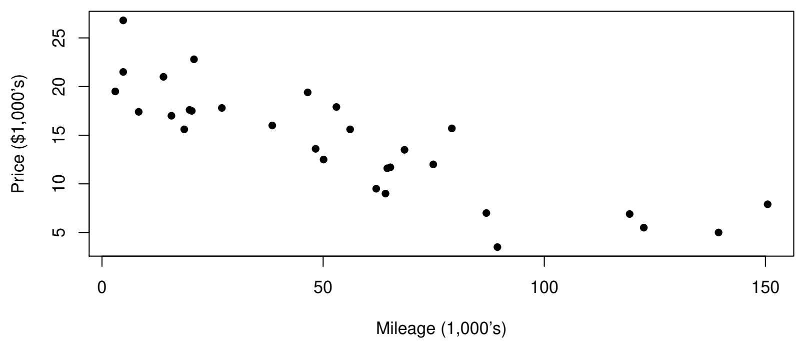
STA 111 Practice Problems

1 Car Prices

The scatterplot and summary data below show the price and mileage of a sample of 30 used Honda Accords. Summary data is $SS_{xx} = 47904.5$, $SS_{yy} = 954.1537$, $SS_{xy} = -5739.531$, $\bar{x} = 54.52667$, and $\bar{y} = 14.27667$.



1.1. What is the correlation?

1.2. What percent of the variation in price is explained by mileage?

1.3. What are the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1?

1.4. What is the equation of the estimated regression line?

1.5. What is the expected price of a used Accord with 50,000 miles on it?

1.6. Would it be appropriate to use your least squares line to make a prediction for the price of an Accord with 200,000 miles on it? Why or why not?

1.7. Construct a 95% confidence lower bound for the expected price of a Honda Accord with 50,000 miles on it.

1.8. In what context would your answer from 1.7 be useful?

1.9. Construct a 95% lower bound for the price of a *particular* Honda Accord with 50,000 miles on it.

1.10. In what context would your answer from 1.9 be useful?

1.11. Comment on the difference in your answers between 1.7 and 1.9.

2 Brobdingnag

The Child Health and Development Studies investigated 1,236 pregnancies between 1960 and 1967 among women on the Kaiser Foundation Health Plan in San Francisco.[1] Researchers were interested in modeling the weight of the infants in ounces (bwt) using the pregnancy time in days (gestation), whether the child was first born (parity, 0 means first pregnancy), the mother's height in inches (height), and whether the mother was a smoker (smoke, 1 means smoker). Summary regression output is provided below.

[1] Data from *OpenIntro Statistics (4th edition) by Diez, Barr, and Cetinkaya-Rundel*

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-85.420	13.680	-6.244	0.000
gestation	0.443	0.029	15.413	0.000
parity	-3.460	1.045	-3.311	0.001
height	1.335	0.181	7.370	0.000
smoke	-8.469	0.938	-9.031	0.000

2.1. Write out the equation of the estimated regression line.

2.2. Interpret the slope of gestation in the context of the problem.

2.3. Interpret the slope of smoke in the context of the problem.

2.4. Data from one of the births is shown in the table below. What is the residual?

bwt	gestation	parity	height	smoke
128	279	0	64	1

2.5. Construct a 95% confidence interval for β_{height} and provide a concise, careful interpretation in the context of the problem.

2.6. Martin Van Buren Bates (born 1837) "The Kentucky Giant" was a $\approx$ 7 feet 9 inch tall schoolteacher who enlisted as a private in the Confederate Army and was quickly promoted to Captain, presumably due to his sense of humor. In 1871 he married Anna Swan, a Canadian who was $\approx$ 7 feet 11 inches tall, giving them the world record for tallest married couple. They had two children, but sadly both children died shortly after being born. Their second child holds the record for heaviest birth at $\approx$ 352 ounces (22 pounds). Suppose Anna did not smoke and her pregnancy was 290 days. What is your prediction of for the baby's birth weight?

2.7. What are limitations of your prediction in 2.6?

3 Intensive Care Unit

The ICU data contains information for a sample of 200 patients from a hospital's Intensive Care Unit (ICU), the unit of the hospital that treats life-threatening cases. Researchers are interested in predicting patient Survival (0 = patient died, 1 = patient lived), based on their age (years), systolic blood pressure (mm of Hg) and admission status (0 = elective admission, 1 = emergency admission).[2] A multiple logistic regression model is fit.

	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	3.679	1.307	2.816	0.0048
Age	-0.034	0.011	-3.124	0.002
Systolic Blood Pressure	0.013	0.006	2.210	0.027
Emergency	-2.288	0.758	-3.017	0.003

3.1. Write out the estimated logistic regression model.

3.2. Interpret $\hat{\beta}_{\text{Age}}$ in the context of the problem.

3.3. Interpret $\hat{\beta}_{\text{Emergency}}$ in the context of the problem.

3.4. Is Age a significant predictor? In other words, should we reject $H_o : \beta_1 = 0$? Write a brief paragraph with your answer in the context of the problem, supported by appropriate evidence (test statistic, p-value, etc).

3.5. An 87 year old patient is admitted in an emergency with a systolic blood pressure of 80 mm of Hg. What is the predicted probability of survival?

3.6. The patient described in 3.5 did not survive. Is this result surprising?

3.7. A 59 year old patient is admitted in an emergency with a systolic blood pressure of 48 mm of Hg. What is the predicted probability of survival.

3.8. The patient described in 3.7 survived. Is this result surprising?

3.9. A patient is admitted in an emergency situation with a systolic blood pressure of 80. How young / old must the patient be in order to have a probability of survival of at least 0.50 based on the logistic regression model?

[2] Data from *Stat2: Models for a World of Data by Cannon et al*

4 D.K.'s Jungle Parkway

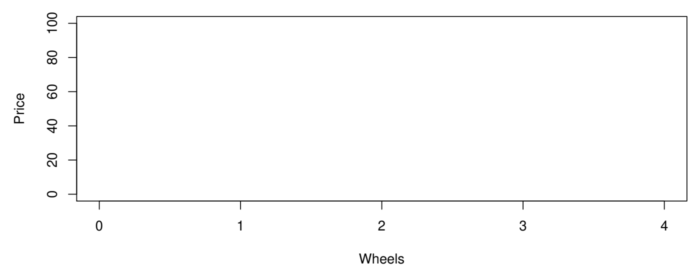
You are interested in predicting the price of Wii MarioKart games on eBay. You take a sample of $n = 141$ auctions for Wii MarioKart and fit a multiple regression model with the sales prices (dollars) as the response and predictors condition (1 = used, 0 = new) and wheels (the number of controllers/wheels included with the game). Summary regression output is provided below.[3]

	Estimate	Standard Error	t value	$\mathbb{P}(>\|t\|)$
(Intercept)	42.370	1.065	39.780	< 0.001
conditionused	-5.585	0.925	-6.041	< 0.001
wheels	7.233	0.542	13.347	< 0.001

4.1. Write out the estimated regression equation.

4.2. Write out the expression (with numbers) for your prediction of the price of a used game with three wheels.

4.3. Draw a picture of the model using the labeled axes below.



4.4. Provide a one-sentence interpretation for $\hat{\beta}_{\text{condition}}$ in the context of the problem.

5 Cherry Trees

A silviculturist is interested in finding the volume of black cherry trees in order to determine their timber yield. This is difficult to measure, so the researcher uses multiple regression to predict volume (cubic feet) based on height (feet) and diameter (inches) for a sample of 31 trees that have been cut down. Both height and diameter are easy to measure so the idea is one can have a prediction for volume based on diameter and height. The multiple regression model has $R^2 = 0.948$.[4]

	Estimate	Standard Error	t value	$\mathbb{P}(>\|t\|)$
(Intercept)	-57.99	8.638	-6.713	< 0.001
Diameter	4.71	0.264	17.816	< 0.001
Height	0.34	0.130	2.607	0.0145

5.1. Write out the fitted regression equation.

5.2. Find a 95% confidence interval for β_{diameter}. *Hint: You will need to use your t table here.*

5.3. Interpret your interval in 5.2 in the context of the problem.

5.4. Provide an interpretation of R^2 in the context of the problem.

5.5. A particular tree has a height of 74 feet and a diameter of 14.5 inches. What is your prediction for this tree's volume?

5.6. If the true volume of the tree is 36.3 cubic feet, how much did the model overestimate or underestimate the volume of the tree?

5.7. State the assumptions of multiple linear regression.

[3] Data from *OpenIntro Statistics (4th edition) by Diez, Barr, and Cetinkaya-Rundel*

[4] Data from *OpenIntro Statistics (4th edition) by Diez, Barr, and Cetinkaya-Rundel*