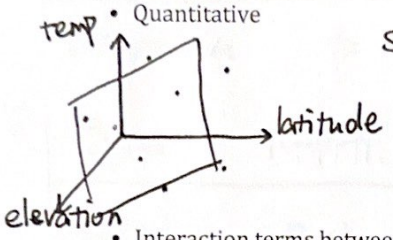
SLR y and x are both quantitative

Multiple Regression (Ch. 13)

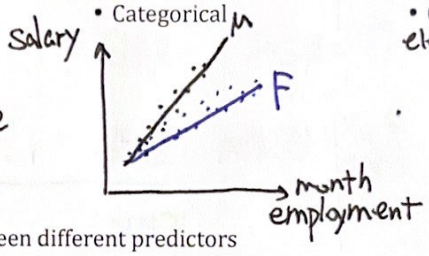
Basics of Multiple Regression

In multiple regression, we have one response variable Y (quantitative) and several predictor variables, $X_1, X_2, X_3, \ldots, X_p$, where p = # of predictors. The predictors can be:

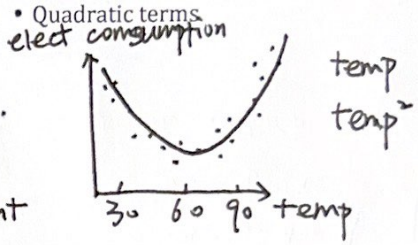
- Quantitative



- Categorical



- Quadratic terms



- Interaction terms between different predictors

$X_1X_2 \quad X_1X_3 \quad X_3X_4 \quad X_1X_2X_3 \quad \cdots$

In practice, we almost always use computer software to compute this. Instead, we will focus on:

- Reading computer output
- Interpreting coefficients for each predictor
- Picking the simplest model that does a good job for predicting y.

Multiple Regression Model:

- **Equation:** $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$

  Predictor variables: $X_1\ X_2 \cdots X_p$

  coefficients of variables: $\beta_1\ \beta_2 \cdots \beta_p$

  α = constant term

  p = # of predictors in model

  # of parameters: $p+1$

- **Model Assumptions:**

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

- **Fitted Equation:**

$$\hat{y} = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

- **Estimators:**

$a, b_1, b_2, \ldots, b_p$ (for $\alpha, \beta_1, \beta_2, \ldots, \beta_p$) and $S = \sqrt{MSE}$ (for σ)

- **Coefficient of Determination, R^2:**

$$R^2 = \frac{SSR}{SST}$$ (SSR → Regression, SST → Total)

% of variability in y explained by the regression model.

Are there any good predictor variable in the model?

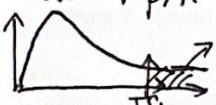
- **ANOVA for Multiple Regression:**

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ → no good predictors

H_a: at least one $\beta_i \neq 0$ → at least one good predictors.

Source	df	SS	MS	F
Regression	p	SSR	$MSR = \frac{SSR}{p}$	$\frac{MSR}{MSE}$
Error	n−p−1	SSE	$MSE = \frac{SSE}{n-p-1}$	
Total	n−1	SST		

TS: $F_{obs} = \frac{MSR}{MSE}$ p-val: $F_{obs} \sim F_{p, n-p-1}$ under H_0



$P(Z \geq F_{obs})$, where $Z \sim F_{p,n-p-1}$

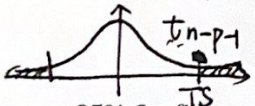
Is this predictor variable good?

- **t-test for Individual Predictors:**

$H_0: \beta_i = 0$ $H_a: \beta_i \neq 0$

TS: $t_{obs} = \frac{b_i - 0}{s.e._i} = \frac{b_i}{s.e._i}$

p-val: $2P(Z \geq |t_{obs}|)$ where $Z \sim t_{n-p-1}$



* These t-tests tell us if that pred variable X_i provides significant information about y AFTER all other pred variables are accounted for.

- **95% Confidence Interval for β_i:**

$b_i \pm t_{n-p-1, 0.025} \cdot s.e._i$ ← standard error for b_i.

↓ # from t-table w/ desired α conf & df from ERROR.

If CI does NOT include zero, then $\beta_i \neq 0 \Rightarrow X_i$ is a good predictor

Important Issues in Multiple Regression

Oversaturated Models: Suppose we collect data on a bunch of different variables that could be used as predictor variables. We should **not** just blindly add predictors to the model. Why?

- Bigger models are harder to interpret
  simpler models are better
- Sample size should be at least 5 to 20 times larger than # predictors.

oversaturated model — when $n = p+1$ (sample size = # parameters)

$R^2 = 100\%$

Even though the model predicts perfectly for data set, it's NOT useful for any population!

Adjusted R^2: One of the shortcomings of R^2 is that it **only** increases or stays the same if a new predictor x_{p+1} is added to the model. This is true even if the new predictors are bad. How do we know if the new predictor is actually useful? We look at **adjusted** R^2.

ANOVA vs t tests: We should always perform ANOVA **first** to see if there are *any* good predictors in our model. If there are not (i.e. p-value > 0.05), then we do not proceed. However, a small ANOVA p-value by itself is not as useful in this case, because it could be the case that only one or a few of the β_i's are not 0. We should look at the p-values of the **individual** predictors determined by the t-test for β_i.

Multicollinearity: We want the predictors (x's) to be correlated with the response (y). But if several of the predictors are highly correlated with each other, they are not adding anything new to predict y. Each x may be a good predictor by itself, but they should not be used together in the model.