Multiple Regression Analysis of Prostate-Specific Antigens

Yemisi Obasemo and Monica Orme

STA 206

UCDAVIS
UNIVERSITY OF CALIFORNIA

Agenda

01 Introduction & Project Objectives

02 Data Analysis Approach

03 Model Selection

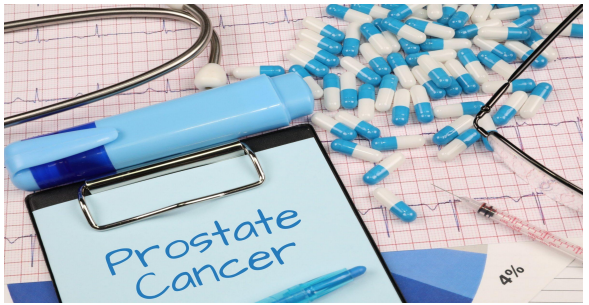
04 Results of the Final Model

05 Conclusion

06 Discussion

UCDAVIS

Introduction

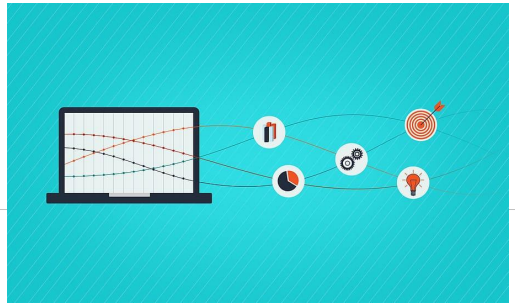


- Prostate cancer is a disease that can severely harm the health of American men - it's the second leading cause of cancer death in men in the U.S. after lung cancer.
- A university medical center urology group was interested in determining the association between prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer.
- Data were collected on 97 men who were about to undergo radical prostatectomies
- Establishing an association can help scientific researchers determine what factors significantly impact PSA levels and the risk of prostate cancer

UCDAVIS

Project Objectives

1. Investigate which factors significantly affect PSA levels
2. Determine which regression model best explains the variability in PSA levels
3. Assess the reliability of the final model's results through data analysis



UCDAVIS

Prostate Dataset

- 97 observations (male patients)
- 7 continuous variables
  - log_PSA_level - logarithmic transformed PSA_levels
  - PSA_level - Serum prostate-specific antigen level (mg/ml)
  - sqrt_cancer_volume - square root transformed prostate cancer volume
  - cancer_volume - Estimate of prostate cancer volume (cc)
  - weight -Prostate weight (gm)
  - age - Age of patient (years)
  - benign_prostatic_hyperplasia - amount of benign prostatic hyperplasia (cm^2)
  - capsular_penetration - Degree of capsular penetration (cm)
- 3 categorical variables
  - seminal_vesicle_invasion - Presence or absence of seminal vesicle invasion: 1 if presence; 0 if absence
  - gleason_score - Pathologically determined grade of disease using total score of two patterns (summed scores were either 6,7,or 8 with higher scores indicating worse prognosis)
  - bph_presence - Presence or absence of benign prostatic hyperplasia: 1 if presence; 0 if absence

Data Analysis Approach

1. Conduct Exploratory data analysis (EDA)
2. Perform preliminary model fitting on the entire dataset
3. Conduct model selection and validation to select a final model
4. Perform statistical inference on the final model

EDA Process

- Verify the distributions of the variables through boxplots, histograms, pie charts, and scatter plots
- Determine if any variables should be transformed
  - Transformed PSA_levels to log_PSA_levels based on the Box-Cox procedure
  - Applied the square root transformation to cancer_volume due to non-linearity with the response variable
- Fit a preliminary regression model using all data
- Check for influencing cases
  - Removed the 32nd observation since its cook's distance value was extreme compared to the that of the other observations and was greater than 1

Model Selection

- Split the data 50/50 into training and validation and checked that the distributions are similar. There are 96 dataset left after the influential case has been removed. Each dataset has 48 datasets.
- Fit the model using the training data
- Do subset regression and get the AIC, BIC, R_squared_adjusted and C_p
- Get the one best model for all model sizes and name it model 1 and we used the function regsubsets in the leaps library to get a summary for the one best model.
- Fit the none-model with no X variable since the regsubsets function does not fit it.
- We can see that the SSE was decreasing as the r_squared and r-squared-adjusted was increasing

Model Selection(continued)

- Use the forward or backward stepwise to select the final best model (model 1) .
- The final best model based on the forward or backward stepwise is “log_PSA_level ~ sqrt_cancer_volume + seminal_vesicle_invasion + benign_prostatic_hyperplasia + gleason_score.
- Use the training data to rerun the first model (best model)
- The coefficients of the chosen X variables in model 1 are significant
- Do model diagnostic for model 1
- We can see a linear relationship between the log PSA level and the chosen X variables and we can say that model 1 is a good model.

UCDAVIS

Model Selection and Validation for model 1

- Obtain $Press_p$ value (22.08743) for model 1.
- Obtain the estimated regression and standard errors of model 1 built on both the training and validation datasets.

	Train Est1	Valid Est1	Train s.e1.	Valid s.e1.
(Intercept)	1.21095161	0.72558271	0.2881334	0.27247089
sqrt_cancer_volume	0.27208481	0.55928274	0.1106114	0.11834660
seminal_vesicle_invasion1	0.83689816	0.27523053	0.2712690	0.45596768
benign_prostatic_hyperplasia	0.09201267	0.09061299	0.0321479	0.04098196
gleason_score7	0.24184981	0.06129269	0.2292620	0.27426810
gleason_score8	0.68955502	0.62178729	0.2898624	0.46898977

- Most of the estimated coefficients as well as their standard errors agree quite closely on the two data sets.

	SSE1	R2_adj1
train_sum1	17.37095	0.5690635
valid_sum1	29.13088	0.5956635

- The SSE values are quite far, but the adjusted R squares are very close
- Find the *SSE*/*n* under the training model, and compare it to the *MSPEv* .The MSPE is not much larger than the SSE divided by n, so it doesn't overfit the data as much.

Model Selection (continued)

- We chose another model (log_PSA_level regressing onto sqrt_cancer_volume, seminal_vesicle_invasion, benign_prostatic_hyperplasia, capsular_penetration, and gleason_score) based on the r_adjusted criterion and name it model 2.
- We repeat the procedures we took above for model 1 and apply it to model 2. Then, compare the two models.
- We used the training data to rerun the second model and found out that the models' coefficients except the ones corresponding to capsular_penetration and gleason_score7 were significant.
- Do model diagnostics for model 2
- From the QQ-plot and fitted versus residuals plot, we can see an almost linear relationship between the log PSA level and the X variables.

Model Selection and Validation process for model 2

- Obtained $Press_p$ value (27.32679) for model 2.
- Obtain the estimated regression and standard errors of model 2 built on both the training and validation datasets.

	Train Est2	Valid Est2	Train s.e.2	Valid s.e.2
(Intercept)	1.13545838	0.68231775	0.29532119	0.28624706
sqrt_cancer_volume	0.33892559	0.58966284	0.12572802	0.13199138
seminal_vesicle_invasion1	0.97124078	0.39928821	0.29646044	0.51423252
benign_prostatic_hyperplasia	0.08699184	0.09149602	0.03237966	0.04135697
capsular_penetration	-0.04686808	-0.02830124	0.04229674	0.05253867
gleason_score7	0.21998164	0.08946019	0.22949283	0.28146025
gleason_score8	0.66489101	0.68592116	0.28993499	0.48766607

- Most of the estimated coefficients as well as their standard errors agree quite closely on the two data sets.

	SSE2	R2_adj2
train_sum2	16.86587	0.5713886
valid_sum2	28.93100	0.5888768

- The SSE values are quite far, but the adjusted R squares are very close
- Find the *SSE*/*n* under the training model, and compare it to the *MSPEv.* The MSPE is not much larger than the SSE divided by n, so it doesn't overfit the data as much.

Comparison between model 1 and model 2

- We can see that the variable “capsular_penetration” was not present in model 1 but present in model 2
- However, the coefficient corresponding to this variable was not significant in model 2, which makes it a justifiable reason for not being present in the “best model” (Model 1).
- The $Press_p$ for Model 1 (22.08743) appears better than the $Press_p$ for model 2 (27.32679), since it is smaller, this implies that model 1 fits the data better
- Most of the estimated coefficients as well as their standard errors for the training and validation data are closer to each other in model 1 compared to model 2.

Comparison between model 1 and model 2 (continued)

- The SSE values and adjusted R squares are farther from each other in the training and validation data in model 2 compared to model 1, which implies that there's more variability between the training and validation data in model 2

- The difference between the MSPEv and SSE/n in model 2 is greater than that of model 1 and this implies that model 1 fits the data better than model 2

- Based on the comparison above, our final preferred model is Model 1, which is the final model chosen by forward or backward stepwise selection

Final Model - Results and Methodology

Final Multiple Regression Model (Model 1):

$$\log(\text{PSA\_levels}) = 1.211 + 0.092*X_{\text{benign\_prostatic\_hyperplasia}} + 0.272*X_{\text{sqrt\_cancer\_volume}} + 0.837*X_{\text{seminal\_vesicle\_invasion\_1}} + 0.242*X_{\text{gleason\_score7}} + 0.69*X_{\text{gleason\_score8}}$$

Model Assumptions:

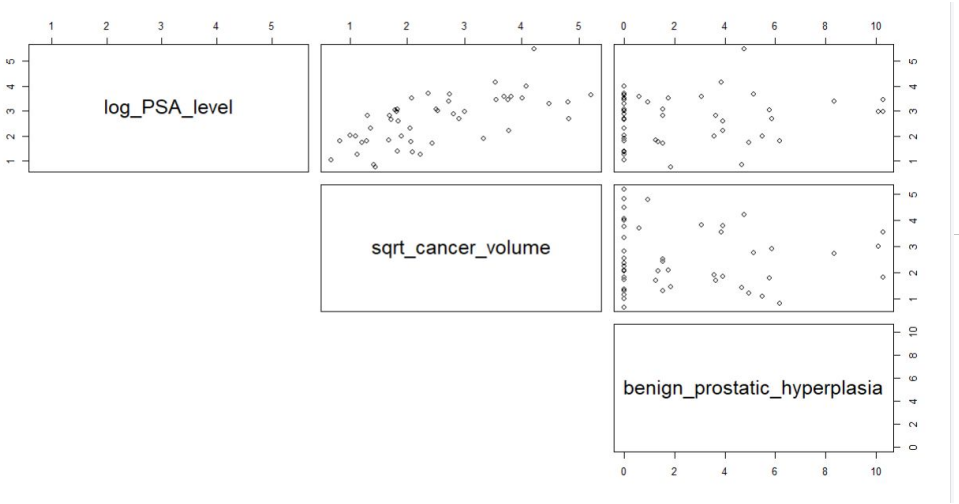
1. Linearity between the quantitative variables and the response variable
2. Errors are normally distributed with mean 0 and constant variance
3. The errors are uncorrelated

UCDAVIS

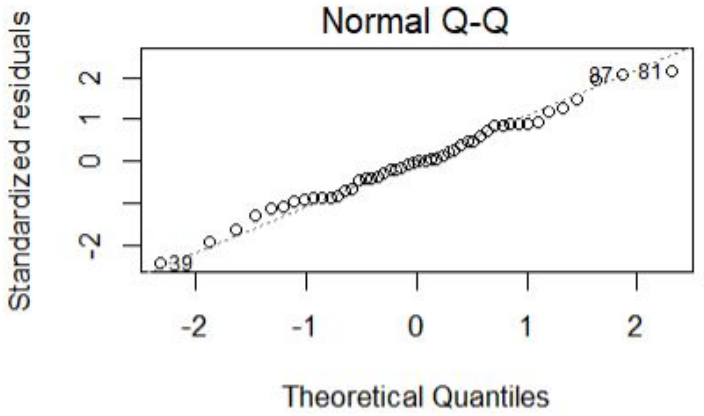
Final Model - Summary Output

```
lm(formula = log_PSA_level ~ sqrt_cancer_volume + seminal_vesicle_invasion +
    benign_prostatic_hyperplasia + gleason_score, data = data.c)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47054 -0.43833 -0.01212  0.43671  1.33867

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.21095    0.28813   4.203 0.000135 ***
sqrt_cancer_volume          0.27208    0.11061   2.460 0.018099 *
seminal_vesicle_invasion1   0.83690    0.27127   3.085 0.003592 **
benign_prostatic_hyperplasia 0.09201   0.03215   2.862 0.006535 **
gleason_score7              0.24185    0.22926   1.055 0.297499
gleason_score8              0.68956    0.28986   2.379 0.021986 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6431 on 42 degrees of freedom
Multiple R-squared:  0.6149,    Adjusted R-squared:  0.5691
F-statistic: 13.41 on 5 and 42 DF,  p-value: 7.866e-08
```

Final Model - Diagnostics

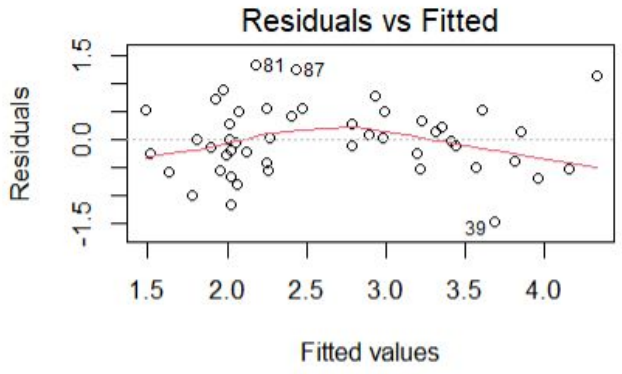


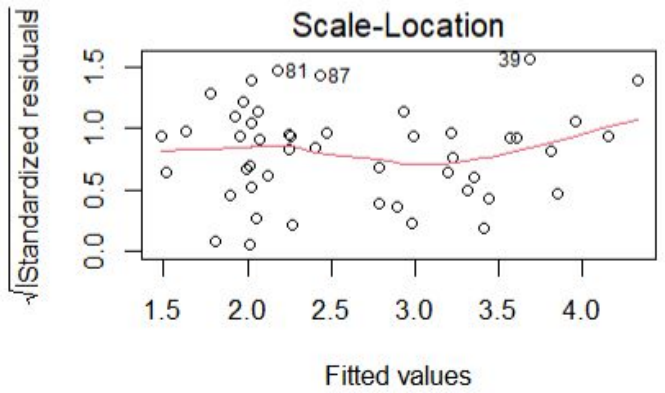
Linearity



Normality

Final Model - Diagnostics





Constant Variance and Uncorrelatedness of the Errors

Final Model - Results and Interpretation

Final Multiple Regression Model (Model 1):

$$\log(\text{PSA\_levels}) = 1.211 + 0.092*X_{\text{benign\_prostatic\_hyperplasia}} + 0.272*X_{\text{sqrt\_cancer\_volume}} + 0.837*X_{\text{seminal\_vesicle\_invasion\_1}} + 0.242*X_{\text{gleason\_score7}} + 0.69*X_{\text{gleason\_score8}}$$

Goodness of Fit Metrics:

- Adjusted R^2 - 0.5691
  - 56.91% of the variation in the log-transformed response can be explained by the model
- MSE - 0.4136
  - The average squared difference between the actual log-transformed PSA levels and the fitted values was low

Final Model - Interpretation of the Coefficients

Coefficient Name	Beta Coefficient Value (β_i)
Intercept	1.211
benign_prostatic_hyperplasia	0.092
sqrt_cancer_volume	0.272
seminal_vesicle_invasion_1	0.837
gleason_score7	0.242
gleason_score8	0.69

Coefficient Interpretation: The change in the average log-transformed PSA levels by the coefficient value with a unit increase in the corresponding predictor when all other predictors are held constant

UCDAVIS

Final Model - Statistical Inference

Inference of the Model Coefficients through T-Test

Coefficient Name	p-value
Intercept	0.000135
benign_prostatic_hyperplasia	0.018
sqrt_cancer_volume	0.0036
seminal_vesicle_invasion_1	0.0065
gleason_score7	0.2975
gleason_score8	0.022

- Decision Rule: A coefficient is considered significant if its p-value is below the significance level of 0.05
- Conclusion: The intercepts and all of the coefficients except for gleason_score7 have a significant effect on the change in the mean response

UCDAVIS

Conclusion

- Model 1 is reliable for predicting the PSA levels of a typical or average man based on the goodness of fit metrics and the model's assumptions were satisfied
- The factors that significantly affect the change in the log-transformed PSA levels:
  - seminal vesicle invasion
  - square-root transformed cancer volume
  - benign prostatic hyperplasia
  - gleason score (with only two levels based on a gleason score of 8) based on the t-tests on the model's coefficients.
- The coefficient corresponding to a gleason score of 7 is unreliable for prediction and interpretation since its standard error is high
- This model shouldn't be used to predict the PSA levels of a man with extreme levels of cancer volume and/or benign hyperplasia since extreme observations weren't contained in building this model.

Discussion - Possible Limitations

- The training and validation datasets aren't the largest since both are less than 50
- Benign Prostatic Hyperplasia (BPH) and Seminal Vesicle Invasion had several 0 values, which may have skewed the results of the final model.
- We attempted to address the issue with BPH by adding a categorical variable to account for the presence and absence of BPH,we could've also added a categorical variable to account for the presence and absence of seminal vesicle invasion and determine if this predictor would've been significant to include in the model.
- The additional categorical variable did not really have an effect on the model as it wasn't chosen as part of the X variables in the preferred final model. The adjusted R^2 also did not quite improve.
- The adjusted R^2 for the final model was moderate at 56.91%. Although more than half of the variance in the response is explained by the model, this goodness of fit isn't strong.
- It may have been worth checking to see if other regression models such as ridge regression and if adding interaction terms could have provided a better fit.