Contents

5 Regression

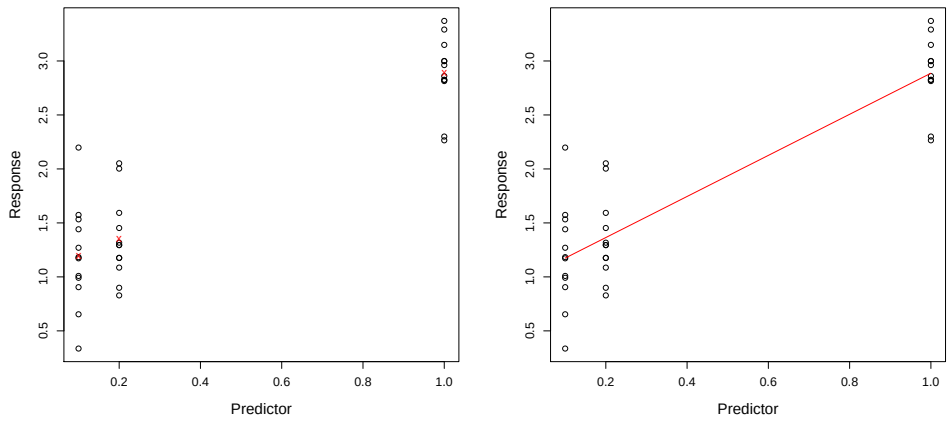
5.1 Introduction

A major goal of statistics is to identify the relationship between *predictor* variables and *response* variables. We first started this quest when we studied samples from two populations, say A and B. In this case, the predictor variable (though we did not mention it) is merely a variable indicating whether the observed response came from population A or B. If there is a difference in population means, μ_A and μ_B, then knowing whether the predictor indicates population A or B provides valuable information about the data Y_i.

In ANOVA, we generalized to multiple (> 2) samples or treatments. In this case, the predictor variable indicates one of several treatments. Sometimes the treatments may be represented by numeric values, for example the initial conditions in the virus data (HW3), but we have never actually *used* these numeric values for anything but to characterize the treatment.

So, up until now, the relationship between predictor an response has been simple. The predictor simply chooses the mean of the response from a finite collection of treatment means. In Fig. 5(a), the ANOVA analysis fits a population mean to all three levels of the predictor variable (small red "x"'s).

In regression, a precise functional relationship is postulated to exist between numeric predictors and the numeric response. As for ANOVA, the predictor indicates the mean of the response, but the function predicts the mean for *any* value of the predictor, not just the values used in the experimental treatments. For example, the simple regression model posits a linear relationship between the predictor and response (red line of



(a) (b)

Figure 5: Comparison of ANOVA and Simple Linear Regression. (a) ANOVA fits one population mean to each level of the predictor. (b) Regression postulates a *functional relationship*, in this case a line, between a numeric predictor and the mean numeric response.

Fig. 5(b)).

We shall focus predominately on *linear* regression, where the functional relationship is linear in the population parameters.

Two Parts to Linear Regression

As with any statistical model, we make assumptions in order to accomplish the three goals (estimate, inference, predict) of statistics. It takes relatively few assumptions to estimate parameters, but far more assumptions to do inference or prediction.

1. **Estimation:** Unbiased errors, $E[\epsilon] = 0$, needed to summarize the data and estimate model parameters.
2. **Inference:** Distributional assumptions (normality) required to derive estimator properties, test hypotheses, estimate confidence intervals.

5.2 Simple Linear Regression

5.2.1 Model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

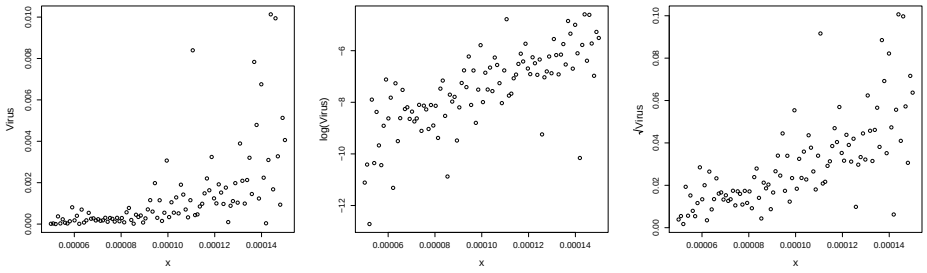
where Y_i is a random (*independent* or *response*) variable, x_i are known or observable (*dependent* or *predictor*) variables, β_0, β_1 are fixed, unknown population parameters, and $E[\epsilon_i] = 0$.

Given the last assumption, the *population regression function* is

$$E[Y_i \mid x_i] = \beta_0 + \beta_1 x_i$$

where we condition on x_i in the expectation to make the dependence on x_i explicit. Note, however, that x_i is *not* (yet) a random variable, so the conditional notation is abusive. (Later, we will generalize to random predictor X_i, where the conditional expectation is legitimate.)

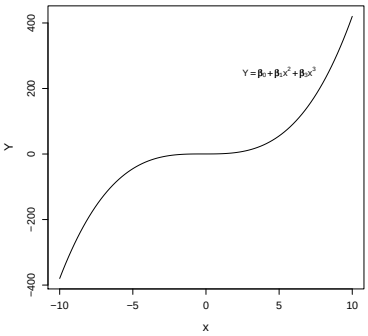
There are two population parameters β_0 and β_1. Further, the population regression function $\beta_0 + \beta_1 x$ is linear in the population parameters.



(a) raw data (b) log transformed data (c) square-root transformed data

Figure 6: Virus data transformed

Linear Model



What does it mean for the model to be linear? It is more flexible than you may think. A model is linear if the regression function is a linear function in the unknown parameters. The figure is the linear regression function $\beta_0 + \beta_1 x^2 + \beta_2 x^3$, which is nonlinear in the predictors, but not the parameters. The following are all linear models.

$$\begin{aligned} E[Y_i \mid x_i] &= \beta_0 + \beta_1 x_i \\ E[\log(Y_i) \mid x_i] &= \beta_0 + \beta_1 x_i^2 \\ E[\text{logit}(Y_i) \mid x_i] &= \beta_0 + \beta_1 / x_i \end{aligned}$$

Notice, both the response and the predictors can be modified in nonlinear ways. The following are not valid right-hand-sides for the simple linear regression model:

$$\begin{aligned} &\beta_0 + \beta_1^2 x_i \\ &\beta_0 + \log(\beta_1) x_i \\ &\text{...etc....} \end{aligned}$$

Is it reasonable to assume a linear model? No, probably not. The available theory may not support a linear model, however it is very convenient mathematically and often does a remarkable job at approximating what is actually complex and nonlinear. *We assume a linear relationship adequately captures the true relationship.*

$$E[Y_i \mid x_i] \approx \beta_0 + \beta_1 x_i$$

5.2.2 Least Squares Estimation

Problem Setup

We observe $(x_1, y_1), \ldots, (x_n, y_n)$ and hypothesize a linear relationship between X and Y. The points will not fall on a perfect straight line because of randomness, but we "summarize" the relationship as a line that is somehow best characterizing the relationship.

Example: A mathematical model of virus growth in cell culture was developed. Unfortunately, it is sufficiently complex that the actual formula relating parameters and response is unknown. In other words, the model can be viewed as a "black box", taking in parameters and putting out responses. A sensitivity analysis is a statistical method for detecting how parameters affect a response, which is particularly useful for "black box" models. Parameters are varied methodically over a range and the response produced by the model is regressed on the parameters, viewed in this context as predictors. A plot of the response (amount of virus at day 2 of the cell culture) in relation to one predictor, the fraction of produced virus that are infectious, generically called x in Fig. 6.

This data shows a classic pattern of increasing variability with predictor level. The log transform or square-root transform can fix the non-constant variance (see Fig. 6). Perhaps the log transform is better in this case.

To summarize the data, we can compute the summary statistics

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} y_i \qquad \text{sample means}$$

$$S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad S_{YY} = \textstyle\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{sums of squares}$$

$$S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \text{sum of cross produces}$$

Least Squares Estimates of β_0, β_1

We seek a straight line that comes as close as possible to all points in some sense. We need not make any statistical assumptions to come up with this line. It is a purely mathematical argument.

Definition: (regression) residual

Let $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ be the difference between the observed y_i and that predicted $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ by the estimated regession equation.

Definition: residual sum-of-squares

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2$$

Sometimes, statisticians write SSE (sum of squared error) for RSS.

Definition: least-squares estimators

The least squares estimators of β_0 and β_1 are those which minimize RSS

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{\beta_0,\beta_1} RSS$$

We will find formula for the LS estimators, but first we prove a lemma that will help.

Lemma 21. *Consider $x_1, \ldots, x_n$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then, $\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$.*

Proof.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2
\end{aligned}$$

and clearly we can see that $a = \bar{x}$ minimizes this sum. □

Theorem 22. *The least squares estimators are*

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

Proof. It is clear that $\hat{\beta}_0 = \overline{y_i - \beta_1 x_i} = \bar{y} - \beta_1 \bar{x}$. To minimize over β_1, we are left with

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \beta_1 x_i - \bar{y} - \beta_1 \bar{x})^2 &= \sum_{i=1}^{n} [(y_i - \bar{y}) - \beta_1 (x_i - \bar{x})]^2 \\ &= S_{YY} - 2\beta_1 S_{XY} + \beta_1^2 S_{YY} \end{aligned}$$

Take the derivative and set to 0,

$$\begin{aligned} \frac{\partial}{\partial \beta_1} \left(S_{YY} - 2\hat{\beta}_1 S_{XY} + \hat{\beta}_1^2 S_{XX} \right) &= 0 \\ 2\hat{\beta}_1 S_{XX} - 2S_{XY} &= 0 \\ \hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \end{aligned}$$

Taking second derivatives verifies that this is a minimum. □

Example: estimating $\hat{\beta}_0$ and $\hat{\beta}_1$

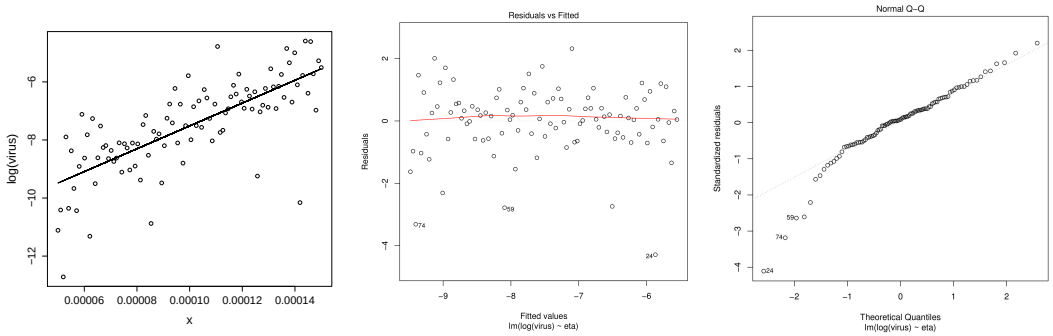
Here is a `R` session demonstrating how to fit the virus data. The fit is shown in Fig. 7(a).

```
> d <- read.table("etaVvirus.txt",header=T)
> head(d, n=2)
           eta        virus
1 9.141414e-05 0.0007092100
2 1.500000e-04 0.0040598946
> v.fit <- lm(log(virus) ~ eta, data=d)
> summary(v.fit)

Call:
lm(formula = log(virus) ~ eta, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2893 -0.5040  0.0906  0.6029  2.3200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -11.4494     0.3788  -30.22   <2e-16 ***
eta         39323.1036  3636.9549   10.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.06 on 98 degrees of freedom
Multiple R-squared: 0.544,Adjusted R-squared: 0.5393
F-statistic: 116.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

Applying `plot()` to the fitted object produces a number of diagnostic plots. The residuals plot Fig. 7(b) shows no obvious trends in the residuals, but there are a few very negative residuals that `R` highlights as outliers. The probability plot Fig 7(c) shows these outliers are a part of an over-extended left tail relative to the normal distribution.

Alternative Estimation

There are other ways to pick a best fitting line. Perhaps we could minimize the sum of squared horizontal distances between points and the line. Or, we could use another distance measure.



(a) Fit (b) Residuals (c) Probability Plot

Figure 7: Diagnostic plots for the virus fit shown in (a). (b) plots residuals against the fitted values $\hat{Y}_i$. (c) is a probability plot of the residuals.

However, if x_i is predictor and y_i is the response, then it is reasonable to try to minimize the distance from our prediction (the estimated regression line) to the observed y_i.

How good is the fit?

The total variability in the data Y, as measured by sums of squares can be partitioned into a contribution coming from the variation in the regression function itself and variation in the measurements around the regression function. The partitioning equation is

$$\text{total sum of squares} = \text{regression sum of squares} + \text{residual sum of squares}$$

where the residual sum of squares (RSS) captures the extra variation observed in the data around the fitted line, and the regression sum of squares represents the amount of variability in Y_i captured by the line. For example, if $\hat{y}_i = \bar{y}$ for all i, then the fitted line is flat and regression sum of squares is 0. In this case, the line does not account for any of the variability in Y. On the other hand if RSS$= 0$, then the points lie on a straight line and the line itself explains all of the variation in Y, i.e. knowing x_i, we can perfectly predict y_i.

The same equation in mathematical notation is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

and can also be shown to be

$$S_{YY} = \frac{S_{XY}^2}{S_{XX}} + \text{RSS}.$$

Coefficient of Determination

Definition: coefficient of determination

The coefficient of determination r^2 is

$$r^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{S_{XY}^2}{S_{XX}S_{YY}}$$

is a measure of the fraction of total variation in Y that can be explained by the regression function.

Notice that $r^2 \in [0, 1]$. $r^2 = 1$ when there is a perfect linear relationship between x and Y. $r^2 = 0$ when the best fitting regression line has $\hat{\beta}_1 = 0$, i.e. there is no relationship between x and Y, so that Y is independent of level x.

Interpretting β_0 and β_1

Recall that $\beta_0 = E[Y \mid x = 0]$ is the expected observation when the predictor $x = 0$. This quantity is usually less meaningful than β_1, which is the slope of the line. The slope tells use how we expect Y to change given a particular change in the level x.

5.2.3 Inference

We now add additional assumptions about the model in order to perform inference.

The Conditional Normal Model

Suppose

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

for all $i = 1, \ldots, n$ are independent. (Notice, they are no longer iid, because they are not "identically distributed" because the mean is changing.)

Equivalently, we could state $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and impose restriction

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

These models include the following hypotheses:

1. Independent observations
2. Common variance for all levels x_i
3. Normal distribution
4. Linear mean

Likelihood

Now that we have a fully specified probability model for our data $y = (y_1, \ldots, y_n)$ (now we use small y to indicate that these data are an observed realization of the random vector Y with elements Y_i), we can write down the likelihood

$$\begin{aligned}
f(y \mid \beta_0, \beta_1, \sigma^2) &= f(y_1, \ldots, y_n \mid \beta_0, \beta_1, \sigma^2) \\
&= \textstyle\prod_{i=1}^n f(y_i \mid \beta_0, \beta_1, \sigma^2) && \text{by independence} \\
&= \textstyle\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right] && \text{normal distribution} \\
&= \textstyle\frac{1}{\sigma^2 (2\pi)^{n/2}} \exp\left[-\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 / (2\sigma^2)\right]
\end{aligned}$$

Maximum Likelihood Estimation

With a likelihood in hand, the MLEs become accessible to us. Recalling the procedure, we compute first the log likelihood.

$$\log f(y \mid \beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

To maximize the log likelihood we start by observing that only the last term involves β_0 and β_1, so to maximize with respect to β_0 and β_1 is equivalent to *minimizing* the last sum

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Yet, we minimized this expression to obtain our LS estimates, thus the mles are

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}}\end{aligned}$$

To find the mle of σ^2, we take the partial derivative and set it equal to 0. The result is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Thus, we have justified the LS estimates from another direction. However, note that the LS estimates exist without assuming the normal conditional model.

Unbiased Sample Variance

The MLE estimate $\hat{\sigma}^2$ is biased. An unbiased estimator is

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We can remember this formula by recognizing there are n data points and 2 are lost to estimate the parameters β_0 and β_1. Thus, if we divide the sum-of-squares by the remaining degrees of freedom $n-2$, we get a unbiased estimator.

Sampling Distributions

Theorem 23. *Assume the conditional normal distribution, then sampling distributions of $\hat{\beta}_0, \hat{\beta}_1$, and S^2 have the following properties:*

1. $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{nS_{XX}}\sum_{i=1}^{n} x_i^2\right)$
2. $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$
3. $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{X}}{S_{XX}}$
4. $(\hat{\beta}_0, \hat{\beta}_1)$ *and* S^2 *are independent*
5. $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Proof. This is a partial proof.

First, notice that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of independent normal random variables, therefore they are normally distributed. We need to find their mean and variance.

We prove that the LS estimates are unbiased by using $E[y_i] = \beta_0 + \beta_1 x_i$.

$$\begin{aligned}E[\hat{\beta}_0] &= \frac{1}{n(\sum_i x_i^2) - (\sum_i x_i)^2}\left[\left(\sum_i x_i^2\right)\sum_i E[y_i] - \left(\sum_i x_i\right)\sum_i x_i E[y_i]\right] \\ &= \frac{1}{n(\sum_i x_i^2) - (\sum_i x_i)^2}\left[\left(\sum_i x_i^2\right)\left(n\beta_0 + \beta_1\sum_i x_i\right) - \left(\sum_i x_i\right)\left(\beta_0\sum_i x_i + \beta_1\sum_i x_i^2\right)\right] \\ &= \beta_0\end{aligned}$$

Similarly, $\hat{\beta}_1$ is unbiased. Notice that we have only used $E[\epsilon_i] = 0$ to get this result.

The variance of any linear estimator $\sum_{i=1}^n d_i Y_i$ when the sample $Y_1, \ldots, Y_n$ is iid, with common variance $\text{Var}(Y_i) = \sigma^2$, is

$$\text{Var}\left(\sum_{i=1}^n d_i Y_i\right) = \sum_{i=1}^n d_i^2 \text{Var}(Y_i) = \sigma^2 \sum_{i=1}^n d_i^2$$

Using the variance formula for linear estimators, we have

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{S_{XX}^2} = \frac{\sigma^2}{S_{XX}}$$

and

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{nS_{XX}} \sum_{i=1}^n x_i^2$$

To show independence, we notice

$$\hat{\epsilon}_i = \sum_{j=1}^n \left[\delta_{ij} - (c_j + d_j x_j)\right] Y_i$$

where $\delta_{ij} = 1$ iff $i = j$ and c_j and d_j are the fixed coefficients for the linear estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.

To compute covariance, use the lemma for covariance of two linear combinations of random variables.

You can show, using the same lemma, that $\text{Cov}(\hat{\epsilon}_i, \hat{\beta}_0) = \text{Cov}(\hat{\epsilon}_i, \hat{\beta}_1) = 0$, which implies independence of $\hat{\epsilon}_i$ with $\hat{\beta}_0$ and $\hat{\beta}_1$. Therefore, S^2, which is a function of $\hat{\epsilon}_i$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$. □

Hypothesis Testing for $\hat{\beta}_0$

Because of the above theorem, we can immediately test

$$H_0 : \beta_0 = \beta$$

using statistic

$$t_0 = \frac{\hat{\beta}_0 - \beta}{S\sqrt{\sum x_i^2 / nS_{XX}}} \sim t_{n-2}$$

by independence of $\hat{\beta}_0 \sim N(\cdot, \cdot)$ and $S^2 \sim \chi^2_{n-2}$. The details follow exactly that of the derivation of Student's t-test.

Similarly, hypothesis

$$H_0 : \beta_1 = \beta$$

can be tested with statistic

$$t_1 = \frac{\hat{\beta}_0 - \beta}{S/\sqrt{S_{XX}}} \sim t_{n-2}$$

Generally, although $\hat{\beta}_0$ and $\hat{\beta}_1$ are not independent, they are tested separately against these marginal distributions. Together (t_0, t_1) follows a so-called bivariate t-distribution, but this distribution is not often used for inference.

Confidence Intervals

Confidence intervals are trivial to compute. Suppose $t_{n-2,\alpha/2}$ is the critical value for the upper tail of a t-distribution with $n-2$ degrees of freedom. Then, a $100(1-\alpha)\%$ confidence interval for β_1 is

$$\hat{\beta}_1 - t_{n-2,\alpha/2} \frac{S}{\sqrt{S_{XX}}}, \hat{\beta}_1 + t_{n-2,\alpha/2} \frac{S}{\sqrt{S_{XX}}}$$

Estimating Population Mean at Unobserved Level x_0

Suppose we observe $Y_1, \ldots, Y_n$ for levels $x_1, \ldots, x_n$ and we want to predict the expected of Y at previously unobserved x_0. For example, we may wish to predict the expected score in STAT430 given the score in STAT330.

We know $E[Y \mid x_0] = \beta_0 + \beta_1 x_0$ is the population mean at x_0. A reasonable estimate is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0$$

In fact, this estimate is unbiased

$$E[\hat{\beta}_0 + \hat{\beta}_1 x_0] = E[\hat{\beta}_0] + E[\hat{\beta}_1] x_0 = \beta_0 + \beta_1 x_0.$$

Furthermore,

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \text{Var}(\hat{\beta}_0) + x_0^2 \text{Var}(\hat{\beta}_1) + 2x_0 \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \cdots = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)$$

Finally, since $\hat{\beta}_0, \hat{\beta}_1$ are linear estimators, then reproductive property indicates

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N \left(\beta_0 + \beta_1 x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right] \right)$$

Because S^2 is independent of $(\hat{\beta}_0, \hat{\beta}_1)$, it is also independent of $\hat{\beta}_0 + \hat{\beta}_1 x_0$, so we can proceed to hypothesis testing, e.g. testing

$$H_0 : \beta_0 + \beta_1 x_0 = \beta$$

using

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - \beta}{S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}} \sim t_{n-2}$$

or confidence interval construction

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}, \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

Experimental Design in Estimation of $\beta_0 + \beta_1 x_0$

Notice that the width of the confidence interval depends on $x_1, \ldots, x_n$, i.e. the levels we selected in our experimental design at which to general sample data $Y_1, \ldots, Y_n$. This observation implies that we can modify the experimental design in order to get tighter confidence bounds on our estimate of the population mean

$$\beta_0 + \beta_1 x_0$$

at the untried value x_0.

In particular, we can reduce the variance by increasing our sample size n and by selecting $x_1, \ldots, x_n$ such that their average $\bar{x} = x_0$ or $\bar{x} \approx x_0$.

Predicting Observation at Unobserved Level x_0

Suppose we want to predict an unobserved random variable Y. For example, suppose we want to predict the performance in STAT430 of a student who received 92% in STAT330. Using the above procedure for estimating means, we can predict the mean performance of all STAT430 students with 92% in STAT330, but how do we predict the performance of a single such student?

Definition: prediction interval

A $100(1-\alpha)\%$ *prediction interval* for an unobserved random variable Y based on observed data X is a random interval $[L(X), U(X)]$ with

$$P(L \leq Y \leq U) \geq 1-\alpha$$

Let us suppose we want to predict observation Y_0 at $x = x_0$ having observed already $(x_1, y_1), \ldots, (x_n, y_n)$. Note that Y_0 is independent of previous data and thus independent of the estimates $\hat{\beta}_0, \hat{\beta}_1, S^2$.

$Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0$ has a normal distribution by the reproductive property, with

$$\begin{aligned} \mathrm{E}[Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0] &= \beta_0 + \beta_1 x_0 - \beta_0 - \beta_1 x_0 = 0 \\ \mathrm{Var}(Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) &= \mathrm{Var}(Y_0) + \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right) \end{aligned}$$

Because S^2 and $Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0$ are independent, confidence intervals and t-tests can be constructed in the usual fashion. Note, the variance for predicting unobserved Y_0 is bigger than the variance for estimating the mean $\beta_0 + \beta_1 x_0$ at a novel level x_0. There is extra uncertainty in trying to preict a random variable such as Y as compared to a population parameter $\beta_0 + \beta_1 x_0$ that explains this difference.

Bivariate Normal Distribution

The bivariate normal distribution plays a prominent role in some forms of simple linear regression, so we shall describe it now.

Definition: bivariate normal distribution A length 2 vector of random variables has a bivariate normal distribution, and we write

$$(X, Y) \sim BivariateNormal(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho),$$

if the joint pdf is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp\left\{ \frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \frac{x-\mu_X}{\sigma_X} \frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right\}$$

Notice that $X \in (-\infty, \infty), Y \in (-\infty, \infty), \mu_X \in (-\infty, \infty), \mu_Y \in (-\infty, \infty), \sigma_X \in (0, \infty), \sigma_Y \in (0, \infty)$. Also, one can derive the bivariate normal distribution by hypothesizing two independent standard normal Z_1, Z_2 variables and applying the following change-of-variable:

$$\begin{aligned} X &= \sqrt{\frac{1+\rho}{2}}\sigma_X Z_1 + \sqrt{\frac{1-\rho}{2}}\sigma_X Z_2 + \mu_X \\ Y &= \sqrt{\frac{1+\rho}{2}}\sigma_Y Z_1 - \sqrt{\frac{1-\rho}{2}}\sigma_Y Z_2 + \mu_Y \end{aligned}$$

Properties of Bivariate Normal

1. $X \sim N(\mu_X, \sigma_X^2)$
2. $Y \sim N(\mu_Y, \sigma_Y^2)$
3. $\mathrm{Corr}(X, Y) = \rho = \frac{\mathrm{Cov(X,Y)}}{\sigma_X \sigma_Y}$
4. $aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$
5. $Y \mid X \sim N(\mu_X + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1-\rho^2))$

Bivariate Normal Model

Suppose you observe an iid sample of bivariate normal random vectors $(X_i, Y_i) \overset{\text{iid}}{\sim} \text{BivN}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Notice that now the levels X_i are themselves random.

Then, we know

$$\begin{aligned} \text{E}[Y \mid X] &= \mu_y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X) \\ \text{Var}(Y \mid X) &= \sigma_Y^2(1 - \rho^2), \end{aligned}$$

which implies

- The conditional expectation is linear with $\beta_0 = \mu_y - \frac{\rho\sigma_Y\mu_X}{\sigma_X}$ and $\beta_1 = \frac{\rho\sigma_Y}{\sigma_X}$.
- The variance is independent of level X.

For all intents and purposes, the bivariate normal model reduces to the Conditional Normal Model. Therefore, henceforth, we will focus on the conditional normal model and keep in mind that it also applies for the bivariate normal model.

5.3 Multiple Linear Regression

We now extend our linear regression model to include multiple predictors

$$E\left[Y_i \mid x_{i\cdot}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}.$$

for p population parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ and $p-1$ predictors $x_{i1}, x_{i2}, \ldots, x_{i,p-1}$.

5.3.1 Introduction

Matrix Notation

We define

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}$$

so that the linear regression can be written as

$$Y = X\beta.$$

Least Squares Estimation

The residual sum-of-squares that is minimized to find the least squares estimators is

$$\begin{aligned} \text{RSS} &= ||Y - X\beta||^2 \\ &= (Y - X\beta)^T(Y - X\beta) \\ &= \left(Y^T - \beta^T X^T\right)(Y - X\beta) \\ &= Y^TY - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta \end{aligned}$$

where ^{T} is the transpose operator. We minimize the RSS by taking the partial derivatives with respect to each component of β and setting the resulting equations to 0. We can perform this calculation in matrix notation (see Matrix Calculus), so

$$\frac{d\text{RSS}}{d\beta} = -Y^TX - Y^TX + \beta^T\left(X^TX + X^TX\right) = -2Y^TX + 2\beta^TX^TX$$

Setting the result to 0 at $\hat{\beta}$, we have

$$\begin{aligned} -Y^TX + \hat{\beta}^TX^TX &= 0 \\ \hat{\beta}^TX^TX &= Y^TX \\ X^TX\hat{\beta} &= XY \\ \hat{\beta} &= (X^TX)^{-1}XY \end{aligned}$$

whenever the inverse matrix $(X^TX)^{-1}$ exists.

Lemma 24. *X^TX is nonsingular (i.e. $(X^TX)^{-1}$ exists if and only if rank$(X) = p$.*

The rank, is the number of linearly independent columns in X. The columns of X are linearly *dependent* if there exists $a = (a_1, \ldots, a_p)^T \neq 0$ such that

$$\sum_{j=1}^{p} a_j x_{\cdot j} = 0$$

5.3.2 Random Vectors

Properties of Random Vectors

Before we continue, we need some basic results regarding random vectors.

Definition: *random vector*

A random vector is a vector of random variables. Specifically, suppose $Y_1, \ldots, Y_n$ are random variables, then

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

is a random vector.

Definition: *mean of random vector*

We can define the mean of a random vector as the component-by-component mean:

$$E[Y] = \begin{pmatrix} E[Y_1] \\ E[Y_2] \\ \vdots \\ E[Y_n] \end{pmatrix}$$

Definition: *variance of random vector*

And the variance of a random vector is a matrix

$$\text{Cov}(Y) = \Sigma_{YY} = \{\sigma_{ij}\}$$

where i, j entry $\sigma_{ij} = \text{Cov}(Y_i, Y_j)$. Notice, the variance of a random vector is a symmetric matrix with positive diagonal entries (the diagonal holds the variances).

Linear Functions of Random Vectors

We now concern ourself with the properties of linear functions of random vectors

$$Z = c + AY$$

where c is a vector of constants and A is a matrix of constants. In general, Y has dimension $n \times 1$, but A may have dimension $m \times n$, implying Z and c are both of dimension $m \times 1$.

The following theorems are proven by showing equivalence of left and right sides, term-by-term.

Theorem 25. *The expectation continues to be a linear function, so*

$$E[Z] = c + AE[Y]$$

Theorem 26. *The covariance of Z is*

$$Cov(Z) = \Sigma_{ZZ} = A\Sigma_{YY}A^T$$

given $Cov(Y) = \Sigma_{YY}$.

Quadratic Form

A prevalent form in our future calculations is the so-called quadratic form, a scalar variable that can be written in terms of a matrix of constants.

Definition: *quadratic form*

Let A be a symmetric $n \times n$ matrix and X a $n \times 1$ vector. Then,

$$X^T AX = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j$$

is a quadratic form.

Example:

The numerator in the sample variance of $X_1, \ldots, X_n$ is a quadratic form. Recall, the numerator of sample variance is

$$\sum_{i=1}^{n}(X_i - \bar{X})^2$$

We have to do some tricks with matrix notation. Let $\mathbf{1}$ be a column of n 1's. Then, we can write the sample mean in matrix notation

$$\bar{X} = \frac{1}{n}\mathbf{1}^T X.$$

We can get a vector containing only sample means using

$$\begin{pmatrix} \bar{X} \\ \bar{X} \\ \vdots \\ \bar{X} \end{pmatrix} = \frac{1}{n}\mathbf{1}\mathbf{1}^T X,$$

and the vector of deviances is

$$X - \mathbf{1}\bar{X} = \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix} = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X$$

where I is the $n \times n$ identity matix. Thus, if $B = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then we have written vector $X - \mathbf{1}\bar{X}$ as BX. Further,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = (BX)^T BX = X^T B^T BX,$$

proving our premise that the sample variance numerator has a quadratic form with matrix $A = B^T B$.

Our final theorem provides the expectation of the quadratic form

Theorem 27. *If random vector X has $E[X] = \mu$ and Cov$(X) = \Sigma$ and A is a fixed matrix, then*

$$E[X^T AX] = \textit{trace}(A\Sigma) + \mu^T A\mu$$

Example:

Let's check the expectation of the numerator of the sample variance for a sample $X_1, \ldots, X_n$ of independent random variables with constant variance σ^2. In other words, Cov$(X) = \sigma^2 I$, a constant times the identity matrix. First, we will simplify $B^T B$.

$$\begin{array}{rcll} B^T B & = & \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) & \text{definition} \\ & = & \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) & (A+B)^T = A^T + B^T \\ & = & I - 2\frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T & \text{distributive property: } (A+B)C = AC + AB \\ & = & I - \frac{1}{n}\mathbf{1}\mathbf{1}^T & \mathbf{1}\mathbf{1}^t\mathbf{1}\mathbf{1}^t \text{ is a matrix of } n\text{'s} \end{array}$$

$$\begin{array}{rcl} E[X^T B^T BX] & = & \text{trace}\left[\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\Sigma\right] + \mu^T\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mu \\ & = & \sigma^2\text{trace}\left[I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right] + \mu^T\left(\mu - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mu\right) \\ & = & \sigma^2\left[\text{trace}(I) - \text{trace}(\frac{1}{n}\mathbf{1}\mathbf{1}^T)\right] + \mu^T\left(\mu - \frac{1}{n}\mathbf{1}n\mu\right) \\ & = & \sigma^2(n-1) + \mu^T(\mu - \mu) \\ & = & \sigma^2(n-1) + 0 \end{array}$$

We have proven once again that the sample variance, where $X^T B^T BX$ is divided by the degrees of freedom $n - 1$, is an unbiased estimator of σ^2. This argument is used over and over to prove F tests for hypotheses in multiple linear regression.

5.3.3 Estimator Properties

Multiple Linear Regression Model

The multiple linear regression model is

$$Y = X\beta + \epsilon$$

where ϵ has $E[\epsilon] = \mathbf{0}$ and Var$(\epsilon) = \sigma^2 I$, i.e. the errors are unbiased, have constant variance, and are independent. We will add a normality assumption later to do inference.

Theorem 28. *The least squares estimators are unbiased if $E[\epsilon] = 0$.*

Proof.

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T Y \\
&= (X^T X)^{-1} X^T (X\beta + \epsilon) \\
&= \beta + (X^T X)^{-1} \epsilon \\
E[\hat{\beta}] &= \beta + (X^T X)^{-1} E[\epsilon] \\
&= \beta
\end{aligned}$$

□

Theorem 29. *The variance of the least squares estimators is*

$$Cov(\hat{\beta}) = \sigma_{\hat{\beta}\hat{\beta}} = \sigma^2 (X^T X)^{-1}$$

if $E[\epsilon] = 0$ *and Var*$(\epsilon) = \sigma^2 I$.

Proof.

$$\begin{aligned}
\text{Cov}(\hat{\beta}) &= (X^T X)^{-1} X^T \Sigma_{\epsilon\epsilon} X (X^T X)^{-1} && \text{theorem 26} \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} && \Sigma_{\epsilon\epsilon} = \sigma^2 I \\
&= \sigma^2 (X^T X)^{-1} && \text{defn inverse matrix: } C^{-1} C = I
\end{aligned}$$

□

Estimation of σ^2

For simple linear regression, we used

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

as our estimate of sample variance. This generalizes to multiple linear regression as

$$S^2 = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i = \widehat{E}[Y_i] = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{p-1} x_{i,p-1}$. We divide by $n - p$ because p degrees of freedom have been lost to estimate the p parameters in β.

We will now write this estimator in matrix notation and show that it is unbiased for population variance σ^2. The arguments are very similar to the sample variance calculations shown earlier. With these two demonstrations under your belt, it will be assumed that you know how to compute expectations of sums-of-squares.

$$\begin{aligned}
\hat{\epsilon} &= Y - \hat{Y} \\
&= Y - X\hat{\beta} \\
&= Y - X(X^T X)^{-1} X^T Y \\
&= (I - P) Y
\end{aligned}$$

where $P = X(X^T X)^{-1} X^T$.

Lemma 30. *Both matrix P and $I - P$ are symmetric and projection matrices.*

1. $P = P^T = P^2$
2. $(I - P) = (I - P)^T = (I - P)^2$

Proof. We will show the first. The second is shown similarly.

$$\begin{array}{rcll} P^T & = & \left[X(X^TX)^{-1}X^T\right]^T & \text{definition} \\ & = & X(X^TX)^{-1}X^T & (AB)^T = B^TA^T \text{ and } (A^{-1})^T = (A^T)^{-1} \\ & = & P & \\ P^2 & = & X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T & \\ & = & X(X^TX)^{-1}X^T & A^{-1}A = I \end{array}$$

□

These properties are all we need to verify our proposed population variance estimator, S^2.

$$\begin{array}{rcll} (n-p)S^2 & = & [(I-P)Y]^T(I-P)Y & X^TX = \sum_i X_i^2 \\ & = & Y^T(I-P)^T(I-P)Y & (AB)^T = B^TA^T \\ & = & Y^T(I-P)Y & I-P \text{ is symmetric projection matrix} \\ (n-p)E[S^2] & = & E[Y^T(I-P)Y] & \\ & = & \text{trace}((I-P)\Sigma_{YY}) + (E[Y])^T(I-P)(E[Y]) & \text{expectation of QF} \\ & = & \sigma^2\text{trace}(I-P) + (X\beta)^T(I-X(X^TX)^{-1}X^T)X\beta & \text{definition of P} \\ & = & \sigma^2[\text{trace}(I) - \text{trace}(P)] & \text{trace}(A-B) = \text{trace}(A) - \text{trace}(B) \\ & & + (X\beta)^T(X\beta - X(X^TX)^{-1}X^TX\beta) & (A+B)C = AC + BC \\ & = & \sigma^2(n - \text{trace}\left(X(X^TX)^{-1}X^T\right) + (X\beta)^T(X\beta - X\beta) & I \text{ is } n \times n \\ & = & \sigma^2(n - \text{trace}\left(X^TX(X^TX)^{-1}\right) + 0 & \text{trace}(AB) = \text{trace}(BA) \text{ for conformable matrices} \\ & = & (n-p)\sigma^2 & X^TX \text{ is } p \times p \end{array}$$

5.3.4 Inference

On β_j

We now add another hypothesis in order to discuss inferences relevant to the multiple regression problem.

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Noticing that $\hat{\beta} = (X^TX)^{-1}XY$ is a linear combination of the random observations Y_i (i.e. it is a linear estimator), we conclude

Theorem 31.

$$\hat{\beta} \sim N\left(\beta, \sigma^2(X^TX)^{-1}\right)$$

Proof. Trivial application of the reproductive property of normal distributions. □

The above theorem implies

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2 c_{jj}\right)$$

where c_{jj} is the ith diagonal entry of $(X^TX)^{-1}$. Thus, to test $H_0 : \beta_j = b_j$, we have

$$t = \frac{\hat{\beta}_j - b_j}{S\sqrt{c_{jj}}} \sim t_{n-p}$$

and CI

$$\hat{\beta}_j \pm t_{n-p}(1-\alpha/2)S\sqrt{c_{jj}}$$

Assessing Assumptions

We have mentioned previously the importance of examining residuals to check for assumptions. We know $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, but can we assume the $\hat{\epsilon}_i$ have a similar distribution? Warning: we have indeed made this assumption for simple linear regression, but it was a lazy assumption. Let's see why.

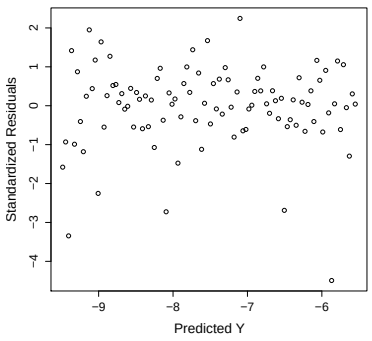


Figure 8: Standardized Residuals

The estimated residual

$$\hat{\epsilon} = Y - \hat{Y} = (I - P)Y$$

is a random variable with variance (by Thm 26)

$$\Sigma_{\epsilon\epsilon} = (I - P)(\sigma^2 I)(I - P)^T = \sigma^2(I - P)$$

Clearly, the residuals are not independent (off-diagonal terms may not be 0) and they do not have constant variance. However, $I - P$ is a known matrix, so we can easily standardize the residuals.

Definition: *standardized residual*

The standardized residual is

$$SR_i = \frac{Y_i - \hat{Y}_i}{S\sqrt{1 - p_{ii}}}$$

where p_{ii} is the ith diagonal entry of P and S is the square root of S^2.

A plot of SR_i against predicted $\hat{Y}_i$ for the simple linear regression shown in Fig. 7(a) is shown in Fig. 8. It is not much changed from the plot of unstandardized residuals in Fig. 7(b), but now we can draw more reliable conclusions. We expect 95% of the residuals to lie between -1.98 and 1.98 (`qt(0.975, df=98)`), so there appear to be a few particularly small values. There is no evident trend in the residuals, and the variance seems roughly constant. Notice, the probability plot of Fig. 7(c) used the standardized residuals. The extended right tail verifies the presence of small outliers. If you tried the square root transform, it would be even worse.

Partitioning Variance

As for simple linear regression, we can partition variance

$$\begin{aligned} \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\ SS_{tot} &= SS_E + SS_{reg} \end{aligned}$$

only $\hat{Y}_i$ is now a bigger linear function $X\hat{\beta}$.

We can define the **coefficient of determination** exactly as before

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$

Again, this can be interpretted as the percentage of variation (around the mean) that is explained by the regression function.

Testing Subsets of Regression Parameters

To generalize inference in the multiple linear regression context, it is of interest to be able to test hypotheses like

$$H_0 : \beta_A = 0$$

where the parameter vector

$$\beta = \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix}$$

has been partitioned into two parts, one of length p_A and the other β_B of length p_B, such that $p_A + p_B = p$. We can also partition the design matrix

$$X = [X_A : X_B]$$

of dimensions $n \times p_A$ and $n \times p_B$. Since the terms of the linear model can be reordered in any way we choose, we reorder β and the columns of X, such that the parameters to test in H_0 come first. The linear model can be now written as

$$Y = X_A\beta_A + X_B\beta_B + \epsilon.$$

We will use further partitions of variability to test such generic hypotheses. Introduce notation

$$SS_{reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = R(\beta_1, \ldots, \beta_{p-1} \mid \beta_0) = R(\beta) - R(\beta_0) = \sum_{i=1}^{n}\hat{Y}_i^2 - \sum_{i=1}^{n}\bar{Y}^2$$

So, $R(\beta_1, \ldots, \beta_{p-1} \mid \beta_0)$ is the increase in regression sum-of-squares attributed to adding $\beta_1, \ldots, \beta_{p-1}$ to the model with β_0 only. We can generalize this notation to any non-overlapping subsets of the parameters.

$$R(\beta_A \mid \beta_B) = R(\beta_A, \beta_B) - R(\beta_B)$$

$R(\beta_A \mid \beta_B)$ is the increase in regression sum-of-squares obtained by adding β_A to the model with β_B. Since every addition to the regression sum-of-squares is achieved by removing from the residual sum-of-squares, we can also view this as the decrement in residual sum-of-squares achieved by adding β_A to the model.

Let's explore this quantity further.

$$\begin{aligned}
R(\beta_A \mid \beta_B) &= \sum_{i=1}^{n}\hat{Y}_i^2 - \sum_{i=1}^{n}\hat{Y}_{i,B}^2 \\
&= (X\hat{\beta})^T(X\hat{\beta}) - (X_B\hat{\beta}_B)^T(X_B\hat{\beta}_B) \\
&= \left[X(X^TX)^{-1}X^TY\right]^T\left[X(X^TX)^{-1}X^TY\right] \\
&\quad - \left[X_B(X_B^TX_B)^{-1}X_B^TY\right]^T\left[X_B(X_B^TX_B)^{-1}X_B^TY\right] \\
&= Y^TX(X^TX)^{-1}X^TY - Y^TX_B(X_B^TX_B)^{-1}X_B^TY.
\end{aligned}$$

Notice that $X(X^TX)^{-1}X^T$ is our friend P. Under $H_0 : \beta_A = 0$, there is another projection matrix $P_B = X_B(X_B^TX_B)^{-1}X_B^T$ for the smaller model. Notice, if H_0 is *not* true, then P_B is *not* a projection matrix because $(1 - P_B)Y = Y - \hat{Y}_B = \epsilon + X_A\beta_A \neq \epsilon$. Under the iid, constant variance assumptions *and* $H_0 : \beta_A = 0$,

$$\begin{aligned}
E[R(\beta_A \mid \beta_B)] &= E[Y^TX(X^TX)^{-1}X^TY] - E[Y^TX_B(X_B^TX_B)^{-1}X_B^TY] \\
&= \sigma^2(p - p_B) = p_A\sigma^2
\end{aligned}$$

by our results for expectations of quadratic forms. Following arguments that we have seen before, we have two estimates of population variance σ^2, and statistic

$$\frac{R(\beta_A \mid \beta_B)/p_A}{S^2} \sim F(p_A, n - p)$$

under H_0.

In conclusion, we have a statistic for testing any subset β_A of the population parameters.

Prediction

As for simple linear regression, we can also predict $E[Y_i \mid x_0] = \hat{Y}(x_0)$ at a new $x_0 = (1, x_{01}, x_{02}, \ldots, x_{0,p-1})$ or a new observation Y_i at x_0.

For predicting the mean, we have

$$\hat{Y}(x_0) = x_0^T\hat{\beta}$$

with variance

$$\text{Var}(\hat{Y}(x_0)) = x_0^T\sigma_{\hat{\beta}\hat{\beta}}x_0 = \sigma^2x_0^T(X^TX)^{-1}x_0$$

yields CI

$$x_0^T \hat{\beta} \pm t_{n-p}(1-\alpha/2) S \sqrt{x_0^T (X^T X)^{-1} x_0}$$

For predicting an observation, we predict the same value $\hat{Y}(x_0)$, but

$$Y_0 = \hat{Y}(x_0) + \epsilon_0,$$

which has variance

$$\text{Var}(Y_0) = \text{Var}(\hat{Y}(x_0)) + \text{Var}(\epsilon_0)$$

by independence of ϵ_0 from all previous observations $Y_1, \ldots, Y_n$. The CI is

$$x_0^T \hat{\beta} \pm t_{n-p}(1-\alpha/2) S \sqrt{x_0^T (X^T X)^{-1} x_0}$$

Comparing Models

Choosing from among multiple alternative regression models is a complicated topic worthy of its own class. We will just touch on the subject. We have a number of summaries of a model fit that can be used to rank methods

- R^2: the coefficient of determination either stays the same or increases as we add new predictors to the model.
- S^2: the mean residual sum-of-squares estimates of population variance under the true model, and overestimates it under the wrong model. Thus, a ranking of models by this statistic finds better fitting models.
- $\bar{R}^2$: the *adjusted* R^2 is defined as

  $$\bar{R}^2 = 1 - \frac{SS_E/(n-p)}{SS_{tot}/(n-1)}$$

  and provides the same ranking as S^2, though $\bar{R}^2$ increases with better fit.

None of these methods avoids the problem of over-fitting. One can include a predictor in the model that does not provide a significantly better fit, but nevertheless improves $\bar{R}^2$. These predictors are fitting/accounting for *quirks* of the particular dataset at hand. When the same models are used to fit other datasets with different quirks, it is found these "quirk" terms do not fit the new datasets well. Thus, overfitting is a problem to be avoided.

One solution applicable when the dataset is large enough is to fit the model only to a subset of the data, say of size $m < n$. After fitting, the $n - m$ left-over data points are used to assess the fit. If these new data points are well-explained by the fitted model, then the model is considered good. Overfitted models would not well-explain the reserved data.

5.4 Logistic Regression

Binary Response

We now consider data where the response is binary (e.g. "yes"/"no" or 0/1). There is only one probability model for this case, $Y_i \sim$ Bernoulli.

If we continue inside the framework of multiple linear regression we run into problems:

$$Y = X\beta + \epsilon$$

with $E[Y_i] = P(Y_i = 1) := P_i = X_i^T \beta$. There are several problems. First, nothing in our construction as guaranteed that $X_i^T \beta \in [0, 1]$, as it must be for a probability. Second $\epsilon_i = Y_i - X_i^T \beta$ is discrete, depending

on the value of Y_i. Therefore, the errors are not normal. Further, $\text{Var}(\epsilon_i) = P_i(1 - P_i)$ is not constant if the probability of success depends on the observation. However, P_i is constant, then the predictors have no impact on the mean response, which defeats our purpose for studying predictors.

We need another framework. We will use the likelihood framework. Recall that all the results of the multiple linear regression can be obtained under the assumption of iid normal random variables, in which case the least squares estimator are the maximum likelihood estimators.

5.4.1 Data

Data

Consider grouped data

$$\begin{array}{cccccc} n_1 & r_1 & x_{11} & \cdots & x_{1,p-1} \\ & & \vdots & & \\ n_s & r_s & x_{s1} & \cdots & x_{s,p-1} \end{array}$$

where there are s distinct predictor combinations, and of the n_i trials for the ith set of predictors, r_i resulted in a success.

5.4.2 Model

Logistic Function

Instead of using a linear model for the expected value of the response, we use a nonlinear model

$$P(Y_i = 1) = P(x_i) = \frac{1}{1 + e^{-x_i^T \beta}}.$$

This function is called the *logistic function*, and we choose it for convenience. Notice that the logistic function maps the real line to the interval $[0, 1]$ exactly as we require. Futher, if $\beta_j > 0$, as x_{ij} increases, the probability of success $P(x_i)$ increases. The coefficients β_j are not linearly related to the mean response, but they do have a predictable effect on it.

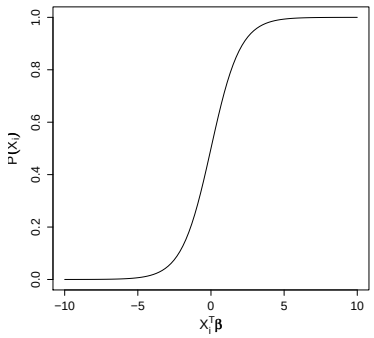


Figure 9: Logistic Function

The inverse function is the log-odds function (aka logit function)

$$\log\left[\frac{P(x_i)}{1 - P(x_i)}\right] = x_i^T \beta$$

exists on the interval $(-\infty, \infty)$. The logistic model views the log odds of success as a linear function in the population parameters.

Estimation

To obtain estimates $\hat{\beta}$ of the population parameters, we maximize the likelihood. The likelihood for the n_i observations with predictors of group i is

$$L(\beta, x_i; r_i) = [P(x_i)]^{r_i} [1 - P(x_i)]^{n_i - r_i} = \left[\frac{1}{1 + e^{-x_i^T \beta}}\right]^{r_i} \left[\frac{e^{-x_i^T \beta}}{1 + e^{-x_i^T \beta}}\right]^{n_i - r_i}$$

Since all observations are independent, the overall likelihood is a product of such terms

$$L(\beta, X) = \prod_{i=1}^{s} [P(x_i)]^{r_i} [1 - P(x_i)]^{n_i - r_i}$$

There is no point in taking derivatives and trying to find analystic solutions. Numerical iterative solutions are required.

Ungrouped Data

The data need not be grouped in order to perform logistic regression. In this case, the data are

$$\begin{array}{cccc} Y_1 & x_{11} & \cdots & x_{1,p-1} \\ & \vdots & & \\ Y_n & x_{n1} & \cdots & x_{n,p-1} \end{array}$$

and the likelihodo becomes

$$L(\beta, X) = \prod_{i=1}^{n} \left[P(x_i)\right]^{Y_i} \left[1 - P(x_i)\right]^{1-Y_i}.$$

Still no analytic solutions.

5.4.3 Inference

Lack-of-Fit

Since the logistic function was used for convenience it is important to determine whether it makes sense at all. The most complicated model is

$$Y_i = P_i + \epsilon_i$$

where P_i is an arbitrary probability associated with each treatment combination i of the predictors. Our logistic model posits $P_i = P(x_i)$ as defined above, which is a constrained version of the above model.

The likelihood under the complicated model is

$$L(P) = \prod_{i=1}^{n} P_i^{Y_i} (1 - P_i)^{1-Y_i}$$

which can be analytically maximized to yield

$$\hat{P}_i = Y_i.$$

Notice, this is a perfectly fitting model because $Y_i - \hat{Y}_i = Y_i - \hat{P}_i = 0$. The residual sum-of-squares is 0.

To compare the fits of this model and the logistic model, we use the likelihood ratio test, a generally applicable test obtained from likelihood theory. The likelihood ratio test is based on the likelihood ratio statitic

$$\lambda(\beta) = -2 \ln \left[\frac{L(\hat{\beta})}{L(\hat{P})} \right]$$

When the model in the numerator is nested in the model in the denominator and the numerator model is the true model, i.e.

$$H_0 : P_i = P(x_i)$$

then $\lambda(\beta) \sim \chi^2_{\Delta \text{df}}$ is asymptotically true as $n \to \infty$. The change in degrees of freedom Δdf is the difference in the number of free parameters between the two models.

$$\Delta \text{df} = n - p$$

(Note: this was stated incorrectly in class as $n - p - 1$, which explains my confusion then!) There are n parameters P_i in the big model and p parameters β in the nested model.

We can tell a model is nested inside another if the nested model imposes one or more constraints on the parameters of the bigger model. In this case, the nested model constrains all of the bigger model parameters to follow the logistic function.

Testing Subset of Coefficients

Generally, we hope to not reject the H_0 above so that we can proceed with inference under the logistic regression model. If logistic regession is justified, then we may wish to determine which predictors impact the mean response. We can test any subset of predictors, i.e.

$$H_0 \beta_A = 0$$

by using the same likelihood ratio test with statistic

$$\lambda(\beta_A \mid \beta_B) = -2 \ln \left[\frac{L(\hat{\beta}_B)}{L(\hat{\beta})} \right] \sim \chi^2_{p_A}$$

where the degrees of freedom is the number of constrained parameters, p_A, under H_0.

Standard Errors and CI

Furthermore, the likelihood framework gives us estimates of the variance of the MLEs $\hat{\beta}$. Asymptotically, as $n \to \infty$,

$$\text{Cov}(\hat{\beta}) = \left[-\frac{\partial^2 \ln L(\hat{\beta})}{\partial \beta^2} \right]^{-1}$$

Notice, that $\frac{\partial^2 \ln L(\hat{\beta})}{\partial \beta^2}$ is the Hessian matrix of the log likelihood, evaluated at the MLE $\hat{\beta}$. So, the inverse of the Hessian is the covariance matrix of the MLE.

If d_{jj} is the jth diagonal of the covariance matrix, then

$$X^2 = \frac{\hat{\beta}_j^2}{d_{jj}} \sim \chi^2_1$$

asymptotically if $H_0 : \beta_j = 0$ is true. The above represents a second way to test if $\beta_j = 0$. It may not give the same result as the subset method above, as they are both asymptotic results from different starting points.

Furthermore, we can form confidence intervals for β_j as

$$\hat{\beta}_j \pm Z_{1-\alpha/2} \sqrt{d_{jj}},$$

but since these CI are only asymptotic, it may be more sensible to just report the standard error of $\hat{\beta}_j$ as $\sqrt{d_{jj}}$.

5.5 Poisson Regression

5.5.1 Data

Count Data

Suppose the observed Y_i are counts, so $Y_i \in \{0, 1, \ldots\}$. We now have multiple probability models possible for the data. One of the most common is

$$p(y) = \frac{e^{-\mu} \mu^y}{y!},$$

the Poisson distribution, but I note that other models (e.g. negative binomial) could be used. This distribution has only one parameter, $\mu = E[Y]$, which we now speculate depends on predictors.

5.5.2 Model

Model

The general model we pose is

$$Y_i = \mu(x_i, \beta) + \epsilon_i$$

where, as usual, $\mu(x_i, \beta) = E[Y_i]$. Notice, $\mu(x_i, \beta) \geq 0$. As for logistic regression, the mean is restricted, this time to be on the positive real line. We need a function that maps the real line $x_i^T \beta$ to the positive real line. The most common function is

$$\mu(x_i, \beta) = e^{x_i^T \beta}.$$

Its inverse is

$$x_i^T \beta = \log(\mu(x_i, \beta)).$$

To estimate the parameters, we need a likelihood. Herein enters the model for the counts, namely Poisson, so

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{(\mu(x_i, \beta)^{Y_i} e^{-\mu(x_i, \beta)}}{y_i!} \right].$$

Again, there is no analytic solution for the MLEs $\hat{\beta}$.

5.5.3 Inference

Inference

There is no difference in inference from the logistic model. We can perform a lack-of-fit test or test any subset of the parameters with likelihood ratio tests. We can obtain the covariance matrix of $\hat{\beta}$ from the inverse Hessian.

5.6 Generalized Linear Model

Generalized Linear Model in R

The generalized linear model is a framework for all multiple regression models we have considered. Notice that in all the models, we propose a function $f(\cdot)$ that maps the linear expression $x_i^T \beta$ to $E[Y_i]$. For multiple linear regression, the function was the identity $f(x_i^T \beta) = x_i^T \beta$. For logisitic regression, the function was the logistic function

$$f(x_i^T \beta) = \frac{1}{1 + e^{-x_i^T \beta}}.$$

For Poisson regression, the function was the exponential function

$$f(x_i^T \beta) = e^{x_i^T \beta}.$$

In the generalized linear model, the inverse of $f(\cdot)$ is called the *link function*. Two things must be defined for a generalized linear model: (1) the probability model for the response Y_i, and (2) the link function. In all cases, it finds $\hat{\beta}$ by maximizing the log likelihood numerically.

Below are `glm` calls in `R` for the three models we have considered.

```
> # same as lm(Y ~ X1 + X2, data=d)
> glm(Y ~ X1 + X2, data=d, family=gaussian(link="identity"))
> # logistic regression for binary response
> glm(Y ~ X1 + X2, data=d, family=binomial(link="logit"))
> # poisson regression for count data
> glm(Y ~ X1 + X2, data=d, family=poisson(link="log"))
```

5.7 Nonlinear Regression

Nonlinear Regression

Nonlinear regression is useful when there is some theoretical relationship known or hypothesized between predictors and response. The objective is to test this proposed relationship and/or estimate the parameters of that relationship. The model is

$$Y_i = f(x_i, \beta) + \epsilon_i$$

where Y_i is once again a continuous response. An example is

$$Y_i = \alpha e^{\beta x_i} + \epsilon_i$$

for a single predictor x_i.

If we hypothesize $\epsilon \sim N(0, \sigma^2 I)$, then minimizing the residual sum-of-squares

$$SS_E = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - f(x_i, \hat{\beta}))^2$$

is equivalent to maximizing the likelihood

$$L(\beta) = \frac{1}{(2\pi\sigma^2)^n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - f(x_i, \hat{\beta}))^2 \right].$$

Note, however, that numerical methods are required to find $\hat{\beta}$ that minimizes SS_E. In addition, the estimated values are no longer necessarily unbiased, although, because they are MLEs, they will be unbiased as $n \to \infty$. As before,

$$S^2 = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

estimates σ^2, but the MLE is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

For testing coefficients, one can use the likelihood ratio testing framework. To test

$$H_0 : \beta_j = b_j$$

let $f(x_i, \beta)$ be the full model with all parameters and $f_j(x_i, \beta_{-j})$ be the restricted model with $\beta_j = b_j$. We estimate the MLE $\hat{\beta}$ under the full model, and the MLE $\hat{\beta}_{-j}$ under the restricted model H_0. Then

$$\begin{aligned}
\lambda(\beta_j \mid \beta_{-j}) &= -2 \ln \left[\frac{L(\hat{\beta}_{-j})}{L(\hat{\beta})} \right] \\
&= -2 \left[\ln L(\hat{\beta}_{-j}) - \ln L(\hat{\beta}) \right] \\
&= -2 \left\{ -\ln \left[\sum_{i=1}^{n} (Y_i - f(x_i, \hat{\beta}_{-j}))^2 \right] + \ln \left[\sum_{i=1}^{n} (Y_i - f(x_i, \hat{\beta}))^2 \right] \right\} \\
&= -2 \left[\ln(SS_E(\hat{\beta})) - \ln(SS_E(\hat{\beta}_{-j})) \right]
\end{aligned}$$

where $SS_E(\hat{\beta})$ is the residual sum-of-squares from the full model and $SS_E(\hat{\beta}_{-j})$ is the residual sum-of-squares from the restricted model with $\beta_j = b_j$.