# Soc3811 Third Midterm Exam

**SEMI-OPEN NOTE:**

**One sheet of paper, signed**

**& turned in with exam booklet**

**Bring Your Own Pencil with Eraser**

**and a Hand Calculator!**

# Prediction with a regression equation

**Use a regression equation to estimate the scores on the dependent variable for different values of the independent variable(s)**

Prediction the midterm scores (Y) from number of beers consumed the night before the test (X)

$$\hat{Y}_i = 93 - 4.8X_i$$

| $X_i$ | $\hat{Y}_i$ |
|---|---|
| **0** | |
| **1** | |
| **3** | |
| **12** | |

# t-tests for Hypotheses about β

**Bivariate & multivariate regression use the same t-test for hypotheses about the linear effect of an independent variable on the dependent variable.**

**Two-tailed hypothesis pair:**

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

**One-tailed hypothesis pair:**

$$H_0 : \beta_j \leq 0$$

$$H_1 : \beta_j > 0$$

**t-test where $b_j$ is the sample regression coefficient & denominator is the standard error of the sampling distribution of $\beta_j$**

$$t_{N-K-1} = \frac{b_j - \beta_j}{s_{b_j}}$$

# t-test for regression slope

**Test a hypothesis about the population regression slope (β), using the sample estimate of the standard error for $b_{YX}$**

Conduct a one-tail hypothesis test:

$$H_0 : \beta_{YX} \geq 0$$

$$H_1 : \beta_{YX} < 0$$

Estimate the regression equation ($s_b$ in parentheses):

$$\hat{Y}_i = 14.36 - 4.83 X_i \qquad R^2_{YX} = 0.26$$

$$(2.38) \qquad (2.34)$$

$$t = \frac{b_{YX} - \beta_{YX}}{s_b} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

| α | 1-tail | 2-tail |
|---|---|---|
| .05 | 1.65 | ±1.96 |
| .01 | 2.33 | ±2.58 |
| .001 | 3.07 | ±3.29 |

**Decision:** ________________ **Prob. of Type 1 error:** ____________

**Conclusion:** ____________________________________

# More Examples

**State** **two-tail** **hypothesis tests:**

$$H_0 : \beta_{YX} = 0$$
$$H_1 : \beta_{YX} \neq 0$$

**Estimate eq.**

$$\hat{Y}_i = -3.59 \quad - 6.27\, X_{1i} \quad + 2.84\, X_{2i} \qquad R^2_{YX} = 0.13$$
$$(1.48) \quad (3.89) \qquad (0.90)$$

**Calculate 1st t-test for $\beta_1$ :**

$$t = \frac{b_1 - \beta_1}{s_{b1}} = \underline{\hspace{6cm}}$$

**Decision:** ____________________ **Prob. of Type I error:** ____________

**Calculate 2nd t-test for $\beta_2$ :**

$$t = \frac{b_2 - \beta_2}{s_{b2}} = \underline{\hspace{5cm}}$$

**Decision:** ________________ **Prob. of Type I error:** _____________

**Conclusions:** ______________________________________________________

# Calculate Coefficient of Determination ($R^2$)

**Use ratios of sums of squares to calculate the coefficient of determination, a proportional measure of the variance in Y predicted by its linear relationship with X**

$$R^2_{YX} = \frac{SS_{TOTAL} - SS_{ERROR}}{SS_{TOTAL}} = \frac{SS_{REGRESSION}}{SS_{TOTAL}}$$

Find R-square for regression of new Florida votes for George Bush on the number of “pregnant chads”

**$SS_{REGRESSION}$ = ______**

**$SS_{ERROR}$ = 750**

**$SS_{TOTAL}$ = 2750**

**$R^2$ = __________**

# Mean Squares

**In bivariate regression, calculate the mean (average) sums of squares per degree of freedom, for the three components in R-square**

$$\frac{SS_{TOTAL}}{N-1} = \frac{SS_{REGRESSION}}{1} + \frac{SS_{ERROR}}{N-2}$$

$$= MS_{REGRESSION} + MS_{ERROR}$$

**The ratio of these two mean squares is the F-statistic with 1 and N-2 degrees of freedom:**

$$F_{1,N-2} = \frac{MS_{REGRESSION}}{MS_{ERROR}}$$

# Test Hypothesis about $\rho^2$

**In the population, does the linear relationship of the variables statistically "explain" some nonzero proportion of the variation in the dependent variable.**

**In other words, is Rho-square greater than 0?**

$$H_0: \rho^2_{YX} = 0$$

$$H_1: \rho^2_{YX} > 0$$

$$\hat{Y} = 13.75 + 0.76X$$

$$R^2 = 0.019 \qquad N = 200$$

$SS_{REGRESSION}$ = 63

$SS_{ERROR}$ = 3,253

$SS_{TOTAL}$ = 3,316

**How many *df* for F?**

________________________

# ANOVA table and *F*-test

Use sums of squares and degrees of freedom in the summary ANOVA table to compute the F-ratio to test $H_0$:

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Regression | 63 | | | |
| Error | 3,253 | | | |
| Total | 3,316 | | -------------------- | |

| $\alpha$ | $df_R$, $df_E$ | c.v. |
|---|---|---|
| .05 | 1, ∞ | 3.84 |
| .01 | 1, ∞ | 6.63 |
| .001 | 1, ∞ | 10.83 |

**Decision:** ____________________

**Probability of Type I error:** ________________

**Conclusion:** ____________________________________________

# F-test for Multiple $R^2$

**In the test of $\rho^2$ for multiple regression, the degrees for freedom for $SS_{REGRESSION}$ = k and for $SS_{ERROR}$ = N-k-1**

**Test the null hypothesis for these data from an equation with four independent variables and sample N=151:**

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Regression | 587 | | | |
| Error | 5,220 | | | |
| Total | 5,807 | | -------------------- | |

| $\alpha$ | $df_R, df_E$ | c.v. |
|---|---|---|
| .05 | 4, ∞ | 2.37 |
| .01 | 4, ∞ | 3.32 |
| .001 | 4, ∞ | 4.62 |

**Decision:** ____________________

**Probability of Type I error:** ______________

**Conclusion:** ____________________________________________________

# Difference in $\rho^2$ for Nested Equations

**Use F-test to decide whether adding predictors to a second, nested regression equation increases $\rho^2$:**

$$H_0: \rho_2^2 - \rho_1^2 = 0$$

$$H_1: \rho_2^2 - \rho_1^2 > 0$$

where subscripts "1" and "2" refer to the equations with fewer and more predictors, respectively

**The F-statistic tests the increase in sample R-squares relative to the difference in the two equations' *df*:**

$$\mathbf{F}_{(\mathbf{K}_2-\mathbf{K}_1),(\mathbf{N}-\mathbf{K}_2-1)} = \frac{(\mathbf{R}_2^2 - \mathbf{R}_1^2) / (\mathbf{K}_2 - \mathbf{K}_1)}{(1-\mathbf{R}_2^2)/(\mathbf{N}-\mathbf{K}_2-1)}$$

# Test Difference in $\rho^2$ of 2 Equations

**A regression equation predicting test scores has $R^2$ = 0.14 with five predictors; after adding two predictors to that equation, the $R^2$ increases to 0.17. Sample N = 300.**

$$H_0:\ \rho_2^2 - \rho_1^2 = 0$$

$$H_1:\ \rho_2^2 - \rho_1^2 > 0$$

$$\mathbf{F}_{(7-5),(300-7-1)} = \frac{(\mathbf{R}_2^2 - \mathbf{R}_1^2)/(\mathbf{K}_2 - \mathbf{K}_1)}{(1-\mathbf{R}_2^2)/(\mathbf{N} - \mathbf{K}_2 - 1)} =$$

| $\alpha$ | $df_R$, $df_E$ | c.v. |
|---|---|---|
| .05 | 2, ∞ | 3.00 |
| .01 | 2, ∞ | 4.61 |
| .001 | 2, ∞ | 6.91 |

**Decision: ____________ Prob. Type I error: ____________**

**Conclusion: ______________________________________________**

# Adjust Multiple $R^2$ for *df*

**For multiple regression, adjust $R^2$ by a degree of freedom for each of the *K* predictors:**

$$R^2_{adj} = R^2 - \left( \frac{(K)(1 - R^2)}{(N - K - 1)} \right)$$

**Assuming N = 500, what are the adjusted $R^2$?**

| Eq. | $R^2$ | K | Adj. $R^2$ |
|---|---|---|---|
| A: | 0.083 | 3 | |
| B: | 0.145 | 5 | |
| C: | 0.372 | 11 | |

# Standardize Multiple Regression b's

**To compare effects of different predictors on a dependent variable, change each b into $\beta^*$ by multiplying b by the ratio of two standard deviations:**

$$\beta^*_{YX_i} = b_{YX_i}\left(\frac{s_{X_i}}{s_Y}\right) \Rightarrow \hat{Z}_Y = \beta^*_1 Z_1 + \beta^*_2 Z_2 + \ldots + \beta^*_K Z_K$$

| Variable | Std. dev. |
|---|---|
| $Y$ | 20 |
| $X_1$ | 8 |
| $X_2$ | 10 |
| $X_3$ | 4 |

$$\hat{Y} = 13.75 + 0.76X_1 - 1.24X_2 + 2.83X_3$$

$$\Downarrow$$

$$\hat{Z}_Y = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Interpretation:.____________________

__________________________________

__________________________________

# Categories into Dummies Variables

**Change REGION into 4 dummy (0-1) predictor variables:**

| REGION | NORTH | SOUTH | EAST | WEST |
|---|---|---|---|---|
| 1. North | | | | |
| 2. South | | | | |
| 3. East | | | | |
| 4. West | | | | |

**Use 3 of 4 region dummies in equation & predict scores:**

$$\hat{\mathbf{Y}}_{\mathbf{i}} = 43 + 58.5\,\mathbf{D}_{\mathbf{NORTH}} + 22.2\,\mathbf{D}_{\mathbf{SOUTH}} - 6.8\,\mathbf{D}_{\mathbf{EAST}} \qquad \mathbf{R}^2_{\mathbf{adj}} = 0.225$$

$$(8.3)\quad (7.6) \qquad\qquad (4.2) \qquad\qquad (2.4)$$

**North:** $\hat{Y}_i = 43 + 58.5(\_\_) + 22.2(\_\_) - 6.8(\_\_) = \_\_\_\_$

**South:** $\hat{Y}_i = 43 + 58.5(\_\_) + 22.2(\_\_) - 6.8(\_\_) = \_\_\_\_$

**East:** $\hat{Y}_i = 43 + 58.5(\_\_) + 22.2(\_\_) - 6.8(\_\_) = \_\_\_\_$

**West:** $\hat{Y}_i = 43 + 58.5(\_\_) + 22.2(\_\_) - 6.8(\_\_) = \_\_\_\_$

# ANCOVA

**ANCOVA includes both continuous & dummy variables in a multiple regression equation**

**Y = statistics test score**

**X = hours spent studying**

**D = sociology major (1) or nonmajor (0)**

$$\hat{Y}_i = 77.2 + 3.5X_i - 6.4D_{MAJORi} \qquad R^2_{adj} = 0.156$$

$$(12.3)\ (1.66)\qquad (3.05)$$

**Calculate the expected test scores for**

**Sociology major studying 6 hours:**

$$\hat{Y}_i = 77.2 + 3.5(\_\_) - 6.4(\_\_) = \_\_\_\_\_\_$$

**Nonsoc major studying 4 hours:**

$$\hat{Y}_i = 77.2 + 3.5(\_\_) - 6.4(\_\_) = \_\_\_\_\_\_$$