Multiple Regression

Project Analysis for Today

Getting your data!

Review of Bivariate Regression

Utopian model for regression

$$\text{Ave}(Y|X) = \text{Intercept} + \text{Slope}\,(X)$$
$$= \beta_0 + \beta_1 (X)$$

where we assume

(a) Independence
(b) Constant variance around Ave(Y|X)
(c) Normally distributed around Ave(Y|X)

Standard error and inference

$$\text{SE(slope estimate)} \approx \frac{\sigma}{\sqrt{n}} \frac{1}{\text{SD}(X)}$$

Use SE to form confidence intervals and test H_0: β_1= some constant.

R-squared (R^2)

$$R^2 = \frac{\text{Variation captured by fitted model}}{\text{Variation in Response}} = \%\ \text{explained variation}$$

Prediction accuracy

Given the RMSE (or estimated SD of the errors), form a 95% prediction interval as (prediction) ± 2 *RMSE*.

The previous interval is only accurate for predictions in the range of the observed data. Extrapolation beyond that range is less accurate.

Leverage, influence, and outliers

Observations with unusual values of the predictor are said to be *leveraged*. Removing *influential* observations lead to changes in the fitted model.

Questions?

Use of software?

About regression?

Review Example: Housing Prices and Crime

Philadelphia housing prices Phila.jmp, page 62

"How do crime rates impact the average selling price of houses?"

– Initial plot shows that Center City is "leveraged" (unusual in X).

– All data
- Initial regression with all data finds $577 impact per crime (p 64).
- Residuals show lack of normality (p 65).
- Plot using JMP button associated with fit in *Fit Y by X* view.

– Set aside Philadelphia temporarily
- Regression has much steeper decay, $2289/crime (p 66).
- Residuals remain non-normal (p 67).

– Scatterplot without Philadelphia suggests curvature (p 66-68).

– Alternative analysis with transformation
- Suggests Center City may be not so unusual. (pages 68-70)

– Why is CC an outlier?
- What do we learn from this one observation?
- Should it be included in the analysis, or excluded?

Multiple Regression

Illustrative Application: Separating the factors that drive sales

– Which factor is the most important determinant of business growth?
Advertising? Product loyalty? Price?

– Complicated because of the relationships among the predictors.

Model

Add several other predictors

"Y" = $\beta_0 + \beta_1$ "X_1" + ... + β_k "X_k" + Error

Sales = $\beta_0 + \beta_1$ Adv$ + β_2 Price + Error

with same three other assumptions

- Independence
- Constant variance σ^2 about regression line
- Normally distributed errors about the regression line.

Discussion of equation

– Slopes measure effect of each predictor "holding others fixed"

– Same slope β_j for each X_j regardless of values of other factors

– Factors combine additively (a.k.a., an additive model)

Marginal slope versus partial slope

– Marginal: "simple" regression slope

– Partial: multiple regression slope, adjusted for levels of other factors.

– Draw the "graph" with variables as "nodes"

Determinants of SE for slope

– Relationships/correlation among predictors increase the SE of slopes.

– SE(slope estimate for X_j) $\approx \frac{\sigma}{\sqrt{n}} \frac{1}{\text{SD(Adjusted } X_j)}$

Goodness-of-fit and R^2

– R^2 = Proportion of variation in response captured by the fitted model.

– R^2 = Squared correlation of Y and predicted values from fitted model.

– Judge changes in R^2 by looking at *what is left over*...

Easy... 0.50 $\Rightarrow$ 0.51 Hard... 0.98 $\Rightarrow$ 0.99

Leverage plots

– Important graphical diagnostic for multiple regression.

– Reduces multiple regression to sequence of simple regressions.

– "Do I like the shown simple regression model?"

Inference in Multiple Regression

Does this predictor improve a model containing the others?
– Answer this using the t-ratio for a partial slope.
– Large values (i.e., |t| > 2) imply a significant improvement.

Does the model, taken collectively, explain significant variation?
– Answer this using the F-ratio from the anova table.
– Large values (i.e., p-value for F < 0.05) imply significant variation "explained" or represented by the fitted model.

Example of Multiple Regression

Automobile design Car89.jmp, page 109

"What is the predicted mileage for a 4000 lb. design, and what characteristics of the design are crucial?"

"How much does my 200 pound brother owe me for gas for carrying him 3,000 miles to California?" (Oops, it's urban mileage in example)

– Initial one-predictor model
- Transform response to gallons per 1000 mile scale.
- Cannot compare R^2's since two model use different dependent variables (MPG and GPM)
- Effect of scaling from GPM to GP1000M.
- RMSE = 4.23 (p 111)
- Skewness in residuals from regression with *Weight*. (p 112)
- Prediction @ 4000 lbs = 63.9, ⇑ 200 lbs for 3000 miles ≈ 8.2 gals

– Add variable for *Horsepower* (p 117)
- R^2 increases from 77% to 84% (added variable is significant, t=7.21)
- RMSE drops to 3.50
- Predictors are related, both increase together, higher SE for *Weight*.
- Picture explains the increase in SE due to restricted range (p 120).
- ⇑ 200 lbs for 3000 miles ≈ 5.3 gals

– Add a predictor less correlated with Weight, use *HP/Pound* (p 123)
- *Weight* and *HP/Pound* less related, more distinct properties of these cars.
- Predicted consumption = 64.3

Residual plots
– Show residuals plotted on fitted values
– Inspect for deviations from assumptions (such as lack of constant variance)

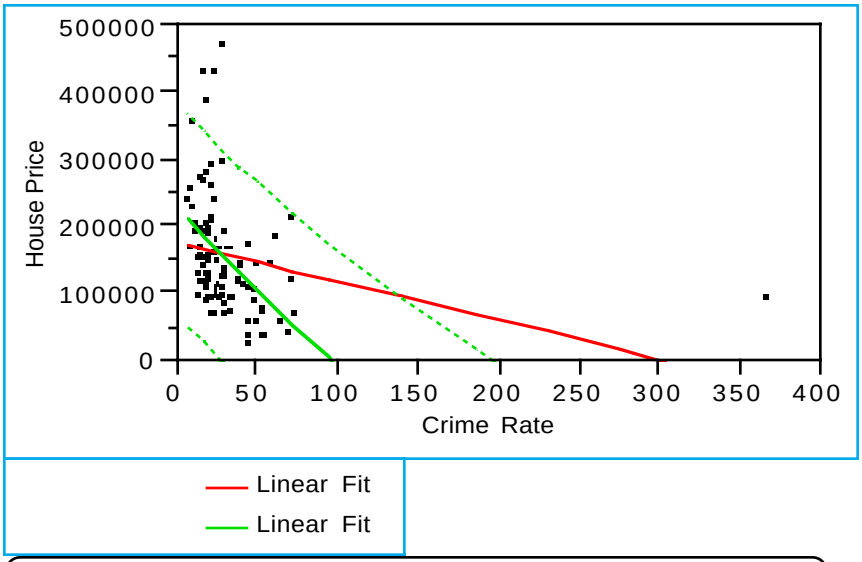
Leverage plots (p 125)
– New diagnostic plot, designed for multiple regression
– Show leveraged observations in *multiple* regression.
– Reveal outliers that exert effects on fit that are hard to see otherwise.

Next steps for this model…
– What other factors are important for the design?
– How small can we make the RMSE?
– How do we avoid "false positives" by searching over many predictors?

Next Time

More multiple regression
Effects of correlation among the predictors.

House Price By Crime Rate



Linear Fit

House Price = 176629 – 576.908 Crime Rate

Summary of Fit

RSquare	0.062
RSquare Adj	0.053
Root Mean Square Error	84325.05
Mean of Response	157835.6
Observations (or Sum Wgts)	99.000

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	176629.4	11245.6	15.71	<.0001
Crime Rate	-576.9	226.9	-2.54	0.0126

Linear Fit

House Price = 225234 – 2288.69 Crime Rate

Summary of Fit

RSquare	0.184
RSquare Adj	0.176
Root Mean Square Error	78861.53
Mean of Response	158464.5
Observations (or Sum Wgts)	98.000

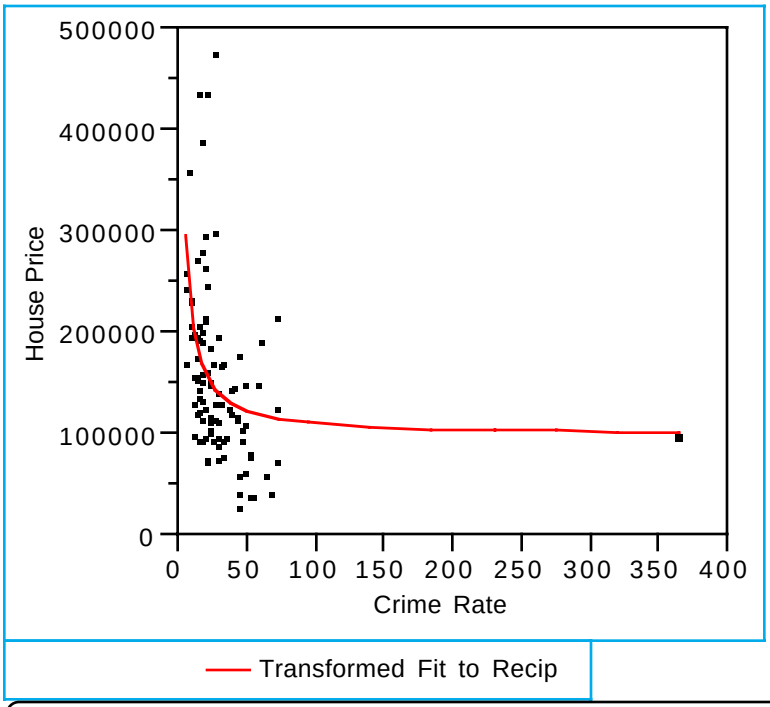
Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	225233.6	16404.0	13.73	<.0001
Crime Rate	-2288.7	491.5	-4.66	<.0001

Note: Point for Philadelphia is not included in the fitted model!

House Price By Crime Rate



Transformed Fit to Recip

House Price = 98120.1 + 1298243 Recip(Crime Rate)

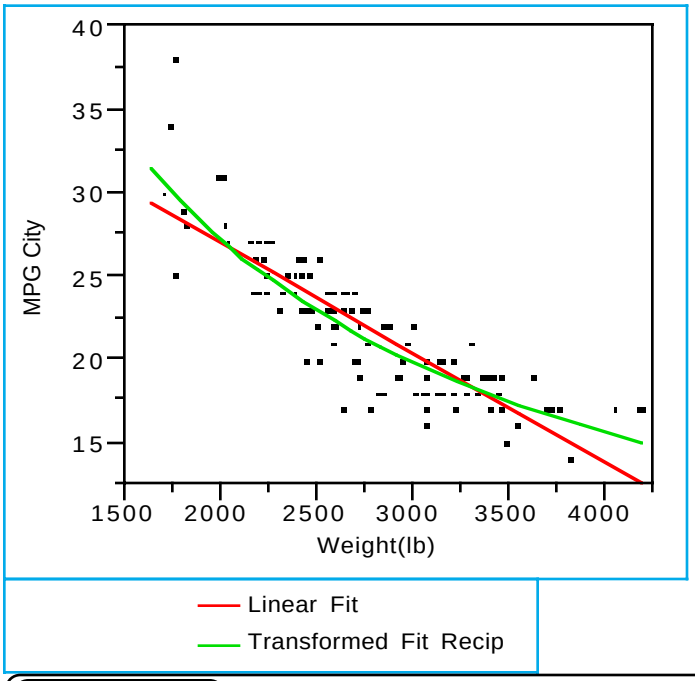
Summary of Fit

RSquare	0.170
RSquare Adj	0.161
Root Mean Square Error	79564.54
Mean of Response	158464.5
Observations (or Sum Wgts)	98.000

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	98120.1	15820.7	6.20	<.0001
Recip(Crime Rate)	1298242.7	293170.8	4.43	<.0001



Linear Fit

MPG City = 40.1183 – 0.00655 Weight(lb)

Summary of Fit

RSquare	0.742
RSquare Adj	0.740
Root Mean Square Error	2.168
Mean of Response	21.759
Observations (or Sum Wgts)	112.000

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	40.1183	1.0523	38.12	<.0001
Weight(lb)	-0.0066	0.0004	-17.79	<.0001

Transformed Fit Recip

Recip(MPG City) = 0.00943 + 0.00001 Weight(lb)

Summary of Fit

RSquare	0.765
RSquare Adj	0.763
Root Mean Square Error	0.004
Mean of Response	0.048
Observations (or Sum Wgts)	112.000

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0094323	0.002055	4.59	<.0001
Weight(lb)	0.0000136	7.19e-7	18.94	<.0001

Response: GP1000M City

Summary of Fit

RSquare	0.765
RSquare Adj	0.763
Root Mean Square Error	4.233
Mean of Response	47.595
Observations (or Sum Wgts)	112.000

Lack of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.4323	2.0545	4.59	<.0001
Weight(lb)	0.0136	0.0007	18.94	<.0001

Response: GP1000M City

Summary of Fit

RSquare	0.841
RSquare Adj	0.838
Root Mean Square Error	3.500
Mean of Response	47.595
Observations (or Sum Wgts)	112.000

Lack of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.6843	1.7270	6.77	<.0001
Weight(lb)	0.0089	0.0009	10.11	<.0001
Horsepower	0.0884	0.0123	7.21	<.0001