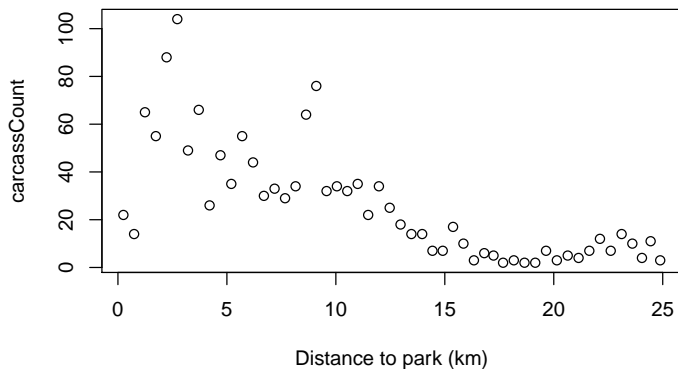


Chapter 6 Practice with GLM

1. More road kill: count of amphibian carcasses. Predictor is distance to a natural park.



```
> RdKill.fit1 <- glm(carcassCount ~ dist2Pk, family = poisson)
> summary(RdKill.fit1)$coef
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	4.316	0.04322	99.9	0.00e+00
dist2Pk	-0.106	0.00439	-24.1	1.28e-128

```
> RdKill.fit2 <- glm(carcassCount ~ dist2Pk, family = quasipoisson)
> summary(RdKill.fit2)
```

```
Call:
glm(formula = carcassCount ~ dist2Pk, family = quasipoisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.110  -1.695  -0.471   1.421   7.334

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.3165     0.1194  36.16 < 2e-16
dist2Pk     -0.1059     0.0121  -8.73  1.2e-11

(Dispersion parameter for quasipoisson family taken to be 7.63)

Null deviance: 1071.4 on 51 degrees of freedom
Residual deviance: 390.9 on 50 degrees of freedom
AIC: NA

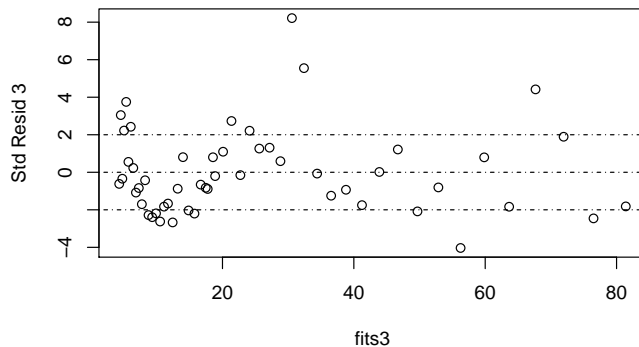
Number of Fisher Scoring iterations: 4
```

- (a) Why do we have z tests in the first output, but t tests in the second?
Because in a Poisson GLM, the variance is known to equal the mean, and it does not have to be estimated. In the second output, the unexplained variance or overdispersion is estimated from the residuals and it is used to increase the SE's.
- (b) How do the standard errors differ? Which make more sense?
By a factor of $\sqrt{7.63}$, the estimated overdispersion.
- (c) Why might we see overdispersion in these data?
Because there are lurking variables which were not measured having some effect on the response.

- (d) Why do residuals get divided by $\sqrt{y_i}$ for standardizing?
Poisson variance = mean, so we are dividing by an estimated SD.
- (e) We need to ask the researchers what is odd about the first two observations. Since they aren't available, I data snooped through the other predictors to see if these rows are unusual in any of the other predictors. They are high in olive (ha of olive groves) and in n.patch, the number of habitat patches. I'll enter an indicator for high patchiness.

```
> patch.80 <- I(RoadKills$n.patch > 80) + 0
> RdKill.fit3 <- update(RdKill.fit1, . ~ . + patch.80)
> fits3 <- fitted(RdKill.fit3)
> plot(y = (carcassCount - fits3)/sqrt(fits3), x = fits3, ylab = "Std Resid 3")
> abline(h = c(-2, 0, 2), lty = 4)
> summary(update(RdKill.fit3, family = quasipoisson))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.554	0.1097	41.52	7.52e-40
dist2Pk	-0.125	0.0113	-11.04	6.87e-15
patch.80	-1.602	0.4045	-3.96	2.42e-04



Is the overdispersion gone?

No, still lots of points over 2 SD from 0.

Other problems?

Yes, there is a zigzag below 20 which our model cannot explain.

Summarize the distance effect on road kill amphibian counts.

An approximate 95% CI for slope of distance is (-0.147, -0.103). Backtransforming from log scale gives a multiplicative effect for increasing distance by 1 km of $e^{-0.125 \pm 0.022} = (0.86, 0.90)$, meaning counts decrease by 10 to 14%. Note that the SE's are adjusted for overdispersion which is now estimated as 5.5

2. In 1996 a poll was taken asking how liberal or conservative voters perceived Clinton to be. The response had seven categories from very liberal (1) to very conservative (7). We will build a multinomial model based on predictors: partyID (0 =strong Democrat to 6 = strong Republican) and education (1 through 7).

```
> nes96 <- read.table("http://www.stat.washington.edu/quinn/classes/536/data/nes96r.dat",
+ header = TRUE)
> names(nes96)[6] <- "partyID"
> Clinton.fit <- polr(ordered(ClinLR) ~ partyID + educ, data = nes96)
> summary(Clinton.fit)
```

```
Call:
polr(formula = ordered(ClinLR) ~ partyID + educ, data = nes96)
```

```
Coefficients:
                Value Std. Error t value
partyID  -0.369      0.0284  -12.98
educ      -0.158      0.0372   -4.26
```

Intercepts:			
	Value	Std. Error	t value
1 2	-4.111	0.235	-17.493
2 3	-2.001	0.205	-9.745
3 4	-0.767	0.196	-3.911
4 5	0.396	0.198	2.002
5 6	1.320	0.217	6.077
6 7	2.455	0.283	8.661

Residual Deviance: 2917.34
AIC: 2933.34

Interpret the coefficients and “Intercepts”.

The *partyID* variable is taken as continuous with estimated slope for our “latent” variable z of -0.369 ($SE = 0.028$). Education, similarly has a slope of -0.158 ($SE = 0.04$). From the t -ratios I see that the *partyID* effect is about 3 times as strong as the education effect. The two combine to create a linear function which gets split into levels by the “intercepts” as in model 6.13 (no overall intercept or we’d have 6.12). With two predictors, I don’t think we can use model 6.11. If z_i is less than -4.1 , we predict a Clinton rating of 1. Between -4.11 and -2 we predict a 2, and so forth until we get to a prediction of 7 if the linear function exceeds 2.46 . The point of having SE ’s attached to these is not to test if one might be zero, but just to indicate how tightly they are estimated. Each consecutive interval is of length 1 to 2, and SE ’s are all $.2$ to $.3$. Intercepts are ordered, and the estimates are not independent.

Prediction:

An independent (*partyID* = 3) with *educ* = 1 would have a score of $-0.3688 \times 3 - 0.1584 \times 1 = -1.26$ and falls into the 3 level of *ClintonLR*, whereas a Republican might have $-0.3688 \times 5 - 0.1584 \times 1 = -2.002$ right on the dividing line between 2 and 3. A strong Democrat would rank him as $-0.1584 \times 1 = -0.16$ (ranking him as a 4, more conservative), and we’re not going to get any rankings in the levels above 4.

This model does not do well in the extremes. To see this, we can compare predicted and observed outcomes using the `table` command:

```
> ttemp <- rbind(table(nes96$ClinLR), table(predict(Clinton.fit)))
> rownames(ttemp) <- c("observed", "predicted")
> xtable(ttemp)
```

	1	2	3	4	5	6	7
observed	109	317	236	160	67	36	19
predicted	0	536	283	125	0	0	0

- Why is overdispersion for logistic regression an issue in Chapter 6, when it was not in Chapter 5?

In Chapter 5 we had Bernoulli responses with a single 0 or 1 in each row of data. In §6 each row represents a cluster of similar responses. The variance should be $n_i p_i (1 - p_i)$, but it could be higher if lurking variables are having unexplained effects. Those could mess with Chapter 5 data as well, but we’re not able to look for it without the replication.