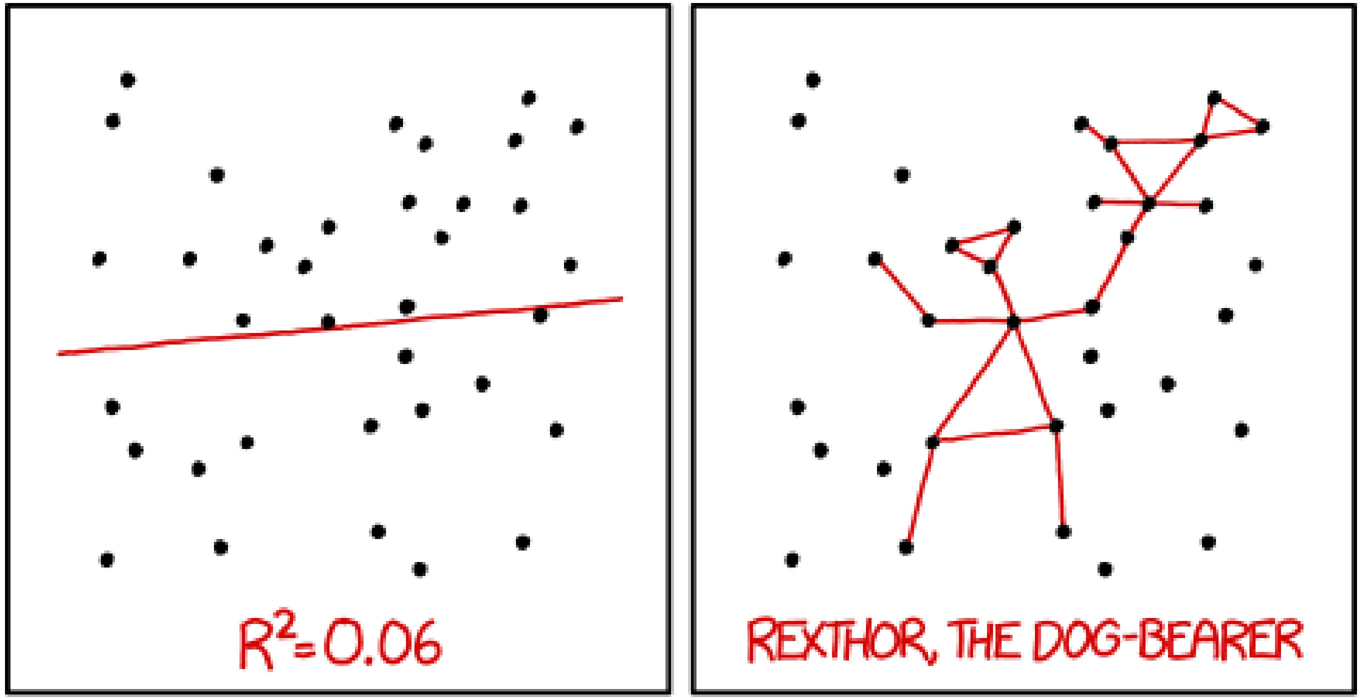


I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Day 1 - What are we studying today?

1. What are different types of Regression Analysis?
2. How to fit Simple linear regression?
3. What are key assumptions underlying Linear Regression?
4. How to validate if model assumes each assumption?
   - Residual Plots
   - Normal QQ Plots
   - Auto-correlation
   - Constant Variance
   - Outliers - Leverage
5. What is Multiple Linear Regression?
6. What is Interaction between variables?
7. What is Polynomial Regression?
8. What are different components of regression summary & what they mean?
9. How to balance in Model Accuracy and Interpret ability?
10. Subset Selection (variables)
    1. Bias Variance Trade-off
    2. Under-fitting vs over-fitting
11. How to choose the Optimal Model (Indirect Estimate) from train metrics?
12. How to compare between different models via Indirect Estimates?
13. How to choose the Optimal Model (Direct Estimate) from test metrics?
    1. Leave On Out – Cross Validation
    2. K-fold Cross Validation
14. 2. Shrinkage / Regularization on Training
    1. Ridge - L2
    2. Lasso - L1
15. 3. Dimension Reduction

Linear Regression

Regressing the response variable on the predictor variables

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

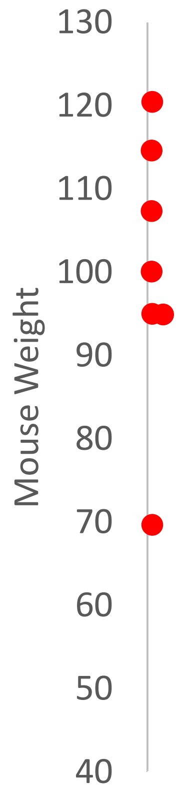
estimated values

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

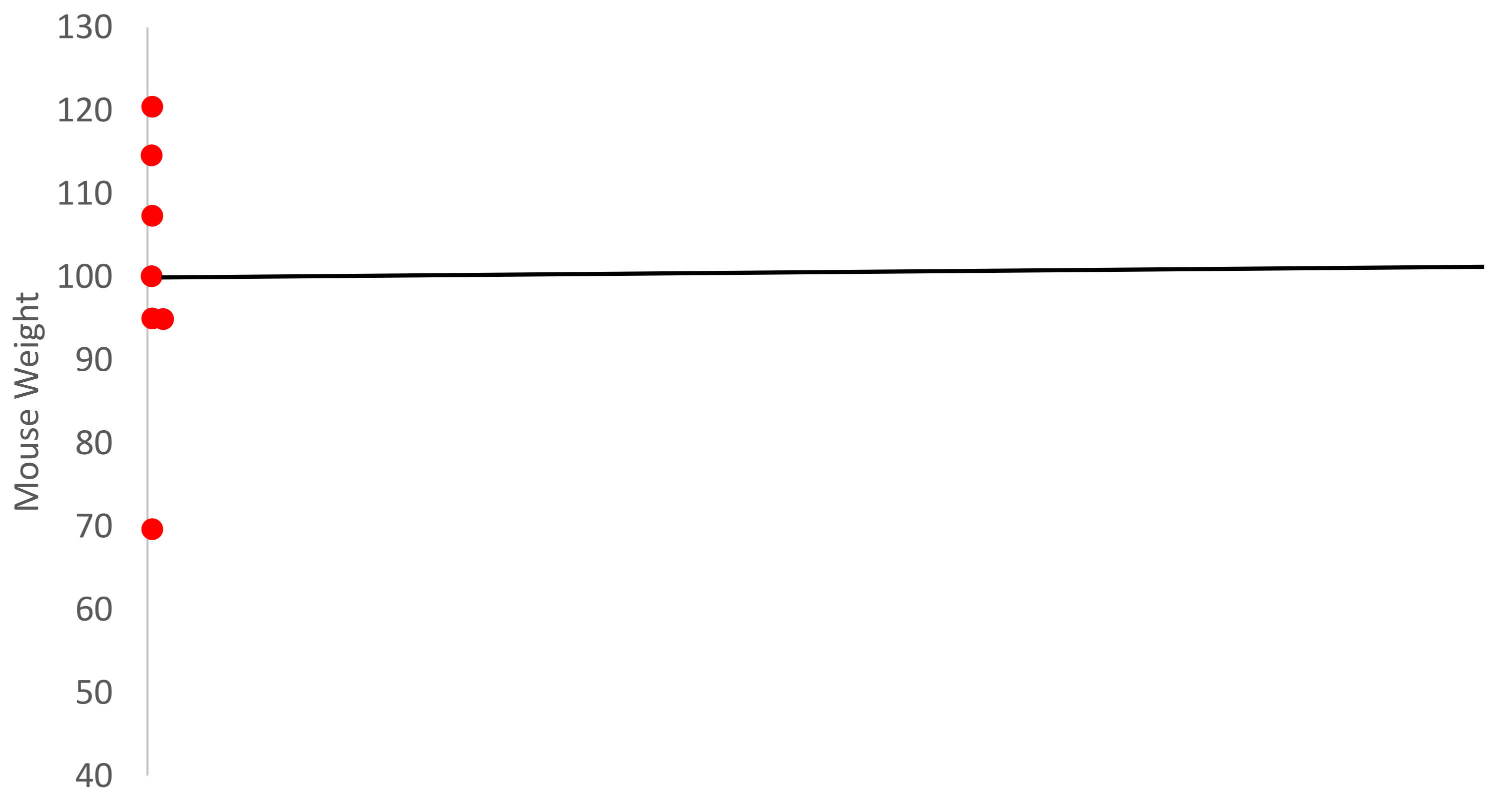
Varieties Of Regression Analysis

Type of regression	Typical use
Simple linear	Predicting a quantitative response variable from a quantitative explanatory variable.
Polynomial	Predicting a quantitative response variable from a quantitative explanatory variable, where the relationship is modeled as an nth order polynomial.
Multiple linear	Predicting a quantitative response variable from two or more explanatory variables.
Multilevel	Predicting a response variable from data that have a hierarchical structure (for example, students within classrooms within schools). Also called *hierarchical*, *nested*, or *mixed* models.
Multivariate	Predicting more than one response variable from one or more explanatory variables.
Logistic	Predicting a categorical response variable from one or more explanatory variables.
Poisson	Predicting a response variable representing counts from one or more explanatory variables.
Cox proportional hazards	Predicting time to an event (death, failure, relapse) from one or more explanatory variables.
Time-series	Modeling time-series data with correlated errors.

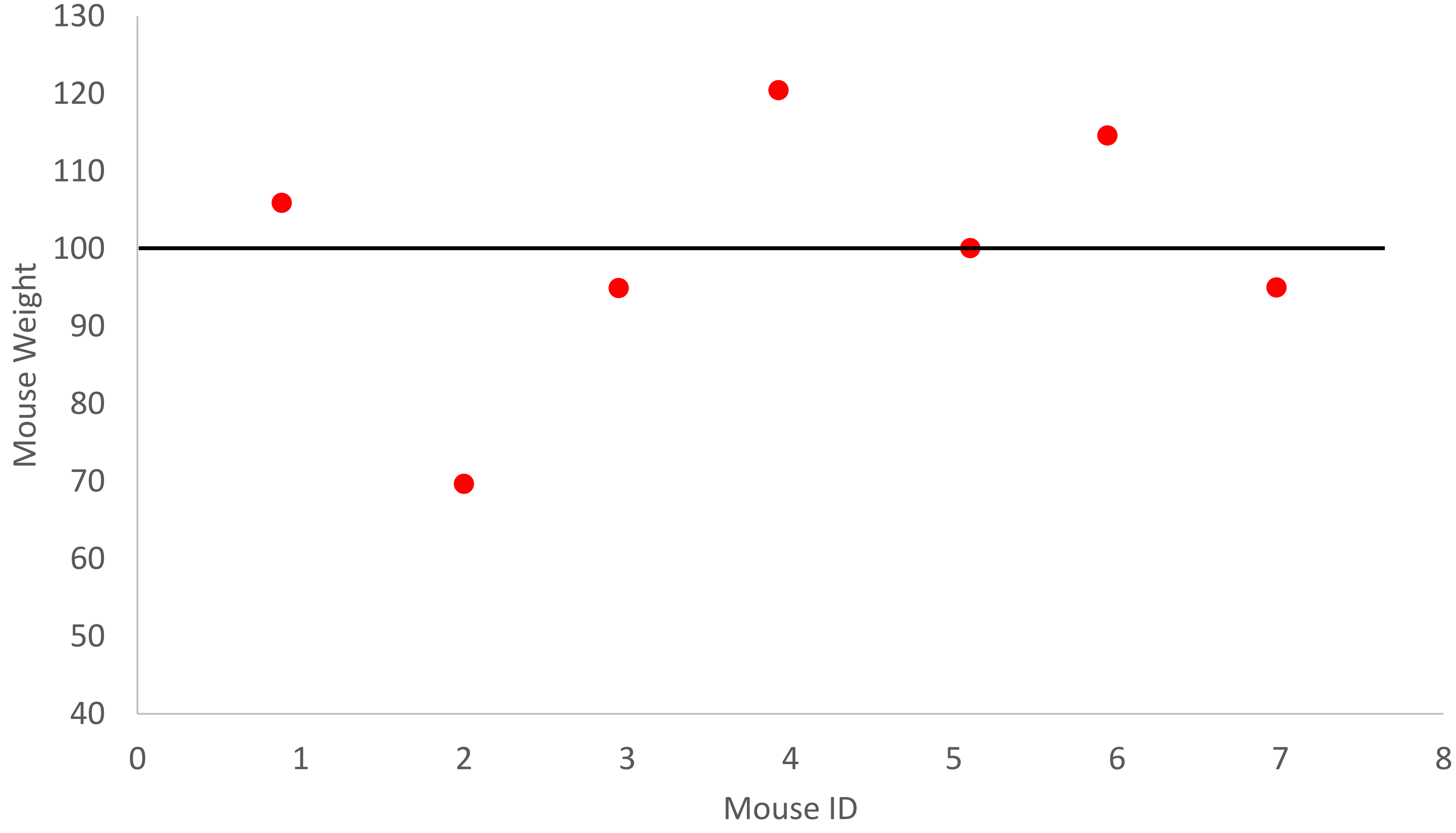
Mouse Weight of Different Observations



Average Mouse Weight

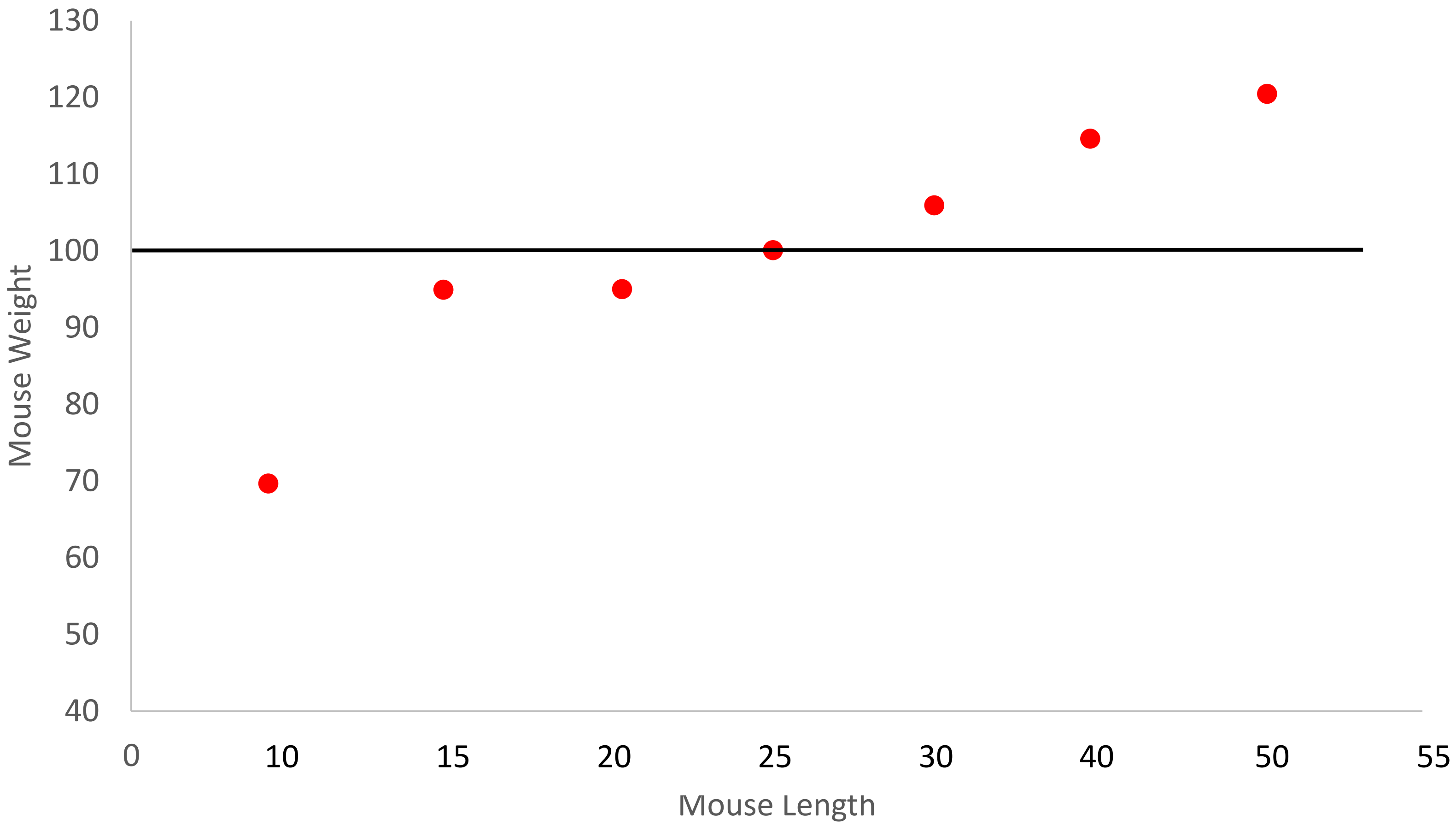


Average Mouse Weight explained by additional Variable – Mouse ID

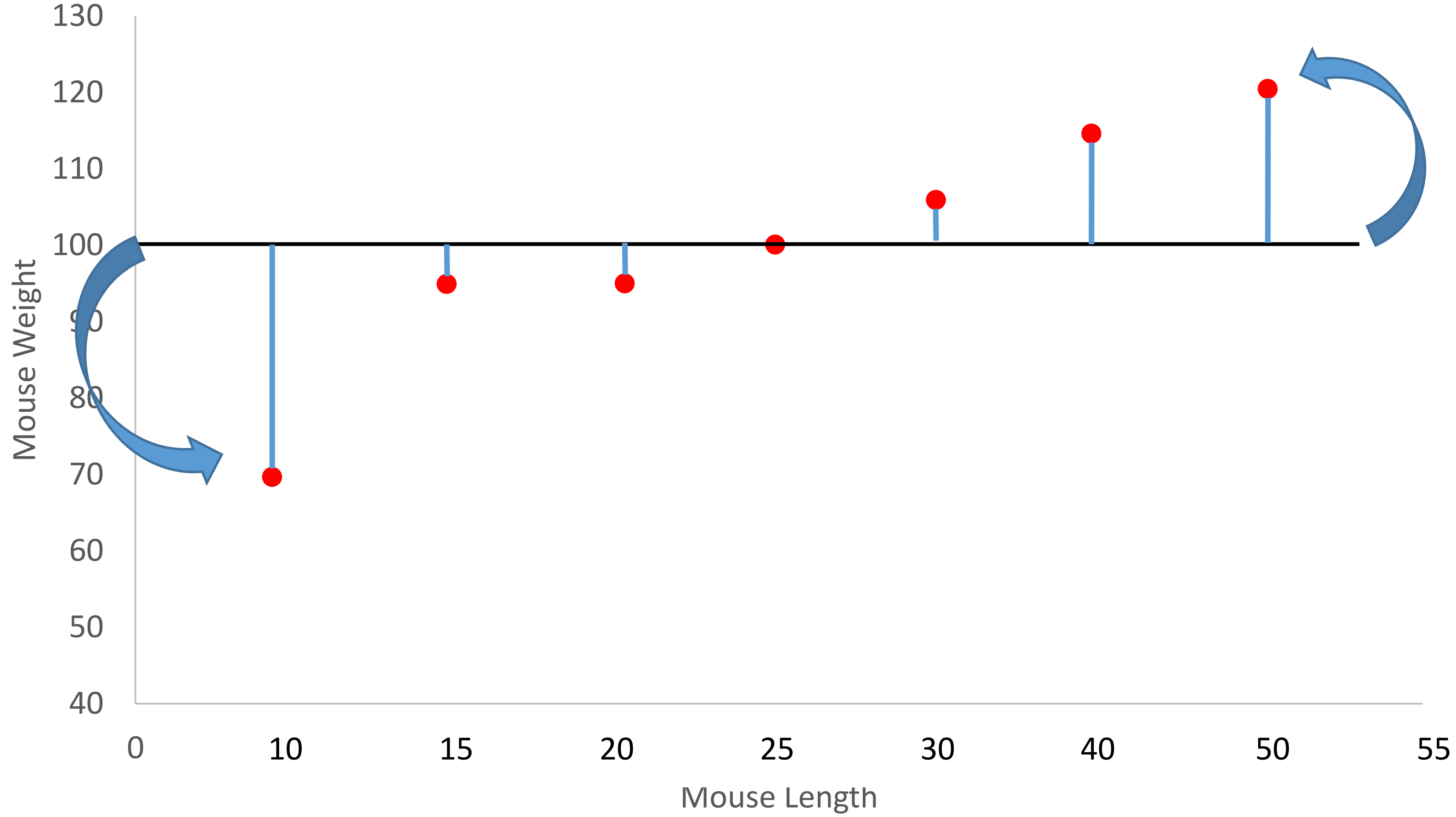


As it is, Didn't explain any variability about Mouse Weight

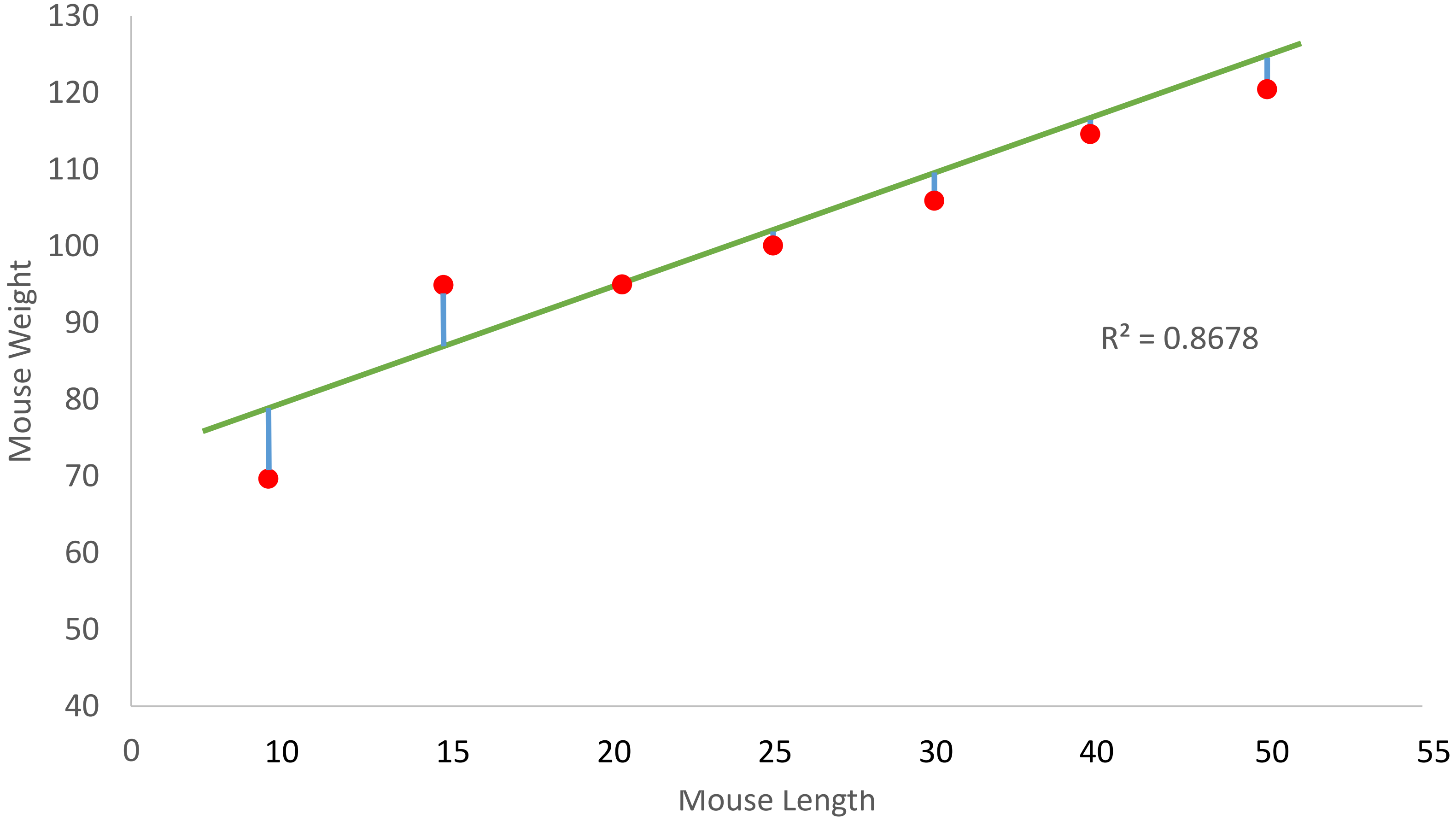
Average Mouse Weight explained by additional Variable – Mouse Length



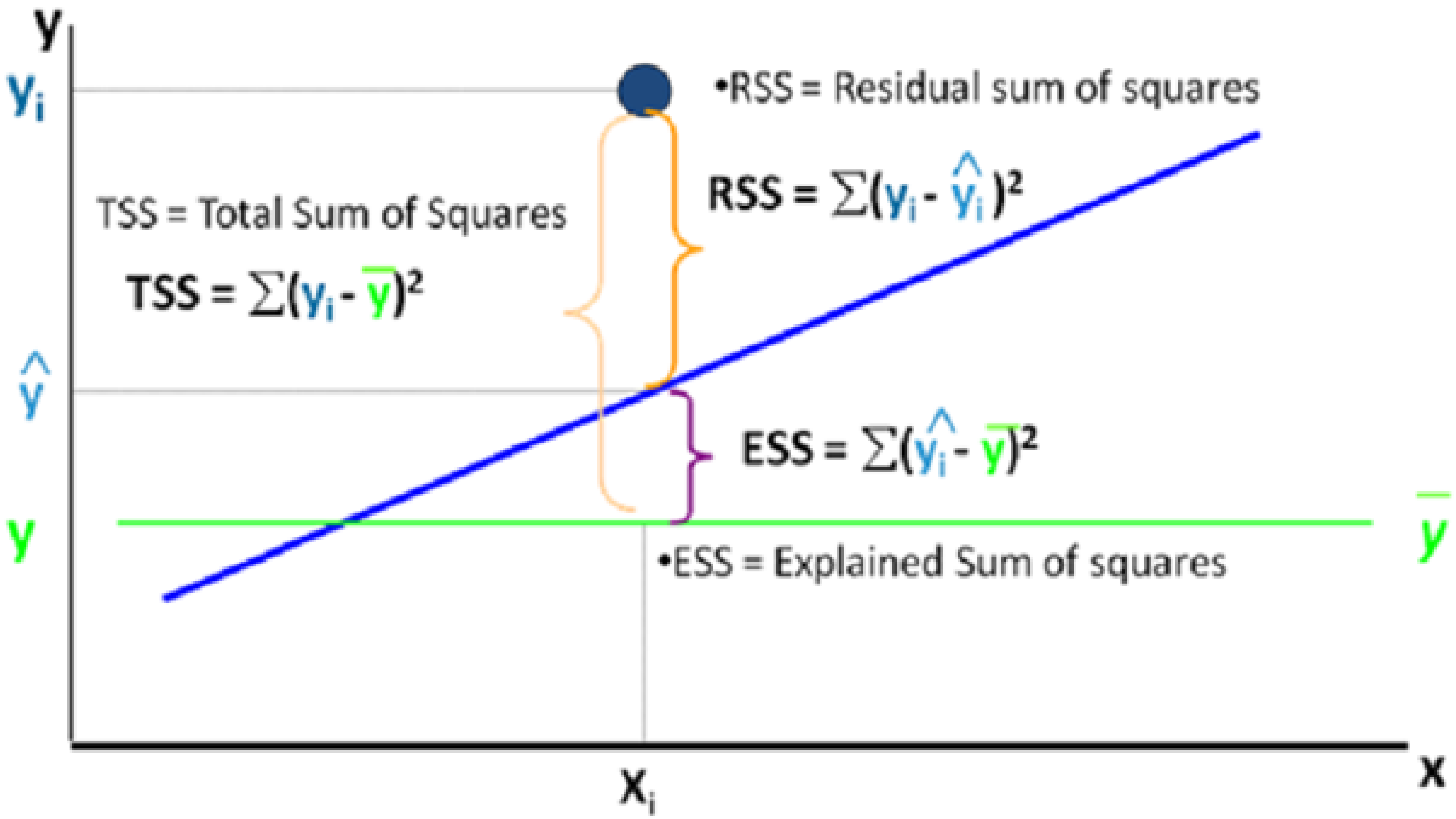
Average Mouse Weight explained by additional Variable – Mouse Length



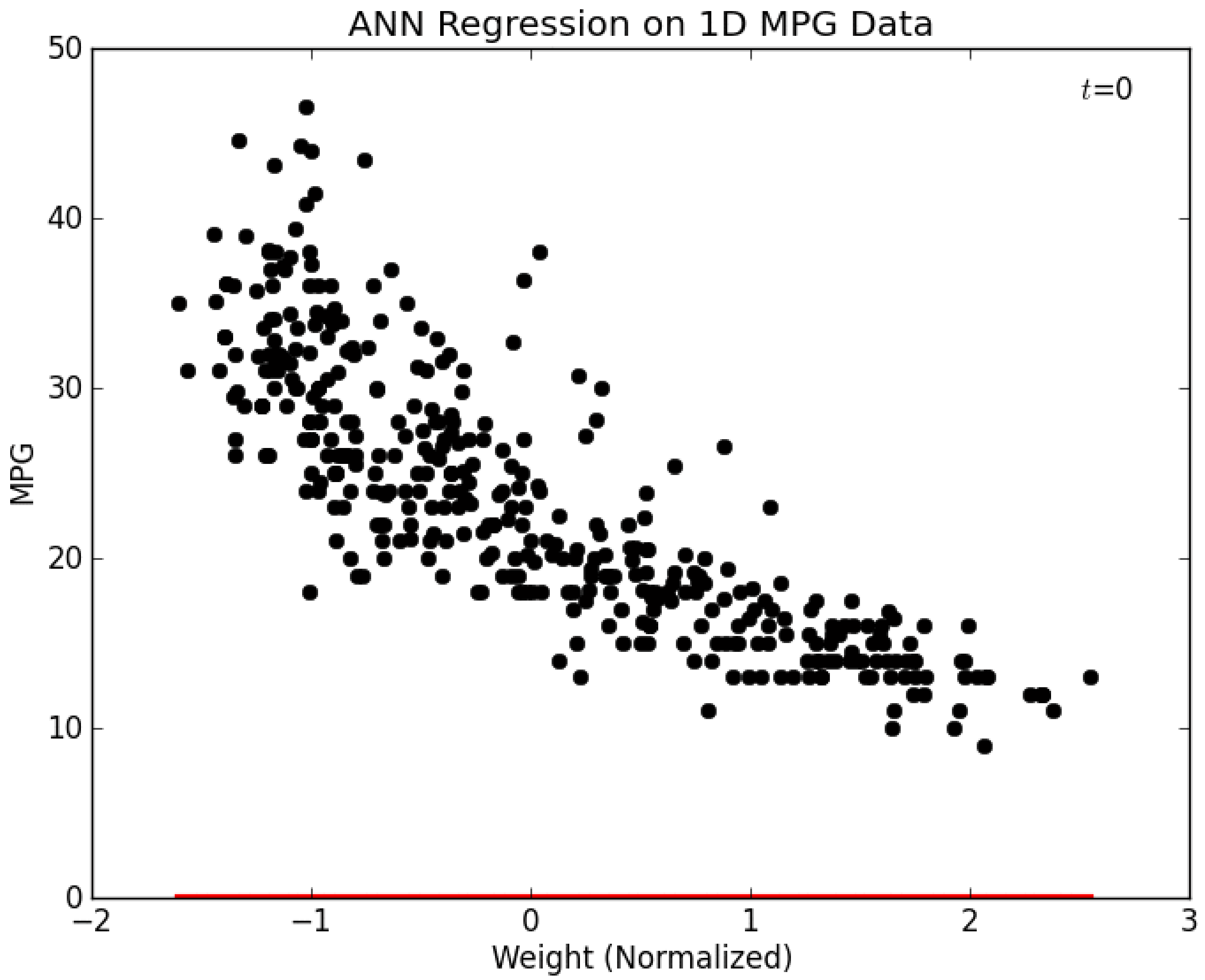
Average Mouse Weight explained by additional Variable – Mouse Length



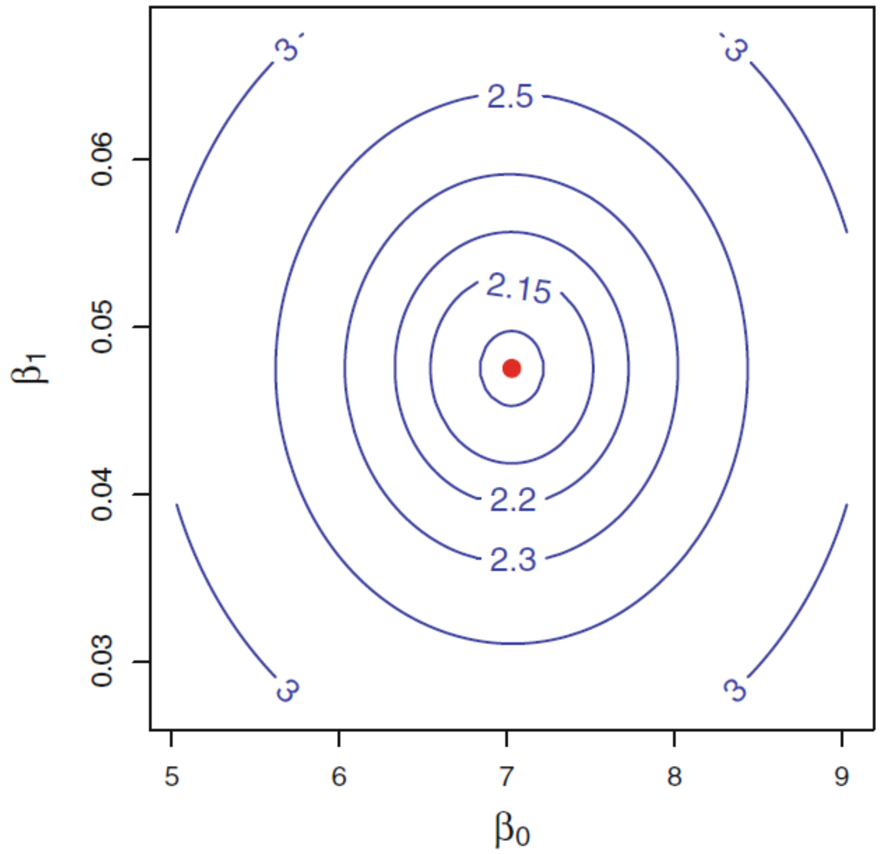
TSS – RSS – ESS in nutshel

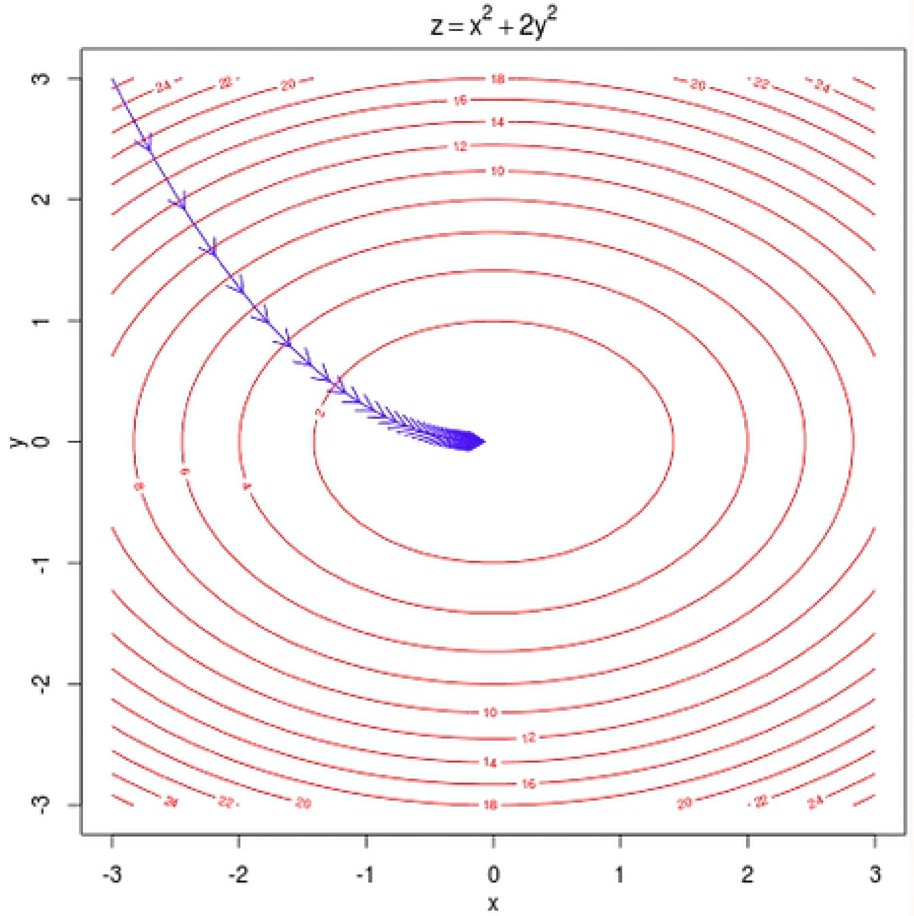


OLS Fit



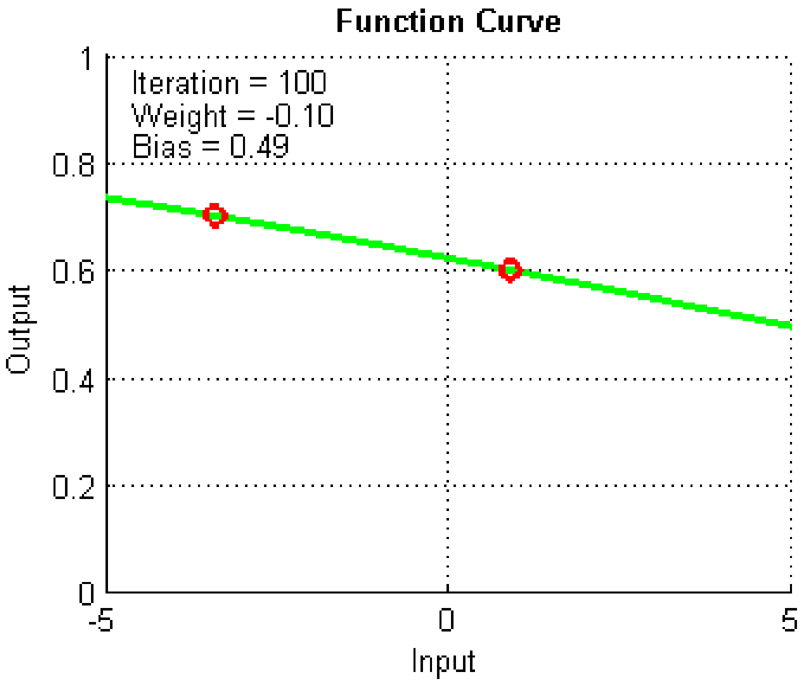
ANN Regression on 1D MPG Data
t=0
50
40
30
20
10
0
MPG
−2
−1
0
1
2
3
Weight (Normalized)

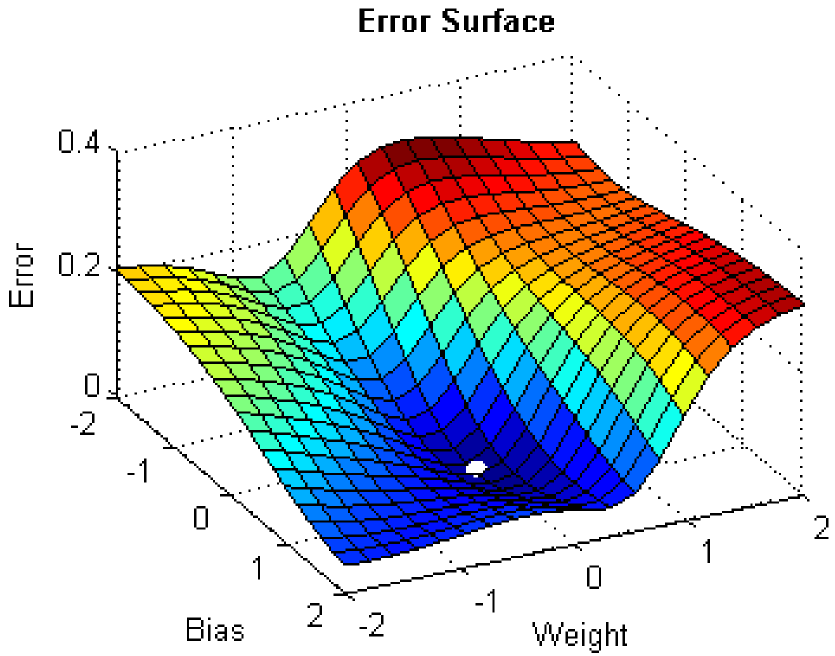


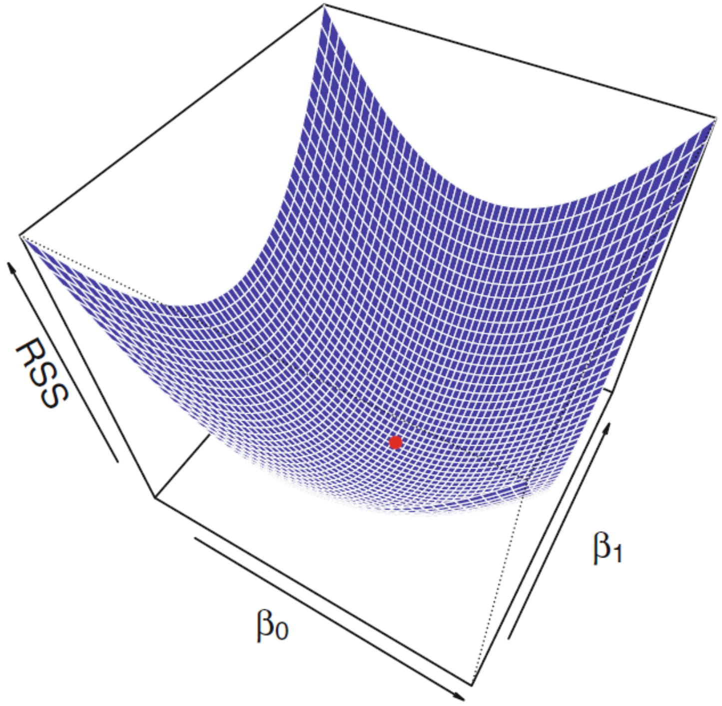


Contour plots for the RSS values as a function of the parameters β for various regressions involving the Credit *data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS.*
Left: *A contour plot of RSS for the regression of* balance *onto* age *and* limit*. The minimum value is well defined.*

OPTIMIZATION OF PARAMETERS







A contour plot of RSS for the regression of balance *onto* rating *and* limit. *Because of the collinearity, there are many pairs with a similar value for RSS.*

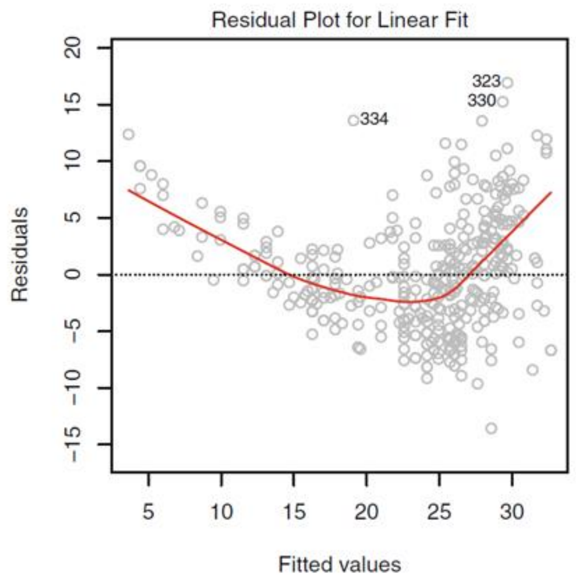
OLS Syntax – Statsmodel or R

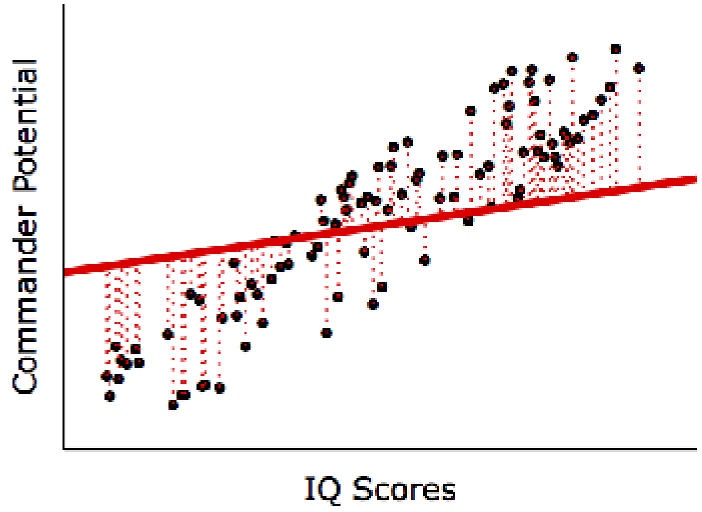
Symbol	Usage
`~`	Separates response variables on the left from the explanatory variables on the right. For example, a prediction of `y` from `x`, `z`, and `w` would be coded `y ~ x + z + w`.
`+`	Separates predictor variables.
`:`	Denotes an interaction between predictor variables. A prediction of `y` from `x`, `z`, and the interaction between `x` and `z` would be coded `y ~ x + z + x:z`.
`*`	A shortcut for denoting all possible interactions. The code `y ~ x * z * w` expands to `y ~ x + z + w + x:z + x:w + z:w + x:z:w`.
`^`	Denotes interactions up to a specified degree. The code `y ~ (x + z + w)^2` expands to `y ~ x + z + w + x:z + x:w + z:w`.
`.`	A placeholder for all other variables in the data frame except the dependent variable. For example, if a data frame contained the variables `x, y, z, and w`, then the code `y ~ .` would expand to `y ~ x + z + w`.
`-`	A minus sign removes a variable from the equation. For example, `y ~ (x + z + w)^2 – x:w` expands to `y ~ x + z + w + x:z + z:w`.
`-1`	Suppresses the intercept. For example, the formula `y ~ x -1` fits a regression of `y` on `x`, and forces the line through the origin at `x=0`.
`I()`	Elements within the parentheses are interpreted arithmetically. For example, `y ~ x + (z + w)^2` would expand to `y ~ x + z + w + z:w`. In contrast, the code `y ~ x + I((z + w)^2)` would expand to `y ~ x + h`, where `h` is a new variable created by squaring the sum of `z` and `w`.
function	Mathematical functions can be used in formulas. For example, `log(y) ~ x + z + w` would predict `log(y)` from `x`, `z`, and `w`.

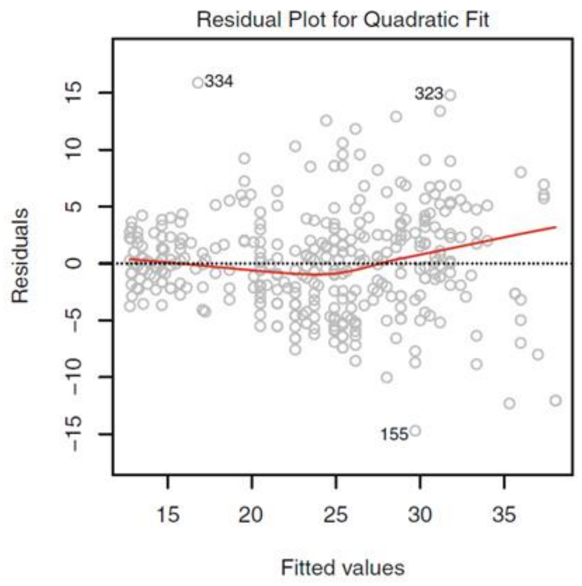
KEY ASSUMPTIONS UNDERLYING REGRESSION ANALYSIS

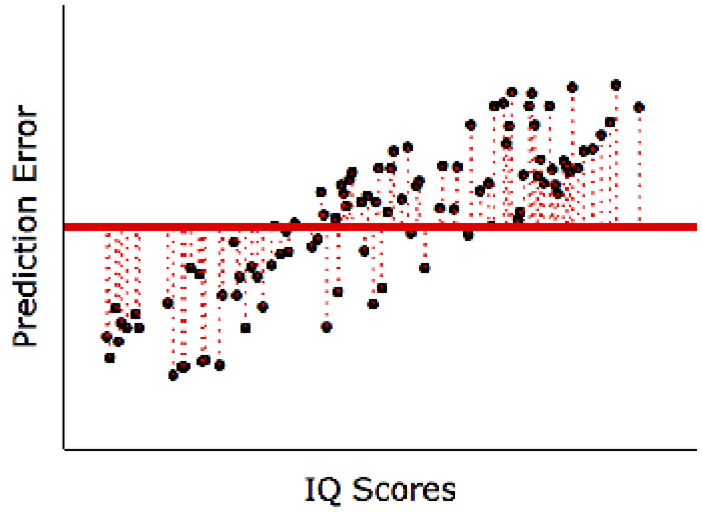
1. **Linearity** - Assumes that trend is linear (validate it with residual plot)
2. **Sample Size** to Observation ratio - Ideally observations should be 20**x** the no of predictors to learn signal
3. **Normality** – Residuals are normally distributed (check with histogram or in Residual plot distribution)
4. **Independence** - Target values are independent of each other
5. **No Multi-collinearity** (Check this by Variance Inflation Factor (VIF))
6. **Homoscedasticity –** Variance of the Target doesn't vary with the levels of the IV
7. **No Outliers** (1-1 scatter-plot) – Check Cook's Distance

Residual Plot - Linearity Check



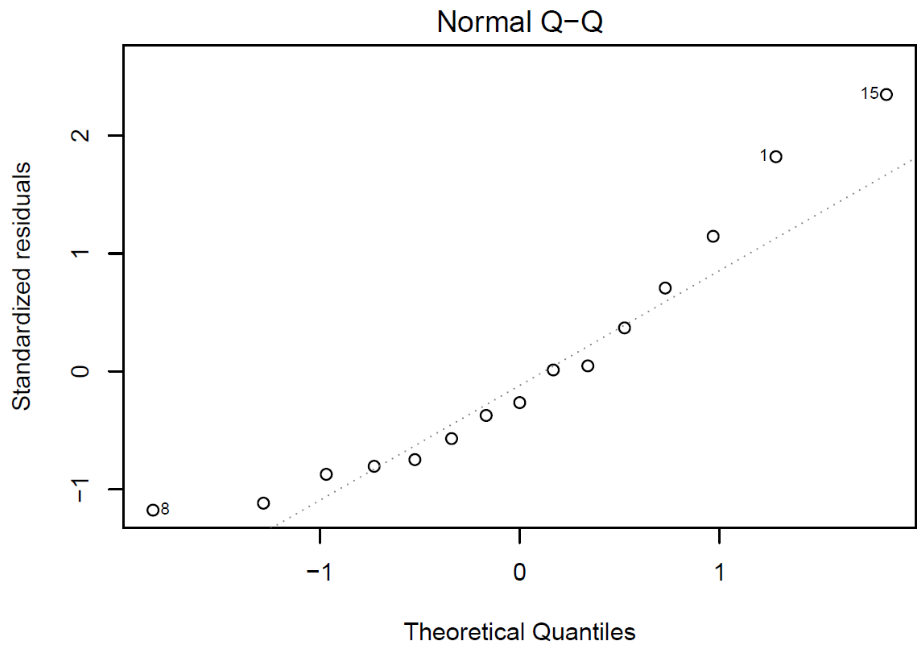


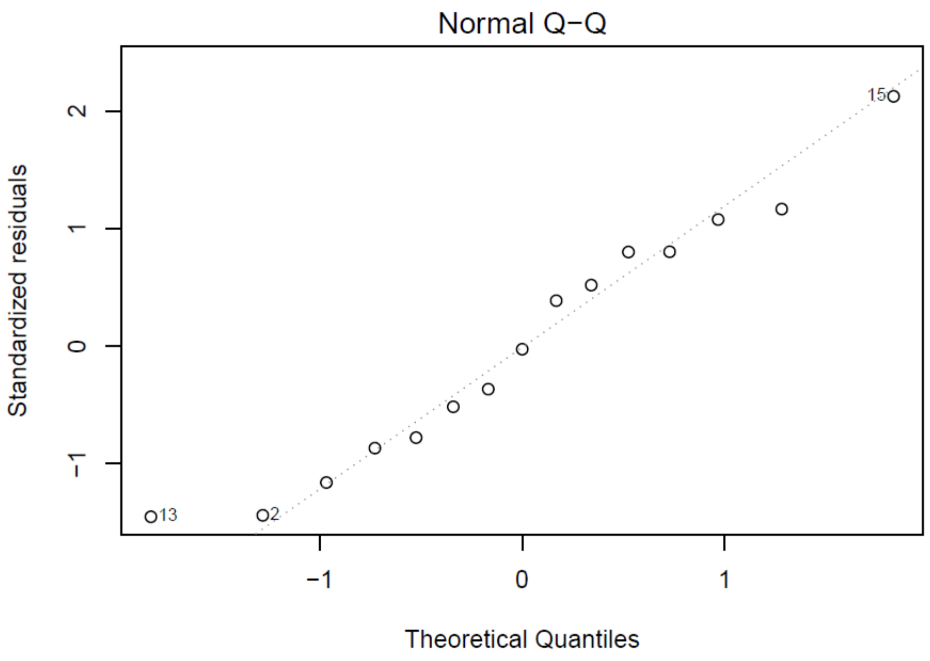




Normal Q-Q Plot – Normality Check

- It is a probability plot of the standardized residuals against the values that would be expected under normality.
- If you've met the normality assumption, the points on this graph should fall on the straight 45-degree line.





Auto Correlation / Independence Check

Use your understanding of how the data was collected to rule out autocorrelation in scenarios.

Durbin Watson is used test measure to detect autocorrelation in residuals .
It can lead to underestimates of the standard error and can cause you to think predictors are significant when they are not

The **Durbin Watson test** reports a test statistic, with a value from 0 to 4, where:

- 2 is no autocorrelation.
- 0 to <2 is positive autocorrelation (common in time series data).
- >2 to 4 is negative autocorrelation (less common in time series data)

A **rule of thumb** is that test statistic values in the range of 1.5 to 2.5 are relatively normal

StatsModel

Durban Watson Wiki

Heteroscedasticity - Variance Check

variance of residuals should not increase with fitted values of response variable. we want to check if the model thus built is unable to explain some pattern in the response variable (Y), that eventually shows up in the residuals.

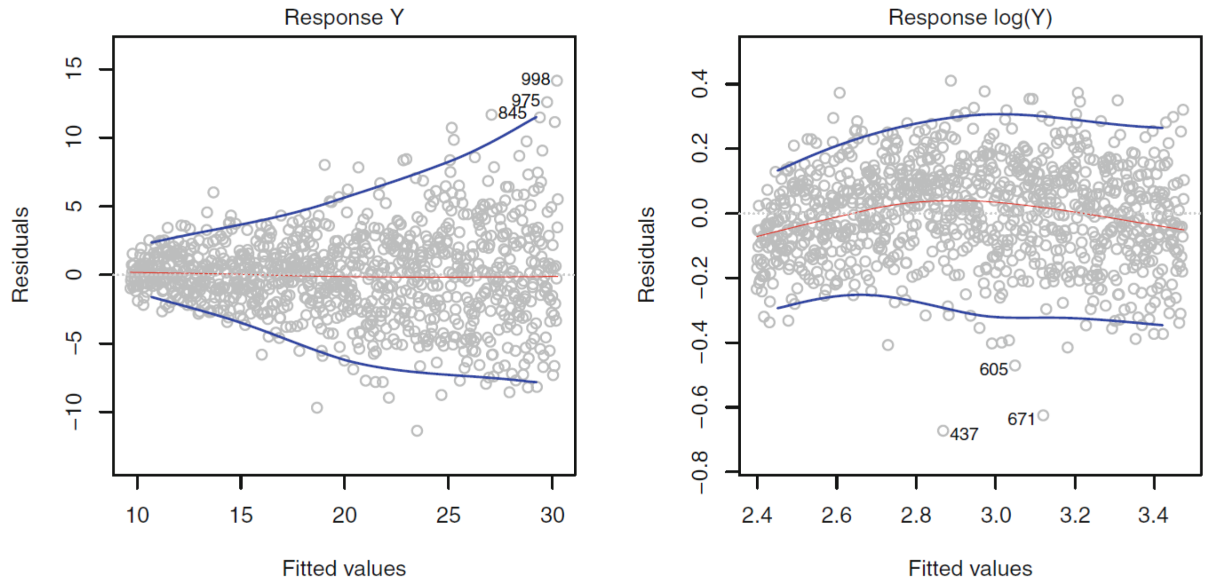
How to test heteroscedasticity ?

1. **Breush-Pagan** test

It tests whether the variance of the errors from a regression is dependent on the values of the independent variables. If the test statistic has a p-value below an appropriate threshold (e.g. $p<0.05$) then the null hypothesis of homoskedasticity is rejected and heteroskedasticity assumed

Non Constant Variance (NCV) – test

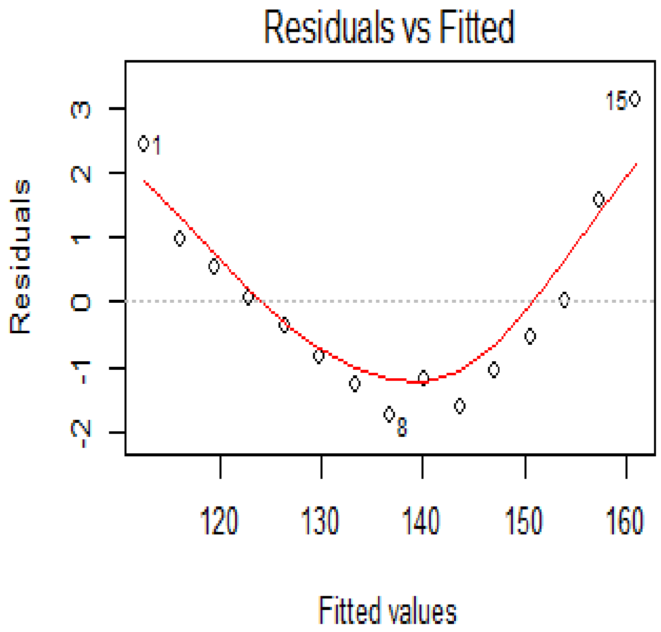
Check p values of above test to reject null hypothesis that Heteroscedasticity exists. One can perform **Box-Cox transformation** on Dependent Variable to rectify it

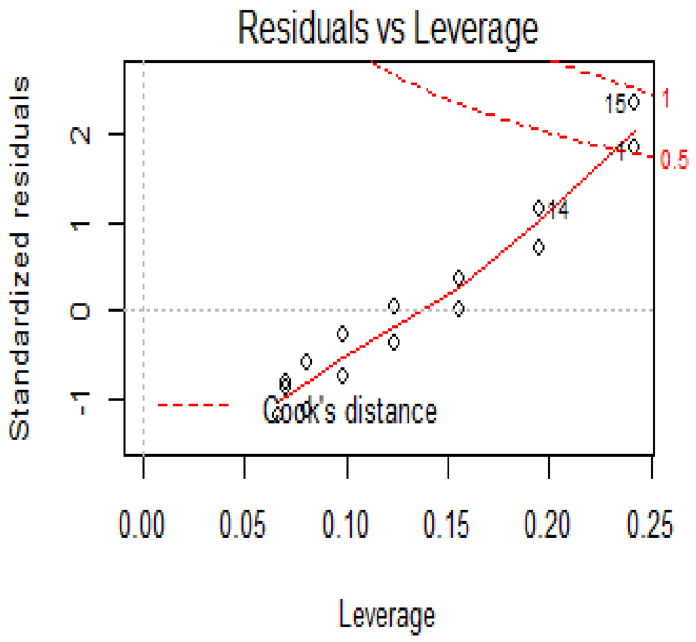


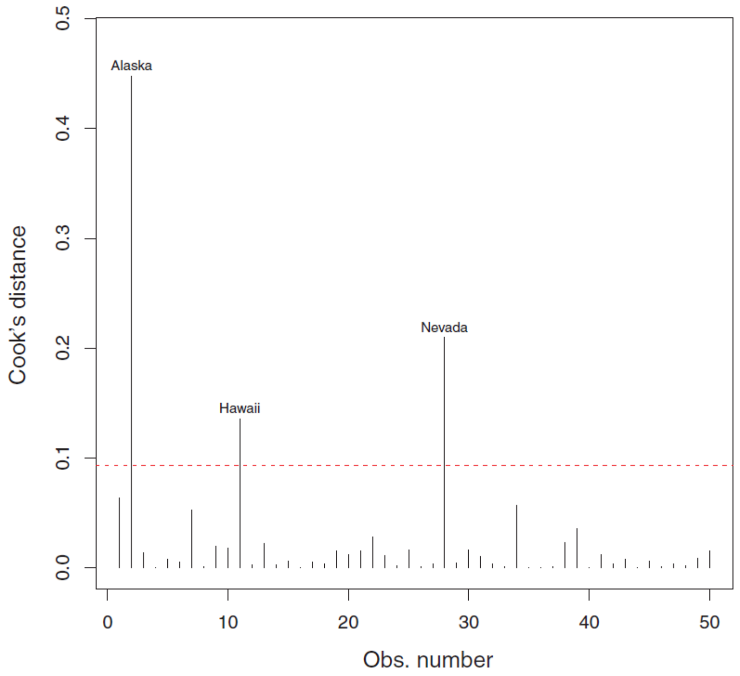
Outliers - Influential Observations – Leverage Check

Outlier is an observation that isn't predicted well by the fitted regression model
Influential observation is an observation that has a disproportionate impact on the determination of the model parameters

Cook's Distance is used to estimate of the influence of a data point especially the outliers when performing a least-squares regression analysis. A common rule of thumb is that an observation with a value of Cook's D over 1.0 has too much influence







[Statsmodel]

Multiple Linear Regression

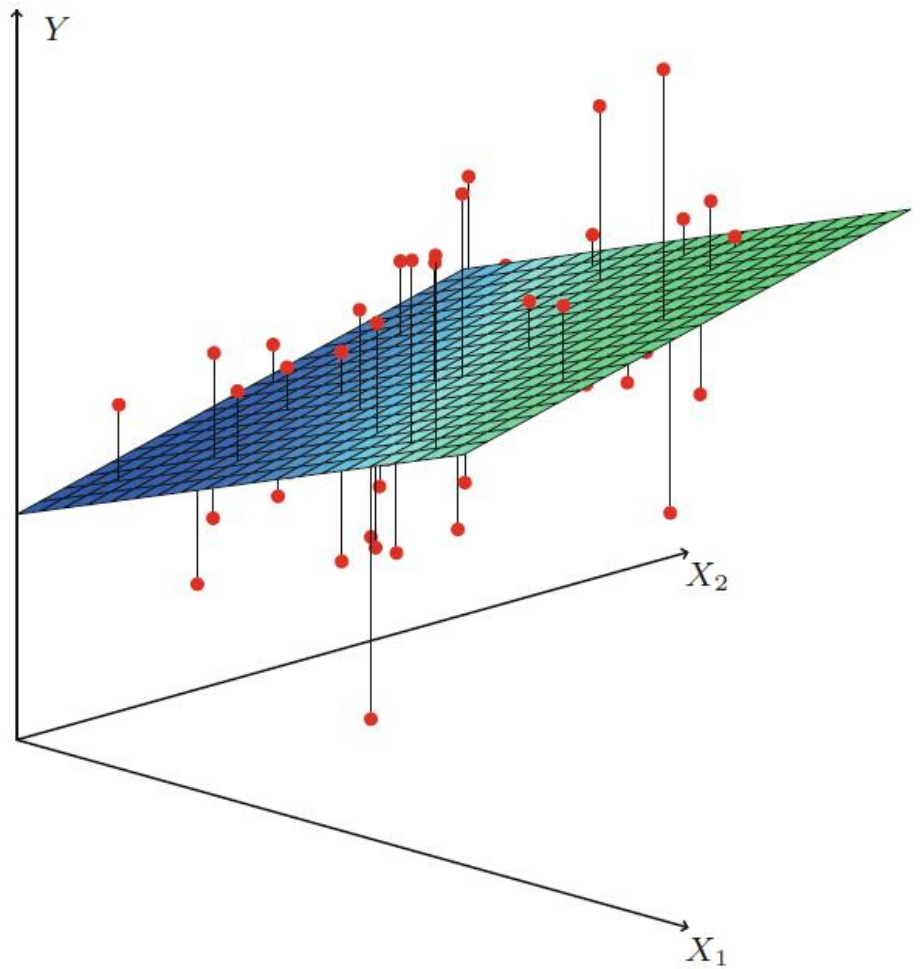
Regressing the response variable on the ***more than one*** *predictor variables*

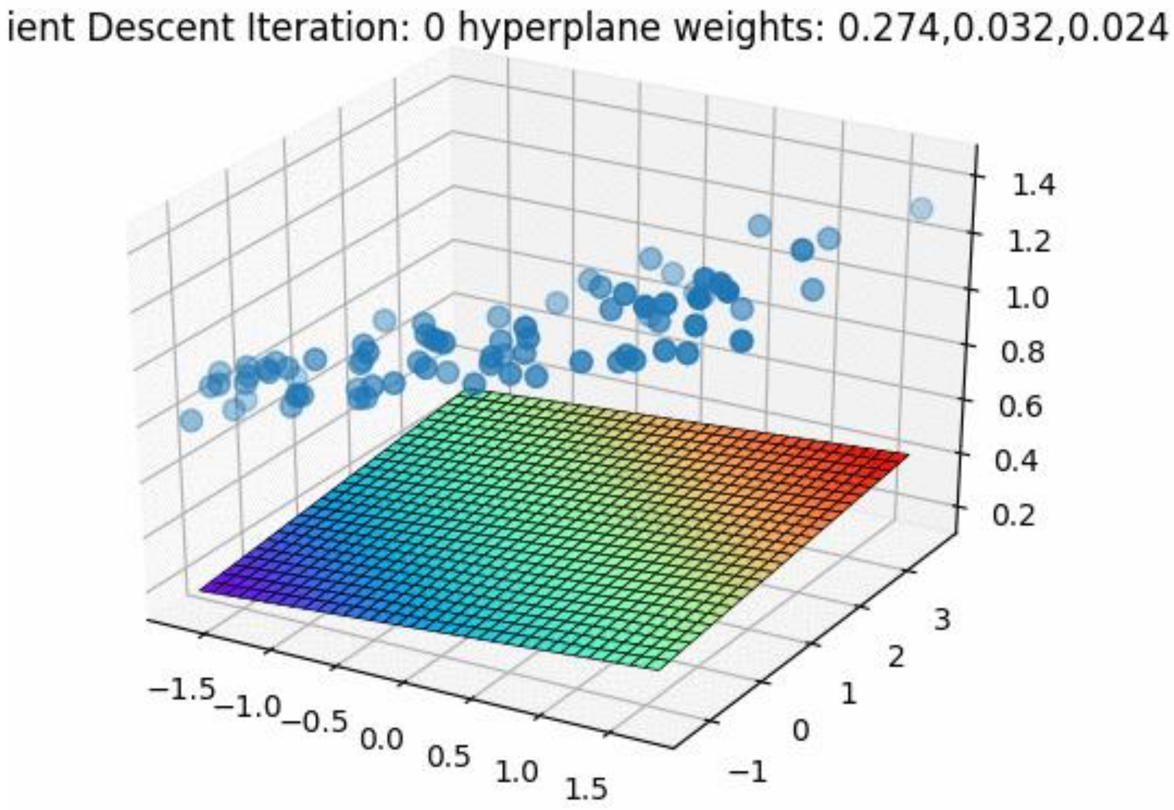
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

estimated values

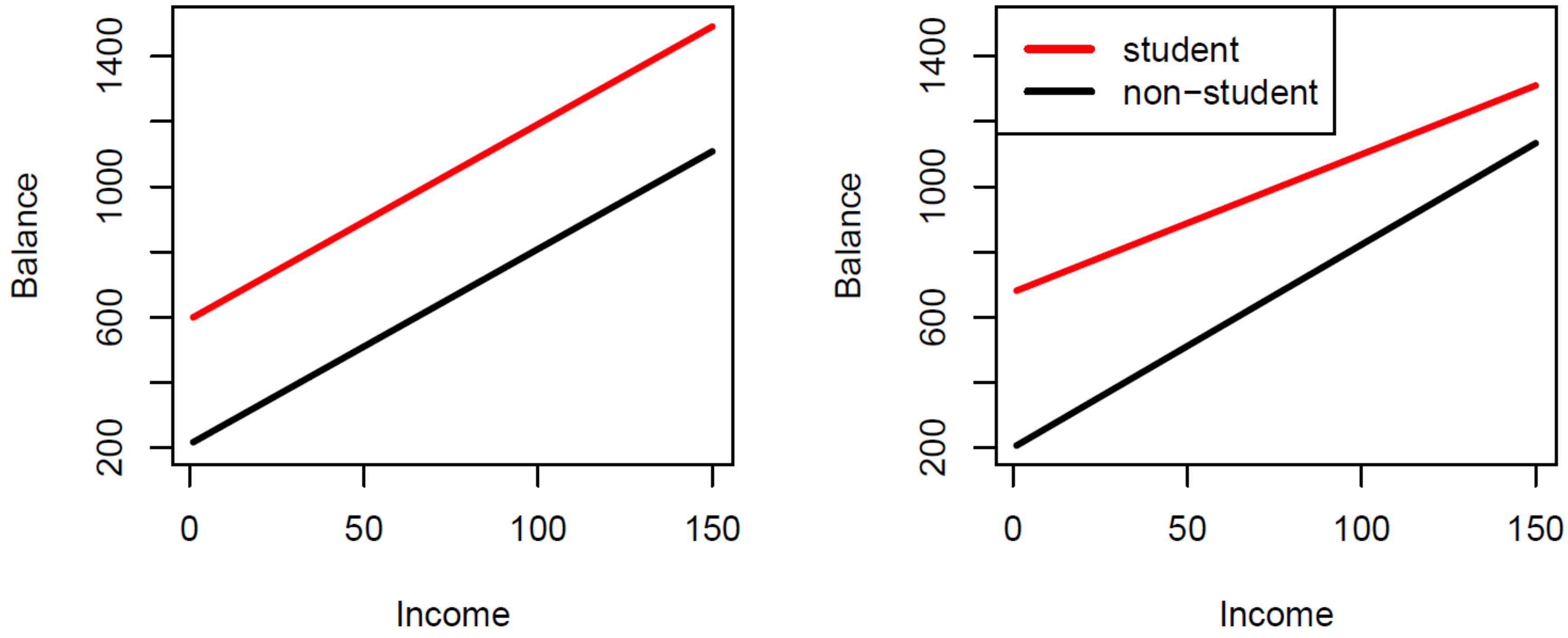
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Fitting Plane in Multiple LR

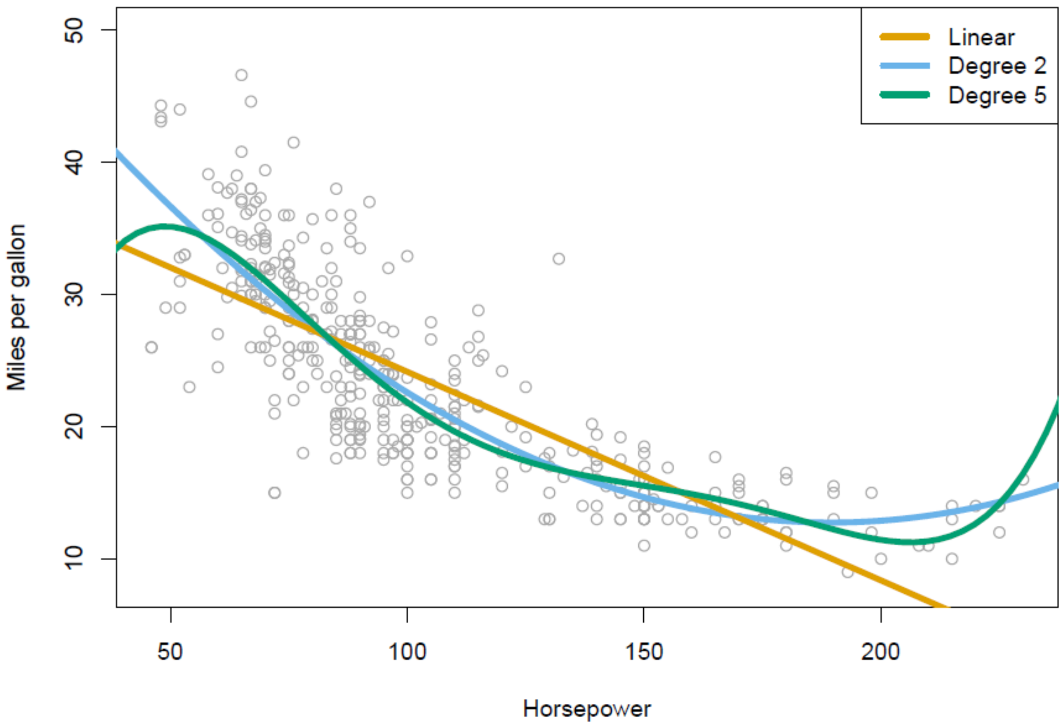


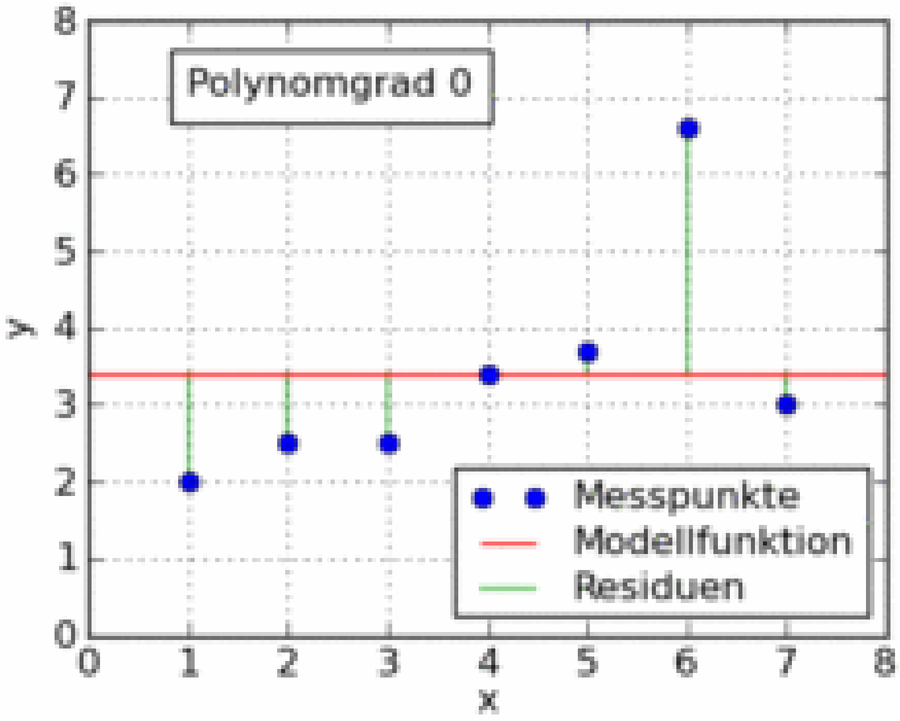


Interaction between Variables



Non-linear Effects of Predictors





REGRESSION SUMMARY

Residuals summary

Residuals:

Min	1Q	Median	3Q	Max
-15.807	-2.862	-0.670	1.418	57.370

Coefficients

Coefficients:

	2	3	4	5
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-141.400	15.480	-9.13	1.19e-14 ***
height	3.891	0.262	14.880	< 2e-16 ***
Calorie	0.01249	0.003644	3.429	0.000899 ***
exercise.level Sedentary	-0.8732	2.500	-0.349	0.727685
exercise.level Very Active	-0.8815	2.478	-0.356	0.722860

7

Signif. codes: '***' p <= 0.001 ; '**' p <= 0.01 ; '*' p <= 0.05 ; '.' p <= 0.1; ' ' p <= 1

Residual standard error: 8.968 on 95 degrees of freedom

9

Model quality summary

Multiple R-squared: 0.8058 ; **Adjusted R-squared:** 0.7976

F-statistic: 98.53 on 4 and 95 DF, **p-value**: < 2.2e-16

1. **Residuals** – Distribution has to around zero mean for best fit.

2. **Coefficients**
   a. **Estimate** – Change in Y for unit increase in variable
   b. **Std. Error** – SE of **Estimates** (dispersion of coefficient) – gives precision
   c. **t-Value** – Coefficient estimate / SE of Estimate Measure of the precision with which the regression coefficient is measured. If a coefficient is large compared to its SE, then it is probably different from 0
   d. **p-Value** - probability that the variable is NOT relevant. A small p-value indicates that it is unlikely that a relationship between target and predictor exists due to chance.

REGRESSION SUMMARY Cont..

Residuals summary

Residuals:

Min	1Q	Median	3Q	Max
-15.807	-2.862	-0.670	1.418	57.370

Coefficients

Coefficients: 2 3 4 5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-141.400	15.480	-9.13	1.19e-14 ***
height	3.891	0.262	14.880	< 2e-16 ***
Calorie	0.01249	0.003644	3.429	0.000899 ***
exercise.level Sedentary	-0.8732	2.500	-0.349	0.727685
exercise.level Very Active	-0.8815	2.478	-0.356	0.722860

7

Signif. codes: '***' p <= 0.001 ; '**' p <= 0.01 ; '*' p <= 0.05 ; '.' p <= 0.1; ' ' p <= 1

Residual standard error: 8.968 on 95 degrees of freedom

9

Model quality summary

Multiple R-squared: 0.8058 ; **Adjusted R-squared:** 0.7976
F-statistic: 98.53 on 4 and 95 DF, **p-value**: < 2.2e-16

3. **Std. error of Residuals** – SE of **Residuals**
4. **Multiple R-Squared** (*Coefficient of determination*) It determines how well model fits the actual data. It also means the proportion of variance in target explained by model. Thus 0 means regression does not explain any variability in the target variable and a number close to 1 does explain all the observed variance in the response variable.
5. **Adjusted R-Squared** – Subsequent variable inclusion in regression will over-estimate via Multiple R-Squared thus has to be penalize to balance the no. of variables taken into account. Adjusted R^2 increases only if the new term improves the model more than what would be expected by chance.
6. **F-statistic** – Overall model fit by testing whether the predictor variables, taken together, predicts the response variable above chance levels. The further F-statistic is from 1, the higher the likelihood of the existence of relationship between dependent and independent variables.
7. **Overall P-value**

Linear Model Selection and Regularization

MODEL SELECTION

Why might we want to use another fitting procedure instead of least squares?
Alternative fitting procedures can yield better

1. Prediction **Accuracy**
2. Model **Interpretability**

But how to chose which model is best among alternative?
And what are alternatives in balancing above two points?

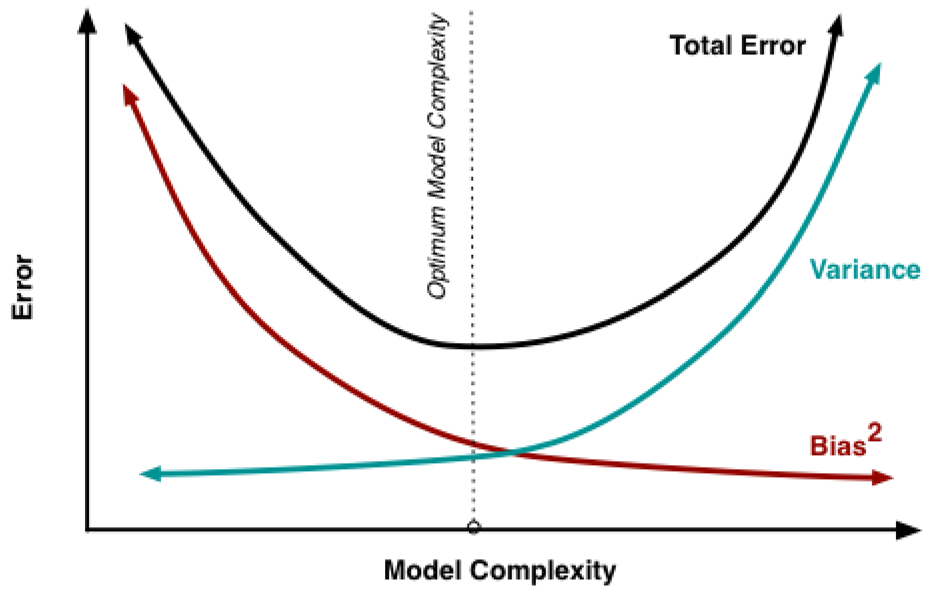
Three important classes of methods of Balancing Accuracy and Interpretability are

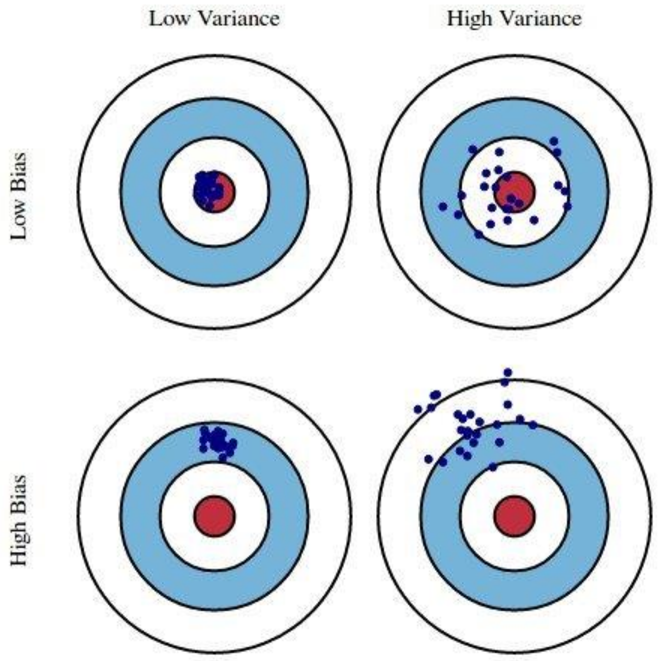
1. **Subset Selection** – Selecting Model built on only subset of variables
2. **Shrinkage** – Penalizing estimated coefficients of model built on all variables to interpretability
3. **Dimension Reduction -** P*rojecting all* predictors into a *M*-dimensional subspace, where $M < p$ and thus using those projections

1. Subset Selection

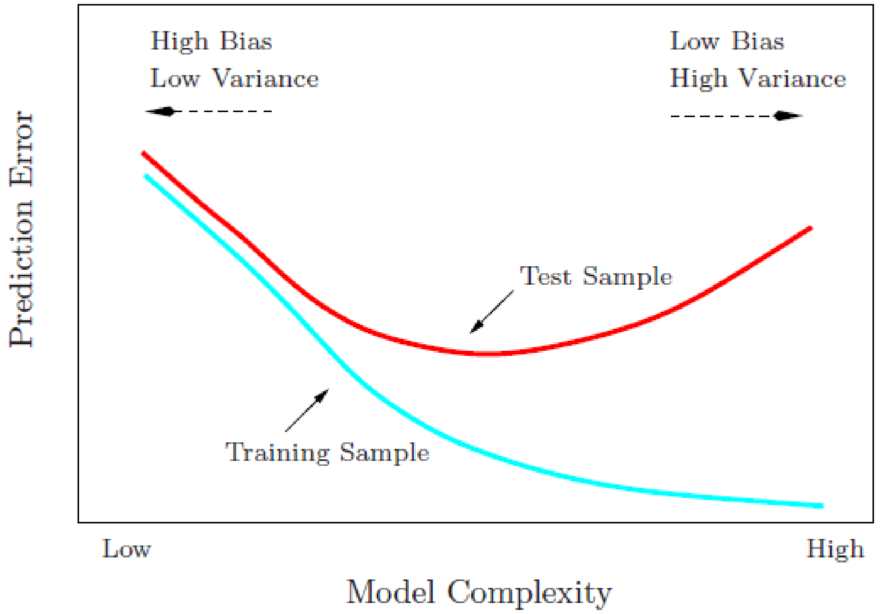
Balance between Prediction **Accuracy** vs Model **interpretability**

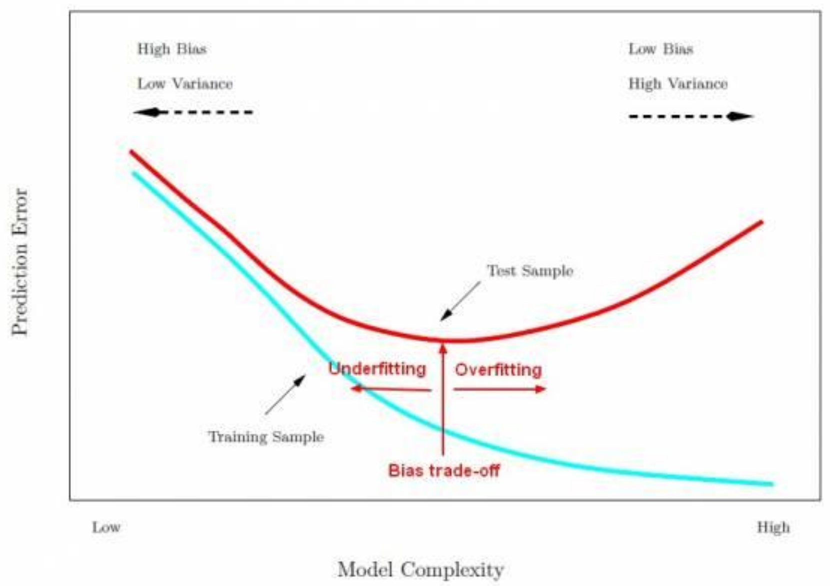
Bias Variance Tradeoff





Under-fitting vs Over-fitting





Choosing the Optimal Model which balance Accuracy and Interpretability–

1. ***Indirect Estimate -*** *Adjustment* to the training error by restricting rampant variable use
2. ***Direct Estimate –*** *Check on subset of dataset known as test to limit learning noise*

Choosing the Optimal Model - Indirect Estimate

Estimate test error by making an *adjustment* to the training error to account for the bias due to over-fitting.

But before going into model building lets discuss metrics which can be used to compare between performance of different models.

Mallow's Cp

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

C_p statistic adds a penalty to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error

Akaike information criterion

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

For least squares models, C_p and AIC are proportional to each other

Bayesian information criterion

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right)$$

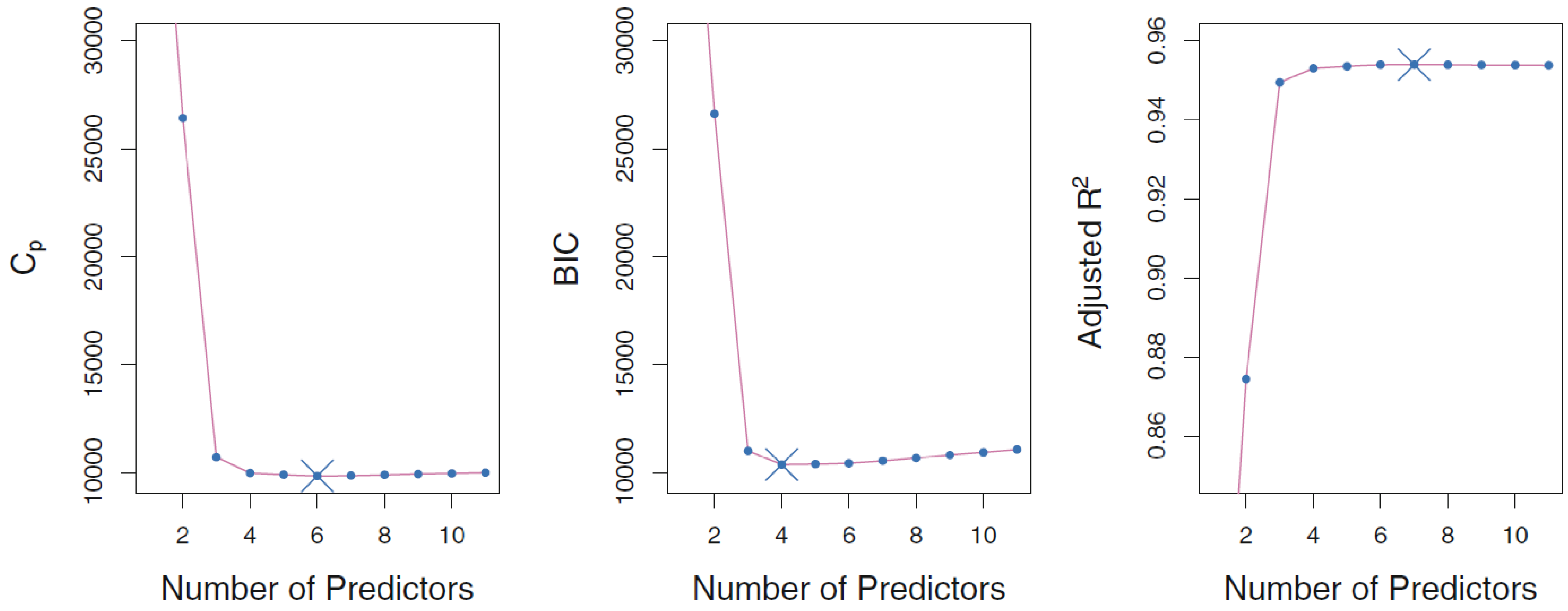
Replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$, thus penalize more

Adjusted R-Square

$$1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

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

Indirect Estimate - Comparison



Indirect Estimate - Subset & Stepwise Selection

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

Rating dropped from combination of 4 variables in Best Subset

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Hybrid Approach

After adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.

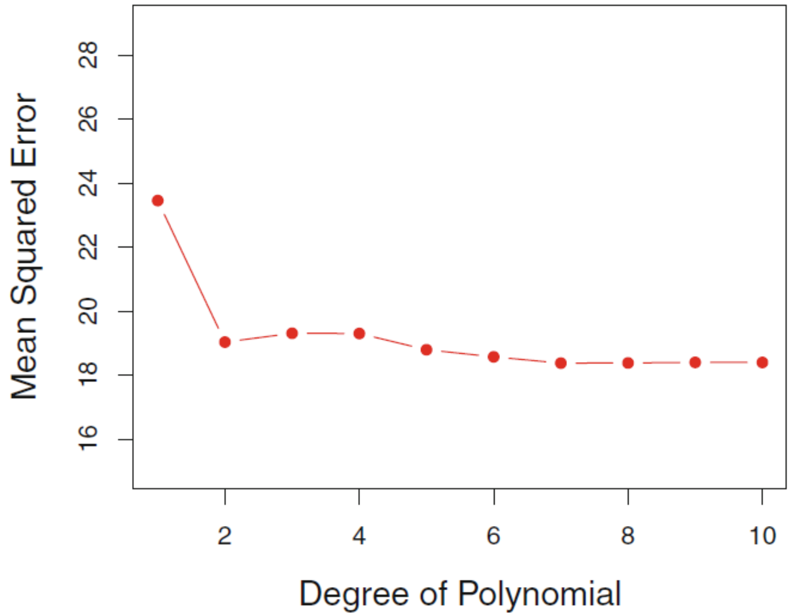
Choosing the Optimal Model – Direct Estimate

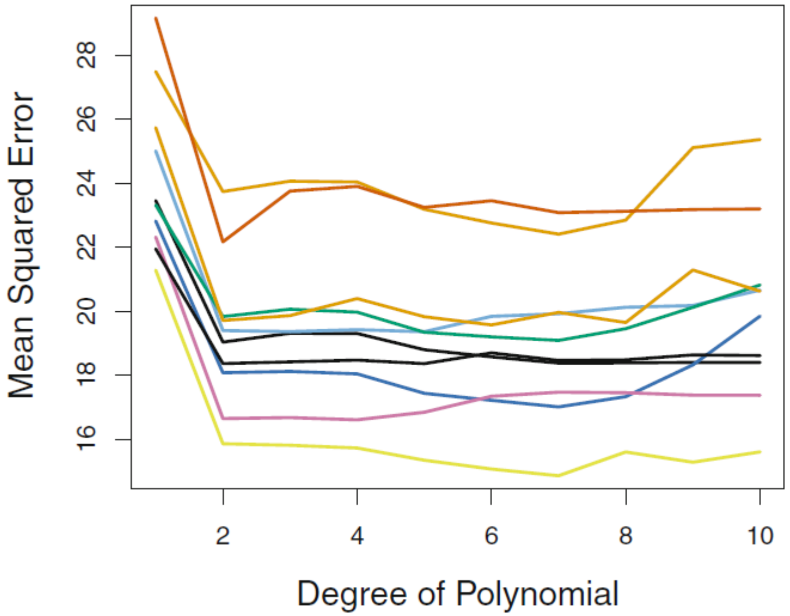
Directly estimate the test error using the **Validation** set and **Cross-validation** methods

Validating model on subset of training data known as Validation set
Checking the model fit using k-fold cross validation

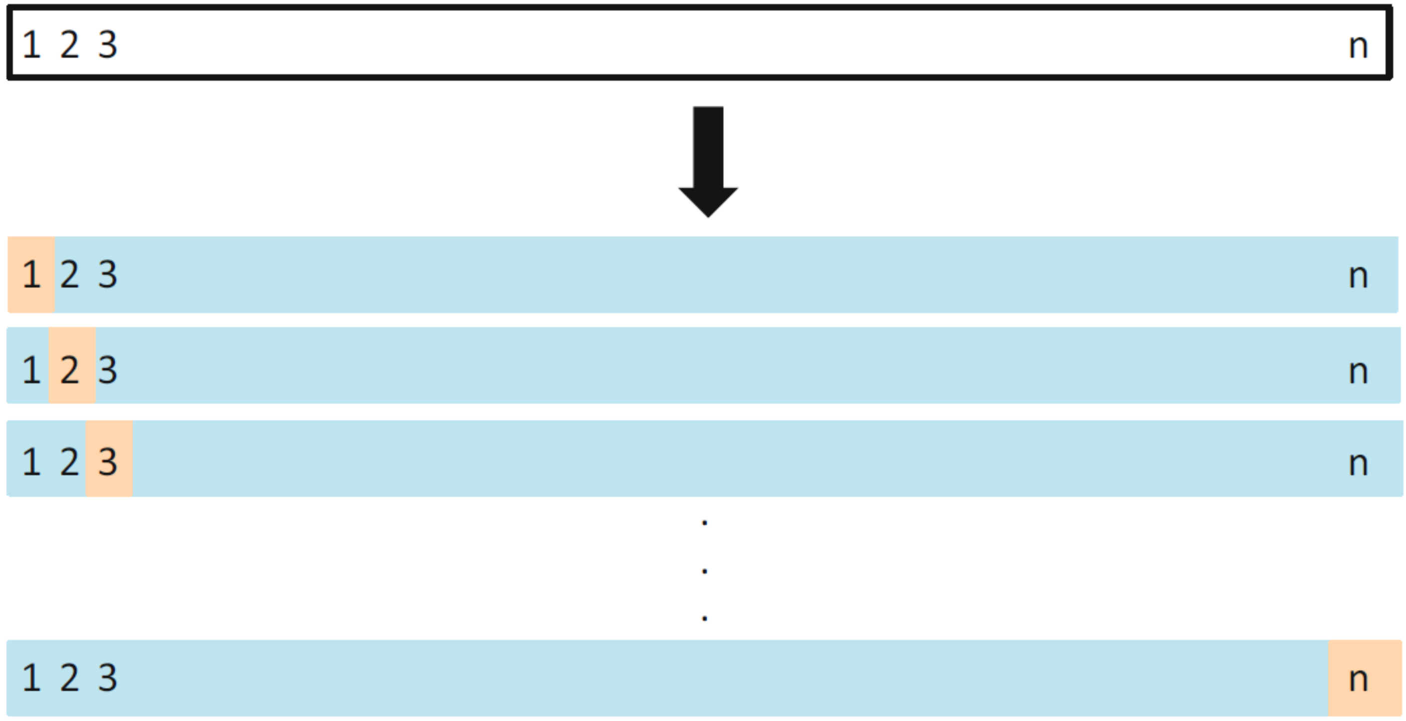
When different models show almost same error at different predictors then one should follow *"one-standard-error rule"* to select model with minimum no of predictors

VARIANCE IN MODEL

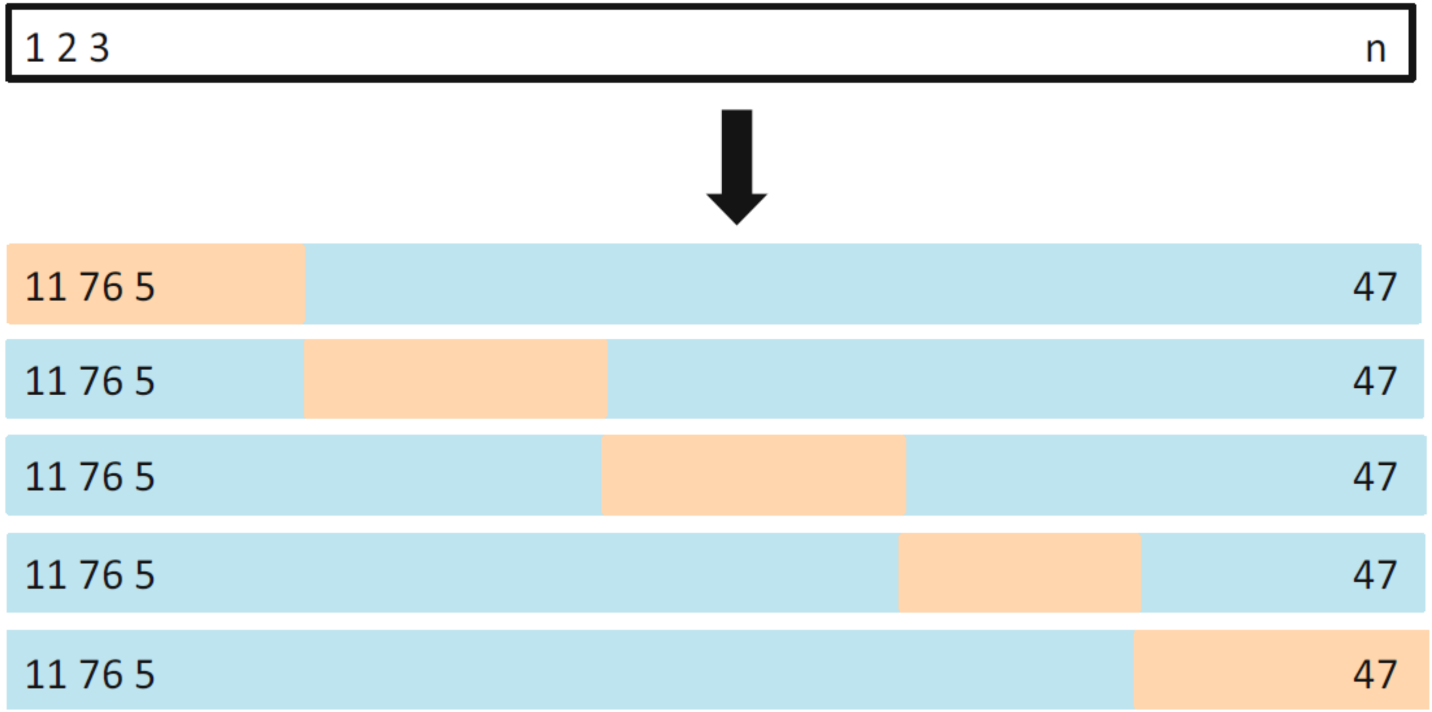




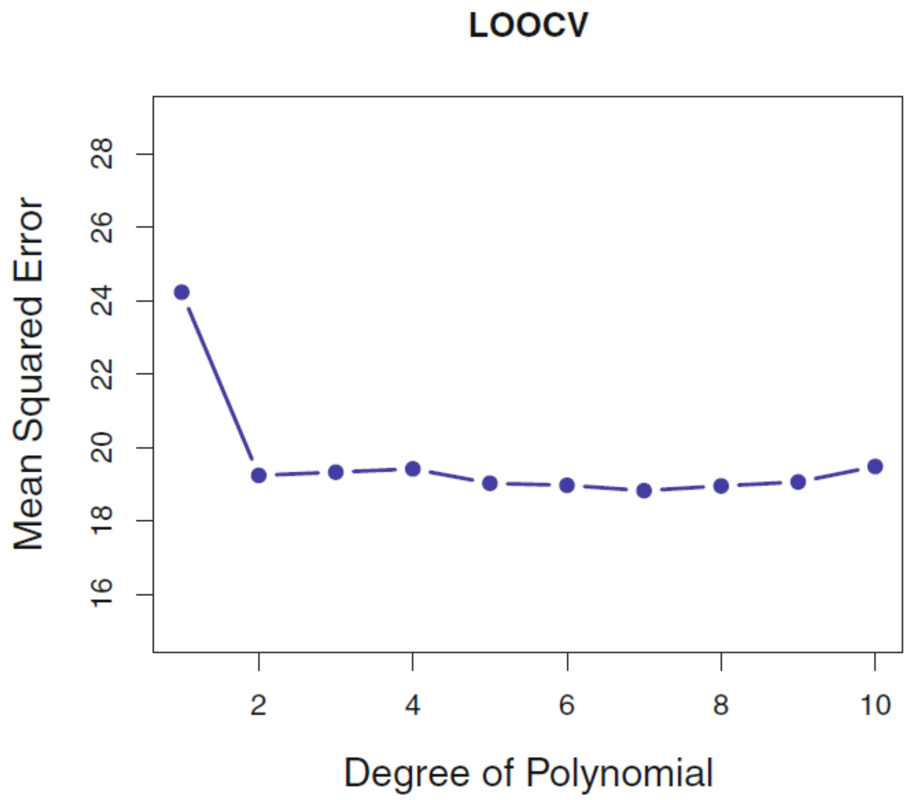
LEAVE ON OUT – CROSS VALIDATION

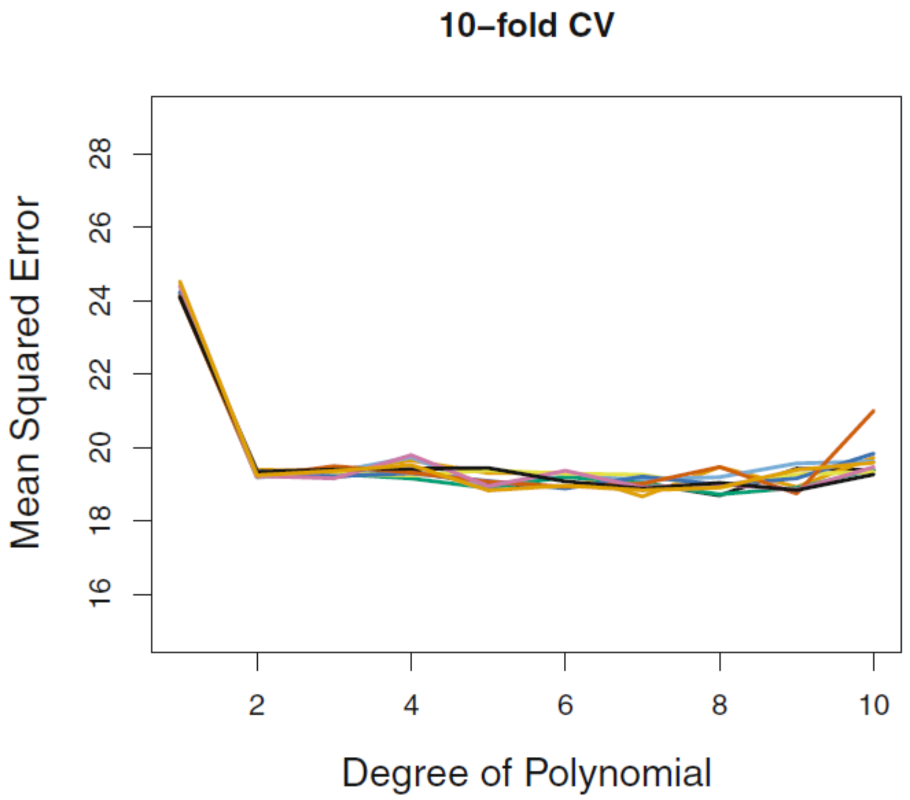


K – FOLD CROSS VALIDATION



DIFFERENCE IN LOO-CV & K-FOLD CV





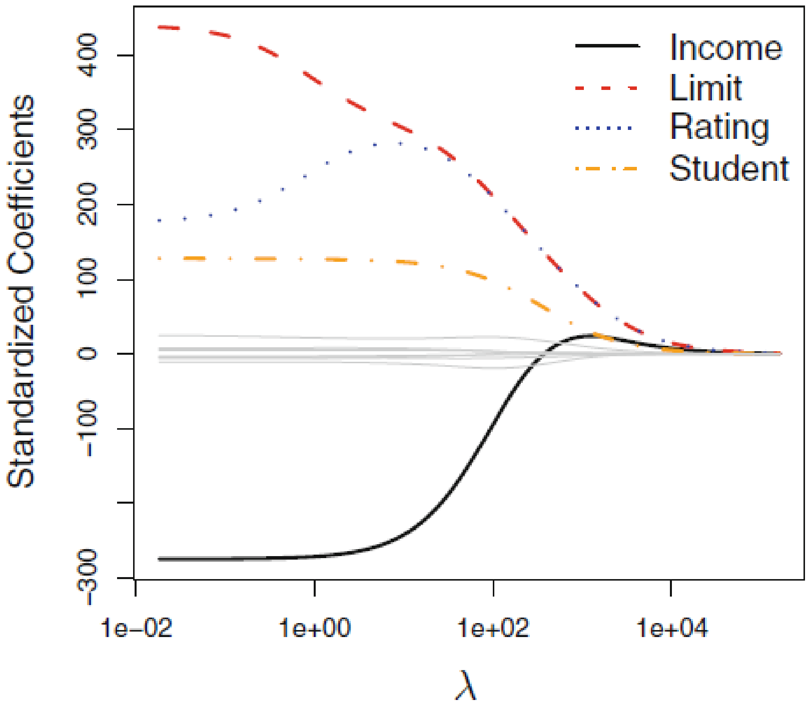
2. Shrinkage

Shrinkage Methods - Ridge

Motivation: - Instead of subset of variables use all but shrinks the coefficient estimates towards zero thus reduce their variance

In Ridge regression the coefficients ridge are estimated by minimizing below error term, which is extension of OLS. This makes the coefficients small as it has penalty component on it which is parameterized by *lambda (shrinkage penalty)*
the shrinkage penalty is applied to *β*1*, . . . ,* *βp*, but not to the intercept *β*0.

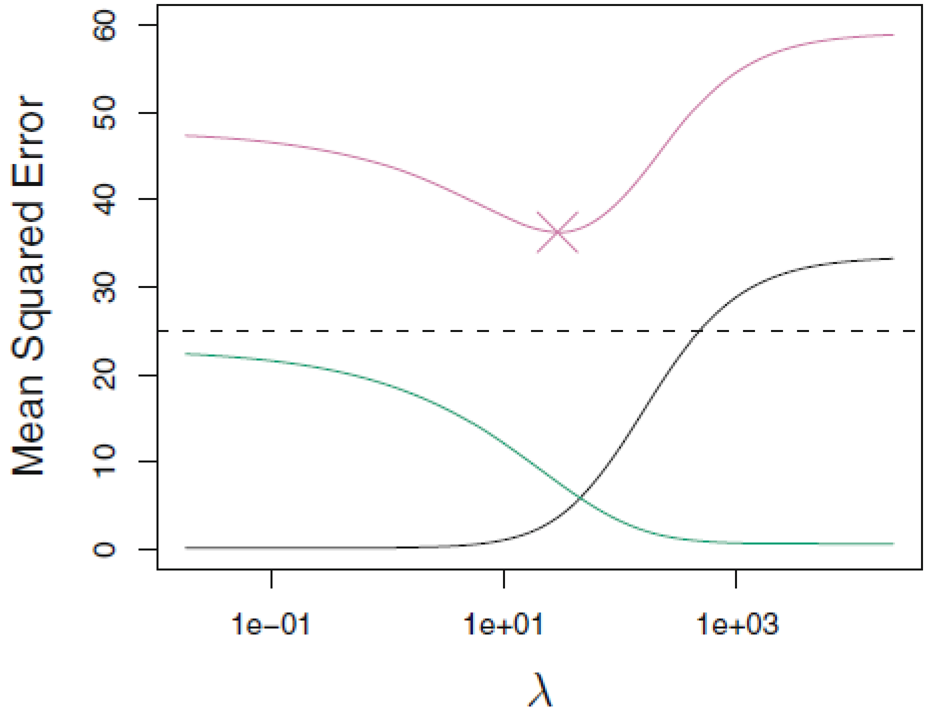
$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$



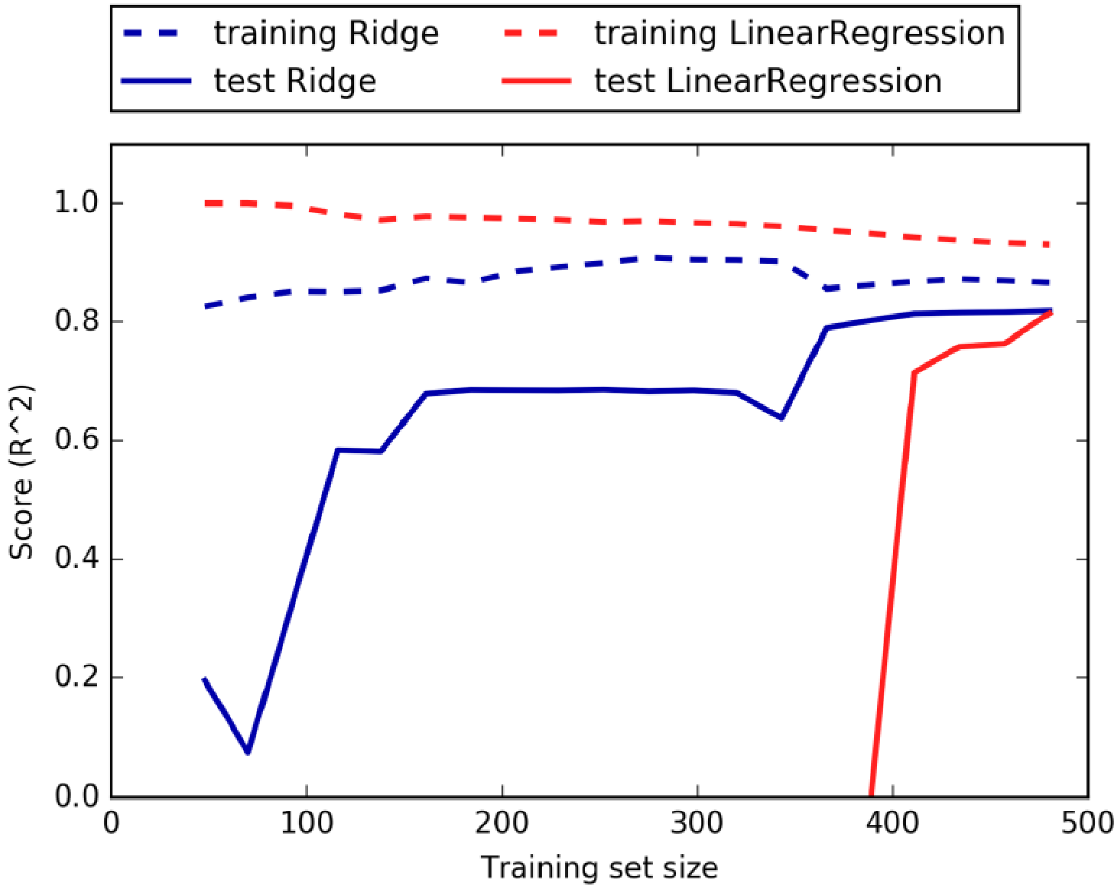
Why Does Ridge Regression Improve Over Least Squares?

Ridge regression's advantage over least squares is rooted in the *bias-variance trade-off*.

As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.



Training Data Size and its Impact on Ridge & Training Error

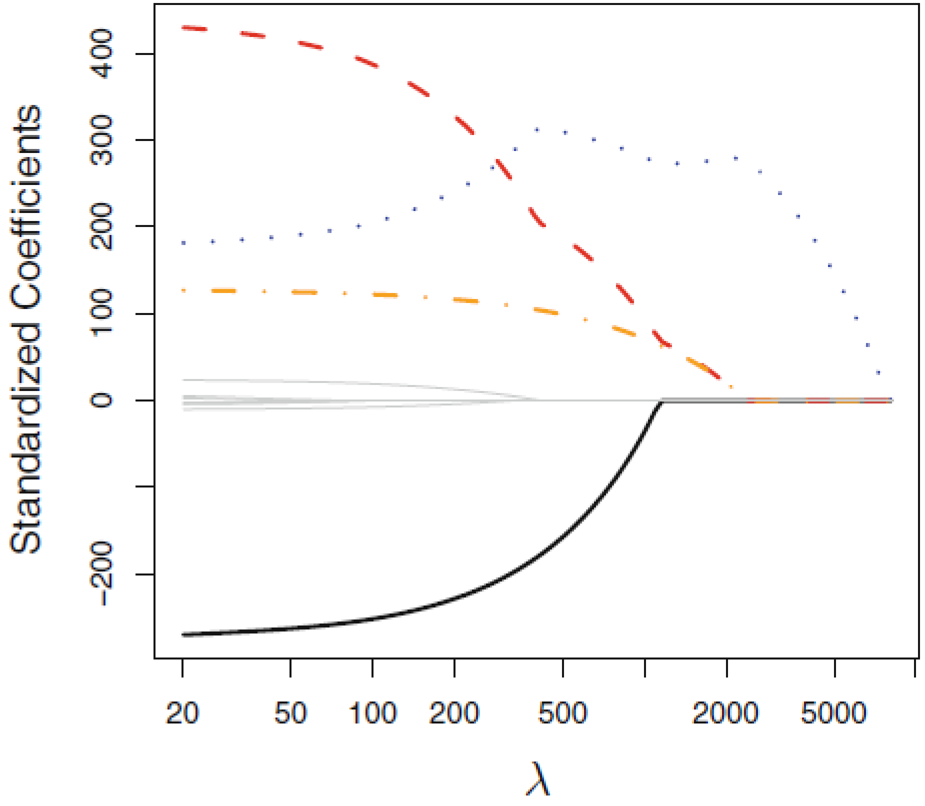


Shrinkage Methods - Lasso

Motivation: - Unlike Best subset, stepwise Ridge fails to do feature selection thus model interpretation becomes challenge

Lasso penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large, thus enables variable selection

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



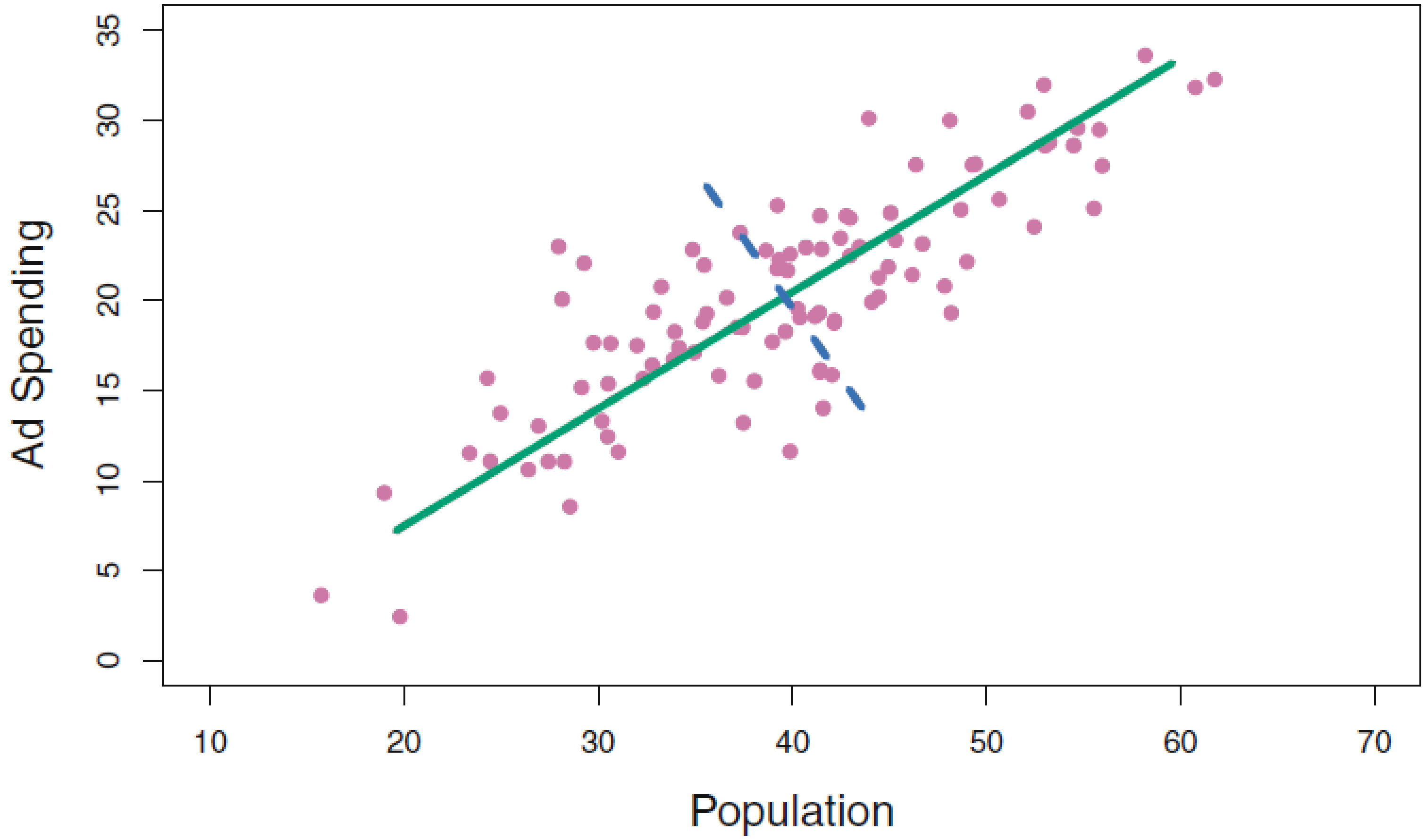
Which one is better? Ridge or Lasso

Lasso perform better in a setting where a relatively small number of predictors have substantial coefficients. Ridge regression will perform better when the response is a function of many predictors.

Since this is not known a priori so cross-validation can be used in order to determine which approach is better on a particular data set.

3. Dimension Reduction

Dimension Reduction Methods ***- transform*** the predictors



Green solid line indicates the 1st principal component

Blue dashed line indicates the 2nd principal component

Step1 - Transformed predictors $Z_1, Z_2, \ldots, Z_M$ are obtained
Step2 - Model is fit using these M predictors.


hungover