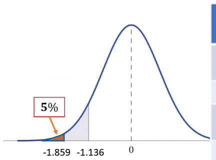
- Step 2: Compute the t-value
  - ○ Sample size (n)= 9,
  - ○ Sample mean $(\bar{X})$ = 0.055 = 5.5%
  - ○ Sample stdev (s)= 0.039 = 3.9%
  - ○ Hypothesized mean (μ_0) =0.07 = 7%
  - ○ $T = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{5.5\%-7\%}{\frac{3.9\%}{\sqrt{9}}} = -1.136$



- Step 3: Make a decision
  - ○ Method 1: Critical value method:
    - Critical value = T.INV(0.05,8) or qt(0.05,8) = -1.859 < -1.136 (t-value)
    - Conclusion: We do not reject the null hypothesis, that is sufficient evidence for rejecting the null hypothesis

Step 3: Make a decision

Method 1: Critical value method by hand

- Critical value = T. INV(0.975, 99) or qt(0.975,99) = 1.98 < |-2.2682(t-value)|
- Conclusion: We reject the null hypothesis

Method 2: p-value method by hand

- T. DIST(-2.2682, 99,1) or pt(−2.2682, 99) = 0.0127
- p-value = 2*0.0127= 0. 0254 < 0.05
- Reject H_0 at α= 0.05 & 0.1 *Do not buy sheets from the supplier!*
- What about α= 0.01? → Do not reject! -> *Hire the supplier!*

Compute test statistic:

- ○ Given n = 400, 23 complaints had late responses

$$\hat{p} = \frac{23}{400} = 0.0575$$

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.0575 - 0.075}{\sqrt{\frac{0.075(1-0.075)}{400}}} = -1.329$$

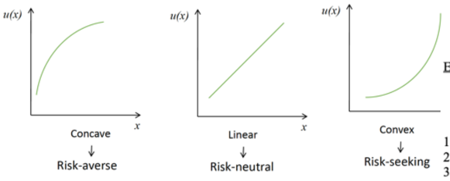
- ○ Critical value method:
  $z_\alpha = z_{0.05}$ = NORM.S.INV(0.05) or qnorm(0.05) = -1.645
  Z=-1.329 > -1.645
- ○ p-value method:
  p-value = NORM.S.DIST(-1.329,1) or pnorm(-1.329) = 0. 0920 > 0.05 (= α)
- ○ Decision:
  Do not reject! The proportion of late responses is not statistically sign 7.5%.

The **dependent variable (Y)** is the variable that predict.

- ○ Also called the **response** or target variable

We use one or more **independent variables (X)** dependent variable

- ○ Also called **explanatory** or **predictor** variables

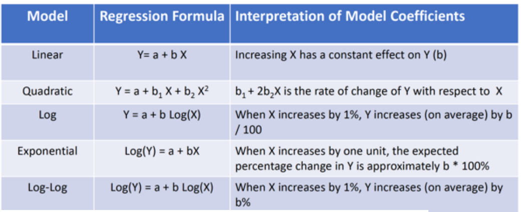


D. Model validity:

1. Linear relationship between X and Y
2. The variance of the dependent variable is constant (constant error variance)
3. The residual (error term) follows a normal distribution with mean = 0 and residuals are independent
4. Independent variables are independent (no multicollinearity)

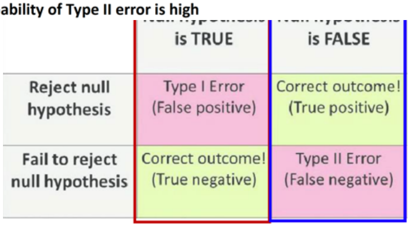
Assumption 2: Constant Error Variance

- Assumption 2 concerns variation around the population regression line.
  - ○ It states that the variation of the *Y*s about the regression line is the *same*, regardless of the values of the *X*s.
    - The technical term for this property is **homoscedasticity**.
    - A simpler term is **constant error variance**.
  - ○ This assumption is often questionable—the variation in *Y* often increases as *X* increases.
  - ○ **Heteroscedasticity** means that the variability of *Y* values is larger for some *X* values than for others.
    - A simpler term for this is **nonconstant error variance**.



Model	Regression Formula	Interpretation of Model Coefficients
Linear	Y= a + b X	Increasing X has a constant effect on Y (b)
Quadratic	Y = a + b_1 X + b_2 X^2	b_1 + $2b_2$X is the rate of change of Y with respect to X
Log	Y = a + b Log(X)	When X increases by 1%, Y increases (on average) by b / 100
Exponential	Log(Y) = a + bX	When X increases by one unit, the expected percentage change in Y is approximately b * 100%
Log-Log	Log(Y) = a + b Log(X)	When X increases by 1%, Y increases (on average) by b%

Ideally probability of Type I error should be low. However, **if α is set very low, then the probability of Type II error is high**



	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

Mean Absolute Error (MAE) = $\frac{1}{n}\sum_{i=1}^{n}|e_i|$. Gives the magnitude of the average absolute error. On average how much did I miss by?

Root Mean Squared Error (RMSE) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}{e_i}^2}$. Gives standard error of estimate in linear regression, computed on validation set.

Mean Absolute Percentage Error (MAPE) = $100 \times \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right|$. Gives a percentage score of how predictions deviate (on average) from the actual values.

- Test statistic: T = b / s_b
  - ○ Confidence interval: b ± t-multiple * s_b
- Interpretations
  - ○ **Small p value**: Reject H_0
    - Strong evidence that β ≠ 0
    - **Independent variable is meaningful**

population slope (β):

$$\boxed{EMV = \sum_{i=1}^{k} v_i p_i}$$

p_i =probability of outcome i

v_i = payoff with outcome i

Standard error of the estimate (s_e)

- ○ Essentially the standard deviation of the residuals

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

$e_i = Y_i - \hat{Y}_i$ (observed – fitted) is the residual of the i^th observation

1) For Explanatory task: may use everything! Select variables via stepwise selection

For Predictive task: Independent variables must proceed dependent and should be available at the time of prediction. Thus, # of bids or bidders are not included

Explanatory Model

A. Statistical interpretation
B. Statistical significance: p-value <
C. Model fit: R^2
D. Four Validity

1. Linear relationship between x and y
2. The variance of the dependent variable is constant (constant error variance) (homoscedasticity)
3. The residual (error term) follows a normal distribution with mean = 0 and residuals are independent
4. Independent variables are independent (no multicollinearity)

Multicollinearity:
- Fairly strong linear relationship among two or more independent variables
- Effects model interpretation
Use Correlation and Variance Inflation Factor (VIF)

Customer Complaints: CI

1. **Multicollinearity**
2. Omitted Variable Bias (OVB)
3. Outliers
4. Simpson paradox

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 0.0575 \pm 1.96\sqrt{\frac{0.0575(1-0.0575)}{400}} = [0.035, 0.080]$$

- Assumption 3 is equivalent to stating that the residuals are normally distributed and independent.
- Random residuals, no pattern or trend when plotting residuals

- Apply the law of total probability

$$P(P_{Good}) = P(P_{Good} \cap A_{Good}) + P(P_{Good} \cap A_{Bad})$$
$$= P(P_{Good}|A_{Good})P(A_{Good}) + P(P_{Good}|A_{Bad})P(A_{Bad}) = \frac{1}{2}$$
$$P(P_{Bad}) = 1 - P(P_{Good}) = \frac{1}{2}$$

- Apply Bayes' rule:

$$P(A_{Good}|P_{Good}) = \frac{P(A_{Good} \cap P_{Good})}{P(P_{Good})}$$
$$= \frac{P(P_{Good}|A_{Good})P(A_{Good})}{P(P_{Good})} = 0.64$$
$$P(A_{Good}|P_{Bad}) = \frac{P(P_{Bad}|A_{Good})P(A_{Good})}{P(P_{Bad})} = 0.16$$

Question 1

a) What is the sampling distribution of the average income from the parking lot over 45 days?

Normal ($850, $150/√45)

b) Write out the null and the alternative hypotheses for the manager. Clearly state them in terms of your parameter.

Let μ = the true daily mean income
H_o: $\mu \geq \$850$
H_a: $\mu < \$850$

Example of wrong answers:
H_0: The attendant is not cheating, $\mu \geq \$800$, $38,250, $36,000, ≤ 2,250, ...
H_a: The attendant is cheating, $\mu > \$850$, ...

c) What is the risk of Type I and Type II error? Clearly state them in English.

Type I: The mean income is $850 (the attendant is not cheating), but we conclude that it is less than $850 (the attendant is cheating) and fire the attendant.

Type II: The mean income is less than $850 (the attendant is cheating), but we conclude that it is equal to $850 (the attendant is not cheating) and keep the attendant.

Which one is worse? Type II
Type I: The loss is $1,000
Type II: The loss is $50/day. If we keep him more than 20 days, the loss will be greater than $1,000

How does this influence our significance level we need?
Set the type II error low -> use high $\boldsymbol{\alpha}$ (0.1 or higher)

d) ASSUME that the p-value is 0.0127. Write out the meaning of this p-value in English.

Assuming that the true daily mean income is $850, there is a 1.27% chance that on average the attendant turns in $800/day or less, over the course of 45 days.

Example of wrong answers:
Assuming that the true mean is $850, there is 1.27% chance that on average we get $800 or higher. (It should be an average over 45 days)
Assuming that H_0 is true, there is 1.27% chance that on average we get the result this extreme or more. (You need to specify the extreme value)

(e) Set up a test statistic for the hypothesis test and compute the value
Since the true standard decision is known, we use Z statistics:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{45}}} = \frac{800 - 850}{\frac{150}{\sqrt{45}}} = -2.2$$

(f) Using a 5% level of significance, should the manager fire the attendant? (Use the critical value method)

$$\boldsymbol{|-2.2361| > Z_{0.95} = 1.645}$$

We reject the null and fire the attendant

Question 2

a) Given the variables specified above, name two variables for which it would be reasonable to convert to dummy variables to be used in a regression model.

Age, History, or Catalogs; Gender and Close are already dummy variables in their current forms.

b) Is this model valid as an explanatory model? Why or why not. Explain your answer.

The model is not valis. The residual plot shows that the residual is non-normal with non-constant variance.

b) Provide an interpretation of the p-value for the Salary.

The small p-value implies that the coefficient of Salary variable is not equal to zero. Therefore, we should include Salary in our regression model.

c) Specify two pieces of information from these results that support the belief that this model is an improvement over the initial model?

1. The multiple-R and the R-square values are improved.
2. All p-values are small.
3. Nonlinear patterns in fitted vs. actual Log(AmountSpent) have disappeared and residuals vs. fitted values shows that the variance of residual is now constant (residual is homoscedastic).

Note: You cannot use the standard error to compare the two models.

d) Provide an economic interpretation for the coefficient of the Catalogs = 18 variable.

On average the customers who received 18 catalogs spend 57% more compared to the customers who received 6 catalogs, when the rest of conditions remains the same

e) Predict AmountSpent for a customer with a household income of $100,000 who lives close to stores that sell similar merchandise, was a high spender in the previous year, and received 6 catalogs. You may round each coefficient to the hundredths place. You must show your work to receive full credit.

f)

Log(AmountSpent)=5.9962 -0.2741 + 1.3796+0.0871=7.1888
AmountSpent = $e^{7.1888}$

Question 3

(a) What is the price elasticity of demand (the change in demand in relation to a change in its price) on routes on which Southwest is present?

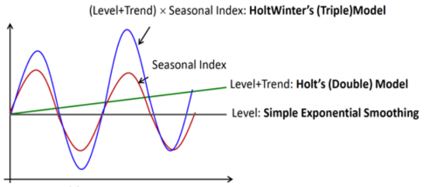
Price Elasticity = Coef of log(FARE) + Coef of SW*log(FARE) = -0.2703-0.3535= -0.6238

(b) Predict demand for a route on which Southwest does not fly and the fare is $500. You may round each coefficient to the hundredths place. You must show your work to receive full credit.

Log (PAX(demand)) = 10.52 +-0.27*ln(500) = 8.8421
PAX = exp(8.8421)= 6919

(D)
Where should the power company build the plant?
The decision depends on the first stage decision and the geologist's prediction. Suppose that the geologist is hired. The plant should be built at Pleasantville if he predicts an earthquake at Chico. Otherwise, it should be built at Chico. If the geologist is not hired, it should be built at Chico.

Should the company hire the geologist? What is the EVI?
The tree shows that the geologist's information is very valuable. The information lowers the expected cost to $17.7m from $18m. The EVI is (-17.7+1.5) – (-$18 m) = $1.8 mil. Therefore, the company would be justified in paying the geologist's fee.


(Level+Trend) × Seasonal Index: HoltWinter's (Triple)Model
Seasonal Index
Level+Trend: Holt's (Double) Model
Level: Simple Exponential Smoothing

- Interpretable
  - Each coefficient corresponds to one component
  - Easy to generate forecasts into the future
- Flexible
  - Easily incorporate external factors (other than time and seasonal factors) into models

    $Y_t = a + b_T * T + b_1 * S_1 + b_2 * S_2 + \dots b_{M-1} * S_{M-1} + c_1 E_1 + \dots + c_N E_N + \varepsilon$
    - Temperature, precipitation, repackaging, introducing a new product
- Inflexible
  - **Stationarity assumption:** Assumes that mean, trend, and seasonality are all constant over time
  - **Static model**: Doesn't allow for changes over time