Stat 102

Introduction to Business Statistics

Class 11

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Transformations of confidence intervals
- Prediction in regression
  1. Confidence bands for the true regression line
  2. Prediction intervals for a new observation
- Summary
- Next time

Last time

- The SRM structure and assumptions
- Utopia – the best situation possible
- Checking the assumptions and the modeling process
- Standard errors for the slope
- Confidence intervals and test statistics

Confidence intervals for the slope and intercept

We know from last class that we can create confidence intervals (approx. 95%) for the true slope and intercept, β_1 and β_0:

$$\{b_1 - 2 \times std.err(b_1), b_1 + 2 \times std.err(b_1)\}.$$

$$\{b_0 - 2 \times std.err(b_0), b_0 + 2 \times std.err(b_0)\}.$$

Transforming the ends of confidence intervals

- If the new Chevy truck is 250 lbs lighter than its predecessor, then how much gas is it expected to save over the course of a year that involves driving 12,000 city miles? Provide a 95% CI.
- If gas costs $2.00 per gallon then how much money is expected to be saved by the weight reduction?
- Base your answer on the model that has GPM City as the Y variable.

Confidence interval output

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	0.0071961	0.00187	3.85	0.0001*	0.0035222	0.0108699
Weight(lb	1.3944e-5	4.797e-7	29.07	<.0001*	0.0000130	0.0000149

Create the confidence interval by clicking inside the parameter estimates table and choosing: Columns. Then add the lower and upper 95% values.

Transforming the ends of confidence intervals

- The theoretical difference in gallons consumed for driving 1 mile is $250\beta_1$.
- We will estimate this with the quantity $250b_1$, that is $250 \times 1.3944\text{e-}5 = 0.003486$, which is measured in gallons per mile.
- To get the 95% CI we use $250(b_1 - 2se(b_1), b_1 + 2se(b_1))$.
- Note that we create the CI for β_1 first, then multiply both ends by 250.
- From the JMP output this equals $250 \times (0.0000130, 0.0000149) = (0.00325, 0.003725)$.
- But this is the gas saved when driving 1 mile. To get the interval for driving 12,000 miles multiply both ends of the interval by 12,000 to get (39.0, 44.7) gallons
- To get the cost savings if gas is \$2.00 per gallon, multiply the ends of this interval by \$2 per gallon, to get (\$78.00, \$89.40).

General transformations

In general, so long as your transformation function is strictly increasing or decreasing you can simply multiple the ends of the confidence interval by the transformation to get the interval on the transformed scale.

Big picture

The model for data generation in regression is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

There are two types of interval associated with the fitted line:

1. Confidence bands for the regression.

$$y_i = \boxed{\beta_0 + \beta_1 x_i} + \epsilon_i.$$

2. Prediction intervals for a new observation.

$$y_i = \boxed{\boxed{\beta_0 + \beta_1 x_i} + \boxed{\epsilon_i}}.$$

Confidence bands for the regression

- The systematic (signal) part of the SRM:

$$E(Y|X) = \beta_0 + \beta_1 X.$$

- In English: what do you think the mean of Y is, for a given value of X, aka what's the height of the true regression line?
- Example: for cars that weigh 3500lbs, what is their expected fuel economy in GPM City?
- We estimate this with the least squares regression line:

$$\hat{y} \approx b_0 + b_1 3500.$$

Confidence bands for the regression

- Provide a 95% CI for this expectation.
- To do this we need the standard error of

$$\hat{y}_i = b_0 + b_1 x_i.$$

- We can't just add the standard errors of b_0 and b_1 because this is a linear combination of random variables.
- But, if you did the math you would get (not in Stine):

$$se(\hat{y}) = RMSE \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}.$$

Confidence bands for the regression

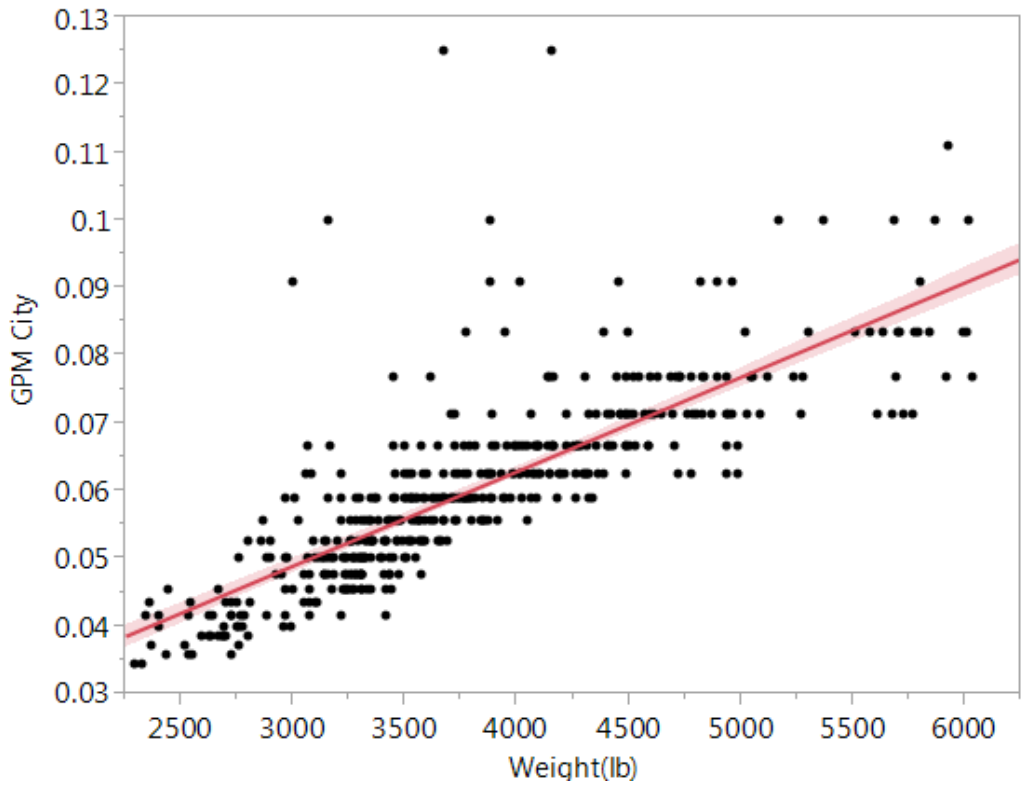
Three key observations from this formula:

1. This is a function of x_i.
2. As x_i moves away from $\bar{x}$ then this standard error increases (the statistical extrapolation penalty).
3. As n gets large the standard error goes to zero. The confidence bands will collapse and capture the true regression line.

Confidence bands for the regression

In JMP from the Linear Fit use:

- Confid Shaded Fit to draw all the intervals (lots of intervals equals a band)
- Mean Confidence Limit Formula to save the edges of the bands into the spreadsheet



Prediction intervals for a **new** observation

- If we were to draw one new observation, then based on the fitted regression line, where do you think its y-value will be?
- Notice, that this is not a question about the average value of y, but the single realized new value.
- Example: I am about to build a car that will weighs 3500lbs. Provide a 95% PI for the fuel economy in GPM City, for this vehicle.

Prediction intervals for a **new** observation

We are trying to estimate the quantity:

$$y_{new} = \beta_0 + \beta_1 x_{new} + \epsilon_{new}.$$

To construct the interval we need to estimate this quantity and obtain its standard error:

$$\hat{y}_{new} = b_0 + b_1 x_{new},$$

$$se(\hat{y}_{new}) \approx RMSE \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{(n-1)s_x^2}}.$$

Notice that the forecast is the same as before $(b_0 + b_1 x_{new})$ but the standard error has changed to incorporate the additional ϵ term.

Prediction bands for the regression

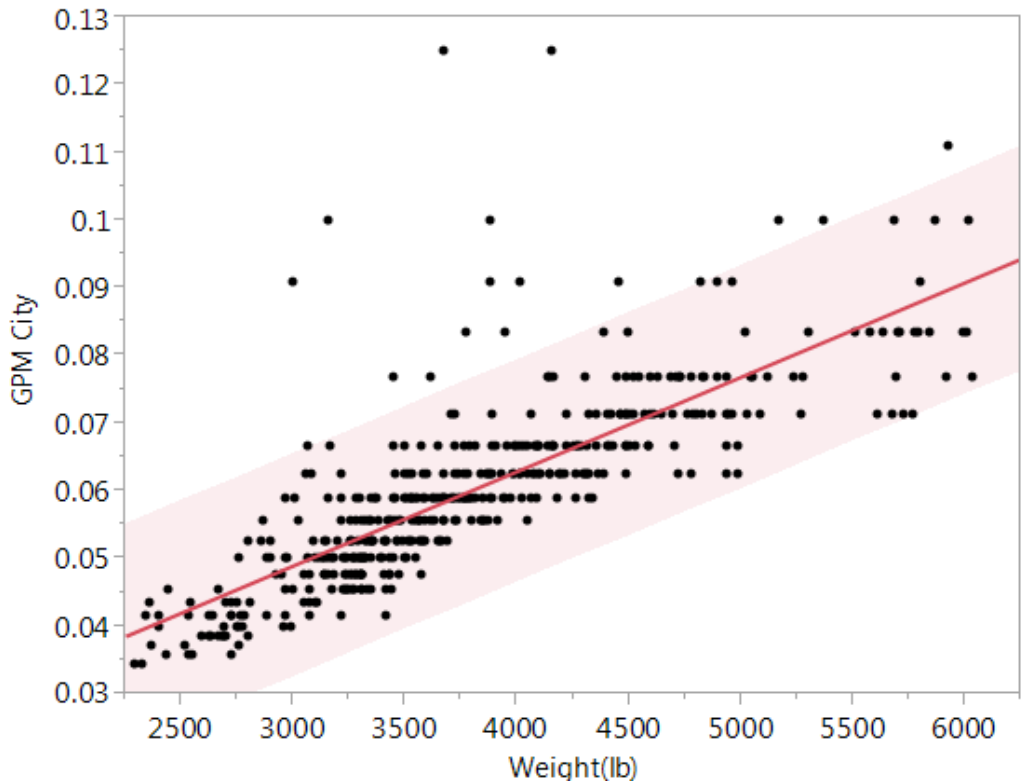
Three key observations from this formula:

1. This is a function of x_i.
2. As x_i moves away from $\bar{x}$ then this standard error increases (the statistical extrapolation penalty).
3. As n gets large the standard error goes to RMSE $\approx \sigma_\epsilon$, **not** zero. The 95% prediction intervals will become parallel to the true regression line at a distance of approximately +/-2RMSE.

Prediction intervals for a new observation

In JMP from the Linear Fit use:

- Confid Shaded Indiv to draw all the prediction intervals
- Save predicteds to get the forecast
- Indiv Confidence Limit Formula to save the edges of the prediction intervals into the spreadsheet



Prediction in JMP

- After having saved the intervals into the spreadsheet, create a new row and just add the x-value that you want to predict at.
- The intervals will be automatically populated.

For a new car that weighs 3500lbs we have

$$\hat{y}_{new} = 0.0560.$$

The 95% PI:

$$(0.0400, 0.072).$$

Back on the more familiar MPG scale (that is the 1/y transform) we have a 95% PI of:

$$\left(\frac{1}{0.0400}, \frac{1}{0.072}\right) = (13.98, 25.00).$$

Module summary

1. Transformations of confidence intervals
2. Confidence bands for the regression
3. Prediction intervals for a new observation

Next time

- Regression diagnostics and fixes for common problems