Propensity Score Methods

Andrew Zieffler

Educational Psychology

University of Minnesota

Driven to Discover SM

Is there an Effect of Treatment?

The **method of comparison** is perhaps the most important idea in science: *To determine whether a treatment has an effect, compare what happens with and without the treatment.*

Confounding

Differences between the control group and the treatment group other than the treatment can be responsible for observed differences in outcome between the two groups

Confounding can:

- Hide a real effect
- Produce the **spurious** appearance of a treatment effect (when the real cause is due to a difference between the treatment and control groups other than the treatment)

If the treatment group predominantly contains individuals who would do well (or who would do poorly) whether or not they received treatment, we cannot separate the effect of the assignment from the effect of the treatment: Confounding still could be responsible for an apparent treatment effect, or could obscure a real treatment effect.

The *New York Times*, reported that the college entry test "coaching industry is playing on parental anxiety," arguing that **coaching does not improve test scores**.

Study design: Questionnaires sent to 1409 Harvard University freshmen in the fall of 1987. Of those surveyed, 69% said they had received no coaching, and 14% said they had received coaching. The verbal and mathematical SAT scores of the students were:

Average SAT Scores

	Not coached	Coached	Difference
Verbal	649	611	38
Mathematical	685	660	25

Method of comparison: The scores of the students who were coached were, on average, lower than those of the students who were not coached...*coaching does not help*.

Alternative explanation: The students who sought coaching were academically weaker, in general, than those who did not (after all, why did they seek help?)

In that case, we would *expect the students who sought coaching to do worse than those who did not*, unless coaching were **so** effective that it more than wiped out the natural difference.

The propensity to seek coaching is likely confounded with any effect of coaching...from these data we cannot say whether coaching is helpful.

How to Deal with Confounding

To prevent confounding, the treatment and control groups should be alike in *every regard that can affect the outcome, except for the treatment*

Confounding caused by individual differences can be reduced by using the method of comparison with **randomization to assign subjects to treatment or control**.

Randomized assignment *tends to balance differences* between the treatment and control groups, so that the overall difference between the outcomes for the two groups *can be attributed more reliably to the treatment* itself.

In observational studies, regression methods can help cope with confounding by the use of model-based evaluation and adjustment of included covariates.

To estimate treatment effects the outcome of interest is regressed on the covariates, including an indicator variable for treatment status and interactions between the treatment variable and each of the covariates.

A statistically significant coefficient of treatment or statistically significant coefficient of an interaction involving the treatment variable indicates a treatment effect.

Propensity Score Methods

Propensity score methods (PSM) can be used to *create groups* of treated and control units *that have similar characteristics* so that comparisons can be made within these matched groups

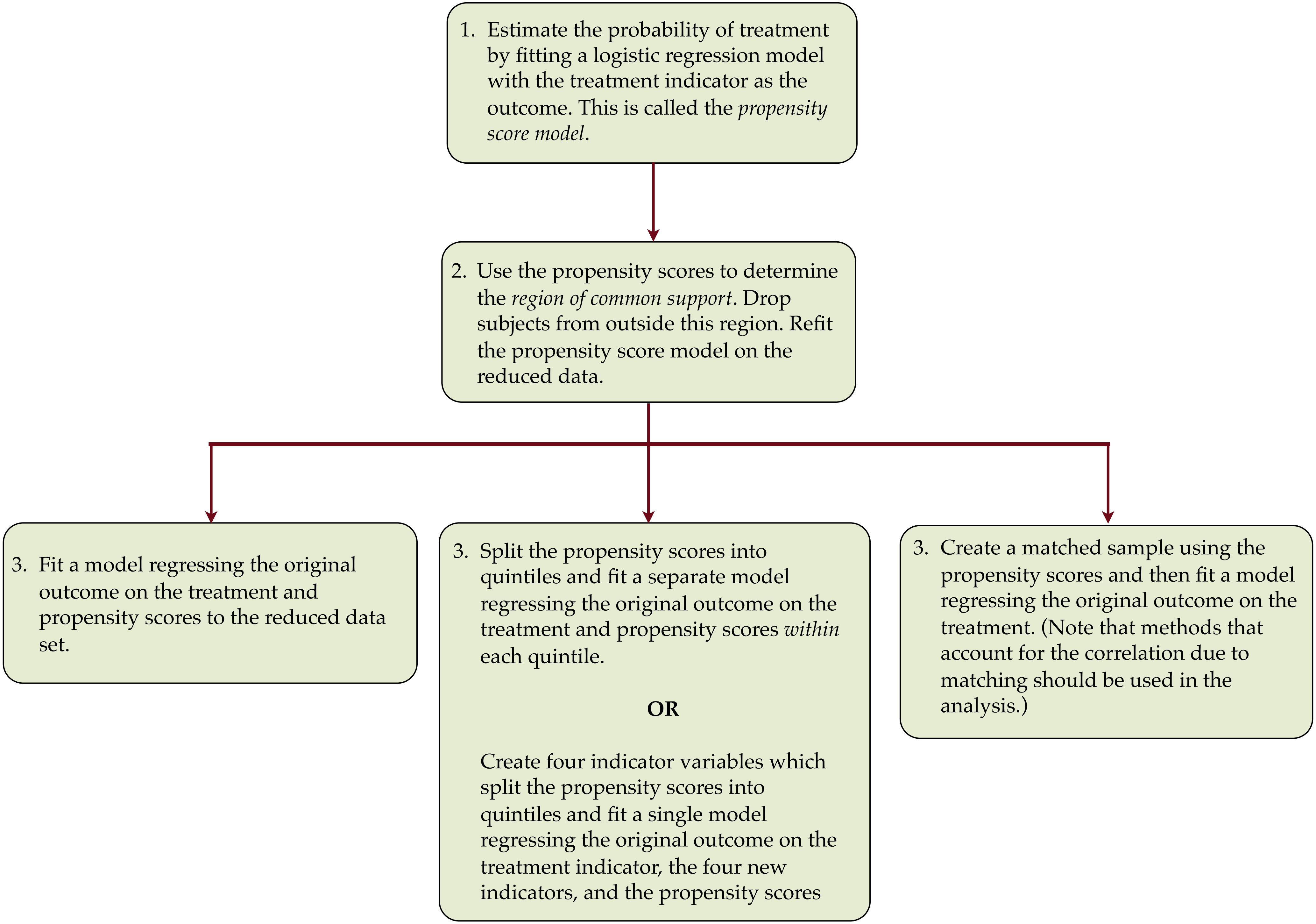
The propensity score is defined as the **conditional probability** of receiving treatment given a set of observed covariates.

$$P(Z = 1|\mathbf{X})$$

The propensity score is a **balancing score**, meaning that conditional on the propensity score the **distributions of the observed covariates** are independent of the assignment to treatment.

Propensity score methods, similar to regression, **cannot** control for unmeasured confounding

- Propensity score has direct scientific interest in studies that focus on **determinants of treatment initiation** or **persistence with treatment**. Consideration of the propensity score can broaden one's perspective to include barriers to treatment.
- Matching can be viewed as a principled approach to eliminate extreme observations that may be unduly influential and problematic in a multivariate analysis because of minimal covariate overlap
- Use of the propensity score provides an effective way to reduce the dimensionality of the covariates before modeling leading to improved estimation
- Consideration of the propensity score focuses on the real possibility that the effectiveness of a treatment may vary according to the strength of the indication for its use
- Propensity scores computed in one study (with more detailed covariate information) can sometimes be used in other studies to calibrate/correct for the measurement error



1. Estimate the probability of treatment by fitting a logistic regression model with the treatment indicator as the outcome. This is called the *propensity score model*.

2. Use the propensity scores to determine the *region of common support*. Drop subjects from outside this region. Refit the propensity score model on the reduced data.

3. Fit a model regressing the original outcome on the treatment and propensity scores to the reduced data set.

3. Split the propensity scores into quintiles and fit a separate model regressing the original outcome on the treatment and propensity scores *within* each quintile.

OR

Create four indicator variables which split the propensity scores into quintiles and fit a single model regressing the original outcome on the treatment indicator, the four new indicators, and the propensity scores

3. Create a matched sample using the propensity scores and then fit a model regressing the original outcome on the treatment. (Note that methods that account for the correlation due to matching should be used in the analysis.)

Reading and Examining the Data

```
# Read in the data
> glow = read.csv(file = "http://www.tc.umn.edu/~zief0002/Data/GLOW2.csv")

> head(glow)

  SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT      BMI PREMENO MOMFRAC
1      1       1     14         0  62   70.3    158 28.16055       0       0
2      2       4    284         0  65   87.1    160 34.02344       0       0
3      3       6    305         1  88   50.8    157 20.60936       0       1
4      4       6    309         0  82   62.1    160 24.25781       0       0
5      5       1     37         0  61   68.0    152 29.43213       0       0
6      6       5    299         1  67   68.0    161 26.23356       0       0

  ARMASSIST SMOKE RATERISK FRACSCORE FRACTURE BONEMED BONEMED_FU
1         0     0        2         1        0       0          0
2         0     0        2         2        0       0          0
3         1     0        1        11        0       0          0
4         0     0        1         5        0       0          0
5         0     0        2         1        0       0          0
6         0     1        2         4        0       0          0
```

BONEMED indicates whether or not a woman was taking one of 11 different bone medications at the time of enrollment in the study.

BONEMED_FU indicates whether or not a woman was taking one of 11 different bone medications at follow-up.

Since use is self-reported, we have no information about compliance or dosage

```
# Read in the data
> glow$BONETREAT = ifelse( (glow$BONEMED + glow$BONEMED_FU) == 2, 1, 0)

> head(glow)

  SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT      BMI PREMENO MOMFRAC
1      1       1     14         0  62   70.3    158 28.16055       0       0
2      2       4    284         0  65   87.1    160 34.02344       0       0
3      3       6    305         1  88   50.8    157 20.60936       0       1
4      4       6    309         0  82   62.1    160 24.25781       0       0
5      5       1     37         0  61   68.0    152 29.43213       0       0
6      6       5    299         1  67   68.0    161 26.23356       0       0

  ARMASSIST SMOKE RATERISK FRACSCORE FRACTURE BONEMED BONEMED_FU BONETREAT
1         0     0        2         1        0       0          0         0
2         0     0        2         2        0       0          0         0
3         1     0        1        11        0       0          0         0
4         0     0        1         5        0       0          0         0
5         0     0        2         1        0       0          0         0
6         0     1        2         4        0       0          0         0
```

BONETREAT indicates whether or not a woman was taking one of 11 different bone medications at both enrollment *and* follow-up.

Taking a bone medication at both enrollment and follow-up will be the definition of treatment that we will use.

What is the effect of BONETREAT on fractures?

Since the use of bone medication is self-reported, we have no information about compliance or dosage, however the variable BONETREAT can act as a reasonable proxy.

```
               | glow$BONETREAT
glow$FRACTURE  |          0 |          1 | Row Total |
---------------|------------|------------|------------|
             0 |        297 |         78 |        375 |
               |      0.385 |      1.246 |            |
               |    79.200% |    20.800% |    75.000% |
               |    77.749% |    66.102% |            |
               |    59.400% |    15.600% |            |
---------------|------------|------------|------------|
             1 |         85 |         40 |        125 |
               |      1.154 |      3.737 |            |
               |    68.000% |    32.000% |    25.000% |
               |    22.251% |    33.898% |            |
               |    17.000% |     8.000% |            |
---------------|------------|------------|------------|
  Column Total |        382 |        118 |        500 |
               |    76.400% |    23.600% |            |
---------------|------------|------------|------------|
```

```
   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  500
```

```
> cor(glow$FRACTURE, glow$BONETREAT)

[1] 0.1142131
```

```
> glm.a <- glm(FRACTURE ~ BONETREAT, data = glow,
     family = binomial(link = "logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2511     0.1230 -10.170   <2e-16 ***
BONETREAT     0.5833     0.2301   2.535   0.0113 *
---
```

```
> exp(coef(glm.a))

(Intercept)   BONETREAT
  0.2861953   1.7918552
```

The unadjusted relationship suggests that taking bone medication is positively related to having a fracture. The odds of having a fracture for women taking bone medication is almost double that of women who are not taking bone medication.

What is the effect of BONETREAT on fractures?

Clearly, the effect of bone medication is to increase the probability of a fracture (you should say this sarcastically). This real effect cannot yet be determined. It is likely that women with a very high risk of fractures are the ones taking the medication and that is making this effect look bad.

Build a Propensity Model to Balance the Treatment and Non-Treatment Groups

1. Estimate the probability of treatment by fitting a logistic regression model with the treatment indicator as the outcome. This is called the *propensity score model*.

To build the PS model, we need to choose covariates that we can regress on BONETREAT.

The chosen covariates are what the treatment and non-treatment groups will be balanced on.

How should these covariates be chosen?

1. Choose covariates that predict the differences in BONETREAT (i.e., covariates related to assignment of treatment).
2. Choose covariates that predict the differences in FRACTURE (i.e., covariates related to the outcome).
3. Choose covariates that confound the effect of BONETREAT (i.e., covariates that confound the effect of treatment).
4. Choose *all* measured covariates (i.e., the kitchen-sink approach).

Early adopters choice. Often including as many covariates, interactions, and higher-order polynomial terms as possible

Problems: Overfitting; Sample size reduced

Simulations suggest (1) and (3) are best (in terms of MSE of the estimate and bias)

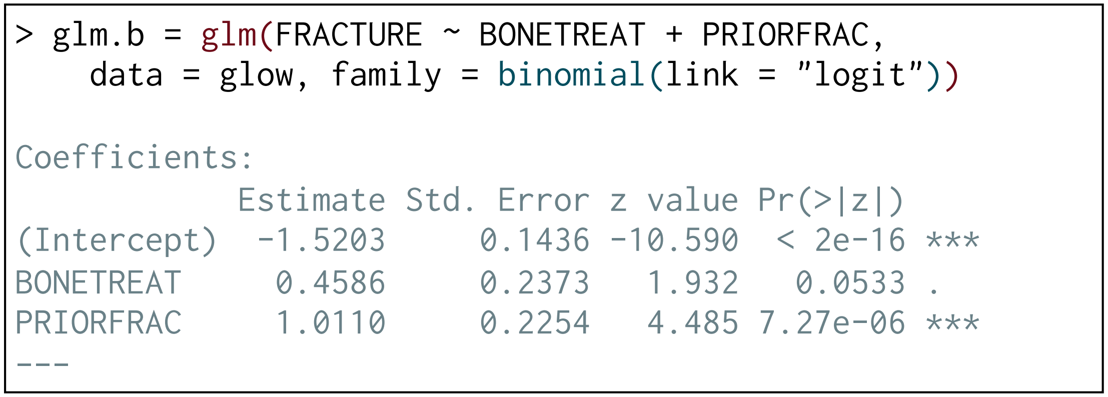
Choose Covariates that confound with BONETREAT

We will call a predictor a confounder of the effect of BONETREAT, if the adjusted effect of BONETREAT (β_1) is different than the unadjusted effect of BONETREAT (θ_1) by more than ~10%.

$$\Delta\hat{\beta}\% = 100 \times \frac{\left(\hat{\theta}_1 - \hat{\beta}_1\right)}{\hat{\beta}_1} > 10$$

Recall the **unadjusted** effect of BONETREAT was

$$\hat{\theta}_1 = 0.5833$$



```
> glm.b = glm(FRACTURE ~ BONETREAT + PRIORFRAC,
    data = glow, family = binomial(link = "logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5203     0.1436 -10.590  < 2e-16 ***
BONETREAT     0.4586     0.2373   1.932   0.0533 .
PRIORFRAC     1.0110     0.2254   4.485 7.27e-06 ***
---
```

$$\Delta\hat{\beta}\% = 100 \times \frac{(0.5833 - 0.4586)}{0.4586}$$

$$\Delta\hat{\beta}\% = 27.192$$

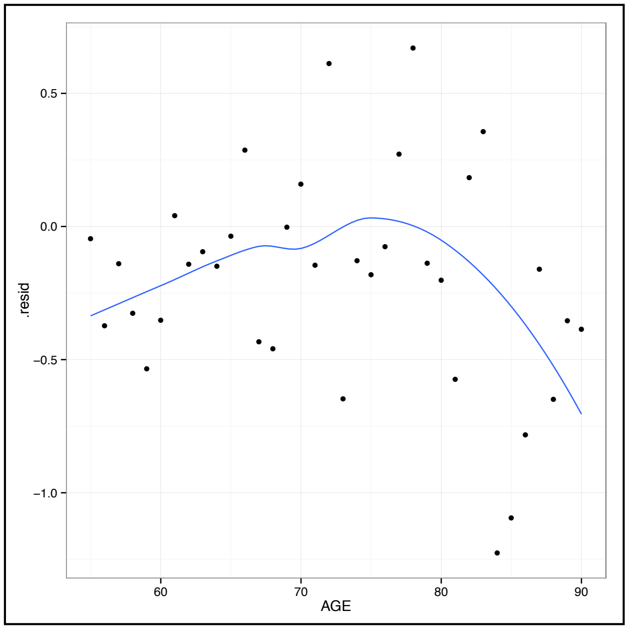
We will consider PRIORFRAC a confounder of the effect of BONETREAT, since the adjusted effect of BONETREAT is ~27.2% different than the unadjusted effect of BONETREAT (27.2 > 10).

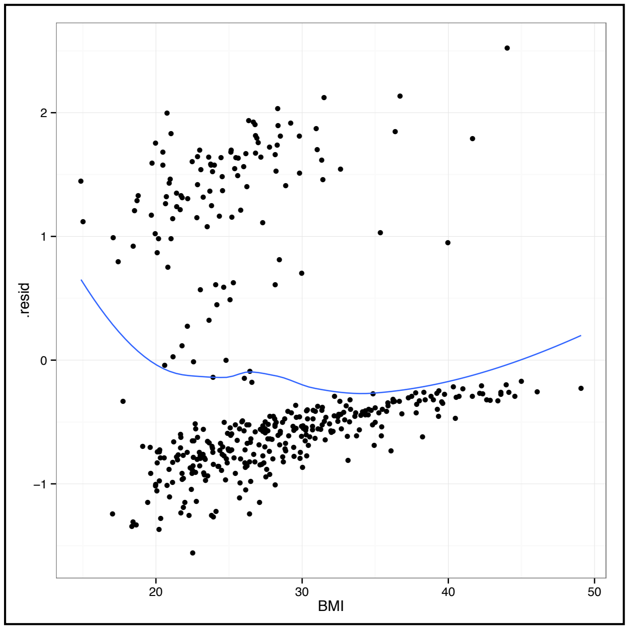
Predictor	**Adjusted effect of BONETREAT**	**Percent change in effect of BONETREAT**	**Confounder?**
PRIORFRAC	**0.4586**	**27.192**	**x**
AGE	**0.4281**	**36.239**	**x**
WEIGHT	0.5772	1.049	
HEIGHT	**0.4748**	**22.839**	**x**
BMI	**0.6465**	**–9.778**	**x**
PREMENO	0.5939	–1.778	
MOMFRAC	0.5849	–0.280	
ARMASSIST	0.6103	–4.423	
SMOKE	0.5718	2.005	

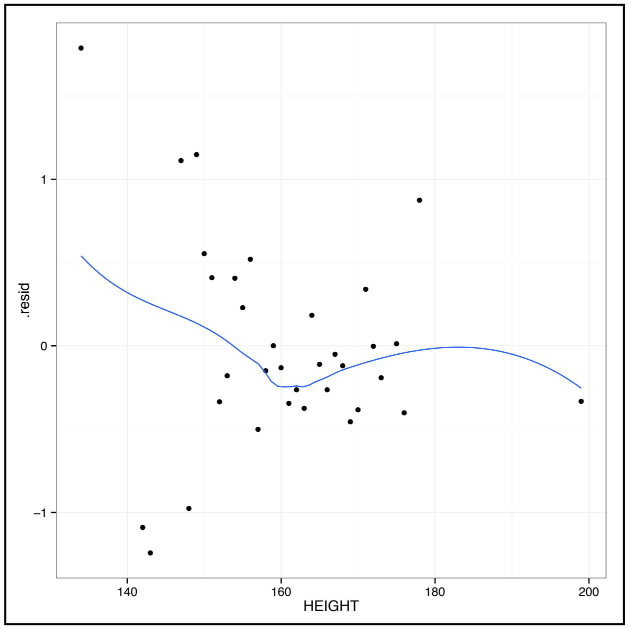
Fit Logistic Model with BONETREAT as Outcome and Confounders as Predictors

```
> glm.psm = glm(BONETREAT ~ PRIORFRAC + AGE + HEIGHT + BMI,
    data = glow, family = binomial(link = "logit"))
```

Examine each of the predictors for any potential non-linear terms (polynomials) that may need to be fitted.

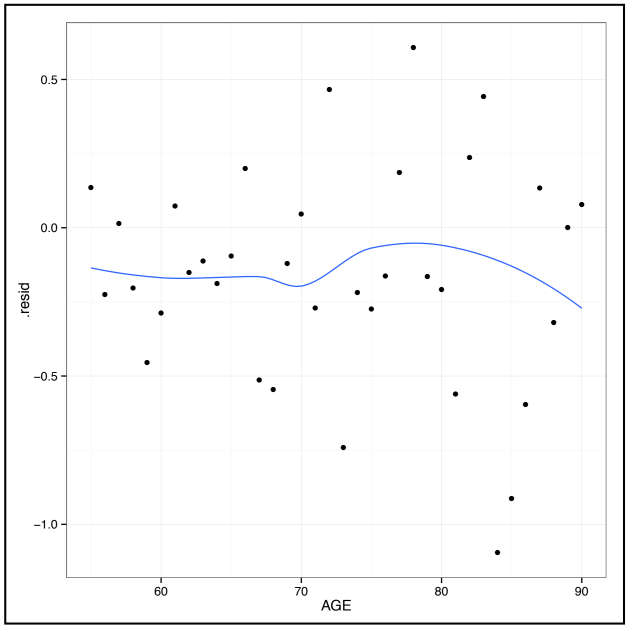


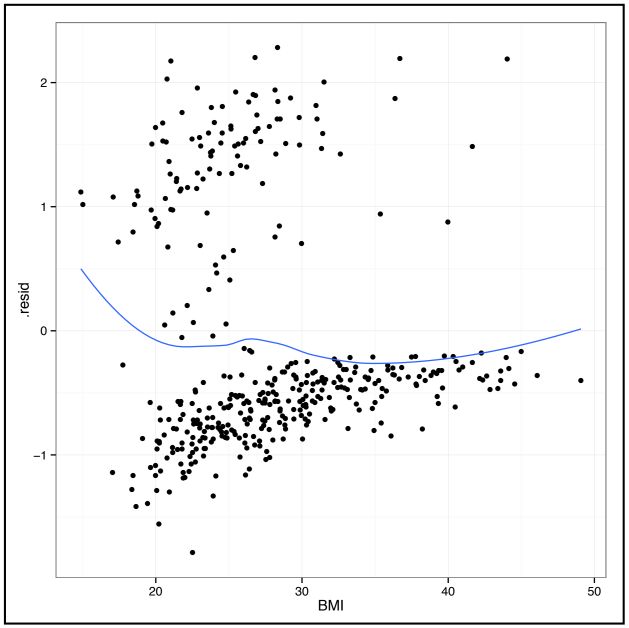


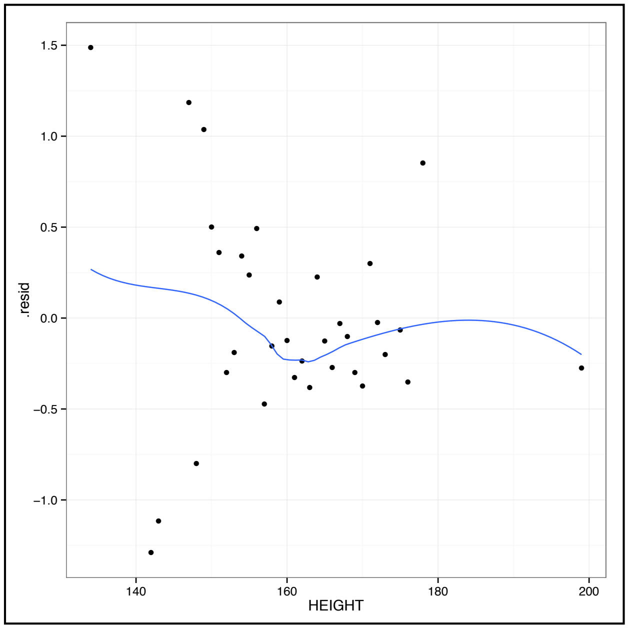


The relationship with AGE seems potentially quadratic.

```
> glm.psm = glm(BONETREAT ~ PRIORFRAC + AGE + I(AGE ^ 2) + HEIGHT + BMI,
    data = glow, family = binomial(link = "logit"))
```







The model's residuals vs. AGE are suggest better fit.
The quadratic relationship is also consistent with clinical practice.

Adding in higher-order terms for BMI and HEIGHT did not improve the fit.

No interaction terms were statistically significant,

```
> glow$PS = fitted(glm.psm)
```

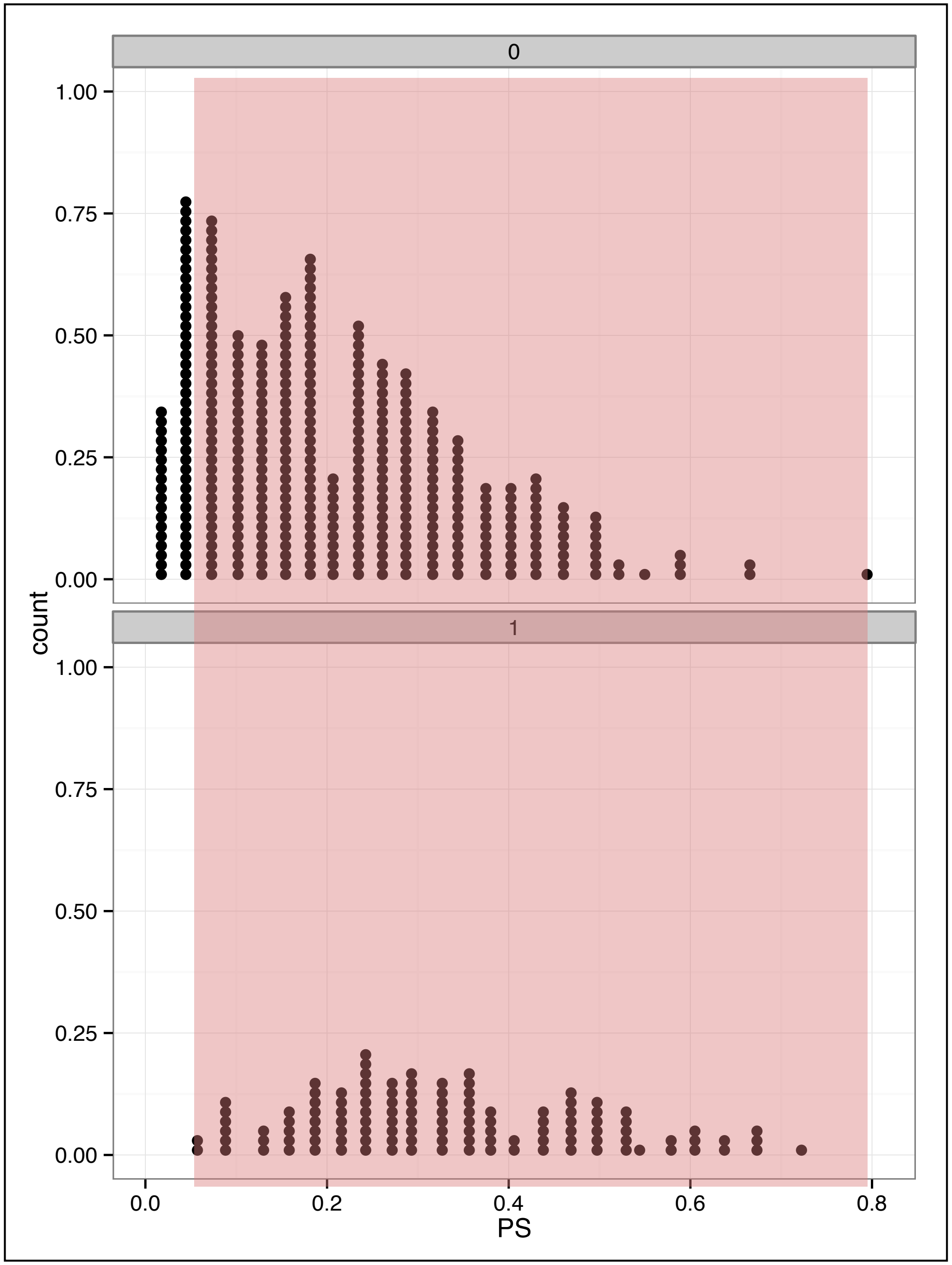
Add propensity scores to the glow data frame.

Examine the Region of Common Support

2. Use the propensity scores to determine the *region of common support*. Drop subjects from outside this region. Refit the propensity score model on the reduced data.

It has been suggested that cases that fit the following criteria be eliminated:

- Any subject in the non-treatment group ($z = 0$) with PS *less than* the smallest PS in the treatment group
- Any subject in the treatment group ($z = 1$) with PS *larger than* the largest PS in the non-treatment group



```
> describeBy(glow$PS, glow$BONETREAT)

group: 0
  var    n mean    sd median trimmed   mad  min  max range skew kurtosis   se
1   1 382 0.21 0.14   0.18    0.19 0.15 0.01 0.79  0.79 0.84     0.51 0.01
-----------------------------------------------------------------------------
group: 1
  var    n mean    sd median trimmed   mad  min  max range skew kurtosis   se
1   1 118 0.33 0.16   0.31    0.33 0.16 0.05 0.72  0.68  0.4    -0.53 0.01
```

Keep all observations that have propensity scores such that, $0.05 \leq \text{PS} \leq 0.79$

```
> glow2 = glow[glow$PS >= 0.05 & glow$PS <= 0.79, ]

> table(glow2$BONETREAT)

  0   1
339 117
```

44 observations were eliminated.

Re-Fit the Propensity Score Model to the Observations in the Region of Common Support

```
> glm.psm2 = glm(BONETREAT ~ PRIORFRAC + AGE + I(AGE ^ 2) + HEIGHT + BMI,
    data = glow2, family = binomial(link = "logit"))
```

Fitted Propensity Score Model for Treatment Variable Bone Medication

Predictor	B	SE	*z*	*p*	95% CI	
Prior Fracture	0.536	0.259	2.07	0.038	0.027	1.043
Age	0.610	0.221	2.76	0.006	0.188	1.057
Age^2	–0.004	0.002	–2.70	0.007	–0.007	–0.001
Height	–0.065	0.019	–3.35	<0.001	–0.103	–0.027
BMI	–0.128	0.027	–4.81	<0.001	–0.182	–0.077
(Intercept)	–9.446	8.154	–1.16	0.247	–25.783	6.285

Examine Discrimination of PSM

We should discrimination of the PSM.

```
> library(verification)

> roc.area(glow2$BONETREAT, fitted(glm.psm2))

$A
[1] 0.7092

$n.total
[1] 456

$n.events
[1] 117

$n.noevents
[1] 339

$p.value
[1] 7.347183e-12
```

The area under the ROC curve is 0.71. This suggests acceptable to poor discrimination.

Since the goal of the PSM is to balance, we want poor discrimination.

If the PSM has good discrimination, it may suggest a very narrow region of common support, which might adversely affect the sample size available for the analysis.

Propensity Score Analysis #1

3. Fit a model regressing the original outcome on the treatment and propensity scores to the reduced data set.

Note that the functional form of the PS covariate was checked prior to fitting the model. The linear predictor seemed sufficient.

```
> glow2$PS = fitted(glm.psm2)

> psa1 = glm(FRACTURE ~ BONETREAT + PS, data = glow2, family = binomial(link = ”logit”))
```

Fitted Models to Assess the Treatment Effect of Bone Medication

Predictor	B	SE	*z*	*p*	OR
Bone medication	0.465	0.234	1.99	0.047	1.59
(Intercept)	–1.159	0.127	–9.10	<0.001	0.31
Bone medication	0.366	0.248	1.47	0.141	1.44
Propensity score	0.937	0.770	1.22	0.224	2.55
(Intercept)	–1.377	0.223	–6.18	<0.001	0.25

The estimated odds ratio of treatment is 1.44, adjusting for propensity scores.

There is an increased odds of fracture for women on bone medication, but the increase is not statistically significant.

The analysis does not account for potential imbalance of the covariates in the two distributions.

```
> t.test(AGE ~ BONETREAT, data = glow2)

t = -3.0538, df = 209.9, p-value = 0.002552


> t.test(PRIORFRAC ~ BONETREAT, data = glow2)

t = -2.2296, df = 183.591, p-value = 0.02698


> t.test(BMI ~ BONETREAT, data = glow2)

t = 4.7736, df = 214.784, p-value = 3.345e-06


> t.test(HEIGHT ~ BONETREAT, data = glow2)

t = 3.1309, df = 177.295, p-value = 0.002039
```

All four covariates are not balanced. This suggests there is likely residual confounding and we still cannot capture the true effect of treatment.

3. Split the propensity scores into **quintiles** and **fit a separate model** regressing the original outcome on the treatment and propensity scores *within* each quintile.

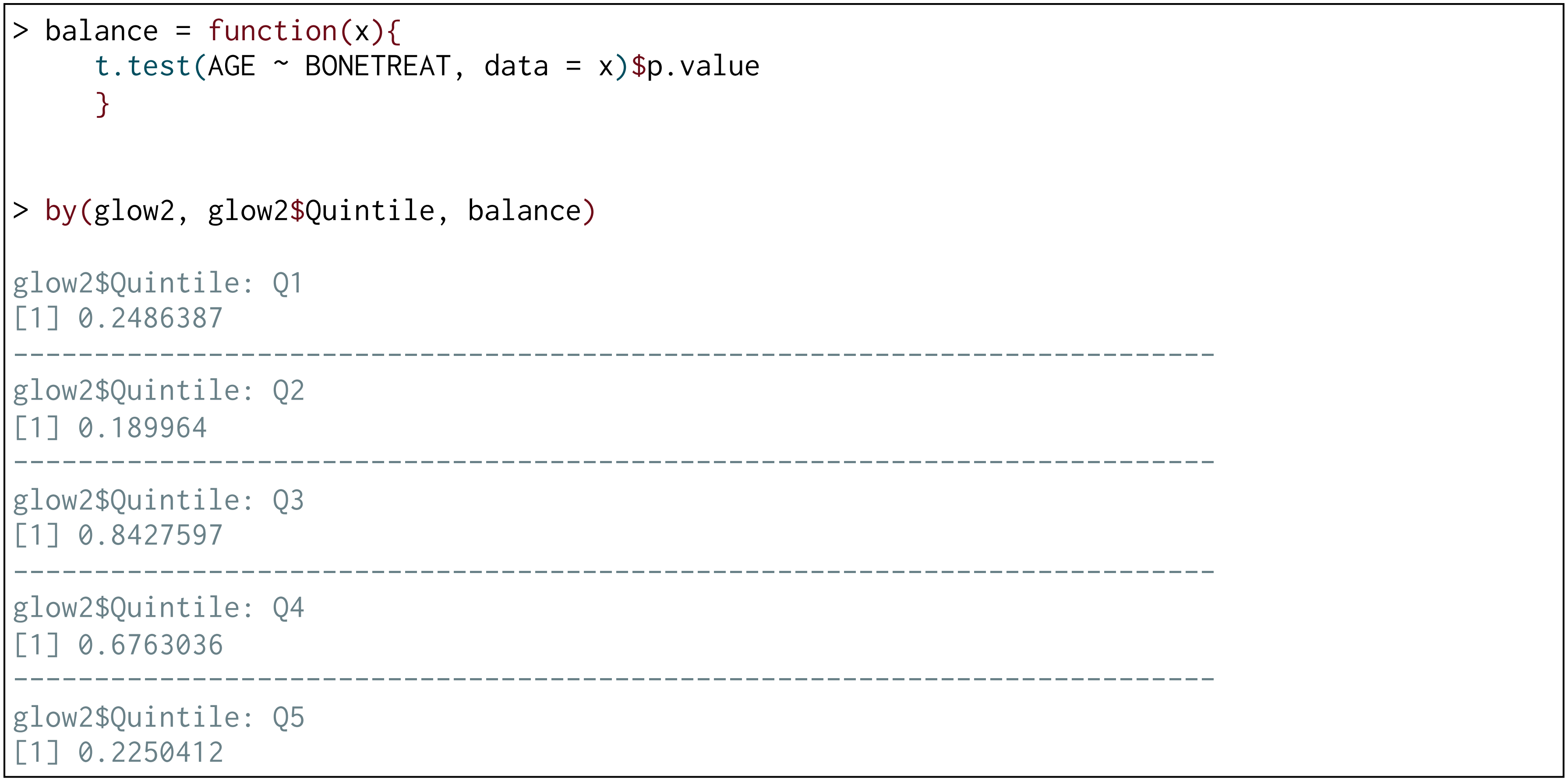
```
> glow2$Quintile = cut_number(glow2$PS, n = 5, labels = c("Q1", "Q2", "Q3", "Q4", "Q5"))


> head(glow2)

  BONETREAT        PS Quintile
1         0 0.1823510       Q2
2         0 0.1052590       Q1
3         0 0.4147955       Q5
4         0 0.2796232       Q4
5         0 0.2023555       Q3
6         0 0.3652338       Q4
```

The function `cut_number()` is in the **ggplot2** library.

Examine Covariate Balance: AGE



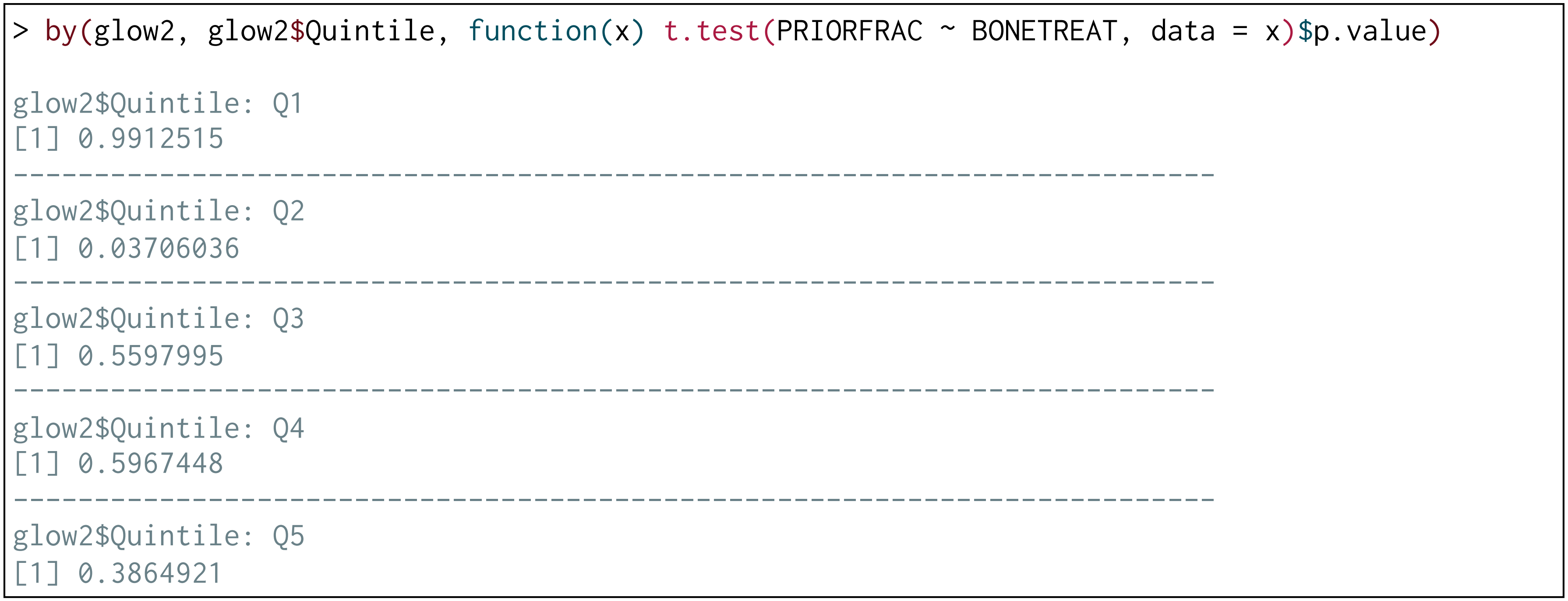
```
> balance = function(x){
     t.test(AGE ~ BONETREAT, data = x)$p.value
     }


> by(glow2, glow2$Quintile, balance)

glow2$Quintile: Q1
[1] 0.2486387
---------------------------------------------------------------------------
glow2$Quintile: Q2
[1] 0.189964
---------------------------------------------------------------------------
glow2$Quintile: Q3
[1] 0.8427597
---------------------------------------------------------------------------
glow2$Quintile: Q4
[1] 0.6763036
---------------------------------------------------------------------------
glow2$Quintile: Q5
[1] 0.2250412
```

All are **non-significant** suggesting the covariate AGE has been **balanced** across BONETREAT within quintiles.

Examine Covariate Balance: PRIORFRAC



```
> by(glow2, glow2$Quintile, function(x) t.test(PRIORFRAC ~ BONETREAT, data = x)$p.value)

glow2$Quintile: Q1
[1] 0.9912515
---------------------------------------------------------------------------
glow2$Quintile: Q2
[1] 0.03706036
---------------------------------------------------------------------------
glow2$Quintile: Q3
[1] 0.5597995
---------------------------------------------------------------------------
glow2$Quintile: Q4
[1] 0.5967448
---------------------------------------------------------------------------
glow2$Quintile: Q5
[1] 0.3864921
```

All are **non-significant** suggesting the covariate AGE has been **balanced** across BONETREAT within quintiles.

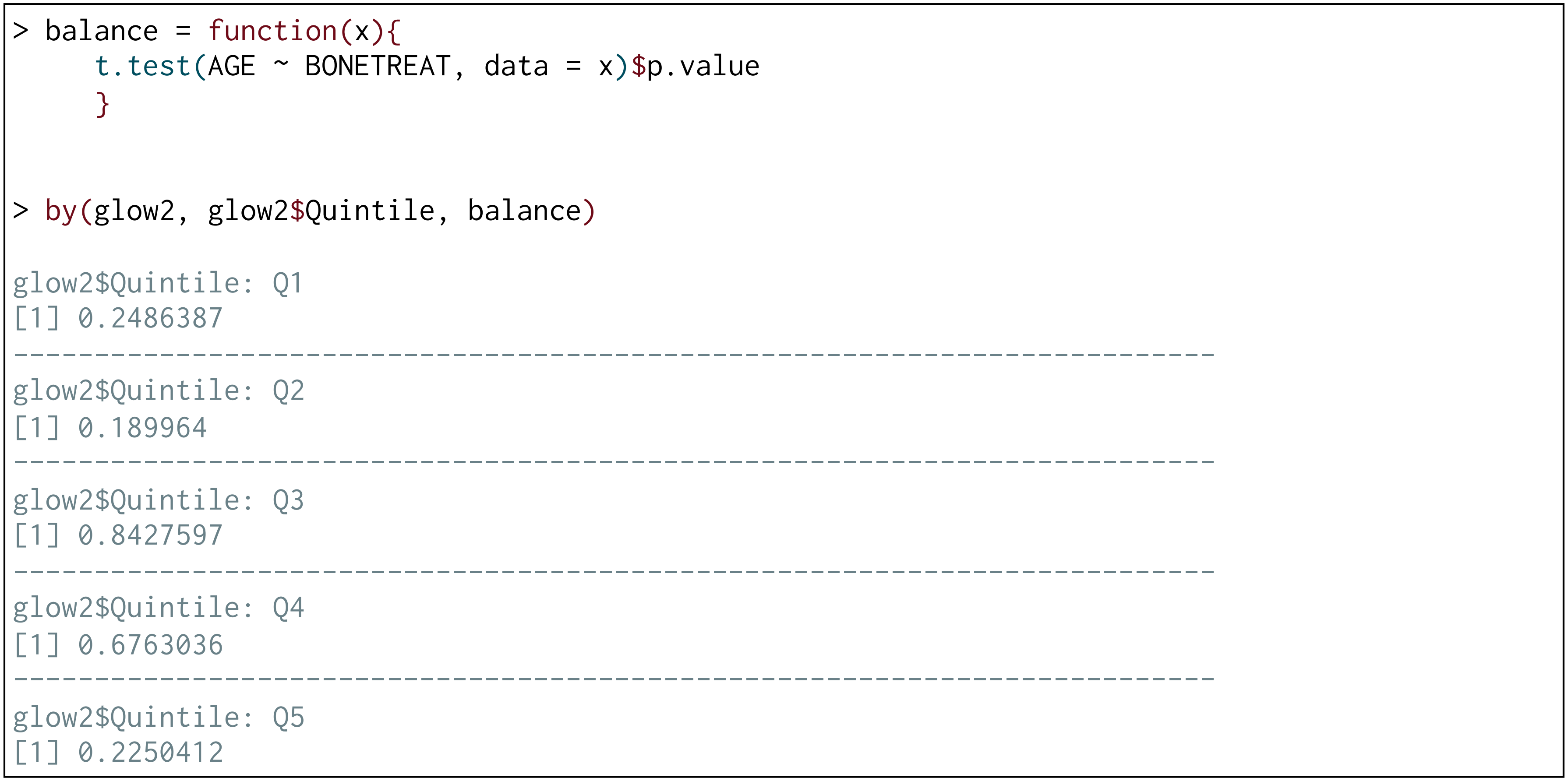
Examine Covariate Balance: AGE

```
> balance = function(x){
     t.test(AGE ~ BONETREAT, data = x)$p.value
     }


> by(glow2, glow2$Quintile, balance)

glow2$Quintile: Q1
[1] 0.2486387
---------------------------------------------------------------------------
glow2$Quintile: Q2
[1] 0.189964
---------------------------------------------------------------------------
glow2$Quintile: Q3
[1] 0.8427597
---------------------------------------------------------------------------
glow2$Quintile: Q4
[1] 0.6763036
---------------------------------------------------------------------------
glow2$Quintile: Q5
[1] 0.2250412
```

Examine Covariate Balance: AGE



```
> balance = function(x){
     t.test(AGE ~ BONETREAT, data = x)$p.value
     }


> by(glow2, glow2$Quintile, balance)

glow2$Quintile: Q1
[1] 0.2486387
---------------------------------------------------------------------------
glow2$Quintile: Q2
[1] 0.189964
---------------------------------------------------------------------------
glow2$Quintile: Q3
[1] 0.8427597
---------------------------------------------------------------------------
glow2$Quintile: Q4
[1] 0.6763036
---------------------------------------------------------------------------
glow2$Quintile: Q5
[1] 0.2250412
```

3. Create **four indicator variables** which split the propensity scores into quintiles and **fit a single model** regressing the original outcome on the treatment indicator, the four new indicators, and the propensity scores