Digital Assistant

Ken Ye

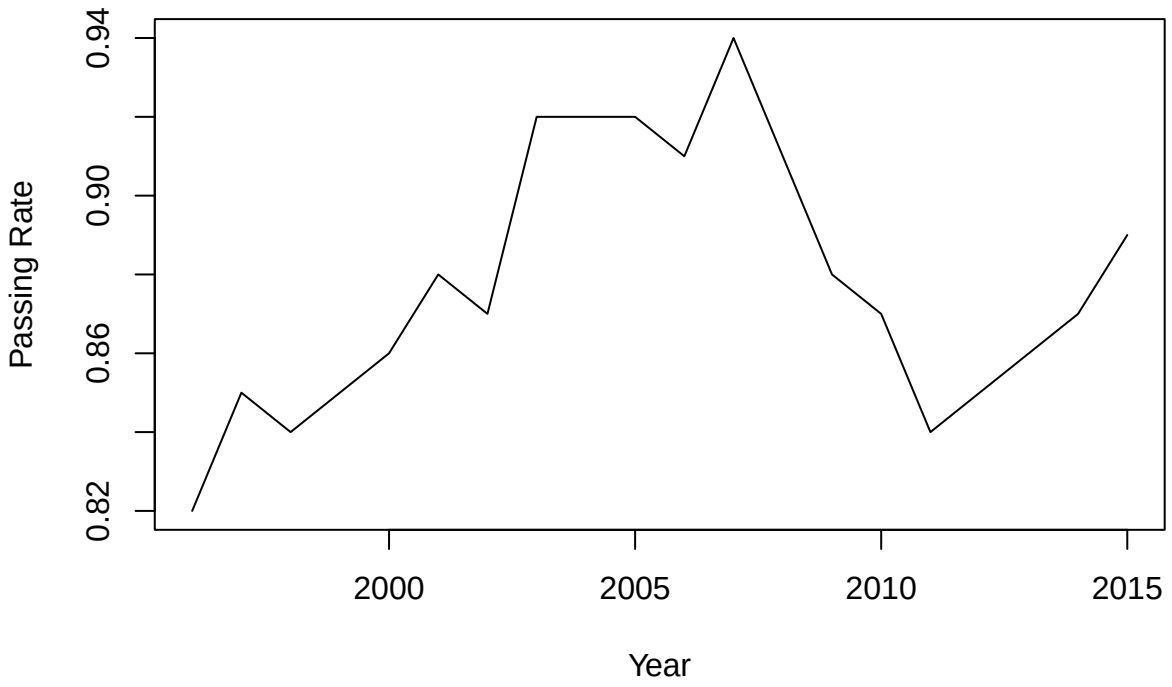
2023-10-18

Step 1: Load Data

```
# Load the dataset into a data frame
df <- read.table("data.txt", header = TRUE)
```

Step 2: Plot and Summarize the Data

```
# Plot the data
plot(df$Year, df$Pct, type = "l", xlab = "Year", ylab = "Passing Rate")
```




```
# Summarize the data
summary(df$Pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8200  0.8500  0.8700  0.8775  0.9100  0.9400
```

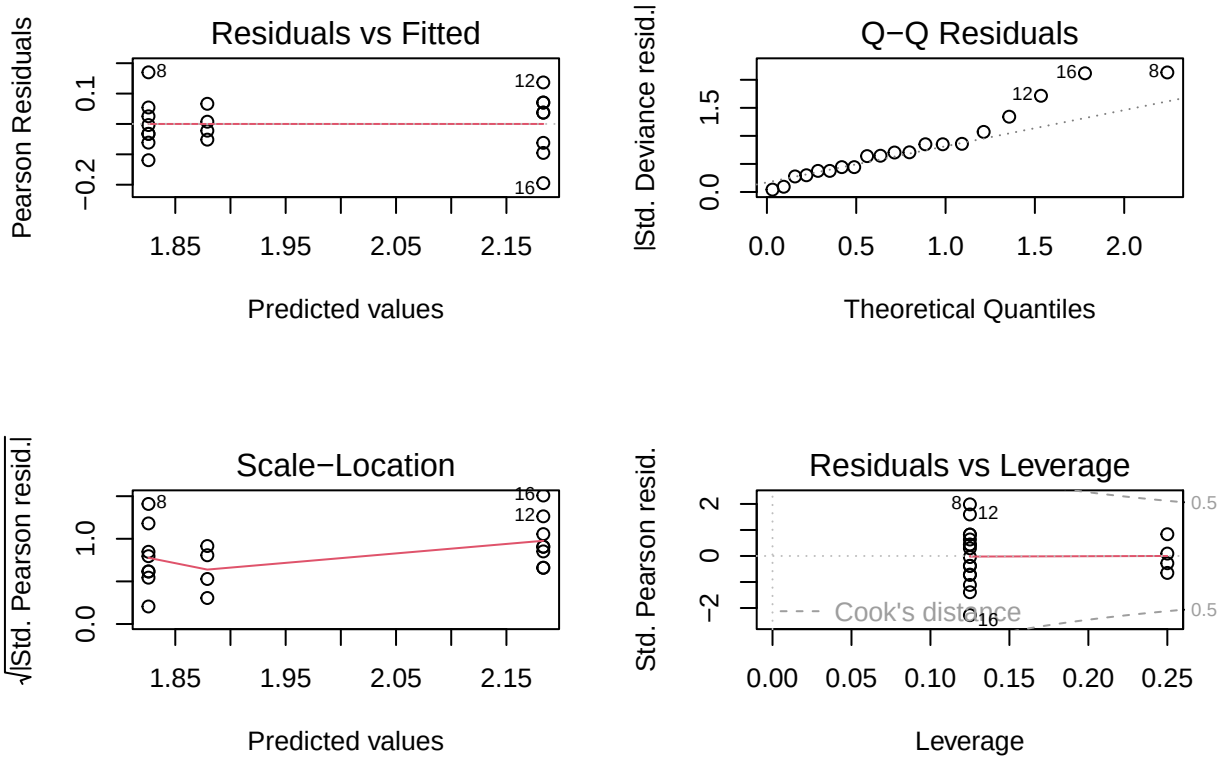
Step 3: Generate Derived Variables

```r
# Create a new column that classifies each year into one of three periods
df$Period <- ifelse(df$Year <= 2003, "tp1",
                    ifelse(df$Year > 2011, "tp3", "tp2"))

# View the data frame
df
```

```
##    Year    N  Pct Period
## 1  1996 6964 0.82    tp1
## 2  1997 7173 0.85    tp1
## 3  1998 7348 0.84    tp1
## 4  1999 7311 0.85    tp1
## 5  2000 7048 0.86    tp1
## 6  2001 6802 0.88    tp1
## 7  2002 7074 0.87    tp1
## 8  2003 6751 0.92    tp1
## 9  2004 7056 0.92    tp2
## 10 2005 7051 0.92    tp2
## 11 2006 7006 0.91    tp2
## 12 2007 7090 0.94    tp2
## 13 2008 7194 0.91    tp2
## 14 2009 7226 0.88    tp2
## 15 2010 7335 0.87    tp2
## 16 2011 7337 0.84    tp2
## 17 2012 7303 0.85    tp3
## 18 2013 7482 0.86    tp3
## 19 2014 7601 0.87    tp3
## 20 2015 7839 0.89    tp3
```

Step 4: Quasi-binomial

```r
# Fit a quasi-binomial model with the Period variable
model <- glm(Pct ~ Period, data = df, family = quasibinomial)

# Implement diagnostics
par(mfrow = c(2, 2))
plot(model)
```




```
# Display tables of parameter summaries
summary(model)
```

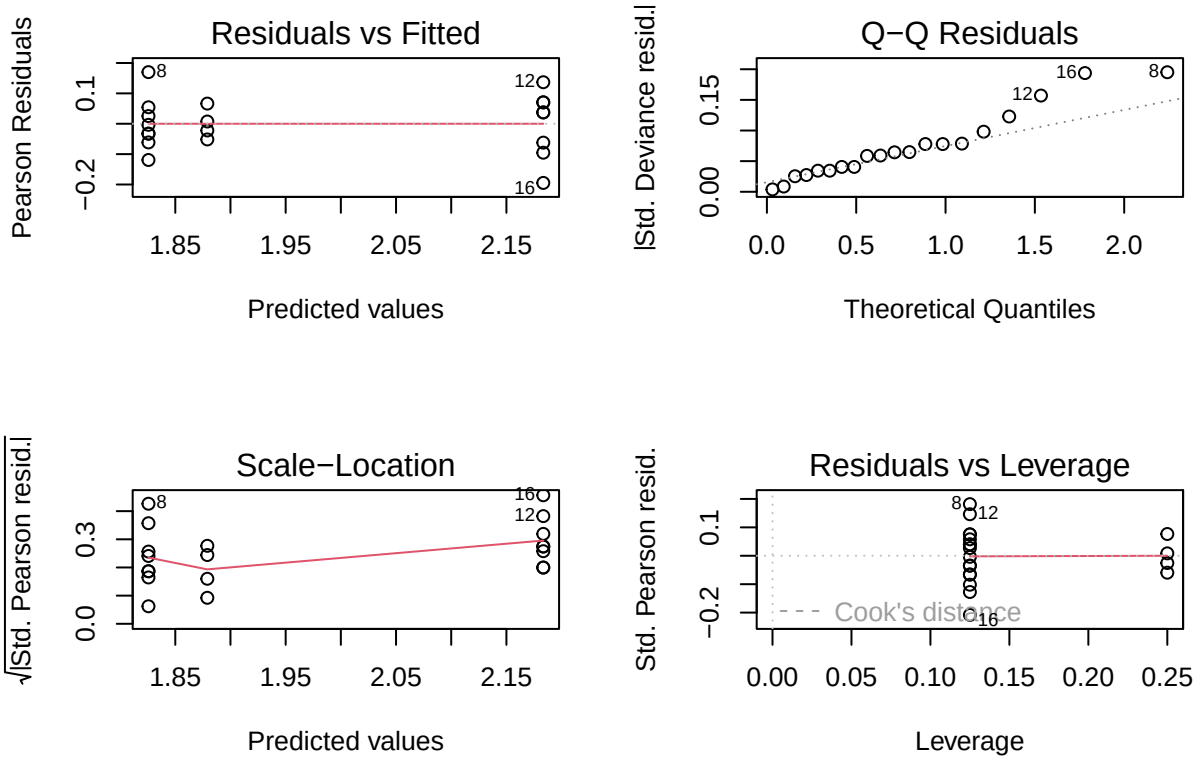
```
##
## Call:
## glm(formula = Pct ~ Period, family = quasibinomial, data = df)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.82571    0.09364  19.497 4.54e-13 ***
## Periodtp2    0.35770    0.14242   2.512   0.0224 *
## Periodtp3    0.05332    0.16432   0.324   0.7495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.008382931)
##
##     Null deviance: 0.20174  on 19  degrees of freedom
## Residual deviance: 0.14370  on 17  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Step 5: GLM

```r
# Fit a GLM with the Period variable
model <- glm(Pct ~ Period, data = df, family = binomial)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
# Implement diagnostics
par(mfrow = c(2, 2))
plot(model)
```



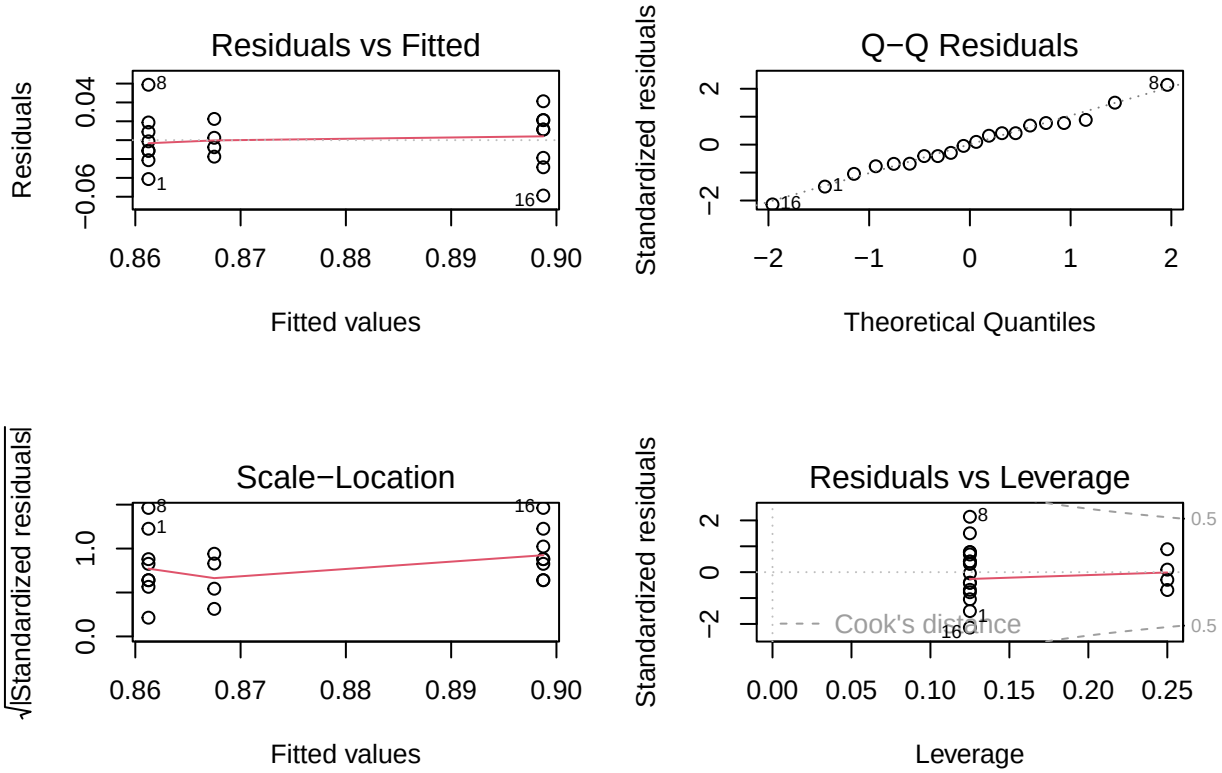

```r
# Display tables of parameter summaries
summary(model)
```

```
##
## Call:
## glm(formula = Pct ~ Period, family = binomial, data = df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.82571    1.02276   1.785   0.0742 .
## Periodtp2    0.35770    1.55553   0.230   0.8181
## Periodtp3    0.05332    1.79472   0.030   0.9763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.20174  on 19  degrees of freedom
```

```
## Residual deviance: 0.14370  on 17  degrees of freedom
## AIC: 11.235
##
## Number of Fisher Scoring iterations: 5
```

Step 6: LM

```r
# Fit a linear regression model with the Period variable
model <- lm(Pct ~ Period, data = df)

# Implement diagnostics
par(mfrow = c(2, 2))
plot(model)
```




```r
# Display tables of parameter summaries
summary(model)
```

```
##
## Call:
## lm(formula = Pct ~ Period, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.058750 -0.017813  0.000625  0.019375  0.058750
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.86125    0.01038  82.981   <2e-16 ***
## Periodtp2    0.03750    0.01468   2.555   0.0205 *
## Periodtp3    0.00625    0.01798   0.348   0.7324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02936 on 17 degrees of freedom
## Multiple R-squared:  0.2948, Adjusted R-squared:  0.2119
## F-statistic: 3.554 on 2 and 17 DF,  p-value: 0.05135
```

Step 7: Hypothesis Tests

```r
# Fit the three models
model1 <- lm(Pct ~ 1, data = df)
model2 <- lm(Pct ~ Period, data = df)
model3 <- lm(Pct ~ Year + Period, data = df)

# Test the three hypotheses using F tests
anova(model1, model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: Pct ~ 1
## Model 2: Pct ~ Period
## Model 3: Pct ~ Year + Period
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1     19 0.020775
## 2     17 0.014650  2 0.0061250 3.3556 0.06068 .
## 3     16 0.014602  1 0.0000475 0.0520 0.82248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Test the stepwise constant hypothesis
model_sc <- lm(Pct ~ Year + Period + Year:Period, data = df)
summary(model_sc)
```

```
##
## Call:
## lm(formula = Pct ~ Year + Period + Year:Period, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.020000 -0.005893 -0.002321  0.002357  0.035714
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -21.276071   4.884253  -4.356 0.000658 ***
## Year             0.011071   0.002443   4.532 0.000469 ***
## Periodtp2       44.400714   6.921209   6.415 1.61e-05 ***
```

```
## Periodtp3       -4.031929  15.068561  -0.268 0.792930
## Year:Periodtp2  -0.022143   0.003455  -6.410 1.63e-05 ***
## Year:Periodtp3   0.001929   0.007489   0.258 0.800531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01583 on 14 degrees of freedom
## Multiple R-squared:  0.8311, Adjusted R-squared:  0.7708
## F-statistic: 13.78 on 5 and 14 DF,  p-value: 5.49e-05
```

```r
# Test the piecewise linear hypothesis
model_pl <- lm(Pct ~ Year + Period + I((Year - 2003)*(Year > 2003)) + I((Year - 2011)*(Year > 2011)), da
summary(model_pl)
```

```
##
## Call:
## lm(formula = Pct ~ Year + Period + I((Year - 2003) * (Year >
##     2003)) + I((Year - 2011) * (Year > 2011)), data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.020000 -0.005893 -0.002321  0.002357  0.035714
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -21.276071   4.884253  -4.356 0.000658 ***
## Year                              0.011071   0.002443   4.532 0.000469 ***
## Periodtp2                         0.048571   0.016018   3.032 0.008959 **
## Periodtp3                         0.023571   0.029364   0.803 0.435541
## I((Year - 2003) * (Year > 2003))  -0.022143   0.003455  -6.410 1.63e-05 ***
## I((Year - 2011) * (Year > 2011))   0.024071   0.007489   3.214 0.006243 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01583 on 14 degrees of freedom
## Multiple R-squared:  0.8311, Adjusted R-squared:  0.7708
## F-statistic: 13.78 on 5 and 14 DF,  p-value: 5.49e-05
```