Chapter 6 - Gaussian Response Models

Peter He

November 15 2019

1 Gaussian Response Models

A Gaussian Response Model is a model for which the distribution of a variate Y, given a vector of covariates $\mathbf{x}= (x_1, \dots, x_k)$ is of the form

$$Y \sim G(\mu(\mathbf{x}), \sigma(\mathbf{x}))$$

where μ and σ are functions of $\mathbf{x}$. From now on, assume $\sigma(\mathbf{x})$ is constant.

1.1 Example

Assume we have a set of data, the size of buildings (in m^2), x, and the selling price in (\$ per m^2) y. We might consider a model where the price of a building of size x_i is represented by a random variable Y_i, with

$$Y_i \sim G(\beta_0 + \beta_1 x_i, \sigma) \qquad \text{for } i = 1, \dots, n \text{ independently}$$

2 Simple Linear Regression

Consider the case where there is a single covariate x, and consider the model with independent Y_i's such that

$$Y_i = G(\mu(x_i), \sigma)$$

where $\mu(x_i) = \alpha + \beta x_i$. The likelihood function for (α, β, σ) is

$$L(\alpha, \beta, \sigma) = \prod_{i=1}^{n} -\frac{1}{2\pi\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right]$$

By finding the log likelihood function, doing partial derivatives, and solving a system of linear equations, we get the maximum likelihood estimates

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}(S_{yy} - \hat{\beta} S_{xy})$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, S_{yy}\sum_{i=1}^{n}(y_i - \bar{y})^2, S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

These values for $\hat{\alpha}$ and $\hat{\beta}$ can also be shown to be the least squares estimates, the parameters for the line of best fit.

2.1 Distribution of the estimator $\hat{\beta}$

The maximum likelihood estimator corresponding to $\hat{\beta}$ is

$$\tilde{\beta} = \frac{1}{S_{xx}}\sum_{i=1}^{n} x_i(Y_i - \bar{Y})$$

So $\tilde{\beta}$ is a random variable with a Gaussian distribution. It can be shown that

$$\hat{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

2.2 Confidence Intervals for β and tests of no relationship

Although we have the maximum likelihood estimate $\hat{\sigma}^2$, we will estimate σ^2 using

$$s_e^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})$$

since it is unbiased. It can be shown that

$$\frac{\tilde{\beta} - \beta}{S_e/\sqrt{S_{xx}}} \sim t(n-2)$$

So, a $100p\%$ confidence interval for β is

$$\hat{\beta} \pm as_e/\sqrt{S_{xx}}$$

where $P(|T| \leq a) = p, T \sim t(n-2)$. To test $H_0 : \beta = 0$, we use the test statistic

$$\frac{|\hat{\beta} - 0|}{S_e/\sqrt{S_{xx}}}$$

so the p-value is given as

$$2\left[1 - P(T \leq \frac{\hat{\beta} - 0}{s_e\sqrt{S_{xx}}})\right]$$

2.3 Confidence Intervals for the mean response $\mu(x) = \alpha + \beta x$

We are interested in estimating the quantity $\mu(x) = \alpha + \beta x$ since it represents the mean response at a specific value of x. The maximum likelihood estimator of $\mu(x)$ is

$$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x = \bar{Y} + \tilde{\beta}(x - \bar{x})$$

We can see that $\tilde{\mu}(x)$ is also a Gaussian random variable:

$$\tilde{\mu}(x) \sim G\left(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

So we have the pivotal quantity

$$\frac{\tilde{\mu}(x) - \mu(x)}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}} \sim t(n-2)$$

which can be used to find confidence intervals for $\mu(x)$ as usual. Thus, a $100p\%$ confidence interval for $\mu(x)$ is given by

$$\hat{\mu}(x) \pm a s_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

where $P(|T| \leq a) = p$.

2.3.1 Example

Assume we have the same data from example 1.1. Recall that we suggested a linear regression model. For the given data, say we have

$$\begin{gathered} n = 30, \bar{x} = 0.9543, \bar{y} = 548.9700 \\ S_{xx} = 22.9453, S_{xy} = -3316.6771, S_{yy} = 489,624.723 \end{gathered}$$

From this, we get

$$\begin{aligned} \hat{\beta} &= \frac{S_{xy}}{S_{xx}} = -144.5459 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 686.9159 \\ s_e^2 &= \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy}) = 364.6199 \\ s_e &= 19.0950 \end{aligned}$$

We can plot $y = \hat{\alpha} + \hat{\beta}x$, which is called the fitted regression line.

Say a building, with the size of the building being $x = 4.47m^2$, costs \$75. We can use the model to determine if this price is too high. Using our model, our expected price is

$$\hat{\mu}(4.47) = \hat{\alpha} + \hat{\beta}(4.47) = \$40.79$$

which is much lower than \$75. What if the person trying to sell you this building says there is uncertainty in this estimate, and wants a confidence interval for $\mu(4.47)$ instead? To find a 95% confidence interval, we have that $P(T \leq 2.0484) = 0.975$ for $T \sim t(28)$, so we get

$$\begin{aligned}\hat{\mu}(4.47) &= \pm 2.0484 s_e\sqrt{\frac{1}{30} + \frac{(4.47-\bar{x})^2}{S_{xx}}} \\ &= \$40.79 \pm \$20.58 \\ &= [\$11.21, \$70.37]\end{aligned}$$

for which \$75 is outside the interval.

2.4 Prediction Interval for an Individual Response, Y at x

Before, these confidence intervals were for the expected value of Y for some value x. What if we wanted an interval for an individual for a certain value of x? All else being equal, prediction intervals are wider. We get that

$$Y - \tilde{\mu}(x) \sim G(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}})$$

Since

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$$

, we have

$$\frac{Y - \tilde{\mu}(x)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

Our estimate is still the same as before, $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$. The resulting prediction interval is

$$\hat{\alpha} + \hat{\beta}x \pm aS_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

This interval is called a *prediction interval* for Y, instead of a confidence interval for $\mu(x)$. This is because Y is not a parameter, but a future observation.

2.4.1 Example

Assume we have the same data from example 1.1. Let us find a 95% prediction interval for Y when $x = 4.47$. From $P(T \leq 2.0484) = 0.975$ when $T \sim t(28)$, we obtain

$$\begin{aligned} &\tilde{\mu}(4.47) \pm 2.0484 s_e \sqrt{1 + \frac{1}{30} + \frac{(4.47 - \bar{x})^2}{22.945}} \\ &= \$40.79 \pm \$49.04 \\ &= [-\$8.25, \$89.83] \end{aligned}$$

2.5 Confidence Intervals for the Variance

The pivotal quantity $\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$ can be used to create confidence intervals for σ^2. A $100p\%$ confidence interval for σ^2 is given by:

$$\left[\frac{(n-2)S_e^2}{b}, \frac{(n-2)S_e^2}{a}\right]$$

where $P(U \leq a) = \frac{1+p}{2} = P(U > b)$, where $U \sim \chi^2(n-2)$.

2.6 Summary

- $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ is the maximum likelihood estimate of β
- $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ is the mle of α
- $s_e^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})$ is the unbiased estimate of σ^2
- $\frac{\tilde{\beta}-\beta}{S_e\sqrt{S_{xx}}} \sim t(n-2)$ is used to find confidence intervals for β and to conduct hypothesis tests for $H_0 : \beta = 0$
- $\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$ is used to find confidene intervals for $\mu(x)$
- $\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$ is used to find prediction intervals for an individual response, Y at some value x

3 Checking the Assumptions

There are two main components in Guassian linear response models:

1. The assumption that Y_i (given any covariates x_i) is Gaussian with constant standard deviation σ.

2. The assumption that $E(Y_i) = \mu(x_i)$ is a linear combination of observed covariates with unknown coefficients.

If there is only one x covariate, a plot of the fitted line superimposed on the scatterplot of the data will suffice. If there are two or more, we use residual plots to check the model assumptions. Three plots that can be use are:

1. Plot points $(x_i, \hat{r}_i), i = 1, \ldots, n$. If the model is satisfactory if the points lie more or less horizontally within a constant band around the line $\hat{r}_i = 0$. In other words the points look randomly scattered.

2. Plot points $(\hat{\mu}_i, \hat{r}_i)$. Same idea as before.

3. Check the qqplot of the residuals $\hat{r}_i$.

We often prefer to use standardize residuals, $\hat{r}_i^* = \frac{\hat{r}_i}{s_e}$, which scale the residuals down.

- If there is fanning out of funneling in, the constant variance assumption has been violated.

- If there is a happy or sad (quadratic) pattern, the linearity assumption has been violated

- If there are any systemic patterns, independence assumption has been violated.

3.1 Sample Correlation Coefficient

A simple one number summary, which measures the strength of a linear association (only reliable if the asusmptions hold). For a data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we define

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

r has no units and is bounded between -1 and 1. The correlation between x and y is the same between y and x. r is also relation to $\hat{\beta}$:

$$\hat{\beta} = r\frac{s_y}{s_x}$$

4 Comparing the Means of Two populations

We will compare means from two different populations using confidence intervals and hypothesis testing. We say population j has mean μ_j and variance σ_j^2 for $j = 1, 2$. Similarly, define the same for $n_j, \bar{y}_j$, and s_j^2. When dealing with two populations, infering about the relationship between their means will mean we are inferring about their difference $\mu_1 - \mu_2$. The type of confidence interval or test statistic to be used depends on the information given from the two populations. There are four main cases. The first three assume independent samples.

1. σ_1 and σ_2 are known.
2. σ_1 and σ_2 are unknown, $\sigma_1 = \sigma_2$
3. σ_1 and σ_2 are unknown, $\sigma_1 \neq \sigma_2$
4. Paired (dependent) samples.

We will mainly focus on the last three cases, where σ_1 and σ_2 are unknown.

4.1 General Model - Notation

The hypothesis of interest is $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$. We will let

- $Y_i = Y_{1i}$ for $i = 1, 2, \ldots, n_1$
- $Y_{n_1+i} = Y_{2i}$ for $i = 1, 2, \ldots, n_2$
- $E(Y_i) = \mu_1$ for $i = 1, \ldots, n_1$
- $E(Y_{n_1+i}) = \mu_2$ for $i = 1, \ldots, n_2$, and
- $Var(Y_i) = \sigma^2$ for $i = 1, \ldots, n_1 + n_2$

In doing so, we see that this model is just a special case of the Gaussian response model

$$Y_i \sim G(\mu(x_i), \sigma)$$

where $\mu(x_i) = \mu_1$ for $i = 1, \ldots, n_1$ and $\mu(x_i) = \mu_2$ for $i = n + 1, \ldots, n_1 + n_2$. The natural estimator for $\mu_1 - \mu_2$ is $\bar{Y}_1 - \bar{Y}_2$. If Y_1 and Y_2 are both normally distributed, then

$$\bar{Y}_1 = \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

4.2 σ_1 and σ_2 known, independent samples

If the two samples are independent of each other, then an exact two-sided $100p\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{(1+p)/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

4.3 σ_1 and σ_2 are unknown, assumed to be equal

Define the pooled estimator of the variance as

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 - 1 + n_2 - 1}$$

It can be shown that

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$$

, so we get

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

4.3.1 Confidence Intervals for a Difference in Means

By the previous result, a $100p\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{y}_1 - \bar{y}_2 \pm a s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

where $P(T \le a) = \frac{1+p}{2}$ and $T \sim t(n_1+n_2-2)$

4.3.2 Tests of Hypotheses for No difference in Means

We use the test statistic $D = \frac{|\bar{Y}_1 - \bar{Y}_2 - 0|}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ to test $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$. The observed value of the test statistic is $d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, and the p-value is calculated as $2[1 - P(T \le d)]$ where $T \sim t(n_1+n_2-2)$.

4.3.3 Example - Airbag

An important property of air bags is the permeability of the woven fabric. Question: is there any difference in permeability at 0°C and 20°C? A researcher takes two independent samples of permeability measurements and obtains the following results:

0°C	20°C
70	40
85	60
92	50
80	45
60	

We would like to know if temperature if having an effect. Assume that the variances are roughly equal. To conduct a hypothesis test, we first need a model.

Let Y_{1i} be the permeability of the ith airbag at 0°C for $i = 1, 2, 3, 4, 5$, and let Y_{2i} be the same at 20°C for $i = 1, 2, 3, 4$. Assume Y_{1i}, $i = 1, 2, 3, 4, 5$ is a random sample from $G(\mu_1, \sigma)$ and independently $Y_{2i}, i = 1, 2, 3, 4$ is a random sample from $G(\mu_2, \sigma)$ distribution. From the data, we find that

- $\bar{y}_1 = 77.4, s_1^2 = 158.80$
- $\bar{y}_2 = 48.75, s_2^2 = 72.92$

By the table, $P(T \leq 2.3646) = 0.975$ when $T \sim t(5 + 4 - 2) = t(7)$. We also find that

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(4)(158.8) + (3)(72.92)}{7} \\ &= 121.99 \end{aligned}$$

So, our confidence interval is

$$\begin{aligned} \bar{y}_1 - \bar{y}_2 + as_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &= 77.4 - 48.75 \pm \sqrt{121.99}\sqrt{\frac{1}{5} + \frac{1}{4}} \\ &= 28.65 \pm (2.3646)(11.04)(0.45) \\ &= 28.65 \pm 17.52 \\ &= [11.13, 46.17] \end{aligned}$$

Since 0 is outside the interval, the p-value must be less than 5%. To find the exact p-value, we use the test statistic

$$\frac{|\bar{Y}_1 - \bar{Y}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

. D has a t-distribution with degrees of freedom being $n_1 + n_2 - 2$ assuming $H_0 : \mu_1 - \mu_2 = 0$ is true. So,

$$\begin{aligned} d &= \frac{|\bar{y}_1 - \bar{y}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{|77.4 - 48.75|}{11.04\sqrt{0.2 + 0.25}} \\ &= 3.87 \end{aligned}$$

Thus,

$$\begin{aligned} p - value &= 2[1 - P(T \leq 3.87)] \qquad \text{where } T \sim t(n_1 + n_2 - 2) \\ &= 2[1 - 0.995] \\ &= 0.01 \end{aligned}$$

According to Table 5.1 guidelines, we have very strong evidence against H_0 based on the observed data. In the context of this example, this suggest that it appears as though temperature has an effect on the permeability of airbags.

4.4 Unequal and Unknown Variances

If n_1 and n_2 are both large, then the pivotal quantity

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \dot{\sim} G(0,1)$$

can be used to construct confidence intervals and test hypotheses for the mean difference $\mu_1 - \mu_2$. For example, an approximate 95% confidence interval for $\mu_1 - \mu_2$ would be given by

$$\bar{y}_1 - \bar{y}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

4.4.1 Example - 5th grade math test scores

Say we were given the data for test scores:

- District 1: $n_1 = 278, \bar{y}_1 = 60.2, s_1 = 10.16$
- District 2: $n_2 = 375, \bar{y}_2 = 58.1, s_1 = 9.02$

Assume that the scores from District 1 have a $G(\mu_1, \sigma_1)$ distribution and the scores from District 2 have a $G(\mu_2, \sigma_2)$ distribution. Assume $\sigma_1 \neq \sigma_2$.
Since n_1 and n_2 are far greater than 30, we can construct an approximate 95% confidence interval for $\mu_1 - \mu_2$ as follows:

$$\begin{aligned}
&60.2 - 58.1 \pm 1.96\sqrt{\frac{10.16^2}{278} + \frac{9.02^2}{345}} \\
&= 2.1 \pm 1.527 \\
&= [0.573, 3.627]
\end{aligned}$$

Since 0 is not in the interval, it appears District 2 kids are even worse at math than District 1 kids.

4.5 Paired Data

Often, studies designed to compare means are conducted with pairs of units, where the responses within a pair are not independent. For example we can have the i-th student take two tests, and compare the means of the two tests. If student Y_i does well on test 1, we would expect that they would do well on test 2 as well, so we would expect $Cov(Y_{1i}, Y_{2i}) > 0$. It turns out if Y_1 and Y_2 are dependent Normal variables, then the random variable $Y_1 - Y_2$ is also Normal with $E(Y_1 - Y_2) = \mu_1 - \mu_2$. We can take the difference in these test scores and end up with a single sample of differences. If we do this and obtain $\bar{y}$ and s^2, we can use the test statistic

$$D = \frac{|\bar{Y} - 0|}{S/\sqrt{n}}$$

to test $H_0 : \mu = 0$. The main advantage to pairing is that if the data is positively correlated, we reduce the variability, and the confidence intervals get shorter. Make sure that $\sigma_{12} = Cov(Y_{1i}, Y_{2,i}) > 0$!

5 Errors in Hypothesis testing

As an aside, there are two types of errors in testing:

1. Type 1 Error: We reject H_0 when it is actually true. $P(\text{Type 1 Error}) = \alpha$

2. Type 2 Error: We fail to reject H_0 when it is actually false. $P(\text{Type 2 Error}) = \beta$