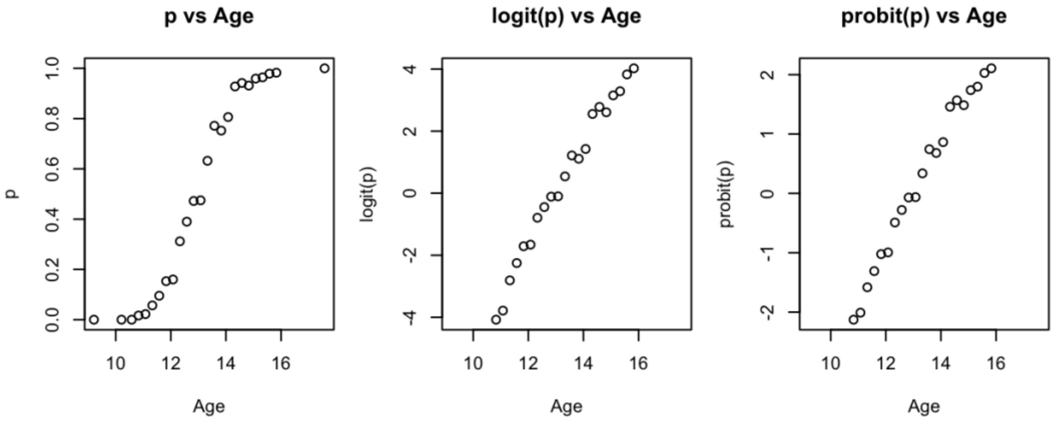
Chong Woon Kiat A0209349X - ST5213 Assignment 1

(a)



Logit and probit transformation are both similar and almost linear, slight polynomial transformation of Age may be needed for the line to be completely linear. Gradient of logit is about 1.8 times of probit.

(b)

Binomial GLMs are fitted using the logit link and probit link and the linear predictor is a polynomial in Age of degree one, two and three. Deviance test is used to compare the nested models with the same link function, whereas AIC is used to compare the models with different link functions. Summary of ANOVA table:

		Logit			Probit		
	Resid Df	Resid Deviance	Drop in Deviance	Pr(>Chi)	Resid Deviance	Drop in Deviance	Pr(>Chi)
NULL	24	3693.9			3693.9		
Age	23	26.7	3667.2	< 2.2e-16*	22.9	3671.0	< 2.2e-16*
Age^2	22	23.2	3.5	0.061318	15.1	7.7	0.005405*
Age^3	21	15.0	8.2	0.004288*	14.1	1.1	0.304149

For the logit model, the drop-in-deviance test statistic for Age^2 and Age^3 is 3.5 (p-value = 0.061) and 8.2 (p-value = 0.0043) respectively on 1 df, suggesting weak evidence that Age^2 is insignificant but Age^3 is highly significant. However, Age^2 is kept in the model due to hierarchy principle and the drop-in-deviance test statistic for $Age^2 + Age^3$ is 3.5 + 8.2 = 11.7 (p-value of .0029) on 2 df, suggesting that $Age^2 + Age^3$ is significant. Hence, model with third order in age is the preferred logit model.

For the probit model, the drop-in-deviance test statistic for Age^2 and Age^3 is 7.7 (p-value of 0.0054) and 1.1 (p-value of 0.30) respectively on 1 df, indicating that model with up to second order in age is adequate.

Model	AIC	Deviance	Df	Pr(>Chi)
$\text{Logit}(p) = -165.40 + 33.52\,\text{Age} - 2.33\,\text{Age}^2 + 0.056\,\text{Age}^3$	107.1	15.0	21	0.82
$\text{Probit}(p) = -20.18 + 2.19\,\text{Age} - 0.048\,\text{Age}^2$	105.2	15.1	22	0.86

Since deviance follows χ_{df} approximately, the best logit model and probit model have deviance of 15.0 (p-value 0.82) and 15.1 (p-value 0.86) respectively, the hypothesis that the models are adequate is not rejected. However, the probit model is the preferred model as it has a lower AIC of 105.2.

(c)

Model	AIC
Logit(p) = – 2.59 + 4.43 log(Age – 9) – 2.15 log(18 – Age)	104.7
Probit(p) = – 20.18 + 2.19 Age – 0.048 Age^2	105.2

The new logit model with term log(Age – 9) and log(18 – Age) has a lower AIC than the probit model and hence a better model. This logit model gives a direct interpretation of log-odds of success and also a closed form solution, whereas for the probit model, there is no close form solution when solving for sample proportion p.
However, the downside of this logit model is that it only works within the Age range of 9 to 18.

(d)
To obtain the decile ages at, the cloglog function is solved numerically for p = 0.1, 0.2, …, 0.9,
$cloglog(p) = -131.416 + 24.727x + -1.560x^2 + 0.0334x^3$

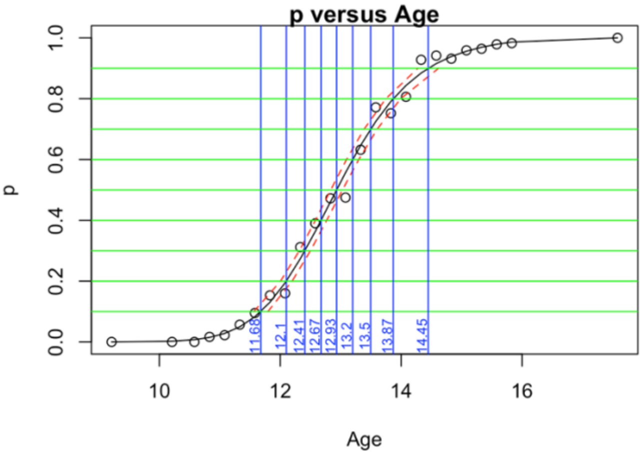
With delta method, $x = h(\boldsymbol{\beta}) \approx h(\boldsymbol{\mu}) + \nabla h(\boldsymbol{\mu})(\boldsymbol{\beta} - \boldsymbol{\mu})$ and hence the standard error $\sqrt{Var[h(\boldsymbol{\beta})]} = \sqrt{\nabla h(\boldsymbol{\mu})^T \Sigma \nabla h(\boldsymbol{\mu})}$ where Σ is the variance covariance matrix of $\boldsymbol{\beta}$.

The MLE, $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}\ , \boldsymbol{I}(\boldsymbol{\beta})^{-1})$ approximately for large samples and the approximation still holds with $\boldsymbol{\beta}$ substituted by $\widehat{\boldsymbol{\beta}}$, hence $\boldsymbol{\beta} \approx \widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma} \approx \boldsymbol{I}(\widehat{\boldsymbol{\beta}})^{-1}$.
$\nabla h(\widehat{\boldsymbol{\beta}}) = [\frac{\partial h}{\partial \beta_0}, \cdots, \frac{\partial h}{\partial \beta_3}]^T$ is obtained by applying partial differentiation with respect to β implicitly to the cloglog function:

$$
\begin{aligned}
0 &= 1 + \beta_1 \frac{\partial x}{\partial \beta_0} + 2\beta_2 x \frac{\partial x}{\partial \beta_0} + 3\beta_3 x^2 \frac{\partial x}{\partial \beta_0} \\
0 &= x + \beta_1 \frac{\partial x}{\partial \beta_1} + 2\beta_2 x \frac{\partial x}{\partial \beta_1} + 3\beta_3 x^2 \frac{\partial x}{\partial \beta_1} \\
0 &= \beta_1 \frac{\partial x}{\partial \beta_2} + x^2 + 2\beta_2 x \frac{\partial x}{\partial \beta_2} + 3\beta_3 x^2 \frac{\partial x}{\partial \beta_2} \\
0 &= \beta_1 \frac{\partial x}{\partial \beta_3} + 2\beta_2 x \frac{\partial x}{\partial \beta_3} + x^3 + 3\beta_3 x^2 \frac{\partial x}{\partial \beta_3}
\end{aligned}
\quad \Rightarrow \quad
\begin{aligned}
\frac{\partial x}{\partial \beta_0} &= \frac{-1}{\beta_1 + 2\beta_2 x + 3\beta_3 x^2} \\
\frac{\partial x}{\partial \beta_1} &= \frac{-x}{\beta_1 + 2\beta_2 x + 3\beta_3 x^2} \\
\frac{\partial x}{\partial \beta_2} &= \frac{-x^2}{\beta_1 + 2\beta_2 x + 3\beta_3 x^2} \\
\frac{\partial x}{\partial \beta_3} &= \frac{-x^3}{\beta_1 + 2\beta_2 x + 3\beta_3 x^2}
\end{aligned}
$$

Hence, the decile ages $h(\widehat{\boldsymbol{\beta}})$ and their standard errors $\sqrt{\nabla h(\widehat{\boldsymbol{\beta}})^{T} \boldsymbol{I}(\widehat{\boldsymbol{\beta}})^{-1} \nabla h(\widehat{\boldsymbol{\beta}})}$ is given by:

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Age	11.68	12.10	12.41	12.67	12.93	13.20	13.50	13.87	14.45
Standard Error	0.059	0.049	0.047	0.048	0.049	0.052	0.055	0.063	0.087



(a)
Firstly, with the null model, full model and model with all main terms as starting point, the best models with respect to AIC are selected using both forward selection and backward elimination method:

Starting model	Parameters in final model selected using AIC	AIC
Null	Age + Region + Gender + Age:Gender	478.6
Full	Gender + Race + Age + Region + Gender:Race + Gender:Age + Race:Age + Gender:Race:Age	478.0
All main terms	Gender + Race + Age + Region + Gender:Race + Gender:Age	475.0

Since AIC tends to favor large models, drop-in-deviance test is further applied on the models selected using AIC to check if any term can be added or dropped. Any parameter with p-value of drop-in-deviance test statistic < 0.05 on 1 df will be added or dropped one at a time.

Starting model	Parameters in final model selected using AIC and deviance test	AIC
Null	Age + Region + Gender + Age:Gender	478.6
Full	Gender + Race + Age + Region + Gender:Race	477.0
All main terms	Gender + Race + Age + Region + Gender:Race	477.0

The final model with lowest AIC is then selected. let p denote the probability of being satisfied, the model is given by logit(p) ~ Gender + Race + Age + Region + Gender:Race
Logit(p) = 0.431 + 0.490 GenderM + 0.213 RaceW + 0.363 Age>44 + 0.128 Age35-44 – 0.349 RegionMW – 0.436 RegionNE – 0.313 RegionNW – 0.025 RegionP – 0.261 RegionS – 0.148 RegionSW – 0.380 GenderM:RaceW

where Female (for Gender), Other (for Race), <35 (for Age), MA (for Region) are the baseline.

(b)
A Logistic model with parameter (Region + Race + Gender*Age) is fitted and the coefficients are given in table below:

- Odds of an employee being satisfied is independent of race as wald test statistic for age is 0.003/0.062=0.05, p-value 2P(Z>0.05) = 0.96.
- The odds of a <35 year-old, female from region Mid-Atlantic region (baseline) being satisfied is exp(0.511) = 1.67
- Given the same age group and gender, the odds of an employee from region X is Y times that of an employee from Mid-Atlantic, where X and Y are given in the table on the right.

X	Y
MidWest	exp(-0.356) = 0.70
NorthEast	exp(-0.444) = 0.64
NorthWest	exp(-0.307) = 0.74
Pacific	exp(-0.02) = 0.98
Southern	exp(-0.266) = 0.77
SouthWest	exp(-0.147) = 0.86

- Given the same region, the odds comparison for (i) female from different age group, (ii) male from different age group, (iii) male and female from the same age group:
  (i) the odds of a female employee aged 35-44 and >44 being satisfied is exp(0.289)=1.34 and exp(0.564)=1.76 times that of a female employee aged <35 respectively.
  (ii) the odds of a male employee aged 35-44 and >44 being satisfied is exp(0.289-0.238)=1.05 and exp(0.564-0.285)=1.32 times that of a male employee aged <35 respectively.
  (iii) the odds of a male employee aged <35, 35-44 and >45 being satisfied is exp(0.307)=1.36, exp(0.307-0.238)=1.07, exp(0.307-0.285)=1.02 times that of a female employee from the same age group respectively.

Parameter	Est.	StdErr	z	Pr(>|z|)	Parameter	Est.	StdErr	z	Pr(>|z|)
(Intercept)	0.511	0.114	4.46	8.07E-06*	RaceW	0.003	0.062	0.05	0.9624
RegionMW	-0.356	0.104	-3.44	5.84E-04*	GenderM	0.307	0.066	4.68	2.85E-06*
RegionNE	-0.444	0.104	-4.27	1.93E-05*	Age>44	0.564	0.099	5.71	1.13E-08*
RegionNW	-0.307	0.104	-2.94	3.28E-03*	Age35-44	0.289	0.099	2.92	0.0035*
RegionP	-0.02	0.125	-0.16	0.874	GenderM:Age>44	-0.285	0.115	-2.47	0.0134*
RegionS	-0.266	0.108	-2.47	0.0136*	GenderM:Age35-44	-0.238	0.116	-2.05	0.0402*
RegionSW	-0.147	0.108	-1.36	1.74E-01*					

(c)
A probit model (Region + Gender*Race + Gender*Age) is fitted with equation given by
probit(p) = 0.234 + 0.354 GenderM + 0.119 RaceW + 0.329 Age>44 + 0.175 Age35-44 – 0.215 RegionMW – 0.271 RegionNE – 0.193 RegionNW – 0.019 RegionP – 0.163 RegionS – 0.093 RegionSW – 0.212 GenderM:RaceW – 0.149 GenderM:Age>44 – 0.137 GenderM:Age35-44

The MLE for linear predictor, $x^T\hat{\beta} \sim N(x^T\beta\,, x^T\Sigma x)$ approximately for large samples and the approximation still holds with β substituted by $\hat{\beta}$, hence $\beta \approx \hat{\beta}, \Sigma \approx I(\hat{\beta})^{-1}$.

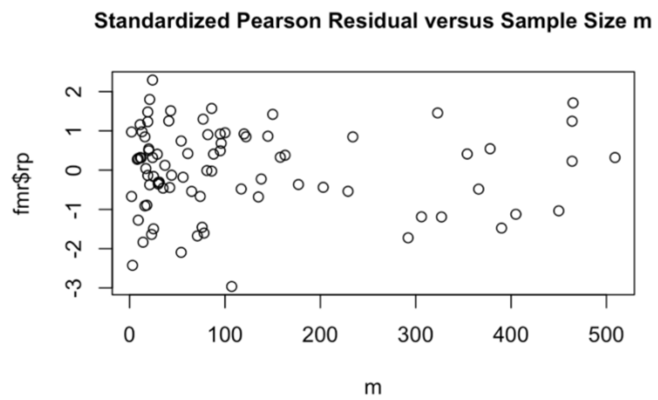
The probability of employees being satisfied, p is given by p = $\Phi\left(x^T\hat{\beta}\right)$ where Φ is the cumulative distribution function of $Z{\sim}N(0,1)$. Since Φ is a monotone function, the 95% confidence interval can be constructed by

$$\left(\Phi\left(x^T\hat{\beta} - z_{\frac{.05}{2}}\sqrt{x^T I\left(\hat{\beta}\right)^{-1}x}\right), \quad \Phi\left(x^T\hat{\beta} + z_{\frac{.05}{2}}\sqrt{x^T I\left(\hat{\beta}\right)^{-1}x}\right)\right)$$

The linear predictor, $x^T\hat{\beta}$ of probability of satisfaction for a female white employee aged 35-44 working in the Pacific region is = 0.5098 and its 95% confidence interval of $x^T\hat{\beta}$ is (0.5098-1.96*0.0730,0.5098+1.96*0.0730) = (0.3668, 0.6528) and therefore the 95% confidence interval of probability of satisfaction $\Phi\left(x^T\hat{\beta}\right)$ is (Φ(0.3668), Φ(0.6528)) = (0.6431, 0.7431)

(d)
An informal goodness of fit test can be used to judge if the model is lack of fit (when sample sizes of all groups are large).
The deviance of the above probit model is 82.0 with a degree of freedom of 70. It follows χ^2_{70} distribution approximately, the p-value = P($\chi^2_{70} > 82.0$) = 0.155. Hence, there is not enough evidence to reject the hypothesis that the model is adequate.

Since the deviance = 82.0 > E(χ^2_{70}) $= 70$.
This suggests some form of overdispersion which could be due to several reasons. For example, the satisfaction of employee might not be independent of each other, or there is missing of important explanatory variables in the model.



From the standardized Pearson residual against sample size plot, the spread of the residuals remains constant with increasing size. This suggests that we can apply the quasi-likelihood approach by applying a constant dispersion parameter ϕ to the variance function $V(\mu_i)$, where the new variance function $v(\mu_i) = \phi V(\mu_i)$ with ϕ estimated by $\frac{deviance}{n-d} = \frac{82.0}{70} = 1.17$.

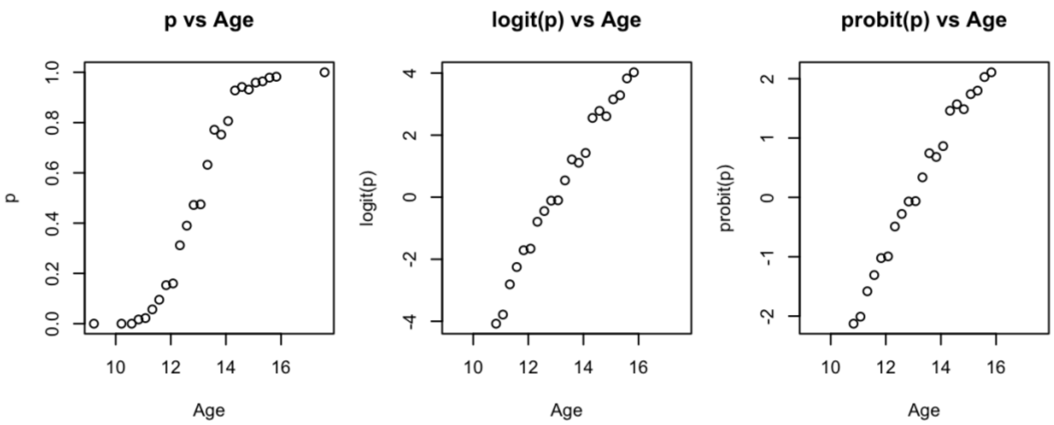
The quasi-likelihood estimates of the parameters are the same as the MLE while the standard error are $\sqrt{\phi}$ times that of MLEs. Hence the new confidence interval is given by

$$\left(\Phi\left(x^T\hat{\beta} - 1.96\sqrt{\phi\, x^T I\left(\hat{\beta}\right)^{-1}x}\right), \quad \Phi\left(x^T\hat{\beta} + 1.96\sqrt{\phi\, x^T I\left(\hat{\beta}\right)^{-1}x}\right)\right)$$

Hence, the confidence interval in (c) is revised to (0.6387, 0.7468) which is wider than the original interval.

APPENDIX #TASK 1

```
> menarche <- read.table("menarche.txt", header=TRUE)
> #menarche

> menarche$p <- Menarche/Total
> attach(menarche)
>
> par(mfrow=c(1,3),mar=c(4,4,4,1))
>
> plot(Age,p, main = "p vs Age")
>
> logit <- function(p) log(p/(1-p))
> plot(Age,logit(p),main = "logit(p) vs Age")
>
> probit <- function(p) qnorm(p)
> plot(Age,probit(p),main = "probit(p) vs Age")
>
```




```
> y.Bin <- cbind(Menarche, Total - Menarche)
> fm_logit <- glm(y.Bin ~ Age , data = menarche, family = binomial)  # logit link
> fm_logit2 <- glm(y.Bin ~ Age + I(Age^2) , data = menarche, family = binomial)
> fm_logit3 <- glm(y.Bin ~ Age + I(Age^2) + I(Age^3) , data = menarche, family = binomial)
> anova(fm_logit,fm_logit2,fm_logit3, test = "Chisq")
```

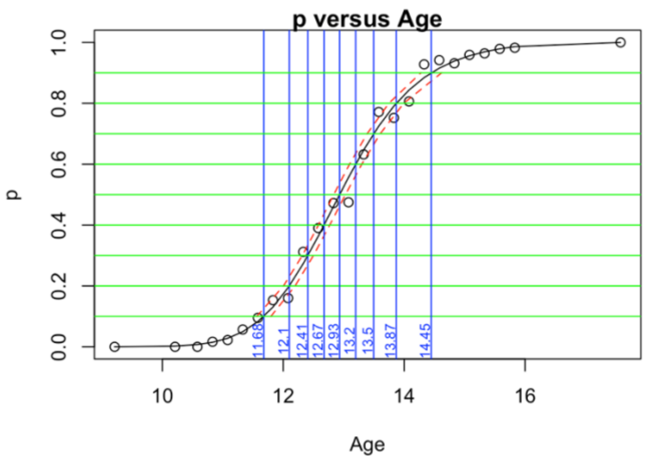
```
Analysis of Deviance Table

Model 1: y.Bin ~ Age
Model 2: y.Bin ~ Age + I(Age^2)
Model 3: y.Bin ~ Age + I(Age^2) + I(Age^3)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        23     26.703
2        22     23.202  1   3.5014 0.061318 .
3        21     15.044  1   8.1575 0.004288 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> fm_probit <- glm(y.Bin ~ Age , data = menarche, family = binomial(link="probit"))  # logit link
> fm_probit2 <- glm(y.Bin ~ Age + I(Age^2), data = menarche, family = binomial(link="probit"))
> fm_probit3 <- glm(y.Bin ~ Age + I(Age^2) + I(Age^3), data = menarche, family = binomial(link="probit"))
> anova(fm_probit,fm_probit2,fm_probit3, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: y.Bin ~ Age
Model 2: y.Bin ~ Age + I(Age^2)
Model 3: y.Bin ~ Age + I(Age^2) + I(Age^3)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        23     22.887
2        22     15.149  1   7.7387 0.005405 **
3        21     14.093  1   1.0559 0.304149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pchisq(fm_logit3$deviance,fm_logit3$df.residual,lower.tail = FALSE)
[1] 0.8207102
> pchisq(fm_probit2$deviance,fm_probit2$df.residual,lower.tail = FALSE)
[1] 0.8557755
>
> AIC(fm_logit,fm_logit2,fm_logit3,fm_probit,fm_probit2,fm_probit3)
           df      AIC
fm_logit    2 114.7553
fm_logit2   3 113.2539
fm_logit3   4 107.0963
fm_probit   2 110.9392
fm_probit2  3 105.2006
fm_probit3  4 106.1446

> menarche$Age9 <- log(Age-9)
> menarche$Age18 <- log(18-Age)
>
> fm_c <- glm(y.Bin ~ Age9 + Age18 , data = menarche, family = binomial)
> summary(fm_c)

Call:
glm(formula = y.Bin ~ Age9 + Age18, family = binomial, data = menarche)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.60468  -0.41869  -0.02285   0.56209   1.42200

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5935     2.0240  -1.281  0.20006
Age9          4.4347     0.6568   6.752 1.46e-11 ***
Age18        -2.1467     0.7370  -2.913  0.00358 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   14.659  on 22  degrees of freedom
AIC: 104.71

Number of Fisher Scoring iterations: 5
```

```
> AIC(fm_logit,fm_logit2,fm_logit3,fm_probit,fm_probit2,fm_probit3,fm_c)
           df      AIC
fm_logit    2 114.7553
fm_logit2   3 113.2539
fm_logit3   4 107.0963
fm_probit   2 110.9392
fm_probit2  3 105.2006
fm_probit3  4 106.1446
fm_c        3 104.7103

> fm_cloglog <- glm(y.Bin ~ Age + I(Age^2) + I(Age^3), data = menarche, family=binomial(link="cloglog"))
> summary(fm_cloglog)

Call:
glm(formula = y.Bin ~ Age + I(Age^2) + I(Age^3), family = binomial(link = "cloglog"),
    data = menarche)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.63537  -0.34567  -0.00701   0.44577   1.43010

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -131.41614   31.54465  -4.166  3.1e-05 ***
Age           24.72671    6.90946   3.579 0.000345 ***
I(Age^2)      -1.55971    0.50277  -3.102 0.001921 **
I(Age^3)       0.03335    0.01215   2.744 0.006071 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   14.604  on 21  degrees of freedom
AIC: 106.66

Number of Fisher Scoring iterations: 7

> beta <- coef(fm_cloglog)
> beta0 <- beta[1]
> beta1 <- beta[2]
> beta2 <- beta[3]
> beta3 <- beta[4]
>
> clogclog <- function(p) log(-log(1-p))
>
> (I <- range(Age))
[1]  9.21 17.58
> f <- function(x, p) {
+   predict(fm_cloglog, data.frame(Age=x))-clogclog(p)}
>
> a <- rep(0, 9)
> p <- seq(9)/10
>
```

```
> for (ii in 1:9) {
+   output <- uniroot(f, interval=I, p=p[ii])
+   a[ii] <- output$root
+ }
>
> sd_a <- rep(0,9)
> ul <- rep(0, 9)
> ll <- rep(0, 9)
>
> for (ii in 1:9) {
+   x <- a[ii]
+   dm <- beta1 + 2*beta2*x + 3*beta3*x^2 # denominator
+   h <- c(-1/dm, -x/dm, -x^2/dm, -x^3/dm)
+   sd_x <- sqrt(h %*% vcov(fm_cloglog) %*% h)
+   sd_a[ii] <- as.numeric(sd_x)
+   ll[ii] <- x + -1 * qnorm(0.975) * sd_a[ii]
+   ul[ii] <- x + 1 * qnorm(0.975) * sd_a[ii]
+ }
>
> round(rbind(p,ul,a,ll,sd_a),3)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]
p     0.100  0.200  0.300  0.400  0.500  0.600  0.700  0.800  0.900
ul   11.795 12.196 12.499 12.768 13.029 13.299 13.604 13.993 14.619
a    11.679 12.100 12.407 12.675 12.932 13.197 13.496 13.869 14.448
ll   11.562 12.004 12.315 12.581 12.835 13.096 13.388 13.746 14.277
sd_a  0.059  0.049  0.047  0.048  0.049  0.052  0.055  0.063  0.087

>
> par(mar = c(4,4,1,1))
> plot(p ~ Age, menarche, ylim = c(0,1), main="p versus Age")
> lines(Age, predict(fm_cloglog, menarche, type = "response"))
> lines(ul,p,lty = 2,col ="red")
> lines(ll,p,lty = 2,col ="red")
> for (ii in 1:9) {
+ abline(v = a[ii], col = "blue")
+ abline(h = p[ii], col = "green")
+ text(a[ii]-0.1, 0.02, round(a[ii],2), col = "blue",srt=90, cex=0.7)
+ }
```



APPENDIX #TASK 2

```
> sat <- read.csv("satisfaction.csv", header=TRUE)
> sat$Gender <- factor(sat$Gender)
> sat$Race <- factor(sat$Race)
> sat$Age <- factor(sat$Age)
> sat$Region <- factor(sat$Region)
> y <- cbind(sat$Satisfied, sat$Notsatisfied)
>
> fm0 <- glm(y ~ 1, sat, family = binomial)
> fmfull <- glm(y ~ Gender * Race * Age * Region, sat, family = binomial)
>
> library(MASS)
>
> ############# from NULL model ###############
> fm1 <- stepAIC(fm0, scope = list(lower=formula(fm0), upper = formula(fmfull)), trace=0)
> summary(fm1)

Call:
glm(formula = y ~ Age + Region + Gender + Age:Gender, family = binomial,
    data = sat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.65274  -0.57551   0.08098   0.77298   2.80528

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.51310    0.10583   4.848 1.25e-06 ***
Age>44            0.56492    0.09832   5.746 9.15e-09 ***
Age35-44          0.28941    0.09902   2.923 0.003469 **
RegionMW         -0.35644    0.10363  -3.440 0.000583 ***
RegionNE         -0.44377    0.10383  -4.274 1.92e-05 ***
RegionNW         -0.30654    0.10424  -2.941 0.003274 **
RegionP          -0.02008    0.12510  -0.160 0.872495
RegionS          -0.26646    0.10800  -2.467 0.013617 *
RegionSW         -0.14722    0.10829  -1.359 0.174004
GenderM           0.30751    0.06505   4.728 2.27e-06 ***
Age>44:GenderM   -0.28504    0.11505  -2.478 0.013228 *
Age35-44:GenderM -0.23819    0.11614  -2.051 0.040283 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 204.141  on 83  degrees of freedom
Residual deviance:  89.564  on 72  degrees of freedom
AIC: 478.57

Number of Fisher Scoring iterations: 3

> add1(fm1, scope=fmfull, test="Chisq")
Single term additions

Model:
```

```
y ~ Age + Region + Gender + Age:Gender
              Df Deviance    AIC    LRT Pr(>Chi)
<none>            89.564 478.57
Race           1   89.562 480.57 0.0022   0.9624
Gender:Region  6   85.707 486.72 3.8565   0.6961
Age:Region    12   82.011 495.02 7.5532   0.8190
> drop1(fm1, test="Chisq")
Single term deletions

Model:
y ~ Age + Region + Gender + Age:Gender
           Df Deviance    AIC    LRT  Pr(>Chi)
<none>          89.564 478.57
Region      6  129.487 506.50 39.923 4.716e-07 ***
Age:Gender  2   97.525 482.53  7.961   0.01867 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ############## from FULL model ################
> fm2 <- stepAIC(fmfull, scope = list(lower=formula(fm0), upper=formula(fmfull)), trace=0)
> summary(fm2)

Call:
glm(formula = y ~ Gender + Race + Age + Region + Gender:Race +
    Gender:Age + Race:Age + Gender:Race:Age, family = binomial,
    data = sat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.70493  -0.54593   0.04255   0.68952   1.91353

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.43267    0.13267   3.261 0.001109 **
GenderM                 0.42175    0.13830   3.050 0.002292 **
RaceW                   0.11515    0.11488   1.002 0.316165
Age>44                  0.29136    0.24845   1.173 0.240899
Age35-44                0.16824    0.18939   0.888 0.374377
RegionMW               -0.35068    0.10373  -3.381 0.000723 ***
RegionNE               -0.43973    0.10396  -4.230 2.34e-05 ***
RegionNW               -0.31346    0.10445  -3.001 0.002691 **
RegionP                -0.02765    0.12542  -0.220 0.825501
RegionS                -0.26632    0.10807  -2.464 0.013726 *
RegionSW               -0.15199    0.10839  -1.402 0.160826
GenderM:RaceW          -0.15432    0.15742  -0.980 0.326940
GenderM:Age>44          0.40600    0.34238   1.186 0.235694
GenderM:Age35-44        0.05246    0.27222   0.193 0.847187
RaceW:Age>44            0.29131    0.27122   1.074 0.282779
RaceW:Age35-44          0.15985    0.22242   0.719 0.472331
GenderM:RaceW:Age>44   -0.72764    0.36473  -1.995 0.046045 *
GenderM:RaceW:Age35-44 -0.34139    0.30302  -1.127 0.259897
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 204.141  on 83  degrees of freedom
Residual deviance:  77.029  on 66  degrees of freedom
AIC: 478.04

Number of Fisher Scoring iterations: 4

> add1(fm2, scope=fmfull, test="Chisq")
Single term additions

Model:
y ~ Gender + Race + Age + Region + Gender:Race + Gender:Age +
    Race:Age + Gender:Race:Age
              Df Deviance    AIC    LRT Pr(>Chi)
<none>            77.029 478.04
Gender:Region  6   74.277 487.29 2.7513   0.8394
Race:Region    6   72.422 485.43 4.6065   0.5952
Age:Region    12   69.114 494.12 7.9150   0.7917
> drop1(fm2, test="Chisq")
Single term deletions

Model:
y ~ Gender + Race + Age + Region + Gender:Race + Gender:Age +
    Race:Age + Gender:Race:Age
                Df Deviance    AIC    LRT  Pr(>Chi)
<none>              77.029 478.04
Region           6  114.982 503.99 37.953 1.147e-06 ***
Gender:Race:Age  2   81.513 478.52  4.484    0.1062
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> (fm2 <- update(fm2, .~. -Gender:Race:Age))

Call:  glm(formula = y ~ Gender + Race + Age + Region + Gender:Race +
    Gender:Age + Race:Age, family = binomial, data = sat)

Coefficients:
     (Intercept)          GenderM            RaceW           Age>44         Age35-44
         0.36557          0.56396          0.21225          0.63416          0.30006
       RegionMW         RegionNE         RegionNW          RegionP          RegionS
        -0.35227         -0.44284         -0.31602         -0.03098         -0.26701
       RegionSW    GenderM:RaceW   GenderM:Age>44  GenderM:Age35-44      RaceW:Age>44
        -0.15149         -0.33804         -0.22862         -0.21890         -0.12023
  RaceW:Age35-44
        -0.02326

Degrees of Freedom: 83 Total (i.e. Null);  68 Residual
Null Deviance:      204.1
Residual Deviance: 81.51        AIC: 478.5
> add1(fm2, scope=fmfull, test="Chisq")
Single term additions

Model:
y ~ Gender + Race + Age + Region + Gender:Race + Gender:Age +
    Race:Age
```

```
             Df Deviance    AIC    LRT Pr(>Chi)
<none>              81.513 478.52
Gender:Region    6   78.524 487.53 2.9887   0.8103
Race:Region      6   76.557 485.57 4.9564   0.5494
Age:Region      12   73.638 494.65 7.8750   0.7948
Gender:Race:Age  2   77.029 478.04 4.4845   0.1062
> drop1(fm2, test="Chisq")
Single term deletions

Model:
y ~ Gender + Race + Age + Region + Gender:Race + Gender:Age +
    Race:Age
          Df Deviance    AIC    LRT Pr(>Chi)
<none>          81.513 478.52
Region       6  119.902 504.91 38.389 9.43e-07 ***
Gender:Race  1   88.861 483.87  7.348 0.006713 **
Gender:Age   2   86.967 479.98  5.454 0.065424 .
Race:Age     2   81.968 474.98  0.455 0.796680
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> (fm2 <- update(fm2, .~. -Race:Age))

Call:  glm(formula = y ~ Gender + Race + Age + Region + Gender:Race +
    Gender:Age, family = binomial, data = sat)

Coefficients:
     (Intercept)          GenderM            RaceW          Age>44         Age35-44
         0.38044          0.57082          0.19110         0.53249          0.28352
        RegionMW         RegionNE         RegionNW         RegionP          RegionS
        -0.35201         -0.44213         -0.31555        -0.03042         -0.26747
        RegionSW    GenderM:RaceW   GenderM:Age>44  GenderM:Age35-44
        -0.15246         -0.34257         -0.23853        -0.22200

Degrees of Freedom: 83 Total (i.e. Null);  70 Residual
Null Deviance:      204.1
Residual Deviance: 81.97        AIC: 475
> add1(fm2, scope=fmfull, test="Chisq")
Single term additions

Model:
y ~ Gender + Race + Age + Region + Gender:Race + Gender:Age
            Df Deviance    AIC    LRT Pr(>Chi)
<none>            81.968 474.98
Race:Age       2   81.513 478.52 0.4546   0.7967
Gender:Region  6   79.020 484.03 2.9477   0.8154
Race:Region    6   77.322 482.33 4.6457   0.5900
Age:Region    12   74.197 491.21 7.7703   0.8028
> drop1(fm2, test="Chisq")
Single term deletions

Model:
y ~ Gender + Race + Age + Region + Gender:Race + Gender:Age
          Df Deviance    AIC    LRT  Pr(>Chi)
<none>          81.968 474.98
```

```
Region       6  120.229 501.24 38.261 9.986e-07 ***
Gender:Race  1   89.562 480.57  7.594  0.005856 **
Gender:Age   2   87.941 476.95  5.973  0.050453 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> (fm2 <- update(fm2, .~. -Gender:Age))

Call:  glm(formula = y ~ Gender + Race + Age + Region + Gender:Race,
    family = binomial, data = sat)

Coefficients:
 (Intercept)      GenderM        RaceW       Age>44      Age35-44     RegionMW
     0.43098      0.49025      0.21345      0.36339      0.12782     -0.34945
    RegionNE     RegionNW      RegionP      RegionS     RegionSW  GenderM:RaceW
    -0.43637     -0.31317     -0.02452     -0.26123     -0.14810     -0.38016

Degrees of Freedom: 83 Total (i.e. Null);  72 Residual
Null Deviance:      204.1
Residual Deviance: 87.94        AIC: 477
> add1(fm2, scope=fmfull, test="Chisq")
Single term additions

Model:
y ~ Gender + Race + Age + Region + Gender:Race
              Df Deviance    AIC    LRT Pr(>Chi)
<none>            87.941 476.95
Gender:Age     2   81.968 474.98 5.9734  0.05045 .
Race:Age       2   86.967 479.98 0.9743  0.61438
Gender:Region  6   85.207 486.22 2.7343  0.84138
Race:Region    6   83.691 484.70 4.2505  0.64282
Age:Region    12   79.786 492.80 8.1552  0.77289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> drop1(fm2, test="Chisq")
Single term deletions

Model:
y ~ Gender + Race + Age + Region + Gender:Race
            Df Deviance    AIC    LRT  Pr(>Chi)
<none>          87.941 476.95
Age          2  138.243 523.25 50.302 1.194e-11 ***
Region       6  126.031 503.04 38.090 1.079e-06 ***
Gender:Race  1   97.516 484.53  9.575  0.001972 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


>
> ############# from ALL-MAIN-TERM model ###############
> fm3 <- glm(y ~  Gender + Race + Age + Region , sat, family = binomial)
> fm3 <- stepAIC(fm3, scope=list(lower = formula(fm0), upper=formula(fmfull)), trace=0)
> summary(fm3)

Call:
```

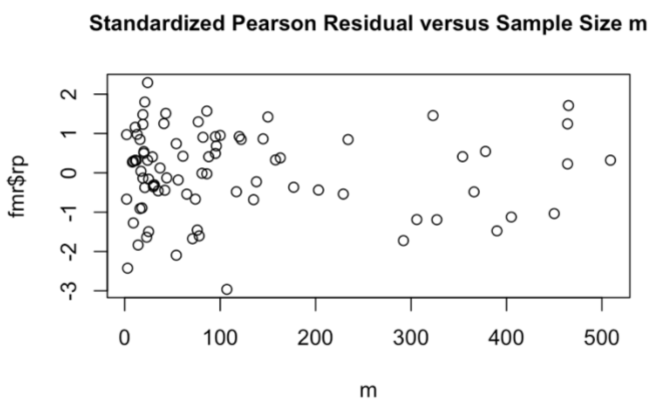
```
glm(formula = y ~ Gender + Race + Age + Region + Gender:Race +
    Gender:Age, family = binomial, data = sat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5674  -0.5234   0.1401   0.7540   2.5049

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.38044    0.12374   3.075 0.002108 **
GenderM            0.57082    0.11639   4.904 9.37e-07 ***
RaceW              0.19110    0.09190   2.079 0.037575 *
Age>44             0.53249    0.09956   5.348 8.88e-08 ***
Age35-44           0.28352    0.09915   2.860 0.004243 **
RegionMW          -0.35201    0.10369  -3.395 0.000687 ***
RegionNE          -0.44213    0.10389  -4.256 2.08e-05 ***
RegionNW          -0.31555    0.10441  -3.022 0.002511 **
RegionP           -0.03042    0.12532  -0.243 0.808209
RegionS           -0.26747    0.10806  -2.475 0.013313 *
RegionSW          -0.15246    0.10837  -1.407 0.159463
GenderM:RaceW     -0.34257    0.12455  -2.750 0.005952 **
GenderM:Age>44    -0.23853    0.11638  -2.050 0.040409 *
GenderM:Age35-44 -0.22200    0.11641  -1.907 0.056504 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 204.141  on 83  degrees of freedom
Residual deviance:  81.968  on 70  degrees of freedom
AIC: 474.98

Number of Fisher Scoring iterations: 4

> add1(fm3, scope=fmfull, test="Chisq")
Single term additions

Model:
y ~ Gender + Race + Age + Region + Gender:Race + Gender:Age
              Df Deviance    AIC    LRT Pr(>Chi)
<none>             81.968 474.98
Race:Age       2   81.513 478.52 0.4546   0.7967
Gender:Region  6   79.020 484.03 2.9477   0.8154
Race:Region    6   77.322 482.33 4.6457   0.5900
Age:Region    12   74.197 491.21 7.7703   0.8028
> drop1(fm3, test="Chisq")
Single term deletions

Model:
y ~ Gender + Race + Age + Region + Gender:Race + Gender:Age
            Df Deviance    AIC    LRT  Pr(>Chi)
<none>           81.968 474.98
Region       6  120.229 501.24 38.261 9.986e-07 ***
Gender:Race  1   89.562 480.57  7.594  0.005856 **
Gender:Age   2   87.941 476.95  5.973  0.050453 .
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> (fm3 <- update(fm3, .~. -Gender:Age))

Call:  glm(formula = y ~ Gender + Race + Age + Region + Gender:Race,
    family = binomial, data = sat)

Coefficients:
  (Intercept)       GenderM         RaceW        Age>44      Age35-44      RegionMW
      0.43098       0.49025       0.21345       0.36339       0.12782      -0.34945
     RegionNE      RegionNW       RegionP       RegionS      RegionSW  GenderM:RaceW
     -0.43637      -0.31317      -0.02452      -0.26123      -0.14810      -0.38016

Degrees of Freedom: 83 Total (i.e. Null);  72 Residual
Null Deviance:      204.1
Residual Deviance: 87.94        AIC: 477
> add1(fm3, scope=fmfull, test="Chisq")
Single term additions

Model:
y ~ Gender + Race + Age + Region + Gender:Race
              Df Deviance    AIC    LRT Pr(>Chi)
<none>            87.941 476.95
Gender:Age     2   81.968 474.98 5.9734  0.05045 .
Race:Age       2   86.967 479.98 0.9743  0.61438
Gender:Region  6   85.207 486.22 2.7343  0.84138
Race:Region    6   83.691 484.70 4.2505  0.64282
Age:Region    12   79.786 492.80 8.1552  0.77289
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> drop1(fm3, test="Chisq")
Single term deletions

Model:
y ~ Gender + Race + Age + Region + Gender:Race
            Df Deviance    AIC    LRT  Pr(>Chi)
<none>          87.941 476.95
Age          2  138.243 523.25 50.302 1.194e-11 ***
Region       6  126.031 503.04 38.090 1.079e-06 ***
Gender:Race  1   97.516 484.53  9.575  0.001972 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> >
```

```
> #(b)
>
> fmb <- glm(y ~ Region + Race + Gender*Age, sat, family = binomial)
> summary(fmb)

Call:
glm(formula = y ~ Region + Race + Gender * Age, family = binomial,
    data = sat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.65106  -0.56894   0.08189   0.78005   2.81081

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.511043   0.114498   4.463 8.07e-06 ***
RegionMW         -0.356402   0.103630  -3.439 0.000584 ***
RegionNE         -0.443731   0.103837  -4.273 1.93e-05 ***
RegionNW         -0.306753   0.104337  -2.940 0.003282 **
RegionP          -0.019846   0.125193  -0.159 0.874044
RegionS          -0.266436   0.108003  -2.467 0.013627 *
RegionSW         -0.147186   0.108297  -1.359 0.174116
RaceW             0.002911   0.061781   0.047 0.962417
GenderM           0.307105   0.065606   4.681 2.85e-06 ***
Age>44            0.564434   0.098860   5.709 1.13e-08 ***
Age35-44          0.289319   0.099040   2.921 0.003486 **
GenderM:Age>44   -0.284831   0.115133  -2.474 0.013363 *
GenderM:Age35-44 -0.238310   0.116171  -2.051 0.040231 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 204.141  on 83  degrees of freedom
Residual deviance:  89.562  on 71  degrees of freedom
AIC: 480.57

Number of Fisher Scoring iterations: 4

>
>
>
```

```
> #(c)
>
> fm_c <- glm(y ~ Region + Gender*Race + Gender*Age, sat, family = binomial(link="probit"))
> summary(fm_c)

Call:
glm(formula = y ~ Region + Gender * Race + Gender * Age, family = binomial(link = "probit"),
    data = sat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5695  -0.5206   0.1472   0.7450   2.5033

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.23447    0.07569   3.098 0.001951 **
RegionMW          -0.21527    0.06268  -3.435 0.000593 ***
RegionNE          -0.27060    0.06283  -4.306 1.66e-05 ***
RegionNW          -0.19275    0.06310  -3.055 0.002253 **
RegionP           -0.01915    0.07559  -0.253 0.800007
RegionS           -0.16342    0.06536  -2.500 0.012403 *
RegionSW          -0.09337    0.06540  -1.428 0.153390
GenderM            0.35365    0.07174   4.930 8.24e-07 ***
RaceW              0.11915    0.05717   2.084 0.037162 *
Age>44             0.32854    0.06091   5.394 6.90e-08 ***
Age35-44           0.17535    0.06136   2.858 0.004267 **
GenderM:RaceW     -0.21222    0.07672  -2.766 0.005673 **
GenderM:Age>44    -0.14884    0.07116  -2.092 0.036462 *
GenderM:Age35-44  -0.13724    0.07193  -1.908 0.056404 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 204.141  on 83  degrees of freedom
Residual deviance:  82.001  on 70  degrees of freedom
AIC: 475.01

Number of Fisher Scoring iterations: 4


> sat[sat$Age=="35-44"&sat$Region=="P"&sat$Gender=="F"&sat$Race=="W",]
   Satisfied Notsatisfied Gender Race   Age Region
56        20           10      F    W 35-44      P
> (X <- model.matrix(fm_c)[56,])
     (Intercept)        RegionMW        RegionNE        RegionNW         RegionP
               1               0               0               0               1
         RegionS        RegionSW         GenderM           RaceW          Age>44
               0               0               0               1               0
        Age35-44   GenderM:RaceW  GenderM:Age>44 GenderM:Age35-44
               1               0               0               0

> (y <- predict(fm_c,data.frame(Region="P",Age="35-44",Gender="F",Race="W"),type="response"))
0.6949113
> (Xb <- predict(fm_c,data.frame(Region="P",Age="35-44",Gender="F",Race="W")))
```

```
0.5098201
> V <- vcov(fm_c)
> pnorm(Xb+qnorm(0.975)*sqrt(t(X)%*%V%*%X))
         [,1]
[1,] 0.7430621
> pnorm(Xb-qnorm(0.975)*sqrt(t(X)%*%V%*%X))
         [,1]
[1,] 0.6431255
>
> #(d)
> pchisq(fm_c$deviance,fm_c$df.residual,lower.tail = FALSE)
[1] 0.1545469
>
> m <- sat$Satisfied+sat$Notsatisfied
> library(boot)
> fmr <- glm.diag(fm_c)
>
> par(mfrow=c(1,1),mar=c(4,4,4,1))
> plot(m,fmr$rp,main = "Standardized Pearson Residual versus Sample Size m", cex.main=1.0)
```




```
> (k <- fm_c$deviance/fm_c$df.residual)
[1] 1.171446

> pnorm(Xb+qnorm(0.975)*sqrt(k*t(X)%*%V%*%X))
         [,1]
[1,] 0.7468429
> pnorm(Xb-qnorm(0.975)*sqrt(k*t(X)%*%V%*%X))
         [,1]
[1,] 0.6387249
```