Predicting Diabetes: Key Physical and Socioeconomic Health Factors

Phoenix Williams

Denver

Research Question & Rationale

What physical and socioeconomic health factors have the strongest ability to predict diabetes risk?

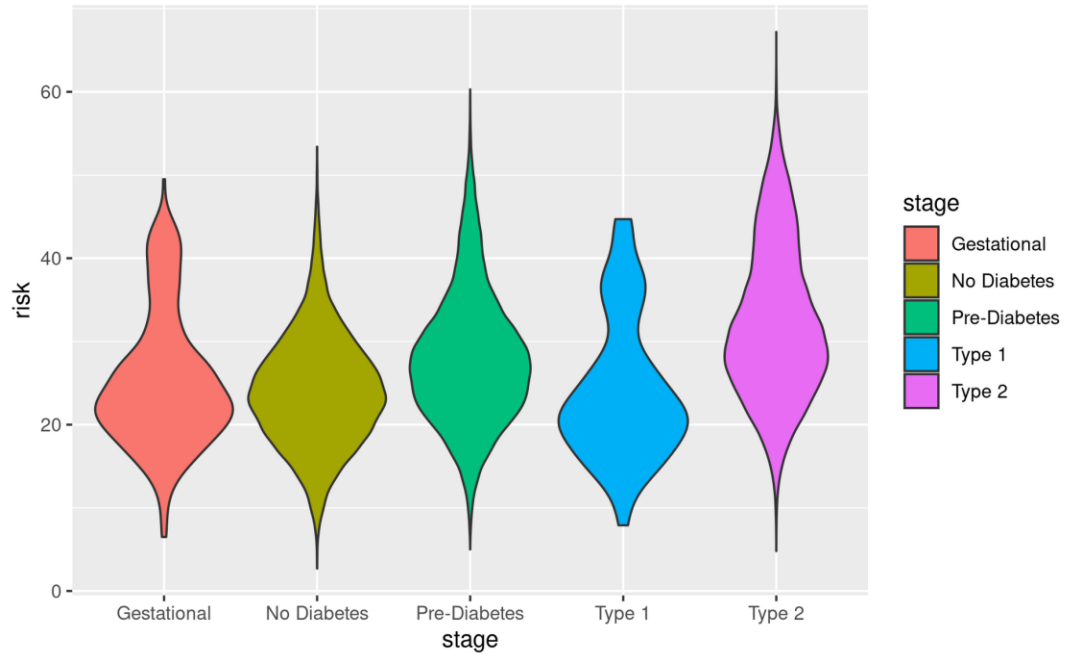
- In the US alone (2021):
  - 11.6% of the population or 38.4 million people of all ages had diabetes
  - 3.4% of US adults or 8.7 million adults had undiagnosed diabetes or were unaware that they had diabetes (this is 22.8% of all US adults with diabetes)
  - Disproportionately affects the elderly and people of race/ethnicity groups
  - Substantially increased rate of undiagnosed diabetes in women
- People I know struggle with diabetes & I want to help them better
- We know that socioeconomic health, mental health & physical health increase & decrease in tandem

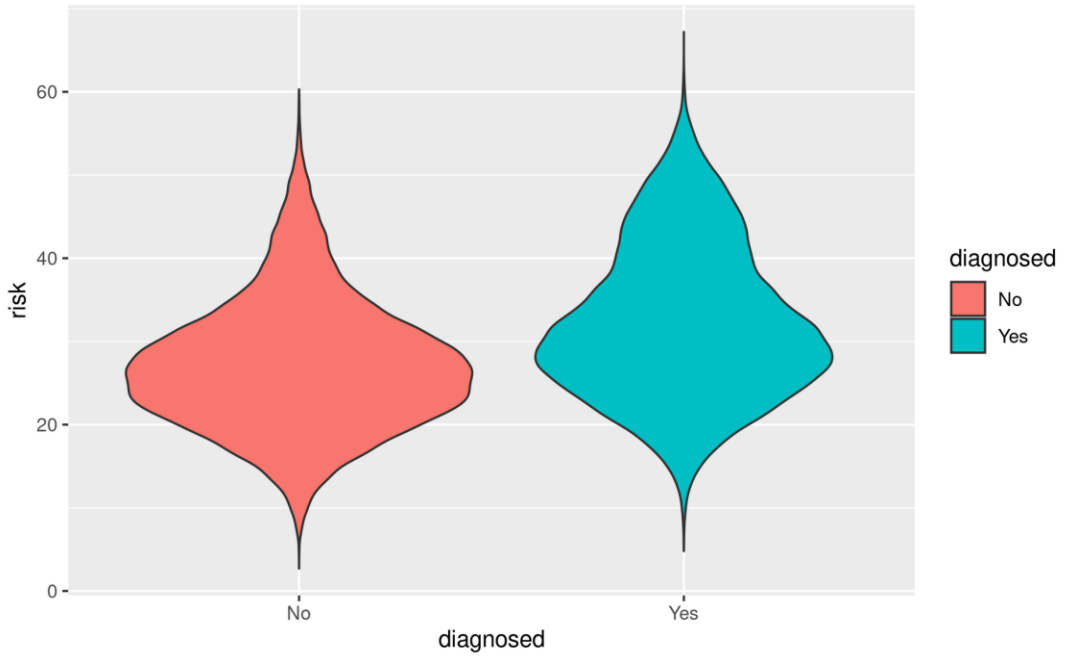
Data Source: https://www.cdc.gov/diabetes/php/data-research/index.html

Data Summary

- Source: https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset?resource=download
- 100,000 observations, 31 variables
- 3 response variables: risk score (numeric), diabetes stage (factor), diagnosed (factor)
- 28 possible predictor variables:
  - Identity: age, gender, and ethnicity
  - Socioeconomic: education level, income, employment
  - Physical health: smoker status, weekly alcohol consumption, weekly physical activity, diet score, daily hours of sleep, daily screen time, BMI, waist to hip ratio, systolic & diastolic blood pressure, heart rate
  - Family History: diabetes, hypertension, cardiovascular
  - Blood Work: total, HDL, & LDL cholesterol, triglycerides, fasting glucose, postprandial glucose (after a meal), insulin, hba1c

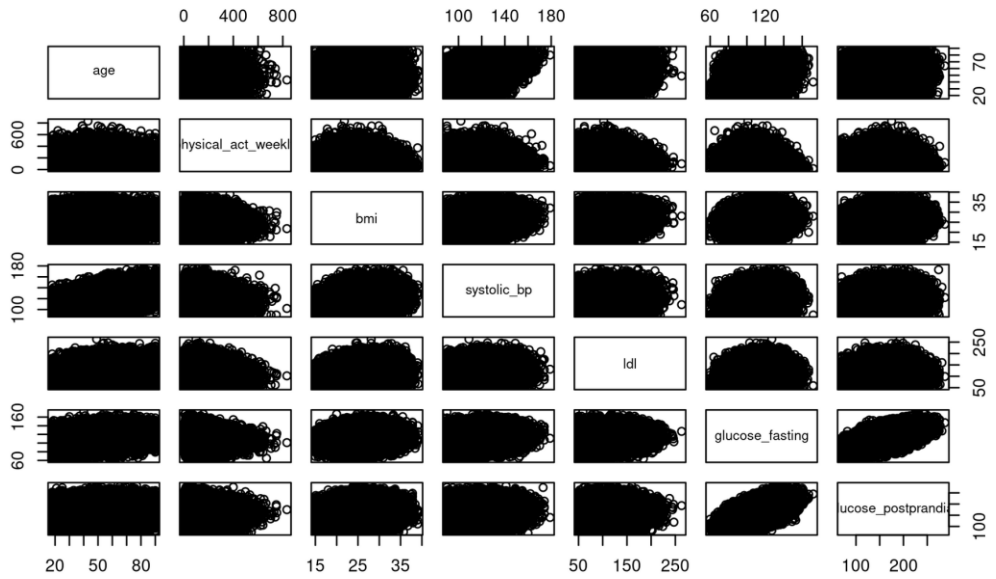
Data Exploration - Response





Data Exploration - Predictors

```
>     cor(diabetes_clean$risk, diabetes_clean[,sapply(diabetes_clean, class) !="factor"])
          age physical_act_weekly diet_score sleep_daily screen_time_daily       bmi
[1,] 0.4959239           -0.348121 -0.1448913 0.003136257         0.0712838 0.3138135
     waist_hip systolic_bp diastolic_bp         hr total_chol        hdl       ldl
[1,] 0.2416505    0.323591    0.1360903 0.09204547  0.1979946 -0.1744592 0.2277722
     triglycerides glucose_fasting glucose_postprandial   insulin     hba1c risk
[1,]     0.1804844       0.4699382            0.2770409 0.1422104 0.3299472    1
>     cov(diabetes_clean$risk, diabetes_clean[,sapply(diabetes_clean, class) !="factor"])
          age physical_act_weekly diet_score sleep_daily screen_time_daily      bmi
[1,] 70.12422           -266.2703  -2.338274   0.0311083          1.594438 10.19923
     waist_hip systolic_bp diastolic_bp     hr total_chol       hdl      ldl
[1,] 0.1025604    41.88407     10.11734 6.9828   57.43545 -16.23131 68.91613
     triglycerides glucose_fasting glucose_postprandial  insulin    hba1c     risk
[1,]      70.93421        57.89483             77.66064 6.383999 2.433476 82.11087
```



Variable Selection

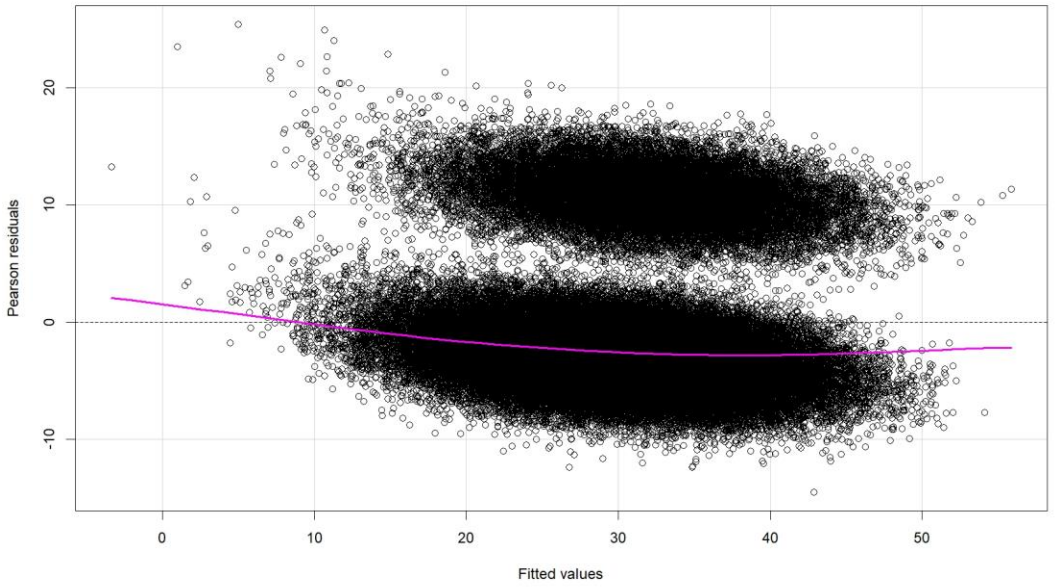
Selection Method	*p*	Predictor Variables	AIC	BIC	Mallow's C_p	Adjusted R^2	RMSE
Backwards Elimination & Stepwise Regression	11	Age, Physical Activity, Diet Score, Screen Time, BMI, Total Cholesterol, HDL, LDL, Triglycerides, Fasting Glucose	365169.4	649073.3	8.357682	0.5306875	6.207368
Forward Selection	9	Age, Fasting Glucose, Physical Activity, BMI, HDL, Diet Score, Screen Time, Triglycerides	365168.1	649052.9	7.059133	0.5306842	6.207452

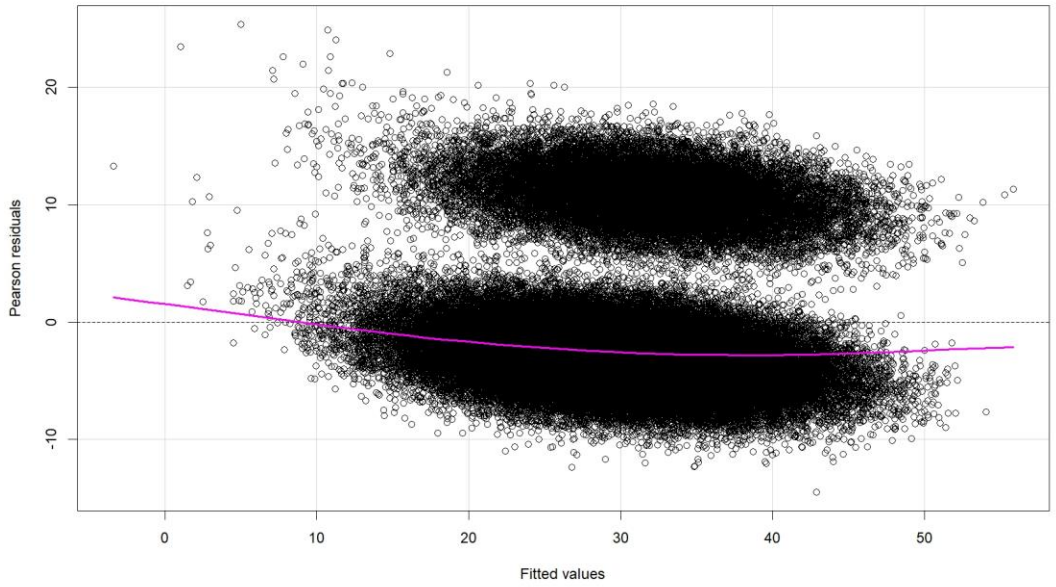
Collinearity

```
+                        glucose_fasting )))),3)
                     age physical_act_weekly diet_score screen_time_daily    bmi total_chol    hdl    ldl triglycerides glucose_fasting
age                1.000               0.003     -0.003            -0.005  0.093      0.310 -0.016  0.281         0.039           0.232
physical_act_weekly 0.003               1.000     -0.002             0.001 -0.072     -0.010  0.018 -0.014        -0.027          -0.162
diet_score         -0.003              -0.002      1.000             0.002 -0.201     -0.039  0.042 -0.046        -0.080          -0.073
screen_time_daily  -0.005               0.001      0.002             1.000 -0.003      0.001 -0.001  0.001        -0.003           0.035
bmi                 0.093              -0.072     -0.201            -0.003  1.000      0.200 -0.212  0.238         0.406           0.151
total_chol          0.310              -0.010     -0.039             0.001  0.200      1.000 -0.042  0.906         0.084           0.091
hdl                -0.016               0.018      0.042            -0.001 -0.212     -0.042  1.000 -0.324        -0.085          -0.081
ldl                 0.281              -0.014     -0.046             0.001  0.238      0.906 -0.324  1.000         0.098           0.105
triglycerides       0.039              -0.027     -0.080            -0.003  0.406      0.084 -0.085  0.098         1.000           0.087
glucose_fasting     0.232              -0.162     -0.073             0.035  0.151      0.091 -0.081  0.105         0.087           1.000


> vif(back_elim)
              age physical_act_weekly          diet_score   screen_time_daily                 bmi          total_chol                 hdl                 ldl
         1.163110            1.032338            1.045646            1.001520            1.343760            9.372796            1.893527           10.286002
    triglycerides     glucose_fasting
         1.198077            1.112661
> vif(forw_select)
              age     glucose_fasting physical_act_weekly                 bmi                 hdl          diet_score   screen_time_daily       triglycerides
         1.063911            1.112647            1.032334            1.310197            1.050039            1.045635            1.001508            1.198068
```

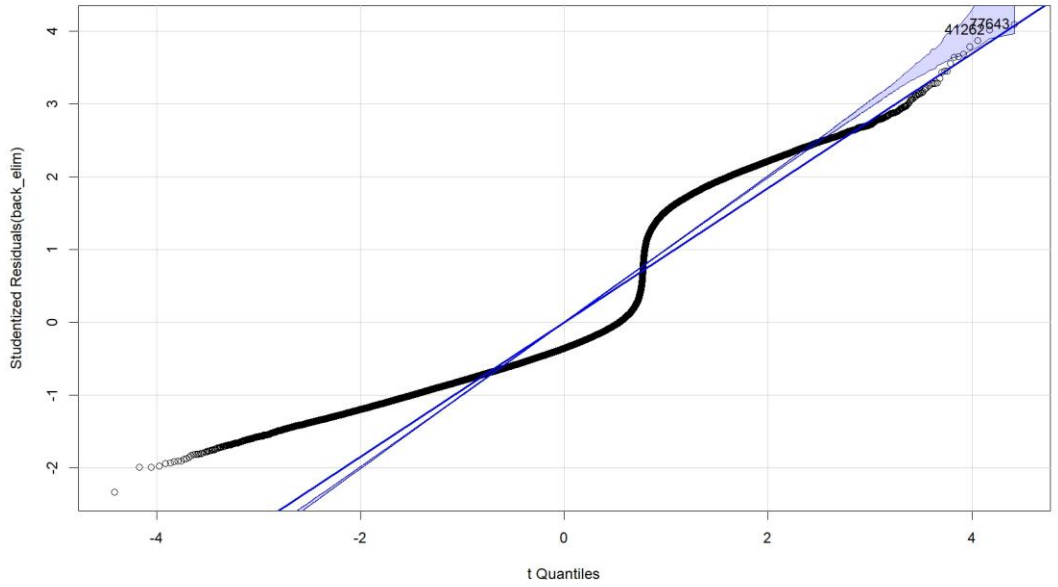
Error Assumptions – Residuals vs. Fitted

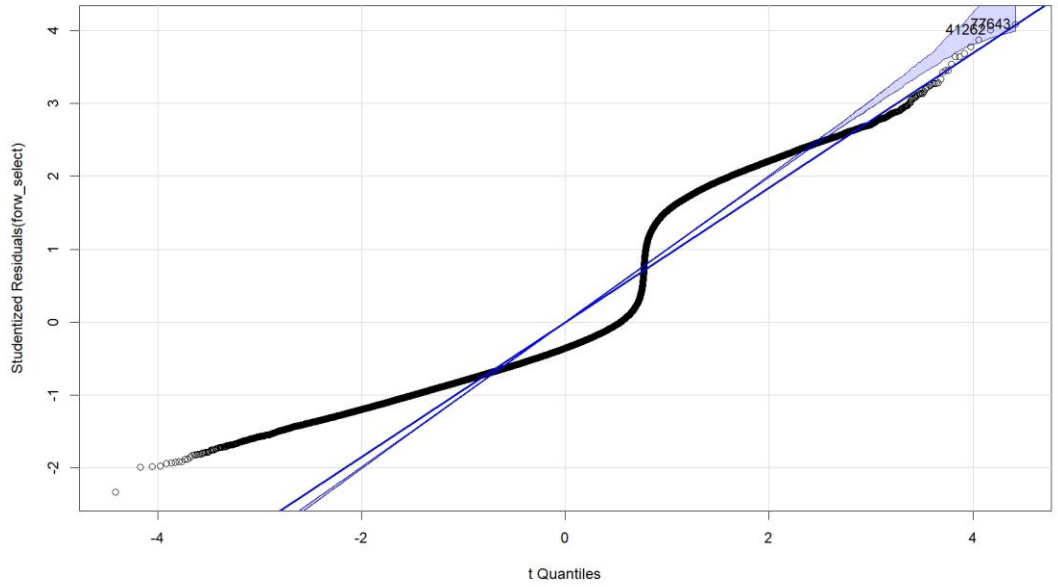




Denver

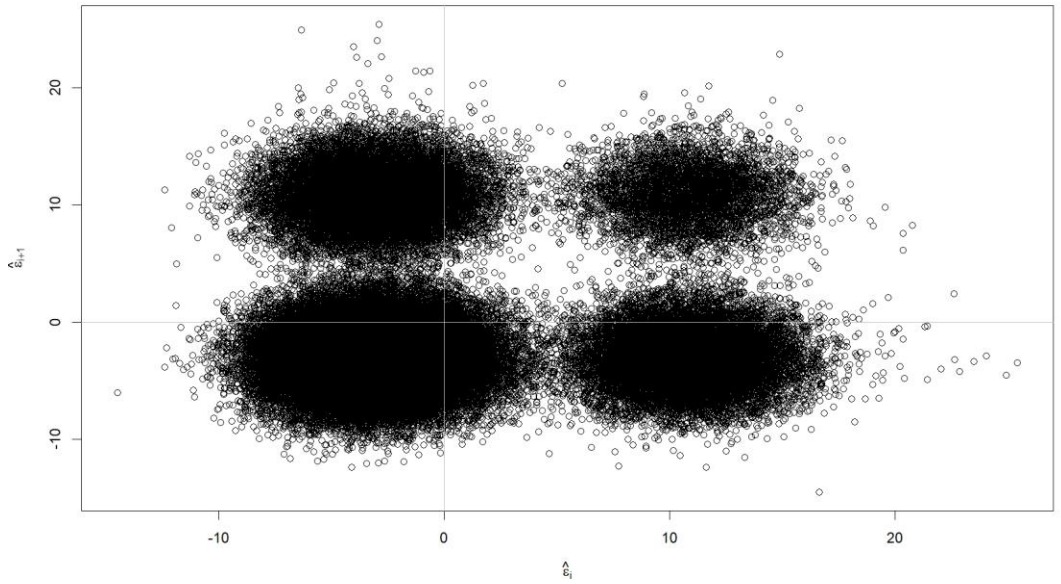
Error Assumptions – QQ Plots

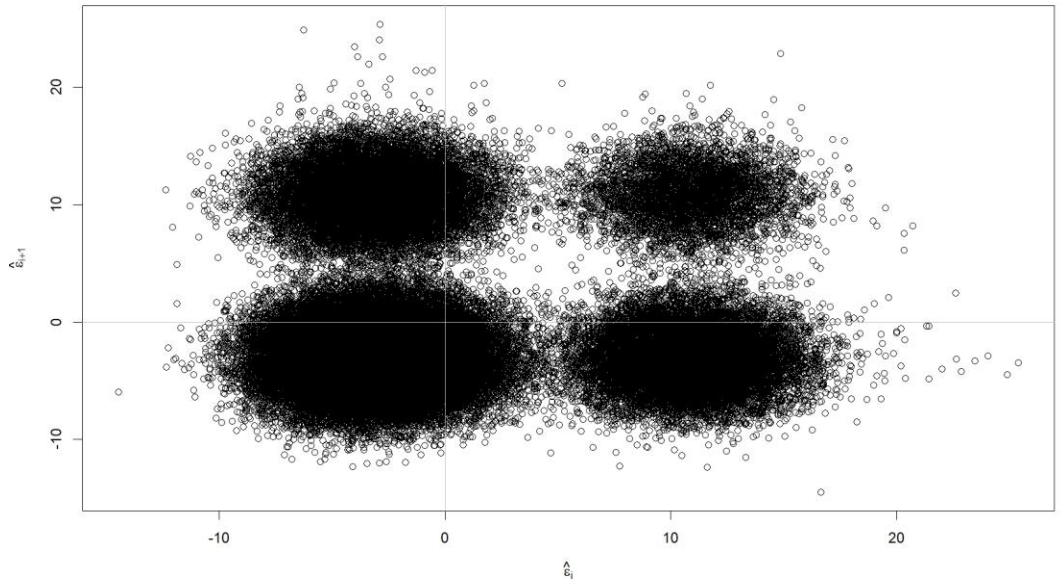




Denver

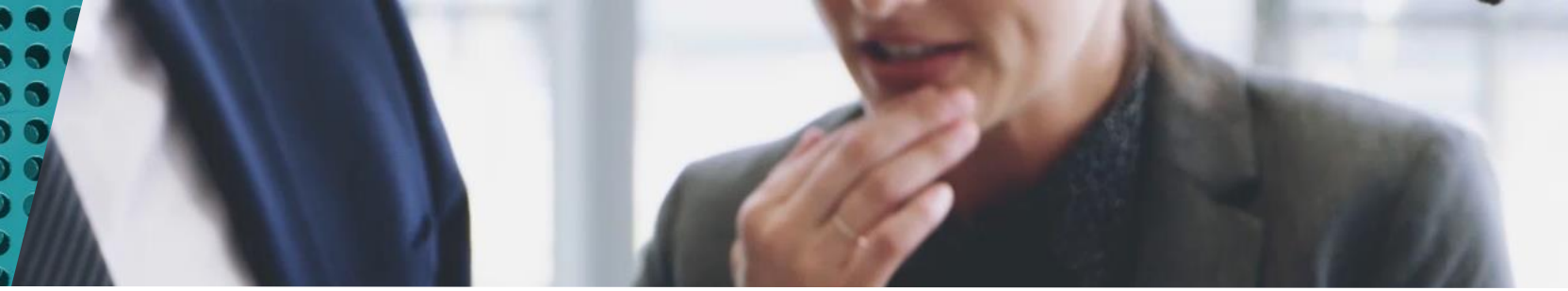
Error Assumptions – Serial Correlation





Final Model… Chosen by all regressor coefficients having p-values of 0 effectively

$$\widehat{E(risk)} = -4.4161 + 0.2401 \cdot age + 0.1887 \cdot fasting\ glucose - 0.0311 \cdot weekly\ physical\ activity + 0.3830418 \cdot bmi - 0.0874 \cdot HDL\ Cholesterol - 0.4309 \cdot diet\ score + 0.2365 \cdot daily\ screen\ time + 0.0116 \cdot triglycerides$$



Final Thoughts

- Need to reassess variable selection accounting for the factor variables
- Examine sources for flawed error assumptions more closely
  - Expecting separation comes from differences in diagnostic status or stage
- Run more statistical tests on error assumptions and regressors