Fox, Module 23 Logistic regression practice problems

** Exercise 1.2: Logistic regression

A generalized linear model with a logit link function and a binomial distribution of the response variable can not be solved by pencil and paper. This exercise uses a *logit transformation of the response variable*, which is similar, though not the same.

- A GLM with a logit link function uses a logit transformation of the fitted response variable. The fitted values are derived by maximum likelihood estimation.
- This exercise uses a logit transformation of the observed response variable. The fitted values are derived by ordinary least squares regression.

An actuary regresses the policy renewal rate on the years insured, using the data in the table below.

Years Insured = explanatory variable	1	2	3	4	5
Renewal Rate = response variable	70%	80%	85%	88%	90%

If the actuary uses classical regression analysis with no transformation of the explanatory variable:

A. What is B, the ordinary least squares estimator of β (the coefficient of years insured)?
B. What is A, the ordinary least squares estimator of α (the intercept of the regression equation)?
C. What is s^2, the ordinary least squares estimator of σ^2_ε (the variance of the error term)?
D. What is the fitted value of the renewal rate for X = 5 years insured?
E. Why is the regression equation a poor predictor?
F. How might a transformation of the response variable improve the analysis?

If the actuary uses a logit transformation of the response variable before fitting the regression equation:

G. What is B, the ordinary least squares estimator of β (the coefficient of years insured)?
H. What is A, the ordinary least squares estimator of α (the intercept of the regression equation)?
I. What is the fitted value of the renewal rate at X = 5?
J. If the response variable has a binomial distribution with 100 observations in each cell, what is the variance of the error term at X = 5?

Part A: The regression equation is $Y_j = \alpha + \beta X_j + \epsilon_j$ (classical regression analysis). We solve for B as

$$B = \sum(x_i - \overline{x})(y_i - \overline{y}) \div \sum(x_i - \overline{x})^2$$

- $\sum(x_i - \overline{x})^2 = 4 + 1 + 0 + 1 + 4 = 10$
- $\sum(x_i - \overline{x})(y_i - \overline{y}) = 0.252 + 0.026 + 0 + 0.054 + 0.148 = 0.480$

So B = 0.480 / 10 = 0.048.

Part B: We solve for A as

$$A = \overline{y} - B \times \overline{x} = 0.826 - 0.048 \times 3 = 0.682$$

Part C: The fitted value for X = 5 is 0.682 + 0.048 × 5 = 0.922

Part D: The least squares estimator s^2 of σ^2_ε = RSS / (n -k -1). RSS is the sum of squared residuals, $\sum(y - \hat{y})^2$

The squared residual for each point are shown below:

Observed X	*Observed Y*	*Fitted Y*	*Residual*	*Squared Residual*
1	0.7000	0.7300	-0.0300	0.000900
2	0.8000	0.7780	0.0220	0.000484
3	0.8500	0.8260	0.0240	0.000576
4	0.8800	0.8740	0.0060	0.000036
5	0.9000	0.9220	-0.0220	0.000484
Total			0.0000	0.002480

The least squares estimator for σ^2_ε is 0.002480 / (5 - 1 - 1) = 0.000827

Part E: The estimated β is 0.048 = 4.8% and the fitted value at 5 years is 0.922 = 92.2%.

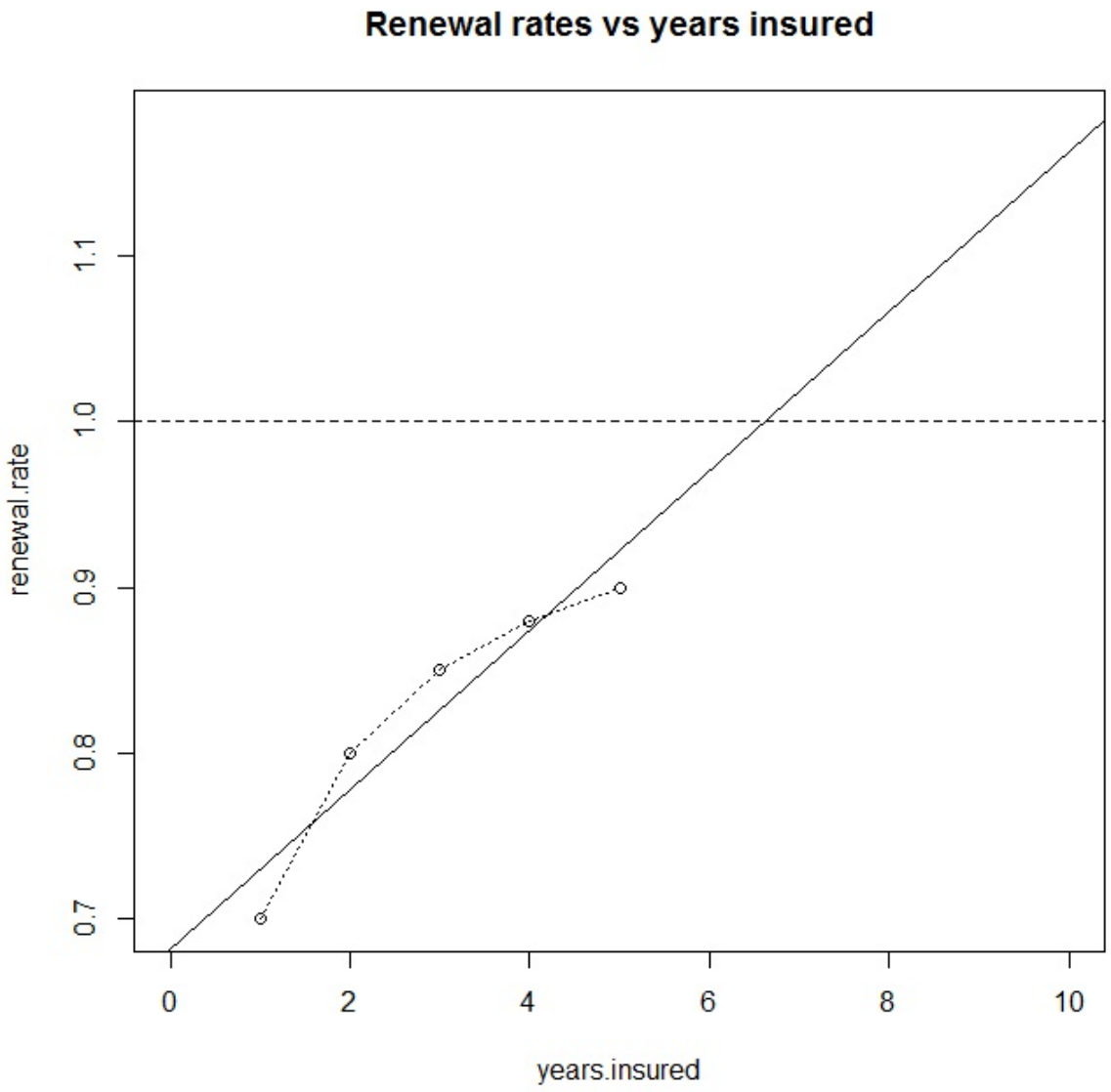
- At 6 years, the fitted value is 92.2% + 4.8% = 97.0%.
- At 7 years, the fitted value is 92.2% + 2 × 4.8% = 101.80%.

At 7+ years insured, the fitted renewal rates are all above 100%. Actual renewal rates can not be above 100% and rarely exceed 95%. The observed renewal rates are reasonable; the fitted renewal rates are not.

The graphic below clarifies the problem.

- The five circles are the five data points.
- The upward sloping diagonal line is the linear regression on renewal rate on years insured.
- The curved dotted line shows the true concave relation of renewal rate to years insured.
- The horizontal dashed line at Y = 1 shows the maximum possible renewal rate.

The concave relation of renewal rate to years insured indicates that the renewal rate flattens as years insured increases and may not exceed 96% even for long-term policyholders. The straight regression line extrapolates the large renewal rate increases from 1 to 2 years insured and 2 to 3 years insured to get a large β (slope coefficient), which wrongly implies renewal rates above 100% for 7 or more years insured. The concave line relating renewal rate to years insured suggests that the renewal rate flattens at about 95%. [An exact figure cannot be inferred from this small graphic, but an upper bound of about 95% seems reasonable.]



Part F: Transformations have two purposes:

- Major purpose: Transformations may change non-linear relations into linear relations.
- Secondary purpose: Transformations may change skewed distributions into symmetric distributions.

The optimal transformation depends on the curvature of the relation between two variables. The Box-Cox transformations and Tokay's ladder of transformations show the general rules; see the textbook chapters and the discussion forum postings on those topics. The logit transformation is specific to response variables that are probabilities.

Part G: The regression equation is $Y_j = \alpha + \beta X_j + \epsilon_j$ (classical regression analysis), but we use the logits (the log odds) of the observed y-values, or $ln(y / (1 - y))$, shown in the table below.

Observed X	*Observed Y*	*Logit of Y*
1	0.7000	0.8473
2	0.8000	1.3863
3	0.8500	1.7346
4	0.8800	1.9924
5	0.9000	2.1972
Mean		1.6316

We solve for B as

$$B = \sum(x_i - \overline{x})(y_i - \overline{y}) \div \sum(x_i - \overline{x})^2$$

- $\sum(x_i - \overline{x})^2 = 4 + 1 + 0 + 1 + 4 = 10$
- $\sum(x_i - \overline{x})(y_i - \overline{y}) = 1.569 + 0.245 + 0 + 0.361 + 1.131 = 3.306$

So B = 3.306 / 10 = 0.3306.

Part H: With the logit transformation, the value of A (the least squares estimator of α) is

$$A = \overline{y} - B \times \overline{x} = 1.6316 - 0.3306 \times 3 = 0.6398$$

Part I: The fitted value of the logit of Y at X = 5 is

$$0.6398 + 0.3306 \times 5 = 2.2928$$

The fitted value of the renewal rate at X = 5 is 1 / (1 + exp(-2.2928)) = 90.828%

Take heed: The logistic regression gives the fitted value of the logit of the renewal rate. To get the fitted value of the renewal rate, we use the logistic function: $y = 1 / (1 + e^{-x})$. Some textbooks show the logistic function after multiplying both numerator and denominator by e^x to give $e^x / (1 + e^x)$.

Part I: The variance of a binomial distribution with probability π and N observations is $\pi \times (1 - \pi) / N$.

If π derived by the logistic regression is 90.828% at x = 5 years insured. With 100 observations at X = 5 years, the variance of the error term is

$$\pi \times (1 - \pi) / 100 = 90.828\% \times (1 - 90.828\%) / 100 = 0.0833\%$$

Jacob: The computation of the variance differs for logistic regression. With no transformation, we compute the RSS over all observations and divide by degrees of freedom. With the logit transformation, the variance depends on the observed value, not on the RSS.

Rachel: The computation of the variance depends on the assumptions.

- Classical regression analysis assumes a normal distribution and the same variance for all points. The variance is independent of the mean, but since it is assumed to be equal for all observations, we derive it from the RSS.

- Logistic regression assumes a binomial distribution for the response variable. The variance differ at each point, so we can't use the RSS, but it is a function of the fitted value. A GLM relates the variance to the fitted value. A transformation with classical regression computes an initial estimate of the variance by assuming the fitted value is close to the observed value.