Applied Regression Analysis Classnotes

Dustin Leatherman

November 24, 2019

Contents

1 Session 1 - Summary and Review

1.1 Relationships

1.1.1 Functional

Mathmatical formula
$Y = f(x)$

1.1.2 Statistical

Not a perfect relationship
$Y = f'(x) + \epsilon$
observations = trials = case
quadratic = curvilinear
Linear models means that the *slope* is not raised to any powers

- $\hat{y} = \beta_0 + \beta_1 X^2$ is linear
- $\hat{y} = \beta_0 + \beta_1^2 X$ is **not** linear

1.2 Basic Concepts

- Tendency of Y to vary with X in a *systematic fashion*
- scatter of points around a curve of a statistical relation

- Prob. Distr of Y for each **Level** of X
- means of Y's distr. to vary for each value of X
  - each point on the regression line can be represented as $\mu_{Y|X_i}$

NOTE: Sir Francis Galton came up with the term "regression"

Regression function of Y on X: Means of the prob. distr. have a systematic relation to the Level of X

Regression curve: graph of the regression function

1.2.1 Construction of Regression models

1. Selection of pred. vars
2. Functional form of the regression relation
   - Summary plots
   - Scatter plots
3. Scope of Model
   - **Scope**: What is the Domain? (range of X's)
   - Making predictions outside the Domain is considered extrapolation and is dangerous

1.2.2 Uses of Regression Analysis

1. Description
2. Control
3. Prediction (most abused)

Association does not imply causation!

1.3 Simple Linear Regression (SLR)

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- One Predictor
- Linear in the Parameters
- Linear in the Predictor Variables

- SLR = First-Order Model (term from outside of statistics)

Y_i: Value of the response for the ith trial β_0: Parameters (unknown. estimate these) X_i: Value of the predictor of the ith term (known) ϵ_i: random error term of the ith observation

1.3.1 Properties of ϵ_i

- $E(\epsilon_i) = 0$
- $\sigma^2(\epsilon_i) = Var(\epsilon_i) = \sigma^2$
- ϵ_i and ϵ_j are uncorrelated

1.3.2 Properties

1. Y_i is the sum of two components. It is **random** because its composed of a random term. constant term: $\beta_0 + \beta_1 X_i$ random term: ϵ_i
2. $E(Y_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) \rightarrow E(\beta_0 + \beta_1 X_i) + E(\epsilon_i) \rightarrow \beta_0 + \beta_1 X_i$
3. Y_i falls short of regression function by ϵ_i
4. $Var(\epsilon_i) = \sigma^2$ error terms have constant variance
5. Since error terms are uncorrelated, then responses (Y_i and Y_j are uncorrelated)

1.3.3 Alternative forms of SLR

1. $Y_i = \beta_0 X_0 + \beta_1 X_i + \epsilon_i$, where $X_0 = 1$
2. $Y_i = \beta_0 + \beta_1(X_i - \bar{x}) + \beta_1\bar{x} + \epsilon_i \rightarrow (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x})\epsilon_i)$\$ $\rightarrow \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon_i$

1.3.4 Method of Least Squares

1. Goal Find estimators of β_0 and β_1

   For each (X_i, Y_i): $Y_i - (\beta_0 + \beta_1 X_i)$ $Q = \sum_1^n [Y_i - \beta_0 - \beta_1 X_i]^2$

   b0 and b1 are estimators of β_0 & β_1 that minimize Q for given data (X_i Y_i), i = [1, n]

1.3.5 Gauss-Markov Theorem

1. Proof First, lets find the value of b_0 by taking the partial derivative of Q with respect to β_1

$$
\begin{aligned}
Q &= \sum_1^n [Y_i - \beta_0 - \beta_1 X_i]^2 \\
\frac{dQ}{d\beta_1} &= -2 \sum_1^n X_i [Y_i - \beta_0 - \beta_1 X_i] \\
\rightarrow \sum_1^n X_i (Y_i - b0 - b1 X_i) &= 0 \\
\rightarrow \sum_1^n X_i Y_i - b_0 \sum_1^n x_i - b_1 \sum_1^n x_i^2 &= 0 \\
\rightarrow \sum_1^n Y - i - n b_0 - b_1 \sum_1^n x_i &= 0 \\
\rightarrow \sum_1^n Y_i - b_1 \sum_1^n x_i &= n b_0 \\
\rightarrow \bar{Y} - b_1 \bar{x} &= b_0
\end{aligned}
\tag{1}
$$

Once b_0 is found, lets use it to find the value of b_1. Replace values of b_0 with the equation above.

$$
\begin{aligned}
&\sum_1^n X_iY_i - b_0\sum_1^n x_i - b_1\sum_1^n x_i^2 = 0 \\
&\rightarrow \sum_1^n X_iY_i - (\bar{Y} - b_1\bar{x})\sum_1^n x_i - b_1\sum_1^n x_i^2 = 0 \\
&\rightarrow \sum_1^n X_iY_i - (\frac{\sum_1^n Y_i}{n} - b_1\frac{\sum_1^n x_i}{n})\sum_1^n x_i - b_1\sum_1^n x_i^2 = 0 \\
&\rightarrow \sum_1^n X_iY_i - \frac{\sum_1^n x_i \sum_1^n y_i}{n} + b_1\frac{(\sum_1^n x_i)^2}{n} - b_1\sum_1^n x_i^2 \\
&\rightarrow \sum_1^n x_iy_i - \frac{\sum_1^n x_i \sum_1^n y_i}{n} = b_1[\sum_1^n x_i^2 - \frac{(\sum_1^n x_i)^2}{n}]
\end{aligned}
$$

$$
= ... = \frac{\sum_1^n (x_i - x)(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2} \tag{2}
$$

2. Properties
   (a) $E(b0) = \beta_0$ & $E(b1) = \beta_1$
   (b) b0 & b1 are more precise than any other unbiased estimators of β_0 and β_1 that are linear functions of Y_i

1.3.6 Residual

Difference between the observation and the estimated value $_i = Y_i - \hat{Y}_i$, i == [1, n]

1. $\sum_i^n e_i = 0$
2. $\sum_i^n e_i^2$ is a minimum
3. $\sum_i^n Y_i = \sum_i^n \hat{Y}_i$

1. Goal Estimate σ^2 know $E(S^2) = E(\frac{\sum(Y_i - \bar{Y})^2}{n-1})$
   - numerator == sum of squares
   - n - 1 == df
   - S^2 = Mean Square = $\frac{SS}{df}$

2. SSE SSE = $\sum(Y_i - \hat{Y}_i)^2 = \sum e_i^2$
   - SSE = Sum of Square Error = Residual Sums of Squares
   - MSE = SSE / n - 2
   - df of SSE = n - 2
   - E(MSE) = σ^2

1.4 Normal Error Regression Model

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon \approx iidN(0, \sigma^2)$, i = [1, n] so $Y_i \approx N(\beta_0 + \beta_1 X_i, \sigma^2)$
To find MLE's of β_0 & β_1 i.e. $\hat{\beta}_0$ & $\hat{\beta}_1$ $L(\beta_0, \beta_1 m \sigma^2) = \prod pdf$

- MLE of β_0: $\hat{\beta}_0 = b_0$
- MLE of β_1: $\hat{\beta}_1 = b_1$

2 Session 2 - Inferences in Regression and Correlation Analysis (2019/09/18)

Model = $Y_i = \beta_0 \beta_1 X_i + \epsilon_i$

2.1 Properties

- $\epsilon_i \approx iidN(0, \sigma^2)$
- $Y_i \approx iidN(\beta_0 + \beta_1 X_i, \sigma^2)$
- X_i: known constant
- β_0 & β_1 are parameters to investigate

2.2 β_1

2.2.1 Inferences

$H_0 : \beta_1 = 0$ (implies no linear association) $H_1 : \beta_1 \neq 0$
This hypothesis test determines if there is a relationship

2.2.2 Sampling Distribution

$b_1 = \frac{\Sigma((x_i - \bar{x})(y_i - \bar{y}))}{\Sigma(x_i - \bar{x})^2}$

- $E(b_1) = \beta_1$
- $Var(b_1) = \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}$

2.2.3 PROOF: b_1 is a linear combination of Y's

- $b_1 = \frac{\Sigma((x_i - \bar{x})(y_i - \bar{y}))}{\Sigma(x_i - \bar{x})^2}$
- $b_1 = \frac{\Sigma((x_i - \bar{x})y_i - \bar{y}\Sigma(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$
- $b_1 = \frac{\Sigma((x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}$

Let $K_i = \frac{x_i - \bar{x}}{\Sigma(x_i - \bar{x})^2}$
Facts about K_i

- $\Sigma K_i = \Sigma \frac{x_i - \bar{x}}{\Sigma(X_i - \bar{x})^2} = 0$
- $\Sigma K_i^2 = \Sigma(\frac{x_i - \bar{x}}{\Sigma(X_i - \bar{x})^2)^2} = \frac{1}{\Sigma(x_i - \bar{x})^2}$
- $b_1 = \Sigma K - iY_i$

Therefore b_1 is a linear combination of $\mathrm{Y_i}$

2.2.4 Properties

- $E(\hat{\beta}_1) = E(\Sigma K_i Y_i) = \Sigma K_i E(Y_i) = \Sigma K_i(\beta_0 + \beta_1 X_i) = \beta_1 \Sigma K_i X_i = \beta_1$

More detailed proof of $\Sigma K_i X_i = 1$ exists in notes. It was a sidebar in class.

- $\beta_1 \approx N(\beta_1, \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2})$
- $\frac{b_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}}} \approx N(0, 1)$

Recall E(MSE) $= E(\frac{SSE}{n-2}) = \sigma^2$
Thus $\frac{b_1 - \beta_1}{\sqrt{\frac{MSE}{\Sigma(x_i - \bar{x})^2}}} \approx t_{n-2}$ NOTE: a T Distribution is a standard normal distribution divided by a chi-square distribution scaled by its DF

1. Solving the Hypothesis Test

   Recall

   $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

   Test Statistic

   $t* = \frac{b_1}{\sqrt{\frac{MSE}{\Sigma(x_i-\bar{x})^2}}} = \frac{b_1}{SE_{b1}} \approx t_{n-2}$

   Then p-value can be calculated

2.3 β_0

$b_0 \approx N(\beta_0, \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i-\bar{x})^2}])$

If Y_i are not exactly normal, b_0 and b_1 are approx. normal. Thus the t statistic provides some level of confidence.

2.4 Spacing of X Levels

- The greater the spread of x, the larger $\Sigma(x_i - \bar{x})^2$
- Var(b_1) and Var(b_0) decrease

2.5 Prediction of new observations

Let a new observation be defined as Y_0

2.5.1 Interval Estimation of $E(Y_0)$

- X_0: level of x we want to estimate the mean response
- $E(Y_0)$: mean response when $X = X_0$
- $\hat{Y_0} = b_0 + b_1X_0$: Point estimate of $E(Y_0)$

2.5.2 Sampling Distribution

$\hat{Y_0} \approx N(E(Y_0), \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}])$

$\hat{Y_0} \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE(\frac{1}{n} + \frac{(X_0-\bar{x})^2}{\Sigma(x_i-\bar{x})^2})}$

NOTE: **confidence interval == mean prediction interval == single value**

2.5.3 Prediction

$\hat{Y}_1$: predicted individual outcome drawn from the distr. of Y

Assumptions

- $E(Y_1)$: estimated by $\hat{Y}_1$
- Var(Y_1): estimated by MSE

$Var(pred) = Var(\hat{Y}_1) + Var(\hat{Y}_0) = \sigma^2[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}]$

100(1 − α)% Prediction Interval

- $\hat{Y}_1 \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\Sigma(x_i-\bar{x})^2})}$

2.6 ANOVA Approach to Regression

Partition the Total Sums of Squares

1. When ignoring the predictor variable, Variation is based on $Y_i - \bar{Y}$ deviations.

   $SSTo$: Total Sums of Squares (or TSS) Therefore, $SSTo = \Sigma(Y_i - \bar{y})^2$

2. When using the predictor variable, variation based on $Y_i - \hat{Y}_i$ deviations. i.e. residuals

   SSE: Error Sum of Squares Therefore, $SSE = \Sigma(Y_i - \hat{Y}_i)^2$

SSR: Regression Sum of Squares $SSR = \Sigma(Y_i - \bar{y})^2$

NOTE: SSR = SSTo - SSE **OR** SSTo = SSR + SSE. proof is in notebook. record here if needed

Degrees of Freedom (df)

- SSto: n - 1. $Y_i - \bar{y}$
- SSE: n - 2. $Y_i - \hat{Y}_i$
- SSR: 2 - 1 = 1. $\hat{Y}_i - \bar{y}$

NOTE:

- $E(MSE) = \sigma^2$
- $E(MSR) = \sigma^2 + \beta_1^2\Sigma(x_i - \bar{x})^2$

Source	SS	df	MS	F Statistic
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	SSE	n - 2	$MSE = \frac{SSE}{n-2}$	
Total	SSTo	n - 1		

$F*$ is the test statistic for
$H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$
$F* \approx F_{1,n-2}$ if H_0 is true $(t*)^2 = F*$ if $F* \approx F_{1,n-2}$

3 Session 3 - General Linear Testing & Model Selection (2019/09/25)

3.1 General Linear Test Approach

Full Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \approx iidN(0, \sigma^2)$

This can be fit by either Least Squares or Maximum Likelihood

Notes F = `Full Model` `R =` Reduced Model

$$\begin{aligned} SSE(F) &= \Sigma[Y_i - (b_0 + b_1 X_i)]^2 \\ &= \Sigma(Y_i - \hat{Y}_i)^2 \\ &= SSE \end{aligned} \tag{3}$$

3.1.1 Reduced Model

$$\begin{aligned} &H_0 : \beta_1 = 0 \text{ if } H_0 \text{ then } Y_i = \beta_0 + \epsilon_i \\ &H_A : \beta_1 \neq 0 \end{aligned} \tag{4}$$

Test Statistic: $SSE(F) \leq SSE(R)$

The more parameters in the model, the better the fit **thus** smaller deviations around the fitted regression model.

A small diff suggests H_0 holds. $(SSE(R) - SSE(F))$

$$F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} \tag{5}$$

Note: The full model has less variation because the hope is that the predictor (X) helps explain the spread in the response (Y).

p-value $= P(F_{df_R - df_F, df_F} \geq F^*)$ For SLR and testing the null hypothesis $(H_0 : \beta_1 = 0)$,

$$
\begin{aligned}
F^* &= \frac{\frac{SSTo - SSE}{(n-1)-(n-2)}}{\frac{SSE}{n-2}} \\
&= \frac{SSR}{MSE} \\
&= \frac{MSR}{MSE}
\end{aligned}
\tag{6}
$$

This is exactly like the ANOVA table!

3.1.2 Coefficients of Determination (R^2)

Goal: Quantify how much variation in the repsonse is explained by the model. **Def**: The proportion of variation in Y explained by regressing Y on X.

$R^2 = \frac{SSR}{SSTo} = 1 - \frac{SSE}{SSTo}$

Properties

- $0 \leq R^2 \leq 1$
- $R^2 = 1$ indicates a perfect fit
- $R^2 = 0 \rightarrow b_1 = 0$ thus a horizontal line **OR** a non-linear pattern

A high R^2 value does NOT indicate

- useful predictions can be made
- estimated regression line is a good fit
- x and y are related

3.1.3 Coefficient of Correlation: $r = \pm\sqrt{R^2}$

A measure of the linear association between Y and X when Y and X are random variables. **Properties**

- $-1 \leq r \leq 1$
- sign of correlation matches sign of slope

3.2 Assessing the Quality of a Model

Diagnostics for X (predictor variable)

1. Dot Plot
2. Sequence Plot $X_1, ..., X_n$. No pattern is good
3. Stem-and-Leaf plot ($<$ 100 observations)
4. Box Plot
5. Histogram

3.2.1 Residuals (observed error)

$e_i = Y_i \hat{Y}_i$

Properties

- $\bar{e} = \frac{\Sigma e_i}{n} = 0$
- $S_e^2 = \frac{\Sigma(e_i - \bar{e})^2}{n-2} = \frac{\Sigma e_i^2}{n-2} = \frac{SSE}{n-2} = MSE$
- e_i's are **not** independent random variables.
  - If large n, the dependence of e_i is relatively unimportant and can be ignored

Standardized vs Studentized

- Standardized $= \frac{Y_i - \bar{y}}{\sigma}$
- Studentized $= \frac{Y_i - \mu}{\frac{\sigma}{n}}$

Semi-studentized Residuals $e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$

3.2.2 Residual Plots

Residual Plot Form

1. Tests
   (a) Non-linearity of regression function A pattern indicates linear regression not appropriate

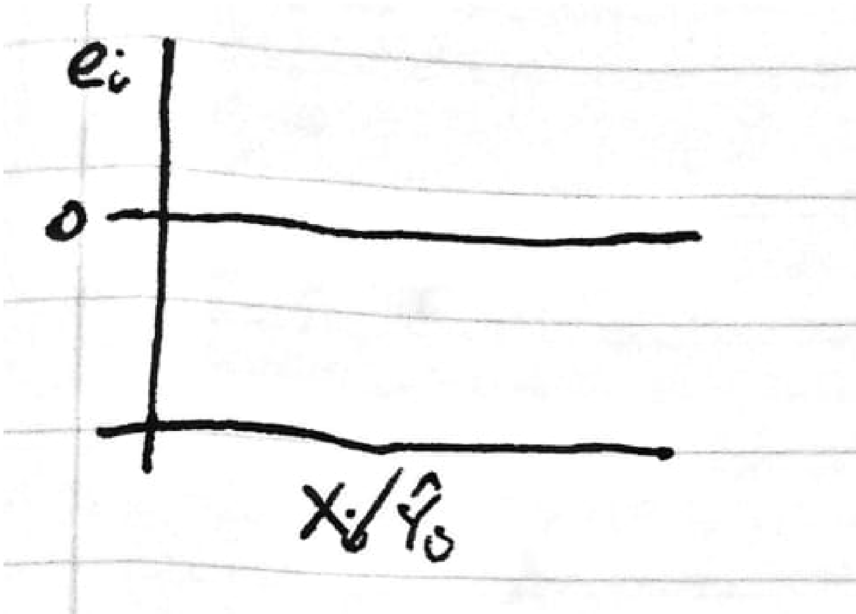

Figure 1: Empty Residual Plot

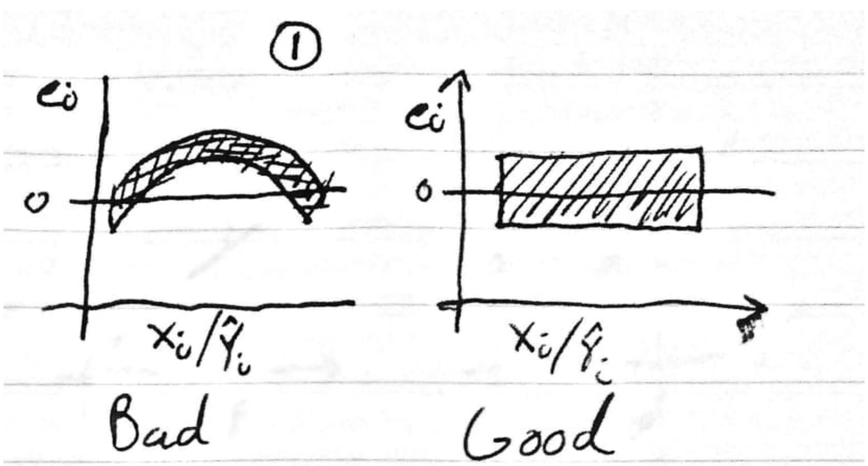

Figure 2: Plots 1

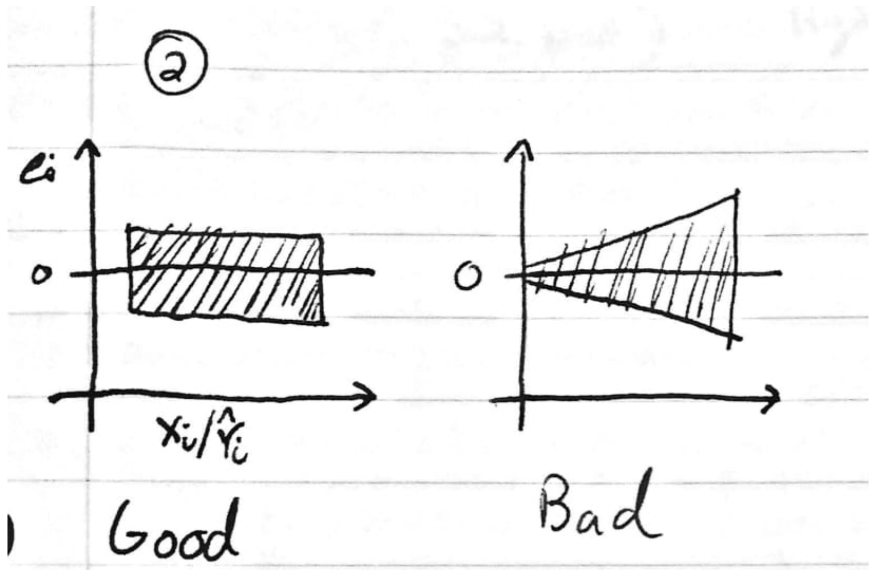

Figure 3: Plots 2

(a) Non-constancy of error terms Fanning indicates different variances for different values of $X_i or \hat{Y}_i$

(b) Presence of outliers Graph Semi-studentized residuals on a Residual plot **OR** a Box Plot
if $|e_i^*| \geq 4$, outlier

(c) Non-independence of error terms (more of a concern with time-series) No pattern is good. Error terms safe to assume independent.

(d) Normality of Error Terms

- Use a normal probability plot. The closer the points the fall on a straight line, the closer they are to a normal distribution.

(e) Omission of Important Predictors? A Pattern indicates that there might be a relationshup between the residuals and some other predictor. This can be used to determine whether a predictor shoudl be used before modeling it. Probably not as necessary anymore since it is easy to run and compare models.

3.2.3 Test of Randomness

1. Durbin-Watson Test

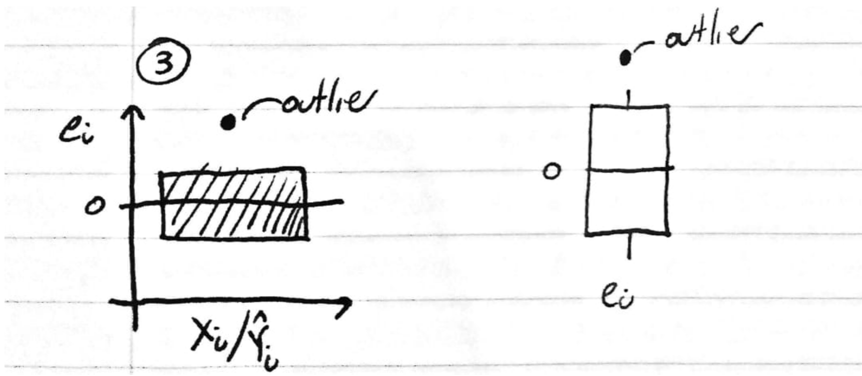

Figure 4: Plots 3

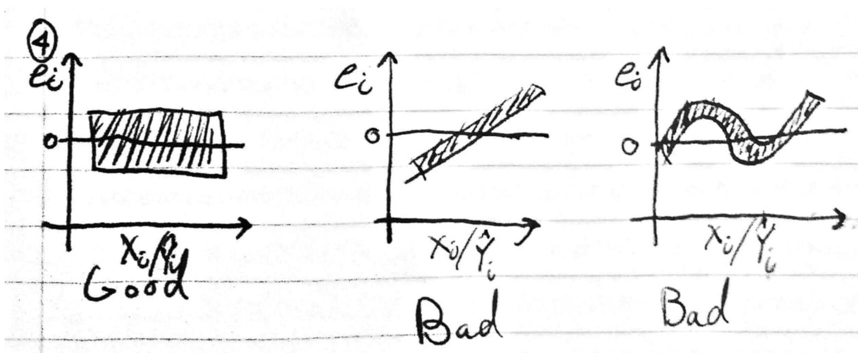

Figure 5: Plots 4

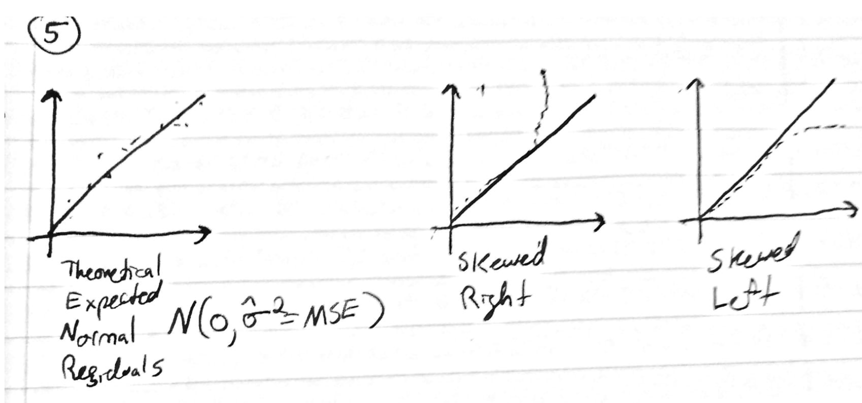

Figure 6: Plots 5

$$
\begin{aligned}
H_0 &: \phi = 0 \text{where } \phi \text{ is an autocorrelation coefficient} \\
H_A &: \phi > 0 \text{most assume positive correlation}
\end{aligned} \tag{7}
$$

```
lmtest::dwtest(modle)
```

2. Shapiro-Wilk Test for Normality Not writing much here because I know it already

```
shapiro.test()
```

3.2.4 Constant Variance

1. Brown-Forsyth Test Robust since it uses Median

```
lawstat::levene.test()
```

2. Breusch-Pagan Test Sensitive to departures from Normality
   $log(\sigma^2) = \gamma_0 + \gamma_1 x_i$

$$
\begin{aligned}
H_0 &: \gamma_1 = 0 \\
H_A &: \gamma_1 \neq 0
\end{aligned} \tag{8}
$$

```
lmtest::bptest()
```

NOTES: Heteroscedascity means non-constant variance

4 Session 4 - Transformations & Inference (2019/10/02)

4.1 Transformations

If non-normality and unequal error variance:

1. Transform Y: $Y' = f(Y)$
2. Transform X: $X' = f(X)$

If non-linearity (rarer)

1. Transform X: $X' = f(X)$

In order to determine which transformation to choose, look at the raw data and make a judgement call.

In Class Example

$Y_i' = log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i \equiv Y_i = exp(\beta_0 + \beta_1 X_i + \epsilon_) i$

A 1 unit increase in X is associated with a $exp(\beta_1)$ multiplicative effect on the **geometric** mean. This link explains in detail the impact of log transformed variables.

Geometric mean $= (\Pi x_i)^{\frac{1}{n}}$

$\hat{Y}_i = log(Y)$

$X_i' = \sqrt{x}$

$\hat{Y}_i = 4.896 + 4.325 X_i' \rightarrow exp(4.235) = 75.528$

For each 1 unit increase in X', the estimated increase in the geometric mean price is 75.53 times its previous value.

4.1.1 Box-Cox Transformations

There is a value λ that is the optimal transformation to the response for equal variance and normality. It is optimal in the sense that it finds the value of λ which produces the smallest SSE for Y_i.

$Y_i^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i \text{where} i \sim \text{iid } N(0, \sigma^2)$

λ	2	0.5	0	-0.5	-1
$Y' = Y^\lambda$	Y^2	$\sqrt{Y}$	log(Y)	$\frac{1}{\sqrt{Y}}$	$\frac{1}{Y}$

```
lindia::gg-boxcox(model)
```

4.2 Simultaneous Inference

Goal: Try to estimate more than one mean response at a time.

$(0.95)^3 = 0.857375$

4.2.1 Working-Hotelling Procedure

Based on the confidence band for the regression line.

100(1 - α)% simultanous confidence limits for g mean responses $E(Y_h)$

$Y_h \pm W\sqrt{MSE(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2})} \text{where } W^2 = 2F_{1-\alpha,2,n-2}$

```
qf(1 - $\alpha$, 2, n - 2)
```

4.2.2 Bonferonni Procedure

$Y_h \pm B\sqrt{MSE(\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\Sigma(X_i-\bar{X})^2})}$where $B = t_{1-\frac{\alpha}{2g},n-2}$

```
qt(1 - alpha / 2g, n - 2)
```

5 Session 5 - Prediction & Linear Algebra in Regression

5.1 Simultaneous Intervals

5.1.1 Confidence

Using the Bonferonni adjustment, The simultanous confidence interval for mean winning percentage for RunDiff of $X_h = -100, 0, 100$ has a confidence level $= 1 - \frac{\alpha}{g}$ where $\alpha = .05$ and $g = 3$

This is good for a smaller number of predictors. i.e. $g < 10$

5.1.2 Prediction

Bonferroni: $\hat{Y}_h \pm t_{1-\frac{\alpha}{2g},n-2}\sqrt{\text{MSE}(1 + \frac{1}{n} + \frac{(x_h-\bar{x})^2}{\Sigma(x_i-\bar{x})^2})}$ level $= 1 - \frac{\alpha}{g}$

Scheffe: $\hat{Y}_h \pm S\sqrt{\text{MSE}(1 + \frac{1}{n} + \frac{(x_h-\bar{x})^2}{\Sigma(x_i-\bar{x})^2})}$ where $S = \sqrt{gF_{1-\alpha,g,n-2}}$

Scheffe is more efficient with a larger **g** (i.e. g > 10). An in-class example showed that this was not the case so the jury is still out.

5.2 Inverse Prediction ("Calibration")

First, construct a model where $Y = X$

Goal: Make a prediction of X that was used to predict a new value of Y.

$$\begin{aligned} \hat{Y_i} &= \beta_0 + \beta_1 X_i + \epsilon_i \text{ where } \epsilon_i \sim \text{ iid } N(0, \sigma^2) \\ \hat{Y} &= b_0 + b_1 x \end{aligned} \tag{9}$$

We are given $Y_{h(new)}$, so what is $X_{h(new)}$?

$X_h(\hat{new}) = \frac{Y_{h(new)}-b_0}{b_1}$

$X_{h(\hat{new})} \pm t_{1-\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{b_1^2}(1 + \frac{1}{n} + \frac{(x_{h(new)}-\bar{x})^2}{\Sigma(x_i-\bar{x})^2})}$

```
investr::calibrate(model, Y, interval = "Wald")
```

The approximate confidence interval is appropriate if the following quantity is small (i.e. $< .1$):
$\frac{t^2_{1-\frac{\alpha}{2},n-2} MSE}{b_1^2 \Sigma(X_i-\bar{X})^2}$

5.3 Linear Algebra in Regression

5.3.1 Review

$\underset{(nX1)}{\vec{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix}$

$\underset{(1\times n)}{\vec{Y^T}} = \begin{bmatrix} Y_1 & ... & Y_n \end{bmatrix}$

Design Matrix

$\underset{(n\times 2)}{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ ... & ... \\ 1 & x_n \end{bmatrix}$

$\underset{(2\times n)}{x^T} = \begin{bmatrix} 1 & ... & 1 \\ x_1 & ... & x_n \end{bmatrix}$

1. Matrix Addition & Subtraction

$$\begin{aligned} Y_i =& E(Y_i) + \epsilon_i \\ \vec{Y} =& E(\vec{Y}) + \vec{\epsilon} \\ E(\vec{Y}) =& \begin{bmatrix} E(Y_1) \\ ... \\ E(Y_n) \end{bmatrix} \\ \vec{\epsilon} =& \begin{bmatrix} \epsilon_1 \\ ... \\ \epsilon_n \end{bmatrix} \end{aligned} \tag{10}$$

2. Matrix Multiplication

$$\underset{(1\times n)(n\times 1)}{\vec{Y}^T\vec{Y}} = \begin{bmatrix} Y_1 & ... & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ ... \\ Y_n \end{bmatrix} = \sum_1^n Y_i^2$$

$$\underset{(2\times n)(n\times 2)}{X^T X} = \begin{bmatrix} 1 & ... & 1 \\ x_1 & ... & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ ... & ... \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{bmatrix} \tag{11}$$

$$\underset{(2\times n)(n\times 1)}{X^T\vec{Y}} = \begin{bmatrix} 1 & ... & 1 \\ x_1 & ... & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ ... \\ Y_n \end{bmatrix} = \begin{bmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{bmatrix}$$

3. Special Matrices **Symmetric**: $A = A^T$ This implies a square matrix. i.e. n x n

Diagonal: $\underset{(n\times n)}{A} = \begin{bmatrix} a_{11} & 0 & ... & 0 \\ 0 & a_{22} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & a_{nn} \end{bmatrix}$

Identity Matrix: $\underset{(n\times n)}{I} = \begin{bmatrix} 1 & ... & 0 \\ ... & 1 & ... \\ 0 & ... & 1 \end{bmatrix}$

Scalar: $gI = \begin{bmatrix} g & ... & 0 \\ ... & g & ... \\ 0 & ... & g \end{bmatrix}$ where g is a scalar value

One vectors

$$\underset{(n\times 1)}{\vec{1}} = \begin{bmatrix} 1 \\ ... \\ 1 \end{bmatrix}$$

$$\underset{(n\times n)}{J} = \begin{bmatrix} 1 & ... & 1 \\ ... & 1 & ... \\ 1 & ... & 1 \end{bmatrix} \tag{12}$$

$$\underset{(1\times n)(n\times 1)}{\vec{1}^T\vec{1}} = n$$

$$\underset{(n\times 1)(1\times n)}{\vec{1}\vec{1}^T} = \begin{bmatrix} 1 \\ ... \\ 1 \end{bmatrix} \begin{bmatrix} 1 & ... & 1 \end{bmatrix} = J$$

4. Inverse of a Matrix

$$
\begin{aligned}
\underset{(2\times 2)}{A} &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\
\underset{(2\times 2)}{A^-1} &= \frac{1}{det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}
\end{aligned} \tag{13}
$$

Application to Regression

$$
\begin{aligned}
\underset{2\times 2}{(X^T X)^-1} &= \frac{1}{det(X^T X)} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} = ... = \begin{bmatrix} \frac{\Sigma x_i^2}{n\Sigma(x_i-\bar{x})^2} & -\frac{\Sigma x_i}{n\Sigma(x_i-\bar{x})^2} \\ -\frac{\Sigma x_i}{n\Sigma(x_i-\bar{x})^2} & \frac{n}{n\Sigma(x_i-\bar{x})^2} \end{bmatrix} \\
det(X^T X) &= n\Sigma x_i^2 - (\Sigma x_i)^2 \\
&= n\Sigma x_i^2 - \frac{n(\Sigma x_i)^2}{n^2} \\
&= n[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}] \\
&= n\Sigma(x_i - \bar{x})^2
\end{aligned} \tag{14}
$$

Side Note

$$
\begin{aligned}
\Sigma x_i &= n\bar{x} \\
\Sigma(x_i - \bar{x})^2 &= \Sigma x_i^2 - n\bar{x}^2 \\
\Sigma x_i^2 &= \Sigma(x_i - \bar{x})^2 + n\bar{x}^2 \\
(X^T X)^-1 &= \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i-\bar{x})^2} & -\frac{\bar{x}}{\Sigma(x_i-\bar{x})^2} \\ -\frac{\bar{x}}{\Sigma(x_i-\bar{x})^2} & \frac{1}{\Sigma(x_i-\bar{x})^2} \end{bmatrix}
\end{aligned} \tag{15}
$$

5. Matrix Rules

$$
\begin{aligned}
A+B &= B+A \\
(A+B)+C &= A+(B+C) \\
(AB)C &= A(BC) \\
C(A+B) &= CA+CB \\
\\
(A^T)^T &= A \\
(A+B)^T &= A^T+B^T \\
(AB)^T &= B^TA^T \\
\\
(AB)^{-1} &= B^{-1}A^{-1} \\
(A^-1)^{-1} &= A \\
(A^T)^{-1} &= (A^{-1})^T
\end{aligned}
\tag{16}
$$

5.3.2 Expectations

$$
\begin{aligned}
\underset{(n\times 1)}{\vec{Y}} &= \begin{bmatrix} Y_1 \\ ... \\ Y_n \end{bmatrix} \\
\underset{(n\times 1)}{E(\vec{Y})} &= \begin{bmatrix} E(Y_1) \\ ... \\ E(Y_n) \end{bmatrix} \\
\vec{\epsilon} &= \begin{bmatrix} \epsilon_1 \\ ... \\ \epsilon_n \end{bmatrix} \\
E(\vec{\epsilon}) &= \vec{0}
\end{aligned}
\tag{17}
$$

5.3.3 Variance-Covariance Matrix

$$
\sigma^2(\vec{Y}) = \begin{bmatrix} Var(Y_i) & ... & Cov(Y_1, Y_n) \\ ... & ... & ... \\ Cov(Y_n, Y_1) & ... & Var(Y_n) \end{bmatrix}
\tag{18}
$$

When Y_i independent, the off diagonals are 0 meaning $\sigma^2(\vec{Y}) = \sigma^2 I$
Aside

$$\begin{aligned} Var(Y) =& E[(Y - E(Y))^2] \\ \sigma^2(\vec{Y}) =& E[(\vec{Y} - E(\vec{Y}))(\vec{Y} - E(\vec{Y}))^T] \end{aligned} \tag{19}$$

Let $\underset{(p\times 1)}{\vec{W}} = \underset{(p\times n)(n\times 1)}{A\vec{Y}}$ where A is a matrix of **constants** and Y is a **random vector**

$$\begin{aligned} E(A) =& A \\ E(\vec{W}) =& AE(\vec{Y}) \\ \sigma^2(\vec{W}) =& E[(\vec{W} - E(\vec{W}))(\vec{W} - E(\vec{W}))^T] \\ =& E[(A\vec{Y} - AE(\vec{Y}))(A\vec{Y} - AE(\vec{Y}))^T] \\ =& E[A(\vec{Y} - E(\vec{Y}))(A(\vec{Y} - E(\vec{Y}))^T] \\ =& E[A(\vec{Y} - E(\vec{Y}))(\vec{Y} - E(\vec{Y}))^T A^T] \\ =& AE[(\vec{Y} - E(\vec{Y}))(\vec{Y} - E(\vec{Y}))^T]A^T \\ =& A\sigma^2(\vec{Y})A^T \end{aligned} \tag{20}$$

5.3.4 Multivariate Normal Distribution

$$\begin{aligned} \underset{(p\times 1)}{\vec{Y}} &= \begin{bmatrix} Y_1 \\ ... \\ Y_p \end{bmatrix} \\ \underset{(p\times 1)}{\vec{\mu}} &= \begin{bmatrix} \mu_1 \\ ... \\ \mu_p \end{bmatrix} \end{aligned} \tag{21}$$

$\underset{(p\times p)}{\Sigma}$ = Variance-Covariance Matrix

$f(\vec{Y}) = \frac{1}{(2\pi)^{\frac{P}{2}}\sqrt{det(\Sigma)}} exp(-\frac{1}{2}(\vec{Y} - \vec{\mu})^T \Sigma^{-1}(\vec{Y} - \vec{\mu}))$

If $Y_1, ..., Y_p$ are jointly normally distributed (i.e in the multivariate normal distr.), then $Y_k \sim N(\mu_k, \sigma_k^2)$ where $k = [1, p]$

Recall the Linear Regression equation $Y_i \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.

$\underset{(n\times 1)}{\vec{Y}} = \underset{(n\times 2)(2\times 1)}{X\vec{\beta}} + \vec{\epsilon}$ where $\underset{(n\times 1)}{\vec{\epsilon}} \sim N_n(\vec{0}, \sigma^2 I)$

N_n is a dimensions of a multivariate normal

$\vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

$E(\vec{Y}) = X\vec{\beta}$

5.3.5 Least Squares Estimation

Normal Equations from Week 2

$$\begin{aligned} nb_o + b_1\Sigma x_i &= \Sigma Y_i \\ b_0\Sigma X_i + b_1\Sigma X_i^2 &= \Sigma X_i Y_i \end{aligned} \tag{22}$$

$X^T X\vec{b} = X^T\vec{Y}$
So?
Least Squares Estimator: $\vec{b} = (X^T X)^{-1}X^T\vec{Y}$
$\underset{(n\times 1)}{\hat{\vec{Y}}} = \underset{(n\times 2)(2\times 1)}{X\vec{b}} = (X^T X)^{-1}X^T\vec{Y}$

1. Hat Matrix $H = (X^T X)^{-1}X^T$

   The Hat Matrix is important for computing diagnostics for the model such as Cook's Distance.

   Properties

   - symmetric ($H^T = H$)
   - Idempotent ($HH = H$)

2. Residuals $E_i = Y_i - \hat{Y}_i \to \vec{Y} - \hat{\vec{Y}} = \vec{Y} - X\vec{b} = \vec{Y} - H\vec{Y} = (I - H)\vec{Y}$

   $\sigma^2(\vec{e}) = \sigma^2(I - H)$

   This is estimated by: $MSE(I - H)$

6 Session 6 - Sums of Squares and Multiple Linear Regression

6.1 Sum of Squares

$$\underset{(1\times n)(n\ times 1)}{\vec{Y}^T\vec{Y}} = \Sigma Y_i^2 \tag{23}$$

Quadratic Form: Contains squares of observations **and** their cross products. These are known as second-degree polynomials.

Quadratic forms scaled by σ^2 allow us to treat the random variable Y as an observation of χ^2_{n-1} distribution.

This is unlike $\sigma^2(A\vec{Y}) = A\sigma^2(\vec{Y})A^T$ since that is squaring a matrix of **constants** whereas $\vec{Y}^T\vec{Y}$ squares a matrix of **random variables** i.e. Y

6.1.1 SSE

$$\begin{aligned} SSE =& \Sigma e_i^2 \\ =& \vec{e}^T \vec{e} \\ =& \vec{Y}^T (I - H) \vec{Y} \end{aligned} \tag{24}$$

6.1.2 SSTo

$$\begin{aligned} SSTo =& \Sigma (Y_i - \bar{Y})^2 \\ =& \Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n} \\ =& \vec{Y}^T (I - \frac{1}{n} J) \vec{Y} \end{aligned} \tag{25}$$

6.1.3 SSR

$$\begin{aligned} SSR =& \Sigma (\hat{Y}_i - \bar{Y})^2 \\ =& \vec{Y}^T (H - \frac{1}{n} J) \vec{Y} \end{aligned} \tag{26}$$

6.2 Mean Estimates σ^2

6.2.1 Mean Responses

$\hat{Y_h} = b_0 + b_1 X_h$

so? we would like

$$\underset{(1\times1)}{\hat{Y_h}} = \begin{bmatrix} 1 & X_h \end{bmatrix} \vec{b} \tag{27}$$

Let $\vec{X_h} = \begin{bmatrix} 1 \\ X_h \end{bmatrix}$

Then, $\hat{Y_h} = \vec{X_h}^T \vec{b}$

This is an estimate of the mean response!

6.3 Variance of $\hat{Y_h}$

$$\begin{aligned} \underset{(1\times1)}{Var(\hat{Y_h})} =& Var(\vec{X_h}^T \vec{b}) \\ =& \vec{X_h}^T Var(\vec{b}) \vec{X_h} \\ =& \vec{X_h}^T \sigma^2 (X^T X)^{-1} \vec{X_h} \\ =& \sigma^2 \underset{(1\times2)(2\times2)(2\times1)}{X_h^T (X^T X)^{-1} \vec{X_h}} \end{aligned} \tag{28}$$

6.4 Multiple Regression Models

$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i$ where $\epsilon_i \sim iidN(0, \sigma^2)$

$E(Y_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i,p-1}$

$Y_i \sim indepN(E(Y_i), \sigma^2)$.

The parameters of this model are $\{\beta_0, \ldots, \beta_{\mathrm{p}}\}$. Thus there are **p** regression coefficients.

6.4.1 Interpretation

Using the model, $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$

let's interpret the coefficients.

β_0: The mean response of Y when $X_1 = 0, X_2 = 0$

β_1: For a fixed value of X_2, the associated increase in mean response in Y is β_1 for every 1 unit increase in X_1. **This is known as a partial effect**

β_2: For a fixed value of X_1, the associated increase in mean response in Y is β_2 for every 1 unit increase in X_2.

β_k: Associated change in mean response of Y for every 1 unit increase in X_k, given all other predictors are held constant.

6.4.2 Aside: Multi-Collinearity

Multicollinearity occurs when two or more predictors are highly correlated.

- Standard Errors blow up which makes test statistic small, which makes p-values high. This affects the ability for us to make **inferences**
- Multicollinearity is acceptable when using models for **prediction** but not when using them for **inference**.

6.4.3 Matrix Notation

$$\begin{aligned} \underset{(n\times 1)}{\vec{Y}} &= \underset{(n\times p)(p\times 1)}{X\vec{\beta}} + \underset{(n\times 1)}{\vec{\epsilon}} \\ \underset{(n\times n)}{Var(\vec{\epsilon})} &= \sigma^2 I \end{aligned} \tag{29}$$

1. Fitted Values $\hat{Y}_i = b_0 + b_1 X_{i,1} + ... + b_{p-1} X_{i,p-1}$

   Residuals: $e_i = Y_i - \hat{Y}_i$

2. Least Squares Estimators

   $\underset{(p\times 1)}{\vec{b}} = \underset{(p\times n)(n\times p)}{(X^T X)^{-1}} \underset{(p\times n)(n\times 1)}{X^T \vec{Y}}$

6.4.4 ANOVA Table

Source	SS	DF	MS	F	p-value
Regression	SSR $= \Sigma(\hat{Y}_i - \bar{Y})^2$	p - 1	$MSR = \frac{SSR}{p-1}$	$F^* = \frac{MSR}{MSE}$	$P(F_{p-1,n-p} \geq F^*)$
Error	SSE $= \Sigma(Y_i - \hat{Y}_i)^2$	n - p	$MSE = \frac{SSE}{n-p}$		
Total	SSto $= \Sigma(Y_i - \bar{Y})^2$	n - 1			

6.4.5 Omnibus F-Test for Regression Relation

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = ... = \beta_p = 0 \\ H_A &: \text{ at least one } \beta_k \neq 0 \end{aligned} \tag{30}$$

Test statistic: $F^* = \frac{MSR}{MSE}$. If H_0 is true, $F^* \sim F_{p-1,n-p}$

6.4.6 Coefficient of Multiple Determination

$R^2 = 1 - \frac{SSE}{SSTo}$

The issue with R^2 is that it increases with the number of predictors **irrespective** of the predictor improving the model.

$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTo}{n-1}}$

6.4.7 Coefficient of Multiple Correlation

$R = \sqrt{R^2}$

6.4.8 Inferences in β_k

$$\begin{aligned} H_0 &: \beta_k = 0 \\ H_A &: \beta_k \neq 0 \end{aligned} \tag{31}$$

Test Statistic: $t^* = \frac{b_k}{SE_{bk}}$
If H_0 is true, then $t^* \sim t_{n-p}$
p-value $= 2P(t_{n-p} \geq |t|)$

```
2 * (1 - pt(abs(t.star), n - p))
```

$100(1 - \alpha)SE_{bk}$

7 Session 7 - Multiple Regression & Qualitative\/Quantitative Predictors

7.1 Multiple Regression

7.1.1 Extra Sums of Squares

Def: The marginal reduction in SSE when one or several predictors are added to the regression model, **given** other predictors are already in the model.

$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$
$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$

These are equivalent because any reduction in SSE implies an increase in SSR per the ANOVA definition: $SSTo = SSR + SSE$

1. Multiple Predictors

   $SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$
   $SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$

   $SSR(X_1, X_2) = SSR(X_2) + SSR(X_1|X_2)$

Source	SS	df	MSE
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2\|X_1$	$SSR(X_2\|X_1)$	1	$MSR(X_2\|X_3)$
$X_3\|X_1, X_2$	$SSR(X_3\|X_1, X_2)$	1	$MSR(X_3|X_1, X_2)$$
Error	$SSE(X_1, X_2, X_3)$	n - 4	$MSE(X_1, X_2, X_3)$
Total	SSTo	n - 1	

2. Hypothesis Test - $\beta_k = 0$ $H_0 : \mu_k = 0$
   $H_A : \mu_k \neq 0$

   This is the $\mu_k X_k$ dropped from the model.

   Test Statistic: $t^* = \frac{b_k}{SE_{bk}}$ $df = n - p$

   (a) Full model $Y_i = \mu_0 + \mu_1 X_1 + ... + \mu_{p-1} X_{i,p-1} + \epsilon_i$
   "p - 1" predictor variables
   $SSE(F) = SSE(X, ..., X_{p-1})$

   (b) Reduced Model $Y_i = \mu_0 + \mu_1 X_1 + ... + \mu_{p-2} X_{i,p-2} + \epsilon_i$
   "p - 2" predictor variables
   $SSE(R) = SSR(X, ..., X_{k-1}, X_{p-1})$

$$F^* = SSE(R) - SSE(F) = \frac{\frac{df_R - df_F}{SSE(F)}}{df_F} = \frac{\frac{SSE(X_1,...,X_{k-1},X_k,...,X_{p-1}) - SSE(X,...,X_{p-1})}{n-(p-1)-(n-p)}}{\frac{SSE(X,...,X_{p-1})}{n-p}}$$

3. Hypothesis Test - $\beta_0 = ... = \beta_k = 0$

   (a) Reduced Model $Y_i = \mu_0 + \mu_1 X_{i1} + ... + \mu_{k-1} X_{i,k-1} + \mu_k X_{ik} + ... + \mu_{p-1} X_{i,p-1} + \epsilon_i$
   "p - g - 1" predictors **OR** "p - g" regression coefficients
   $F^* = \frac{\frac{SSE(X_1,...,X_{k-1},X_k,...,X_{p-1}) - SSE(X,...,X_{p-1})}{n-(p-g)-(n-p)}}{\frac{SSE(X,...,X_{p-1})}{n-p}} = \frac{SSR(X_k,...,X_{k+(g-1)}|X_k,...,X_{k+g},...,X_{p-1})}{MSE(X_1,...,X_{p-1})}$
   If H_0 is true, $F^* \sim F_{g,n-p}$

4. R^2 R^2: Coefficient of multiple determination

   - proportion of variation in Y explained by the regression of Y on $X_1, ..., X_{p-1}$

   <u>Ex</u>

   $Y_i = \mu_0 + \mu_1 X_{i,1} + \mu_2 X_{i,2}$

   $SSE(X_2)$: variation when only X_2 is in the model.

   $SSE(X_1, X_2)$: variation when both X_1, X_2 are in the model.

   Marginal reduction in variation when X_1 is added to the model?
   $\frac{SSE(X_2) - SSE(X_1,X_2)}{SSE(X_2)}$

   $R^2_{Y_1/Y_2} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$

   $R^2_{Y_2/Y_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$

   3 predictors

   $R^2_{Y3|2,1} = \frac{SSR(X_3|X_1,X_2)}{SSE(X_1,X_2)}$

   Recipe for correlation coefficient:

   (a) Take sqrt of partial R^2

   (b) Sign of partial correlation = sign of correlation corresponding coefficient

7.2 Multi-collinearity

Predictors that are highly correlated with each other. **10 N values per predictor**

7.2.1 Effects

1. There is no unique sum of squares that can be assigned to the predictor variable

2. May inflate standard error of b_k least square error.

It does not greatly impact the value of predictions.
ETA^2 tells R^2 given the previously given variable R^2

7.3 Polynomial Regression Models

- true curvilinear response

- true curvilinear response is unknown but a polynomial function provides a good approximation to the true function.

One prediction variable **and** second order:
$Y_i = \mu_0 + \mu_1 X + \mu_2 X^2 + \epsilon_i$ where $X_i = x_i - \bar{x}$
$E(Y) = \mu_0 + \mu_1 X_1 + \mu_2 X^2$
Two parameters **and** second order:
$x_{i,1} = x_{i,1} - \bar{x_1}$
$X_{i,2} = x_{i,2} - \bar{x_2}$
$Y_i = \beta_0 \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1}^2 + \beta_4 X_{i,2}^2 + \beta_5 X_{i,1} X_{i,2} + \epsilon_i$
<u>Strategy?</u> Fit higher order models and compare to reduced models.

```
summary(model)
```

$\hat{Y} = b_0 + b_1 x + b_2 x^2$ $\hat{Y} = b_0' + b_1' x + b_2' x^2$
$b_0' = b_0 - b_1 \bar{x} - b_2 \bar{x}^2$
$b_1' = b_1 - 2 b_2 \bar{x}$
$b_2' = b_2$
Why do this? Solving a regression model with a non-linear $E(Y_i)$

8 Session 8 - Interaction Models & Model Selection

8.1 Interaction Regression Models

- p: # of regression coefficients. i.e. parameters

- p - 1: predictor variables

8.1.1 Additive Effects

$$E(Y) = \sum_{i=1}^{p-1} f_K(x_k) \tag{32}$$

but $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2\ \beta_3 X_1 X_2$ is **not** additive since $X_1 X_2$ is an interaction term

Consider the following:

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2\ \beta_3 X_1 X_2 + \epsilon_i$ where $\epsilon_i \sim iidN(0, \sigma^2)$

A one-unit **increase** in X_2 for a fixed value of X_1, results in an associated change of $\beta_1 + \beta_3 X_1$ units in mean response Y.

- $\beta_3 = 0 =>$ additive model
- $\beta_3 > 0 =>$ reinforcement or synergistic interaction*
- $\beta_3 < 0 =>$ interference or antagonistic interaction*

*if β_1 and β_2 are negative, these terms flip

parallel lines indicate **additive** terms, otherwise <u>interactive</u>

Aside

> To avoid multicollinearity between predictors, center variables!
> $X_{ik} = X_{ik} - \bar{X}_k$
>
> Does Standardizing also help reduce multicollinearity? Yes, but makes interpretation more difficult. This is done in PCA and as I've seen, interpreting PCA can be hairy or a best guess.

- Try to identify possible interactions ahead of time prior to fitting the model.
- When looking at removing **one** term, the t statistic is sufficient to rule out a parameter.

8.1.2 Qualitative Predictors

Qualitative Predictor with two classes. i.e. two values This is sometimes called: Indicators, Binary, dummy variables

For representing C classes, use $C - 1$ indicator variables.

<u>Example</u>

Let C = 4 $C = \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix}$

$$
\begin{aligned}
Y_i =& \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ \beta_3 X_3 + \beta_4 X_4 + \epsilon_i \text{ where} \\
X_2 =& \begin{cases} 1, & A \\ 0, & else \end{cases} \\
X_3 =& \begin{cases} 1, & B \\ 0, & else \end{cases} \\
X_4 =& \begin{cases} 1, & C \\ 0, & else \end{cases}
\end{aligned} \tag{33}
$$

if $X_2 = X_3 = X_4 = 0$, indicates effect of $C = D$ on mean Y_i

$$
\begin{aligned}
A : E(Y) =& (\beta_0 + \beta_2) + \beta_1 X_i \\
B : E(Y) =& (\beta_0 + \beta_3) + \beta_1 X_i \\
C : E(Y) =& (\beta_0 + \beta_4) + \beta_1 X_i \\
D : E(Y) =& (\beta_0) + \beta_1 X_i
\end{aligned} \tag{34}
$$

D is considered the **baseline category**

1 Qualitative variable and 1 Quantitative variable in the same model is known as **ancova**: Analysis of Covariance. ANCOVA assumes that each group has the same slope.

Interpret β_0: The diff in mean response of Y between A and D group for a given value of X_1

Estimate $\beta_3 - \beta_4$

1. $b_3 - b_4$
2. $Var(b_3 - b_4) = var(b_3) + var(b_4) - 2cov(b_3, b_4)$

If doing time series, one can use indicator variables to model time periods

8.2 Model and Variable Selection

8.2.1 Criterion for Model Selection

p: # of parameters (regression coefficients)

1. R_p^2 or SSE_p criterion. Both indicate the same thing.
   $R_p^2 = 1 - \frac{SSE_p}{SSTo}$
   Look for

- *High* R_p^2
- *Small* SSE_p

2. $R_{a,p}^2$ or MSE_p criterion

   $R_p^2 = 1 - \frac{\frac{SSE_p}{n-p}}{\frac{SSTo}{n-1}} = 1 - \frac{MSE_p}{\frac{SSTo}{n-1}}$

   Look For:

   - *High* $R_{a,p}^2$
   - *Small* MSE_p

3. Mallows' C_p Criterion

   $C_p = \frac{SSE_p}{MSE(X_1,...,X_{p-1})} - (n-2p)$

   $MSE(X_1, ..., X_{p-1})$: MSE for the model with **all** potential predictors of interest.

   For largest possible value of P, $C_p = p$

   proof

$$\begin{aligned} MSE =& \frac{SSE}{n-p} \\ \frac{SSE_p}{\frac{SSE_p}{n-p}} =& n-p-(n-2p) = p \end{aligned} \tag{35}$$

Look for

- *Small* C_p or $C_P \leq p$. This means the model has a small amount of bias.

Recall $MSE(Y) = Bias^2(Y) + Var(Y)$

1. AIC_p or SBC_p Criterion

   - AIC_p: Akaike's Information Criterion - $nln(SSE_p) - nln(n) + 2p$
   - SBC_p: Schwartz' Bayesian Information Criterion - $nln(SSE_p) - nln(n) + pln(n)$

   Look for

   - *Small* SSE_p
   - *Small* AIC_p and/or SBC_p

2. $PRESS_p$ Criterion

   Prediction Sum of Squares

   $PRESS_P = \sum_1^n (Y_i - \hat{Y_{i(i)}})^2$

   $\hat{Y_{i(i)}}$

   (a) Ignore the ith case
   (b) Fit model on remaining $n-1$ cases
   (c) Find Fitted value based on deleted ith case

   This is **not** the same as bootstrapping, mostly because there is no resampling oging on.

```
leaps::regsubsets(formula, data, method="exhaustive", nbest=30)

#+NAME: fortify_leaps
fortify.regsubsets <- function(model, data, ...){
  require(plyr)
  stopifnot(model$intercept)
  models <- summary(model)$which
  rownames(models) <- NULL
  model_stats <- as.data.frame(summary(model)[c("bic","cp","rss","rsq","adjr2")
  dfs <- lapply(coef(model, 1:nrow(models)), function(x) as.data.frame(t(x)))
  model_coefs <- plyr::rbind.fill(dfs)
  model_coefs[is.na(model_coefs)] <- 0
  model_stats <- cbind(model_stats, model_coefs)
  # terms_short <- abbreviate(colnames(models))
  terms_short <- colnames(models)
  model_stats$model_words <- aaply(models, 1, function(row) paste(terms_short[r
  model_stats$size <- rowSums(summary(model)$which)
  model_stats
}

get_model_coefs <- function(model){
  models <- summary(model)$which
  dfs <- lapply(coef(model, 1:nrow(models)), function(x) as.data.frame(t(x)))
  model_coefs <- plyr::rbind.fill(dfs)
  model_coefs[is.na(model_coefs)] <- 0
  model_coefs
}
```

9 Session 9 - Model and Variable Selection & Assessing Diagnostics

9.1 Model and Variable Selection

9.1.1 Automatic Search Procedures

1. Backward Selection

   Full Model -> reduce parameters to "smallest" AIC

   ```
   step(model.full, direction = "backward")
   ```

2. Forward Selection

   Intercept-only model -> add parameters to "smallest" AIC

   ```
   step(lm.null, scope = list(lower, upper), direction = "forward")
   ```

3. Step-wise

   Intercept-only model -> add one -> subtract\/add one for "smallest" AIC

   ```
   step(lm.null, scope = list(upper), direction = "both")
   ```

9.1.2 Model Validation

1. Collect new data t ocheck the model and it's predictive validity

   $MSPR = \frac{\sum_1^{n^*}(Y_i - \hat{Y}_i)^2}{n^*}$

   If MSPR approximately your Model's MSE, then your model is not necessarily biased. If the difference is large, MSPR is a good indicator on how well it predicts.

   Defintions

   - MSPR: Mean Square Prediction Error
   - Y_i: Value of the response variable in the ith validation case
   - $\hat{Y}_i$: Predicted value of the ith validation case using the model you previously built.
   - n^*: number of cases in the validation dataset.

2. Compare results with theoretical expectations empirical results, and simulation results.

3. Use a holdout sample to check the model and its predictive ability. This is standard practice for predictive models

9.2 Assessing Diagnostics

9.2.1 Added-variable Plots

Also known as:

- Partial Regression Plots
- Adjusted Variable Plots

These plots show:

- Marginal Importance of this variable in reducing residual variability
- May provide info about the nature of the marginal regression relation for predictor variable X_k under consideration

1. Example $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

   Goal: What is X_i's effect given that X_2 is in the model?

$$\begin{aligned} \hat{Y_i}(X_2) &= b_0 + b_2 X_{i2} \\ e_i(Y|X_2) &= Y_i - \hat{Y_i}(X_2) \end{aligned} \tag{36}$$

   fitted values + residuals from the model with only X_2

$$\begin{aligned} \hat{X_{i1}}(X_2) &= b_0^* + b_2^* X_{i2} \\ e_i(X_1|X_2) &= X_{i1} - \hat{X_i}1(X_2) \end{aligned} \tag{37}$$

   fitted value + residuals from the model with X_1 as the response and X_2 as the predictor.

2. Reading Plots

```
car::avPlots(model)
```

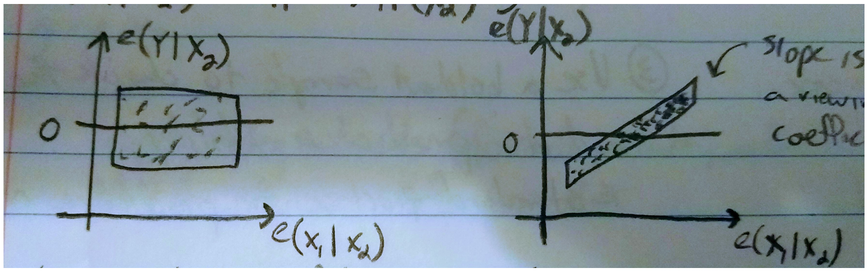

Figure 7: Partial Residuals vs Fitted Values

(a) Partial Residuals X_1 vs Fitted Values
Notice the even distribution of residuals around y = 0. X_1 provides no useful information given X_2 is in the model.

(b) Partial Residuals X_2 vs Fitted Values
Notice the pattern. X_1 may be a good addition to the model given X_2 is already in the model.

Goal: Identify outlying Y observations. i.e. which Y observations are influential on our own regression model?

- Residuals: $e_i = Y_i - \hat{Y}_i$
- Semi-studentized Residuals: $e^* = \frac{e_i}{\sqrt{MSE}}$
- Studentized Residuals: $R_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$

  h_{ii}: the ith diagonal value from the hat matrix H

```
rstandard(model)
```

10 Session 10 - Outliers & Weighted Least Squares

10.1 Outliers

10.1.1 Identifying Outlying Y Observations

- Use Studentized Deleted Residuals to identify *outlying Y Observations*

Residuals: $e_i = Y_i - \hat{Y}_i$ **Semi-studentized Residuals**: $e_i^* = \frac{e_i}{\sqrt{MSE}}$ **Studentized Residuals**: $r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$

h_{ii}: Standard Error of e_i. aka Standard Error of the ith residual

Deleted Residuals: $d_i = Y_i - \hat{Y_{i(i)}} = \frac{e_i}{1-h_{ii}}$
Studentized Deleted Residuals (rstudent):

$$\begin{aligned} t_i =& \frac{d_i}{SE_{d_i}} \\ =& \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \\ =& e_i\sqrt{\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}} \end{aligned} \tag{38}$$

10.1.2 What is an Outling Y Observation?

$|t_i| > t_{1-\frac{\alpha}{2n}, n-p-1}$

```
qt(1 - alpha / 2n, n - p - 1)
```

- The "- 1" is the residual that is being deleted

10.1.3 Identifying Outlying X Observations

- Use leverage values. i.e. "hat matrix leverage values"

h_{ii}: leverage (in terms of X values)

1. $0 \le h_{ii} \le 1, i = [1, n]$
2. $\sum_1^n h_{ii} = p$ (number of parameters in the model)

   Recall: $Var(e_i) = MSE(1 - h_{ii})$

- The larger h_{ii}, $Var(e_i)$ **decreases**, thus making close to Y$_i$$

  How large is a large h_{ii}?

- if $h_{ii} > s\bar{h} = \frac{2p}{n}$, the cases are outlying cases in terms of X.

10.2 Influential Cases

How influential are "new" cases?
$h_{new,new} = X_{new,new}^T(X^TX)^{-1}X_{new,new}$
If $h_{new,new}$ is much larger than h_{ii}, there may be some extrapolation. There are no set guidelines for this.

10.2.1 Identifying Influential Cases

1. Influence of the ith case on a single fitted value, .

    - Use DFFITS (Difference of Fits)

$$
\begin{aligned}
DFFITS_i =& \frac{\hat{Y}_i - Y_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} \\
=& e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}} \sqrt{\frac{h_{ii}}{1-h_{ii}}} \\
=& t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}}
\end{aligned}
\tag{39}
$$

    Notes

    - $MSE_{(i)}$: calculated with the ith case removed
    - t_i: Studentized Deleted Residuals

    What is influential?

    - Small - Med Dataset: $|DFFFITS_i| > 1$
    - Large Dataset: $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$

2. Influence of the ith case on all fitted values

    - Cooks Distance

$$
\begin{aligned}
D_i =& \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y_{j(i)}})^2}{pMSE} \\
=& \frac{e_i^2}{pMSE} [\frac{h_{ii}}{(1-h_{ii})^2}]
\end{aligned}
\tag{40}
$$

    Notes

    - $Y_{j(i)}$: fitted value when the ith case is left out

    What is an influential case? Compare D_i to $F_{p,n-p}$

    - If $P(F_{p,n-p} \leq D_i) < 0.1, 0.2$, the ith case has very little influence.

- If $P(F_{p,n-p} \leq D_i) > 0.5$, the ith case has major influence.

3. Influence of the ith case on the regression coefficients

   - DFBETAS
     $(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} C_{kk}}}$ **Notes**:
     - C_{kk}: Diagonal term of $(X^T X)^{-1}$
     - $Var(\vec{b} = \sigma^2 (X^T X)^{-1} = \sigma^2 C_{kk}$

10.3 Variance Inflation Factors

- used to assess Multicollinearity
  $VIF = \frac{1}{1-R_k^2}$
  Notes
  - R_k^2 is R^2 from ‘lm(Xk ~ X1 + ... + X(k-1) + X(k + 1) + ... + X(p - 1))‘
    * This is a mishmash of math and R

  min $\mathrm{VIF_k} = 1$ max $\mathrm{VIF} = \infty$

  - Sometimes (rarely) signs flip
  - multicollinearity causes increase variance

Interpretation

- VIF > 4, mild/moderate multicollinearity
- VIF > 10, severe multicollinearity
- Ideal? $V\bar{I}F$ close to 1

If experiencing high multicollinearity, check for correlation between response and each predictor.

10.4 Weighted Least Squares

- Good use if Variance is Unequal
  Possible Weight: $W_i = \frac{1}{\sigma_i^2}$

10.4.1 Iteratively Reweighted Least Squares

1. Fit regular least squares model and analyze results
2. Estimate the variance function or the standard deviation function by regressing e_i^2 or $|e_i|$ on the predictors.
3. Use the fitted values from the estimated $Var(\hat{V_i})$ or estimate std. dev $(\hat{S_i})$ function to obtain weights w_i.
4. Estimate regression coefficients use the weights. So?
    - e_i^2 estimates σ_i^2
    - $|e_i|$ estimates σ_i

$W_i = \frac{1}{(\hat{S_i})^2}$ using $|e_i|$
OR
$W_i = \frac{1}{\hat{V_i}}$ using e_i^2

11 Extra Curricular - Weighted Least Squares, Ridge, and Robust Regression

11.1 Weighted Least Squares

- Useful for models with heteroskedasticity (non-constant variance)

$$\begin{aligned}
\vec{b} &=(X^TX)^{-1}X^TY \\
\underset{(n\times n)}{\vec{b_w}} &=(X^TWX)^{-1}X^TWY \\
W &= \begin{bmatrix} w_1 & 0 & ... & 0 \\ 0 & w_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & w_n \end{bmatrix}
\end{aligned} \tag{41}$$

- OLS is a special case of WLS where W = J = 1.
- $w_i = k(\frac{1}{\sigma_i^2})$. if error variances known (rare)
- $w_i = \frac{1}{(\hat{s_i})^2}$. if using fitted standard error

- $w_i = \frac{1}{(\hat{v_i})}$. if using fitted variance

Using the weights to estimate regression coefficient is called *Iteratively Reweighted Least Squares*. Typically done until coefficients have stablized.

Notes

- R^2 does not have a clearcut meaning for WLS.

11.2 OLS with Heteroskedasticity

OLS can still be used with unequal error variances via White's Estimator. This leverages something called the *Robust Covariance Matrix*.

$$\begin{aligned}
\sigma^2(b) =& (X^TX)^{-1}(X^T\sigma^2(e)X)(X^TX)^{-1} \\
S^2(b) =& (X^TX)^{-1}(X^TS_0X)(X^TX)^{-1} \\
\underset{(n\times n)}{S_0} =& \begin{bmatrix} e_1^2 & 0 & ... & 0 \\ 0 & e_2^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & ... & e_n^2 \end{bmatrix}
\end{aligned} \tag{42}$$

e_i: OLS estimator of the residuals squared.

11.3 Ridge Regression

- Useful for cases with severe Multicollinearity.

  What to do when you have multicollinearity?

  1. If only estimating and no conf intervals, nothing
  2. Center predictor variables
  3. Drop Predictors
     - Downside: some predictors not acccounted for and there is a relationship affecting the response that is not being represented in the model.
  4. Add cases that break multicollinearity.
  5. PCA

Definition: Modifies OLS to allow biased estimators to lower variance.
Recall $MSE = Var(Y) + (Bias(Y))^2$
$E(b^R - \beta)^2 = \sigma^2(b^R) + (E(b^R) - \beta)^2$ where b^R: biased estimator

Least Squares Normal Equations given by: $r_{XX}b = r_{XY}$ r_{XX}: correlation matrix of X variables r_{XY}: Vector of coefficients of simple correlation variables between Y and each X Variable

Ridge Standardized Regression: $(r_{XX} + eI)b^R = r_{XY}$

\$ $\underset{(p-1)\times 1}{b^R} = \begin{bmatrix} {b_1}^{\mathrm{R}} \\ \dots \\ {b_{\mathrm{p\ -\ 1}}}^{\mathrm{R}} \end{bmatrix}$

$b^R = (r_{XX} + cI)^{-1} r_{XY}$

A **biasing** constant $c \geq 0$ can be chosen.

- bias increases as c increases. likewise variance decreases
- There is always some value c where b^R has a smaller MSE than OLS b.
  - Optimal Values c varies by application and is unknown

Ridge Trace: Method often used to determine c. This is combined with VIF

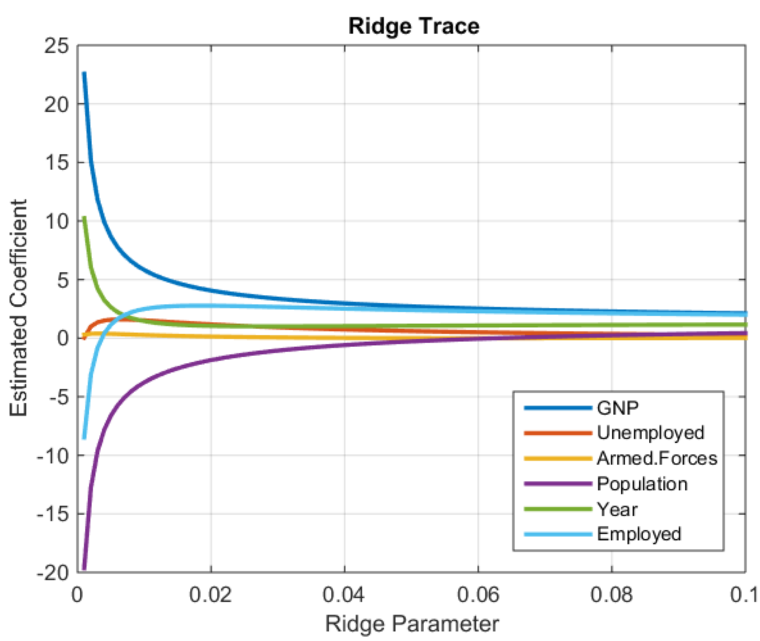

Figure 8: Ridge Trace Example

look for

- spots where the line smooths out
- where least change in b_k^r happens

finding c is a bit of an art.

this formula can be used to convert standardized coefficients to unstandardized coefficients.

$b_k = (\frac{s_y}{s_k})b_k^r$

s_y: standard dev of y s_k: standard error of b_k

11.4 robust regression

- reduce influential cases

uses **iteratively reweighted least squares** where w_i dampens influential cases instead of heteroskedasticity.

u: Scaled residual 0.345|4.685: tuning constants that are robust for 95% of normal data. Huber:

$$w = \begin{cases} 1 & |u| \leq 1.345 \\ \frac{1.345}{|u|} & |u| > 1.345 \end{cases} \tag{43}$$

Bisquare:

$$w = \begin{cases} [1 - (\frac{u}{4.685})^2]^2 & |u| \leq 4.685 \\ 0 & |u| > 4.685 \end{cases} \tag{44}$$

Huber is often used to obtain starting weights for Bisquare.

11.4.1 u

- Semi-studentized residuals could be used but they are not resistant to outliers
- Mean Absolute Deviation (MAD) often used.

  $MAD = \frac{1}{0.6745} med(|e_i - med(e_i)|)$

  $u_i = \frac{e_i}{MAD}$

  0.6745 is used to make this an unbiased estimate for σ from a normal distribution.

11.5 Regression Tree (Non-parametric Method)

- Split X's into distinct regions r and run a regression on each region.
- "Growing a tree" is finding the number of regions r and the boundaries/split points between them.
- If the variance of the residuals in each region seem constant, splitting may not be necessary.
- The best split point minimizes $SSE = \sum_{k=1}^{r} SSE(R_{rk})$
- Once the optimal r is chosen, each r is subdivided to find the most optimal SSE.
- The chosen number of regions is done through validation studies, such as choosing the tree that minimizes MSPR