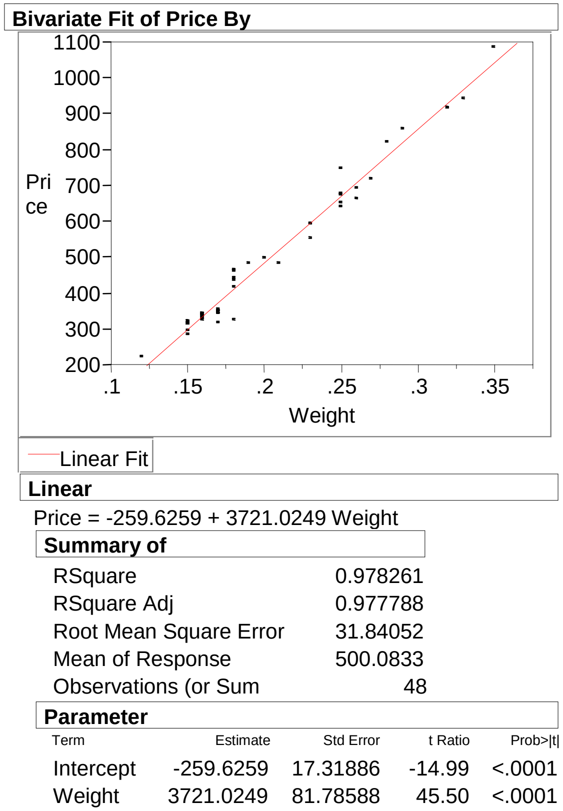
H_0: $\beta_1 = 0$ is the usual null hypothesis of interest. Why?

[2] These are obtained by simply substituting $RMSE$ for σ_ε in the formulas for the standard deviations of their sampling distributions.

JMP provides the details: estimates, SE's, t statistics, and p-values for inference about β_0 and β_1

Example

Consider inference for β_0 and β_1 in the diamond regression.



Bivariate Fit of Price By

Linear Fit

Linear

Price = -259.6259 + 3721.0249 Weight

Summary of

RSquare	0.978261
RSquare Adj	0.977788
Root Mean Square Error	31.84052
Mean of Response	500.0833
Observations (or Sum	48

Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-259.6259	17.31886	-14.99	<.0001
Weight	3721.0249	81.78588	45.50	<.0001

Here, β_0 and β_1 are estimated by

$$\hat{\beta}_0 \approx -259.6 \qquad \text{and} \qquad \hat{\beta}_1 \approx 3721.0$$

The standard errors of these estimates are

$$SE(\hat{\beta}_0) \approx 17.3 \qquad \text{and} \qquad SE(\hat{\beta}_1) \approx 81.8$$

Approximate 95% confidence interval limits for β_0 and β_1 are

$$-259.6 \pm 2\ (17.3) \qquad \text{and} \qquad 3721.0\ \pm 2\ (81.8)$$

Should the hypothesis H_0: $\beta_1 = 0$ be rejected?

Yes, because $t = 45.5 > 2$ or because p-value < .0001

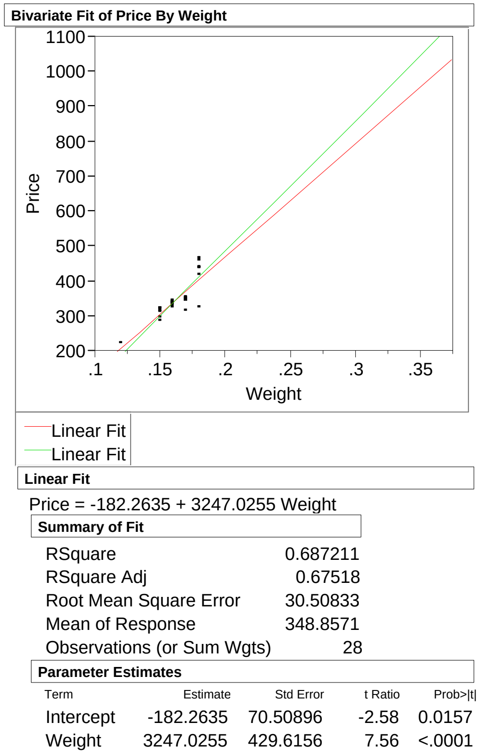
Should the hypothesis H_0: $\beta_1 = 3800$ be rejected?

No, because $|t| = |3721.0 - 3800|/81.8 = .96 < 2$

Why is it interesting that H_0: $\beta_0 = 0$ can be rejected?

Suppose that instead of the full diamond.jmp data set, we only had the 28 observations for which Weight ≤ .18.

The LS regression with these 28 observations yields



```
Linear Fit
Price = -182.2635 + 3247.0255 Weight
Summary of Fit
RSquare                        0.687211
RSquare Adj                     0.67518
Root Mean Square Error         30.50833
Mean of Response               348.8571
Observations (or Sum Wgts)           28
Parameter Estimates
Term          Estimate   Std Error   t Ratio   Prob>|t|
Intercept   -182.2635    70.50896     -2.58    0.0157
Weight      3247.0255    429.6156      7.56    <.0001
```

How has the regression output changed?

Confidence Intervals for the Regression Line

"Where does the true population regression line lie?"

> "What is the *average* price for *all* diamonds of a chosen weight?"

After running a regression based on $(x_1, y_1), \ldots, (x_n, y_n)$, each point $(x, \hat{y}_x)$ on the LS regression line

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

is an estimate of the corresponding point $(x, \mu_{y|x})$ on the true regression line

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

The estimate $\hat{y}_x$ is a statistic with a sampling distribution.

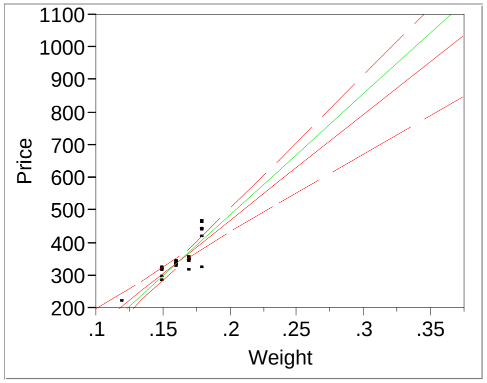
An astonishing fact again comes to the rescue – the sampling distribution of $\hat{y}_x$ is approximately normal with mean $\mu_{y|x}$ and a standard error $SE(\hat{y}_x)$ that can be computed.

An approximate 95% CI for $\mu_{y|x}$ is obtained as

$$\hat{y}_x \pm 2SE(\hat{y}_x)$$

JMP provides[3] a graph of the exact 95% CIs for $\mu_{y|x}$ over the whole line.

For the regression on the smaller diamond data, these confidence bands are seen to be



What happens to the 95% CI for the true regression line x as you get farther away from $\overline{x}$?

This phenomenon can be thought of as a "Statistical Extrapolation Penalty".[4]

Note that the LS line for the full data set is contained within the confidence bands. Why is this reasonable?

[3] After executing the Fit Line subcommand, right click next to "—Linear Fit" and select Confid Curves Fit from the Pop-up menu to obtain this plot.

[4] Even though the intervals widen as we extrapolate, as the output shows, this penalty is rather optimistic because it assumes that the model we have fit is correct for all values of x. If you price a big diamond with this model, you'll see that the interval is not nearly wide enough!

Predicting Individual Values with a Regression

"Where will a future value of the response y lie?"

"How much might I pay for a specific 1/4 carat diamond?"

After running a regression based on $(x_1, y_1), \ldots, (x_n, y_n)$, each point $(x, \hat{y}_x)$ on the LS regression line

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

is an estimate of the corresponding future point (x, y_x) generated by the SRM

$$y_x = \beta_0 + \beta_1 x + \varepsilon_x$$

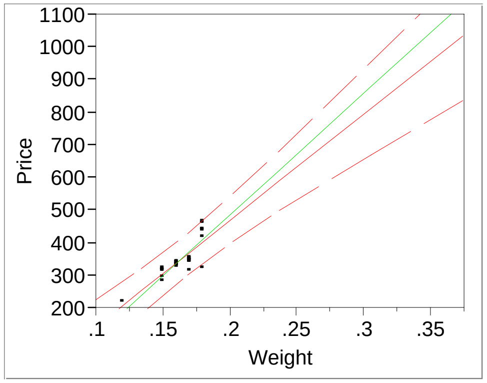
What's the difference between y_x and $\mu_{y|x}$ on pg 3-7?

To accommodate the extra variation of y_x due to ε_x, an approximate 95% prediction interval (PI) for y_x is obtained as

$$\hat{y}_x \pm 2\sqrt{SE(\hat{y}_x)^2 + RMSE^2}$$

This interval has the interpretation that

JMP provides[5] a graph of the exact 95% PIs for y_x over the whole line. For the regression on the smaller diamond data, these prediction bands are seen to be



These prediction bands are wider than the confidence bands for the true regression line on pg 3-8. Why is this reasonable?

Extrapolate with caution!

If x is not in the range of the data, predicting y_x is especially dangerous because the linear model may fail. Consider pricing the Hope Diamond (at 45.5 carats) with this model.

Another example: Average systolic blood pressure in people is well approximated by $y \approx 118 + .0043\, x^2$ for $20 \le x \le 60$ where y = blood pressure and x = age.

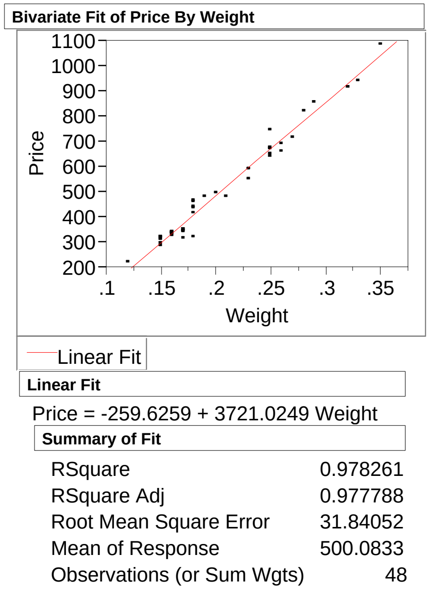
But when $x = 1000$, $y =$

[5] After executing the Fit Line subcommand, right click next to "—Linear Fit" and select Confid Curves Indiv from the Pop-up menu to obtain this plot.

R^2 Index of Performance

The next piece of output that we'll consider from the full diamond regression, is

RSquare = .978



Linear Fit

Price = -259.6259 + 3721.0249 Weight

Summary of Fit

RSquare	0.978261
RSquare Adj	0.977788
Root Mean Square Error	31.84052
Mean of Response	500.0833
Observations (or Sum Wgts)	48

This number is called R^2 and is widely interpreted as

"the proportion of variation explained by the regression"

The intuition behind R^2 is based on the decomposition[6]

$$\text{Response}_i = \text{Signal}_i + \text{Noise}_i$$

or

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

An Amazing Identity

When the signal comes from a regression, it turns out that

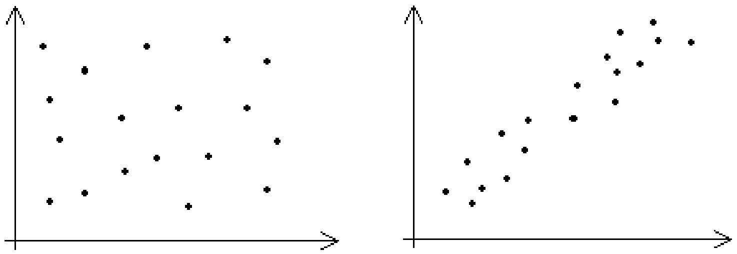
$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

This identity is often written as

$$\textit{Total SS} = \textit{Model SS} + \textit{Residual SS}$$

where SS reads "Sum of Squares" . Sometimes[7] the Residual SS is called the Error SS, but since it comes from the residuals, that name seems better.

What do these *SS* quantities measure?



[6] This language comes from electrical engineering. In that context, the signal might come from a radio or TV station. The unwanted things that contaminate your reception are called noise.

[7] In particular, JMP calls the *Residual SS* the *Error SS*.

R^2 is defined by either of the two expressions

$$R^2 = \frac{Model\ SS}{Total\ SS} = 1 - \frac{Residual\ SS}{Total\ SS}$$

which is “the proportion of the total variation explained by the regression”.

Note how R^2 compares *Total SS* and *Residual SS* to capture the usefulness of using x to predict y. (p 92) [8]

R^2 is often used as a measure of the “effectiveness” of a regression.

Advice: Resist the temptation to think of R^2 in absolute terms. Regression is a statistical tool for extracting information from data. Its value depends on the value of the information provided.

RMSE also provides useful information about the effectiveness of a model.

Do R^2 and *RMSE* answer the same question about a model?

Curious (but useful) Fact: In simple regression $R^2 = r^2$, where as in Stat 603, r is the sample correlation.

[8] These page numbers refer to the BAR casebook.

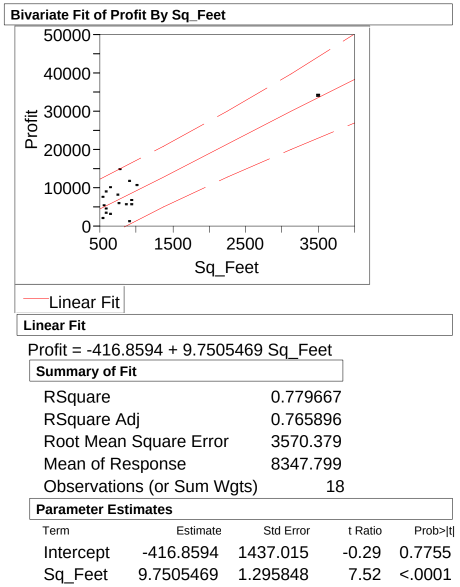
The Impact of Outliers: Another Example

Outliers can impact inferences from regression in a dramatic fashion.

Value of Housing Construction (p 89)

The data set cottage.jmp gives the profits obtained by a construction firm and the square footage of the properties.

The scatterplot shows that the firm has built one rather large “cottage”. This is an “outlier” in the sense that it is very different from the rest of the points. (p 90)



```
Linear Fit
Profit = -416.8594 + 9.7505469 Sq_Feet
Summary of Fit
RSquare                        0.779667
RSquare Adj                    0.765896
Root Mean Square Error         3570.379
Mean of Response               8347.799
Observations (or Sum Wgts)           18
Parameter Estimates
Term        Estimate    Std Error   t Ratio   Prob>|t|
Intercept   -416.8594   1437.015      -0.29   0.7755
Sq_Feet     9.7505469   1.295848       7.52   <.0001
```

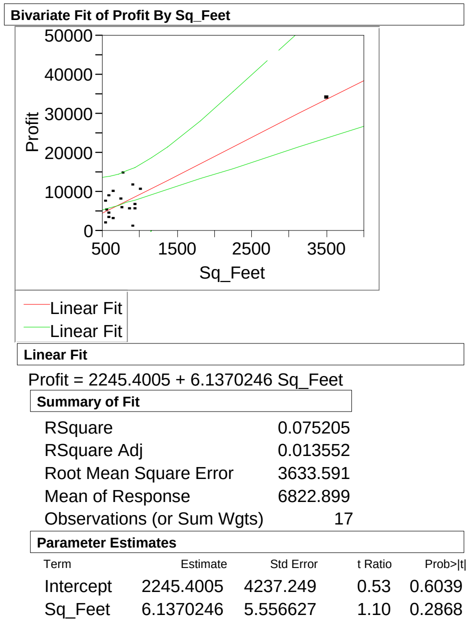
What do R^2 and *RMSE* tell us about this model?

What is the interpretation of the 95% CI for the slope?

What is the interpretation of the 95% CI for the intercept?

Without the Large Property

How does the fitted model change when we set aside the large cottage and refit the model without this one? (p 94)



Bivariate Fit of Profit By Sq_Feet

🟥	Linear Fit
🟩	Linear Fit

Linear Fit

Profit = 2245.4005 + 6.1370246 Sq_Feet

Summary of Fit

RSquare	0.075205
RSquare Adj	0.013552
Root Mean Square Error	3633.591
Mean of Response	6822.899
Observations (or Sum Wgts)	17

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2245.4005	4237.249	0.53	0.6039
Sq_Feet	6.1370246	5.556627	1.10	0.2868

What has happened to R^2? To *RMSE*? To the CI for the slope?

Which version of this model should the firm use to estimate profits on the next large cottage it is considering building?

What additional information about these construction projects would you like to have in order to make a decision?

Leverage and Outliers

The dramatic effects of removing the outlying large “cottage” in this last example illustrates the impact of a *leveraged* outlier.

Leverage: points that are far from the rest of the data along the x-axis are said to be leveraged. BAR gives the formula for leverage on page 63.

Heuristic: Moving away from the center of the predictor impacts the possible effect of a single observation in regression much like moving your weight out to the end of a see-saw. As your weight moves farther from the fulcrum, you can lift more weight on the other side.

Leverage is a property of the values of the predictor, not the response.

Leveraged points are not necessarily bad and in fact improve the accuracy of your estimate of the slope.[9] Just recognize that you are giving some observations a bigger role than others.

[9] In particular, leveraged observations are those that contribute the most to the variation in the predictor. Since these points spread out the values of the predictor, they make it easier to estimate the slope of the regression.

Take-Away Review

Inference for regression benefits from the same sort of "astonishing facts" that made inference for means possible in Stat 603.

In particular, the sampling distribution of the slope is approximately

$$\hat{\beta}_1 \sim N(\beta_1 , \frac{\sigma_\varepsilon}{\sqrt{n}} \times \frac{1}{SD(x)})$$

So, we can form confidence intervals as before, forming the intervals as before, namely as [estimate ± 2 SE(estimate)].

We can use these same ideas to construct confidence intervals for the *average* of the response for any value of the predictor as well as for a specific response.

The R^2 summary measures the proportion of the variation of the response "explained" by the model; the *RMSE* shows the SD of the noise that remains.

Next Time

Getting more data gives you better estimates of the slope and intercept, but has little impact on the accuracy of prediction.

The only way to improve R^2 and reduce *RMSE* is to add more predictors. This is the domain of multiple regression.