1 Introduction

Notation: The outcome y is the main variable of interest whose mean we want to explain in terms of other variables; aka the response, output, dependent variable. The covariate x represents the other variables of interest which potentially explain/predict the outcome in some sense; aka the predictor, input, independent variable, feature.

Quantitative description: To describe data, we can use univariate summaries (eg mean or variance of outcome or covariate) or bivariate summaries (eg covariance or correlation).

1. Mean/expectation: We focus on continuous rv in this course,

$$E[Y] = \int yf(y)\, dy$$

Recall the linearity of expectation and for observations $y_1, \ldots, y_n$, the sample mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and is an estimate of the population mean which cannot be observed in practise.

2. Variance:

$$\text{Var}(Y) = E[(Y - E[Y])^2] = E[Y^2] - E[Y]^2$$

Recall $\text{Var}(aY + b) = a^2\text{Var}(Y)$ and for independent rvs X, Y, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. For observations $y_1, \ldots, y_n$, the (unbiased) sample variance is $s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$. Intuitively, we divide by $n - 1$ rather than n since $\bar{y}$ is an estimate rather than an actual observation, we lost one degree of freedom.

3. Covariance:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Recall $\text{Cov}(X, X) = \text{Var}(X)$ and $\text{Cov}(aY + c, bX + d) = ab\text{Cov}(X, Y)$ and $\text{Cov}(U + V, X + Y) = \text{Cov}(U, X) + \text{Cov}(U, Y) + \text{Cov}(V, X) + \text{Cov}(V, Y)$. Using the last property,

$$\text{Var}(X + Y) = \text{Cov}(Y + X, Y + X) = \text{Var}(Y) + \text{Var}(X) + 2\text{Cov}(X, Y)$$

and recall that $\text{Cov}(X, Y) = 0$ for independent rvs X, Y. For observations $(y_1, x_1), \ldots, (y_n, x_n)$, the sample covariance is $\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$

4. Correlation: the correlation coefficient is,

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

and sample correlation is,

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

where $S_{xy} = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$ (note the exclusion of the $\frac{1}{n-1}$) and analogously for S_{xx}, S_{xy}. Correlation gives the strength of a linear relationship but does not characterize it. For example, the correlation of rvs X, Y is equal to the correlation of rvs X, Y' where $Y' = 2Y$, even though the relationship between X, Y is not the same as between X, Y'.

Normal distribution: Recall $Z \sim N(\mu, \sigma^2)$ with parameters μ, σ^2 has pdf

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z - \mu)^2}{2\sigma^2}\right]$$

and moments $E[Z] = \mu, \text{Var}(Z) = \sigma^2$. As well, for independent $Z_i \sim N(\mu_i, \sigma^2)$, $U = \sum_{i=1}^n (a_i Z_i + b_i)$ is normally distributed,

$$U \sim N\left(\sum_{i=1}^n (a_i\mu_i + b_i), \sum_{i=1}^n a_i^2\sigma_i^2\right)$$

Chi-square distribution: Recall $X \sim \chi^2_\nu$ with ν degrees of freedom has pdf with support $X > 0$,

$$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$$

and moments $E[X] = \nu, \text{Var}(X) = 2\nu$ and for $Z_i \sim N(0,1)$ iid,

$$X = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$$

t-distribution: Recall $Y \sim t_\nu$ with ν degrees of freedom has pdf

$$f(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\frac{\nu}{2}} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

and moments $E[Y] = 0$ if $\nu > 1$ and N/A otherwise, and $\text{Var}(Y) = \frac{\nu}{\nu-2}$ if $\nu > 2$ and ∞ otherwise. As well, for independent $Z \sim N(0,1)$ and $X \sim \chi^2_\nu$,

$$\frac{Z}{\sqrt{X/\nu}} \sim t_\nu$$

2 Simple Linear Regression

We want to: characterize the relationship between x, y, predict y given x, evaluate how the mean of y changes when x increases by a. We can do this (in the case of 1 covariate) using a simple linear regression model,

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

where ϵ_i is the error term for the ith observation. And for cases of multiple covariates, we use a multiple linear regression,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

This course focuses on extending simple linear regression to multiple linear regression: mathematically (derive estimators of unknown parameters $\beta_0, \beta_1, \ldots$), practically (how to fit these models in R), how to choose and compare a model (which x_{ij} to include), and how to evaluate the appropriateness of the model/assumptions (model diagnostics).

For the ith observation, y_i is the outcome, x_i is a covariate, and i indexes the observations in the sample.

Simple Linear Regression Model: Suppose there is 1 covariate. Consider,

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

where ϵ_i is iid $N(0, \sigma^2)$. Alternatively, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and are independent of each other. Note that technically Y_i is the rv and y_i is the outcome but here, we treat y_i as a rv sometimes in notation. In this model, $\beta_0, \beta_1, \sigma^2$ are fixed, unknown parameters, ϵ_i is an unobserved random variable error term (denoted R_i in Stat 332), and y_i, x_i are the observed data. We treat x_i as fixed but y_i is not, since it is the sum of a fixed piece and a random error component. We call β_0, β_1 the regression coefficients.

Since x_i is fixed, we can consider observations y_i as conditioning on x_i. So,

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i$$

by linearity since the mean of the error term is 0. β_0 can be interpreted as the intercept; $E[y_i|x_i = 0] = \beta_0$. As well, β_1 can be interpreted as a slope; $E[y_i|x_i = x^*] = \beta_0 + \beta_1 x^*$ and $E[y_i|x_i = x^* + 1] = \beta_0 + \beta_1 x^* + \beta_1$, so $\beta_1 = (E[y_i|x_i = x^*] - E[y_i|x_i = x^* + 1])/1$. In other words, the mean difference comparing a population with x to a population with x a unit higher (alternatively, with a 1 unit change in the covariate).

Graphically, simple linear regression is a line in 2D space where ϵ_i is the difference between the line at x_i and the actual point y_i; there is variability around the line.

Four assumptions about this model: linearity ($E[y_i|x_i] = \beta_0 + \beta_1 x_i$, or $E[\epsilon_i] = 0$), independence (error terms are iid but note that covariates may or may not be independent of each other), normality (error terms are normally distributed, thus y_i are normal), equal variance (aka homoskedasticity, all error terms have the same variance σ^2).

2.1 Estimation

Since β_0, β_1 cannot be observed, we are interested in estimating them (aka finding the line of best fit). We want to find the estimators $\hat{\beta}_0, \hat{\beta}_1$. Note that in Stat 332, estimators are denoted by a tilde, but in this course, it is a hat. Estimators are random variables (rvs) since it is a function of the outcomes, which are themselves rvs. We consider two possible objective functions,

1. Least Squares: We minimize the sum of squares between the y_i and the line at x_i,

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

To do this, we compute the partial derivatives and set to 0,

$$0 = \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) \implies 0 = \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i$$

Re-arranging, we get,

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

With respect to β_1,

$$0 = \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) \implies 0 = \sum_{i=1}^{n} y_i x_i - \beta_0 n\bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2$$

Substituting in $\beta_0 = \bar{y} - \beta_1 \bar{x}$ and re-arranging,

$$\beta_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum y_i(x_i - \bar{x})}{\sum x_i(x_i - \bar{x})} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

since $\sum \bar{y}(x_i - \bar{x}) = \bar{y} \sum (x_i - \bar{x}) = \bar{y}(\sum x_i - n\bar{x}) = 0$ and $\sum (\bar{x}^2 - x_i \bar{x}) = 0$. Thus, the LS estimators are,

$$\boxed{\hat{\beta}_1^{LS} = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0^{LS} = \bar{y} - \hat{\beta}_1^{LS} \bar{x}}$$

2. Maximum Likelihood Estimation: Recall y_i is independent $N(\beta_0 + \beta_1 x_i, \sigma^2)$. The likelihood (aka probability of unknown parameters given the observed y) is,

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2 | y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{2\sigma^2} \right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{2\sigma^2} \right)$$

The log likelihood is easier to work with,

$$\ell(\beta_0, \beta_1, \sigma^2 | y) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$$

To solve for $\max[\ell(\beta_0, \beta_1, \sigma^2 | y)] = \max[\mathcal{L}(\beta_0, \beta_1, \sigma^2 | y)]$, we solve a system of three equations,

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_0} &= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i]) \right) = 0 \\
\frac{\partial \ell}{\partial \beta_1} &= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i]) \right) = 0 \\
\frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2 = 0
\end{aligned}$$

Solving the first two equations is equivalent to minimizing the sum of squares. That is, under the assumption of normality,

$$\hat{\beta}_0^{LS} = \hat{\beta}_0^{ML} \qquad \hat{\beta}_1^{LS} = \hat{\beta}_1^{ML}$$

From here on out, we call the LS/ML estimators $\hat{\beta}_0, \hat{\beta}_1$.

We define the fitted values,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and the residuals,

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Note that the residuals e_i (which are rvs since y_i are rvs) are different from the random errors $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$, which are based on the true, unobservable β_0, β_1.

Back to Maximum Likelihood Estimation, we can rearrange the third equation to solve for $\hat{\sigma}^2_{ML}$,

$$\hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^n (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2}{n} = \frac{\sum_{i=1}^n e_i^2}{n}$$

However, in practise, we typically use a different estimator,

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}}$$

which is unbiased,

$$E[\hat{\sigma}^2] = \sigma^2$$

Intuitively, this is because (asserted without proof),

$$\frac{1}{\sigma^2} \sum_{i=1}^n e_i^2 \sim \chi^2_{(n-2)} \qquad E[\chi^2_\nu] = \nu$$

but often the distinction between $\hat{\sigma}^2$ and $\hat{\sigma}^2_{ML}$ does not matter when $n \geq 50$. In this course, we will typically use $\hat{\sigma}^2$. Informally, the $n-2$ comes from 2 fewer degrees of freedom since we are using $\hat{\beta}_0, \hat{\beta}_1$ which are functions of the y_i. For that reason, e_i are not independent of each other. See R code for how to fit a linear model.

2.2 Inference

In two different samples, we will get different estimates of β_0, β_1. We can characterize this uncertainty and variability in our estimates using standard errors (SE), confidence intervals (CI), and hypothesis tests (HT).

Properties of $\hat{\beta}_1$: Recall y_i are independent $N(\beta_0 + \beta_1 x_i, \sigma^2)$ and $\hat{\beta}_1 = \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \sum_{i=1}^n w_i y_i$ where $w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$ are fixed wrt y. Hence, $\hat{\beta}_1$ is a linear combination of independent Normals,

$$\hat{\beta}_1 \sim N\left(\sum_{i=1}^n w_i (\beta_0 + \beta_1 x_i), \sigma^2 \sum_{i=1}^n w_i^2 \right)$$

So, the mean is,

$$\begin{aligned}
E[\hat{\beta}_1] &= \sum_{i=1}^n w_i (\beta_0 + \beta_1 x_i) \\
&= \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n x_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= 0 + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta_1
\end{aligned}$$

Hence, this estimator is unbiased. The variance is,

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= \sigma^2 \sum_{i=1}^{n} w_i^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right]^2 \\
&= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{[\sum_{j=1}^{n}(x_j - \bar{x})^2]^2} \\
&= \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{[\sum_{j=1}^{n}(x_j - \bar{x})^2]^2} \\
&= \sigma^2 \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Thus,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \implies \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

and it can be shown using similar steps that,

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

Confidence Intervals: Using the known distribution values for $Z \sim N(0,1)$,

$$0.95 = P(-1.96 \leq Z \leq 1.96) = P\left(-1.96\frac{\sigma}{\sqrt{S_{xx}}} \leq \hat{\beta}_1 - \beta_1 \leq 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right)$$

Re-arranging,

$$0.95 = P\left(\hat{\beta}_1 - 1.96\frac{\sigma}{\sqrt{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right)$$

So, a 95% CI for β_1 (assuming σ is known) is,

$$\boxed{\hat{\beta}_1 \pm 1.96\frac{\sigma}{\sqrt{S_{xx}}}}$$

In other words, this random interval (random because it is centered around the rv $\hat{\beta}_1$) will cover the true β_1 95% of the time. However, in practise, σ is unknown and must be estimated. We have,

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1) \qquad V = \frac{1}{\sigma^2}\sum_{i=1}^{n} e_i^2 \sim \chi^2_{(n-2)}$$

and it can be shown that Z, V are independent. Recall that for independent $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$, we have that $\frac{Z}{\sqrt{V/\nu}} \sim t_\nu$. Thus,

$$\frac{Z}{\sqrt{V/(n-2)}} \sim t_{(n-2)} \implies \frac{\frac{\hat{\beta}_1 - \beta}{1/\sqrt{S_{xx}}}}{\sqrt{\sum_{i=1}^{n} e_i^2/(n-2)}} \sim t_{(n-2)} \implies \frac{\hat{\beta}_1 - \beta}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}$$

So, using similar steps as before, a 95% CI for β_1 when σ is unknown (and we use $\hat{\sigma}$) is,

$$\boxed{\hat{\beta}_1 \pm q\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \qquad ; q \sim t_{(n-2)}}$$

Using R code (and see Stat 332 work), we compute q using `qt(p=0.05/2, df=n-2, lower.tail=FALSE)`. q is often denoted $t_{n-2,1-\alpha/2}$ for a $1-\alpha$ CI,

$$\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

where $1-\alpha/2$ is used to calculate the p value for HT and $n-2$ is the df for the t.

We can compute CI in R using `confint(Model)`, for some `Model <- lm(y ~ x)`. Note that CI does not represent a probability such as $P(0.235 \leq \beta_1 \leq 0.287) = 0.95$, since β_1 is an (unobservable) constant thus either does or does not fit in the interval. Rather, a CI says that if we draw samples repeatedly and construct intervals in the same way, on average, 95% of such intervals will contain the true β_1.

Standard Error: The standard error of $\hat{\beta}_1$ is the estimated standard deviation of $\hat{\beta}_1$,

$$SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}} \qquad SE(\hat{\beta}_1) = \sqrt{\hat{\text{Var}}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

so we can write a CI for β_1 unknown σ as $\hat{\beta}_1 \pm t_{n-2,1-\alpha/2}SE(\hat{\beta}_1)$. This is the same as in Stat 332.

We can easily find CI for β_0 by using $\hat{\text{Var}}(\hat{\beta}_0)$ shown earlier instead of $\hat{\text{Var}}(\hat{\beta}_1)$ to get $SE(\hat{\beta}_0)$.

Hypothesis Testing: We want to test a null hypothesis $H_0 : \beta_1 = \theta_0$ against an alternative $H_1 : \beta_1 \neq \theta_0$. Often, in linear regression, $\theta_0 = 0$. The goal is to characterize how much evidence we have against H_0 (how extreme is our data relative to H_0). Under H_0 (assuming it is true),

$$\frac{\hat{\beta}_1 - \theta_0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}$$

So the probability *under the null* of a test statistic as extreme (or more) than what we observe is in a two-sided test is,

$$p = P(|T| \geq |t|) = 2P(T \geq |t|) = 2[1 - P(T \leq -|t|)]$$

The observed test statistic t is,

$$t = \frac{\text{estimate} - H_0\text{value}}{\text{standard error}} = \frac{\hat{\beta}_1 - \theta_0}{\sqrt{\text{Var}(\hat{\beta}_1)}}$$

We reject the null hypothesis at the 5% level (ie $p < 0.05$) and if $p > 0.05$), we cannot accept the null, rather we say we do not have enough evidence to reject. We cannot accept the null because all these calculations were under the assumption that H_0 is true.

As well, it is not true that $P(H_0) = p$. Rather, the p value means: under the null hypothesis, the probability of a test statistic as extreme as the one observed is p. That's why a small p is evidence against the null, since it would be very rare to observe this data under the null.

If we reject the null whenever $p < 0.05$, then if the null is true and we repeat the experiment over and over, we only reject the null incorrectly 5% of the time. This is called the Type-I error rate (incorrectly rejecting a true null). Fun fact, the word "null" means that it is a commonly accepted fact that researchers work to nullify.

Note that the t value and p values in `summary()` in R have $\theta_0 = 0$ in the null hypothesis and the alternative is two sided. So for any other θ_0, you cannot use `summary()`.

2.3 Prediction

Estimating mean response: Recall the mean response is $E[Y_i] = \beta_0 + \beta_1 X_i$, since $E[\epsilon_i] = 0$. For an arbitrary x_0 (may or may not have been observed), the mean is $\mu_0 = E[Y_0] = \beta_0 + \beta_1 x_0$. We can estimate this as,

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x})$$

So, by linearity since x_0 is fixed, and since $\hat{\beta}_0, \hat{\beta}_1$ are unbiased estimators,

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

Thus, the estimator of the mean response is unbiased. For variance,

$$\begin{aligned}
\text{Var}(\hat{\mu}_0) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \text{Var}(\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})) \qquad \text{; derived above} \\
&= \text{Var}\left(\left(\sum_{i=1}^{n} \frac{1}{n} y_i \right) + \left(\sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} y_i \right) (x_0 - \bar{x}) \right) \\
&= \text{Var}\left(\sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right) y_i \right) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right)^2 \sigma^2 \qquad \text{; by independence of } y_i \text{ and } \text{Var}(y_i) = \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + 2 \frac{1}{n} \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right) \\
&= \sigma^2 \left(\sum_{i=1}^{n} \frac{1}{n^2} + \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + 2 \sum_{i=1}^{n} \frac{1}{n} \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{2(x_0 - \bar{x})}{n S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2} S_{xx} + 0 \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)
\end{aligned}$$

Since $\hat{\mu}_0$ is a linear combination of normal variables, it is itself normal. Using the mean and variance above,

$$\hat{\mu}_0 \sim N\left(\mu_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right)$$

This implies (for known σ or in the unknown case, $\hat{\sigma}$),

$$\frac{\hat{\mu}_0 - \mu_0}{\sigma \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \sim N(0,1) \qquad \frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \sim t_{(n-2)}$$

In other words, a $100(1-\alpha)\%$ CI for unknown σ is,

$$\boxed{\hat{\mu}_0 \pm t_{n-2,1-\alpha/2} \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}}$$

Note that CIs are wider on the edges (as x_0 gets further away from $\bar{x}$), which makes sense since there is less data at the edges. As well, many points usually fall outside of the mean response CI; what id we do not care about the mean but want to do predictions?

Prediction: Suppose instead of the mean response, we want the predict the response itself (ie one specific value) for a new covariate value,

$$y_{new} = \beta_0 + \beta_1 x_{new} + \epsilon_{new}$$

Define the predicted value $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$ and the prediction error $\hat{y}_{new} - y_{new}$. Note that,

$$E[\hat{y}_{new} - y_{new}] = E[(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) - (\beta_0 + \beta_1 x_{new} + \epsilon_{new})] = \beta_0 + \beta_1 x_{new} - (\beta_0 + \beta_1 x_{new}) = 0$$

Here, $\hat{y}_{new}$ and $-y_{new}$ are independent rvs ($\hat{y}_{new}$ can be expressed as a linear combination of y_i from $i = 1$ to n but y_{new} is a completely new observation that is not included in these/did not form the estimates) and a linear combination of normals. The variance of the prediction error is,

$$\begin{aligned}
\text{Var}(\hat{y}_{new} - y_{new}) &= \text{Var}((\hat{\beta}_0 + \hat{\beta}_1 x_{new}) - y_{new}) \\
&= \text{Var}((\hat{\beta}_0 + \hat{\beta}_1 x_{new})) + \text{Var}(y_{new}) \qquad \text{; independence} \\
&= \left[\sigma^2\left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right] + \text{Var}(y_{new}) \qquad \text{; same derivation as in mean response} \\
&= \sigma^2\left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) + \sigma^2 \qquad \text{; by assumption, since } \text{Var}(\epsilon_{new}) = \sigma^2 \\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)
\end{aligned}$$

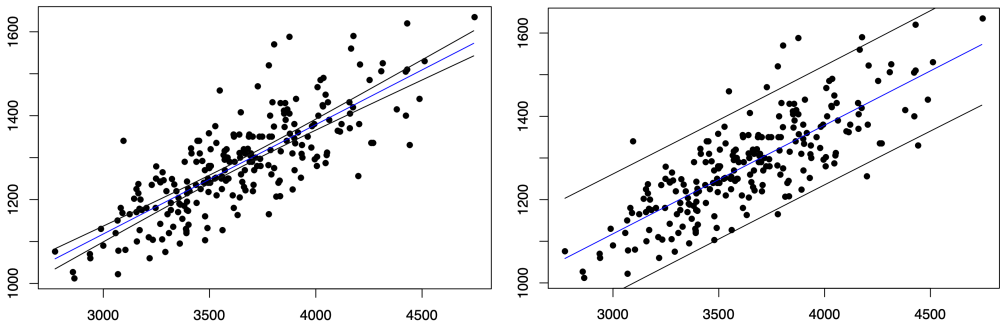
Thus, since the prediction error is normally distributed,

$$\frac{\hat{y}_{new} - y_{new}}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim N(0,1) \qquad \frac{\hat{y}_{new} - y_{new}}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}$$

Thus, a $100(1-\alpha)\%$ prediction interval (for unknown σ) is,

$$\boxed{\hat{y}_{new} \pm t_{n-2,1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

CI vs PI: The prediction interval (right) is wider than the CI for the mean response (left) because of the extra 1 in the square root, which comes from variability of any one observation around the mean line (uncertainty in estimating the line itself is accounted for in the rest of the SE after the 1+).

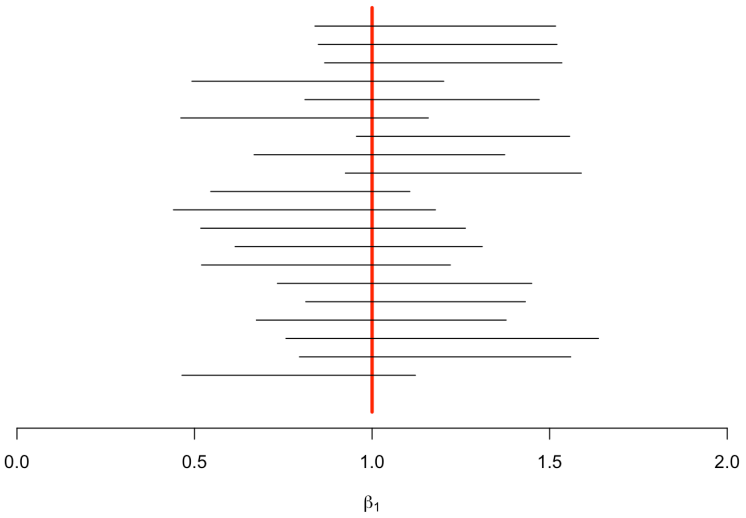


Note that CI is usually for parameters like μ, β_0, β_1 and PI is for a future individual observation, which is a data point that has not yet been observed. For parameters like μ, β_0, β_1, they already exist so it is not a prediction. We use CI for mean response to estimate the mean response for a population of observations at a certain covariate level and PI to predict one specific new response.

2.4 Summary

The important concepts from this section on simple linear regression,

1. The fitted values are $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where $\hat{\beta}_0, \hat{\beta}_1$ are the LS (or ML) estimates of the parameters based on the observed data (x_i, y_i) for $i = 1, \ldots, n$
2. The LS estimates for β_0, β_1 are $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and the ML estimates are the same. These estimators are unbiased. The derivation for unbiasedness only relies on the linearity of expectation and did not need the normality, homoskedasticity, or independence assumptions
3. The ML estimate for σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$ where $e_i = y_i - \hat{y}_i$ are the residuals
4. CI, PI, and HT for known ($\sim N(0,1)$) and unknown σ ($\sim t_{n-2}$).
5. $\hat{\beta}_1$ can be expressed as a linear combination of the outcomes, $\sum_{i=1}^n w_i y_i$ for some w_i, thus is also Normal
6. If we collect a new sample, we would get a different CI. And for $100(1-\alpha)\%$ CI, we have that $100(1-\alpha)\%$ of such intervals will cover the true underlying parameter
7. The p value is the probability of a test statistic as extreme or more than what we observe, *under the null* (ie assuming the null is true)
8. Estimating the mean response: $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} = \hat{\beta}_1(x_0 - \bar{x})$
9. Predicting a new response: $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$
10. A visual interpretation of CI:



Here, the true β_1 is 1 (the red line) and the black lines are possible sampled $100(1-\alpha)$ CIs. We expect $1-\alpha$ proportion of such CIs to cover the true β_1 of 1 as more CIs are sampled.

11. Mean response and predicted response are not two types of response. A response is an outcome (ie y_i). "Mean response" actually refers to the mean of responses $\mu_0 = E[y_i | x_i = x_0] = \beta_0 + \beta_1 x_0$. This is estimated with the estimated mean response $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

    We can also consider a brand new response, y_{new}, which we are trying to predict. The predicted response is our prediction of y_{new}, called $\hat{y}_{new}$.

    μ_0 lies on the line $\beta_0 + \beta_1 x$, which is the true underlying regression line which we do not know. y_{new} varies around this line because of the error term, $\beta_0 + \beta_1 x_{new} + \epsilon_{new}$.

$\hat{\mu}_0$ lies on the fitted regression line $\hat{\beta}_0 + \hat{\beta}_1 x$, which is our best guess for the unknown regression line. As such, there is uncertainty in estimating this line. Even if we had no uncertainty in estimating the regression line, (ie we knew exactly $\beta_0 + \beta_1 x$, there would still be uncertainty in predicting y_{new} since it does not lie on the true regression line because of the error term. That is why there is the 1+ term in the standard error.

Interestingly, the predicted response $\hat{y}_{new}$ lies also on the fitted regression line $\hat{\beta}_0 + \hat{\beta}_1 x$. In other words, if we estimate the mean response at covariate level x^* and predict a new response at covariate level x^*, both will have the same value: $\hat{\mu}_\star = \hat{\beta}_0 + \hat{\beta}_1 x^*$ and $\hat{y}_\star = \hat{\beta}_0 + \hat{\beta}_x^*$. See Quiz 2 Q1f for how this fact can be used to represent a PI in terms of a given CI.

Why does this happen? If we know a new response is not going to fall on the regression line, why is the predicted response the same as the estimate of the mean response? This is because we have no special knowledge of the new observation other than its covariate. Some new observations will be above the line, some below, we do not know. But, their average falls on the regression line. So, the predicted value will still fall on the regression line, as it is an unbiased estimator. Nevertheless, the prediction interval will be much wider since there is variability around the line to account for, in addition to the variability in our estimate of the line.

12. y_{new} is a random variable, which is why it is called a prediction rather than estimation (which is for a fixed parameter). However, like our outcomes, we could observe a realization of it. For example, suppose we drew a random sample of 30 patients in a hospital, and fit our model. We could consider sampling one more patient, with outcome y_{new}, which we would assume is normally distributed from the same model as the one from which our data were drawn. We can make a prediction, construct a prediction interval. And then we could hypothetically measure the outcome on this 31st patient (a realization of this random variable), and see how close our prediction was, whether our prediction interval covered this value, etc.

    A 95% PI can be interpreted as: if we repeatedly drew samples and calculated the PI for a fixed covariate value of x_new, then 95% of the time, we would end up with the true value of y_{new} in our interval.

3 Multiple Linear Regression

3.1 Math Review

Denote vectors and matrices with bold, like in CS 480,

3.1.1 Matrix review

Recall from Math 136,

1. Transpose: $[\boldsymbol{C}^T]_{ij} = [\boldsymbol{C}]_{ji}$
2. $\boldsymbol{C}$ is symmetric if $\boldsymbol{C}^T = \boldsymbol{C}$
3. $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$
4. If square matrix $\boldsymbol{B}$ is nonsingular (aka invertible), then $\boldsymbol{B}\boldsymbol{B}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{B} = \boldsymbol{I}$
5. $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ if both $\boldsymbol{A}, \boldsymbol{B}$ are square and nonsingular
6. $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$
7. Trace (of a square matrix):
   (a) $\operatorname{tr}(\boldsymbol{A}) = \sum_{j=1}^{n} a_{jj}$, sum of the diagonals
   (b) $\operatorname{tr}(\boldsymbol{A} + \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B})$
   (c) $\operatorname{tr}(c\boldsymbol{A}) = c\operatorname{tr}(\boldsymbol{A})$
   (d) $\operatorname{tr}\big(\boldsymbol{A}^T\big) = \operatorname{tr}(\boldsymbol{A})$
   (e) $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$

3.1.2 Matrix calculus

1. Let $z = f(y_1, \ldots, y_k)$ and $\boldsymbol{y} = (y_1, \ldots, y_k)^T$, then,

$$\frac{\partial z}{\partial \boldsymbol{y}} = \begin{bmatrix} \frac{\partial z}{\partial y_1} \\ \frac{\partial z}{\partial y_2} \\ \vdots \\ \frac{\partial z}{\partial y_k} \end{bmatrix}$$

2. A special case of (1): If $z = \boldsymbol{a}^T\boldsymbol{y}$, where $\boldsymbol{a} = (a_1, \ldots, a_k)^T$ is a vector, then,

$$\frac{\partial z}{\partial y} = \boldsymbol{a}$$

3. Another special case of (1): If $z = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}$ (quadratic form) where $\boldsymbol{A}$ is a $k \times k$ matrix, then,

$$\frac{\partial z}{\partial \boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{A}^T\boldsymbol{y}$$

   and if $\boldsymbol{A}$ is symmetric, then $\frac{\partial z}{\partial \boldsymbol{y}} = 2\boldsymbol{A}\boldsymbol{y}$

3.2 Random Vectors

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a random vector (ie, each of $y_1, \ldots, y_n$ are random variables). The mean of $\boldsymbol{y}$ is element-wise,

$$E[\boldsymbol{y}] = \begin{bmatrix} E[y_1] \\ \vdots \\ E[y_n] \end{bmatrix}$$

and a random matrix can be similarly defined element-wise.

3.2.1 Covariance matrix

The variance of a vector $\boldsymbol{y}$ is,

$$\text{Var}(\boldsymbol{y}) = E[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T]$$

where $\boldsymbol{\mu} = E[\boldsymbol{y}]$. In matrix form,

$$\boldsymbol{V} = \text{Var}(\boldsymbol{y}) = \begin{bmatrix} \text{Var}(y_1) & \text{Cov}(y_1, y_2) & \cdots & \text{Cov}(y_1, y_n) \\ \text{Cov}(y_2, y_1) & \text{Var}(y_2) & \cdots & \text{Cov}(y_1, y_n) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(y_n, y_1) & \text{Cov}(y_n, y_2) & \cdots & \text{Var}(y_n) \end{bmatrix}$$

Note that $\boldsymbol{V}$ is symmetric and positive semi-definite, meaning that for all $n \times 1$ vectors $\boldsymbol{a}$ in $\mathbb{R}^n$, $\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a} \geq 0$. As well, the linear regression model does not assume that covariates are independent of each other, so $\boldsymbol{V}$ is not just a diagonal matrix.

The covariance matrix is also often called the variance matrix or the variance-covariance matrix.

3.2.2 Linear combinations

Recall that if $z = \sum_{i=1}^n a_i y_i + c$ and $u = \sum_{i=1}^n b_i y_i + d$ (note that z, y_i, u are rvs),

$$E[z] = \sum_{i=1}^n a_i E[y_i] + c \qquad \text{Cov}(z, u) = \sum_{i=1}^n \sum_{j=1}^n a_i b_j \text{Cov}(y_i, y_j)$$

In matrix form, $z = \boldsymbol{a}^T\boldsymbol{y} + c$ and $u = \boldsymbol{b}^T\boldsymbol{y} + d$ and,

$$E[z] = \boldsymbol{a}^T\boldsymbol{\mu} + c \qquad \text{Cov}(z, u) = \boldsymbol{a}^T\boldsymbol{V}\boldsymbol{b}$$

where $\boldsymbol{\mu} = E[\boldsymbol{y}]$ and $\boldsymbol{V} = \text{Var}(\boldsymbol{y})$.

And for $k \times 1$ random vector $\boldsymbol{z} = (z_1, \ldots, z_k)^T$ of k linear combinations of $n \times 1$ random vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$,

$$z_i = a_{i1}y_1 + a_{i2}y_2 + \cdots + a_{in}y_n$$

for $i = 1, \ldots, k$. We can equivalently write,

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{y}$$

where $\boldsymbol{A}$ is a $k \times n$ matrix with elements a_{ij} and,

$$E[\boldsymbol{z}] = \boldsymbol{A}\boldsymbol{\mu} \qquad \text{Var}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T$$

Note that if $\boldsymbol{z} = \boldsymbol{y}^T\boldsymbol{B}$, then $E[\boldsymbol{z}] = E[\boldsymbol{y}^T]\boldsymbol{B}$; a matrix of constants can be pulled out of both the left and right sides. Using this fact combined with $E[\boldsymbol{A}\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{\mu}$, the $\text{Var}(\boldsymbol{z})$ expression comes from,

$$\begin{aligned}
\text{Var}(\boldsymbol{z}) &= \text{Var}(\boldsymbol{A}\boldsymbol{y}) \\
&= E[(\boldsymbol{A}\boldsymbol{y} - E[\boldsymbol{A}\boldsymbol{y}])(\boldsymbol{A}\boldsymbol{y} - E[\boldsymbol{A}\boldsymbol{y}])^T] \\
&= E[\boldsymbol{A}(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{A}(\boldsymbol{y} - E[\boldsymbol{y}]))^T] \\
&= E[\boldsymbol{A}(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T\boldsymbol{A}^T] \\
&= \boldsymbol{A}E[(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T]\boldsymbol{A}^T \qquad \text{; pulling out constant matrices} \\
&= \boldsymbol{A}\text{Var}(\boldsymbol{y})\boldsymbol{A}^T \\
&= \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T
\end{aligned}$$

3.2.3 Summary of Useful Properties of Random Vectors

For fixed (constant) vectors $\boldsymbol{a}, \boldsymbol{b}$, fixed matrix $\boldsymbol{A}$, and random vector $\boldsymbol{y}$,

1. $E[\boldsymbol{a}] = \boldsymbol{a}$
2. $E[\boldsymbol{a}^T\boldsymbol{y} + \boldsymbol{b}] = \boldsymbol{a}^T E[\boldsymbol{y}] + \boldsymbol{b}$
3. $E[\boldsymbol{A}\boldsymbol{y}] = \boldsymbol{A}E[\boldsymbol{y}]$
4. $E[\boldsymbol{y}^T\boldsymbol{A}] = E[\boldsymbol{y}^T]\boldsymbol{A}$
5. $\text{Var}(\boldsymbol{y})$ is the covariance matrix
6. $\text{Var}(\boldsymbol{y}) = E[(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T]$
7. $\text{Var}(\boldsymbol{a}^T\boldsymbol{y}) = \boldsymbol{a}^T\text{Var}(\boldsymbol{y})\boldsymbol{a}$
8. $\text{Var}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\text{Var}(\boldsymbol{y})\boldsymbol{A}^T$

3.3 Multivariate Normal (MVN) Distribution

Let $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ be a random vector of iid standard normal variables, ie z_i is iid $N(0, 1)$. Then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$ (a linear transformation of $\boldsymbol{z}$, a linear combination of $z_1, \ldots, z_n$) has multivariate normal distribution, ie,

$$\boldsymbol{y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $E[\boldsymbol{y}] = \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{y}) = \boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. Note that $\boldsymbol{y}, \boldsymbol{\mu}$ are $n \times 1$ and $\boldsymbol{\Sigma}$ is $n \times n$. The density function of $\boldsymbol{y}$ is,

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}$$

where $|\boldsymbol{\Sigma}|$ is notation for $\det(\boldsymbol{\Sigma})$. Note that if $n = 1$, this is just the Normal distribution.

Some properties of $\boldsymbol{y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (sometimes MVN is just denoted N if it is clear from context that this is a multivariate case),

1. Linearity: if $\boldsymbol{u} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{d}$, then,
$$\boldsymbol{u} \sim MVN(\boldsymbol{C}\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T)$$
2. Marginal distribution: if $\tilde{\boldsymbol{y}} = (y_1, \ldots, y_m)^T \subset \boldsymbol{y}$ is a vector subset of $\boldsymbol{y}$ (ie $m \leq n$ and $y_1, \ldots, y_m$ are also elements of $\boldsymbol{y}$, although not necessarily in that order), then $\tilde{\boldsymbol{y}}$ is MVN-distributed with mean $(\mu_1, \ldots, \mu_m)^T$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m, \mu_{m+1}, \ldots, \mu_n)^T$ and variance is the sub-matrix of $\boldsymbol{\Sigma}$ containing the $1, \ldots, m$ related elements. Thus, it is easy to identify the MVN distribution of a subvector of $\boldsymbol{y}$, given the $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ of the full vector $\boldsymbol{y}$.

3. Conditional distribution: if $\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T)^T$ (ie vector $\boldsymbol{y}$ is the concatenation of two vectors $\boldsymbol{y}_1, \boldsymbol{y}_2$), then $\boldsymbol{y}_1^T | \boldsymbol{y}_2^T$ is MVN-distributed.
4. Independence: If $\boldsymbol{\Sigma}_{ij} = 0$, then $\boldsymbol{y}_i, \boldsymbol{y}_j$ are independent. Note that this is not generally true; $\text{Cov}(X, Y) = 0$ does not generally imply that rvs X, Y are independent (although the inverse does hold).

Exercise: Let rvs $X_1 \sim N(0, 1)$ and $B \sim \text{BERN}(0.5)$, independent of X_1. Now consider the rv $X_1 = X_1$ if $B = 0$ or $X_2 = -X_1$ if $B = 1$. X_2 is clearly Normally distributed, since we've just swapped half the signs to the the opposite and since it is symmetric, the distribution is unchanged. However, $Y = X_1 + X_2$ is not Normal. We know this because $P(Y = 0) = P(X_1 = -X_2) = P(B = 1) = 1/2$. Y is not Normal because while X_1, X_2 are Normal, $\boldsymbol{x} = (X_1, X_2)^T$ is not multivariate Normal; ie we cannot find a $2 \times k$ matrix $\boldsymbol{A}$ such that $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$ for $\boldsymbol{z} = (Z_1, \ldots, Z_k)^T$ where Z_i are iid $N(0, 1)$.

3.4 Multiple Linear Regression

Mutiple linear regression is useful over simple linear regression in situations where there is more than covariate. Suppose we have a dataset of multivariate data collected from patients at a hospital with satisfaction as the outcome and age, severity of condition, and stress as the covariates. Some questions we might ask, which could be answered by multiple linear regression,

1. Is mean satisfaction associated with stress, conditional on age and severity?
2. How does mean satisfaction differ for older patients vs younger patients with the same severity and stress scores?
3. Given a patient's stress, age, and severity, can we predict their satisfaction?
4. Is the (conditional) association between stress and satisfaction different for older patients than for younger patients?

In multiple linear regression with p covariates (assuming $p < n$, where n is the number of observations),

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where ϵ_i is iid $N(0, \sigma^2)$ for $i = 1, \ldots, n$, and,

$$y_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2)$$

where the y_i's are independent. As well, since x_{ij} is fixed, like in simple linear regression, we can consider observations y_i as conditioning on x_i (ie $y_i | x_{i1}, \ldots, x_{ip}$).

We can also represent this formulation using matrices,

$$\begin{bmatrix} y_i \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

which is denoted using the variables,

$$\boxed{\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

In this case,

$$\boxed{\boldsymbol{\epsilon} \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \iff \boldsymbol{y} \sim MVN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})}$$

where $\boldsymbol{0}$ is a $N \times 1$ zero vector. Like in simple linear regression, $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ are generally unknown and must be estimated. As well, $\boldsymbol{X}$ is often called the design matrix. Specifying the contents of $\boldsymbol{X}$ is all that is needed to define the model, since the other vectors are always present.

Interpreting regression coefficients: Note that,

$$E[y_i|x_{i1},\ldots,x_{ip}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_i$$

although often, we simply write $E[y_i]$. As well,

$$E[y_i|x_{i1} = \cdots = x_{ip} = 0] = \beta_0$$

which says that β_0 is the mean outcome when all covariates are set to 0. However, this is sometimes not interpretable (eg, how can age be 0 in a sample of adults). And,

$$\begin{aligned} E[y_i|x_{i1} = x_1, x_{i2} = x_2, \ldots, x_{ip} = x_p] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_P x_p \\ E[y_i|x_{i1} = (x_1+1), x_{i2} = x_2, \ldots, x_{ip} = x_p] &= \beta_0 + \beta_1 (x_1+1) + \beta_2 x_2 + \cdots + \beta_p x_p \end{aligned}$$

Subtracting the first expectation from the second results in just β_1. More generally, β_j is the difference in mean outcome for a one unit change in the jth covariate, holding all other covariates fixed.

3.5 Estimation

Like in the simple linear regression case, we consider Least Squares and Maximum Likelihood Estimation to estimate $\beta_0, \ldots, \beta_P, \sigma^2$,

1. Least Squares: We want to minimize the sum of squares,

$$\begin{aligned} S(\boldsymbol{\beta}) &= . \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_i))^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

   Setting the partial derivative wrt $\boldsymbol{\beta}$ (using the matrix calculus properties above) to 0 and solving,

$$0 = \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \implies (\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$$

   And if the columns of $\boldsymbol{X}$ are linearly independent (which is usually the case, if the number of covariates is small compared to the number of data points), then,

$$\boxed{\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}$$

   This is a very important result in Statistics. See also the Linear Regression section of my CS 480 notes (although this formula is represented using sums rather than matrix multiplication)

2. Maximum Likelihood: Recall $\boldsymbol{y} \sim MVN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. The likelihood function is,

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2\boldsymbol{I}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\}$$

   and since $\det(\sigma^2\boldsymbol{I}) = (\sigma^2)^n$, the log likelihood is,

$$\ell(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

   And maximizing the log likelihood wrt $\boldsymbol{\beta}$ is equivalent to minimizing $\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ wrt $\boldsymbol{\beta}$, which is exactly the same objective as Least Squares. Thus, $\hat{\boldsymbol{\beta}}_{MLE}$ is the same as the $\hat{\boldsymbol{\beta}}$ derived in LS.

The mean of $\hat{\boldsymbol{\beta}}$, using the properties of random vectors derived earlier, is,

$$E[\hat{\boldsymbol{\beta}}] = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{\beta}$$

Thus, $\hat{\boldsymbol{\beta}}$ is unbiased. For the variance, again using the properties of random vectors from earlier,

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{y})((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{y})\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X})^{-1})^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{y})\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X})^T)^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}$$

Moreover, since $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, then $\hat{\boldsymbol{\beta}}$ is MVN distributed, since it is just a linear transformation of $\boldsymbol{y}$, which is MVN distributed. Thus,

$$\boxed{\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})}$$

Moreover, by property 2 of MVN distributions (see above),

$$\boxed{\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})}$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Next, let the fitted values be $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. This is the multivariate version of $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Expanding,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = [\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called the Hat matrix (or the projection matrix). It maps the outcomes to the vector of fitted values.

Some interesting facts about $\boldsymbol{H}$: it is symmetric ($\boldsymbol{H}^T = \boldsymbol{H}$) and idempotent ($\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$; see Lecture 7 slide 16 for proof). As well, $\boldsymbol{I} - \boldsymbol{H}$ is symmetric and idempotent (see Lecture 8 slide 16 for proof).

The mean of $\hat{\boldsymbol{y}}$ is,

$$E[\hat{\boldsymbol{y}}] = E[\boldsymbol{H}\boldsymbol{y}] = \boldsymbol{H}E[\boldsymbol{y}] = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$$

and the variance is,

$$\text{Var}(\hat{\boldsymbol{y}}) = \text{Var}(\boldsymbol{H}\boldsymbol{y}) = \boldsymbol{H}\text{Var}(\boldsymbol{y})\boldsymbol{H}^T = \boldsymbol{H}\sigma^2\boldsymbol{I}\boldsymbol{H} = \sigma^2\boldsymbol{H}$$

so the distribution of $\hat{\boldsymbol{y}}$ is,

$$\boxed{\hat{\boldsymbol{y}} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H})}$$

Now consider a vector of residuals $\boldsymbol{e}$ (the multivariate version of $e_i = y_i - \hat{y}_i$),

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

this shows that the residuals can also be written as a linear combination of the outcomes. Note that $\boldsymbol{X}^T\boldsymbol{e} = 0$, since,

$$\boldsymbol{X}^T\boldsymbol{e} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}) = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = 0$$

Deriving an estimate for σ using Maximum Likelihood, recall,

$$\ell(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$$

and setting the derivative with respect to σ^2 to 0,

$$\begin{aligned}
0 = \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
\implies n\sigma^2 &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
\implies \hat{\sigma}^2_{ML} &= \frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2 = \frac{1}{n}\boldsymbol{e}^T\boldsymbol{e}
\end{aligned}$$

however, like in simple linear regression, we usually do not use the ML estimate and instead use the unbiased estimator for σ^2,

$$\boxed{\hat{\sigma}^2 = \frac{1}{n-(p+1)}\boldsymbol{e}^T\boldsymbol{e}}$$

We now take a closer look at this unbiased estimator. Recall $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{e} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$. Consider the vector $\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \boldsymbol{e}\end{bmatrix}$, which is a linear combination of $\boldsymbol{y}$. Since $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, then this vector is MVN distributed. We want to find the expectation and variance of this vector. Recall $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. And,

$$\begin{aligned}
E[\boldsymbol{e}] &= E[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}] \\
&= (\boldsymbol{I}-\boldsymbol{H})E[\boldsymbol{y}] \\
&= (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\
&= 0
\end{aligned}$$

Thus, $E\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta} \\ 0\end{bmatrix}$. And it can be shown (see Lecture 7 slides 15, 16) that the variance is $\sigma^2\begin{bmatrix}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{I}-\boldsymbol{H})\end{bmatrix}$. The distribution is then,

$$\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \boldsymbol{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{\beta} \\ 0\end{bmatrix}, \sigma^2\begin{bmatrix}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{I}-\boldsymbol{H})\end{bmatrix}\right)$$

and can also conclude using the theory about marginal MVN distributions that,

$$\boxed{\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I}-\boldsymbol{H}))}$$

and that $\hat{\boldsymbol{\beta}}$ (whose distribution is already described earlier) and $\boldsymbol{e}$ are independent, since the off-diagonal covariance terms in the block matrix variance above are zero matrices and in the case of MVN, this implies independence.

Note that an alternate way to determine $\text{Var}(\boldsymbol{e})$ using the fact that $\boldsymbol{I}-\boldsymbol{H}$ is idempotent is,

$$\text{Var}(\boldsymbol{e}) = \text{Var}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}) = (\boldsymbol{I}-\boldsymbol{H})\text{Var}(\boldsymbol{y})(\boldsymbol{I}-\boldsymbol{H})^T = \sigma^2(\boldsymbol{I}-\boldsymbol{H})$$

3.6 Inference

Recall that in simple linear regression, we showed that $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}$ using the fact that for independent $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$, then $\frac{Z}{\sqrt{V/\nu}} \sim t_\nu$. We want something similar for the multi variate case.

Since $\frac{1}{\sigma^2}\boldsymbol{e}^T\boldsymbol{e} \sim \chi^2_{n-(p+1)}$ (see Lecture 7 slides $21-24$ for a proof using eigen decomposition; outside of the scope of the course), $\frac{1}{\sigma^2}\boldsymbol{e}^T\boldsymbol{e}$ is independent of $\hat{\boldsymbol{\beta}}$ (follows from the fact that $\boldsymbol{e}$ is independent of $\hat{\boldsymbol{\beta}}$ which was shown earlier), and $\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})$, then,

$$\frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma^2 \boldsymbol{V}_{jj}}}}{\sqrt{\left(\frac{1}{\sigma^2}\boldsymbol{e}^T\boldsymbol{e}\right)/(n-(p+1))}} \sim t_{n-(p+1)} \implies \boxed{\frac{\hat{\beta}_j-\beta_j}{\hat{\sigma}\sqrt{\boldsymbol{V}_{jj}}} \sim t_{n-(p+1)}}$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Intuitively, we have $n-(p+1)$ degrees of freedom in the t distribution because we are estimating $p+1$ different β_i values for $i = 0, \ldots, p$.

Hypothesis Testing: Similarly to the simple linear regression case, we want to test a null hypothesis $H_0 : \beta_j = \theta_0$ (often θ_0 is 0) against an alternative $H_1 : \beta_j \neq \theta_0$. We compute the observed test statistic $t^{(obs)}$ under H_0 (assuming it is true),

$$t^{(obs)} = \frac{\hat{\beta}_j-\beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j-\theta_0}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j-\theta_0}{\hat{\sigma}\sqrt{V_{jj}}} \sim t_{n-p-1}$$

and we want to compute $P(|T| \geq |t^{(obs)}|)$, which in R is, `2*pt(|t_obs|, df=n-p-1, lower.tail=FALSE)` and reject if this p value is less than α (eg 0.05). Alternatively, we can find a critical value $t_{n-p-1,1-\alpha/2}$ such that the probability that less than this value is $1-\alpha/2$, implying that the probability to the right is $\alpha/2$. The critical value is a test statistic such that the p value would be exactly α, so we reject if $|t^{(obs)}| \geq t_{n-p-1,1-\alpha/2}$.

As well, recall one-sided tests from Stat 332. If we have $H_0 : \beta_j = \theta_0$ and $H_1 = \beta_j > \theta_0$, we use the same test statistic $t^{(obs)}$, since the null has not changed, but the p value is $P(T > t^{(obs)})$. In R, the command would be `pt(t_obs, df=n-p-1, lower.tail=FALSE)`.

In R, to do a two-sided HT that $H_0 : \beta_j = 0$ on all the coefficients (ie n separate HT; for all $j = 0, 1, \ldots, p$),

```
Model <- lm(data = mydataset, formula = Y ~ X1 + X2 + X3) # implicitly includes intercept
beta.hat <- coef(Model) # numerators of the test statistics
V <- solve(crossprod X) # X is the design matrix
beta.se <- sqrt(sigma(Model)^2 * diag(V)) # denominators of the test statistics (std errs)
beta.se <- sqrt(diag(vcov(Model))) # easier way to do the above using vcov
Tobs <- beta.hat/beta.se # T statistics

pval <- 2 * pt(-abs(Tobs), df = n-p-1) # computes P(|T| > |Tobs|) where T is t(n-p-1)
pval <- 2 * pt(abs(Tobs), df = n-p-1,lower.tail=FALSE) # equivalent to the previous line

# or can just see the pvalues using summary
summary(Model)
```

Confidence Intervals: To construct a $(100(1-\alpha)\%$ CI for a single coefficient β_j, we want,

$$1-\alpha = P\left(-t_{n-p-1,1-\alpha/2} \leq \frac{\hat{\beta}_j-\beta_j}{\hat{\sigma}V_{jj}} \leq t_{n-p-1,1-\alpha/2}\right)$$
$$\implies 1-\alpha = P(\hat{\beta}_j - t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{V_{jj}} \leq \beta_j \leq \hat{\beta}_j + t - n - p - 1, 1-\alpha/2\hat{\sigma}\sqrt{V_{jj}})$$

which implies that a $100(1-\alpha)\%$ CI for β_j is,

$$\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2}SE(\hat{\beta}_j) \implies \boxed{\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{V_{jj}}}$$

3.7 Prediction

Estimating mean response: For an arbitrary row vector of covariates $\boldsymbol{x}_0 = [1, x_{01}, x_{02}, \ldots, x_{0p}]$, the mean (a scalar) is $\mu_0 = E[y_0] = \boldsymbol{x}_0\boldsymbol{\beta}$. We can estimate this as $\hat{\mu}_0 = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$. The mean is (since $\hat{\boldsymbol{\beta}}$ is unbiased),

$$E[\hat{\mu}_0] = E[\boldsymbol{x}_0\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0 E[\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0\boldsymbol{\beta} = \mu_0$$

which shows that $\hat{\mu}_0$ is an unbiased estimator for μ_0. The variance is,

$$\text{Var}(\hat{\mu}_0) = \text{Var}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0\text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0^T = \boldsymbol{x}_0\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T = \sigma^2\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T$$

Since $\hat{\mu}_0 = \boldsymbol{x}_0\boldsymbol{\beta} = \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is a linear combination of $\boldsymbol{y}$ which is Normal, then $\hat{\mu}_0$ is also Normal. Thus,

$$\frac{\hat{\mu}_0 - \mu_0}{\sigma\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim N(0,1)$$

and by the same logic as before,

$$\frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-p-1}$$

and a $100(1-\alpha)\%$ CI is,

$$\boxed{\hat{\mu}_0 \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}}$$

Prediction: For a new response $y_{new} = \boldsymbol{x}_{new}\boldsymbol{\beta} + \epsilon_{new}$ (where $\boldsymbol{x}_{new}$ is a row vector containing the new covariates) which is independent of all the other outcomes, the prediction is $\hat{y}_{new} = \boldsymbol{x}_{new}\hat{\boldsymbol{\beta}}$ which has mean $E[\hat{y}_{new}] = \boldsymbol{x}_{new}\boldsymbol{\beta}$ and variance $\text{Var}(\hat{y}_{new}) = \sigma^2\boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T$. See Lecture 8 slide 21 for the proof, however it is the exact same as the mean and variance of $\hat{\mu}_0$ above except with $\boldsymbol{x}_{new}$ instead of $\boldsymbol{x}_0$.

Furthermore, we know that y_{new} and $\hat{y}_{new}$ are independent (since $\hat{y}_{new}$ is a function of only the observed y not including y_{new}) and normally distributed. So, since $y_{new} \sim N(\boldsymbol{x}_{new}\boldsymbol{\beta}, \sigma^2)$ by assumption,

$$y_{new} - \hat{y}_{new} \sim N(0, \sigma^2 + \sigma^2\boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T)$$

and as before,

$$\frac{y_{new} - \hat{y}_{new}}{\sigma\sqrt{1 + \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T}} \sim N(0,1) \qquad \frac{y_{new} - \hat{y}_{new}}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T}} \sim t_{n-p-1}$$

and a $100(1-\alpha)\%$ prediction interval for y_{new} is,

$$\boxed{\hat{y}_{new} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T}}$$

Like in the simple linear regression case, the extra 1 under the square root captures the variability around the estimated line and the rest captures the variability of estimating the line (mean) itself.

3.8 Categorical Covariates

We now consider categorical covariates, but the outcome is still continuous.

Recall in the multiple linear regression model discussed so far, β_j for $j > 0$ is the mean difference in the outcome for every one unit increase in x_{ij}, holding all other covariates constant. However, this only applies if the covariates are continuous. Even though categorical covariates can be encoded as integers, we do not want to use these integers directly since the numeric values and how widely spaced the integers are imposes additional assumptions on the covariates.

Consider the `fishermen_mercury.csv` dataset (see files) with continuous response `MeHg` and one categorical covariate `fishpart` which can take values N, M, MW, W. Rather than encoding these categories as integers, we instead consider four separate means,

$$\texttt{MeHg}|\{\texttt{fishpart} = N\} \sim N(\gamma_N, \sigma^2) \quad \texttt{MeHg}|\{\texttt{fishpart} = M\} \sim N(\gamma_M, \sigma^2) \quad \cdots$$

In words, the response for data points with N `fishpart` is γ_N, etc. This does not require assumptions on the relative differences between categorical values. The model is then,

$$\texttt{MeHg}_i = \gamma_N I[\texttt{fishpart}_i = N] + \gamma_M I[\texttt{fishpart}_i = M] + \gamma_{MW} I[\texttt{fishpart}_i = MW] + \gamma_W I[\texttt{fishpart}_i = W] + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and are iid and $I[A]$ is 1 if A is true and 0 if A is false (an indicator function). This model essentially has four different intercepts (means) and no slope (ie four horizontal lines when plotting any other non-`fishpart` covariate against the response). An equivalent and more familiar parameterization of this model is,

$$\texttt{MeHg}_i = \beta_0 + \beta_M I[\texttt{fishpart}_i = M] + \beta_{MW} I[\texttt{fishpart}_i = MW] + \beta_W I[\texttt{fishpart}_i = W] + \epsilon_i$$

again with $\epsilon_i \sim N(0, \sigma^2)$ iid. This is the same as the multiple linear regression discussed earlier, where the "covariates" are now indicator functions of the actual covariates and the design matrix would be filled with 0 and 1. The data points with `fishpart` of N is called the referent group, since the indicator is not explicitly included in the model.

The design matrix for this formulation has a column of 1s for the intercept and three columns of 0s and 1s (corresponding to whether the data point has `fishpart` of M, MW, W), a column of 1s. There is no column for the referent group, because from the columns of the three other categories, it can be determined whether the data point is in the referent group (ie if all three columns are 0, then it must be in the referent group; the four indicators for all four categories sum to one since each data point is in exactly one category). As well, the intercept column already is filled with 1s, so there would be linearly dependent columns (and for the $\gamma_N, \gamma_M, \gamma_{MW}, \gamma_W$ model without the intercept, there is no intercept column so this problem does not appear). If the design matrix has linearly dependent columns, then $(\boldsymbol{X}^T\boldsymbol{X})$ would be non-invertible, which would not work with our theory so far. In summary, adding a column for the referent group while there is already an intercept column would be over-parameterizing. See 10:00 in the Lec 10 video for more details.

Furthermore, note that $E[y_i|\texttt{fishpart} = N] = \beta_0, E[y_i|\texttt{fishpart} = M] = \beta_0 + \beta_M, E[y_i|\texttt{fishpart} = MW] = \beta_0 + \beta_{MW}$, etc. This is four unique means described by four unknown parameters $\beta_0, \beta_M, \beta_{MW}, \beta_W$. Relating to the first model,

$$\gamma_N = \beta_0 \qquad \gamma_M = \beta_0 + \beta_M \qquad \gamma_{MW} = \beta_0 + \beta_{MW} \qquad \gamma_W = \beta_0 + \beta_W$$

or alternatively,

$$\beta_0 = \gamma_N \qquad \beta_M = \gamma_M - \gamma_N \qquad \beta_{MW} = \gamma_{MW} - \gamma_N \qquad \beta_W = \gamma_W - \gamma_N$$

so β_0 is the mean of data points with `fishpart` of N, β_M is the mean difference between data points with `fishpart` of M and those of N (the referent group), etc. For example, in an experiment with a placebo, the placebo is the referent group so that the β can represent comparisons to that placebo.

Adding a continuous covariate: Another covariate in the dataset is `weight`, measured in kg, and is continuous. How do we encode a regression model where the expected `MeHg` is linear in `weight` for each level of `fishpart`, with common slope but different intercepts? One way is,

$$\begin{aligned}\texttt{MeHg}_i = \gamma_1 \texttt{weight}_i + \gamma_N I[\texttt{fishpart}_i = N] + \gamma_M I[\texttt{fishpart}_i = M] \\ + \gamma_{MW} I[\texttt{fishpart}_i = MW] + \gamma_W I[\texttt{fishpart}_i = W] + \epsilon_i\end{aligned}$$

where γ_1 is the mean difference in the outcome for a one unit change in `weight`, for any constant value of `fishpart`. And γ_N is the mean outcome when `fishpart` is N and `weight` is 0. Another formulation is,

$$\texttt{MeHg}_i = \beta_0 + \beta_1 \texttt{weight}_i + \beta_M I[\texttt{fishpart}_i = M] + \beta_{MW} I[\texttt{fishpart}_i = MW] + \beta_W I[\texttt{fishpart}_i = W] + \epsilon_i$$

and as usual, ϵ_i is iid $N(0, \sigma^2)$ in both models. In this second model, β_0 is the mean outcome when `weight` is 0 and `fishpart` is N. And, β_1 is the mean difference in outcome with a one unit change in `weight` for any constant value of `fishpart`. Lastly, $\beta_M = E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = M] - E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = N]$; the mean difference in outcome holding `weight` constant and changing `fishpart` from M to N.

The design matrix is the same as before but with an additional column for the `weight`. The first is the design matrix for the formulation without the intercept (note the column for the referent group `fishpart` of N) and the second is the design matrix for the formulation with the intercept (note the lack of the column for the referent group, the intercept takes care of it),

	weight	fishpartnone	fishpartmuscle	fishpartmuscle_whole	fishpartwhole
1	70	0	0	1	0
2	73	0	1	0	0
3	66	0	0	1	0
4	80	0	1	0	0
5	78	0	1	0	0
6	75	0	1	0	0

	(Intercept)	weight	fishpartmuscle	fishpartmuscle_whole	fishpartwhole
1	1	70	0	1	0
2	1	73	1	0	0
3	1	66	0	1	0
4	1	80	1	0	0
5	1	78	1	0	0
6	1	75	1	0	0

Graphically, if plotting `weight` on the x-axis and `MeHg` on the y-axis, if `fishpart` is N, the regression line is $\gamma_N + \gamma_1 \texttt{weight}$, and similarly when `fishpart` is M, MW, W. These four lines are all parallel with slope γ_1 and intercepts γ_N, γ_M, etc. Using the variables in the second model, the common slope would be β_1 and intercepts $\beta_0, \beta_0 + \beta_M, \beta_0 + \beta_{MW}, \beta_0 + \beta_W$.

See R files for how to represent continuous covariates as factors then build a linear model.

3.8.1 Testing

Using this model where the expected `MeHg` is linear in `weight` for each level of `fishpart`, with common slope but different intercepts, how do we test how different groups compare to one another? The β model is easier to work with for testing.

Comparing a group to the referent: Recall N is the referent group. To test whether

$$E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = N] = E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = M]$$

the null hypothesis is $H_0 : \beta_M = 0$. This can be done using the test statistic,

$$\frac{\hat{\beta}_M - 0}{SE(\hat{\beta}_M)} \sim t_{n-(p+1)}$$

and constructing CI for β_M can be done similarly as before.

Comparing a group to another (non-referent): To test whether

$$E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = M] = E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = MW]$$

the null hypothesis is $H_0 : \beta_M = \beta_{MW}$ (ie $\beta_M - \beta_{MW} = 0$). For known σ, the test statistic is,

$$\frac{\hat{\beta}_M - \hat{\beta}_{MW}}{SE(\hat{\beta}_M - \hat{\beta}_{MW})} \sim N(0,1)$$

where,

$$\text{Var}(\hat{\beta}_M - \hat{\beta}_{MW}) = \text{Var}(\hat{\beta}_M) + \text{Var}(\hat{\beta}_{MW}) - 2\text{Cov}(\hat{\beta}_M, \hat{\beta}_{MW}) = \sigma^2(\boldsymbol{V}_{3,3} + \boldsymbol{V}_{4,4} - 2\boldsymbol{V}_{3,4})$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ (the numbers come from "labelling" N, M, MW, W with $1, 2, 3, 4$ respectively. The square root of this variance is the SE. For unknown σ (ie replace σ with $\hat{\sigma}$, the test statistic is $t_{n-(p+1)}$).

Comparing more than two groups at the same time: Suppose we want to know whether the mean `MeHg` varies by `fishpart`, adjusted for weight. The null hypothesis is $H_0 : \boldsymbol{\beta}_\star = (\beta_M, \beta_{MW}, \beta_W)^T = 0$ (ie $\gamma_N = \gamma_M = \gamma_{MW} = \gamma_W$). Recall,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \implies \hat{\boldsymbol{\beta}}_\star \sim N(\beta, \sigma^2\boldsymbol{V}_\star)$$

where $\boldsymbol{V}_\star$ is the corresponding 3×3 sub-matrix of $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$, since $\boldsymbol{\beta}_\star$ is a "subset" of $\boldsymbol{\beta}$ (excluding the β_0, β_1).

We will use the fact that any variance matrix $\boldsymbol{V}$ can be uniquely written as $\boldsymbol{V} = \boldsymbol{L}\boldsymbol{L}^T$ (Cholesky decomposition), where $\boldsymbol{L}$ is a lower triangular matrix with non-negative entries $\boldsymbol{L}_{ii} \geq 0$ on the diagonal. When $\boldsymbol{V}$ is positive-definite (which can be shown is the case), then $\boldsymbol{L}_{ii} > 0$ (meaning $\boldsymbol{L}$ is non-singular since $\det \boldsymbol{L} = \text{tr}(\boldsymbol{L})$.

So, let $\boldsymbol{L}$ such that $\sigma^2\boldsymbol{V}_\star = \boldsymbol{L}\boldsymbol{L}^T$ and define $\boldsymbol{Z} = \boldsymbol{L}^{-1}(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star)$. It can be shown that (see 55:00 in the Lec 9 video, or slide 20) $Z \sim MVN(0, I)$ and,

$$\sum_{j=1}^{q} \boldsymbol{Z}_j^2 = \boldsymbol{Z}^T\boldsymbol{Z} = (\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star)^T(\boldsymbol{L}^{-1})^T\boldsymbol{L}^{-1}(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star) = \frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star)$$

where q is the dimension of the vector $\hat{\boldsymbol{\beta}}_\star$ (in this case, 3) and $\boldsymbol{Z}_j$ is the jth entry of $\boldsymbol{Z}$.

So, under $H_0 : \boldsymbol{\beta}_\star = 0$,

$$\frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star) = \sum_{j=1}^{q} \boldsymbol{Z}_j^2 \sim \chi^2_q$$

since each $\mathbb{Z}_j \sim N(0,1)$. As well, this is independent of,

$$\frac{n-(p+1)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-(p+1)}$$

since the $(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)$ is a function of the $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ is a function of the residuals $\boldsymbol{e}$ and we showed earlier that these are independent. Now, define an F statistic,

$$\boxed{F = \frac{\frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)/q}{\frac{n-(p+1)}{\sigma^2}\hat{\sigma}^2/(n-(p+1))}} = \frac{(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)}{q\hat{\sigma}^2}$$

Recall (from Stat 332), if $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$ are independent, then $W = \frac{X_1/\nu_1}{X_2/\nu_2}$ has an F distribution,

$$W \sim F(\nu_1, \nu_2) \qquad f(w) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left(\nu_1^{\nu_1}\nu_2^{\nu_2}\frac{w^{\nu_1-2}}{(\nu_2+\nu_1 w)^{(\nu_1+\nu_2)}}\right)^{1/2}$$

Thus, under the null,

$$\boxed{F = \frac{(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)}{q\hat{\sigma}^2} \sim F(q, n-(p+1))}$$

where again, q is the length of the $\hat{\boldsymbol{\beta}}_\star$ vector. In R, the code for testing the null hypothesis is `pf(F_obs, df1=q, df2=n-p-1, lower.tail=FALSE)` (note that the F distribution is non-negative). A one-sided F test is not possible, since the numerator of the test statistic can be considered the square of $\hat{\boldsymbol{\beta}}_\star$ in matrix form, so large positive values cannot be distinguished from large negative values.

Comparison to t-test: In the above formulation, when $q = 1$, the F test is equivalent to the two-sided t test. That is, the t test for $H_0 : \beta_j \neq 0, H_1 : \beta_j \neq 0$ is equivalent to an F test for the same hypothesis. This is because the F statistic simplifies to the square of the t statistic, and it can also be shown that if $T \sim t_\nu$, then $T^2 \sim F(1, \nu)$. However, this is only for the two-sided alternative, since the F test cannot perform one-sided tests, as explained earlier. See Quiz 3 Q1 for more details.

3.8.2 Summary

The only thing which changes with categorical covariates compared to the previous theory when we exclusively considered continuous covariates is the design matrix $\boldsymbol{X}$. This matrix obviously can only contain numbers rather than categories like "low", "medium", etc. The way we incorporate data into $\boldsymbol{X}$ (eg, whether we include a column of $0, 1, 2, 3$ corresponding to `fishpart` of N, M, MW, W, or we include 3 columns of indicators for the 3 non-referent categories, etc) has implications on how we interpret our model parameters.

However, the LS estimators $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ are still the same type of object whether or not we are dealing with 0s and 1s in $\boldsymbol{X}$, or just continuous covariates. $\hat{\boldsymbol{\beta}}$ is still just a linear transformation of the continuous outcomes $\boldsymbol{y}$. So, all the previous theory carries over. As well, the variance matrix for $\hat{\boldsymbol{\beta}}$ is computed exactly the same as before, with the only difference being that $\boldsymbol{X}$ could include 0s and 1s instead of just continuous numbers, etc.

3.9 Interaction

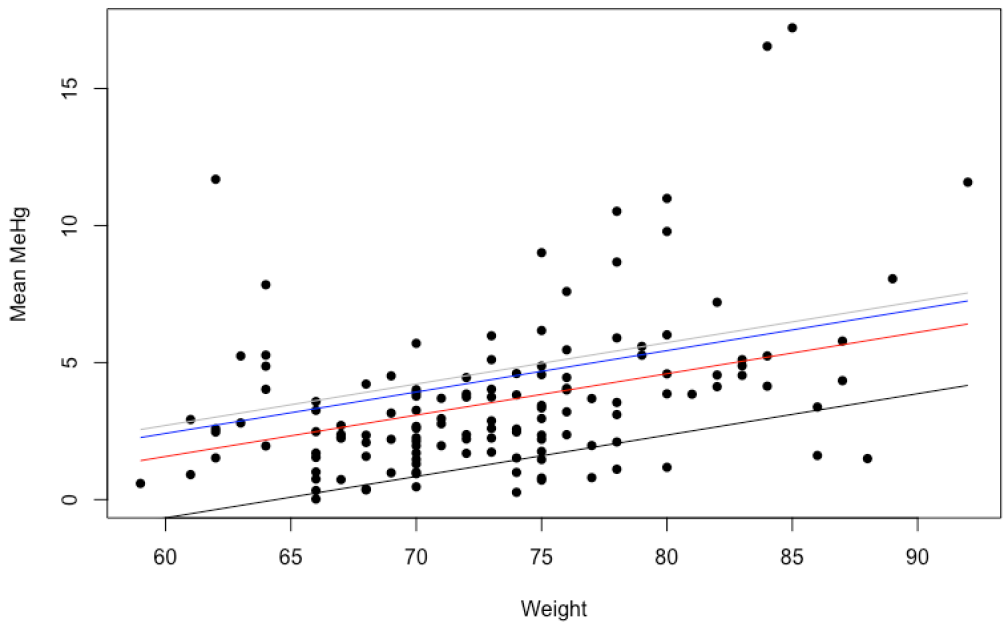
Recall,

$$\texttt{MeHg}_i = \beta_0 + \beta_1\texttt{weight}_i + \beta_M I[\texttt{fishpart}_i = M] + \beta_{MW} I[\texttt{fishpart}_i = MW] + \beta_W I[\texttt{fishpart}_i = W] + \epsilon_i$$

where ϵ is iid $N(0, \sigma^2)$ and the referent group is `fishpart` of N. For simplicity of notation, we will re-write this model as,

$$\texttt{MeHg}_i = \beta_0 + \beta_1\texttt{weight}_i + \beta_M M_i + \beta_{MW} MW_i + \beta_W W_i + \epsilon_i$$

which implies a common slope for `weight` for any value of `fishpart` but different intercepts for different `fishpart`. Plotted in R (see `week5(lec9+10).R` for code), where each line is for a different category,



The black data points are not color-coded by category, but depending on which dots belong to which category, it can be seen that forcing every fitted line to have the same slope can be quite restrictive.

To have both different intercepts and different slopes, we consider the model,

$$\begin{aligned}\texttt{MeHg}_i = \beta_0 + \beta_1 \texttt{weight}_i + \beta_M M_i + \beta_{MW} MW_i + \beta_W W_i \\ + \beta_{1M} \texttt{weight}_i M_i + \beta_{1MW} \texttt{weight}_i MW_i + \beta_{1W} \texttt{weight}_i W_i + \epsilon_i\end{aligned}$$

where ϵ_i is iid $N(0, \sigma^2)$. The product of two different covariates (ie $\texttt{weight}_i M_i$) is called an interaction term. The design matrix for this now has a column of 1s, a column for `weight`, three columns of 0 and 1 for M_i, MW_i, W_i, and three additional columns of $\texttt{weight}_i M_i, \texttt{weight}_i MW_i, \texttt{weight}_i W_i$,

	(Intercept)	weight	fishpartmuscle	fishpartmuscle_whole	fishpartwhole	weight:fishpartmuscle
1	1	70	0	1	0	0
2	1	73	1	0	0	73
3	1	66	0	1	0	0
4	1	80	1	0	0	80
5	1	78	1	0	0	78
6	1	75	1	0	0	75

	weight:fishpartmuscle_whole	weight:fishpartwhole
1	70	0
2	0	0
3	66	0
4	0	0
5	0	0
6	0	0

Note that if $N_i = 1$ (ie `fishpart` is N), then,

$$E[\texttt{MeHg}_i | \texttt{weight}_i = w, N_i = 1] = \beta_0 + \beta_1 \texttt{weight}_i$$

so the mean outcome when `fishpart` is N is linear in the `weight`. If $MW_i = 1$, then,

$$E[\texttt{MeHg}_i | \texttt{weight}_i, MW_i = 1] = \beta_0 + \beta_1 \texttt{weight}_i + \beta_{1MW} \texttt{weight}_i + \beta_{MW} = (\beta_0 + \beta_{MW}) + (\beta_1 + \beta_{1MW}) \texttt{weight}_i$$

The mean outcome is again linear in `weight` but with a different intercept and slope as in the case when $N_i = 1$. Using this model, rather than parallel lines as before with slope β_1, there are four lines with possibly different slope and intercept, allowing more flexibility in how the data is modelled.

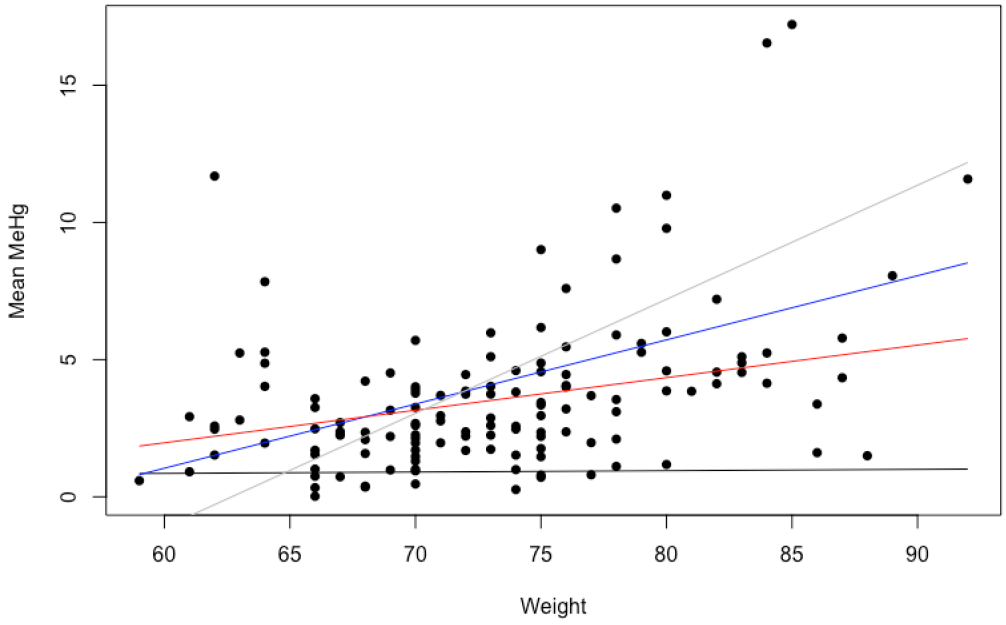
Next, β_1 is the mean difference in response with a one unit increase in `weight`, among those in the group where `fishpart` is N, since,

$$\begin{aligned}\beta_1 &= (\beta_0 + \beta_1(x^* + 1)) - (\beta_1 + \beta_1 x^*) \\ &= E[\texttt{MeHg}_i|\texttt{weight}_i = x^* + 1, N_i = 1] - E[\texttt{MeHg}_i|\texttt{weight}_i = x^*, N_i = 1]\end{aligned}$$

Similarly, $\beta_{1MW} + \beta_1$ is the mean difference in response with a one unit increase in `weight`, in the group where `fishpart` is MW,

$$\begin{aligned}\beta_{1MW} + \beta_1 &= (\beta_0 + \beta_{MW} + \beta_1(x^* + 1) + \beta_{1MW}(x^* + 1)) - (\beta_0 + \beta_{MW} + \beta_1 x^* + \beta_{1MW} x^*) \\ &= E[\texttt{MeHg}_i|\texttt{weight}_i = x^* + 1, MW_i = 1] - E[\texttt{MeHg}_i|\texttt{weight}_i = x^*, MW_i = 1]\end{aligned}$$

Thus, β_{1MW} (a coefficient of an interaction term) is a difference of differences; the increase in the slope of `weight` in the MW group from what it was in the referent group (ie from the referent group slope to the MW group slope). This is a change in the association between `weight` and outcome. This interaction model (note the different slopes and intercepts for the fitted lines for each of the four categories) plotted in R,



3.9.1 Interactions of continuous covariates

Consider,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

where x_{i1}, x_{i2} may be continuous or categorical. The $x_{i1}x_{i2}$ is called an interaction term (or effect modifier) and x_{i1} and x_{i2} are the main effects. Note that,

$$\beta_1 = E[y_i|x_{i2} = 0, x_{i1} = x^* + 1] - E[y_i|x_{i2} = 0, x_{i1} = x^*]$$

So, β_1 is the difference in mean outcomes for a one unit change in x_{i1} when x_{i2} is 0, which may or may not be interpretable.

A better way to interpret this model is: suppose x_{i2} is fixed at x^*. Then,

$$E[y_i|x_{i1}, x_{i2} = x^*] = \beta_0 + \beta_1 x_{i1} + \beta_2 x^* + \beta_3 x_{i1} x^* = (\beta_0 + \beta_2 x^*) + (\beta_1 + \beta_3 x^*) x_{i1}$$

This shows that at every level of x_2, the conditional mean outcome is linear in x_1 and the intercept and slope of x_1 are different. In other words, a change in mean outcome due to a one unit change in x_1 varies with x_2. In practise, to interpret models like this, first choose some reasonable values to fix x_2 to, then report the change in mean outcome for a one unit change in x_1 at these values.

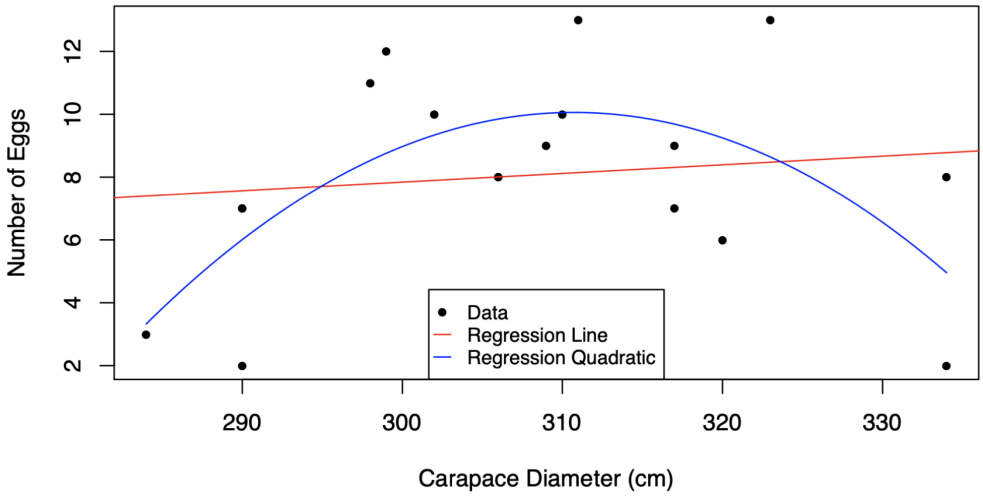
3.10 Non-Linearities

3.10.1 Quadratic Model

Sometimes $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ does not fit the data well. One way to make this model more flexible is to include a quadratic term for x,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

This is still considered a linear regression model, since it is linear to the β coefficients. Linear regression models do not need to be linear in the covariates.



Like in the interaction model, it is hard to explicitly interpret β_1, β_2 in this quadratic model. The change in mean outcome for a one unit change in x_i varies with x_i.

Testing: To test whether the quadratic model is more appropriate than the simple linear model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, we can perform HT on $H_0 : \beta_2 = 0$.

3.10.2 Non-Linear Terms

Beyond polynomial terms, linear regression can be specified quite flexibly,

$$y_i = \sum_{j=1}^{p} \beta_j f_j(x_i) + \epsilon_i$$

where $f_j(\cdot)$ are arbitrary known functions of x_i. Recall generalized linear regression from CS 480 and how kernels can convert data points to a different space to get a non-linear separator in the original space. However, using complex functions can cause overfitting. As well, there is a tradeoff between fit and interpretability.

3.10.3 Hierarchical Principle

Generally, we want to fit hierarchically well-formulated models.

If there is a higher order interaction term, the main effects and lower order interaction terms should also be included. For example, if including x_1x_2, also include x_1 and x_2. If including $x_1x_2x_3$, also include $x_1x_2, x_1x_3, x_2x_3, x_1, x_2$.

If there is a higher order polynomial term, also include the main effects and lower order terms. For example, if including x^3, also include x_2 and x.

Otherwise, there can be unexpected interpretations and implications. For example, suppose we fit the model $y_i = \beta_0 + \beta_2 x_i^2 + \epsilon_i$ and we want to center a covariate to have mean 0 (ie "shift the exposure" by some fixed amount b, where $b = \bar{x}$),

$$y_i = \beta_0 + \beta_2(x_i - b)^2 + \epsilon_i = (\beta_0 + b^2\beta_2) + (-2b\beta_2)x_i + \beta_2 x_i^2 + \epsilon_i$$

And there is now a linear term which was not there before, fundamentally changing the model structure, from just a shift in the data.

4 Model Building

5 Model Diagnostics

6 Extensions

7 Other

7.1 Assignments

Further facts proven on Assignment 1,

1. $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$
2. $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where the first sum is $SS(Total)$, second is $SS(Reg)$, and third is $SS(Res)$
3. $SS(Reg)/SS(Total) = r^2$, where r is the sample correlation between x_i and y_i
4. If $\hat{y}_i = y_i$ for all $i = 1, \ldots, n$, then $r = \pm 1$ (the sample correlation between x_i's and y_i's) and all points lie exactly on the fitted regression line

7.2 Miscellaneous Things

1. CI and HT relationship: recall from Stat 231 that if the same pivotal quantity is used to construct a CI for θ and a test of the hypothesis $H_0 : \theta = \theta_0$, then the parameter value $\theta = \theta_0$ is inside a $100q\%$ CI for θ iff the p value for testing $H_0 : \theta = \theta_0$ is greater than $1 - q$. See the Stat 231 book page 191 for proof.
2. Suppose we construct a single 95% CI. By definition of CI, we expect 95% of CIs constructed using the same procedure to contain the true parameter value across repeated samples, but there is no such assumption for a single CI. This is inherent to the classical/frequentist paradigm of statistics (frequentist meaning that probabilities are interpreted as relative frequencies across repeated trials).

   Alternatively, in the "Bayesian" paradigm (comes from Bayes' theorem), one could construct what are known as "credible intervals", which have an interpretation that is closer to what you want. That is, a credible interval is one within which an unobserved parameter value falls with a particular probability. These Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist CIs treat their bounds as random variables and the parameter as a fixed value.
3. See Lecture 7 slide 15 for how block matrix multiplication works.

7.3 R

```
data <- read.csv("multivariatedata.csv")
pairs(data) # gives a cool visual comparison

Model <- lm(Y ~ X + Y + Z)
X <- model.matrix(M) # get the design matrix X

# See the realestate.csv plot in week5(lec9+10).R for plotting linear models with
# categorical covariates, where data is color-coded and of different shape.
# Also, how to fit a model for: different intercept among categories but same slope,
# different intercept and different slope, different slope and same intercepts.

# fit a quadratic model y_i = beta0 + beta1x_i + beta2(x_i)^2 + error_i
Model <- lm(Y ~ X + I(X^2), data = mydata)
```