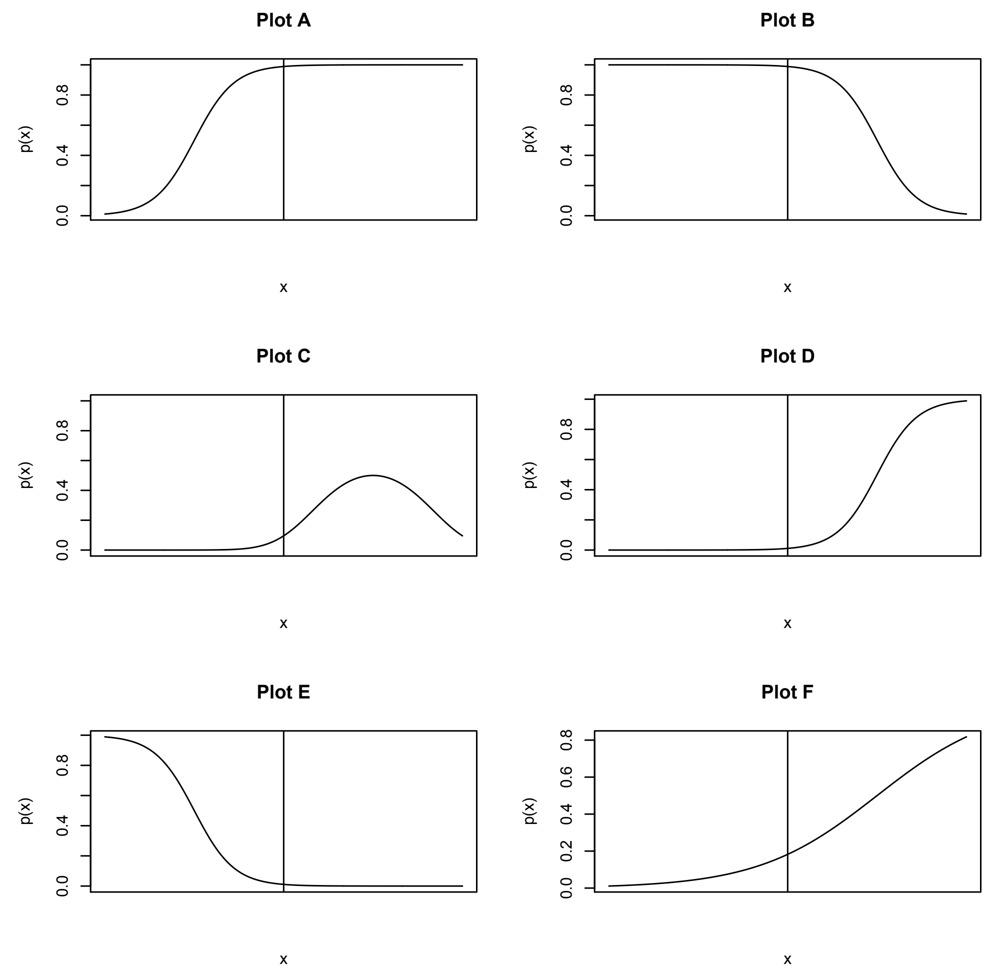
Stat 545 Exam Solutions

December 19, 2016

1. [30 points] A logistic model of the form $P[Y = 1|X = x] = p(x)$ is fit where x is a 1-dimensional continuous variable. The fitted model is

$$\text{logit}(p(x)) \;=\; -4.79 \;+\; 3.52x.$$

Below are plots of $p(x)$ for different logistic models. Determine which one could be the plot of the given fitted model. In each plot, the vertical line corresponds to the axis where $x = 0$.



Solution: Since the logit fit has a positive slope, the fitted probability must be an increasing function of x. This rules out plots B, C, and E. When $x = 0$, the logit is -4.79, which corresponds to a pretty small probability:

$$p(0) \;=\; \frac{e^{-4.79}}{1+e^{-4.79}} \;\leq\; \frac{3^{-4}}{1+3^{-4}} \;\leq\; 3^{-4} = 1/81.$$

This rules out plots A and F, leaviong only plot D as the possible correct answer.

2. **[20 points]** Explain the difference between Pearson residuals and standardized residuals. Give an example of a family (model) where the two are the same and an example where they are different.

Solution: The Pearson residuals are

$$R_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

The standardized residuals are divided by and estimated variance, i.e.

$$S_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i}.$$

For the Poisson family, $\hat{\sigma}_i = \sqrt{\hat{\mu}_i}$, so the two agree in this case. For the binomial family, $\hat{\sigma}_i = \sqrt{\hat{\mu}_i(1 - \hat{\mu}_i)}$, so the two types of residuals do not agree in this case. (Note: This was discussed in lecture.)

3. [**15 points**] Define the linear (or identity) link function for a GLM, and explain why it is seldom used for the binomial or Poisson families.

Solution: The linear link function is the identity:

$$g(\mu_i) = \mu_i = \sum_j \beta_j x_{ij}.$$

It is appropriate for the Gaussian family because the mean can be any real number. For most other families, it is not a good choice since the mean has constraints. For example, for the binomial family we must have $0 \leq \mu_i \leq 1$ which imposes a bunch of linear constraints on the allowable values of the coefficients β_j where the constraints depend on the observed values of the predictor variables x_{ij}, and this would be very unnatural.

4. [**20 points**] Below is some output from the fit of a log-linear model that is created from 3 categorical variables X, Y, and Z. Use this ouput to answer the questions that follow.

```
> summary(fit)

Call:
```

```
glm(formula = n ~ X + Y + Z + X * Y + X * Z + Y * Z, family = poisson())

Deviance Residuals:
      1        2        3        4        5        6        7        8        9
 1.2159   0.2425  -2.0992  -0.9309  -0.8693   1.7998  -3.0255  -0.7586   2.4604

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.3103     0.1814  18.251  < 2e-16 ***
XX2           1.0939     0.2104   5.198 2.01e-07 ***
XX3          -0.2750     0.2604  -1.056  0.29111
YY2           0.6525     0.2172   3.004  0.00266 **
ZZ2          -1.2815     0.3028  -4.233 2.31e-05 ***
XX2:YY2      -3.0822     0.3934  -7.835 4.69e-15 ***
XX3:YY2      -0.6343     0.3119  -2.034  0.04196 *
XX2:ZZ2      -0.9032     0.4042  -2.235  0.02545 *
XX3:ZZ2       0.5676     0.2886   1.967  0.04920 *
YY2:ZZ2       1.2385     0.3019   4.103 4.08e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 248.813  on 11  degrees of freedom
Residual deviance:  31.192  on  2  degrees of freedom
AIC: 106.67

Number of Fisher Scoring iterations: 5
```

(a) Do the results show any evidence of independence or conditional independence between any pair of the variables X, Y, Z?

Solution: First, let's sort out what we have. There are 3 levels for the X variable and 2 levels for Y and Z. The only coefficient which is not significant (at 0.05 level) is $XX2$, which is the difference between the log mean of $X = 2$ minus $X = 1$, all others held the same. Since all the interaction terms are significantly non-zero, there is no evidence for indpendence or conditional independence.

(b) Write an expression for an approximate 95% confidence interval for λ_2^Z using the numbers from the output. You don't need to do any arithmetic.

Solution: That would be the `ZZ2` coefficient. The confidence interval is

$$-1.2815 \ \pm \ 1.960 * 0.3028.$$

The 1.960 is the 0.975 quantile of the standard normal.

(c) Can you give a 95% confidence interval for λ_1^X?

Solution: By definition $\lambda_1^X = 0$ in this formulation of the model. It doesn't make sense to talk about a confidence interval for something know the be 0.

(d) Write a few lines of R-code to compute $P[Z = 1|X = 2\&Y = 2]$.

Solution: Working out the formula,

$$\begin{aligned}
& P[Z = 1|X = 2\&Y = 2] \\
&= \frac{P[X = 2\&Y = 2\&Z = 1]}{P[X = 2\&Y = 2]} \\
& P[X = 2\&Y = 2\&Z = 1] \\
&= \mu_{221}/\sum_{ijk} \mu_{ijk}. \\
& P[X = 2\&Y = 2] \\
&= (\mu_{221} + \mu_{222})/\sum_{ijk} \mu_{ijk}. \\
& P[Z = 1|X = 2\&Y = 2] \\
&= \mu_{221}/(\mu_{221} + \mu_{222}). \\
\mu_{221} &= \exp\left[\lambda + \lambda_2^X + \lambda_2^Y + \lambda_1^Z + \lambda_{22}^{XY} + \lambda_{21}^{XZ} + \lambda_{21}^{YZ}\right].
\end{aligned}$$

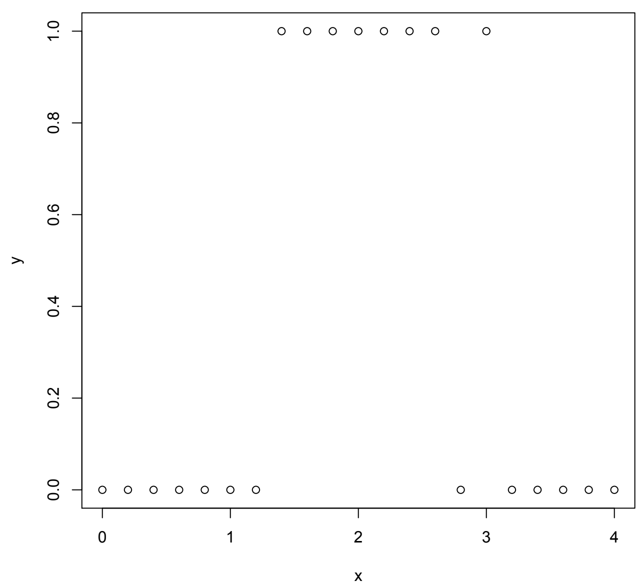
There is a similar experssion for μ_{222}. Note that any term with a subscript value of 1 will be 0. Reading of the coefficient values, here is the R-code:

```
mu221 = exp(3.3103 + 1.0939 + 0.6525 + 0 - 3.0822 + 0 + 0)
mu222 = exp(3.3103 + 1.0939 + 0.6525 - 1.2815 - 3.0822 - 0.9032 + 1.2385)
mu221/(mu222+mu221)
```

The printout from the last statement will be the answer. Actually, we can simplify things because there are a lot of common terms in the exponents which can be factored out. So a better result might be

```
1/(1+exp(- 1.2815- 0.9032 + 1.2385))
```

5. [15 [points] Below is a plot of some data on a binary outcome variable Y that (possibly) depends on a predictor variable X.



Here are the results of fitting a logistic regression model to the data:

```
> summary(fit)

Call:
glm(formula = y ~ x, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0887  -1.0216  -0.8943   1.3662   1.4278

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  -0.7654     0.8878  -0.862    0.389
x             0.1383     0.3734   0.370    0.711

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.910  on 20  degrees of freedom
Residual deviance: 27.772  on 19  degrees of freedom
AIC: 31.772

Number of Fisher Scoring iterations: 4
```

(a) Does the output of the model offer any evidence of dependence between X and Y?

Solution: Assuming the model is correct, the coefficient of `x` is not significant, so there is no evidence of an association.

(b) Do you think the data plot offers any evidence of dependence between X and Y?

Solution: The plot clearly suggests that intermediate values of `x` have a high probability for productin `y` values of 1, i.e., that there is a dependence.

(c) Suggest a better model.

Solution: A logistic model which is quadratic in the `x` would fit better:

$$\text{logit}(x) \;=\; \beta_1 + \beta_2 x + \beta_3 x^2.$$

Note: These data were in fact generated from a quadratic logit model.