Linear Regression

ddebarr@uw.edu

2017-01-19

"In God we trust, all others bring data." – William Edwards Deming

Course Outline

1. Introduction to Statistical Learning
2. Linear Regression
3. Classification
4. Resampling Methods
5. Linear Model Selection and Regularization
6. Moving Beyond Linearity
7. Tree-Based Methods
8. Support Vector Machines
9. Unsupervised Learning
10. Neural Networks and Genetic Algorithms

Agenda

Homework Review

Probability

Chapter 3

Gradient Descent

Robust Regression

Probability

- Probability: the proportion of outcomes that we expect to meet some condition
  - Probability(Flipped Coin Lands on Heads)
  - Probability(Face of a Rolled Die Displays an Even Number)
- Joint Probability for Independent Events: the product of the individual probabilities
  - Reminder: log(Probability1 * Probability2) = log(Probability1) + log(Probability2)
- Fun Fact: the central limit theorem says the sum of a sufficiently large set of independent identically distributed random variables can be modeled as a Gaussian (bell curve) distribution [useful for arithmetic means]

Questions for the Advertising Data

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contributes to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising data?

Simple Linear Regression

$$Y \approx \beta_0 + \beta_1 X \qquad Y = \beta_0 + \beta_1 X + \epsilon$$

read "$\approx$" as *"is approximately modeled as"*

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

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

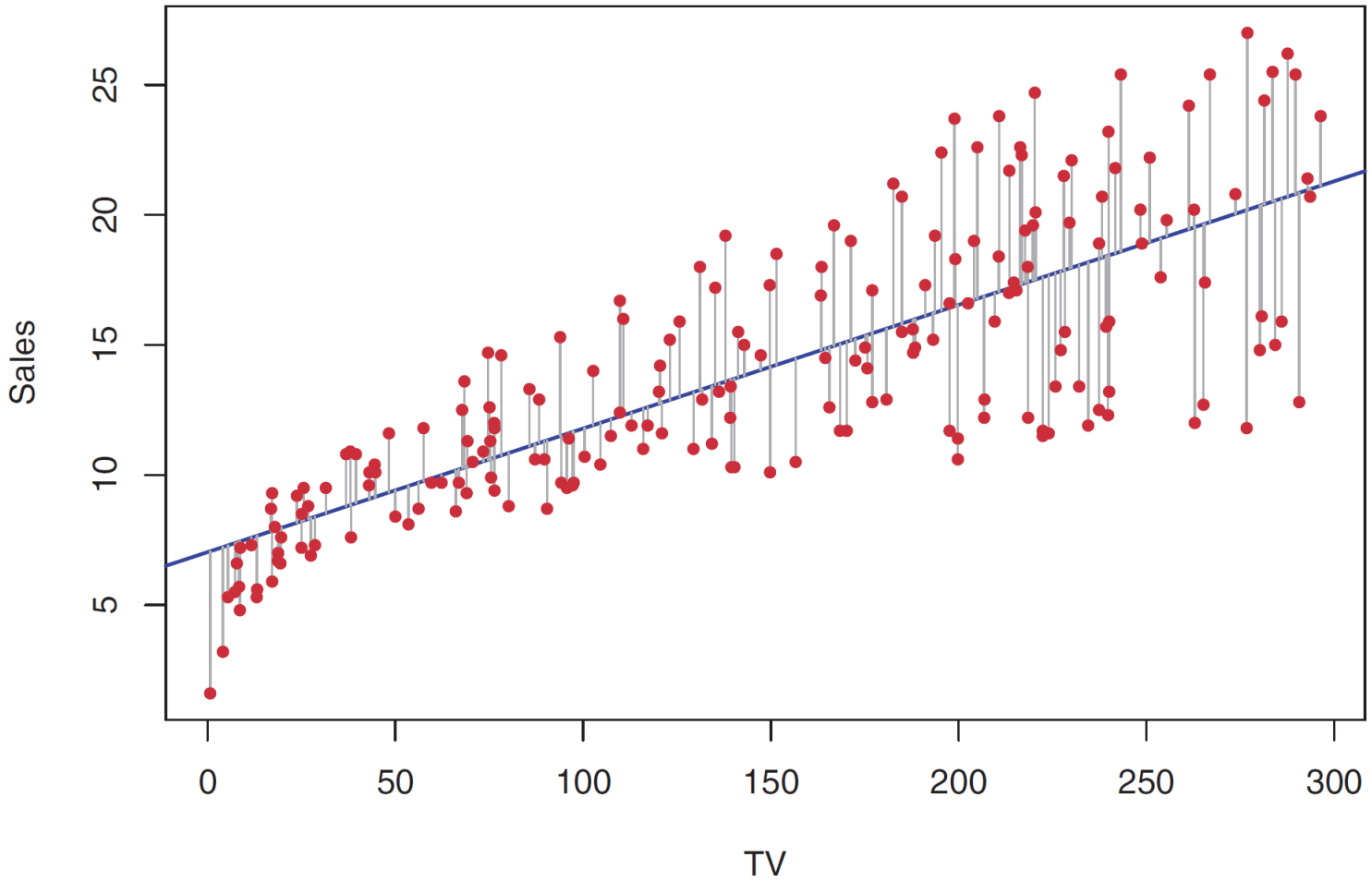
hat symbol, $\hat{}$, to denote the estimated value

The intercept moves the line up and down.

The slope measures the rate of change.

Estimating the Coefficients

$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$$



Residuals

$$e_i = y_i - \hat{y}_i \text{ represents the } i\text{th } \textit{residual}$$

- Residual Sum of Squares (RSS)

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Residual Standard Error (RSE)

$$\sqrt{\text{RSS}/(n-2)}$$

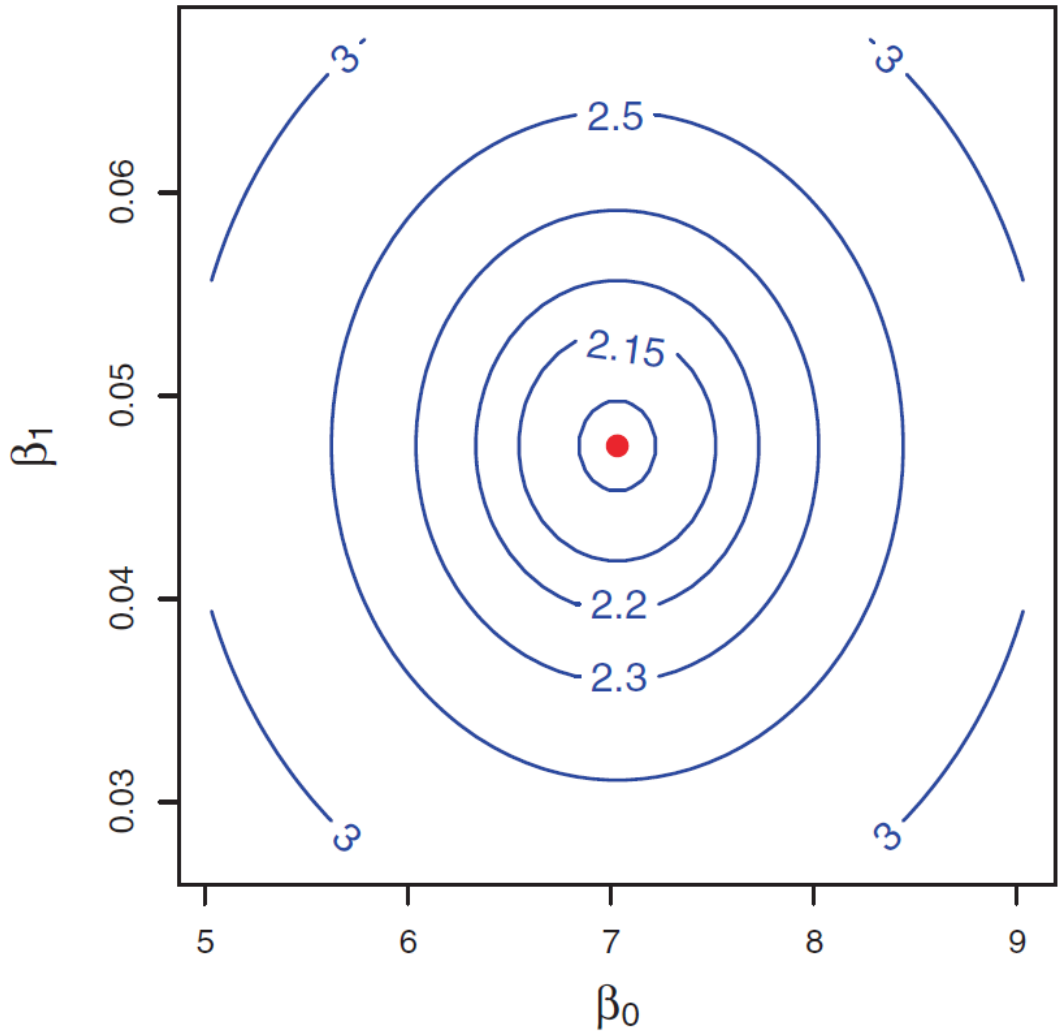
Simple Linear Regression

- Simple linear regression only has one predictor
- Slope and intercept are computed as …

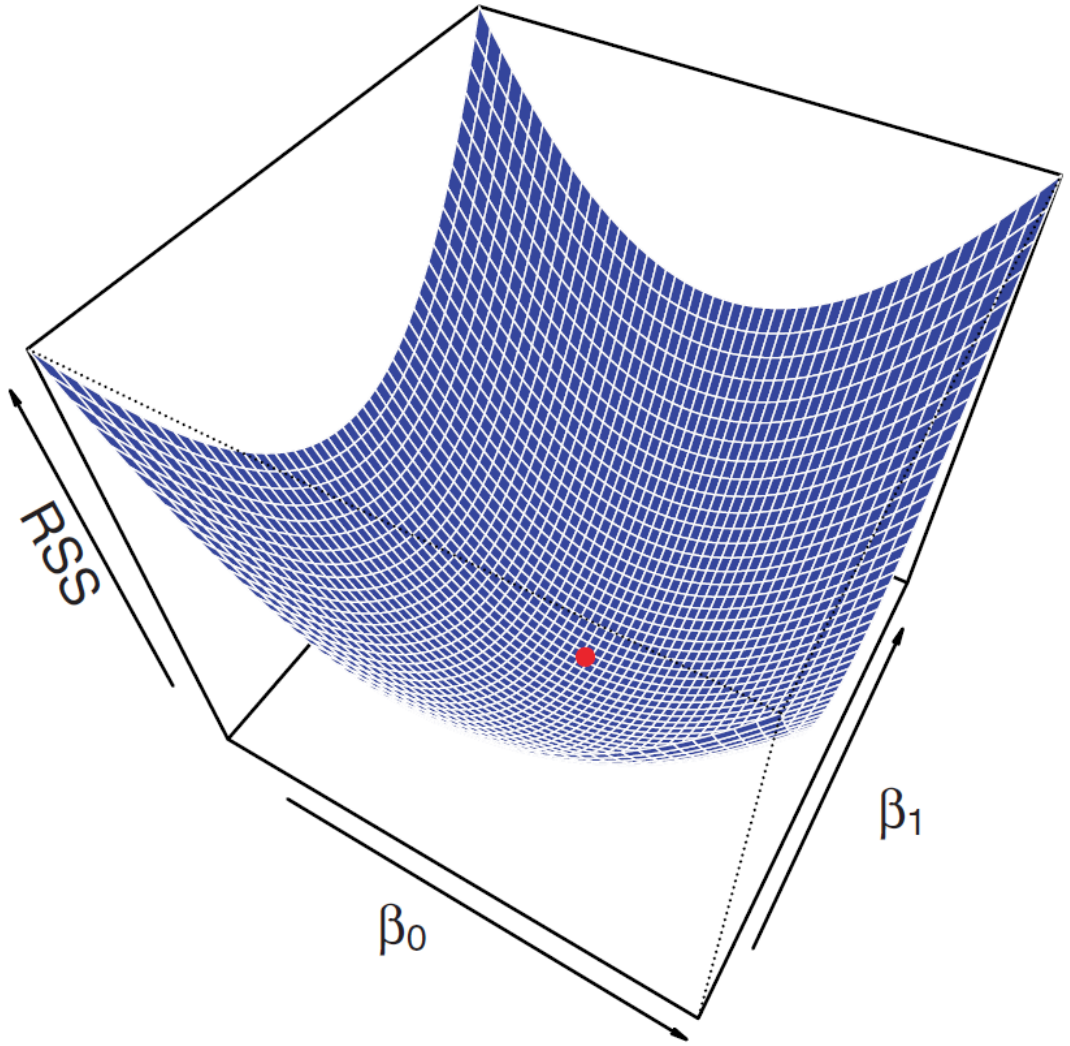
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

RSS as a Function of the Regression Coefficients



Contour Plot



3-D Plot

Derivation of the Maximum Likelihood Estimate for Multiple Regression

The negative log likelihood of the data is proportional to the residual sum of squared errors

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

$$RSS = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)$$

$$RSS = \left(\mathbf{y}^T - \boldsymbol{\beta}^T \mathbf{X}^T\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)$$

$$RSS = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

$$RSS = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

...so...

$$\frac{\partial RSS}{\partial \boldsymbol{\beta}} = 0 - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

... setting the gradient equal to 0 and solving for β ...

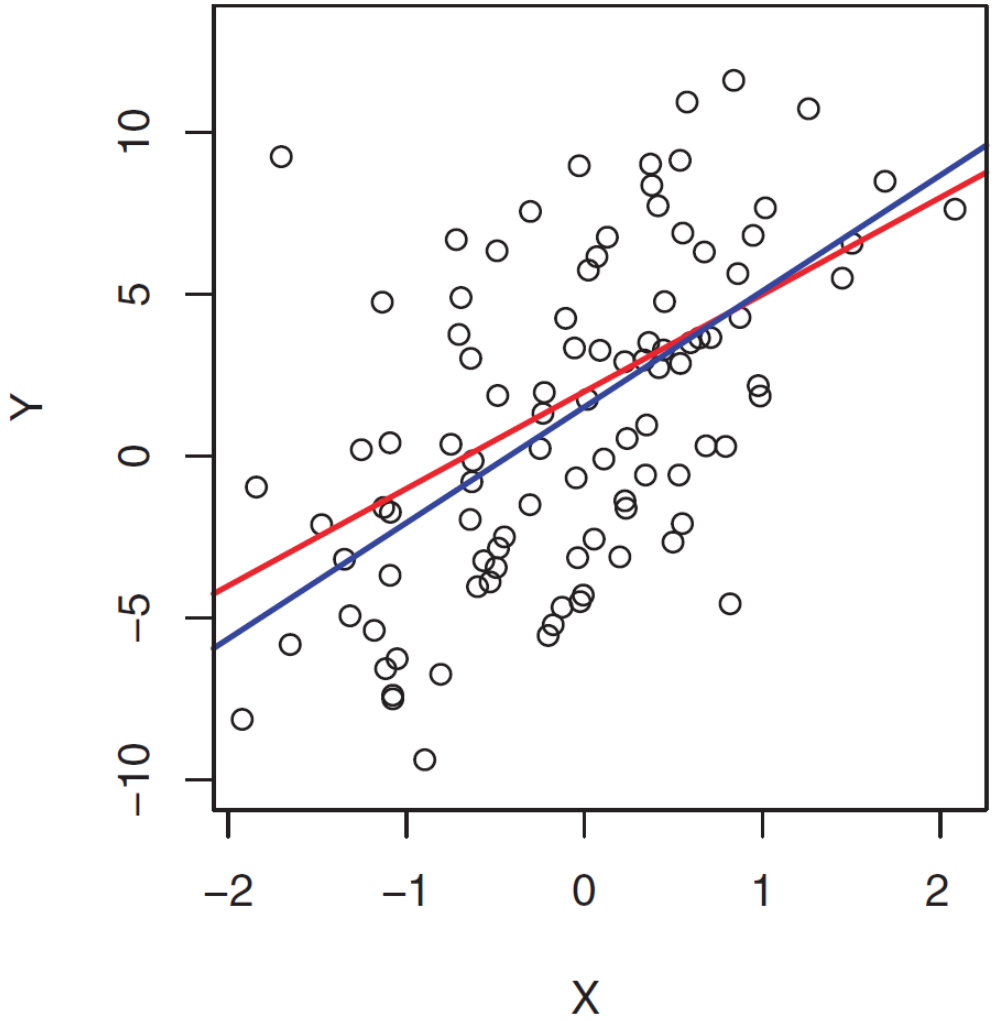
$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0$$

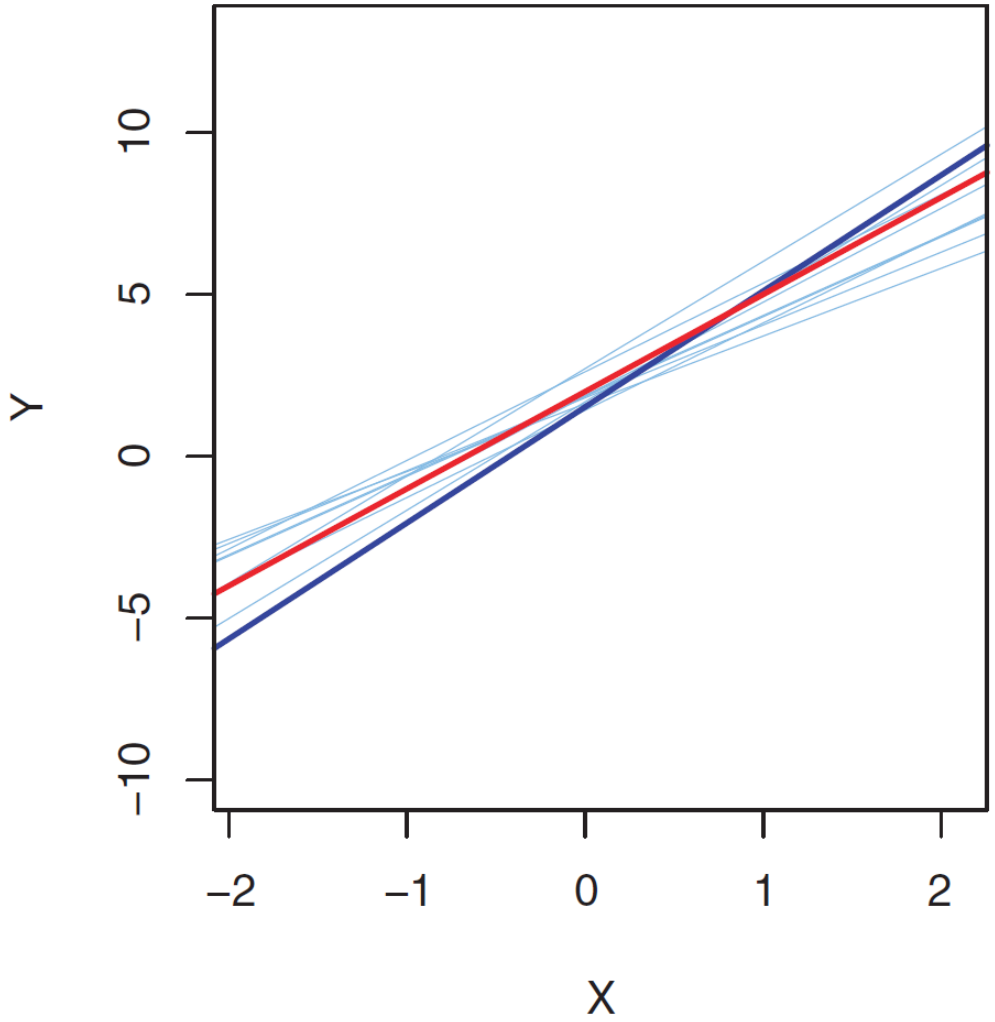
$$2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 2\mathbf{X}^T\mathbf{y}$$

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

$$\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Simulated Regression Problem





Red line is the population regression line f(X) = 2 + 3*x; blue lines are estimates based on random samples

Standard Error of a Mean

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

- The standard error of a mean quantifies our uncertainty about the mean
- We can estimate the lower and upper bounds of a 95% confidence interval for the mean as the 2.5th and 97.5th percentiles of a “t” distribution with mean = 0, standard deviation = SE, and degrees of freedom = n - 1

Standard Error of the Regression Coefficients

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The standard error of a regression coefficient quantifies our uncertainty about the regression coefficient
- We can estimate the lower and upper bounds of a 95% confidence interval for a regression coefficient as the 2.5th and 97.5th percentiles of a “t” distribution with mean = $\hat{\beta}_j$, standard deviation = SE, and degrees of freedom = n - 2

Hypothesis Test for a Regression Coefficient

Null Hypothesis

$$H_0 : \beta_1 = 0$$

Alternative Hypothesis

$$H_a : \beta_1 \neq 0$$

Test Statistic: the ratio of a difference to its standard error

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- The “t test” for the regression coefficient compares “t” to the “t” distribution with mean = 0, standard deviation = SE, and degrees of freedom = n – 2 [to compute a “p value”: the probability of observing a test statistic as extreme (as far from the mean) as the value of “t”]

Evaluating the Coefficients for our First Model

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Additional Statistics for the Model

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Three Simple Linear Regressions

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

Multiple Linear Regression

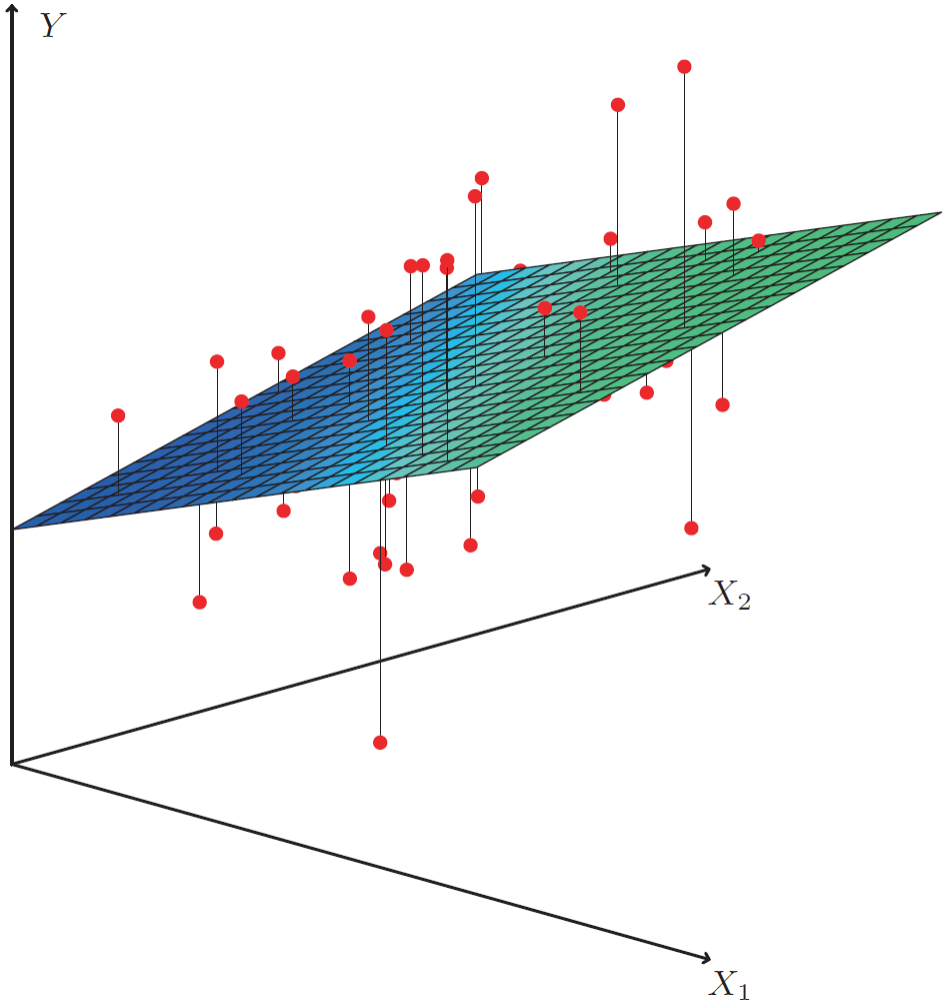
Multiple: more than one predictor

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Simple Multiple Regression Example



Notice that the regression plane cuts through the middle of the observations

Multiple Regression for the Advertising Data

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Notice that the newspaper effect is no longer statistically significant;
and the newspaper budget is positively correlated with the radio budget

Some Important Questions

1. *Is at least one of the predictors* $X_1, X_2, \ldots, X_p$ *useful in predicting the response?*

2. *Do all the predictors help to explain* Y*, or is only a subset of the predictors useful?*

3. *How well does the model fit the data?*

4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

1. Is there a relationship between the Response and Predictors?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

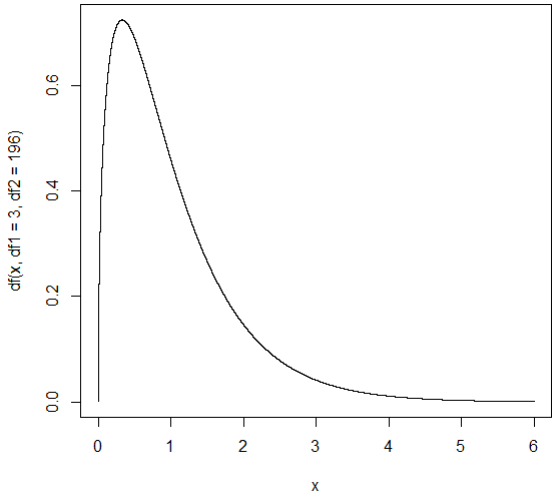
$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}$$

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

$$\text{RSS} = \sum (y_i - \hat{y}_i)^2$$

If the null hypothesis is true, this ratio will be one; otherwise this will be larger than 1

```
> pf(570.3, df1 = 3, df2 = 196, lower.tail = F)
[1] 1.568132e-96
```



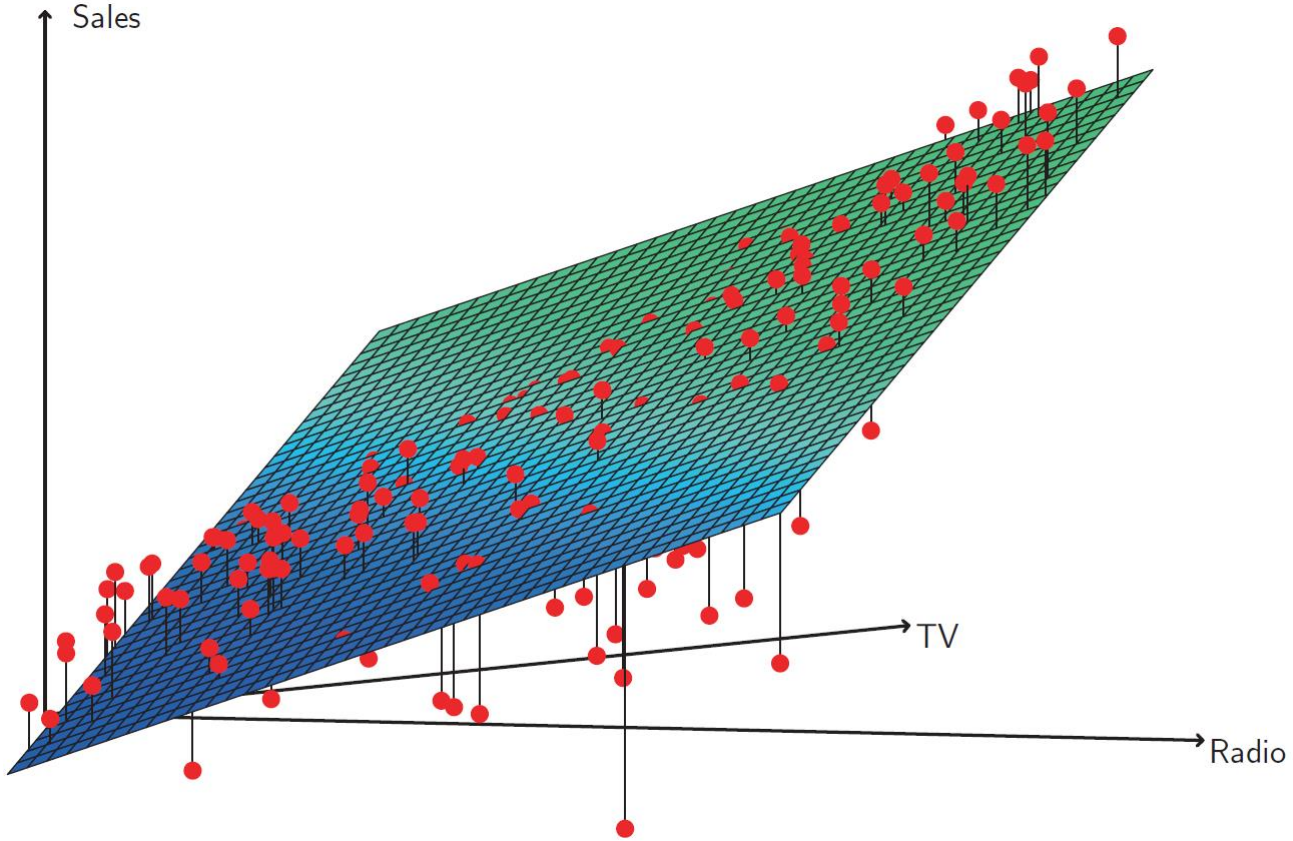
2. Deciding on Important Variables

- Various statistics can be used to evaluate the quality of the model (e.g. assessing various penalties for complexity): Mallow's Cp, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Adjusted R^2 [more later]
- Feature Selection
  - Forward Selection: add one variable at a time, choosing the variable that best reduces the RSS
  - Backward Selection: remove one variable at a time, choosing the variable with the largest p value
  - Mixed Selection: use forward selection, but remove any variable that exceeds a threshold p value

3. Model Fit

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}$$

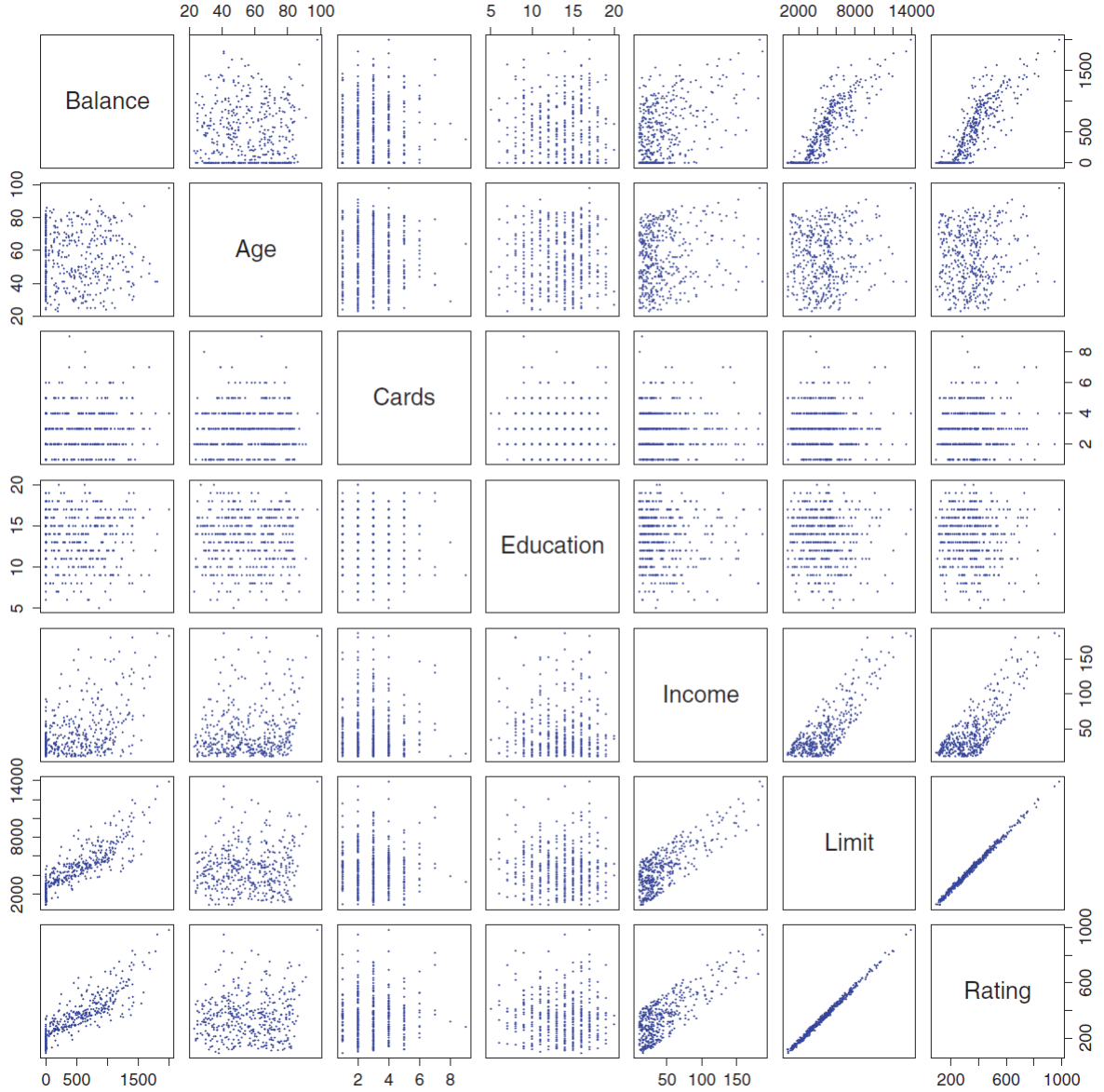


Does this plane look like it splits the observations?

4. Predictions

- Three types of uncertainty
  - Confidence interval: for the prediction of the mean output variable [the mean for a particular input vector]
  - Prediction interval: for the prediction of the output variable
  - Model bias: the error caused by choosing a linear model when the true model [which is unknown] does not match the model used

Quantitative Variables for the Credit Data Set



Example of a Model with a Qualitative Predictor

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

- Interpretation: the average Balance for gender=Male is $509.80, while the average Balance for gender=Female is $19.73 more
- Note: the p value is not significant

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Alternative Coding Scheme for Dummy Var

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

- Interpretation: average overall balance is Beta[0], with Beta[1] added to derive the average Balance for gender=Female and Beta[1] subtracted to derive the average Balance for gender=Male

Qualitative Predictors with More than Two Values

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

Evaluating the Predictors

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	−18.69	65.02	−0.287	0.7740
`ethnicity[Caucasian]`	−12.50	56.68	−0.221	0.8260

This F-test has a p-value of 0.96, indicating that we cannot reject the null hypothesis that there is no relationship between `balance` and `ethnicity`.

Extensions: Interactions

- Add an interaction term

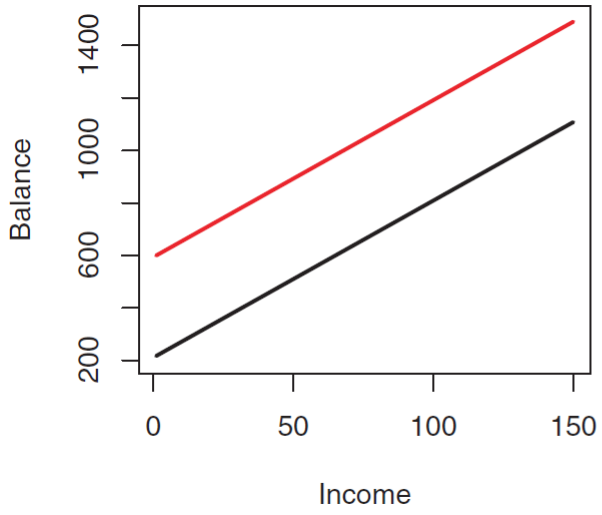
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

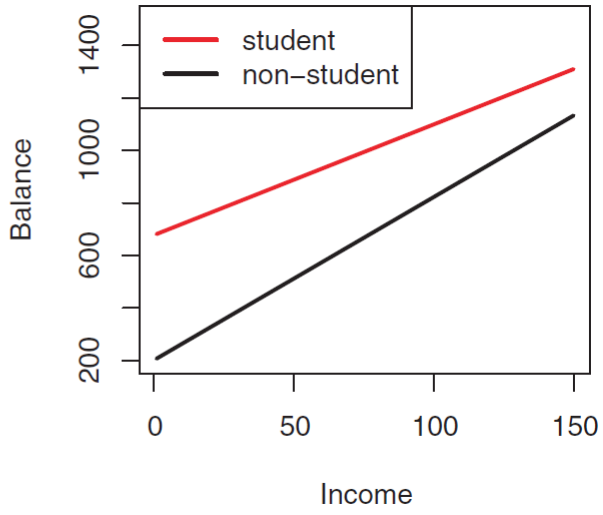
	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Extensions: Interactions

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

$$= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}$$

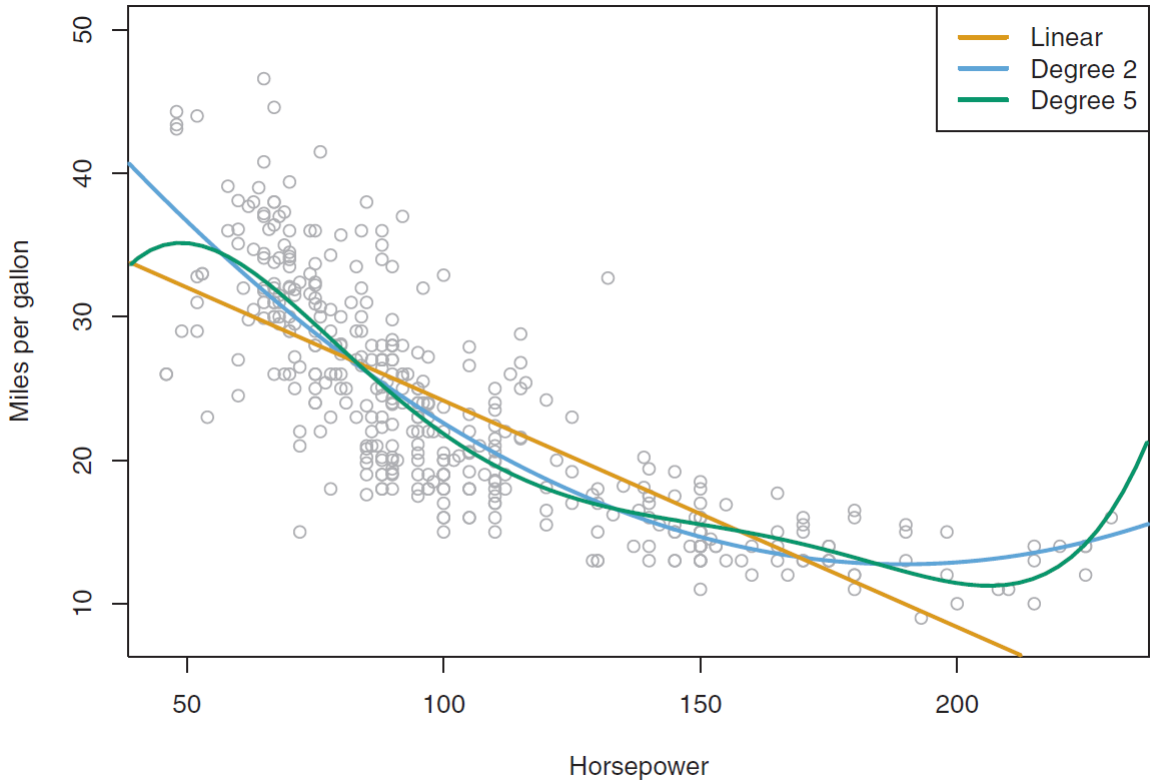




$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}$$

Extensions: Non-Linear Relationships



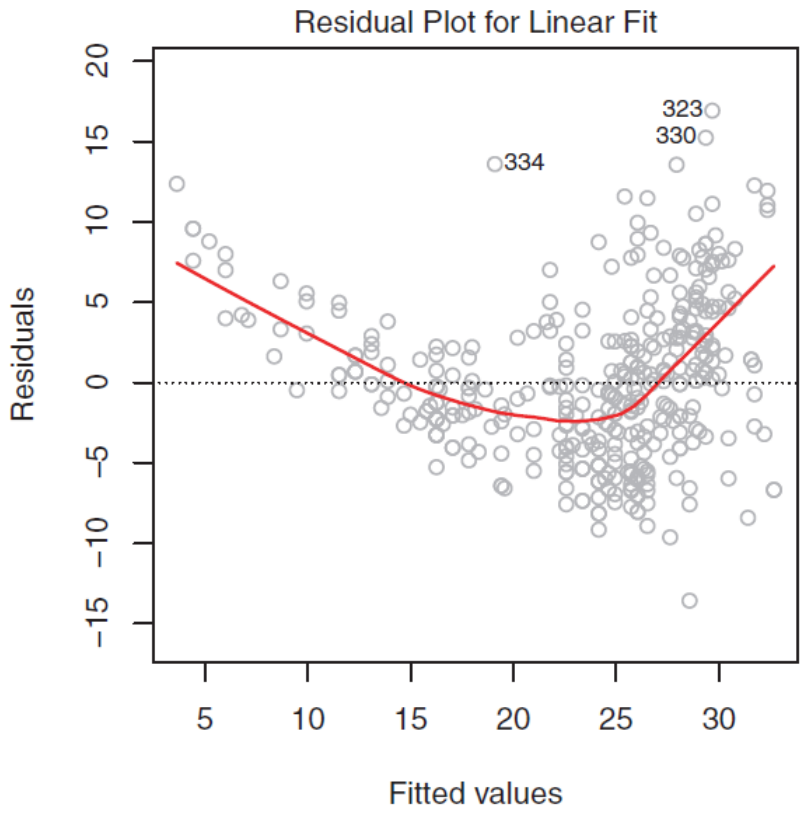
$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

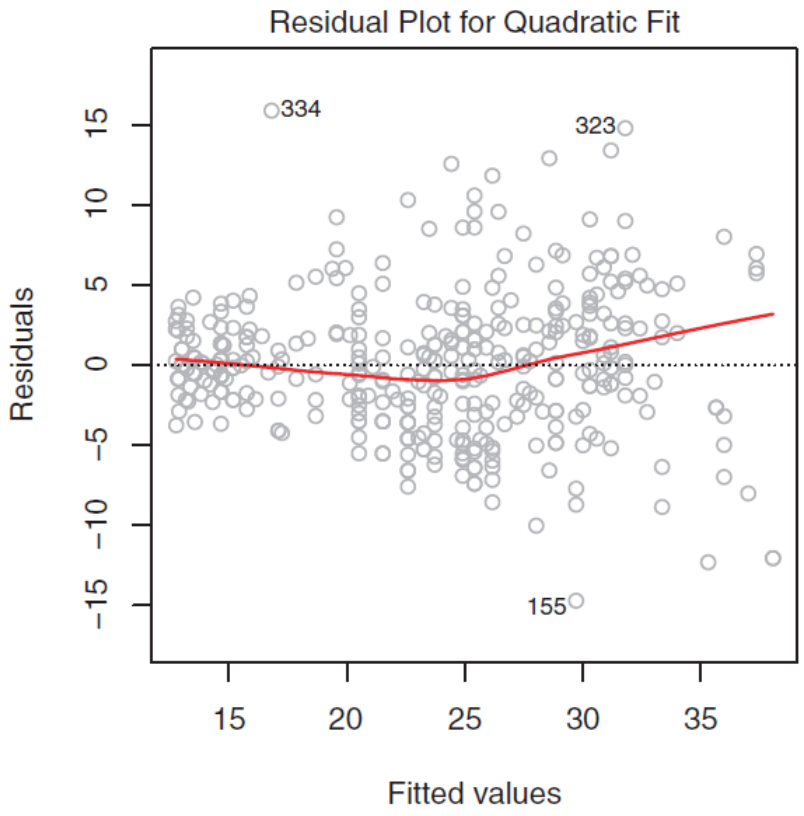
	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

Polynomial Regression

Potential Problem: Non-Linearity of the Data

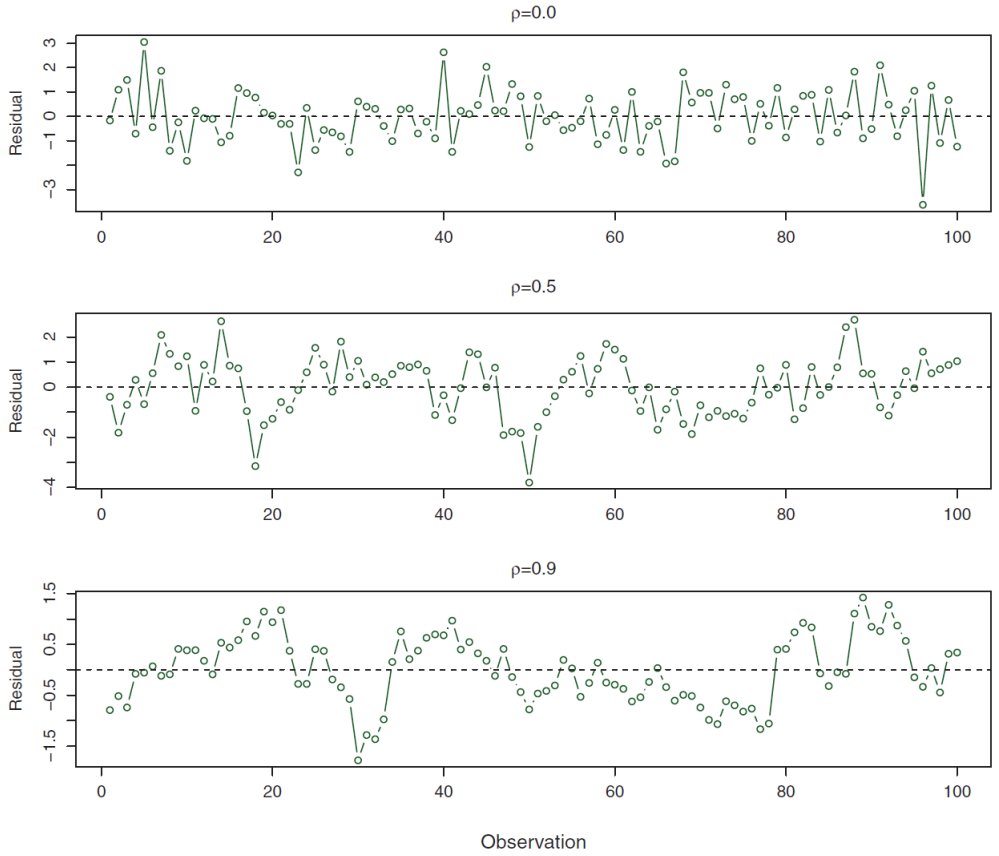
- Consider transform of predictors; e.g. log(x), sqrt(x), x^2, …





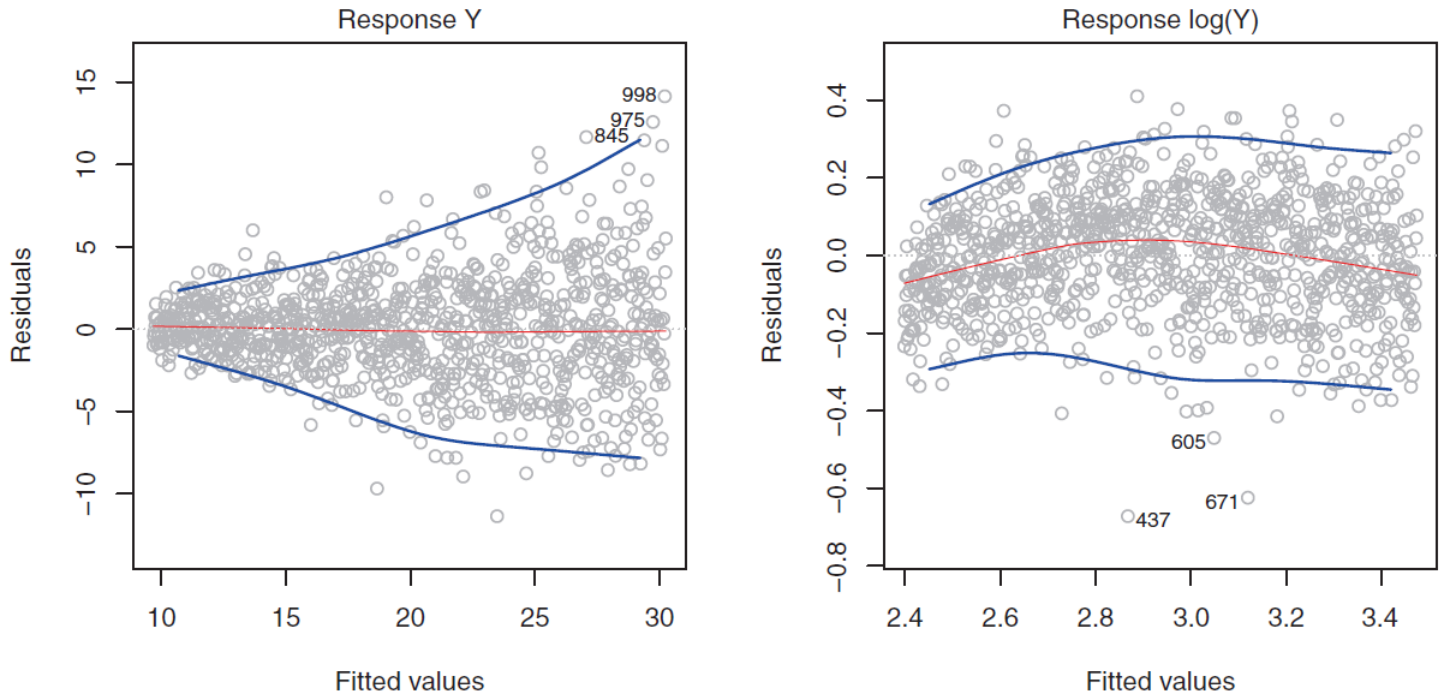
Problem: Correlation of Error Terms

- Can lead to underestimating the error terms
- May observe “tracking” among the residuals (2nd and 3rd plots)



Problem: Non-Constant Variance of Error Terms

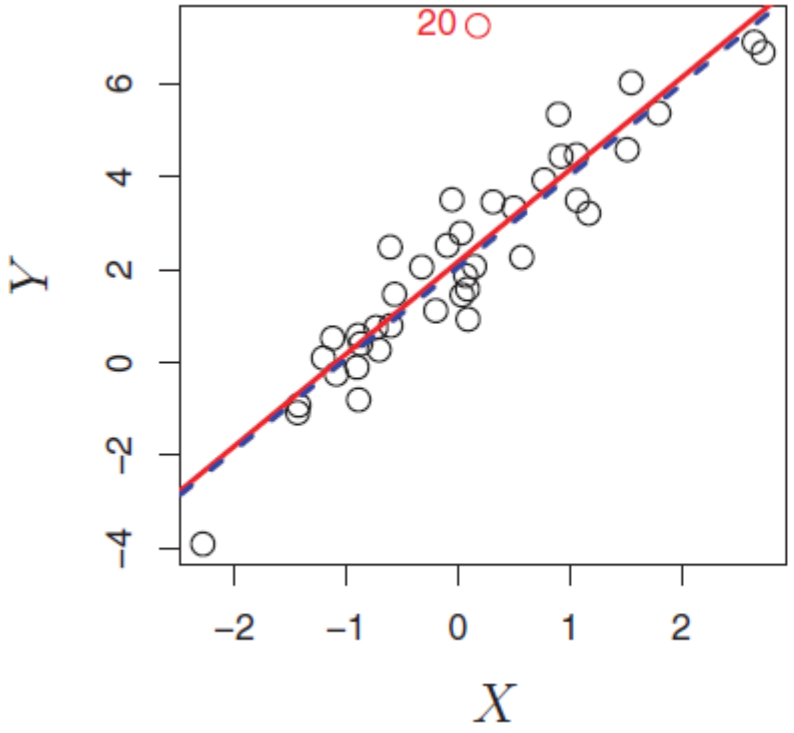
- Consider transforming the output

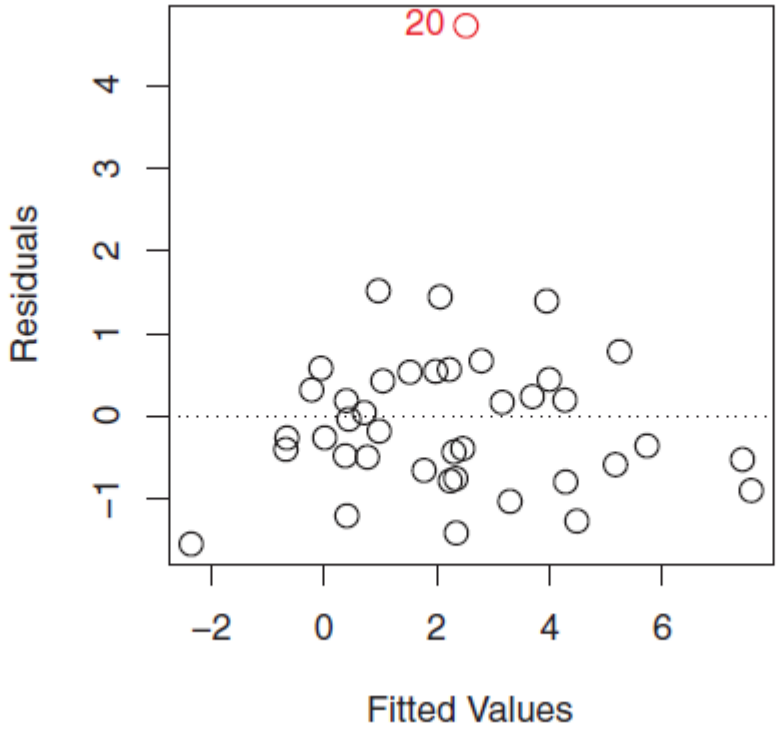


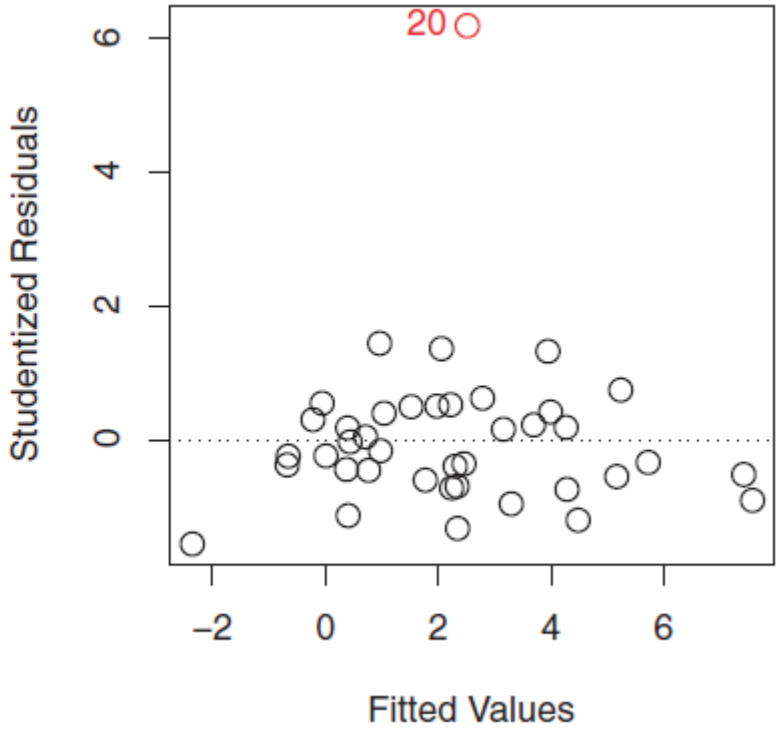
- Use weighted least squares when using average output values

Problem: Outliers

- Unusual output value may increase the RSE and reduce R^2

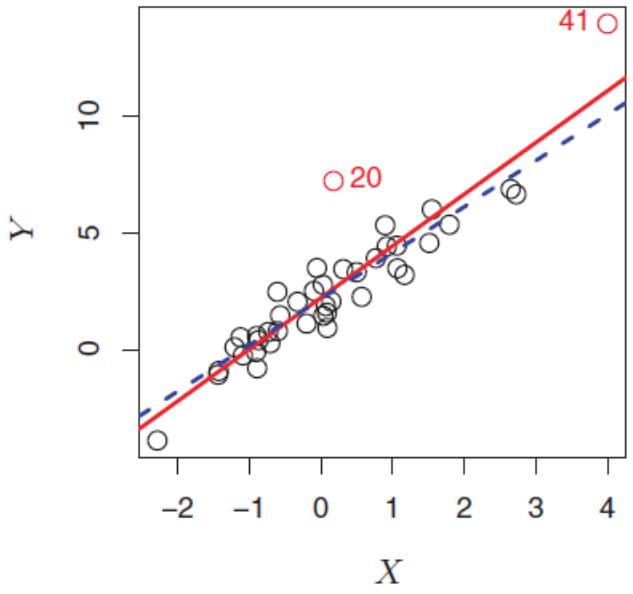


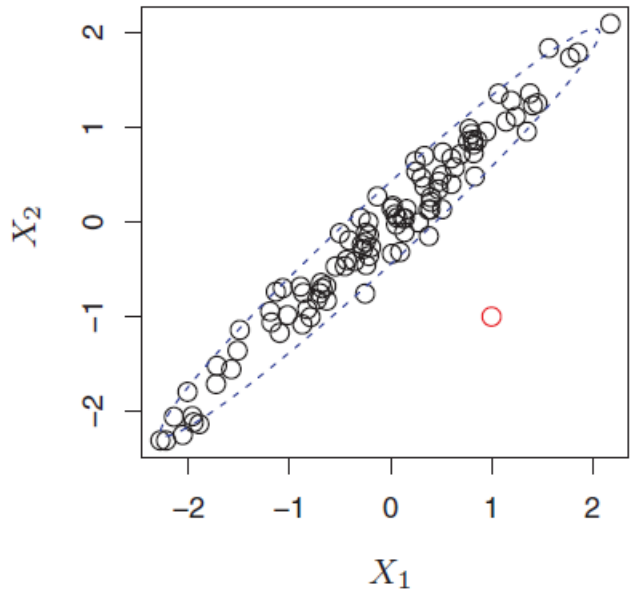


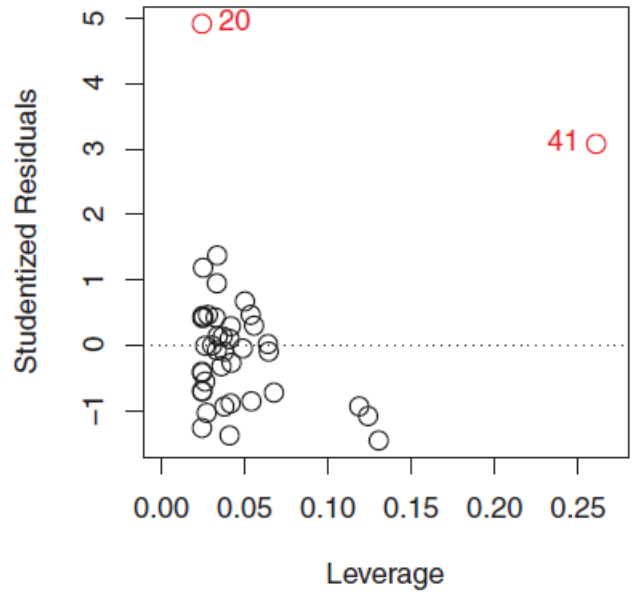


Problem: High Leverage Points

- Unusual input values may modify the model



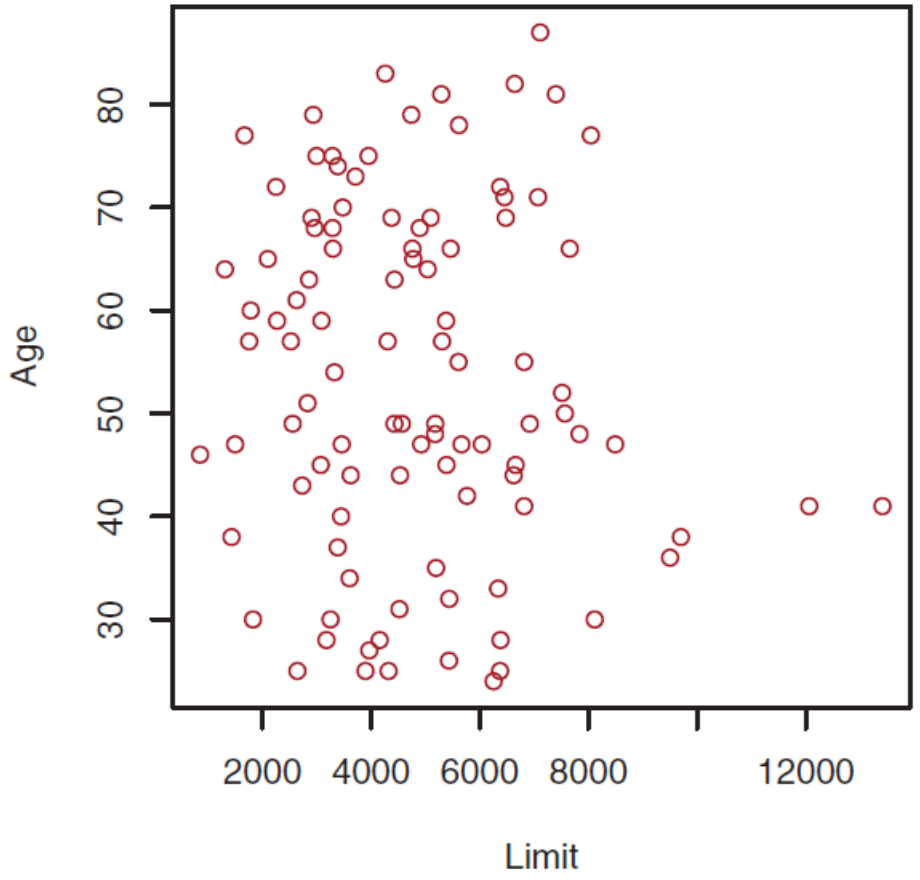


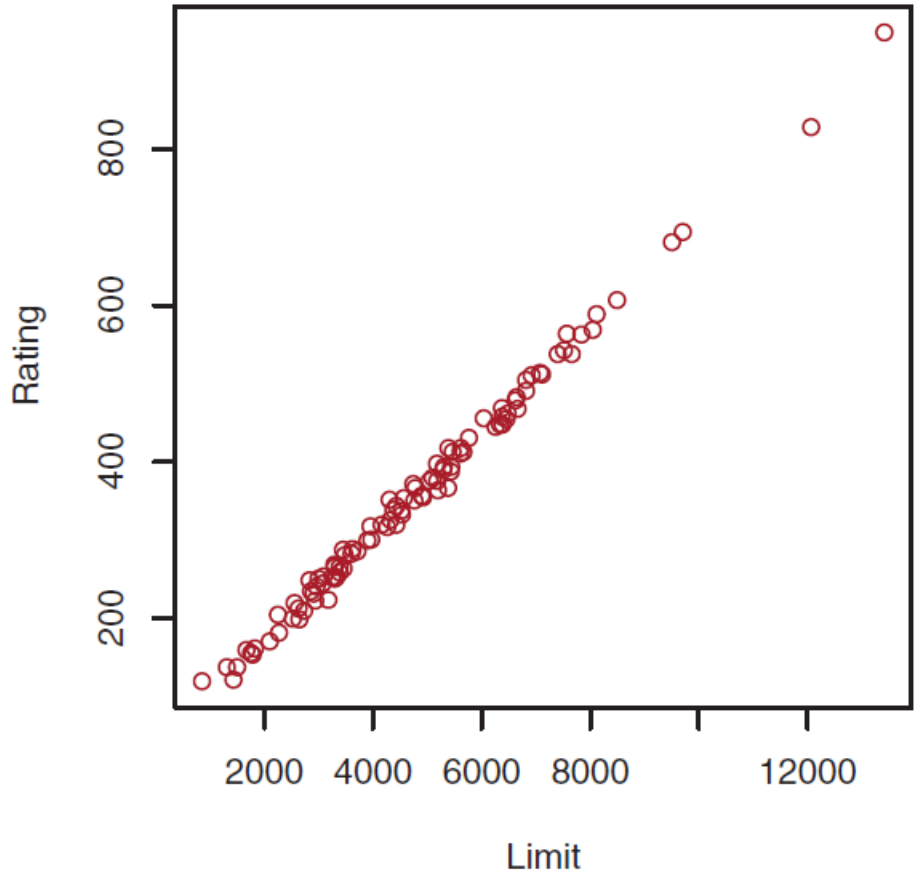


$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$

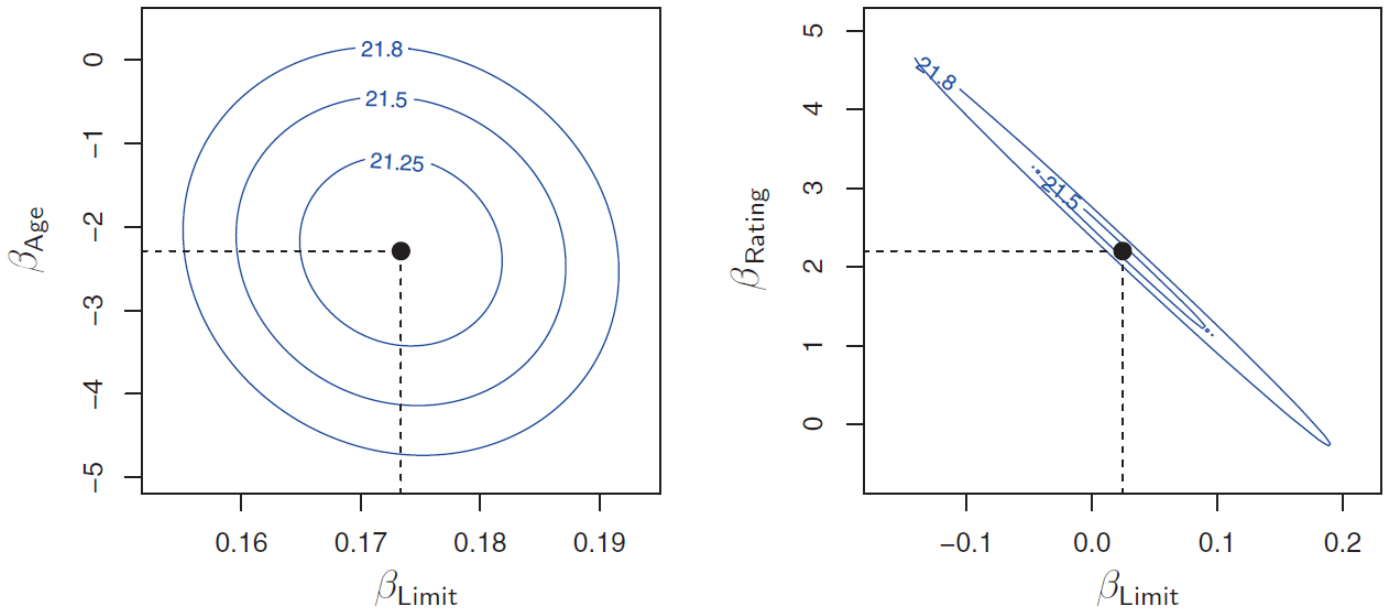
Problem: Collinearity

- Correlated variables: called multi-collinearity if more than two variables are involved





Problem: Collinearity: Example



		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

Problem: Collinearity: Detecting Collinearity

- Variance Inflation Factor [variable X_j predicted by other variables]

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

- Large values indicate a collinearity problem

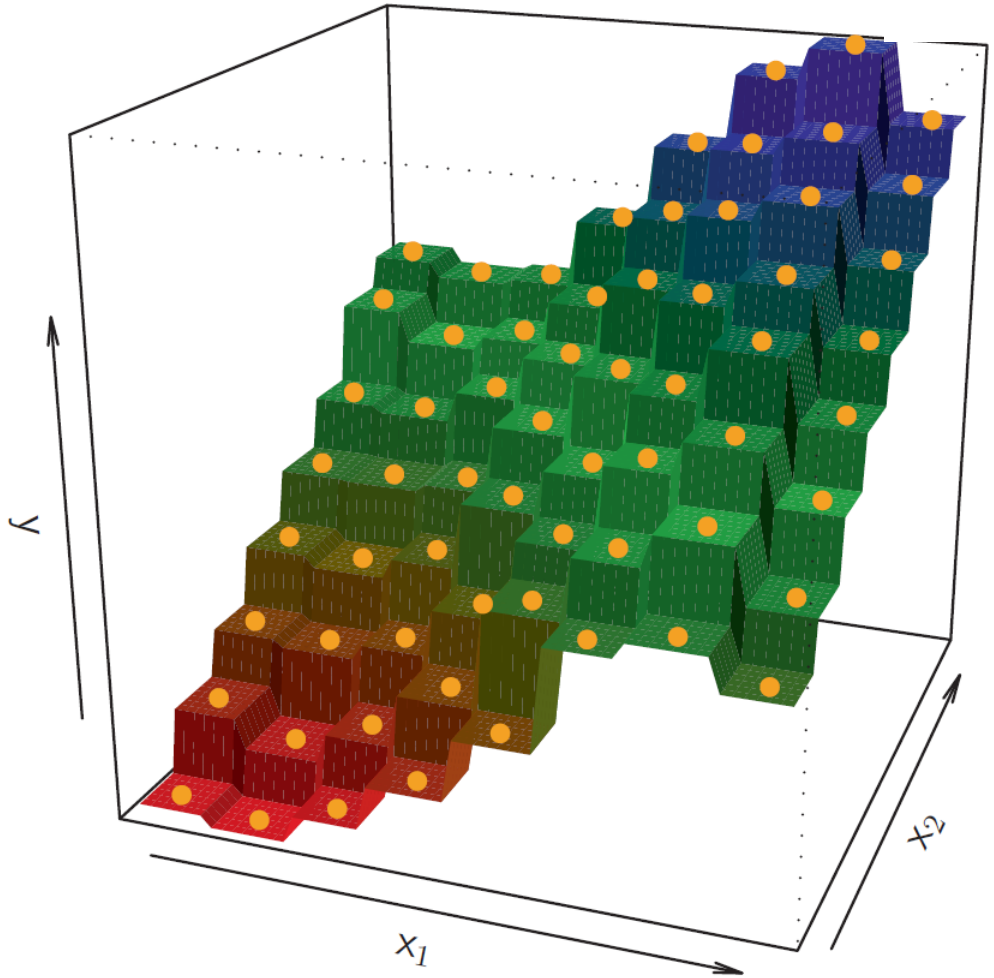
In the `Credit` data, a regression of `balance` on `age`, `rating`, and `limit` indicates that the predictors have VIF values of 1.01, 160.67, and 160.59.

Return to Questions for the Advertising Data

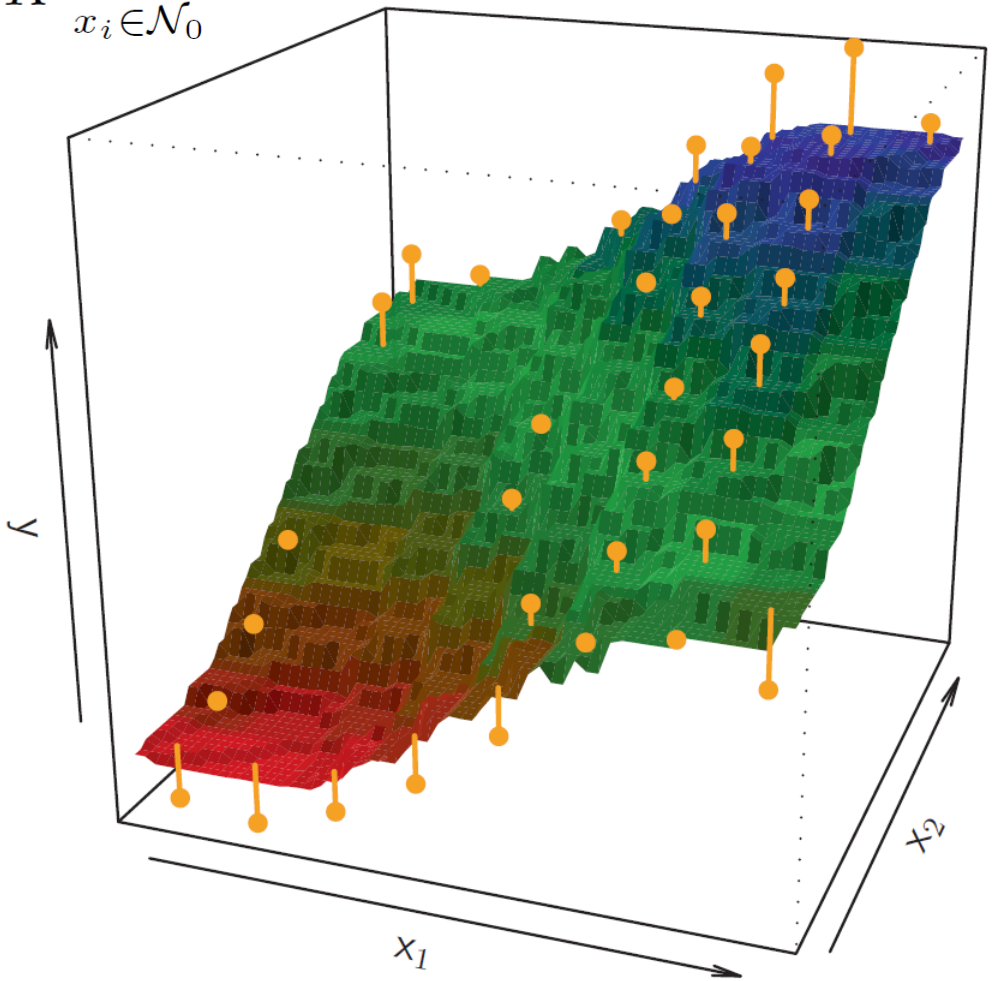
1. Is there a relationship between advertising sales and budget? F test (RSS)
2. How strong is the relationship? RSE; R^2
3. Which media contribute to sales? t test (coefficients)
4. How large is the effect of each medium on sales? confidence interval (coefs)
5. How accurate can we predict future sales? confidence/prediction intervals
6. Is the relationship linear? residual plot
7. Is there synergy among the advertising media? interactions

KNN Regression: Which has higher variance?

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

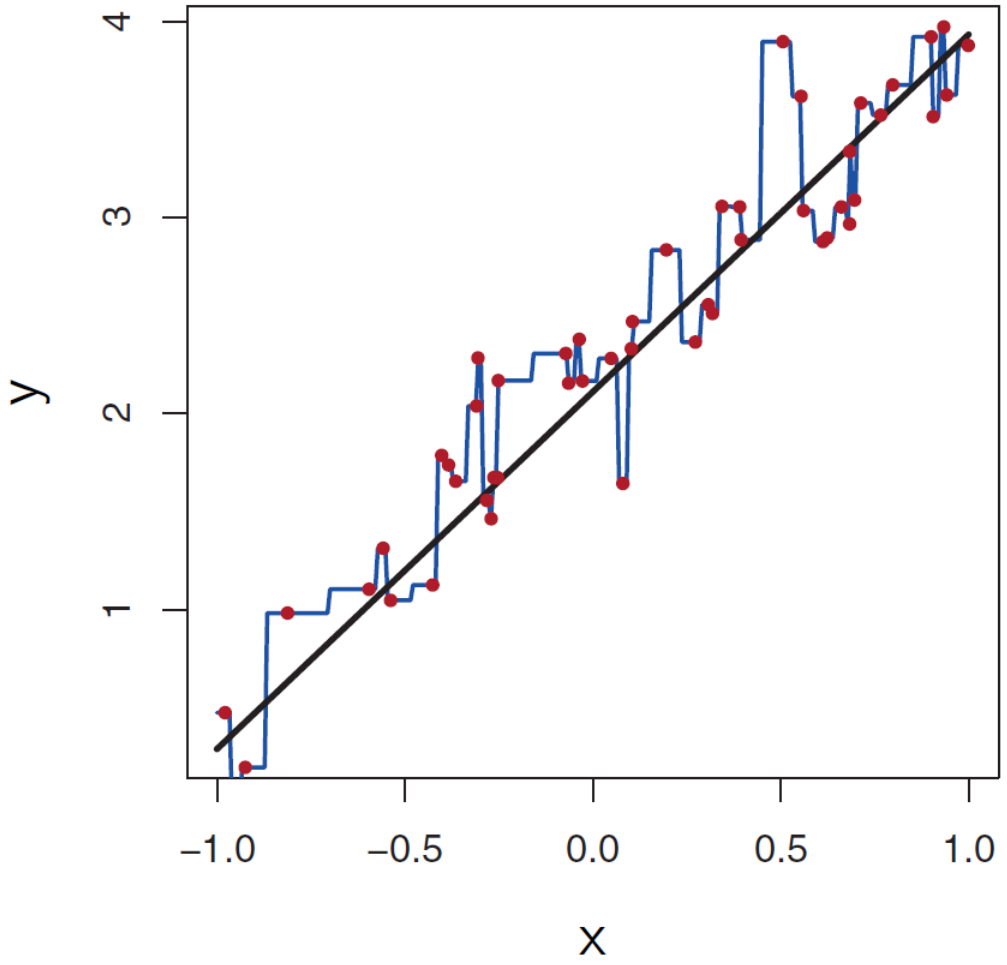


K = 1

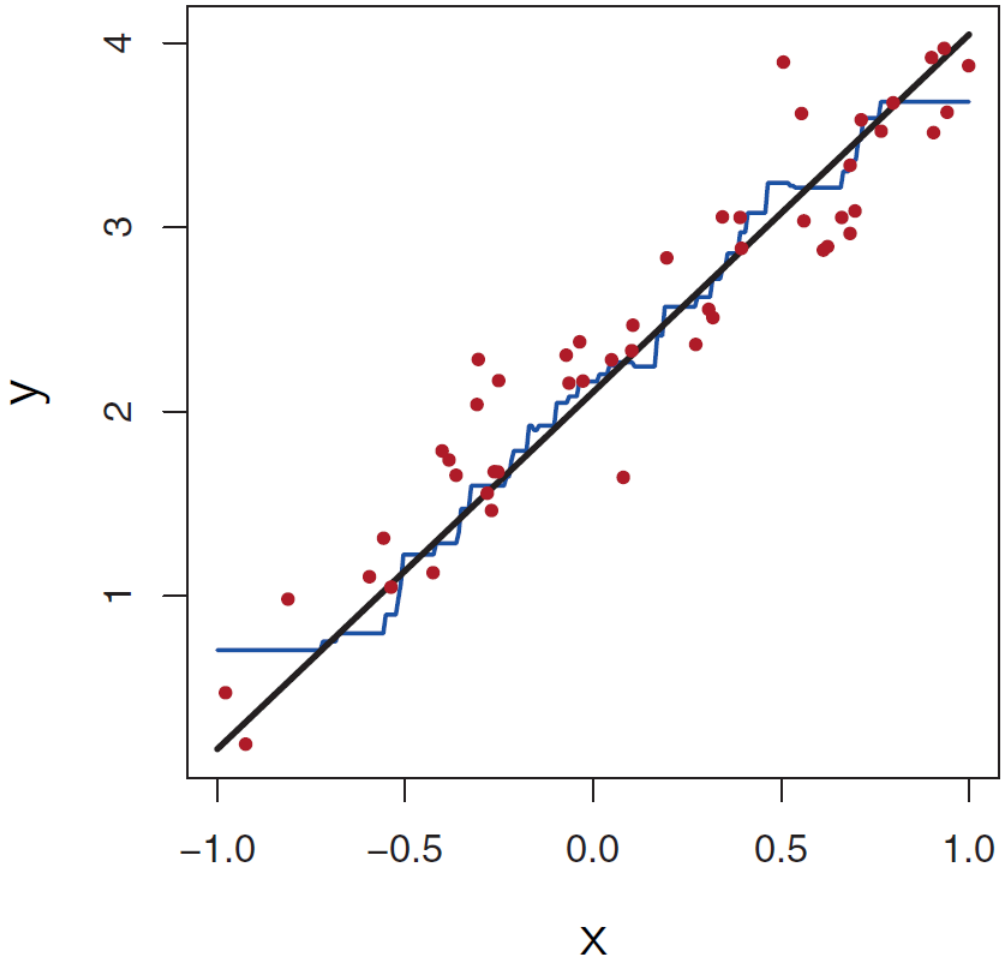


K = 9

KNN Regression: with Only 1 Predictor

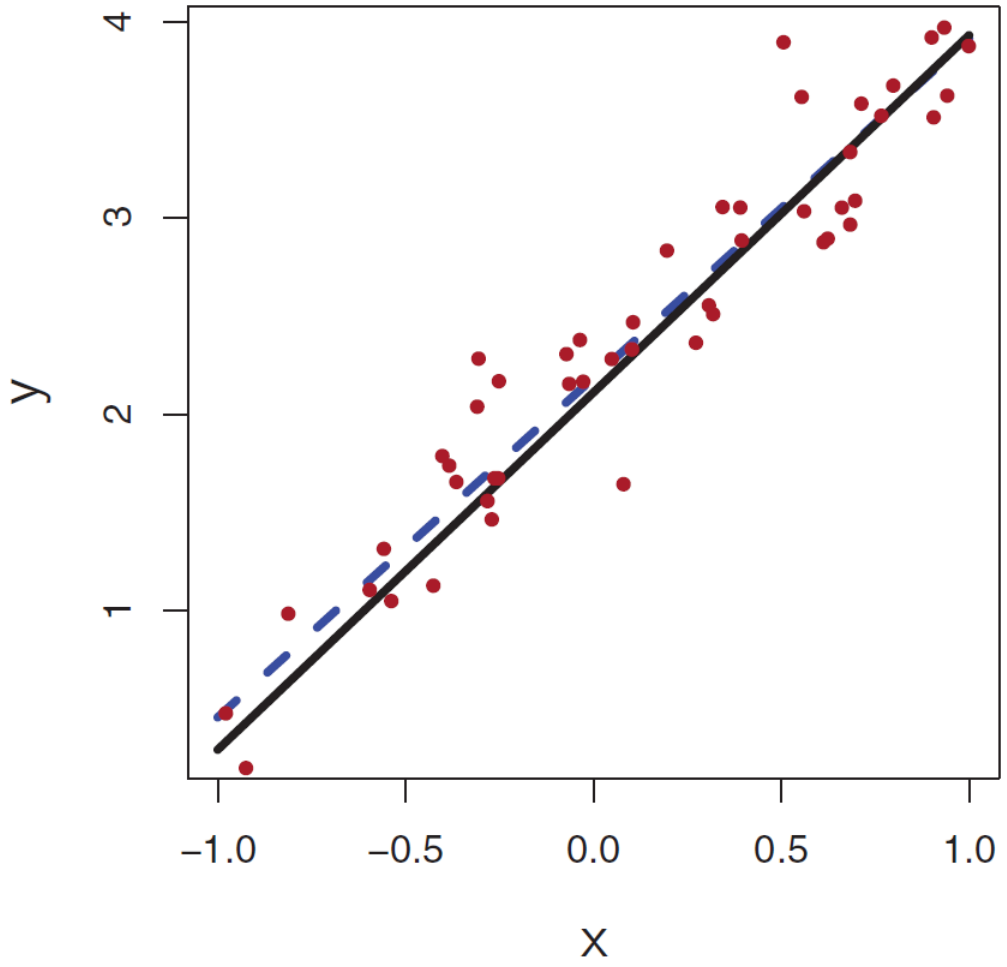


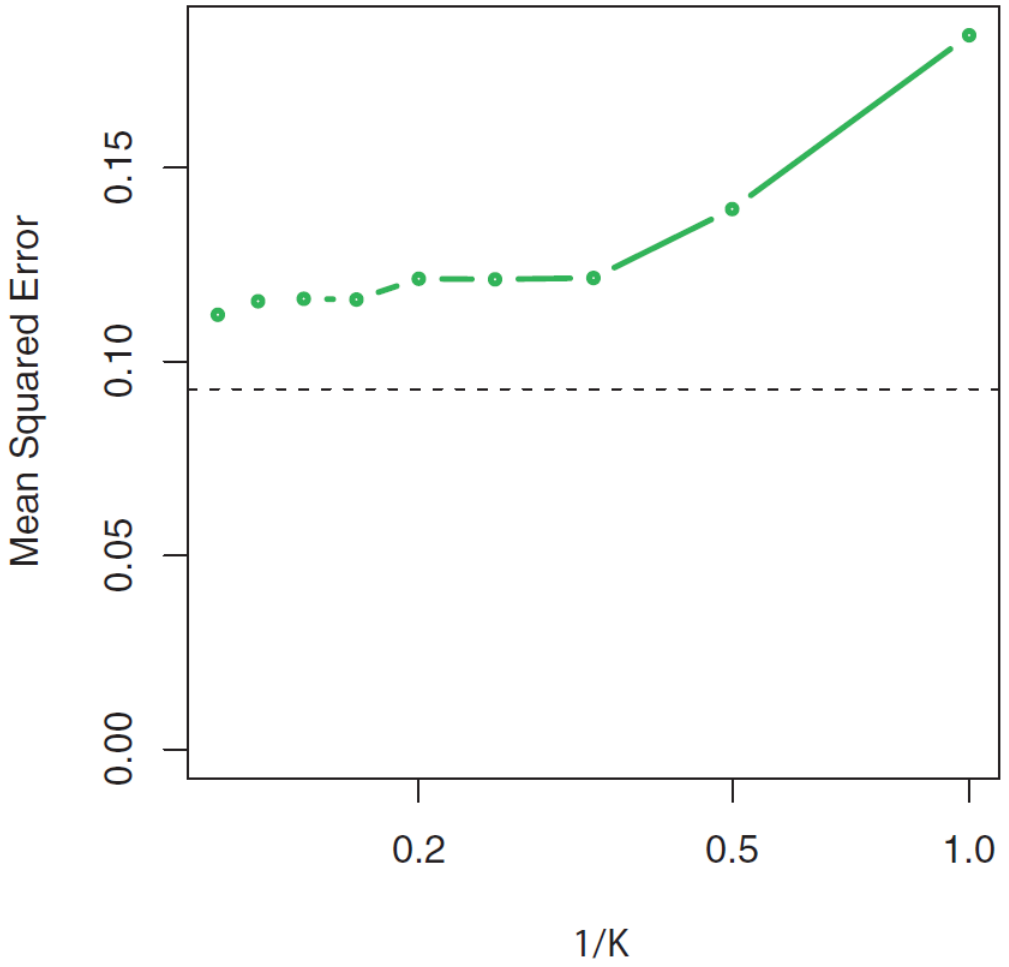
K = 1



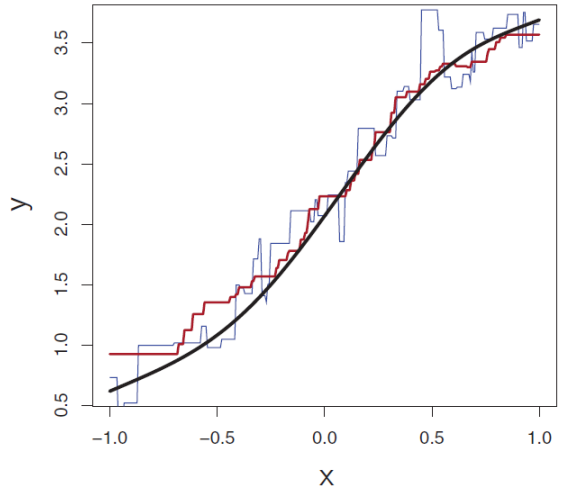
K = 9

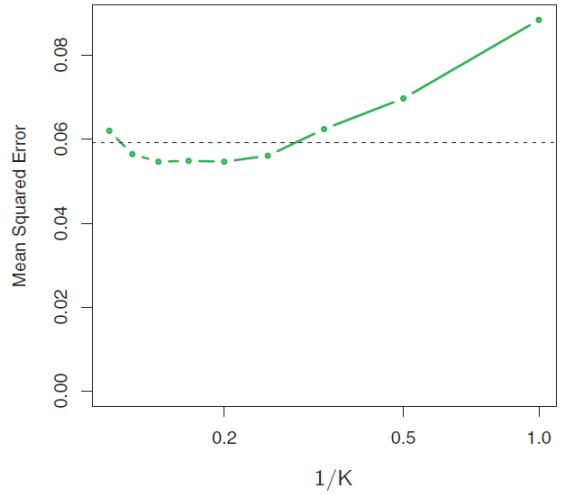
Round 1: Linear Regression versus KNN

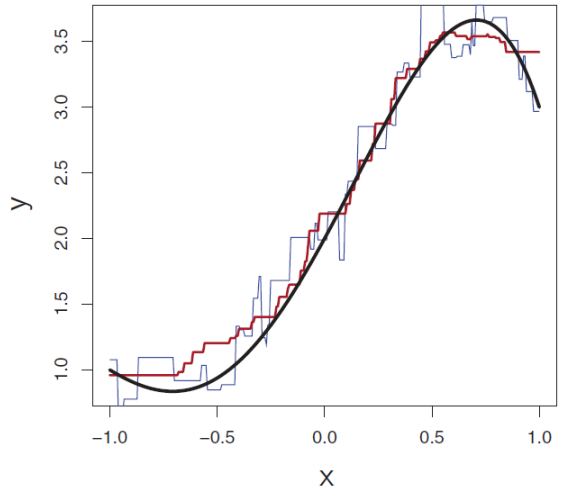


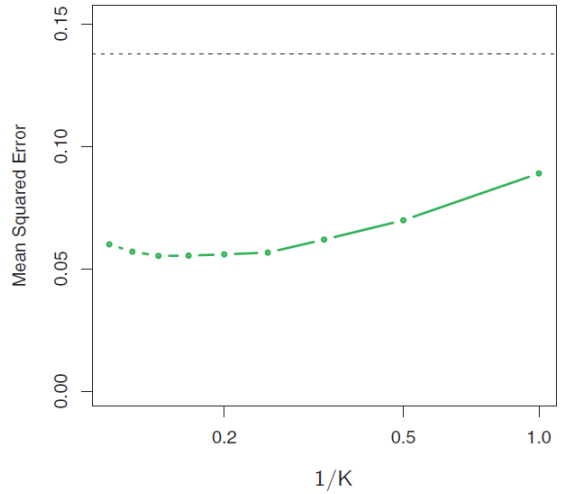


Rounds 2 and 3: Linear Regression v KNN



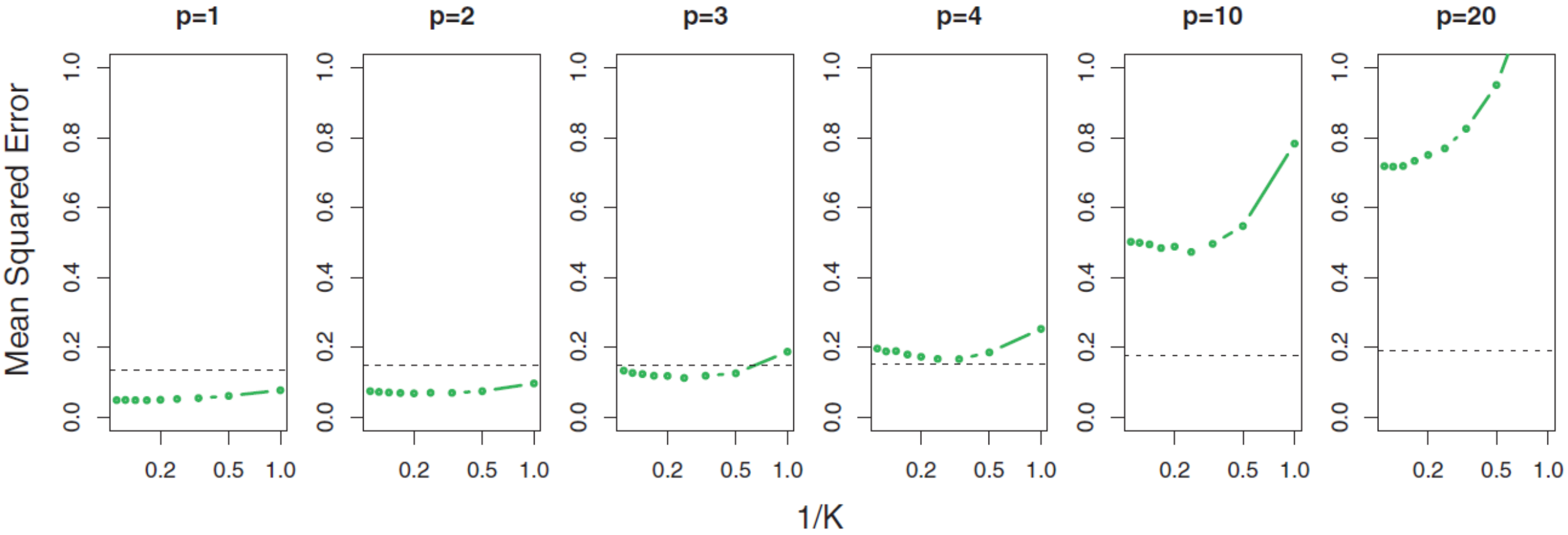






Higher complexity function

Rounds 4 – 8: Linear Regression v KNN



Higher complexity function, but with various quantities of noise

Chain Rule for Gradient Descent

$$-\frac{\partial}{\partial \beta_i}\left(\frac{1}{2}\left(y_i - \hat{f}(x_i)\right)^2\right) = -\frac{\partial}{\partial \hat{f}(x_i)}\left(\frac{1}{2}\left(y_i - \hat{f}(x_i)\right)^2\right)\frac{\partial \hat{f}(x_i)}{\partial \beta_i}$$

- We want to move the weight in the opposite direction of the partial derivative of the loss function with respect to this weight
- See example code near bottom of http://cross-entropy.net/ML210/linear_regression.txt

Gradient for Mean Squared Error Loss

$$-\frac{\partial}{\partial \hat{f}(x_i)}\left(\frac{1}{2}\left(y_i-\hat{f}(x_i)\right)^2\right)=-\frac{1}{2}\frac{\partial}{\partial \hat{f}(x_i)}\left(\left(y_i-\hat{f}(x_i)\right)^2\right)$$

$$=-\frac{1}{2}\left(2*\left(y_i-\hat{f}(x_i)\right)\right)\frac{\partial}{\partial \hat{f}(x_i)}\left(y_i-\hat{f}(x_i)\right)$$

$$=-\left(y_i-\hat{f}(x_i)\right)\frac{\partial}{\partial \hat{f}(x_i)}\left(y_i-\hat{f}(x_i)\right)$$

$$=-\left(y_i-\hat{f}(x_i)\right)*\left(\frac{\partial}{\partial \hat{f}(x_i)}y_i-\frac{\partial}{\partial \hat{f}(x_i)}\hat{f}(x_i)\right)$$

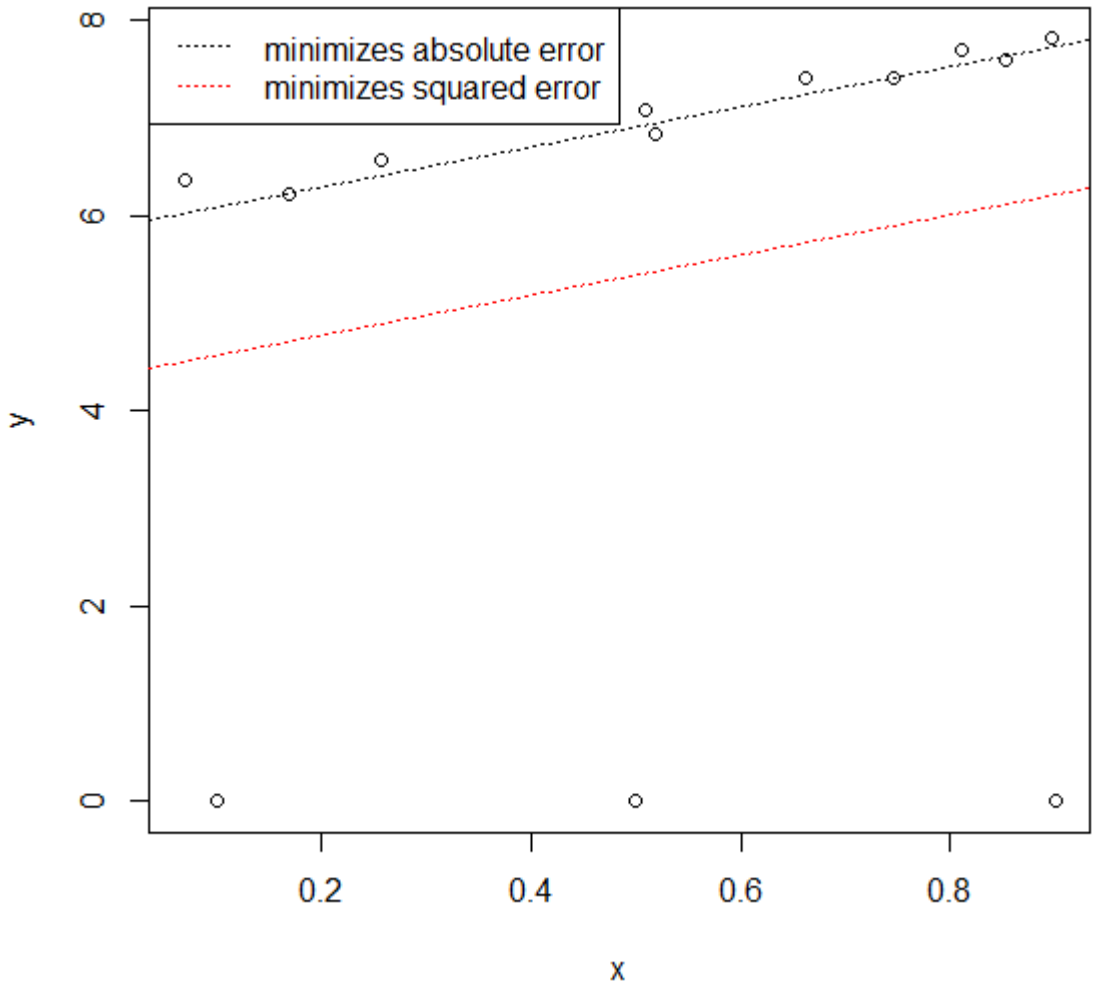
$$=-\left(y_i-\hat{f}(x_i)\right)*(0-1)$$

$$=y_i-\hat{f}(x_i)$$

Robust Regression

- We use Laplacian loss (absolute error) rather than Gaussian loss (squared) error
- A Linear Programming (LP) solver is used to derive the coefficients for Laplacian loss [constrained optimization]

Robust Regression Example



See example code at bottom of http://cross-entropy.net/ML210/linear_regression.txt

Agenda

Homework Review

Probability

Chapter 3

Gradient Descent

Robust Regression