A Model Predicting Number of Prescriptions Filled by Medicare Beneficiaries

An analysis completed by Andi McCollam

December 14, 2016

Background

Prescription drugs are an important component of health care delivery. Through established and innovative pharmaceutical treatments, people are living longer and healthier lives.

In the United States, many people have their prescription drugs paid for through a program called Medicare Part D. Part D is available to anybody who qualifies for Medicare; that is, most people who are over the age of 65 and/or disabled. Part D has been available since 2006, and was launched as part of the Medicare Prescription Drug, Improvement, and Modernization Act of 2003.

The purpose of this analysis is to build a model predicting how many prescriptions a Part D member will fill over the course of one year.

Methods

The data for this model was obtained in the 2010 Chronic Conditions Public Use File, published by the Centers for Medicare & Medicaid Services (CMS). This file contains aggregated data about Medicare claims and beneficiaries from 2010. The data dictionary can be found in Appendix A.

An initial review was made of the complete data set. There were 21,364 rows across the 55 columns listed in the appendix. The data is categorized based on six age categories, two genders, whether or not the member also had Medicaid (dual eligible), and comorbidity indicators for eleven conditions – up to 49,152 categories had all combinations occurred.

Each category had a variety of aggregated outcome variables. It provided information about how many beneficiaries had Medicare Part A, Part B, Part C, and Part D. It provided information about services received, such as skilled nursing care and inpatient hospital admissions. This information took the form of average cost per beneficiary and average number of units per beneficiary. It was further broken out by people who had been enrolled with Medicare Part A, Part B, Part C, and Part D for the full year or for only part of 2010, which reduced the number of categories to 22,003.

For this analysis, the outcome variable of interest was AVE_PDE_PD_EQ_12, the average number of prescriptions per beneficiary in 2010 for people who had been enrolled in Part D for the full year. Columns containing information about partial year members for Part A, Part B, Part C, and Part D were all removed. 4,537 rows did not have any Part D beneficiaries or prescription counts, so they were removed. 767 rows were removed which all had missing values for the same five comorbidities – alzheimers disease, COPD, depression, osteoporosis, and stroke. The remaining 33 columns and 16699 rows were used to construct a predictive model.

Poisson Model

When analyzing count data, such as the number of prescriptions filled by one person in a year, the Poisson distribution is the standard model used. Another discrete distribution, the Binomial distribution, was not explored because it is only appropriate when, "the range of values for which there exist positive probabilities has finite length" (Klugman, Panjer, & Willmot, 2004). That is, it was not appropriate in this case because the number of prescriptions a person could fill in one year has no maximum value.

Since the data did not provide a literal count of all prescriptions filled for Part D beneficiaries, a new column was created by multiplying the average number of prescriptions times the total number of Part D beneficiaries for each category. Since this lead to some categories having inflated prescription counts due to having a lot of beneficiaries, by default an offset factor was included of the log of beneficiary count in my models. The log of this value was used because R uses a log transformation on the dependent variable when family poisson is used in the glm() function; the offset should reflect this.

An initial Poisson model was built using glm() and all covariates except for BENE_COUNT_PA_EQ_12, BENE_COUNT_PB_EQ_12, and BENE_COUNT_PC_EQ_12. These were the beneficiary counts for Part A, Part B, and Part C, which highly correlated with beneficiary counts for part D. The log was used for continuous covariates.

The resulting model had a deviance of $6{,}977{,}394/12{,}015 = 580.72$, far from the ideal value of 1. The AIC was an astonishing 7,115,908. There were no obvious covariates to remove as each had a observed level of significance of less than 2e-16, the lowest value that R reports. That is, each covariate rejected the null hypothesis that it was equal to zero and did not have a relationship with the outcome variable.

Functions stepAIC() and drop1() were used to try to improve the model, but neither removed any variables from the original model. Some exploratory Poisson models were created to see if using demographic covariates only, other service charge amounts only, or the average number of other services only would potentially reduce the deviance and AIC. Every model resulted in deviance values in the hundreds and AIC values in the millions. There was evidence of extreme overdispersion, as illustrated by a Pearson statistic of 583.051 on the original model, and reaching as high as 14,407.48 on the smaller models. With no overdispersion, the Pearson statistic would be close to 1.

To attempt to control for this overdispersion, Quasipoisson models were created. Unfortunately, these new models still had similarly large deviance values.

Linear Model

Next, the data was fitted to a linear model. Since there were 16,699 observations and AVE_PDE_PD_EQ_12 is approximately continuous, this was an appropriate model to use.

First, scatterplots were generated to compare how well the outcome variable correlated with each predictor variable (Appendix B). These plots supported much of what the Poisson models had been indicating, which is that most of the covariates correlate with the number of prescriptions filled. While most comorbidities were independent of each other, linear relationships emerged between the number of different types of visits a patient received across different categories of service and how much they paid for these different services. That is, beneficiaries who received some services generally received others, with the exception of AVE_CA_VST_PB_EQ_12, or general physician services. And beneficiaries who had high average expenses for some services often had high average expenses for others.

StepAIC() was used to stepwise construct a linear model (Model 1) using all covariates, with the exception of BENE_COUNT_PA_EQ_12, BENE_COUNT_PB_EQ_12, BENE_COUNT_PC_EQ_12, and BENE_COUNT_PD_EQ_12, the beneficiary counts for Part A, Part B, Part C, and Part D, and AVE_PDE_CST_PD_EQ_12, the average prescription drug cost, since it so closely relates to prescription drug count. The model had an R^2 value of 0.9351. Its residual plots are in Appendix C with its coefficients and indicate an overall good fit, although the values deviate from normality at higher and lower values, as seen in the QQ-plot.

While this model had high correlations and low heteroscedasticity , it had 25 variables, which violated the value of parsimony. Since the variables all had such high correlation with the outcome and the analyst was familiar with the data on a conceptual level, a manual forward stepwise regression analysis was performed. With only ten covariates, the R^2 was 0.8812. The residual plots are in Appendix C (Model 2). Unfortunately, the Residuals vs. Fitted graph indicates that the data is heteroscedastic, which was supported by the Breusch-Pagan test, which rejected the null hypothesis that enough of the variance is explained by the explanatory variables at the p<2.2e-16 level. Transforming the outcome variable by squaring it, taking the square root, taking the log, or exponentiating it also lead to heteroscedastic results.

To control for heteroscedasticity, the robust variance methods available in the sandwich library were employed and used on Model 2. This robust test gave the same coefficients and coefficient levels of significance as the main model. Considering its high correlation and low standard errors, this was determined to be the best model.

Conclusion

This model indicated that as people have more comorbidities, they tend to purchase more prescription drugs. Diabetes and Depression seem to have the largest impact on this behavior, as they lead to 14.9 and 13.8 more prescriptions per year on average.

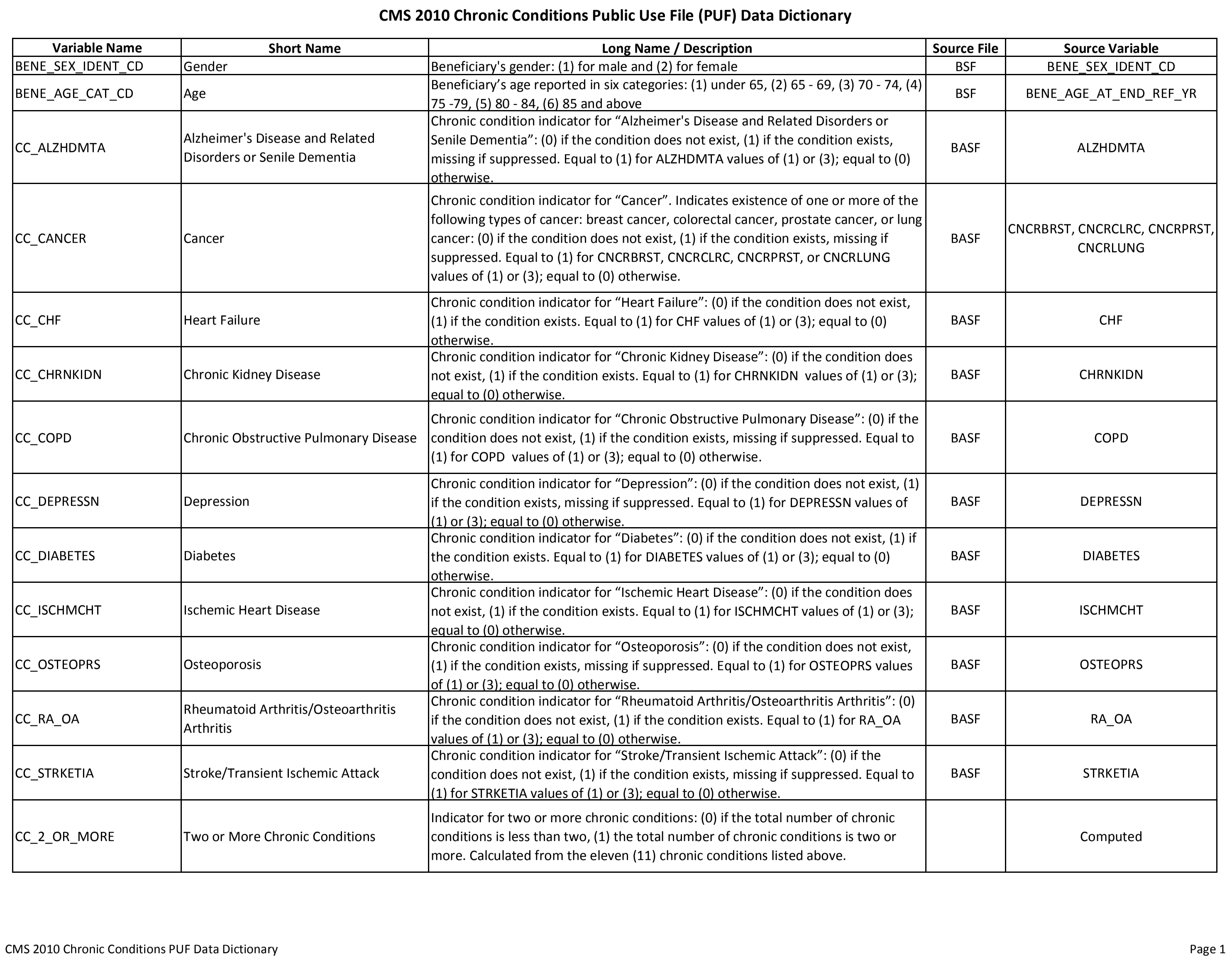
It was interesting that dual eligible beneficiaries – that is, people who were eligible for both Medicare and Medicaid – purchased 26.9 more prescriptions per year on average than people who were on Medicare alone. This might be because people on Medicaid have no out of pocket costs associated with their prescriptions. It might also be due to Medicaid being more often available to people with disabilities, which might also be the cause of more prescriptions.

Works Cited

Centers for Medicare & Medicaid Services. (2012, Oct 9). *Chronic Conditions PUF*. Retrieved Dec 2016, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/Chronic_Conditions_PUF.html

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2004). *Loss Models: From Data to Decisions* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Appendix A



CMS 2010 Chronic Conditions Public Use File (PUF) Data Dictionary

Variable Name	Short Name	Long Name / Description	Source File	Source Variable
BENE_SEX_IDENT_CD	Gender	Beneficiary's gender: (1) for male and (2) for female	BSF	BENE_SEX_IDENT_CD
BENE_AGE_CAT_CD	Age	Beneficiary's age reported in six categories: (1) under 65, (2) 65 - 69, (3) 70 - 74, (4) 75 -79, (5) 80 - 84, (6) 85 and above	BSF	BENE_AGE_AT_END_REF_YR
CC_ALZHDMTA	Alzheimer's Disease and Related Disorders or Senile Dementia	Chronic condition indicator for "Alzheimer's Disease and Related Disorders or Senile Dementia": (0) if the condition does not exist, (1) if the condition exists, missing if suppressed. Equal to (1) for ALZHDMTA values of (1) or (3); equal to (0) otherwise.	BASF	ALZHDMTA
CC_CANCER	Cancer	Chronic condition indicator for "Cancer". Indicates existence of one or more of the following types of cancer: breast cancer, colorectal cancer, prostate cancer, or lung cancer: (0) if the condition does not exist, (1) if the condition exists, missing if suppressed. Equal to (1) for CNCRBRST, CNCRCLRC, CNCRPRST, or CNCRLUNG values of (1) or (3); equal to (0) otherwise.	BASF	CNCRBRST, CNCRCLRC, CNCRPRST, CNCRLUNG
CC_CHF	Heart Failure	Chronic condition indicator for "Heart Failure": (0) if the condition does not exist, (1) if the condition exists. Equal to (1) for CHF values of (1) or (3); equal to (0) otherwise.	BASF	CHF
CC_CHRNKIDN	Chronic Kidney Disease	Chronic condition indicator for "Chronic Kidney Disease": (0) if the condition does not exist, (1) if the condition exists. Equal to (1) for CHRNKIDN values of (1) or (3); equal to (0) otherwise.	BASF	CHRNKIDN
CC_COPD	Chronic Obstructive Pulmonary Disease	Chronic condition indicator for "Chronic Obstructive Pulmonary Disease": (0) if the condition does not exist, (1) if the condition exists, missing if suppressed. Equal to (1) for COPD values of (1) or (3); equal to (0) otherwise.	BASF	COPD
CC_DEPRESSN	Depression	Chronic condition indicator for "Depression": (0) if the condition does not exist, (1) if the condition exists, missing if suppressed. Equal to (1) for DEPRESSN values of (1) or (3); equal to (0) otherwise.	BASF	DEPRESSN
CC_DIABETES	Diabetes	Chronic condition indicator for "Diabetes": (0) if the condition does not exist, (1) if the condition exists. Equal to (1) for DIABETES values of (1) or (3); equal to (0) otherwise.	BASF	DIABETES
CC_ISCHMCHT	Ischemic Heart Disease	Chronic condition indicator for "Ischemic Heart Disease": (0) if the condition does not exist, (1) if the condition exists. Equal to (1) for ISCHMCHT values of (1) or (3); equal to (0) otherwise.	BASF	ISCHMCHT
CC_OSTEOPRS	Osteoporosis	Chronic condition indicator for "Osteoporosis": (0) if the condition does not exist, (1) if the condition exists, missing if suppressed. Equal to (1) for OSTEOPRS values of (1) or (3); equal to (0) otherwise.	BASF	OSTEOPRS
CC_RA_OA	Rheumatoid Arthritis/Osteoarthritis Arthritis	Chronic condition indicator for "Rheumatoid Arthritis/Osteoarthritis Arthritis": (0) if the condition does not exist, (1) if the condition exists. Equal to (1) for RA_OA values of (1) or (3); equal to (0) otherwise.	BASF	RA_OA
CC_STRKETIA	Stroke/Transient Ischemic Attack	Chronic condition indicator for "Stroke/Transient Ischemic Attack": (0) if the condition does not exist, (1) if the condition exists, missing if suppressed. Equal to (1) for STRKETIA values of (1) or (3); equal to (0) otherwise.	BASF	STRKETIA
CC_2_OR_MORE	Two or More Chronic Conditions	Indicator for two or more chronic conditions: (0) if the total number of chronic conditions is less than two, (1) the total number of chronic conditions is two or more. Calculated from the eleven (11) chronic conditions listed above.		Computed

Variable Name	Short Name	Long Name / Description	Source File	Source Variable
DUAL_STUS	Dual Eligibility Status	Beneficiary's dual eligibility status: (0) if the not dual eligible, (1) if dual eligible. Equal to (1) if any of the monthly indicators (DUAL_STUS_CD_01 - DUAL_STUS_CD_12) for a beneficiary has a value of ‘01’, ‘02’, ‘03’, ‘04’, ‘05’, ’06’, ‘08’, or ‘99’ in the calendar year.	BSF	DUAL_STUS_CD_01 - DUAL_STUS_CD_12
BENE_COUNT_PA_LT_12	Count of Beneficiaries (Part A < 12)	Count of beneficiaries enrolled in Medicare Part A for at least 1 month but less than 12 months in the calendar year	BSF	Computed
AVE_MO_EN_PA_LT_12	Average Months of Enrollment (Part A < 12)	Average months of enrollment for beneficiaries enrolled in Medicare Part A for at least 1 month but less than 12 months in the calendar year	BSF	BENE_HI_CVRAGE_TOT_MONS
AVE_PA_PAY_PA_LT_12	Average Medicare Payment for Part A per Beneficiary (Part A < 12)	Average Medicare payment per beneficiary for all Part A services for beneficiaries enrolled in Medicare Part A for at least 1 month but less than 12 months in the calendar year	BASF	Computed
AVE_IP_PAY_PA_LT_12	Average Medicare Payment for IP per Beneficiary (Part A < 12)	Average Medicare payment for inpatient services per beneficiary for beneficiaries enrolled in Medicare Part A for at least 1 month but less than 12 months in the calendar year	BASF	MEDREIMB_IP
AVE_SNF_PAY_PA_LT_12	Average Medicare Payment for SNF per Beneficiary (Part A < 12)	Average Medicare payment for skilled nursing facility services per beneficiary for beneficiaries enrolled in Medicare Part A for at least 1 month but less than 12 months in the calendar year	BASF	MEDREIMB_SNF
AVE_OTH_PAY_PA_LT_12	Average Medicare Payment for other services per Beneficiary (Part A < 12)	Average Medicare payment for the sum of home health agency and hospice services per beneficiary for beneficiaries enrolled in Medicare Part A for at least 1 month but less than 12 months in the calendar year	BASF	MEDREIMB_HH and MEDREIMB_HS
AVE_IP_ADM_PA_LT_12	Average IP Admissions per Beneficiary (Part A < 12)	Average number of inpatient admissions per beneficiary for beneficiaries enrolled in Medicare Part A for at least 1 month but less than 12 months in the calendar year	BASF	IPSTY
AVE_SNF_DAYS_PA_LT_12	Average SNF Covered Days per Beneficiary (Part A < 12)	Average number of skilled nursing facility covered days per beneficiary for beneficiaries enrolled in Medicare Part A for at least 1 month but less than 12 months in the calendar year	BASF	SNF_COVDYS
BENE_COUNT_PA_EQ_12	Count of Beneficiaries (Part A = 12)	Count of beneficiaries enrolled in Medicare Part A for 12 months in the calendar year	BSF	Computed
AVE_PA_PAY_PA_EQ_12	Average Medicare Payment for Part A per Beneficiary (Part A = 12)	Average Medicare payment per beneficiary for all Part A services for beneficiaries enrolled in Medicare Part A for 12 months in the calendar year	BASF	Computed
AVE_IP_PAY_PA_EQ_12	Average Medicare Payment for IP per Beneficiary (Part A = 12)	Average Medicare payment for inpatient services per beneficiary for beneficiaries enrolled in Medicare Part A for 12 months in the calendar year	BASF	MEDREIMB_IP
AVE_SNF_PAY_PA_EQ_12	Average Medicare Payment for SNF per Beneficiary (Part A = 12)	Average Medicare payment for skilled nursing facility services per beneficiary for beneficiaries enrolled in Medicare Part A for 12 months in the calendar year	BASF	MEDREIMB_SNF
AVE_OTH_PAY_PA_EQ_12	Average Medicare Payment for other services per Beneficiary (Part A = 12)	Average Medicare payment for the sum of home health agency and hospice services per beneficiary for beneficiaries enrolled in Medicare Part A for 12 months in the calendar year	BASF	MEDREIMB_HH and MEDREIMB_HS
AVE_IP_ADM_PA_EQ_12	Average IP Admissions per Beneficiary (Part A = 12)	Average number of inpatient admissions per beneficiary for beneficiaries enrolled in Medicare Part A for 12 months in the calendar year	BASF	IPSTY
AVE_SNF_DAYS_PA_EQ_12	Average SNF Covered Days per Beneficiary (Part A = 12)	Average number of skilled nursing facility covered days per beneficiary for beneficiaries enrolled in Medicare Part A for 12 months in the calendar year	BASF	SNF_COVDYS

Variable Name	Short Name	Long Name / Description	Source File	Source Variable
BENE_COUNT_PB_LT_12	Count of Beneficiaries (Part B < 12)	Count of beneficiaries enrolled in Medicare Part B for at least 1 month but less than 12 months in the calendar year	BSF	Computed
AVE_MO_EN_PB_LT_12	Average Months of Enrollment (Part B < 12)	Average months of enrollment for beneficiaries enrolled in Medicare Part B for at least 1 month but less than 12 months in the calendar year	BSF	BENE_SMI_CVRAGE_TOT_MONS
AVE_PB_PAY_PB_LT_12	Average Medicare Payment for Part B per Beneficiary (Part B < 12)	Average Medicare payment per beneficiary for all Part B services for beneficiaries enrolled in Medicare Part B for at least 1 month but less than 12 months in the calendar year	BASF	Computed
AVE_CA_PAY_PB_LT_12	Average Medicare Payment for CA per Beneficiary (Part B < 12)	Average Medicare payment for carrier/physician services per beneficiary for beneficiaries enrolled in Medicare Part B for at least 1 month but less than 12 months in the calendar year	BASF	MEDREIMB_CAR
AVE_OP_PAY_PB_LT_12	Average Medicare Payment for OP per Beneficiary (Part B < 12)	Average Medicare payment for outpatient services per beneficiary for beneficiaries enrolled in Medicare Part B for at least 1 month but less than 12 months in the calendar year	BASF	MEDREIMB_OP
AVE_OTH_PAY_PB_LT_12	Average Medicare Payment for other services per Beneficiary (Part B < 12)	Average Medicare payment for the sum of home health agency services and durable medical equipments per beneficiary for beneficiaries enrolled in Medicare Part B for at least 1 month but less than 12 months in the calendar year	BASF	MEDREIMB_DME and MEDREIMB_HH
AVE_CA_VST_PB_LT_12	Average CA Visits per Beneficiary (Part B < 12)	Average number of carrier/physician visits per beneficiary for beneficiaries enrolled in Medicare Part B for at least 1 month but less than 12 months in the calendar year	BASF	PHSVST
AVE_OP_VST_PB_LT_12	Average OP Visits per Beneficiary (Part B < 12)	Average number of outpatient visits per beneficiary for beneficiaries enrolled in Medicare Part B for at least 1 month but less than 12 months in the calendar year	BASF	OPVST
BENE_COUNT_PB_EQ_12	Count of Beneficiaries (Part B = 12)	Count of beneficiaries enrolled in Medicare Part B for 12 months in the calendar year	BSF	Computed
AVE_PB_PAY_PB_EQ_12	Average Medicare Payment for Part B per Beneficiary (Part B = 12)	Average Medicare payment per beneficiary for all Part B services for beneficiaries enrolled in Medicare Part B for 12 months in the calendar year	BASF	Computed
AVE_CA_PAY_PB_EQ_12	Average Medicare Payment for CA per Beneficiary (Part B = 12)	Average Medicare payment for carrier/physician services per beneficiary for beneficiaries enrolled in Medicare Part B for 12 months in the calendar year	BASF	MEDREIMB_CAR
AVE_OP_PAY_PB_EQ_12	Average Medicare Payment for OP per Beneficiary (Part B = 12)	Average Medicare payment for outpatient services per beneficiary for beneficiaries enrolled in Medicare Part B for 12 months in the calendar year	BASF	MEDREIMB_OP
AVE_OTH_PAY_PB_EQ_12	Average Medicare Payment for other services per Beneficiary (Part B = 12)	Average Medicare payment for the sum of home health agency services and durable medical equipments per beneficiary for beneficiaries enrolled in Medicare Part B for 12 months in the calendar year	BASF	MEDREIMB_DME and MEDREIMB_HH
AVE_CA_VST_PB_EQ_12	Average CA Visits per Beneficiary (Part B = 12)	Average number of carrier/physician visits per beneficiary for beneficiaries enrolled in Medicare Part B for 12 months in the calendar year	BASF	PHSVST
AVE_OP_VST_PB_EQ_12	Average OP Visits per Beneficiary (Part B = 12)	Average number of outpatient visits per beneficiary for beneficiaries enrolled in Medicare Part B for 12 months in the calendar year	BASF	OPVST
BENE_COUNT_PC_LT_12	Count of Beneficiaries (Part C < 12)	Count of beneficiaries enrolled in Medicare Part C (HMO) for at least 1 month but less than 12 months in the calendar year	BSF	Computed
AVE_MO_EN_PC_LT_12	Average Months of Enrollment (Part C < 12)	Average months of enrollment for beneficiaries enrolled in Medicare Part C (HMO) for at least 1 month but less than 12 months in the calendar year	BSF	BENE_HMO_CVRAGE_TOT_MONS

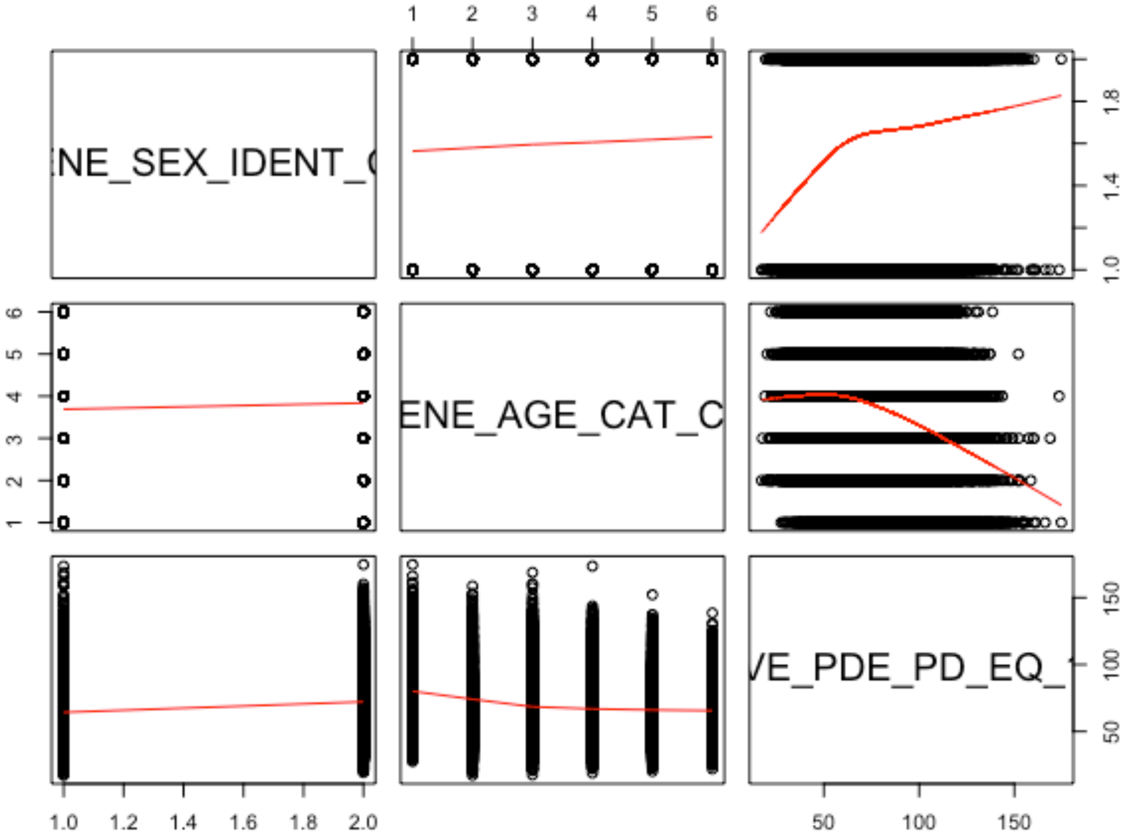
Variable Name	Short Name	Long Name / Description	Source File	Source Variable
BENE_COUNT_PC_EQ_12	Count of Beneficiaries (Part C = 12)	Count of beneficiaries enrolled in Medicare Part C (HMO) for than 12 months in the calendar year	BSF	Computed
BENE_COUNT_PD_LT_12	Count of Beneficiaries (Part D < 12)	Count of beneficiaries enrolled in Medicare Part D for at least 1 month but less than 12 months in the calendar year	BSF	Computed
AVE_MO_EN_PD_LT_12	Average Months of Enrollment (Part D < 12)	Average months of enrollment for beneficiaries enrolled in Medicare Part D for at least 1 month but less than 12 months in the calendar year	BSF	PLAN_CVRG_MOS_NUM
AVE_PDE_CST_PD_LT_12	Average Drug Cost per Beneficiary (Part D < 12)	Average prescription drug cost per beneficiary for beneficiaries enrolled in Medicare Part D for at least 1 month but less than 12 months in the calendar year	PDE	TOT_RX_CST_AMT
AVE_PDE_PD_LT_12	Average Prescriptions per Beneficiary (Part D < 12)	Average number of prescriptions per beneficiary for beneficiaries enrolled in Medicare Part D for at least 1 month but less than 12 months in the calendar year	PDE	PDE files
BENE_COUNT_PD_EQ_12	Count of Beneficiaries (Part D = 12)	Count of beneficiaries enrolled in Medicare Part D for 12 months in the calendar year	BSF	Computed
AVE_PDE_CST_PD_EQ_12	Average Drug Cost per Beneficiary (Part D = 12)	Average prescription drug cost per beneficiary for beneficiaries enrolled in Medicare Part D for 12 months in the calendar year	PDE	TOT_RX_CST_AMT
AVE_PDE_PD_EQ_12	Average Prescriptions per Beneficiary (Part D = 12)	Average number of prescriptions per beneficiary for beneficiaries enrolled in Medicare Part D for 12 months in the calendar year	PDE	Computed

BSF: Beneficiary Summary File
BASF: Beneficiary Annual Summary File
PDE: Prescription Drug Events

Appendix B

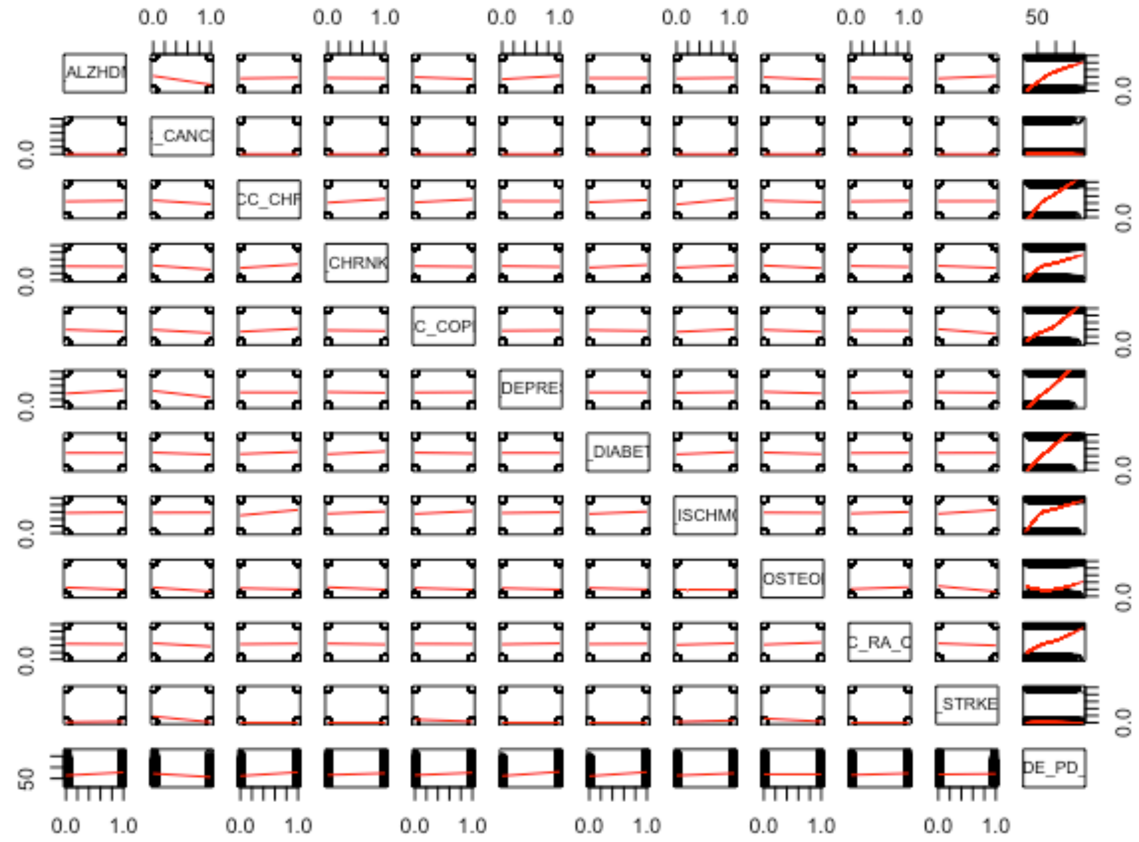
Scatterplot of Demographics

plot(PDP_2010_WD[c(1:2,33)],panel=panel.smooth)



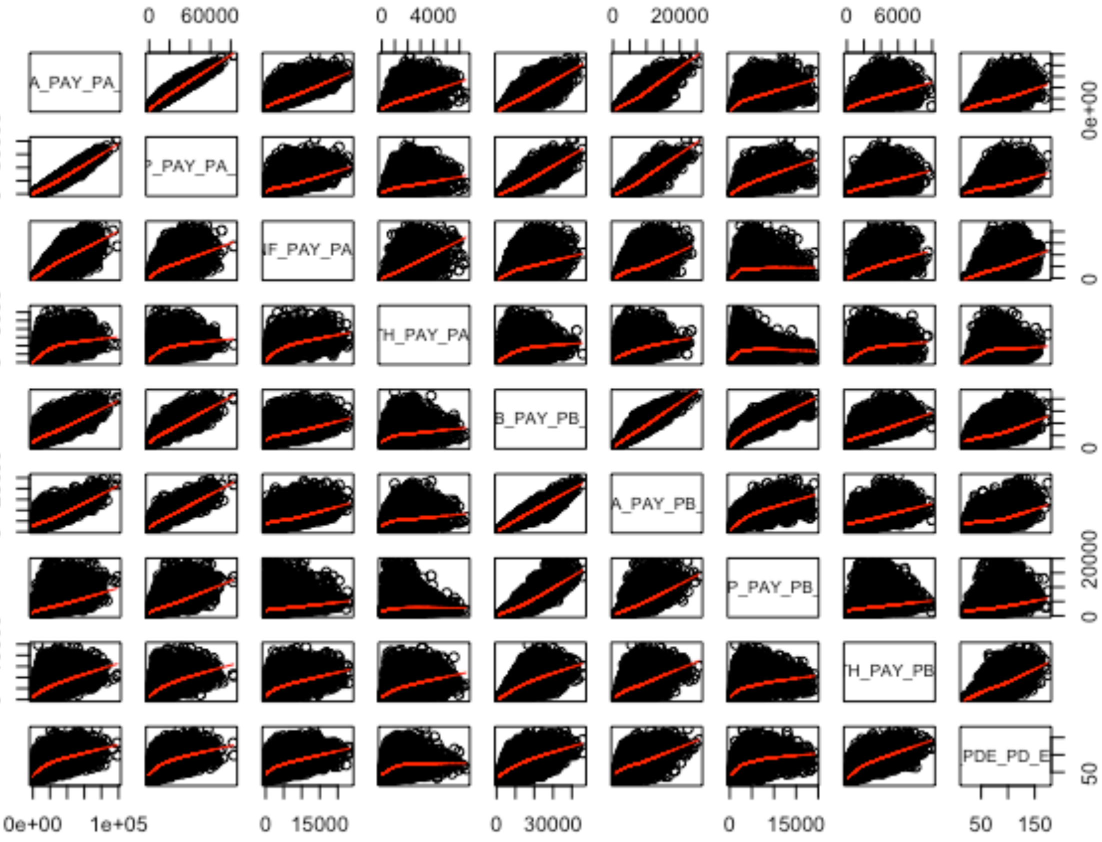
Scatterplot of Comorbidities

```
plot(PDP_2010_WD[c(1:2,33)],panel=panel.smooth)
```



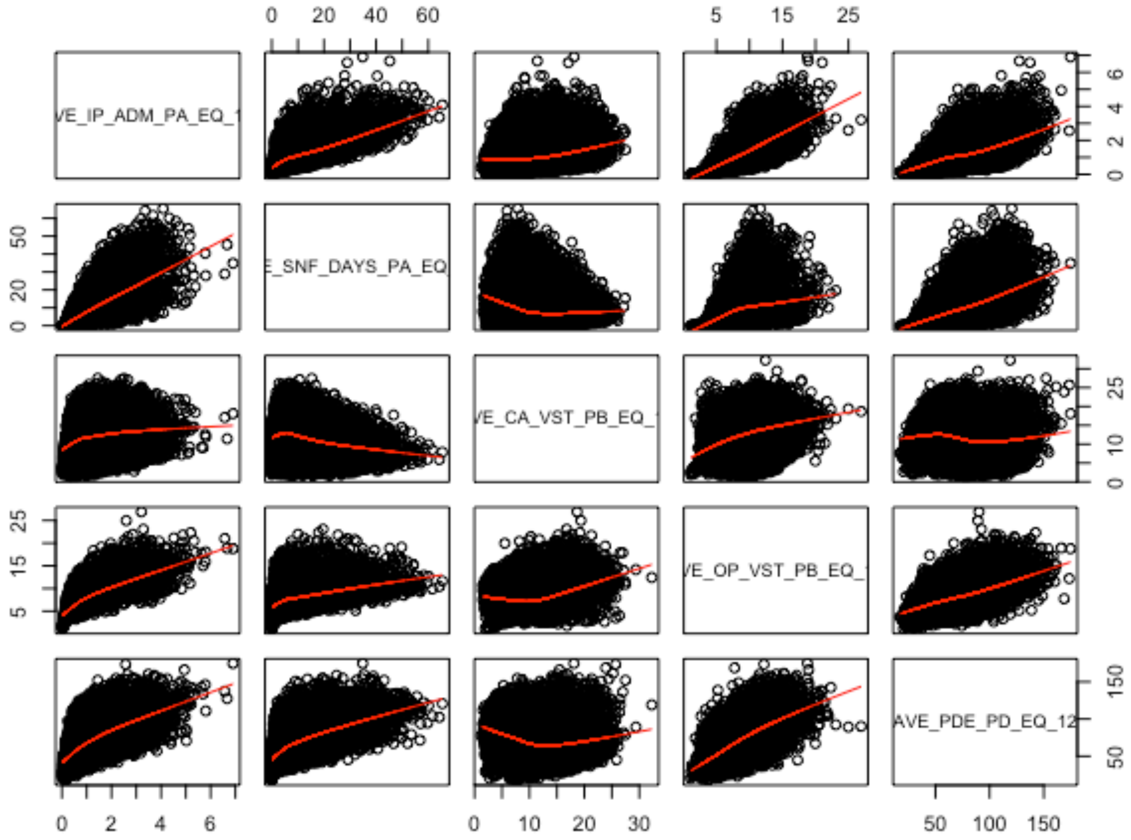
Scatterplot of Payments

```
plot(PDP_2010_WD[c("AVE_PA_PAY_PA_EQ_12","AVE_IP_PAY_PA_EQ_12","AVE_SNF_PAY_PA_EQ_12","AVE_OTH_PAY_PA_EQ_12","AVE_PB_PAY_PB_EQ_12","AVE_CA_PAY_PB_EQ_12","AVE_OP_PAY_PB_EQ_12","AVE_OTH_PAY_PB_EQ_12","AVE_PDE_PD_EQ_12")],panel=panel.smooth)
```



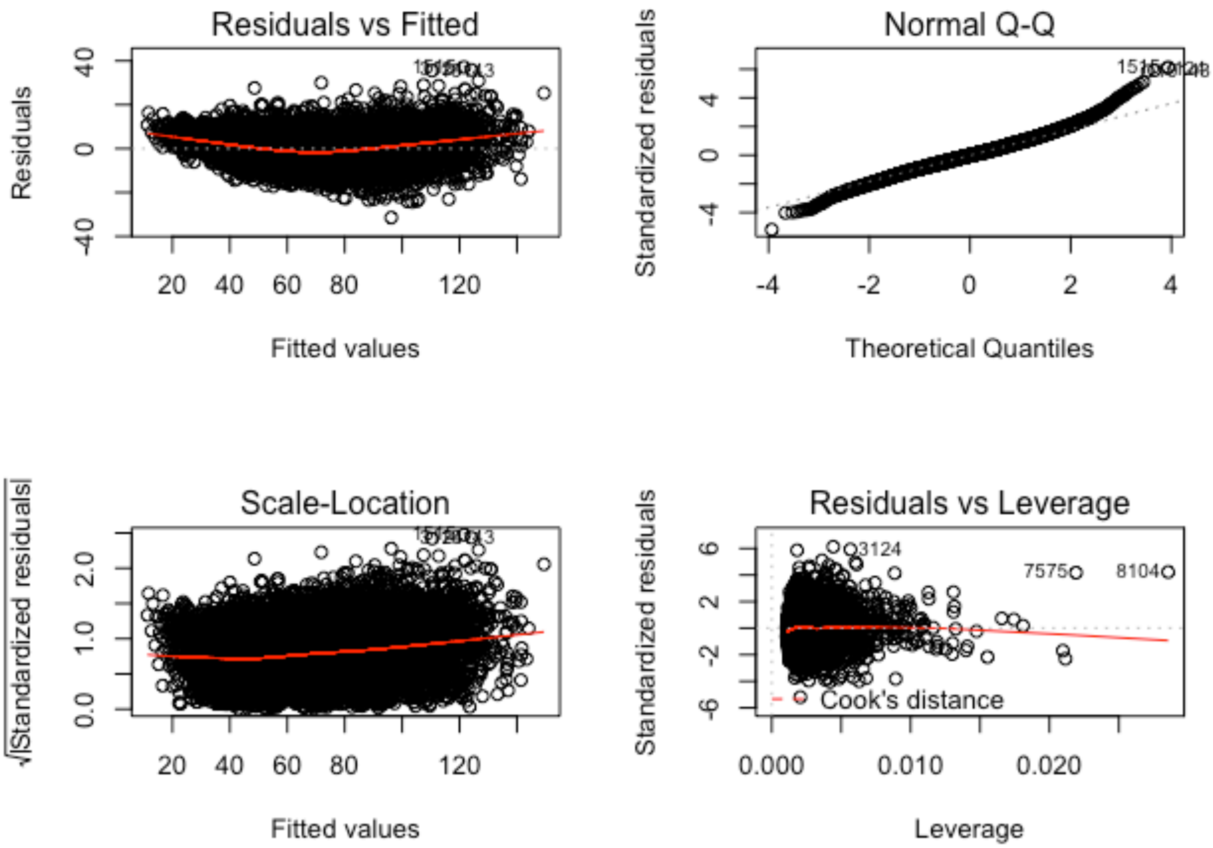
Scatterplot of Services

plot(PDP_2010_WD[c("AVE_IP_ADM_PA_EQ_12","AVE_SNF_DAYS_PA_EQ_12","AVE_CA_VST_PB_EQ_12","AVE_OP_VST_PB_EQ_12","AVE_PDE_PD_EQ_12")],panel=panel.smooth)



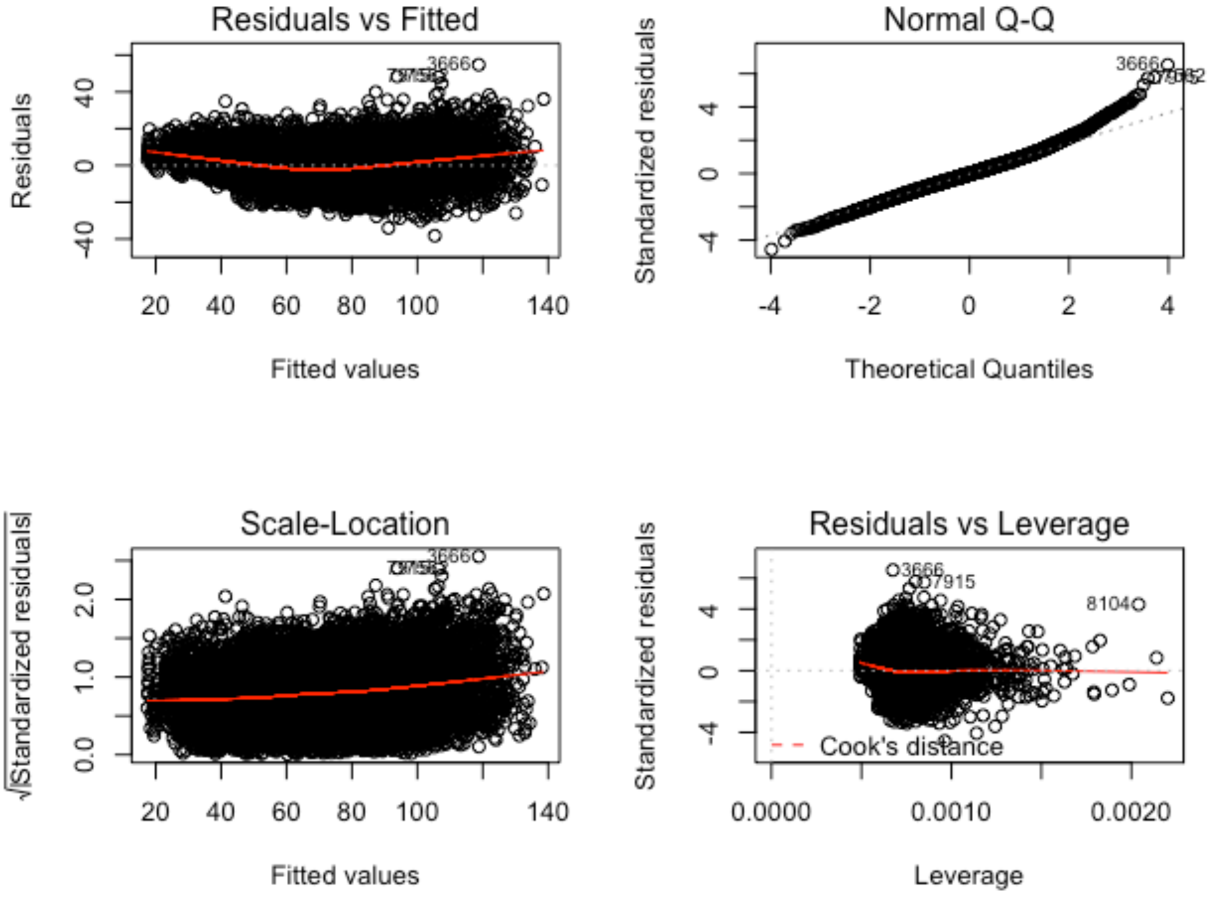
Appendix C

Residual Plots for Model 1



```
lm(formula = AVE_PDE_PD_EQ_12 ~ BENE_SEX_IDENT_CD + BENE_AGE_CAT_CD +
    CC_ALZHDMTA + CC_CANCER + CC_CHF + CC_CHRNKIDN + CC_COPD +
    CC_DEPRESSN + CC_DIABETES + CC_ISCHMCHT + CC_OSTEOPRS + CC_RA_OA +
    CC_STRKETIA + CC_2_OR_MORE + DUAL_STUS + AVE_PA_PAY_PA_EQ_12 +
    AVE_IP_PAY_PA_EQ_12 + AVE_SNF_PAY_PA_EQ_12 + AVE_IP_ADM_PA_EQ_12 +
    AVE_SNF_DAYS_PA_EQ_12 + AVE_PB_PAY_PB_EQ_12 + AVE_CA_PAY_PB_EQ_12 +
    AVE_OP_PAY_PB_EQ_12 + AVE_CA_VST_PB_EQ_12 + AVE_OP_VST_PB_EQ_12,
    data = PDP_2010)
Coefficients:
          (Intercept)     BENE_SEX_IDENT_CD      BENE_AGE_CAT_CD          CC_ALZHDMTA            CC_CANCER
           13.2307969             4.5439760           -1.3829243            9.3696783           -3.3806380
               CC_CHF           CC_CHRNKIDN              CC_COPD          CC_DEPRESSN          CC_DIABETES
           10.0640643             4.7786094            6.2507484           10.3848339           12.4138907
          CC_ISCHMCHT           CC_OSTEOPRS             CC_RA_OA          CC_STRKETIA         CC_2_OR_MORE
            5.1952364             1.5096831            3.7967731            1.8234542            2.3367692
            DUAL_STUS   AVE_PA_PAY_PA_EQ_12  AVE_IP_PAY_PA_EQ_12 AVE_SNF_PAY_PA_EQ_12  AVE_IP_ADM_PA_EQ_12
           20.7534325            -0.0014230            0.0009303           -0.0022327            5.3691718
AVE_SNF_DAYS_PA_EQ_12   AVE_PB_PAY_PB_EQ_12  AVE_CA_PAY_PB_EQ_12
            1.2436299             0.0022586           -0.0019519
  AVE_OP_PAY_PB_EQ_12   AVE_CA_VST_PB_EQ_12  AVE_OP_VST_PB_EQ_12
           -0.0036352            -0.3272736            2.2495017
```

Model 2



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            15.7376     0.2344   67.14   <2e-16 ***
CC_DEPRESSN            13.8080     0.1646   83.87   <2e-16 ***
CC_DIABETES            14.9320     0.1409  105.96   <2e-16 ***
CC_CHF                 10.1062     0.1649   61.30   <2e-16 ***
CC_ALZHDMTA             7.5730     0.1474   51.38   <2e-16 ***
CC_COPD                 6.5127     0.1728   37.68   <2e-16 ***
CC_RA_OA                5.9816     0.1403   42.64   <2e-16 ***
CC_CHRNKIDN             2.3043     0.1828   12.60   <2e-16 ***
CC_ISCHMCHT            3.2038     0.1578   20.30   <2e-16 ***
sqrt(AVE_IP_ADM_PA_EQ_12) 12.0021 0.3764   31.89   <2e-16 ***
DUAL_STUS              26.9456     0.1404  191.88   <2e-16 ***
```

Appendix D

R Code

```
#Needed to import csv in format ASCII
Chronic_Conditions_PUF_2010 <-
read.csv("~/Dropbox/2016_Fall/BS845/Project/2010_ChronicConditions_PUF/Chronic_Conditions_PUF_201
0.csv", encoding="ASCII")
attach(Chronic_Conditions_PUF_2010)
#Creating a subset of the original Medicare data to only show demographics, comorbidities, and
full year counts.
WholeYearCC_2010<-Chronic_Conditions_PUF_2010[,c(1:15,24:30,39:45,48,53:55)]
detach(Chronic_Conditions_PUF_2010);
attach(WholeYearCC_2010)
str(WholeYearCC_2010)
summary(WholeYearCC_2010)

#BENE_COUNT_PD_EQ_12, AVE_PDE_CST_PD_EQ_12, AVE_PDE_PD_EQ_12 all have 4537 NA's.
#What are these rows like?  Can/should they be omitted?
PDNA<-subset(WholeYearCC_2010,is.na(AVE_PDE_PD_EQ_12))
summary(PDNA)

#These NA rows are for people who don't have Part D at all it appears (or Part C as it happens)
#This hits both genders, and all age categories
#Need to remove NA from WholeYearCC_2010
WholeYearCC_2010_WD<-subset(WholeYearCC_2010,AVE_PDE_PD_EQ_12>0)
summary(WholeYearCC_2010_WD)

#Several comorbidities have 767 NA's out of 17466.  Checking to see if these are all suppressed
in the same rows
ComorbidNA<-subset(WholeYearCC_2010_WD,is.na(CC_ALZHDMTA))
summary(ComorbidNA)

#Yep - all 767 NA/suppressed CC_ALZHDMTA are also NA for CC_COPD, CC_DEPRESSN, CC_OSTEOPRS,
CC_STRKETIA
#They're also NA for BENE_COUNT_PC_EQ_12, but I think this is unrelated.
#Removing these rows.  Conveniently, this gets rid of all NA comorbidity values.
WholeYearCC_2010_WD2<-subset(WholeYearCC_2010_WD,!is.na(CC_ALZHDMTA))
summary(WholeYearCC_2010_WD2)

#Think data's clean now!  That was surprisingly painless (I hope).
#Giving working data set an easier name and cleaning up a little.
PDP_2010<-WholeYearCC_2010_WD2
summary(PDP_2010)
rm(ComorbidNA,PDNA,WholeYearCC_2010_WD2,WholeYearCC_2010_WD)
detach()

#Saving this file to my project folder to keep a version untouched.
```

```
str(PDP_2010)
attach(PDP_2010)
write.table(PDP_2010, file="PDP_2010.csv",quote=F, sep=",", na="", row.names=F)

hist(PDP_2010$AVE_PDE_PD_EQ_12, main="Average Prescriptions Filled per Beneficiary, 2010\nFull
Year Part D Only", xlab="Average Prescriptions by Category")


#let's run a quick and dirty poisson to see what happens:

modp1<-glm(AVE_PDE_PD_EQ_12~., family=poisson, data=PDP_2010)
summary(modp1)
#Hm.  glm isn't working with poisson because AVE_PDE_PD_EQ_12 is an average, not an integer.

#Taking a moment to understand the number of full year beneficiaries
#To get the mean and sd for the poisson, I'll need to take a weighted average.
#First, creating a new column with total RX = full year PDP beneficiaries * avg Rx filled for
year

PDP_2010_WD<-PDP_2010
PDP_2010_WD$TotRx<- (BENE_COUNT_PD_EQ_12 * AVE_PDE_PD_EQ_12)

#Now dividing by the total number of beneficiaries to get pop mean of 39.55605
PopAvgPDE<-sum(PDP_2010_WD$TotRx)/sum(PDP_2010_WD$BENE_COUNT_PD_EQ_12)


#Ok...  now what?
#Should I do something with the equation p(Y=k)=(e^-u*u^k)/k! where k=BENE_COUNT_PD_EQ_12?
#Should I check predictor varaibles for colinearity?
#Copying my 11/9 HW - got a 9.9/10 for the poisson I built there.  Output in plots.docx

anova(modp1)
drop1(modp1,test="Chisq")

#The output is telling me to stop using benefit payment amounts for Part A, Part B svcs, and I
agree this is a good idea.
#All of these have Likelihood Ratio Test values of 0, which is I think what happens when R rounds
from a really really low number (aka bad)
#Building a second model excluding these.
modp2<-glm(AVE_PDE_PD_EQ_12~BENE_SEX_IDENT_CD+BENE_AGE_CAT_CD

+CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+CC_OSTEOPRS
+CC_RA_OA+CC_STRKETIA
           +CC_2_OR_MORE+DUAL_STUS+BENE_COUNT_PA_EQ_12+AVE_IP_ADM_PA_EQ_12+AVE_SNF_DAYS_PA_EQ_12

+BENE_COUNT_PB_EQ_12+AVE_CA_VST_PB_EQ_12+AVE_OP_VST_PB_EQ_12+BENE_COUNT_PC_EQ_12+BENE_COUNT_PD_EQ
_12+AVE_PDE_CST_PD_EQ_12
           , family=poisson
```

```
          , data=PDP_2010);
summary(modp2)

drop1(modp2,test="Chisq")

#This is still a terrible model.  Can I do forward stepwise?
#She used stepAIC in class, part of MASS package.

stepAIC(glm(AVE_PDE_PD_EQ_12~AVE_PDE_PD_EQ_12~BENE_SEX_IDENT_CD+BENE_AGE_CAT_CD

+CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+CC_OSTEOPRS
+CC_RA_OA+CC_STRKETIA

+CC_2_OR_MORE+DUAL_STUS+BENE_COUNT_PA_EQ_12+AVE_IP_ADM_PA_EQ_12+AVE_SNF_DAYS_PA_EQ_12

+BENE_COUNT_PB_EQ_12+AVE_CA_VST_PB_EQ_12+AVE_OP_VST_PB_EQ_12+BENE_COUNT_PC_EQ_12+BENE_
COUNT_PD_EQ_12+AVE_PDE_CST_PD_EQ_12
           , family=poisson, data=PDP_2010),direction = "both")


#AIC=Inf happens if using Poisson with non-integer values.

#Does Binomial work?
modp1<-glm(AVE_PDE_PD_EQ_12~., family=binomial, data=PDP_2010);
summary(modp1)
#No, because output needs to be between 0-1

#making matrix of scatter plots to bring to meeting (this will be ugly)
plot(PDP_2010,panel=panel.smooth)
#data too big - this would need to be broken out.  Not even sure if it's necessary...

#~~~~~~~~~~~~~~~~~Meeting with Prof~~~~~~~~~~~~~
#Rounded total RX col added
attach(PDP_2010)
PDP_2010_WD$TotRx<- (BENE_COUNT_PD_EQ_12 * AVE_PDE_PD_EQ_12)
PDP_2010_WD$TotRxRound<- round(PDP_2010_WD$TotRx, digits=0)
detach()

#model built with prof
mod_prof<-glm(TotRxRound~BENE_COUNT_PD_EQ_12+CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_COPD,
family=poisson, data=PDP_2010_WD)
#~~~~~~~~~~~~~~~~Complete~~~~~~~~~~~~~~~~~~~~~~~~~

#Trying with comorbidities only (with bene count)
```

```
mod_comorbid<-
glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKI
DN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+CC_OSTEOPRS+CC_RA_OA+CC_STRKETIA
              , family=poisson, data=PDP_2010_WD)
#using the log() of Bene count gives a better model than not
#Comorbidities all highly significant with and wihtout log(bene), but comorbid
coefficients go in the right direction (i.e. positively corrleate) once the log is
taken.
#Not great models yet, but at least they're working now.

mod_offsettest<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+

CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+C
C_OSTEOPRS+CC_RA_OA+CC_STRKETIA
                    ,family=poisson,data=PDP_2010_WD)
#This worked really well!!!!  Coefficients in the right directions!!!!

#Ok, making a model with all covariates and seeing what happens.
#prof says continuous variables need log transformation with poisson.
#I'm not sure about adding bene counts for Parts A, B, and C, no matter how
predictive.
#I'm going to leave those out.  They correlate with Part D bene count which is
definitely going in there already.

mod_poisson1<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+
                  BENE_SEX_IDENT_CD+log(BENE_AGE_CAT_CD)+

CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+C
C_OSTEOPRS+CC_RA_OA+CC_STRKETIA+
                  CC_2_OR_MORE+DUAL_STUS+

log(AVE_PA_PAY_PA_EQ_12)+log(AVE_IP_PAY_PA_EQ_12)+log(AVE_SNF_PAY_PA_EQ_12)+log(AVE_OT
H_PAY_PA_EQ_12)+log(AVE_IP_ADM_PA_EQ_12)+log(AVE_SNF_DAYS_PA_EQ_12)+

log(AVE_PB_PAY_PB_EQ_12)+log(AVE_CA_PAY_PB_EQ_12)+log(AVE_OP_PAY_PB_EQ_12)+log(AVE_OTH
_PAY_PB_EQ_12)+log(AVE_CA_VST_PB_EQ_12)+log(AVE_OP_VST_PB_EQ_12)+
                  log(AVE_PDE_CST_PD_EQ_12)
                  ,family=poisson,data=PDP_2010_WD)

#Ok, now let's see about using drop1() to make this better.
drop1(mod_poisson1,test="Chisq")

#This didn't drop anything.
```

```
#Yeah...  I'm getting rid of cost of drugs - this corrlates *too* well.
mod_poisson2<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+
                    BENE_SEX_IDENT_CD+log(BENE_AGE_CAT_CD)+

CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+C
C_OSTEOPRS+CC_RA_OA+CC_STRKETIA+
                    CC_2_OR_MORE+DUAL_STUS+

log(AVE_PA_PAY_PA_EQ_12)+log(AVE_IP_PAY_PA_EQ_12)+log(AVE_SNF_PAY_PA_EQ_12)+log(AVE_OT
H_PAY_PA_EQ_12)+log(AVE_IP_ADM_PA_EQ_12)+log(AVE_SNF_DAYS_PA_EQ_12)+

log(AVE_PB_PAY_PB_EQ_12)+log(AVE_CA_PAY_PB_EQ_12)+log(AVE_OP_PAY_PB_EQ_12)+log(AVE_OTH
_PAY_PB_EQ_12)+log(AVE_CA_VST_PB_EQ_12)+log(AVE_OP_VST_PB_EQ_12)
                  ,family=poisson,data=PDP_2010_WD)

#This made fit worse, not sure if that's a bad thing.
#trying drop1() again
#All variables still *highly* correlated.  Deviance even worse than before, no drops
improve AIC.
#I want to compare pay vs. visit/days/admission.  Should not have both of these I
don't think.
#Going to try the model with just pay with model with just visits.
mod_poisson_justpay<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+

log(AVE_PA_PAY_PA_EQ_12)+log(AVE_IP_PAY_PA_EQ_12)+log(AVE_SNF_PAY_PA_EQ_12)+log(AVE_OT
H_PAY_PA_EQ_12)+

log(AVE_PB_PAY_PB_EQ_12)+log(AVE_CA_PAY_PB_EQ_12)+log(AVE_OP_PAY_PB_EQ_12)+log(AVE_OTH
_PAY_PB_EQ_12)
                  ,family=poisson,data=PDP_2010_WD);

mod_poisson_justvisit<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+
                  log(AVE_IP_ADM_PA_EQ_12)+log(AVE_SNF_DAYS_PA_EQ_12)+
                  log(AVE_CA_VST_PB_EQ_12)+log(AVE_OP_VST_PB_EQ_12)
                  ,family=poisson,data=PDP_2010_WD)

#mod_poisson_justpay has lower deviance and lower AIC than mod_poisson_justvisit,
although it also has more variables.
#mod_comorbid looks very similar, with quality somewhere in between (although all 3
models are terrible fits)

mod_poisson_justdemographics_nolog<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+
                   BENE_SEX_IDENT_CD+BENE_AGE_CAT_CD
                   ,family=poisson,data=PDP_2010_WD)

#still no good.
#Let's check out this overdispersion thing graphically
scatter.smooth(log(fitted(mod_poisson2)),log((PDP_2010_WD$TotRxRound-
fitted(mod_poisson2))^2),xlab=expression(hat(mu))
               , ylab=expression(sigma^2==(y-hat(mu))^2))
abline(0,1,lty=2)

#No luck, running a pearson statistic to measure overdisperison.
phi<- sum(resid(mod_poisson2, type = "pearson")^2)/df.residual(mod_poisson2);
phi_justdemo<- sum(resid(mod_poisson_justdemographics_nolog, type =
"pearson")^2)/df.residual(mod_poisson_justdemographics_nolog);
phi_justpay<- sum(resid(mod_poisson_justpay, type =
"pearson")^2)/df.residual(mod_poisson_justpay);
phi_justvisit<- sum(resid(mod_poisson_justvisit, type =
"pearson")^2)/df.residual(mod_poisson_justvisit)
phi_justvisit<- sum(resid(mod_poisson_justvisit, type =
"pearson")^2)/df.residual(mod_poisson)

#Why isn't this working?  Oh, fitted values for glm exclude NA by default.  This
causes error in TotRxRound-fitted()
#length(PDP_2010_WD$TotRxRound)=16699, length(fitted(mod_poisson2)) = 12043

#Right!  I can use StepAIC now that I'm getting AICs!
#then also consider doing add1() to complement drop1()
mod_poisson3<-stepAIC(glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+
                            BENE_SEX_IDENT_CD+log(BENE_AGE_CAT_CD)+

CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+C
C_OSTEOPRS+CC_RA_OA+CC_STRKETIA+
                            CC_2_OR_MORE+DUAL_STUS+

log(AVE_PA_PAY_PA_EQ_12)+log(AVE_IP_PAY_PA_EQ_12)+log(AVE_SNF_PAY_PA_EQ_12)+log(AVE_OT
H_PAY_PA_EQ_12)+log(AVE_IP_ADM_PA_EQ_12)+log(AVE_SNF_DAYS_PA_EQ_12)+

log(AVE_PB_PAY_PB_EQ_12)+log(AVE_CA_PAY_PB_EQ_12)+log(AVE_OP_PAY_PB_EQ_12)+log(AVE_OTH
_PAY_PB_EQ_12)+log(AVE_CA_VST_PB_EQ_12)+log(AVE_OP_VST_PB_EQ_12)
                          ,family=poisson,data=PDP_2010_WD),direction="forward")

#Maybe I should create a column with the sum of comorbidities...
attach(PDP_2010_WD);
```

```
PDP_2010_WD$TotComorbid<-
(CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+
CC_OSTEOPRS+CC_RA_OA+CC_STRKETIA+
                             CC_2_OR_MORE+DUAL_STUS);
detach()

mod_lm_totcomorbid<-lm((AVE_PDE_PD_EQ_12~TotComorbid), data=PDP_2010_WD)
#Oh god, liner model with tot number of comorbidities done on a larf is *much* better.
Much much much.
#Let's try it with the individual comorbidities
mod_lm_comorbid<-
lm(AVE_PDE_PD_EQ_12~CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DI
ABETES+CC_ISCHMCHT+CC_OSTEOPRS+CC_RA_OA+CC_STRKETIA
                    ,data=PDP_2010_WD)

#Cancer is negatively correlated, the others are positively correlated.
#The individual comorbidities which seem to lead to the most prescriptions are
Depression, Diabetes, CHF, COPD, and Alzheimers.
#The fewest prescriptions are Cancer (-), Osteoporosos, and Strketia


summary(lm(AVE_PDE_PD_EQ_12~CC_ALZHDMTA+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIAB
ETES+CC_ISCHMCHT+CC_OSTEOPRS+CC_RA_OA+CC_STRKETIA
                    ,data=PDP_2010_WD))



#I do like the data/coefficients I get from having the comorbidities broken out like
this.

#~~~~~~~~~~~~~~~~
#Ok, should I do anything else with poisson before I forget what I was doing?
#While looking for how to do poisson transformations...
#Found some info about quasi poisson
mod_quasipoisson2<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+
                    BENE_SEX_IDENT_CD+log(BENE_AGE_CAT_CD)+

CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+C
C_OSTEOPRS+CC_RA_OA+CC_STRKETIA+
                    CC_2_OR_MORE+DUAL_STUS+

log(AVE_PA_PAY_PA_EQ_12)+log(AVE_IP_PAY_PA_EQ_12)+log(AVE_SNF_PAY_PA_EQ_12)+log(AVE_OT
H_PAY_PA_EQ_12)+log(AVE_IP_ADM_PA_EQ_12)+log(AVE_SNF_DAYS_PA_EQ_12)+
```

```
log(AVE_PB_PAY_PB_EQ_12)+log(AVE_CA_PAY_PB_EQ_12)+log(AVE_OP_PAY_PB_EQ_12)+log(AVE_OTH
_PAY_PB_EQ_12)+log(AVE_CA_VST_PB_EQ_12)+log(AVE_OP_VST_PB_EQ_12)
                  ,family=quasipoisson(link="log"),data=PDP_2010_WD)

mod_quasipoisson_justpay<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+

log(AVE_PA_PAY_PA_EQ_12)+log(AVE_IP_PAY_PA_EQ_12)+log(AVE_SNF_PAY_PA_EQ_12)+log(AVE_OT
H_PAY_PA_EQ_12)+

log(AVE_PB_PAY_PB_EQ_12)+log(AVE_CA_PAY_PB_EQ_12)+log(AVE_OP_PAY_PB_EQ_12)+log(AVE_OTH
_PAY_PB_EQ_12)
                         ,family=quasipoisson,data=PDP_2010_WD);

mod_quasipoisson_justvisit<-glm(TotRxRound~offset(log(BENE_COUNT_PD_EQ_12))+
                             log(AVE_IP_ADM_PA_EQ_12)+log(AVE_SNF_DAYS_PA_EQ_12)+
                             log(AVE_CA_VST_PB_EQ_12)+log(AVE_OP_VST_PB_EQ_12)
                           ,family=quasipoisson,data=PDP_2010_WD)



#Yeah, this is totally linear - giving up on poisson and finishing the analysis.
#Running scatterplots against AVE_PDE_PD_EQ_12

#demographics - being young and female leads to more Rx
plot(PDP_2010_WD[c(1:2,33)],panel=panel.smooth)

#comorbidities - all increase Rx except for Cancer and Stroke. Osteoporosis only a
little bit
#Cancer negatively correlates with alzheimers, stroke, and depression (surprising);
CHF positively correlates wiht IschmCHT.
plot(PDP_2010_WD[c(3:13,33)],panel=panel.smooth)

#Other Medicare charge amounts- All positively correlate with Rx, some more htan
others.
#Also, they all strongly correlate with each other, some more than others.
plot(PDP_2010_WD[c("AVE_PA_PAY_PA_EQ_12","AVE_IP_PAY_PA_EQ_12","AVE_SNF_PAY_PA_EQ_12",
"AVE_OTH_PAY_PA_EQ_12"
,"AVE_PB_PAY_PB_EQ_12","AVE_CA_PAY_PB_EQ_12","AVE_OP_PAY_PB_EQ_12","AVE_OTH_PAY_PB_EQ_
12","AVE_PDE_PD_EQ_12")],panel=panel.smooth)

#Other Medicare visits - Average number of carrier/physician visits doesn't correlate
with anything, really, including Rx count.
```

```
#Everything else highly correlates.  Especially IP admission with SNF days, but that
makes sense.
plot(PDP_2010_WD[c("AVE_IP_ADM_PA_EQ_12","AVE_SNF_DAYS_PA_EQ_12","AVE_CA_VST_PB_EQ_12"
,"AVE_OP_VST_PB_EQ_12","AVE_PDE_PD_EQ_12")],panel=panel.smooth)

plot(PDP_2010_WD[c("CC_2_OR_MORE","DUAL_STUS","AVE_PDE_PD_EQ_12")],panel=panel.smooth)


#Linear model all of the everything
mod_linear2<-lm(AVE_PDE_PD_EQ_12~BENE_SEX_IDENT_CD+BENE_AGE_CAT_CD+

CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+C
C_OSTEOPRS+CC_RA_OA+CC_STRKETIA+
                    CC_2_OR_MORE+DUAL_STUS+

AVE_PA_PAY_PA_EQ_12+AVE_IP_PAY_PA_EQ_12+AVE_SNF_PAY_PA_EQ_12+AVE_OTH_PAY_PA_EQ_12+AVE_
IP_ADM_PA_EQ_12+AVE_SNF_DAYS_PA_EQ_12+

AVE_PB_PAY_PB_EQ_12+AVE_CA_PAY_PB_EQ_12+AVE_OP_PAY_PB_EQ_12+AVE_OTH_PAY_PB_EQ_12+AVE_C
A_VST_PB_EQ_12+AVE_OP_VST_PB_EQ_12,
                  data=PDP_2010_WD)

#Running stepAIC on it

mod_linear2_step<-stepAIC(lm(AVE_PDE_PD_EQ_12~.-AVE_PDE_CST_PD_EQ_12-
BENE_COUNT_PD_EQ_12-BENE_COUNT_PC_EQ_12-BENE_COUNT_PB_EQ_12-BENE_COUNT_PA_EQ_12,
data=PDP_2010),direction="both",na.rm=T)
mod_linear2_step$anova
summary(mod_linear2_step)

mod_linear2_step2<-stepAIC(lm(AVE_PDE_PD_EQ_12~BENE_SEX_IDENT_CD+BENE_AGE_CAT_CD+

CC_ALZHDMTA+CC_CANCER+CC_CHF+CC_CHRNKIDN+CC_COPD+CC_DEPRESSN+CC_DIABETES+CC_ISCHMCHT+C
C_OSTEOPRS+CC_RA_OA+CC_STRKETIA+
                  CC_2_OR_MORE+DUAL_STUS+

AVE_PA_PAY_PA_EQ_12+AVE_IP_PAY_PA_EQ_12+AVE_SNF_PAY_PA_EQ_12+AVE_OTH_PAY_PA_EQ_12+AVE_
IP_ADM_PA_EQ_12+AVE_SNF_DAYS_PA_EQ_12+

AVE_PB_PAY_PB_EQ_12+AVE_CA_PAY_PB_EQ_12+AVE_OP_PAY_PB_EQ_12+AVE_OTH_PAY_PB_EQ_12+AVE_C
A_VST_PB_EQ_12+AVE_OP_VST_PB_EQ_12,
                data=PDP_2010),direction = "both")
summary(mod_linear2_step2)
```

```
mod_linear2_step2$anova

par(mfrow=c(2,2));
plot(mod_linear2_step2)

#apparently there's a command called getDeltaRsquare() which I forgot about in
rockchalk packate
getDeltaRsquare(mod_linear2_step2)
getDeltaRsquare(mod_lm_comorbid)
#everything is more influential on R2 in comorbid model.  I think linear2_step2 has
too many predictors.

bptest(mod_linear2_step2)

#Getting rid of unwanted predictors.
#Unfortunately, -BENE_COUNT_PC_EQ_12 has null values and is breaking things when I try
to take it out.
mod_linear2_step3<-stepAIC(lm(AVE_PDE_PD_EQ_12~. -AVE_PDE_CST_PD_EQ_12 -
BENE_COUNT_PD_EQ_12 -BENE_COUNT_PB_EQ_12 -BENE_COUNT_PA_EQ_12,
                                data=PDP_2010, na.action=na.exclude),direction =
"both")



#Did my own forward stepwise, like this best of them due to R2=.8812 and Intercept
error = .2344
mod_lm_construct_best<-
lm(AVE_PDE_PD_EQ_12~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_RA_OA+CC_CHR
NKIDN+CC_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD)
summary(mod_lm_construct_best)
getDeltaRsquare(mod_lm_construct_best)
plot(mod_lm_construct_best)
#Testing heteroscedasticity
library("lmtest",
lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
bptest(mod_lm_construct_best)

plot(lm((AVE_PDE_PD_EQ_12)^2~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_RA_
OA+CC_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD))
#this one throws everything out of whack
```

```
plot(lm(sqrt(AVE_PDE_PD_EQ_12)~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_R
A_OA+CC_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD))
bptest(lm(sqrt(AVE_PDE_PD_EQ_12)~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC
_RA_OA+CC_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD))
#this ain't bad.  Still heteroscedastic

plot(lm(log(AVE_PDE_PD_EQ_12)~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_RA
_OA+CC_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD))
bptest(lm(log(AVE_PDE_PD_EQ_12)~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_
RA_OA+CC_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD))
#not bad eihter.  Still heteroscedastic.

plot(lm(exp(AVE_PDE_PD_EQ_12)~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_RA
_OA+CC_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD))
bptest(lm(exp(AVE_PDE_PD_EQ_12)~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_
RA_OA+CC_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD))

#Now this is intesting...  took the exponential of the outcome and there are some
strange effects
#especially due to obs 3666 (173.5) and 8104 (174.7)
#this following code is from Avery's lecture 5
mod_lm_construct_exp<-
lm(exp(AVE_PDE_PD_EQ_12)~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_RA_OA+C
C_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_WD)
mod_lm_construct_exp.dffits<-dffits(mod_lm_construct_exp)
mod_lm_construct_exp.hat<-hatvalues(mod_lm_construct_exp)
id.mod_lm_construct_exp.dffits<-
which(mod_lm_construct_exp.dffits>(2*sqrt((9+1)/16699)))
influence.measures(mod_lm_construct_best)

#Removed influential observations in Excel just to see if this is worth pursuing
#it isn't
mod_lm_construct_exp_nooutliers<-
lm(exp(AVE_PDE_PD_EQ_12)~CC_DEPRESSN+CC_DIABETES+CC_CHF+CC_ALZHDMTA+CC_COPD+CC_RA_OA+C
C_ISCHMCHT+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS, data=PDP_2010_NoOutliers)
mod_lm_construct_exp_nooutliers.dffits<-dffits(mod_lm_construct_exp_nooutliers)
mod_lm_construct_exp_nooutliers.hat<-hatvalues(mod_lm_construct_exp_nooutliers)
id.mod_lm_construct_exp_nooutliers.dffits<-
which(mod_lm_construct_exp_nooutliers.dffits>(2*sqrt((9+1)/16699)))
influence.measures(mod_lm_construct_best)


#Output variable is moderately skewed (0.5-1.0) (Class 6)
```

```
> skewness(AVE_PDE_PD_EQ_12)
[1] 0.5777349
> kurtosis(AVE_PDE_PD_EQ_12)
[1] -0.1198344

#Found a better way to test correlation than the graphs
All_Cor<-cor(PDP_2010_WD, use="complete.obs")
View(All_Cor)
write.table(All_Cor, file="All_Cor.csv",quote=F, sep=",", na="", row.names=T)


summary(mod_lm_construct_exp)
#This is not a good model.  R2=0.001823
mod_lm_construct_exp_Totcomorbid<-
lm(exp(AVE_PDE_PD_EQ_12)~TotComorbid+sqrt(AVE_IP_ADM_PA_EQ_12)+DUAL_STUS,
data=PDP_2010_WD)
summary(mod_lm_construct_exp_Totcomorbid)

#trying sandwich test with previous constructed 10-factor model.
coeftest(mod_lm_construct_best)
summary(mod_lm_construct_best)
```