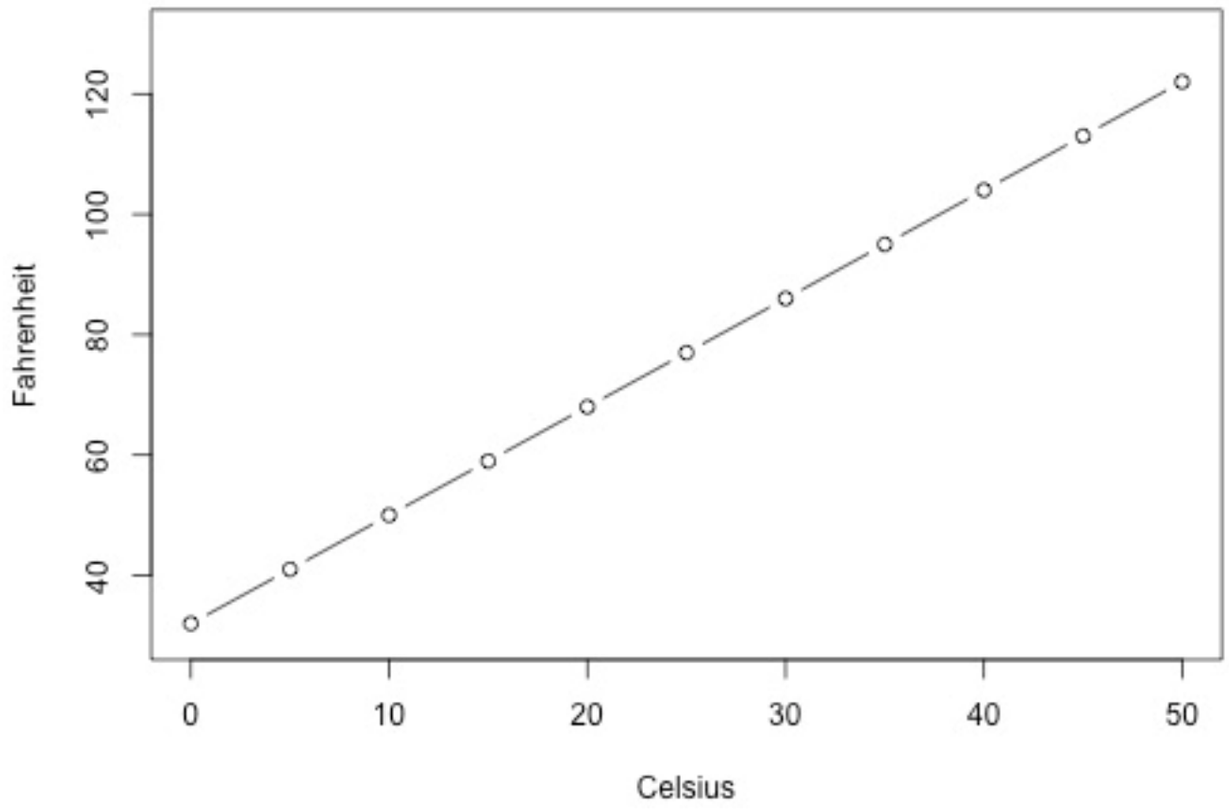
Linear regression

Outline

- Simple linear regression
- Multiple linear regression
- Ridge/lasso regression

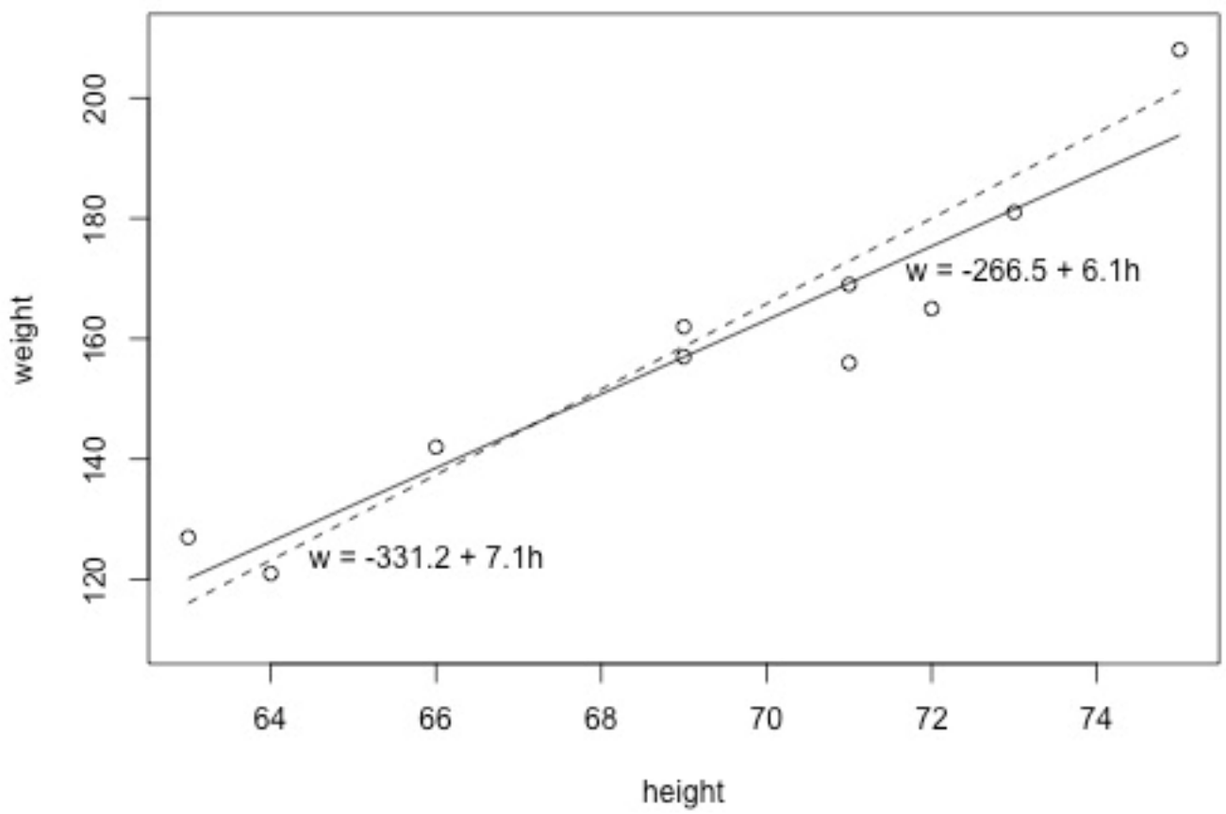
What is Simple Linear Regression?

- supervised machine learning method that uncovers a linear relationship between two continuous variables
  - One variable, x, is regarded as the predictor, features, or independent variable
  - The other variable, y, is regarded as the response, target, or dependent variable



Fahr =9/5Cels+32

What is the "Best Fitting Line"?



- y_i : the observed response for i
- x_i : the predictor value for experimental for i
- $\hat{y}_i$:the predicted response for i

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- β_0 : intercept of the line, the expected value when x is 0
- β_1 : slope of the line, the expected change in y when x shifts by one unit

Minimize the residual or error e_i: $e_i = y_i - \hat{y}_i$

Minimize the residual sum of square (RSS):

$$RSS = \sum_{i=1}^{n} e_i^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

What is the "Best Fitting Line"?

First order derivative of RSS with respect to β is equal to 0

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(2)(-1) \stackrel{!}{=} 0$$

$$\Rightarrow n\hat{\beta}_0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)$$

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

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

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

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

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))(2)(-1)(x_i - \bar{x}) \stackrel{!}{=} 0$$

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

$\bar{x}$: mean of x_i

$\bar{y}$: mean of y_i

variance of x_i

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

covariance between x_i and y_i

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\mathbf{Cov}(x, y)}{\mathbf{Var}(x)}$$

Example

age x	glucose level y	$x_i - \bar{x}$	$y_i - \bar{y}$	$\hat{y}_i$	e_i
43	99	1.83	18	81.70	17.30
21	65	-20.17	-16	73.23	-8.23
25	79	-16.17	-2	74.77	4.23
42	75	0.83	-6	81.32	-6.32
57	87	15.83	6	87.10	-0.10
59	81	17.83	0	87.87	-6.87

How to calculate β_0 , β_1

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

$\bar{x}$ = (43+21+25+42+57+59)/6=41.17

$\bar{y}$ = (99+65+79+75+87+81)/6=81

β_1 =[1.83*18+(-20.17)*(-16)+(-16.17)*(-2)+0.83*(-6)+15.83*6+17.83*0]/[1.83**2+(-20.17)**2+(-16.17)**2+0.83**2+15.83**2+17.83**2]
=478/1240.8334=0.38522

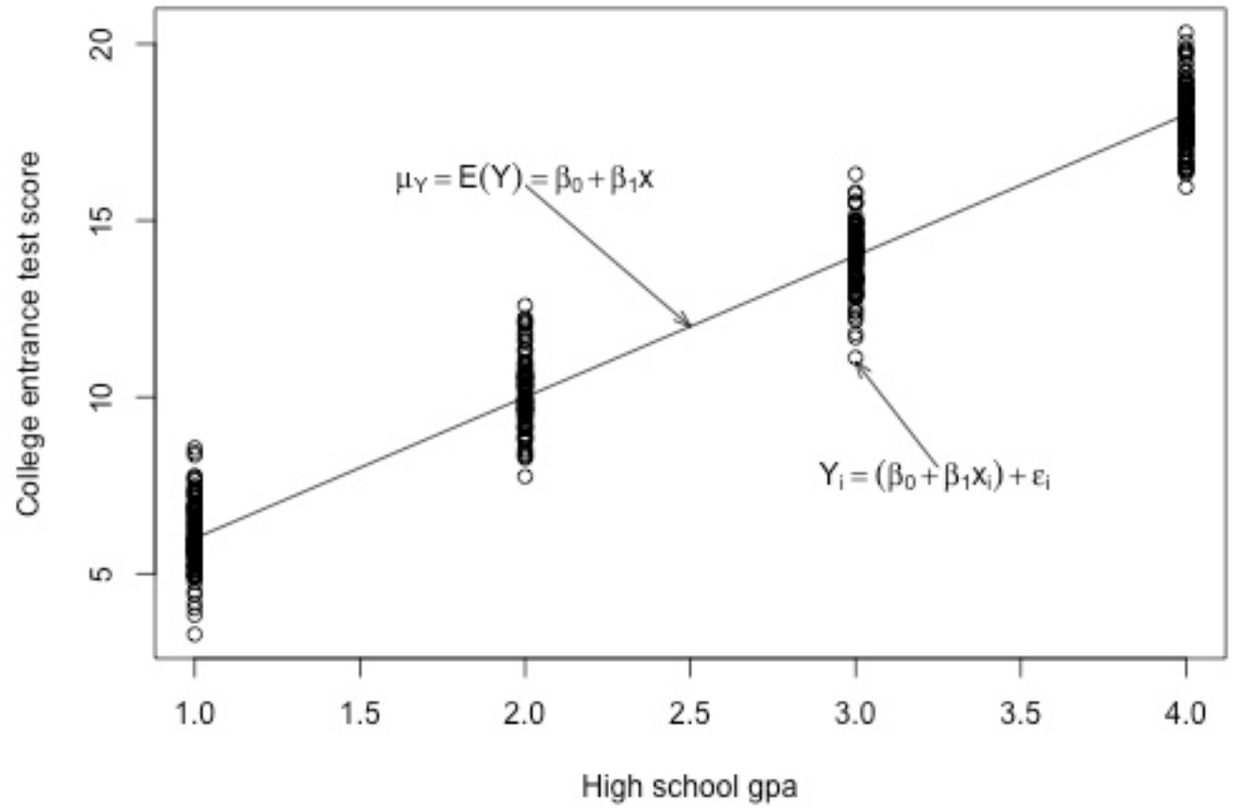
β_0 =81-0.38522*41.17=65.14

$\hat{y}_i$ =65.14+0.38522 x_i

RSS=17.30**2+(-8.23)**2+4.23**2+(-6.32)**2+(-0.10)**2+(-6.87)**2=472.0651

Assumption of simple linear regression

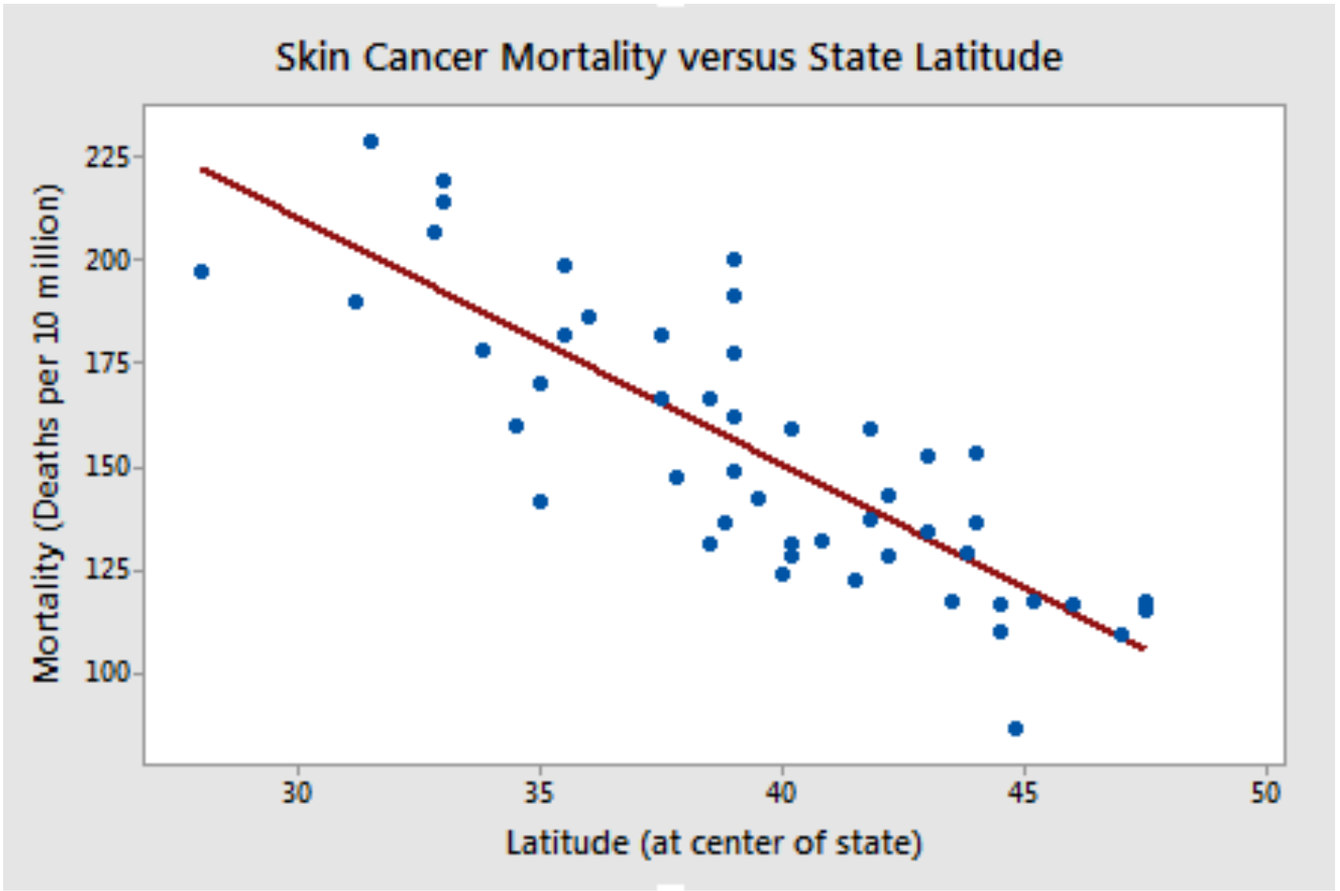
$$Y = \beta_0 + \beta_1 X + \epsilon$$



- Linearity
  - The mean of the response, $\hat{y}_i$, at each value of the predictor, x_i, is a Linear function of the x_i.
- Independent errors
  - The errors, e_i, are Independent.
- Normality
  - The errors, e_i, at each value of the predictor, x_i, are Normally distributed.
- Constant variance
  - The errors, e_i, at each value of the predictor, x_i, have Equal variances (denoted σ^2).

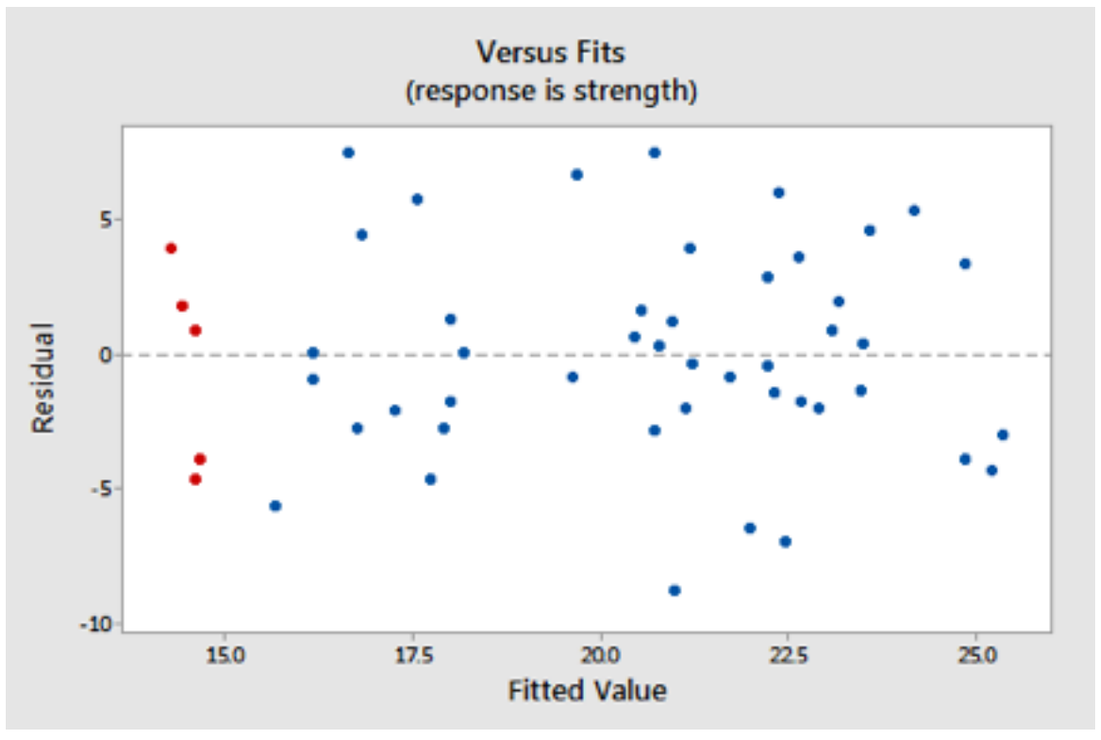
Linearity: Visual Inspection

This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables.

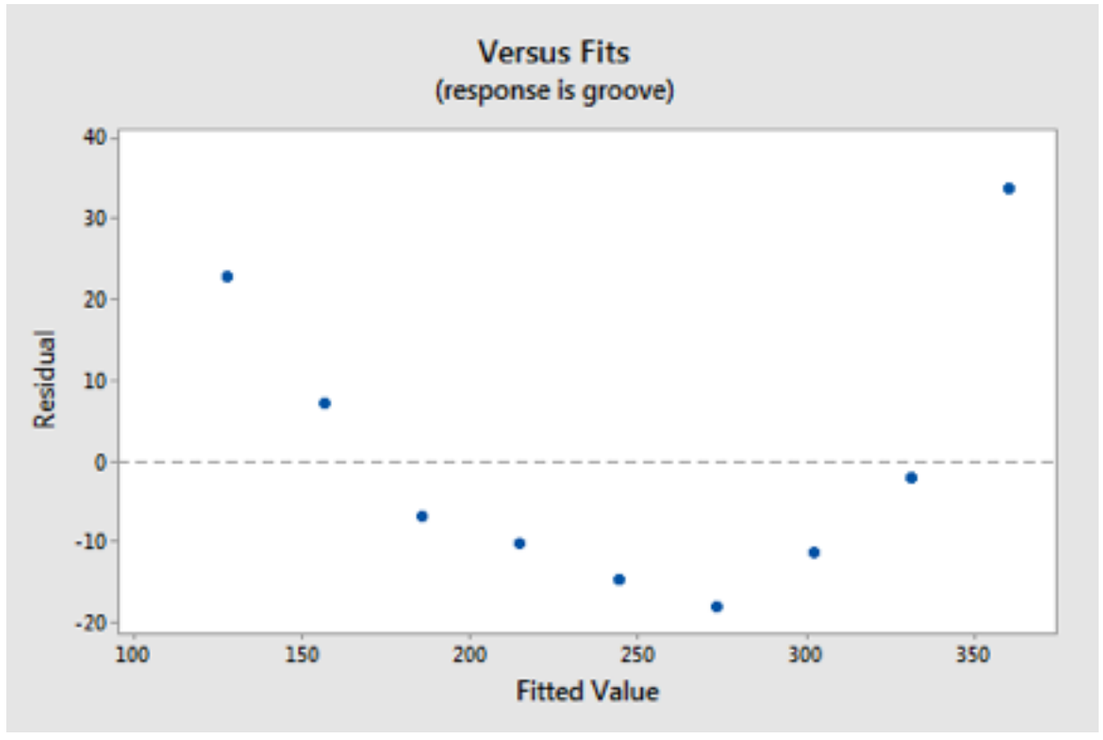


Constant variance: residual plot

This means that different response variables have the same variance in their errors, regardless of the values of the predictor variables
a scatterplot of residual values versus fitted values



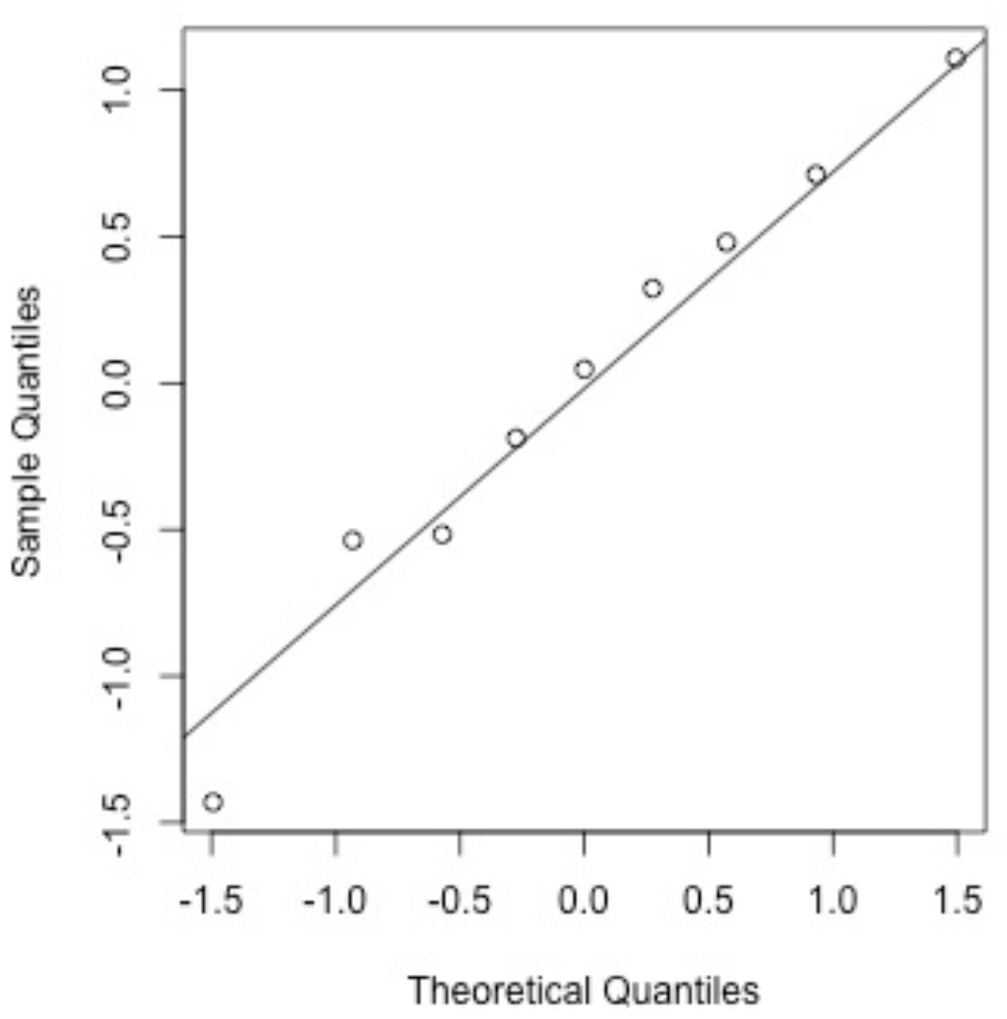
well-behaved



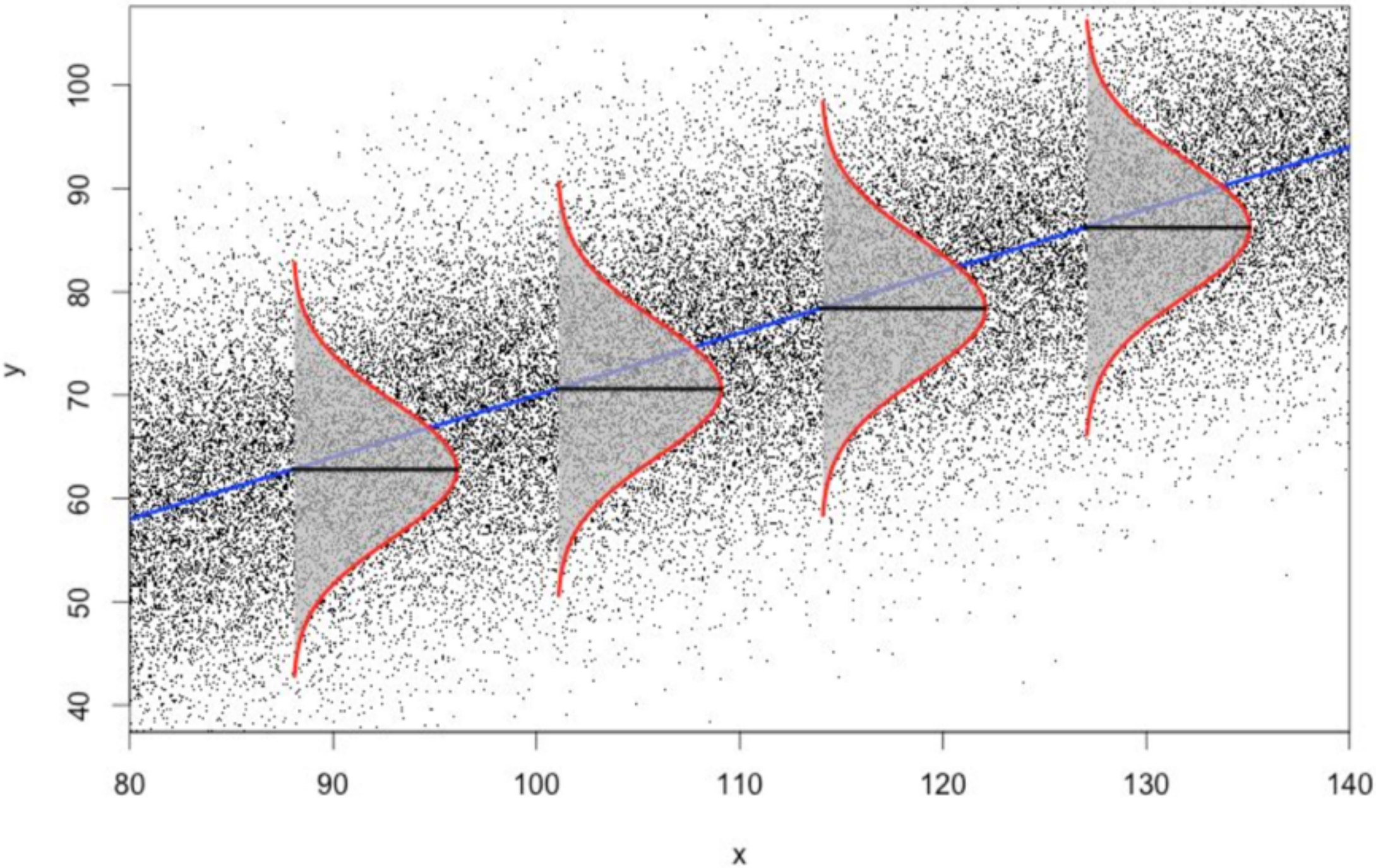
violation

Normality:quantile-quantile plot of the residuals

a scatterplot of residual values versus the corresponding normal theoretical values that preserve the observed quantile



Visualization of All Assumption



Model Assessment

- residual standard error: RSE

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}}$$

The lower of RSE, the better performance of the model

- coefficient of determination R^2
  - total sum of squares

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

  - TSS: the total squared deviation of the response value from its mean

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

  - greater R^2: a better fit
  - R^2: between 0 and 1
  - R^2: how much percentage of the variability in the response variable is explained by the predictors variable

Example

age x	glucose level y	$x_i - \bar{x}$	$y_i - \bar{y}$	$\hat{y}_i$	e_i
43	99	1.83	18	81.70	17.30
21	65	-20.17	-16	73.23	-8.23
25	79	-16.17	-2	74.77	4.23
42	75	0.83	-6	81.32	-6.32
57	87	15.83	6	87.10	-0.10
59	81	17.83	0	87.87	-6.87

RSS=472.0651

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}}$$

RSE=[472.0651/(6-2)]**0.5=10.86

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

TSS=[18**2+(16)**2+(-2)**2+(-6)**2+6**2+0**2]
=656

RSS=472.0651

R^2 =1-472.0651/656=0.28

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

28% of the variability in y is explained by x

What is Multiple Linear Regression?

- supervised machine learning method that uncovers a linear relationship between a set of variables x_i and a single outcome variable y
  - The predictor, features, or independent variables: x_i
  - The response, target, or dependent variable: y

$\beta_0 * 1$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip} \quad \Rightarrow \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

- Minimize RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

First order derivative of RSS with respect to β is equal to 0

square of matrix is equal to matrix multiplied by its transpose

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

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \stackrel{!}{=} 0$$

$$\Rightarrow \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

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

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

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

A Matrix Formulation of the Multiple Regression Model

y: n-dimensional vector $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ Y_n \end{pmatrix}$

x_p: n*(p+1) matrix $X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots X_{1p} \\ 1 & X_{21} & X_{22} & \cdots X_{2p} \\ & & \cdots & \\ & & \cdots & \\ 1 & X_{n1} & X_{n2} & \cdots X_{np} \end{pmatrix}$

β: (p+1)-dimensional vector $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \beta_p \end{pmatrix}$

e: n-dimensional vector $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \epsilon_n \end{pmatrix}$

$$e = y - \hat{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

$$y = \beta x + e$$

Assumption of multiple linear regression

- Linearity
- Independent errors
- Normality
- Constant variance
- No multicollinearity between the predictor variables
  - the predictor variables are unrelated with each other
  - correlations plot of predictor variables
  - VIF: variance inflation factor of each predictor

    example: two variables x1, x2

$$VIF_{x_1x_2} = \frac{1}{1-\rho_{12}^2}$$

  - ρ: correlations between variable x1 and x2
  - Higher correlations, higher VIF, higher uncertainty of the coefficient estimates
  - 1 = not correlated.
  - Between 1 and 5 = moderately correlated.
  - Greater than 5 = highly correlated.
  - VIF>5: remove this predictor from the model

How to calculate VIF for more than 2 variables? Please check this link:
https://online.stat.psu.edu/stat462/node/180/

Model Evaluation: RSE and R^2_{Adj}

RSE:

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

n: number of observations
p: number of predictors
lower RSE: a better fit

- Problem 1: Every time we add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently, a model with more predictors may appear to have a better fit simply because it has more features.
- Problem 2: If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as over-fitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions.

- To avoid adding irrelevant predictor variables, R^2_{Adj}

$$R^2_{Adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

- Adjusted R^2 is always positive. Ranges from 0 to 1 with values closer to 1 indicating a stronger relationship
- Adjusted R^2 is the value of R^2 which has been penalized for the number of variables added to the model, so Adjusted R^2 is always smaller than R^2

Variable/Model Selection

How do we decide what variable to include?

To find balance over- and under-fitting in our modeled relationships

– We want a model that is as simple as possible, but no simpler
– A reasonable model power is traded off against model size
– AIC/BIC measures the balance of this for us

- AIC (Akaike information criterion):balance the fit of the model and complexity of the model
  - The purpose of MLR is prediction:

$$AIC = -2\ln(L) + 2p$$

    - L: Log-likelihood is a measure of model fit. The higher the number, the better the fit. This is usually obtained from statistical output.
    - P: number of predictors in the model
    - Smaller AIC, better model

- BIC(Bayesian Information Criterion): balance the fit of the model and complexity of the model
  - The purpose of MLR is descriptive (such as find the most meaningful relative predictors that influence the target):

$$BIC = -2\ln(L) + p\ln(n)$$

    - n: number of observations
    - Smaller BIC, better model

Stepwise procedure

- Forward selection
  - Start with a null model with just one intercept, then sequentially add the predictor that most improves the fit

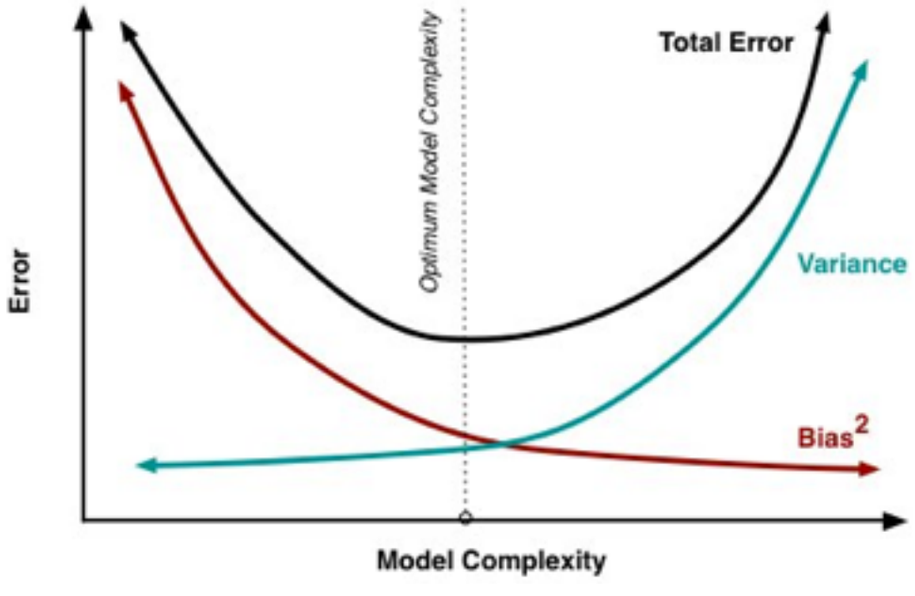
- Backward selection
  - Start with a model that includes all predictors, then sequentially remove the predictor that has the least impact on the fit

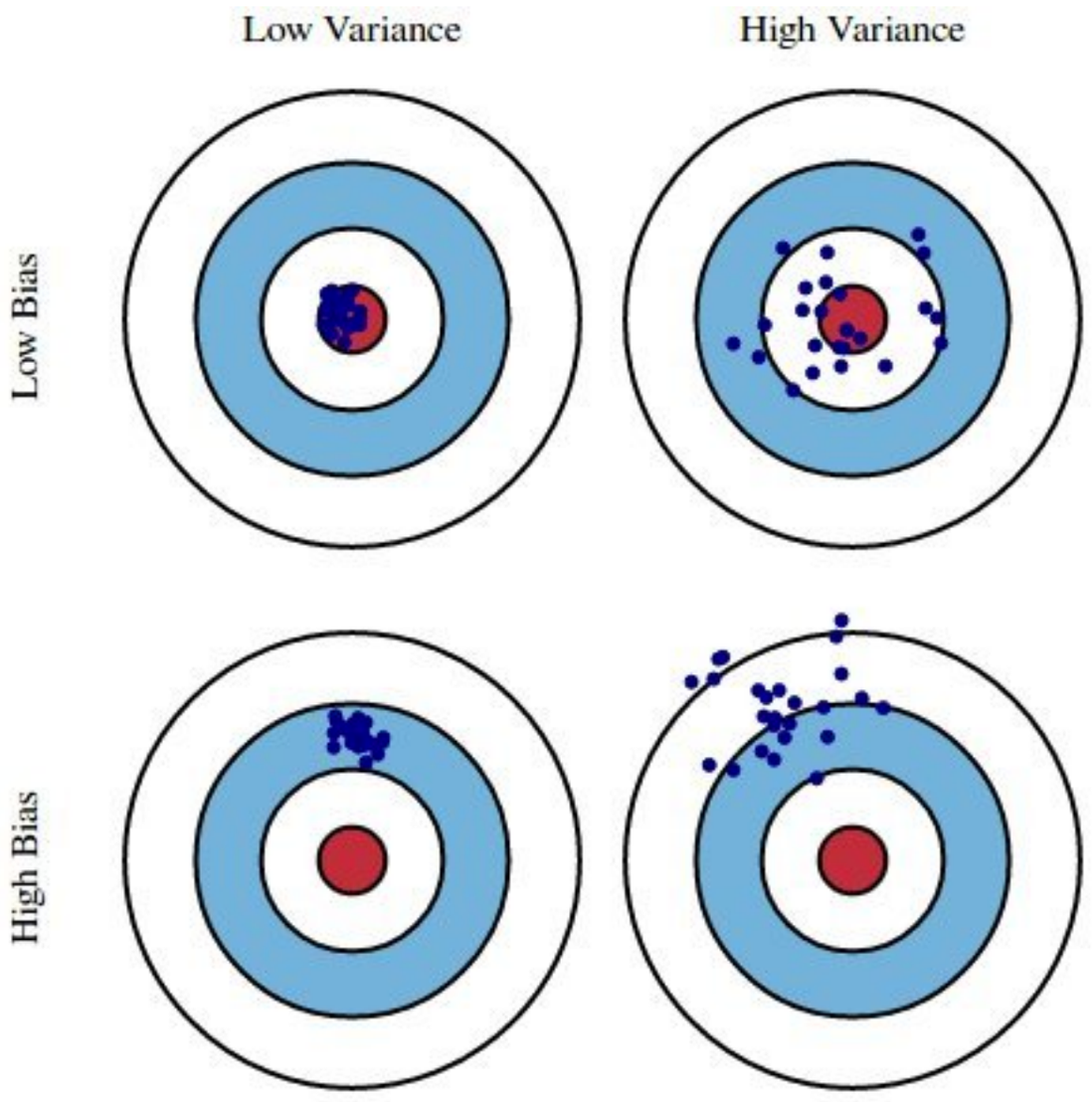
- Both selection
  - Start with either a model that only has an intercept term or a model that includes all predictors, then sequentially add or remove the predictor that that has the most/least impact on the model, respectively

- Select a model based on
  - AIC
  - BIC

Trade-off between Bias and Variance

- Error due to Bias: The error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. Bias measures how far off in general these models' predictions are from the correct value.
- Error due to Variance: The error due to variance is taken as the variability of a model prediction for a given data point. The variance is how much the predictions for a given point vary between different realizations of the model.





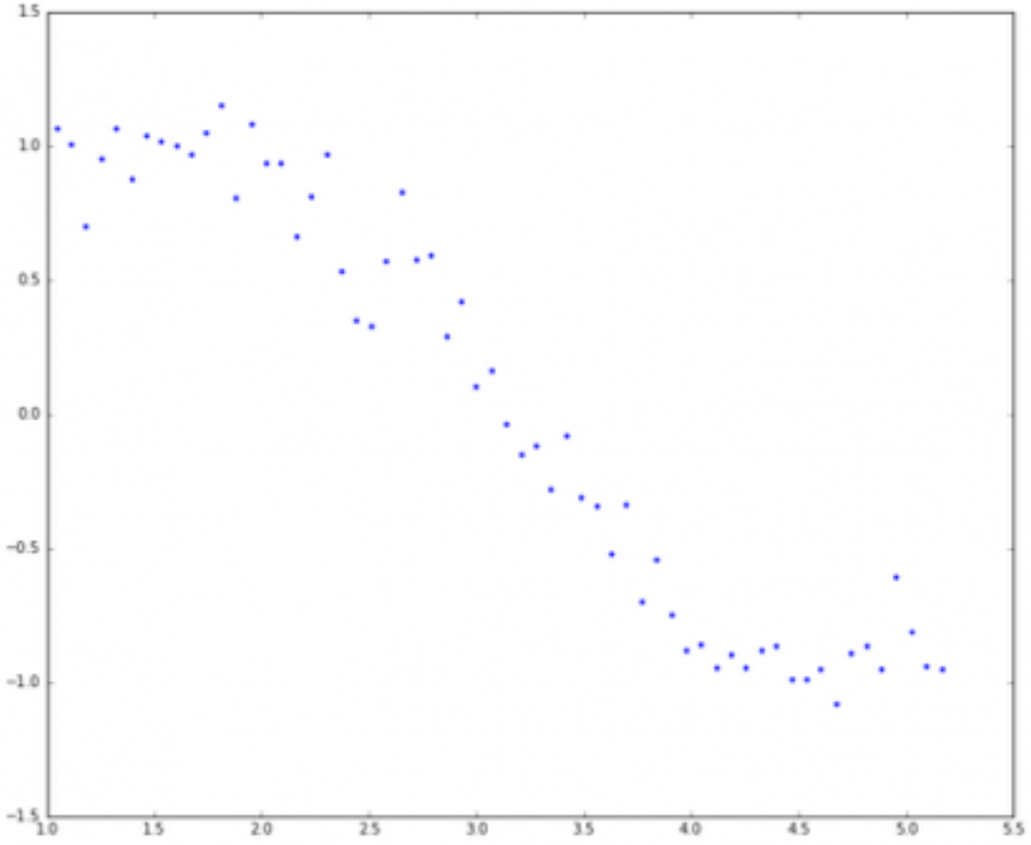
- Low bias (high accuracy)
- Low variance (high precision)

Ridge/Lasso

- Ridge and Lasso regression are powerful techniques generally used for creating parsimonious models in presence of a ‘large’ number of features. Here ‘large’ can typically mean either of two things:
  - Large enough to enhance the tendency of a model to overfit (as low as 10 variables might cause overfitting)
  - Large enough to cause computational challenges. With modern systems, this situation might arise in case of millions or billions of features

- Overfitting problem
  - Lets try to estimate y using polynomial regression with powers of x from 1 to 15 (n=15)

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \varepsilon.$$



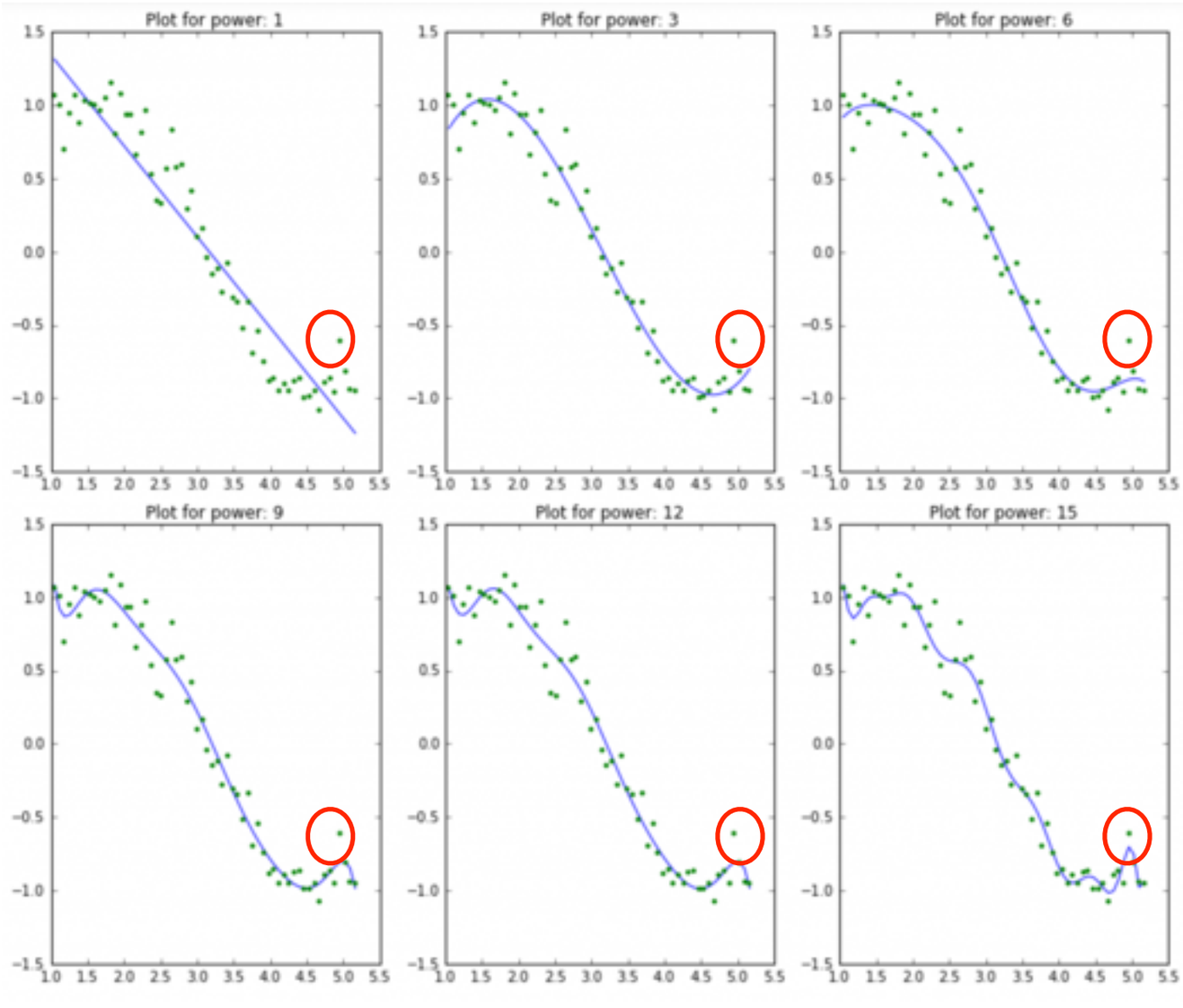
	x	y	x_2	x_3	x_4	x_5	x_6
0	1.047198	1.065763	1.096623	1.148381	1.202581	1.259340	1.318778
1	1.117011	1.006086	1.247713	1.393709	1.556788	1.738948	1.942424
2	1.186824	0.695374	1.408551	1.671702	1.984016	2.354677	2.794587
3	1.256637	0.949799	1.579137	1.984402	2.493673	3.133642	3.937850
4	1.326450	1.063496	1.759470	2.333850	3.095735	4.106339	5.446854

	x_7	x_8	x_9	x_10	x_11	x_12	x_13
0	1.381021	1.446202	1.514459	1.585938	1.660790	1.739176	1.821260
1	2.169709	2.423588	2.707173	3.023942	3.377775	3.773011	4.214494
2	3.316683	3.936319	4.671717	5.544505	6.580351	7.809718	9.268760
3	4.948448	6.218404	7.814277	9.819710	12.339811	15.506664	19.486248
4	7.224981	9.583578	12.712139	16.862020	22.366630	29.668222	39.353420

	x_14	x_15
0	1.907219	1.997235
1	4.707635	5.258479
2	11.000386	13.055521
3	24.487142	30.771450
4	52.200353	69.241170

Why Penalize the Magnitude of Coefficients?

- Lets make 15 different linear regression models with each model containing variables with powers of x from 1 to 15
- We would expect the models with increasing complexity to better fit the data and result in lower RSS values. This can be verified by looking at the plots generated for 6 models.
- As the model complexity increases, the models tends to fit even smaller deviations in the training data set. Though this leads to overfitting, lets keep this issue aside for some time and come to our main objective, i.e. the impact on the magnitude of coefficients.



	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8	coef_x_9	coef_x_10	coef_x_11	c
model_pow_1	3.3	2	-0.62	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
model_pow_2	3.3	1.9	-0.58	-0.006	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
model_pow_3	1.1	-1.1	3	-1.3	0.14	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
model_pow_4	1.1	-0.27	1.7	-0.53	-0.036	0.014	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
model_pow_5	1	3	-5.1	4.7	-1.9	0.33	-0.021	NaN	NaN	NaN	NaN	NaN	NaN	N
model_pow_6	0.99	-2.8	9.5	-9.7	5.2	-1.6	0.23	-0.014	NaN	NaN	NaN	NaN	NaN	N
model_pow_7	0.93	19	-56	69	-45	17	-3.5	0.4	-0.019	NaN	NaN	NaN	NaN	N
model_pow_8	0.92	43	-1.4e+02	1.8e+02	-1.3e+02	58	-15	2.4	-0.21	0.0077	NaN	NaN	NaN	N
model_pow_9	0.87	1.7e+02	-6.1e+02	9.6e+02	-8.5e+02	4.6e+02	-1.6e+02	37	-5.2	0.42	-0.015	NaN	NaN	N
model_pow_10	0.87	1.4e+02	-4.9e+02	7.3e+02	-6e+02	2.9e+02	-87	15	-0.81	-0.14	0.026	-0.0013	NaN	N
model_pow_11	0.87	-75	5.1e+02	-1.3e+03	1.9e+03	-1.6e+03	9.1e+02	-3.5e+02	91	-16	1.8	-0.12	0.0034	N
model_pow_12	0.87	-3.4e+02	1.9e+03	-4.4e+03	6e+03	-5.2e+03	3.1e+03	-1.3e+03	3.8e+02	-80	12	-1.1	0.062	-(
model_pow_13	0.86	3.2e+03	-1.8e+04	4.5e+04	-6.7e+04	6.6e+04	-4.6e+04	2.3e+04	-8.5e+03	2.3e+03	-4.5e+02	62	-5.7	0
model_pow_14	0.79	2.4e+04	-1.4e+05	3.8e+05	-6.1e+05	6.6e+05	-5e+05	2.8e+05	-1.2e+05	3.7e+04	-8.5e+03	1.5e+03	-1.8e+02	1
model_pow_15	0.7	-3.6e+04	2.4e+05	-7.5e+05	1.4e+06	-1.7e+06	1.5e+06	-1e+06	5e+05	-1.9e+05	5.4e+04	-1.2e+04	1.9e+03	-(

- the size of coefficients increase exponentially with increase in model complexity. That is why putting a constraint on the magnitude of coefficients can be a good idea to reduce model complexity.

Shrinkage/Regularization

- What does a large coefficient signify? It means that we're putting a lot of emphasis on that feature, i.e. the particular feature is a good predictor for the outcome. When it becomes too large, the algorithm starts modeling intricate relations to estimate the output and ends up overfitting to the particular training data.
- Ridge/Lasso fit a model with all the predictors and shrink the coefficient estimates to 0 relative to the least square estimates
  - reduce estimated variance of coefficients
  - perform variable selection
- Ridge regression: balance the trade-off between fitness and complexity of the model by regularizing the coefficients (l2 penalty)
- python: alpha α, R: lambda λ

objective: minimize the following function

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- λ is the shrinkage parameter
- λ controls the size of the coefficients
- λ controls amount of regularization
- As λ ↓ 0, we obtain multiple linear regression
- As λ ↑ ∞, we have βˆ =0 (intercept-only model)

First order derivative of objective with respect to β is equal to 0

$$\hat{\beta}_{\lambda}^{\text{ridge}} = (\mathbf{Z}^{\top}\mathbf{Z} + \lambda \mathbf{I}_p)^{-1}\mathbf{Z}^{\top}\mathbf{y}$$

- Z is standardized X
- y is centered
- There is a trade-off between the penalty term and RSS. Maybe a large β would give you a better RSS but then it will push the penalty term higher. This is why WE might actually prefer smaller β with worse RSS

Standardization

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

$$X_{changed} = \frac{X - \mu}{\sigma}$$

$$z_i=(x_i-\mu)/\sigma$$

$y_i=y_i$- mean value of y_i

$$\hat{\beta}_{\lambda}^{\text{ridge}} = (\mathbf{Z}^{\top}\mathbf{Z} + \lambda \mathbf{I}_p)^{-1}\mathbf{Z}^{\top}\mathbf{y}$$

Ridge/Lasso

- Lasso: l1 penalty

objective: minimize the following function

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

First order derivative of objective with respect to β is equal to 0

$$\hat{\beta}^{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\beta}_j^{\text{lasso}}(\lambda_1) = \begin{cases} \hat{\beta}_j^{\text{OLS}} - \frac{1}{2}\lambda_1 & \text{if} \quad \beta_j > 0 \\ \hat{\beta}_j^{\text{OLS}} + \frac{1}{2}\lambda_1 & \text{if} \quad \beta_j < 0 \end{cases}$$

The ellipses correspond to the contours of RSS.
The circle corresponds to the ridge constraint.
The diamond corresponds to the lasso constraint.
Minimize the ellipse size and circle/diamond simultaneously.
The coefficient estimate is given by the point at which the ellipse and the circle /diamond touch.
For the lasso, there are "corners" in the diamond.
If the ellipses hit one of these corners, then the coefficient corresponding to the axis is shrunk to zero.
While the circle does not have “corners” and ellipses never hit, then the coefficient corresponding to the axis is never shrunk to zero.

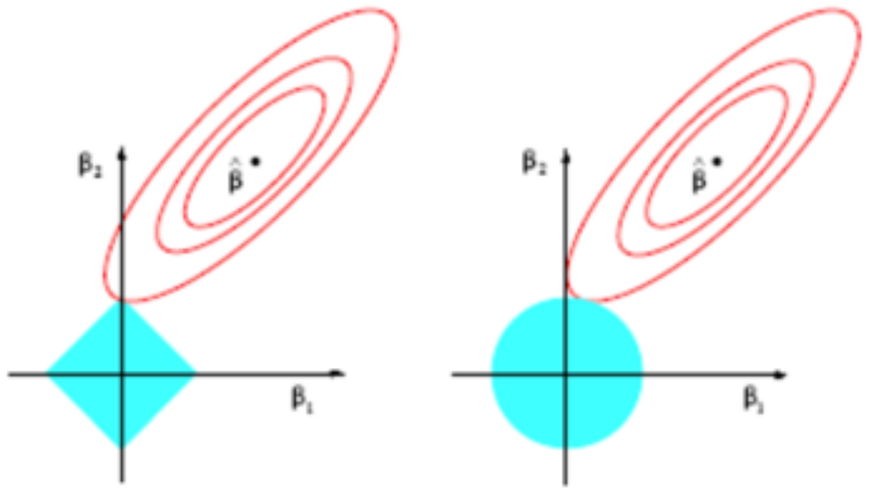


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

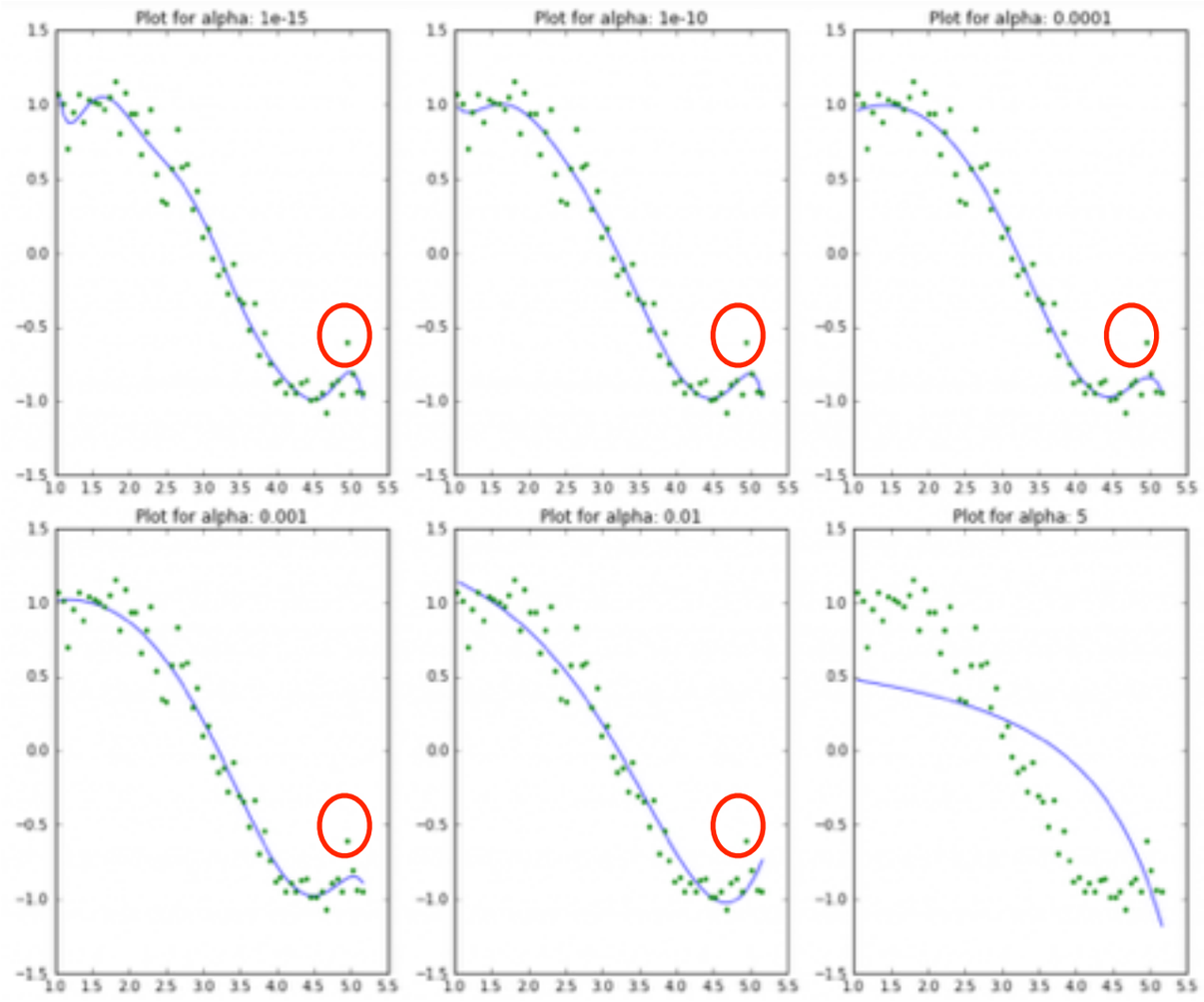
Ridge/Lasso

- λ is the parameter which balances the amount of emphasis given to minimizing RSS vs minimizing sum of square of coefficients or absolute value of coefficients . λ can take various values:
  - λ= 0:
    - The objective becomes same as multiple linear regression.
    - We'll get the same coefficients as multiple linear regression.
  - λ = ∞:
    - The coefficients will be zero. Because of infinite weightage on square of coefficients, anything less than zero will make the objective infinite.
  - 0 < λ < ∞:
    - The magnitude of λ will decide the weightage given to different parts of objective.

- Coefficients β_i
  - The ridge coefficients are a reduced factor of the multiple linear regression coefficients and thus never attain zero values but very small values
  - The lasso coefficients become zero in a certain range and are reduced by a constant factor, which explains lower magnitude in comparison to ridge.

Ridge

- As the value of λ increases, the model complexity reduces. Though higher values of λ reduce overfitting, significantly high values can cause underfitting as well (eg. λ = 5). Thus λ should be chosen wisely. A widely accept technique is cross-validation, i.e. the value of λ is iterated over a range of values and the one giving higher cross-validation score is chosen.



ridge

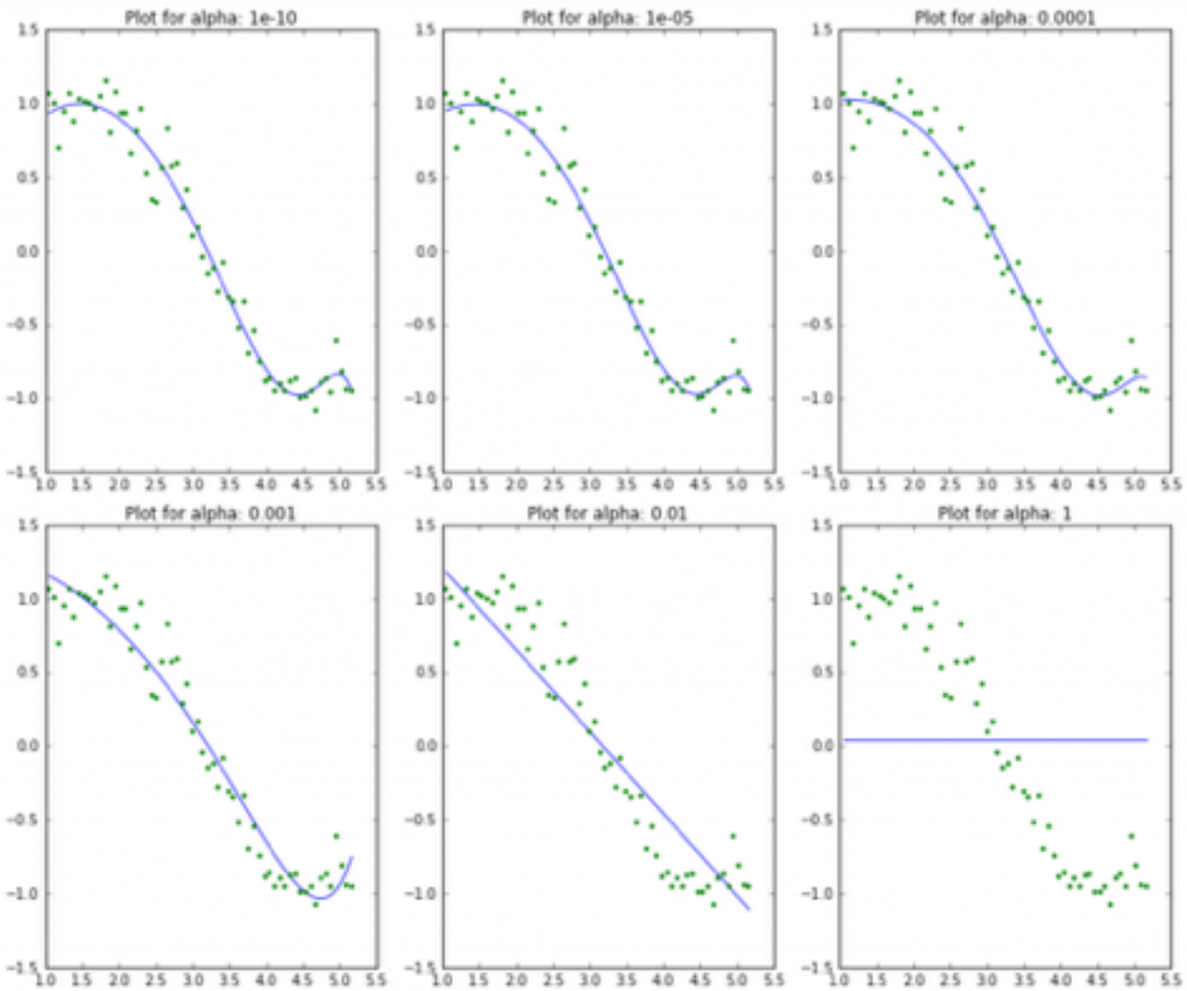
	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8	coef_x_9	coef_x_10	coef_x_11	co
alpha_1e-15	0.87	95	-3e+02	3.8e+02	-2.4e+02	66	0.96	-4.8	0.64	0.15	-0.026	-0.0054	0.00086	0.0
alpha_1e-10	0.92	11	-29	31	-15	2.9	0.17	-0.091	-0.011	0.002	0.00064	2.4e-05	-2e-05	-4.2
alpha_1e-08	0.95	1.3	-1.5	1.7	-0.68	0.039	0.016	0.00016	-0.00036	-5.4e-05	-2.9e-07	1.1e-06	1.9e-07	2e-
alpha_0.0001	0.96	0.56	0.55	-0.13	-0.026	-0.0028	-0.00011	4.1e-05	1.5e-05	3.7e-06	7.4e-07	1.3e-07	1.9e-08	1.9
alpha_0.001	1	0.82	0.31	-0.087	-0.02	-0.0028	-0.00022	1.8e-05	1.2e-05	3.4e-06	7.3e-07	1.3e-07	1.9e-08	1.7
alpha_0.01	1.4	1.3	-0.088	-0.052	-0.01	-0.0014	-0.00013	7.2e-07	4.1e-06	1.3e-06	3e-07	5.6e-08	9e-09	1.1
alpha_1	5.6	0.97	-0.14	-0.019	-0.003	-0.00047	-7e-05	-9.9e-06	-1.3e-06	-1.4e-07	-9.3e-09	1.3e-09	7.8e-10	2.4
alpha_5	14	0.55	-0.059	-0.0085	-0.0014	-0.00024	-4.1e-05	-6.9e-06	-1.1e-06	-1.9e-07	-3.1e-08	-5.1e-09	-8.2e-10	-1.3
alpha_10	18	0.4	-0.037	-0.0059	-0.00095	-0.00017	-3e-05	-5.2e-06	-9.2e-07	-1.6e-07	-2.9e-08	-5.1e-09	-9.1e-10	-1.6
alpha_20	23	0.28	-0.022	-0.0034	-0.0006	-0.00011	-2e-05	-3.6e-06	-6.6e-07	-1.2e-07	-2.2e-08	-4e-09	-7.5e-10	-1.4

polynomial regression

model_pow_14	0.79	2.4e+04	-1.4e+05	3.8e+05	-6.1e+05	6.6e+05	-5e+05	2.8e+05	-1.2e+05	3.7e+04	-8.5e+03	1.5e+03	-1.8e+02	1
model_pow_15	0.7	-3.6e+04	2.4e+05	-7.5e+05	1.4e+06	-1.7e+06	1.5e+06	-1e+06	5e+05	-1.9e+05	5.4e+04	-1.2e+04	1.9e+03	-2

- This straight away gives us the following inferences:
  - The RSS increases with increase in λ, this model complexity reduces
  - An λ as small as 1e-15 gives us significant reduction in magnitude of coefficients. Compare the coefficients in the first row of this table to the last row of polynomial regression table.
  - High λ values can lead to significant underfitting. Note the rapid increase in RSS for values of λ greater than 1
  - Though the coefficients are very very small, they are NOT zero

Lasso



lasso

	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8	coef_x_9	coef_x_10	coef_x_11	co
alpha_1e-15	0.96	0.22	1.1	-0.37	0.00089	0.0016	-0.00012	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.4
alpha_1e-10	0.96	0.22	1.1	-0.37	0.00088	0.0016	-0.00012	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.4
alpha_1e-08	0.96	0.22	1.1	-0.37	0.00077	0.0016	-0.00011	-6.4e-05	-6.3e-06	1.4e-06	7.8e-07	2.1e-07	4e-08	5.3
alpha_1e-05	0.96	0.5	0.6	-0.13	-0.038	-0	0	0	0	7.7e-06	1e-06	7.7e-08	0	0
alpha_0.0001	1	0.9	0.17	-0	-0.048	-0	-0	0	0	9.5e-06	5.1e-07	0	0	0
alpha_0.001	1.7	1.3	-0	-0.13	-0	-0	-0	0	0	0	0	0	1.5e-08	7.5
alpha_0.01	3.6	1.8	-0.55	-0.00056	-0	-0	-0	-0	-0	-0	-0	0	0	0
alpha_1	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
alpha_5	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0
alpha_10	37	0.038	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0

HIGH SPARSITY

- Apart from the expected inference of higher RSS for higher λ, we can see the following:
  - For the same values of λ, the coefficients of lasso regression are much smaller as compared to that of ridge regression (compare row 1 of the 2 tables).
  - For the same λ, lasso has higher RSS (poorer fit) as compared to ridge regression
  - Many of the coefficients are zero even for very small values of λ

ridge

	rss	intercept	coef_x_1	coef_x_2	coef_x_3	coef_x_4	coef_x_5	coef_x_6	coef_x_7	coef_x_8	coef_x_9	coef_x_10	coef_x_11	co
alpha_1e-15	0.87	95	-3e+02	3.8e+02	-2.4e+02	66	0.96	-4.8	0.64	0.15	-0.026	-0.0054	0.00086	0.0
alpha_1e-10	0.92	11	-29	31	-15	2.9	0.17	-0.091	-0.011	0.002	0.00064	2.4e-05	-2e-05	-4.
alpha_1e-08	0.95	1.3	-1.5	1.7	-0.68	0.039	0.016	0.00016	-0.00036	-5.4e-05	-2.9e-07	1.1e-06	1.9e-07	2e-
alpha_0.0001	0.96	0.56	0.55	-0.13	-0.026	-0.0028	-0.00011	4.1e-05	1.5e-05	3.7e-06	7.4e-07	1.3e-07	1.9e-08	1.9
alpha_0.001	1	0.82	0.31	-0.087	-0.02	-0.0028	-0.00022	1.8e-05	1.2e-05	3.4e-06	7.3e-07	1.3e-07	1.9e-08	1.7
alpha_0.01	1.4	1.3	-0.088	-0.052	-0.01	-0.0014	-0.00013	7.2e-07	4.1e-06	1.3e-06	3e-07	5.6e-08	9e-09	1.1
alpha_1	5.6	0.97	-0.14	-0.019	-0.003	-0.00047	-7e-05	-9.9e-06	-1.3e-06	-1.4e-07	-9.3e-09	1.3e-09	7.8e-10	2.4
alpha_5	14	0.55	-0.059	-0.0085	-0.0014	-0.00024	-4.1e-05	-6.9e-06	-1.1e-06	-1.9e-07	-3.1e-08	-5.1e-09	-8.2e-10	-1.
alpha_10	18	0.4	-0.037	-0.0055	-0.00095	-0.00017	-3e-05	-5.2e-06	-9.2e-07	-1.6e-07	-2.9e-08	-5.1e-09	-9.1e-10	-1.
alpha_20	23	0.28	-0.022	-0.0034	-0.0006	-0.00011	-2e-05	-3.6e-06	-6.6e-07	-1.2e-07	-2.2e-08	-4e-09	-7.5e-10	-1.

Ridge/Lasso

- **Common:**
  - Ridge or lasso are forms of regularized linear regressions, which adds tuning parameter λ to a model to induce smoothness in order to prevent overfitting.

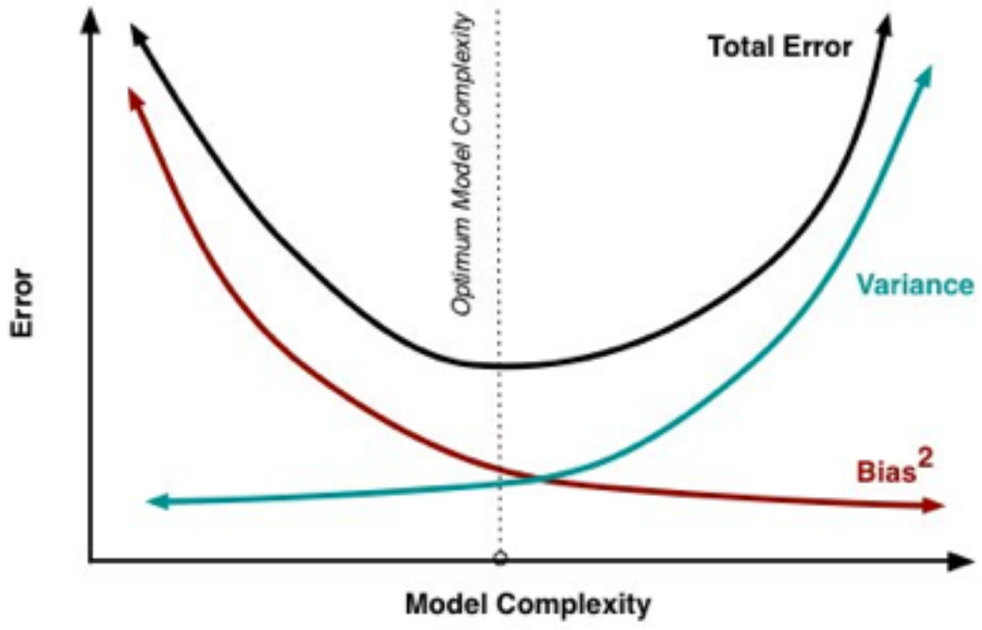
- **Difference:**
  - Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 while as lasso l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it's the sum of the absolute values of the coefficients.
  - In the lasso, this makes it easier for the coefficients to be zero and therefore easier to eliminate some of the input variable as not contributing to the output.

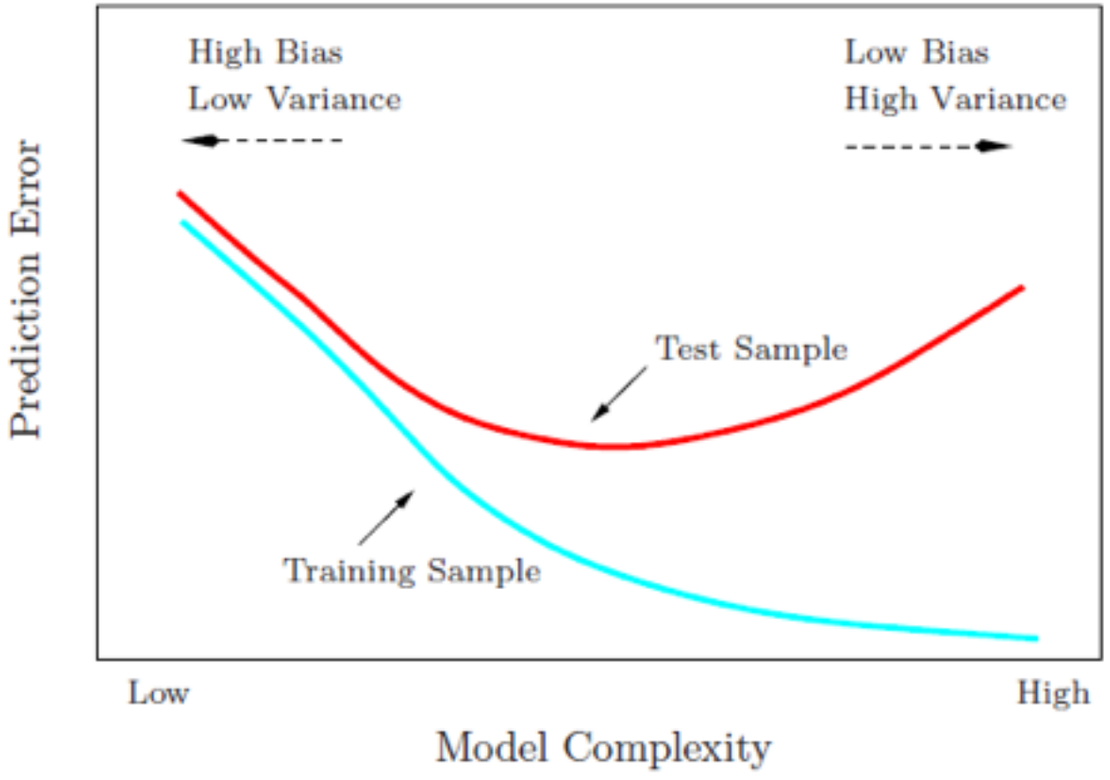
- **Typical use cases:**
  - Ridge: It is majorly used to prevent overfitting. Since it includes all the features, it is not very useful in case of exorbitantly high number features, say in millions, as it will pose computational challenges.
  - Lasso: Since it provides sparse solutions, it is generally the model of choice for modeling cases where the number features are in millions or more. In such a case, getting a sparse solution is of great computational advantage as the features with zero coefficients can simply be ignored.

- **Presence of Highly Correlated Features**
  - Ridge: It generally works well even in presence of highly correlated features as it will include all of them in the model but the coefficients will be distributed among them depending on the correlation.
  - Lasso: It arbitrarily selects any one feature among the highly correlated ones and reduced the coefficients of the rest to zero. Also, the chosen variable changes randomly with change in model parameters. This generally doesn’t work that well as compared to ridge regression.

Training/Test Data Set

- training set: a set of data used to discover potentially predictive relationships
- training error: the error of applying the statistical learning method to the observations used in its training
- test set: a set of data used to assess the strength and utility of a predictive relationship
- test error: the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method





Select Tuning Parameter: λ

K-folder cross validation: estimates can be used to select best model, and to give an idea of the test error of the final chosen model. (5 or 10-folder)

- Randomly divide the data into K equal-sized parts.
- Leave out part k, put it to the side
- fit the model to the other K – 1 parts (combined) as the training set
- Use part k as the test set to estimate the prediction error
- Repeat this process K times, once each for the different splits of the data
- The cross-validation error can be obtained by computing the weighted average of K folders

$$CV_K = \sum_{k=1}^{K} \frac{n_k}{N} MSE_k$$

- Select the λ in which the cross-validation error is smallest

Advantages and Disadvantages of linear regression

Advantages

- Simple to implement and fast to run
- For linear relationship between inputs and output, performance is better
- Easy to interpret by the coefficient

Disadvantages

- Limited to linear relationship
  - By its nature, linear regression only looks at linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them. Sometimes this is incorrect. For example, the relationship between income and age is curved, i.e., income tends to rise in the early parts of adulthood, flatten out in later adulthood and decline after people retire.

- Only Look at the mean of the dependent variable
  - Linear regression looks at a relationship between the mean of the dependent variable and the independent variables. For example, if we look at the relationship between the birth weight of infants and maternal characteristics such as age, linear regression will look at the average weight of babies born to mothers of different ages. However, sometimes we need to look at the extremes of the dependent variable, e.g., babies are at risk when their weights are low, so we would want to look at the extremes in this example.

Advantages and Disadvantages

- Sensitive to outliners
  - Outliers are data that are surprising. Outliers can be univariate (based on one variable) or multivariate. If we are looking at age and income, univariate outliers would be things like a person who is 118 years old, or one who made $12 million last year. A multivariate outlier would be an 18-year-old who made $200,000. In this case, neither the age nor the income is very extreme, but very few 18-year-old people make that much money.
  - Outliers can have huge effects on the regression.

- Data must be independent
  - Linear regression assumes that the data are independent. That means that the scores of one subject (such as a person) have nothing to do with those of another. This is often, but not always, sensible. Two common cases where it does not make sense are clustering in space and time.
  - A classic example of clustering in space is student test scores, when you have students from various classes, grades, schools and school districts. Students in the same class tend to be similar in many ways, i.e., they often come from the same neighborhoods, they have the same teachers, etc. Thus, they are not independent.
  - Examples of clustering in time are any studies where you measure the same subjects multiple times. For example, in a study of diet and weight, you might measure each person multiple times. These data are not independent because what a person weighs on one occasion is related to what he or she weighs on other occasions.